FIT1043 Introduction to Data Science Module 1

Data Science and Data in Society

2018 Lecture 2
Monash University

Unit Schedule: Modules

Module	Week	Content
1.	1 2	Overview and look at projects **(Job) roles, and the impact**
2.	3	Data business models / application areas
3.	4 5	Characterizing data and "big" data Data sources and case studies
4.	6 7	Resources and standards Resources case studies
5.	8 9 10	Data analysis theory Regression and decision theory Data analysis process
6.	11 12	Issues in data management GUEST SPEAKER & EXAM INFO.

Who did their homework?

From Section 1.1:

- watch *Cukier's TED talk on "Big Data"*
- watch the video on *"Big Data" by Tim Smith*
- read *"What is Data Science?"* the O'Reilly pamphlet

Remember, the only content you *should* explore further from Alexandria is marked:

Remember, the only content you *need* for the exam is in the lecture.

Aside: Visualising Big Data

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

Roles of a Data Scientist (ePub section 1.5)

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 1:
Reviewing *What is Data Science?*

O'Reilly pamphlet describing what one does

What is Data Science? (O'Reilly)

Quick 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

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

Jargon

Has lots of jargon and tech talk, so would be difficult to read!

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

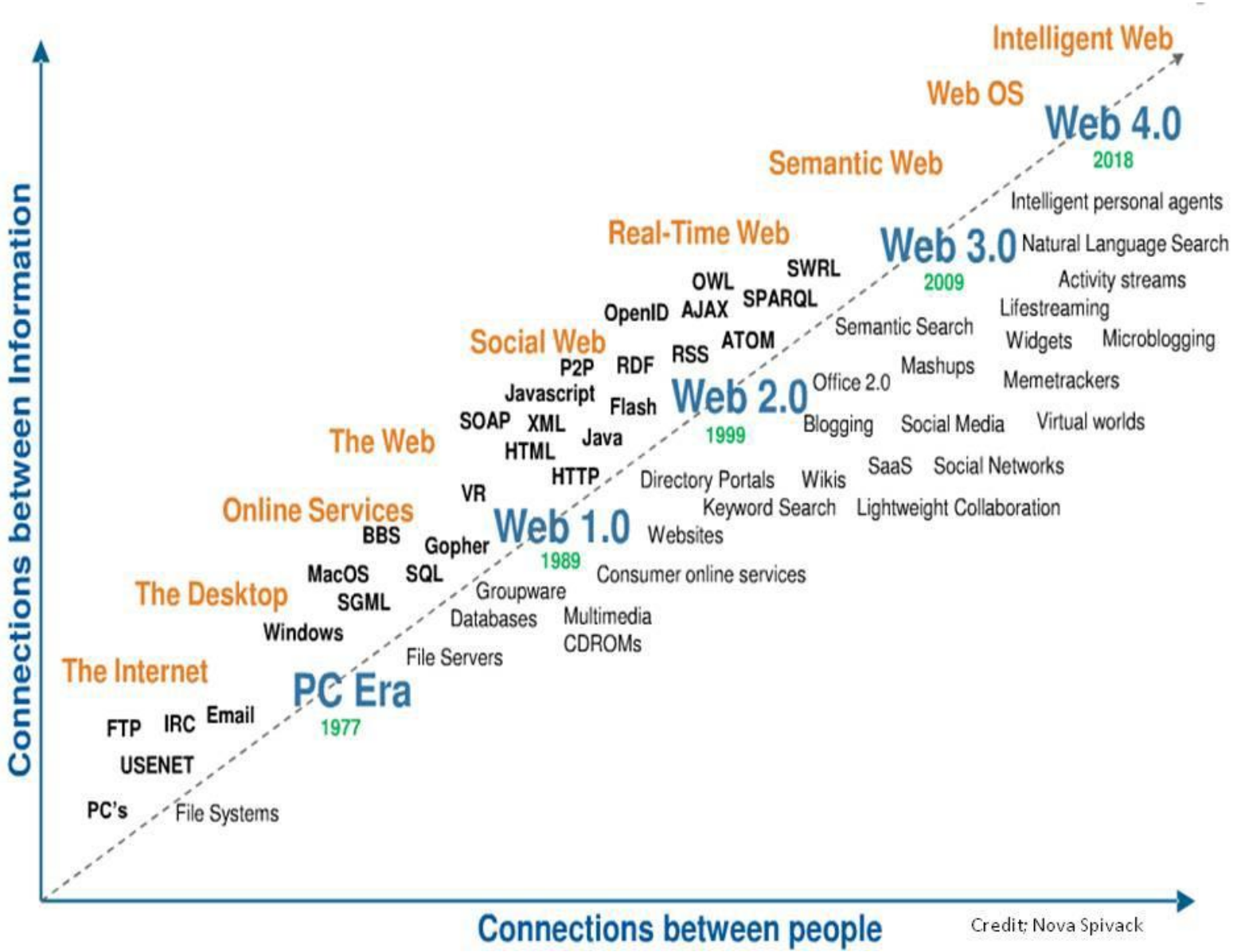
Author's (or colleague's) imagination: data exhaust, data conditioning, dictatorship of data, data jiujitsu, wild data

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 ...”
- "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.”
- “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?*

What do Data Scientist do according to Loukides' article?

"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

What is the Difference Between ...

A *quote from Quora from Jason Widjaja*:

> **Data analysts** *are primarily people who develop insights with data. ....*
>
> **Data scientists** *are primarily people who develop data models and products, that in turn produce insights. ...*
>
> **Data engineers** *are primarily people who manage data infrastructure, automate data processing and deploy models at scale. ...*

(Note the use of the word “primarily”!)

see also *Job Comparison* – Data Scientist vs Data Engineer vs Statistician

MARS Question - **Y66M8T**

Consider the definition given for data science, is the boundary between data science, data engineering and data analysis fixed?

A. TRUE
B. FALSE

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.*

Visualising big data isn't easy but can be very important!

Roles of a Data Scientist 2:

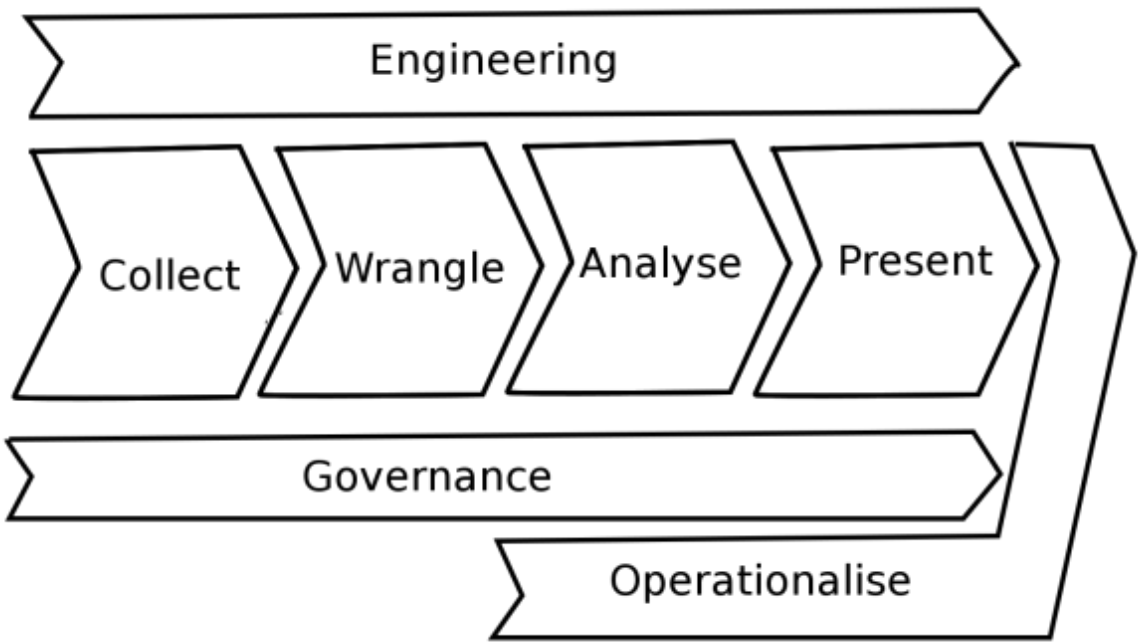
Reviewing *Doing Data Science*

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

Standard Value Chain

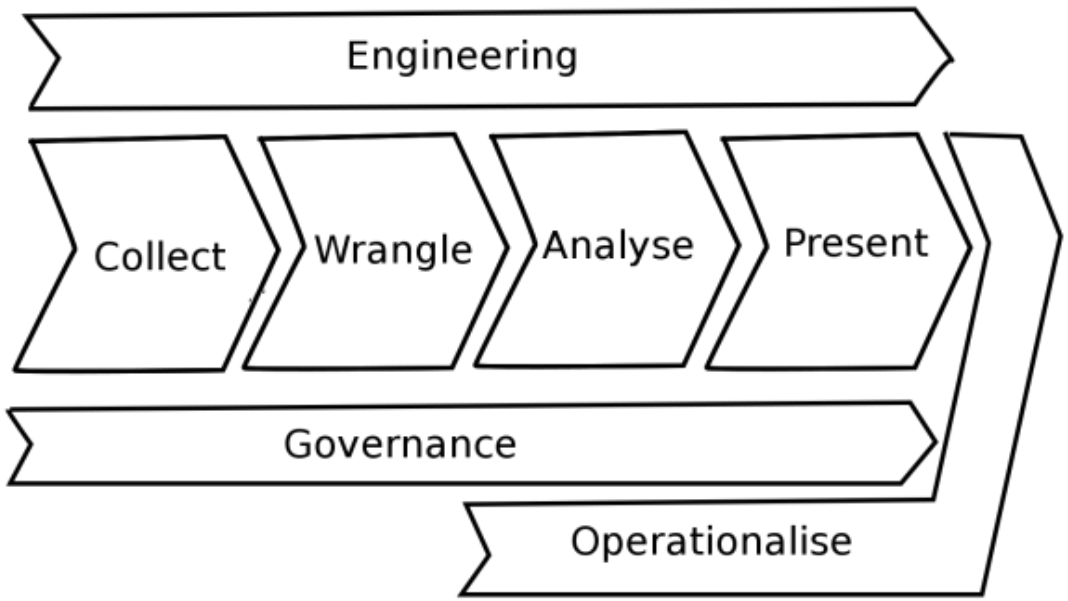
from *Doing Data Science* by Schutt and O'Neil, 2013, (available digitally through library)

Chapter 1 of the book provides the following visualisation of the standard value chain for a data science project:



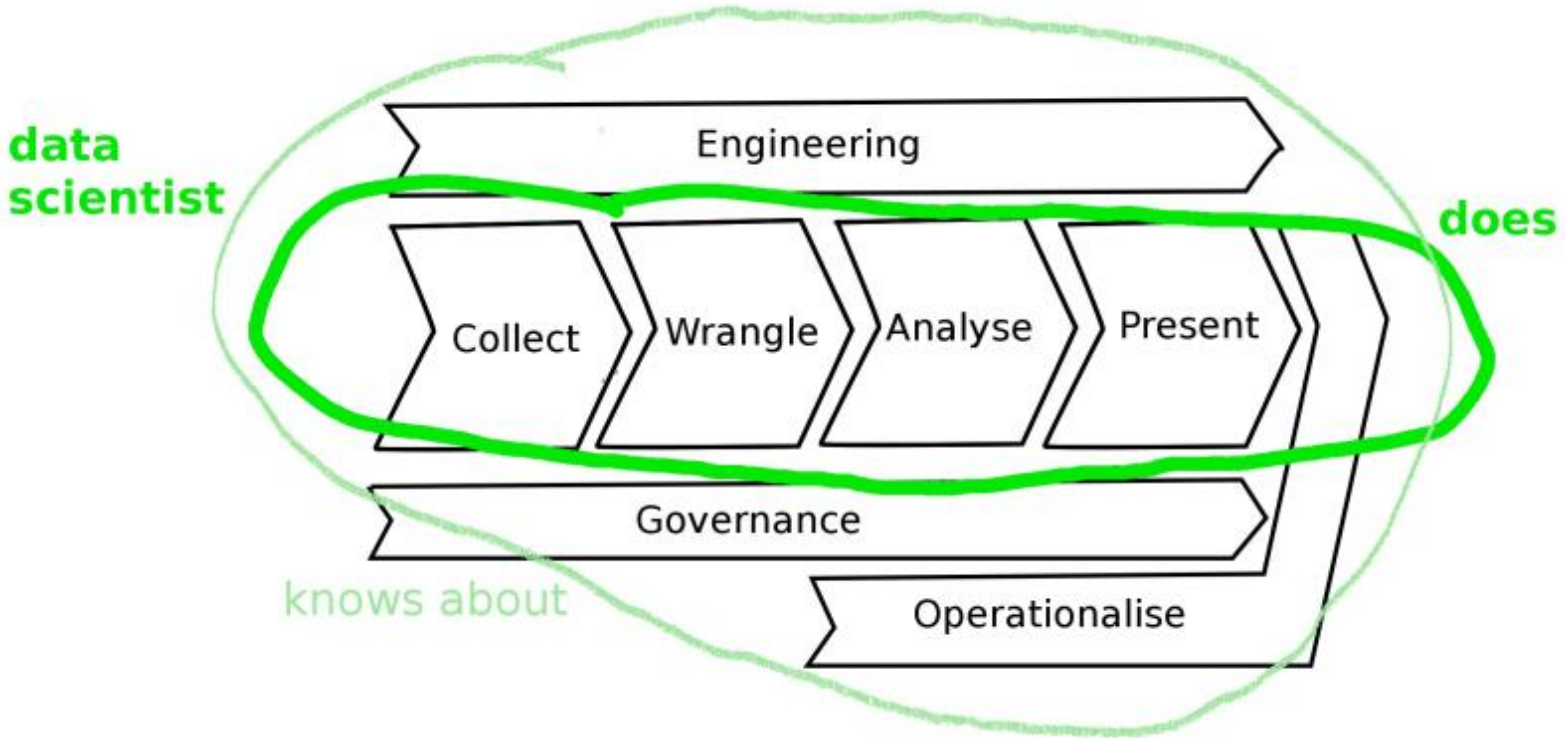
From *Doing Data Science*

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



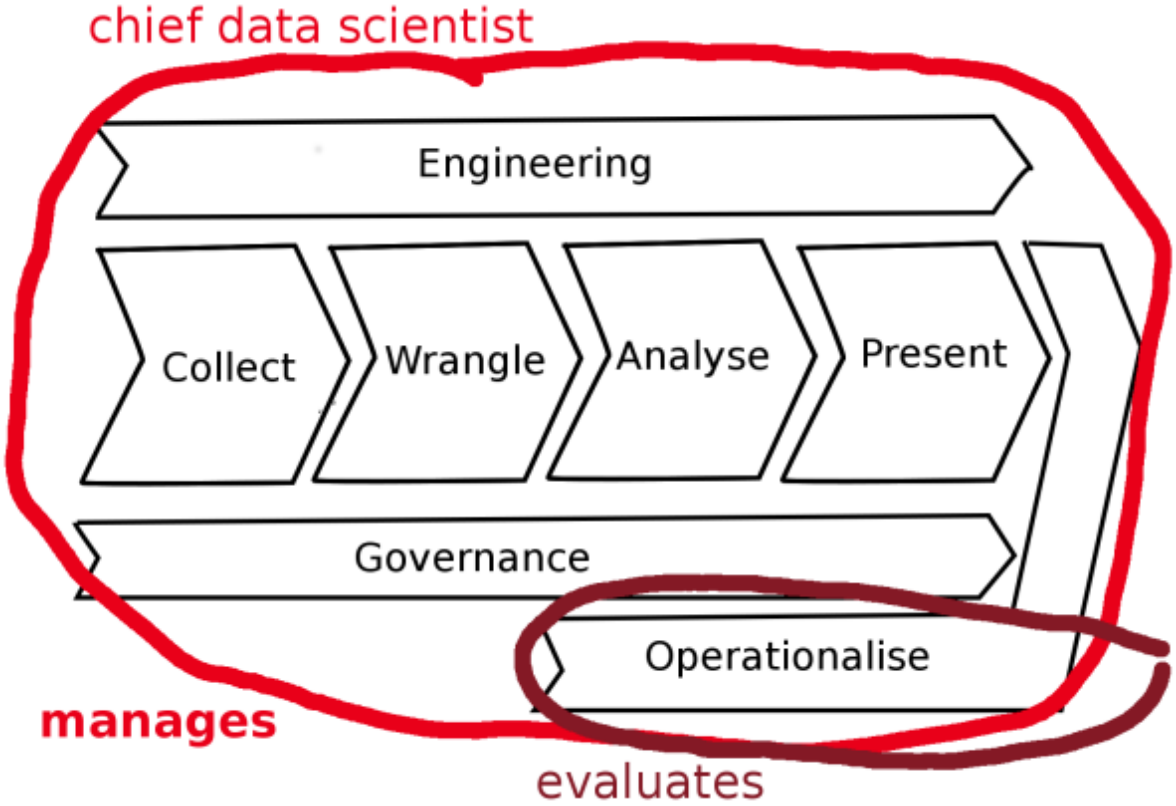
Profile: typical data scientist has a different mix of skills as well as domain knowledge

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 3:

Reviewing *Analyzing the Analyzers*

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, algorithms

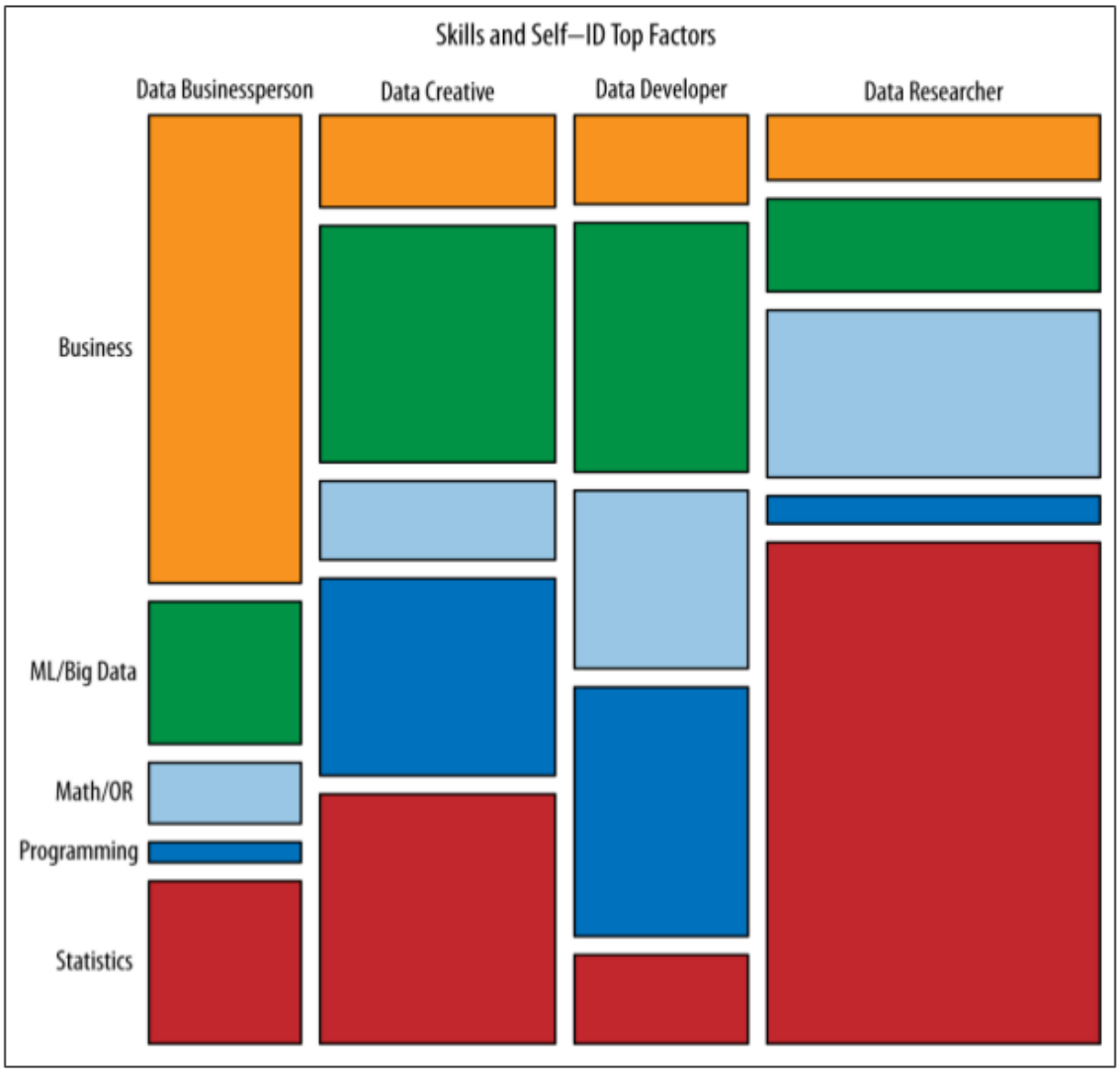
Programming: systems administration, back end programming, front end programming

Statistics: visualisation, temporal statistics, surveys and marketing, spatial statistics, science, data manipulation

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

Mapping Styles to Skills

Analyzing the Analyzers, Harris, Murphy and Vaisman, 2013



X-axis:
different roles

Y-axis:
different skills

which might you be?

Roles of a Data Scientist 4:

Interviews from *Data Analytics Handbook*

From *Data Analytics Handbook*

The *Data Analytics Handbook* is a four volume set of long interviews from industry and academic professionals in the field.

Volume 1 deals with practitioners:

- 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?

Suggested Reading

From *Data Analytics Handbook* read the interviews of

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

Lessons from the DA Handbook

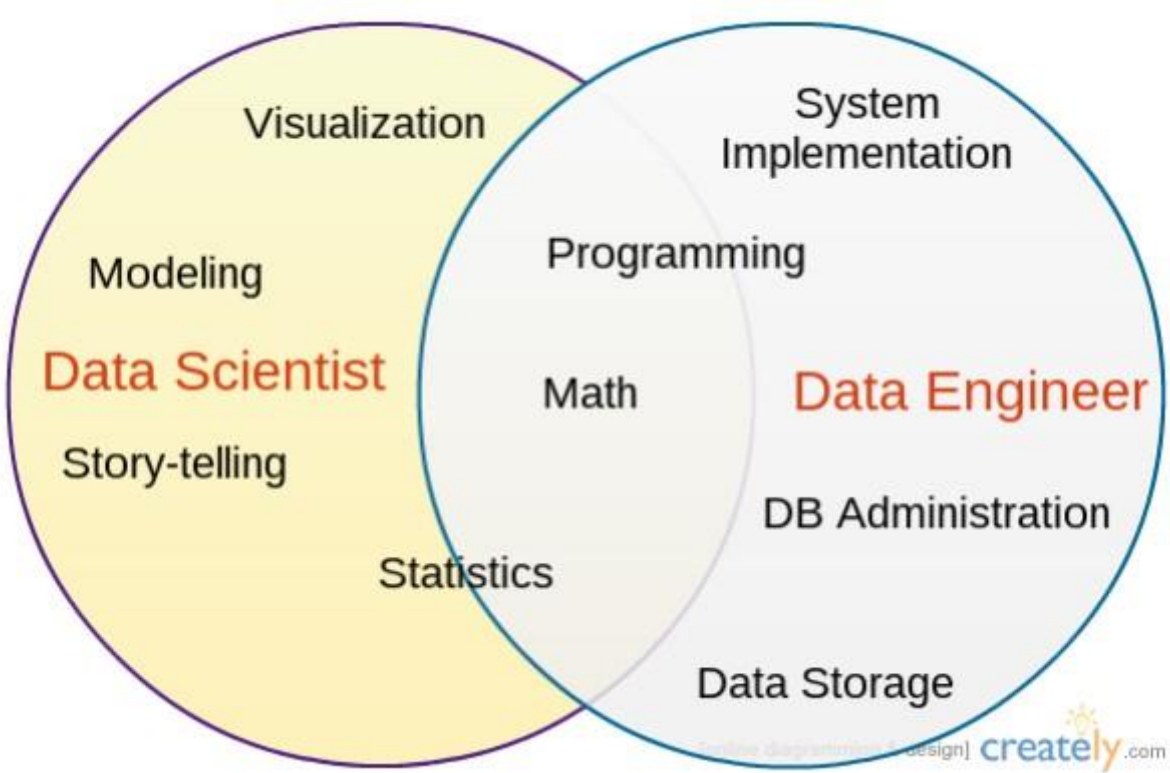
1. Communication skills are underrated.
2. The biggest challenge for a data analyst is the Collection and Wrangling steps.
3. A data scientist is better at statistics than a software engineer and better at software engineering than a statistician.
4. The data industry is still nascent and the roles less well defined so you get to interact with many parts of the company from engineering to business intelligence to product managers.
5. Keep a curiosity about working with data, a quality as important as your technical abilities.

Roles of a Data Scientist Addendum:

Other Views

some infographics

Data Scientists versus Data Engineers



see *Udacity on data careers*

Career as Data Scientist

To become a specialist you need:

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

This unit provides an introduction and background only.

Impact of Data Science (ePub section 1.7)

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:

- social life (Facebook),
- career (LinkedIn),
- search history (Google, *etc.*),
- health and medical (Fitbit, TBD),
- music (Apple), ...

This provides many, many advantages:

- *e.g.* personal agents, computerised support for health

But also some disadvantages:

- *e.g.* security and privacy breaches

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

MARS Question - **Y66M8T**

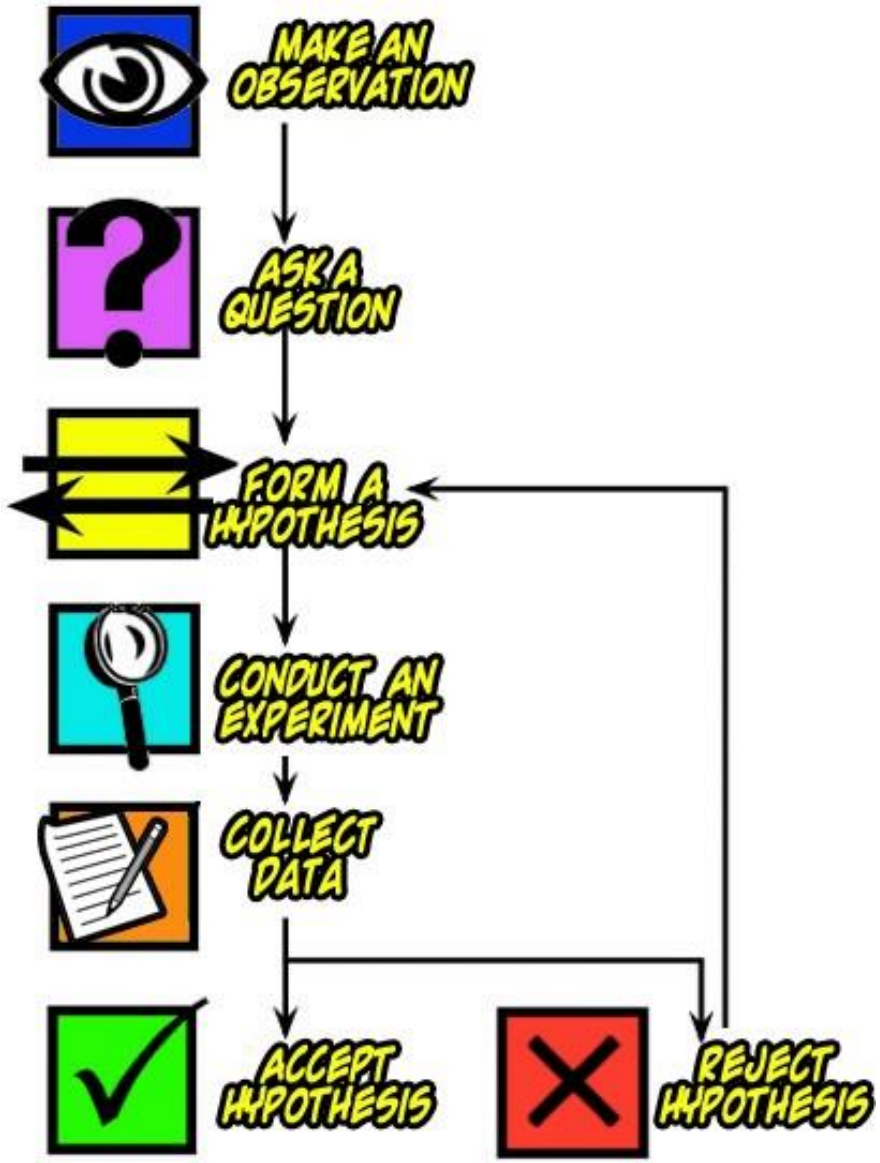
What role has the internet had in the development of data science?

A. the first big users of data science were internet giants
B. source of data for use
C. avenue for data science tools
D. all of the options

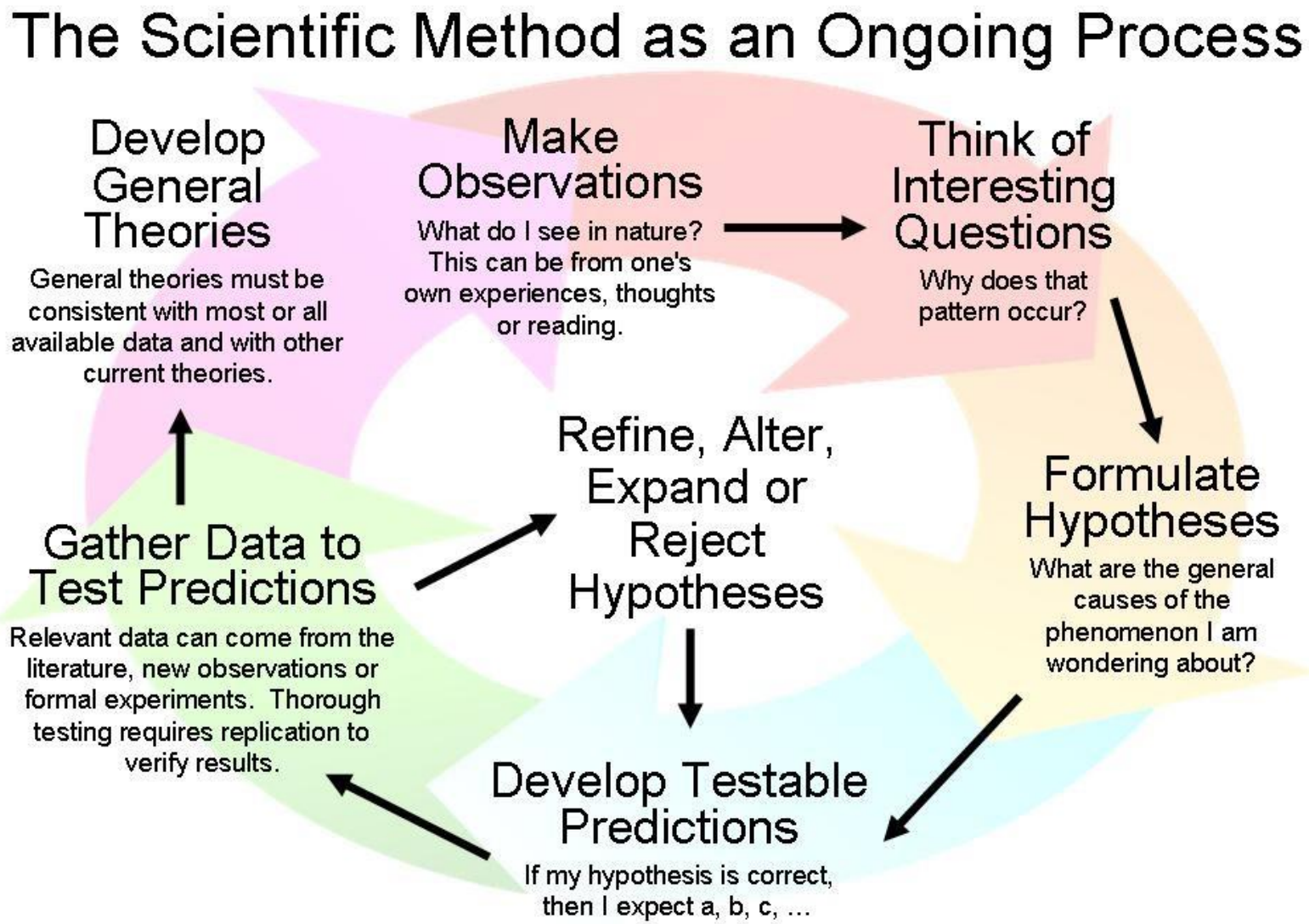
Impact of Data Science: Science

scientific method holds true, but broadens technology

The Scientific Method



The Scientific Method



from Wikipedia *Scientific method*

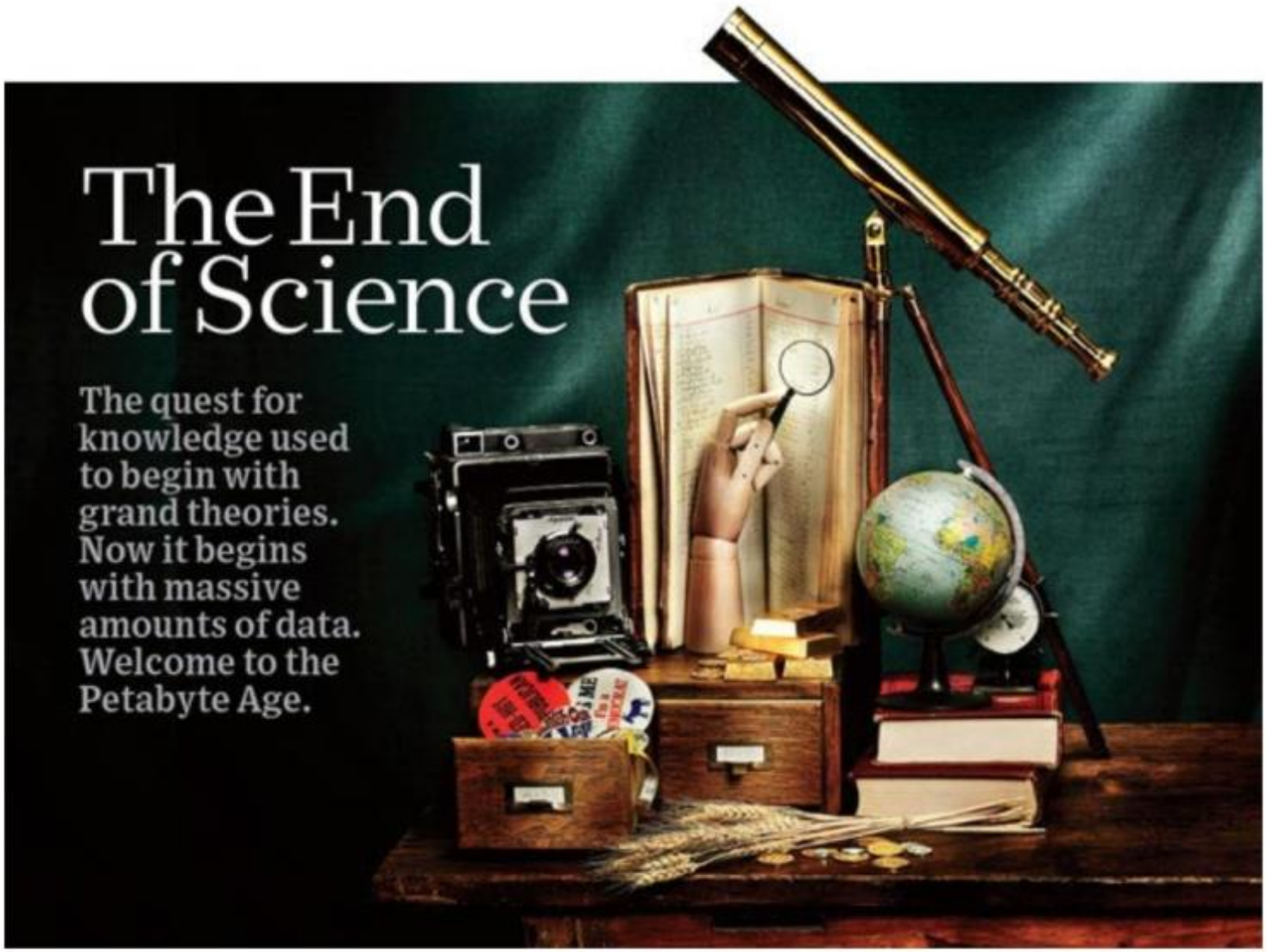
Scientific Method and Data Science

What is the relationship?

How does Data Science affect the Scientific Method?

The End of Theory

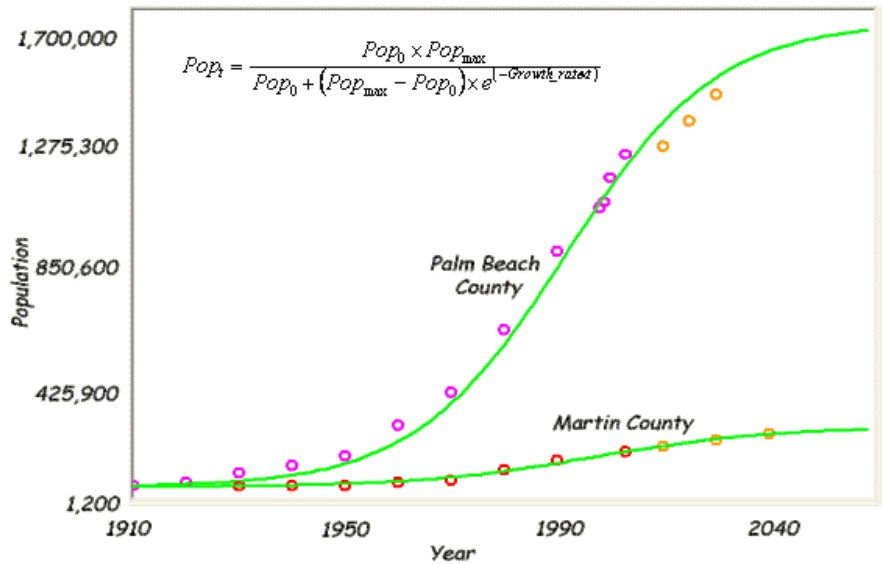
Chris Anderson's blog in Wired 23/05/2008



To Understand the Issues ...

What is a model?

A simple model of population growth:



A complex model of obesity:

- *Obesity Systems Map*

The End of Theory (cont.)

Science is largely driven by labourious studies to find complex causal models. 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. ...*
> ...
> *The new availability of huge amounts of data [...] offers a whole new way of understanding the world. Correlation supersedes causation, ...*

NB. When Google is delivering an advert, it doesn't need to be right, it just needs a good guess, so causality, models, etc., are not important.

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
- new data sources and tools for collecting data has been provided
  - crowd sourcing
  - social media
  - survey

Impact of Data Science:
Social good

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

longitudinal studies:= a method in which data is gathered for the same subjects repeatedly over a period of time

Car Industry Evolution, 1760s – Today = Driven by Innovation + Globalization

Early Innovation (1760s-1900s) = European Inventions

***1768** = First Self-Propelled Road Vehicle (Cugnot, France)*

***1876** = First 4-stroke cycle engine (Otto, Germany)*

***1886** = First gas-powered, 'production' vehicle (Benz, Germany)*

***1888** = First four-wheeled electric car (Flocken, Germany)*

Streamlining (1910s-1970s) = American Leadership

***1910s** = Model T / Assembly Line (Ford)*

***1920s-1930s** = Car as Status Symbol... Roaring '20s / First Motels*

***1950s** = **Golden Age...** Interstate Highway Act (1956)... 8 of Top 10 in Fortune 500 in Cars or Oil (1960)*

Modernization (1970s-2010s) = Going Global / Mass Market

***1960s** = Ralph Nader / Auto Safety*

***1970s** = Oil Crisis / Emissions Focus*

***1980s** = Japanese Auto Takeover Begins...*

***1990s – 2000s** = Industry Consolidation; Asia Rising; USA Hybrid Fail (Prius Rise)*

DAIMLERCHRYSLER

HYUNDAI

KIA

SAIC

***Late 2000s** = Recession / Bankruptcies / Auto Bailouts*

Re-Imagining Cars (Today) = USA Rising Again?

***DARPA Challenge (2004, 2005, 2007, 2012, 2013)** = Autonomy Inflection Point?*

***Today** =*

Ford GM TESLA

+

Google

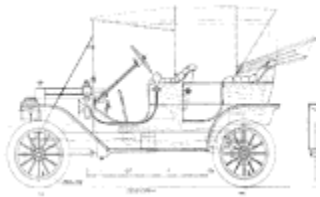
+

UBER DiDi

Source: KPCB Green Investing Team, Reilly Brennan (Stanford), Piero Scaruffi, Inventors.About.com, International Energy Agency, Joe DeSousa, Popular Science, Franz Haag, Harry Shipler / Utah State Historical Society, National Archives, texasescapes.com, Federal Highway Administration, Matthew Brown, Forbes, Grossman Publishers, NY Times, Energy Transition, UVA Miller Center for Public Affairs, The Detroit Bureau, SAIC Motor Corporation, Hyundai Motor Company, Kia Motors, Toyota Motor Corporation, DARPA, Chris Urmson / Carnegie Mellon,

Car Computing Evolution Since Pre-1980s = Mechanical / Electrical → Simple Processors → Computers

Pre-1980s Analog / Mechanical
Used switches / wiring to route feature controls to driver



1980s (to Present) CAN Bus (Integrated Network)
New regulatory standards drove need to monitor emissions in real time, hence central computer

1990s (to Present) OBD (On-Board Diagnostics) II Monitor / report engine performance; Required in all USA cars post-1996

1990s-2010s Feature-Built Computing + Early Connectivity
Automatic cruise control... Infotainment... Telematics... GPS / Mapping...

Today = Complex Computing
Up to 100 Electronic Control Units / car...
Multiple bus networks per car (CAN / LIN / FlexRay / MOST)... Drive by Wire...

Today = Smart / Connected Cars
Embedded / tethered connectivity...
Big Tech = New Tier 1 auto supplier (CarPlay / Android Auto)...

Tomorrow = Computers Go Mobile?...
Central hub / decentralized systems?
LIDAR...
Vehicle-to-Vehicle (V2V) / Vehicle-to-Infrastructure (V2I) / 5G...
Security software...

"The Box"
(Brooks & Bone)

Source: KPCB Green Investing Team, Darren Liccardo (DJI); Reilly Brennan (Stanford); Tom Denton, "Automobile Electrical and Electronics Systems, 3rd Edition," Oxford, UK: Tom Denton, 2004; Samuel DaCosta, Popular Mechanics, Techmor, US EPA, Elec-Intro.com, Autoweb, General Motors, Garmin, Evaluation Engineering, Digi-Key Electronics, Renesas, Jason Aldag and Jhaan Elker / Washington Post, James Brooks / Richard Bone, Shareable

MARS Question - **Y66M8T**

Referring to the two slides on the car industry.

First, they underwent a digitization process, followed by a _____ process

MARS Question - **Y66M8T**

Using a word or short phrase, name a non-automotive industries that have had similar developments in recent decades.

How do you expect the datafication process to change <*these industries*>?
More effective in _____

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)

In preparation for the tutorial

- Watch this *TED talk by Hans Rosling* showing the importance of visualising data.

Next: Module 2
Data Models in Organisations