Introduction to Data Science

FIT1043

Monash University

to FIT1043:
Introduction to data science

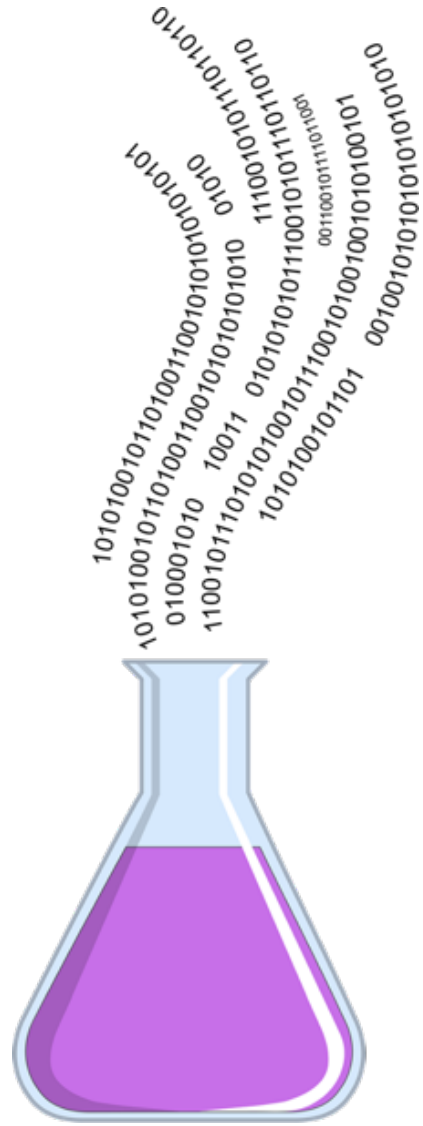


Image modified from: https://pixabay.com/en/beaker-glass-science-lab-23417/
and https://openclipart.org/detail/202770/binary-code-wave

Before we get started

Do the following:

1. get out your device
2. login to Moodle and download these slides from the FIT1043 page
3. complete *this survey*
4. download *this ePub* from Alexandria

About this Unit

Resources

1. review of *Moodle*
   - contains Unit Orientation, Assessments and Discussion Forums
   - as well as Lecture Notes, which contain active links to recommended videos & readings
2. review of *Alexandria* + ePubs
   - contains LOTS of additional resources and exercises
   - use as an online textbook

Resources

1. review of *Moodle*
   - contains Unit Orientation, Assessments and Discussion Forums
   - as well as Lecture Notes, which contain active links to recommended videos & readings
2. review of *Alexandria* + ePubs
   - contains LOTS of additional resources and exercises
   - use as an online textbook
3. additional textbook:
   - No "perfect" *Introduction to Data Science* textbook available
   - But a good introductory text available for purchase is: *The Art of Data Science* by Peng & Matsui

Resources

1. review of *Moodle*
   - contains Unit Orientation, Assessments and Discussion Forums
   - as well as Lecture Notes, which contain active links to recommended videos & readings
2. review of *Alexandria* + ePubs
   - contains LOTS of additional resources and exercises
   - use as an online textbook
3. additional textbook:
   - No "perfect" *Introduction to Data Science* textbook available
   - But a good introductory text available for purchase is: *The Art of Data Science* by Peng & Matsui
4. be aware also of the:
   - library services available
   - special consideration policies
   - disability support available

Getting Started

1. no tute this 1st week
2. but there are activities to do, so check them on Moodle
   - Module 1: Data Science and Data in Society

Getting Started

1. no tute this 1st week
2. but there are activities to do, so check them on Moodle
   - Module 1: Data Science and Data in Society
3. how these classes are run
   - watch videos and read background material between classes
   - prepare for labs/tutes
   - bring a device to lectures to participate
   - see Module 7 in ePub for Data Science Resources

Contacts

Need help?

1. ask questions during tutorials and lectures
   - please interrupt me with questions!
2. check for relevant Discussions on Moodle
   - note in particular the “Assessments” discussion threads
3. attend the consultation hour of the tutors or the lecturer
   - consultation hours in Moodle
4. send email to tutor or lecturer

Motivation for the Unit

Data Science is in its growth phase:

- every academic & industry community wants to claim credit
- huge community of “leading international experts,” “highly sought-after consultants,” and “thought leaders” to confuse you with advice, blogs, guidelines, ...
- huge growth in software and services

We try and cover the full extent of what makes Data Science:

- background and context
- leading review articles, lectures, introductions
- academic surveys and national programmes

Prerequisites

You will need:

- high school level of mathematics and statistics
- a “critical mindset”:
  - you will read/view a variety of material
  - different levels of quality and standards
  - some sales, some educational, some journalistic
- basic exposure to information technology and internet businesses:
  - software, science or business computing
  - MS Excel
  - laptop or tablet, ...
  - Amazon, Google, Twitter, ...
- basic exposure to “business thinking”

Warning

Alexandria links to a LOT of content:

- videos, blogs, articles, ...
- there is way too much for you to read it all in detail!

Strategy:

- limit your time per week
- get the big picture from articles/videos
- find out what is out there
- focus in on the details when you need something for assessment (or on stuff you want for your own development)

Unit Schedule: Modules

Module	Week	Content
1.	1	overview and look at projects
	2	(job) roles, and the impact
2.	3	data business models
	4	application areas and case studies
3.	5	characterising data and "big" data
	6	data sources and case studies
4.	7	resources and standards
	8	resources case studies
5.	9	data analysis theory
	10	data analysis process
6.	11	issues in data management
	12	data management frameworks

Any Other Questions?

... before we get started?
... otherwise, use the discussion board!

Your Background

Describe a little bit about yourself: (link only works for the lecturer ;-)

FIT1043: Introduction to Data Science

Module 1:
Data Science and Data in Society

Lecture 1: Overview and a Look at Projects

Monash University

Unit Schedule: Modules

Module	Content
1.	**overview and look at projects** **(job) roles, and the impact**
2.	data business models application areas and case studies
3.	characterising data and "big" data data sources and case studies
4.	resources and standards resources case studies
5.	data analysis theory data analysis process
6.	issues in data management data management frameworks

Overview of Data Science (ePub section 1.1+1.3)

a quick overview of the context

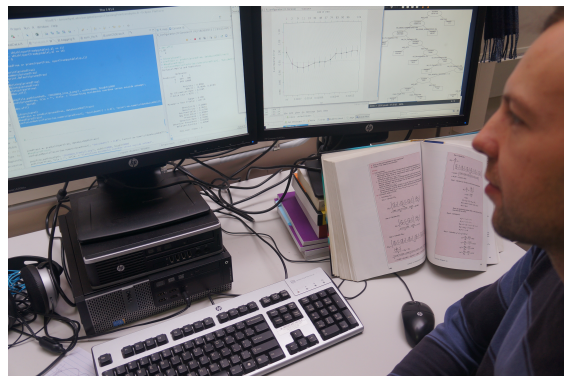
Who are the Data Scientists?

Vote by clicking on the *survey form*

person A

person B

person C



person D

What is Data Science?

“name contains the word ‘science’, so it can’t be one”

- *Note: this is an old joke ...*

What is Data Science?

‘‘name contains the word ‘science’, so it can’t be one”

- *Note: this is an old joke ...*

“data science is what a data scientist does”

- *a circular definition!*

What is Data Science?

‘‘name contains the word ‘science’, so it can’t be one”

- *Note: this is an old joke ...*

“data science is what a data scientist does”

- *a circular definition!*

“data science is the technology of handling and extracting value from data”

- *less circular and a bit more useful*

What is Data Science?

“name contains the word ‘science’, so it can’t be one”

- *Note: this is an old joke ...*

“data science is what a data scientist does”

- *a circular definition!*

“data science is the technology of handling and extracting value from data”

- *less circular and a bit more useful*

“machine learning on big data”

- *useful, but too narrow!*

Machine Learning Definition

(well understood and agreed on)

Machine Learning (ML) is concerned with the development of algorithms and techniques that allow computers to *learn*.

- concerned with building computational artifacts
- but the underlying theory is statistics

Why use Machine Learning?

Machine learning is useful when:

- Human expertise is not available
  e.g. Martian exploration
- Humans cannot explain their expertise (as a set of rules), or their explanation is incomplete and needs tuning
  e.g. speech recognition
- Many solutions need to be adapted automatically
  e.g. user personalisation
- Situation changes over time
  e.g. junk email
- There are large amounts of data
  e.g. discover astronomical objects
- Humans are expensive to use for the work
  e.g. handwritten zipcode recognition

Why use Machine Learning?

because you do not want to be this poor guy!

- sifting through all the data by hand

Why use Machine Learning?

Other reasons for needing Machine Learning:

- the information society
- information warfare
- information overload
- information access

Exercise: Google these to find out about them!

Data Science Examples

Some famous data science projects and investigations:

1. Google's spell checker and *translation engine*
   - we'll learn about these in Module 5
2. Amazon.com's *recommendation engine*
3. Public health: *"saturated fat is not bad for you after all"*
   - many more of this type of investigation will be coming ...
4. Microsoft's *predictive analytics for traffic*

Another Example: Melbourne Datathon 2016

Background for this is Section 1.2 of the ePub

- *Seek.com* is an online jobs website, which provided data and tasks for a data science competition.
- The tasks were:
  - label prediction: predict if a job is in the "Hotel and Tourism" category
  - data exploration: what useful information can be discovered from the data that Seek can use.
- See their own description of the *business context and dataset*.

Datathon Questions

- how did Seek come up with their prediction task?
- why is it important to them?
- did a data scientist come up with the task?
- all Datathon participants had to destroy their copies of the data at the end of the Datathon: why?
- how would you present results of exploratory analysis to Seek.com management? see *one such presentation by the 4Quarters team*

Datathon Questions, cont.

- how much data is there?
- what software/systems could you use to do the prediction task?
- could you introduce/find auxiliary data to do the prediction better? is that “cheating”?
- how would you estimate how well your predictions are going?
- how would Seek.com “fairly” evaluate participants in the Datathon?

Historical Context

Links to resources providing historical background to data science:

- *Wolfram Alpha: computable knowledge history*
- *Cloud Infographic: Evolution Of Big Data*
- *The Web Technology timeline*
- *A brief history of Data Science*

The Rise of Big Data

in *Foreign Affairs*, by Cukier and Mayer-Schoenberger

Data Science interest is related to the arrival of "Big Data"

- data collection has changed:
  - lots of data, but more messy
  - don't look for perfect models – settle for finding patterns
  - examples: Google's *language translation* and *flu trends*
- datafication:
  - taking all aspects of life and turning them into data
  - e.g. NYC using big data to improve public services and lower costs
- the "information society" has come of age
  - and data brokers have started amassing huge data about individuals: *big data could become Big Brother*

Homework

From Section 1.1:

- watch *Cukier's TED talk on "Big Data"*
- watch the video on *"Big Data" by Tim Smith*

The Data Science Process (ePub section 1.2)

what happens in a Data Science project?

- illustrating the process
  - a quick walkthrough illustrating the steps
- the standard value chain
  - our model of the process

The Data Science Process: Illustrating the Process

a quick walkthrough illustrating the steps

The Data Science Process

Many different tasks come together to complete a Data Science project

- a data scientist should be familiar with most, but doesn't need to be an expert in all

Not all are labelled as Data Science

- some from other field such as computer engineering, business, ...

Pitching ideas for data science projects to investors/managers.

"Young Business Man Holding a Tablet" by Pic Basement, CC-BY 2.0

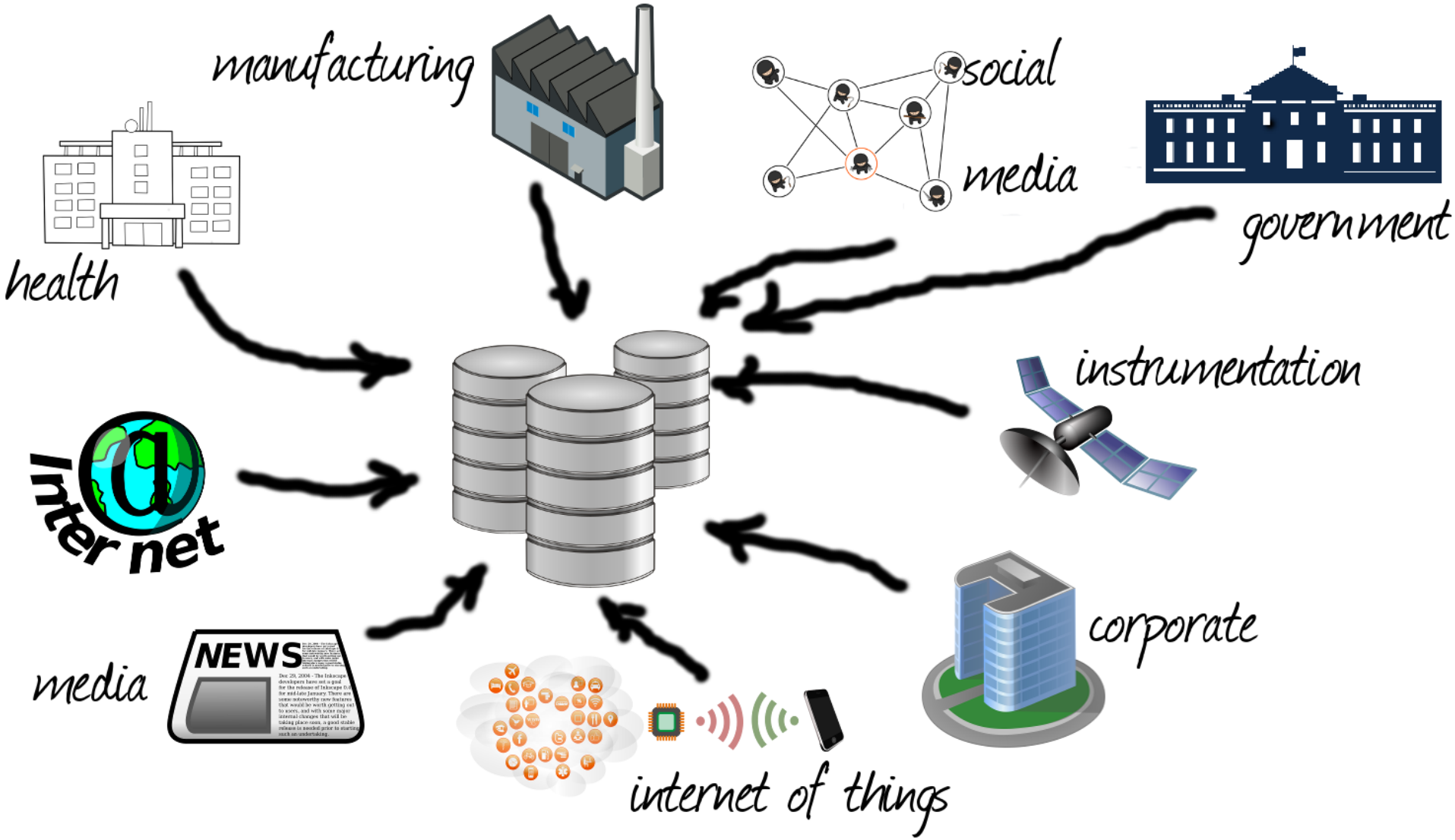
CC BY

Collecting data: researchers preparing to x-ray a patient.

by Stephen Ausmus acquired from USDA ARS, public domain.

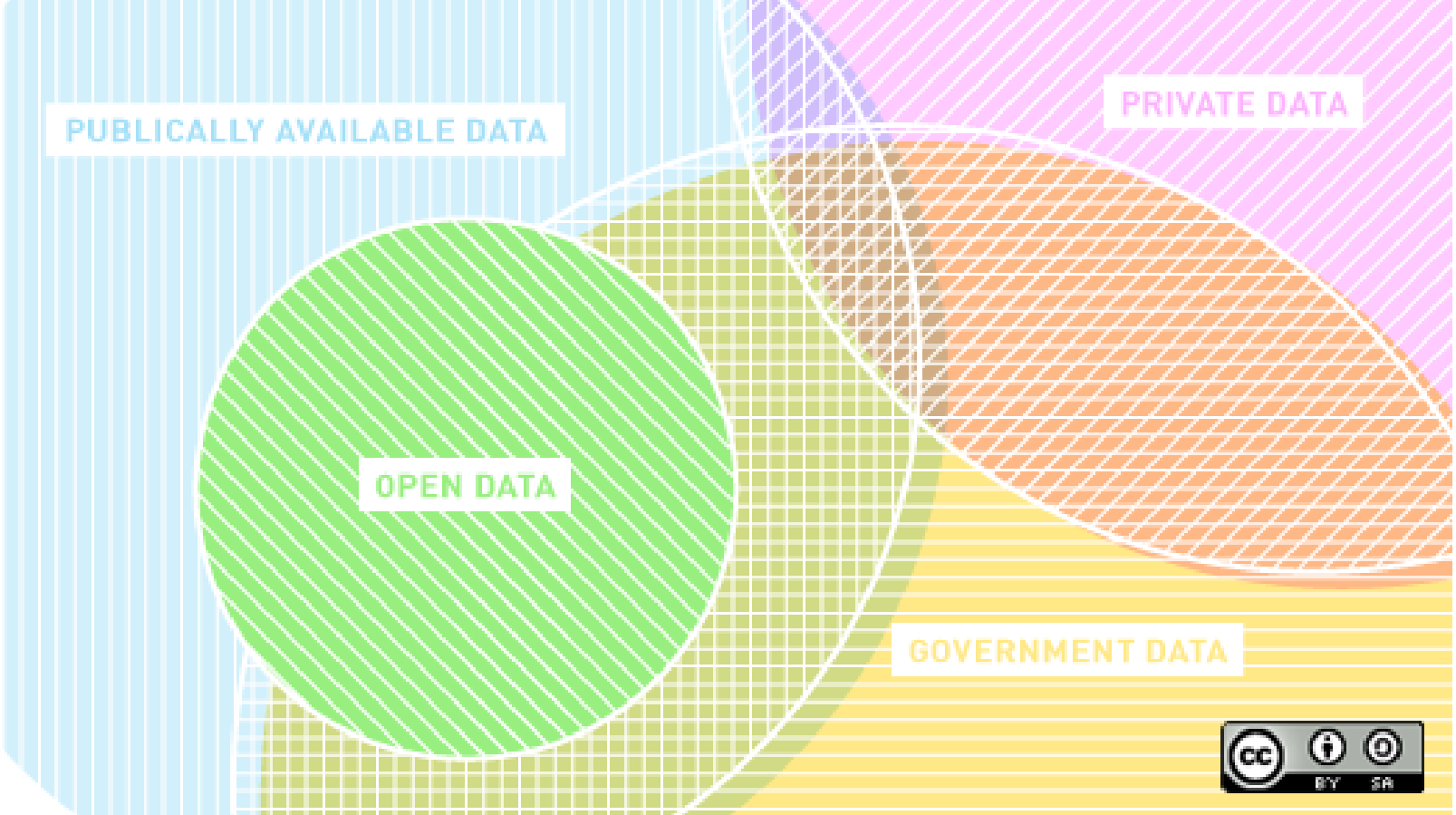
Monitoring: Scientists watch over data collected by the gravimeter & magnetometer instruments.

by NASA/GSFC/Jefferson Beck, CC-BY 2.0



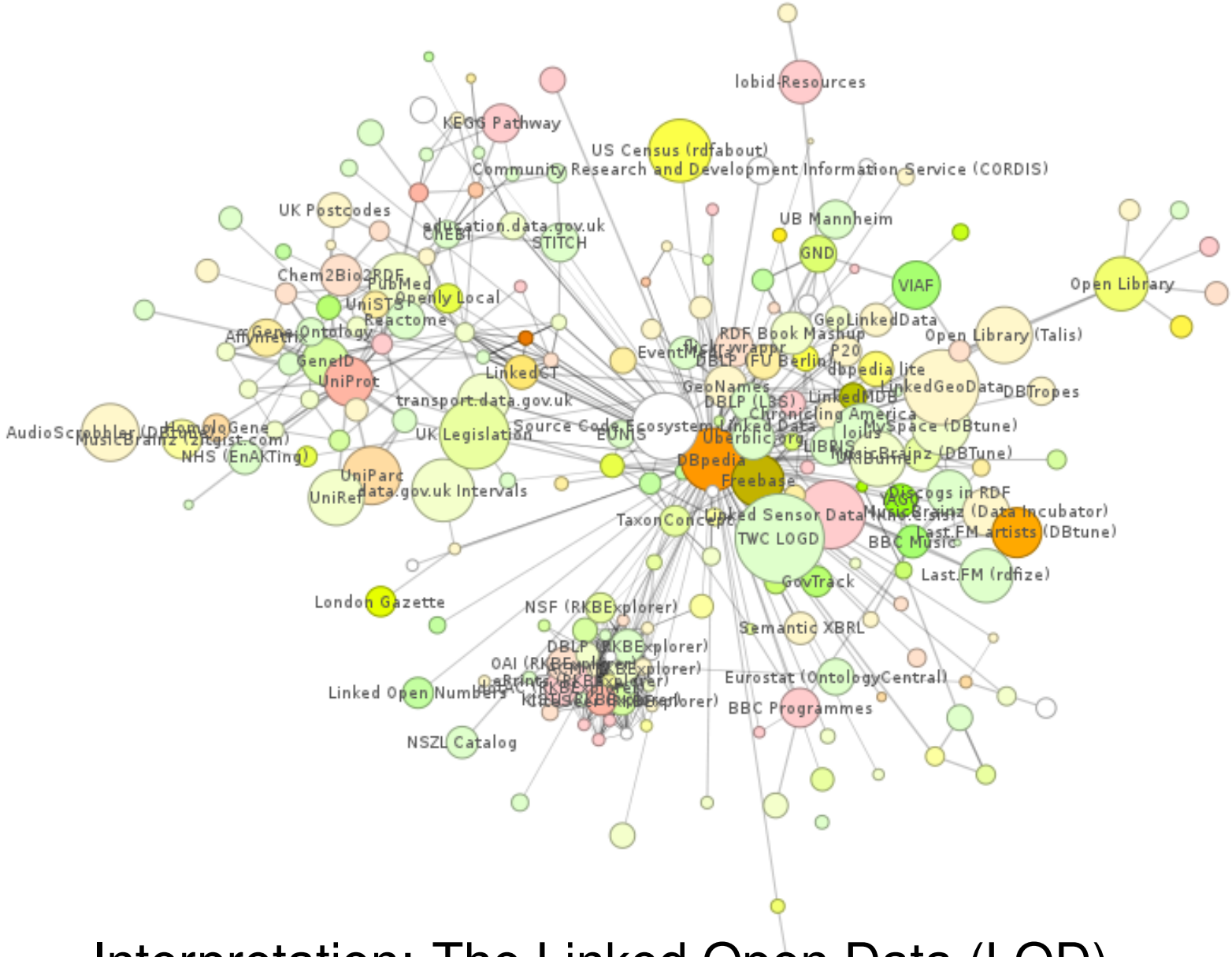
Integration: Data can come from many different sources.

icons from by Openclipart.org, public domain



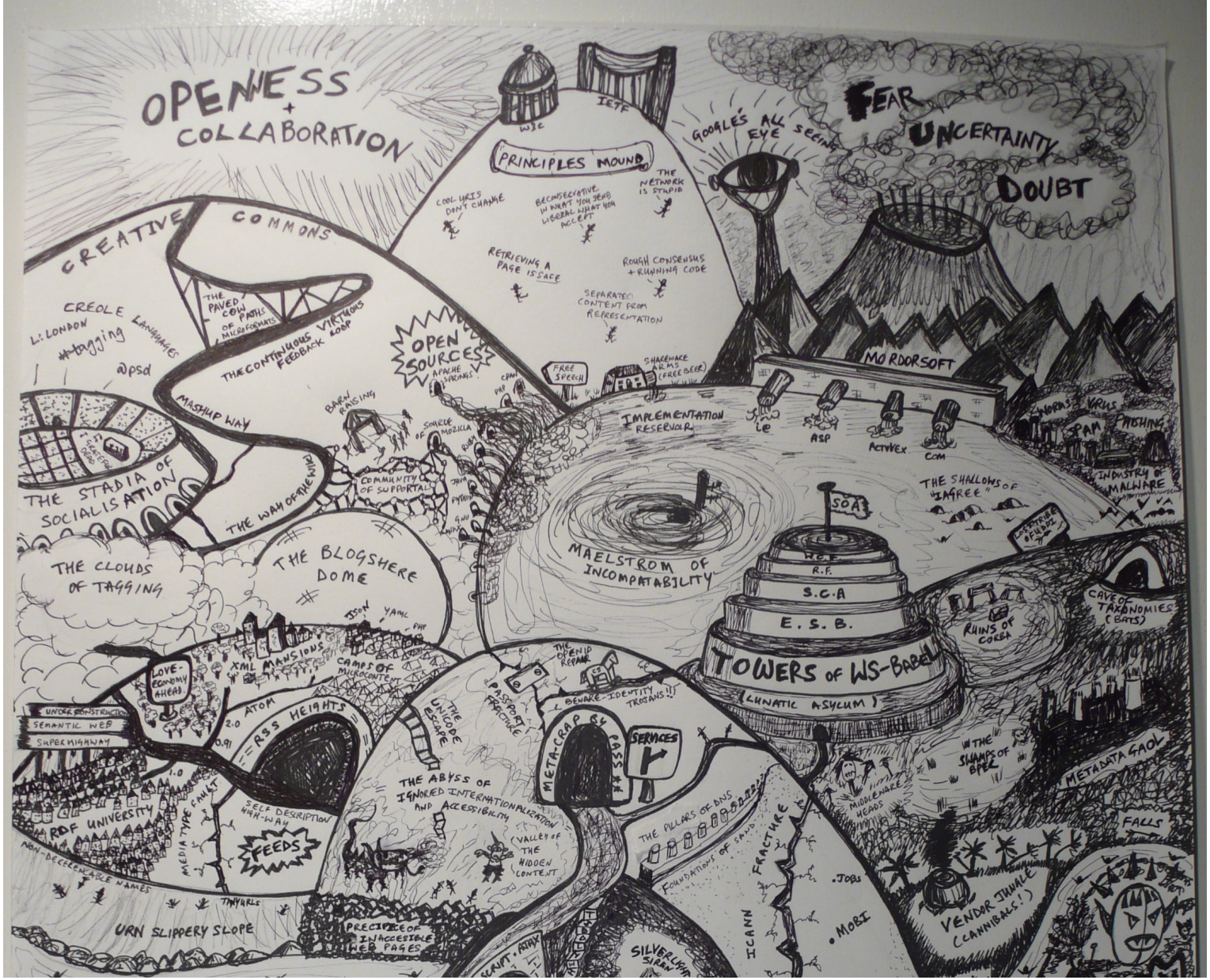
Some of the best data is Open (publicly available) Data.

by Libby Levi for opensource.com, CC-BY-SA 2.0



Interpretation: The Linked Open Data (LOD) graph can ascribe meaning (semantics) to data.

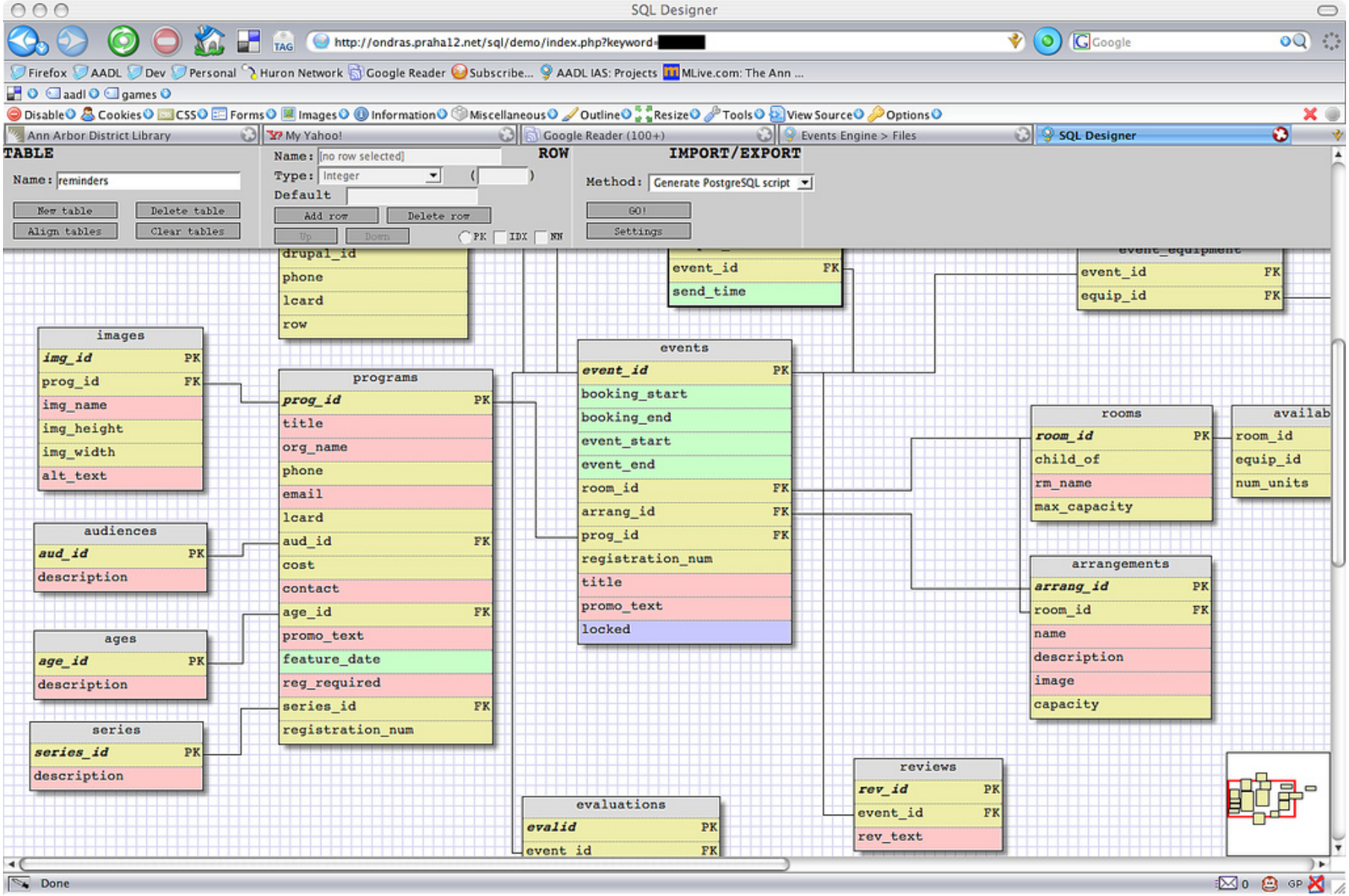
by Open Knowledge, CC-BY-SA 2.0



Navigation: of data standards and formats

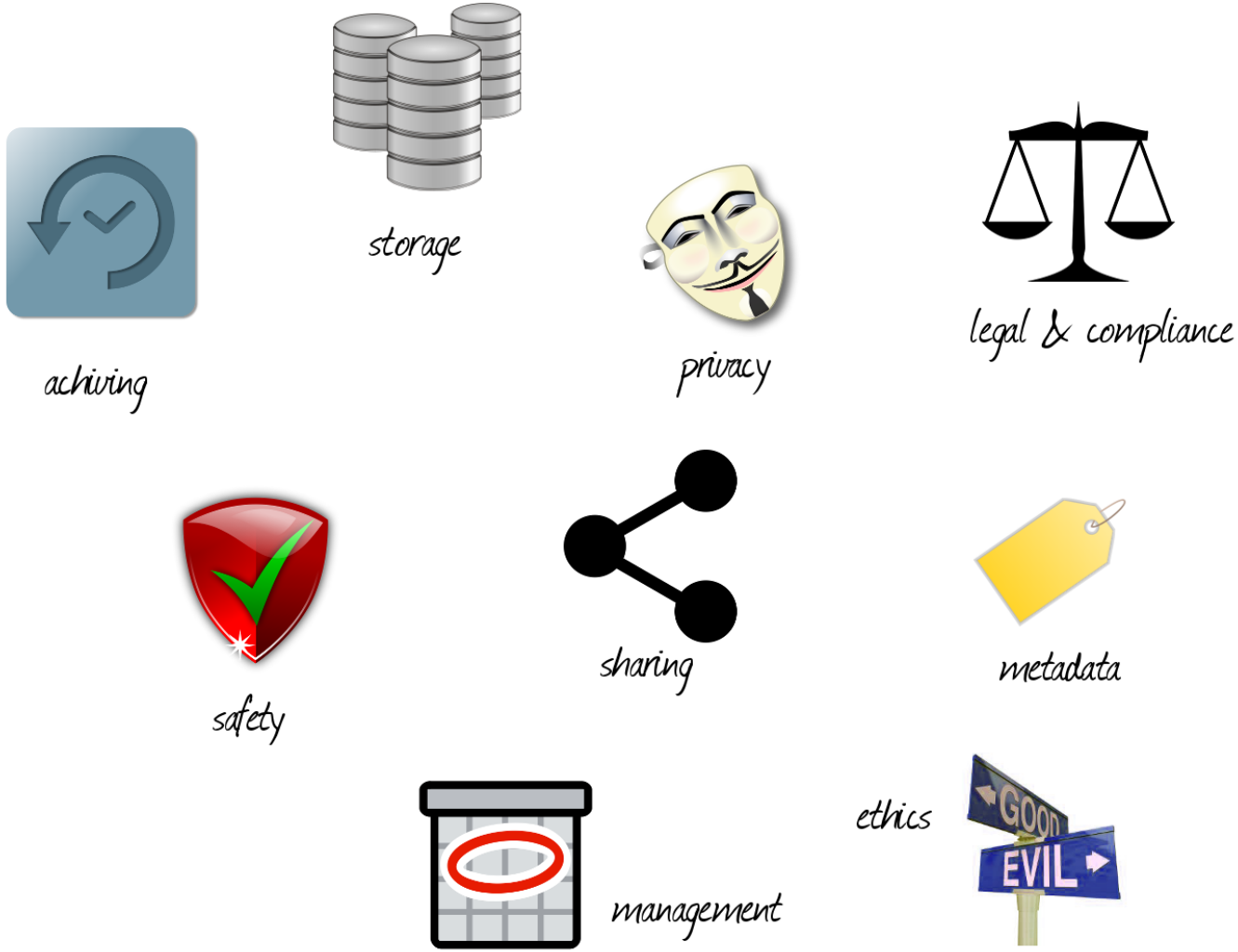
"The Web is Agreement" cropped, by Paul Downey, CC-BY 2.0

CC BY



Understanding a database schema.

by Eric, Sql Designer, CC-BY-SA 2.0



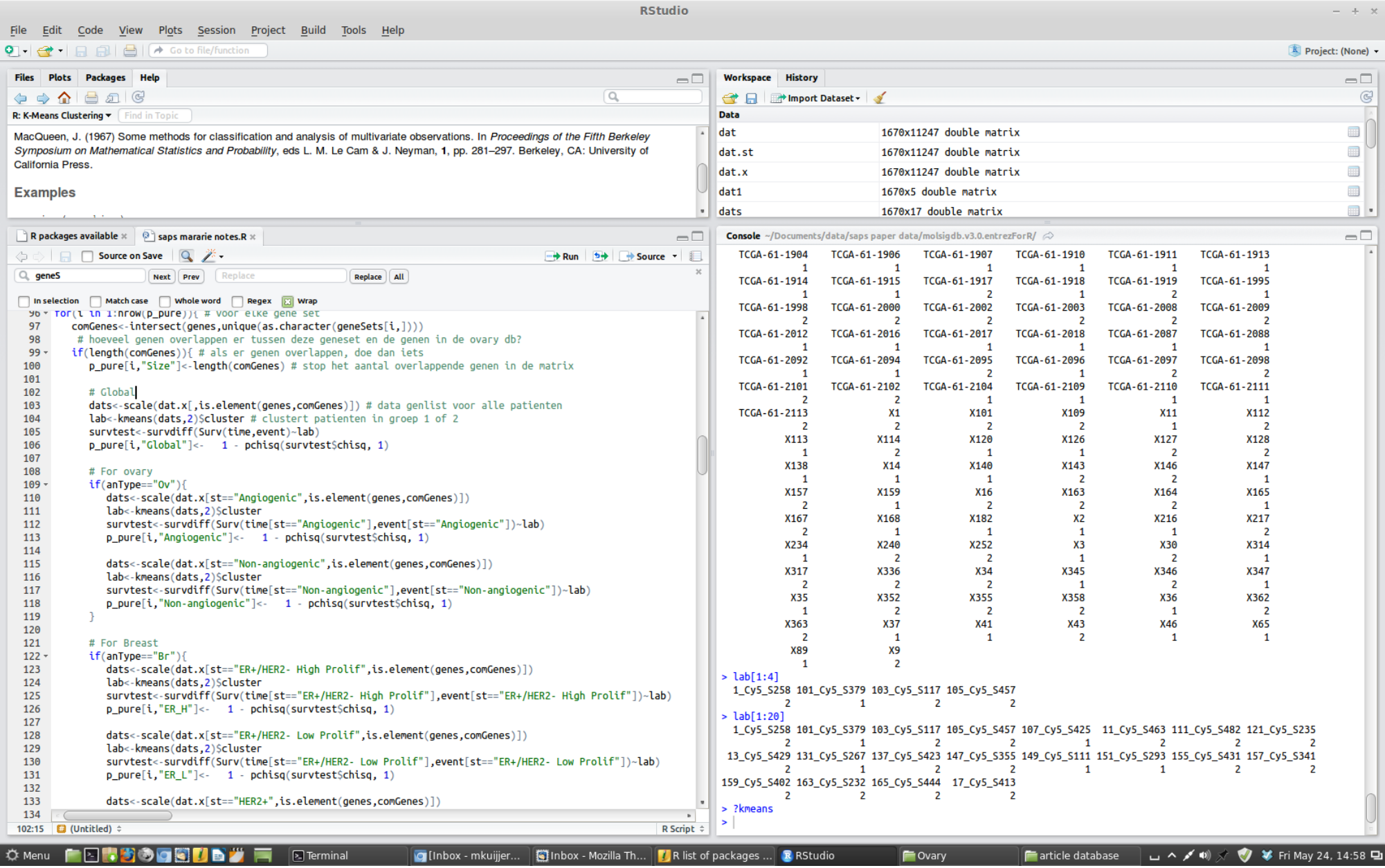
Governance cares for the data and its subjects.

icons from by Openclipart.org, public domain;

Good and Evil by AJC ajcann.wordpress.com, CC-BY-SA 2.0

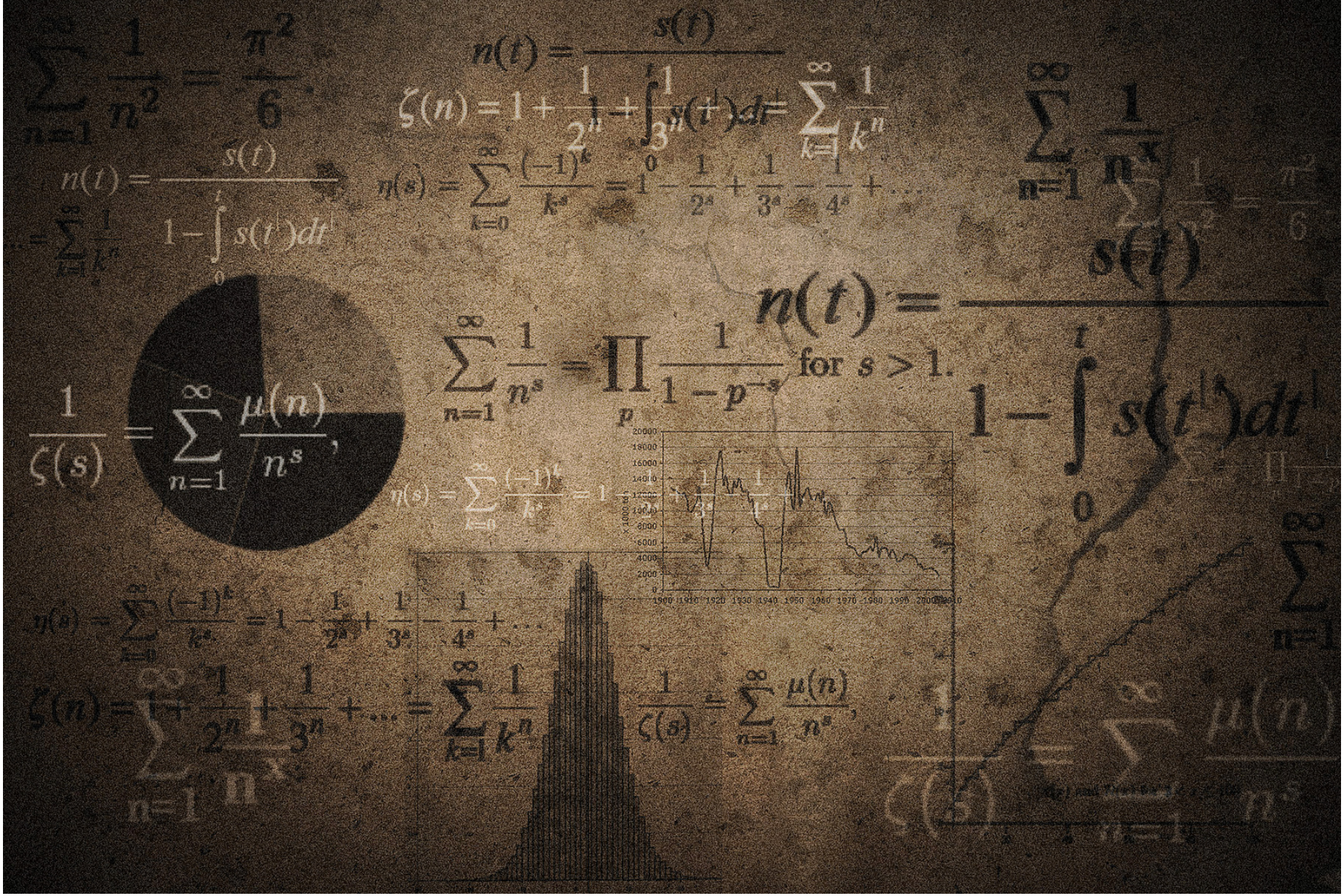
Data engineers make the back-end work

by Intel Free Press, CC-BY 2.0



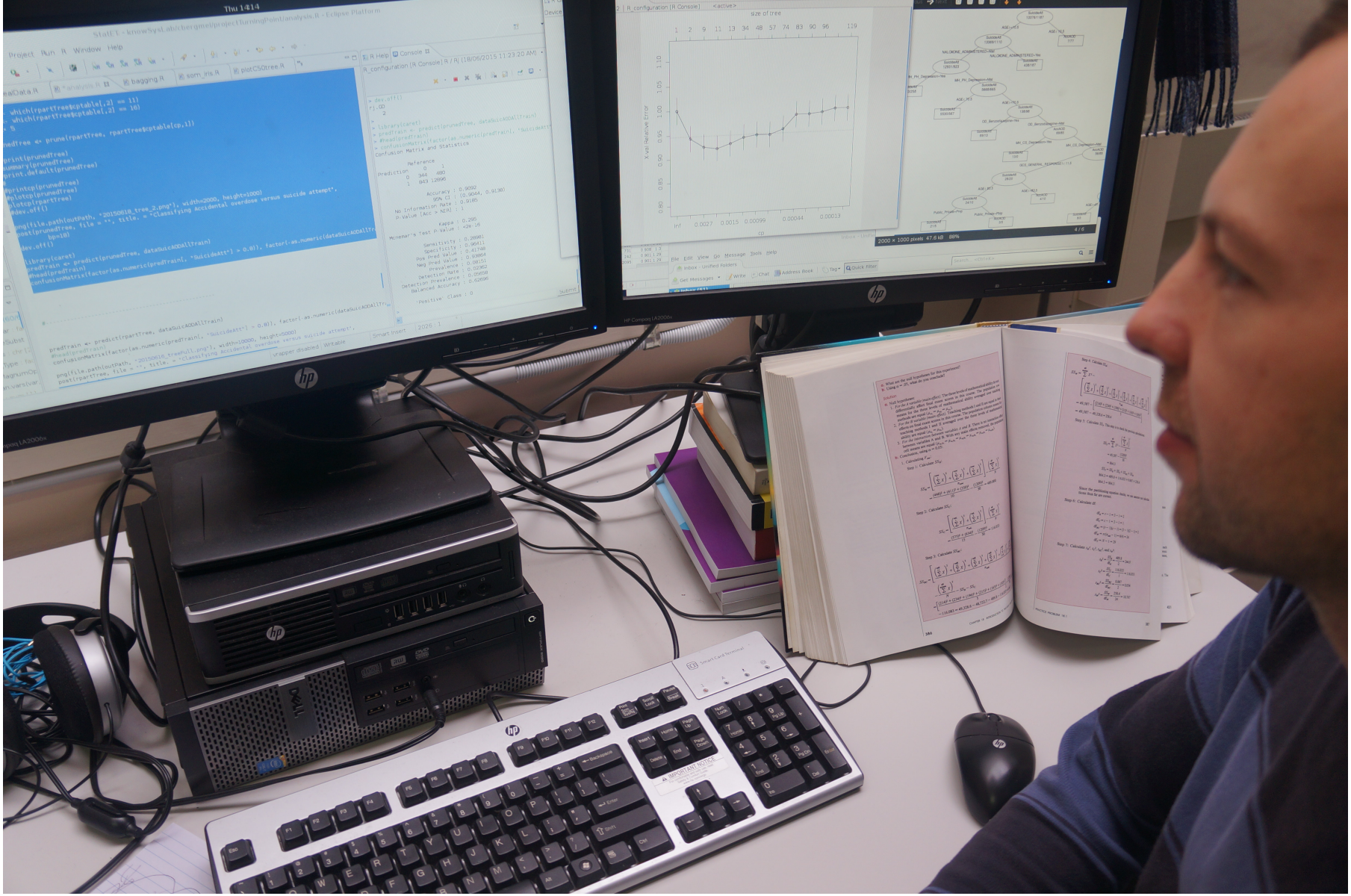
Inspecting and cleaning the data.

"rstudio" by marrarie, CC-BY-SA 2.0

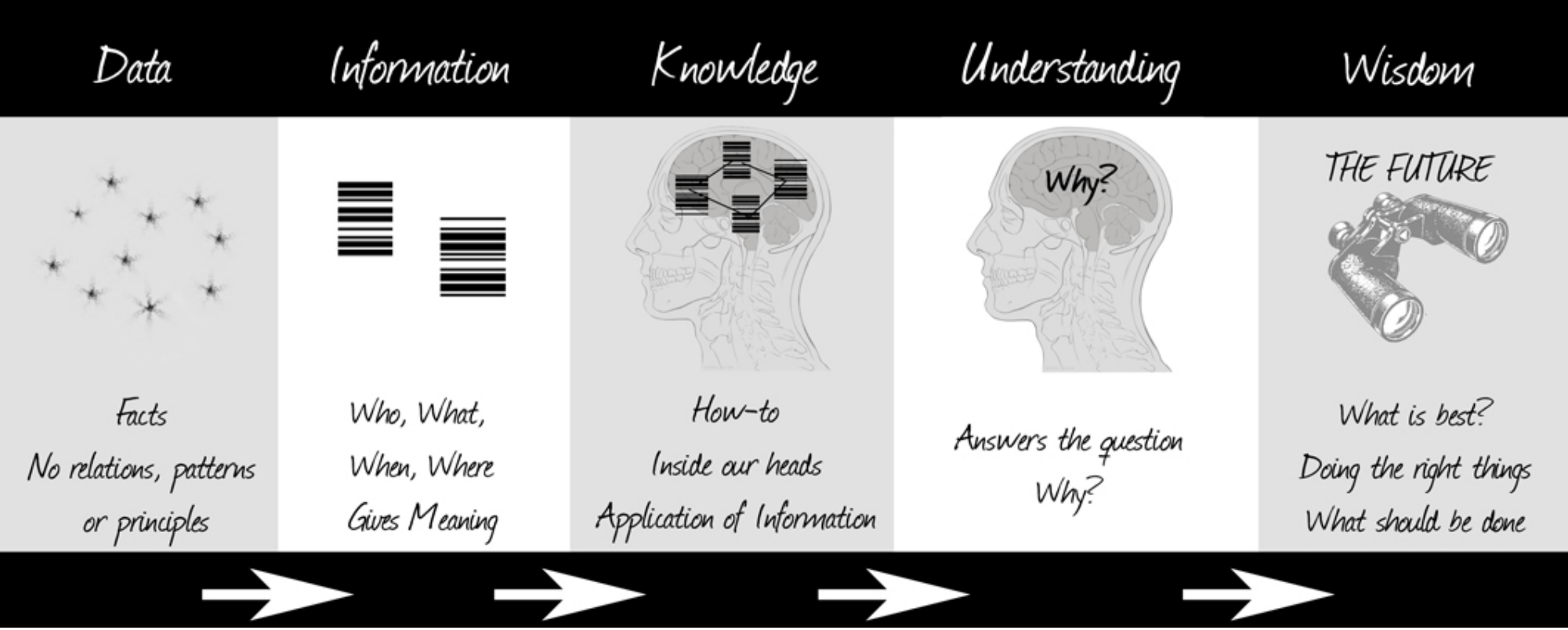


Proposing a conceptual/mathematical/functional model.

"Mathematics" by Tom Brown, CC-BY 2.0



Analyst building models with his favourite tool.



Analysis, statistics and/or machine learning works on the data.

"From Data to Wisdom" by Nick Webb, CC-BY 2.0

Choosing appropriate visualizations for the data. Many different options exist!

"Visualization Matrix" cropped, by Lauren Manning, CC-BY 2.0

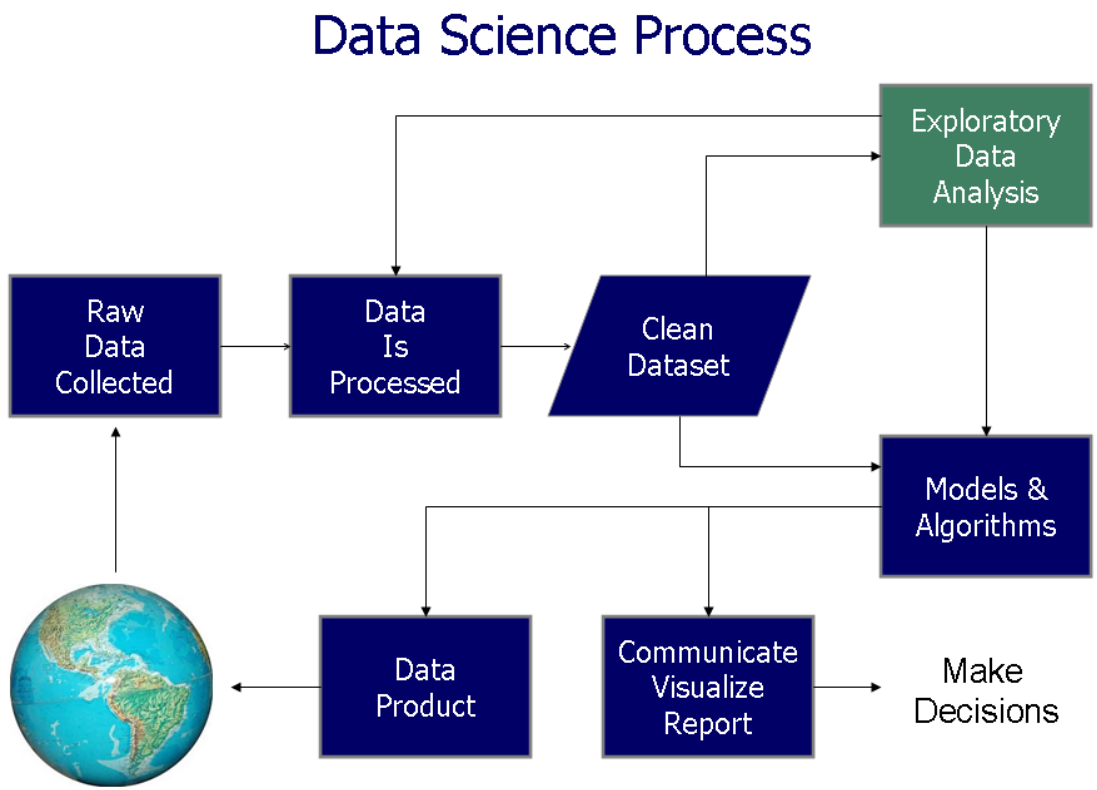
CC BY

Visualising data to interpret it and present results.

by Stephen Ausmus acquired from USDA ARS, public domain.

The Data Science Process

- A flowchart showing the overall data science process



by Farcaster, CC-BY-SA 3.0

Operationalization: putting the results to work.

"Illustration of Strategy“ by Denis Fadeev, CC-BY-SA 3.0

The Data Science Process:
Our Standard Value Chain

our model of the process

Parts of a Data Science Project

Collection: gathering the data

Engineering: storage and computational resources across full lifecycle

Governance: overall management of data across full lifecycle

Wrangling: data preprocessing, cleaning

Analysis: discovery (learning, visualisation, *etc.*)

Presentation: arguing the case that the results are significant and useful

Operationalisation: putting the results to work, so as to gain benefits or value

We call this the Standard Value Chain.

Interpretations of the Parts

Following *Jeff Hammerbacher's* UC Berkeley 2012 course notes, we will interpret these four entities:

- business analyst
- programmer
- enterprise
- web company

Interpretations: the Business Analyst

Collection: copy and paste into Excel

Engineering: use Excel to store and retrieve

Wrangling: use Excel functions, VBA

Analysis: charts

Interpretations: the Programmer

Collection: web APIs, scraping, database queries

Engineering: flat files

Wrangling: Python and Perl, *etc.*

Analysis: Matplotlib in Python, R

Interpretations: the Enterprise

Collection:	application databases, intranet files, server logs
Engineering:	Teradata, Oracle, MS SQL Server
Wrangling:	Talend, Informatica
Analysis:	Cognos, Business Objects, SAS, SPSS

Interpretations: the Web Company

Collection:	application databases, server logs, crawl data
Engineering:	Hadoop/Hive, Flume, HBase
Wrangling:	Pig, Oozie
Analysis:	dashboards, R

What is Data Science?
(ePub section 1.3)

how can we define or circumscribe data science?

Wikipedia Definitions

Data Science is the extraction of knowledge from data, which is a continuation of the field of data mining and predictive analytics.

Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate.

Pivotal Definition

Data Science: 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.

NIST Big Data Working Group Definition

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.

Journal of Data Science
Definition

Data Science is almost everything that has something to do with data: collecting, analyzing, modeling. . . . . yet the most important part is its applications — all sorts of applications.

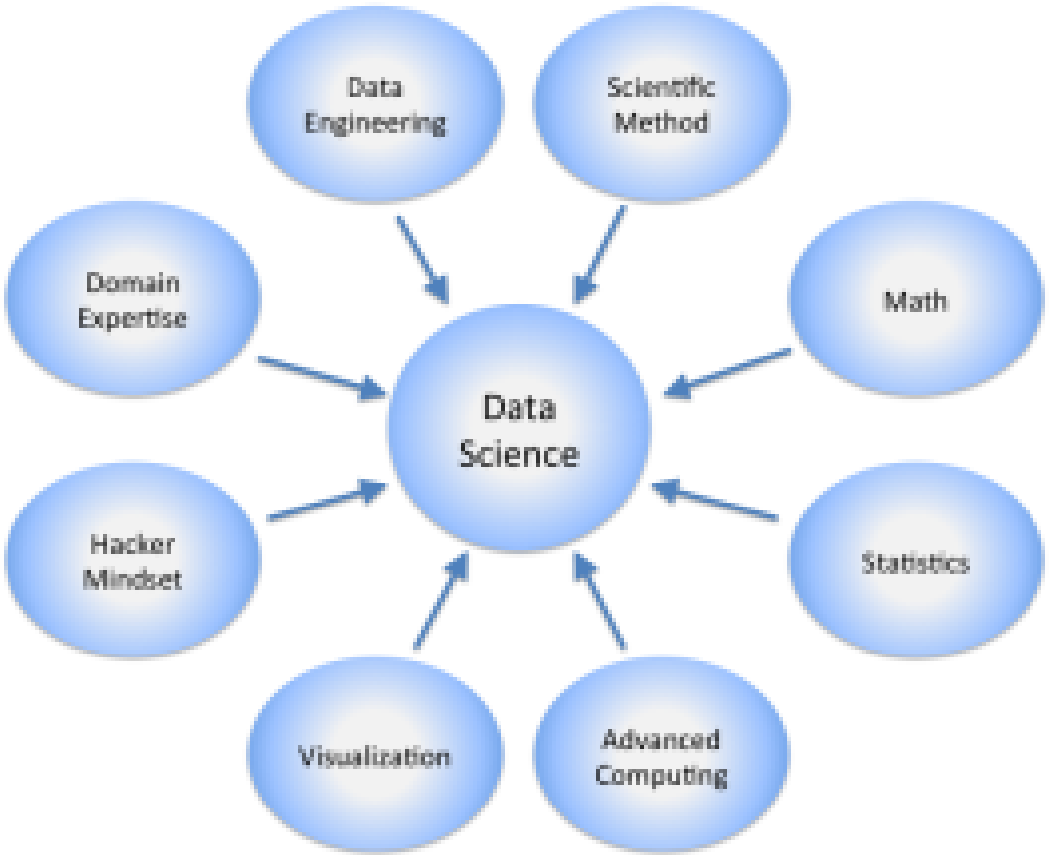
Definitions: Summary

narrow: “machine learning on big data”

broad: extraction of knowledge/value from data “through the complete data lifecycle process”

- broad concern with the different stages
- focus on the learning/knowledge discovery

Data Science Relationships



Related: Data Engineering

building scalable systems for storage, processing data

- *e.g.* Amazon Web Services, Teradata, Hadoop, ...
- databases, distributed processing, datalakes, cloud computing, GPUs, wrangling, ...
- huge, continuous improvement ....

Related: Data Analysis

performing analysis and understanding results

- *e.g.* R, Tableau, Weka, Microsoft Azure Machine Learning, ...
- machine learning, computational statistics, visualisation, ...
- huge, continuous improvement ....

Related: Data Management

managing data through its lifecycle

- *e.g.* ANDS, Talend, Master Data Management, ...
- ethics, privacy, providence, curation, backup, governance, ...
- huge, continuous improvement ....

Fits and Starts

- Data Analysis (John Tukey) in 1962
- Expert Systems in the 1980's
- Machine Learning in the 1980's
- Data Mining in the 1990's
- see *Business Week's "Database Marketing"* cover story September 1994

Data Science Emerges $\sim$2000

- data analysis came of age 1990's
- William Cleveland publishes in 2001 *"Data Science: An Action Plan for ... the field of Statistics"*
- data engineering came of age 2000's (Dot.Com boom)
- (digital) data management came of age 2000's (Dot.Com boom)
- the data/information society
- business pressure on decision making
- "data" as a valuable asset
- Dot.Com companies show the way

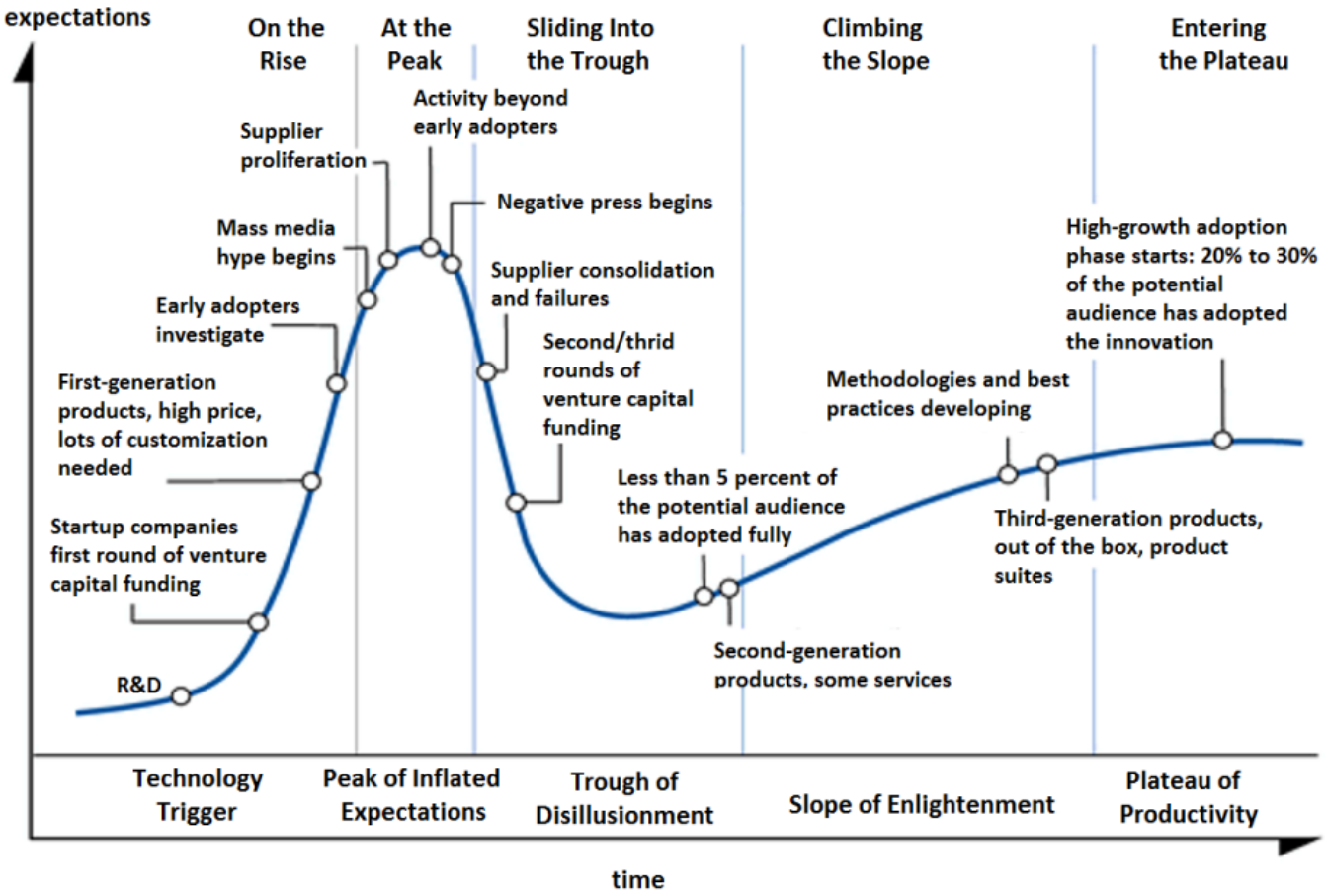
see also David Donoho's *"50 years of Data Science"* (PDF paper)

Homework

- look up Dot.Com boom in Wikipedia

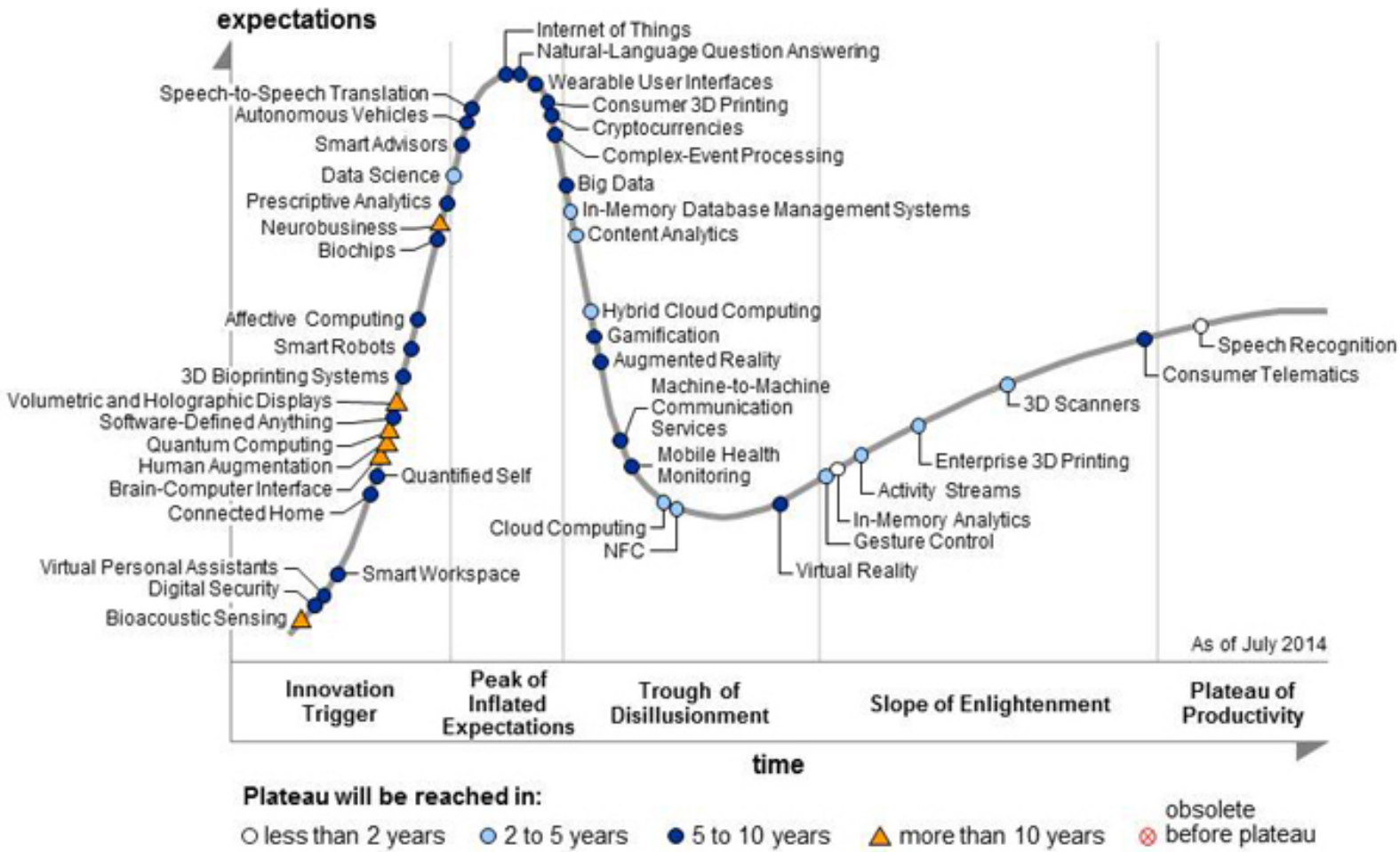
The Hype Cycle

- Gartner's Hype Cycle© attempts to quantify the level of maturity of various technologies:



Hype Cycle for 2014

- Can you spot Data Science?



Career as Data Scientist

Data science currently provides a strong career path.

To become a specialist you need:

- solid machine learning and statistics
- related mathematics (1st+2nd year in many degrees)
- solid programming/prototyping (Python, Java)
- Unix experience (Linux, Mac OSX)

See also:

- the infamous Metromap: *becoming a data scientist*
- and *Modern Data Scientist* on Pinterest

This unit provides an introduction and background (but not the core analyst skills) to become a data scientist

Data Science Research

Data Science is seeing major growth at universities internationally

Many research programs exist, including:

- US National Institute of Standards' Big Data Working Group (2013-2015)
- US National Academy of Sciences' Committee on the Analysis of Massive Data (2013)
- Alan Turing Institute for Data Science at London's new Knowledge Quarter (near National Library, 2016-)

End of Week 1