Introduction to Data Science

Wray Buntine

Introduction to Data Science

Wray Buntine

Version: Pre-release 0.3

Generated by Alexandria (https://www.alexandriarepository.org) on July 24, 2016 at 8:28 pm AEST

Contents

1
Data Science and Data in Society

Introduction

This is our first module of six for the Introduction to Data Science unit. In this module we look at data science from a top level perspective: what is Data Science, what does a data scientist do (roughly), how is it presented to the business world and the public, and what impact is it having.

(https://youtu.be/j0Lx75BxDg8)

Aims of this module

Our aims for this module are as follows:

- Evaluate different definitions of data science.
- Discuss and place in context the emergence of data science and its relationship to other fields.
- Interpret the lifecycle of a data science project.
- Classify participants in a data science project.
- Self-reflect on own capabilities and goals as a data scientist.
- Identify the major impacts of data science on society.

How to study for this module

In this module we draw on a lot of material in the public domain, in general magazines and journals, and papers given at conferences. So there are a lot of videos, online magazine entries and blogs. We also have a few videos we took of industry professionals.

Data Science is a new field so it is hard to find critical analysis, or top down pedagogical presentation of the material. Therefore, we ask that you review the material with a critical eye, and we will use the class

discussion forums to air our views about these aspects of the field.

Please note:

- Reference items marked with a single "johny look it up" icon,  , should be viewed as *suggested reading*, not essential nor important for assessment.
- Reference items marked with a two "johny look it up" icons,  ,  should be viewed as *important reading*, considered important for assessment.

In this module you should also start your first assessment "Data Science and Me."

1.1
Overview of Data Science

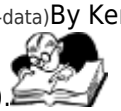
We will start with a look at the popular business content related to data science. This lacks the technical depth but presents the motivations, issues and broader implications well. This section introduces Data Science by means of a number of videos, an audio track and some magazine articles. Nowadays if you went looking for videos or general articles on Data Science you would easily find thousands. Which should you watch or read? Time is precious! We recommend the following.

"Big Data is Better Data" by Kenneth Cukier

Kenneth Cukier is the "Data Editor" for *The Economist*, and has made a name for himself with books and videos on big data. His website (http://www.cukier.com/) has more of the material. "Big Data is Better Value" is a talk in the popular TED talk series.

- "Big Data is Better Data" (http://www.ted.com/talks/kenneth_cukier_big_data_is_better_data) by Kenneth Cukier (streamed video, 16 mins, transcript 2300 words) 

Here is a related magazine article by Cukier, which really serves as a primer/taster for his book in the popular press, "Big Data" (NYT review) (http://www.nytimes.com/2013/06/11/books/big-data-by-viktor-mayer-schonberger-and-kenneth-cukier.html?pagewanted=all&_r=1). This article is published in *Foreign Affairs* and extends the content from the TED talk (so is optional to review).

- "The Rise of Big Data" (https://www.foreignaffairs.com/articles/2013-04-03/rise-big-data)By Kenneth Neil Cukier and Viktor Mayer-Schoenberger (magazine article, 5000 words, 30 mins). 

You can also (optionally) see a 45 minute talk he presented about his book ("Big Data: A Revolution That Will Transform How We Live, Work, and Think" (http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544227751)) at a Google seminar, arguably the company responsible for making the world aware of Big Data and Data Science (it is one technology plank for their business model). The material for the two (article and video) is mildly overlapping.

(https://www.youtube.com/watch?v=bYS_4CWu3y8)

What is Data Science," by Mike Loukides

This is a short pamphlet produced by O'Reilly, the publisher of many popular tech books that any geek would recognise. See for instance their "Data Science Starter Kit" collection (http://shop.oreilly.com/category/get/data-science-kit.do). O'Reilly have a large presence in Data Science, and run regular conferences such as Strata+Hadoop World, (http://strataconf.com/) which have tutorials, visionary talks, business pitches and even some academic talks.

- "What is Data Science?" (O'Reilly) (http://cdn.oreilly.com/radar/2010/06/What_is_Data_Science.pdf) (free PDF, approx 4300 words, 25 mins) 

There are many more such pamphlets and popular books along these lines,

- "*Field Guide to Data Science*", Booz Allen Hamilton Inc. (http://www.boozallen.com/insights/2013/11/data-science-field-guide) (the free PDF or e-Book is linked at the bottom of the page).
- "Big Data at Work" is a book by Thomas H. Davenport, 2014

but for our purposes the O'Reilly one does a good introduction.

The CERN view of Big Data

This is an introductory lesson on big data from TED-Ed looking at the CERN (http://home.web.cern.ch/) angle on big data. CERN is a French acronym for the European Organization for Nuclear Research. They were instrumental in starting up the internet and have been influential in the rise of big data. This "lesson" could well be targeted at high school students, but it does a good summary of CERN's role.

- "Big Data" - Tim Smith (http://ed.ted.com/lessons/exploration-on-the-big-data-frontier-tim-smith) (TED-ED video, 6

mins)

"Data Science and its Far-reaching Uses and Implications" by Phil Brierley

This is a radio broadcast which includes an interview of Phil Brierley. He is interviewed on Data Science from minutes 14:30 to 32:15 of the podcast. Phil is a successful data science consultant and gives a low key, hype-free introduction.

- http://rrrfm.libsyn.com/byte-into-it-15-april-2015 (web page for podcast, see 14:30-32:15) 

1.2
Melbourne Datathon 2016

In April 2016 the Melbourne Data Science Meetup (http://www.meetup.com/Data-Science-Melbourne/?chapter_analytics_code=UA-49664189-1) organisation ran a datathon. We will use this to illustrate:

- a business context (Seek.com's business);
- a data set;
- candidate Data Science projects;
- discussion of different techniques to predict the business category of a job advert;
- exploratory data analysis.

The datathon itself is recorded as follows:

- the Datathon website (http://www.datasciencemelbourne.com/datathon2016/) itself
- Kaggle competition website (https://inclass.kaggle.com/c/melbourne-datathon-2016) for the datathon
- file repository of various presentations and background info. (http://www.meetup.com/Data-Science-Melbourne/files/?chapter_analytics_code=UA-49664189-1) on the meetup website

Business Context and Data

Seek.com manages a job search system. Potential employers place their job descriptions on the Seek.com website, and candidate employees search the site to look for jobs. Jobs are stored in a database and are categorised into broad classes and subclasses (e.g., "Healthcare & Medical Jobs' and "Dental"). The database has fields such as location, salary, description, etc., so that searchers can do faceted search (e.g., to focus on location, salary ranges, key words, etc.). The users keywords and click history are recorded to help suggest relevant jobs to them. Seek provides a conceptual schema for the database with tables for jobs, job searches, job impressions (results), and click history.

- Seek reports details here "Datathon 2016, Marketplace Analytics" Show "Seek-Datathon-Presentation-v5.pdf" (https://docs.google.com/viewer?url=https%3A%2F%2Fwww.alexandriarepository.org%2Fwp-content%2Fuploads%2F20160724151640%2FSeek-Datathon-Presentation-v5.pdf&hl=en_GB&embedded=true)
  Download (PDF, 730KB) (https://www.alexandriarepository.org/wp-content/uploads/20160724151640/Seek-Datathon-Presentation-v5.pdf)

  (800k 16 slide PDF file)

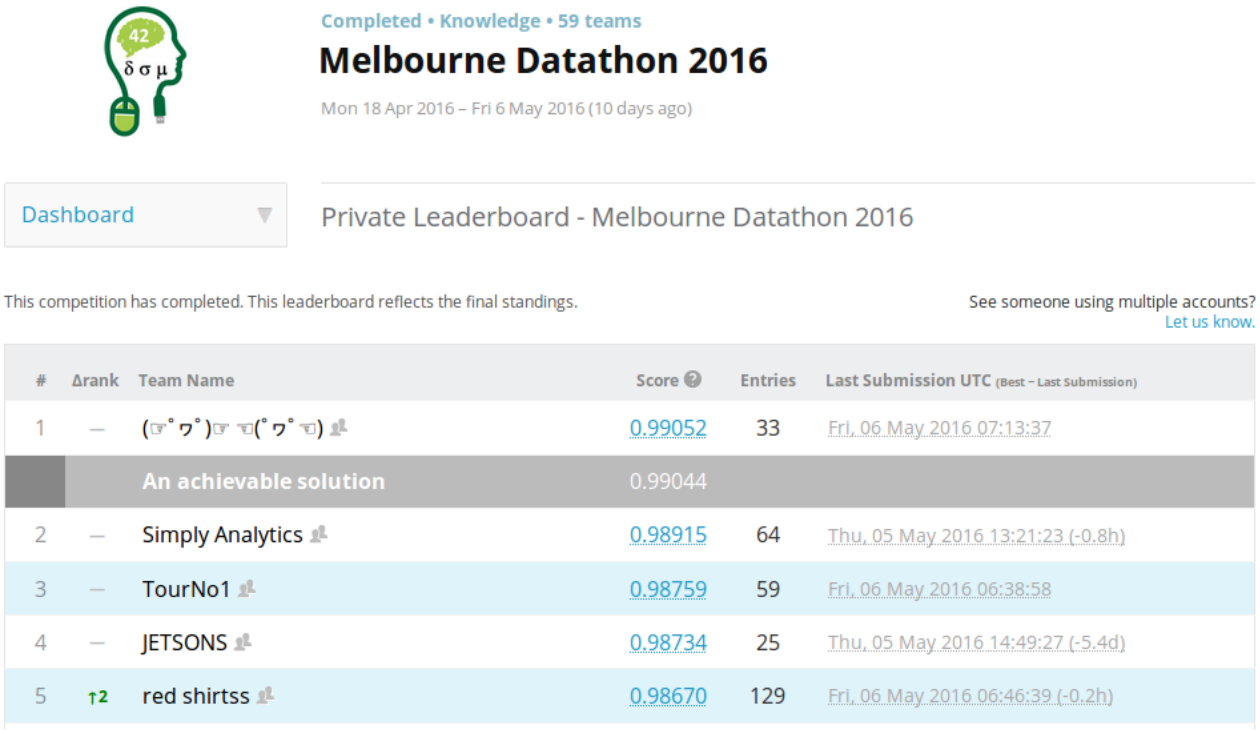
The live system has approximately 100,000 job adverts and 2 million candidates have searched.

Seek would like to consider a number of tasks:

- How might they better suggest jobs to individuals? *i.e.*, what constitutes a good match?
- How can they better classify jobs? Note, jobs are self classified by the employers, but mistakes are made and jobs in the wrong class won't get responses, so no commission is earned by Seek. This is discussed below under **Predictive Analysis**.
- Generally, what interesting knowledge/information exists in the data that Seek could use to improve customer (prospective employer and employee) results and experience? This is discussed below under **Exploratory Analysis**.

Predictive Analysis

Datathon participants were invited to study the data and predict if a job is in the 'Hotel and Tourism' category. For this, a separate test set was held out and a competition run on the Kaggle.com (https://inclass.kaggle.com/) in-class platform. At the end of the competition the leaderboard looked like so:



Completed • Knowledge • 59 teams

Melbourne Datathon 2016

Mon 18 Apr 2016 – Fri 6 May 2016 (10 days ago)

Dashboard ▼ Private Leaderboard - Melbourne Datathon 2016

This competition has completed. This leaderboard reflects the final standings.

See someone using multiple accounts? Let us know.

#	Δrank	Team Name	Score	Entries	Last Submission UTC (Best – Last Submission)
1	–	(☞ﾟヮﾟ)☞ ☜(ﾟヮﾟ☜)	0.99052	33	Fri, 06 May 2016 07:13:37
		An achievable solution	0.99044		
2	–	Simply Analytics	0.98915	64	Thu, 05 May 2016 13:21:23 (-0.8h)
3	–	TourNo1	0.98759	59	Fri, 06 May 2016 06:38:58
4	–	JETSONS	0.98734	25	Thu, 05 May 2016 14:49:27 (-5.4d)
5	↑2	red shirtss	0.98670	129	Fri, 06 May 2016 06:46:39 (-0.2h)

PrivateLeaderboard - Melbourne Datathon 2016

We will just consider the top three entries:

3rd place team: Simply Analytics: this entry was an initial leader. One of the author's, Grant McKinnon has a blog entry describing their system (http://www.grant-mckinnon.com/?p=17), where he says the following:

> Speaking to others following this competition and hearing their advanced methods and intelligent feature creation and seeing their mouths drop after hearing that my final solution was nothing more than a bag of words on the Title and Abstract fields and a Weighted Average of a Logistic Regression and Naive Bayes makes me want to emphasise one important point in Predictive Modelling.
>
> You should never dismiss the simple ideas that have worked for years.

Though it should be noted they used sophisticated evaluation methods (such as "cross validation") to estimate the quality of their solution, and probably selected their final entry from a range of different methods. As described on the blog, they used Amazon Web Services (a cloud system) for the computation, and the system is implemented with Python scripts to pre-process the data and then run a standard algorithm ("linear regression") as a subroutine wrapped with an optimizer. A key parameter to the algorithm was selected using a sophisticated search and estimation method ("grid search" with "cross

validation"). If you are new to Data Science, all these terms should be mysterious jargon!

2nd place team: An achievable solution: this entry was submitted by competition organiser Phil Brierley who is a well-known Data Science expert with a large custom collection of tools. It was entered towards the end using his own state of the art system.

1st place team: (☞□□□)☞ ☜(□□□☜): this entry was the final leader. They knew from previous experience that Phil Brierley would release a high performing system right at the end so they withheld their best solution till the end, so as not to let on that they had a special trick up their sleave. They presented their trick at the awards night. They collected two different kinds of auxiliary data:

- **target categories for searchers**: they analysed searchers' typical target jobs; if a searcher who usually clicks on "accounting" jobs responds to a particular job advert, it is likely the job advert is not "hospitality".
- **business classifications**: they searched for a business name in the job advert and then looked it up in the ABS database of businesses, which has business classifications attached.

With these auxiliary data, they could tag job adverts with the extra information. This clearly helped their predictive algorithms. One of the underlying themes of this unit is that uncovering new and useful sources of data is a key strategy for success.

Exploratory Analysis

Datathon participants were invited to study the general data and develop a presentation showing useful information. Some can be viewed from the "Top 5 pitches Melbourne Datathon 2016" stored on the repository website.

We can consider two different presentations. The first is a Monash student effort:

- "Monash Bubble"
  Download (PDF, 22.35MB) (https://www.alexandriarepository.org/wp-content/uploads/20160724152504/Monash-Bubble.pdf)

  (23Mb, 23 pages of slides) a presentation by a Monash MDS team.

This shows an interesting and insightful collection of R (?) graphics exploring different facets of the data including:

- locations of jobs and searchers
- top queries by location
- mobile usage
- search time of day
- job classification by state
- supply and demand
- salary by classification

as well as a motivating

- study of Data Science jobs

The second is a professional effort (team was practicing data scientist and business intelligence expert, as well as a developer and a student):

- "Team 4Quarters" Show "4.-4Quarters-compressed-for-filesize.pdf"

(https://docs.google.com/viewer?url=https%3A%2F%2Fwww.alexandriarepository.org%2Fwp-content%2Fuploads%2F20160724152934%2F4.-4Quarters-compressed-for-filesize.pdf&hl=en_GB&embedded=true)
Download (PDF, 946KB)
(https://www.alexandriarepository.org/wp-content/uploads/20160724152934/4.-4Quarters-compressed-for-filesize.pdf)

(0.97Mb, 23 pages of slides) a presentation by a another team.

This is more focused on particular business questions such as:

- Which Jobs have more Clicks in relation to Salaries?
- How do Job Classes compare?
- What are the most popular Jobs for the Users?
- How many Jobs get through to Impressions?
- How many Clicks does each Job get?

1.3
The Data Science Process

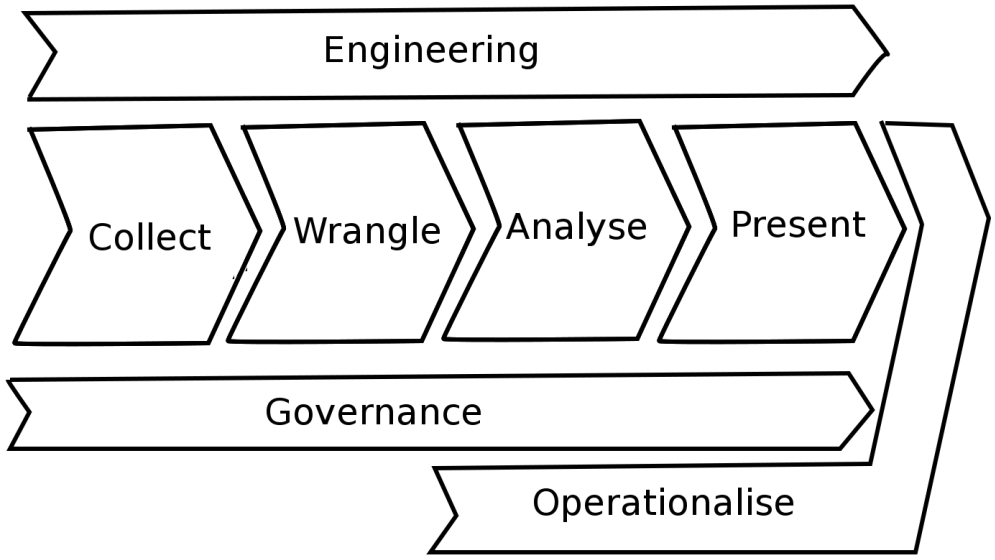
This short (8:20 min) talk introduces Data Science and the life cycle or process of a Data Science project. We present this before actually defining Data Science, so that one can understand the nature of it before considering the formalities.

(https://www.alexandriarepository.org/wp-content/uploads/20150629140309/DataScience_Process_Draft4.mp4)

The major steps of a Data Science project are as follows:

Collection: obtaining and collecting data from various sources, instruments or providers.
Engineering: processing and storing data, managing the databases, computers and hardware.
Governance: all aspects of data management such as security, metadata, etc.
Wrangling: transforming and cleaning data prior to analysis.
Analysis: analysis in its many forms.
Presentation: visualisation and summarisation, to present the case for "value".
Operationalisation: putting the results of analysis to work to obtain value.

Note that there are different versions of this scheme and different names for the parts, however we refer to this particular sequence of steps as the **Standard Value Chain**. This is illustrated in the figure below, which diagramatically also represents the timeline, left to right.



Standard Value Chain

1.4
What is Data Science

Definitions

This subsection assembles a number of different definitions on what data science is. *It is important not to get lost in the details or the arguments for one or the other. Just be aware there are some differences!*

Wikipedia's entry on Data Science (http://en.wikipedia.org/wiki/Data_science) has the following to say (I have left their embedded links in):

> In general terms, **Data Science** is the extraction of knowledge (http://en.wikipedia.org/wiki/Knowledge) from data (http://en.wikipedia.org/wiki/Data), which is a continuation of the field data mining (http://en.wikipedia.org/wiki/Data_mining) and predictive analytics, also known as knowledge discovery and data mining (KDD). It employs techniques and theories drawn from many fields within the broad areas of mathematics (http://en.wikipedia.org/wiki/Mathematics), statistics (http://en.wikipedia.org/wiki/Statistics), information theory (http://en.wikipedia.org/wiki/Information_theory) and information technology (http://en.wikipedia.org/wiki/Information_technology), including signal processing (http://en.wikipedia.org/wiki/Signal_processing), probability models (http://en.wikipedia.org/wiki/Probability_models), machine learning (http://en.wikipedia.org/wiki/Machine_learning), statistical learning (http://en.wikipedia.org/wiki/Statistical_learning_theory), data mining (http://en.wikipedia.org/wiki/Data_mining), database (http://en.wikipedia.org/wiki/Database), data engineering (http://en.wikipedia.org/wiki/Data_engineering), pattern recognition and learning (http://en.wikipedia.org/wiki/Pattern_recognition_and_learning), visualization (http://en.wikipedia.org/wiki/Data_visualization), predictive analytics (http://en.wikipedia.org/wiki/Predictive_analytics), uncertainty modeling (http://en.wikipedia.org/wiki/Uncertainty_modeling), data warehousing (http://en.wikipedia.org/wiki/Data_warehousing), data compression (http://en.wikipedia.org/wiki/Data_compression), computer programming (http://en.wikipedia.org/wiki/Computer_programming), and high performance computing (http://en.wikipedia.org/wiki/High_performance_computing). Methods that scale to Big Data (http://en.wikipedia.org/wiki/Big_Data) are of particular interest in data science, although the discipline is not generally considered to be restricted to such data. The development of machine learning (http://en.wikipedia.org/wiki/Machine_learning), a branch of artificial intelligence (http://en.wikipedia.org/wiki/Artificial_intelligence) used to uncover patterns in data from which predictive models can be developed, has enhanced the growth and importance of data science.

Note Wikipedia's entry on big data (https://en.wikipedia.org/wiki/Big_data) is far more extensive and provides better reference in some perspectives, for instance applications and industry.

Pivotal (http://pivotal.io), an IT innovation company, defines Data Science (via Michael Brand) as the following:

> The use of statistical and machine learning techniques on big multi-structured data in a distributed computing environment to identify correlations and causal relationships, classify and predict events, identify patterns and anomalies and infer probabilities, interest and sentiment.

Thus, they tie it much more specifically to big data and distributed computing, which is in agreement with several of the corporate players in the area.

The American National Institute of Standards and Technology (NIST) has a Big Data Working Group

(http://bigdatawg.nist.gov) that looks at the issues from the prospective of big data:

> **Big Data** consists of extensive datasets primarily in the characteristics of volume, variety, velocity, and/or variability that require a scalable architecture for efficient storage, manipulation, and analysis.
>
> The **Big Data paradigm** consists of the distribution of data systems across horizontally coupled, independent resources to achieve the scalability needed for the efficient processing of extensive datasets.
>
> **Big Data engineering** includes advanced techniques that harness independent resources for building scalable data systems when the characteristics of the datasets require new architectures for efficient storage, manipulation, and analysis.
>
> The **data science paradigm** is extraction of actionable knowledge directly from data through a process of discovery, hypothesis, and hypothesis testing.
>
> **Data science** is the empirical synthesis of actionable knowledge from raw data through the complete data lifecycle process.
>
> A **data scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes through each stage in the data lifecycle.

Interestingly, the NIST notion of Data Science seems closer to the field of Data Mining (http://en.wikipedia.org/wiki/Data_mining) that developed in the 1990s, whereas what they view that a data scientist does, and the Wikipedia definition has a broader view that includes data engineering and computational paradigms such as databases.

The Journal of Data Science (http://www.jds-online.com/about) takes a similar broad view again, saying:

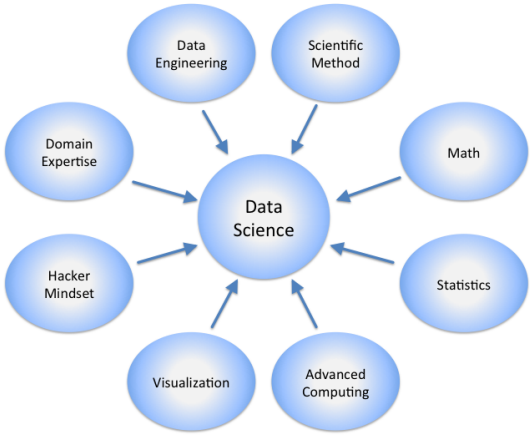
> By "Data Science", we mean almost everything that has something to do with data: Collecting, analyzing, modeling..... yet the most important part is its applications - all sorts of applications.

The DataScientists.Net website (http://www.datascientists.net/aboutds.htm) takes a more functional view but it is similar:

> The three components involved in data science are **organising, packaging** and **delivering** data (the OPD of data). Organising is where the physical location and structure of the data is planned and executed. Packaging is where the prototypes are build, the statistics is performed and the visualisation is created. Delivering is where the story gets told and the value is obtained. However what separates data science from all other existing roles is that they also need to have a continual awareness of What, How, Who and Why. A data scientist needs to know what will be the output of the data science process and have a clear vision of this output. A data scientist needs to have a clearly defined plan on how will this output be achieved within the restraints of available resources and time. A data scientist needs to deeply understand who the people are that will be involved in creating the output. And most of all the data scientist must know why there is a motivation behind attempting to manifest the creative visualisation.

While earlier definitions mentioned applications, this definition talks about the *delivery* of data instead.

This definition talks about organising, whereas data management is the more common term.



Depicts a mash-up of disciplines from which Data Science is derived. CC BY-SA Calvin Andrus, 13/07/2012.

There are also a lot of visual depictions of data science, and one theme is the various fields that contribute to Data Science, as shown in the mash-up.

For Big Data, the visual artists have gone crazy combining waterfalls, waves, highways and tunnels with digital themes. If you wish to browse, see Tumblr's BigDataPix.tumblr.com (http://bigdatapix.tumblr.com/) (this is obviously not study material).

Now there are, of course, hundreds of other blogs, images and videos and recent books all proclaiming, dissecting and variously defining Data Science. But of these various definitions, we use the general view:

> By Data Science, we mean almost everything that has something to do with data and its use in extracting value.

Now some have also called this Dataology (http://www.paper.edu.cn/en_releasepaper/content/4432156), but we prefer the term Data Science. As we study what Data Scientists do in this unit, we will extend verbs used to "do with data" to include not only collection and analyzing, but also visualisation, management, and the general delivery of data products too.

Fields contributing to data science

Arguably, there are three broad areas one needs to study as a Data Scientist, though these are not mutually exclusive nor exhaustive, so they overlap somewhat and do not cover the full space. Moreover, they are fields with a long history, whose recent success has been instrumental in allowing Data Science to develop.

Data Engineering includes techniques that harness resources for building scalable data systems when the characteristics of the datasets require architectures for collection, storage, manipulation, and analysis.

Data Analysis includes techniques that pre-process, statistically analyse, and then present results of analysis, for instance in visualisations. There are many variations of these field and specialisations of it exist in many application domains.

Data Management includes techniques to manage data through its lifecycle and in accordance with regulations and best practices to address issues such as privacy, security, and appropriate access. Often the terms governance and curation are used too.

These do not include business models, as indeed business models and applications cross all areas. Data Science is usually most closely associated with Data Analysis, and indeed it is in this area where the real value is created from data. However, the other areas are equally important for the successful functioning

of a Data Science project.

So what is Data Science?

We will use the NIST definition, assembled as follows:

> A **data scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes through each stage in the data lifecycle. Moreover, **Data science** is the empirical synthesis of actionable knowledge from raw data through the complete data lifecycle process.

Of the relevant fields contributing to Data Science, a data scientist needs to be familiar with data engineering and data management for their work, but their core competence is the use of data analysis, the theory of which is covered by the field of Statistics. However, they are more than just a data analyst as they are involved in the broader data lifecycle process.

History of data science

Three carefully presented time-lines relevant to Data Science appear in the business literature. The first was published by Gil Press in the *Forbes* (business) magazine. The second was produced by Wolfram Alpha covering a broader area relevant to their system. The third is IBM's perspective on big data, the IBM inspired timeline. IBM is one of the few computing companies with a long enough history to provide this sort of perspective.

"A Very Short History of Data Science," by Gil Press

This is a review of major events in the science, technology and business community over the last few decades that led to the emergence of Data Science. There are many of these timelines, as indeed everyone hopes to get themselves recorded as playing a role. In truth, thousands of us scientists have been doing data science for many decades, but now, at long last, the ideas have reached the shakers of movers of industry and government.

- "A Very Short History Of Data Science" (http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/) by Gil Press (magazine article, 3800 words, 20 mins) 

You may not be able to understand this in full detail. Some of the developments chronicled you may not know about, and some are emergent areas discovered independently by many smart folks, but one lucky scientist gets the credit in the popular press! But skim over it and get an idea of events.

"Timeline of Systematic Data and the Development of Computable Knowledge" on Wolfram Alpha

This is a broader timeline on the theme of data and knowledge represented on computer, covering the

key ideas and some of the companies involved. You can go back in time to explore the contributions from pre-history of languages, libraries, maps and codices. This really does put things into perspective. The first half of the 1900's was really about establishing codes, whereas the 60's onwards saw computerisation taking hold. Older digital aficionados should recognise everything post 1959. From here, your really understand how data sets and meta data we take for granted were established.

- "Timeline of Systematic Data and the Development of Computable Knowledge" (http://www.wolframalpha.com/docs/timeline/computable-knowledge-history-6.html) on Wolfram Alpha (infographic)


Now, pre-1959 is mainly for interest. Most of the subsequent entries are critical for some aspects of Data Science. Some of the key technologies here were demonstrated at Engelbart's Mother of All Demos (https://en.wikipedia.org/wiki/The_Mother_of_All_Demos) in 1968.

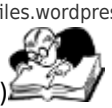
Big Data according to IBM

IBM has played an important role in all things computing since the early 1900s. Their own history of the field is presented in an infographic:

- "The Evolution of Big Data" (http://cloudtweaks.com/2013/11/cloud-infographic-evolution-of-big-data/) from IBM (infographic)


Berkeley Data Science view

Jeff Hammerbacher in his introduction to a Berkeley Data Science course gives another timeline in his lecture slides

- "Introduction to Data Science" (https://berkeleydatascience.files.wordpress.com/2011/01/20110118berkeley.pdf) by Jeff Hammerbacher, Mike Franklin (PDF, see pages 11-30)


A functional development

We have already argued that Data Science, broadly speaking, is related to three communities, analysis, engineering and management, although these are neither mutually exclusive nor exhaustive. In earlier times, perhaps prior to 2000, these communities developed somewhat independently, though there were of course some interactions.

Data Engineering: this community started out serving business with big iron and data bases, with transaction processing to keep data consistent, and with business intelligence layered on top to report on what was happening in an organisation. With the development of relational databases (http://en.wikipedia.org/wiki/Relational_database) and standards such as SQL, companies such as Oracle flourished. Business intelligence (http://en.wikipedia.org/wiki/Business_intelligence) (BI) was initially descriptive statistics to give historical accounting, and the field gradually expanded to include more analytic methods. Business Analytics (http://en.wikipedia.org/wiki/Business_analytics) is the full analytic version of BI that is usually considered allied to Statistics but within a business school, and is part of the broader Data Analysis field.

Data Analysis: this community started out with Statistics (http://en.wikipedia.org/wiki/Statistics), sometimes developed by physicists. Traditional statistics developed techniques for confirmatory analysis, such as hypothesis testing, which is recognised as a core technique for the scientific method. With the advent of computers, the community gradually broadened to include areas such as machine learning and its applications in robotics, natural language and image processing, pattern recognition, business and economic statistics, bioinformatics, and many other application and methodology areas. Moreover, visualisation was recognised early on as critical for both presentation and exploratory analysis, and with the rise of the Internet methods for visualisation exploded as the audience expanded. Data mining (http://en.wikipedia.org/wiki/Data_mining), which emerged in the 1990's, was the first major attempt to combine with the data engineering community. Eventually, computer savvy statisticians and analysts developed platforms for their craft including SAS, R, Weka and so forth, and visionary data analysts proclaimed the growing importance of their methods.

Data Management: this community initially grew out of the need for curation and management of assets of various kinds, whether they be books or historical artifacts. With the advent of the digital world, this extended to digital assets including data, and extended to become the field of information science (http://en.wikipedia.org/wiki/Information_science). The field started to include data analysis as the need for understanding large collections became apparent, for instance, methods for automated classification of content. Businesses became well aware for the need to perform management in order to deal with privacy, security, regulatory demands and disaster management.

There are, perhaps, several related factors that emerged as drivers of Data Science from 2000 onwards. Some of these trends are documented in the classic McKinsey Global Institute report "Big data: The next frontier for innovation, competition, and productivity" (http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation), which takes 156 pages to document the business case.

Business pressure on decision making: businesses function in a continuously changing environment.

- The more data an organization has, the more onerous the task of keeping order to the data.
- Businesses saw a need to convert to data driven decision making in line with earlier advances such as just-in-time manufacturing. Operations became too big for the intuitive leader with the right hunches to operate without analytic guidance.
- Daily operation of businesses require them to reach insights faster and act on them in real time.

Managing the expanse of corporate data to achieve the goal of being data driven was then recognised as a key problem. Analytics was the means by which predictions and prescriptions could be made.

Moore's Law changed the landscape: according to Moore's Law (http://en.wikipedia.org/wiki/Moore%27s_law), which held up until recently, computing power and storage capacity is growing exponentially (so doubling in constant time). Moreover, according to Bell's Law (http://en.wikipedia.org/wiki/Bell%27s_law_of_computer_classes), different computing platforms and ecosystems will continue to emerge, the laptop, the mobile phone, cloud computing *etc.* These offer up new kinds of data, and new kinds of data processing. So sometime in the period 2000-2010 companies could affordably store all their data digitally and make it readily accessible to processing, while concurrently new. complementary forms of data became available. Some of the numbers here and the types of data across businesses are given on Exhibits 5-8 in the McKinsey report. Concurrently, consumers started creating their own data in social networks, and the internet of things started to emerge with the convergence of cheap computing, networking and the right standards. This is documented on Exhibits 9-10 in the McKinsey report.

Data itself is an asset: some pathfinding organisations recognised early on that data is an asset and started building up data resources. Examples include Bloomberg L.P. (http://en.wikipedia.org/wiki/Bloomberg_L.P.), which provided a terminal (http://en.wikipedia.org/wiki/Bloomberg_Terminal) for the finance world to view data from

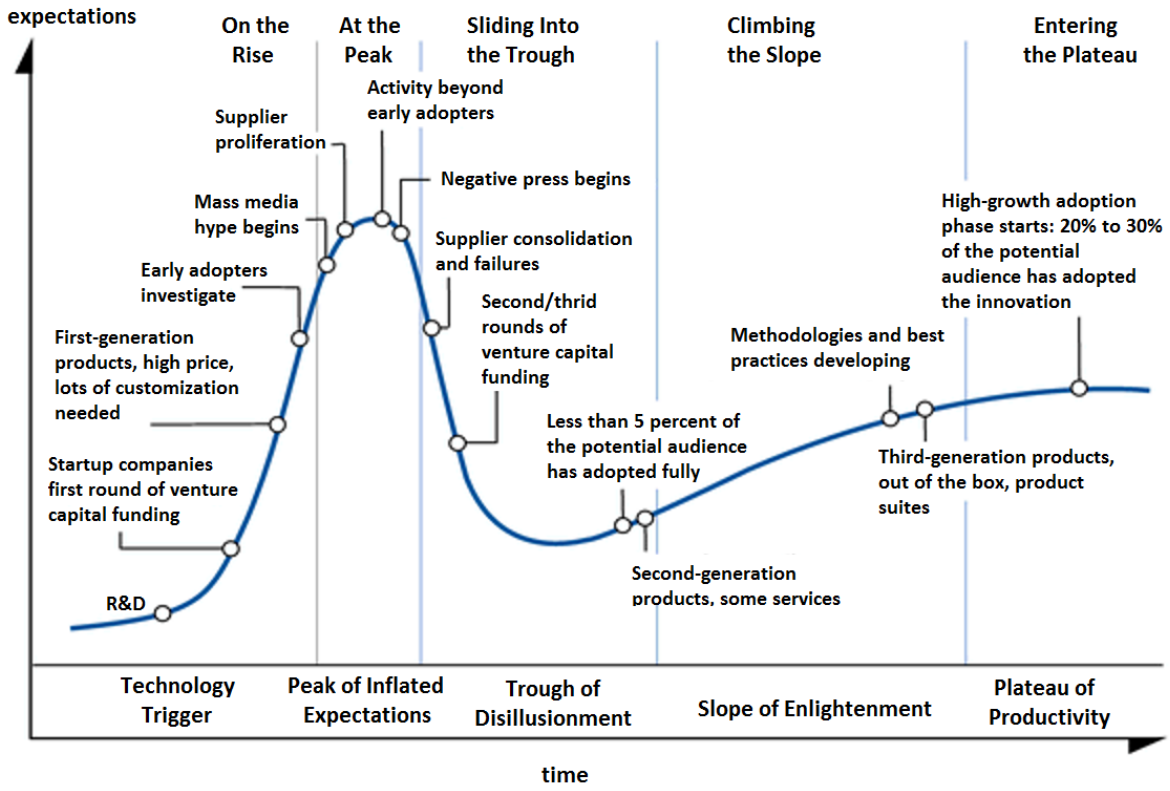
Wall Street and other centres, and LexisNexis (http://en.wikipedia.org/wiki/LexisNexis) and Reuters (http://en.wikipedia.org/wiki/Reuters). Data analysts soon discovered that a major hurdle to building a business using their craft was corporate perspective in the business world. Corporations understand the value of their data, so preferred to use in house analysts and prevent external players. As Tomasz Tunguz says (http://tomtunguz.com/data-isnt-a-business-model-its-much-more-fundamental/),

> "Data is the most valuable outcome of building a successful product. It's the insight, the secret, the keys to the kingdom. Don't sell the keys to the kingdom. Data provides economies of scale and insights used to develop huge barriers to entry and it should be kept within an organization. Internal data use is the path to building a huge business."

Dot-Com companies grew: some survivors of the Dot-com bubble (http://en.wikipedia.org/wiki/Dot-com_bubble) and subsequent Internet companies went on to become huge businesses, Amazon, Google, Twitter, E-Bay and so forth. Perhaps their distinguishing feature was that they became companies with a business model centered on their data. They extracted business advantage using their data, and also pioneered some of the techniques, such as map-reduce, that went on to become staples of the Data Science community. The Dot-Com giants demonstrated to the rest of the business world that the age of the data-centric company had arrived. CEOs and the business community now recognised data as a valid asset to be valued, managed and exploited.

The hype cycle

Those of us living in Silicon Valley in the late 1990s lived through the Dot-com bubble (http://en.wikipedia.org/wiki/Dot-com_bubble). We saw the huge related booms in commercial real-estate, computing hardware and services, and so forth. Those stuck holding stock when the bubble burst lost out. Economic bubbles (http://en.wikipedia.org/wiki/Economic_bubble) come and go, and have counterparts in the technology world and in the academic world, as there are some fundamental laws of psychology at play. Gartner has documented the so-called hype cycle of technologies to help us understand this phenomena and as of November 2014 saw Data Science still rising in its "hype," but Big Data is already past its peak. The hype cycle is very nicely documented on Wikipedia (https://en.wikipedia.org/wiki/Hype_cycle). They give the following figure to explain it:



Hype Cycle (general) By NeedCokeNowOlga Tarkovskiy. CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0), via Wikimedia Commons

- "Gartner's 2014 Hype Cycle for Emerging Technologies" (http://www.gartner.com/newsroom/id/2819918) (web page, 1000 words, 6 mins) 
- "Big Data Meets Trough Of Disillusionment: Gartner" (http://www.informationweek.com/big-data/software-platforms/big-data-meets-trough-of-disillusionment-gartner/d/d-id/898939), on Information Week 11/18/2013 (online magazine, 700 words, 4 mins) 

So while we see Data Science is currently near its peak in hype. As with all believers, we claim the field has established itself and will have a strong, lasting impact in both the academic and commercial arenas.

A completely different style of analysis, but related to the hype cycle, is understanding whether a new technology should be adopted, assessed, trialed, or whether one should "wait and see". This sort of analysis is done by the Technology Radar analysis of ThoughtWorks (https://www.thoughtworks.com/radar). For instance, you can browse their report for Report for May 2015 (https://assets.thoughtworks.com/assets/technology-radar-may-2015-en.pdf).

Government sponsored scientific disciplines

As with the emergence of any major field, significant government sponsored efforts have begun.

In the USA on March 29 2012 (http://bigdatawg.nist.gov/WhiteHouse_big_data_press_release.pdf),

> ... the Obama Administration today announced a "Big Data Research and Development Initiative." By improving our ability to extract knowledge and insights from large and complex collections of digital

> data, the initiative promises to help solve some the Nation's most pressing challenges.

This lead to the NIST Big Data Working Group that has release its initial report on the NIST Big Data Interoperability Framework on April 6 2015 (http://www.nist.gov/itl/bigdata/20150406_big_data_framework.cfm). The report comprises a suite of documents (http://bigdatawg.nist.gov/V1_output_docs.php) covering definitions, a reference architecture, use cases and a standards roadmap.
Related is a report released 2013 by the US National Academy of Sciences which is a large description of the research area. They say:

> Data mining of massive data sets is transforming the way we think about crisis response, marketing, entertainment, cybersecurity and national intelligence. Collections of documents, images, videos, and networks are being thought of not merely as bit strings to be stored, indexed, and retrieved, but as potential sources of discovery and knowledge, requiring sophisticated analysis techniques that go far beyond classical indexing and keyword counting, aiming to find relational and semantic interpretations of the phenomena underlying the data.

The details of the report are not important to us at this stage, but the existence of the report indicates the importance of the field.

- "Frontiers in Massive Data Analysis" (http://www.nap.edu/catalog/18374/frontiers-in-massive-data-analysis) by Committee on the Analysis of Massive Data; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; National Research Council (click through to download a PDF of 190 pages).

In many places, the report effectively states that Data Science is a new interdisciplinary field, for instance (page 15):

> This effort goes well beyond the province of a single discipline, and one of the main conclusions of this report is the need for a thoroughgoing interdisciplinarity in approaching problems of massive data.

In the UK, January 25 2013, Science minister David Willetts sets out plans (http://www.theguardian.com/science/2013/jan/25/government-technology-science-funding) for spending, and "Big Data" was listed as one of the eight major technologies to be supported. Subsequently, on September 4th 2014, the government announced (https://www.gov.uk/government/news/new-turing-institute-at-londons-knowledge-quarter-announced-by-chancellor) the new headquarters of the £42 million Alan Turing Institute for Data Science (http://en.wikipedia.org/wiki/Alan_Turing_Institute) will be at London's new Knowledge Quarter - with 'spurs' around the country:

> The world-class research institute, dedicated to British computer pioneer and WW2 Enigma code-breaker Alan Turing, will work with universities across the country to focus on new ways of collecting, organising and analysing large sets of data - commonly known as big data.

A Retrospective

First in series of blog posts by Evangelos Simoudis covering insightful applications. From an investor's perspective, Simoudis will discuss the changing landscape of data analytics, the value of insight-generation applications, and key areas in big data that are attracting the most VC interest.

- "Insightful applications: The next inflection in big data" (https://www.oreilly.com/ideas/insightful-applications-the-next-inflection-in-big-data), blog by Evangelos Simoudis,

February 2, 2016 (2000 words, 15 mins) 

1.5
Roles of a Data Scientist

With put together various sources to look at data scientists in their job.

Doing Data Science

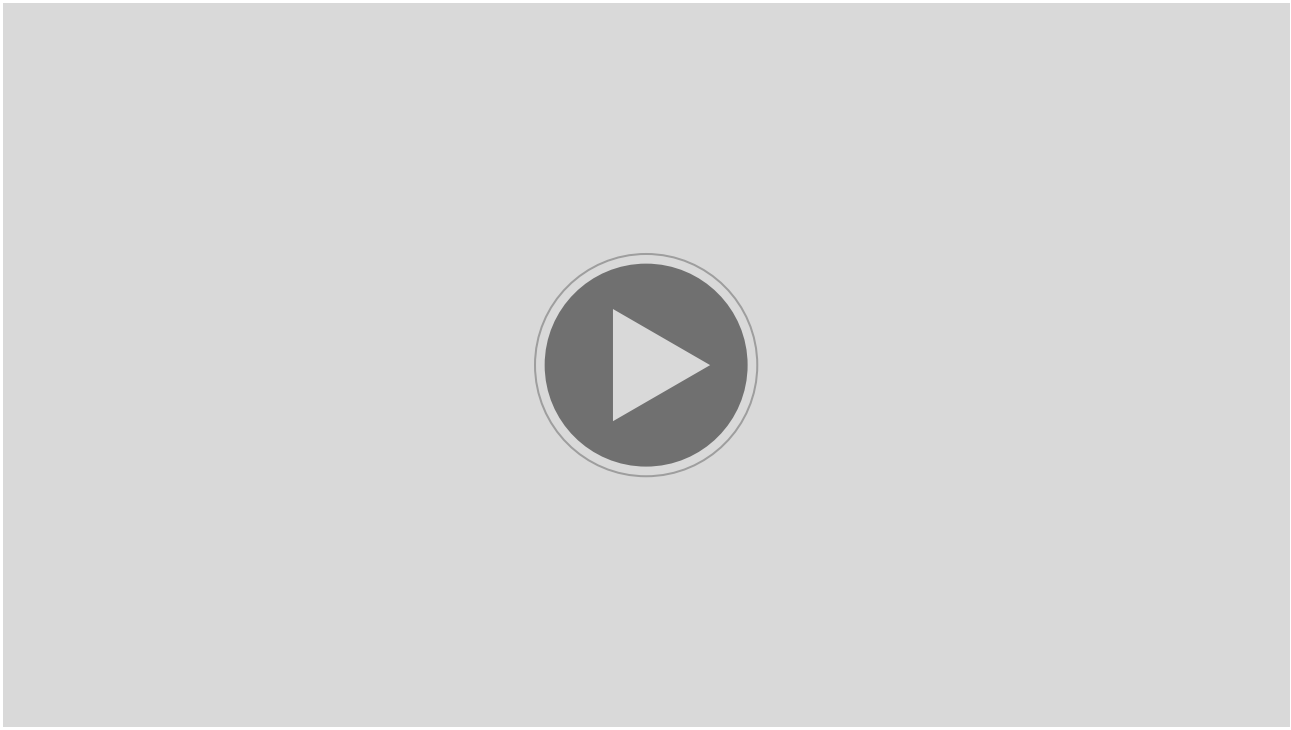
First, we will look at a general introduction to the field from Rachel Schutt and Cathy O'Neil. Their book developed out of a unit they taught at Columbia University (NY, NY), and many of the later chapters work through specific case studies using standard software such as R, which will make great individual study chapters for later. But for now we will take in their big picture of the field. In this chapter they start with some of the usual discussions about the field, but then they go on to talk about the skills matrix of data scientists and some of the roles in the job.

- Whole of Chap. 1 of *Doing Data Science*, Schutt and O'Neil, 2013 (http://search.lib.monash.edu/primo_library/libweb/action/search.do;jsessionid=317892714D7648DEBBBE84C82C5C3C08?fn=search&ct=search&initialSearch=true&mode=Basic&tab=default_tab&indx=1&dum=true&srt=rank&vid=MON&frbg=&vl%28freeText0%29=Doing+Data+Science&scp.scps=scope%3A%28catelec%29%2Cscope%3A%28catau%29%2Cscope%3A%28MUA%29%2Cscope%3A%%20border=) (approx 8000 words, 45 mins). The Monash Library has the e-book for student use, and chapters can be read online individually. 

Interviews with industry professionals

Second, we have done a few interviews with industry professionals here in Australia. We've grouped their segments together around themes, and give a good cross section of industry, government, consulting and research. Now these five professionals are leaders rather than workers, but they give you an idea of what is happening and what sorts of work their people do.

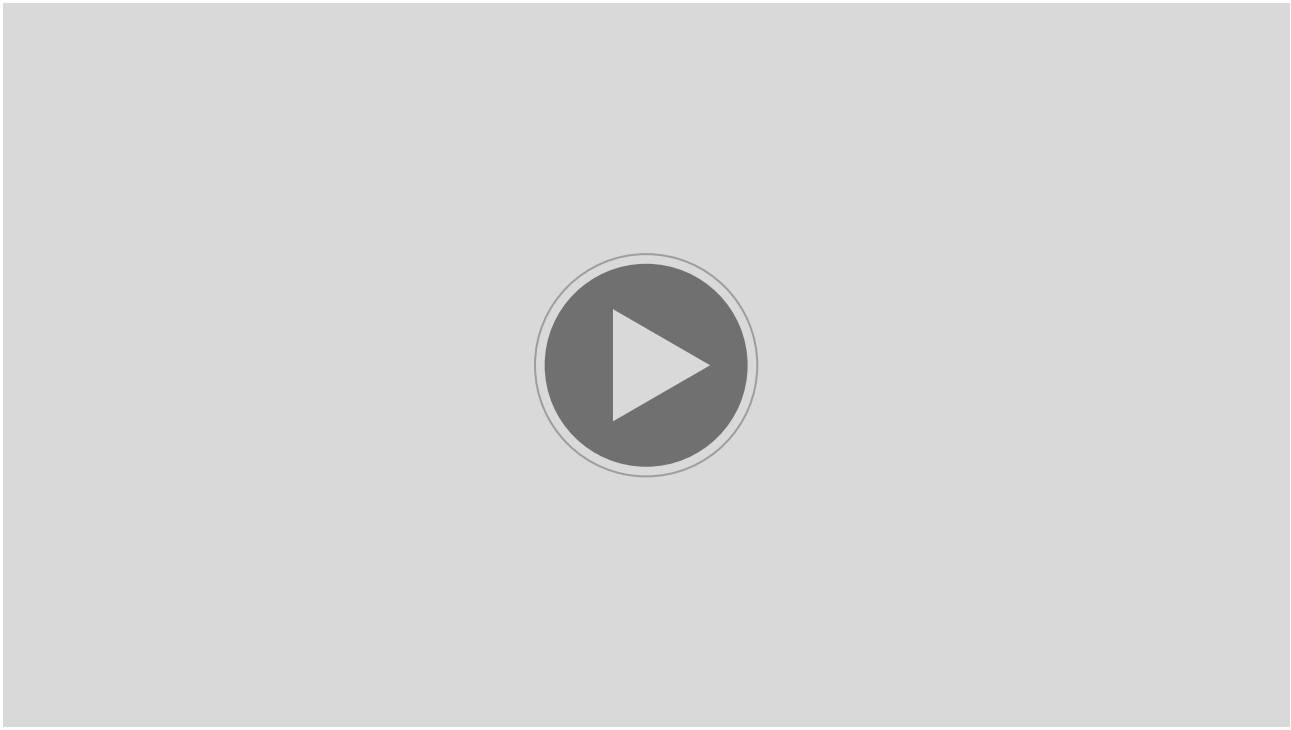
Watch Con Nidras (Head of Customer and Channel Analytics - National Australia Bank (NAB)), Associate Professor Michael Brand (Faculty of Information Technology - Monash University and former data scientist at Pivotal) and Dr Rami Mukhtar (CEO - Ambiata) talk about their **experiences as data scientists**.



(https://www.alexandriarepository.org/wp-content/uploads/20150629094719/FIT5145_module_1_being_data_scientist_combined.mp4.mp4)

Alternatively, you can download the transcript for Being a data scientist (https://www.alexandriarepository.org/wp-content/uploads/20150701100058/transcript_FIT5145_module_1_being_data_scientist.pdf).

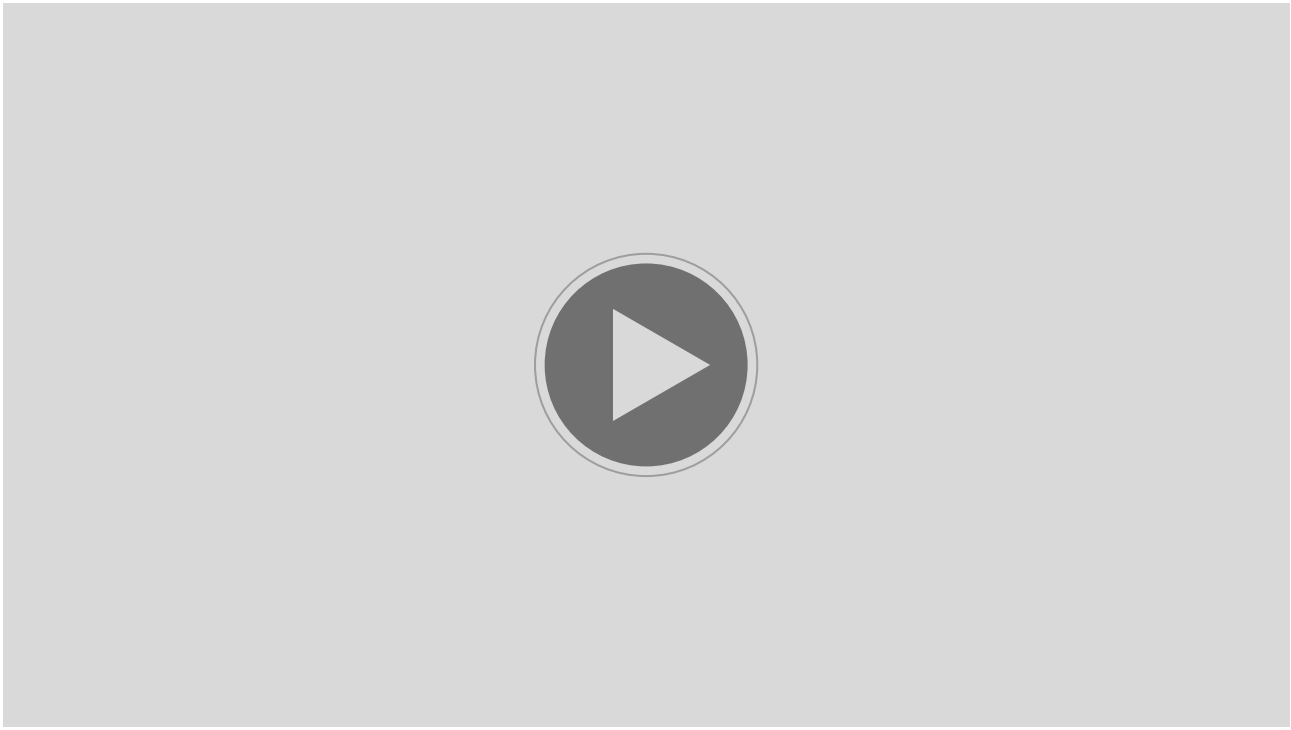
Watch Con Nidras (Head of Customer and Channel Analytics - National Australia Bank (NAB)), Associate Professor Chris Bain (Director of information services - The Alfred Hospital), Dr Fang Chen (Research Group Manager - National ICT Australia (NICTA)) and Associate Professor Michael Brand discuss the various aspects and issues of **working on a data science project**.



(https://www.alexandriarepository.org/wp-content/uploads/20150626201716/FIT5145_module_1_data_science_projects_combined.mp4.mp4)

Alternatively, you can download the transcript for Data science projects (https://www.alexandriarepository.org/wp-content/uploads/20150701100101/transcript_FIT5145_module_1_data_science_projects.pdf).

Watch Associate Professor Chris Bain (Director of information services - The Alfred Hospital) discuss the challenges of **working with data**.



(https://www.alexandriarepository.org/wp-content/uploads/20150629074030/FIT5145_module_1_working_with_data_combined.mp4.mp4)
Alternatively, you can download the transcript for Working with data
(https://www.alexandriarepository.org/wp-content/uploads/20150701100104/transcript_FIT5145_module_1_working_with_data.pdf).

Analyzing the Analyzers

In another O'Reilly short pamphlet, authors Harris, Murphy and Vaisman interviewed 250 data scientists to extract information about their skills sets and how they identified with their work.

- "Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work" (O'Reilly) (http://www.oreilly.com/data/free/files/analyzing-the-analyzers.pdf) by Harlan D. Harris, Sean Patrick Murphy, and Marck Vaisman (PDF, 30 pages, don't read, reference only)

First, using factor analysis, they identified four job categories for data scientists, and in each category a number of roles. These were identified from the sorts of skills people had rather than their current job title.

Category	Roles		
data developer	developer	engineer	
data researcher	researcher	scientist	statistician
data creative	jack-of-all-trades	artist	hacker
data businessperson	leader	business person	entrepreneur

The skills people had were categorised into the following areas:

- **Business:** product development, business
- **Machine learning/Big data:** unstructured data, structured data, machine learning, big and distributed data
- **Mathematics/Operations research**: optimisation, mathematics, graphical models, Bayesian and

Monte Carlo statistics, algorithms, simulation
- **Programming**: systems administration, back end programming, front end programming
- **Statistics**: visualisation, temporal statistics, surveys and marketing, spatial statistics, science, data manipulation

With these skill areas, Harris *et al.* were then able to create a map of skill sets to job categories, see their Figure 3-3 on page 13.

With this:

- most data scientists had at least moderate skills in all areas, so fairly good breadth
- most data scientists had one area of expertise or depth

The broad job categories for people calling themselves data scientists can then be discusses as follows:

- **Data developers:** were relatively strong in all areas except Statistics, Harris *et al.* say they are "people focused on the technical problem of managing data - how to get it, store it, and learn from it."
- **Data researchers**: had little Programming, Harris *et al.* say they are people with an academic research background, using their training to "understand complex processes"
- **Data businesspersons:** had little Mathematics/Operations research and even less Programming, Harris *et al.* say they are "most focused on the organization and how data projects yield profit."
- **Data creatives:** were relatively strong in all areas except Mathematics/Operations research, Harris *et al.* say they are " the broadest of data scientists, those who excel at applying a wide range of tools and technologies to a problem, or creating innovative prototypes at hackathons."

Data Analytics Handbook

Data Analytics Handbook is a four volume set of long interviews from industry and academic professionals in the field. Edition 1 deals with practitioners so is most relevant to this section.

- https://www.teamleada.com/handbook, Volume 1, with 7 interviews (approx 10,000 words, 1 hour ... careful, there is a lot, read a selection). 

They have "Top 5 Takeaways" at the beginning of the volume that are so good we duplicate them here. Read their edition to get the details!

1. Communication skills are underrated.
2. The biggest challenge for a data analyst isn't modelling, it's cleaning and collecting.
3. A Data Scientist is better at statistics than a software engineer and better at software engineering than a statistician.
4. The data industry is still nascent, if you want to work with a variety of stakeholders in a more free-form role, the time to do so is now. (as of early 2014)
5. Both roles require a curiosity about working with data, a quality more important than your technical abilities.

Lifehacker

Lifehacker is an interesting e-magazine on "how to get things done". They have an *IT Pro* section and recently interviewed a data scientist.

- "Career Spotlight: What I Do as a Data Scientist" (http://lifehacker.com/career-spotlight-what-i-do-as-a-data-scientist-1684793405), by Andy Orin (1700 words, 9 mins)

How to become a data scientist

While its hard to take any infographic too seriously, following XKCD's own contribution (https://xkcd.com/1273/), this infographic offers another alternative of the roles a Data Scientist plays, and the skills they need to know.

- "How To Become A Data Scientist in Simple Steps: The Infographic" (http://fossbytes.com/become-data-scientist-simple-steps-infographic/) on Fossbytes.com (infographic)

1.6
Activity: Data Science Job Trends

Data Science Job Trends

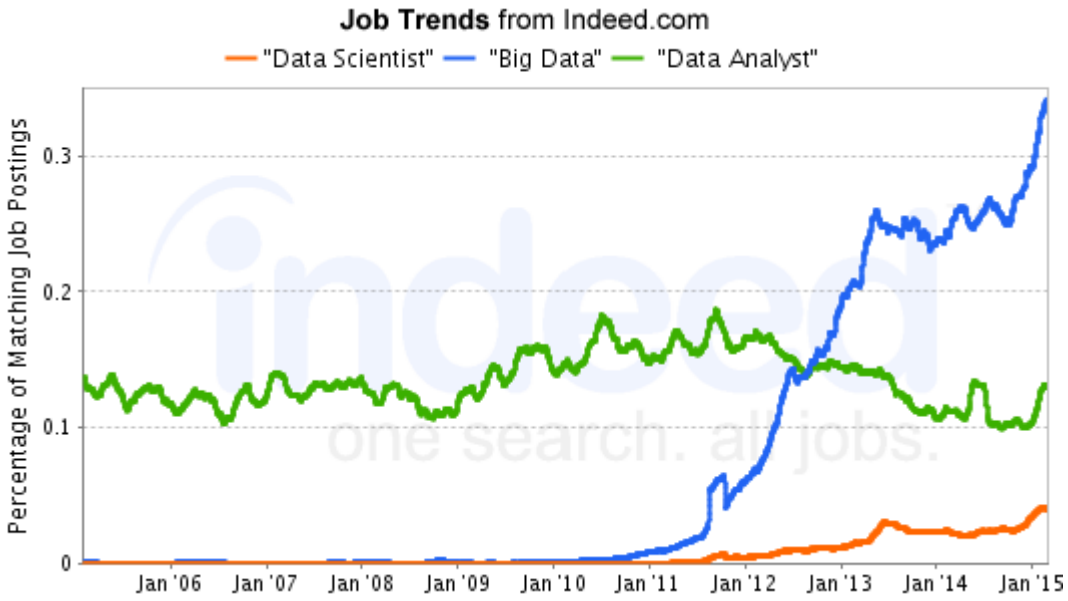
Let's explore some of the roles and skills in Data Science using job trends.
One resource is indeed.com: http://www.indeed.com/jobtrends

Step 1

Search for:

```
"Data Scientist", "Big Data", "Data Analyst"
```

keeping the double quotes in the query, to see:



Step 2

Add Data Science (in quotes) to the search terms, what is the result?

Step 3

Compare these results to Australia, search for the same exact terms (one at a time) on *e.g.* Seek.com.au (http://www.seek.com.au)
How many ads (the number of jobs found is in pink in top left corner) for:

```
"Data Scientist"
"Big Data"
"Data Analyst"
"Data Science"
```

What's the difference with US trends? For Seek.com.au you cannot look at trends over time, but you can compare the relative strength now for different key words.

Step 4

Search some of the following Australian sites for

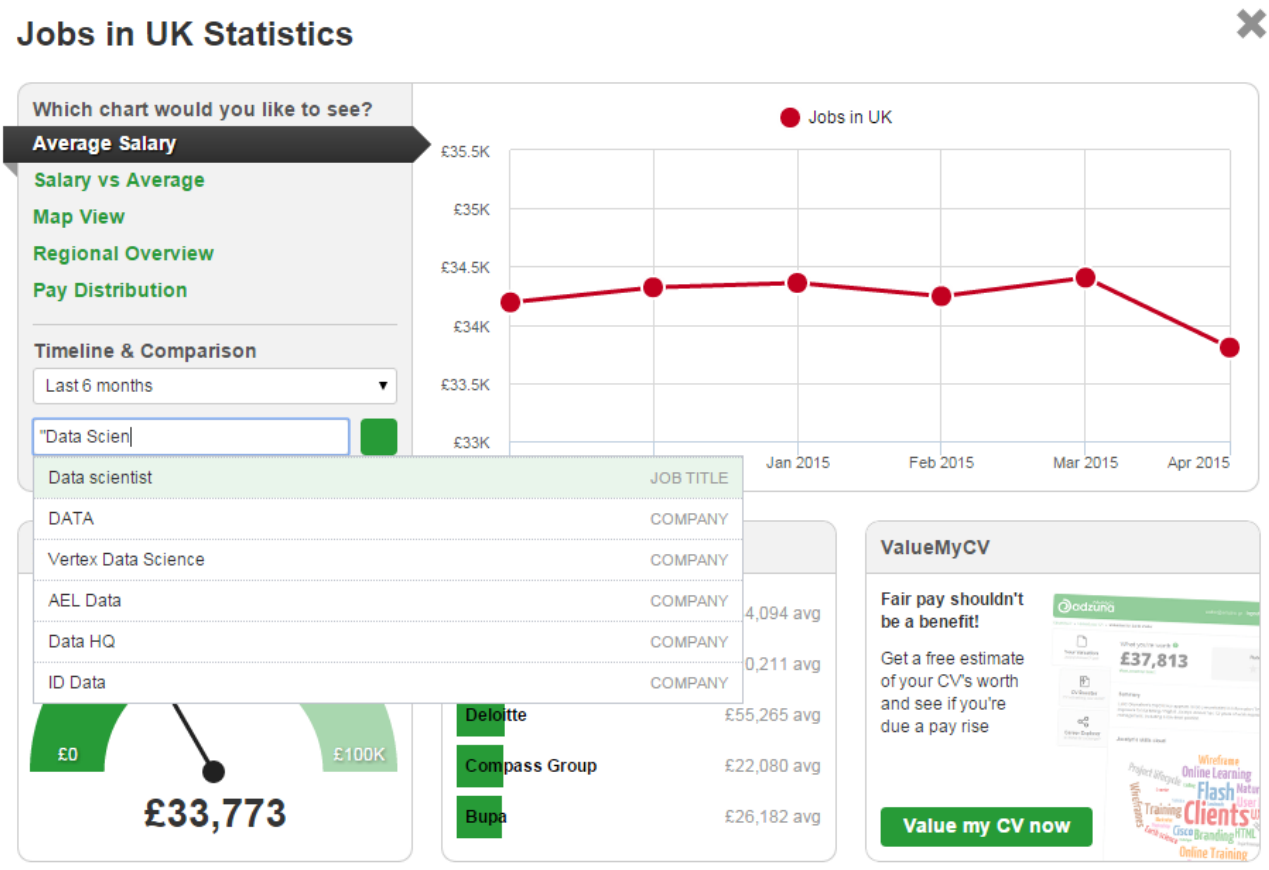
`"Data Scientist" or "Data Science"`

- http://www.livesalary.com.au/
- http://au.hudson.com (http://au.hudson.com/portals/au/documents/Salary%20Guides/SalaryTables2015-Aus-ICT.pdf) (e.g. SalaryTables)
- http://www.abs.gov.au/
- http://lmip.gov.au/

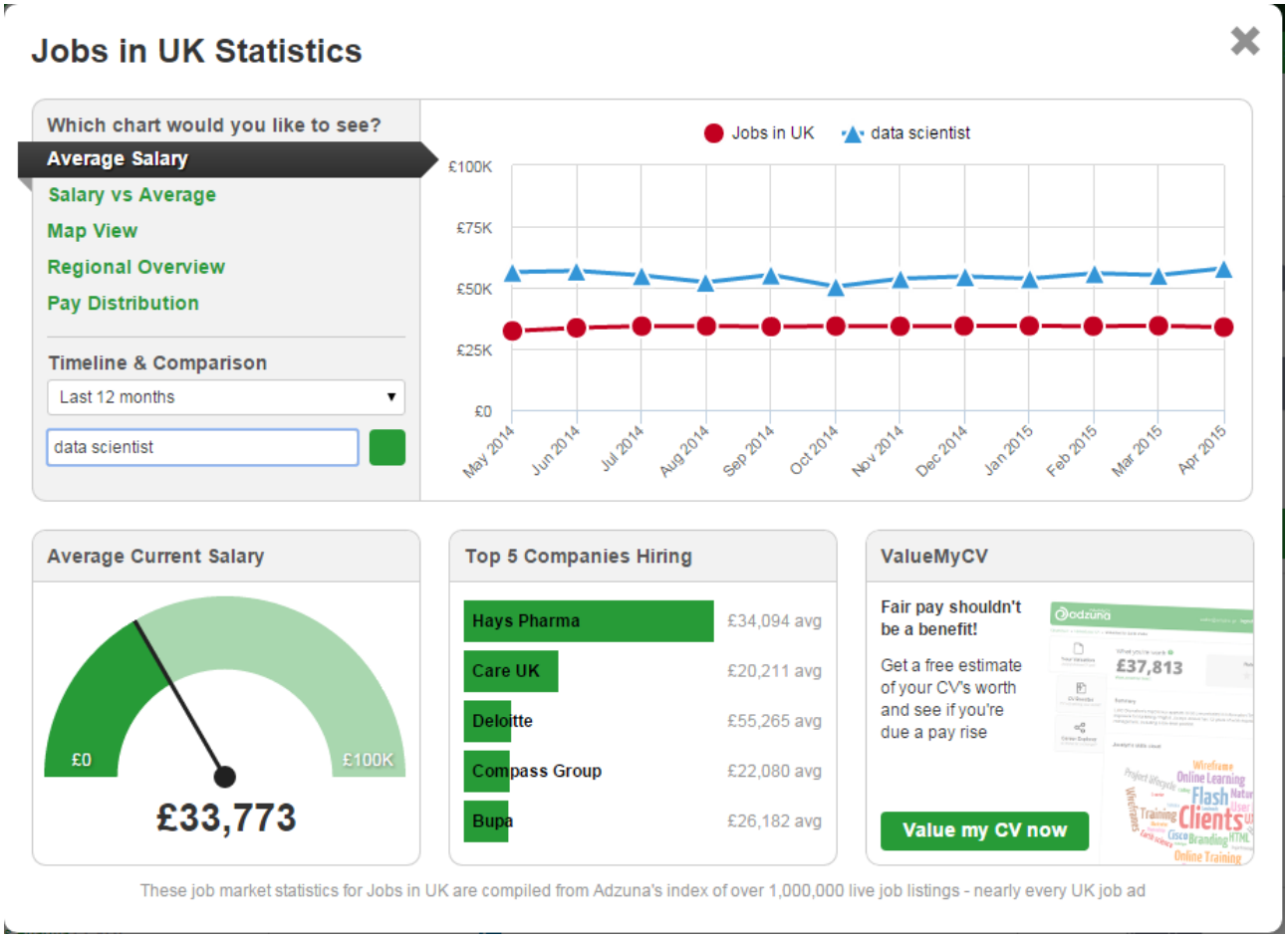
What did you find? Are there any related terms you should try, like "business analyst"?

Step 5

There are many options for job searches, one is Adzuna, it offers a stats/map tool: http://www.adzuna.co.uk/jobs/uk#stats



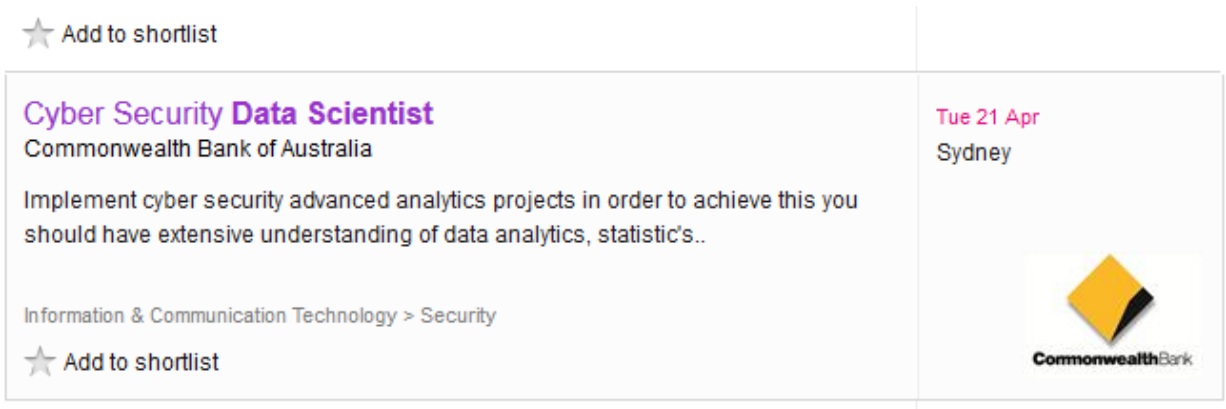
Start typing 'Data Scien...' (shown above) to see that "Data Scientist" is a 'known' job in the UK (above). Select it to see:



There is also a 'Map View' and 'Pay' and 'Salary' *etc*. tools at left. What is the average salary in the UK, and the average salary for a "Data Scientist" in the UK? Which companies are hiring?

Step 6

There is an Adzuna in Australia, and it had a map/stats tool. You can still search for jobs on Adzuna Australia or using Seek. What are some ways you can get the data from the search results page? What are some ways you can clean up (wrangle) this data? (*e.g.* if you copy from the screen you get a lot of 'Add to shortlist' in Seek *etc.*). Here's the summary view of a job from Seek:



Step 7

There is a famous (or infamous) "Metromap" (*i.e.*, analogous to a London metro map) for data science skills. http://nirvacana.com/thoughts/becoming-a-data-scientist/

This is a useful resource because it includes a lot of the phrases (for skills) you will see on a job advertisement, except for general business, IT and communication skills, etc. However, it is extensive. Do you think a typical data scientist would have all of these skills? Note each segment in the map is a general category, such as "data munging", "text mining", *etc*. So, we're going to analyse some of the job advertisements and we want to pull out the skill sets related to data science, *i.e.*, those that could be on the Metromap.

Step 8

Here's the full advertisement from the CBA:

Cyber Security Data Scientist

- **Be a part of a high profile Security Program with significant investment**
- **Leading edge technology**
- **Opportunity to work on greenfield projects**

Your Team

The Digital Protection Group (DPG) protects the bank and our customers from theft, losses and risk events, through effective and proactive management of cyber security, privacy and operational risk. This role reports to the Senior Manager Cyber Security Reporting and Analytics. You will be working with a team of cyber security professionals including Data Engineer to support the objectives of DPG. The Information Security Governance and Assurance team provides a number of services to our customers and a number of capabilities that are leveraged across the Group:

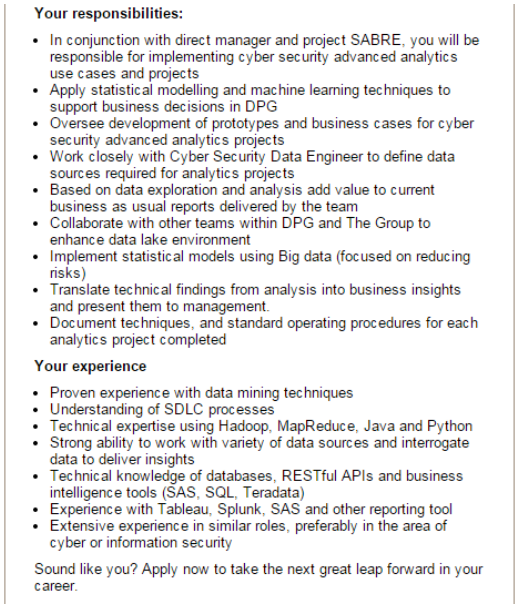
- Cyber Security Reporting & Analytics capabilities
- Archer GRC capabilities
- Material Supplier Information Security Governance Services
- Supplier & Third Party Information Security Assessment Services
- Cyber Security Compliance management Services
- Cyber Security Consultancy services including Cyber Readiness
- Archer development services

Your contribution:

Your role will be key to implement cyber security advanced analytics projects that will support the Governance and Assurance team in making decisions based on data discovery and insights. In order to achieve the you should have extensive understanding of data analytics, statistical models, SQL, and strong ability to analyse complex data to find insights.

Note the keywords, the skills, the industry, also what is missing ($, qualifications). This is an image (or two). Can you get information out of it in digital form? What about from the source data (the web pages). If you wanted to compare salaries, skills, locations, companies automatically, with an algorithm, what are your options? You could outsource digital coding of the page (*e.g.* see Mechanical Turk), to divide and conquer, or do it yourselves...

Find job data on Seek pertaining to "Data Scientist." What companies/organisations? What skills? What qualifications, pay, location...



Your responsibilities:

- In conjunction with direct manager and project SABRE, you will be responsible for implementing cyber security advanced analytics use cases and projects
- Apply statistical modelling and machine learning techniques to support business decisions in DPG
- Oversee development of prototypes and business cases for cyber security advanced analytics projects
- Work closely with Cyber Security Data Engineer to define data sources required for analytics projects
- Based on data exploration and analysis add value to current business as usual reports delivered by the team
- Collaborate with other teams within DPG and The Group to enhance data lake environment
- Implement statistical models using Big data (focused on reducing risks)
- Translate technical findings from analysis into business insights and present them to management.
- Document techniques, and standard operating procedures for each analytics project completed

Your experience

- Proven experience with data mining techniques
- Understanding of SDLC processes
- Technical expertise using Hadoop, MapReduce, Java and Python
- Strong ability to work with variety of data sources and interrogate data to deliver insights
- Technical knowledge of databases, RESTful APIs and business intelligence tools (SAS, SQL, Teradata)
- Experience with Tableau, Splunk, SAS and other reporting tool
- Extensive experience in similar roles, preferably in the area of cyber or information security

Sound like you? Apply now to take the next great leap forward in your career.

Pick **five** advertisements and identify several major terms from the job advertisement (In some cases, there may be too many to include all).
Is this manual approach scalable?
Usually the key terms are 'tagged' for you, e.g. 'qualifications' or 'responsibilities' - what if they're not?

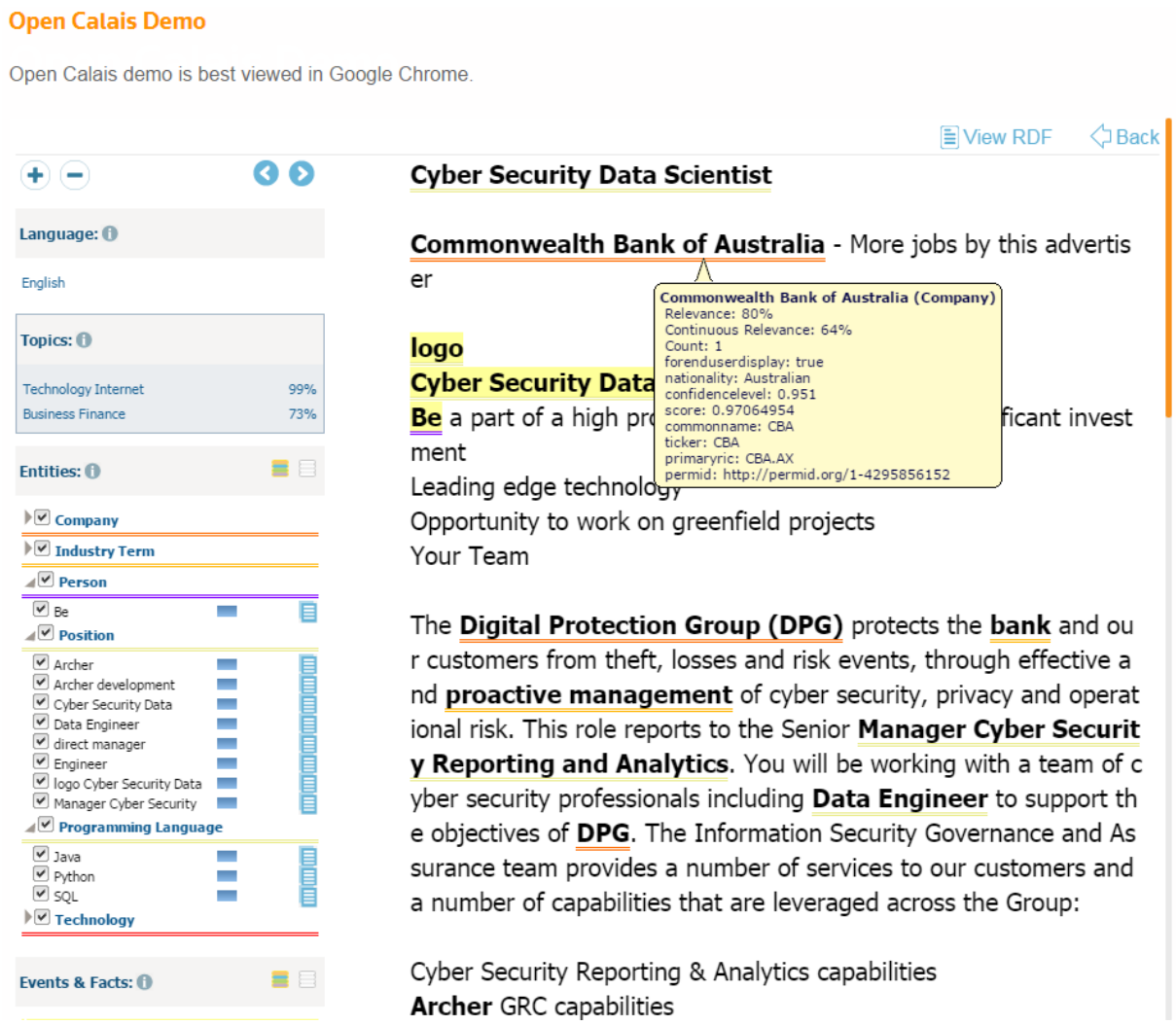
What if companies use different synonyms? For instance "munging" versus "wrangling."

What happens if you search for *e.g.* 'Apple' - is this Apple Inc, apple (horticulture) or a location (New York is 'the big apple')?
Can computers make sense of ads like this, can they process them, summarise, extract...?

Step 9

Open Calais has an online demo which processes and tags text, as below it 'knows' that the CBA is a company, and is important (relevance 80%), and has detected several programming languages (Java, Python, SQL?). But tagged ('thinks') Archer is a position. What is Archer GRC? How would you tag it?

http://new.opencalais.com/opencalais-demo/



Pick a job you like and paste it into OpenCalais then explore its tagging of the text. What is accurate, and what is not? Look at the top right (above) to see "View RDF". What is RDF?

Step 10

The process of tagging the text by identifying entities is part of the role of Information Extraction (https://en.wikipedia.org/wiki/Information_extraction) and Text Mining (https://en.wikipedia.org/wiki/Text_mining). Why is this difficult?

Appendix

Here are some other interesting links:

http://www.kdnuggets.com/2014/12/data-science-skills-most-demand.html
(http://www.google.com/url?q=http%3A%2F%2Fwww.kdnuggets.com%2F2014%2F12%2Fdata-science-skills-most-demand.html&sa=D&sntz=1&usg=AFQjCNFfj-_hh9Rb8QR5tNLIERur2tOqMQ)

http://www.oreilly.com/data/free/files/2014-data-science-salary-survey.pdf

http://www.payscale.com/research/US/Job=Data_Scientist,_IT/Salary

http://au.indeed.com/hire/research

http://insight.seek.com.au/news/how-to-hire-a-data-scientist

http://insight.seek.com.au/news/get-business-ready-for-big-datas-golden-ticket

http://blog.revolutionanalytics.com/2011/07/growth-in-data-related-jobs.html

And a visualisation of skills:
http://insights.dice.com/2015/07/01/dice-data-how-tech-skills-connect/?CMPID=AF_SD_UP_JS_AV_OG_DNA
=

1.7
Impact of Data Science

Here we present some short sections relevant to the impact and the types of data science. In some cases, they raise more questions than they answer, so we use them as discussion starters.

Your life on the cloud

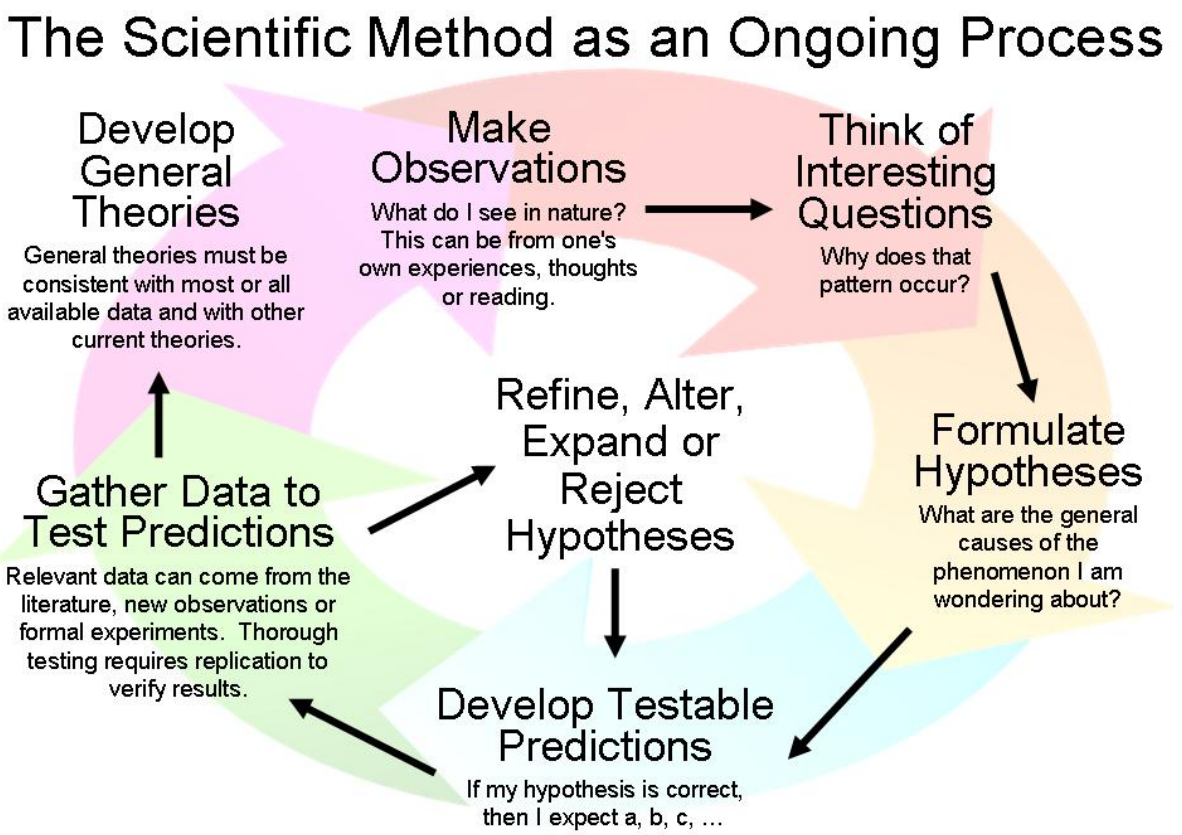
Alistair Croll talks about what he believes is the major consumer impact for the rise of big-data and data science:

> In 10 years, every human connected to the Internet will have a timeline. It will contain everything we've done since we started recording, and it will be the primary tool with which we administer our lives.

- "Year Zero: Our life timelines begin" (http://radar.oreilly.com/2015/03/year-zero-our-life-timelines-begin.html) on O'Reilly (2400 words, 15 mins) 
- "Year Zero: How We'll Run Our Lives in Ten Years' Time" (https://www.youtube.com/watch?v=EtTNYarshrU) (embedded Youtube video, 5:30 mins)  - this is a version of the blog given as a Strata+Hadoop 2015 talk

The end of science

Chris Anderson, editor of Wired, wrote an opinion piece 6/2008 in *Wired* claiming "the end of the scientific method". First, its best to quickly consider what it is: the Wikipedia page on Scientific method (https://en.wikipedia.org/wiki/Scientific_method) is a good introduction where the following figure sums it up:



12

"Scientific Method 3" by Whatiguana - Own work. Licensed under CC BY-SA 4.0 via Wikimedia Commons

Anderson's opinion piece is then worth looking at:

> "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete,"
>
> (http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory) on O'Reilly (1300 words, 8 mins) 

The piece is best viewed as a position statement rather than a professional scientific proposal. One interpretation of the article might be that with the flood of data, we can now search for correlations (in a statistically sound manner, or course!) rather than follow the time-honored scientific method (which intrinsically involves building testable causal models).

Philosophers commented (see the opinion piece (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2711825/) from Massimo Pigliucci from CUNY), Slashdot garnered popular feedback (http://science.slashdot.org/story/08/06/25/146250/google-begat-the-end-of-the-scientific-method), and also statisticians responded, with Andrew Gelman, Bayesian author from Columbia University, making a comment in his blog (http://andrewgelman.com/2008/06/26/the_end_of_theo/). Not mentioned here, but something we will discuss later in the unit are some of the more recent problems with the scientific method, its gamification for economic gain.

Data science for social good

Technology and "improving the third world" have a long and sometimes chequered interaction, though also dramatic successes such as one laptop per child (http://one.laptop.org/). Critics say the most important contributions there are clean water, sanitation and education, not technology. However, looking at it more

broadly, and viewing "social good" from an international perspective, clearly many opportunities exist: funding and organisational applications, helping underprivileged pockets of first-world countries, *etc*. The **Data Science for Social Good** **(http://dssg.io)** movement is championed by Rayid Ghani (http://www.rayidghani.com/wordpress/) at the University of Chicago.

Dr. Eric Horvitz, Distinguished Scientist & Managing Director at Microsoft, gave a 40 minute wide-ranging technical talk on data science with the theme of "social good." It was an invited talk at KDD 2014 in New York (http://www.kdd.org/kdd2014/), and has been recorded and presented on the VideoLectures.NET (http://VideoLectures.NET) site for scientific content. But we will look at the last part now and consider some of the more technical sections later in the unit.

- "Data, Predictions, and Decisions in Support of People and Society" (http://videolectures.net/kdd2014_horvitz_people_society/) see the final section 46:51-53:00 mins (video with slides, 6 mins) 

This section starts off by considering the task of predicting earthquake locations from mobile phone connection data, and goes onto to discussing applications generally.

21 ways how big data will improve your life

While its hard to take any infographic too seriously, following XKCD's own contribution (https://xkcd.com/1273/), this infographic, along with the usual fluff about big data and words beginning with "V", does a reasonable job of gathering together a number of different societal impacts. You have to page way down the bottom to find the impact section.

- 21 Ways How Big Data Will Improve Your Life (https://datafloq.com/read/21-ways-will-big-data-improve-life/247) (infographic) 

How the data revolution will change the world

This is another journalistic piece by Elana Levine on the World Economic Forum website as she discusses views of Jeremy Howard, data science entrepreneur, raised at a recent WEF meeting.

- "How the data revolution will change the world" (https://agenda.weforum.org/2014/09/data-revolution-changing-world/), (1800 words, 12 mins)

Big data - 2020 vision

This is a short 5 minute talk by SAP manager John Schitka at the Strata-Hadoop 2014 conference. You will have to pass through the first 1 minute of big data statements but then the speaker gives some projections about use of data science and related computing technologies, and by 2 mins starts to discuss the future of healthcare, self driving cars.

- "Big Data - 2020 Vision" (http://player.oreilly.com/videos/9781491900345?toc_id=192995) on Strata (video, 5 minutes; note first 2 mins is review) 

He raises the question, what is going to change about the way industries and life works?

Data science is a competitive sport

This is a short ABC documentary video from the *Catalyst* programme describing the *Kaggle* system (founded by Jeremy Howard, mentioned above). The underlying idea is the use of collaboration, crowd-sourcing (https://en.wikipedia.org/wiki/Crowdsourcing) and gamification (https://en.wikipedia.org/wiki/Gamification) to solve science and industry problems based around data science competitions.

- "Lucrative Algorithms" (http://www.abc.net.au/catalyst/stories/3296837.htm) (video, 6 mins) 

Arguably this was an important initial part of the data science culture, students we encouraged to engage in these competitions. Many companies and organisations are now doing this, for instance see the Artificial Intelligence for Development website (http://ai-d.org) mentioned by Horvitz above.

Impact on business operations

How does data science affect the day to day operations of businesses? How do you enable employees with access to data and the ability to analyze it? A global survey conducted by the Economist Intelligence Unit and Alteryx considered these questions and developed the following infographic.

- "Humanizing Big Data" (http://visual.ly/humanizing-big-data) by Alteryx (infographic, 5 mins)

1.8
Activity: Motion Charts

Motion charts

Motion Charts are interactive multi-dimensional data visualisations, originally introduced to the world as GapMinder by Hans Rosling and made famous by his TED talks. The GapMinder technology was bought by Google and the name changed. There are a few GapMinder videos around showing various data sets in action.

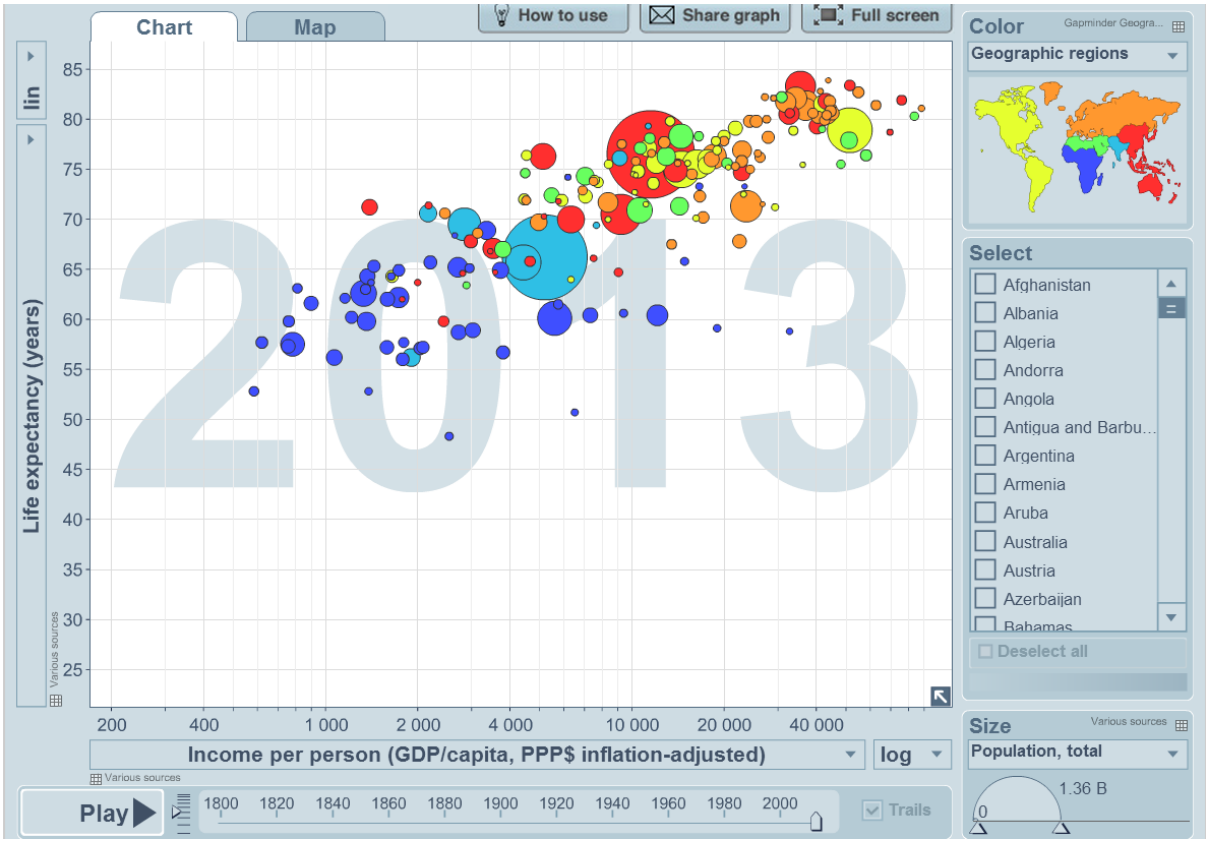
GapMinder

You can see swine flu in the news:
http://www.gapminder.org/videos/swine-flu-alert-news-death-ratio-tuberculosis/

or cancer:
http://www.gapminder.org/videos/liver-cancer-statistics/

or even the original TED talk:
http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen
(http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen?language=en)

Or you can interact with the data yourself:
http://www.gapminder.org/world (wait for it to load, there's a lot of data)



Step 1

Click 'Play >' to run the chart, you will see data from 1800 on. Note the axes ('Income..' and 'Life Expectancy'), also the colours and sizes of the bubbles - there are four attributes plus time, and there are controls on all four sides of the chart, plus the 'bubbles'. See *Gap Minder Controls* PDF (http://www.gapminder.org/GapminderMedia/wp-uploads/tutorial/Gapminder_World_Guide.pdf) for details. Note that many of these 'bubbles' don't move initially - why? (Have a look at the ones that do for a clue).

Step 2

'**Select**' Australia (right side) and 'Play', when does data for Australia begin?

Step 3

Compare Australia and New Zealand ('Select' NZ and 'Play'), when does data for NZ begin?
Where would you rather live?

Another chart

We can investigate other issues, for example Health. Which country do you think has the lowest infant mortality rate, Singapore, Sweden or Cuba? The most doctors per population? (The most money spent on health?)

Step 1

For the Y-axis select: 'Health' > ' ' (**NB.** top menu item appears blank?) > 'Infant mortality rate (per 1,000 births)'.

Step 2

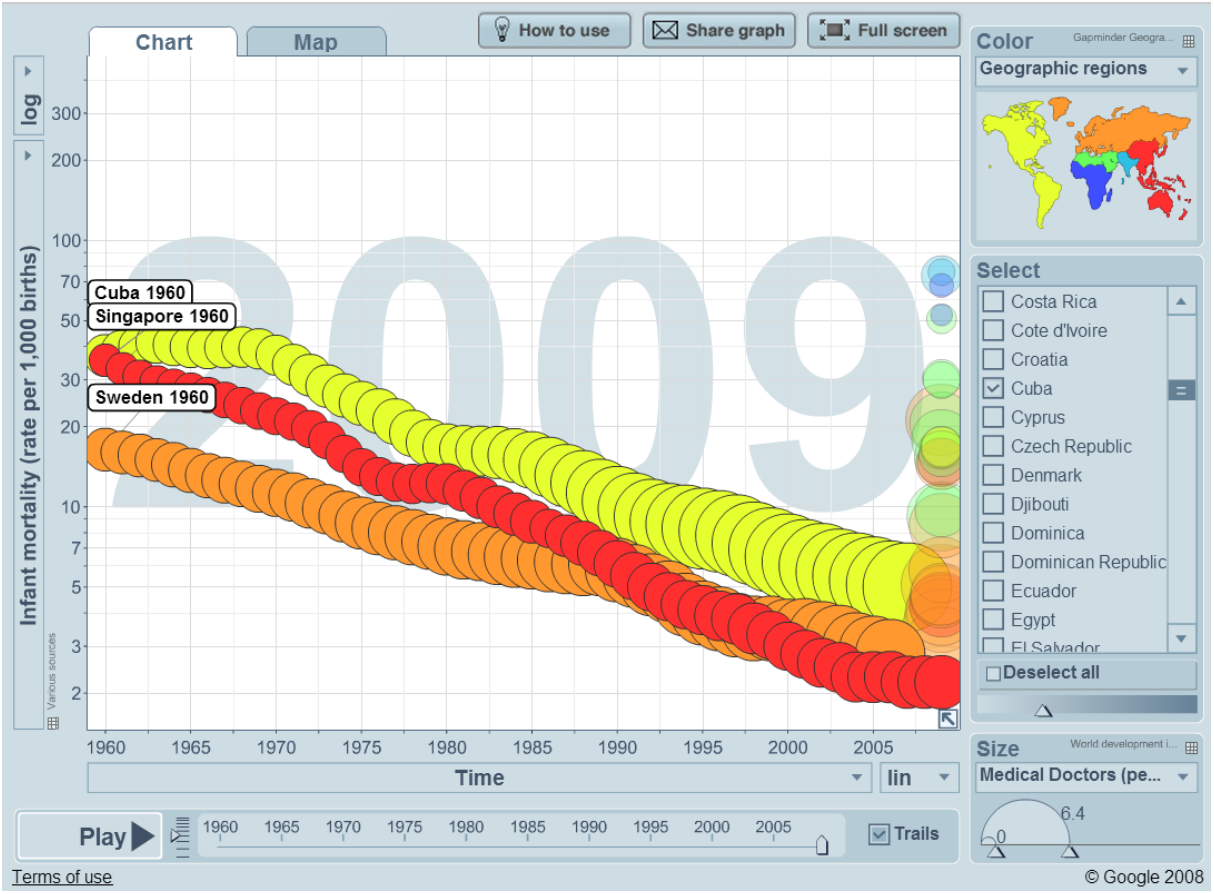
Change the X-axis to 'Time'

Step 3

Change the 'Size' control to 'Indicators' > 'Health' > 'Health Economics' > 'Medical Doctors (per 1000 people), like this:

Step 4

Select Sweden, Singapore and Cuba and select 'Trails' (Bottom right), then 'Play >'



Singapore passed Sweden in the 90s. Cuba has a lot of doctors. You may have noticed that there are a lot of controls and even more indicators (hundreds). You can do the math, 100s of indicators 4 or 5 different ways is astronomical. The trick then is in choosing what makes sense to model or compare.

Another chart

Step 1

In which country do people live the longest on average today: Botswana, Egypt or Cuba?
(Start by changing the Y-axis from 'Infant mortality..' to 'Life Expectancy', choose the countries then play.)

What happened to Botswana around 1990? War?

Step 2

Use the 'Size' control. select 'Indicators' > 'Health' > 'HIV' > 'Adults with HIV... ' Is that (war?) an explanation, the only factor?

Gap Minder Controls

Click on image to get a readable version.

(https://www.gapminder.org/GapminderMedia/wp-uploads/tutorial/Gapminder_World_Guide.pdf)

Motion charts in Google

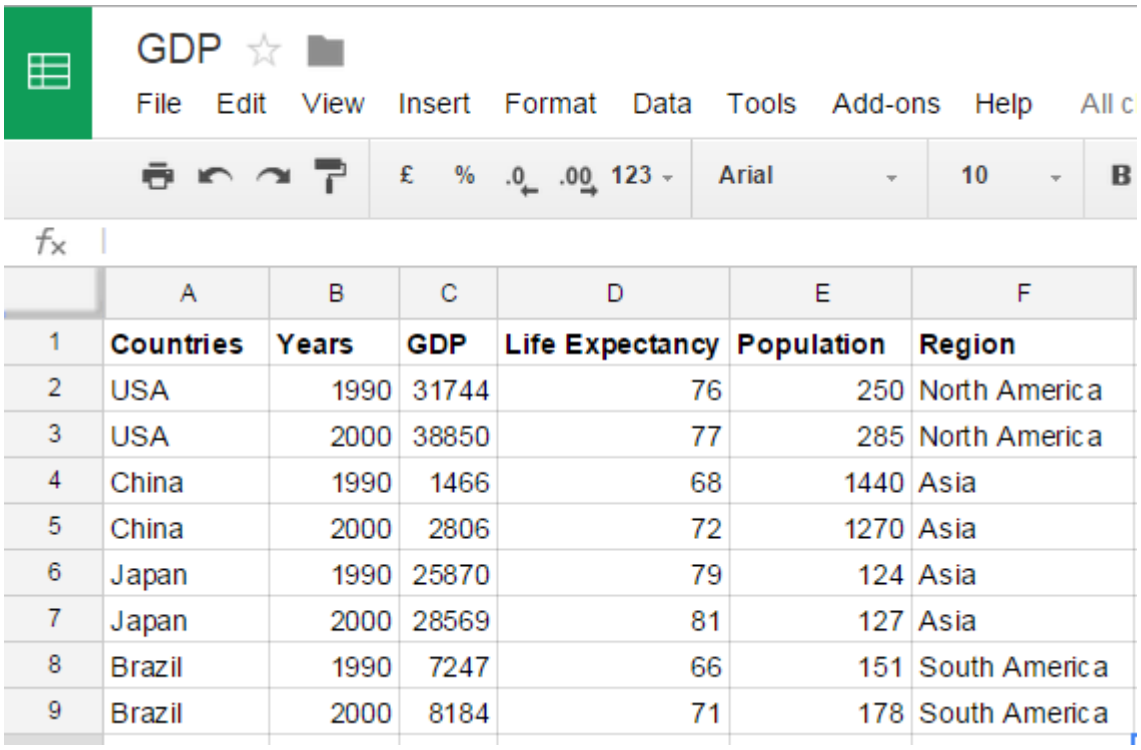
If you have gmail (and GoogleDoc) accounts, you can make your own Motion Charts but the data must be in a specific format:

Left to Right: name/category, date/time, 1 to 4 columns of data (x, y, colour, size)
https://developers.google.com/chart/interactive/docs/gallery/motionchart
e.g.

Countries	Years	GDP	Life Expectancy	Population	Region
USA	1990	31744	76	250	North America
USA	2000	38850	77	285	North America
China	1990	1466	68	1440	Asia
China	2000	2806	72	1270	Asia
Japan	1990	25870	79	124	Asia
Japan	2000	28569	81	127	Asia
Brazil	1990	7247	66	151	South America
Brazil	2000	8184	71	178	South America

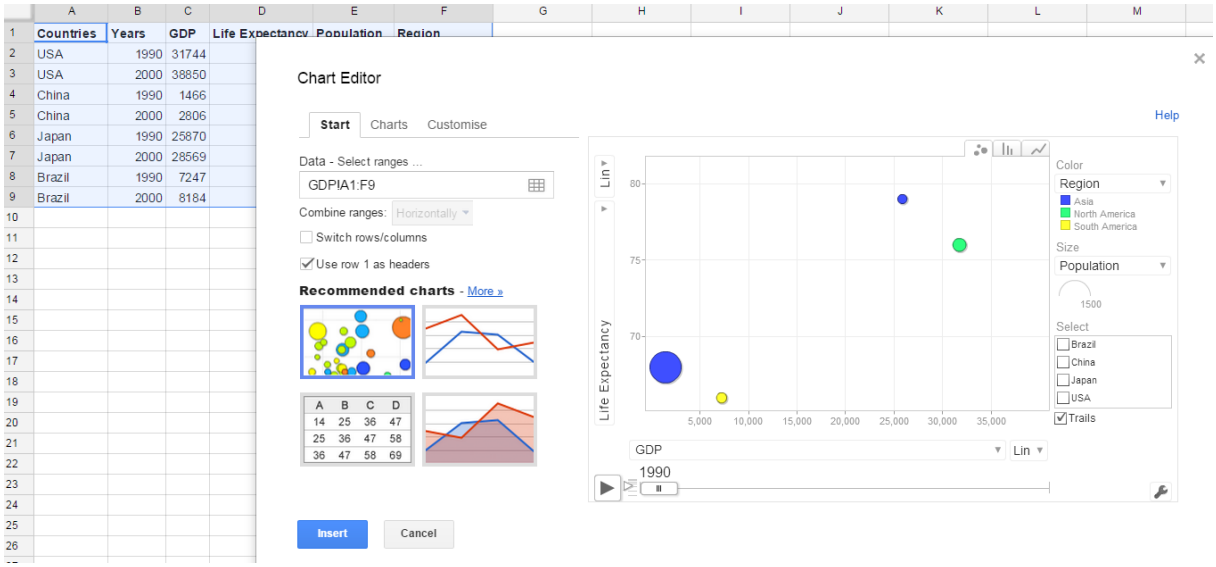
Step 1

Create a spreadsheet in GoogleDocs and type or paste the above data:



Step 2

Select all the data (from A1:F9) and using the 'Insert' menu, choose 'Motion Chart'



Step 3

Now 

to add the chart to your sheet then 'play'  to run. There are a lot of controls in Motion Charts, on all 4 sides plus the bubbles in the middle of the chart. Note that the bubbles move smoothly from 1990 to 2000 but that there are no intermediate years - what's going on?

Step 4

Add Australia to the dataset, you can research the GDP etc. or approximate. Run, test.

Step 5

Now get a big dataset from GapMinder: http://www.gapminder.org/data/ (or from US labor stats, see: www.bls.gov/osmr/pdf/st110110.pdf (http://www.bls.gov/osmr/pdf/st110110.pdf)). Explore and assess:

- How do you rate GapMinder as a tool, as used by Rosling, for social good, for education?
- GapMinder is nearly 10 years old (Google bought it circa 2008), what's available now?

Critiques

> "The one danger of great data tools like these, however, is that they create such beautiful graphs that it is easy to forget that what you are looking at are correlations, not necessarily anything causal." STEVEN D. LEVITT
> http://freakonomics.com/2008/06/16/gapminder-is-cool/ (http://freakonomics.com/author/stevenlevitt/)

Who is Steven Levitt? What's the difference between correlation and causality?
(We will look at this in more detail in a later module).
https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

You can see another critique of motion charts here:
http://www.juiceanalytics.com/writing/better-know-visualization-motion-charts/

Bill Gates says Rosling's work persuaded him to give billions to the developing world:
http://www.telegraph.co.uk/culture/tvandradio/10431350/Hans-Rosling-the-man-who-makes-statistics-sing.html

Optionally, have a look at this BBC documentary on GapMinder: "Don't Panic - The Truth About Population."
http://www.gapminder.org/videos/dont-panic-the-facts-about-population/

1.9
Activity: SAS Registration

SAS (Statistical Analysis System)

SAS software is widely used in many domains, including Science, Business and Medicine. It's a required skill according to some: "SAS and/or R - In-depth knowledge of at least one of these analytical tools" (see KD Nuggets (http://www.kdnuggets.com/2014/11/9-must-have-skills-data-scientist.html)). According to Gartner it's the best analytics software (http://www.kdnuggets.com/2015/02/gartner-2015-magic-quadrant-advanced-analytics-platforms.html).

We're going to use SAS Studio and SAS Visual Analytics (and possibly Enterprise Miner) later in this unit, but first we need create an account and a profile:

Step 1

Go to **https://odamid.oda.sas.com/SASODARegistration/** use your Monash email (gmail), name, country, then submit.

Step 2

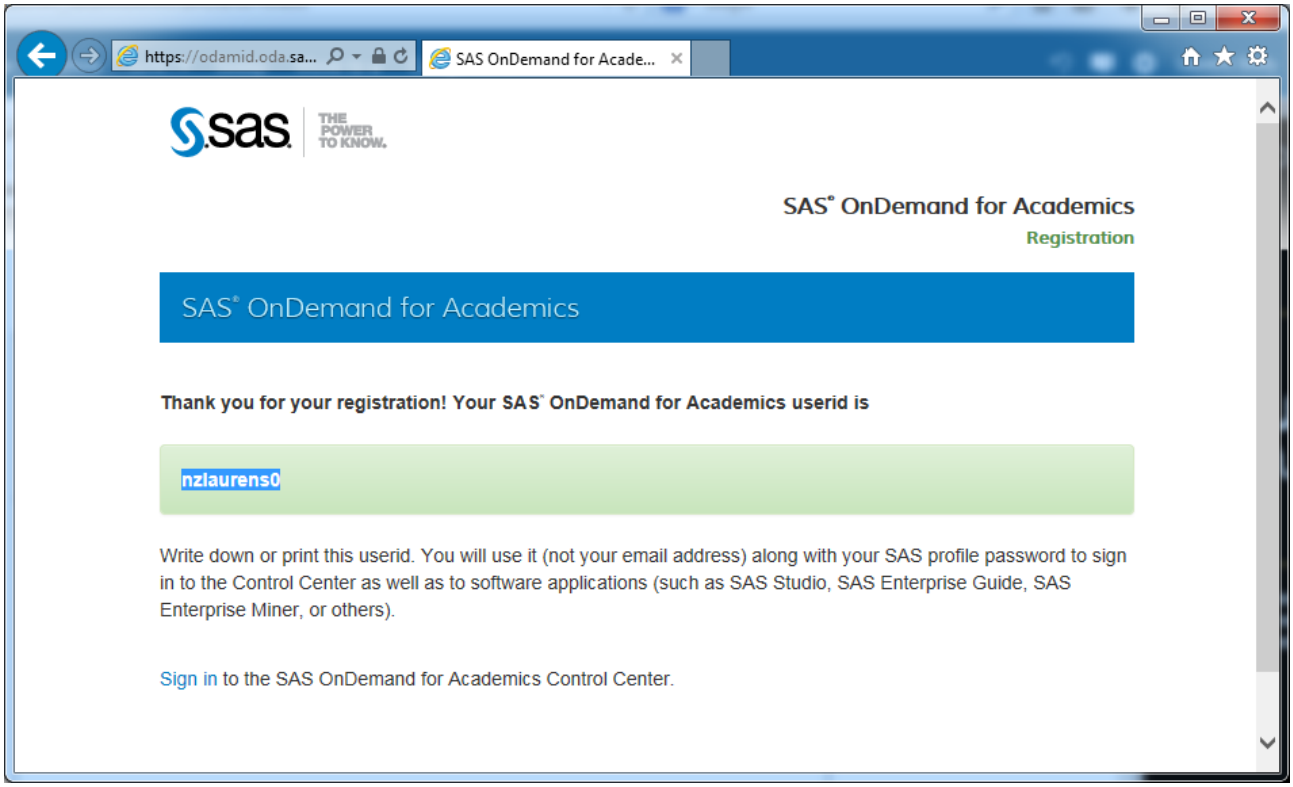
Check Monash email for a reply and a link: 'Please activate your SAS Profile to join SAS OnDemand for Academics'

Step 3

Follow the link to complete registration, use the same email, create a password, agree to terms and 'Create Account'.

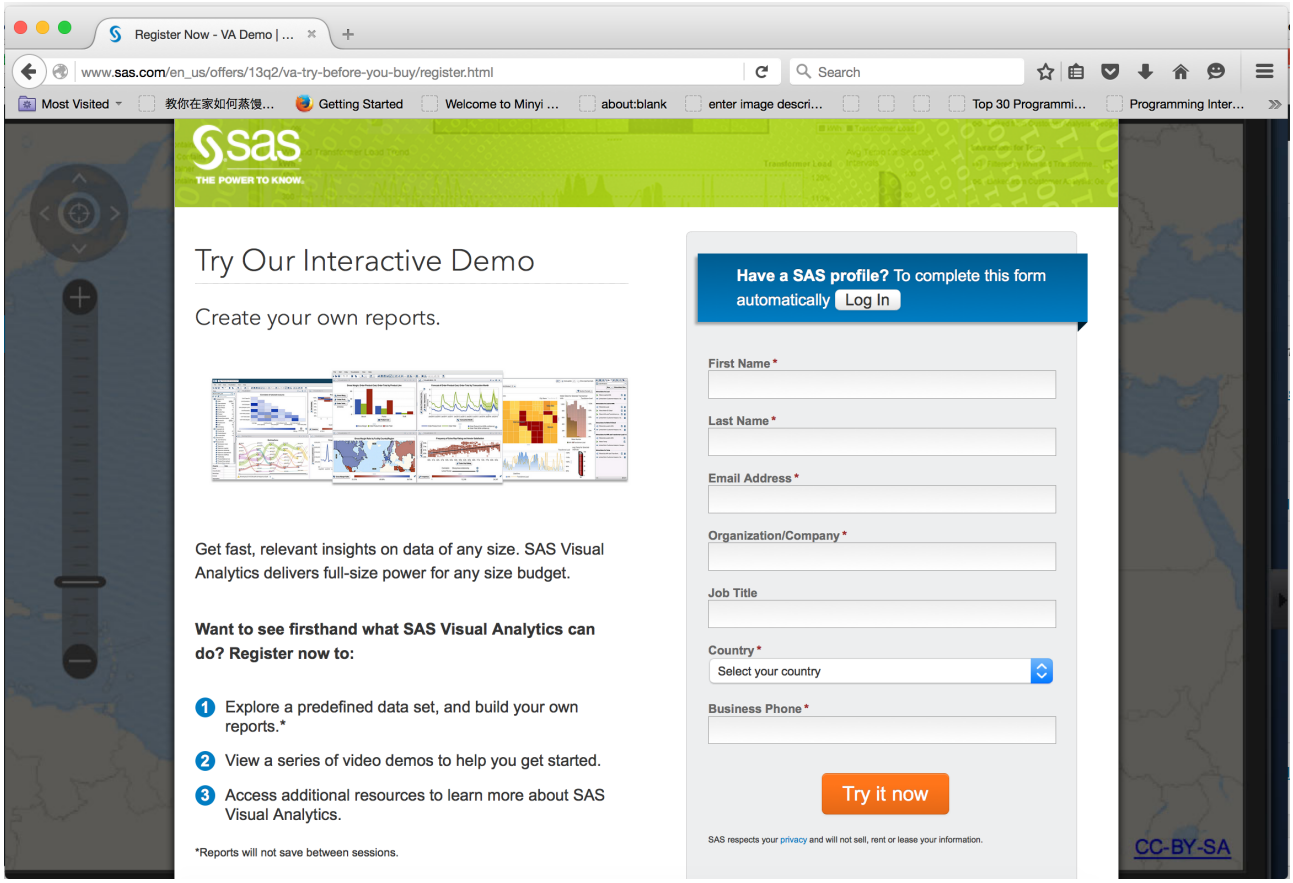
Step 4

Wait for user ID, copy and save, DON'T 'Sign in' here (see below). Keep a record of email used, userid (and password).



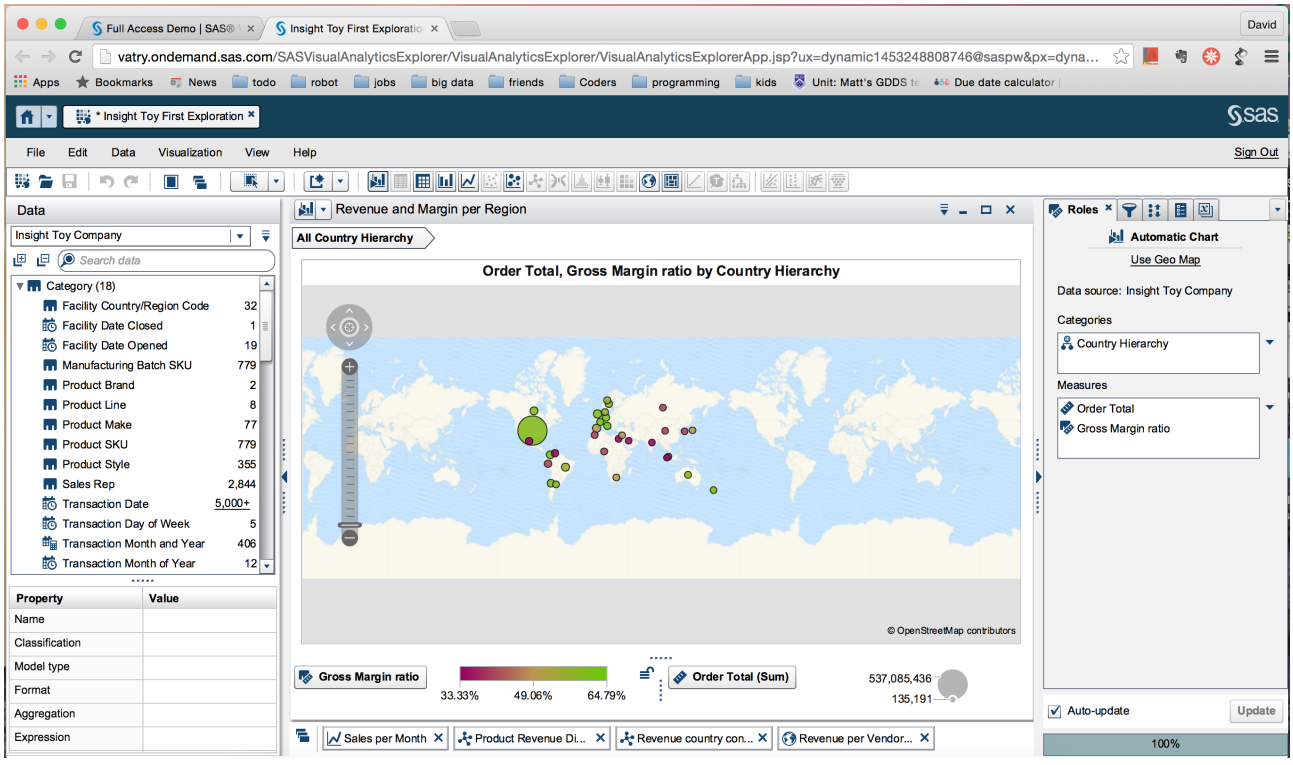
Step 5

To complete your SAS Profile go to: http://www.sas.com/en_us/software/business-intelligence/visual-analytics/demo.html and click 'Create your own' then 'Log in' (see below) to populate form (and PS, if you don't want to risk sales calls, use '0's for phone number).



Step 6

Click 'Try it now' and then 'Launch SAS Visual Analytics' (and wait for it to load):



You can also sign in to SAS Studio:

> https://odamid.oda.sas.com/SASStudio

The next time you use SAS Visual Analytics you can go directly to:

> http://www.sas.com/en_us/software/business-intelligence/visual-analytics/demo.html
>
> (you may like to keep the above two links alongside your registration details: username and password)

2
Data Models in Organisations

Introduction

This is our second module of six for the Introduction to Data Science unit. In this module we look at data from an organisational and business perspective: the value chain for data in an organisation, different business models and different applications, and data science projects. We will look at a wide variety of ways of modelling the organisational and business aspects of the problem.

(https://youtu.be/DqWwN-IjSbs)

Aims of this module

- Describe value chains or life cycles for data (in an organisation) in a general sense.
- Classify different business models for big data companies.
- Classify different data science projects within one organisation.
- Classify different broad business areas addressed by data science.
- Analyse a data science project with regards to the enabling factors that made the project work, and its business value.

How to study for this module

In this module we again draw on material in the public domain such as interviews and videos, online magazine entries and blogs. We also have also written some material to tie together various kinds of models. In addition to studying and viewing the material, you will also do activities using SAS and Python. Python is one of the two main languages covered in the course (the other is R). It is an important language for many general data science duties because of its powerful scripting capability. So skill with Python is an asset and worth your practice.

In this module you should continue with your first project "Data Science and Me," and start the second project "Business and Data Case Study."

Please note:

- Reference items marked with a single "johny look it up" icon,  , should be viewed as *suggested reading*, not essential nor important for assessment.
- Reference items marked with a two "johny look it up" icons,  , should be viewed as *important reading*, considered important for assessment.

2.1
Data and Decision Models

This section looks at a number of different models we can use when working with data and making decisions. These are not models in the statistical sense, like linear regression. They are conceptual models that help us understand our data and its organisational or decision making context.

Life cycles and value chains

There are several process models relevant to data science. In Module 1 we looked at the tasks performed in a data science project. There are two kinds of models used for this:

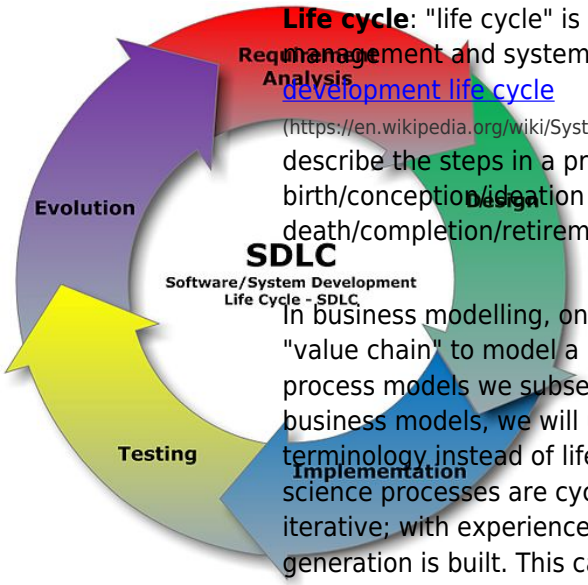


figure by Cliffydcw (Own work) CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0), via Wikimedia Commons

Life cycle: "life cycle" is term used in biology, project management and systems development (*e.g.*, systems development life cycle (https://en.wikipedia.org/wiki/Systems_development_life_cycle)) to describe the steps in a process, usually from birth/conception/ideation to death/completion/retirement.

In business modelling, one uses a similar term called "value chain" to model a process. Because the data process models we subsequently present are primarily business models, we will use the value chain terminology instead of life cycle. Many of the computer science processes are cycles because they are iterative; with experience a second and third generation is built. This can also happen in the Data Science world too. Nevertheless, we will use "value chain."

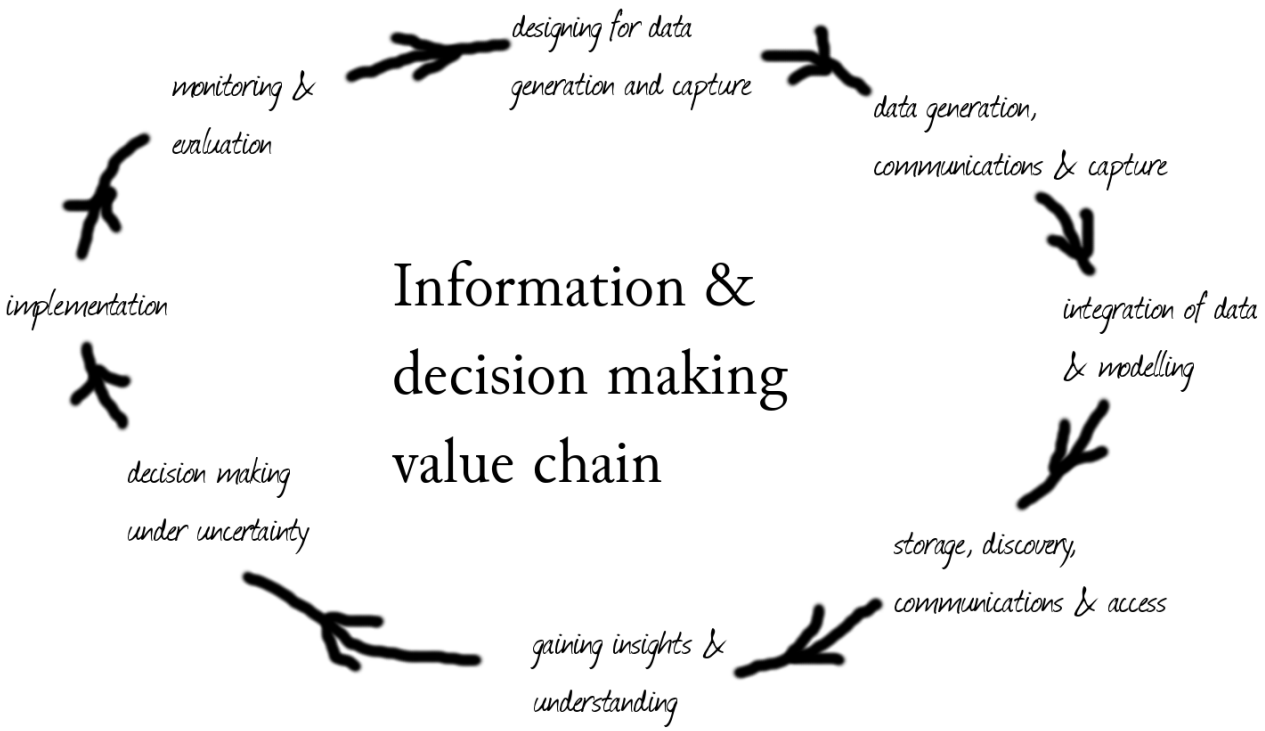
Value chain: "value chain (https://en.wikipedia.org/wiki/Value_chain)" is a business term used to describe the series of activities done to create an item of value. For instance, for the production of a technical product, the *engineering value chain* refers to a series of engineering activities and capabilities contributing to value creation, as shown below.

figure by Yufeng Zhang (Own work) CC BY-SA 3.0, via Wikimedia Commons

Process modelling of data in an organisation

A number of people have proposed value chains or life cycles for the use of data. There is no one right

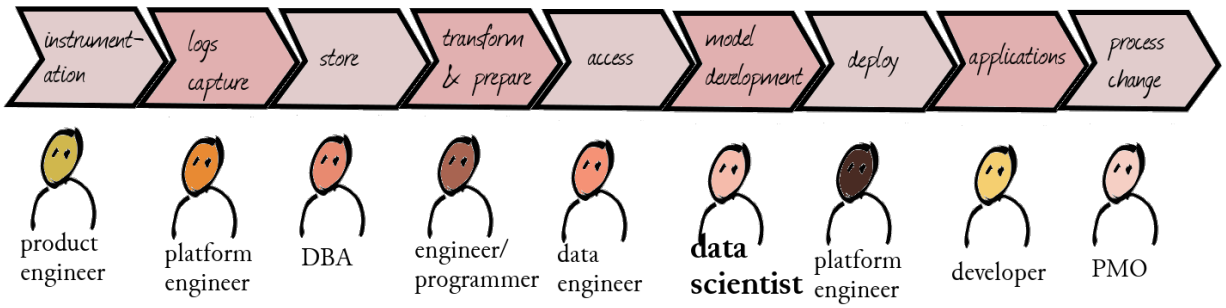
answer to which is the best. They are simply models. These generally build on the tasks we covered previously of what is done in a data science project.



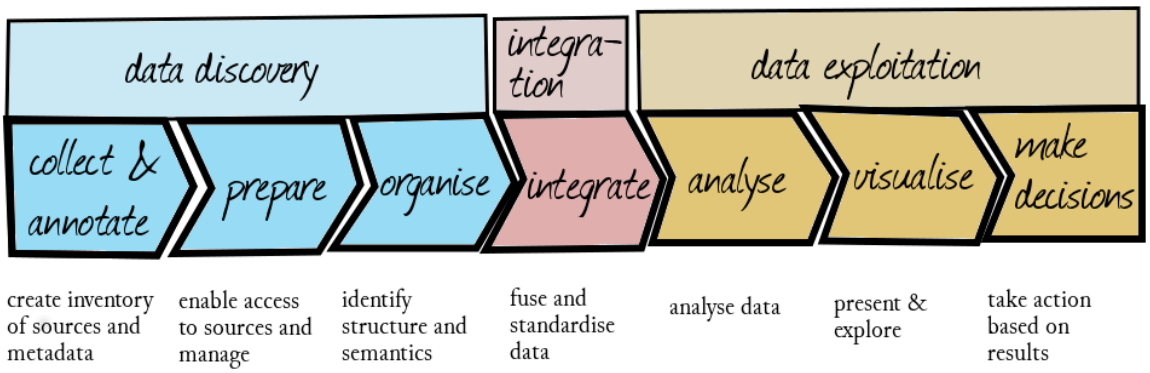
CSIRO's Data Value Chain

CSIRO, [in reviews of their organisation (see slide 15)](http://www.slideshare.net/elliotduff/austrade-v5), gives a value chain for data acquired by instrumentation that is used in monitoring a system. The system could be a field of wheat, and the instrumentation a swarm of sensors. Here instruments are developed and fielded, data is acquired and stored, analysis is done, then after analysis some changes/decisions are made to the system, finally monitoring and evaluation then informs the next cycle.

[Pivotal](http://pivotal.io), a big data company presents the following value chain for performance data from a large organisation (see slide 12 in their "[Data Science + Data Engineering](http://www.slideshare.net/emcacademics/pivotal-data-sciencedataengineeringsecretweaponsofthestrategicenterprise)"). In the figure below the roles are also added. Here we see the role of the data scientist as being very focused: they develop the model to be used (using data analysis).



Pivotal's Data Value Chain



Miller and Mork's Data Value Chain

Another value chain model is presented by Miller and Mork (2013):

- "From Data to Decisions: A Value Chain for Big Data," Miller, H.G. ; Mork, P., *IT Professional*, 15(1), 2013, (1100 words, 6 minutes) http://dx.doi.org/10.1109/MITP.2013.11 (available by search through Monash digital library or search for thePDF by article title directly on the web). 

Their Figure 1 is adapted below. They describe the various steps in some detail.The main difference with their value chain is the integration step in the middle that standardises the data in a common format. This may be required in large organisations were data may have been collected independently according to different schema in different parts of the organisation. More often, however, integration gets replaced by wrangling or cleaning where the disparate data is fused somehow prior to analysis. The integration was not planned for in the earlier pre-data-science world and is expensive to do thereafter.

For our purposes, we will use the **Standard Value Chain** that was first presented in section **Introduction to Data Science** of module **Data Science and Data in Society**. This is defined as follows:

> **Collection**: obtaining and collecting data from various sources, instruments or providers.
> **Engineering**: processing and storage, managing the databases, computers and hardware.
> **Governance**: all aspects of management of data such as security, metadata, etc.
> **Wrangling**: transforming and cleaning data prior to analysis.
> **Analysis**: analysis in its many forms.
> **Presentation**: visualisation and summarisation, to present the case for "value".
> **Operationalisation**: putting the results of analysis to work to obtain value.

One can see the rough correspondence between this, and those above. We will refer to the Standard Value Chain in later sections.

Analytic levels

Analytic levels are very general terms used in the business community. They were described in a report by SAS in the form of a simple two-page infographic. They are a simplified model of the levels of information provided to decision makers. The lower levels cover standard business intelligence reporting, whereas the higher levels reflect actual analytics. These are broad qualitative terms that are useful in high-level discussions and in discussions with management.

- "Eight Levels of Analytics" (https://www.sas.com/news/sascom/analytics_levels.pdf) by SAS (2 page infographic)

The top three levels in this model are:

- **Forecasting:** estimating the trend.
- **Predictive modelling:** more extensive modelling of future behaviour.
- **Optimisation:** given constraints and priorities, attempting to optimising future plans based on predictive models.

These levels are useful terms in general discussion as they cover broad classes of technology capability. A more common version of this is the descriptive-predictive-perscriptive levels of business analytics (https://en.wikipedia.org/wiki/Business_analytics):

- **Descriptive Analytics**: gain insight from historical data with reporting, scorecards, clustering, *etc*., answering questions like "how am I doing?"
- **Predictive analytics**: (https://en.wikipedia.org/wiki/Predictive_analytics) predictive modeling using statistical and machine learning techniques, text mining, *etc*., answering questions like "what will happen next year?"
- **Prescriptive analytics** **(https://en.wikipedia.org/wiki/Prescriptive_analytics)** recommend decisions using optimization, simulation, *etc*., answering questions like "what offerings should I make to maximise expected return next year?"

Influence diagrams

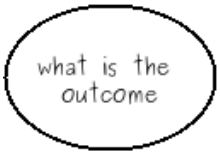
Influence Diagrams (https://en.wikipedia.org/wiki/Influence_diagram) are a more mathematically oriented approach to modelling decisions. In their full treatment one introduces probabilistic inference, decision theory, Bayesian reasoning and model elicitation. We will not cover this mathematics here. Rather we will simply introduce the graphical formalism. This stresses the actions, outcomes and uncertainties in a problem. Moreover, it is a good starting point for the causal and correlation models we will do in the later module **Data Analysis Process**. So influence diagrams are the starting point for a good decision model because they encourage you to think about:

- what are the knowns and unknowns?
- what actions can you take?
- what is the outcome, or how do you get value from your decisions?
- what sort of additional information should you obtain, and how costly is it?

Indeed, these are the key things that should be going through the decision makers mind.

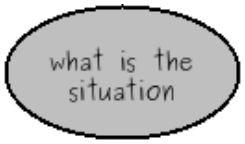
So influence diagrams have 4 kinds of nodes:

Chance variable: A ***chance variable*** is an uncertain quantity, whose value you do not know, but you should find out it's value in the future. Moreover, you cannot control it directly. It is represented as:



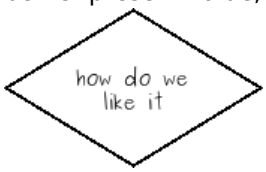
Known variable: A ***known variable*** has a value you do know. It might have formerly been a chance

variable. It is represented so:

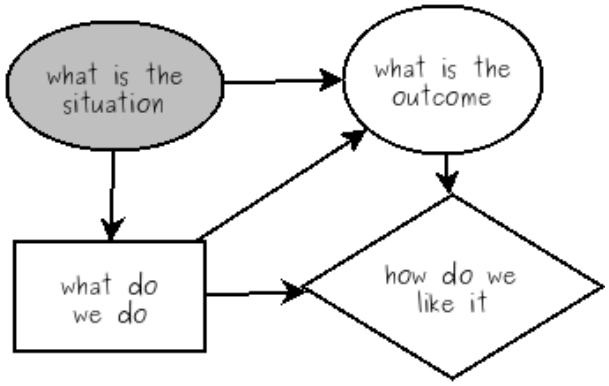


Decisions: A *decision* is a variable that you as the decision maker have the power to modify directly. It could be whether to invest in a new project, how much to invest, whether to go or not, *etc.* It is represented:

Objectives: An ***objective*** is a measure of the value of your satisfaction with possible outcomes. It might be net present value, lives saved, dollars, or some other measure of worth. It is represented:



These sorts of nodes are placed together in a directed graph, which is then called an influence diagram. The standard influence diagram has no fully directed cycles, though dynamic variations allow cycles forward in time. See the prototypical *Simple Influence Diagram*. The arcs indicate "can influence" and for outcome and value nodes, "can cause".



The Simple Influence Diagram

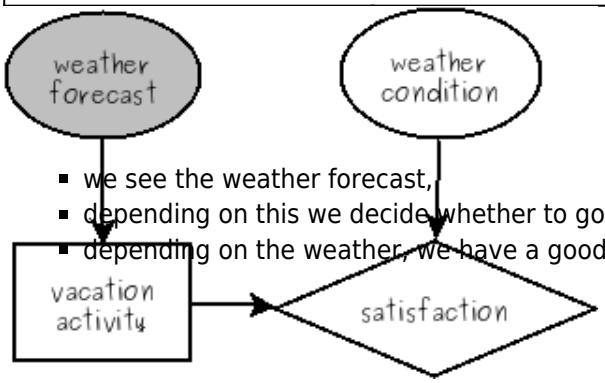
So this influence diagram says:

- we assess the situation ... it is known,
- we take an action,
- we then observe an outcome (after our action), which depends on the situation and our action,
- then we obtain some value from the outcome, which may also depend on the action.

Moreover, the structure of influences represented by the arcs in the graph is also important: which knowns and unknowns influence what. When do we connect an arc between two nodes. Basically, one uses the idea of "cause". The node at the source of an arc is a "cause" for the node at the destination. The following table gives the details of when to connect an arc to a particular kind of node.

chance variable	known variable	decision	objective

Connect from node A to the chance variable, if, holding all other nodes fixed in value, changing the value of A (perhaps by force) would influence the chance variable.	None essentially, but to be informative you can connect arcs to it as if it was unknown currently.	Which other nodes, whose values must be known at the time of the decision, are looked at to make the decision?	Which other nodes, whose values must be known at the time the objective is computed, are looked at to compute the value?

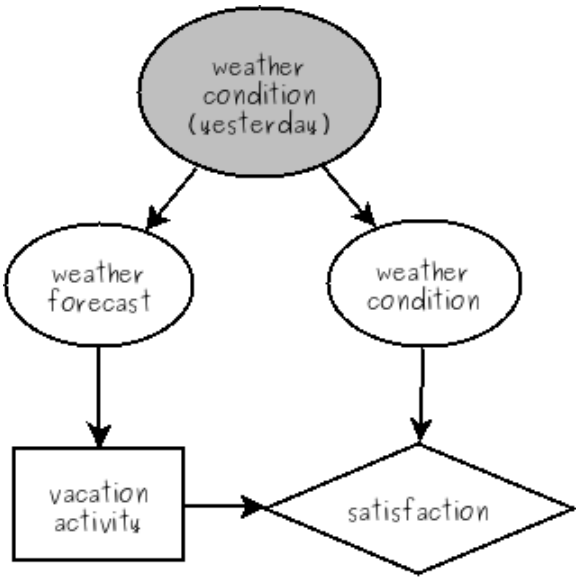


Last Minute Vacation Scenario

So a simple scenario, which is a variation on the above, is in the influence diagram given as the *Last Minute Vacation Scenario*. It works as follows:

- we see the weather forecast,
- depending on this we decide whether to go to the beach for the weekend, or not,
- depending on the weather, we have a good or bad (*i.e.*, wet/cold) time at the beach.

Now, you can question this structure. That is what good modelling is about, questioning and checking assumptions, trying to get the significant decisions, objectives and variables into the scenario. You may say:



Modified Last Minute Vacation Scenario

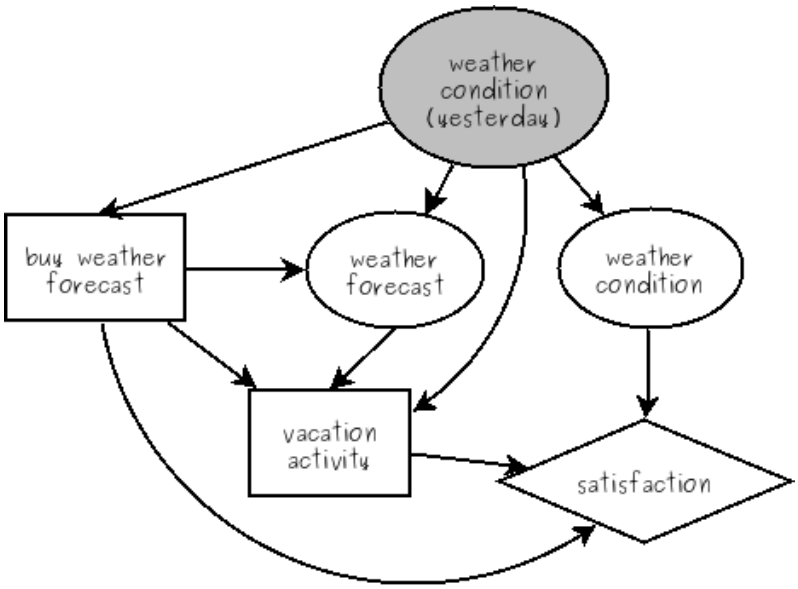
> Shouldn't the actual weather be somehow related to the weather forecast. The Bureau of Meteorology don't make random forecasts.

Indeed this is true. However, the thing to note is that the correlation between the weather forecast and today's weather is because they are both influenced by yesterday's weather. Atmospheric physics is such that today's weather will have some relationship to yesterday's. Moreover, the Bureau of Meteorology bases their weather forecast for today on yesterday's weather too.

So to model this situation, we could add yesterday's weather to the diagram, as shown in the *Modified Last Minute Vacation Scenario*. Here we know yesterday's weather but still have not received the forecast for today.

Now, additionally, let us suppose that weather forecasts were expensive. Suppose, for instance, they were not broadcast on free television and radio. Rather, we had to pay maybe $10 or $100 to obtain a forecast.

Then we have an extra decision to make: shall we buy a weather forecast before deciding whether to take the vacation?



Last Minute Vacation Scenario Buying Forecasts

This situation is represented in the *Last Minute Vacation Scenario Buying Forecasts*. We have two decisions to make, the second being influenced by the outcome of the first. We could choose not to buy a forecast and use our own hunches, basing the vacation decision on our knowledge of yesterday's weather. However, if we need a long term forecast (say the next 4 days), then our hunch could be poor, and a $800 long weekend at a remote beach could be ruined, so we may be willing to pay even $50 for the forecast. Another scenario is that we do buy the forecast, it predicts good weather, we go on the vacation, but the weather is terrible. This represents the worst case scenario: we pay money for the forecast but we have a bad vacation.

So you can see all these decisions also depend on:

- how reliable the weather forecaster is,
- how good our own hunches are,
- how bad it would be to miss out on the vacation when a bad weather prediction did not eventuate,
- how bad it would be when a good weather prediction failed and you had bad whether on the vacation.

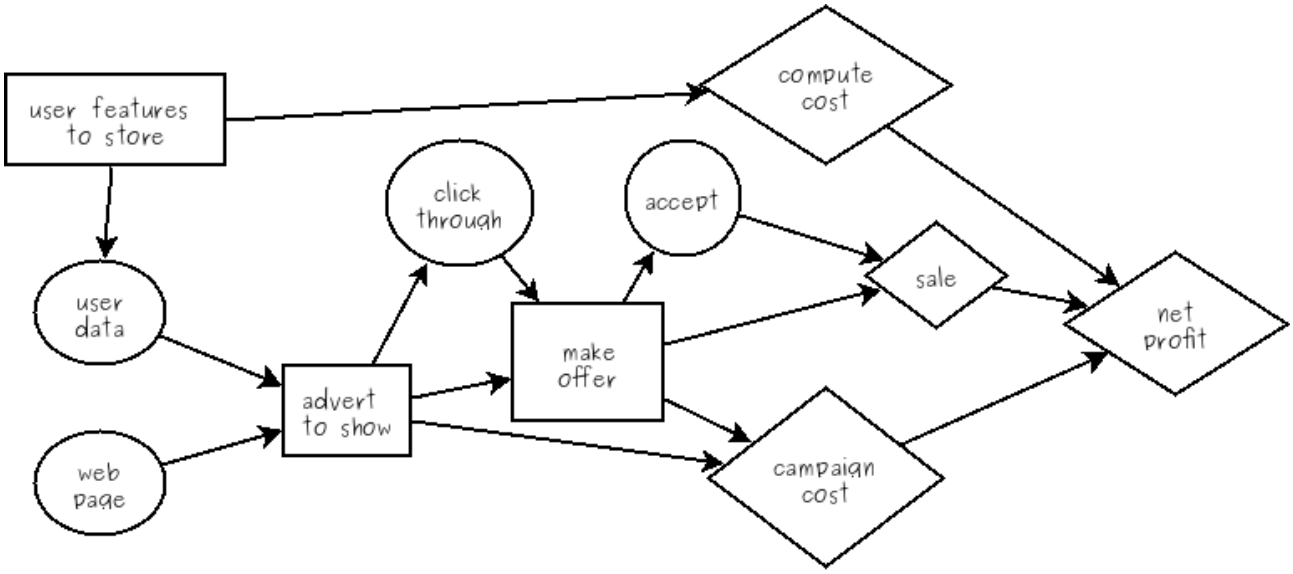
Building a model

An introductory lecture by Dr. John Hanson gives us a complementary view of influence diagrams and how they can be constructed.

- "Influence Diagrams (https://www.youtube.com/watch?v=ExVhiEaKD1o)" as part of the DSCI 300 course taught at UCSD (Youtube, 13 mins)

Example: online advertising model

Lets look at a more relevant and realistic example, online advertising (https://en.wikipedia.org/wiki/Online_advertising). This is an area with enormous variety, with high performance computational infrastructure handling a campaign at high speed. While the engineering aspects of this are an entire subject in themselves, a high level view of the decision choices, for a particular simplified variant, is given in the model below.



Internet Advertising Model

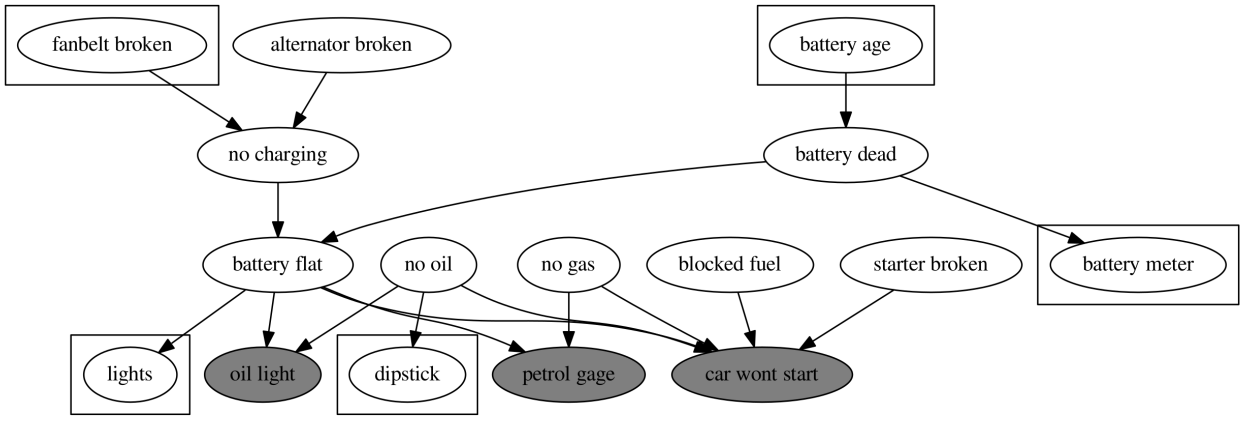
For well targeted online advertising, one has three decisions to make:

Which features about the user to store: the web pages they have visited historically can be characterised simply and stored (rather than full details of the web pages themselves); store too much it becomes expensive and slow, store too little there is inadequate differentiation. So a short number of features are to be maintained for each user.
What advert to show them this time: given the user features, what is a good advert to show. In practice, this decision is programmed to work at internet response times.
If they click through, what particular offer to make: so the advert got their interest; what will we offer them.

The profit derived from the advertising is then derived from three objective variables: the cost of tracking the user and maintaining a feature set for them ("compute cost"), the particular advertising campaign cost, and the net from the sale. These combine to give the "net profit".

Example: diagnosing your car

So you are going to work in the morning and you car wont start. A simplified view of the situation before you is given in the *Car Wont Start* diagram. This is a simple example, but it is characteristic of many problems in engineering and medicine.

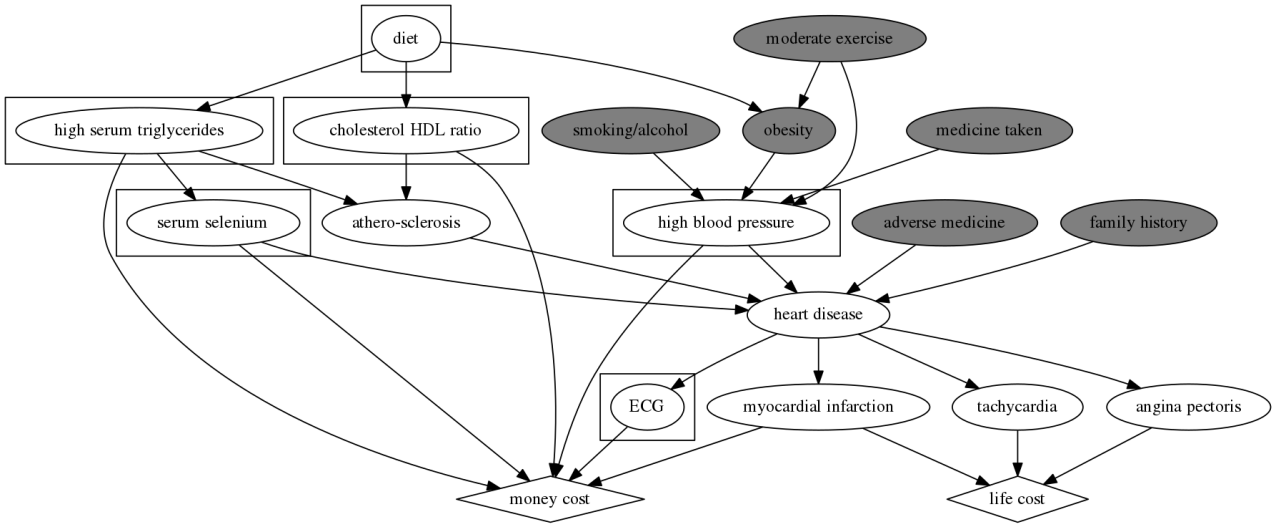


Car Wont Start

You are sitting in the drivers seat and you immediately look to see what the petrol gage reads and whether the oil light is on. So you have assessed the three grey variables in the model and you know their value. With that in mind, you can choose to get out of the car and examine some other variables, to start diagnosing the fault. The boxed ovals in the model combine a decision and a chance variable: where you can decide to test/determine the enclosed variable. You can fetch your battery meter and measure the battery. You can turn the lights on and get out and check if they are OK. You can open the hood and check the oil level with the dipstick. In each case you decide whether to measure the variable and then see its effect on your diagnostics. Each test has a cost: testing the lights is simple, especially at night time; testing with the battery meter is harder (you need to find it in the garage). If you are remotely familiar with how a car works, you should be able to figure out how each variable effects the others. From this graph, one can imagine designing an optimal strategy for figuring out why your car would not start.

Example: diagnosing heart disease

Now the *Car Wont Start* diagram is really a standard alone inference task, without data. Let us look at a more realistic task, diagnosing heart disease, shown in the figure *Heart Disease*.



Heart Disease, adapted from [Bayesian Approach in Medicine and Health Management](http://www.intechopen.com/books/current-topics-in-public-health/bayesian-approach-in-medicine-and-health-management) (http://www.intechopen.com/books/current-topics-in-public-health/bayesian-approach-in-medicine-and-health-management), see Figure 4.

While for the *Car Wont Start* model we (most drivers at least) understand the relationships quite well, but for the Heart Disease model, much less so. In fact, while a medical specialist could probably tell you the directions of influence ("obesity increases blood pressure", *etc.*), they could not turn this model into a set of probability rules or tables so we can precisely quantify how each variable affects the other. But they could possibly provide us with a data set and we can use statistical machine learning tools to develop such a model.

The grey variables could be obtained by the physician with a simple interview with the patient. Note again the boxed ovals that double as decision nodes and variables. The ECG, a diagnostic measurement, may cost $300 to take. Similarly for measuring serum selenium. In this model we make the diet variable a decision/variable combination: perhaps a dietitian is asked in to do a proper survey. A myocardial infarction is a heart attack, and depending in the severity can be bad to fatal, but is very costly regardless.

Let us suppose we were able to developed a model using data. From here you can understand the difference between predictive and prescriptive analytics. Predictive analytics would answer the question, given a particular patient's conditions and tests available, what is the probability of a sever myocardial infarction. Prescriptive analytics is far more interesting, and would be suitable as a "what if" tool for a patient. This can answer questions like:

- What combinations of lifestyle changes could I make to lower my chance of a myocardial infarction by 50%?
- I hate the side effects of the blood pressure medicine. Suppose I stopped taking the medicine, what combinations of lifestyle changes could I make so that my chance of a myocardial infarction remains unchanged?
- Which combinations of tests should I have taken so I properly understand the current state of my heart health, assuming I have a fixed budget available (according to my health management plan)?

Summary

The mathematical theory of influence diagrams is designed to help you put the various probabilities and costs together. Using influence diagrams is more a state of mind than the tendency to start drawing little diagrams or to start solving probabilistic equations. What it teaches us is that it is important to recognise the key variables in a problem:

- different objectives that yield value,
- different decisions or actions that can be made, and their influence on other variables,
- important knowns and unknowns that influence objectives or decisions,
- different decisions that can be made which influence what we know, thus improving the quality of our subsequent decisions.

Resources

Here are two useful resources for further studying influence diagrams. Be warned, each school and company using these models changes the notation slightly, but the underlying ideas are the same. The two are as follows:

- Introductory slides "Influence Diagrams (http://slideplayer.com/slide/4626155)" by Dr Yan Liu of Wright State University. Slides numbered 13-29 introduce the model by means of simple examples.
- The Wikipedia page Influence Diagrams (https://en.wikipedia.org/wiki/Influence_diagram) gives a fairly complete treatment.

2.2 Activity: Getting Started with Python notebooks

iPython notebooks (.ipynb)

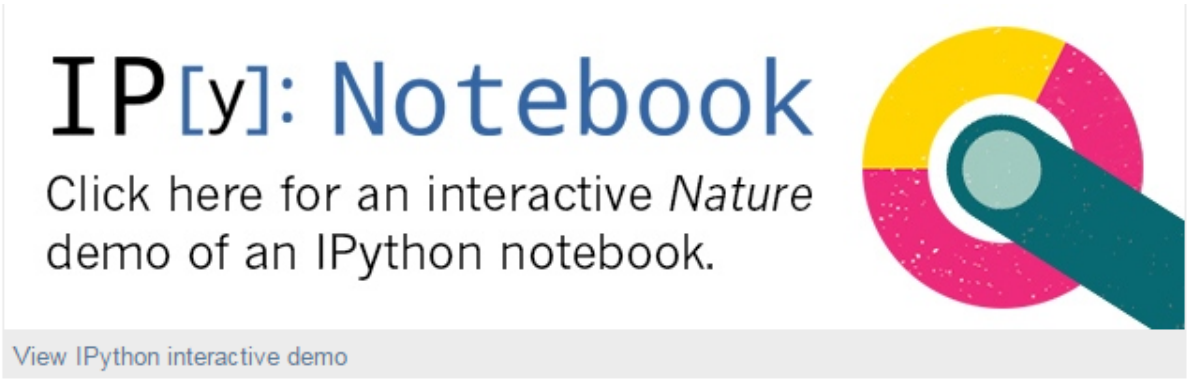
These provide an interface to the iPython interpreter through a web browser. An iPython notebook also allows you to interweave formatted text (including mathematical equations) and graphics with executable and editable Python code. Notebooks provide an ideal platform for writing documents that not only explain concepts but also allow the reader to interact with the calculations (and even modify the code). This is invaluable both for pedagogical purposes and for recording, duplicating, and communicating research results. On with the show...

Step 1

See this article in Nature, it is an introduction to iPython notebook that includes links to examples. There is also a notebook embedded in the site:
http://www.nature.com/news/interactive-notebooks-sharing-the-code-1.16261

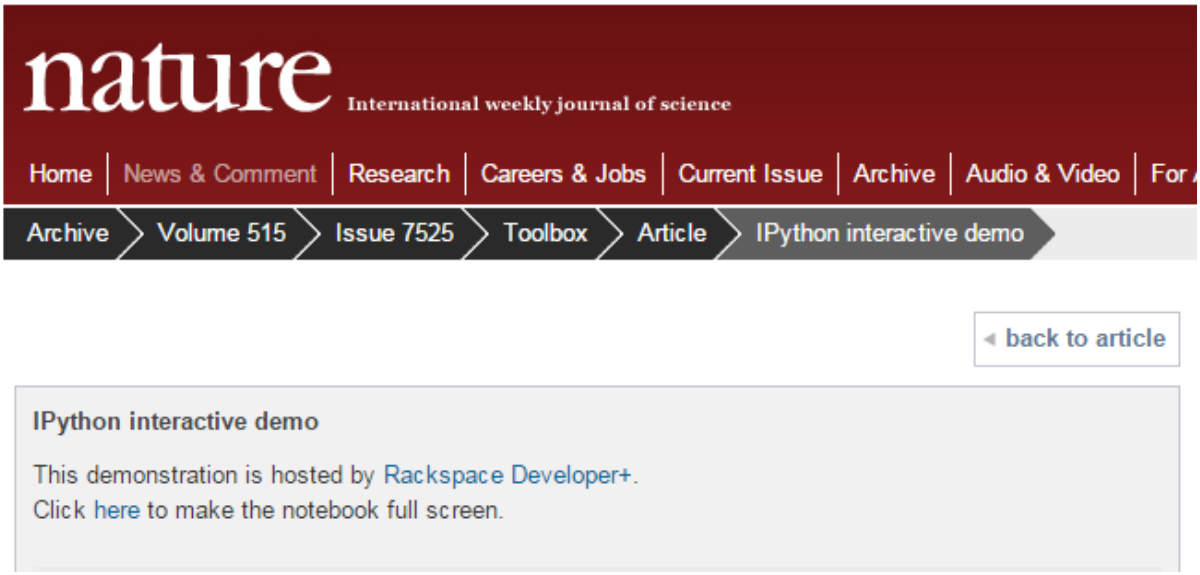
Step 2

Now follow the link to the live demo, as shown below or use the URL:
http://www.nature.com/news/ipython-interactive-demo-7.21492?article=1.16261



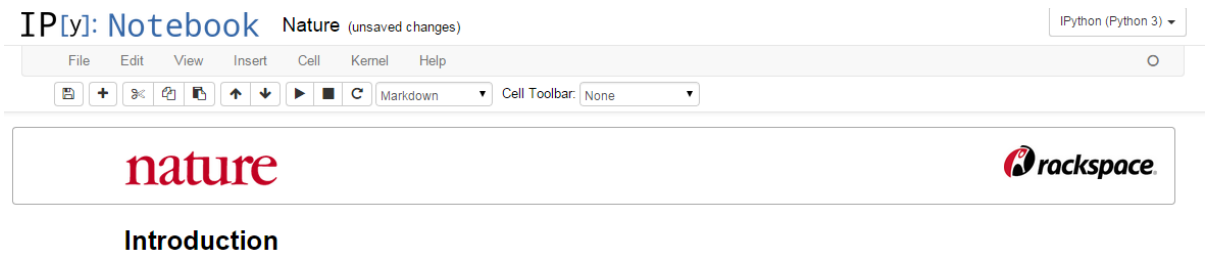
Step 3

Expand to full screen ('Click here' below):



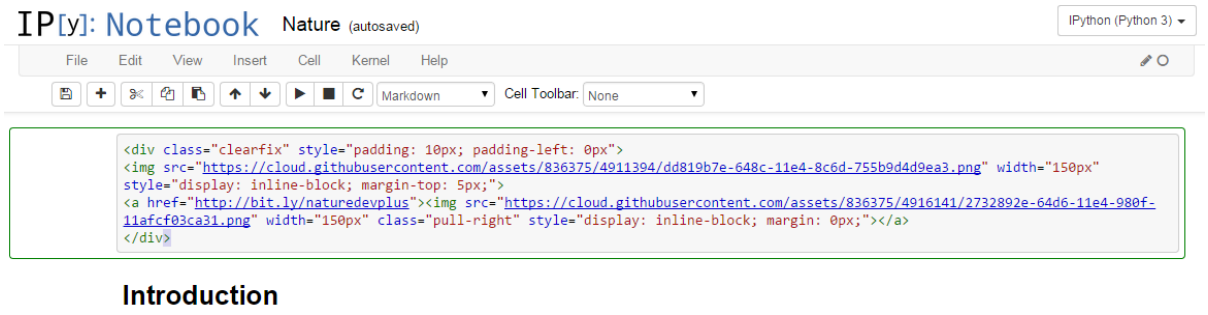
Step 4

And this is an ipython Notebook (.ipynb) online:



Step 5

You can see the first section or cell, '**nature**' (above), is highlighted (has a border), if you double click you can see what's going on in the background (it's HTML):



Step 6

Use Shift-Enter to 'run' the cell and return it to it's previous state.

Step 7

The next 4 cells are similar, text/html. Use Shift-Enter to navigate down to the 6th cell, which is python

code:

A full tutorial for using the notebook interface is available here.

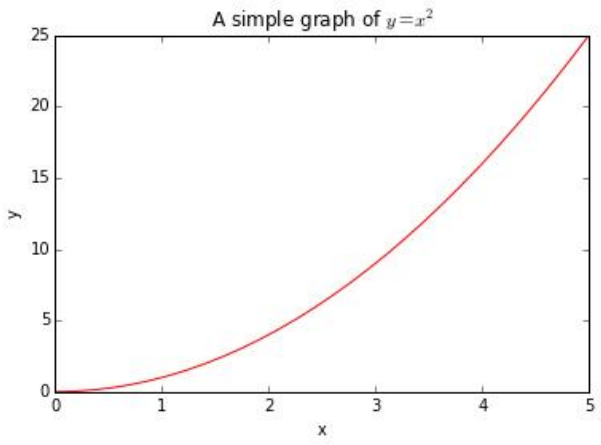
In [ ]:
```
# Import matplotlib (plotting) and numpy (numerical arrays).
# This enables their use in the Notebook.
%matplotlib inline
import matplotlib.pyplot as plt
import numpy as np

# Create an array of 30 values for x equally spaced from 0 to 5.
x = np.linspace(0, 5, 30)
y = x**2

# Plot y versus x
fig, ax = plt.subplots(nrows=1, ncols=1)
ax.plot(x, y, color='red')
ax.set_xlabel('x')
ax.set_ylabel('y')
ax.set_title('A simple graph of $y=x^2$');
```

Above, you should see a plot of $y = x^2$.

What is numpy?
What is %matplotlib inline?

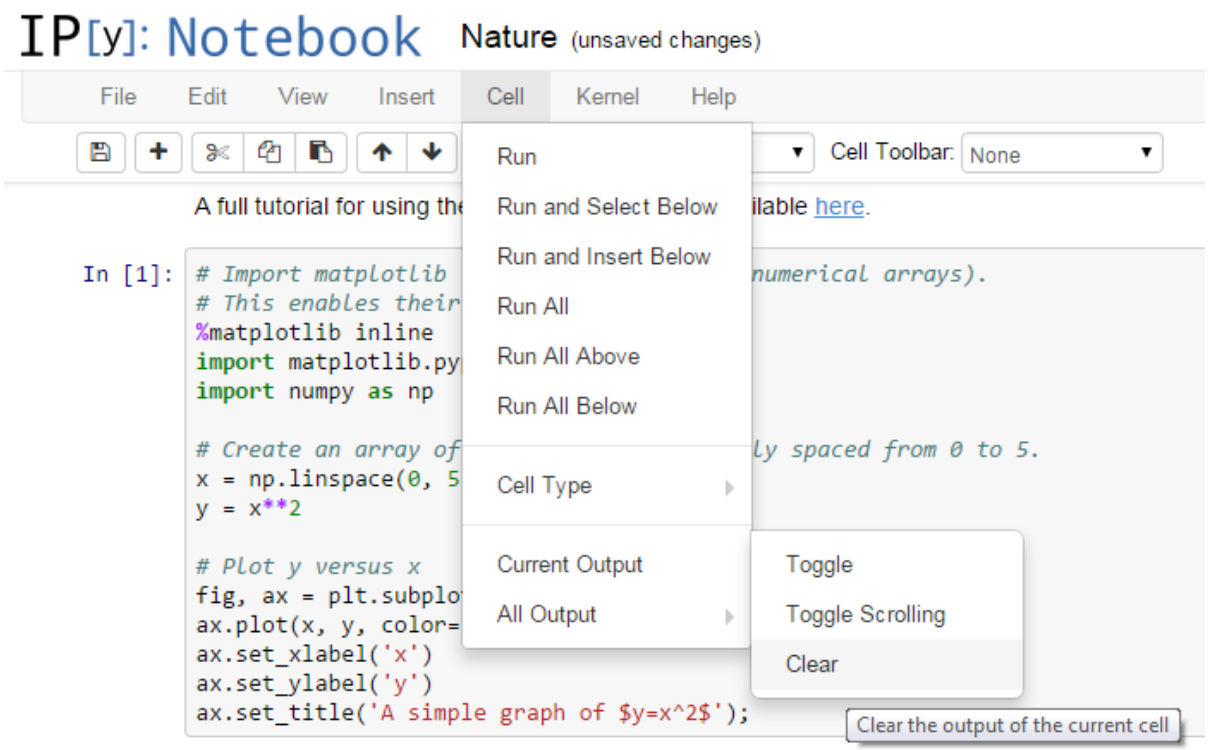
Step 8

Now Shift-Enter to run the code and see the output:



Step 9

You can make the plot bigger or smaller, save (as img), or clear. Do the latter by using the menu: Cell > Current Output > Clear (below, you can also 'Clear All' cells, a good idea when you reload a nb). Now run again.



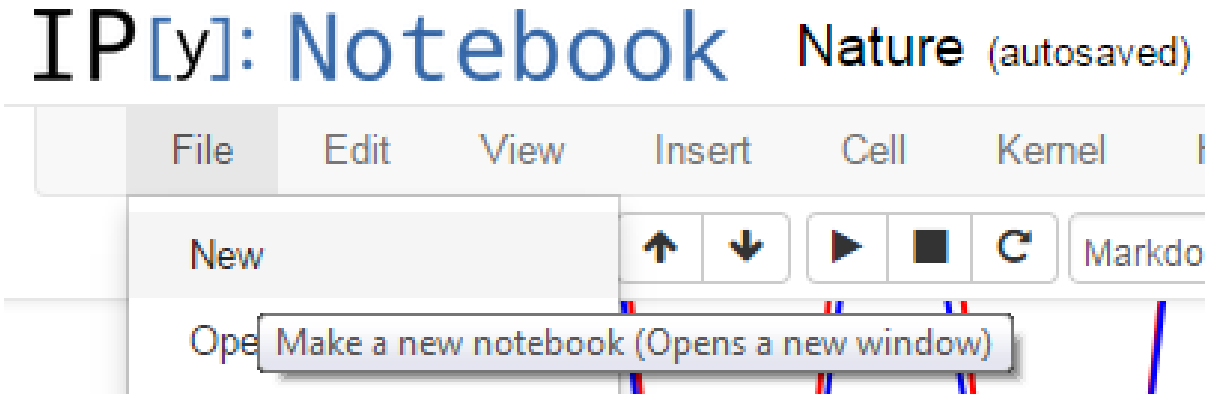
Step 10

The 'In [1]:' (above left) indicates the first cell has run. If there were output it would be Out [1], or in this case a graph. The 'In [*]:' (below) indicates processing.

```
In [*]: # Import matplotlib (plotting), skimage (image processing) and interact (user interfaces)
        # This enables their use in the Notebook.
        %matplotlib inline
        from matplotlib import pyplot as plt
```

Step 11

Continue to run the other code cells. Note also that plots 2 and 3 (Aliasing & Galaxies) are interactive. Now (thanks to Nature) you can make your own Notebook. Use File > New



Step 12

Copy the following code into the first cell, Shift-Enter to run:

```
import pandas as pd
df = pd.DataFrame({'Name' : ['Mike Hussey','Aaron Finch', 'Brad Hogg' , 'Steve
Smith' , 'George Bailey', 'Mitchell Johnson', 'Shaun Marsh',  'Glenn
Maxwell','Pat Cummins' , 'Mitchell Starc', 'David Warner'],
'Age' : [39,28,44,25,32,33,31,26,22,25,28],
'IPLSal' :[310,662,103,828,672,1340,455,1240,207,1030,1140]})

df
```

The output, colour coded, is below, we have imported a library, created a data structure (df), and filled it with some cricket player data, and after that, displayed the data

```
In [1]: import pandas as pd

        df = pd.DataFrame({'Name' : ['Mike Hussey','Aaron Finch', 'Brad Hogg' , 'Steve Smith' ,
                           'George Bailey', 'Mitchell Johnson', 'Shaun Marsh',  'Glenn Maxwell',
                           'Pat Cummins' , 'Mitchell Starc', 'David Warner'],
                           'Age' : [39,28,44,25,32,33,31,26,22,25,28],
                           'IPLSal' :[310,662,103,828,672,1340,455,1240,207,1030,1140]}) # in $1,000s

        df
```

Out[1]:

	Age	IPLSal	Name
0	39	310	Mike Hussey
1	28	662	Aaron Finch
2	44	103	Brad Hogg

What is pandas?
What is a DataFrame?

Step 13

First thing to do with data... look at it, so make a plot. Copy the following into the 2nd cell and run:

```
import matplotlib.pylab as plt
%matplotlib inline
plt.scatter(df['Age'], df['IPLSal'])
# and plot to see data
plt.show()
```

What is %matplotlib inline?
We now have 3 views of the same dataset, the raw data embedded in the code (or in a file), the dataframe (fairly similar), and the plot. What information do you gain/lose in these different views?

Step 14

We can see a general trend, from left to right, which you could probably draw a line through to show an estimate (linear regression).
Or we can use Python:

```
from scipy.stats import linregress
slope, intercept, r_value, p_value, std_err = linregress(df['Age'],df['IPLSal'])
# Here's our function:
line = [slope*xi + intercept for xi in df['Age']]
# plot it up
plt.plot(df['Age'],line,'r-', linewidth=3)
plt.scatter(df['Age'], df['IPLSal'])
plt.show()
```

In the above code what is 'slope'?
What is 'intercept'?

Step 15

Use python to display slope and intercept:

```
In [12]: slope
Out[12]: -27.178343949044589

In [13]: intercept
Out[13]: 1548.8535031847136
```

Why is slope negative, why 27? What does this mean in $ terms?
Now use them in an equation (y = ax + b) to test.

Step 16

We can see from the data that George Bailey is 32 and got $672K. Create a new cell and use python to calculate y = ax + b

```
In [16]: y = slope*32 + intercept
         y
Out[16]: 679.14649681528681
```

What does this result tell us (compare it to the original data), why choose Bailey?
Try a few more values (e.g. 20, 40, 50... 10).

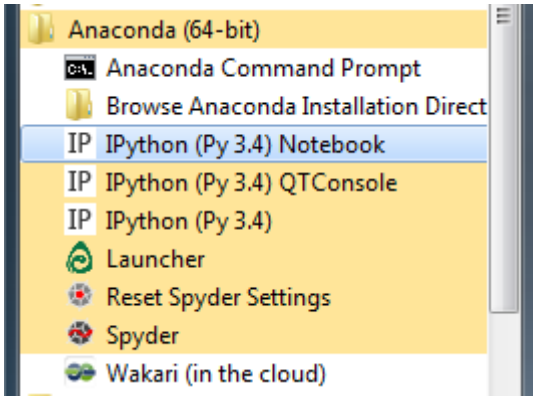
Step 17

Other cells... apart from code you can 'run' html, LaTex etc. and to save your notebook you can use File > Download as > .ipynb

Step 18

To run your saved notebook (or create some new ones) there are several options:

- Use Nature, but this is a temporary service, plus notebooks get deleted anyway
- Use a similar online service such as Wakari or SageMath, the latter works well (and can also run R code)
- Use standard Python (www.python.org (http://www.python.org)) and install the packages you need
- Use a prepackaged Python (e.g. Canopy or Anaconda), the latter is good because it can run .ipynb 'out of the box' (below), see: http://continuum.io/downloads#py34.
  Run (or install then run) Anacondas 'IPython (Py 3.4) Notebook':



2.3
Business Models with Data

This section looks at the business models used in the data science world. This is best viewed with reference to the value chains presented in the previous section. The simplest kind of busineses provides tools for a particular step in the value chain. Other businesses act as data producers, consumers or resellers. These businesses are fairly tradition businesses, excepting that the tools or services provided happen to be for software or data, rather than for donuts or lawn mowers.

A more recent phenomona, however, are the businesses that are uniquely based on their data. We refer to these as *data-based businesses*. They have spurred people to think afresh as to what sorts of businesses one can build out of data. In some sense these do more than just sell data. At the end of this section we discuss these data-based businesses.

Value chains for big data

We have previously looked at value chains in Data Science, in section **Data and Decision Models** of the current module. To get a better understanding of business models in the data science world, it pays to step back. We can do this by looking at the broader big data area which includes the hardware and systems community as well for the data engineering task.

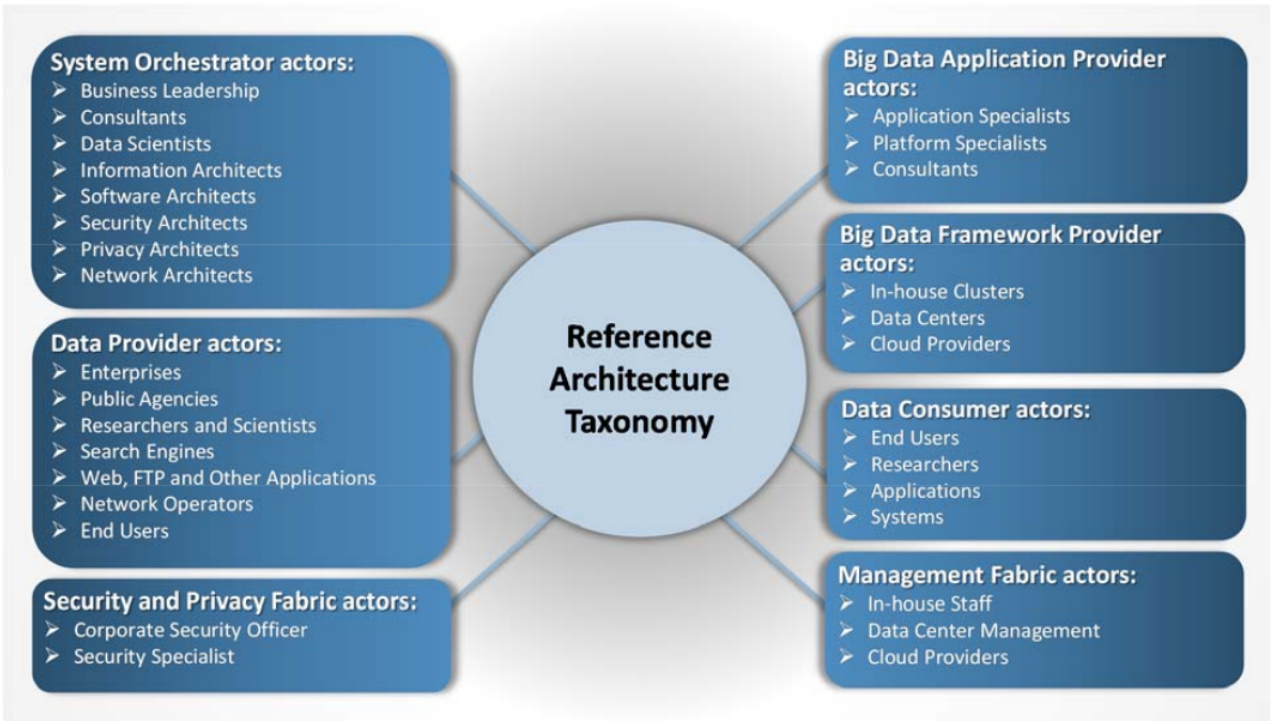
NIST's Big Data Working Group has developed a functional model that distinguishes the major actors: data providers, application providers (who handle the data value chain), data consumers, and framework providers (who provide the computing infrastructure).

NIST Reference Architecture

The reference architecture is presented in:

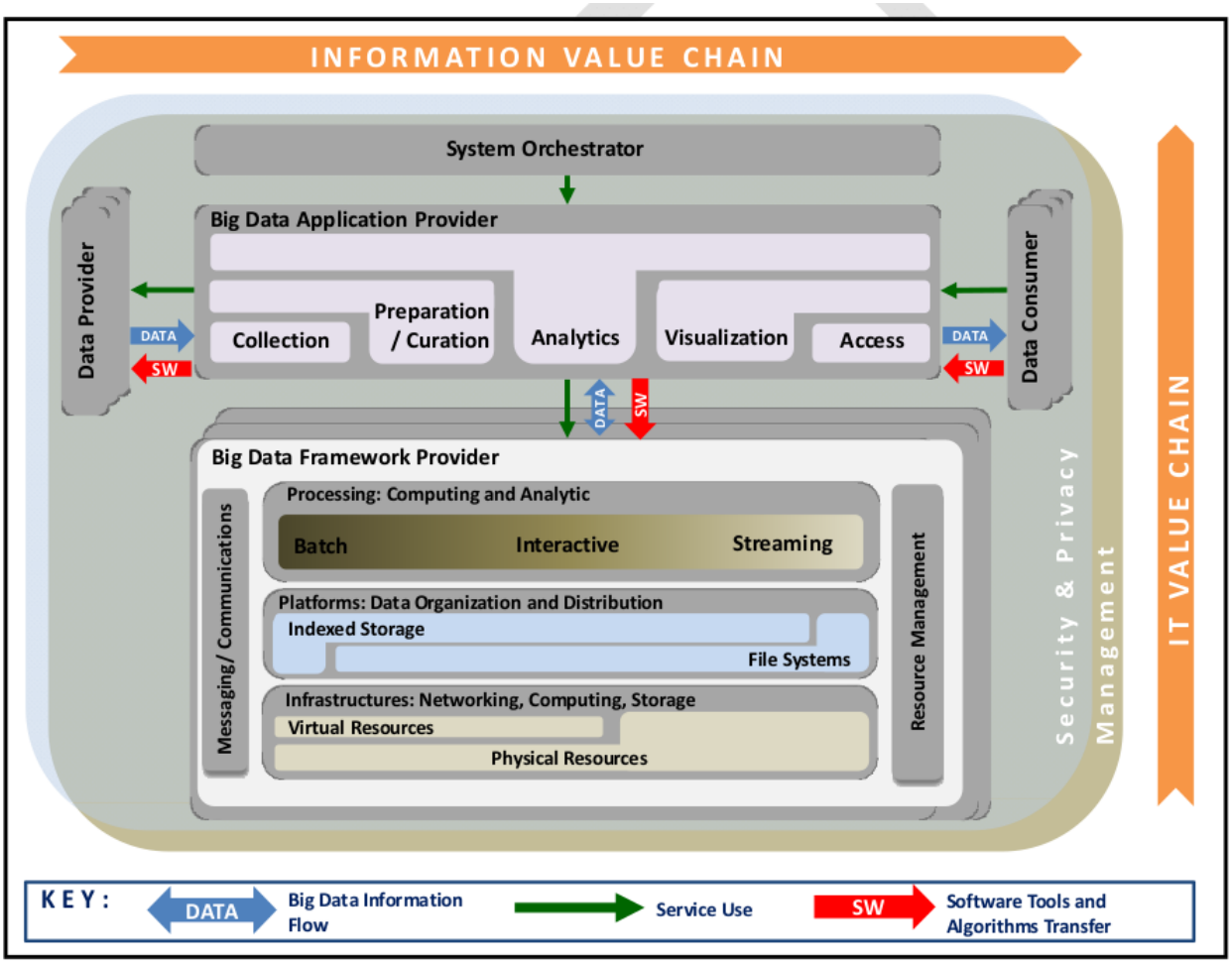
- "NIST Big Data Interoperability Framework: Volume 6, Reference Architecture" (http://dx.doi.org/10.6028/NIST.SP.1500-6) see sections 2.3, 3, 4.1, 4.2, 4.3, 4.5 (17 pages in all)  . This covers things in way more detail than we need for this unit. Note 4.4 discusses the "Big Data Framework Provider" which/who provides the computational capability, and we will discuss this further in the **Data Types and Storage** module.

Their model presents a richer information value chain (see Figure 2 on page 11) that clearly breaks up the data engineering components from the main Data Science ones, and it uses a variation of our Standard Value Chain: collection, preparation/curation, analytics, presentation and access.
Their model also has a different take on the landscape, which they call the *Reference Architecture Taxonomy*, and is centred around the different roles played in a project. So look at the Taxonomy in the figure and understand the different actors. This describes all the diifferent roles that can apply in the broader view of a big data project.



NIST Reference Architecture Taxonomy (from Fig. 1 of "Volume 6: Reference Architecture")

The relationship between these different actors is summed up in *Big Data Reference Architecture*. So the data value chain starts with the data sources, and through wrangling and analytics produces data products for the data consumer, and runs left to right in the figure. The bottom half of the figure is the computational or big-data processing side. In the figure they also introduce the *IT value chain* which runs vertically in the picture. The IT value chain starts at the bottom with hardware, storage systems and software infrastructure, and finally builds up to the data science part running along the top.



NIST Big Data Reference Architecture (from Fig 2 of "Volume 6: Reference Architecture")

So this reference architecture provides us with a perspective from which to consider the business/organisational roles that exist in a larger project.

The big data landscape

For further insights into the business world of Data Science, a number of bloggers and authors, primarily in the venture funding area, have laid out their understanding of the field in a sequence of graphics called "landscapes" or "ecosystems". Note interacting systems of software are often referred to as ecosystems (though, sometimes also tool-chains). We will explore these because they lay out the different businesses and providers that exist in the Data Science world. Its important to understand these with respect to the Reference Architecture from NIST above, as the reference architecture lays out the different roles required. Each clearly requires difference kinds of support. The support for a project can take different forms:

- software as a service,
- "rent a cloud" facilities to provide the big data computing facilities,
- specific software or tool chains, and
- data as a service.

Now, below we mention 5 different "landscapes" or "ecosystems." None of these is the "true way" or the best in any sense. We cover them just to offer some alternative views. We should not attempt to over analyse these, and we will not attempt to develop a consensus model.

The "big data landscapes" exists in three versions, one each year 2012-2014, and have been produced by different individuals, the first by Dave Feinleib on Forbes (http://www.forbes.com/sites/davefeinleib/2012/06/19/the-big-data-landscape/). Here we reference the Pinterest.com (http://www.pinterest.com) versions as three infographics:

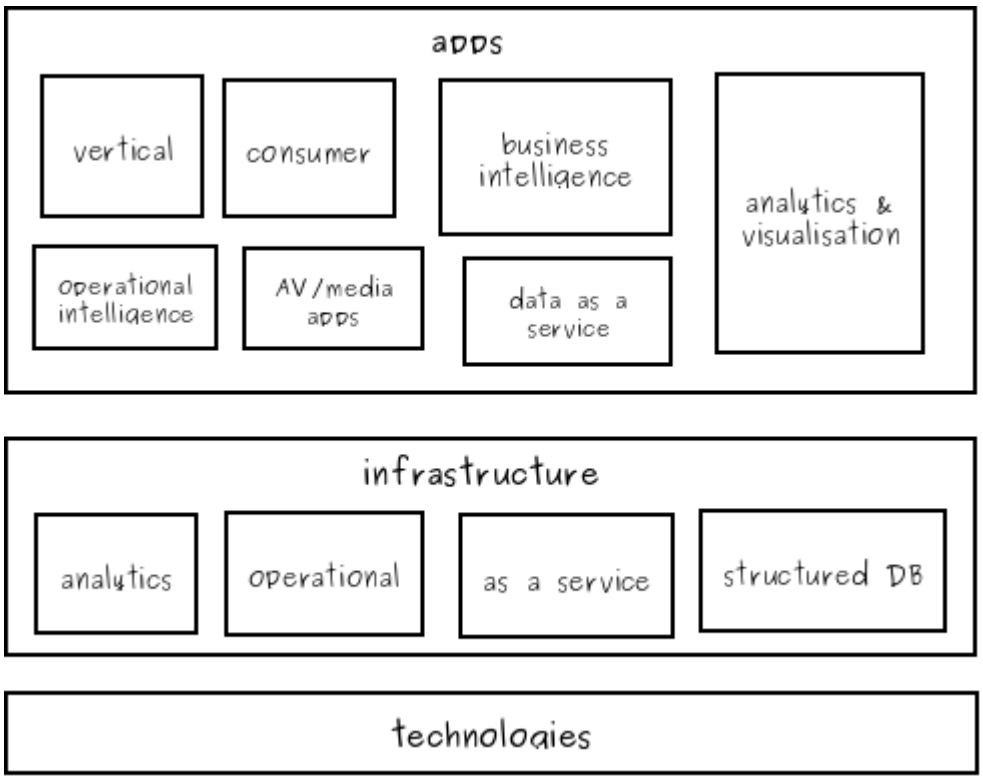
- Big Data Landscape 1.0 (2012) (https://www.pinterest.com/pin/439382507368166865) by Dave Feinleib on a Forbes blog
- Big Data Landscape 2.0 (2013) (https://www.pinterest.com/pin/141933825729010069/) by Matt Turck and Shivon Zilos of Bloomberg Ventures up on Slideshare.net
- Big Data Landscape 3.0 (2014) (https://www.pinterest.com/pin/293226625711471725) by Matt Turck of FirstMark Capital on Slideshare.net; note this one is too large but the Slideshare version (http://www.slideshare.net/mjft01/big-data-landscape-matt-turck-may-2014) allows you to go into fullscreen mode and thus view all the detail

Feinleib subsequently set up a website based on the concept at The Big Data Landscape (http://www.bigdatalandscape.com/) with a blog and review of the top 100 companies (http://www.bigdatalandscape.com/bigdata100), a useful contribution. Turck is a venture capitalist so his landscape is based on start-up companies.

A related graphic is The Data Science Ecosystem in one Tidy Infographic (http://www.datavizualization.com/blog/the-data-science-ecosystem-in-one-tidy-infographic) by Renette Youssef from CrowdFlower.

Different landscapes

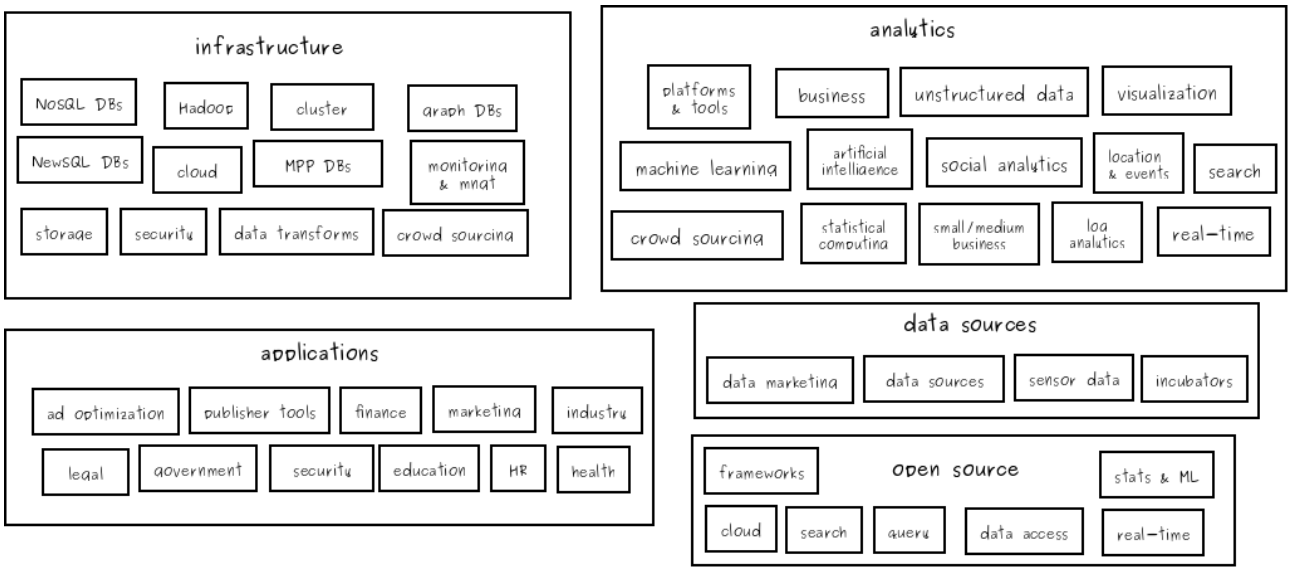
From these we get three different organisational frameworks for the landscape or ecosystem in the Data Science and big data world.



Feinleib's Big Data Landscape

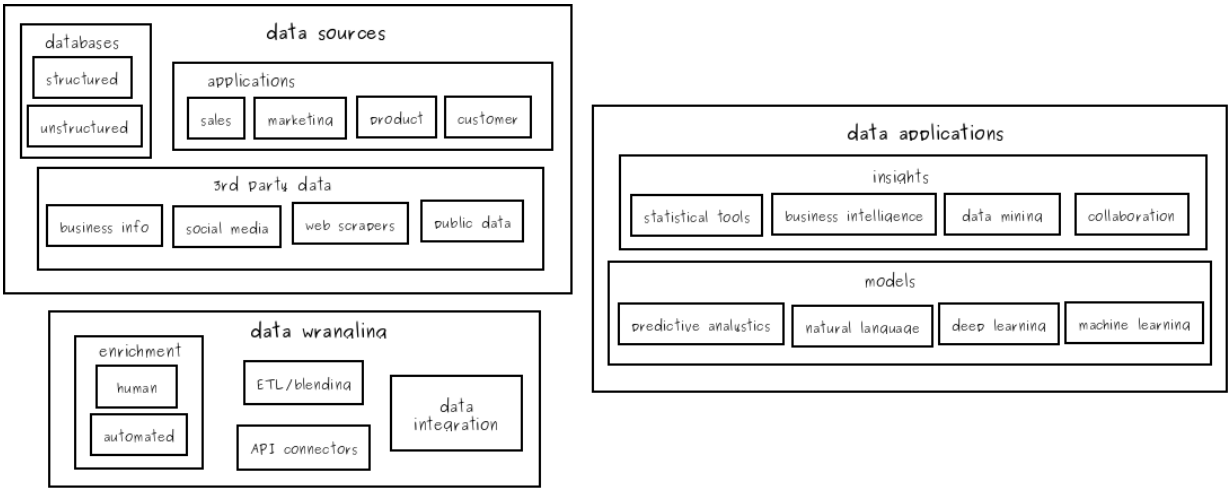
Feinleib's Landscape has the structure outlined in the figure, based around a simple split of apps

versus infrastructure.



Turck's Big Data Landscape

Turck's Landscape corresponds roughly to the Standard Value Chain with top level areas corresponding to collection ("data sources"), engineering ("infrastructure"), analytics and operationalisation ("applications"). Within each area we see the different kinds of data and different verticals addressed. The open source area covers many of the other areas.



Crowdflower Data Science Ecosystem

The **Crowdflower ecosystem** on the other hand is based on parts of the Standard Value Chain with collection and engineering ("data sources"), wrangling, and analytics ("data applications").

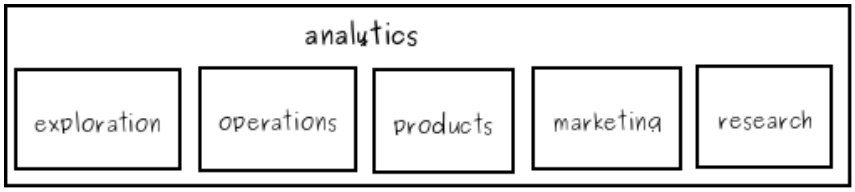
Analysis

A statistical analysis of the landscape is presented by Piatetsky in [Big Data Landscape, v 3.0, analyzed](http://www.kdnuggets.com/2014/05/big-data-landscape-v30-analyzed.html) (http://www.kdnuggets.com/2014/05/big-data-landscape-v30-analyzed.html), which is Turck's landscape. Their breakdown in terms of number of companies is roughly as follows: analytics (36%), infrastructure (28%), applications (18%), open source (9%) and data sources (9%). Since Turck's landscape is based on start-ups, Piatetsky pointed out that only 4.5% of the companies had moved beyond the start-up phase as at the time of his analysis. They also noted the most popular areas (in terms of number of companies):

Analytics: Data Visualization, Unstructured Data

Infrastructure: NoSQL and NewSQL databases
Applications: Ad optimization and Marketing.

A final characterisation of just analytics itself, a single area in Turck's landscape, is proposed by Tunguz in Which of the Five Types of Data Science Does Your Startup Need? (http://tomtunguz.com/data-science-types/) Tomaz Tonguz is an insightful venture capitalist working in Data Science. One difference, however, is that Tunguz is classifying data scientists, not companies.



Tunguz's Analytics

Operations, products and marketing are the verticals, and exploration is general modelling.

Models for the data-based business

So while the above landscape and ecosystem models revolve around either the value chain of data, or servicing traditional vertical businesses and areas such as marketing and customer relations, some businesses developing especially since the rise of the internet are primarily based around using data. Data is the core of their business, and the business model involves leveraging the data to obtain value. This sounds abstract, so let us look at some simple examples.

Bloomberg Terminal

According to Wikipedia:

> The **Bloomberg Terminal** **(https://en.wikipedia.org/wiki/Bloomberg_Terminal)** is a computer system provided by Bloomberg L.P. (https://en.wikipedia.org/wiki/Bloomberg_L.P.) that enables professionals in finance and other industries to access the **Bloomberg Professional** service through which users can monitor and analyze real-time financial market data and place trades on the electronic trading platform. The system also provides news, price quotes, and messaging across its proprietary secure network.

This is an *information brokering service*, it provisions basic data to users and also provides additional information, insight and analysis based on the data. Note Bloomberg predates the internet. Most data-based organisations that existed prior to the internet were in the finance industry, or, like the media business, servicing the finance industry.

Amazon.com

According to Wikipedia:

> **Amazon.com, Inc. (Wikipedia)** **(https://en.wikipedia.org/wiki/Amazon.com)** is an American electronic

> commerce company with headquarters in Seattle, Washington. It is the largest Internet-based retailer in the United States. Amazon.com started as an online bookstore, but soon diversified, selling DVDs, Blu-rays, CDs, video downloads/streaming, MP3 downloads/streaming, software, video games, electronics, apparel, furniture, food, toys and jewelry. The company also produces consumer electronics.

Moreover, Amazon also acts as a channel for other, primarily low end, retailers to sell their similar wares through similar systems as their own. Amazon Marketplace (https://en.wikipedia.org/wiki/Amazon_Marketplace) (Wikipedia)

> is Amazon.com's fixed-price online marketplace which enables sellers to offer new and used items alongside Amazon's regular offerings. Customers can buy those items directly from third-party sellers. The Marketplace uses Amazon.com's software infrastructure. Amazon.com charges the buyer's credit card and sends his or her payment to the seller, but does not pass along any credit-card information.

So what is Amazon doing with its data that is unique? Amazon does *information-based differentiation,* it satisfies customers by providing superior information and a superior range to support purchasing decisions. This differentiates them significantly from many standard book stores, DVD stores, and so forth. It is also an *information-based delivery network*, it enables advertising and fosters the market place for its secondary retailers in the Amazon marketplace. It can direct customers to them.

Business models

This material is drawn from an article in the Harvard Business Review:

- "What a Big-Data Business Model Looks Like" (http://hbr.org/2012/12/what-a-big-data-business-model) by Ray Wang in the Harvard Business Review (1000 words, 6 minutes). 

Wang argues there are three different styles of business models for the data-based business.

Information-based differentiation: providing improved customer service or improved products.
Information-based brokering: the provision of data and value added information.
Information-based delivery networks: using information to deliver services, marketing, and advertising.

Further examples are in Wang's article.

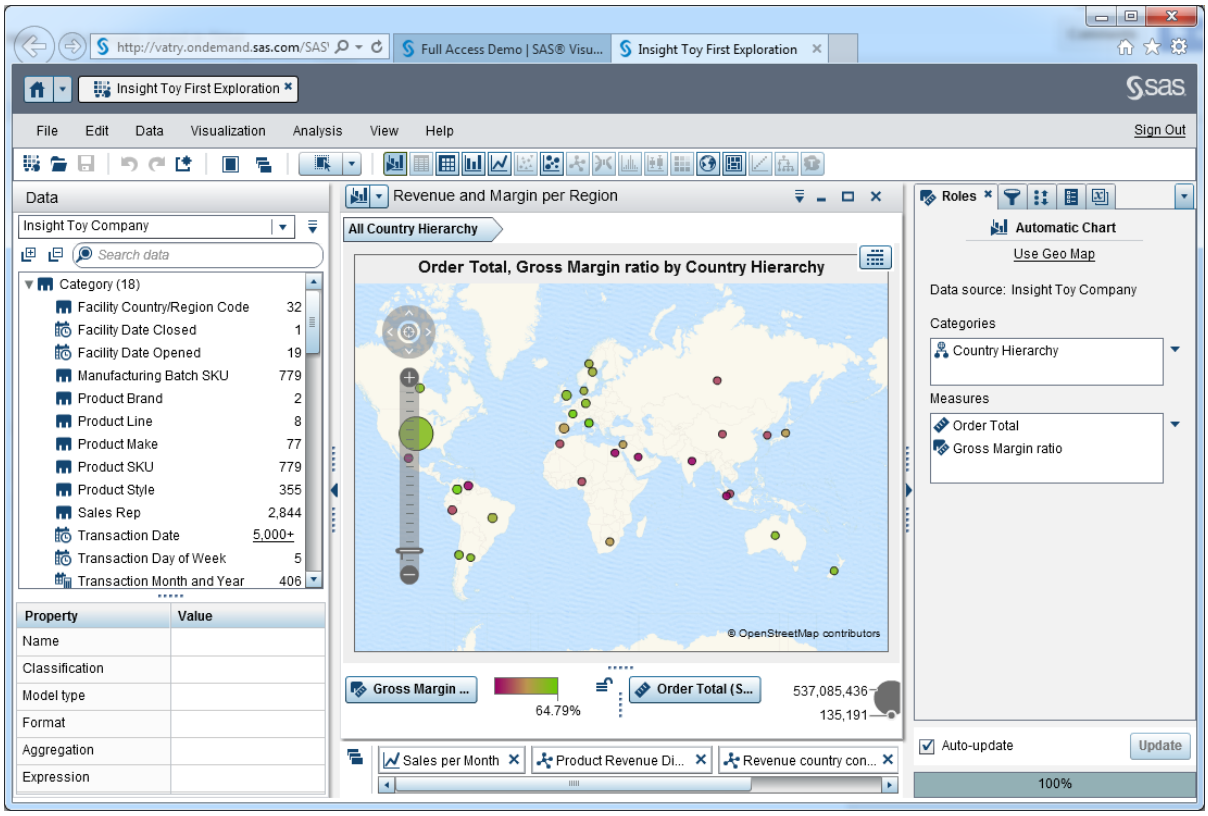
2.4
Activity: SAS Visual Analytics

SAS Visual Analytics

Statistical Analysis System (SAS) is one of the most commonly used commercial products for statistical analysis and visual analytics.
SAS provides good support for data science and is commonly used in areas such as business intelligence and data mining.
We are going to do some data exploration and visualisation using one of the datasets provided by SAS:

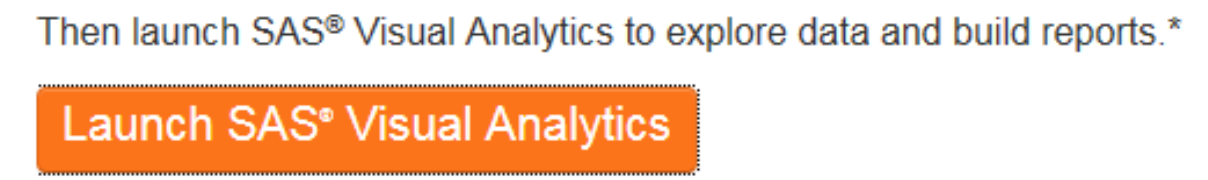
'The Insight Toy Company' which is '...made up of 1.4 million rows and 60+ columns'

This activity follows that provided by SAS, it is one of many, if you have an interest in any of the others then you are encouraged to explore further e.g. Text Analytics, HeatMaps, Decision Trees (see page 3 of the SAS Visual Analytics Startup Guide (SAS)
(http://www.sas.com/software/visual-analytics/demos/explore/SAS-Visual-Analytics-Startup-Guide.pdf)),



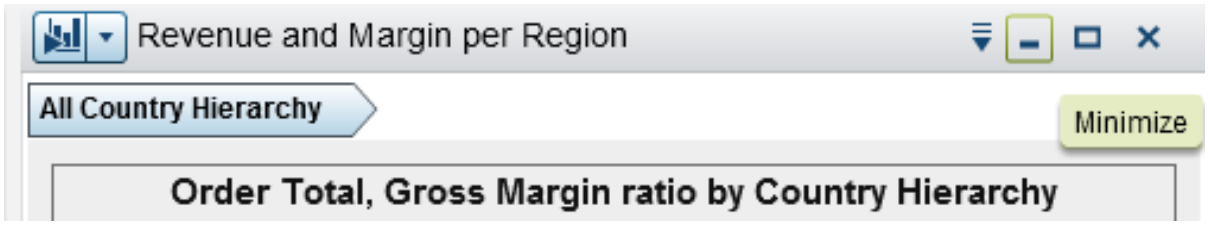
Step 1

Sign in to SAS and 'Launch SAS Visual Analytics' :
http://www.sas.com/software/visual-analytics/demos/full-access.html



Step 2

Start a new visualization on a blank workspace by simply minimizing the current one.

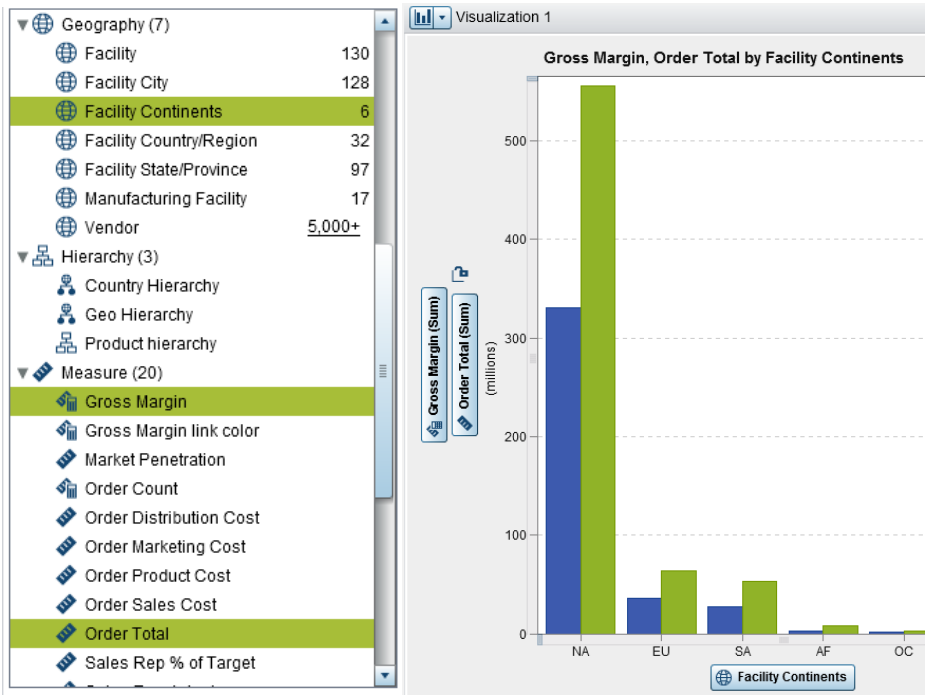


Step 3

Select 'Bar Chart' from menu

Step 4

On the Left (below) are the data by Category, Geography, Hierarchy, Measure etc. Use CTRL to select multiple items, then drag 'Facility Continents', 'Gross Margin' and 'Order Total' onto the central, blank visualisation:



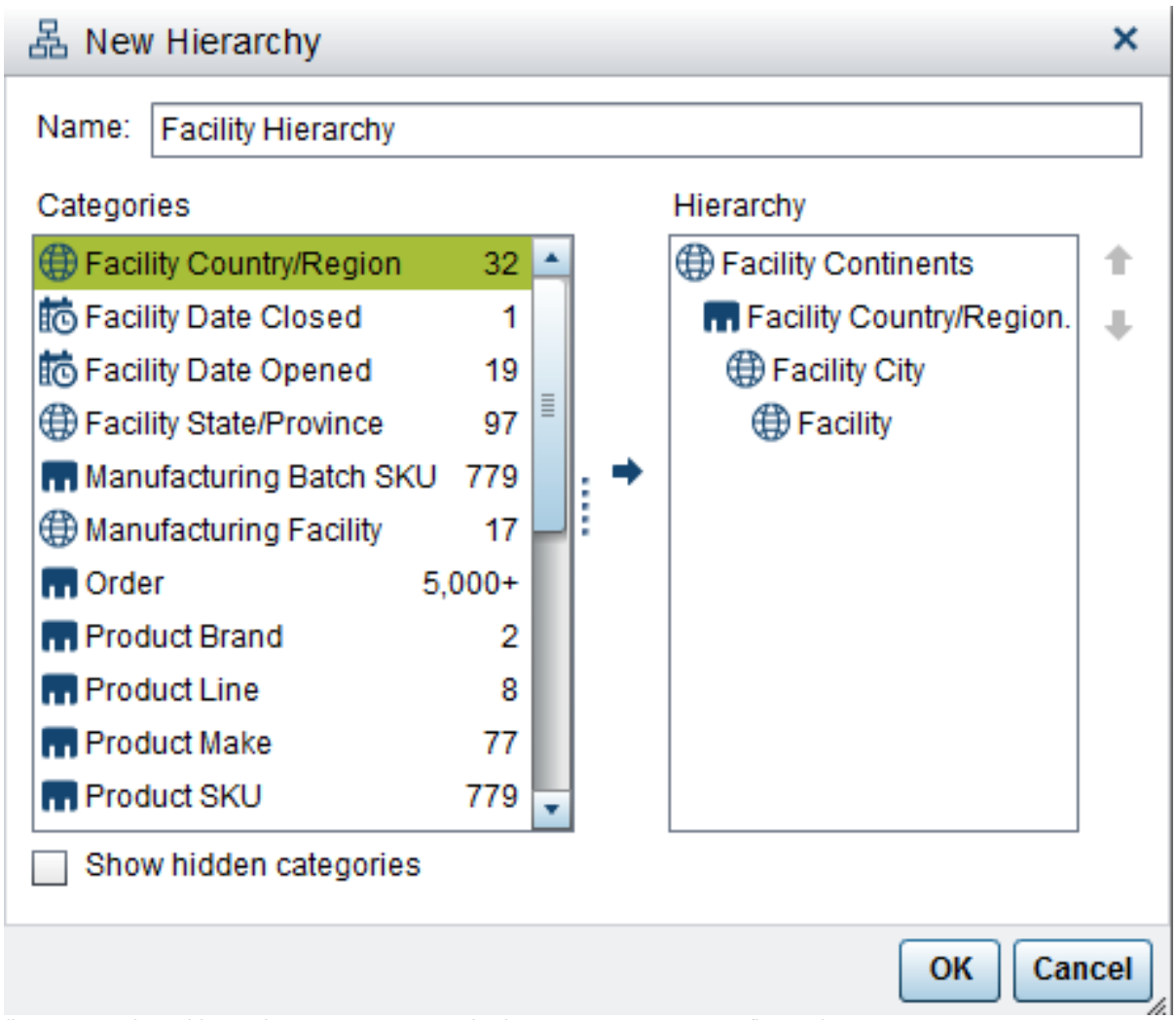
(https://www.alexandriarepository.org/wp-content/uploads/20150624011951/sas-va-fig6.png)And we can see some data from various regions (e.g. EU is Europe).
How many regions are there? What are they?

Step 5

Now we can investigate or 'drill-down' by adding a hierarchy.
From the 'Data' tab select 'New Hierarchy' and name it 'Facility Hierarchy'

Step 6

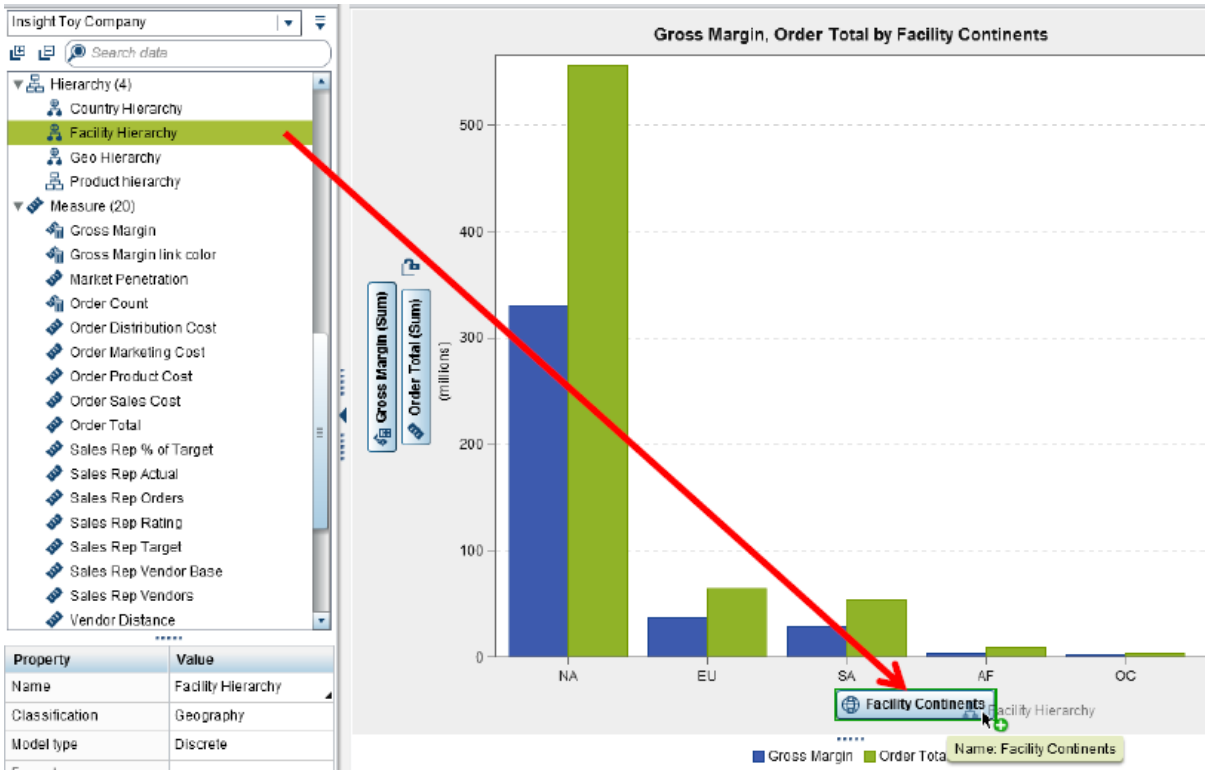
Drag (or double click) the following onto the Hierarchy IN ORDER: Facility Continents, Facility Country/Region, Facility City and Facility (then OK to continue).



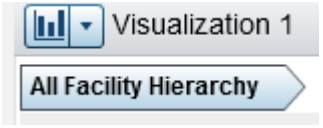
(https://www.alexandriarepository.org/wp-content/uploads/20150624011951/sas-va-fig7.png)

Step 7

Using a hierarchy: Drag the newly created 'Facility Hierarchy' from top left onto the bar chart, on top of 'Facility Continents'



Not much will change but you should see there is an added 'breadcrumb' trail icon at top of the bar chart



and that the regions (e.g. EU) are now underlined. What we have effectively done is add depth (or hierarchy) which we can now drill down into.

Step 8

Click on North America (**NA**) to drill down and see… how many countries?

Step 9

Drill down through **US**, which city has the lowest Order Total and what is that total?
Note that you can use 'Undo' & 'Redo' from the menu to go backwards (and forwards).

Step 10

Edit your hierarchy to add sales reps and drill down to find the best sales rep in Melbourne.
What is their ID and Order Total? What is their name?

Optional reading: How to Be a Data Scientist Using SAS
http://support.sas.com/resources/papers/proceedings14/1486-2014.pdf

Further activities, see: SAS-Visual-Analytics-Startup-Guide (103 pages)
http://www.sas.com/software/visual-analytics/demos/explore/SAS-Visual-Analytics-Startup-Guide.pdf

2.5 Application Areas

This section reviews different applications were Data Science has or is being applied. The material presented comes from reports, articles and some are extracted from videos.

In reviewing an application you need to consider its relationship to:

- the Standard Value Chain is how are different steps handled;
- the Data Science landscape or ecosystem with section **Business Models in Data** of module **Data Models in Organisations** which parts of the ecosystem are in play, which software and companies are involved;
- and what factors makes this kind of application/project be successful.

We will introduce more relationships in later modules.

MGI analysis of applications

The classic McKinsey Global Institute report on Big Data from 2011 can be found at:

- "Big data: The next frontier for innovation, competition, and productivity" (http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation) report from MGI (download the full report as PDF/EPUB/MOBI, just scan through **one** of the application areas listed below) 

While much of the analysis and prediction is now dated, their discussion of different application areas still makes insightful reading. They cover five main areas and each one is long, about 5000-6000 words, so 20-30 minutes reading each. You should pick *one area of interest* to read.

- Health: pages 39-51
- Government: pages 54-63
- Retail: pages 64-75
- Manufacturing: pages 76-84
- Location technology: pages 85-95

Future of data in medicine and health

Eric Schadt of the Icahn School of Medicine at Mount Sinai gave a longer invited talk at KDD 2014 (http://www.kdd.org/kdd2014/), but the first 11:30 minutes is an excellent introduction to data driven approaches in medicine.

- "A Data Driven Approach to Diagnosing and Treating Disease" (http://videolectures.net/kdd2014_schadt_treating_disease/), on VideoLectures.NET (video, see time 00:00-11.27) 

U.S. Department of Health and Human Services has released data sets on obesity and is challenging data

scientists to explore the data. The webpage gives a lucid description of goals and possibilities, and what other kinds of data could be used.

- "Big Data for a Big Problem: Putting Data To Work To Tackle Obesity" (http://www.hhs.gov/idealab/2015/07/08/big-data-big-problem-putting-data-work-tackle-obesity/) blog entry from the U.S. Department of Health and Human Services and their challenge webpage "U.S. Obesity Data Challenge" (https://www.challenge.gov/challenge/u-s-obesity-data-challenge/)

Big data in New York

This documentary news piece in the New York Times describes the use of Data Science in New York, apparently promoted by Mayer Bloomberg himself.

"The Mayor's Geek Squad (http://www.nytimes.com/2013/03/24/nyregion/mayor-bloombergs-geek-squad.html)" on New York Times 24/03/2013 (2600 words, 13 minutes) 

Manufacturing

This is a Forbes cover story taken from another McKinsey report, this time about applications in manufacturing. The actual report is quite lengthy, and this Forbes article covers the main points for us.

- "Ten Ways Big Data Is Revolutionizing Manufacturing" (http://www.forbes.com/sites/louiscolumbus/2014/11/28/ten-ways-big-data-is-revolutionizing-manufacturing/) on Forbes 28/11/2014 (1400 words, 8 minutes) 

Data-Intensive Science

The Fourth Paradigm: Data-Intensive Scientific Discovery (http://research.microsoft.com/en-us/collaboration/fourthparadigm/) is a Microsoft promoted book inspired by the work of computer science legend Jim Gray. The book itself has many contributed chapters, but Jim Gray's main chapter summarises the key ideas of data-driven science that interest us.

- "Jim Gray on eScience: A Transformed Scientific Method" (http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf) book chapter from Microsoft Research (PDF, 2500 words, 13 minutes)

How do people spend their time online

People's activity online is an important source of data, and an important subject for business models involving data. Twitter, a data-centric company, built their business around providing an online tool. The following infographic prepared by GO-Gulf provides some of the details.

- "How People Spend Their Time Online" (http://www.go-gulf.com/blog/online-time/) by GO-Gulf (infographic on a blog) 

Operations analytics

The abundance and growth of machine data, sensors, meters, GPS devices, machines *etc.* in factories, on ships, in satellites, is another major driver and application area for Data Science. IBM provides the following infographic.

- "Operations Analytics" (http://www.ibmbigdatahub.com/infographic/operations-analytics) by IBM (infographic on a blog)

General Application Areas

The NIST Big Data Working Group has developed a list of use cases, reported in the 260 page PDF file "Volume 3, Use Cases and General Requirements" (http://dx.doi.org/10.6028/NIST.SP.1500-3), where contributors prepare a small fact sheet in each use case, in areas such as health, government, defense, science, retail and manufacturing. The descriptions are quite detailed, so this is best viewed as a resource rather than a set of documents to study. We suggested that you select an area of interest to you for further investigation, as there are too many to cover. So pick a few of the cases studies and review them.

NIST uses a set of criteria to organise their analysis, which we will convert to our own terminology. Their criteria can be viewed by looking at the beginning of Appendix A. Their analysis of individual cases is tabulated in Appendices A, B, C and D. The Table of Contents, pages v and vi of the report, gives a listing of all cases, and discussion and background about the cases is given in the main body, pages 5 to 42. We use the following aspects of their analysis, see page 43 and 44 and appendix pages A-3 to A-5.

Data Sources: where does the data come from?
Volume: how much data?
Velocity: how does it change over time?
Variety: what different kinds of data are there?
Software: what major software is used in the various steps?
Analytics: what sorts of analysis are done?
Transformation: what sorts of wrangling is done?
Data Consumer: operationalisation, what happens to the results?
Security & Privacy: aspects of governance
Lifecycle Management: how is governance managed?

So we can see the NIST framework describes:

- major steps from the Standard Value Chain,
- some of the key dimensions of the data (which we cover more in the next module)
- and, the software used.

2.6
Analysis and Interviews

Analysis from Provost and Fawcett

In their discussion article:

- "Data Science and its Relationship to Big Data and Data-Driven Decision Making," (http://online.liebertpub.com/doi/abs/10.1089/big.2013.1508) Foster Provost and Tom Fawcett. *Big Data*. March 2013, 1(1): 51-59 (8 page PDF, 6000 words, 30 minutes) 

Provost and Fawcett discuss the new science and its parts. Given the material we have covered in the first two modules, this is an excellent summary and insightful discussion.

This is a really good article! You should really read it all, but it is quite long.

Analysis from a Business Investor

In his discussion article:

- "You're Thinking About Investing In Big Data? Consider These 5 Things," (https://blog.pivotal.io/big-data-pivotal/features/youre-thinking-about-investing-in-big-data-consider-these-5-things?utm_source=social&account_id&utm_medium=TWITTER&PivotalBigData&utm_campaign=Big%20Data&20160206) Jeff Kelly in the *Pivotal Blog*. February 03 2016, blog page, 900 words, 5 minutes) 

Jeff Kelly mentions 5 critical points. We can reinterpret them for Data Science:

1. What is the business use case for data science effort?
2. How will you measure success?
3. Who is leading the initiative, the business or IT?
4. How do you plan to address the Data Science skills gap (finding skilled staff is hard)?
5. How do you plan to overcome institutional resistance (changing the way of doing business is challenging)?

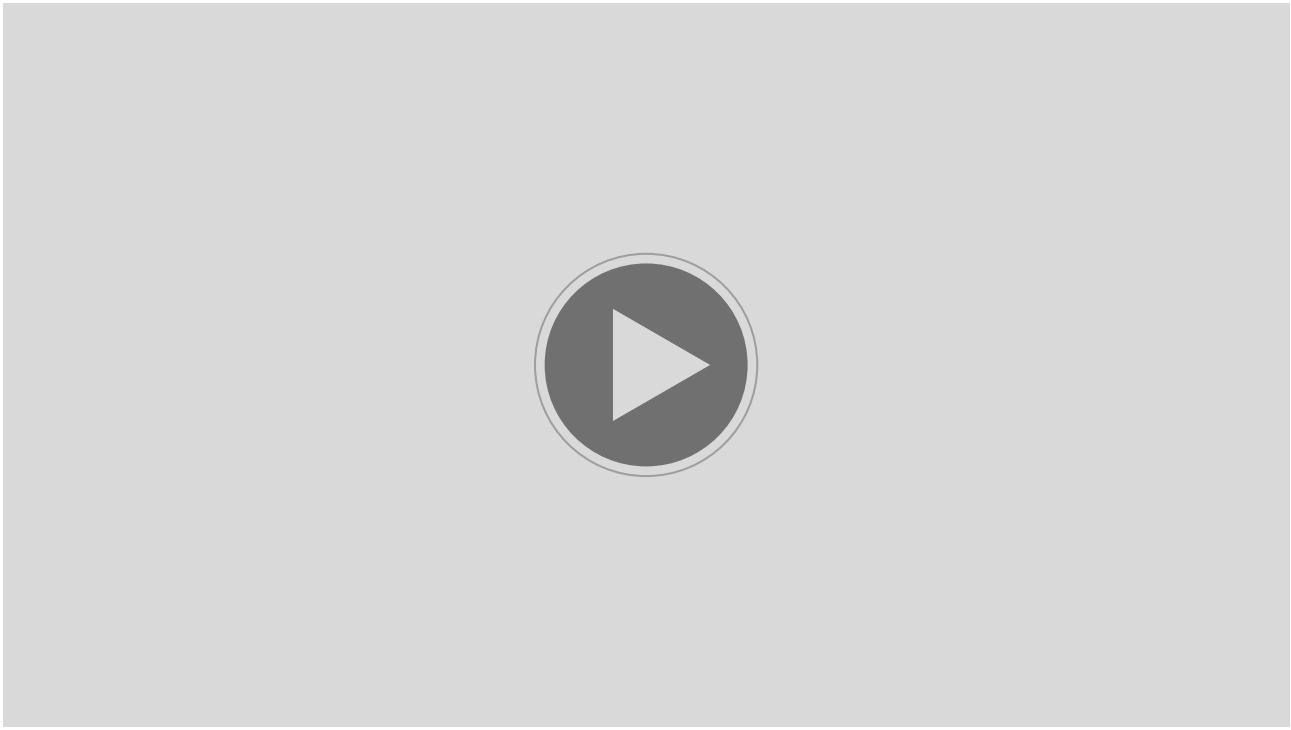
See if you can understand what he means in each case. These are the key business decisions in starting up a data science effort. You see issues like whether to use R or Python, and other technical issues, are not that important here!

Interviews with industry professionals

We continue with a few interviews with industry professionals here in Australia.

Watch Con Nidras (Head of Customer and Channel Analytics - National Australia Bank (NAB)), Associate Professor Michael Brand (Faculty of Information Technology - Monash University and former data scientist at Pivotal), Associate Professor Chris Bain (Director of information services - The Alfred Hospital) and Dr

Fang Chen (Research Group Manager - National ICT Australia (NICTA)) talk about their experiences with **data models**.



(https://www.alexandriarepository.org/wp-content/uploads/20150629085912/FIT5145_module_2_data_models_in_org_combined.mp4.mp4)
Alternatively, you can download the transcript for Data models in an organisation (https://www.alexandriarepository.org/wp-content/uploads/20150701100107/transcript_FIT5145_module_2_data_models_in_org.pdf).

Data Analytics Handbook, Volume 2, Business Leaders

Volume 2 of the informative *Data Analytics Handbook* series interviews Business Leaders. There are 9 interviews in all.

- https://www.teamleada.com/handbook, Volume 2 (approx 10,000 words, 1 hour)

Videos from the Institute of Analytics Professionals of Australia

The IAPA (http://www.iapa.org.au/) has prepared the Value of Data Series videos with speakers from a cross section of industry.

- Evan Stubbs, SAS (http://www.iapa.org.au/Article/VideoTheValueOfDataSeriesEvanStubbsSAS) (video, 16 minutes)
- Robert Hillard, Deloitte (http://www.iapa.org.au/Article/VideoTheValueOfDataSeriesRobertHillardDeloitteAust) (video, 15 mins)

Further discussions of business strategy

Additional material as reference on strategies for applying data science

- "Winning The Analytics Race With Data Science" (http://www.forbes.com/sites/teradata/2015/05/07/winning-the-analytics-race-with-data-science/) by Scott Langfeldt in Forbes
- "The analytics imperative: embedding data and analytics in the business model" (http://blog.kpmg.ch/analytics-imperative-embedding-data-analytics-business-model/) by Teodor Pistalu on KPMG's blog

- "Five Factors In Building Giants Of The Big Data Era" (http://techcrunch.com/2014/08/30/five-factors-in-building-giants-of-the-big-data-era/) by Navin Chaddha on TechCrunch.com

- "How we scaled data science to all sides of Airbnb over 5 years of hypergrowth" (http://venturebeat.com/2015/06/30/how-we-scaled-data-science-to-all-sides-of-airbnb-over-5-years-of-hypergrowth/) by Riley Newman on VentureBeat

- "Australian Public Service Better Practice Guide for Big Data" (http://www.finance.gov.au/sites/default/files/APS-Better-Practice-Guide-for-Big-Data.pdf) is a 30 page report describing better practice targeting Australian Government agencies

2.7
Activity: Decision Modelling

In this activity we will develop some simple decision models using influence diagrams presented in the **Data and Decision Models** section of the **Data Models in Organisations** module.

Traffic accident analysis

Consider the following scenario for which we want to build an influence diagramme.

> You wish to build a model for road accidents. While this is not a spatial model (*i.e.*, based on a map) it should be spatially aware. The model is about whether you will have an accident on a particular section of street or street corner at a particular point of time. What sorts of variables might you consider, and what sorts of impacts would an accident have so you can evaluate its cost? For now, consider just chance and known variables and objectives. Don't include decision variables in the problem. To distinguish between knowns and unknowns, lets assume the knowns are about long term properties of the street, such as location, and the unknowns are about current (transient) properties such as weather, presence of roadworks, *etc. An important class of these transient variables describe features of the accident itself, such as lives lost or people injured.*

One important group of unknown variables should classify the accident: what vehicles were their and what people or property was harmed by the accident? Some of these variables might also be objectives or should connect to objectives? Clearly there are also many variables as inputs. For instance, how busy is this section of street at a given point of time, is their currently an "event" (like a football final) that could affect traffic, and what sort of street is it, dual carriageway, dead-end, one way, *etc.*? Now you need to be realistic about your variables. For instance, "one driver is drunk" is not a realistic variable to use as it is not one we can readily measure, though "there is a busy pub nearby" or "current traffic likely comes from a drinking venue" could be measured. Our simple model should have no more than 8 nodes total, but no more than 16. Moreover, their is no correct answer to this question. There are good and better answers, just as their are realistic and not so realistic variables to measure.

1. What nodes objectives will you need? Clearly, "dollar damage to private property" might be one variable. But is there a total cost, or live's lost? What other objectives?
2. What other variables do you need? Some will describe the accident and be input to the objectives, some will be long term properties of the location, and some will be transient properties.
3. What arcs do you place between nodes?

Draw up the mode, and call this model A. Be warned, you could spend weeks on this problem if you wanted to do a thorough job. Try and keep it simple but with a good breadth.

You can draw it using drawing software, but it might be easier to use pen and paper and take a picture.

- *dia* works well on Linux
- *Google Drawings* is an adequate browser-based system
- LucidCharts.com (http://LucidCharts.com) is another web-based system

> Some variables could be obtained from the police officer attending the incident. Of your other variables, which variables can you directly measure, which variables will you have to estimate from data, and which variables will you have to ask an expert to help you gather.
>
> For instance, to get general traffic volume at time of day at the nearest traffic light, one would use the road sensor data for the light. Weather data we get for time of day and location from BOM. Note, if we are doing traffic analysis about future accidents, the less we ask the police office at an accident site, the better. We want to gather as much informative data as we can about each location.

Write a few sentences about each variable and its worthiness and difficulty as a data source, and the strategy you might take to gather the data.

> Suppose you work for the local traffic planning office: add a single decision node, which is "install a speed camera at the location". Connect the decision node up to other appropriate nodes.

Now extend the previous model with the extra decision, and call the extension model B. Again, there are many potential choices to make, so try to keep it simple.

3
Data Types and Storage

This is our third module of six for the Introduction to Data Science unit. This module discusses the actual problems with big data and the 3 (or 4) V's. In this module, different kinds of data and their properties, and simple models of big data processing will be covered.

(https://www.youtube.com/watch?v=l3z7TX4pX_4)

Aims of This Module

- Explain the 4 V's, velocity, volume, variety, veracity.
- Describe the different kinds of data and their typical storage requirements.
- Explain basic issues and constraints of different data types, storage, streaming, in-memory processing.

How to study for this module

In this module we again draw on material in the public domain such as interviews and videos, online magazine entries and blogs. We have also written some material to tie together various kinds of models. As well as studying and viewing the material, we have some activities around this material.

Please remember:

- Reference items marked with a single "johny look it up" icon,  , should be viewed as *suggested reading*, not essential nor important for assessment.

- Reference items marked with a two "johny look it up" icons,  should be viewed as *important reading*, considered important for assessment.

3.1
Characterizing Data

Overview

Clearly, "big data" is a loaded term. How big is big? While this is in some sense a diversion from the real business of Data Science, Patricia Florissi, VP and Global Sales CTO of EMC Corporation, discusses the issues:

- "Big Ideas: How big is big data?" (https://www.youtube.com/watch?v=eEpxN0htRKI) (9 mins)  , learn how "EMC Solutions help create value by merging Big Data with Cloud Computing in this hand-drawn animation"

While some of this is a sales pitch for EMC, it does a good job of discussing the industry perspective and issues that are relevant.

The Wikipedia entry for big data (https://en.wikipedia.org/wiki/Big_data) is extensive and some of the sections should be reviewed.

- Wikipedia on Big data (https://en.wikipedia.org/wiki/Big_data), see the section **Definition** plus the sections **Characteristics, Architecture and Technologies** (1500 words, 8 mins)

The V's

All self-respecting data scientists need to know a lot of words beginning with "V". The phrase "Volume, Velocity, and Variety" started with a report in 2001 by Gartner analyst Doug Laney, "3D Data Management: Controlling Data Volume, Velocity, and Variety" (http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf) (of historical interest only). Somewhere along the way extra V's were added, Veracity, Variability, ... A brief and adequate summary of all the "V"s is:

- "Understanding Big Data: The Seven V's" (http://dataconomy.com/seven-vs-big-data/) by Eileen McNulty on the Dataconomy news portal (1100 words, 5 mins)  .
- "Why The 3V's Are Not Sufficient To Describe Big Data" (https://datafloq.com/read/3vs-sufficient-describe-big-data/166) by Mark van Rijmenam on the Datafloq portal (1300 words, 6min)  and read the comments at the bottom by Doug Laney for a discussion on definitions.

There are, of course, way too many infographics for this, and this one by IBM is useful for perspective.

- "The Four V's of Big Data," (http://www.ibmbigdatahub.com/infographic/four-vs-big-data) by IBM (infographic) 

Data formats

Wolfram Alpha's history of systematic data gives key dates for recording/storage/coding methods for storing data and metadata, for instance, the Medical Subject Headings (MeSH) created in 1963. In computer science we talk about data formats, which could mean primitive file or disk formatting. But above this level we need formats to store information rather than single letters or floating point numbers. For this there exists metadata (we use the Wikipedia entry "Metadata_standards#Metadata" as the definition is preferable):

> **Metadata** (https://en.wikipedia.org/wiki/Metadata_standards#Metadata) is often defined as *data about data*. It is "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource".

The important word here is "structured", which could mean a computer can process it without statistical inference or supposition. Wikipedia presents three types of metadata (see **Metadata** (https://en.wikipedia.org/wiki/Metadata_standards#Metadata)):

- *Descriptive metadata* describes an information resource for identification and retrieval through elements such as title, author, and abstract.
- *Structural metadata* documents relationships within and among objects through elements such as links to other components (e.g., how pages are put together to form chapters).
- *Administrative metadata* helps to manage information resources through elements such as version number, archiving date, and other technical information for purposes of file management, rights management and preservation.

Other relevant terms to understand data formats (from Wikipedia):

> Machine-readable data (https://en.wikipedia.org/wiki/Machine-readable_data) is data (or metadata) which is in a format that can be understood by a computer. There are two types; human-readable data that is marked up so that it can also be read by machines (examples; microformats, RDFa) or data file formats intended principally for processing by machines (RDF, XML, JSON).

> Markup language (https://en.wikipedia.org/wiki/Markup_language) is a system for annotating a document in a way that is syntactically distinguishable from the text. The idea and terminology evolved from the "*marking up*" of paper manuscripts, i.e., the revision instructions by editors, traditionally written with a blue pencil on authors' manuscripts.

> A digital container (or wrapper) format (https://en.wikipedia.org/wiki/Digital_container_format) is a file format whose specification describes how different elements of data and metadata coexist in a computer file. Containers are frequently used in multimedia applications.

We can say markup language is one form of descriptive metadata done via annotation. It is also popular with programming languages for embedded documentation.

Having data be *machine-readable* is critical for automated processing. However, "understandable" is also somewhat of a nebulous concept. A computer can certainly parse text, display PDF and uncompress JPEG but can it understand it the data? Not really. For instance, the free text inside a document is not able to be automatically processed to "extract information". It requires natural language processing (NLP) tools to extract the information from the text, which is typically done with 70-90% accuracy. Moreover, the image of a page in a book, used in some earlier documents stored in PDF, cannot be readily converted into the text and needs OCR tools to complete a 99% accurate conversion to text before NLP tools could be

applied. The other parts of a document, for instance the structured components and metadata can sometimes be automatically processed without additional conversions. However, older PDF documents, for instance, are notoriously difficult to extract the document structure from.

The important role of metadata is to store information needed for the description and the administration of the data. A simple example of this is the metadata embedded in a web page.

```
<?xml version="1.0"?>
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
<head>
    <meta http-equiv="X-UA-Compatible" content="IE=9"/>
    <title>Machine Learning - Monash University</title>
    <meta http-equiv="Content-Type" content="text/html; charset=UTF-8"/>
    <meta name="date.review" scheme="ISO8601" content="2013-10-09"/>
    <meta name="dc.description" content="The Machine Learning research program investigates t
    <meta name="description" content="The Machine Learning research program investigates the
    <meta name="keywords" content="machine learning,data mining,bayesian networks,mml"/>
    <meta name="dc.title" content="Machine Learning Research Flagship"/>
    <meta name="dc.identifier" scheme="uri" content="http://www.infotech.monash.edu.au/resear
    <link rel="stylesheet" href="MachineLearning-MonashUniversity_files/reset.css" type="text
    .....
</link>
<script src="MachineLearning-MonashUniversity_files/ga.js" async="" type="text/javascript"/>
...

<!--OTHER HEADERS-->
<!-- BRANCH SPECIFIC HEADERS -->
<link rel="stylesheet" href="MachineLearning-MonashUniversity_files/faculty.css" type="text/c
<!-- PAGE SPECIFIC HEADERS -->
<link rel="stylesheet" href="MachineLearning-MonashUniversity_files/flagships.css" type="text
</head>
<body id="content_page">
```

Monash Web Page XML

This is a snapshot of a Monash University webpage. The key thing of interest here is the <meta> tags. The tags whose name has a "dc." prefix means Dublin Core (https://en.wikipedia.org/wiki/Dublin_Core), which is a metadata standard for describing web, library and museum resources. There are many other metadata standards (https://en.wikipedia.org/wiki/Metadata_standards). You can see the entries such as "date.review" with a value "2013-10-09", "charset=UTF-8", "dc.title", "keywords", *etc.* Other Dublin Core fields can be "dc.rights" (*i.e.*, copyright *etc.*), "dc.language", "dc.format". So language, title, keywords, date and charset are all descriptive metadata that can be used in retrieval functions,. Language, date.review and rights are administrative metadata that can be used for managing the data.

A final aspect of this file is the initial entry on the second line xmlns="http://www.w3.org/1999/xhtml". This is structural metadata that says the file is in XHTML format. Actually, the original was not correct XHTML - a common problem is improper adherence to standards. Even prestigious organisations like the National Institute for Health in the USA distributes XML data on bibliographies that does not confirm.

We will look at the formats more when considering standards in the module **Data Resources, Processes, Standards and Tools**.

Kinds of data

There are, of course, many different kinds of data available. Here are some examples that are not necessarily mutually exclusive:

- **social:** microblogging, friends networks, *etc.*,

- **transactional**: primarily corporate data in structured records on purchases, inventory, customers, *etc.*,
- **audio, image and video:** both in the consumer space and the entertainment space,
- **mobile and telecom**: including networks and services (*e.g.*, cell tower data, GPS data),
- **documents and text**: offices, government, education, healthcare and corporate have huge volumes; documents can often be structured text and images,
- **IT/OT:** "OT" stands for Operational Technology so it is the data you get from instrumentation on the running of a factory or warehouse. Typically this is *sensor* data but is often collected via a lowend computing device, and more and more is integrated with IT,
- **search and internet**: intranet and internet search engines, web marketing, web logs and the public web, *etc.*,
- **bibliographic and intellectual property**: patents, library databases, bibliographic data such as Thompson Reuters, Scopus, *etc.*,
- **publishing**: academic and trade journals, *etc.*,
- **IoT**: standing for Internet of Things,
- **LOD**: standing for "Linked Open Data" and less formal variants, which is structured, networked content such as Freebase or PubChem,
- **geospatial and science**: many forms, but the geospatial data is sometimes distributed in open form via government services,
- **government**: broad category often includes much of the above, and sometimes has public (or open) access.

Note that the commercial world of Data Science is usually only concerned with the first six, however they continue to use other data, notably geospatial data and linked open data as important auxiliary resources. Publishing and bibliographic data is also vital in the medical and pharmaceutical business.

Dimensions of data

Infographics are a natural means to describe the dimensions of data:

- "Data Science Matters" (http://datascience101.wordpress.com/2013/11/15/data-size-matters-infographic/) (infographic)  from the datascience@berkeley Blog, for a perspective on the different storage sizes over time.
- "Intelligence by Variety - Where to Find and Access Big Data" (http://dataconomy.com/intelligence-by-variety-where-to-find-access-big-data/) (infographic)  from Kapow software presents the different data sources and the nature of their volume, velocity, and variety.
- "60 Seconds - Things That Happen On Internet Every Sixty Seconds" (http://www.go-gulf.com/blog/60-seconds/) (infographic)  from GO-Gulf which presents data created, and its counterpart "60 Seconds - Things That Happen Every Sixty Seconds Part 2" (http://www.go-gulf.com/blog/60-seconds-v2/) (infographic)  which presents transactions made.

Growth laws

To understand, from a technology viewpoint, how big data is inevitable, and its only going to get bigger, there are four so-called laws of technology that talk about the inevitablity of growth. These are really predictions based on observation, but most have stood the test of time.

Moore's Law

The most famous growth law, Moore's Law (http://en.wikipedia.org/wiki/Moore%27s_law), from Intel founder Gordon Moore states that:

> over the history of computing hardware (https://en.wikipedia.org/wiki/History_of_computing_hardware), the number of transistors (https://en.wikipedia.org/wiki/Transistor) in a dense integrated circuit (https://en.wikipedia.org/wiki/Integrated_circuit) has doubled approximately every two years

For our purposes, what this means is that things like memory capacity and speed of computers, more or less, does the same, doubles every two years.

Koomey's Law

A corollary of Moore's Law, Koomey's Law (http://en.wikipedia.org/wiki/Koomey%27s_law), is about energy consumption:

The implications of Koomey's law are that the amount of battery needed for a fixed computing load will fall by a factor of 100 every decade.

This really implies the inevitability of the Internet of Things, and for our purposes means than Data Science type analysis will become more and more available to the general public with lower cost computing (in terms of power).

Bell's Law

Based on Moore's law (but also Koomey's Law), computing visionary Gordan Bell gave Bell's Law (http://en.wikipedia.org/wiki/Bell%27s_law_of_computer_classes), which argues that every ten years or so a new class of computer system will emerge to fundamentally change the way we do business. Recent "classes" along this line have been mobile computing, the Cloud, and the Internet of Things. Each is arguably critical for Data Science. These new classes open up whole new avenues for collecting, storing, processing and distributing data.

Zimmerman's Law

Zimmerman is the creator of Pretty Good Privacy (PGP), the most widely used email encryption software in the world. Zimmerman's Law (http://en.wikipedia.org/wiki/Phil_Zimmermann#Zimmermann.27s_Law) was stated in a blog he did:

> The natural flow of technology tends to move in the direction of making surveillance easier, and the ability of computers to track us doubles every eighteen months.

Thus, privacy decreases as ability and to store and track data increases. Given that corporations are driven by tasks like advertisement targeting and personalisation to track personal data, they have some incentive to maintain surveillance.

3.2
Activity: Big Data

This activity is intended to explore aspects of big data, to probe your understanding of the sizes and speeds involved.

A. Big Data Sets

> The amount of data stored is increasing exponentially, in line with Moore's law (which indirectly says the available storage should increase exponentially).

Given 6 massive datasets (below) try to pick which is the biggest, 2nd and 3rd biggest. Note they are listed here in alphabetical order:

- Business email (1 year)
- Facebook posts (1 year)
- Google's search index
- Kaiser Permanente Medical Records
- Large Hadron Collider data (1 year)
- YouTube uploads (1 year)

Compare your guesses with original graphic and key, which is here
http://www.wired.com/2013/04/bigdata . Any surprises?

How big is 3 million terabytes anyway? It is 3 x 1,000 x 1,000 terabytes which is an Exabyte. See
https://en.wikipedia.org/wiki/Exa-
Which of these sets overlap? (not in the graphic, in reality). Which of these sets do you have access to? What kind of access? For example, not the medical data, probably not even the structure, or, put another way, you have some idea of how big the Kaiser medical dataset is, are you able to discover anything else about it? When you 'google the web' do you have full access to their index?

B. Public Data Sets

Start with the Enron email dataset story here: http://www.salon.com/2003/10/14/enron_22/
There was also a 2005 documentary "Enron: The Smartest Guys in the Room" (and a scandal, and a court case).
Where can you get the Enron email dataset? How big is it? How many versions are there?

As Mick Dundee might say ... "That's not a dataset, this is a dataset"

> Common Crawl Corpus 541TB https://aws.amazon.com/datasets/41740

How long would that take to download this 541TB? Where would you store it anyway, and what is in it? What has been done with this dataset? See if you can find anything bigger (public and private) than the Common Crawl Corpus.
Many of these big data sets are text.

Optional

Reports on email statistics:
http://www.radicati.com/wp/wp-content/uploads/2013/04/Email-Statistics-Report-2013-2017-Executive-Summary.pdf

also
http://sourcedigit.com/4233-much-email-use-daily-182-9-billion-emails-sentreceived-per-day-worldwide/

C. Big Data Transfer

Transferring (downloading) a movie, e.g. The Matrix (7.8GB).

Method	Approx Speed	Days	Hours	Minutes
Dialup	56 Kbps	13		
Wireless	512 Kbps	1.5		
DSL	640 Kbps	1		
Cable	1.5 Mbps	11.5		
T1	1.54 Mbps	11		
Ethernet	10 Mbps	2		
Fast Ethernet	100 Mbps	10.5		
Gigabit Ethernet	1000 Mbps	1		

How many time faster than Wireless is Gigabit Ethernet?
What is 'SneakerNet'? Look it up if you don't know. How fast would it be?
What's faster than Gigabit Ethernet?

Speed test your Internet connection, e.g.
http://www.ozspeedtest.com or http://www.iinet.net.au/internet/broadband/speed-test/.
Pick a data set from above (e.g. Hubble or LHC) and compute how long would it take to download using your connection?

Suppose you were searching entries in the dataset for some pattern ("cats" in youtube videos; particular sub-atomic traces in the large Hadron collider; particular galaxy shapes in Hubble images). Clearly there's a bottleneck in downloading to your computer to do the search. What's the solution?

700 terabytes you say... "The Square Kilometer Array (SKA), a planned array of thousands of telescopes in South Africa and Australia... the array will scan the skies for radio waves coming from the earliest galaxies known. JPL is involved with archiving the array's torrents of images: 700 terabytes of data are expected to rush in every day. That's equivalent to all the data flowing on the Internet every two days." http://www.jpl.nasa.gov/news/news.php?release=2013-299

Coincidentally... "Breakthrough in storing 700 terabytes of data in 1 gram of DNA"

http://www.smh.com.au/technology/sci-tech/breakthrough-in-storing-700-terabytes-of-data-in-1-gram-of-dna-20130123-2d89q.html

How have these ideas impacted Big Data and Data Science. What new "computer classes" in Bell's sense might have a future impact on Data Science?

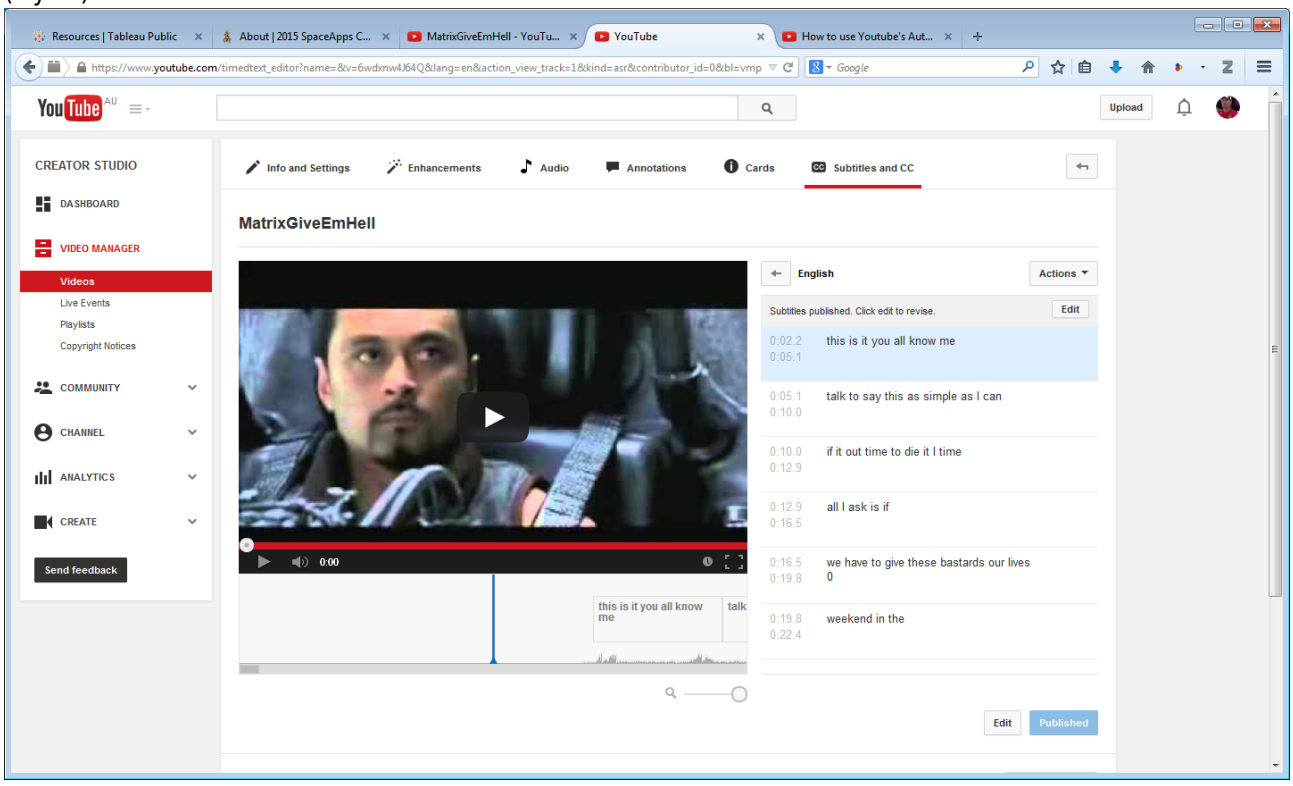
D. Data Formats

Consider the various forms that data can take, e.g. a movie (The Matrix) could be compressed, edited, it could be audio only, stills only (screenshots), transcript only, *etc.* If there wasn't a transcript you could make one:

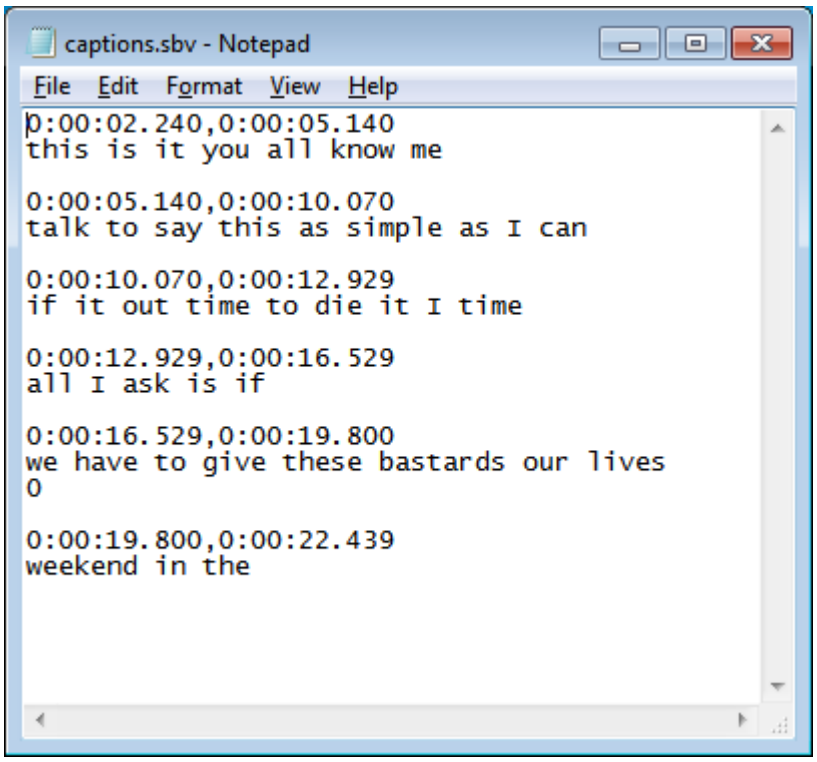
- By hand (listen and write)!?
- Automated speech recognition, YouTube does this (see below).
- Outsource, *e.g.* to Mechanical Turk?

Why would you do this? Consider http://moviesfortheblind.com/

Below is a small excerpt (27seconds) from 'The Matrix', at right you can see YouTube is autotranscribing (try it!):



https://www.youtube.com/watch?v=6wdxnw4J64Q (if you want to watch it).
And here's the result:



This is not too bad for (New Zealand) English. Small problem with the shouting/swearing.
What is the factor of reduction in size, from movie clip to transcript, and what is lost?

So now we've seen the scale of data and datasets, from a small transcript, to a movie, to a collection of movies (YouTube), and in various formats, and of varying quality, both fidelity and veracity. Consider with respect to the 4 Vs - what happens to data every time it's moved, copied, edited, merged, compressed, transcribed, translated, *etc*.? One description of the Internet is that it's like a library ... with all the books on the floor. Sometimes it's better, sometimes it's worse.

3.3
Data Case Studies

Medical informatics

Corporations tend to shy away from exposing their internal details for proprietary and competitive reasons. While healthcare has to keep their data private, they are happy to reveal the kinds of data they have and what they do with it. Therefore, we have assembled here some detailed material about medical informatics that makes a good in-depth case study that all should have a look at.

To understand the business and operating context:

- "When Health Care Gets a Healthy Dose of Data" (http://sloanreview.mit.edu/case-study/when-healthcare-gets-a-healthy-dose-of-data) from *MIT Sloan Management Review* (8000 words, 20 mins)  , long but gives a full background **NB**. to obtain you are required to become a "member" of the SMR get a free account

To understand the kind of data one has in a hospital:

- an extensive report on "Medical Data" (https://www.alexandriarepository.org/wp-content/uploads/20150701214439/MedicalData-v1.pdf) prepared by Bahadorreza Ofoghi (PDF, 2800 words, 15 mins)

Note as additional background, various interviews were given in other sections that included Dr. Chris Bain of Alfred Health:

- Module 1, **Roles of a Data Scientist**; Module 2, **Analysis and Interviews**; Module 4, **Interviews on Standards and Tools**; Module 5, **Data Analysis Case Studies**; Module 6, **Interviews on Data Management**

There are a huge range of developments here. For reference (not study) IEEE has a recent collection of articles around data, medicine and health.

- "Special Report: (http://spectrum.ieee.org/static/hacking-the-human-os) Hacking the Human OS" (http://spectrum.ieee.org/static/hacking-the-human-os) in *IEEE Spectrum* in May 2015 (use Monash Library portal, search for *IEEE Spectrum*, and locate Vol 52 Issue 5 for all the content).

The Enron email corpus

What is the Enron Corpus?

The Enron Corpus (https://en.wikipedia.org/wiki/Enron_Corpus) is a collection of real emails communicated among real individuals and organizations from a single business. It is perhaps the first such corpus made available.

The Enron corpus was first made available to the public during the legal investigations into Enron Corp. led by the Federal Energy Regulatory Commission in the US in May 2002 during a process called

electronic discovery (https://en.wikipedia.org/wiki/Electronic_discovery). The data set contains in excess of 619,000 email messages among Enron employees as well as others at external organizations. Most of the email messages in the Enron data set were communicated between senior managers of the Enron Corp. Emails in this collection contain the email addresses of the sender and recipient/s, message body, date, time, and subject of the email. Attachments have been excluded from the first publicly available Enron corpus; however, they can now be accessed from the Electronic Discovery Reference Model's website (http://edrm.net/).

After requests made from a number of affected Enron employees for their privacy, a smaller version of the Enron data set has been made available online by William Cohen from Carnegie Mellon University. In William Cohen's smaller Enron data set version, invalid email addresses were simplified. There is also a relational version of the Cohen's Enron data set that has been made available in the MySQL format. This database was created by Andrew Fiore and Jeff Heer at University California, Berkeley (http://bailando.sims.berkeley.edu/enron_email.html). A substantial amount of processing has been performed by the database creators to remove duplicates and normalize names.

An even smaller version of the Enron data set has been generated and titled The EnronSent Corpus by Will Styler. This corpus is available at http://savethevowels.org/enronsent/ and according to Will "it has been cleaned specifically for use with conventional corpus linguistics tools (such as grep, python), and an attempt has been made to remove as much non-human generated text as possible from the raw messages in the original data".

Data analysis with the Enron dataset

The Enron data set has been the subject of a number of different data analytics projects in recent years. Two most prevalent types of analysis on this data set include:

1. Social network analysis and visualization, and
2. Thematic structure analysis (better known as topic modelling in the natural language processing domain).

The social network analysis and visualization work on the Enron data set involves visualization of communication flow to mainly understand who sent an email to whom. One of the first such analyses was carried out using the software tool Enron Corpus Viewer (http://hci.stanford.edu/~jheer/projects/enron/) (follow link to obtain images of the system). Enron Corpus Viewer performs the social network visualization on the individuals who communicated messages in the Enron data set. In the network visualization pane, each user is shown by a circle and each link between two user circles represents an email communication. Some email meta data, such as sender and recipient information is shown in the pane at the right hand side. Enron Corpus Viewer also has an e-mail message viewer and support for user community structure analysis with the aim of understanding which groups of users have had lots of email communications with each other.

A more dynamic visualization of the Enron data set can also be found in the video:

- "Enron Email Network Visualization" (https://www.youtube.com/watch?v=F_w_buGM3No) by Michael Griscom (Youtube, 4 mins)

This visualization also shows users with dots and email communications with edges or links between them. The varying dimension in this video is the time of communications.

Other analysis, visualisation of the graphs, that you can optionally review includes:

- "Hub and Spoke 'theory' of illicit projects" (https://www.sciencenews.org/article/information-flow-can-reveal-dirty-deeds), on ScienceNews (600 words, 4 mins)
- "Plots of edge distributions" (http://konect.uni-koblenz.de/networks/enron) at Uni-koblenz.de (various figures, 5 mins)

NIST use cases

NIST's Big Data Working Group has done analysis of use cases to look at the various characteristics of data, mostly described in their Volume 3, Use Cases and General Requirements (http://dx.doi.org/10.6028/NIST.SP.1500-3).

The use cases cross application boundaries, health, manufacturing, science and so forth. A number of them are "big science" so are specialised views of Data Science. The case studies are listed below as a handy example of understanding data sets. There are too many of these for it to be worthwhile looking at them all. You should consider an area of interest to you (science, defence, healthcare, government, ...) and review a few of these. The main point of this is to understand the analysis they do.

They are as follows:

- Government Operation
  - Census 2010 and 2000 - Title 13 Big Data; National Archives and Records Administration Accession NARA, Search, Retrieve, Preservation; Statistical Survey Response Improvement (Adaptive Design); Non-Traditional Data in Statistical Survey Response Improvement (Adaptive Design)
- Commercial
  - Cloud Eco-System, for Financial Industries (Banking, Securities & Investments, Insurance) transacting business within the United States; Mendeley - An International Network of Research; Netflix Movie Service; Web Search; IaaS (Infrastructure as a Service) Big Data Business Continuity & Disaster Recovery (BC/DR) Within A Cloud; Eco-System; Cargo Shipping; Materials Data for Manufacturing; Simulation driven Materials Genomics
- Defense
  - Large Scale Geospatial Analysis and Visualization; Object identification and tracking from Wide Area Large Format Imagery (WALF) Imagery or Full Motion Video (FMV) - Persistent Surveillance; Intelligence Data Processing and Analysis
- Healthcare and Life Sciences
  - Electronic Medical Record (EMR) Data; Pathology Imaging/digital pathology; Computational Bioimaging; Genomic Measurements; Comparative analysis for metagenomes and genomes; Individualized Diabetes Management; Statistical Relational Artificial Intelligence for Health Care; World Population Scale Epidemiological Study; Social Contagion Modeling for Planning, Public Health and Disaster Management; Biodiversity and LifeWatch
- Deep Learning and Social Media
  - Large-scale Deep Learning; Organizing large-scale, unstructured collections of consumer photos; Truthy: Information diffusion research from Twitter Data; Crowd Sourcing in the Humanities as Source for Big and Dynamic Data; CINET: Cyberinfrastructure for Network (Graph) Science and Analytics; NIST Information Access Division analytic technology performance measurement, evaluations, and standards
- The Ecosystem for Research
  - DataNet Federation Consortium DFC; The 'Discinnet process', metadata big data global experiment; Semantic Graph-search on Scientific Chemical and Text-based Data; Light source beamlines
- Astronomy and Physics
  - Catalina Real-Time Transient Survey (CRTS): a digital, panoramic, synoptic sky survey; DOE Extreme Data from Cosmological Sky Survey and Simulations; Large Survey Data for

Cosmology; Particle Physics: Analysis of LHC Large Hadron Collider Data: Discovery of Higgs particle; Belle II High Energy Physics Experiment;

- Earth, Environmental and Polar Science
  - EISCAT 3D incoherent scatter radar system; ENVRI, Common Operations of Environmental Research Infrastructure; Radar Data Analysis for CReSIS Remote Sensing of Ice Sheets; UAVSAR Data Processing, Data Product Delivery, and Data Services; NASA LARC/GSFC iRODS Federation Testbed; MERRA Analytic Services MERRA/AS; Atmospheric Turbulence - Event Discovery and Predictive Analytics; Climate Studies using the Community Earth System Model at DOE.s NERSC center; DOE-BER Subsurface Biogeochemistry Scientific Focus Area; DOE-BER AmeriFlux and FLUXNET Networks;
- Energy
  - Consumption forecasting in Smart Grids

They have developed a table summarising use cases with fields such as Volume, Velocity, Software, Analytics, Data Sources, *etc.*, in the Appendices of their long report.

Case studies

In this section, we provide a number of resources that can be used to study different kinds of data. The variety here is quite large. Some of the resources are quite extensive, and some are more incomplete and perhaps suggestive, giving you an area to consider and you may have to follow-up to find additional resources. These resources are for use in the Case Study activity with this module.

- "Visualizing the world's Twitter data - Jer Thorp" (http://ed.ted.com/lessons/mapping-the-world-with-twitter-jer-thorp), a TEDYouth 2012 Talk, former *New York Times* data artist-in-residence Jer Thorp (video, 6mins) 
- Foster Provost talks about "Is Bigger Really Better? Predictive Analytics with Fine-grained Behavior Data," (https://www.youtube.com/watch?v=1jzMiAfLH2c) (10 minute, Youtube, from 2013 Strata-Hadoop) discussing analytics for banking. 
- "Bringing Data Science to the Speakers of Every Language" (http://videolectures.net/kdd2014_munro_data_science), by Robert Munro of Idibon (video, 34 mins)
- *Internet of Things* is an emerging area, but there are many resources. Some of our favorite:
  - "What is the Internet of Everything (IoE)" (http://singularityhub.com/2015/04/21/the-internet-of-everything-a-19-trillion-opportunity/) by Peter Diamandis of the X PRIZE Foundation (900 words, 5 mins)
  - "How Big Will The Internet of Things Be?" (https://datafloq.com/read/how-big-will-the-internet-of-things-be/523) by Mark van Rijmenam of Datafloq (600 words plus infographic, 6 mins)
  - The blog entry "5 reasons why IoT needs data analytics" (http://www.internetofbusiness.co.uk/insight/2016/01/21/5-reasons-why-iot-needs-data-analytics/) by Barclay Ballard, jan. 2016 gives some other background.
  - "11 Top IoT Infographics to Help you Learn About the Internet of Things" (http://iotcompanydirectory.com/11-top-iot-infographic-designs/) at the IoT Directory website
  - A special web report (http://raconteur.net/internet-of-things-2016) in the Raconteur online magazine.
- The "National Map" (https://www.youtube.com/watch?v=6s_n9A7BARs) (Youtube, 2 mins)  is a website for map-based access to Australian spatial data from government agencies. The website is

http://nationalmap.gov.au/.

- Using Big Data to Understand the Human Condition: The Kavli HUMAN Project (http://kavlihumanproject.org/).
- Legal software to support classification (what topics are a document about?) in the legal domain, developed by FTI Technology, called Ringtail (http://www.ftitechnology.com/ringtail-ediscovery-software). One thing it does is "predictive coding", which is using machine learning tools to train a system to recognise what
  themes are in a document. A related bit of software without the machine learning or statistical prediction, but supporting
  all the tedious annotation and some linguistic analysis is NVivo from QSR International (http://www.qsrinternational.com/nvivo-products).
- "Why smart statistics are the key to fighting crime" (http://www.ted.com/talks/anne_milgram_why_smart_statistics_are_the_key_to_fighting_crime) a TED talk by Anne Milgram (video, 13 minutes or 2300 words transcript)  .
- "The big idea my brother inspired" (https://www.ted.com/talks/jamie_heywood_the_big_idea_my_brother_inspired) a TEDMED talk by Jamie Heywood (video 17 mins or 3400 words transcript).
- "Coca-Cola's Unique Challenge: Turning 250 Datasets Into One" (http://sloanreview.mit.edu/article/coca-colas-unique-challenge-turning-250-datasets-into-one/), from *MIT Sloan Management Review* (2000 words, 11 mins, has a firewall, to read full article you need to register)
- "Helping the Republican Party Use Data and Engineering to Win the US Senate" (http://player.oreilly.com/videos/9781491924143?toc_id=210834) by Azarias Reda from *Strata+Hadoop World* 2015 (video, 35 mins) 
- "Style Stalking; The Stochastic Patterns that Drive Fashion Trends" (http://player.oreilly.com/videos/9781491900345?toc_id=192993), by Karen Moon from *Strata+Hadoop World 2014* (video, 10 minutes) 
- Open Internet comments: the Federal Communications Commission asked for public comments about the issue of keeping the Internet free and open with comments closing 19 July 2014. The response was huge.
  - Blog post "A Fascinating Look Inside Those 1.1 Million Open-Internet Comments" (http://www.npr.org/blogs/alltechconsidered/2014/08/12/339710293/a-fascinating-look-inside-those-1-1-million-open-internet-comments) by Elise Hu from NPR (700 words, 4 mins)
  - "What can we learn from 800,000 public comments on the FCC's net neutrality plan?" (https://sunlightfoundation.com/blog/2014/09/02/what-can-we-learn-from-800000-public-comments-on-the-fccs-net-neutrality-plan/) a statistical analysis and visualisation by Bob Lannon and Andrew Pendleton on sunlightfoundation.com's blog (2200 words, 12 mins) 
  - The official website with the comments https://www.fcc.gov/files/ecfs/14-28/ecfs-files.htm; the source files are huge, so do not download, the data format is described.

Note Wikipedia's big data (applications) (https://en.wikipedia.org/wiki/Big_data#Applications) entry also mentions a good number of successful applications too.

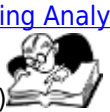
3.4
Big Data Processing

This section presents the different kinds of databases and computation used to handle big data.

Business imperatives

Classically, business data such as corporate transactions was stored in the form of files on file systems. Initially, data could be stored in text files or in binary form and processed sequentially. When data became larger and more prevalent, the need arose for a standardised way to store and access data. This led to the invention of the Relational Database Management System (RDBMS (https://en.wikipedia.org/wiki/Relational_database_management_system)). RDBMSs store data with associated schemata that described the data and allowed access to it via a standardised Structured Query Language (SQL) (https://en.wikipedia.org/?title=SQL).

More recently, however, this paradigm has hit several walls:

- **Businesses function in a continuously changing environment:** This environment, reflected in database schemata, must therefore continuously change. The more data an organization has, the more onerous the task of schema adaptation. In the world of big data, the level of effort to maintain a classic RDBMS has become unmanageable. One solution has been to "go schema-less" in one of several types of NoSQL databases (https://en.wikipedia.org/wiki/NoSQL).
  - "What is NoSQL Database?" (https://www.youtube.com/watch?v=pHAItWE7QMU) by Hasan Mir (Youtube, 8 minutes) 
- **Businesses are moving to data driven decision-making:** This requires complex analytical queries to be run on massive amounts of data, a usage pattern never envisioned by the designers of SQL. One example of a NoSQL database management method optimised for a specific type of such complex analytical queries, namely queries regarding data interaction patterns, is the graph database (https://en.wikipedia.org/wiki/Graph_database).
  - "How does a graph database differ from a relational database?" (https://www.youtube.com/watch?v=41qdmKIIMz0) by David Mizell (Youtube, 3.5 minutes) 
- **Need to reach insights faster and act on them in real time:** The store-first-query-later methodology is unable to cope with this need, and tools had to be developed to handle data in real time. This kind of data is called a stream (https://en.wikipedia.org/wiki/Stream_%28computing%29). This has led to streaming analytics.
  - "Streaming Analytics" (https://www.youtube.com/watch?v=H4Br7iOUkko) from Software AG (Youtube, 3.5 minutes) 

New types of databases

So with different database types evolved, like NoSQL and graph databases, RDBMSs themselves had to evolve. Initially these were designed to work on large mainframe machines, as storage requirements rose,

the solution of always enlarging the machine's storage and its processing power ceased to be viable. There is a distinction between **scalable** (https://en.wikipedia.org/wiki/Scalability) systems and large scale systems. Scalable systems can keep expanding. As more capacity is needed, it can be added:

> ***Scalability*** is the ability of a system, network, or process to handle a growing amount of work in a capable manner or its ability to be enlarged to accommodate that growth.

The first step taken by most organisations to mitigate the effect of ever increasing storage was to separate their data according to its type, storing it in a large number of RDBMSs, usually one per department. Today, this interim solution is among the greatest stumbling blocks for organisations on their way to harnessing the power of data, as it means that **data is siloed** (https://en.wikipedia.org/wiki/Information_silo), storage methods are inconsistent across an organisation, data is not accessible in ways that are conducive to analysis outside its primary storage purpose, and data scientists spend much of their time "wrangling" the data.

A more modern solution to the problem of data silos is the use of distributed databases, allowing one database to be stored across many (potentially small and cheap) boxes.

- "Distributed Databases" (https://www.youtube.com/watch?v=ah_tJg4sR5U) by Harifa Ahmed (animation, 2 minutes) 

With the right combination of a distributed database and software tools, this allows for distributed analytics.

- "Distributed Analytics: A Primer" (http://thomaswdinsmore.com/2014/05/01/distributed-analytics-primer/) by Thomas W. Dinsmore (blog, 1000 words, 5 minutes) 

Analysis of this kind, which does not require one to move data out of the database for it to be analysed, is known as in-database analytics (https://en.wikipedia.org/wiki/In-database_processing).

Another direction in which RDBMSs had to evolve is regarding their speed and ability to scale up to millions of simultaneous users. Consider the volumes of transactions that need to be handled by an online shop such as Amazon's. Even without the need to handle complex analytical queries of the sort discussed above, latency had to be reduced dramatically. This led to the advent of **in-memory databases** (https://en.wikipedia.org/wiki/In-memory_database). Once it was viable to store limited amounts of data in memory and analyze it as such. Much of the traditional operating procedures of data science has been geared towards this (*e.g.*, sample from your on-disk database in an amount that fits your memory, then analyze using R). The more modern approach is for in-memory databases to be distributed, replicated and persistent, and in-memory analytics tools need to run, correspondingly, on distributed memory.

Most recently, processing trends have focused on creating platforms that incorporate multiple paradigms under one roof, allowing for more flexibility: decisions on how to handle the data can be made at analysis time, rather than at storage time. The *de facto* standard in this area is Hadoop

- "What is Big Data and Hadoop?" (https://www.youtube.com/watch?v=FHVuRxJpiwI) from Learning Tree International (Youtube, 8 minutes) 

for distributed storage, either using Hadoop's Map-Reduce-based ecosystem or just using Hadoop's HDFS. An example of the latter is Spark:

- "What is Apache Spark" (https://www.youtube.com/watch?v=cs3_3LdCny8) by Hasa Mir (Youtube, 2.5 minutes)

Spark provides its own ecosystem of tools (Spark SQL for relational data, Spark Streaming for streamed data, MLlib for machine learning, GraphX for graph analytics), independent of the Hadoop ecosystem.

Today, users expect their data to be not only distributed, but distributable in a cloud.

Despite major trends and apparent convergences in this space, it is important to note that every existing platform has limitations that make them unsuitable for particular businesses. For example, even though Hadoop is sometimes considered synonymous with Big Data, Map-Reduce is not used by, for example, Google, the very company that initially pushed it into existence.

- "Google Dumps MapReduce in Favor of New Hyper-Scale Analytics System" (http://www.datacenterknowledge.com/archives/2014/06/25/google-dumps-mapreduce-favor-new-hyper-scale-analytics-system/) by Yevgeniy Sverdlik (blog, 400 words, 2 minutes)

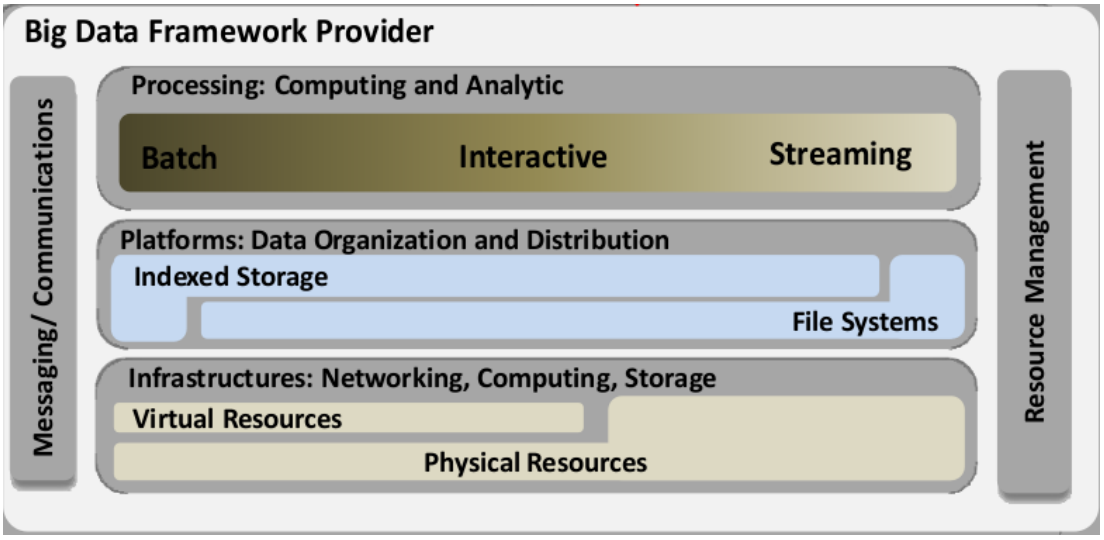
The big data processing world continues to be in flux, with new trends and technologies appearing in it daily.

NIST's computational framework

NIST provides a framework for the computational aspects of big data processing.

- "NIST Big Data Interoperability Framework: Volume 6, Reference Architecture" (http://dx.doi.org/10.6028/NIST.SP.1500-6) by NIST Working Group on Big Data (PDF 60 pages total) see Section 4.4 only (16 pages in all)

There overall framework is summed up by the *Big Data Framework Provider* figure.



NIST Big Data Framework Provider (extracted from Fig. 2 in "Volume 6: Reference Architecture")

Suggested Reading

- "Udemy Industry Insights: Hadoop" (https://www.udemy.com/hadoop-tutorial/#interview) 22min audio tutorial and infographic on Hadoop.
- "Bi isn't big data and big data isn't BI" (http://www.slideshare.net/mrm0/bi-isnt-big-data-and-big-data-isnt-bi) is a thought provoking presentation on the topic of big data and big data processing. There is a 3000 word transcript at the bottom of the page should you wish to read what was said with the 76 slides. Probably a one hour talk originally. Assumes familiarity with a lot of the material we have covered.
- "A Scalability Analysis on Big Data" (http://fatihhamurcu.blogspot.com.au/2015/02/scale-your-vision-about-scalability.html) by Fatih Hamcurcu for a more extensive discussion of **scalability**
- http://hadoop.apache.org/ - The Hadoop homepage
- http://spark.apache.org/ - The Spark homepage
- "Pivotal: A New Platform for a New Era" (https://www.alexandriarepository.org/wp-content/uploads/20150623133745/Spark_Training_Public_Facing.pdf) training slides from Pivotal (see slides 6-20, 60-65 for comparisons and discussions).
- "Transactional streaming: If you can compute it, you can probably stream it," (https://www.oreilly.com/ideas/transactional-streaming-if-you-can-compute-it-you-can-probably-stream-it) blog entry by John Hugg on O'Reilly, Jan. 2016, goes into more detail on streaming.
- "The world beyond batch: Streaming 101," (https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-101) long long blog on O'Reilly by Tyler Akidau in Aug. 2015 covering streaming ... worth a glance at the beginning at least to see what it is about.

4
Data Resources, Processes, Standards and Tools

This is our fourth module of six for the Introduction to Data Science unit. This module will investigate some of the major resources available for a data science project including analysis and visualisation tools, development environments, different kinds of government and free data resources as well commercial data. Moreover, some examples will be given of the creative combination of resources to address complex problems. Finally, some standards, as far as they exist, will be presented.

(https://www.youtube.com/watch?v=cE5izu8nR14)

Aims of This Module

- Identify and describe the kinds of resources, software and data, that can be available for a project.
- Describe the importance of data fusion, and why it is a driver for finding resources.
- Identify the major standards available and platforms (iPython, R).
- Analyse case studies of using/combining data resources.
- Describe issues with open data.

How to study for this module

In this module we again draw on material in the public domain such as interviews and videos, online magazine entries and blogs. We have also written some material to tie together various kinds of models. As well as studying and viewing the material, we have some activities around this material.

Please remember:

- Reference items marked with a single "johny look it up" icon,  , should be viewed as *suggested reading*, not essential nor important for assessment.
- Reference items marked with a two "johny look it up" icons,  ,  should be viewed as *important reading*, considered important for assessment.

4.1
Introduction to Resources

This section gives some examples of the creative use of data, data fusion, open data, and the industry need for general standards.

Using data

One major source of creativity in the Data Science world is thinking of different ways of combining existing data sets or producing new data sets to produce "value." Sometimes new companies are started, and other times a new app appears on mobile phones. This is not Data Science in the usual business sense (*i.e.*, about doing data-drive decision making in marketing, distribution, human-resources, *etc.)* Clever use of data is often times not about big data, and sometimes there is no real statistical analysis. We will have a look at some instances of this sort of creativity here.

GovHack

One example in the creativity ethos is the GovHack (https://www.govhack.org/) competitions, which are hackathons run on government data in Australia and New Zealand. On its website GovHack says of itself:

> GovHack is an event that draws people together to innovate with Open Government Data. The best teams have a mix of skill so we encourage every one to come along including entrepreneurs, developers, designers, digital media creators, artists, story tellers, researchers and open data enthusiasts.

Some descriptions are:

- An extensive write-up describing some of the data sources, "GovHack: All your data are belong to us" (http://csironewsblog.com/2015/07/03/govhack-all-your-data-are-belong-to-us/), from CSIRO (article, 650 words, 4 mins)  also giving some of their own hacks
- "The 10 industries with the worst gender pay gap revealed" (http://www.smartcompany.com.au/people/47571-the-10-industries-with-the-worst-gender-pay-gap-revealed.html) in SmartCompany (article, 550 words, 4 mins) describes one of the results.
- One of many newspaper covers, a bit less on detail, "Helpful 'hackers' to decode government data jumble" (http://www.sunshinecoastdaily.com.au/news/helpful-hackers-decode-government-data-jumble/2667552/) from Sunshine Coast Daily (article, 400 words, 2 mins)

Using government data in New York City

This is really an entertaining story about the availability of government data about New York City. Now the data is not always easy to get, but when you do get it it invariably helps to map it.

- "How we found the worst place to park in New York City - using big data" (http://www.ted.com/talks/ben_wellington_how_we_found_the_worst_place_to_park_in_new_york_city_using_big_data?language=en)

  by Ben Wellington, a TEDXNewYork talk, using NYC gov data (12 mins or 2600 words) 

Data fusion for traffic prediction

In a larger talk by Eric Horvitz at KDD 2014 (http://www.kdd.org/kdd2014/) (a data mining conference), he mentions a number of problems.

In one section he describes how a Microsoft team looked at the traffic problems in Seattle. This is a great example of prescriptive analytics because they have to combine prediction with optimisation to make use of their many different data sources.

- "Data, Predictions, and Decisions in Support of People and Society," (http://videolectures.net/kdd2014_horvitz_people_society) Eric Horvitz of Microsoft giving a 40 minute wide-ranging technical talk on data science with theme of "social good".
    - see the traffic section at 7.44-11:06 minutes 

The key issue here is how the combination of different data is used in their task.

Web-scale pharmacovigilance

In another study, a team with Eric Horvitz of Microsoft used pairs of drugs and were interested in testing if the drugs, when taken together by a patient, tended to cause hyperglycemia. Now the problem being addressed here is drug interactions, where a combination of drugs causes problems not seen with just one or the other. The presentation skips over a lot of details, so while viewing the video, follow the separate explanation below which is linked to the times.

- "Data, Predictions, and Decisions in Support of People and Society," (http://videolectures.net/kdd2014_horvitz_people_society) Eric Horvitz of Microsoft, see the web-scale pharmacovigilance section at 38:43-43:30 minutes, and see explanatory notes below.

Details of this analysis are as follows:

- defines a score called reporting ratio (RR) at 40:40 minutes
- computes this score for 62 pairs of drugs for which ground truth of "interactions causing hyperglycemia" has been reported elsewhere
- for different "cut off" values of the score:
    - *predict* which drugs have interactions causing hyperglycemia
    - compare predicted versus actual to get false positive (https://en.wikipedia.org/wiki/False_positives_and_false_negatives) and true positive (https://en.wikipedia.org/wiki/Sensitivity_and_specificity) rates for this cutoff as a pair (FP,TP)
    - plot the point (FP,TP)
    - this is an example of what is called ROC analysis (https://en.wikipedia.org/wiki/Receiver_operating_characteristic) (you do not really have to understand ROC to understand the logic here, but its an important concept for future, and well explained in this longish article)
- the resulting plot is given at 41:14 minutes, this shows how efficient "reporting ratio from web queries" is as a signal to predict if a pair of drugs causes hyperglycemia
    - *e.g.*, a high cutoff means
- the same logic is used to compute similar signals for acute renal failure, upper GI bleed, acute liver injury and acute myocardial infarction, again from web queries; this is plotted in blue at 41:59 minutes
    - the comment about the "multi-item Gamma Poisson shrinker" can be ignored; it is a data pre-processing step to make the data more robust

- a complementary study is done with AERS data, which is data from FDA's Adverse Event Reporting System (http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/) which comes from doctor's reports; this is plotted in red at 41:59 minutes
- one can see that the AERS-derived RR signal is complementary to the WEB-derived RR signal, so he then fuses the two signals to show, at 43:12 minutes, that they combine to produce an even better signal.

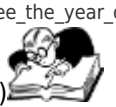
So there are a number of aspects to this:

- the "true" data for the drug interaction problem is available from AERS;
- web queries are used as "proxy data" (https://www.ncdc.noaa.gov/news/what-are-proxy-data), a term coming from environmental science, *e.g.*, "tree ring width as a proxy for rainfall"; we do not know how well the proxy data performs until we calibrate it with known truth;
- fusing the proxy data with the true data provides an even better prediction;
- the basic ideas here are quite simple: no sophisticated data processing, no sophisticated statistics (other than the shrinker).

The reference for the journal article behind the talk is "Toward Enhanced Pharmacovigilance using Patient-Generated Data on the Internet," (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4111778/) by White, Harpaz, Shah, DuMouchel and Hortvitz 2014.

Open data

A complementary development to the World Wide Web pioneered by Prof. Sir Tim Berners-Lee is the Semantic Web (https://en.wikipedia.org/wiki/Semantic_Web). While this is not considered successful in the same way, the semantic web has lead to the development of properly structured databases and knowledge bases, properly stuctured in the sense that their metadata is well developed and they are machine-readable in the sense discussed in the section **Characterizing Data** in the module **Data Types and Storage**. The best use of this technology is in the Open Data movement (https://en.wikipedia.org/wiki/Open_data), presented in the following:

- "The year open data went worldwide" (http://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide/) a TED talk by Prof. Sir Tim Berners-Lee (video, 6 mins or 1000 words) 
- "Linked Open Data" (https://vimeo.com/36752317) (video on Vimeo, 4 mins) 
- "The New Data Republic: Not *Quite* a Democracy" (http://sloanreview.mit.edu/article/the-new-data-republic-not-quite-a-democracy) by Sam Ransbotham, *MIT Sloan Management Review* (900 words, 5 minutes)

Data wrangling

Trifacta, a company providing software to support data wrangling, provide a number of good videos describing the data science process, with particular focus on data wrangling. They stress the fact that wrangling is one of the most time consuming tasks, and in the spirit of data science as a rapid prototyping task, they see efficient support for data wrangling as an important goal.

- "Trifacta Leads in the Wisdom of Crowds Study on End User Data Preparation" (https://www.youtube.com/watch?v=Ugy-PSWKCK0) by Trifacta (Youtube, 48 mins); this is a long sales pitch but

starts with a good description documenting the importance of wrangling in organisations, followed by an extensive demonstration of Trifacta software. We do not need to know all the details of this.

- "Using Learning to Accelerate Data Wrangling" (https://www.brighttalk.com/webcast/12327/157621) by Trifacta (video, 60 mins, but only view the first 15 mins or so)  , subsequent content goes into far more detail about their product then we need to know currently. **Warning: only view this video if you are willing to register.**

A call for standards

So with the more wide-scale availability of data, and the growth of Data Science as an academic and a professional area, the need for standards arises. The issues and some examples are presented by Kirk Bourne.

- "Raising the Standard in the Big Data Analytics Profession" (https://www.mapr.com/blog/raising-standard-big-data-analytics-profession) by Dr. Kirk Bourne (blog, 1100 words, 6 mins) 

4.2
Activity: Data Wrangling with SAS

Wrangling Data with SAS Studio

Step 1

"Patients.txt" is a small, messy, medical data file. The first two lines are:

001M11/11/1998 88140 80 10
016F11/13/1998 84120 78 X0

Some of the data here is obvious, there's a date in there: '11/11/1998'. Some is not so obvious, but you can guess, 'M' or 'F' could be Male or Female. Some is mysterious... it's patient data, there are 8 variables:

VariableName	Description	VariableType	ValidValues	e.g.
PATNO	Patient Number	Character	Numerals	016
GENDER	Gender	Character	'M' or 'F'	
VISIT	Visit Date	MMDDYY10.	Any valid date	
HR	Heart Rate	Numeric	40 to 100	
SBP	Systolic Blood Pressure	Numeric	80 to 200	
DBP	Diastolic Blood Pressure	Numeric	60 to 120	
DX	Diagnosis Code	Character	1 to 3 digits	
AE	Adverse Event	Character	'0' or '1'	

From left to right we have Name, Description, Type and Values (or range). Heart Rate (HR) for example is Numeric, Range is 40-100. What do you think the variable type for VISIT means (MMDDYY10.)?

Refer to the second record of data (above), identify the values and complete the table above (the e.g. column).
PATNO has been done.

Step 2

Print out or open the *Patients.txt* (https://www.alexandriarepository.org/wp-content/uploads/20150629135742/Patients.txt) (you can open it with a text editor e.g., WordPad or NotePad, but do not change it). Scan the data and look for any problems, make a note of line # or ID and values, aim for 10 or more. Problems might include:

- no data given for a variable, which is called "missing data"
- illegal but repairable data for a variable
- *etc.*

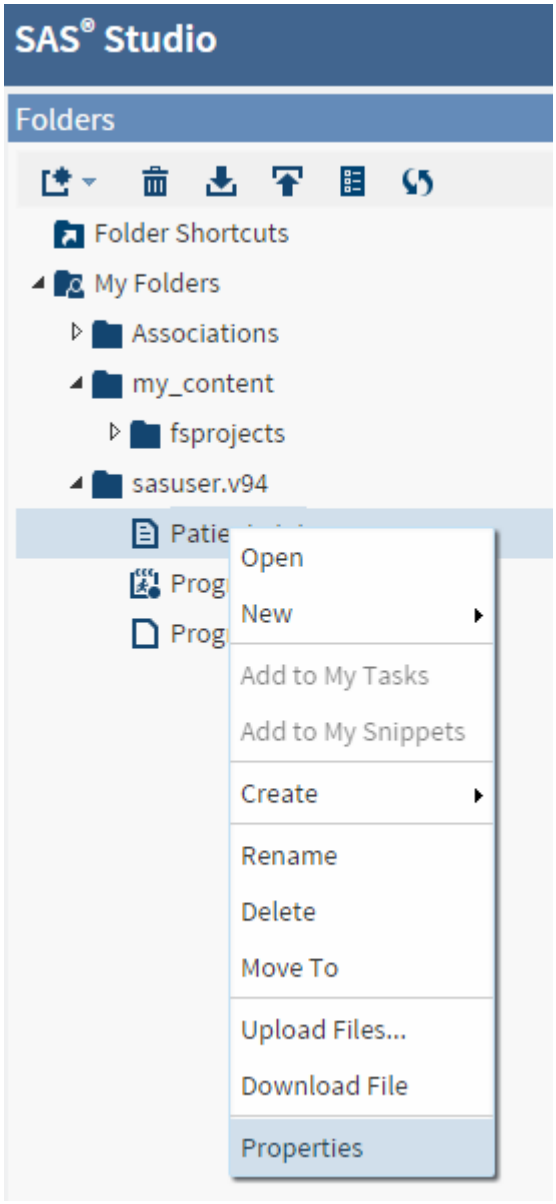
Now classify these errors, aim for 3 or 4 types of error (it is essential that you 'know your data').

Step 3

Run SASStudio https://odamid.oda.sas.com/SASStudio, login and upload the *Patients.txt* (https://www.alexandriarepository.org/wp-content/uploads/20150629135742/Patients.txt) file (see menu option 'Upload files' below, you can also see the file is in the sasuser folder)

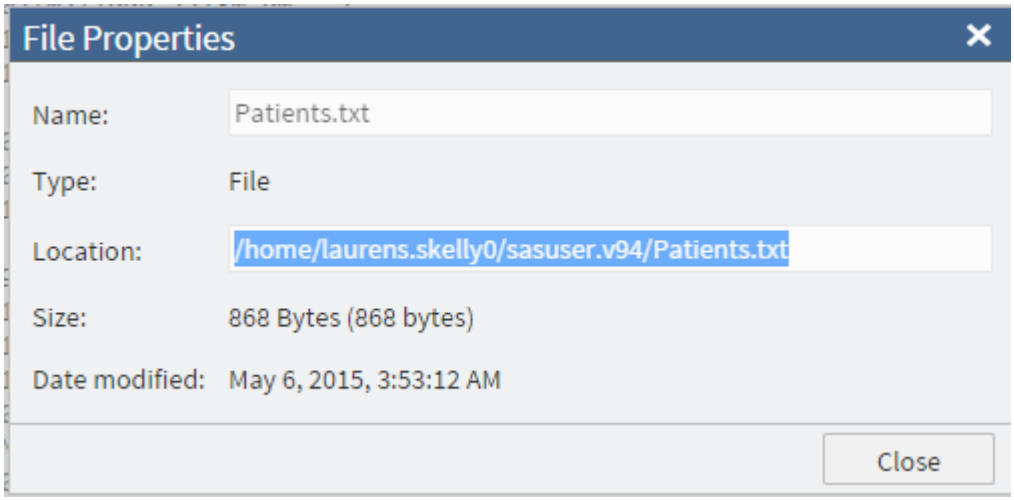
Step 4

Right click on the file to get the properties (you can also select 'Open' to view the data).



Step 5

Select 'Properties' then copy the entire 'Location' path (as below), this tells us where the raw data is



Step 6

Import the file into SAS (see "Enter your code here", use F3 to run) using the following program example, you will have to use your own path from above in line 2:
infile "home/..." etc.

```
data patients;
infile "/home/laurens.skelly0/sasuser.v94/patients.txt" PAD; *** use your own
path here;
input @1 patno $3.
@4 gender $1.
@5 visit mmddyy10.
@15 hr 3.
@18 sbp 3.
@21 dbp 3.
@24 dx $3.
@27 ae $1.;
label patno = "patient number"
gender = "gender"
visit = "visit date"
hr = "heart rate"
sbp = "systolic blood pressure"
dbp = "diastolic blood pressure"
dx = "diagnosis code"
ae = "adverse event?";
format visit mmddyy10.;
run;
proc print data=patients;
run;
```

Starting at line 3 ("input..") are commands to describe the data format (location, name, type).
What data type is Gender, what size, and where does it occur in a line of data?
What data type is HR, what size, and where does it occur in the data?
How many records were imported? Where is the data saved?
The last two lines print the imported data (see the 'Results' tab below):

CODE | LOG | RESULTS

Obs	PATNO	GENDER	VISIT	HR	SBP	DBP	DX	AE
1	001	M	11/11/1998	88	140	80	1	0
2	016	F	11/13/1998	84	120	78	X	0
3	033	X	10/21/1998	68	190	100	3	1
4	004	F	01/01/1999	101	200	120	5	A

Already we can see some problems, there's an 'X' in the Gender column, you could fix this manually using a text editor for example, but there may be others and there are tools available in SAS.

Step 7

Browse/view data using FREQ function e.g. look at the variable GENDER using this code:

```
proc freq data=patients;
title "frequency counts";
tables gender / nocum nopercent;
run;
```

You can paste the above 4 lines of code below your previous code, then select (highlight), then run:

```
28 run;
29
30 PROC FREQ DATA=Patients;
31 TITLE "FREQUENCY COUNTS";
32 TABLES GENDER / NOCUM NOPERCENT;
33 RUN;
```

You can see a summary of an attribute (Gender) for the entire data set:

FREQUENCY COUNTS

The FREQ Procedure

GENDER	
GENDER	Frequency
2	1
F	12
M	13
X	1
f	2
Frequency Missing = 1	

We can see 12 'F' and 13 'M' but also a few problems. Which ones should you fix?

Step 8

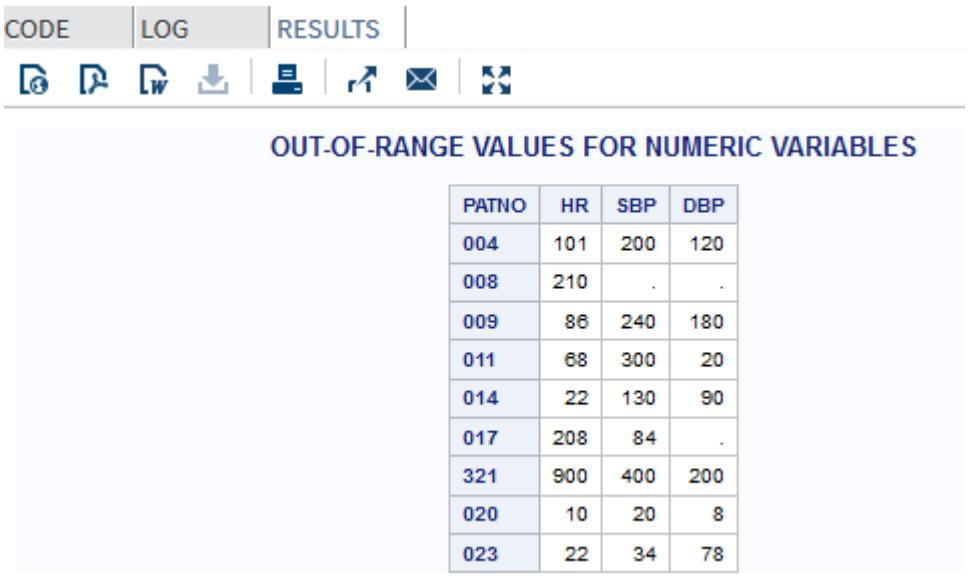
Do the same summary for the AE and DX variables, how many missing, how many invalid?

Step 9

List data ranges:

```
proc print data=patients;
where hr not between 40 and 100 and
hr is not missing or sbp not between 80 and 200 and
sbp is not missing or dbp not between 60 and 120 and dbp is not missing;
title "out-of-range values for numeric variables";
id patno;
var hr sbp dbp;
run;
```

How many missing values are there?
How many suspicious data values are there?



OUT-OF-RANGE VALUES FOR NUMERIC VARIABLES

PATNO	HR	SBP	DBP
004	101	200	120
008	210	.	.
009	86	240	180
011	68	300	20
014	22	130	90
017	208	84	.
321	900	400	200
020	10	20	8
023	22	34	78

There is not much you can do about 'bad data' (like 900) but some can be repaired:

Step 10

To change the 'f' to 'F' we could use: 'gender = upcase(gender);' or do it when reading the data in Step 6 above: '@4 gender $upcase1.' or do this, add the following below your previous code, select it, then run:

```
data cleanpatients changed;
set patients;
if gender = 'f' then gender = 'f';
output cleanpatients ;
```

```
run;
```

Where is the data saved now?
How can you demonstrate that data has been repaired?

Step 11

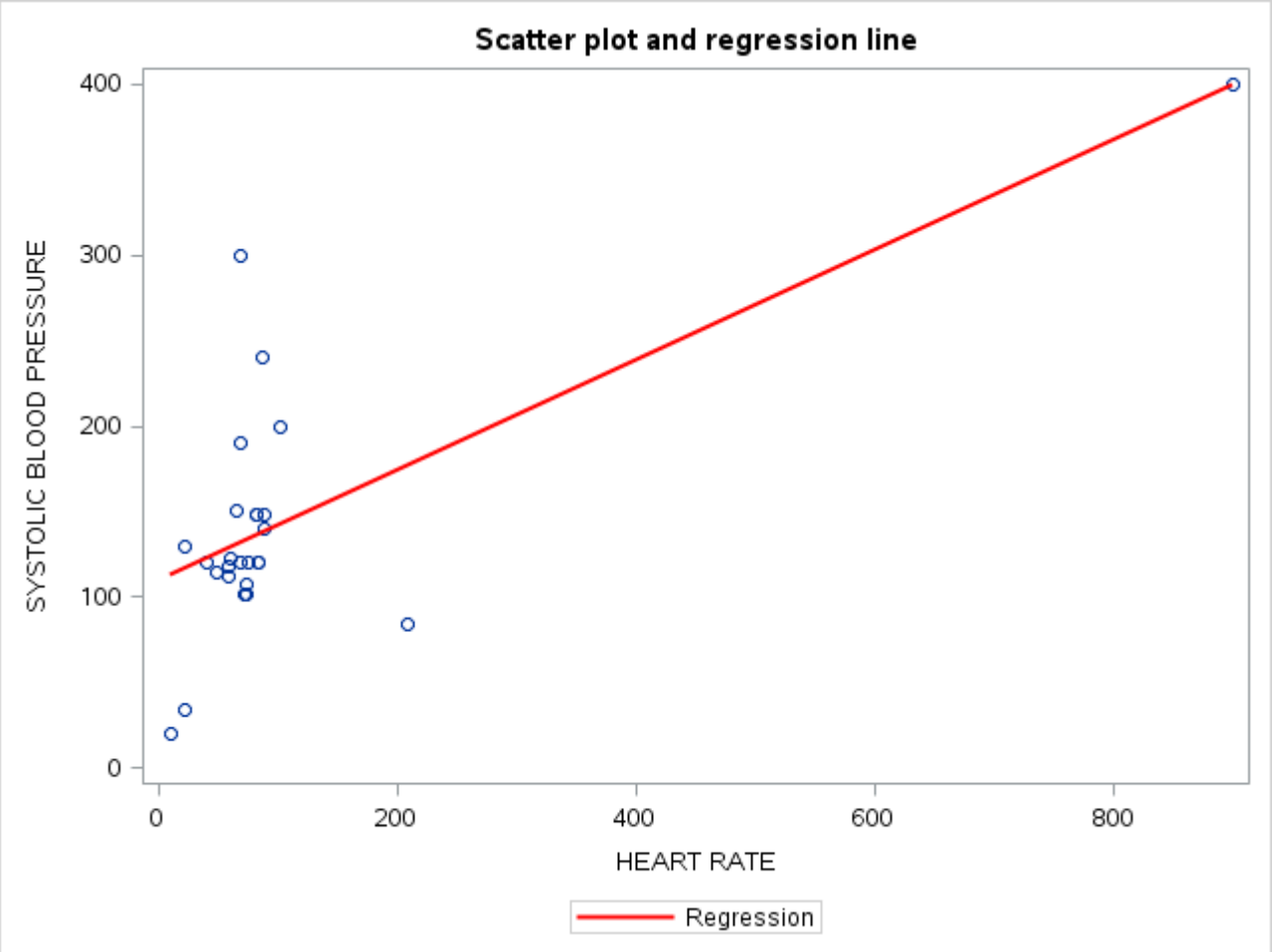
There is other data that we can argue is fixable, e.g., the date '98. What could it be and why?
How about 'XX5'?

Of course much of this is preventable, what could you do (at the data entry stage) to prevent many of these problems?
(If you haven't come across them yet research 'Regular expressions').

Step 12

Another way to explore data is with a plot:

```
proc sgplot data=patients;
reg x=dbp y=sbp/ lineattrs=(color=red thickness=2);
title "Scatter plot and regression line";
run;
```



You can see what bad data does to data analysis, that HR 900 has to go:

```
data cleanpatients;
set patients;
if hr <= 200;
run;
```

Where is the cleaned data now? How many records?
Plot again to see current state.
What are the pros and cons of deleting vs. ignoring 'bad' data?

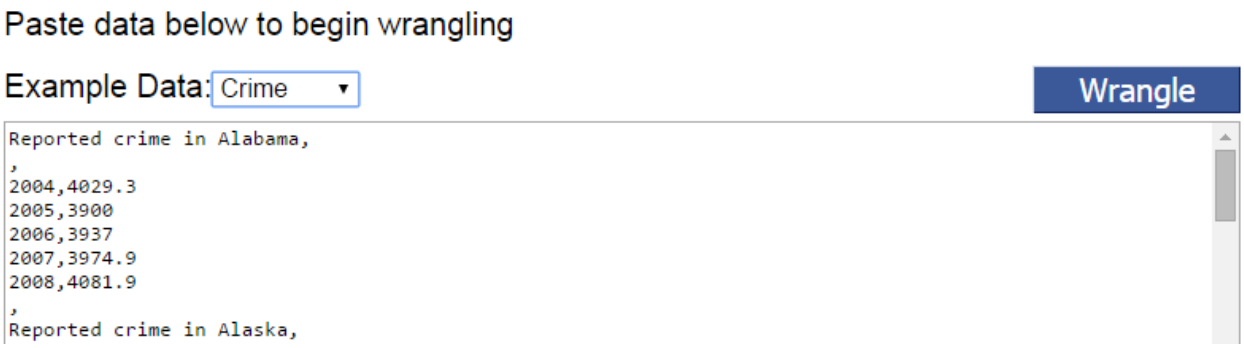
4.3
Activity: DataWrangler

Wrangling Data with DataWrangler

DataWrangler is a Stanford project for data cleaning and preprocessing. The team who built it has later on migrated it to a commercial version (Trifacta (http://Trifacta.com)) but you can still use their tool (and data) online.

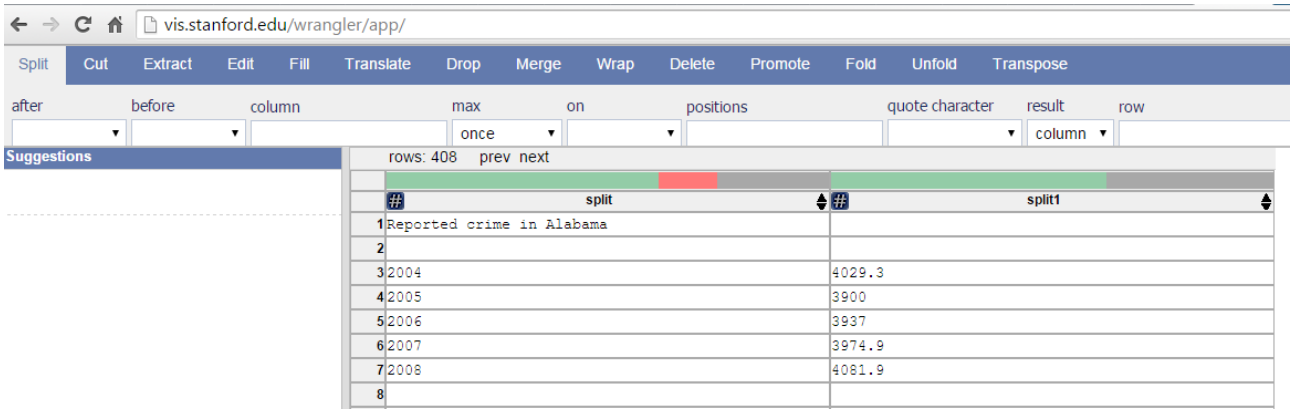
Step 1

Go to: http://vis.stanford.edu/wrangler/app/ . Select "Crime" from the Example Data drop list. Click Wrangle (as below).



Step 2

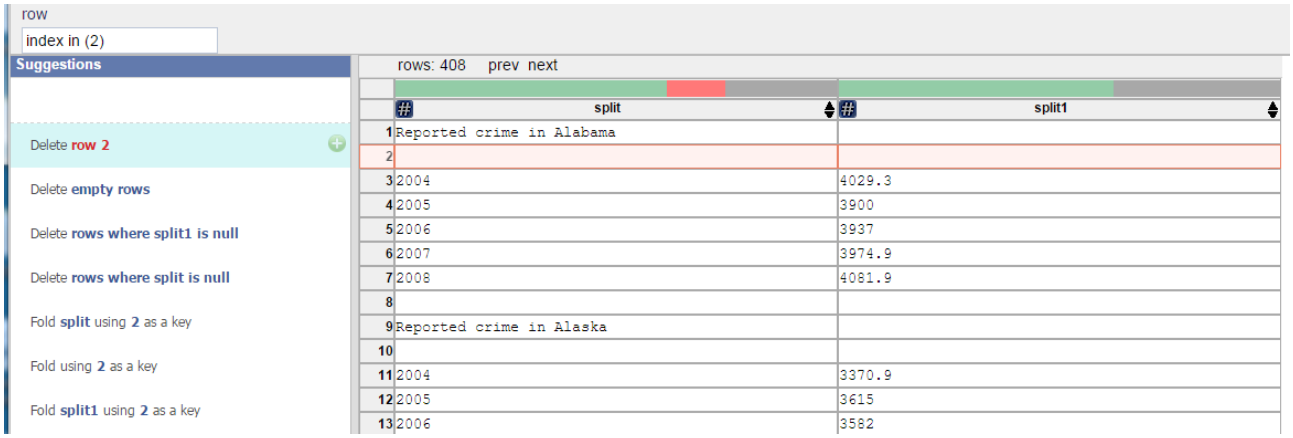
Note (above) that there are lines with ',' only (line 2, 8 etc.). DataWrangler has cleaned these:



You can see below, in the Script window on the left (before we did anything) 'newline' and ',' have been 'split'. Note also that there is a rollback feature, 'Undo Transform'.

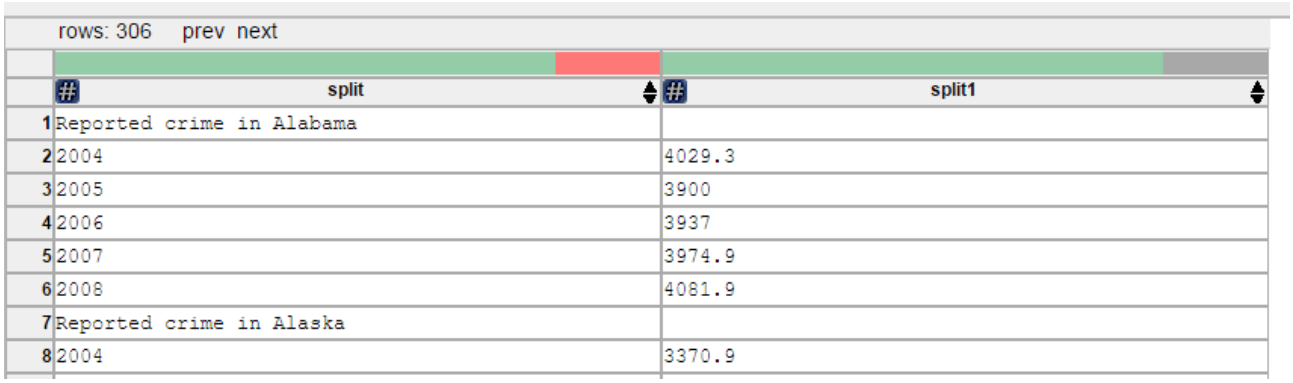
Step 3

Back to the data, note that e.g. the second line is empty (and the 8th etc.), if you select one of them (by clicking in the numbered column beside the data), Wrangler suggests the following:



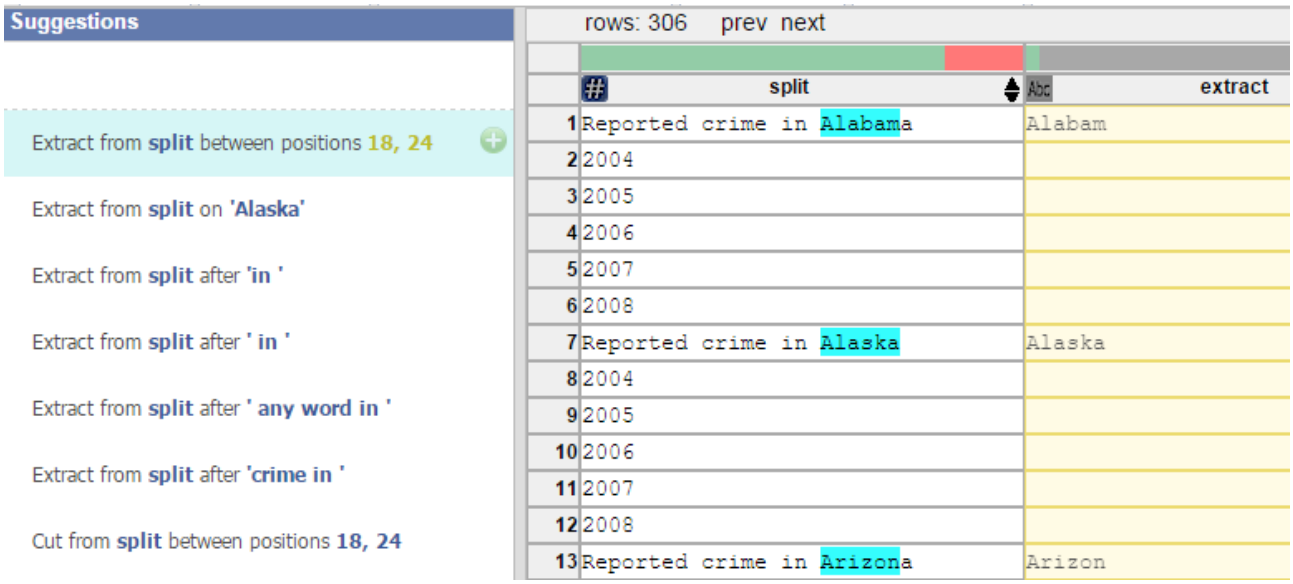
Step 4

We don't want to just 'Delete **row 2**, so try the second suggestion: 'Delete empty rows'..

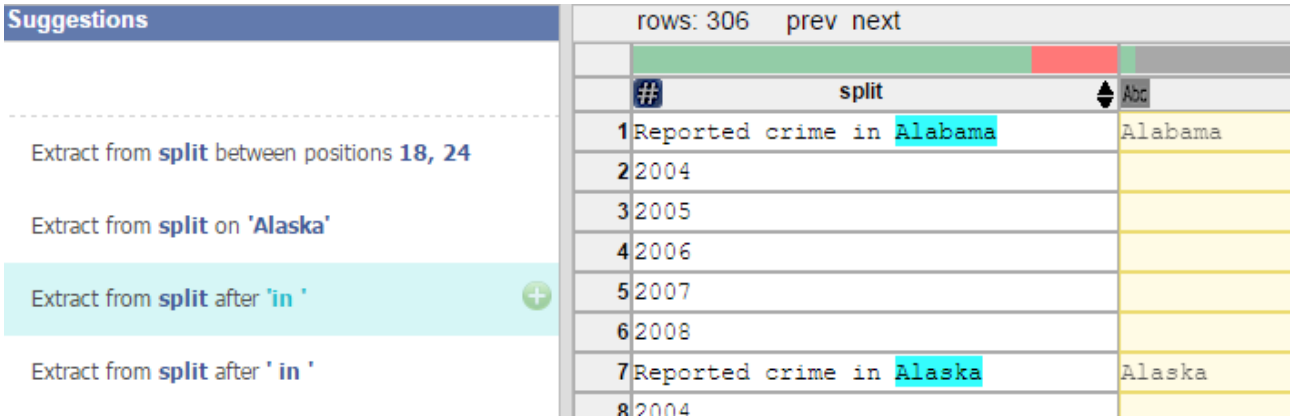


Step 5

Now we can extract the State, select any of the states in column 1, e.g. Alaska (click and drag in the second data column) and see the suggestions generated (and a preview in yellow):

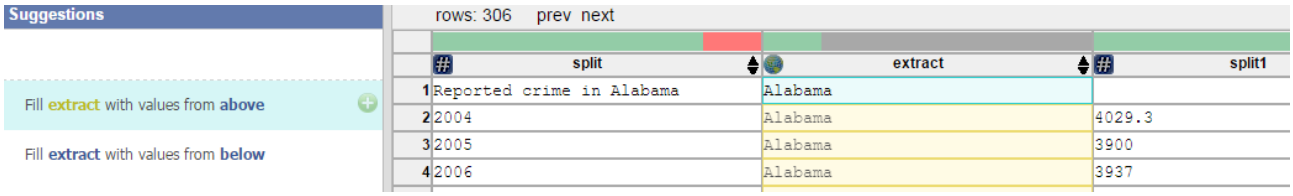


The preview is not quite right, 'Alabam'? We want to split after 'in ' - here's the preview:



Step 6

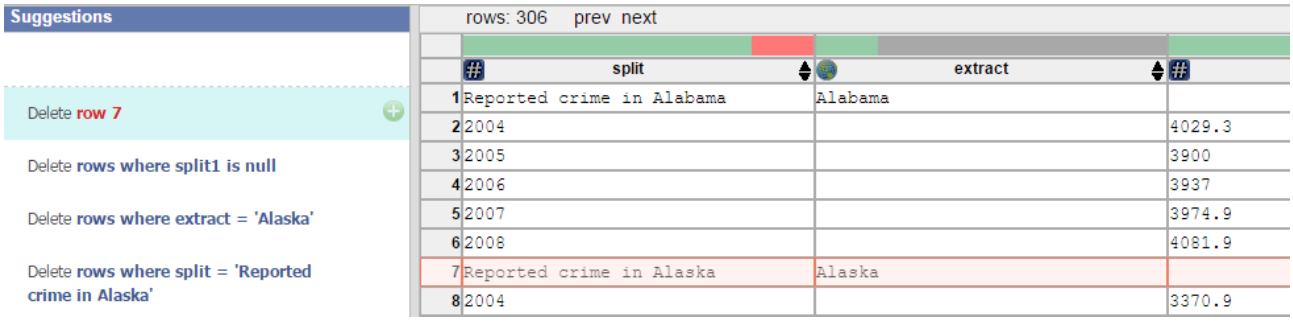
Select the third rule to extract the states. Now we want to copy the States down so in the second column click on the column title 'extract' to see this preview:



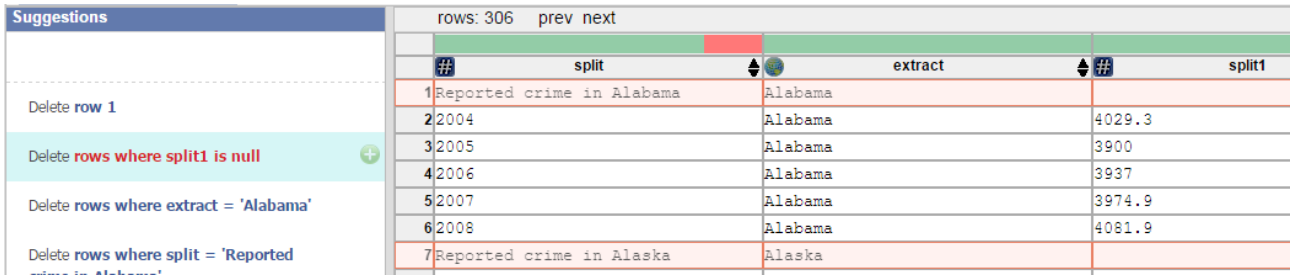
And the first suggestion is what we want '..values from above'.

Step 7

We don't need the original state record now, there's no data (see under the third column 'split1'). Select any of the original state rows (1st, 7th etc.) to see this suggestion:



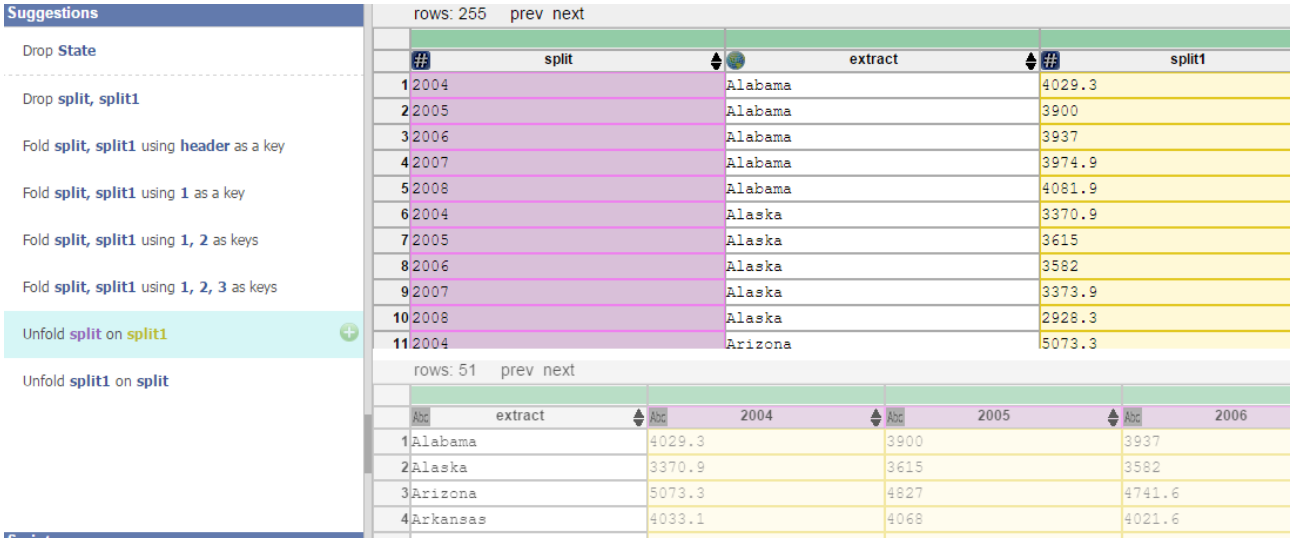
Delete just row 7? That could take a while, row by row. Preview suggestion 2:



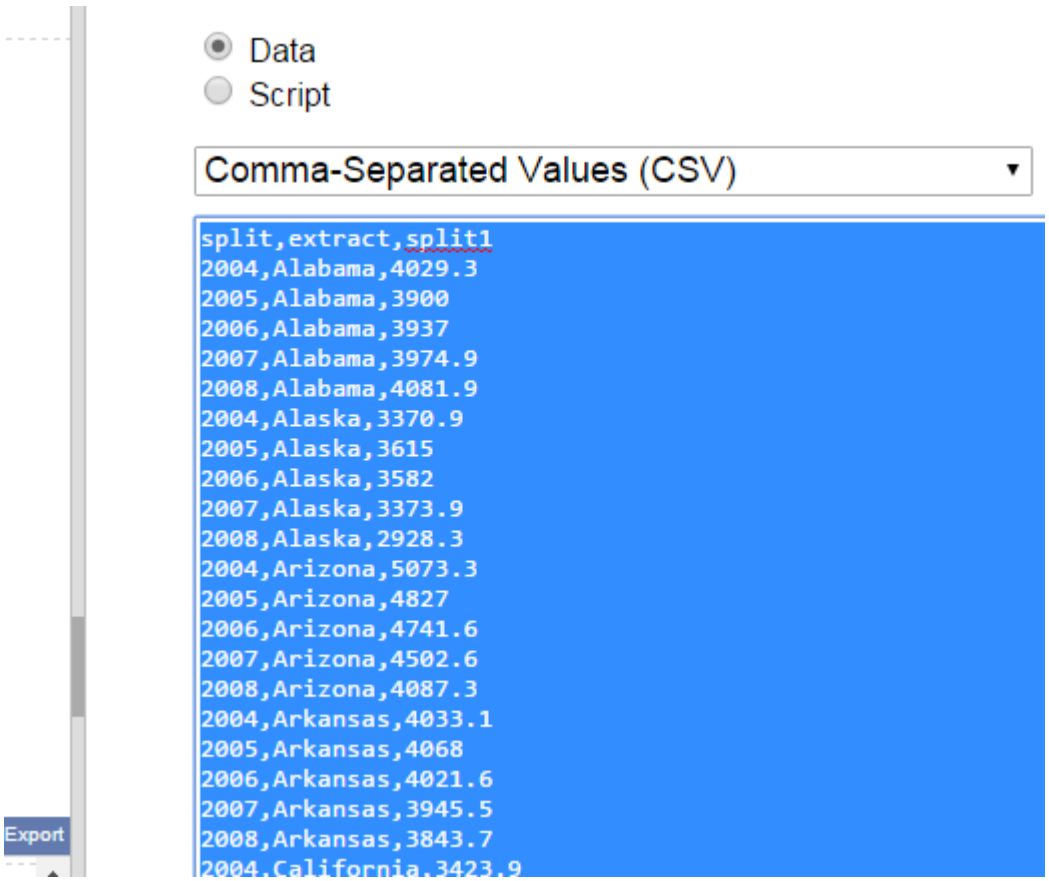
That's what we want.

Step 8

Now we have tidy data, you can continue and explore. Select column 1, CTRL, column 3 then you can see a preview of data:



You can export you transformed data, also the script (using the rather small Export option above the script):



Step 9

Paste your own data and Wrangle it, or try *Patients.txt* (https://www.alexandriarepository.org/wp-content/uploads/20150629135742/Patients.txt).

4.4
Activity: Data Wrangling with iPython

Data Wrangling
(https://www.alexandriarepository.org/wp-content/uploads/20150709142519/Data-Wrangling.ipynb) (.ipynb) or Data Wrangling with iPython

First, you need to get yourself familiar with iPython notebooks. If your Python coding is a bit rusty, best work cooperatively with someone else., or do one of the suggested Python tutorials before hand. You do not need extensive Python skills for this activity.

Start by reading the Nature article and playing on their sandbox:

- "Interactive notebooks: Sharing the code" (Nature) (http://www.nature.com/news/interactive-notebooks-sharing-the-code-1.16261) , Nov 2014
- iPython interactive demo (on Nature) (http://www.nature.com/news/ipython-interactive-demo-7.21492?article=1.16261), the companion sandbox

Once familiar, you have several options for coding in iPython notebooks:

1. use the Nature sandbox, but beware that it get's wiped regularly: http://www.nature.com/news/ipython-interactive-demo-7.21492?article=1.16261
2. use Cloud Sage Math (CSM), requires an account (free) https://cloud.sagemath.com
3. use a software package like Anaconda on your computer: http://continuum.io/downloads#py34

Either of the latter two are preferred (Sagemath can also do 'R' and allows multiple users = collaboration).
The file "Patients.txt" must be in the same directory as this code, otherwise you will need to know and use the path to it (*e.g.* folder\data\).
Run each code cell, one at a time and in sequential order (top down and Shift-Enter or play '>').
Note code, comments, output and questions.

Step 1

Create a new notebook (in either CSM or Anaconda) then copy the code below into a cell and run:

```
import pandas as pd
arr = pd.read_fwf('Patients.txt', colspecs='infer', header=None)
# ask pandas to guess data types... 'infer'
print(arr)
```

What is pandas?
How many rows and columns did pandas 'infer'?

Step 2

Dimensions of array, how big is it? Try this: arr.shape

30 is correct but why 2?

Step 3

Try another way, forced or fixed width:

```
import numpy as np
# force types & columns, add names
dt = np.dtype([('ID', np.str_, 3), ('Gender', np.str_, 1),
('Visit', np.str_, 10,), ('HR',int),('SBP',int),('DBP',int),('DX', np.str_, 3),
('AE', np.str_, 1)])
data = np.genfromtxt('Patients.txt', dt, delimiter=[3]+[1]+[10]+[3]*4+[1])
print(data)
```

What is numpy?
How many columns discovered now?

Step 4

Plot that data:

```
%matplotlib inline

import matplotlib.pyplot as plt
plt.scatter(data['DBP'], data['SBP'])
# and plot to see data
plt.show()
```

What is %matplotlib inline?
How many suspect data points do you see?

Step 5

Another type of plot, similar data

```
plt.scatter(data['HR'],data['DBP'], c=data['SBP'], s=40, cmap='hot')
```

There is one extreme value here ('white hot'), which record is it in (ID)?

Step 6

Try another data structure, DataFrame, and display data:

```
df = pd.DataFrame(data)
df
```

There are a lot of -1 values, OK for this data but what can you do if we look at e.g. temperatures?

Step 7

we can see a few 'f' so change all 'f' (and 'm'), using to uppercase

```
df['Gender'] = df['Gender'].map(str.upper,["f","m"])
```

How many records remain?
(you can 'Insert' from the menu above and use Python to tell you)
What are the valid ranges for HR, SBP & DBP?

Step 8

remove some extremes, destructive

```
df = df[df['DBP'] >= 60]  #
```

How many records are gone?

Step 9

Now for the Heart Rate (HR) data, what's valid? We could refer to the valid range (HR 40-100, above) then remove anything outside this range (but this is also destructive and more serious than 'f' vs 'F'), what if 30 is possible? Try this:

```
df = df[df['HR'] > 30]
df.shape
```

So now how many records are left?
(and if we keep 'cleaning' data like this, how many valid records will remain...)

Step 10

```
df.plot(kind='bar', stacked = True)
# df.plot(kind='barh', stacked=True)
```

If that 900 HR record had to go, how would you do it?
How can you save your data? (look up DataFrame.to_csv)
What other formats are possible - JSON, XML?
Finally, how does Python compare to other languages for data processing and visualisation?

Step 11

- Read: "Evaluation of Open Source Data Cleaning Tools: Open Refine and Data Wrangler (https://www.alexandriarepository.org/wp-content/uploads/20150629132348/p12-plarsson.pdf)"
- Which tool do you prefer (SAS, DataWrangler, Python, other) and why?

4.5
Standards and Issues

Standards can be about many different aspects of the process of Data Science. In this section many different standards are collected and discussed. In choosing which standards to include, and which to ignore, and the different related issues, we have to restrict ourselves. Many domains have their own specific standards: the publishing business has NewsML (http://NewsML.org), Chemistry has ChemML (https://en.wikipedia.org/wiki/Chemical_Markup_Language), the list goes on, so these are largely ignored here.

Standards for the Data Science process

In a follow up from Kirk Bourne's blog "Raising the Standard", the Data Science 101 blog presents the various versions of the Data Science process. We have discussed this process several times before, in the section **Introduction to Data Science** module **Data Science and Data in Society** and the section **Data and Decision Models** module **Data Models in Organisations**. Starting from the (hopefully) familiar software development lifecycle, Cross Industry Standard Process for Data Mining (CRISP-DM) and various Data Science workflows are discussed.

- " (http://101.datascience.community/tag/process/)Data-Driven Software Engineering" (http://101.datascience.community/tag/process/) on DataScience101 (500 words, 5 mins) 

It is important to be aware of different views here, and when in doubt, prefer simplicity, but CRISP-DM, developed in the late 90's within the data mining community, is the process with a long history of development, so it is the most important to study.

- "Cross Industry Standard Process for Data Mining" (https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining), on Wikipedia (1000 words, 6 mins)


Just so you do not think the process is easy or simple, have a look at the most extensive process diagramme developed to date, by Fatih Hamurcu. Now, one would rarely take all the routes and diversions listed on this process diagramme. Nevertheless, it is nice to see the possibilities listed.

- "Data Science life-cycle" (http://fatihhamurcu.blogspot.com.au/2015/04/data-science-life-cycle.html) (infographic, scrutinize for 5 minutes)

A statistics task related to the discovery process of CRISP-DM that has previously been discussed in the statistics community is called exploratory data analysis (EDA) (https://en.wikipedia.org/wiki/Exploratory_data_analysis), from Wikipedia:

> John W. Tukey ... held that too much emphasis in statistics was placed on statistical hypothesis testing (confirmatory data analysis); more emphasis needed to be placed on using data to suggest hypotheses to test. In particular, he held that confusing the two types of analyses and employing them on the same set of data can lead to systematic bias owing to the issues inherent in testing hypotheses suggested by the data. ..... Many EDA techniques have been adopted into data mining and are being taught to young students as a way to introduce them to statistical thinking.

Standard data formats for semi-structured data

Digital containers (see *Data types and Storage: Characterizing Data*) are common formats for multimedia data, and there are many different kinds. For semi-structured data consisting of text, there are a number of formats that predominate: Extensible Markup Language (XML) and JavaScript Object Notation (JSON), although there are others. It is therefore important to see examples of these and understand their broad capabilities, though you might not learn details about them until later units.

- Samples of JSON and XML (https://en.wikipedia.org/wiki/JSON#Samples) (on Wikipedia, just the subsection, 200 words, 1 min) 

XML is a complex standard, and the subject of entire books, whereas JSON is a simple alternative designed for lightweight processing in Javascript web applications.

- Overview of JSON (https://en.wikipedia.org/wiki/JSON) (Wikipedia, read the first two sections **History** and **Data types, syntax and example** only, 1200 words, 7 mins) 

XML is is sufficiently complex that, initially, it is best to look at some examples, and only learn the detail and specification at a later date when needed. XML uses many internet standards as a basis such as character codes, escaping, comments, its own schema formats and translation systems. XML and HTML have a common ancestor in SGML, so if you are familiar with HTML you will find it easier to learn XML.

All major programming languages have extensive support for processing XML and JSON, with routines for reading and writing. For Python:

- "A Python guide for open data file formats" (http://okfnlabs.org/blog/2013/10/17/python-guide-for-file-formats.html), from Open Knowledge Labs is a brief summary of major semi-structured data formats and Python ingest library calls (blog, 1200 words, 7 mins) 

Many semi-structured datasets are distributed in XML so it can be important to know how to work with these.

Resource Description Framework (RDF) is another style of language for representing (subject, verb, object) triples, which is used to represent semantics. It is a core representation language for Linked Open Data and the Semantic Web discussed in section **Introduction to Resources** of module **Data Resources, Processes, Standards and Tools**. RDF can be represented in different formats, for instance as XML or simply as line delimited lists. For our purposes, we just need to see the overview:

- Resource Description Framework (RDF) (https://en.wikipedia.org/wiki/Resource_Description_Framework#Overview) (Wikipedia, just read the Overview, 400 words, 2 mins) 

For an introduction to these different formats see:

- "A Primer on XML, RDF, JSON, and Metadata" (http://geekdoctor.blogspot.com.au/2011/01/primer-on-xml-rdf-json-and-metadata.html) on the blog of John Halamka (1000 words, 5 mins, start at "XML" half a page down) 

- A data scientists blog discussing JSON, XML and RDF, "Data Science and the Open World Assumption" (http://www.datasciencecentral.com/m/blogpost?id=6448529%3ABlogPost%3A274306), by Kurt Cagle (blog, 1500 words, 8 mins).

For detailed tutorial on these formats as well as reference pages, see http://www.w3schools.com/ (this, however, is far more detail than what we need in this overview, so only use as a reference).

Supporting open data

The Wikipedia overview of open data is a good place to get the basic concepts:

- open data (https://en.wikipedia.org/wiki/Open_data) (Wikipedia, read Overview only, 600 words, 3 mins)

A more detailed presentation of Open Data from a business perspective has been prepared by McKinsey Global Institute: "Open data: Unlocking innovation and performance with liquid information" (http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information) by McKinsey Global Institute. The Executive Summary is still too long for our purposes (6500 words, 40 minutes). The summary report contains major recommendations, example domains and some example descriptions of data sets.

MGI put the potential financial benefit of open data in the billions of dollars, and report the following findings (or claims) about the impact of open data:

- Open data enhances the value potential of big data analytics and provides additional opportunities.
- Open data helps businesses raise productivity, and create new products and services, helping the consumer especially.
- Open data creates new risks, including threats to reputation and loss of control over confidential information.
- Governments have a central role to play as a source of open data and as a regulator.
- Open data supports public understanding of policy and improves citizen engagement and participation.
- Open data also enables collaborations across sectors in both public and private settings, for instance disaster response and education.

A complementary study has been done by the Royal Society in the UK more from a science perspective. We will refer to their main findings only.

- Summary "Science as an open enterprise" (https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe-summary.pdf) by the Royal Society June 2012 (PDF, 2900 words, 15 mins). This is a 5 page report which goes in to some detail, and one should just scan it quickly.

Their report website (https://royalsociety.org/policy/projects/science-public-enterprise/report/) also has the full report. This is a useful reference with many more examples of open data repositories, discussions of management and policy, and more detail on the findings. For science, they make the following recommendations:

- Scientists need to better publicise their data, and universities, research organisations and journals should support this.
- Governments should provide supporting policy and industry sectors work with regulators.
- Governments should support the development of software tools and skilled personnel,

- Security and privacy should be appropriately managed.

Data APIs and software as a service

An API is a "application programmer interface". With the proliferation of computing devices on the internet and of data, it is natural that internet based APIs and software as a service will proliferate too. Some of the best known APIs are things like access to Twitter data, Yammer data, many of the big vendors such as Tableau, Google services, DropBox and so-forth. Many of the big open data providers also have APIs.

- "Data, APIs to accelerate application economy" (http://www.zdnet.com/article/data-apis-to-accelerate-application-economy-ca-tech/) an article on ZDNet 11/11/2014 (600 words, 3 mins) 

ProgrammableWeb provides a large catalogue of over 500 data-focused APIs as well as news, code snippets and other support. Some supply data, some provide software-as-a-service analytics or visualisation, and some are more application focused such as for marketing.

- "ProgrammableWeb API Category: (http://www.programmableweb.com/category/data) Data" (http://www.programmableweb.com/category/data) on ProgrammableWeb.com (use the search features to explore some examples)

To see some of the top predictive analytics API's have a look at the following article:

- "Top 30 Predictive Analytics Software API" (http://www.predictiveanalyticstoday.com/top-predictive-analytics-software-api/) on PredictiveAnalyticsToday (scan through the first 10, 400 words, 5 mins) 

Some good wrangling software appears at:

- "Seven free data wrangling tools," (http://blog.varonis.com/free-data-wrangling-tools/) by Andy green (Blog, 500 words, 5 mins) 

Open source software

Open source software (https://en.wikipedia.org/wiki/Open-source_software) is widely used. It is not always the best choice, but sometimes there are distinct advantages. Open source tools play a major role in the Data Science community, starting for instance with Hadoop and R, but there are many more. While proponents claim (A) lower cost (B) better security (C) no vendor lock-in and (D) better quality, antagonists can in some cases claim (B) worse security (C) good vendor support for non-free software and (D) lower quality. It does very much depend on the tools in question.

One good way of exploring the open source world is to see which projects are winning awards:

- "Bossie Awards 2014: The best open source big data tools" (http://www.infoworld.com/article/2688074/big-data/big-data-164727-bossie-awards-2014-the-best-open-source-big-data-tools.html) by InfoWorld (set of 22 short web pages, 10 mins) 

Some open source tools are considered important for the data scientist.

- "For data scientists, the big money is in open source" (http://www.techrepublic.com/blog/big-data-analytics/data-scientists-can-find-big-money-in-open-source/) by Matt Asay in TechRepublic (article, 900 words, 5 mins) 

Predictive models

The Predictive Model Markup Language (PMML) is a standard developed in XML for representing predictive models. It was developed by the independent, vendor led consortium The Data Mining Group (http://www.dmg.org/) (DMG). The standard is given on the DMG website but is too much detail for our purposes. Instead, see the summary on Wikipedia and scan the product list on DMG website.

- The Wikipedia page provides a good summary, PMML (https://en.wikipedia.org/wiki/Predictive_Model_Markup_Language) (read the introduction and *Components* section only, 700 words) 
- A list of products working with PMML is the PMML Powered page (http://www.dmg.org/products.html) on DMG site.

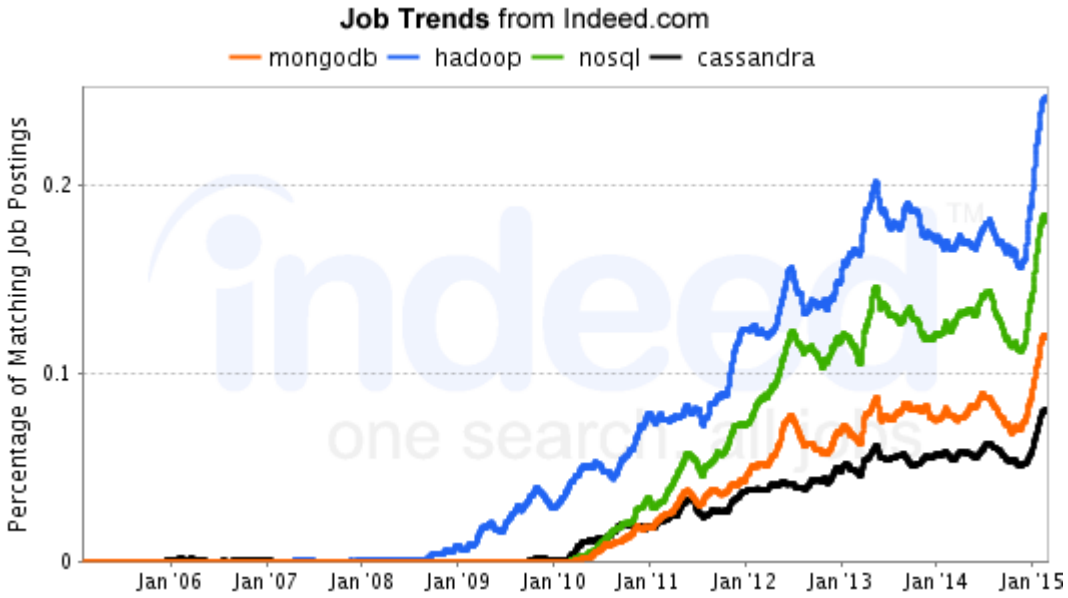
4.6
Activity: Software/Tools Trends

Data Science Tools and Trends

Similar to the Seek.com and Indeed.com activity earlier, here we will explore different software tools and programming languages, and their popularity over time. For this we can use Indeed job trends (http://www.indeed.com/jobtrends) and Google trends (http://www.google.com/trends). Refer to the MetroMap as a guide to remind yourself of different technologies.

Step 1

Search for: cassandra, hadoop, nosql, mongodb on Indeed.com.

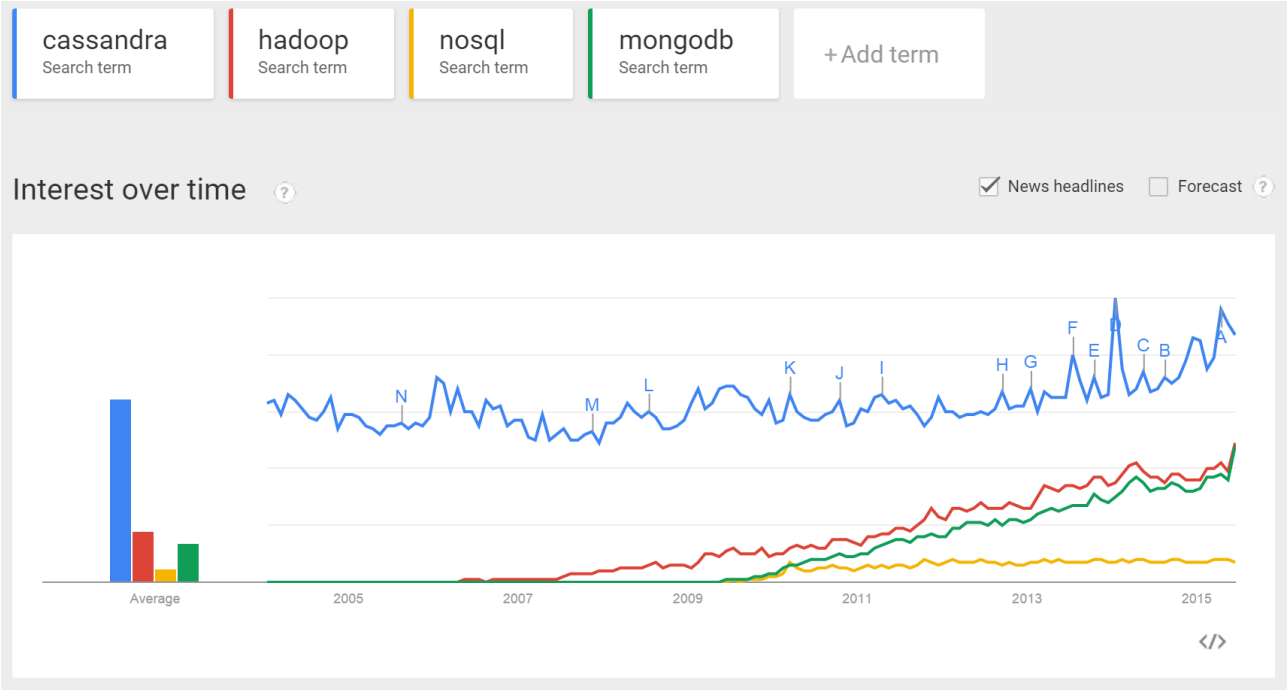


Step 2

Add SPARK to the search terms. Why did SPARK dip in 2014/2015?

Step 3

Do the same search in Google trends: cassandra, hadoop, nosql, mongodb.



Are we making fair comparisons i.e. when you search with Indeed it's job ad trends, what are you getting with Google trends?
If you explore the cassandra plot you can see that there's a reason why the trend lines don't match those from Indeed - what is it?

Step 4

Redo the Google trends search using 'Apache Cassandra' - still noticeably different to Indeed, what's the cause now?

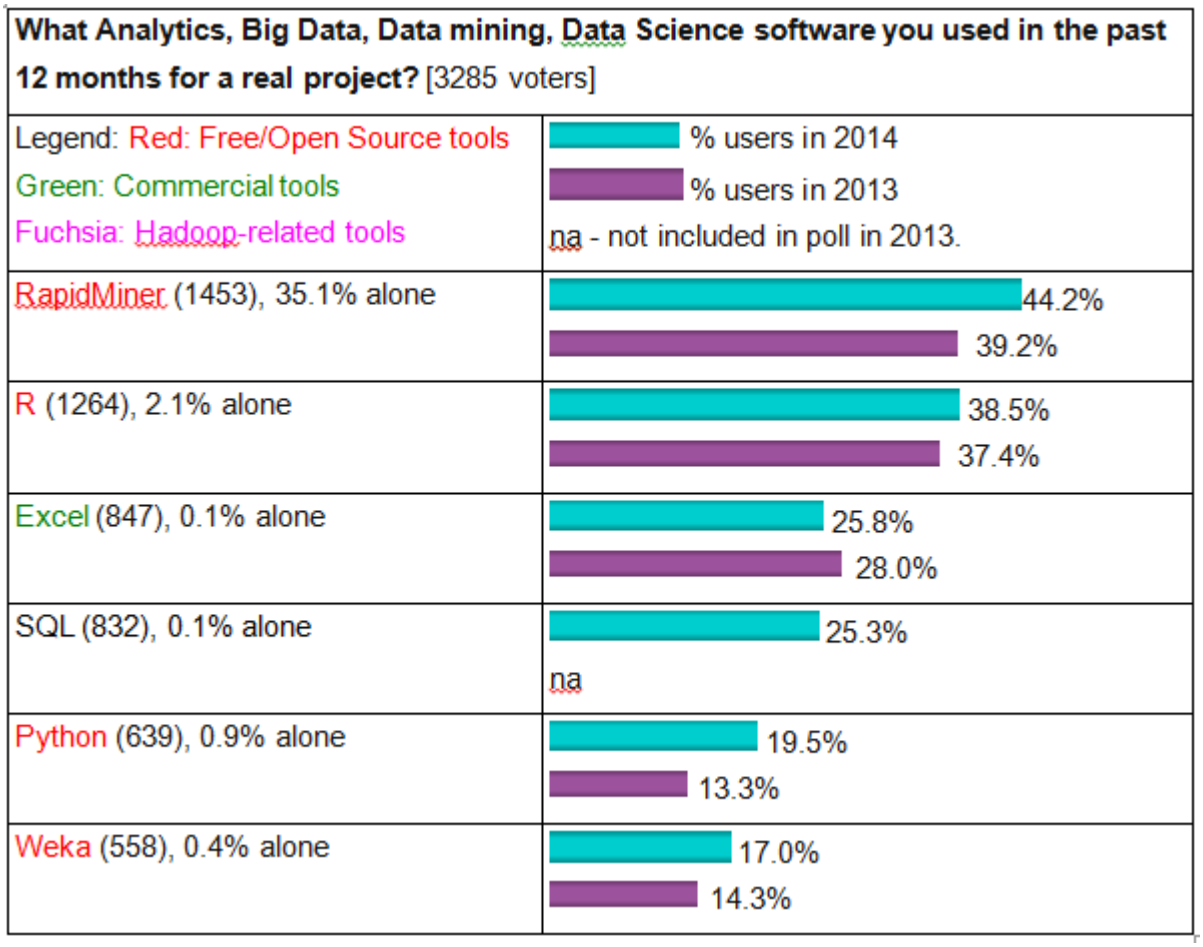
Step 5

Plot the technologies you (and others) discovered in the Seek ads activity (e.g. Java, Python, SQL from the CBA ad) using the tool of your choice. Compare them to trends from Indeed and Google - are there any obvious differences, what about between countries?

Step 6

There are other sources of tool trends and popularity, e.g., KDNuggets:

"The following table shows results of the poll, with Tool (User-votes), % alone.
% alone is the percent of tool voters used only that tool alone. For example, just 0.9% of Python users have used only Python, while 35.1% of RapidMiner users indicated they used that tool alone."



http://www.kdnuggets.com/2014/06/kdnuggets-annual-software-poll-rapidminer-continues-lead.html

What is so good about RapidMiner? What does it tell you about various tools if a ⅓ of all RapidMiner respondent use *only* that tool (compare this to 0.1% for Excel)?

What is Weka?

Optional Reading

Open source awards: "Although Hadoop is more popular than ever, MapReduce seems to be running out of friends."

http://www.infoworld.com/article/2688074/big-data/big-data-164727-bossie-awards-2014-the-best-open-source-big-data-tools.html

A business perspective, **Tableau Public** (http://www.tableausoftware.com/public/) ranked first

http://www.kdnuggets.com/2014/06/top-10-data-analysis-tools-business.html

Another one has SAS on top:

http://www.kdnuggets.com/2015/02/gartner-2015-magic-quadrant-advanced-analytics-platforms.html

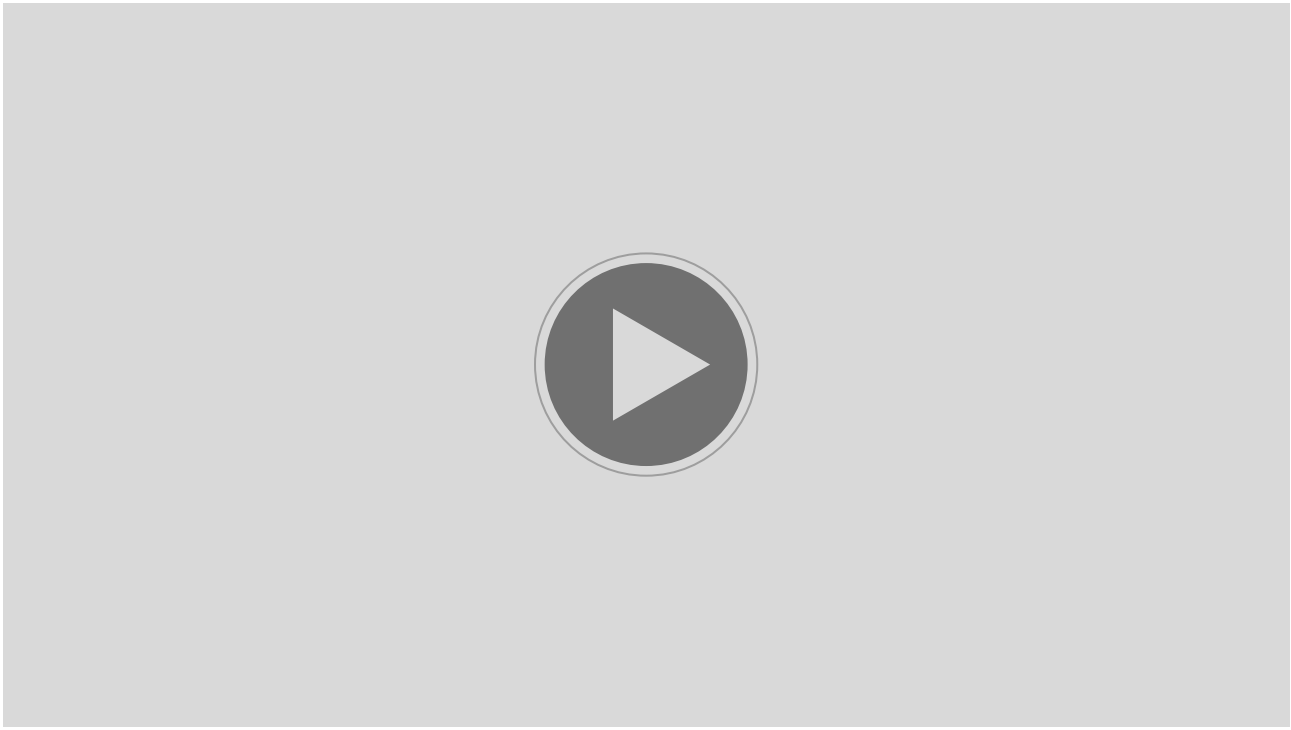
Predictions, not all tools...
http://pivotal.io/agile/webinar/top-data-science-trends-for-2015-webinar

See tools section:
http://www.oreilly.com/data/free/files/2014-data-science-salary-survey.pdf

4.7
Interview on Software and Tools

We continue with a few interviews with industry professionals here in Australia.

Watch Con Nidras (Head of Customer and Channel Analytics - National Australia Bank (NAB)), Associate Professor Michael Brand (Faculty of Information Technology - Monash University and former data scientist at Pivotal) and Dr Rami Mukhtar (CEO - Ambiata) talk about **Software and Tools** (4 mins)



(https://www.alexandriarepository.org/wp-content/uploads/20150629074039/FIT5145_module_4_software_and_tools_combined.mp4.mp4)

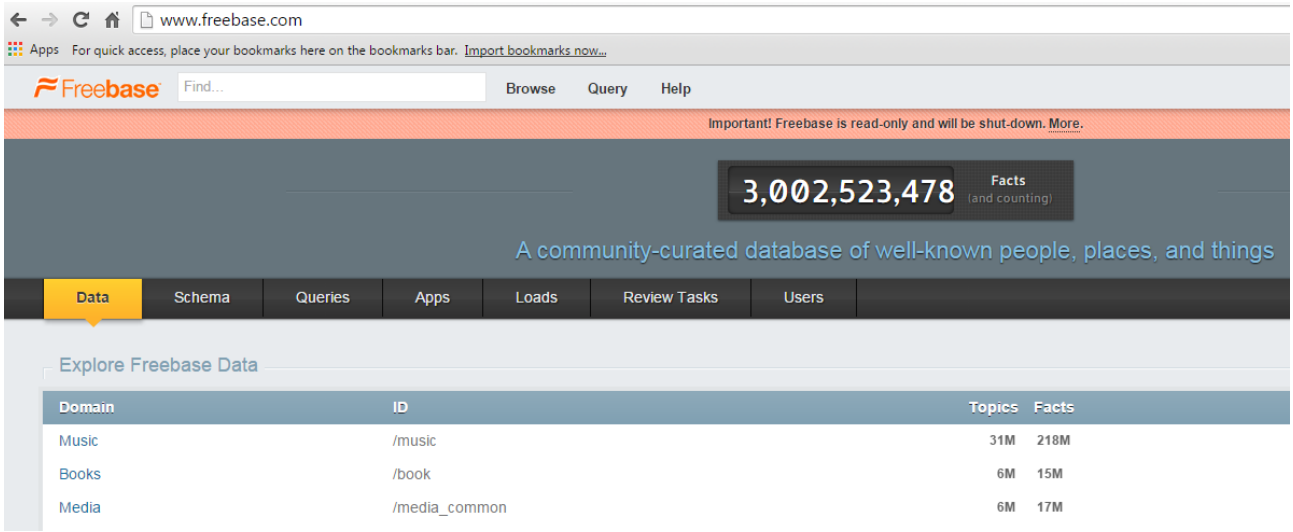
Alternatively, you can download the transcript for Software and Tools

(https://www.alexandriarepository.org/wp-content/uploads/20150701100111/transcript_FIT5145_module_4_software_and_tools.pdf).

4.8
Case Studies of Data and Standards

Freebase

Freebase, (http://www.freebase.com) a well known example of Linked Open Data, is a large knowledge-base which consists of different types of machine-readable data. These data are composed in a collaborative way by the community members of Freebase (http://www.freebase.com). The data are organized in a structured way and the majority of them are harvested from a number of online sources such as individual wiki contributions. The main idea behind the development of a large-scale resource like Freebase is to provide users with a global resource which allows for access to common information in a more effective fashion. Freebase was developed by Metaweb, an American software company, and it has been publicly available since March 2007 and acquired by Google in 2010. Google's internal development, the Knowledge Graph (https://en.wikipedia.org/wiki/Knowledge_Graph), was powered partly by Freebase. Freebase is being phased out by Google.



Currently, Freebase contains over 39 million topics concerning real-world entities such as people, locations, and any other things. It used to have an Application Programming Interface (API) which was retired on June 30, 2015. Instead, Freebase has launched a new API for entity search which is now powered by Google's Knowledge Graph.

The data in Freebase has been used for a variety of data analytics projects as well as research projects. As a result of its abundance, in particular, the textual data in Freebase has been utilized for training statistical and machine learning-based natural language processing components, such as Named Entity Recognition (https://en.wikipedia.org/wiki/Named-entity_recognition) (NER, *i.e.*, finding mentions of people's names, places, dates, companies, etc.) and Natural Language Question Answering (https://en.wikipedia.org/wiki/Question_answering). (https://www.textrazor.com/)One of the projects that has made use of Freebase data for its statistical training of NER is TextRazor (https://www.textrazor.com/). See the main page of TextRazor to get an example of the use.

- TextRazor (https://www.textrazor.com/): finding mentions of companies, people's names, locations, *etc*.

(webpage) 

Medical Data Dictionaries

Medical Data Dictionaries describes some of the resources used in medical informations.

- "Medical Data Dictionaries" (https://www.alexandriarepository.org/wp-content/uploads/20150630095337/DataDictionaries-v1.pdf) by Bahadorreza Ofoghi

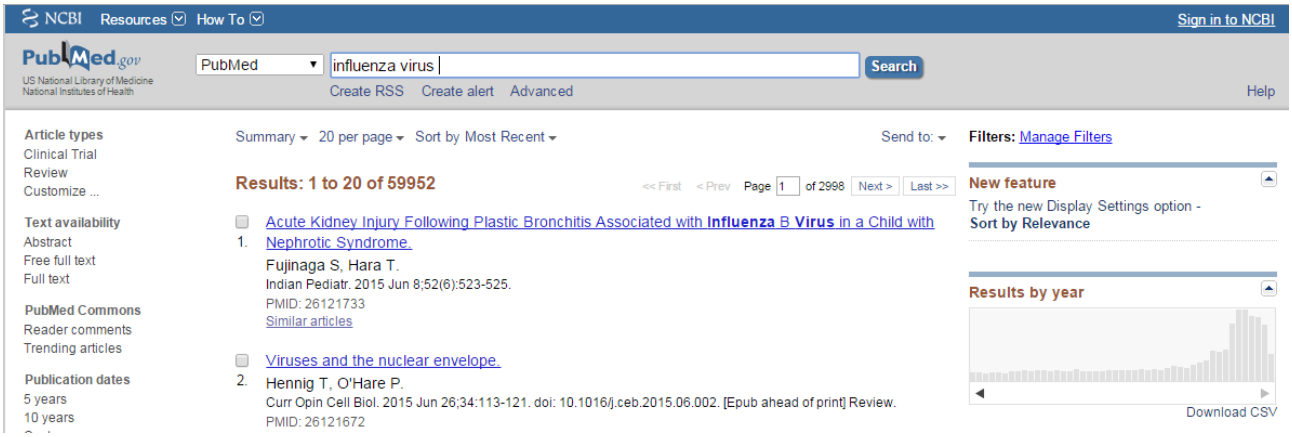
  (PDF, 1000 words, 15 minutes) 

Publishing Repositories

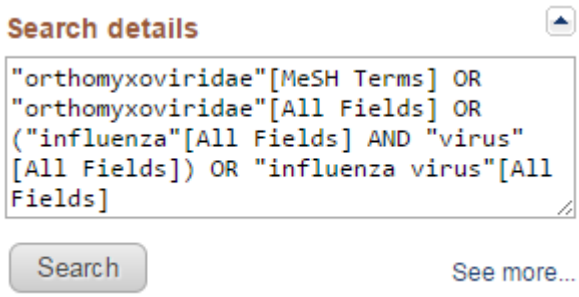
Today, there are a number of online repositories of peer-reviewed and published research materials as well as patents. We look at a few of them in the following sections.

PubMed

PubMed (http://www.ncbi.nlm.nih.gov/pubmed) is a freely available search engine which primarily provides users with access to the MEDLINE (https://en.wikipedia.org/wiki/MEDLINE) database. MEDLINE is a database of research references and article abstracts mainly on life sciences and biomedical topics. The US National Library of Medicine at the National Institutes of Health maintains the MEDLINE database. PubMed has developed its own Query Language that enables users and connecting APIs to more effectively seek information on the MEDLINE database. Below is a screenshot of the PubMed's online search system returning some results for the query "Influenza virus".



The query is then translated in to a more structured query language as shown below:



PubMed searches can be refined according to several categories such as species, year, text availability, article type, and some PubMed commons (e.g., reader comments, etc.).

ACM

ACM stands for Association for Computing Machinery (https://www.acm.org/) and, according to their website, is the world's largest educational and scientific computing society. Among a number of resources that ACM provides is the ACM Digital Library (http://dl.acm.org/) which is a large and the most comprehensive collection of full-text scientific research papers, articles, and bibliographic records in the areas of Computing and Information Technology. This digital library currently consists of more than 407,000 full-text articles with more than 18,000 papers being added each year.

Patents

Each year, a number of patents around the world are recorded in a number of electronic repositories. The data about these patents is usually structured and contains some information regarding the invention title, the name(s) of the applicant(s), filing date, PCT (Patent Cooperation Treaty (http://www.wipo.int/pct/en/texts/articles/atoc.htm)) number, WIPO (World Intellectual Property Organization (http://www.wipo.int/)) number, earliest priority date, and abstract. A few organizations and online repositories take care of patents and their relevant information in different areas of the globe. In Europe, the European Patent Office (https://www.epo.org/index.html) (EPO) is in charge of managing all the data and information regarding the European patents. In the US, the United States Patent and Trademark Office (http://www.uspto.gov/) (USPTO) manages the similar types of data concerning the US patents. In Australia, the AusPat (http://pericles.ipaustralia.gov.au/ols/auspat/) plays the similar role.

As mentioned, one of the important data fields regarding each patent is its WIPO number that is given by the (World Intellectual Property Organization (http://www.wipo.int/)). According to their website, *WIPO is the global forum for intellectual property services, policy, information and cooperation*. They have international systems for patents, trademarks, and design systems.

DBLP

The DBLP (http://dblp.uni-trier.de/) is the online Computer Science Bibliography where users can browse a large list of bibliographic items ordered by author, journals, series, and monographs. On 18/06/2015, the DBLP announced that they have now indexed around 3 million research publications. According to DBLP, *in each of the past three years, more than 325,000 new publications have been added to the database*. Currently, this online database has indexed publications from more than 4,000 conferences, and 1,400 journals. The bibliographic items can be downloaded/exported using different formats, i.e., BibTex, RIS, RDF, and XML.

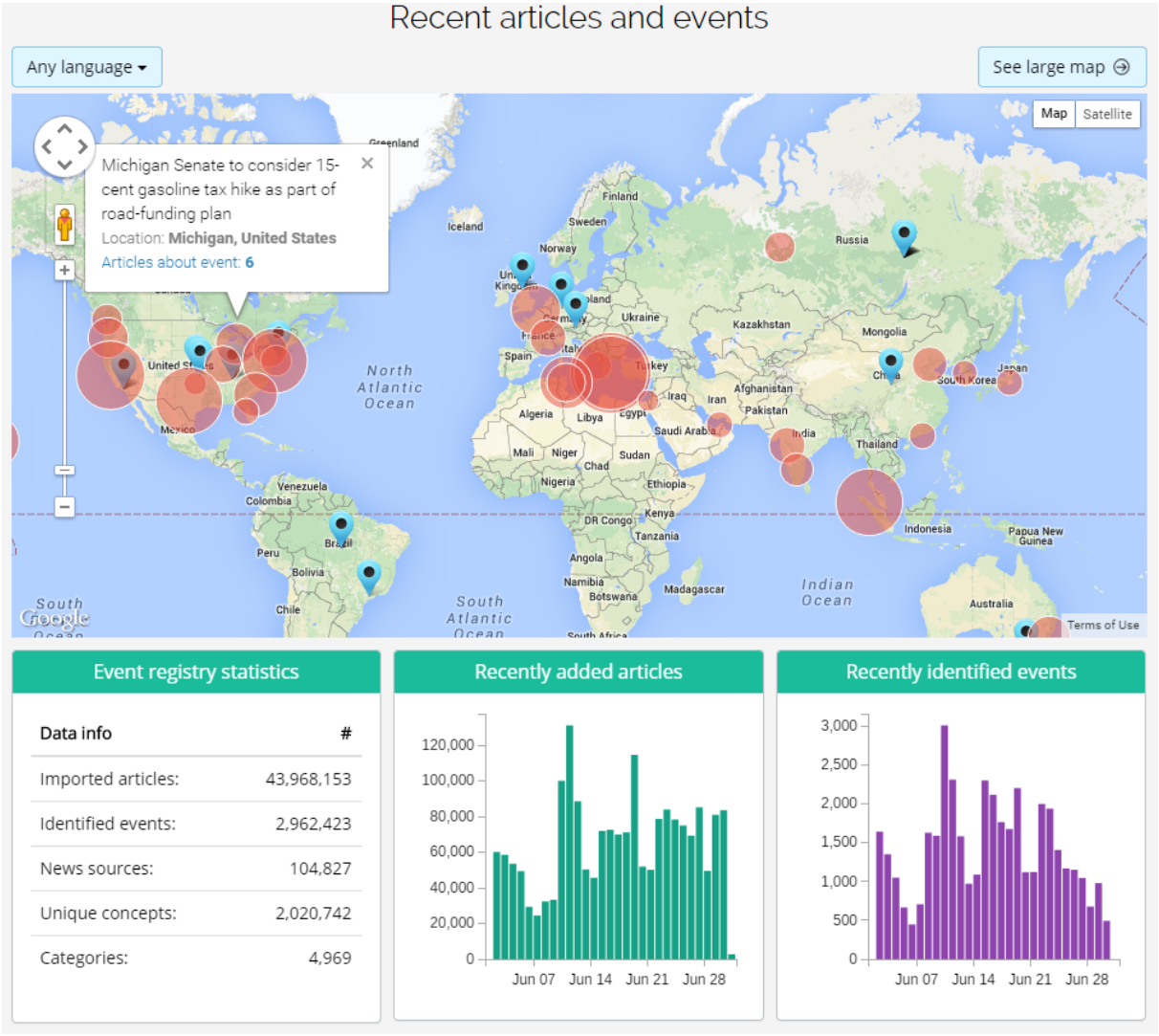
DCMI

The DCMI, the short form for The Dublin Core Metadata Initiative (http://dublincore.org/), is another online metadata library that supports best practices across a range of business models. The DCMI has an international participation scope which is open to every individual with an interest in metadata. The metadata in the DCMI has been structured using the RDF format (http://www.w3.org/RDF/) which enables linked data and semantic interoperability. The metadata concerning each concept or object in the DCMI contains information such as the title, creator, date issued, publisher, identifier, and language.

EventRegistry and NewsML

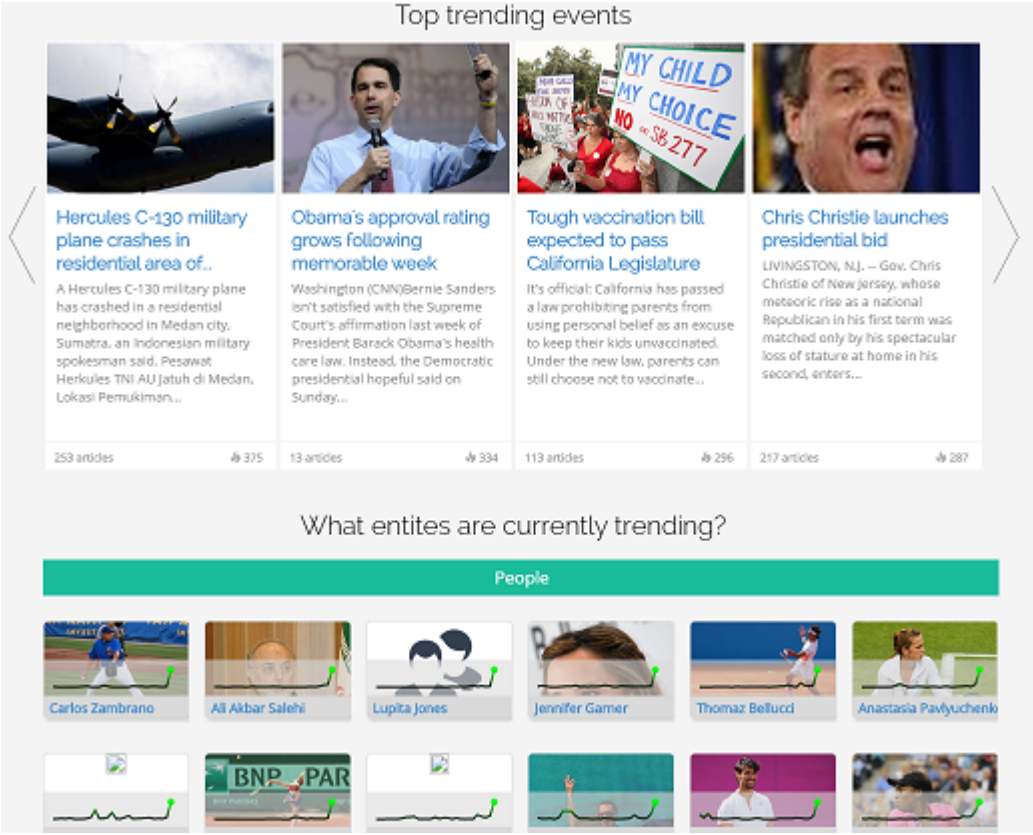
Online news is one of the current and rich resources of information today. Such information can be processed by both humans and machines to make critical decisions everyday, such as forecasting financial trends or planning for environmental disasters.

One of such online news systems is EventRegistry, (http://eventregistry.org/) developed at the Jozef Stefan Institute (http://www.ijs.si/ijsw/JSI). This system provides a real-time collection of news articles that are published by news outlets and agencies globally on a regular basis. EventRegistry (http://eventregistry.org/) facilitates analysis of the news articles in such a way that it is possible to identify events mentioned in the news as well as any relevant information about them in an automated and effective fashion. Then, such data is extracted from the text of the news articles and is stored in a searchable form.



A screenshot of the EventRegistry map and relevant statistics of news articles and identified events.

EventRegistry also provides users with a current list of trending entities (people, places, etc.) as well as popular and/or trending events in the news, as shown below:



A screenshot of the EventRegistry trending events.

The news data can be in different forms and formats, such as text, audio, video, and sound. To exchange such diverse data, there is a need for a standard exchange format over the network/internet. One such single format for exchanging text, images, video, audio news is NewsML-G2 (https://iptc.org/standards/newsml-g2/) using which it is also possible to exchange events or sports data. A number of news agencies as well as system vendors around the world have now registered with NewsML-G2 so that their news articles can be accessed through the data format that NewsML-G2 has to offer. Using NewsML-G2 it is possible to:

- Receive facts about an event from the event organizer,
- Publish facts concerning a specific event by a news provider,
- Publish facts of one to many events by event listings,
- Store facts about knowledgeable events in electronic archives, and
- Add information regarding the coverage of an event by a news provider.

This data format allows for storage, retrieval, and transfer of photo objects in a more effective and secure way. It incorporates all the digital pixel, metadata tags, as well as rights and licenses on a specific image.

There are similar data formats for the other types of news data that NewML-G2 makes use of. Can you find any of them?

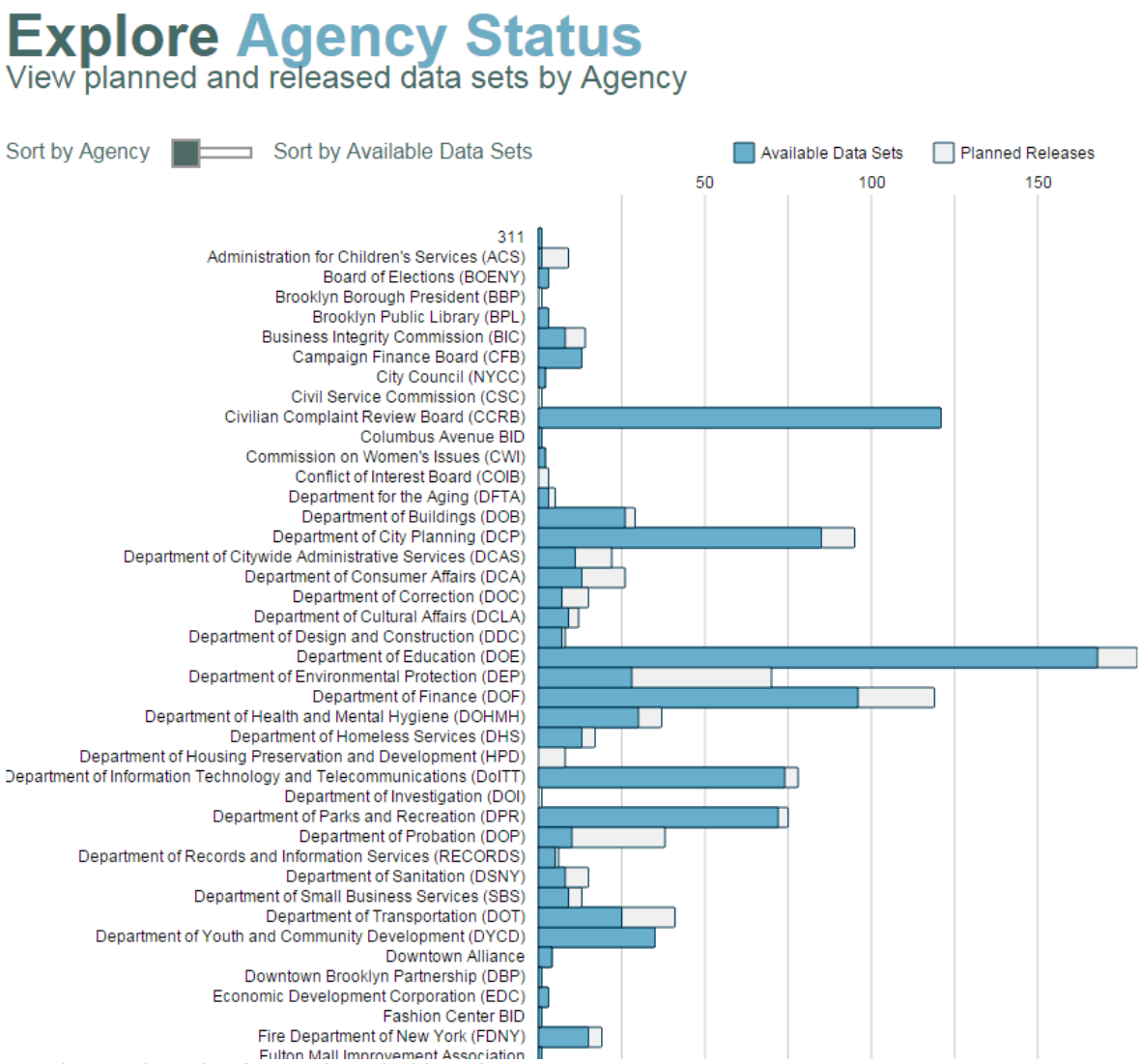
Data.gov and NYC Open Data

The US Government has developed an open online data store at Data.gov (http://www.data.gov/). In this online electronic data store, you will find data, tools, and resources which can assist you with conducting research in different fields. It also allows you to develop web and mobile applications, and design data visualizations. Currently, there are in excess of 141,000 data sets in Data.gov from 77 US Federal agencies and 321 publishers. Below is a screenshot from Data.gov in which you can see the results of a search for *organization:ed-gov AND type:dataset* published by *Office of Safe and Healthy Students*.

A screenshot from Data.gov in which you can see the results of a search.

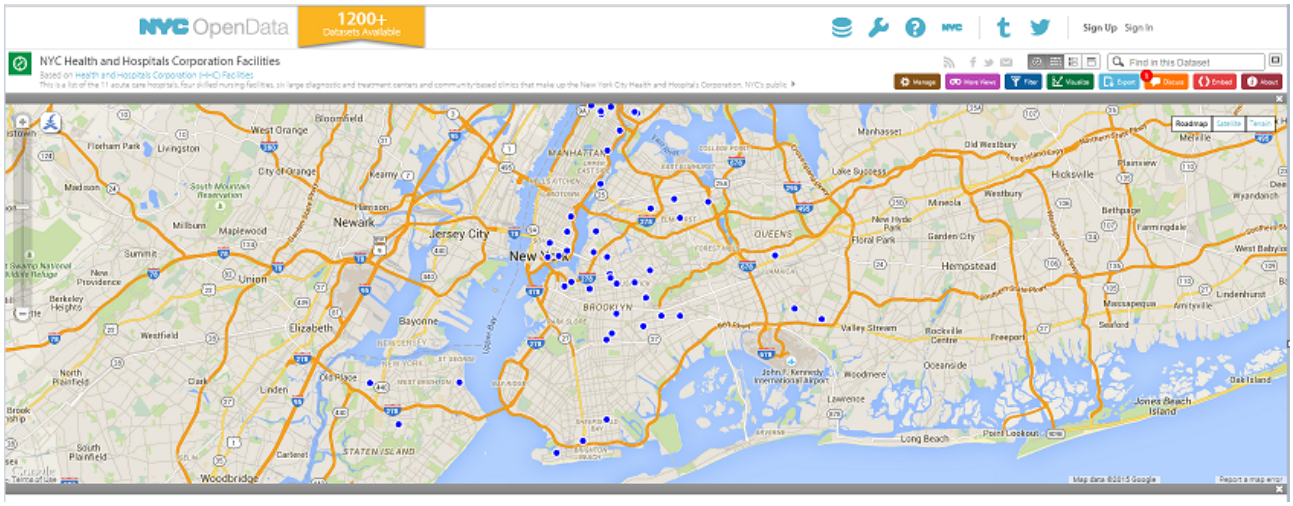
The open government data provided by Data.gov are currently used by a number of software (web or mobile) applications, such as AIRNow, Bankrank.org, Citymapper, etc. in which are mostly free and do not require registration for use.

Another open data repository is NYC Open Data (https://data.cityofnewyork.us/dashboard), which, as of July 2013, includes nearly 1300 data sets. The NYC Open Data has an Open Data Plan that was released in September 2013. The plan outlines the timeline of data releases for different agencies through to a deadline in 2018. Currently, there are 345 data sets listed in the Open Data Plan. As shown below, NYC Open Data releases the current status of each data provider agency in terms of the number of data sets that they have made (and are going to make) available.



Source: https://data.cityofnewyork.us/dashboard

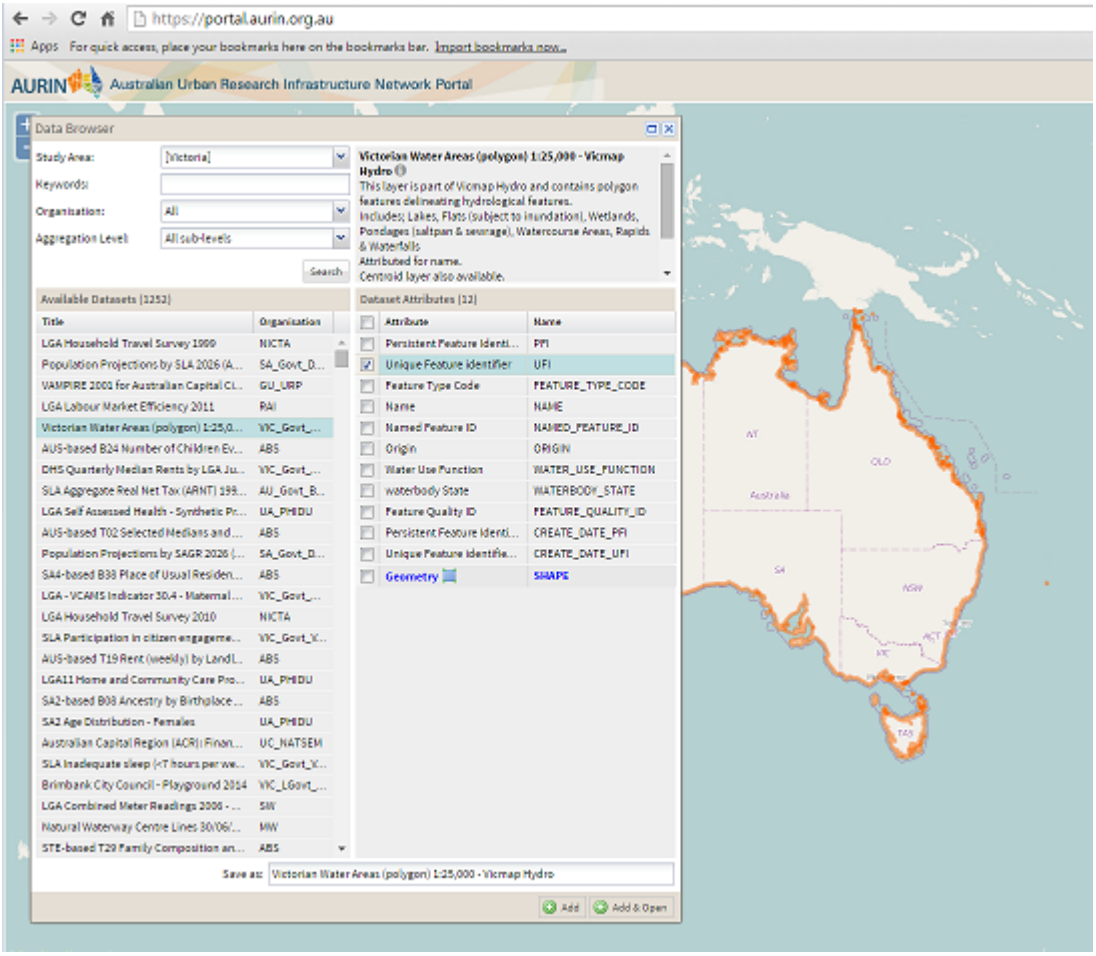
Below is a sample data set output on NYC Open Data. The data comes from NYC Health and Hospitals Corporation Facilities and shows the locations of such facilities on the map of New York City and its suburbs.



Source: https://data.cityofnewyork.us/dashboard

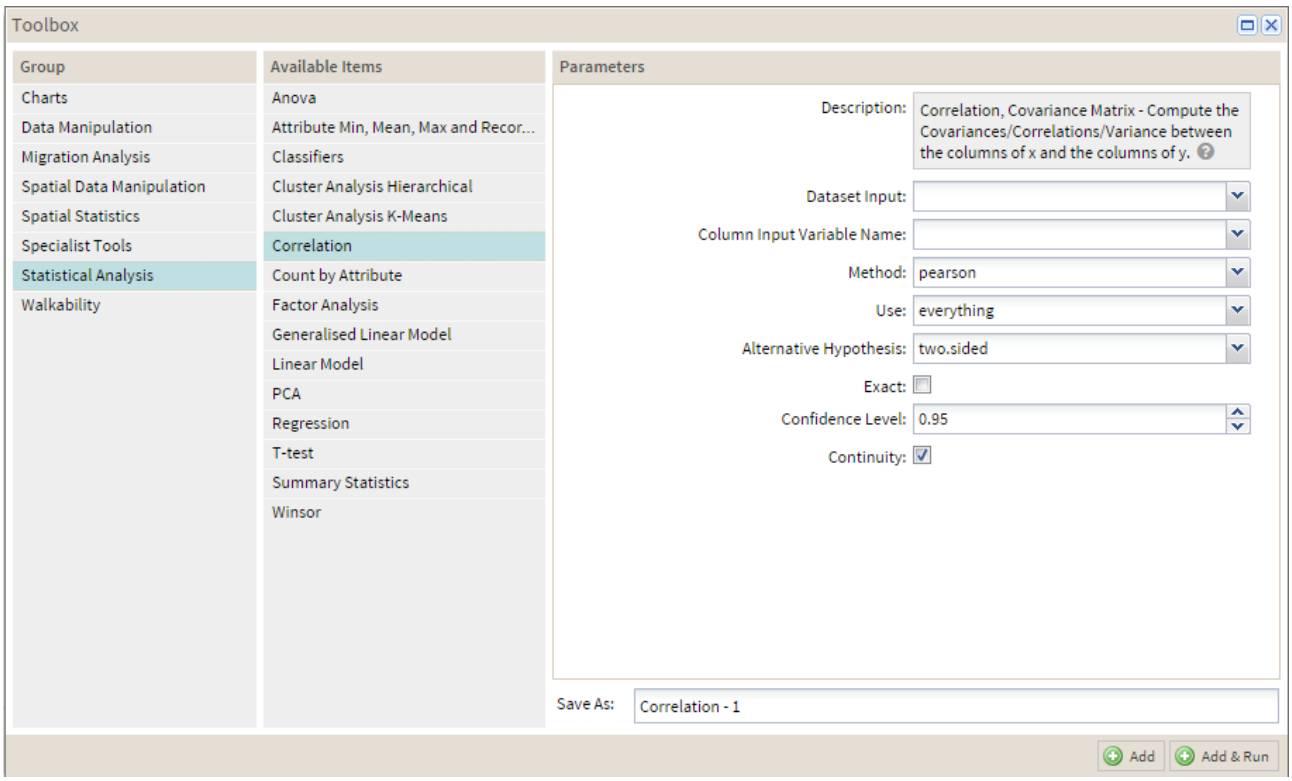
AURIN

Australia's Urban Intelligence Network (http://aurin.org.au/) (AURIN) is funded by the Australian Government to build the e-research infrastructure, enable better understanding of the current state of Australia's cities and towns, and meet the challenges they currently (or in the future will) face. This Australian national collaboration is led by The University of Melbourne (http://www.unimelb.edu.au/) and has the AURIN Portal (https://apps.aurin.org.au/gate/index.html) which delivers access to diverse data from multiple sources. The AURIN Portal (https://apps.aurin.org.au/gate/index.html) also facilitates data integration and data interrogation by means of open source e-research tools. Below is a snapshot of the portal where it is possible to select from a large list of data sets and visualize their specific features on relevant maps.



Source: https://portal.aurin.org.au/

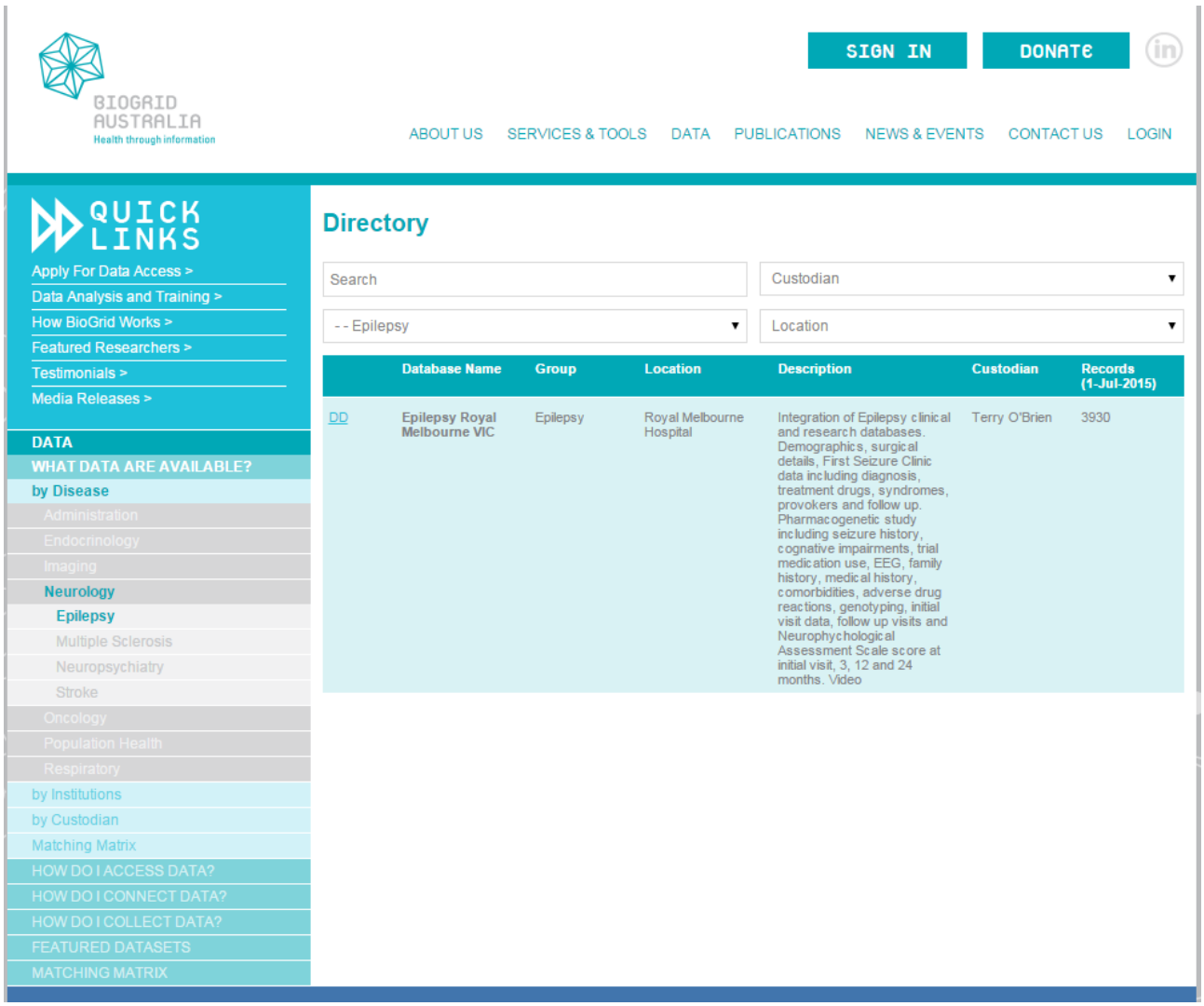
Using the AURIN Portal (http://aurin.org.au/about/), it is also possible to carry out a number of data analysis tasks such as statistical and machine learning-based analyses. Below, you can see a screenshot of AURIN Portal's dialogue window to perform some statistical analysis on a selected data set.



Source: https://portal.aurin.org.au/

BioGrid Australia

BioGrid Australia (https://www.biogrid.org.au/) Limited offers a data sharing environment for collaborative translational health and medical research. One major strength of BioGrid is its capability of linking real-time and de-identified health data records across diverse institutions, jurisdictions, and diseases. BioGrid has a web-based infrastructure which provides researchers and analysts with ethical access to health-related data while protecting and preserving privacy and intellectual property. The data sets on BioGrid are listed based on both disease types and institution. The image below shows parts of the data set list organized by the names of the providing disease types. Each data set, comes with a description of its schema (data file/table names, data fields etc.) which can be explored by clicking on the **DD** (Data Definition) link at the left hand side of the **Database Name**.

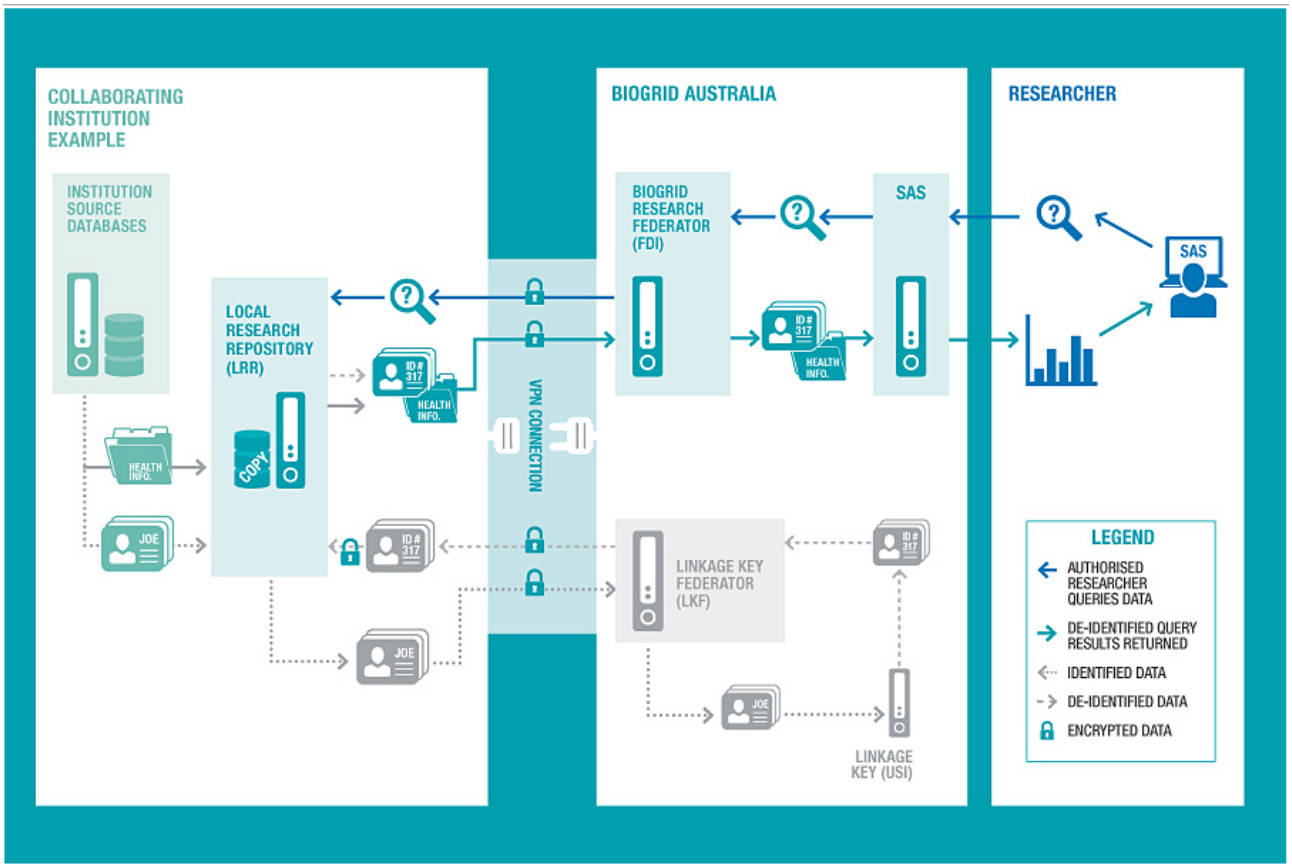


Source: https://www.biogrid.org.au

BioGrid Australia provides users with some data linking as well as analysis capabilities. Using BioGrid, a researcher or analyst can link a database with diverse hospital, university, and government databases. They can also link Biospecimen data to clinical data. In terms of data analysis, BioGrid offers statistical analysis software and services as well as assistance with data exploration, analysis or web reports. The analytics tools that BioGrid has include:

- SAS Enterprise Guide
- SAS Visual Analytics
- SAS Web Report Studio

Below is a diagram that shows the schematic view of how BioGrid facilitates research and analysis of health-related data.



Source: https://www.biogrid.org.au

5
Data Analysis Process

This is our fifth module of six for the Introduction to Data Science unit. This module will cover through case studies and some different aspects of data analysis, including the use of Python as a scripting language for the data analysis process.

(https://www.youtube.com/watch?v=TUWVX_lQufE)

Aims of This Module

- Classify different data analysis tasks in terms of problem requirements such as actions, values and unknowns.
- Evaluate the "data mining process" and the role of scripting languages.
- Determine scripting language requirements for a data science project.
- Analyse the role of sampling, data quality, proxy data and signal to noise (active sampling, using all data, mismatched data) in data science project.
- Contrast the Bias-Variance Dilemma with the Overfitting Problem in the context of sampling.

How to study for this module

In this module we again draw on material in the public domain such as interviews and videos, online magazine entries and blogs. We have also written some material to tie together various kinds of models. As well as studying and viewing the material, we have some activities around this material.

Please remember:

- Reference items marked with a single "johny look it up" icon,  , should be viewed as *suggested reading*, not essential nor important for assessment.
- Reference items marked with a two "johny look it up" icons,  ,  should be viewed as *important reading*, considered important for assessment.

5.1
Introduction to Data Analysis

In this section we present some motivating examples for the core data analysis process, for instance statistical analysis and visualization. While many of the speakers are well known, interestingly, there are no *bona fide* machine learning or statistical computing experts in this sequence. Their talks are perhaps too technical. We have some more technical talks in the later section **Data Analysis Case Studies** of module **Data Analysis Process**.

The wonderful and terrifying implications of computers that can learn

Jeremy Howard is an Australian Data Scientist and entrepreneur. He is the CEO of Enlitic (http://www.enlitic.com/), an advanced machine learning company in San Francisco. Previously, Howard was the president and Chief Scientist at Kaggle (http://www.kaggle.com/), which is a community and competition platform of over 200,000 data scientists. He is also a faculty member at Singularity University (http://singularityu.org/) where he teaches data science and has spoke at the World Economic Forum Annual Meeting 2014 on "Jobs for the Machines (https://www.youtube.com/watch?v=74xPTVQx0Lc)". Watch his talk about computers that learn:

- "The wonderful and terrifying implications of computers that can learn" (https://www.ted.com/talks/jeremy_howard_the_wonderful_and_terrifying_implications_of_computers_that_can_learn) at TED (video 20 mins or 3500 words transcript) 

The unreasonable effectiveness of data

Peter Norvig is an American Computer Scientist. He is a Director of Research (formerly Director of Search Quality) at Google Inc. (https://en.wikipedia.org/wiki/Google) Norvig is also a Fellow and Councilor of the Association for the Advancement of Artificial Intelligence (https://en.wikipedia.org/wiki/Association_for_the_Advancement_of_Artificial_Intelligence) and co-author (with Stuart Russell (https://en.wikipedia.org/wiki/Stuart_J._Russell)) of *Artificial Intelligence: A Modern Approach (https://en.wikipedia.org/wiki/Artificial_Intelligence:_A_Modern_Approach)*, now the leading college text in the field of artificial intelligence. He has over fifty publications in various areas of Computer Science, focusing on artificial intelligence, natural language processing, information retrieval and software engineering. Norvig talks about the use of big data.

- "The Unreasonable Effectiveness of Data" (https://www.youtube.com/watch?v=yvDCzhbjYWs) lecture at Univ. of British Columbia (Youtube, 60 mins)  and has published the respective scientific paper at *IEEE Intelligent Systems*, http://dl.acm.org/citation.cfm?id=1525689.

Knowledge is beautiful

David McCandless is a London-based author, data-journalist, information designer, and the author of the blog and bestselling book *Information is Beautiful*. He works across print, advertising, TV, and web. His information design work has appeared in over forty publications internationally including The Guardian, Wired and Die Zeit. In recent years, McCandless has exploring the use of data visualization and info-

graphics to explore new directions for journalism and to discover new stories in the seas of data.

- "Knowledge is Beautiful" (https://www.thersa.org/discover/videos/event-videos/2015/02/david-mccandless-on-knowledge-is-beautiful/) by David McCandless at the RSA (video, 18 mins)  , is based on his new book.

Mapping ideas worth spreading

Eric Berlow is an ecologist and network scientist who is recognized for his research on food webs and ecological networks and for creative approaches to complex problems. Berlow holds a Ph.D. from Oregon State University in marine ecology. **Sean Gourley** spent the past five years working at Oxford on complex adaptive systems and collective intelligent systems, basically, using data to understand the nature of human conflict. He is the co-founder and CTO of Quid which is building a global intelligence platform.

- "Mapping Ideas Worth Spreading" (http://www.ted.com/talks/eric_berlow_and_sean_gourley_mapping_ideas_worth_spreading) by Eric Berlow and Sean Gourley (TED talk, 8 minutes or 1600 words transcript).

The power of emotions: When big data meets emotion data

Rana El Kaliouby is a contributor to facial expression recognition research and technology development. This is a subset of facial recognition designed to identify the emotion being expressed by the face. El Kaliouby is also co-founder, chief strategy and science officer at Affectiva (http://www.affectiva.com/), the US-based company delivering consumer emotion analytics and insights. Here she discusses the world's largest repository of consumer emotions and presents the challenges and opportunities that this data presents for machine learning as well as data mining and visualization.

Before viewing this, review the definition of clustering (https://en.wikipedia.org/wiki/Cluster_analysis) from the Wikipedia:

> **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters).

This is used in her talk.

- "The power of emotions: When big data meets emotion data" (http://player.oreilly.com/videos/9781491900345?toc_id=192991), by Rana El Kaliouby, keynote talk at Strata (O'Reilly video, 11 mins) 

IBM Watson for healthcare

Martin Kohn is chief medical scientist, care delivery systems at IBM Research. His research work includes healthcare population analytics and the role of expert systems in the clinical decision process. Kohn is one of the lead team members working to bring the power of this supercomputer - known as Watson (http://www.ibm.com/smarterplanet/us/en/ibmwatson/) - to healthcare applications. Watson is a cognitive system with natural language processing abilities. Here Kohn discusses the use of Watson in healthcare,

and we just view a segment.

- "IBM Watson for healthcare" (https://www.youtube.com/watch?v=UFF9bI6e29U), by Martin Kohn of IBM (Youtube, see 1:30-16:40 mins only) 

Make data more human

Jer Thorp is a Vancouver-born data artist in residence at the New York Times (http://www.nytimes.com/) who also teaches in NYU's ITP program. Thorp's software-based art has been featured all over the world. His former career as a data artist explains why his art often brings big data sets to life and is deeply influenced by science. In this TED talk, Thorp discusses graphing an entire year's news cycle, to mapping the way people share articles across the internet.

- "Make data more human" (http://www.ted.com/talks/jer_thorp_make_data_more_human), Jer Thorp (TED, 17:30 mins) 

Six types of analyses every data scientist should know

Jeffrey Leek is an Assistant Professor of Biostatistics at John Hopkins Bloomberg School of Public Health. He works on figuring out how to go from raw data from next generation sequencing machines to results, turning public genomic data into clinically useful tools, and understanding how people use data analysis in real life. Here he discusses analysis types that is an alternative to the SAS analytic levels from section **Data and Decision Models** in module **Data Models in Organisations**, but more focused on the modelling.

- "Six types of analyses every data scientist should know" (http://datascientistinsights.com/2013/01/29/six-types-of-analyses-every-data-scientist-should-know/), by Jeffrey Leek (blog, 600 words, 4 mins) 

5.2
Theory of Data Analysis

In this section we will present different statistical theories of learning. We are not pursuing the standard theory one does in statistics or machine learning, which will start discussing formulas like this:

$$\mathbf{L}(f) \leq \hat{\mathbf{L}}_{\psi}(f) + \frac{2}{\gamma}\mathscr{R}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

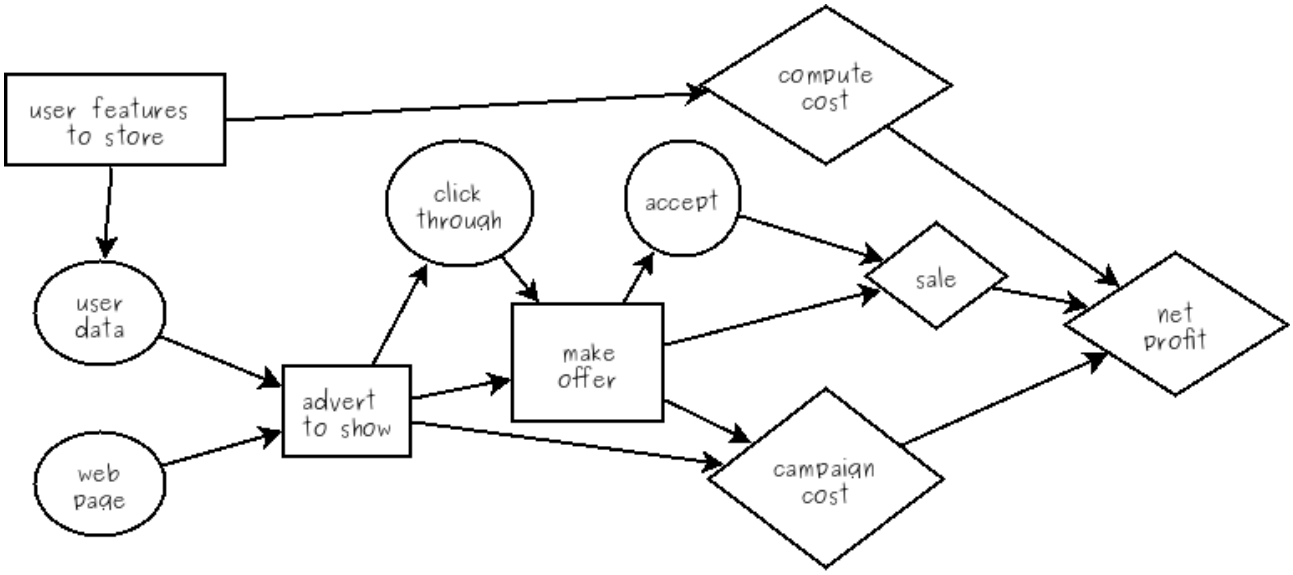
for all $f \in \mathcal{F}$

We will instead look at a general graphical modelling language for learning and look at the ideas of machine learning theory without doing any of the mathematics. Well, admittedly, we will consider some mathematics, but as little as possible. Some of the major principles of statistical learning are quite easy to understand, and this can be done by looking at learning in action, plots and curves. Of course, this means we will not be able to cover particular mathematical properties such as "asymptotic normality of the maximum likelihood estimator" which one might do in a more thorough treatment of theory, but it does give us the intuitions needed before we begin further study.

Graphical models

The first theory we are going to look at is based on graphical models. The influence diagrams we learnt in section **Data and Decision Models** in module **Data Models in Organisations** are one kind of graphical model. We introduce graphical models because they give us a structural/graphical language to talk about different learning problems. Not all learning problems can be usefully represented in graphical models, but quite a lot are.

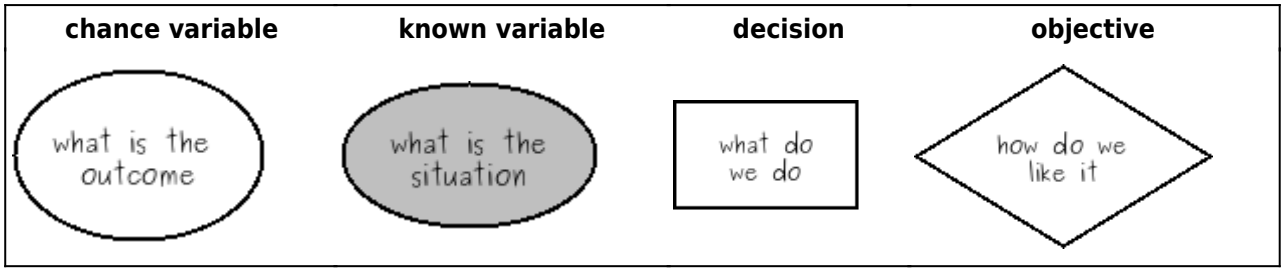
Consider the advertising model given in subsection **Example: online advertising model**.



Internet Advertising Model

The model represents a sequence of decisions, followed by observing evidence, then more decisions, and at various stages costs and the (net from) sale are obtained to allow the net profit to be computed. Now this is not a complete model: for instance in a full system there could be tens of millions of users and millions of websites. Moreover, in a modern implementation, computational considerations dominate, and one must make a "reasonable" offer very quickly; response time is also one of the objectives. However, this simple model is able to convey the key steps of the process.

We see here a combination of most of the node types, shown below.



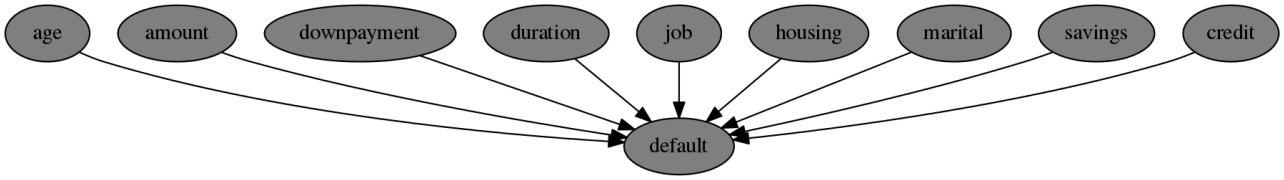
Moreover, the structure of influences represented by the arcs in the graph is also important: which knowns and unknowns influence what. One uses the idea of "cause". The node at the source of an arc is a "cause" for the node at the destination. The following table gives the details of **when** to connect an arc to a node.

chance variable	known variable	decision	objective
Connect from node A to the chance variable, if, holding all other nodes fixed in value, changing the value of A (perhaps by force) would influence the chance variable.	None essentially, but to be informative you can connect arcs to it as if it was unknown currently.	Which other nodes, whose values must be known at the time of the decision, are looked at to make the decision?	Which other nodes, whose values must be known at the time the objective is computed, are looked at to compute the value?

In the section **Data and Decision Models** in module **Data Models in Organisations** we also presented the *Heart Disease* model and the *Car Wont Start* model to illustrate mixing decisions, testing variables and evaluating objectives. Let us look at some more examples to understand learning. When one removes the decision and objective nodes from an influence diagram they are usually called Bayesian networks (https://en.wikipedia.org/wiki/Bayesian_network) or graphical models. Related models are causal loop diagrams (https://en.wikipedia.org/wiki/Causal_loop_diagram), which are influence diagrams with implicit time. Causal loop diagrams are popular in business and medical modelling to show feedback effects. Fault tree diagrams are used in fault tree analysis (https://en.wikipedia.org/wiki/Fault_tree_analysis) and root cause analysis (https://en.wikipedia.org/wiki/Fault_tree_analysis). Fault tree diagrams are a Boolean abstraction of a Bayesian network that are widely used in diagnosing failures in large scale networks.

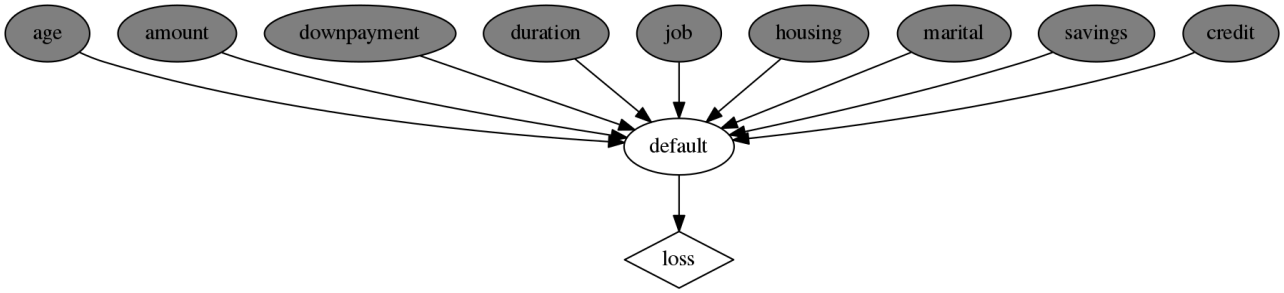
Housing loans

An early use of predictive analytics was assessing the suitability of a customer for a credit card, a house or personal loan. The customer would usually fill out an application form, the bank would enter the details into their computer system and a suitability score would be given for the customer. Depending on this, and depending on the bank's current propensity to take on mortgage clients, they may agree to make the loan. The data for a prototypical situation like this is shown in the *Housing Loan Data* figure, which the bank would have usually drawn from their historic set of mortgages. The target variable here is *default* which indicates whether the customer later on defaulted on the loan or not.



Housing Loan Data

This is a rather uninteresting figure. It basically just enumerates the variables in the task. When it comes for the bank manager to do their prediction, the model they have is shown in the *Housing Loan Prediction* figure.

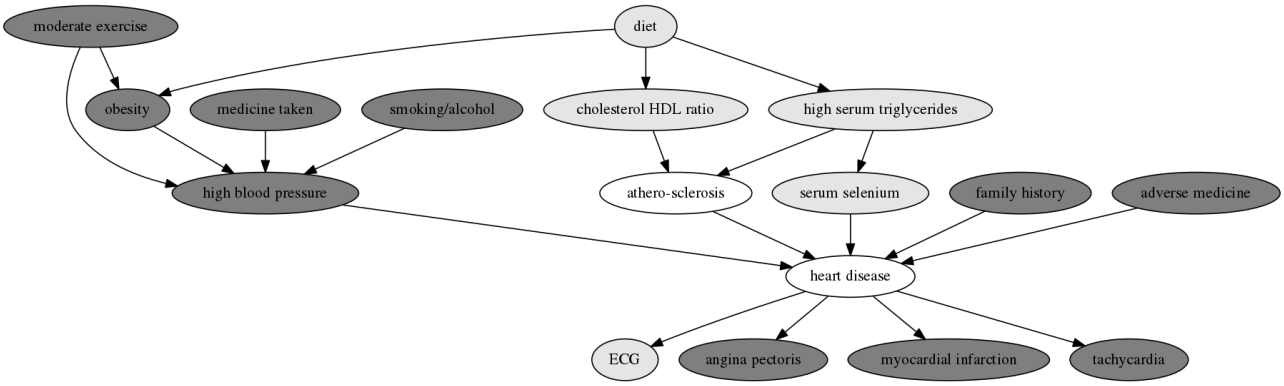


Housing Loan Prediction

Again, this is a rather uninteresting figure. You can see that when doing simple predictive analysis, the Bayesian network model does not lead to a great deal of understanding. In this situation, analysts spend their time considering the choice of variables to use and the prediction algorithm: support vector machine, or logistic regression, *etc*. The relationships between different variables and the prescriptive analysis is simple.

Lifestyle choices for a heart disease patient

The *Heart Disease* model presented the situation where the patient and doctor had to decide which tests to take, and the patient had to decide what lifestyle choices to make. Suppose at particular cardiologist clinic, doctors have dutifully recorded their patient details over time. They can have data looking something like that in the *Heart Disease Data* figure.

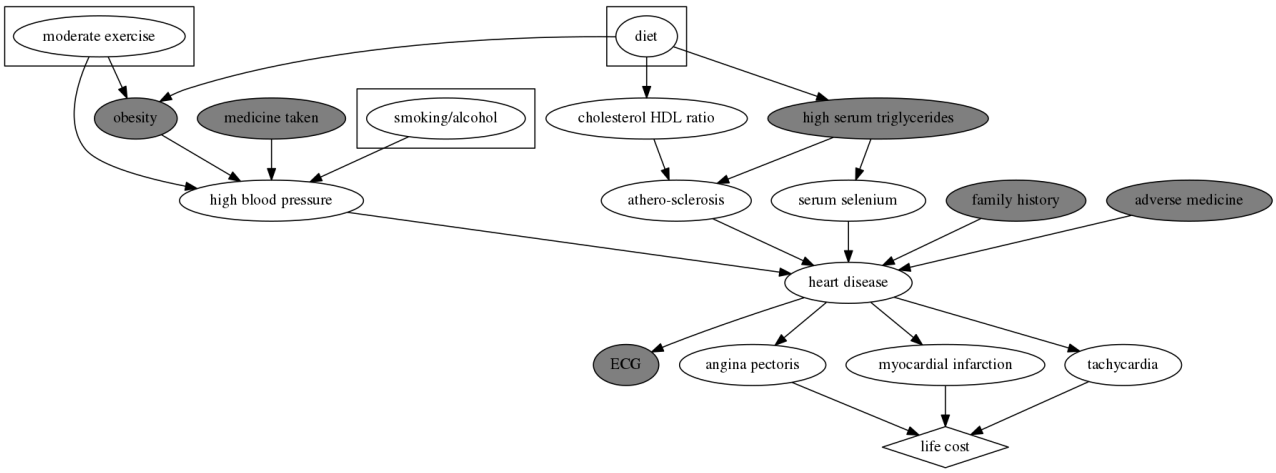


Heart Disease Data

Now in this model, as before, the dark grey nodes are known variables, so they have been recorded by the doctors. The light grey nodes are those for which we do not always have data. So some patients will have had an ECG done, but not all. Some have their serum selenium measured, but not all. Variables that are not included in a particular sample, but are often included in other such data samples, are called [missing data](https://en.wikipedia.org/wiki/Missing_data). Missing data are common in some domains and cause all

sorts of problems for learning algorithms.

This model is being used for a prescriptive analysis task, so the model for the task might look as given in the *Heart Disease and Lifestyle* figure. The patient has his ECG and serum triglycerides measured, and the doctor says: "It looks bad Arnold, you're going to have to make some tough lifestyle choices, no more whiskey and cigars."



Heart Disease and Lifestyle

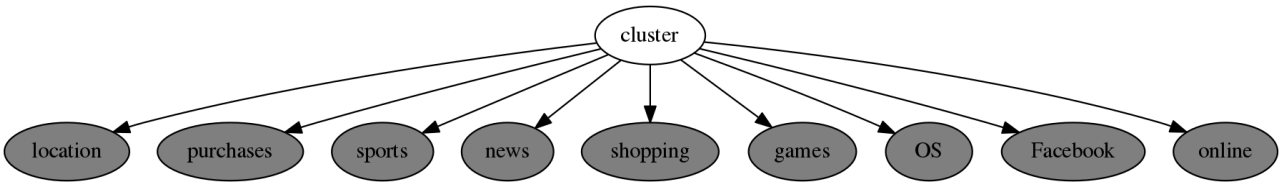
Here, the life cost objective has been added back. The decision variables are now the diet, smoking/alcohol and moderate exercise and the prescriptive analysis involves changing these to see how the lifestyle will be affected.

So we see the structure of this model, for prescriptive analysis, is far more interesting.

Segmenting consumers for advertising

Segmentation (referred to previously as clustering), when applied to customers, is the process of splitting the customers into groups that you believe are similar, in some sense. The hope is that similar customers have similar behaviours, and thus you can target them with the same kinds of advertisements. The general idea is called [market segmentation](https://en.wikipedia.org/wiki/Market_segmentation) and it can be applied to demographic data or previous activities (for instance, purchase history or internet click-throughs) for the customer. The main advantage of segmentation is the simplification it achieves: the customers are now placed in broad buckets and you can do your analysis and prediction per bucket rather than per customer.

For customer segmentation, the problem is just like predictive analysis. except we have no known values for the target variable. We know the variable is "segment" or "cluster", but we do not have values for it. Data for a characteristic example for the online advertising case is shown in the *Customer Segmentation* figure.



Segmentation model

Here, there purchase history and online frequency is somehow characterised, whether they are on Facebook, and their OS they use (Mac, Windows7, Android, *etc.*), as well as some variables giving the kinds of webpages they frequent and their rough location inferred by proxy such as IP address.

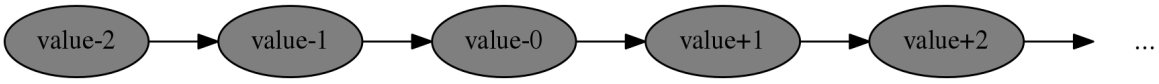
The statistical learning, applied to this data, is given the job of filling in the "cluster" variable values. That is, it is to assign each customer to a cluster. In statistical terms, the "cluster" variable is called a [latent variable](https://en.wikipedia.org/wiki/Latent_variable), which means a state that is existing but not yet developed or manifest, or it is currently hidden or concealed.

We do not consider a resulting analysis or evaluation for this learning task. In principle, one could evaluate the quality of the advertising one does with this segmentation. But this, however, is non-trivial to set up.

Simple dynamic models

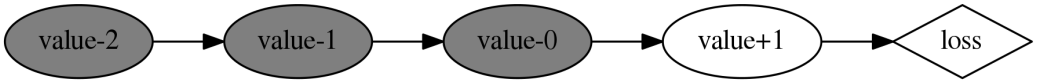
A particularly important and broad class of models are dynamic models. These are used in a context where sequencing is important, for instance for things happening in time. Streaming data often arises from this context. Patient monitoring in hospitals, data from student e-learning environments such as MOOCs, and internet browsing behaviour through a complex website are often modelled with dynamic models.

Data for the simplest kind of dynamic model is presented in figure *Time Series (1st order)*.



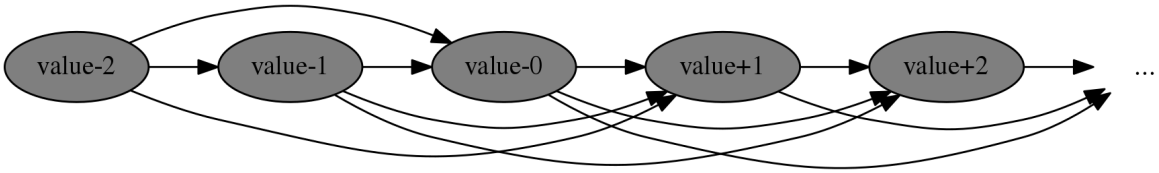
Time Series (1st order)

In this case, the $value_i$ variable is predicted from the $value_{i-1}$ variable. A matching predictive analysis is presented in figure *Time Series (1st order) Prediction.*



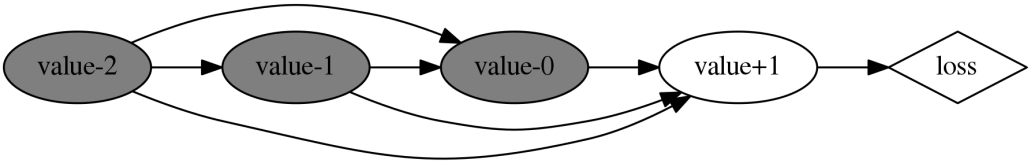
Time Series (1st order) Prediction

Now we can go further back to do a prediction. So the $value_i$ variable could also be predicted from the $value_{i-1}$, $value_{i-2}$ and $value_{i-3}$ variable. What sort of difference does this make? If the value is a location, then the *Time Series (1st order)* model basically says you look at the previous location only. Whereas the *Time Series (3rd order)* model looks at the previous three, which means it can also consider, approximately, the velocity and acceleration. For a car, for instance, the *Time Series (3rd order)* model would be better.



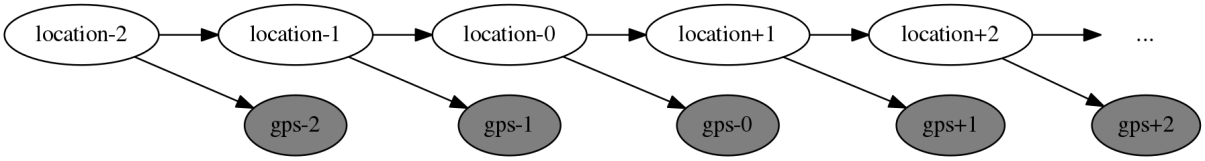
Time Series (3rd order)

Likewise, the matching predictive analysis is presented in figure *Time Series (3rd order) Prediction.*



Time Series (3rd order) Prediction

The next level of complexity is to introduce latent variables. The GPS devices most of us have in our phones have, approximately, the sort of model given the *GPS Location* figure.



GPS Location

This works as follows. The GPS device uses three satellite signals to get an estimate of the current location that is accurate, usually, to within 5 or so metres. This figure represents that we do not know our current location, and only get to measure it using GPS. But we can also use the previously estimate location to estimate our current location, with perhaps a small unknown change in location that might happen due to movement. So the $location_i$ variable is estimated from the $location_{i-1}$ and also considering the gps_i variable.

Further resources on graphical models

Perhaps the best book discussing the combination of graphical models and statistical learning is the text book "Machine Learning: a Probabilistic Perspective" (http://www.cs.ubc.ca/~murphyk/MLbook/) by Kevin Murphy, and this covers both the graphical models and the statistical theory behind learning. It is a textbook for a full year (two semester) unit presented at the masters level.

To illustrate different kinds of models, consider the following.

- A **causal loop diagram** is a dynamic Bayesian network where arcs go from the $t-1$-th (previous) value of the variable to the t-th (current) value of the variable; solid arcs represent positive influence and dashed arcs represent negative influence. A **complex dynamic model for obesity**, presented as a causal loop diagram (http://www.shiftn.com/obesity/Full-Map.html), was developed by an organisation called shiftN. This is a complex model but worth a quick look:
  - Background about the diagram (http://www.shiftn.com/news/detail/interactive_functionality_obesity_systems_map_restored) is on their webpage.
  - Note this causal loop diagram was developed by hand (using expert consultants), but consider what would have to be done to learn this from data.
- A diagram (https://johnroscoe.files.wordpress.com/2013/10/pages-from-edward-tufte-beautiful-evidence-2006-pdf-hi-res.jpg) unintentionally similar to an influence diagram illustrating **the development of cubism and abstract art**, (note the word "Negro" in this case means "African") that appeared on a blog about Edward Tuft (https://johnroscoe.wordpress.com/2013/10/24/a-love-letter-to-edward-tufte-part-four/).
  - "Yet this neat array of 51 causal links neatly summarizes almost fifty years of the most frenetic period in artistic expression the Western world has ever known."

Characterising learning problems

One way of characterising different learning problems is by analysing their basic structure. The main characteristics described above are:

Prediction	Dynamic	Missing DATA	Latent variables
Is the task a simple prediction, or a more complex analysis?	Does the task repeat over space or time?	Do some of the variables missing have missing data? (note they cannot be 100% missing)	Are there latent variables? Note the target variable for a prediction task cannot be latent.

Introduction to learning theory

In this section a number of classic results from learning theory are presented in an intuitive way.

Truth

In a typical scientific or business problem you may have a large amount of data. You may be able to get astronomically more data, given an adequate budget. But what is the underlying "truth". In Physics, we can talk about Newtonian Physics as being a model of physics, but we know it is just a model and not a true reflection of the world (which apparently is more complex). In Medicine, we are told that a pregnant woman should take certain vitamins or avoid certain medicines, and doctors may know the general mechanisms behind the negative effects but the precise quantities influencing subsequent degree of the effects will usually not be known. All of this hidden detail were refer to as the "truth". Philosophers and statisticians can talk at great length about what the Truth (Wikipedia) (https://en.wikipedia.org/wiki/Truth) is and how we might try and obtain some measure of truth from data.

Our concern for the "truth" is when we make a new prediction, or take an action and wait to see an outcome (a cost or value). We can never be perfect, but we hope to be as good as possible. A "true" model is a model that makes probabilistic assessments from which we can make predictions and take actions that are as good as possible given the context. Some examples:

- For an unbiased six-sided die, the "true" model makes each of the six outcomes equally likely. This we assume.
- For the *Heart Disease and Lifestyle* figure above, the "true" model is one that gives probabilities in agreement with the long-run frequencies for patients attending the clinic. This we never know.

Of course, there are always exceptions and challenges (what if the wind blows, what if the economy changes, *etc*.). However, in most cases,, we do not know the "truth" so it is usually some idealistic unknown used as a point of reference but never able to be measured.

Evaluating quality

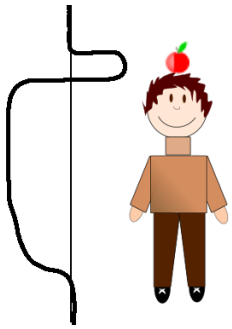
The very idea of "learning" means that you are getting better at something. In some learning scenarios, you have some predictive task and you are trying to do better at the predictions. In other tasks you are making an action, and the outcome then determines the gain or loss.

William Tell's apple shot

To illustrate this, consider the tale of William Tell (https://en.wikipedia.org/wiki/William_Tell), who was forced to

shoot an arrow that was sitting atop his son's head, the so-called "apple shot". This is given in the etching, with the boy, his son, on the far left.

William Tell's Apple Shot by Hans Rudolf Manuel Deutsch (1525-1571) (Sebastian Münster, Cosmographia)

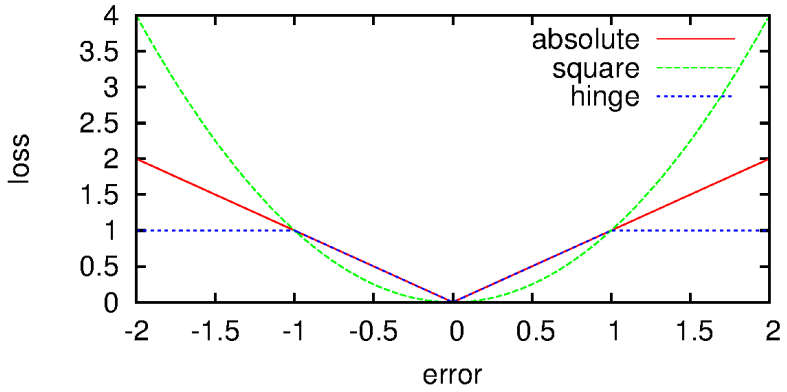


Apple Shot Loss Function

In this case, when William Tell shoots his arrow, the outcome is where the arrow strikes. The loss function in terms of the height of the arrow is presented in the *Apple Shoot Loss Function* figure. This shows the loss as a function of the height, relative to the boy's height and the apple. If it strikes the apple, he win's their freedom so there is a gain, shown as positive (*i.e.*, the loss is drawn to the right). If it strikes lower, he may kill his son, so their is a large loss, shown as negative (*i.e.*, the loss is drawn to the left). Hitting the legs is not so negative, and missing both boy and apple is a zero loss.

Real valued prediction and loss

Evaluation of predictions can be measured in many different ways. If the value being predicted is a real value, then one way of evaluating, measuring quality, is to have a "loss function" (Wikipedia) (https://en.wikipedia.org/wiki/Loss_function). A loss function should return zero if the prediction is perfect, and otherwise the loss function should be positive and be an increasing function of the distance between the prediction and the truth. The greater the distance (or "error"), the higher the loss. A plot of some common mathematical loss functions is given in the *Loss Functions* figure.



Loss Functions

These are as follows:

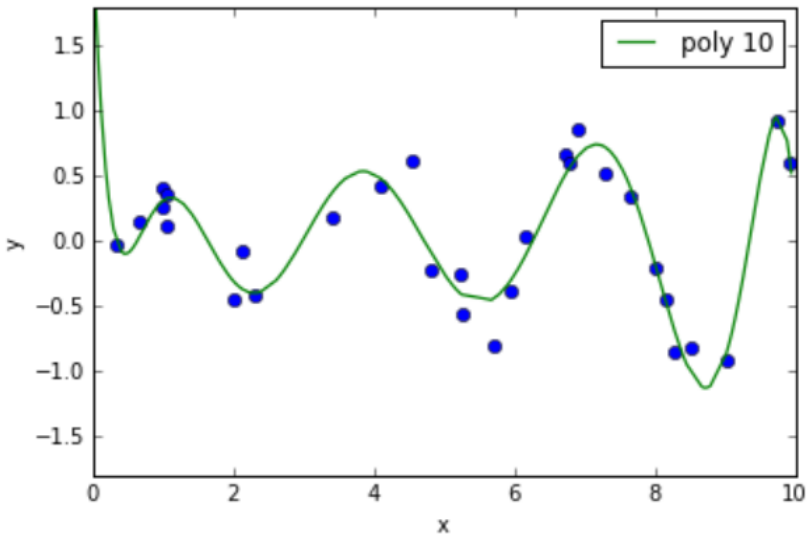
$$\text{absolute}(x) = |x|$$
$$\text{square}(x) = x * x$$
$$\text{hinge}(x) = \begin{cases} |x| & \text{if } |x| \leq 1 \\ 1 & \text{otherwise} \end{cases}$$

The absolute loss takes the absolute value of the error. The hinge loss is like an absolute loss but saturates at 1.0. The square loss squares the error. So the square loss penalises large errors more than the absolute loss, whereas the hinge loss treats all errors with an absolute value greater than 1 the same. Of course all these losses are computed relative to whatever scale the error is measured at. So this kind of loss is relative only; the scale in some sense is meaningless and what matters is the relative loss between different error values.

Convergence and sample sizes

With learning, we expect to improve over time. In statistical theory, "in time" is formalised by "as the training set size gets bigger.

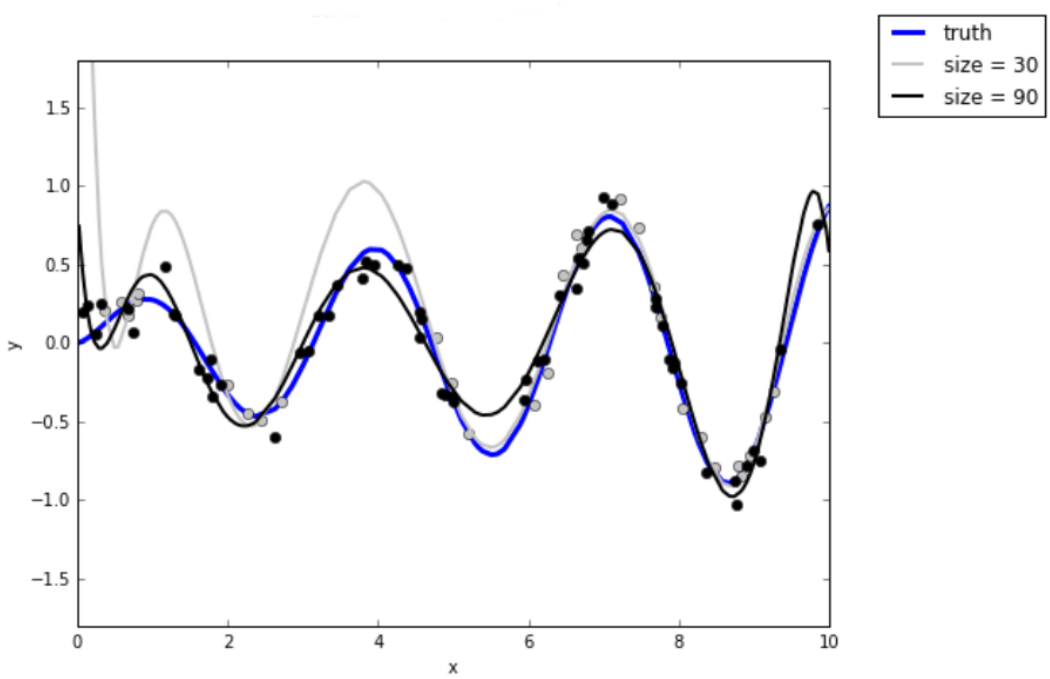
Let us investigate this idea using the simple notion of curve-fitting, sometimes called regression. This works as follows: a set of data points is given on the 2-D plane, data in the form of (X,Y) pairs. This is shown in the figure *Simple Regression*.



Simple Regression

In this figure, the data is plotted with blue dots, and this is the "training set". Our regression algorithm, and the details do not matter too much, just assume we have one, makes an estimate for the "true" line based on the training set as input. It estimates that it follows the green curve. Now you might think you can do better, and perhaps you can. For instance, the upward rise of the blue curve near $x = 0$ seems wrong. But, well, this is what our algorithm produces automatically.

Now lets consider what happens when we increase the training set. This situation is shown in the figure *Regression on a Growing Sample*.



Regression on a Growing Sample

Now this is a complex figure, so lets work through it. Note for this example we know the "truth". The data has been generated from the curve given in thick blue.

1. We get a sample of 30 X values generated uniformly in the region [0,10].
2. We get a corresponding sample of 30 Y values, shown with the grey dots. Note these do not match

the "truth" (the blue line) exactly. This is because we have added some noise to the sample, changing the y value a bit. Then we fit these 30 data points using our regression algorithm. The estimated line is shown in grey.

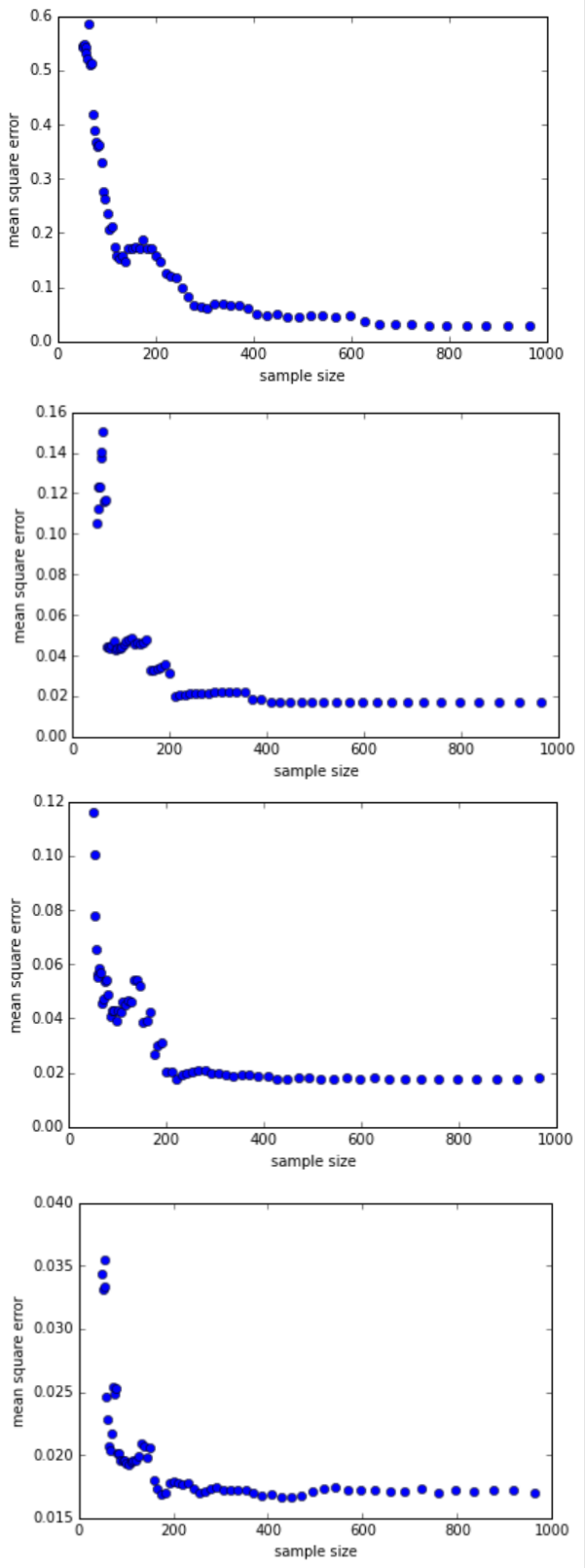
3. Now we extend the sample with 60 more black points, so the data now has 90 points in total. Again, we fit these 90 data points using our regression algorithm. The second estimated line is shown in black.

So we have one "true" line in blue, and two estimated lines in grey and black. Note the black line is generally closer than the grey line. The grey line is really quite bad on the left half of the plot. The 60 extra data points make the black line fit better. This is learning in action.

So now lets extend this same procedure. We will do many different regressions on an increasing sized sample. But doing all these plots will get rather messy. So instead we will distil the quality of the fit into one number. This is the *mean square error*. If your data has the form of a sequence of N values like $(x_1, y_,), (x_2, y_2), \ldots, (x_N, y_N)$, then the mean square error (computed on the training set), MSE_{train}, is computed as

$$MSE_{train} = \frac{1}{N} \sum_{i=1}^{N} (truth(x_i) - y_i)^2$$

This formula is the mean of the squared error.



In the plots we see the mean square error for four different scenarios. In each case the data set starts at 50 sample points, and gradually increases to 1000 sample points. You can see that the MSE_{train} jumps around a lot initially, and gradually settles down to a minimum. This initial erratic behaviour is the the effect of randomness. Sometimes unusual data points (with a large noise) are seen, and initially they can have a bad effect. Note also, the minimum is not zero! It never gets a zero error fit. For some

problems a perfect zero error fit is possible in the long run, but often times it is not. In this case zero error is impossible in the long run because we are trying to fit a 10-th order polynomial and the "true" function is transcendental (Wikipedia) (https://en.wikipedia.org/wiki/Transcendental_function). The fact that the error settles on a minimum as the training set gets very large is referred to as *convergence*. The *rate of convergence* is the speed (measured in size of training set) at which convergence happens.

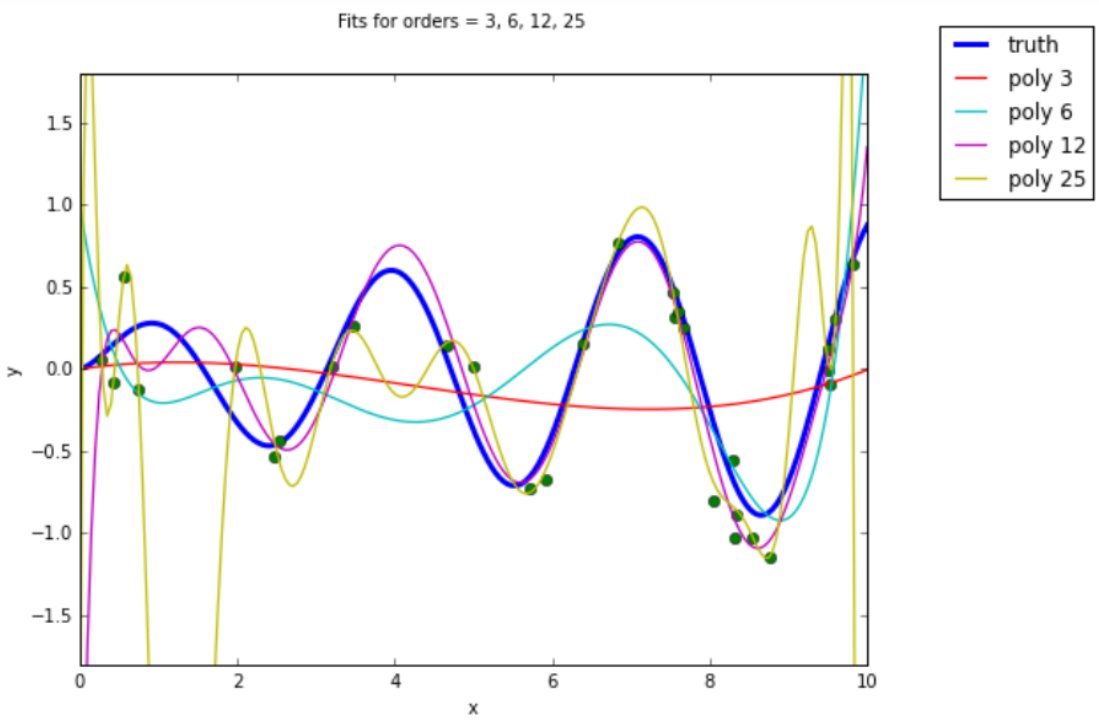
These curves a called *learning curves*. Note they are subtly different to what psychologist call learning curves (Wikipedia) (https://en.wikipedia.org/wiki/Learning_curve). The psychology use generally measures "human experience" on the X-axis, and humans do not learn the same way as machines. But the general idea is similar, though sometimes they may flip the Y-axis so higher is better. For instance, if they measure mean accuracy instead of mean error, then higher is better.

Under-fitting and Over-fitting

Consider the same situation as the regression fit above. Now we can use many different kinds of models to fit the 30 points. The figure below, *Fitting different order polynomials*, shows what happens when you fit polynomials of order 3, 6, 12 and 25 to the 30 data points. For instance the 12-order polynomial takes the form

$$y = a_0 + \sum_{i=1}^{12} a_i x^i$$

and the parameters $a_0, a_1, \ldots, a_{12}$ are set to make the mean square error on the 30 data points minimum.



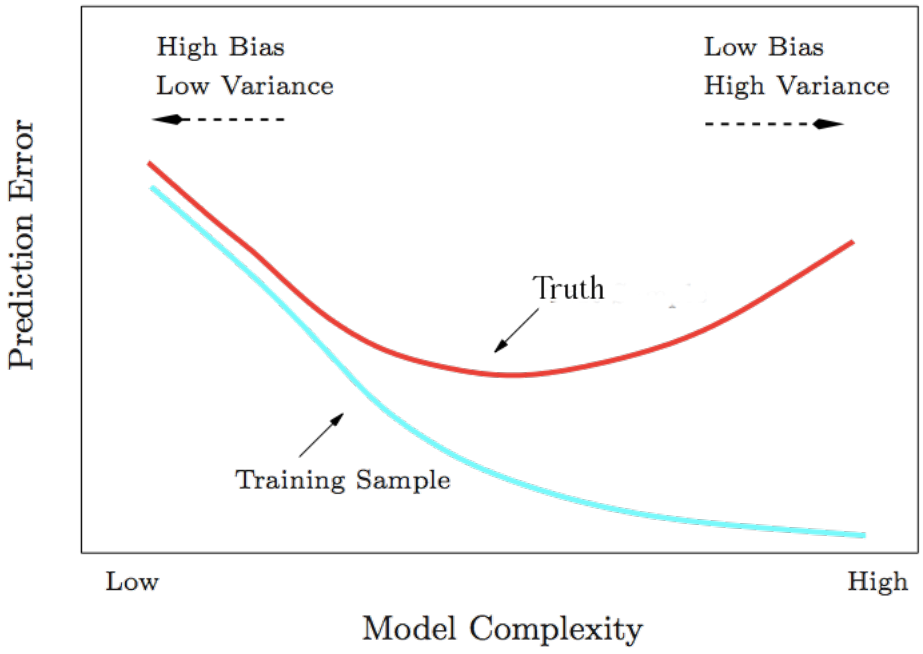
Fitting different order polynomials

One can see that the 3rd order polynomial, in red, just does not have the flexibility to fit this wavy (true) curve. The 12-th order polynomial, in purple, is just about right, but has trouble at the extremes, and the 25-th order polynomial is, well, wild. It has way too many degrees of freedom so includes all sorts of crazy peaks to try and fit the data really closely, in some cases, almost exactly. based on these descriptions, we

say the 3rd-order polynomial is **under-fitting** the data, and the 25-th order polynomial is **over-fitting** the data. Under-fitting and over-fitting are descriptions we use relative to a particular training set and model. In the above example, we compare the fits with the truth, so we know where they apply.

Also, we say the 3rd-order model has **high bias**, which means the model makes very strong assumptions about the data, assumptions can be wrong giving the model large error in the limit of ideal large data sets. In contrast, we say the 12th-order model has **low bias**, which means the model makes weaker assumptions about the data, assumptions giving the model smaller error in the limit of ideal large data sets. High bias and low bias are known, measurable properties of a model, independent of any data set.

The underlying issue here is what kind of **complexity** should we use in our model. if we have too much complexity, then the model has the ability to twist and bend to arbitrary kinds of data, and so it does without any real justification. This is a low bias model. So it will make poor predictions on the future data if there is not large enough data to counteract the effect. If we have too little complexity, then the model does not fit the data well, so should also make poor predictions on future data. This is a high bias model. These distinctions are depicted on the *Bias-Variance Model* figure. This is a cartoon figure to show what happens as we change model complexity, using error as a generic proxy for model quality. While we can observe the training set error, we might not have a test set to estimate true error in an unbiased way, which means we cannot always measure the red curve, so we might not be able to determine where the sweet spot is, where we have just the right amount of model complexity.



Bias-Variance Model

Note in the *Bias-Variance Model* figure, the high bias situation shows a small difference between training set error and "truth", where as the small bias situation shows a large difference between training set error and "truth". Small bias models need less data to fit well, because there is less degrees of freedom in the model. Therefore, the training set error they report is closer to the truth.

It would be nice if we knew just which order polynomial to try, and how much complexity we need. Unfortunately, few data science problems come neatly packaged and tagged with hints and instructions on what to do. In the plot above, we have been fortunate to know what the truth is. In practice we do not. So one of the big challenges in data analysis is to estimate the right level of complexity. This is usually covered in more advanced statistical or data analysis classes, and a variety of techniques exist. Technical

terms used for this rinclude regularisation, Bayesian estimation, minimum description length and cross-validation. These are also the most controversial parts of the field, so heated arguments do happen, and some communities show particular preferences.

Training sets versus test sets

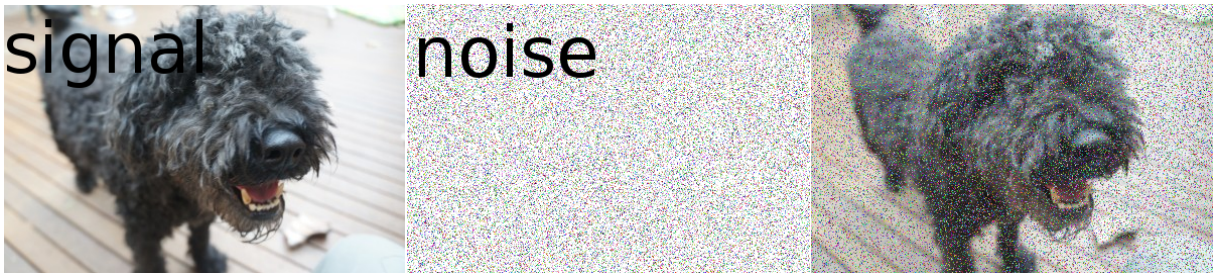
In the above subsection we considered the problem of over-fitting versus under-fitting. Now in learning, we always want more data. The larger our training set, the more accurate model we should be able to build. However, we are also faced with the problem of trying to prevent over-fitting. We would like to know whether we are over-fitting. One way to do that is to hold out a test set. We keep part of the original, full data set aside, and use this, intermittently, to estimate the "true" error we have. However, doing so, we also harm our training, as we have less data to train on. If we have a smaller amount of data, sometimes the problem in data science, then we will not want to hold-out much data. If we hold out too little data, then the error measured on the test set itself will be noisy and thus be a poor estimate of "true" error. So if you are holding out a test set, finding the right compromise between train and test data set sizes is a real problem. Now some methods, such as Bayesian techniques, do not require a test set to operate, but these come with issues of their own. Learning is not an easy task!

Signal versus noise

In image processing and related areas, people talk about signal versus noise. *Signal* is the image or video you want to communicate across a "channel", for instance via radio waves or across the internet. *Noise* (https://en.wikipedia.org/wiki/Noise_%28signal_processing%29) is the unwanted artifacts that are picked up along the way, perhaps due to quantitization, transient power surges, errors in communication, *etc*.

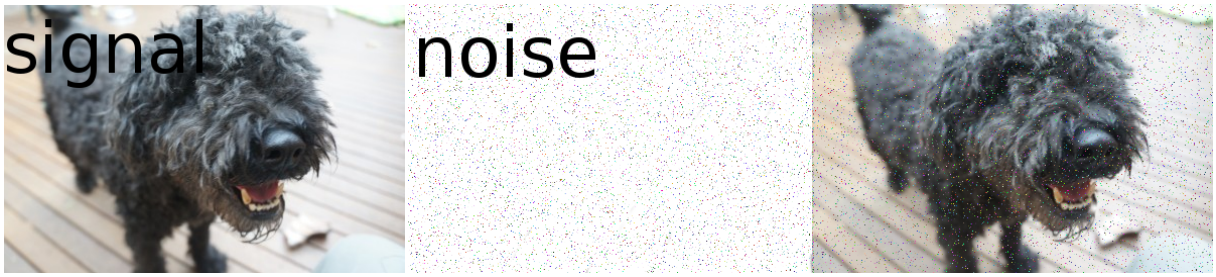
Noise in our problems may be due to measurement error, or it may reflect that the domain is intrinsically uncertain. For instance, if measuring your body temperature under the tongue, you expect the value to be 36.8±0.4 if you are well. The ±0.4 constitutes auxiliary effects: time of day, weather, your activity, *etc*. These are uncertainties in the measurement. Probably ±0.2 would be true measurement error, that tiny thermometer. If measuring the time in minutes it takes you to drive across town to work, then you have a mean travel time but this can vary quite widely due to weather, major events in town, road works, *etc*., that again are uncertainties, not true measurement error. However, when we are collecting data and plotting values on a graph, we do not know about the details behind the uncertainties so they are hard to distinguish from noise. So, often times, we use the term noise to include the effect of all these other unmeasured uncertainties.

Signal versus noise is best illustrated with images. In the figure *Signal plus Noise*, we see noise added to the picture of Ngozi the dog.



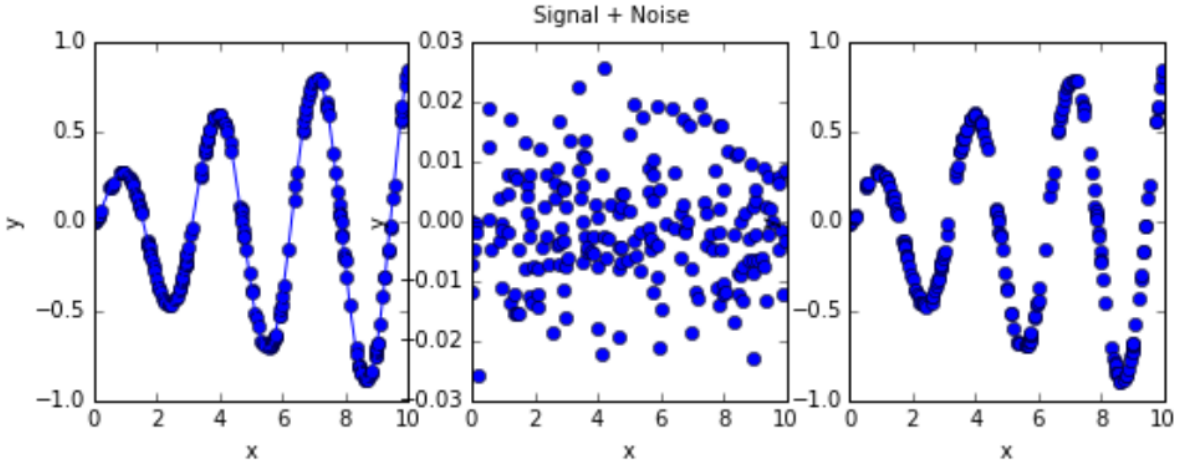
Signal plus Noise

Whereas, in the figure *Signal plus low Noise*, we see less noise added to the picture of Ngozi the dog.

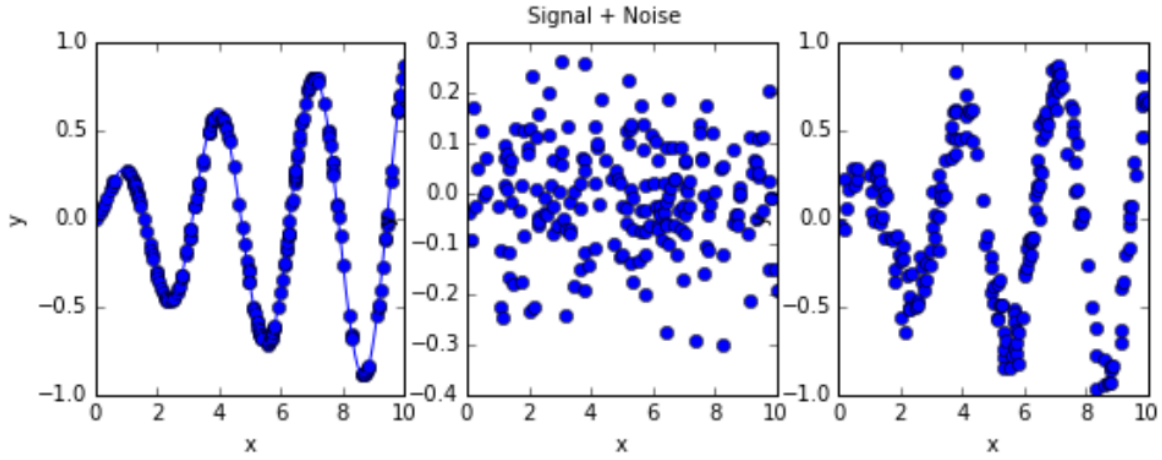


Signal plus low Noise

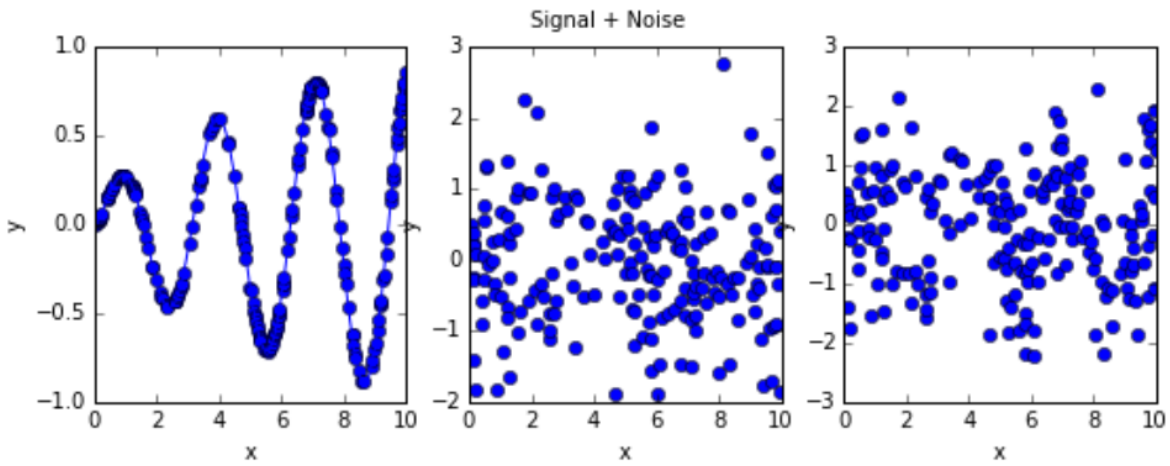
In a perfect world, the problem looks quite simple: if we want to reconstruct the original signal from the received image, "all we have to do" is subtract the noise. In practice, we never know what the noise is. The problem of learning is related to the problem of reconstructing the original signal. To see this, here is a variant of the above idea, but using points on a curve, instead of an image.



Curve plus low Noise



Curve plus medium Noise



Curve plus high Noise

Here we see three different noise levels, standard deviations that are low (0.01), medium (0.1) or high (0.9).

One can see with high noise the signals, the curve, is washed out, it disappears in the noise. With medium level noise, the signal is fairly discernible. With low noise, the data looks exactly like the signal. What constitutes "low" versus "high"? Well, it depends on the level of the signal. So engineers talk about the signal to noise ratio (https://en.wikipedia.org/wiki/Signal-to-noise_ratio), which is a concept we may not use in any formal, well defined sense, but it is a very important concept in an approximate sense (as in the "low, "high" examples above).

When learning, we need to consider the impact of unmeasured uncertainties, noise, *etc*. and what level it might reasonably be.

Correlation versus causation

Correlation (Wikipedia) (https://en.wikipedia.org/wiki/Correlation_and_dependence) is the statistical association between variables. What this means is that the observed values appear to be dependent in some way. For instance:

- "rained last night" and "lawn is wet this morning" are dependent
- "suit is hearts" and "card is royalty" are independent for cards drawn from a full deck
- "suit is hearts" and "card is royalty" are dependent for cards drawn from a full deck with the red queens removed
- "patient has long history of smoking" and "patient has lung cancer" are dependent

Correlation is relatively easy to assess. We observe data in an unbiased way, recording the values of the variables in question. When we have a large enough sample, we compute a measure of correlation (Wikipedia) (https://en.wikipedia.org/wiki/Correlation_coefficient) on the data or a measure of dependence. We cannot get a definitive answer on correlation, but we can often say with confidence if something is correlated or not, given enough data.

Causality (Wikipedia) (https://en.wikipedia.org/wiki/Causality), on the other hand, has a very different meaning. Two events A and B are observed. A is said to *cause* B if A is in some sense responsible for B. In this case B is the *effect* of A. Now, in some cases, cause and effect are easily seen from our understanding of physics. if a car going at speed runs into a garden fence and the fence is demolished before our eyes, then we can quite reasonably say the car caused the fence to be demolished. In other cases, cause and effect are mainly hidden from us. For a long time it was not known that smoking caused cancer because the

mechanism for the effect is complex biology, unknown to all but a very few. In this case we can readily measure a correlation between smoking and cancer, but does smoking *cause* cancer? Perhaps people susceptible to getting addicted to smoking have a gene that also makes them susceptible to lung cancer.

One accepted way of measuring causality is by *intervention*. The basic idea, informally, is as follows:

> To test if A causes B, we hold every other variable fixed, then force A to have a certain value in an observed population, and observe B. If, in this situation, A is correlated with B, then it follows that A causes B.

Here the *intervention* is the act of forcing A to have a certain value. This concept of intervention is the basic idea behind clinical trials in medicine where the subjects are split into a treatment group, that gets the "treatment", and a placebo group that does not. So in this case, by using intervention we then test for causality using correlation.

However, without the act of intervention [correlation does not imply causation (Wikipedia)](https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation). The Wikipedia gives some excellent examples.

> **Example:** The faster windmills are observed to rotate, the more wind is observed to be. Therefore wind is caused by the rotation of windmills. (Or, simply put: windmills, as their name indicates, are machines used to produce wind.)
>
> In this example, the correlation (simultaneity) between windmill activity and wind velocity does not imply that wind is caused by windmills. It is rather the other way around, as suggested by the fact that wind doesn't need windmills to exist, while windmills need wind to rotate. Wind can be observed in places where there are no windmills or non-rotating windmills-and there are good reasons to believe that wind existed before the invention of windmills.
>
> **Example:** Sleeping with one's shoes on is strongly correlated with waking up with a headache. Therefore, sleeping with one's shoes on causes headache.
>
> The above example commits the correlation-implies-causation fallacy, as it prematurely concludes that sleeping with one's shoes on causes headache. A more plausible explanation is that both are caused by a third factor, in this case going to bed drunk, which thereby gives rise to a correlation. So the conclusion is false.

No-free Lunch Theorem

When learning, Wolpert and McCready formalised the idea that "there ain't no such thing as a free lunch" with their [No Free Lunch Theorems (Wikipedia)](https://en.wikipedia.org/wiki/No_free_lunch_theorem). The theorems, called the NFL theorems, are technical and difficult to state in an informal way precisely. But a reasonable explanation which they use is as follows:

> We have dubbed the associated results NFL theorems because they demonstrate that if an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems.

So if a learning algorithm A does well on certain problems, in comparison to others, then there must exist a learning algorithm B that will do better than A on other problems. Or in other words, no one algorithm can be uniformly good in all learning scenarios. This is a theoretical result. In practice, people do find certain algorithms perform well in particular contexts, for instance:

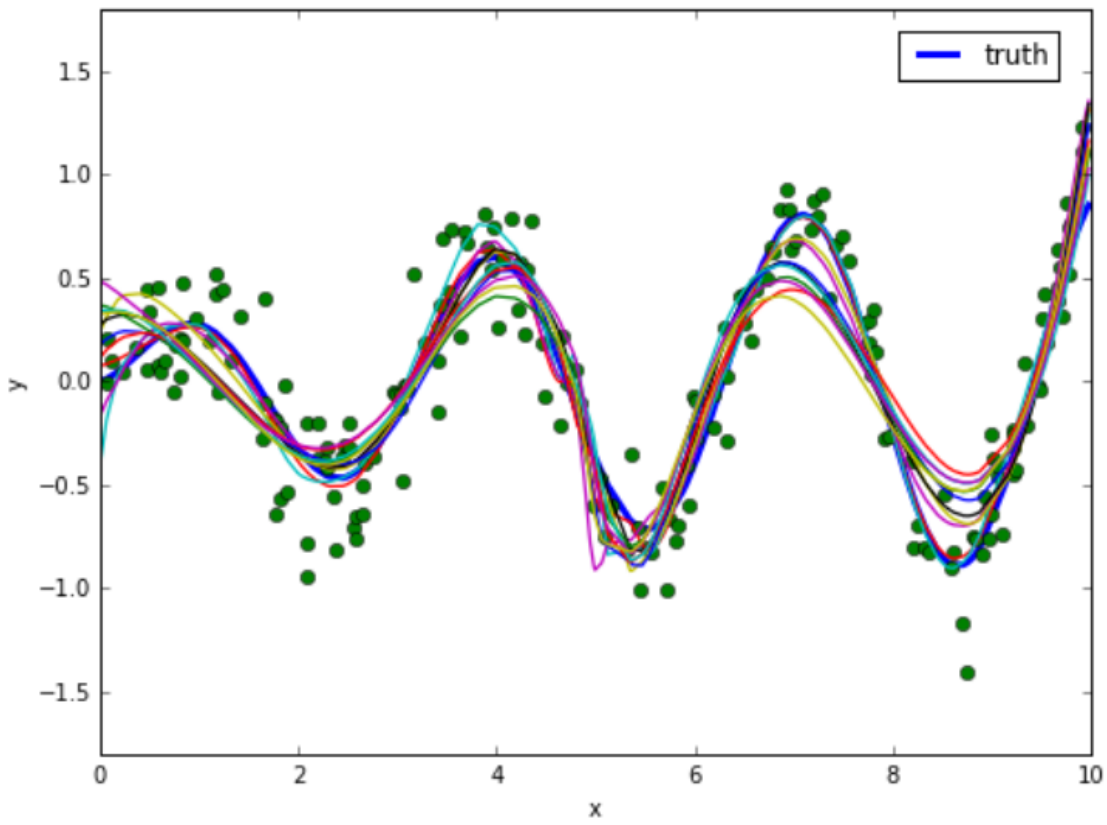
- Naive Bayesian classification performs well for text classification with smaller data sets, and
- linear Support Vector Machines perform well for text classification.

These observations do not contradict the NFL theorems, because they are considering specialised uses of the algorithms.

Ensembles

Two different learning algorithms behave differently with the same data set. So lets now ask a different question. What different fits could reasonably match the finite sample of data we actually have? This is a very important question in practice. For learning, this is the classical "what if" scenario: I don't know the real curve and only have a finite amount of data. Within the assumptions of my algorithm, and fitting the data I have, what different fits are realistic? What variability is there? What are the range of possibilities?

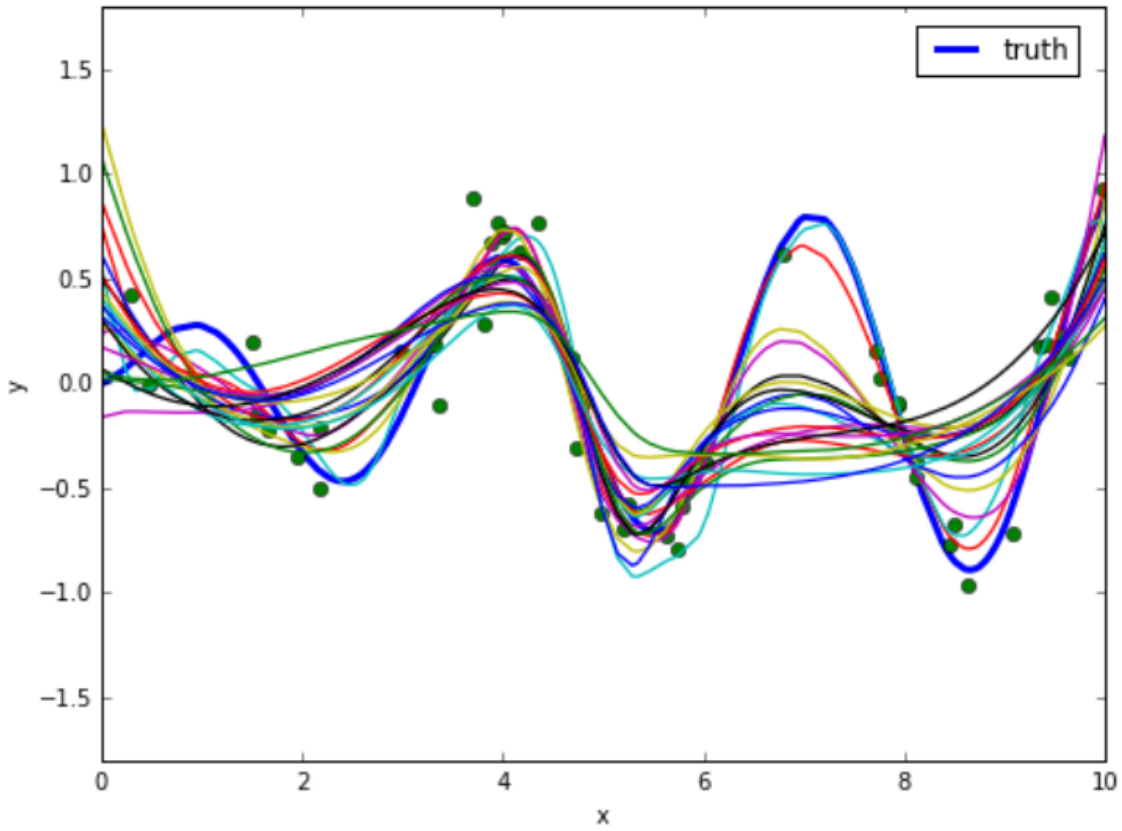
This is the motivation behind generating an *ensemble* of functions to match our data. Figure *Ensemble of Curves (N=200)* illustrates what an ensemble looks like for the regression task in the case where a data set of size 200 is used. Here we see multiple reasonable curves fitting the data. This set of curves helps answer the above questions.



Ensemble of Curves (N=200)

Now in this case, we seem to have a lot of data so the ensemble generally looks like a snake running

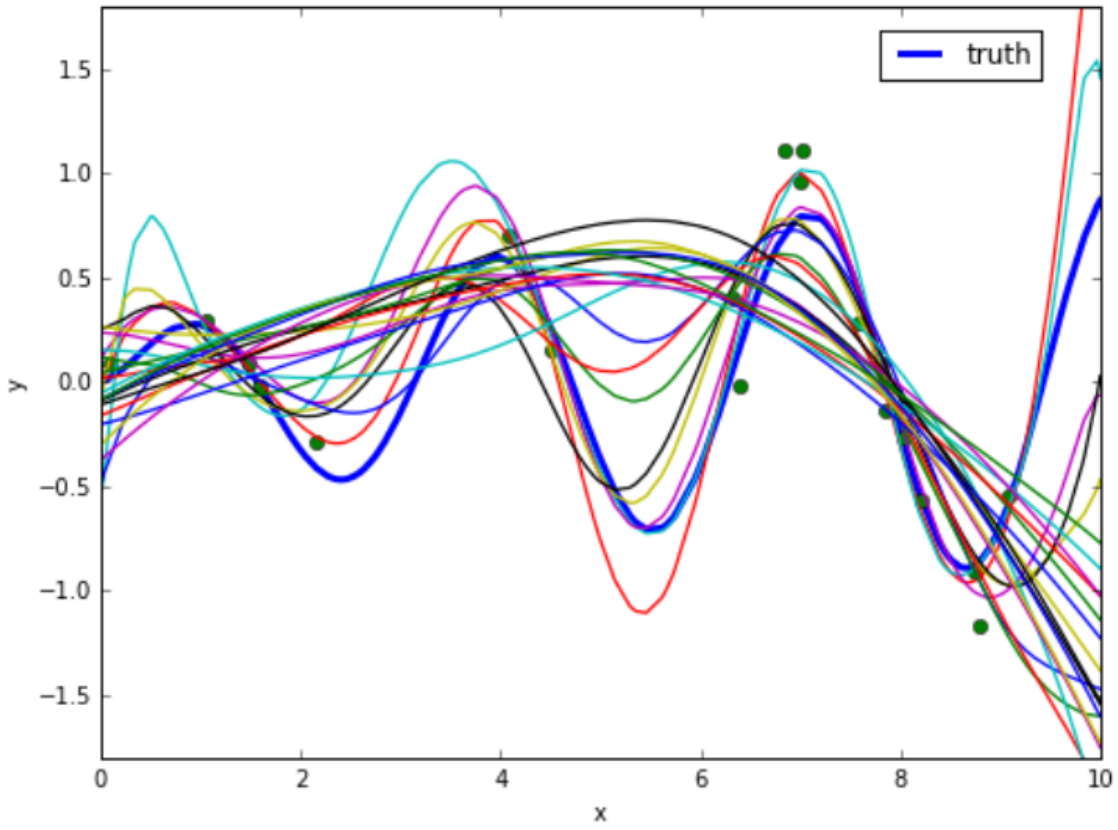
through the centre of the data. It tells us we probably have the "true" curve location accurate on the Y-axis to within about ±0.2. Let us look at another example with less data, this time $N = 50$.



Ensemble of Curves (N=50)

Here the ensemble of curves is starting to show some differentiating behaviour. Notice in the region of $x = 7$ there is only one data point and the ensemble is very uncertain in the Y-axis, perhaps with a range of about 1.2. There is another wider part around $x = 2.5$ due to lack of data. In the region of $x = 4$ there is a lot more data and the ensemble is much narrower in Y. There is also uncertainty around $y = 9$, and we give this a more complex explanation. For all those curves cutting off the peak at $x = 7$, the curves are just continuing more smoothly rather than trying to place a downwards bump at $x = 9$. Finally, a general weakness of regression methods is the end points, in this case $x = 0$ and $x = 10$, where uncertainty also exists.

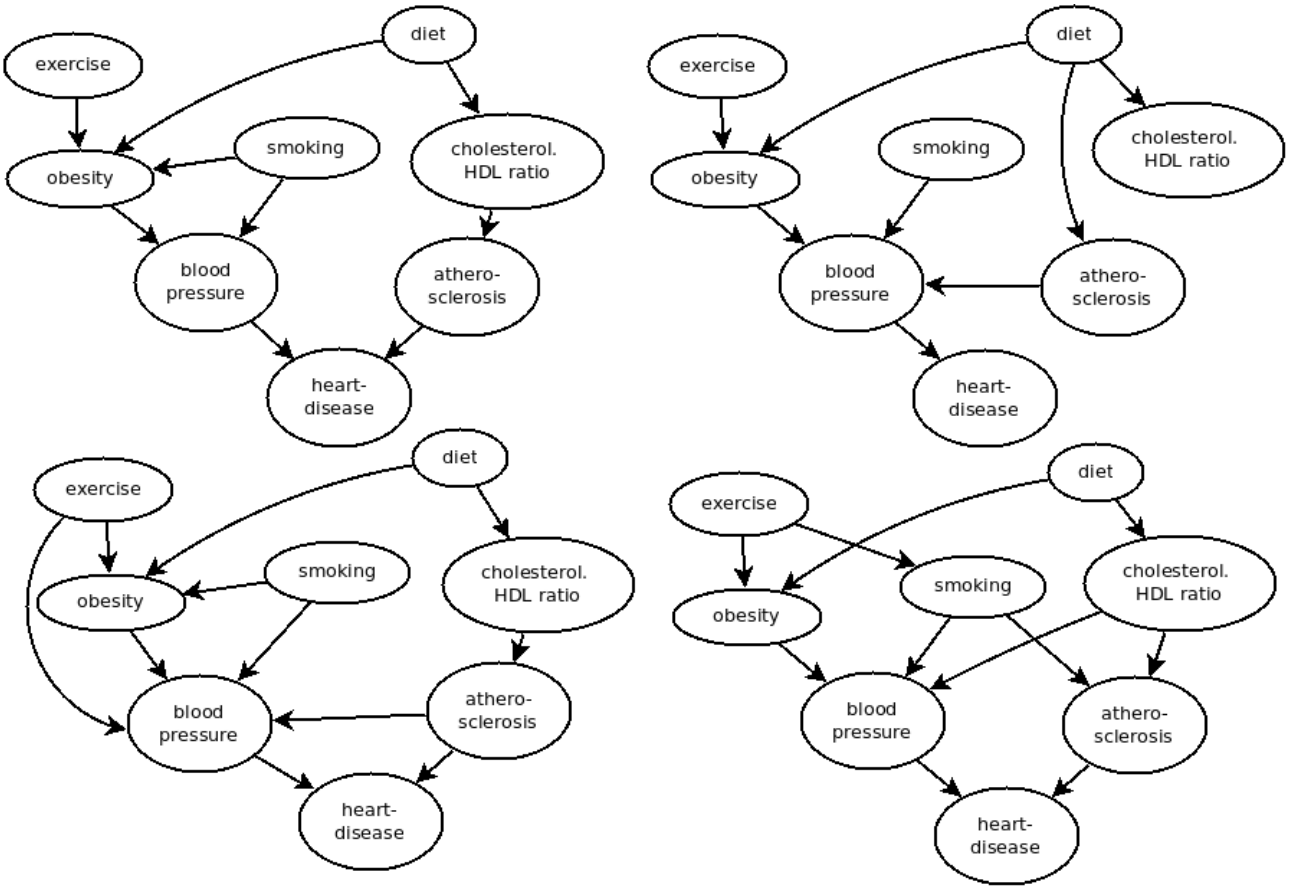
Lets drop now to a smaller data set $N = 20$, almost too small!



Ensemble of Curves (N=20)

Here we see similar behaviour. Where there is little data, the uncertainty in Y is quite large. Moreover, we see things go quite wild at the end point $x = 10$.

Averaging predictions over an ensemble (https://en.wikipedia.org/wiki/Ensemble_learning) turns out to be one of the best ways of improving on a particular method for doing prediction. Working with an ensemble is a way of making multiple bets or hedging against using just one model. So lets suppose you are considering the heart disease problem discussed above in subsection **Lifestyle choices for a heart disease patient**. You have gathered 200 cases after some time obtaining permission from patients at the local clinic. Now you try to build a graphical model. The problem is, your data is not adequate to distinguish between a set of reasonable candidates. Your statisticians tell you that their software says either of the four models in figure *Four Possible Heart Models* could be used.



Four Possible Heart Models

What can you do? Well all the models have an arc from "diet" to "obesity" but not all have an arc from "smoking" to "obesity". In the ensemble approach, you would use all models at once. You make individual predictions with each model in turn and then pool the results.

What is needed to use ensembles is a method for generating the multiple models that make up the ensemble. With a learning algorithm, we can attempt to create an ensemble in a number of ways:

- working with random subsets of the features (columns) of the data set
- training on random subsets of the data
- making some random choices in the algorithm (note so-called Monte-Carlo algorithms do this already)
- use a variety of different algorithms

Imputation

When characterising learning problems above, we mentioned the problem of missing data. Consider the following table for the Housing Loans scenario discussed above. Remember, for this problem the target variable is whether the client defaulted or not, given by DEFAULT. In this situation, the issue is how do we deal with the missing values (indicated by "?") for case 002 of DEFAULT, case 003 of AGE and case 004 of JOB. Some learning algorithms, such as Support Vector Machines or Random Forests, require all values to be known, so they cannot work with missing values. So what shall we do?

ID	Age	Amount	Duration	Job	Housing	Marital	Default
001	43	$200,000	240	A	apartment	yes	no
002	27	$150,000	280	A	apartment	no	**?**

003	**?**	$180,000	240	B	house	yes	no
004	42	$200,000	240	**?**	apartment	yes	no
005	31	$300,000	240	C	house	yes	no

In this case DEFAULT, the target variable, the one to be predicted, is missing for case 002. Therefore, one could discard this case from the dataset. AGE is missing for case 003, so one could guess the AGE. For instance, replace it by the mean of the known ages, which rounds up to 36. Also, JOB is missing for CASE 004. One could replace it by the most frequent JOB, which is A. Alternatively, one could discard all cases with missing values, which removes cases 002, 003 and 004, leaving only two left.

The method for treating the missing values is known as imputation (Wikipedia) (https://en.wikipedia.org/wiki/Imputation_%28statistics%29), and there are several different techniques. In general, imputation is a crude approximation so should be used with caution, for instance when there are few missing values. Some more sophisticated imputation strategies that integrate with learning algorithms do exist.

Semi-supervised learning

A variant of the missing data problem is where only the target variable (the one you want to predict) is missing. Consider the following (different) table for the Housing Loans scenario discussed above. In this situation, we see that cases 002 and 004 have the target missing. Do we just delete these rows from the table and ignore them during learning?

ID	Age	Amount	Duration	Job	Housing	Marital	Default
001	43	$200,000	240	A	apartment	yes	no
002	27	$150,000	280	A	apartment	no	**?**
003	33	$180,000	240	B	house	yes	no
004	42	$200,000	240	A	apartment	yes	**?**
005	31	$300,000	240	C	house	yes	no

Data with the target variable missing is called *unlabelled data*. The target variable, then, is referred to as the label. In many cases unlabelled data is easier to get than labelled data.

Unlabelled data occurs a lot in natural language tasks. Consider the problem of "labelling" proper names in text. We have terabytes of text available on the internet, but very little of it has the proper names properly labelled. So consider the following text about the city of Melbourne, taken from the Wikipedia.

> The metropolis is located on the large natural bay of Port Phillip and expands into the hinterlands toward the Dandenong and Macedon mountain ranges, Mornington Peninsula and Yarra Valley. ... Founded by free settlers from the British Crown colony of Van Diemen's Land on 30 August 1835, in what was then the colony of New South Wales, it was incorporated as a Crown settlement in 1837. It was named "Melbourne" by the Governor of New South Wales, Sir Richard Bourke, in honour of the British Prime Minister of the day, William Lamb, 2nd Viscount Melbourne.

We can "label" the proper names in this text, indicated here by bracketing them like so "[...]":

> The metropolis is located on the large natural bay of [Port Phillip] and expands into the hinterlands toward the [Dandenong] and [Macedon] mountain ranges, [Mornington Peninsula] and [Yarra Valley]. ... Founded by free settlers from the [British Crown] colony of [Van Diemen's Land] on 30 August 1835, in what was then the colony of [New South Wales], it was incorporated as a [Crown] settlement in 1837. It was named "[Melbourne]" by the [Governor of New South Wales], [Sir Richard

> Bourke], in honour of the [British Prime Minister] of the day, [William Lamb, 2nd Viscount Melbourne].

The first paragraph is unlabelled text and the second paragraph is labelled text. So you can see how the term labelled and unlabelled data arose (in natural language processing) and they are now used generally for other kinds of data when referring to the presence or absence of the target variable in the data.

In some cases, unlabelled text can be used to support the predictive task (of predicting the labels). Semi-supervised learning (Wikipedia) (https://en.wikipedia.org/wiki/Semi-supervised_learning) is the learning task where both labelled and unlabelled data is used for building a model.

Generative versus discriminative models

A generative model (Wikipedia) (https://en.wikipedia.org/wiki/Generative_model) is a model that, via sampling, could reproduce similar data sets to the data sample. With a very large sample of data, a good learning algorithm should be able to return a generative model that can generate samples almost indistinguishable from the original.

About the simplest generative model we can have is for normal (Gaussian) data. For this, you have data of the form $x_1, x_2, \ldots, x_N$. Once we estimate the mean and variance for the Gaussian, we can then generate more pseudo-data by sampling using the Gaussian probability density function.

Consider the regression problem discussed in subsection **Convergence and sample sizes**. For this, you have data of the form $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$. A regression model in this case is usually a function from x to an estimate of y, sometimes referred to as $\hat{y}$. So the regression model might be:

$$\hat{y} = 2x^2 + 3.5x - 1.2$$

Now this model is *not generative* because it only tells how to estimate or "generate" y; it says nothing about x. To make regression a generative model for (x, y) data, you need to add a generative model for x. If one is just doing a predictive task, then the extra burden of the generative model for x would seem like unnecessary trouble.

In many cases predictive models are not generative. This applies to the popular models like regression, random forests, support vector machines, *etc.* These models, which discriminate on one variable, are called discriminative models (Wikipedia) (https://en.wikipedia.org/wiki/Discriminative_model). Discriminative models are primarily used for predictive tasks. Prescriptive tasks, which require some optimization and possibly manipulation of possibilities, often use generative models.

5.3
Activity: Regression in iPython

We will use a sequence of iPython notebooks to explore some abstract concepts of learning using linear regression on a two dimensional plot. The linear regression routines themselves are hidden, because the techniques are not important for the concepts being covered. We will, however, be looking at a series of more sophisticated plots. The Python code has been marked in places indicating where you can modify parameters if you wish to change the plots or the repetitions done, for instance.

Note, however, this activity is not intended to be high performance or robust. The implementation of the regression routines is done simply, and it *will not scale* to large numbers of data points (say over 100) or high orders of polynomial regression (say, over 50). So, when experimenting with it, do not stress it too much!

1. The source files are in the zip folder RegressionActivity1.zip (https://www.alexandriarepository.org/wp-content/uploads/20160216144021/RegressionActivity1.zip) (Zip file, 650kb). Unpack the ZIP file into a new directory on your local machine.
2. We will run iPython to operate on that directory. The contents of the zip file are:
   regressiondemo.py
   1.RegressionBasic.ipynb
   2.RegressionStudy.ipynb
   3.RegressionLegendre.ipynb
   4.RegressionLegendreEnsemble.ipynb
   X.RegressionLegendreDetail.ipynb
   The single Python file is a library containing the regression and data creation routines used by the notebooks. The .ipynb file names are prefixed 1,2,3,4,X so you know the order to run them. The last one, X, is not part of the activity. Look at it later if you are curious about the Legendre polynomials.
3. If you run iPython locally, run it using the above directory so it can locate the library "regressiondemo.py". Otherwise, upload the 5 notebooks (".ipynb") and the "regressiondemo.py" file to the same directory in a new project in your SageMath.com account (or similar) and run there.

Now we will work through each notebook in turn.

1.RegressionBasic

This is a simple tutorial on linear regression. Work through Steps 1.1 to 1.4. The last, Step 1.4 will get you to try regressions with different polynomial orders to see what sorts of fits you get. Vary the options and see what happens. Some of you will already be familiar with linear regression. That is good, just do this quickly to ensure we agree on terminology.

What can you see from trying out different amounts of data, and then trying different orders of polynomials? You should learn not to set the order of the polynomial too high! Describe succinctly your advice.

2.RegressionStudy

This tests out the linear regression under different scenarios. Work through Steps 2.1 to 2.3. Once you have worked through it, you can alter the *truefunc()* at Step 2.1, and then "run all" cells to rebuild the

plots.

Here we get to see a story about how well the fitting works:

Step 2.1: how does it work on a given sample with the two extremes, too small an order or too large an order.
Step 2.2: in theory, assuming "infinite" data, how well could it work with small versus large orders.
Step 2.3: revisits the lesson from Step 2.1, but shows how it works with multiple data sets.

3.RegressionLegendre

This tests out the linear regression with smoothed Legendre polynomials under different scenarios. Work through Steps 3.1 to 3.3.

Step 3.1 does the same plot as done with Step 2.3. You will see the smoothed regression behaves quite differently to regular linear regression.
Step 3.2 does a side-by-side comparison between the two versions. You can rerun this with different orders of polynomial. You should see for small orders, the two are indistinguishable. But for higher orders, the smoothed regression damps down the wild behaviour. How do the two compare?
Step 3.3 works on the one data set and does fits to the initial subsets. This mimics how smoothed regression works as a data set grows. You can see how the quality of the fit improves.

4.RegressionLegendreEnsembles

This notebook only has one step, Step 4.1. It generates a set of possible fits for the one data set. Note, normally the regression routines return the best (guess) fit. Here, a range of possibilities are displayed instead. This gives you an idea of the variability in the y-axis in a particular x-region. In regions of x with a lot of data, the range for the the fit should be smaller. When there is no data near a specific x value, the range of corresponding y values displayed is quite large.

5.4
Tools for the Data Analysis Process

From looking at the data science process in section **Standards and Issues** of the module **Data Resources, Processes, Standards and Tools** we know the main tasks during the process are:

- **access:** getting parts of a full dataset needed for our analysis, sometimes included with wrangling.
- **wrangling:** transforming, cleaning and fusing data so that is more appropriate for analysis.
- **visualisation:** interpretation of both data and results can be useful at all stages; visualisation is one of the best methods of determining errors or missing data prior to analysis, and is one of the best methods of interpreting results.
- **statistical analysis**: this is the step of taking some data sets and building a model that is subsequently to be used in a delivery system.

A quick review of the standard internet blogs and question answering sites will give you no shortage of advice on which software systems you should learn "so you can get by".

- "Quora: What are some software and skills that every data scientist should know?" (http://www.quora.com/What-are-some-software-and-skills-that-every-data-scientist-should-know) has 29 answers, too many to read them all, but it is clear that languages such as R and Python are important, some special purpose tools are recommended, as are commercial tools (Tableau, RapidMiner, *etc.*) and internet services.
- "KDnuggets 15th Annual Analytics, Data Mining, Data Science Software Poll: RapidMiner Continues To Lead" (http://www.kdnuggets.com/2014/06/kdnuggets-annual-software-poll-rapidminer-continues-lead.html) gives an analysis of 3000 votes. This is an unscientific poll, and perhaps has a more data mining bent, but is nevertheless probably better than other informal polls.

Our own, admittedly biased and unscientific, summary of some tools is below. There are too many tools to properly cover them all, and many of the popular tools are multi-purpose. Many tools started out as university projects and still retain their open source nature but commercial vendors provide services or upgraded additions. There is also a rich variety of tools becoming available now in cloud platforms, often provided for free (though the use of their cloud is not), and a similar variety available in Software-as-a-Service mode or via APIs. There are many start-up companies in this area now, so the landscape is changing. There are also many *software solutions*, that is software targeted at specific application areas, and we explicitly exclude these, although they are very important in specific business areas.

- **access:** SQL, Hadoop, MS SQL Server, PIG, Spark, ...
- **wrangling:** common scripting languages (Python, Perl),
- **visualisation:** Tableau, Matlab, Javascript, ...
- **statistical analysis**: Weka, SAS, R, ...
- **multi-purpose:** Python, R, KNIME, RapidMiner
- **cloud-based:** Azure ML (Microsoft), AWS ML (Amazon), ...

Note one can also talk about delivery or operationalisation, when some of these tools can be applied, and in other times conversion to software in Java, C++ *etc.* may be done.

Note the Python versus R question is common one and really depends on what you want to do.

- "R vs Python for Data Science: The Winner is ..." (http://www.kdnuggets.com/2015/05/r-vs-python-data-science.html) (1500 words, 10mins) blog post on KDnuggets by Martijn Theuwissen of DataCamp, May 2015.

Scripting languages

An important part of the tools landscape is the use of scripting languages. The Wikipedia entry scripting languages (https://en.wikipedia.org/wiki/Scripting_language) describes these as follows:

> A **scripting language** or **script language** is a programming language that supports **scripts**, programs written for a special run-time environment that can interpret (rather than compile) and automate the execution of tasks that could alternatively be executed one-by-one by a human operator. Environments that can be automated through scripting include software applications, web pages within a web browser, the shells of operating systems (OS), and embedded systems. A scripting language can be viewed as a domain-specific language for a particular environment; in the case of scripting an application, this is also known as an **extension language**. Scripting languages are also sometimes referred to as very high-level programming languages, as they operate at a high level of abstraction, or as **control languages**, particularly for job control languages on mainframes.

The definition of a scripting language is a bit nebulous because many such languages have compiled versions (like Matlab) as well. Of the tools mentioned above, Python, R and Matlab fall into this category. Other well known languages used in this mode are Perl, Visual Basic, Lisp and even the operating systems languages like Bash.

To understand the role of scripting languages, you should review the Wikipedia section on types of scripting languages to understand the uses. StackExchange also tries to answer this question, but acknowledges the difficulty, stating the answers are "opinion based."

- "Types of scripting languages" (https://en.wikipedia.org/wiki/Scripting_language#Types_of_scripting_languages) on Wikipedia (1300 words, 7 mins) 
- "When is a language considered a scripting language?" (http://stackoverflow.com/questions/101055/when-is-a-language-considered-a-scripting-language) on StackExchange (this will lead you to an endless series of discussions and related questions, so added here for those wishing to see more).

The key thing one needs to understand is that a key function of the data scientist, in software engineering terms, is rapid prototyping (https://en.wikipedia.org/wiki/Software_prototyping). In this role, many of the standard processes of software engineering are not followed as closely, processes such as requirements analysis and testing. Data scientists need to rapidly play with their data, understand the issues, and then explore different possibilities. So high level languages are important, as are general purpose tools like Hadoop. These may triple the computing time over a custom designed system, but they can do so with far less human effort. Of course, with really big data, the trade-off between computing time and developer time starts to swing back towards more careful development.

In data science, we are doing a range of activities:

- putting together a processing pipeline,
- testing out different alternatives,
- trying "cheap hacks" for data cleaning to test ideas before investing more effort,
- glueing in custom software that might be difficult or particular to use (for instance, natural language or image processing tools or web services),
- running what is usually GUI systems in command line mode to support a processing pipeline,
- making changes to a pipeline to try out different ideas,

- restarting a half-run pipeline to alter the processing without wasting intermediate results.

For this range of roles, Python is probably considered the superior language because of its broad library support. In some areas, other languages are better. R has better statistical support. Javascript has impressive visualisation tools. Java has excellent natural language processing tools, and Matlab has excellent image processing tools. Old-timers like to use Perl because it is excellent for text munging.

5.5
Activity: Decision trees with BigML

Bank loan approvals using decision trees, built step by step from data

The data we are going to use is bank loan criteria, starting with the simplest case (therefore not realistic). The data format is as follows:
the first line is column or field title, under that is one record of data. So we have Age is 26 and Loan approval is No (sorry).

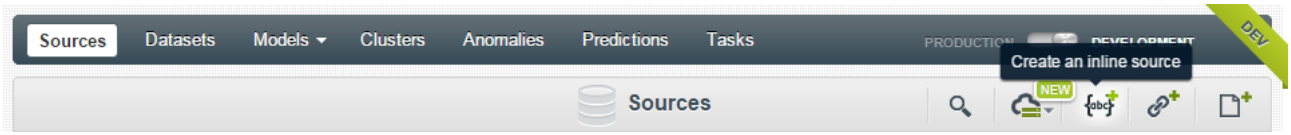
Age Loan
26 No

Step 1

Register to use BigML: https://bigml.com.au/

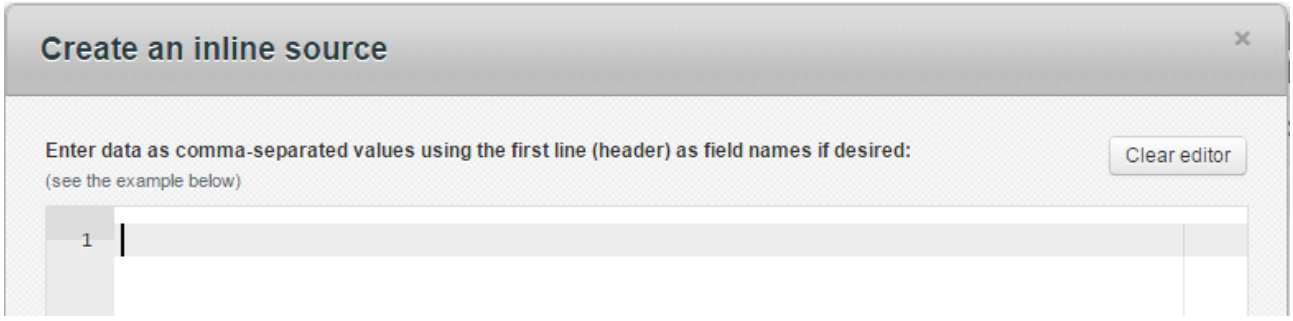
Step 2

To create a data source in BigML click on 'Sources' tab on LHS (Left Hand Side), then 'Create an inline source' (RHS):



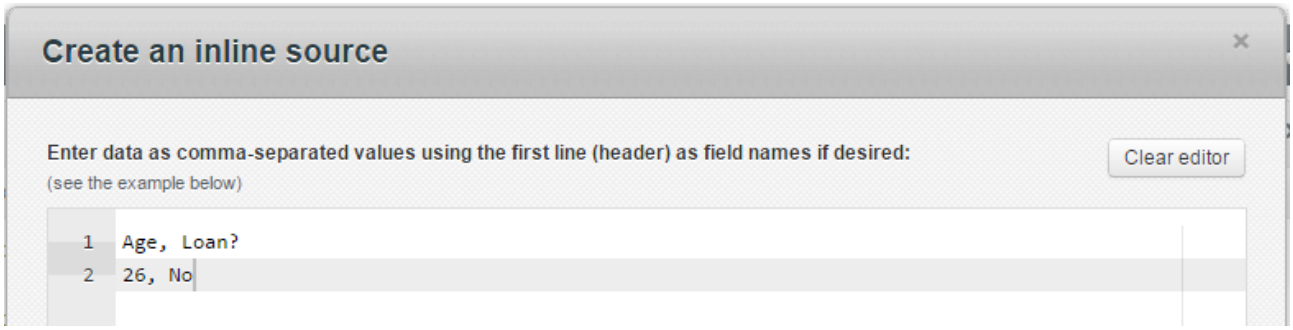
Step 3

Now 'Clear editor' so we can enter our own data.



Step 4

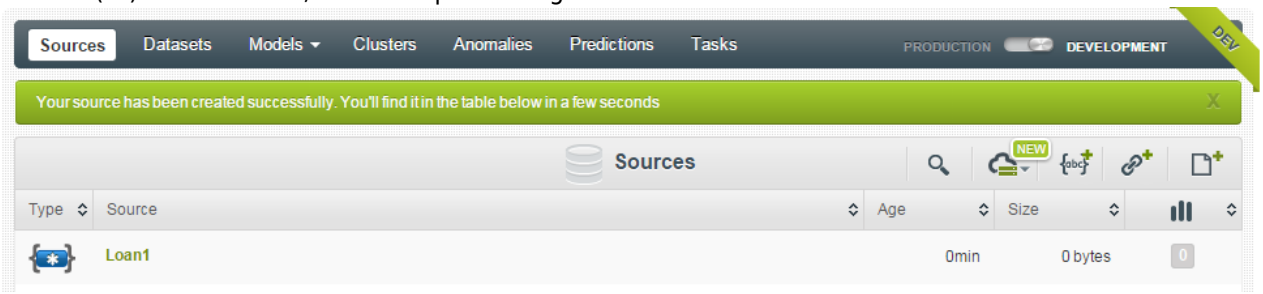
Paste or type the data from above:



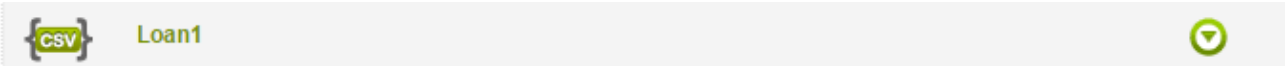
Step 5

Enter file name ('Loan1') and 'Create' (then wait a moment for it to process)

The star ('*') at LHS below, indicates processing...

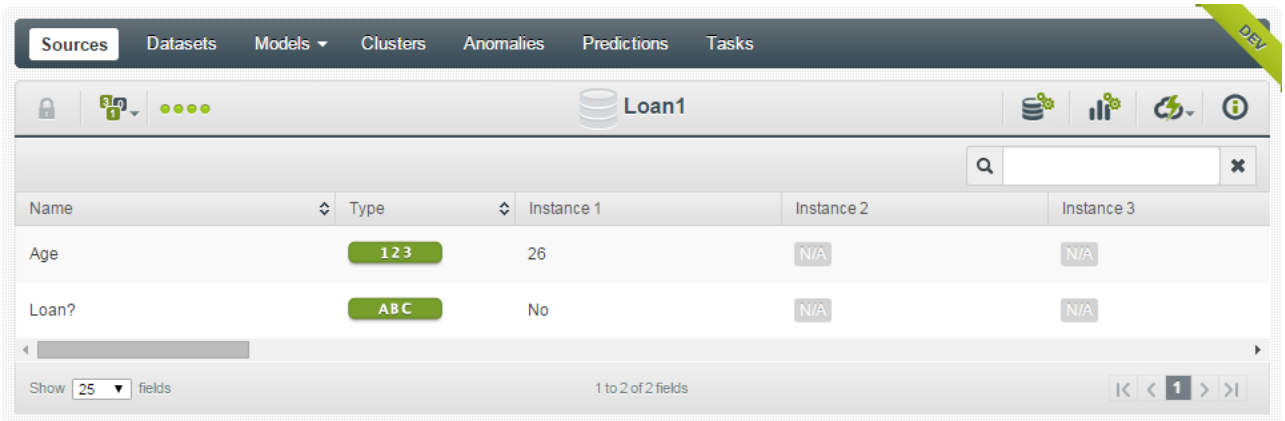


Now it's a .csv file and ready to use (as a data source).



Step 6

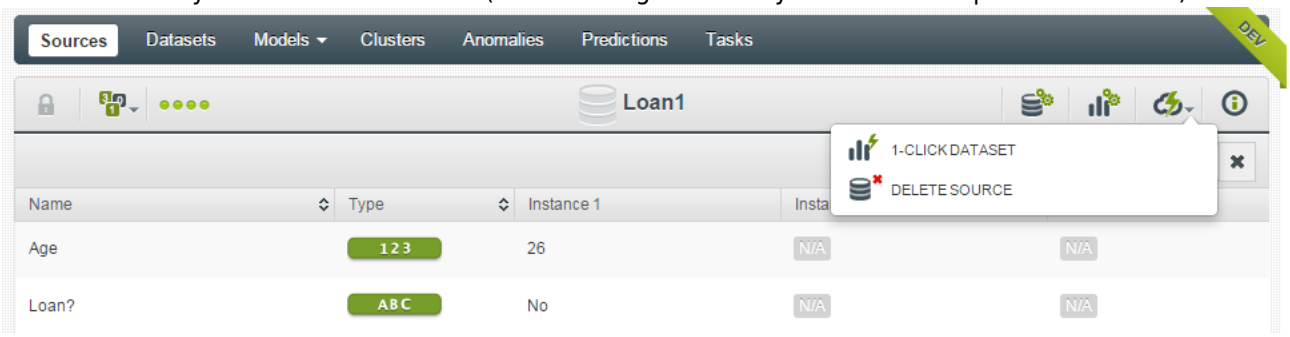
We need to make it into a Dataset we can use (essentially a 'copy'). First let's examine the data, click on Loan1.



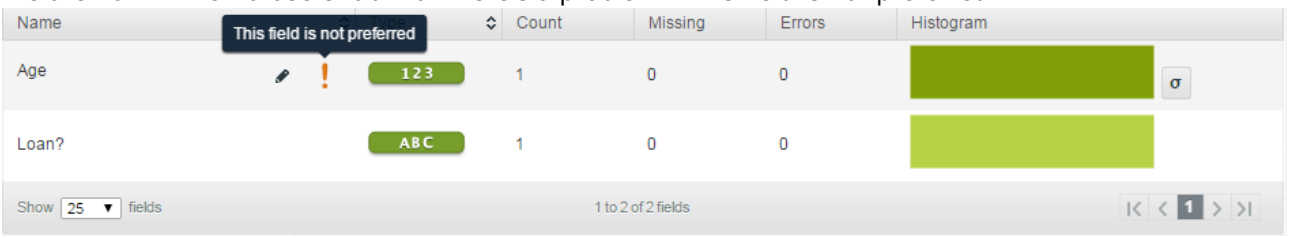
Age is 26, Loan? is 'No' - as entered. What rule can we infer from this data so far?
Data types have been determined by BigML. What are they? What other data types might you expect?

Step 7

On the RHS is a cloud/lightning icon, you can see two options (below), select '1-CLICKDATSET'. If you make mistakes you can 'Delete Source' (note that BigML allows you to create duplicate filenames).



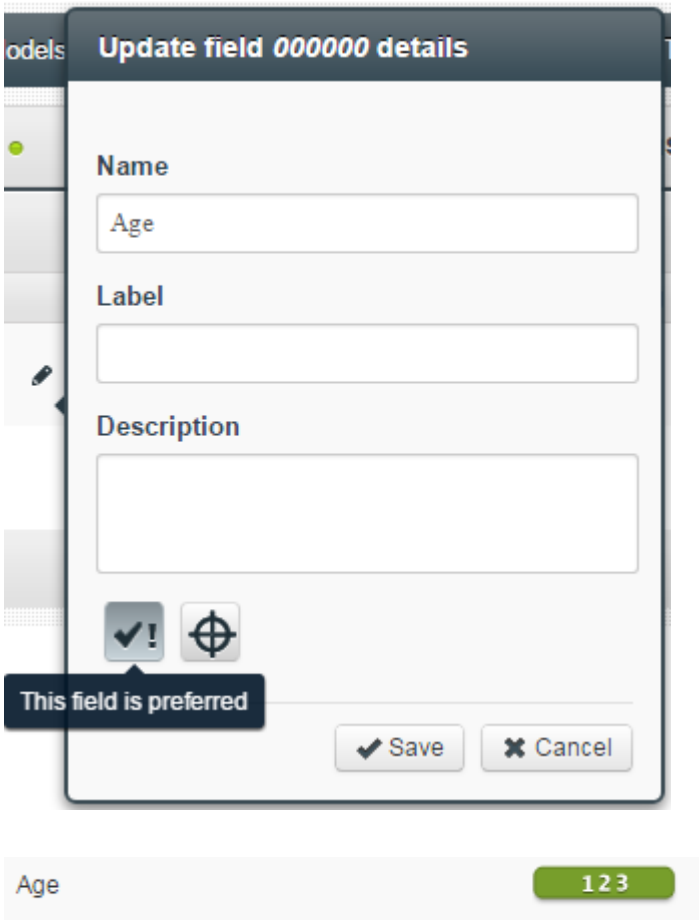
We are now in the Datasets tab. But there's a problem: "This field is not preferred"



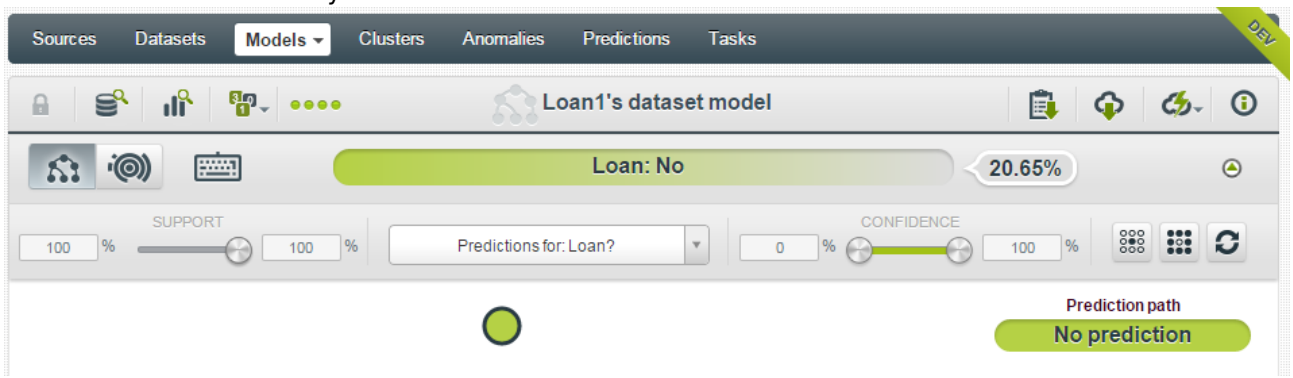
Research this warning, what does it mean?

Step 8

We can override this by editing Age (the small pencil icon to the left of the '**!**' above). This brings up the 'Update... ' dialog (below).
Click on to change the status to 'preferred' then 'Save'

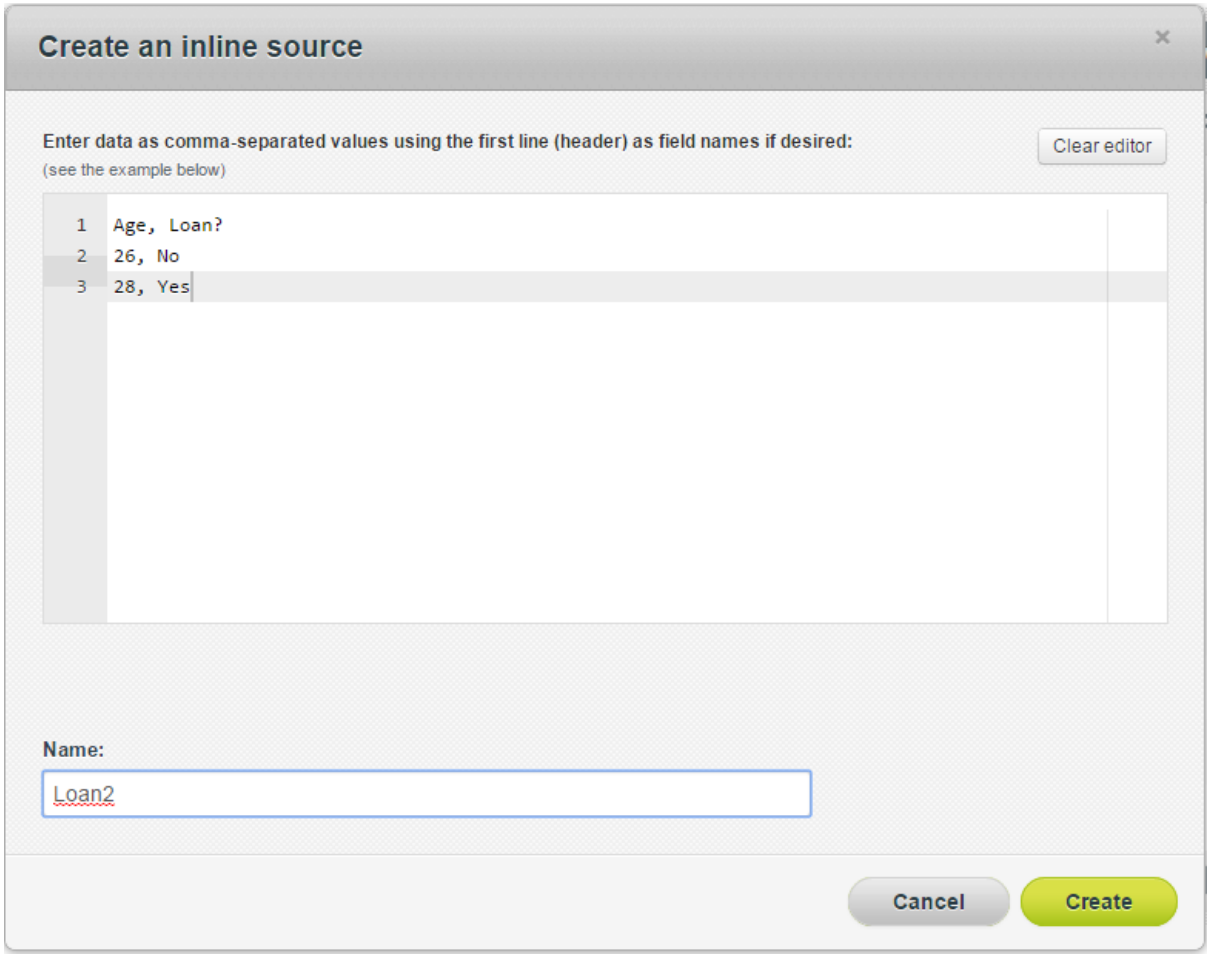


We could now generate a model (tree) with this data but there's not much point - you can't distinguish between data if there's only one record.

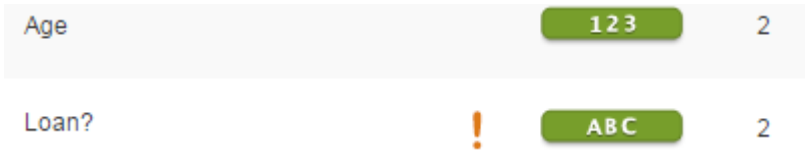


Step 9

Let's add more data. Go back to 'Sources' tab on LHS, then 'Create an inline source' (RHS). You will probably get your previous data, so you can add a line or make a new set. We want to add another record so pick an age (not 26, make it 28) and 'Yes' then name it 'Loan2' and 'Create':

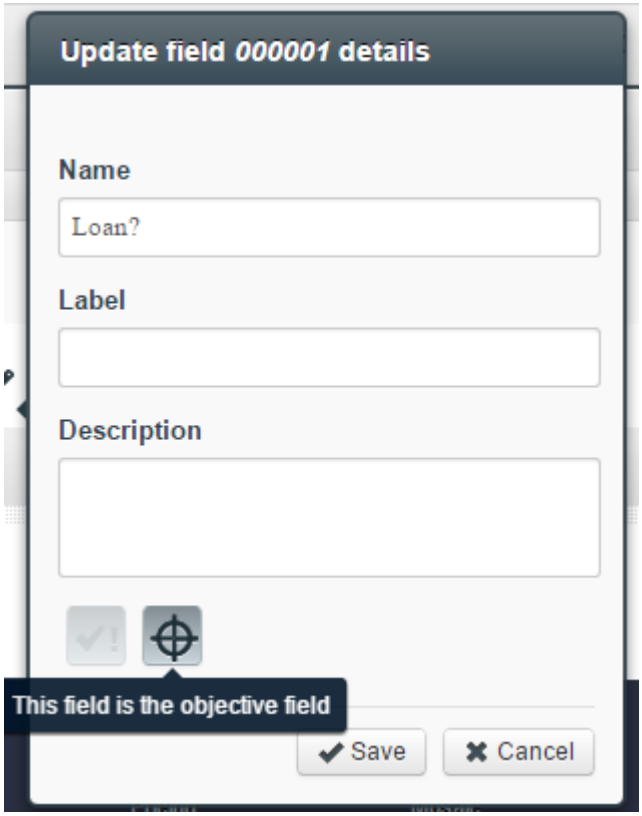


Wait for processing... Now we have two sources 'Loan1' & 'Loan2'.
As before, explore the data, then '1-CLICKDATSET'



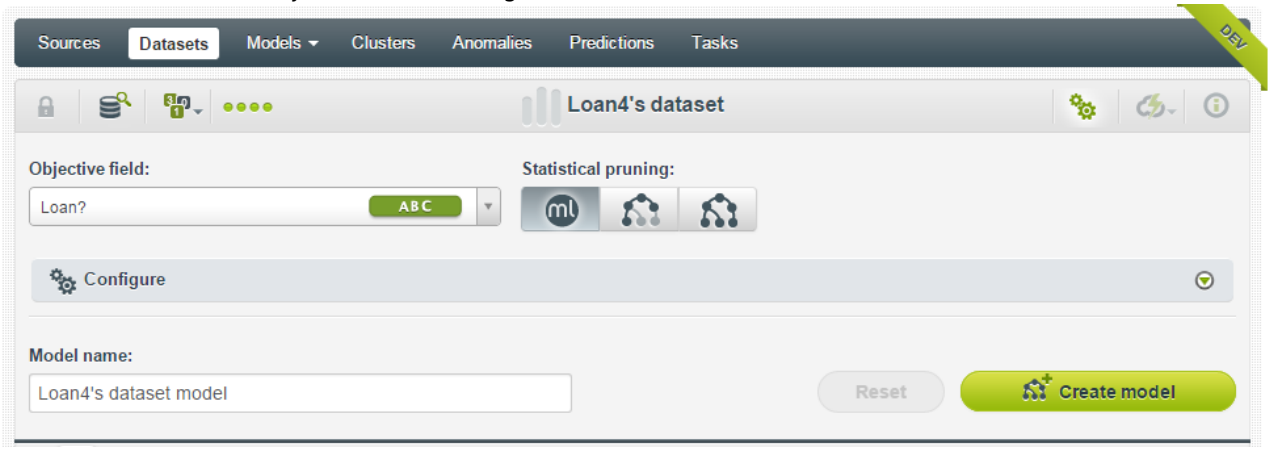
Step 10

'Loan?' is 'not preferred' and, because the last field in a dataset in BigML is usually the default goal (or **objective field**), we need to fix these attributes so edit 'Loan?' to make it both **preferred** and the **objective field** then **Save** (below).



Step 11

You can also verify which field is the 'Objective field' by using 'Configure Model' (top RHS, the **gears** icon). You can see that the 'Objective field' (the goal) is 'Loan?' (below).



Before creating a model for this data what is your prediction of any rules from this data?

Step 12

Now use **Create model**:

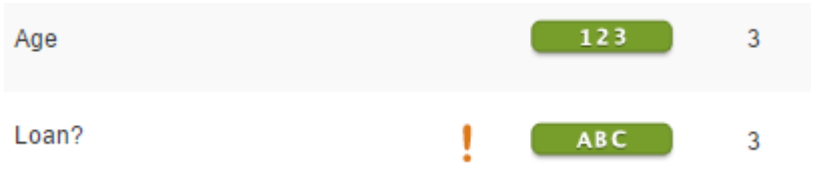
So we can see that there are two tree branches (or paths), based on age, one leads to the 'No' leaf, the other to 'Yes' leaf (for ages 26 and 28 respectively).

Why is the rule <= 27?

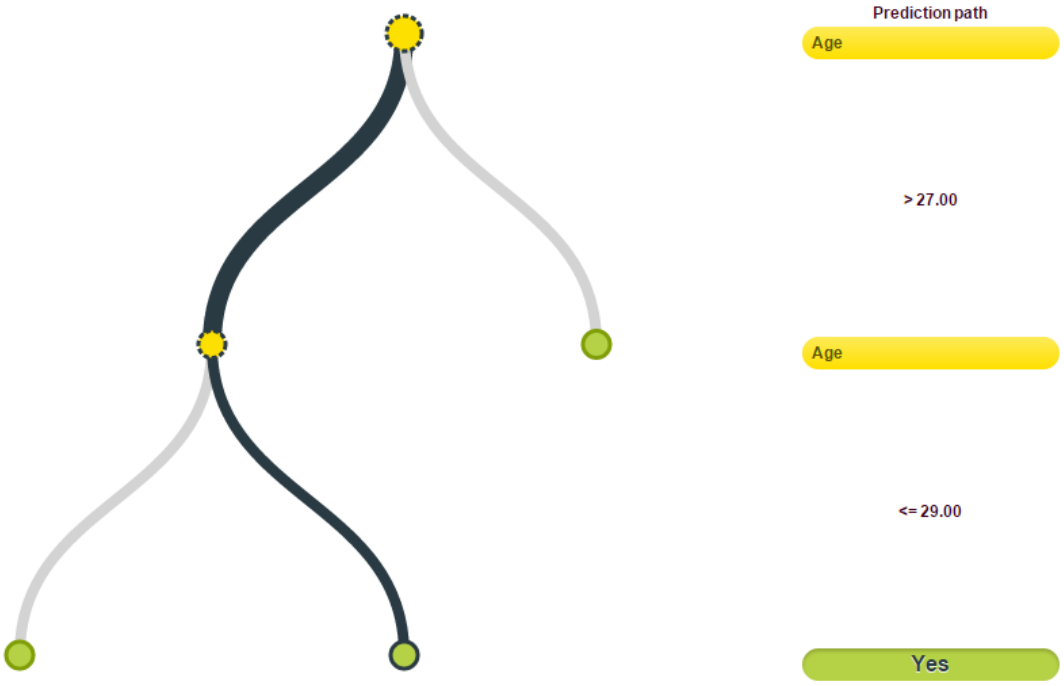
As before, there's not much data here so let's add some, we have:
Age, Loan?
26, No
28, Yes
Given that we assume there is a relationship between age and loan approval, what would happen if we add some non intuitive data, for example (30, No)?

Step 13

As before, create a dataset and edit 'Loan?' to make it **preferred** and **objective.**



Now generate a model:



This makes no sense, > 27 is a rule or boundary (Yes to a loan), therefore <=27 (is No), but <= 29 is another rule (Yes), but > 29 is No! Garbage In Garbage Out (GIGO). This is no way to run a bank. Let's try some other criteria, like income.

Step 14

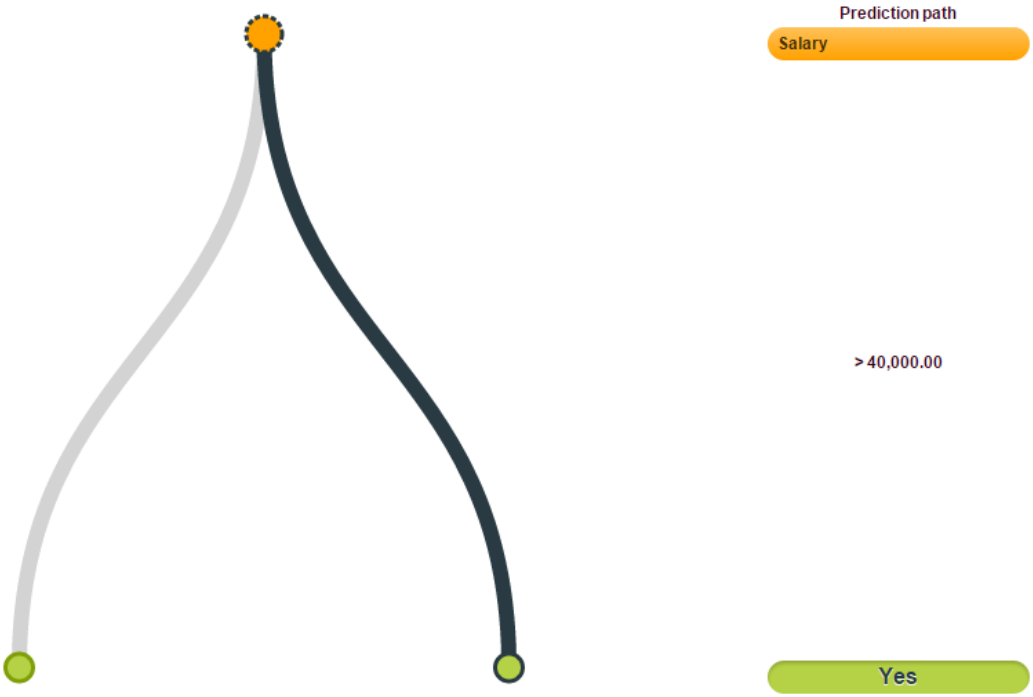
Create a new data source ('Loan4'), and then a new dataset:

Age, Salary, Loan?
26, 35000, No
28, 45000, Yes
30, 30000, No
40, 50000, Yes

We no longer have any 'not preferred' fields, why do you think?

Name	Type	Count
Age	123	4
Salary	123	4
Loan?	ABC	4

Before generating a model, what do you think the prediction and boundary values be?



So the prediction path is salary and the boundary is $40,000.

At this point we can stop, add more data, add more columns, add both, or switch to a 'real' data set, e.g. German bank data (via Penn State) https://onlinecourses.science.psu.edu/stat857/node/222

5.6
Activity: Prediction with BigML

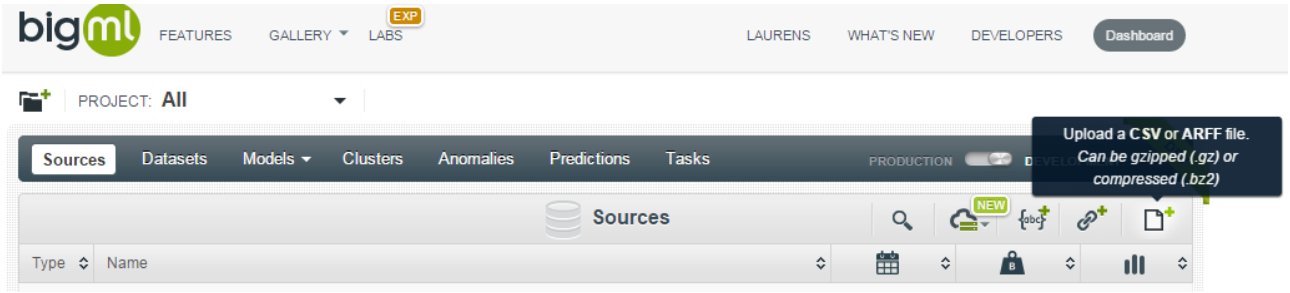
Real Estate Price Prediction using BigML

So now that you have your loan, time to look at houses, this is an extension of the 'Loans do grow on Trees' exercise in BigML. In that exercise the dataset was minimal, for the purpose of introducing BigML, datasources, datasets, datamodels and prediction or decision trees. This dataset is more about quantity, nearly 5,000 houses, but it's still a limited set in the sense that there are only a few attributes:

Bedrooms	Bathrooms	Type	HouseSize (sqft)	LandSize (sqft)	Price
3	2	Single Family Home	1992	6098	220000

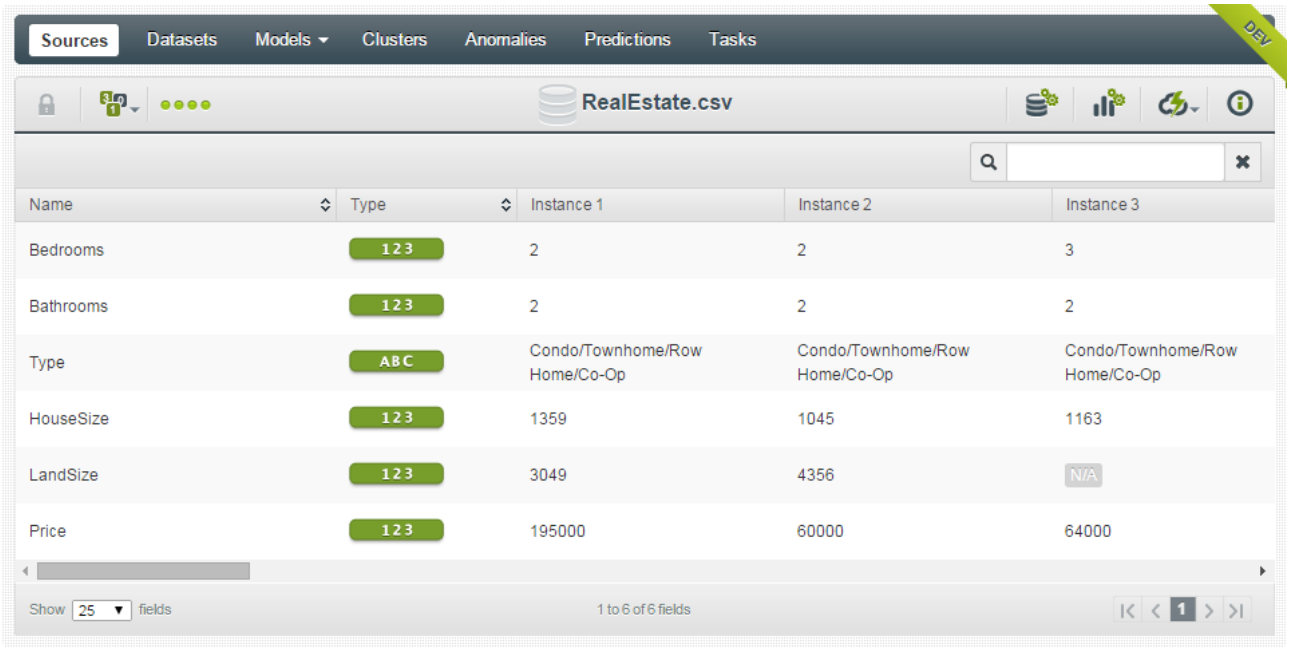
Step 1

Load *RealEstate.csv* *(https://www.alexandriarepository.org/wp-content/uploads/20150629120123/RealEstate.csv)* into BigML using the far right option 'Upload a CSV...'



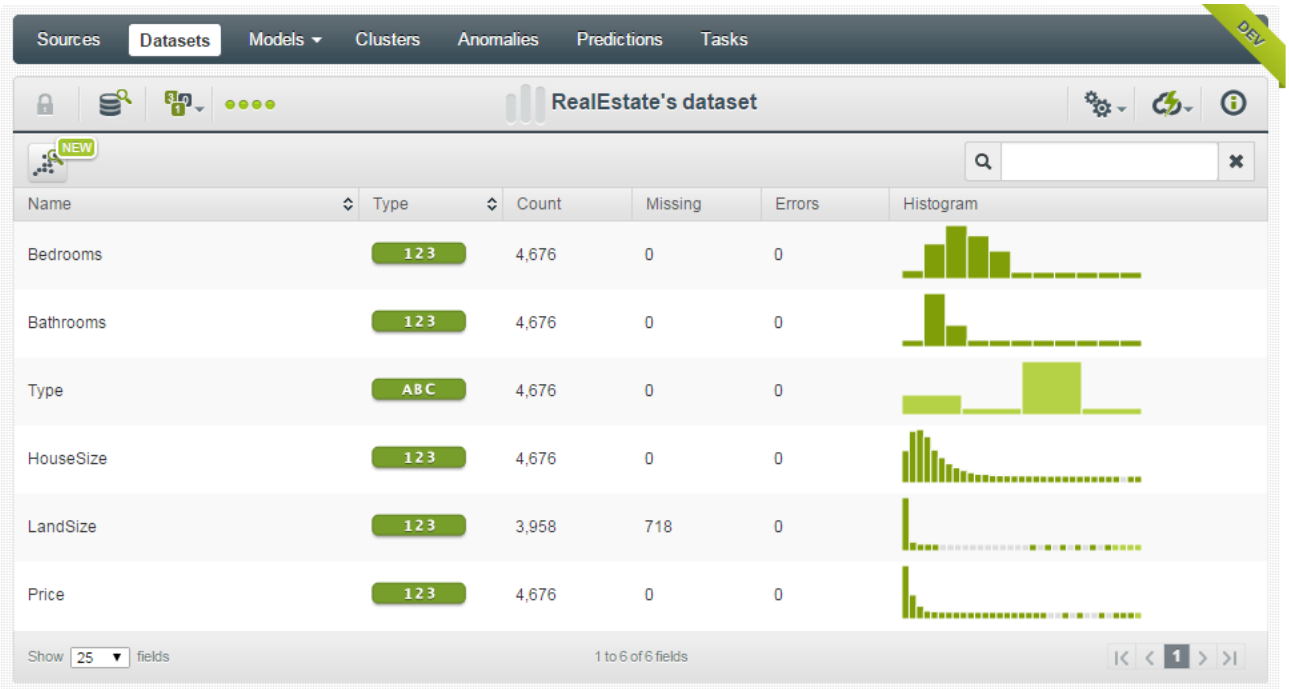
Step 2

Open to explore, which attribute is the likely prediction goal here?



Step 3

Create a dataset (under cloud/lightning icon):



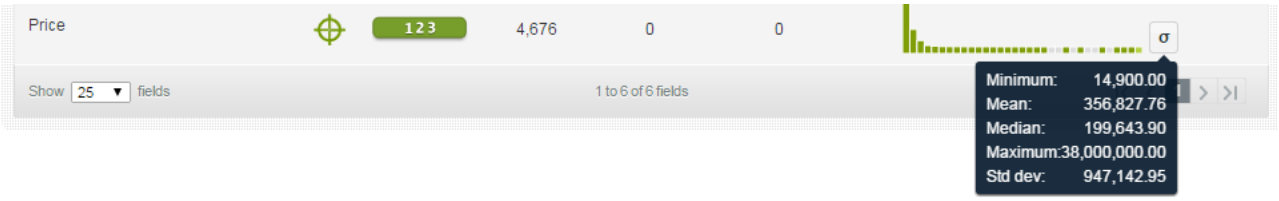
Step 4

Check data, especially the **Objective goal** by using **Configure Model.**

Or you can see the 'crosshairs' (i.e. target) icon:

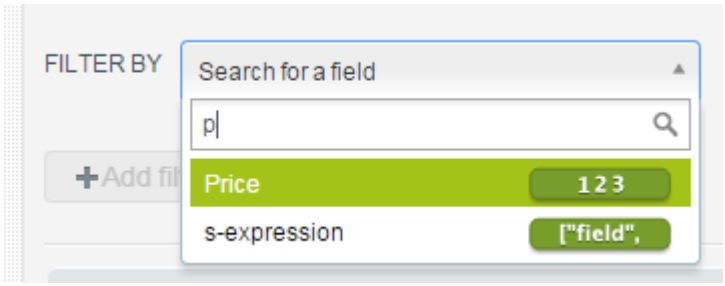
Step 5

Now we are going to filter the data, you can see that there are some extreme prices below ($38,000,000! Also $14,900, probably don't want that house...)



Step 6

Find the **Filter Dataset** option to remove houses outside of budget, select (or find) 'Price'.



Step 7

Change the maximum and minimum to something more sensible like 100,000 to 1,000,000.

Step 8

Now **Create** dataset. How many records remain?

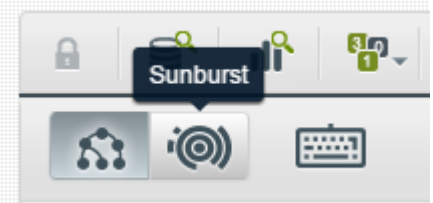
Step 9

Create a predictive model, use '1-CLICK MODEL', by default it's a tree.

Which house attribute is the top level branch based on?

Step 10

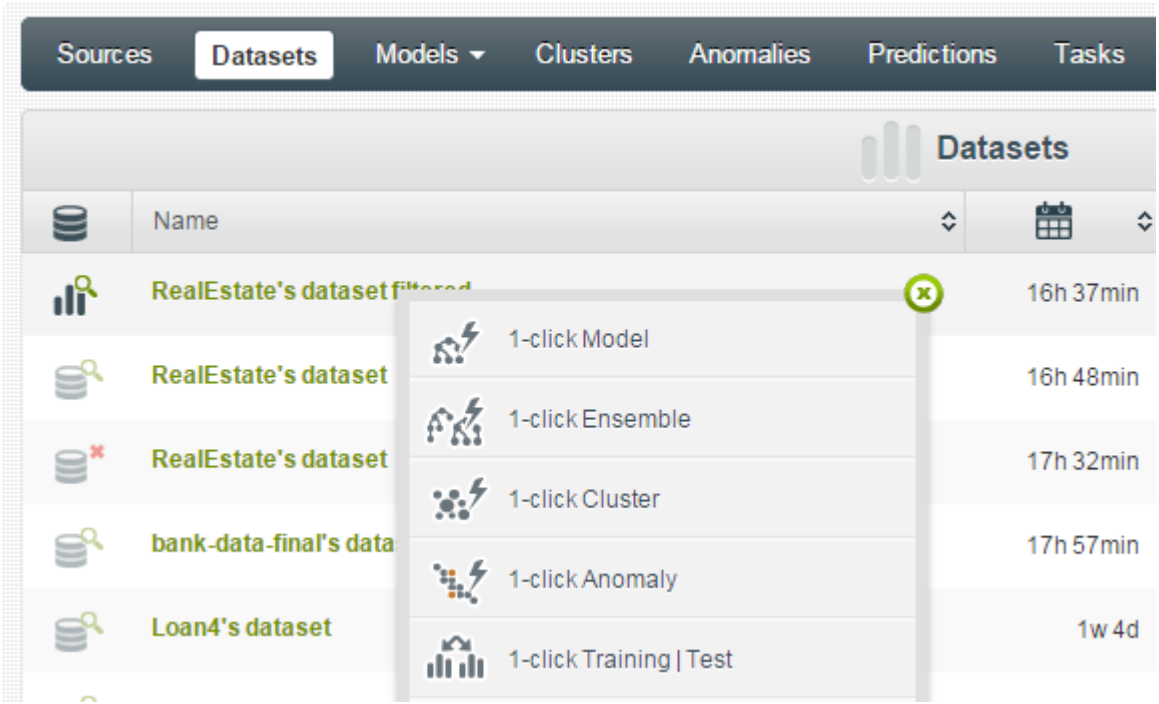
Or you can try **Sunburst** which is a centre out (rather than a top down) view.



Step 11

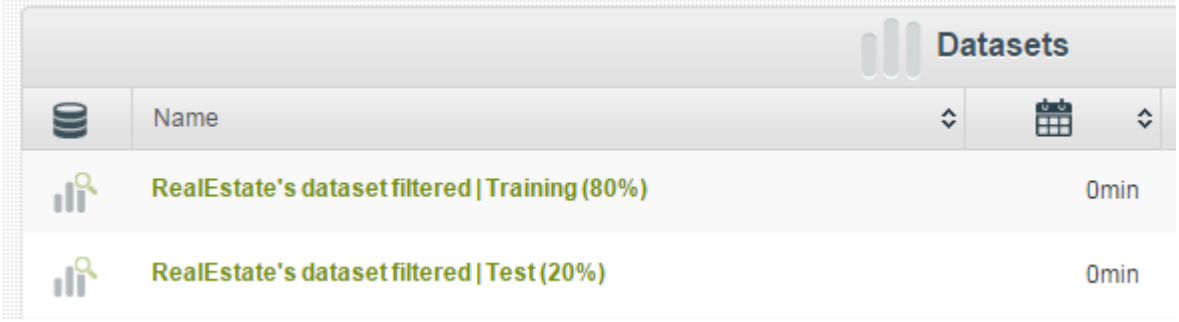
Evaluating the model, now we are going to split the dataset into 'training' and 'testing' subsets. Usually 80% and 20% of the data respectively. Return to the Datasets tab, select our filtered set, in the drop down menu look for

'1-click Training | Test'



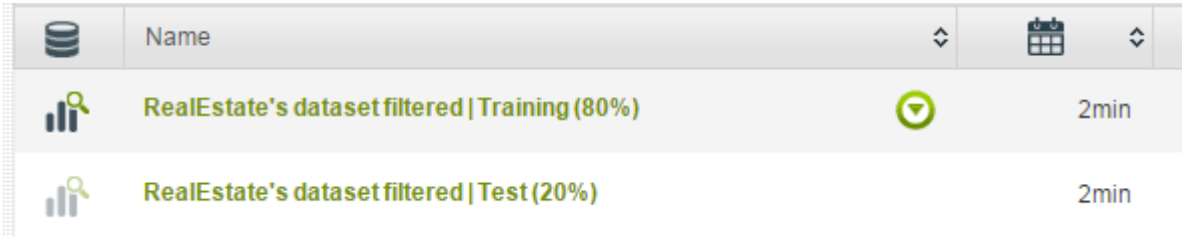
Step 12

And we now have the newly created 80/20 sets. How should the sets be divided, first 80%, last 80%, randomly, what could happen if the data is sorted?



Step 13

Make sure to use the 80% set for this next step.

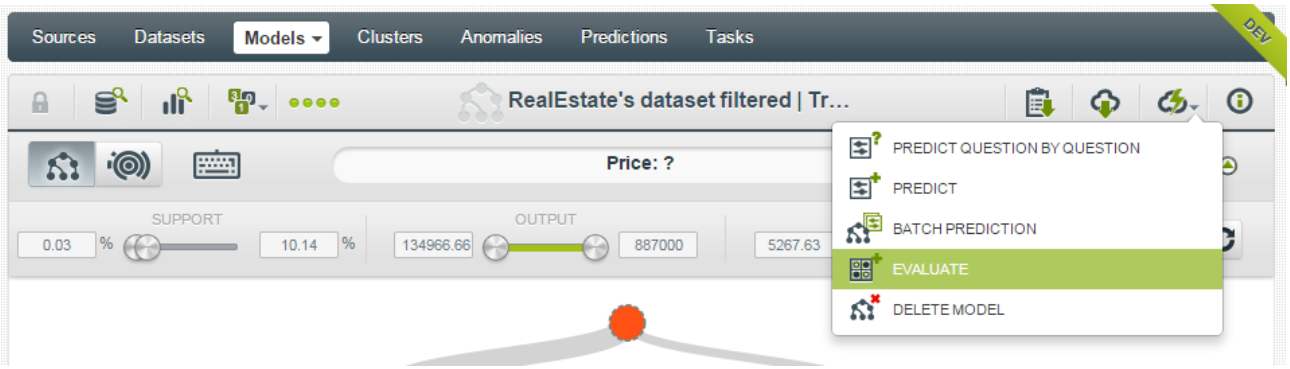


Step 14

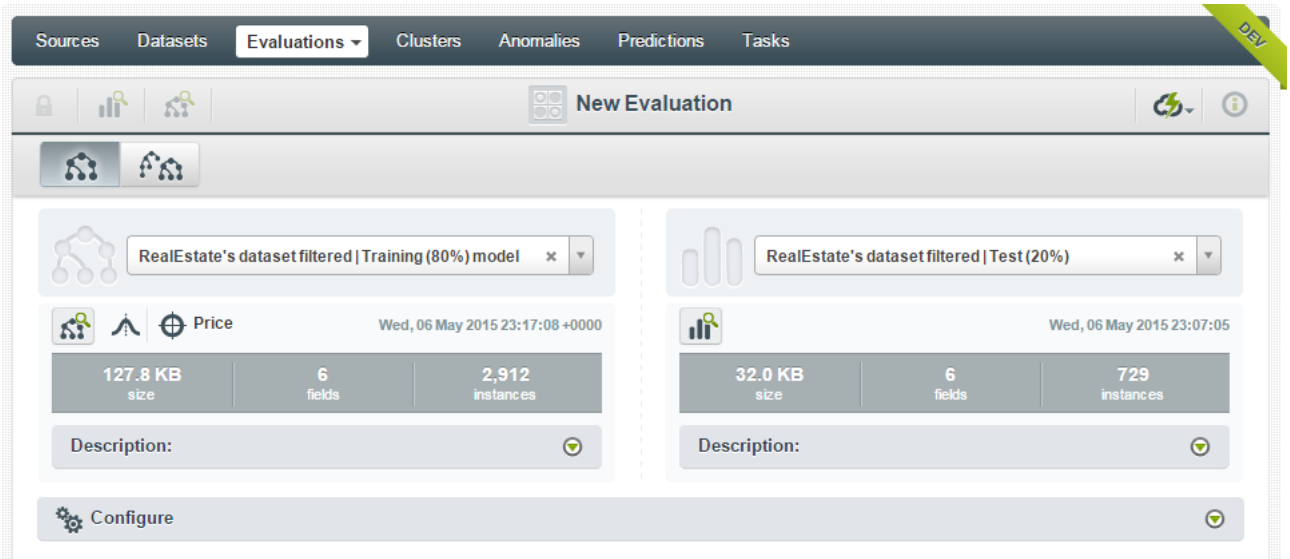
In the drop down menu find '1-click Model'.

Step 15

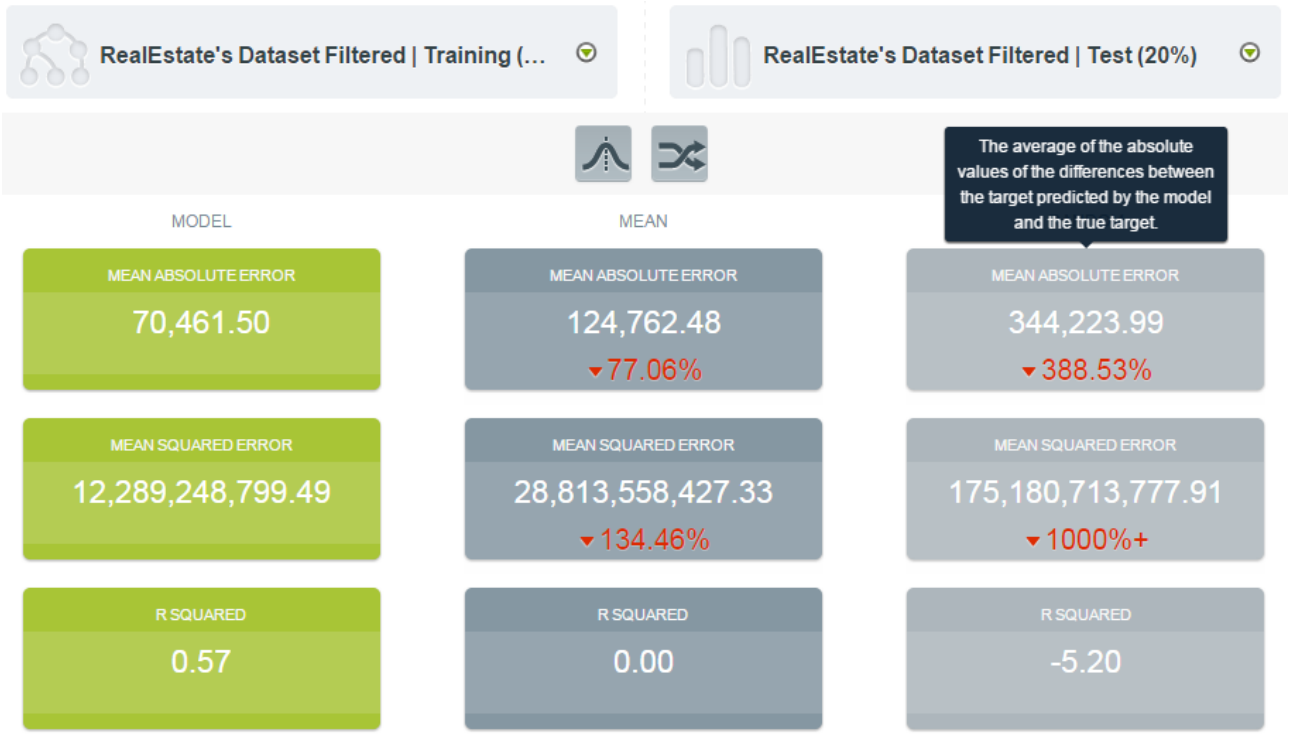
Now we are going to Evaluate. Essentially BigML is going to test itself against the remaining 20% set to see how accurate the 80% model is. What % accuracy would be acceptable here 50%, 80%, 90%, is 100% realistic?



Look for the **EVALUATE** option:



Training set is at left, test set at right: 

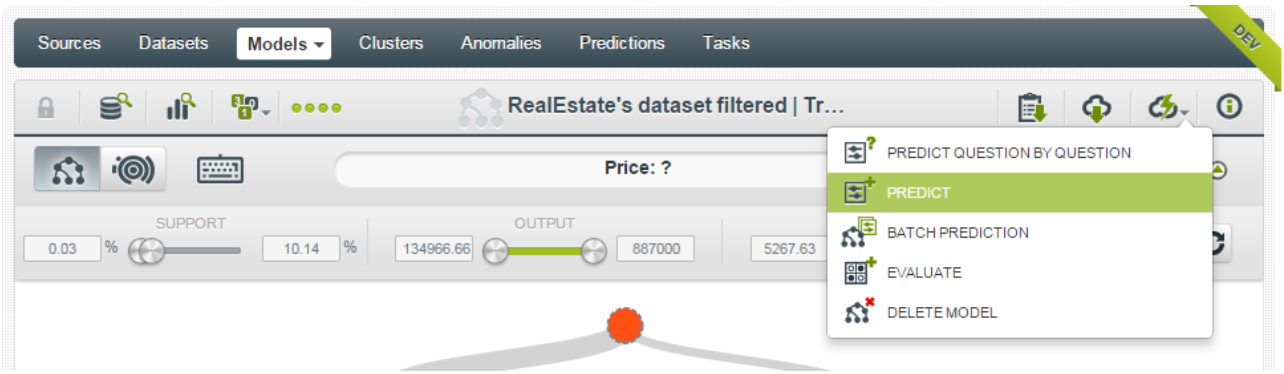


Step 16

Finally, you can now test out the model with new data i.e. a house you are interested in buying (or selling). Go back to the **Filtered Model** then to the **PREDICT** option.

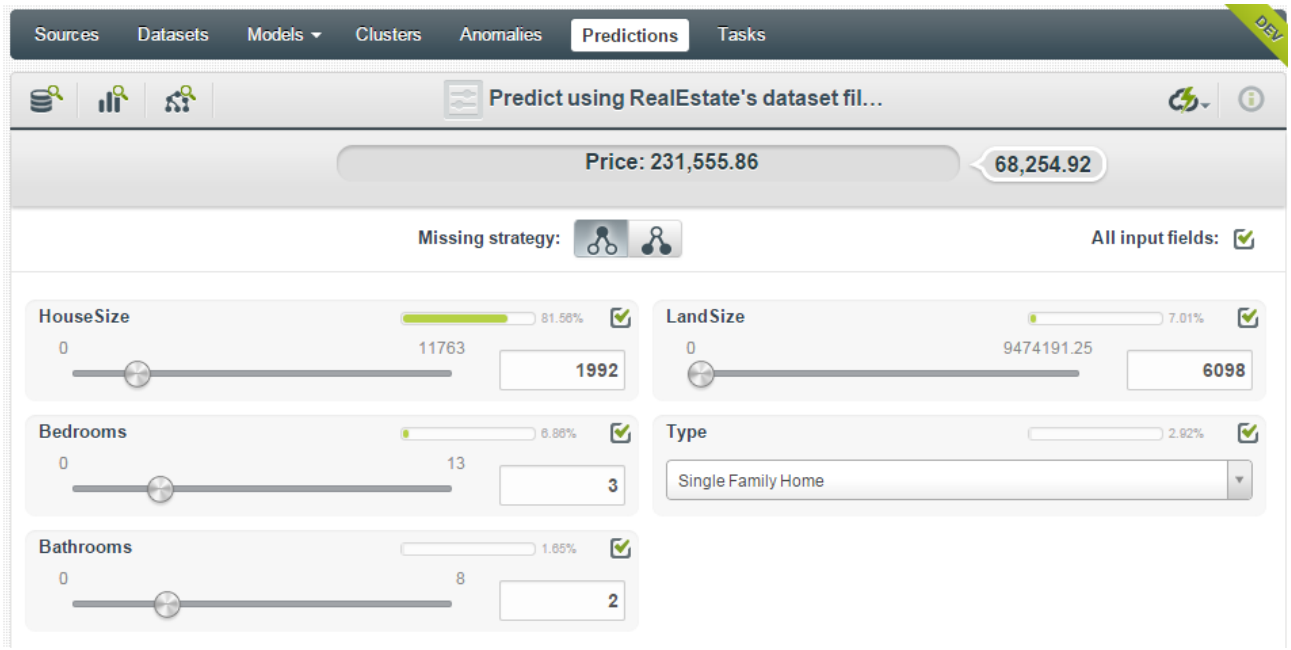
Step 17

If you try a prediction using existing data (i.e. copy a line from the csv or dataset), do you think BigML will get the price exactly right, within a certain % (i.e. range)? Predict the prediction for the following: 3, 2, Single Family Home, 1992, 6098:



Step 18

The result is shown below:



Price 231,555.86

Step 19

The data is from page 1, Price is $220,000:

Bedrooms	Bathrooms	Type	HouseSize (sqft)	LandSize (sqft)	Price
3	2	Single Family Home	1992	6098	220000

So the prediction is fairly accurate but what is the $68,254.92 expected error?
Is this accurate enough? How can you improve accuracy?

5.7
Data Analysis Case Studies

This section presents a number of different case studies that are useful lessons for us because they include more critical or in depth discussion of the issues. Interestingly, they are around a common theme, medicine and health. This was not intended, but it perhaps reflects where the most published critical analysis of data science related activities. Moreover, visualisation is not discussed. Visualisation is often used to get initial insights, and to properly prepare the data, but final delivery of a predictive system does not require it. So in the final summaries, visualisation may have done a lot of the work, and may also get the media attention, but the final delivered system uses a (statistical) predictive analysis which may *claim* (rightly or wrongly) all the credit.

Google flu study

In the section **Introduction to Resources and Standards** in module **Data Resources, Processes, Standards and Tools** we looked at web-scale pharmacovigilance, which was about using web query logs to investigate drug interactions. This idea, of using web-logs to support health, was really initiated with Google Flu Trends (GFT), a web-service based on query logs to indicate flu activity.

- "Google Flu Trends Overview" (https://www.youtube.com/watch?v=6111nS66Dpk), video on Youtube (2 mins)


- "Google Flu Trends: How does this work?" (http://www.google.org/flutrends/about/how.html) on Google (700 words, 4 mins) - Google has suspended this application in mid 2015, so the link is now a historical pointer to old data!
- they have a full article in Nature, too long and detailed for our purposes, but worth a quick scan if you are interested in the science angle "Detecting influenza epidemics using search engine query data" (http://static.googleusercontent.com/media/research.google.com/en//archive/papers/detecting-influenza-epidemics.pdf) (PDF, 7000 words)

For reference there is also a critical report in the journal *Science* March 2014, "The Parable of Google Flu: Traps in Big Data Analysis" (http://dash.harvard.edu/handle/1/12016836) (PDF, 2500 words, 15 mins). This report is based on the observation the GFT had some serious systematic errors in 2009 and 2013 when compared with the CDC's reported counts, the so-called "ground truth". They make the following points:

- Google's search engine and its query logs are constantly evolving system so the stability of the signals Google is using is unclear.
- Google apparently did not use standard time series models. Time series models are a standard technique used in areas like economics and finance to predict current indicators from their past values (days or weeks earlier). For instance one could use last week's "flu counts" to help predict this week's. Google could have augmented their query log signals with CDC's historical count data and made their predictions more robust.
- Because of the proprietary nature of Google's search engine's internals, full disclosure by them of the details of their data (many words were used in their signals) was not done. This makes the usual process of scientific reproducibility very difficult because others cannot work with their data to explore the results.

In the section **Introduction to Resources and Standards** in module **Data Resources, Processes, Standards and Tools** we looked at Horvitz's use of (Microsoft) query logs for pharmacovigilance. In their

case the signal is really differential, contrasting the combined individual search counts. Like Google, they did an evaluation against ground truth, but as yet we see no serious systematic errors in the pharmacovigilance study. Will this situation hold? How robust is this kind of proxy data (query logs)?

The Science article authors talk about "big data", the query logs, versus "small data", the CDC's counts. We know the real issue is that the query logs are proxy data, with veracity issues, whereas the CDC's data is curated data of higher veracity based on laboratory results. Depending on the quality of signal you need, proxy data can be adequate, but it does need to be calibrated. Another interesting question is: when does the proxy data go wrong and what influences it? For instance, maybe Google's query log proxy is most wrong when the profile of the flu is raised in the media due to reports about potential flu-related epidemics such as the swine-flu. If this is the case, then unfortunately it is during these times when the estimates are most needed, so accuracy becomes more important.

Eric Horvitz on data analysis for decision making

Here we continue with more extracts from Horvitz's KDD2014 talk.

- "Data, Predictions, and Decisions in Support of People and Society," (http://videolectures.net/kdd2014_horvitz_people_society) Eric Horvitz of Microsoft giving a 40 minute wide-ranging technical talk on data science with theme of "social good". In this case view the following extracts:
  - Clinical Medicine and Readmissions: 17:21-21.25  (do not follow after this, it gets more technical)
  - Errors, Adverse Events and Deaths + Hospital-Associated Infection: 28:05-36:50 

The first segment discusses the readmissions challenge. The first thing to notice is that the dataset discussed is not a single dataset, but rather a fusion of data from many different parts of the hospital. Not all hospitals fuse this kind of data as there is not a great incentive, in billing terms or for individual departments in the hospital, to do so. This 4 minute segment also talks about the delivery issues of data Science.

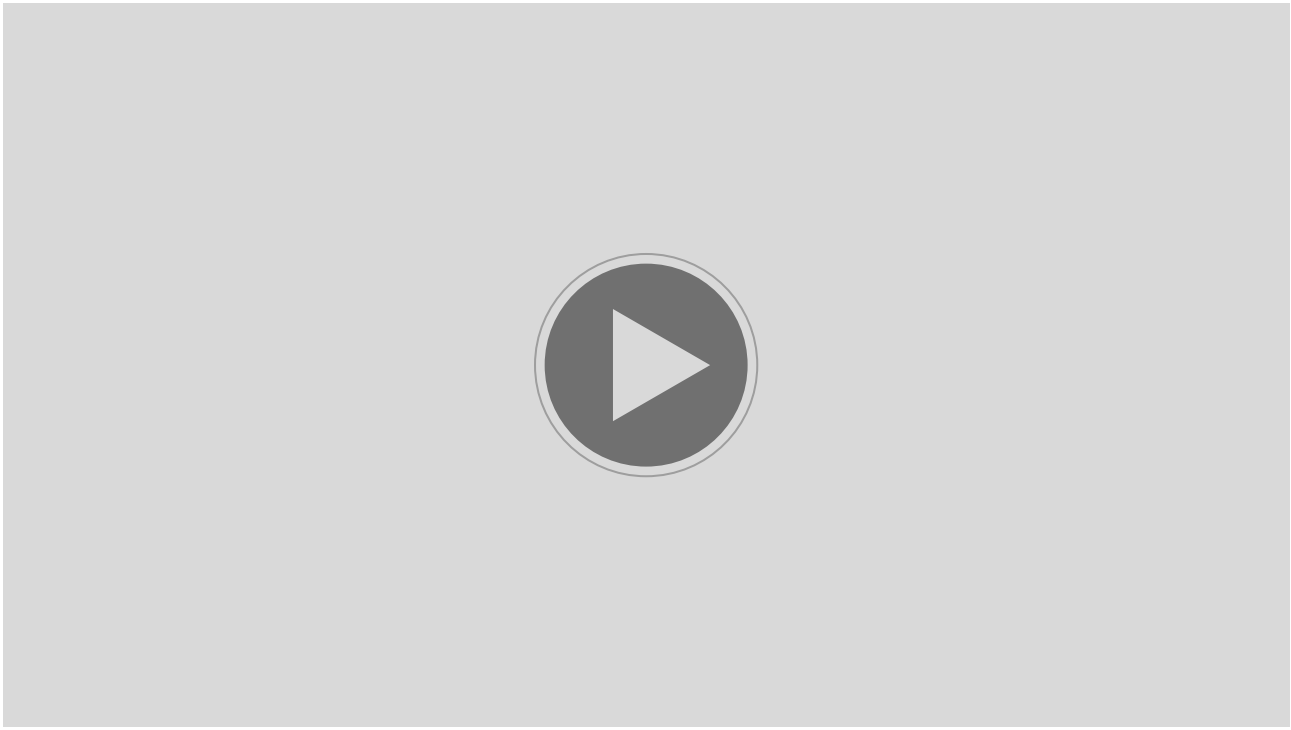
The second segment firstly discusses the prediction of adverse events, *i.e.,* rare anomalies. Anomalies are what the medical expert does not expect, and they are often times bad, so it is good to predict them. Here, a custom dataset was selected to support the prediction task, and therein lies the art.

The second segment secondly discusses predicting hospital-associated infection, an effect that happens over time, during their pathway through the hospital. This is another interesting application where time is a factor and appropriate selection of data is important.

Interviews with an industry professional

We continue with an interview with an industry professional here in Australia.

Watch Associate Professor Chris Bain (Director of information services - The Alfred Hospital) discuss the various aspects and issues of **analysing data:**



(https://www.alexandriarepository.org/wp-content/uploads/20150629093441/FIT5145_module_5_analysing_data_combined.mp4.mp4)
Alternatively, you can download the transcript for Analysing data
(https://www.alexandriarepository.org/wp-content/uploads/20150701100114/transcript_FIT5145_module_5_analysing_data.pdf).

Scientific method

Medical and health research are getting criticism in both the mainstream media and the medical journals. This is now widely disseminated so that industry, research and the public are aware of it. While some may argue this is an internal battle that does not concern us data scientists, it is also an opportunity for us to learn. Medical and health research has some parallels to Data Science: data is collected, analysis is done, results are presented. Clearly, we have more discovery, and the kinds of data we look at differ, but there are similarities. So what is it they are doing wrong that should concern data scientists?

First, here are some of the articles that discuss this, and there are many more. Now the relevant parts of these are discussed below, so these are provided here for reference only and not for study.

- "How science goes wrong" (http://www.economist.com/news/leaders/21588069-scientific-research-has-changed-world-now-it-needs-change-itself-how-science-goes-wrong) on The Economist (article, 1000 words, 5 mins)
- "Battling Bad Science" (http://www.ted.com/talks/ben_goldacre_battling_bad_science) a TED talk by Ben Goldacre (TED, 14 mins, 3000 words) 
- "The Truth Wears Off" (http://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off) by Jonah Lehrer in *The New Yorker* (webpage, 5000 words, 25 mins)
- "Richard Smith: Time for science to be about truth rather than careers" (http://blogs.bmj.com/bmj/2013/09/09/richard-smith-time-for-science-to-be-about-truth-rather-than-careers/) blog on BMJ (blog, 950 words, 5 mins)
- "Offline: What is medicine's 5 sigma?" (http://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736%2815%2960696-1.pdf) by Richard Horton on The Lancet (blog, 700 words, 4 mins)
- "The 10 stuff-ups we all make when interpreting research," (https://theconversation.com/the-10-stuff-ups-we-all-make-when-interpreting-research-30816) by Will Grant and Rod

Lamberts on The Conversation (blog, 1300 words, 7 mins) , while written about scientific research, often applies to Data Science as well. 

Errors fall into a number of categories. Some are to do with industry funding and coercion, some are to do with the academic publishing game, but some are general science errors and these latter ones we will focus on.

Significance testing

The most often discussed issue is flaws with [significance testing](https://en.wikipedia.org/wiki/Statistical_significance). The effect is remarkably simple to understand, but the effect has confounding subsequent effects. Areas such as medicine, health and psychology generally have to do expensive collection to get data so they are often data poor and in need of sensitive statistics to obtain results. A typical statistical test computes a so-called *P value*, which you will learn about elsewhere, but the logic is as follows:

1. You want to test a hypothesis, say "a daily aspirin reduces heart attacks in older men". This is your *hypothesis*.
2. You line up 100 subjects, older men, give 50 of them daily aspirin for 5 years and the other 50 an identical looking placebo. This is your *experimental design*.
3. At the end of the 5 years you count the number that had heart attacks for those taking aspirin HA with those taking the placebo HP. These are your *experimental data*.
4. Now its time for the statistical computation. The statistician takes these number and asks the question:
   Let us assume there is no difference between a daily aspirin and a placebo. Then how surprised would you be in the counts (HA,HP)? This measure of surprise is called a *P value*.
5. The computed P value is measured in units of chance. We will represent them in the form "one in a hundred" represented as 1:100. If the P value is more than 1:100, say 1:200, you the scientist get to publish a paper saying:
   Our study confirms a daily aspirin reduces heart attacks in older men.
   If the P value is less than 1:100, say 1:20, then you could publish a paper saying:
   Our study shows no significant evidence that a daily aspirin reduces heart attacks in older men.
   Alternatively, you may not publish anything at all, since science does not like negative statements.

The *P value* refered to above computes the value of surprise. We will not do the math/stats here but let us look at examples:

- If HP=4, HA=3; this means roughly 4/50 older men had a heart attack in 5 years and for those taking aspirin it is very close, the "surprise" should be small since by your assumption they should be similar.
- If HP=10, HA=1; this means roughly 10/50 older men had a heart attack in 5 years and for those taking aspirin it is considerably less, the "surprise" should be large since by your assumption they should be similar.

So having described P values, we can now describe the issues. In areas such as medicine, health and psychology, where data is hard to get, scientists use smaller data samples and they make their cut-off values quite small, like 1:20 or 1:100. So now let us consider what happens at the larger scale, with teams of scientists and we will discuss the various effects reported in the literature:

Significance chasing: there are two versions of this flaw in science practice:

- Suppose at one large laboratory we have a team of 25 scientists, each doing about 4 experiments

a year. Then the chances are that one of these 4×25 experiments will beat the odds and produce a significant result at 1:100 even though it should not be significant.
- Suppose one scientist does an experiment with 100 different treatments in parallel. Then the chances are that one of these 100 experiments will beat the odds and produce a significant result at 1:100 even though it should not be significant.

This effect was discussed above by Horton's *The Lancet* article, by Lehrer's *The New Yorker* article. It is also extensively discussed in a recent articles, which are provided for reference only, not to study.

- "Scientific method: Statistical errors" (http://www.nature.com/news/scientific-method-statistical-errors-1.14700) in *Nature*, February 2014, P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.
- "I Fooled Millions Into Thinking Chocolate Helps Weight Loss. Here's How." (http://io9.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800) by John Bohannon, an entertaining report of significance chasing, explaining in full detail the second variant of the problem (3000 words).

Horton mentions an area of physics where the scientists use P values of 1:3,000,000. This could not generally be done in areas such as medicine, health and psychology because of the difficulty of getting enough data to achieve that level. While we talk about big data, we do not always have big data, especially, for instance, with rare diseases or where patients/subjects need to be actively recruited.

Ignoring negative results: your colleague has recently published an article saying "a daily aspirin reduces heart attacks in older men" based on such a study. A few years later you do a related follow-up but get remarkably different results. Your results show no such thing. What happens? Well, its a negative result and harder to publish, as mentioned in *The Economist* article above. Your manager may say "ignore it, its a fluke" while really thinking "don't rock the boat!" Smith in *The Lancet*, Horton in the *BMJ* and Goldacre at TED talk about bad incentives and mixed effects from industry.

The decline effect: a scientist some time ago published an article saying "a daily aspirin reduces heart attacks in older men" based on a large study. Over the years negative results gradually trickle in (though initially suppressed) so that meta-analysis (pooled analysis of multiple studies) gradually reduces the effect, as the numbers balance out, until eventually the initially study is shown to no longer hold. What happened? The initial study may have been a fluke, or a bad design. This effect was discussed in Lehrer's *The New Yorker* article.

Inadequate reproducibility: the accepted way of countering the effect of significance chasing is to have an independent team follow up with a second study. In this way, their P value is in some sense a realistic measure because their experiment is run only once whereas the initial result was from a potential pool of 10s-100s of experiments. Reproducibility is important and encouraged, said by Smith in the *BMJ*, but organisationally it can be difficult to achieve, and moreover, as Lehrer in *The New Yorker* says "dramatic findings are also the most likely to get published," so results about reproducibility are hard to publish.

As data scientists, we see the following important lessons:

- When doing discovery, be careful with P-values, or use a more modern statistical technique that accounts for the discovery process in the statistical analysis.
- Record negative results as they may be useful for colleagues studying related data, and you need to learn from the experience. If the negative result goes against prior work of others, then possibly the earlier work was wrong and you may need to double check other results. Many doing experimental work, like data science, will tell you, always check for errors in your own study first before doubting the work of others; perhaps your negative result is your own error.
- For proper evaluation, keep a test set of data aside so you can run independent confirmation of the results from your initial discoveries Most data science challenges use this strategy. If the discovery work is done with active data collection, then you need to reproduce the result after a second data collection step (*i.e.*, to get the test set). A related approach in statistics is the use of cross-validation (https://en.wikipedia.org/wiki/Cross-validation_%28statistics%29).

Correlation does not imply causation

This effect was not discussed in the above articles. The accepted scientific standard for experimental design is blind experiments (https://en.wikipedia.org/wiki/Blind_experiment), which involves taking two groups, one group we treat in a specific way, and one group we do not treat. The "blind" part of the design refers to the fact that subjects are not told which treatment they get to prevent psychological effects (so-called bias) from affecting results. With this active treatment of subjects we get to test if our actions (treat or not) cause a difference. This is accepted as a good proxy for testing causation, as long as our subjects are otherwise similar. In marketing the similar use of two random groups for treatment is called A/B testing (https://en.wikipedia.org/wiki/A/B_testing) and it is widely used by internet companies for advertising. So by good scientific practice, causation is tested well.

However, this is not always possible to do. Data can be hard to get. So scientists still use observational data (where active treatment was not done). The same situation sometimes holds for Data Science. Remarkably, scientists still draw the wrong conclusions about causation and sometimes the effect can be costly. The Wikipedia has a good section on this:

- "Correlation does not imply causation" (https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation) on Wikipedia, and see the **introduction**, **General Pattern** and **Examples** (1600 words, 9 mins)


They mention a number of serious examples which had significant health consequences, and you should review these. The difficult part is looking for the hidden confounding effects that block the correlation from being a cause.

- Testing for causality is a challenging problem from a philosophical and statistical perspective, and when one does not have the luxury of doing active experimentation, some statisticians argue it is impossible, while others argue for particular assumptions. Nonetheless, we need to take great care in studies and use A/B testing or similar where possible if determining causality is important.

Missing the true cost

Ben Goldachre at TED and *The Economist* mention poor quality of experimental designs as being a factor, something they mention journal reviewers should spot but they often do not. This area is considerably more technical, and harder to write up in a non-technical magazine such as *The New Yorker*, so we refer to some medical articles on this. But first a disclaimer. *As we are not medical specialists, we have no idea if the article cited below are correct or not, that is if their analysis holds.* But there is a general principle here for data scientists, so regardless of the veracity of their claims, our lesson still holds. *i.e.*, leave the medical controversy to the medical community.

Cardiovascular mortality: we refer to (but you should not look at):

> *Arch Intern Med*. 2010 Jun 28;170(12):1032-6. doi: 10.1001/archinternmed.2010.184. "Cholesterol lowering, cardiovascular diseases, and the rosuvastatin-JUPITER controversy: a critical reappraisal." de Lorgeril M, Salen P, Abramson J, Dodin S, Hamazaki T, Kostucki W, Okuyama H, Pavy B, Rabaeus M.

de Lorgeril *et al.* claim of the earlier JUPITER study, roughly, as follows (ignoring the specific medical details)

> While the treatment (use of statins) showed a significant decrease in cardiovascular mortality, which is good, it showed no significant decrease in overall mortality.

So measuring "cardiovascular mortality" shows great results for the drug. But a more complete evaluation "overall mortality" shows no difference when using the drug. So it works with one measure, but not with the more important measure.

A related effect is often see with the use of machine learning in Wall Street, trying to use predictive algorithms to develop trading systems for stock. For those in the machine learning community, the results are well known to old-timers: most naive experimental work fails to include "transaction costs" (the transaction fees you pay to the broker). When these are included, they eat away at the profits over time. With every buy or sell, the system loses 1% or some similar amount of the transaction. If a lot of buys/sells are done this amount adds up to much more than 1% of the principle. With transaction costs included, many reportedly successful machine learning trading systems fail to beat simple 'buy and hold' strategies. So the real evaluation measure is money earned from stock minus the transaction costs.

- You have to be careful what your evaluation measure is, and whether it is the one you want. This requires good knowledge of the domain, and mimicking, as far as possible, the different effects of the eventual delivery platform, and looking at the complete costs and benefits. This is what the car industry call "total cost of ownership". This invariably requires a good understanding of the domain, and often is why you need a business process expert in your data science team.

6
Data Curation and Management

This is the last module of six for the Introduction to Data Science unit. This module will introduce students to the basic tasks involved in data curation and management in an organisation. Basic idea is that all other activities that fall under "Data Science" require access and management to the data, so curation and management will be done prior to, in parallel with, and post to any other tasks! This module will emphasise importance, and non-technical aspects of Data Science such as privacy.

(https://www.youtube.com/watch?v=86xUFoexWJc)

Aims of This Module

- Evaluate the relationship between ethics, privacy, storage (i.e., of personal data), security and analysis.
- Identify potential for conflicts among business and legal objectives.
- Determine data management requirements from an internal (data lifecycle) and external (data value chain) perspectives

How to study for this module

In this module we again draw on material in the public domain such as interviews and videos, online magazine entries and blogs. We also have also written some material to tie together various kinds of models. As well as studying and viewing the material, we have some activities around this material.

Please remember:

- Reference items marked with a single "johny look it up" icon,  , should be viewed as *suggested reading*, not essential nor important for assessment.
- Reference items marked with a two "johny look it up" icons,  ,  should be viewed as *important reading*, considered important for assessment.

6.1
Issues in Data Curation and Management

In this section we introduce different issues to do with ethics, information privacy, security and data management in general.

Motivation: What can we learn about you from the Internet of Things

Before discussing privacy, ethics, and security and the related issues under the general heading of data management, let us consider some of the implications. You have almost certainly thought about the problem of losing your credit card details, but consider what happens if someone breaks into your internet connected devices. Charles Givre of Booz Allen Hamilton explored this and the related question of IoT data stored on a car. Using the data from the home-based IoT devices, he could determine when you were not home, and your social media accounts, handy information for traditional thieves or identity thieves.

- "Your Smart Home Knows a Lot About You" (https://www.propublica.org/article/your-smart-home-knows-a-lot-about-you) by Lauren Kirchner (ProPublica blog, 800 words)

Privacy versus confidentiality

Privacy (https://en.wikipedia.org/wiki/Privacy) is a many faceted concept, but in our context privacy is about having control over how one shares oneself with others. Confidentiality, however, is about how information about an individual is treated. So confidentiality is related to information privacy (https://en.wikipedia.org/wiki/Information_privacy). The legals issues here are complex.

Social media and the loss of confidentiality

Jennifer Golbeck discusses how they are able to predict many things about users from the social media. So while we have **explicit data** recorded about us, and we expect that data to be secure, there is also **implicit data** available on us for those with the predictive tools to do the inference.

- "The curly fry conundrum: Why social media "likes" say more than you might think" (https://www.ted.com/talks/jennifer_golbeck_the_curly_fry_conundrum_why_social_media_likes_say_more_than_you_might_think?language=en%20) by Jennifer Golbeck (TED 9 mins, 3700 words transcript) 

To deal with this situation, we should really control our own social media data, but is that possible? Certainly some people think it should be done.

- "Empower consumers to control their privacy in the Internet of Everything" (https://enterprisersproject.com/article/2015/7/empower-consumers-control-their-privacy-internet-everything) by Carla Rudder (blog, 550 words, 3 mins) 

Policies and ethics with big data

There are many more societal questions raised by the use of big data, more than just the loss of information privacy. Eli Collins talks more about these issues, giving some examples, touching on ethics and policy.

- "Pax Data" (http://player.oreilly.com/videos/9781491900345?toc_id=192990) by Eli Collins of Cloudera (O'Reilly video, 10 mins) 

This, however, is a talk about hopes and dreams. How do ideas from ethics and potential policy become widespread practice? Usually this is done through regulations. The process of ensuring you meet regulations is called compliance.

Regulatory compliance

For modern organisations, there are **regulations** in place in areas such as taxation, information privacy and freedom of information, human resources, corporate transparency, banking, *etc.* For instance, in Australia, hospitals are required to have good record keeping in order to satisfy Medicare compliance audits (Medicare is part of the Dept. of Human Services and provides some funding for patients). The **audit** is one way of checking compliance, and various organisations can demand an audit from a company. In Australia, Australian Taxation Office, Medicare, Australian Securities and Investments Commission can all do audits. Other government laws and regulations require transaction data or written records be kept, for instance for *Work Health and Safety* or *Employment*. For instance, tax and employment related records need to be kept for seven years. Note transaction data, which records financial or service transactions or customer interactions, and document data, which records human relations (employment), customer relations such as complaints, are all part of an organisation's data.

The Wikipedia defines auditing as follows:

> Auditing (https://en.wikipedia.org/wiki/Audit) refers to a systematic and independent examination of books, accounts, documents and vouchers of an organization to ascertain how far the financial statements present a true and fair view of the concern.

and regulatory compliance as:

> Regulatory compliance (https://en.wikipedia.org/wiki/Regulatory_compliance) describes the goal that organisations aspire to achieve in their efforts to ensure that they are aware of and take steps to comply with relevant laws and regulations.

In order to meet regulatory compliance, as well as continue the general practices needed for the organization to operate, the organisations have to set up appropriate workflows with their staff, and establish databases or other means to comply.

Data governance

Thus an important part of an organisation's operations are the **management** of its data, which means storing it, backing it up, making it accessible, and finally destroying it once records are no longer needed. Of course, on top of all this, organisations now want to make the data accessible for both business

intelligence and management reporting, both *ad hoc* ("how many of our top salesmen left the company this year") and systematic reports ("sales reports by region, by product class, by quarter"), as well as general predictive analysis and data science. In addition, organisations now understand that they need to properly manage common/overlapping data so that a particular customer's details are not spread over many different departments in isolated databases (the problem of silos mentioned in section **Big Data Processing** of module **Data Types and Storage**.

- "What is Data Governance?" (https://www.youtube.com/watch?v=t4IOS5csv40) by Rand Secure Data (Youtube, 3:15 mins)  explains the role of governance.
- "What is Data Governance?" (https://www.youtube.com/watch?v=sHPY8zIhy60) by Intricity (Youtube, 6 mins)  explains governance strategies

Cybersecurity

Loss of information privacy can happen in many ways. Another possibility is theft, especially of credit card data or other identity or financial data. Here, implications of this are impacting corporations directly, as cybersecurity is now high profile.

- "Target CEO ousted as boards focus on cyber risk mitigation" (http://www.smh.com.au/it-pro/security-it/target-ceo-ousted-as-boards-focus-on-cyber-risk-mitigation-20140506-zr5nm.html) on The Sydney Morning Herald (newspaper article, 1000 words, 5 mins)

In this case, we see cybersecurity being treated as a critical function worthy of upper management attention and becoming part of the core data governance processes, along with the other regulatory compliance requirements. Google, for instance, treats its users' search data in this way: for instance, very few people inside the company have access to the users' search data. They have determined early on to make securing user information privacy one of the core business goals. However, there are people who believe that Google should not keep user search data at all.

Ethical viewpoints

Regarding the management of data for information privacy, Cathy O'Neil claims there are four different groups in the data science world, with different viewpoints on the issues of information privacy and ethics.

- "Four political camps in the big data world" (http://mathbabe.org/2015/04/22/four-political-camps-in-the-big-data-world) by Cathy O'Neil (blog, 700 words, 4 mins) 

While Eli Collins of Cloudera (whose talk was given above) discussed needed policies and actions, these four camps represent the competing interests that can affect what might actually happen. As we see with the cybersecurity example, while some companies now know to protect information privacy as a matter of corporate policy to retain customer respect. However, companies still keep "private" data about individuals as a way of conducting their business, most notably FaceBook and Google use it for targeted advertising. With the advent of Data Science, these organisations are now able to gain implicit data, as discussed by Jennifer Golbeck above, and the implications of this are still to be seen.

Data management in an organisation

To complete this introductory section, we will look at data management practices generally in an organisation, where the second and third videos use the term "data management" in quite different ways. Data management is the broader process, from Wikipedia:

> Data management (https://en.wikipedia.org/wiki/Data_management) is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets.

- "The Difference Between Data Governance & Data Management" (http://blogs.perficient.com/healthcare/blog/2012/06/12/data-governance-vs-data-management/) blog by Pete Stiglich (blog, 550 words) 
- "Top 10 Mistakes in Data Management" (https://www.youtube.com/watch?v=5Pl671FH6MQ) a tutorial from Intricity (a data management company) (Youtube, 6 mins) 
- "How to avoid a data management nightmare" (https://www.youtube.com/watch?v=nNBiCcBlwRA), a video created by NYU Health Sciences Library (Youtube, 5 mins) 

Data management in science/research

In science and research, data management has many different constraints. Business contraints like corporate compliance do not affect many science projects, although issues of information privacy, where human subjects are involved, is similar. The Digital Curation Centre (DCC) in the UK presents many if the issues:

- "Managing Research Data" (http://www.dcc.ac.uk/news/managing-research-data-video) from DCC (video, 12:30 mins) 

Note the DCCs keeps a DC 101 materials (http://www.dcc.ac.uk/training/train-the-trainer/dc-101-training-materials) site with extensive training material.

Data management and Data Science

We note that data management serves multiple roles inside an organisation. Only one of those roles is the provision of data for data science projects. However, many of the issues affecting data management also affect a data science project. Consider the following examples:

- For a medical informatics project, predicting fungal infections from nursing notes, the data science team needs to sign confidentiality and security assurances, and must only work with the data with in the confines of the hospital's informatics department.
- A internet advertising company needs to consider the sorts of implicit and explicit data it is willing to store and use about a user.
- A large retailing company has separate departments from the smaller clothing companies it

recently acquired. It now wishes to conduct market intelligence on new products so wishes to pool the product and sales databases from the constituent companies for a data science project.

- A data science project in medical informatics is successful and the developed predictive system is to go live. This means developing a maintenance plan to keep the system in optimal performance going forward. During the project, they discovered a lot of problems with the initial data set: nurses and doctors had kept some parts of the records in compliance with billing regulations, but not properly reflecting the full diagnosis and treatment listed in the written notes. Part of the project was therefore data cleaning and some fact checking by the resident doctors. Putting their system into production therefore means changing some of the standard practices of the staff.

6.2
Frameworks for Data Management

Data management was introduced in the previous section. Data management frameworks, in general, come from different perspectives, but have a lot of features in common. In this section we review some of the frameworks with particular reference to those aspects relevant to Data Science projects.

The different perspectives we observe are as follows:

Science: science data management has a particular emphasis on supporting science, reproducibility and credibility of scientific work, and producing artifacts of knowledge for the public good. Science teams invest a lot of effort in gathering data, and more and more effort is now being put into its management.

Business/Organisational: the process of data management emphasised in the previous section with governance, compliance, information privacy, *etc.* as discussed.

Digital Curation: the term curation comes from information sciences and includes museums and libraries, where preservation and cataloguing are important, but has a strong overlap with business/organisation and science management practice.

Government: standard government practice includes issues such as transparency and different regulatory practices to businesses and organisations. Government generally needs to keep full records of decision making, but subject to freedom of information requests, and retain records for a fixed period. Government also has multiple layers of security in its operations. Moreover, government is a key producer of data and supports this role as a service to industry and the general public.

Medical: while perhaps a mix of some of the above perspectives, medicine and health are their own unique area with significant privacy issues, conflicting corporate financial constraints, government regulations and the furthering of medical science.

Digital Curation Centre's lifecycle model

The DCC has developed a model available on a single infographic (available on one of two formats):

- "The DCC Curation Lifecycle Model" (http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf) by DCC (PDF, 10 mins to review) 
- "The DCC Curation Lifecycle Model" (http://www.dcc.ac.uk/resources/curation-lifecycle-model) by DCC (webpage, 10 mins to review)

Sequential actions in the DCC lifecycle are given below. The influence of the digital curation field can be seen with the preservation action, which would be renamed as protection or backup in the informations systems world.

Step	Description
Conceptualise	Conceive and plan the creation of data, including capture and storage.

Create or Receive	Create data including administrative, descriptive, structural and technical metadata.
Appraise and Select	Evaluate data and select for long-term curation and preservation, noting policies and legalities.
Ingest	Transfer data to an archive, repository, or data centre.
Preservation Action	Undertake actions to ensure long-term preservation and retention of the authoritative nature of data.
Store	Store the data in a secure manner adhering to relevant standards.
Access, Use and Reuse	Ensure that data is accessible to both designated users and reusers.
Transform	Create new data from the original.

"APS Better Practice Guide for Big Data" from the Australian Public Service

A task force within the APS released a strategy document. Regarding ethics, privacy and security, they say (page 13):

> As with traditional data management and analytics, governance of programs and projects is critical. A strong governance framework will include sensible risk management and a focus on information security, privacy management and agency specific legislative obligations as they relate to data use, acquisition and secrecy.
>
> In a big data environment agencies are required to respect privacy and be responsible for the safe and proper use of data, particularly when the data being used is sensitive. This includes the requirement for agencies to have clear and transparent privacy policies and provide ethical leadership on their use of big data.
>
> As with many other large analysis projects that may identify scope for change, projects can be risky - they require experimentation and discovery, they can deliver unexpected results and sometimes no significant results, in extremely rare cases they can deliver false results. As such, capability development will need to be governed and expectations will need to be managed at the outset of big data projects; this includes stakeholders, technical personnel and the community.

Regarding data management, they say (page 15):

> Commonwealth data, including data used by analytics projects, needs to be authentic, accurate and reliable if it is to be used to support accountability and decision making in agencies. The creation, collection, management, use and disposal of agency data is governed by a number of legislative and regulatory requirements, government policies and plans. These regulatory requirements also apply to the management of all data sources.

Read more about their statements on privacy and security on pages 18-21 of the report.

- "APS Better Practice Guide for Big Data" (http://www.finance.gov.au/blog/2014/04/16/aps-better-practice-guide-for-big-data/) (webpage with links to report, 30 page PDF report)

Science/Research lifecycle models

Science and research data management frameworks are most similar to what is needed for Data Science because, often times, they are doing Data Science.

DataONE is a US funded research organisation, "Data Observation for Earth" with an interest in data collection. They have published an extensive guide intended for the general public on data management. Here they discuss issues of data quality, metadata creation, protection and backups, and discovery of data from external sources. They also provide a tool for preparing data management plans. Note most data management organisations provide facilities for creating data management plans, so view this as one random example.

- "Data Management Guide for Public Participation in Scientific Research" (https://www.dataone.org/sites/all/documents/DataONE-PPSR-DataManagementGuide.pdf), by DataONE 2013 (extensive 15 page reference); review pages 2-3 describing the model and Steps 3-6 on pages 5-12  , especially the best practices, which are Assure, Describe, Preserve and Discover.
- "Example Data Management Plan: Rio Grande Basin Hydrologic Geodatabase Compendium" (https://www.dataone.org/sites/all/documents/DMP_Hydrologic_Formatted.pdf) from DataONE (PDF, 3 pages, 5 mins) scan the plan as the fine detail is not important

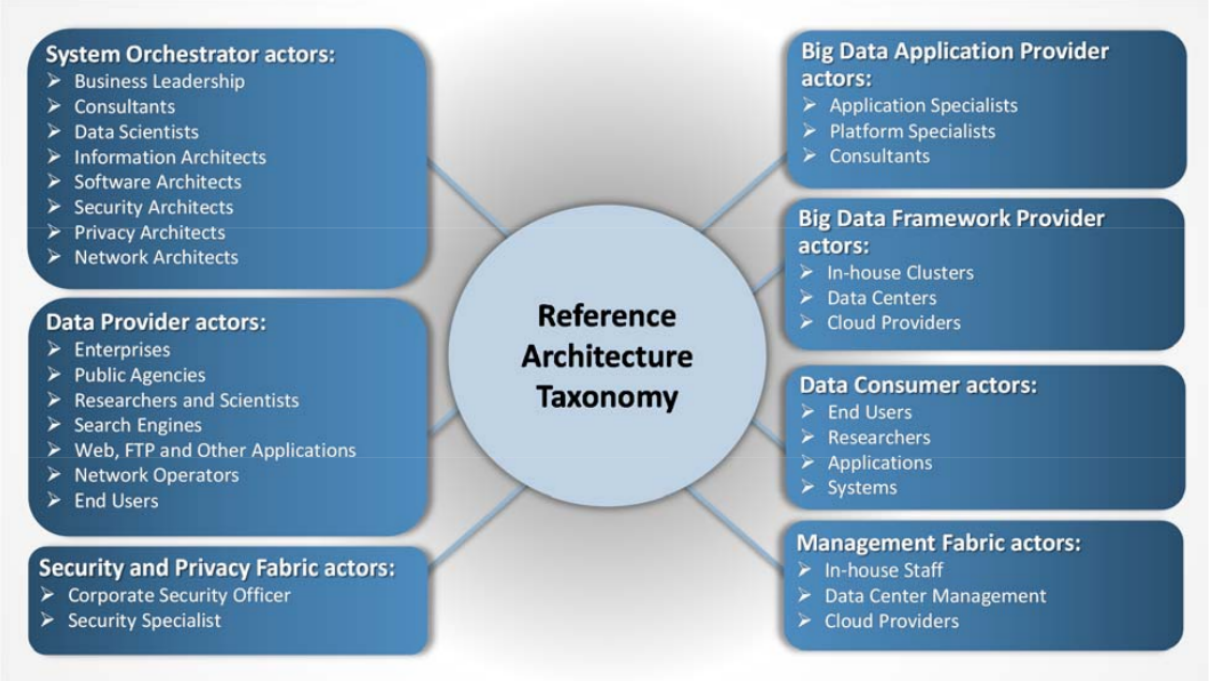
United States Geological Survey (USGS) has over a century of experience in creating data products. Of their lifecycle, the steps Plan and Acquire are most relevant to a data science project covering many aspects of data formats, metadata, *etc.*

- "The United States Geological Survey Science Data Lifecycle Model" (http://www.usgs.gov/datamanagement/why-dm/lifecycleoverview.php) from the USGS (extensive set of nested webpages, about 30 in all); review the Plan and Acquire subpages of the documentation only 

NIST's use case requirements model

NIST presents a model derived from big data use cases, thus it is a model targeted specifically at the data science case, moreover it pays more attention to the data analysis project itself. The NIST model was previously reviewed in the section **Business Models with Data** of module **Data Models in Organisations,** which covered the report "NIST Big Data Interoperability Framework: Volume 6, Reference Architecture" (http://dx.doi.org/10.6028/NIST.SP.1500-6) (included here for reference only).

They separate out data science projects into tasks by a number of actors, given in the figure *NIST Reference Architecture Taxonomy*.



NIST Reference Architecture Taxonomy

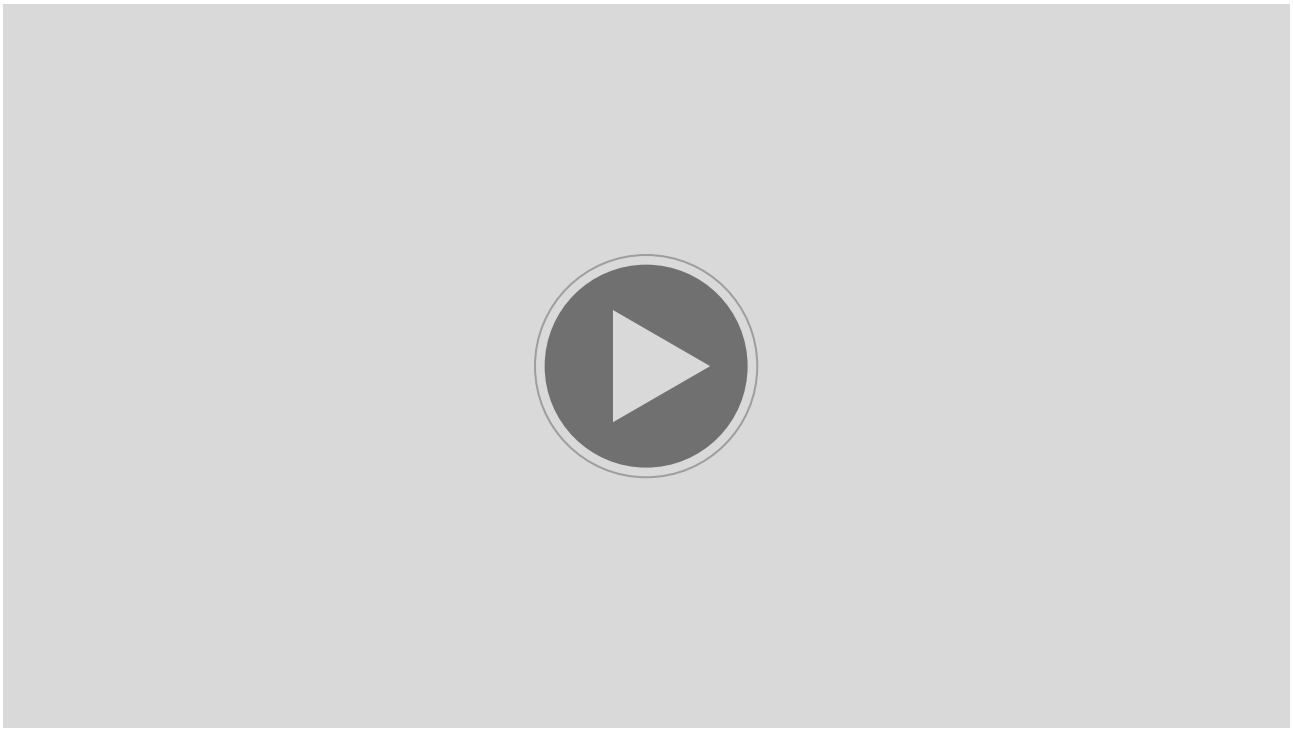
The main actors in a project are the **data provider** and the **application provider**, with examples given in the figure.

The data provider performs and manages the following: Data Capture from Sources, Data Persistence, Data Scrubbing, Data Annotation and Metadata Creation, Access Rights Management, Access Policy Contracts, Data Distribution Application Programming Interfaces, Capabilities Hosting, Data Availability Publication.

The application provider performs the following actions: Collection, Preparation, Analytics, Visualization, and Access.

6.3
Interview on Data Management

Watch Con Nidras (Head of Customer and Channel Analytics - National Australia Bank (NAB)) and Associate Professor Chris Bain (Director of information services - The Alfred Hospital) talk about their experiences with **data management**.



(https://www.alexandriarepository.org/wp-content/uploads/20150629094440/FIT5145_module_6_data_management_combined.mp4.mp4)

Alternatively, you can download the transcript for Data management
(https://www.alexandriarepository.org/wp-content/uploads/20150701100116/transcript_FIT5145_module_6_data_management.pdf).

7
Data Science Resources

General Resources

There are too many resources to name or list them all, or to attempt to do some kind of tracking. It is recommended, however, that you install a news aggregator on your tablet/smart-phone/laptop and enrol in some of the better and more relevant RSS feeds, to keep track.

Magazines

All the big **business and technology magazines** have relevant sections on Data Science or Big Data: Forbe's (http://www.forbes.com/search/?q=data+science), Harvard Business Review (https://hbr.org/search?term=data%20science), O'Reilly (http://radar.oreilly.com/data), ZDNet (http://www.zdnet.com/topic/big-data/), MIT Sloan Management Review (http://sloanreview.mit.edu/tag/data-science/), Information Week (http://www.informationweek.com/big-data.asp), Wired (http://www.wired.com/tag/data-science/), InfoWorld (http://www.infoworld.com/category/data-science/), TechCrunch (big data) (http://techcrunch.com/tag/big-data/) and TechCrunch (data science) (http://techcrunch.com/tag/data-science/), ... Each of these has a particular perspective, which is useful in understanding their contributions. For instance, TechCrunch is a technology startup magazine whereas Forbes targets Fortune 500 companies. The articles in this class of magazines usually are good quality, although they are sometimes "commissioned" journalism or press releases for marketing.

Blogs

Many **technology blogs** focus on Data Science. The following are listed as most popular first: KDNuggets.com (http://www.kdnuggets.com/), DataScienceCentral.com (http://www.datasciencecentral.com) and its offshoot AnalyticBridge.com (http://www.analyticbridge.com), Datafloq.com (https://datafloq.com/), PredictiveAnalyticsToday.com (http://www.predictiveanalyticstoday.com), Dataconomy.com (http://dataconomy.com), 101.DataScience.community (http://101.datascience.community), DataScienceWeekly.org (http://DataScienceWeekly.org). The first, KDNuggets has been in the business for almost two decades. Many of these have email and RSS subscription services and Twitter feeds. Some of these have a low signal to noise ratio so it is easy to get drowned in content. See also Quora's " (http://www.quora.com/What-are-the-best-blogs-for-data-scientists-to-read)What are the best blogs for data scientists to read?" (http://www.quora.com/What-are-the-best-blogs-for-data-scientists-to-read) for more discussion.

There are two weekly newsletters that you should sign up to for great content in your email. The O'Reilly Data Newsletter (http://www.oreilly.com/data/newsletter.html) is more about industry and is essential reading for anyone who wants to remain current. The Data Science Weekly Newsletter (http://datascienceweekly.org/) has more of a technology orientation with, for instance, some popular machine learning content.

Conferences

Most of the blogs are also coupled with **curated information** sources. Another site with curated information is Big Data and Applications Knowledge Repository (http://www.aee.odu.edu/bigdata/index.php). This also has a good list of conferences. But my favorite **Data Science conference series** is Strata+Hadoop World (http://conferences.oreilly.com/strata) which are global and several times a year. Their talks are recorded

and some are available for free afterwards. They also have extensive tutorial segments on all the best technologies. Don't be fooled by Hadoop in the title, it covers way more than Hadoop in terms of big data processing.

Question Answering

A related category are the **question answering sites**: Quora has Data Science (http://www.quora.com/Data-Science) and Big Data (http://www.quora.com/Big-Data) channels, though many other discussions are useful too. A site more in the Slashdot style is Datatau.com (http://www.datatau.com). Related are content curation sites like ScoopIt which has a big data group (http://www.scoop.it/i/big-data).

Infographics and Cheat Sheets

Pinterest.com (http://Pinterest.com) is a site that records **infographics.** *e.g.,* queries for "data science (https://www.pinterest.com/search/boards/?q=data+science)" and "big data (https://www.pinterest.com/search/boards/?q=big+data)". These are seductive, and some certainly informative. However, you have to register with them. Worth the trouble. Datafloq.com (https://datafloq.com/read/?cat=25) also has an infographics section. Some notables here that go way beyond infographics are **cheat sheets**: Machine Learning Cheat Sheet (http://eferm.com/machine-learning-cheat-sheet/) and the Probability Cheat Sheet (http://www.wzchen.com/probability-cheatsheet). These are handy academic references, and also a nice way to find out what you do *not* know.

Data Sets

Many sites give **collections of data sets**, so perhaps the most notable here are: aws.amazon.com data sets (https://aws.amazon.com/datasets), KDNuggets.com awesome public datasets (http://www.kdnuggets.com/2015/04/awesome-public-datasets-github.html), Google's public data directory (http://www.google.com.au/publicdata/directory), Quora.com large data sets (http://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public), and Xiaming Chen's Github'd awesome list (https://github.com/caesar0301/awesome-public-datasets) … The Internet Archive (https://archive.org/index.php) is a long running source of free digital content (books, *etc.*). There are many, many more such sites, especially as governments now support open data.

On Wikipedia

Finally, most terms and concepts are well explained in the **Wikipedia**, often with good diagrams and related discussions. As one delves into the more esoteric aspects of statistics or computer science, the quality of Wikipedia's entries drop's off. Wikipedia's definition of Data Science (https://en.wikipedia.org/wiki/Data_science), for instance, as "a continuation of the field data mining and predictive analytics" would be hotly contested by some, but others would find the distinctions not that important.

NIST Big Data Interoperability Framework Reports

To advance progress in Big Data, the NIST Big Data Public Working Group (NBD-PWG) worked to develop consensus on important, fundamental concepts related to Big Data, including Data Science. Meetings were held in 2013, draft reports released in 2014 and final reports 2015. The results are reported in the NIST Big Data Interoperability Framework series of volumes.

There are seven reports by NIST. These links give the DOI for the document, and return a PDF file.

- Volume 1, Definitions (http://dx.doi.org/10.6028/NIST.SP.1500-1),
- Volume 2, Big Data Taxonomies, (http://dx.doi.org/10.6028/NIST.SP.1500-2)
- Volume 3, Use Cases and General Requirements (http://dx.doi.org/10.6028/NIST.SP.1500-3),
- Volume 4, Security and Privacy (http://dx.doi.org/10.6028/NIST.SP.1500-4),
- Volume 5, Architectures White Paper Survey (http://dx.doi.org/10.6028/NIST.SP.1500-5),
- Volume 6, Reference Architecture (http://dx.doi.org/10.6028/NIST.SP.1500-6),
- Volume 7, Standards Roadmap (http://dx.doi.org/10.6028/NIST.SP.1500-7).

Alternatively, these can be obtained from the NIST database of publications through the NIST Publications Portal. (http://www.nist.gov/publication-portal.cfm)

Other related reports and slide sets are as follows:

- "Big Data Use Cases and Requirements" (http://dsc.soic.indiana.edu/publications/NISTUseCase.pdf) by Geoffrey Fox and Wo Chang; short PDF report giving background on the use cases
- "NIST Big Data Public Working Group NBD-PWG (http://www.slideshare.net/Foxsden/nist-bdwg-bigdataoverviewoct1113a)" 33 page PPT presentation by Geoffrey Fox presented end of 2013, up on Slideshare.net.
- "Big Data Standard Ecosystem and Applications. (http://www.fedsummits.com/oldfedsummits/wp-content/uploads/2015/12/Wo-Chang-ATARC-Federal-Big-Data-Summit-2015-12-08.pdf)" by Wo Chang, final slides in PDF Dec.. 2015.
- "ISO/IEC JTC 1. Preliminary Report 2014. Big data (http://www.iso.org/iso/big_data_report-jtc1.pdf)", report from ISO, some drawn from NIST plus other material.