FIT1043: Introduction to Data Science

Module 1:
Data Science and Data in Society

Lecture 2: (Job) Roles, and the Impact

Monash University

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

Aside: Visualising Big Data

extract from *"Turning powerful stats into art"* by Chris Jordan, starting at minute 1:00

Homework

From Section 1.1:

- read Cukier's *Foreign Affairs* article
- view Cukier's TED talk, "Big Data is Better Data"
- read "What is Data Science?" the O'Reilly pamphlet
- view the CERN video, "Big Data" from Tim Smith

Roles of a Data Scientist (ePub section 1.4)

better understanding the different kinds of data scientists:

- reviewing different writings:
  - from *What is Data Science?* from O'Reilly
  - from *Doing Data Science* from Schutt and O'Neil
  - from *Analyzing the Analyzers* from Harris, Murphy and Vaisman
- interviews
  - from *Data Analytics Handbook*

Roles of a Data Scientist: Reviewing *What is Data Science?*

O'Reilly pamphlet describing what one does

What is Data Science? (O'Reilly)

Outline of *the document*.

1. Introduction.
2. Where data comes from: Moore's law and data wrangling
3. Working with data at scale: big data processing and the role of statistical inference
4. Making data tell its story: visualisation tools
5. Data Scientists: what it is they do

NB. has lots of jargon and tech talk, so would be difficult to read!

Moore's law ::= computer hardware memory/CPU power increases exponentially

Jargon

Your analysis of jargon and sentences:

General: foreclosure data, sugar coat, red herring, putting lipstick on a pig, take it for granted

Tech talk: data mashups, Beautiful Soup, Mechanical Turk, CDDB database, PageRank, citizen science screen scraping, awk, Web 2.0, knowledge discovery

Data Science: Beautiful Soup, NoSQL, BigTable, Amazon's Dynamo, Cassandra, Hbase, Map Reduce, EC2 clusters, Hadoop

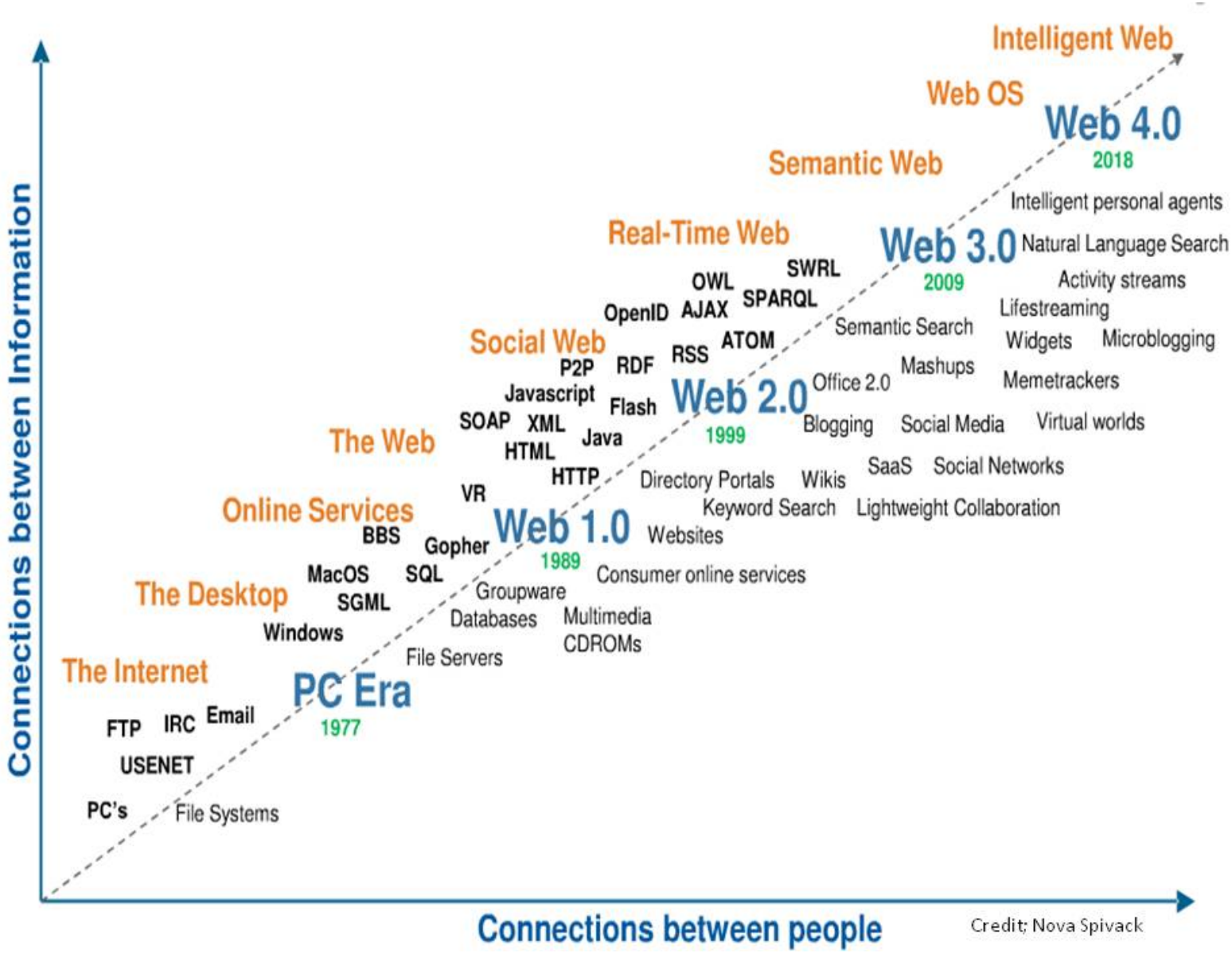
Author's (or colleague's) imagination: data exhaust, data conditioning, dictatorship of data, data jiujitsu, wild data

Phrases

- “common to mashup data from a variety of sources”
- “While we aren’t drowning in a sea of data, we’re finding that almost everything can (or has) been instrumented.”
- “what differentiates data science from statistics is that data science is a holistic approach”
- “Precision has an allure, but in most data-driven applications outside of finance, that allure is deceptive. Most data analysis is comparative ...”
- “In all countries, but particularly in nondemocratic ones, big data exacerbates the existing asymmetry of power between the state and the people.” (from Cukier article)
- "Information is the oil of the 21st century, and analytics is the combustion engine.” (from somewhere else)
- “Much of the data we currently work with is the direct consequence of Web 2.0 ...”

Web X.0

(credit: Nova Spivack)



Phrases (cont.)

- “Data expands to fill the space you have to store it.”
- “‘big data’ is when the size of the data itself becomes part of the problem”
- “statistics is the grammar of data science”
- “If you want to find out just how bad your data is, try plotting it.”
- “the best data scientists tend to be ‘hard scientists,’ particularly physicists, rather than computer science majors” (from DJ Patil)
- “Entrepreneurship is another piece of the puzzle ... they’re all trying to build new products.”
- “The future belongs to companies who figure out how to collect and use data successfully.”

From *What is Data Science?*

"Data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others."

"Data scientists combine entrepreneurship with patience, the willingness to build data products incrementally, the ability to explore, and the ability to iterate over a solution."

"At O'Reilly, we frequently combine publishing industry data from Nielsen BookScan with our own sales data, publicly available Amazon data, and even job data to see what's happening in the publishing industry."

i.e., data mashup, web crawling, text mining, visualisation, ...

From *What is Data Science?*

A quote from *Jeff Hammerbacher*

> *... on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization ...*

hypothesis test ::= statistical test to evaluate a simple claim

regression analysis ::= fitting a curve to real valued data

Hadoop ::= system for partitioning computation across a compute cluster

From *What is Data Science?*

A quote from *Hal Varian*

> *The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades.*

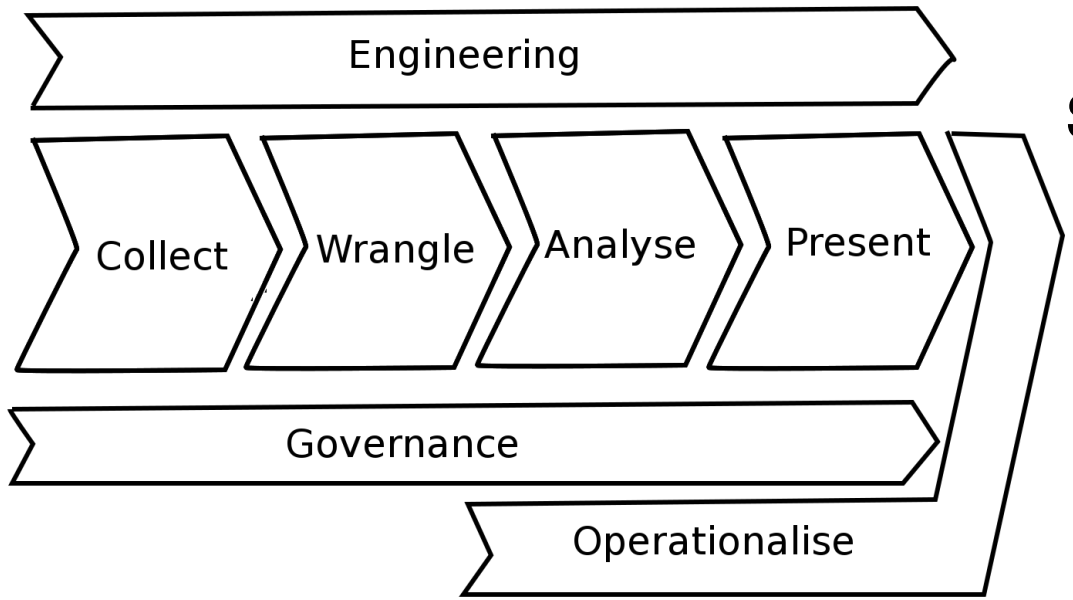
Roles of a Data Scientist: Reviewing *Doing Data Science*

Schutt and O'Neil taught an early course at Columbia U. in NYC

From *Doing Data Science*

Schutt and O'Neil, 2013, Chapter 1, available digitally through library

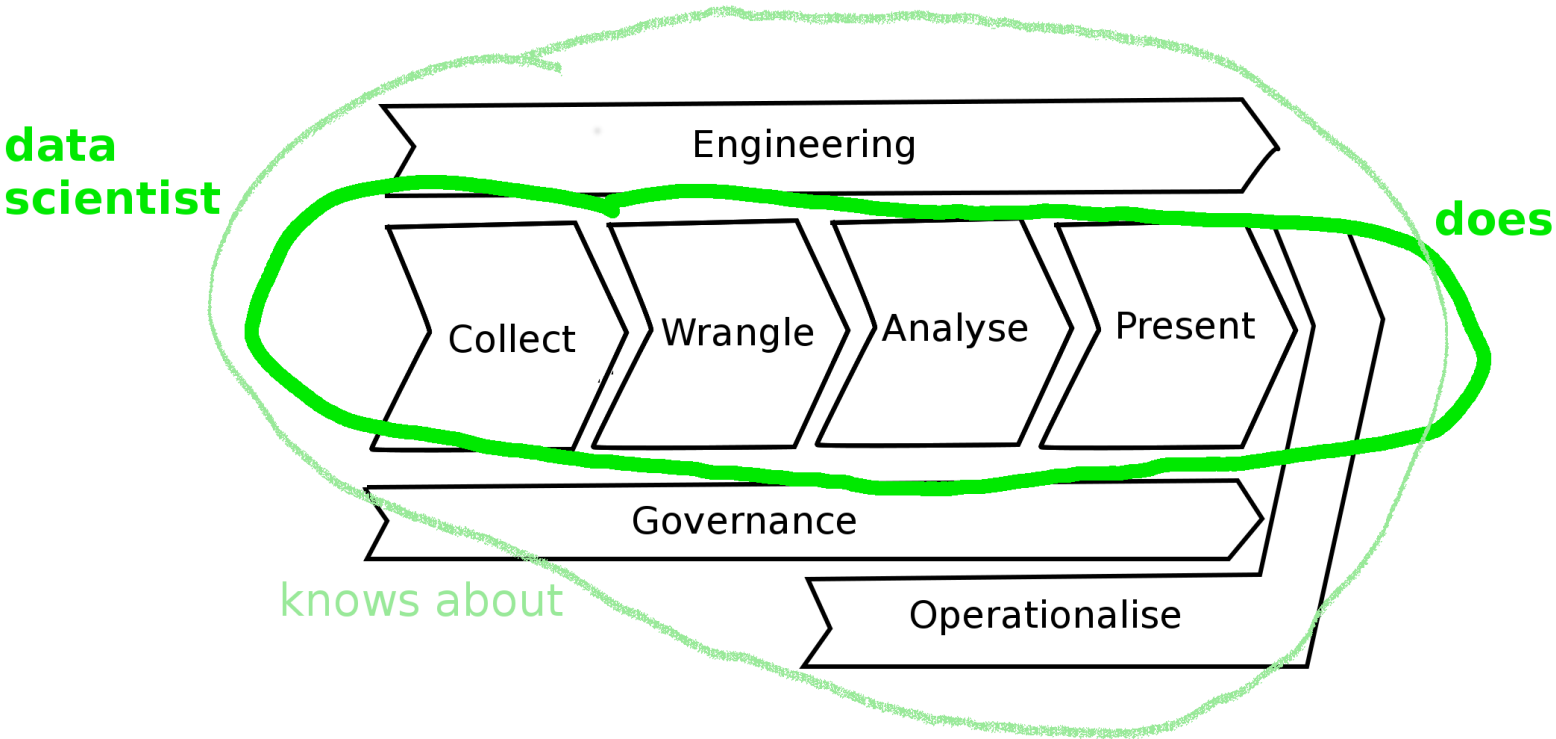
Jobs: require operating the Collect through to Presentation stages of the Standard Value Chain, as well as knowledge of the remaining stages and the domain.



Standard Value Chain

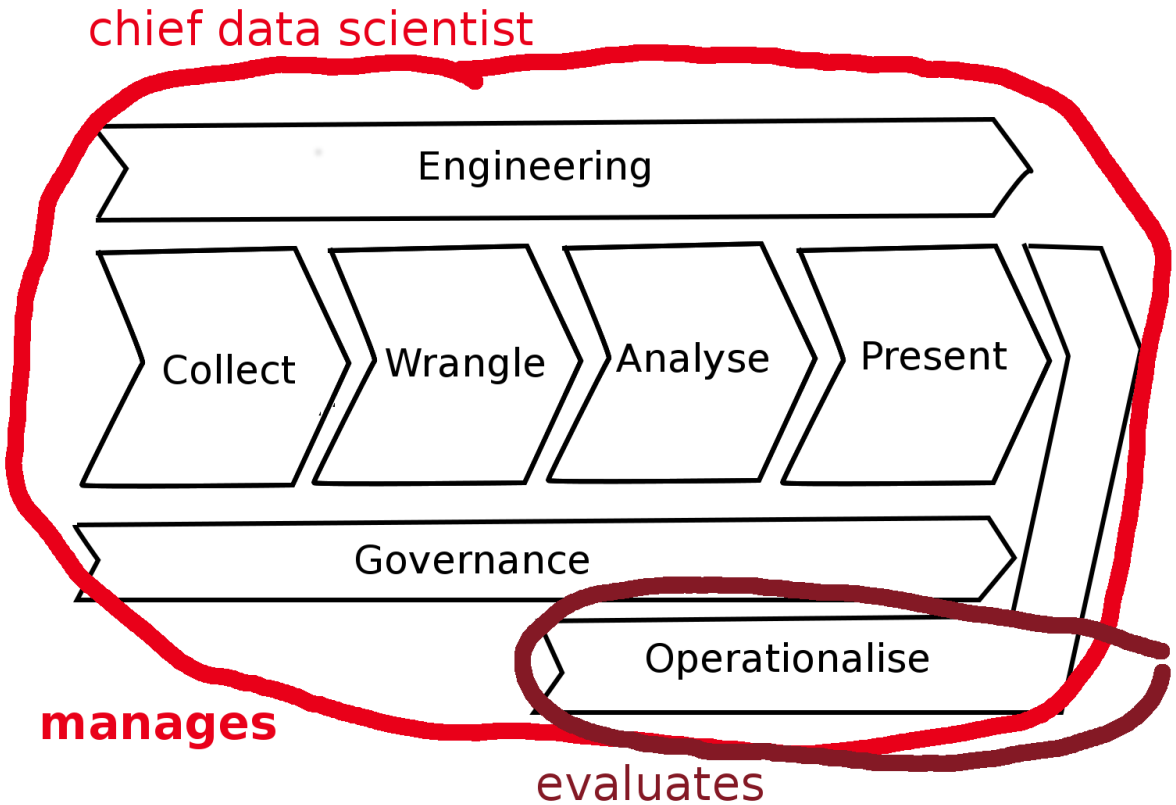
Profile: typical data scientist has a different mix of skills as well as domain knowledge

Doing Data Science (cont.)



Data scientist ::= addresses the data science process to extract meaning/value from data

Doing Data Science (cont.)



Chief data scientist: a form of chief scientist who addresses data management, data engineering and data science goals.

chief scientist ::= corporate position, responsible for science related aspects of a company/organisation

Roles of a Data Scientist: Reviewing *Analyzing the Analyzers*

text mining on data scientists

Skills of Data Scientists

Analyzing the Analyzers, Harris, Murphy and Vaisman, 2013

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

NB. typical data scientist doesn't have to know all of these!

Different styles† of Data Scientists

Analyzing the Analyzers, Harris, Murphy and Vaisman, 2013

Data developers: people focused on the technical problem of managing data – how to get it, store it, and learn from it

Data researchers: people with an academic research background, using their training to "understand complex processes"

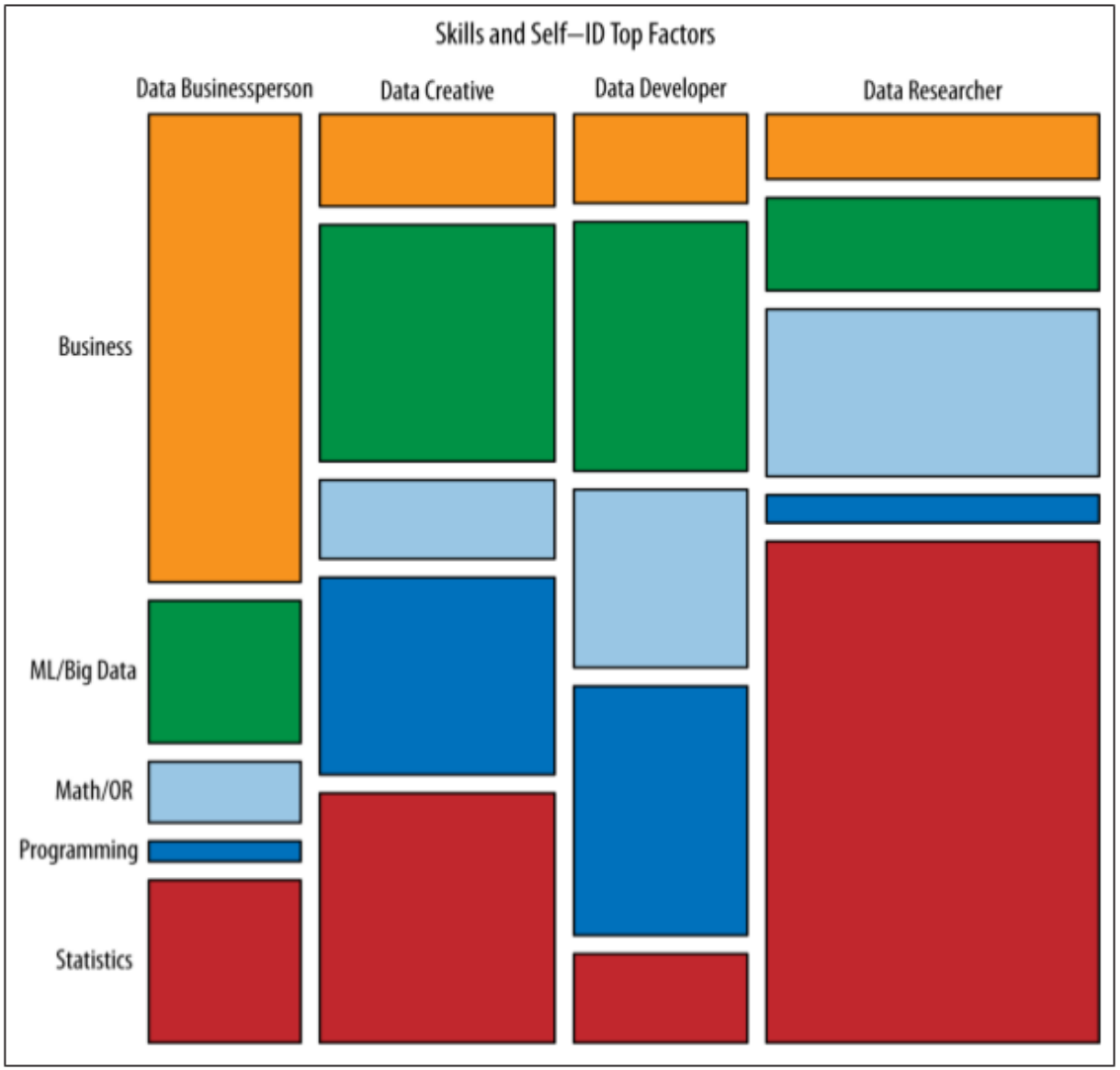
Data businesspersons: most focused on the organization and how data projects yield profit

Data creatives: the broadest of data scientists, those who excel at applying a wide range of tools and technologies to a problem

†data scientists tend to identify with one of these styles

Mapping Styles to Skills

Analyzing the Analyzers, Harris, Murphy and Vaisman, 2013



X-axis:
different roles

Y-axis:
different skills

which might you be?

Homework

From Section 1.4:

- view the 3 Monash videos under “Interviews with Industry Professionals”

Roles of a Data Scientist: Interviews from *Data Analytics Handbook*

a variety of firsthand accounts

From *Data Analytics Handbook*

This pamphlet an interview project by three UC Berkeley graduates:

- What exactly do the sexy “data scientists” do?
- What other professions are there in big data?
- What tools do they use to accomplish their tasks?
- How can I enter the industry if I don’t have a Ph.D. in Statistics?

Discussion in Class

From *Data Analytics Handbook* read the interviews of

- Abraham Cabangbang (2 pp)
- Ben Bregman (2 pp)
- Leon Rudyak (3 pp)

Lessons listed.

Lessons: I

Communication skills are underrated.

Lessons: II

The biggest challenge for a data analyst is the Collection and Wrangling steps.

Lessons: III

A data scientist is better at statistics than a software engineer and better at software engineering than a statistician.

Lessons: IV

The data industry is still nascent and the roles less well defined so you get to interact with many parts of the company from engineering to business intelligence to product managers.

Lessons: V

Keep a curiosity about working with data, a quality as important as your technical abilities.

Impact of Data Science (ePub section 1.6)

some examples of how data science is impacting others:

- your life in the cloud
  - datafication of you
- science and social good
  - scientific method holds true, but broadens technology
- futurology
  - healthcare and automobiles

Impact of Data Science:
Your life in the cloud

datafication of you

Your Life on the Cloud

From *Year Zero: Our life timelines begin*

Our personal information is increasingly stored in the cloud (though perhaps behind firewalls): social life (Facebook), career (LinkedIn), search history (Google, *etc.*), health and medical (Fitbit, TBD), music (Apple), ...

Many, many advantages:

e.g. personal agents

- computerised support for health
- ...

But some disadvantages:

e.g. security and privacy breaches

- ...

please make some suggestions on this form, *results*

Your Life on the Cloud (cont.)

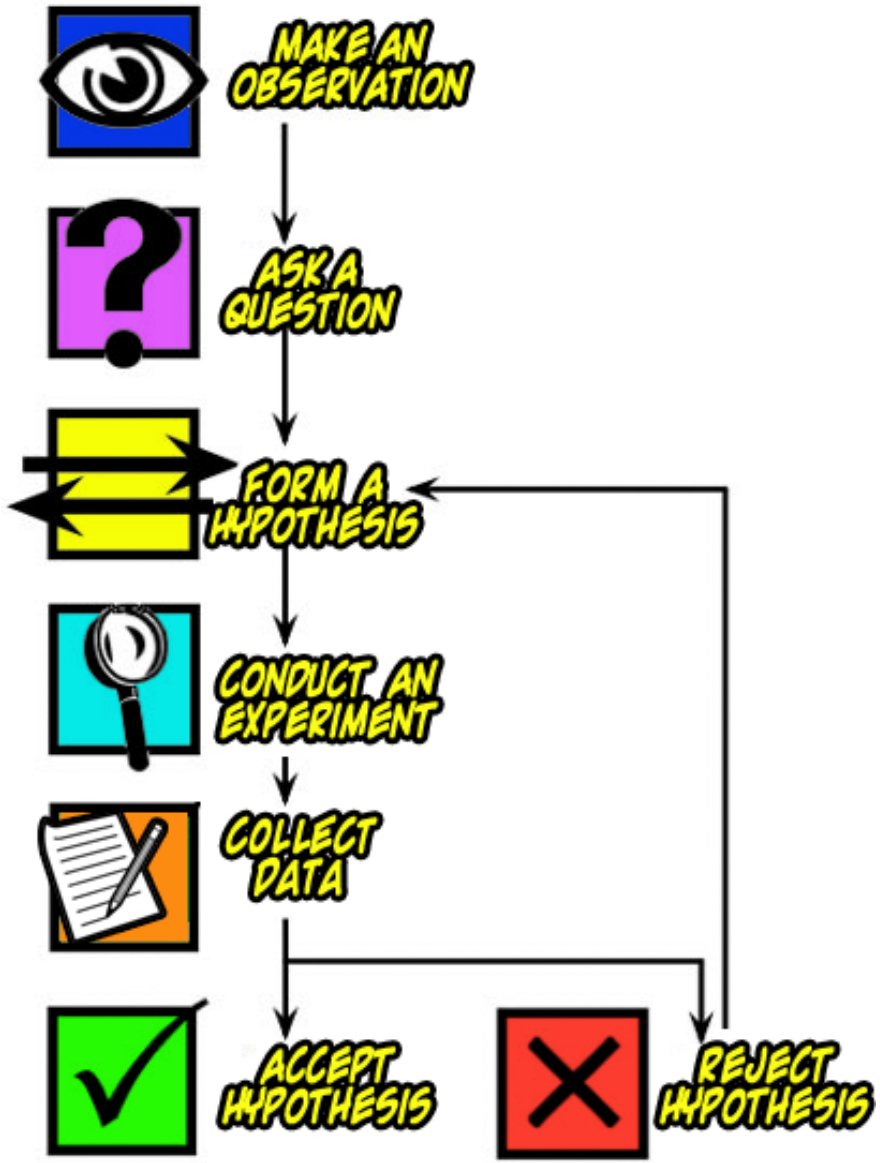
But

- corporate leakage to government (security, tax, *etc.*)
- what if you don't have rights to access/delete data?
- security and privacy breaches
- what if we've changed our ways?
- the department of pre-crime
- corporate mergers
- "the science is settled" and government mandates

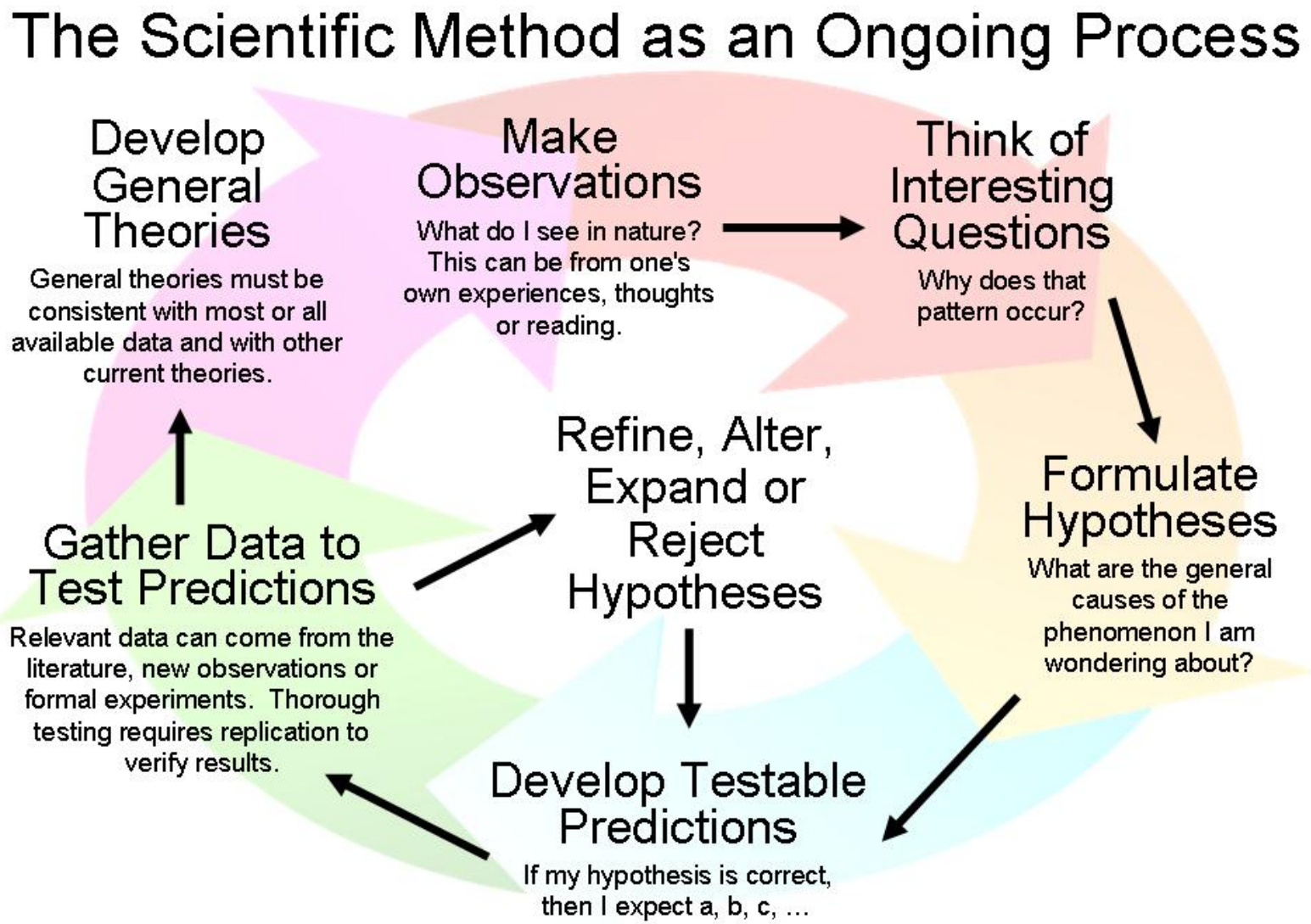
Impact of Data Science: Science

scientific method holds true, but broadens technology

The Scientific Method



The Scientific Method



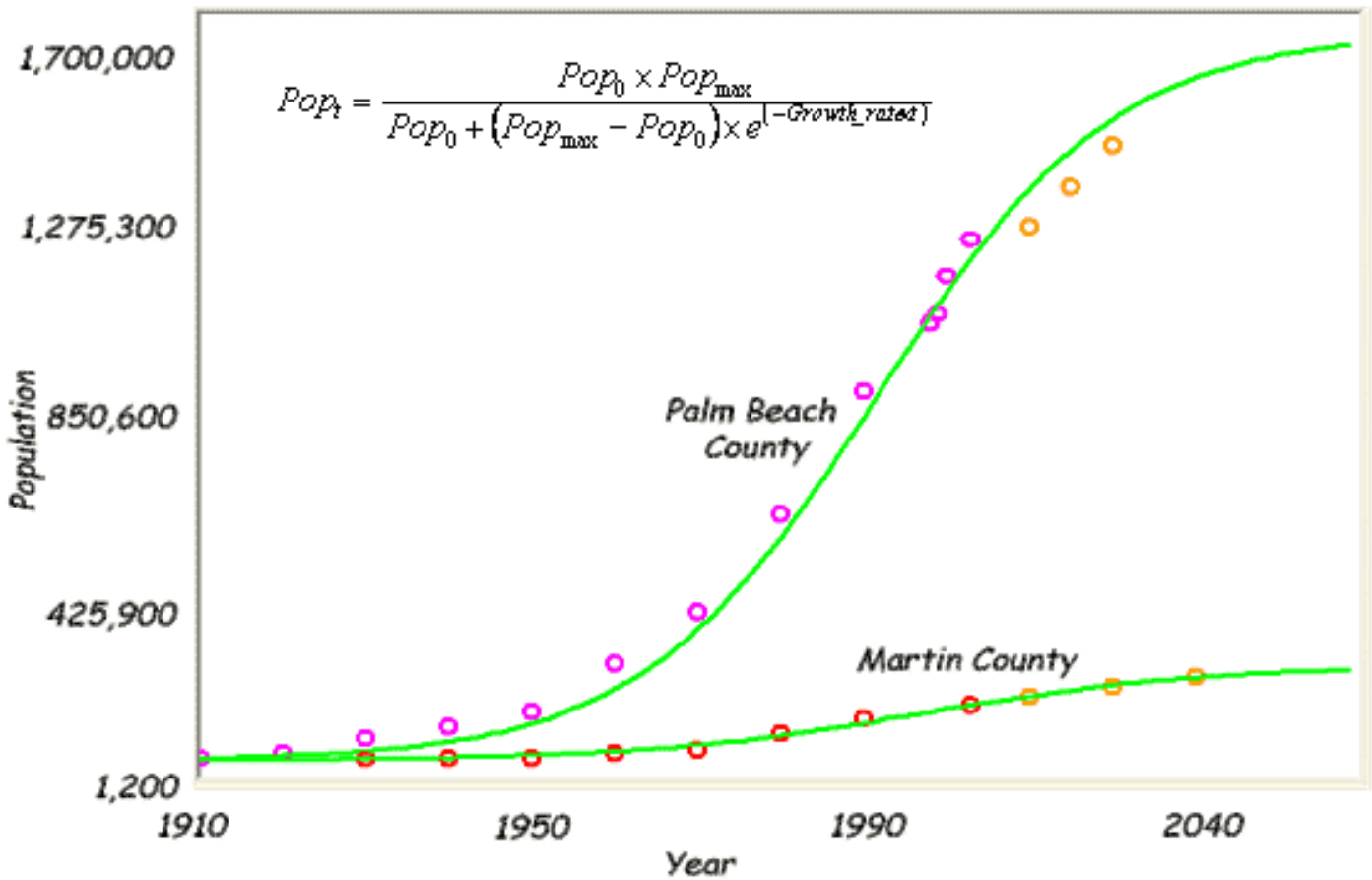
from Wikipedia *Scientific method*

The End of Theory

Chris Anderson's blog in Wired 23/05/2008

A Model of Population Growth

From *Integrating Urban Growth Models*, Pearlstine, Mazzotti, Pearlstine and Mann, 2004



A Complex Model of Obesity

Obesity Systems Map (from `Shiftn.com`)

The End of Theory (cont.)

Science is largely driven by labourious studies to find complex causal models, sometimes using *reductionism*. The intent is to find an explanation that can be used for future prediction.

Chris Anderson (Editor-in-chief of *Wired* magazine) says:

> *Google's founding philosophy is that we don't know why this page is better than that one: If the statistics of incoming links say it is, that's good enough. No semantic or causal analysis is required.*
>
> *...*
>
> *Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen ...*

The End of Theory (cont.)

Chris Anderson sums up:

> *The new availability of huge amounts of data [...] offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.*

The End of Theory (cont.)

Philosopher *Massimo Pigliucci* says:

> *But, if we stop looking for models and hypotheses, are we still really doing science? Science, unlike advertizing, is not about finding patterns–...–it is about finding explanations for those patterns.*
>
> *...*
>
> *science advances only if it can provide explanations*

Data scientist Drew Conway says in some areas the data doesn't exist.

Statistician *Andrew Gelman* says:

> *... you'll still have to worry about ... all the ... reasons why people say things like, "correlation is not causation" and "the future is different from the past."*

Data Science for Science

- fields like physics, bioinformatics and earth science used big data anyway
  - had their own independent data science revolution
- in other areas has raised the profile of data-driven science
- spurred on governments to develop cross-disciplinary programmes
  - *Alan Turing Institute for Data Science* in the UK
- has provided new data sources and tools for collecting data
  - crowd sourcing
  - social media
- allows for citizen/participatory science
  - *DataONE*

Impact of Data Science:
Social good

scientific method holds true, but broadens technology

Data Science for Social Good

Example:
"Data, Predictions, and Decisions in Support of People and Society"
by Eric Horvitz (Distinguished Scientist & Managing Director at Microsoft) see the final section of video 46:51-53:00 mins.

Interactive website *Aid Data* (making development finance data more accessible).

Data Science for Social Good movement training data scientists to support community and charity.

Impact of Data Science: Futurology

some areas where significant impact is to be made in the future

Health Care Futurology

see “Big data – 2020 vision” talk by SAP manager John Schitka

- your stomach can be instrumented to assess contents, nutrients, *etc.*
- your bloodstream can be instrumented too assess insulin levels, *etc.*
- your “health” dashboard can be online and shared by your GP
- health management organisations (HMO) tying funding levels to patient care performance
- GP/HMO will know about your icecream/beer binge last night and you missing your morning run
- longitudinal studies feasible

Automobile Futurology

see “Big data – 2020 vision” talk by SAP manager John Schitka

Self driving cars:

- how does the city replace traffic fine revenue?
- can you drink and drive if the car is automatic?
- what happens to the taxi industry?
- what happens to the auto insurance industry?
- what happens to people still “slef” driving, and their insurance?

Impact of Data Science: Other

Data Science as Competitive Sport

Kaggle:

- crowd-sourced science and data science
- a platform for competitions about predictive modelling and analytics,
- companies and (application) researchers describe their problem post their data
- statisticians and data miners compete to produce the best models

Browse the site and leader boards to get an idea. See also *Wikipedia on Kaggle*.

crowd-sourcing ::= obtaining services/ideas/content by soliciting contributions from large group of people (usually online)

Infographics

21 Ways How Big Data Will Improve Your Life

Explore more aspects of Data Science via infographics at *Pinterest: data science* and *Pinterest: big data*

Next: Module 2
Data Models in Organisations