INM430
Principles of Data Science

Week 01
Introduction & Basic Concepts

Aidan Slingsby

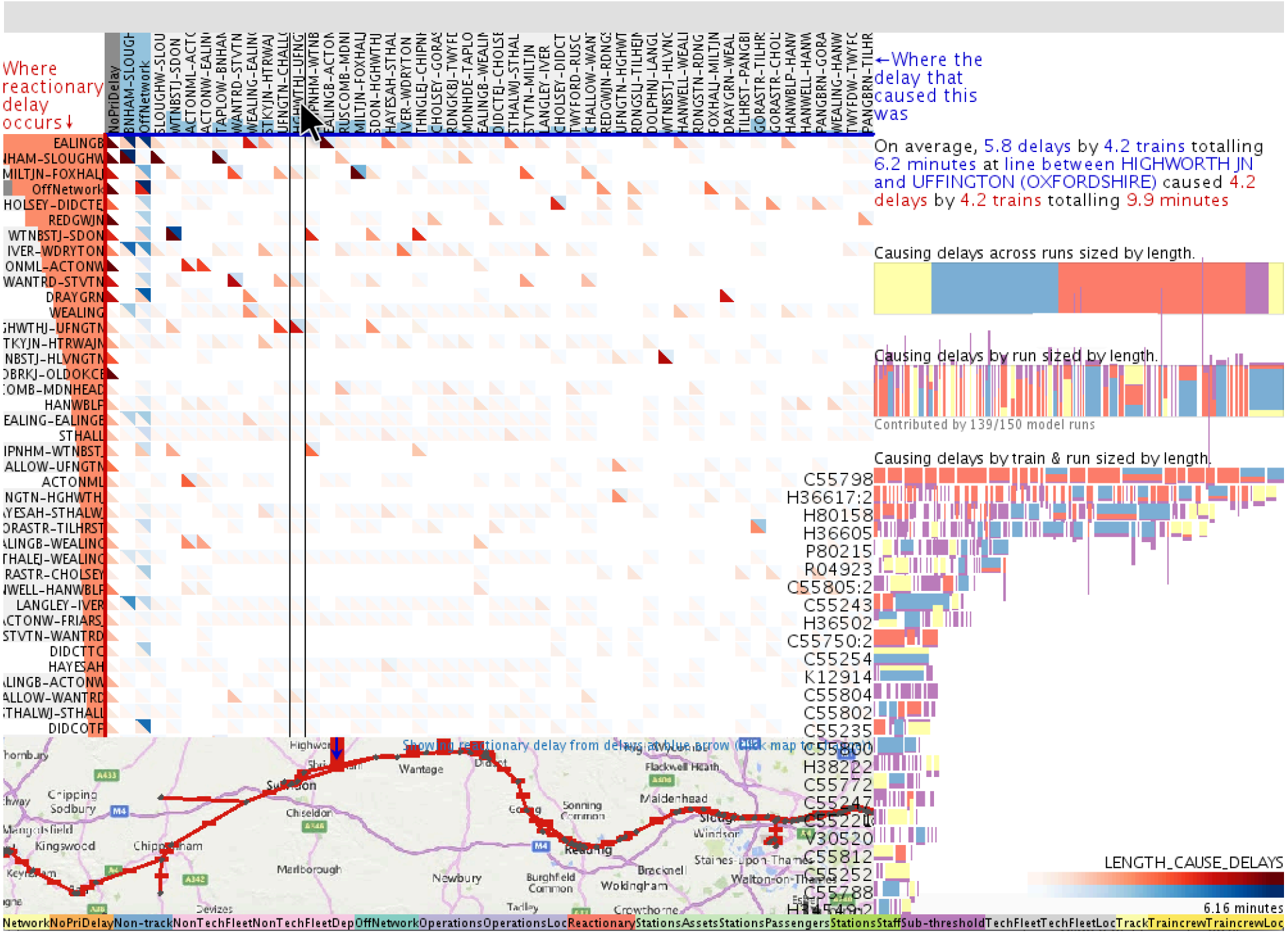
CITY
UNIVERSITY OF LONDON
EST 1894

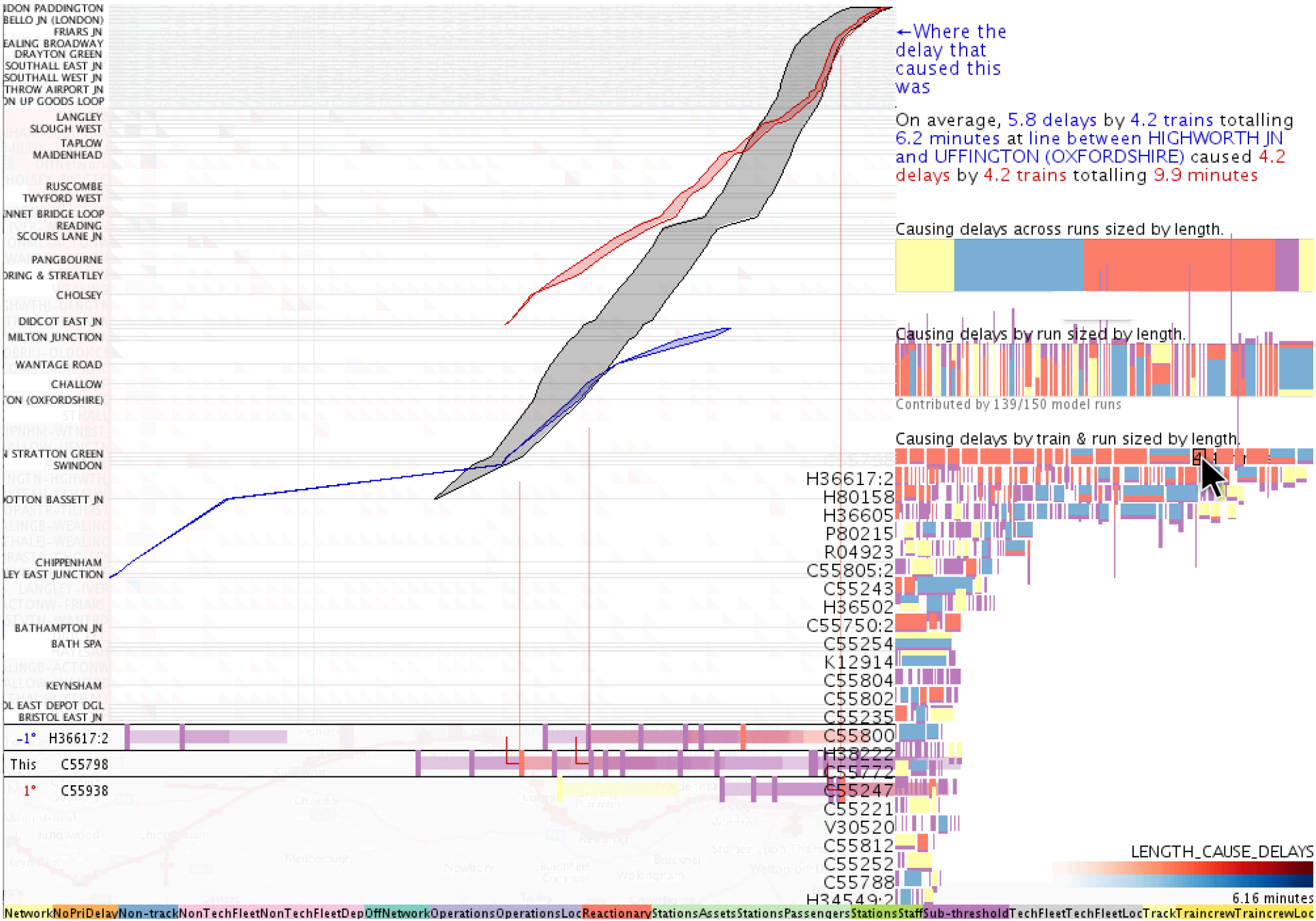
Module Specs – Aims & Skills

The Principles of Data Science (DS) module provides the **glue** to the other modules with an emphasis on the **DS process**, highlighting and exemplifying the main **challenges**, and exposing the students to some of the **tools** to be used in the other modules.

Module Specs – Aims & Skills

- Data Science (DS) process:
  - **Understanding business needs**
  - **Storing, accessing, preprocessing**
  - **Analysing data**
  - **Interpreting & communicating**
- DS process challenges
- Maths and statistics for DS
- Data storage technologies
- Knowledge representation
- Visualisation and visual analytics


Where reactionary delay occurs↓
←Where the delay that caused this was
On average, 5.8 delays by 4.2 trains totalling 6.2 minutes at line between HIGHWORTH JN and UFFINGTON (OXFORDSHIRE) caused 4.2 delays by 4.2 trains totalling 9.9 minutes
Causing delays across runs sized by length.
Causing delays by run sized by length.
Contributed by 139/150 model runs
Causing delays by train & run sized by length.
C55798
H36617:2
H80158
H36605
P80215
R04923
C55805:2
C55243
H36502
C55750:2
C55254
K12914
C55804
C55802
C55235
C55800
H38222
C55772
C55247
C55220
V30520
C55812
C55252
C55788
LENGTH_CAUSE_DELAYS
6.16 minutes
NetworkNoPriDelayNon-trackNonTechFleetNonTechFleetDepOffNetworkOperationsOperationsLocReactionaryStationsAssetsStationsPassengersStationsStaffSub-thresholdTechFleetTechFleetLocTrackTraincrewTraincrewLoc


←Where the delay that caused this was
On average, 5.8 delays by 4.2 trains totalling 6.2 minutes at line between HIGHWORTH JN and UFFINGTON (OXFORDSHIRE) caused 4.2 delays by 4.2 trains totalling 9.9 minutes
Causing delays across runs sized by length.
Causing delays by run sized by length.
Contributed by 139/150 model runs
Causing delays by train & run sized by length.
H36617:2
H80158
H36605
P80215
R04923
C55805:2
C55243
H36502
C55750:2
C55254
K12914
C55804
C55802
C55235
C55800
H38222
C55772
C55247
C55221
V30520
C55812
C55252
C55788
H34549:2
-1° H36617:2
This C55798
1° C55938
LENGTH_CAUSE_DELAYS
6.16 minutes
NetworkNoPriDelayNon-trackNonTechFleetNonTechFleetDepOffNetworkOperationsOperationsLocReactionaryStationsAssetsStationsPassengersStationsStaffSub-thresholdTechFleetTechFleetLocTrackTraincrewTraincrewLoc

Knowledge & Skills

At the end of the module, you will hopefully be able to:

- Formulate **business / research requirements**
- Identify **analytical needs**
- Understand **data source characteristics**
- Evaluate **data representation** options
- Evaluate **data storage** possibilities
- **Wrangle / merge** data sources
- Evaluate **data quality**
- Choose & utilise correct **analytical tools**
- **Communicating** insights

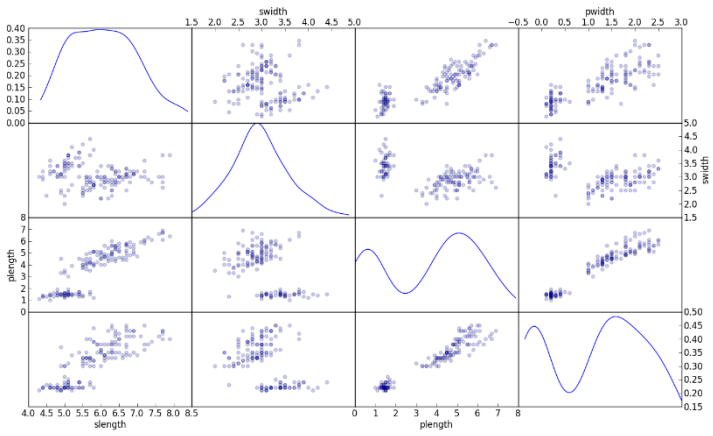
Knowledge & Skills

At the end of the module, you will hopefully have skills to:

GENERATE VALUABLE
INSIGHT
THROUGH
DATA ANALYSIS

Practical Skills – yes, we'll be coding !!

- Data analysis skills in **Python**
  - **scipy, numpy, pandas, scikit-learn,**
  - scientific computing skills
- Statistical analysis methods
- Visual analytics / visualisation skills



Module Information

Lecturers

Aidan Slingsby – a.slingsby@city.ac.uk
Lectures: 7 (and module leader)
Research group: giCentre
Office : A401b (College Building)
Office Hours: Wednesdays 16:00-18:00 (usually)

Michael Garcia Ortiz - michael.garcia-ortiz@city.ac.uk
Lectures: 3
Research group: Artificial Intelligence
Office: A309i (College Building)
Office Hours: Monday, 12:00-12:50
and Tuesday 12:00 - 12:50

Lab assistants

Alex Galkin - Oleksandr.Galkin@city.ac.uk
Data Science Teaching Assistant

Adam White - adam.white.2@city.ac.uk
Research Assistant, Machine Learning Group

Charitos Charitou
PhD student, Machine Learning Group

Kevin Allain
PhD student, giCentre

Lectures & Labs

- **Thursdays:**
  - Lectures **14:00-16:00** **in** **ELG03**
  - Labs/Tutorials:
    - Lab1- **16:00 - 16:50 : ELG10**
    - Lab2- **16:00 - 16:50 : ELG06/ELG07**
    - Lab3- **17:00 - 17:50 : ELG10**
    - Lab4- **17:00 - 17:50 : ELG06/ELG07**

Reading week in week 5 not 6!

Check timetabling: https://sws.city.ac.uk/

Assessment

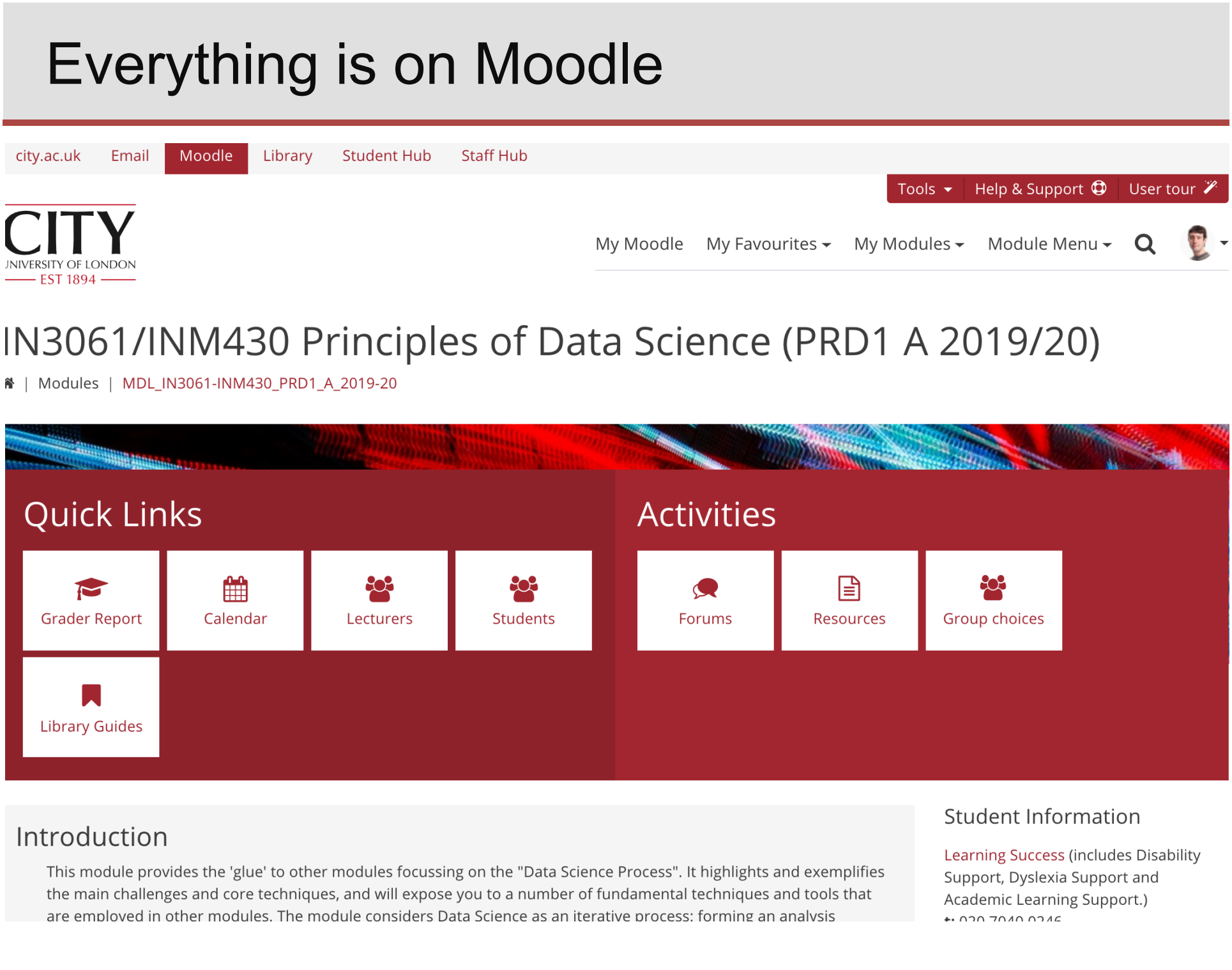
- **No exam**
- **Coursework**
  - **A plan and progress update** (formative feedback)
  - **A piece of analysis (100% of the module mark)** , computational notebook & report
- **Deadlines strictly enforced**
- **We are not allowed to grant extensions**
  - Apply for an EC if you need to: http://goo.gl/9gIIMi

Extenuating Circumstances

- **ECP : form & evidence – see Intranet / Moodle**
  - unforeseeable
  - medical or legal in nature
  - supportable by evidence

- **Extenuating Circumstances Information:**
  http://goo.gl/9gIIMi

Everything is on Moodle



Moodle - http://moodle.city.ac.uk

- Overview, lecture notes and exercises released here
  - weekly overview
  - lecture notes
  - practical exercise(s) : model answers
  - additional docs : external links & resources
- Coursework
  - released and submitted via Moodle only
- **Discussion Forum !!**
  - **Engage and discuss, make it a live platform!**

Textbook?

- No fixed textbook
- Materials on Moodle
- **Several books** to get assistance
- Weekly suggested readings
- Links to online resources

How to Learn …?

- **come to Lectures**
  - engage
- **come to Tutorials / Classes / Labs**
  - try stuff out
  - engage
  - discuss, ask questions
- **Independent Learning**
  - Moodle – links and model answers
  - course text
- **Coursework**
  - keep with the schedule!

Tutorials / Practical Exercises / Labs

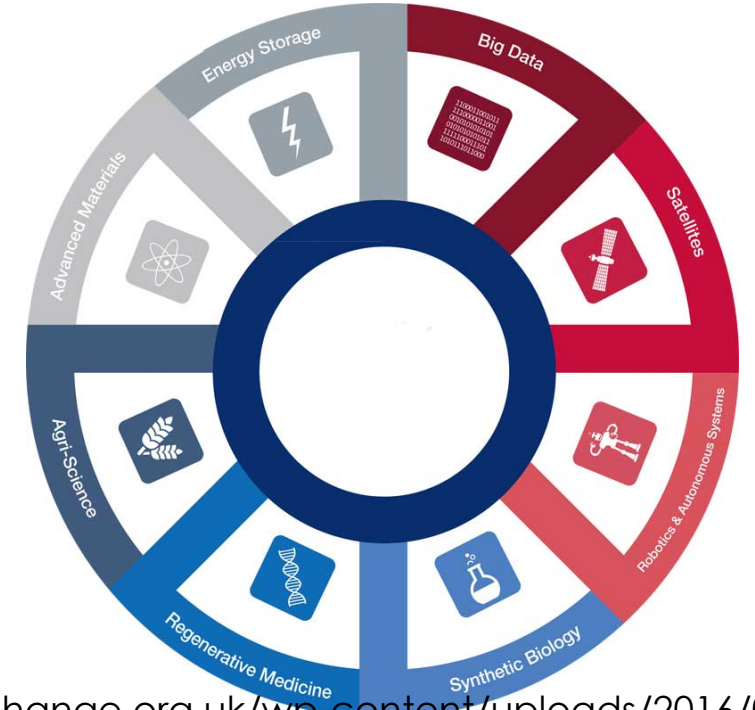
- Important that you …

- complete tasks
- try things out
- discuss approaches
- be curious

Data Science

Some motivation -- *Eight great technologies*

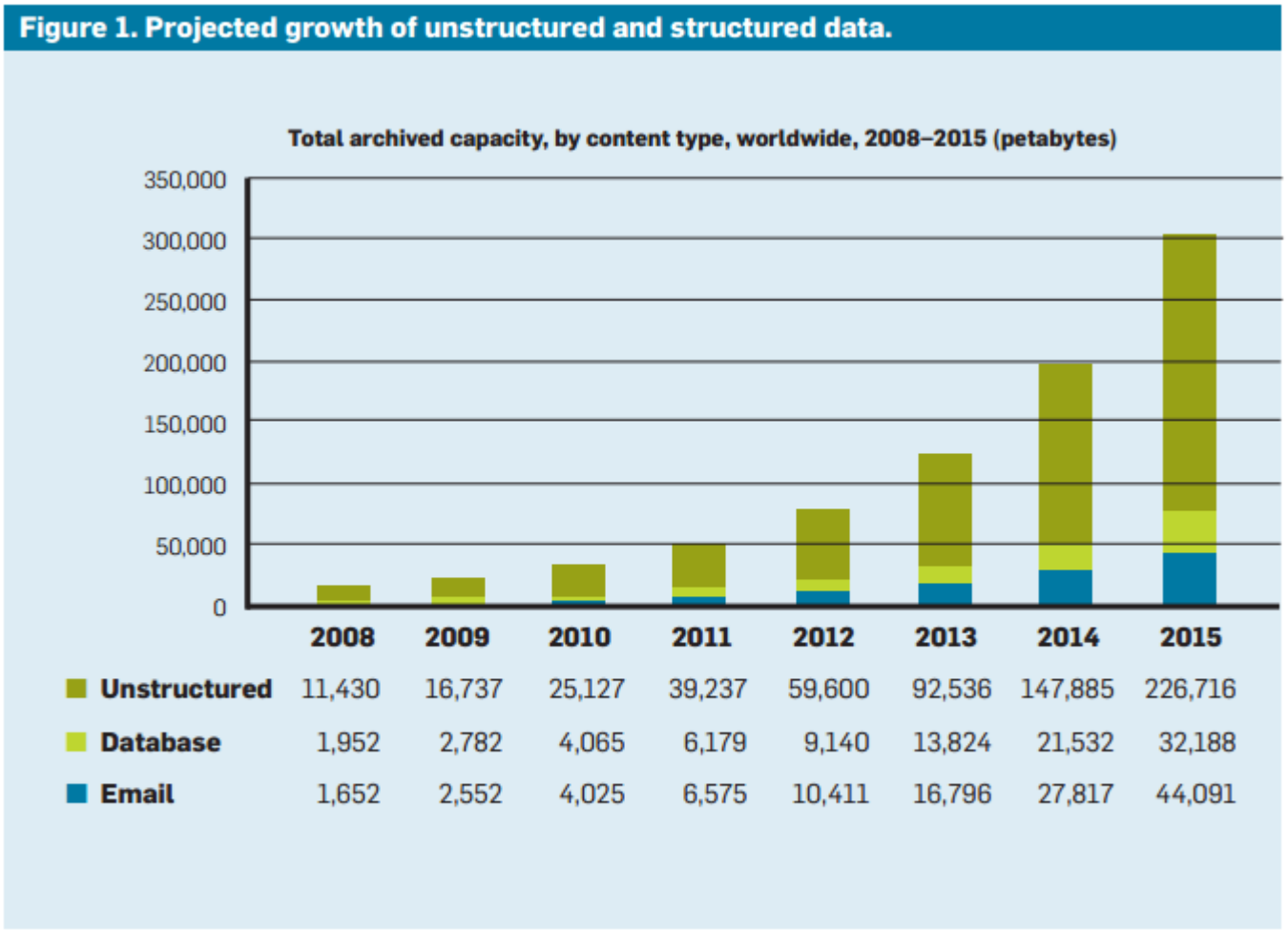
"The next generation of scientific discovery and innovation will be data-driven as previously unrecognised patterns are discovered by **analysing massive and mixed data sets**. ", David Willetts, MP, 2013.



https://policyexchange.org.uk/wp-content/uploads/2016/09/eight-great-technologies.pdf

data?

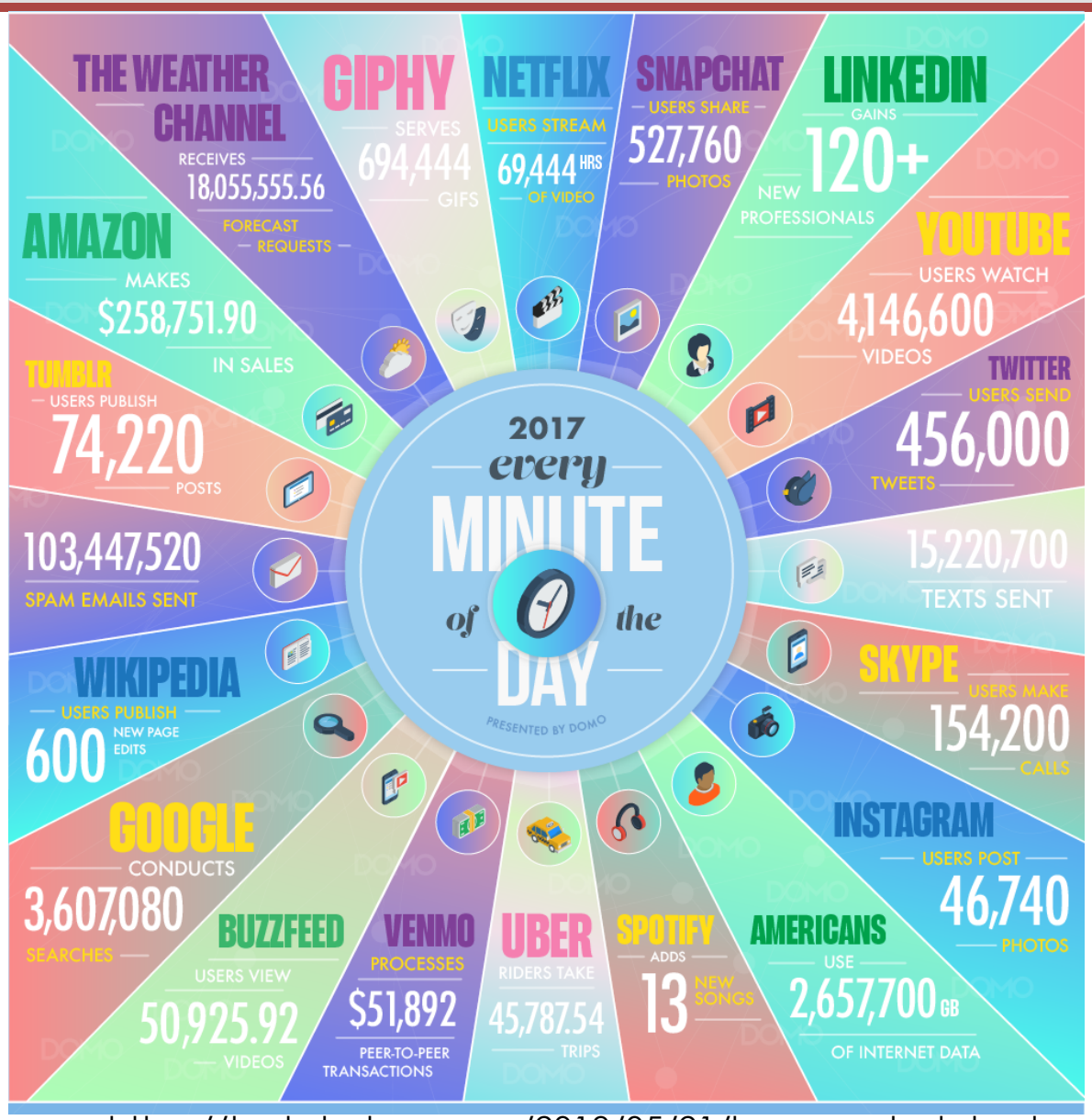
Figure 1. Projected growth of unstructured and structured data.



	2008	2009	2010	2011	2012	2013	2014	2015
Unstructured	11,430	16,737	25,127	39,237	59,600	92,536	147,885	226,716
Database	1,952	2,782	4,065	6,179	9,140	13,824	21,532	32,188
Email	1,652	2,552	4,025	6,575	10,411	16,796	27,817	44,091

From "***Data Science and Prediction***" by Vasant Dhar, *Communications of the ACM*, 2013

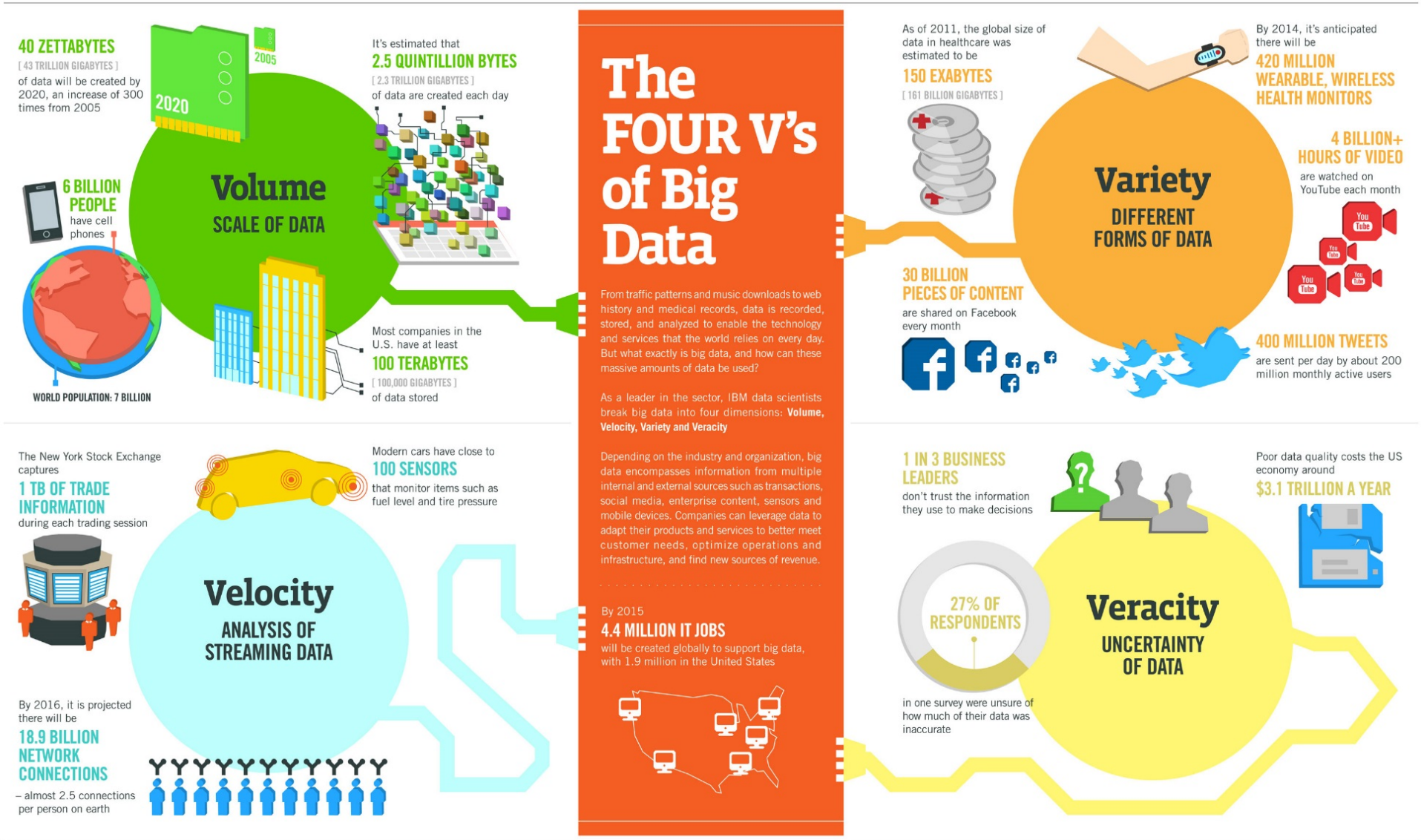
How much data do we create every day?



- We produce 2.5 quintillion bytes of data each day.
- 90% of the data in the world today has been created in the last two years alone.

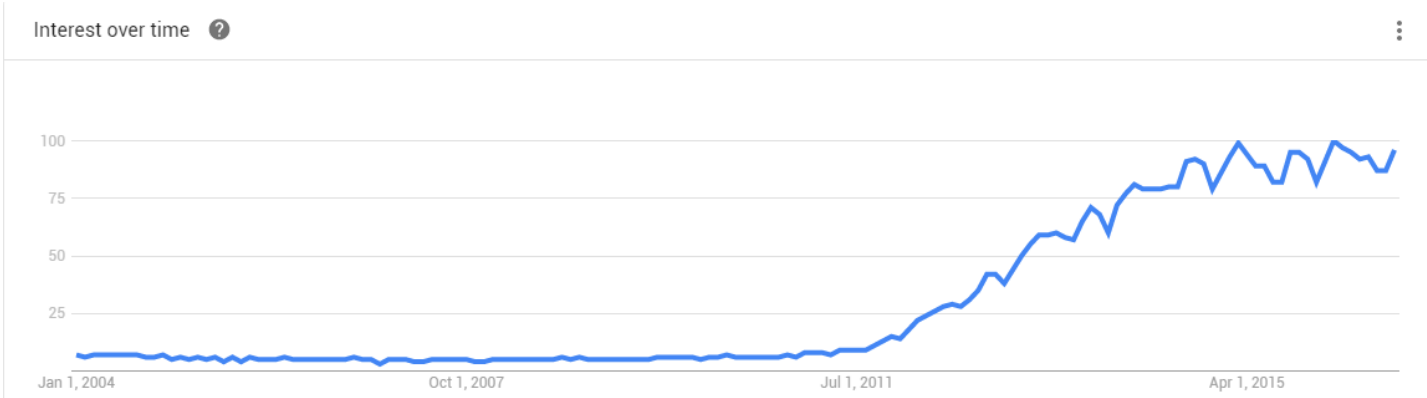
https://techstartups.com/2018/05/21/how-much-data-do-we-create-every-day-infographic/

V's of Big data



More V's added

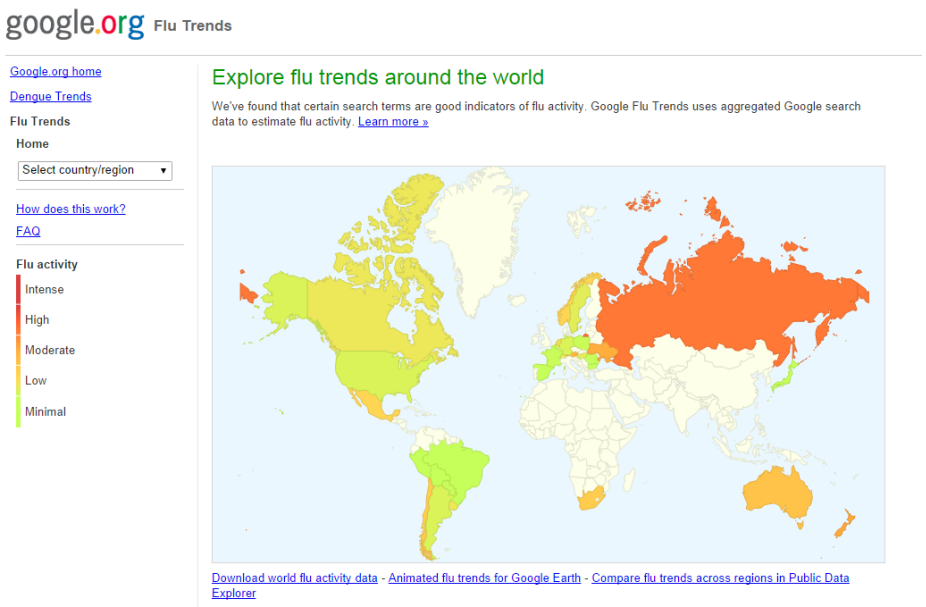
- Volume
- Velocity
- Variety
- Veracity
- Visualisation
- Value



"**big data**" on Google Trends, as of today
http://www.google.com/trends/explore#q=big%20data

Classic example – Google Flu Trends

- http://www.google.org/flutrends/ (stopped last year)



Google says ..

.. **relationship** between how many people **search for flu-related topics** and how many people actually **have flu symptoms**. … to **estimate** how much flu is circulating ….

Video: http://goo.gl/4ysAmw

Ginsberg, Jeremy, et al. "Detecting influenza epidemics using search engine query data." *Nature* 457.7232 (2008): 1012-1014.

Mobile phone data redraws bus routes in Africa

By Jane Wakefield
Technology reporter

1 May 2013

Top Stories

Commons 'bear pit' condemned by Jo Cox's husband
Brendan Cox criticises PM's "sloppy language" after he told MPs the best way to remember the murdered MP is to get "Brexit done".

5 minutes ago

MPs return to Commons after rowdy scenes

35 minutes ago

Laura Kuenssberg: Parliament a place of fear and loathing

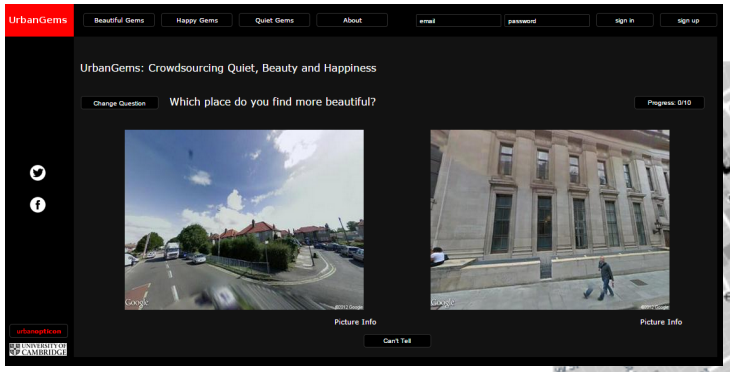
7 hours ago

Features

https://www.bbc.co.uk/news/technology-22357748

Another example - The Shortest Path to Happiness

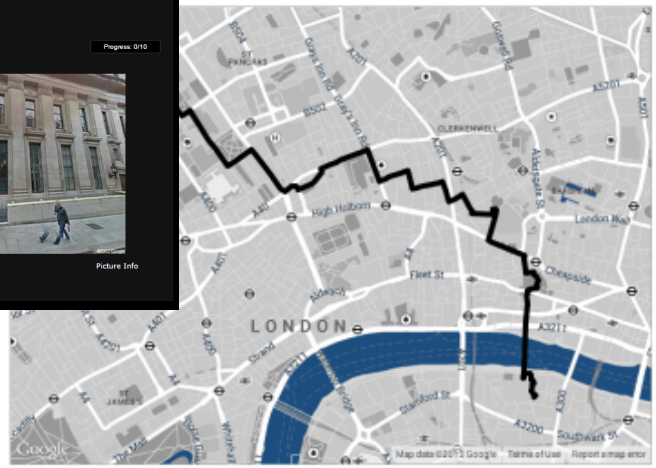
The Shortest Path to Happiness: Recommending Beautiful, Quiet, and Happy Routes in the City



http://urbangems.org/

(a) Shortest

(b) Beauty



(c) Quiet

(d) Happy

Quercia, Daniele, Rossano Schifanella, and Luca Maria Aiello. *ACM conference on Hypertext and social media*, 2014.

SurveyGlyphs

Glyphs for Exploring Crowd-sourced Subjective Survey Classification

A. Kachkaev, J. Wood and J. Dykes, giCentre, City University London

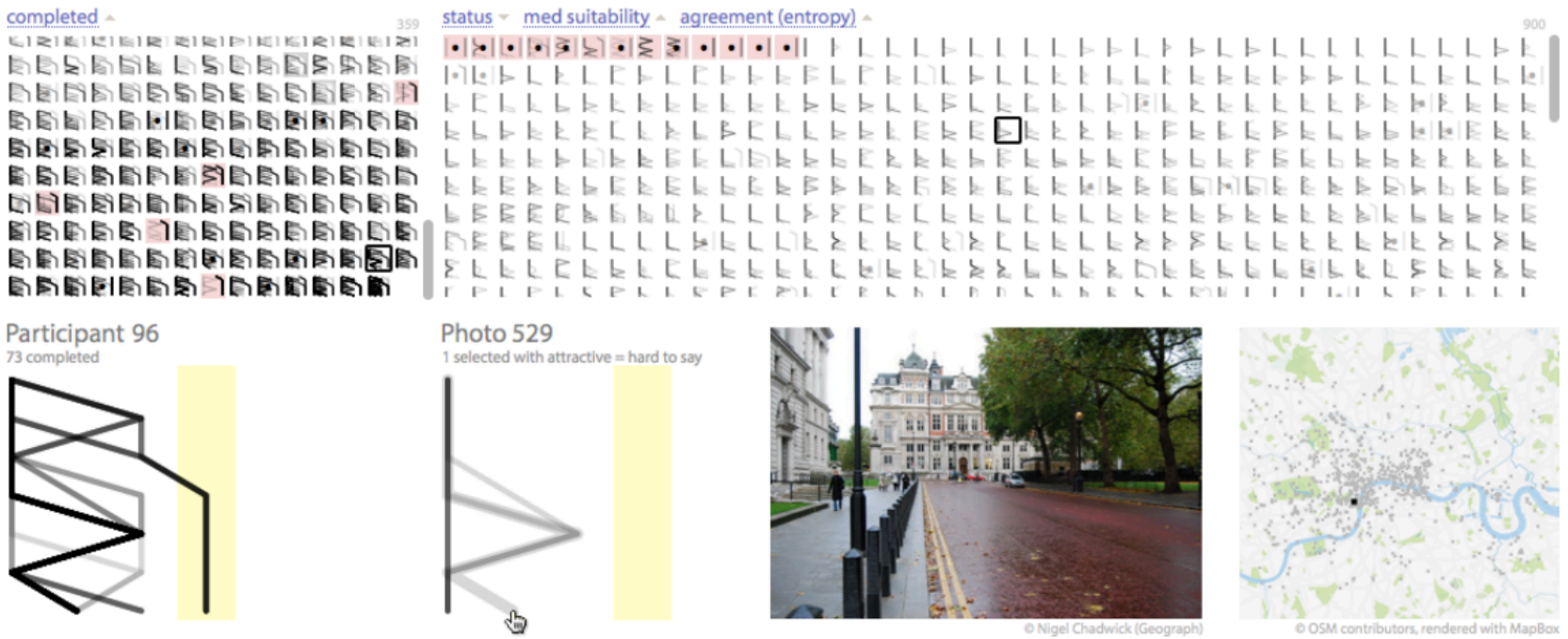


Figure 1: The interface of the survey analysis tool allowing navigation through 8,434 collected responses (49,285 answers) with use of glyphs, linked views and interaction. Available at photoassessment.org/results, demo video at vimeo.com/90299533.

“Mappiness”

https://www.mappinessapp.com/

Quick discussion

The scientists on “Shortest Path to Happiness” did a survey to gather data on how “*beautiful, quiet, and happy*” a place is.

Q: Can you think of alternative data sources that might be a proxy for how “*beautiful, quiet, and happy*” or “how nice to walk” a place is?

Some I can think of:

- Traffic sensors
- CCTV pedestrian counts
- Volume of connections from cell towers
- “Greenness” derived from satellite imagery
- # of businesses/shops

Happiness?

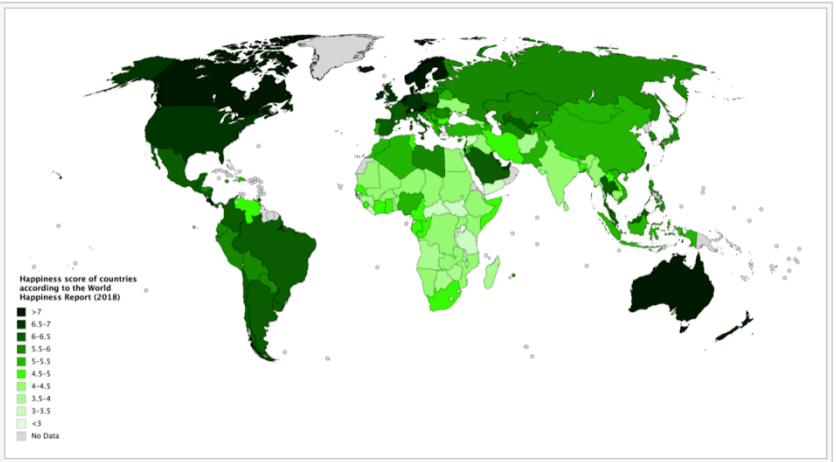
World Happiness Report

From Wikipedia, the free encyclopedia

The **World Happiness Report** is an annual publication of the United Nations Sustainable Development Solutions Network. It contains articles, and rankings of national happiness based on respondent ratings of their own lives,[1] which the report also correlates with various life factors.[2]

As of March 2019, Finland was ranked the happiest country in the world twice in a row.[3][4]

Contents [hide]



Map showing happiness of countries by their score according to the 2018 World Happiness Report.

https://en.wikipedia.org/wiki/World_Happiness_Report

Data Science (as **DS** from now)

- Data science is a **systematic study** of generalizable **extraction of knowledge** from data (by Vasant, 2013)
- Term coined by William S. Cleveland in 2001[*]
- DS is in the process of being defined!
- Lots of criticism on the term itself
  - Where is the science? – **systematic & generalizable**

[*] Cleveland, William S. "Data science: an action plan for expanding the technical areas of the field of statistics.“, (2001)

On the origins ..

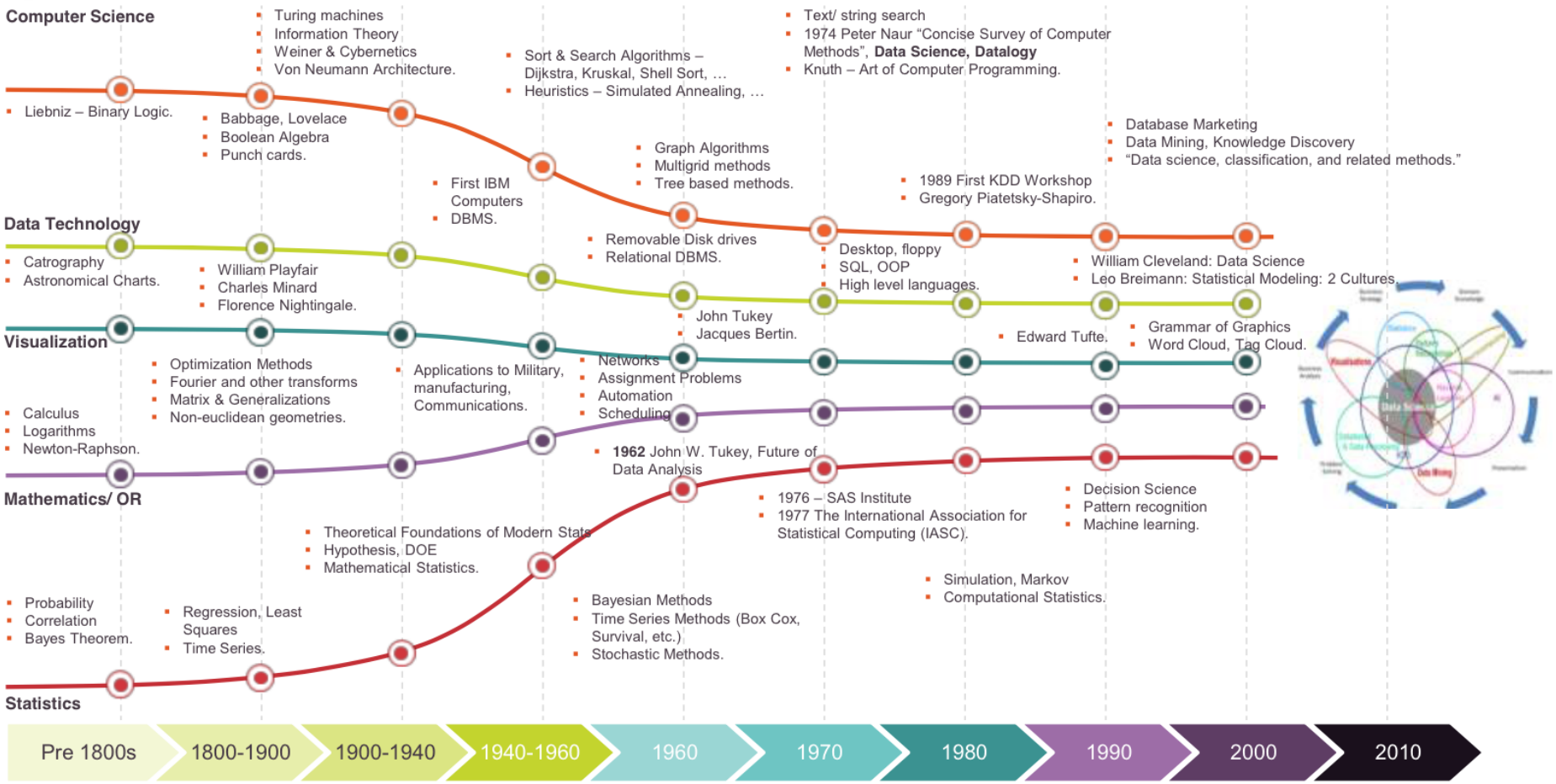
Term coined by William S. Cleveland in 2001[*]

*…knowledge among **computer scientists** about how to think of and approach the analysis of data is limited, just as the knowledge of computing environments by **statisticians** is limited. A **merger of the knowledge bases** would produce a powerful force for innovation.*

[*] Cleveland, William S. "Data science: an action plan for expanding the technical areas of the field of statistics.“, (2001)

DS sits on the shoulder of many fields:

signal processing, mathematics, probability models, machine learning, statistical learning, computer programming, data engineering, pattern recognition and learning, visualization, uncertainty modelling, data warehousing, and high performance computing …

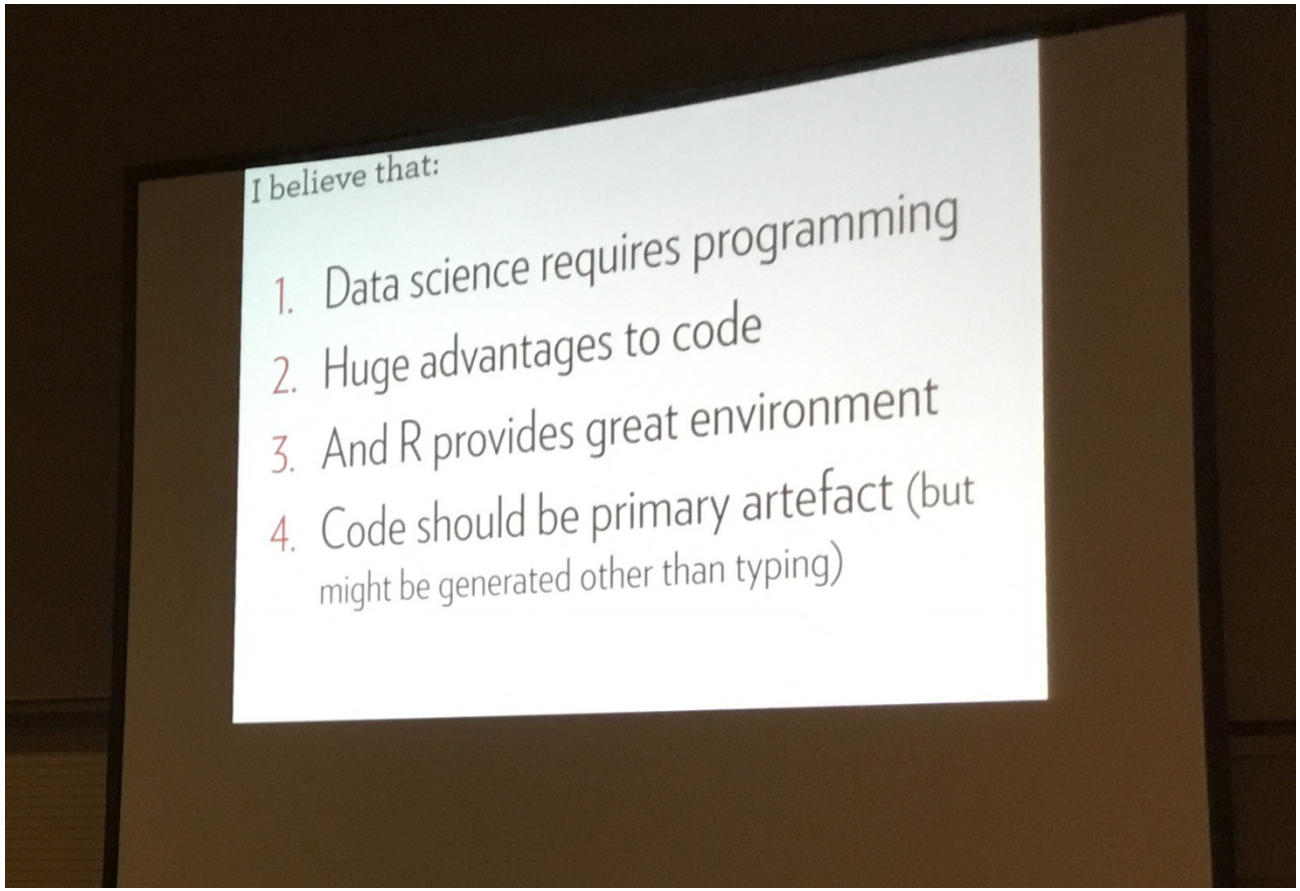


By Capgemini Consulting, http://www.slideshare.net/capgemini/impact-of-big-data-on-analytics

Characteristics of Data Science

- Attitude
  - Data-driven decision-making
  - Bridging IT and business
- Types of analysis
  - Patterns that match data (rather than data that match patterns)
  - Multiple disciplines
- Skills
  - Wrangling and repurposing data
  - Using methods from different places
  - Understanding limitations
  - Understanding business/analytical needs
  - Knowing how to “sell” them to the business

Hadley Wickham's view on DS



Visualization in Data Science (VDS at IEEE VIS 2017)

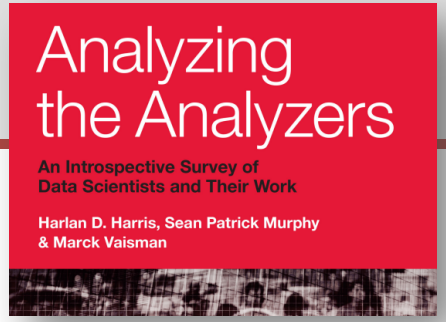
VDS
Sunday, October 1, 2017

Keynote: Challenges in Data Science
Hadley Wickham, RStudio
http://hadley.nz

Data scientist ?

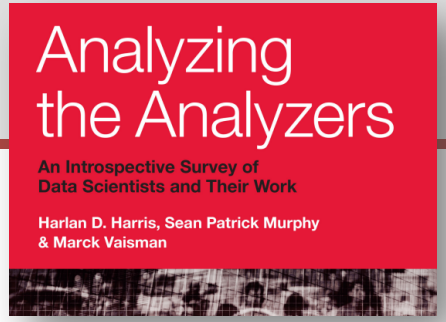
- Sexiest job of the 21st century, according to Harvard Business Review, 2012
- A data scientist is many people in one, someone:
  - who **understands domain** (industry/academia) needs and terminology
  - who is able to **mash-up** several analytical tools
  - who is able to **design and implement** solutions to extract knowledge from the data
  - who can **communicate** findings

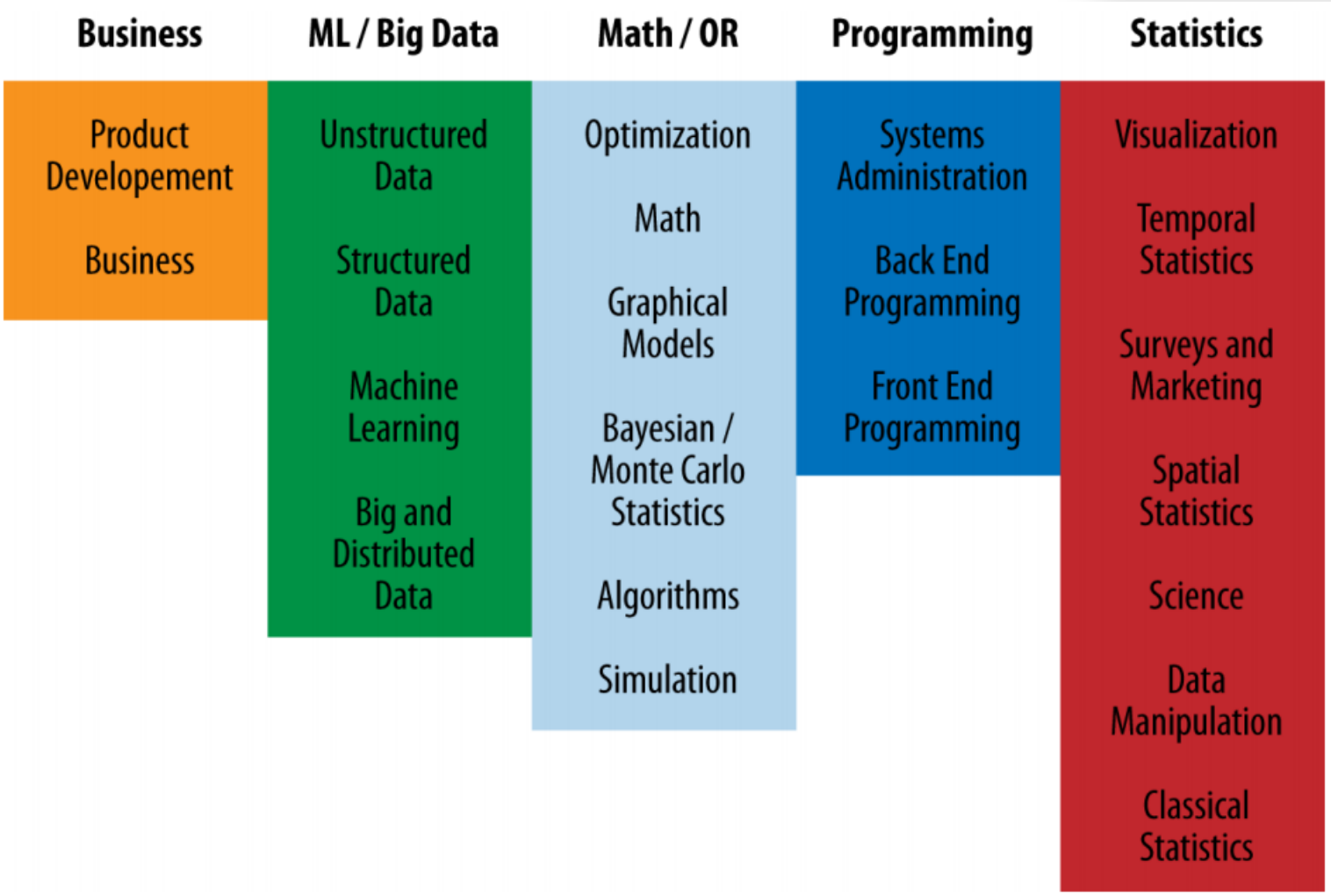
On data analysts – analyst types



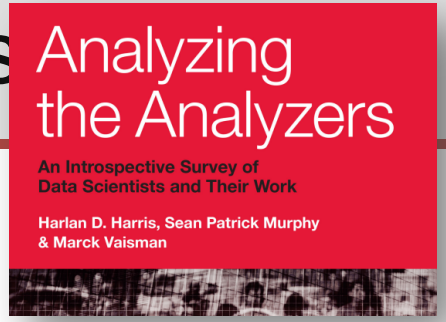
Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

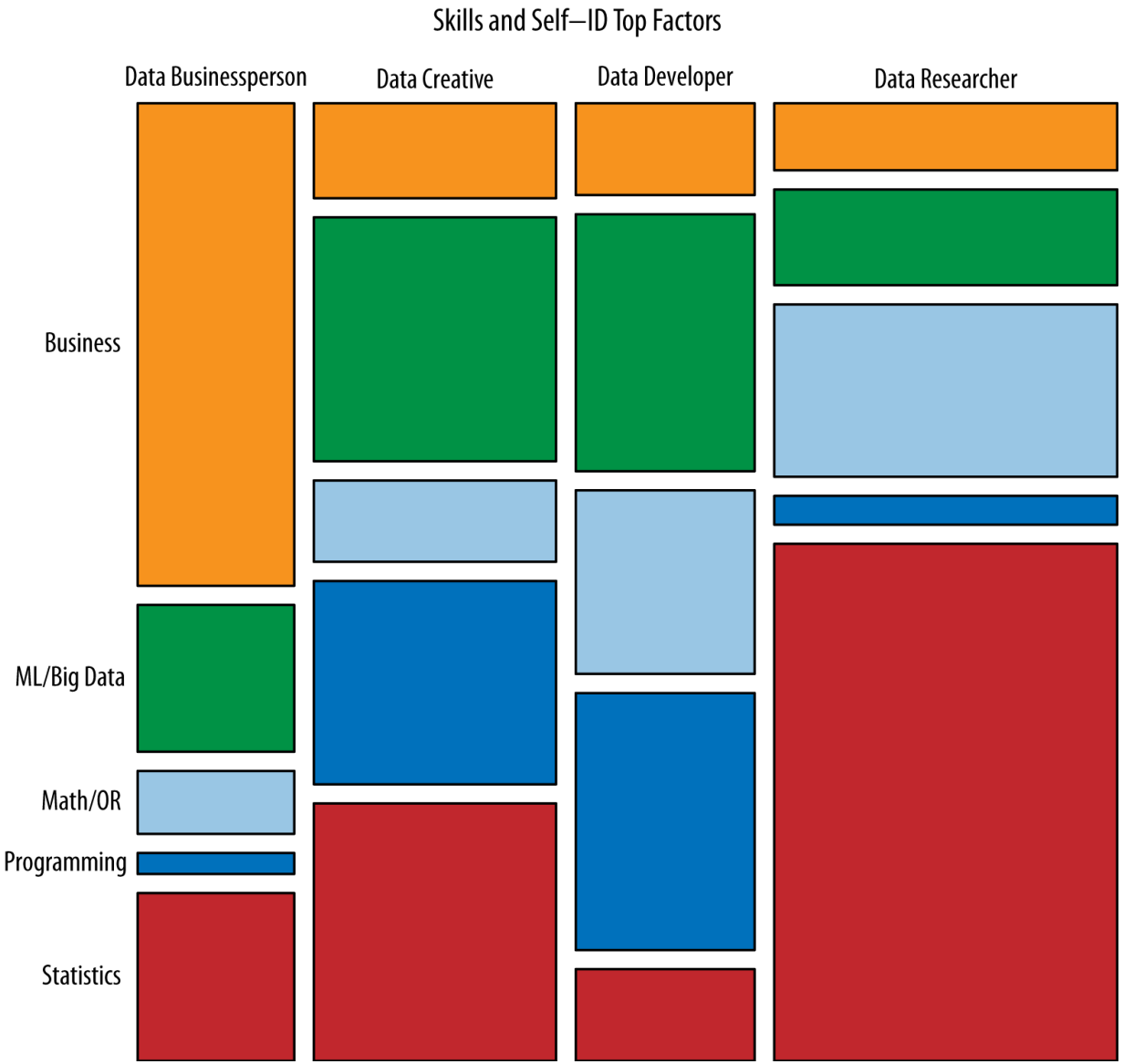
On data analysts – different skills



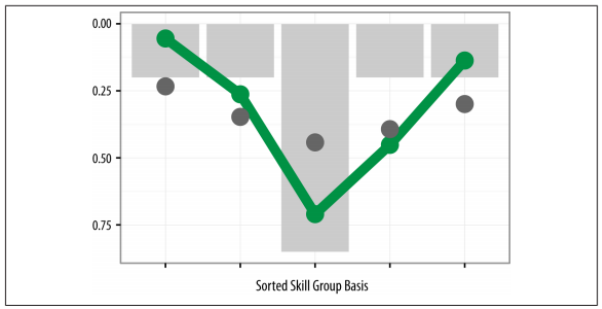


On data analysts – skills vs. types





T-shaped data scientists



DS Process

- **Understand** analytical domain needs
- **Collect** & make data **available**
- Get the **data ready** for analysis
- Exploratively (and visually) **analyse** the data
- **Model** the phenomena (if needed)
- **Evaluate** findings
- **Communicate** findings
- **ITERATE** (from any stage to any other stage)!

DS Process – Formulating the question

- Understand / represent processes
- Domain specific
- Business / Scientific
- Requires domain knowledge
  - Background in the domain
  - Working closely with experts
  - Workshops / brainstorming sessions
- Build data **models** / **systems**
- Evaluate / decide **analytical tools**

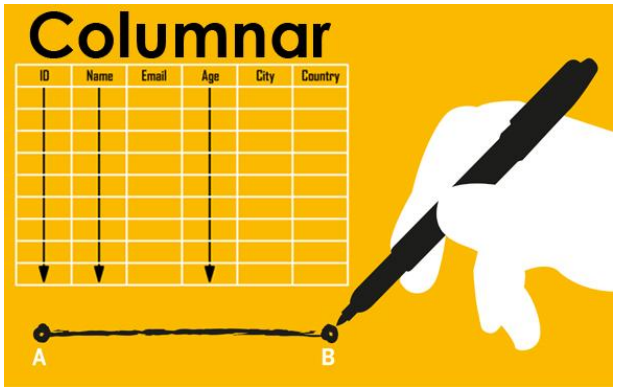
DS Process – Formulating the question

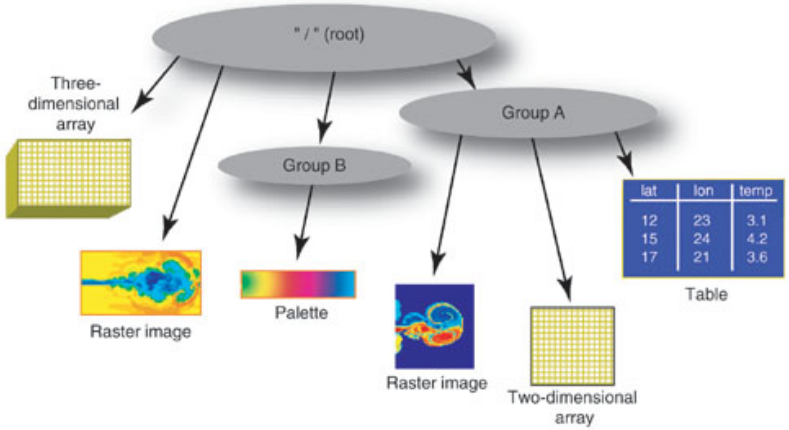
This is a **VERY VERY** important part of the process:

Always strive for **clear**, **well-informed**, **well-scoped** analysis questions right from the start to design your analysis, but also be **agile** and **open to serendipity**!

DS Process – Data collecting, storing, access

- **Storing, accessing** large data
  - **High performance** data storage / access (Hadoop, NoSQL, columnar databases, MongoDB)
  - Using **efficient** data structures, e.g., Hierarchical Data Format HDF
  - Efficient computational mechanisms, e.g., MapReduce





DS Process – Data collection, efficient storing & access

Document Database	Graph Databases
Couchbase MarkLogic mongoDB	Neo4j InfiniteGraph The Distributed Graph Database
Wide Column Stores	**Key-Value Databases**
redis amazon DynamoDB AEROSPIKE riak	accumulo HYPERTABLE INC Cassandra APACHE HBASE Amazon SimpleDB

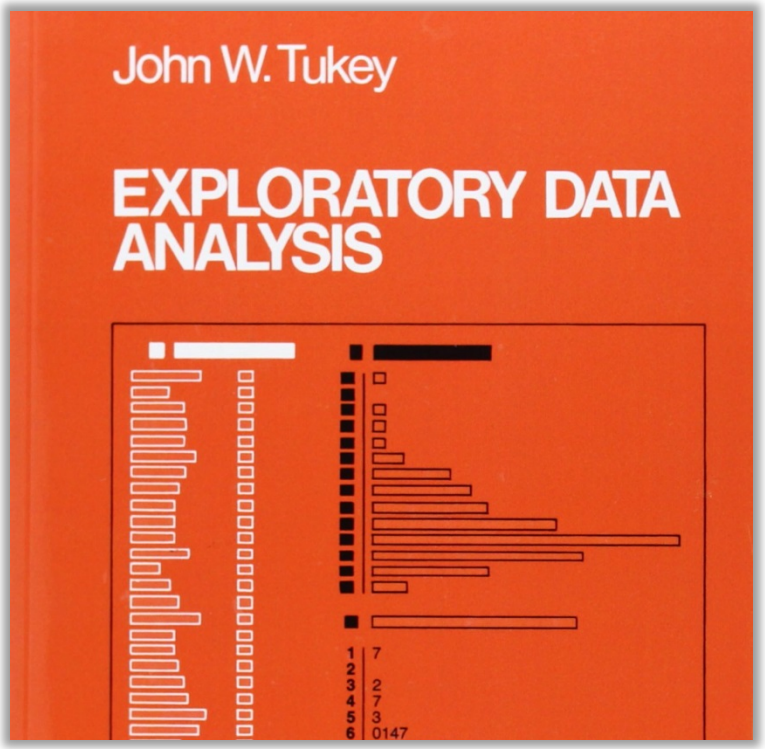
@cloudtxt http://www.aryannava.com

DS Process – Data wrangling & fusion

- **Getting the data ready** to be analysed
- Data is **never perfect** and it is **segregated,** i.e., multiple sources
- Many names: data **wrangling**, data **munging**, data **cleaning**, data **massaging**, data **scrubbing**, **pre-processing**, ….
- Data **fusion**: merging / integrating several data sources
- Handle **missing data**

DS Process – (Exploratively) Analysing data

- **Goal:** Generate / confirm ideas, findings (hypotheses)
- Variety of **analysis tasks** exist
  - Finding **anomalies**
  - Finding **relations**
  - Finding **groups**
  - **Summarizing** information
  - Making **predictions**
  - Understanding **uncertainties**
  - Evaluate **hypotheses**



DS Process – (Exploratively) Analysing data

- Research from many fields
  - Statistics
  - Machine learning
  - Visualisation / visual analytics
  - Knowledge Discovery in Databases (KDD)
  - Data mining
- A selection of analytical tools / methods
    - dimension reduction, clustering, classification, regression, decision trees, neural networks, …
    - Deriving new features
    - **Visual analytics** for interactive exploration
- Need to map methods to tasks

DS Process – Build models

- Build computational models if the problem needs it

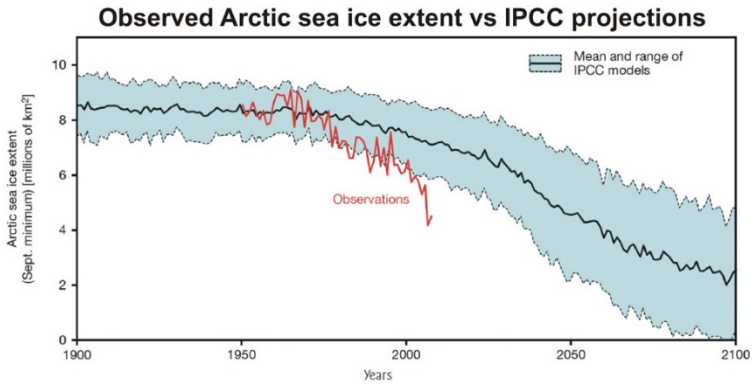
"*A model is our attempt to understand and represent the nature of reality through a particular lens, be it architectural, biological, or mathematical*"

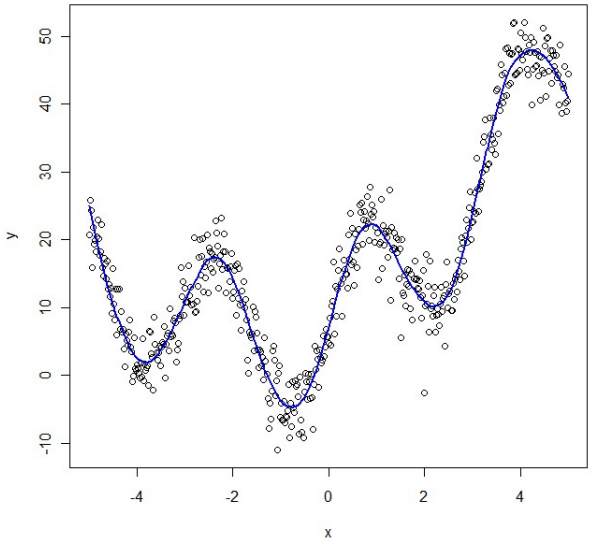
"*A model is an artificial construction where all extraneous detail has been removed or abstracted. Attention must always be paid to these abstracted details after a model has been analyzed to see what might have been overlooked*"

(from Doing Data Science by Schutt and O'Neil , p28)

DS Process – Build models

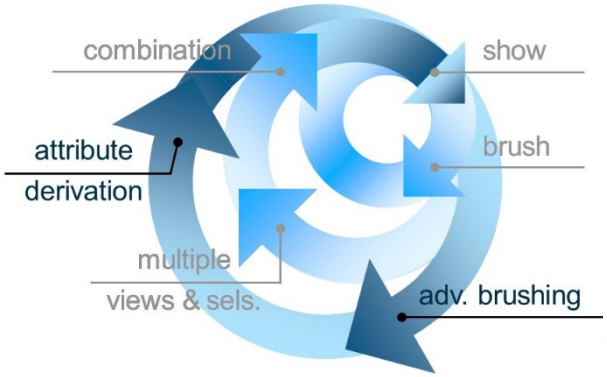
- Statistical models
  - For summarising, representing data
  - For predictions, estimations
- Machine learning methods used, e.g., neural networks
  - Predictive tasks
  - Classification tasks
  - Unsupervised / supervised models



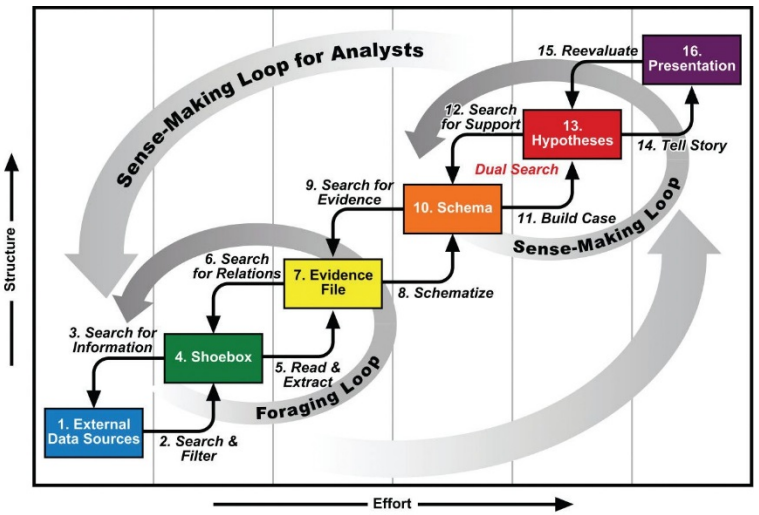


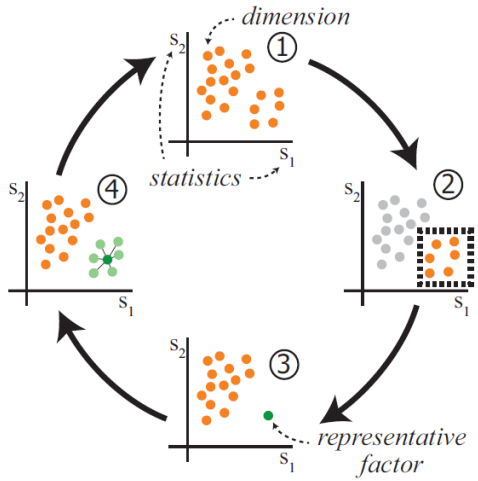
DS Process – ITERATE!

- Iterative process
  - Learn through the process, evaluate your previous steps
  - There might be problems in any steps of the analysis
  - Iterate back to any step and perform tasks again



Hauser, 2013

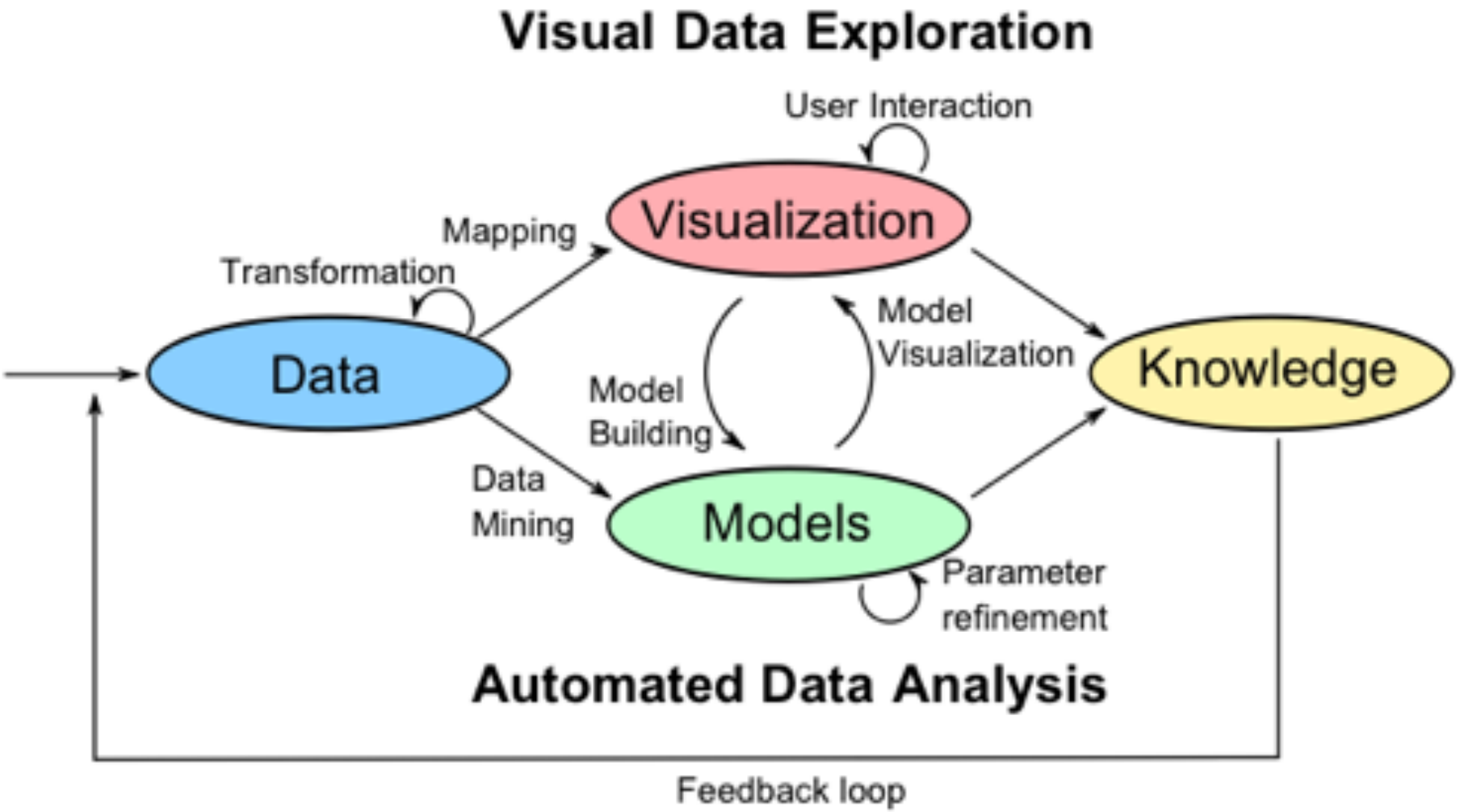




Turkay, 2012

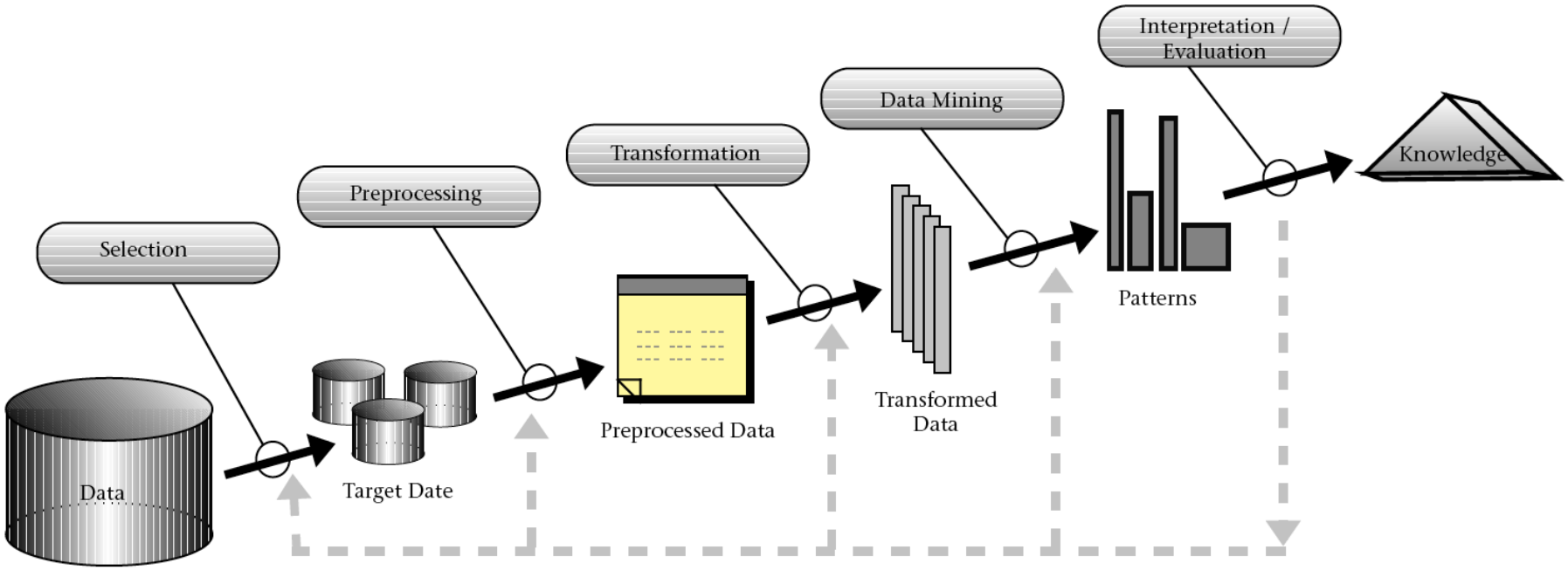
Sense making loop by Pirolli and Card, 2005

Other existing processes – Visual Analytics



Keim, D. A., Kohlhammer, J., Ellis, G., & Mansmann, F. (Eds.). (2010).Mastering the information age-solving problems with visual analytics. Florian Mansmann.

Other existing processes – Knowledge Discovery in Databases (KDD)



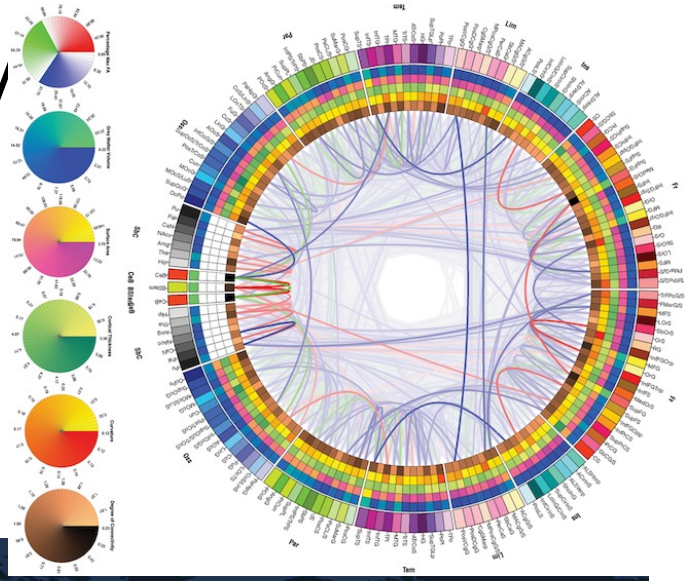
From: U. Fayyad, G. P.-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. AI Magazine, 17(3):37-54, Fall 1996.

DS Process (again)

- Understand domain needs
- Collect & make data available
- Get the data ready for analysis
- Exploratively (and visually) analyse the data
- Model the phenomena (if needed)
- Evaluate findings
- Communicate findings
- ITERATE (from any stage to any other stage)!

DS Application areas

- Most of the businesses nowaday
  - Finance
  - Retail
  - Internet companies
- Social network analysis
- Natural sciences
  - Biology
  - Physics
- Social sciences
- Health
- Digital cities
- ……

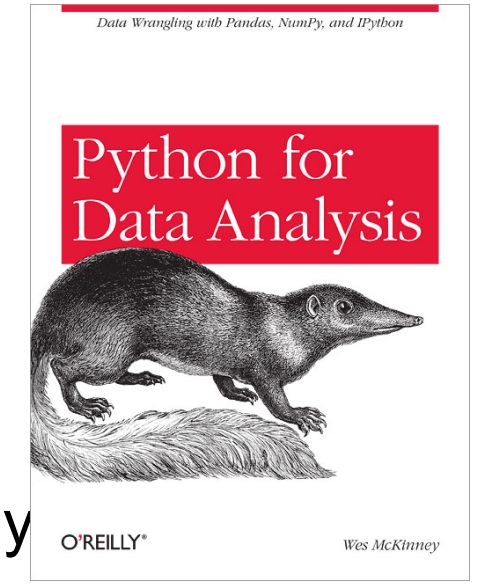


Wrapping up …

Data science is a **process** starting with formulating a question that can be answered with data, and **iteratively** **collecting**, **cleaning**, **analysing** and **modelling** the data, while **communicating** the answers to the relevant audience **along this iteration**.

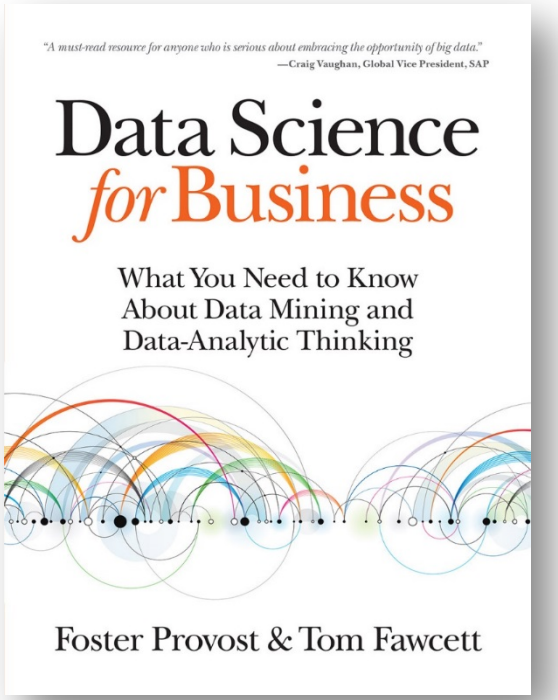
Technologies –off to the lab session

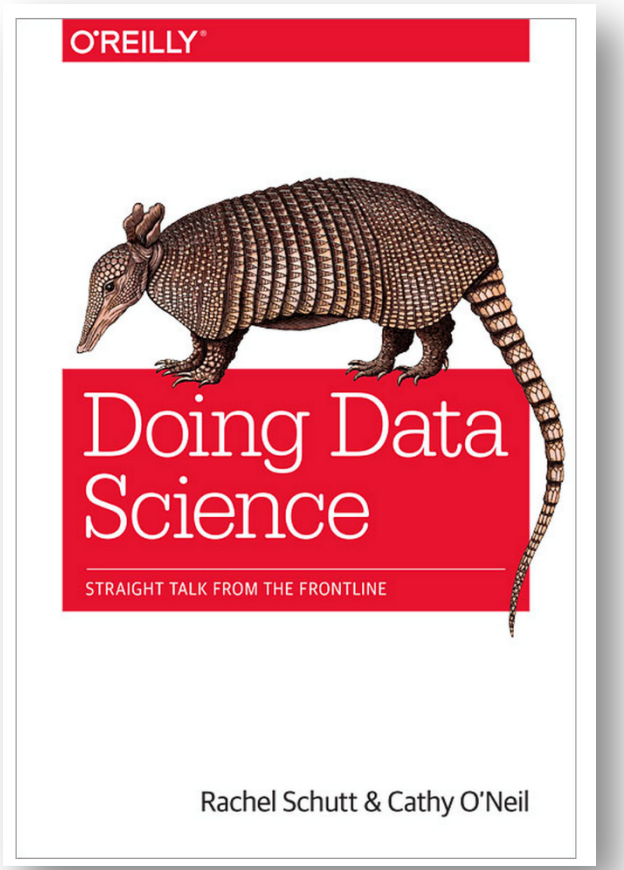
- **Python** (Anaconda)
- **Pandas** for Statistical Computing
- **Scikit-learn** for machine learning
- **Numpy**
- **Matplotlib**
- A good resource:
  - "Python for Data Analysis" by Wes McKinney

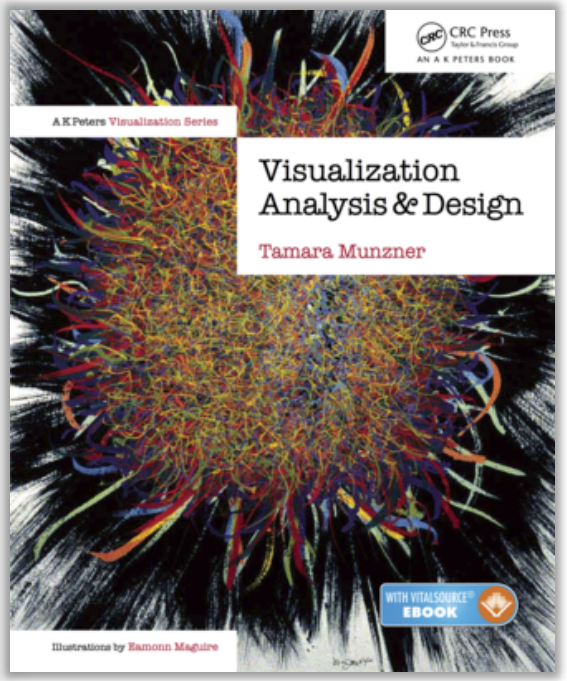


- Python Programming resources:
  - Beginners in Python : http://www.codecademy.com/en/tracks/python
  - Python for Non-programmers: http://en.wikibooks.org/wiki/Non-Programmer%27s_Tutorial_for_Python_3
  - http://www.learnpython.org/
  - https://wiki.python.org/moin/IntroductoryBooks

Further reading …







Some suggested readings

- **What is data science?,** by Mike Loukides : http://radar.oreilly.com/2010/06/what-is-data-science.html
- **Data science as a scientific practice :** Dhar, Vasant. "Data science and prediction." *Communications of the ACM,* 56.12 (2013): 64-73.
- **On Google's influenza epidemic application:** Ginsberg, Jeremy, et al. "Detecting influenza epidemics using search engine query data." *Nature* (2008)
- **On 3V's of data:** Laney, D. (2001), '3D Data Management: Controlling Data Volume, Velocity, and Variety' , Technical report, META Group .
- First chapter called "**What Is Data Science?"** from "*Doing Data Science*", by Schutt and O'neil, 2014.
- **On challenges and analysts roles** : Kandel, Sean, et al. "Enterprise data analysis and visualization: An interview study." Visualization and Computer Graphics, IEEE Transactions on 18.12 (2012): 2917-2926.

Some interesting reads online

- Interviews with Data Scientists: http://www.datasciencenceweekly.org/data-scientist-interviews
- An extensive collection of DS related books: https://www.goodreads.com/shelf/show/data-science
- Nate Silver's FiveThirtyEight blog (good interactive visualisations): http://fivethirtyeight.com/
- …..

Next week …

DATA CHARACTERISTICS & WRANGLING

: Earlier stages of the DS process