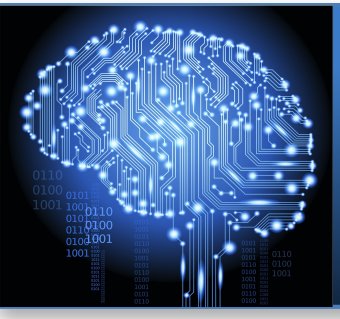


FIT1043 Lecture 1
Introduction to Data Science

Mahsa Salehi*

Faculty of Information Technology, Monash University

Semester 2, 2022

*with material from © Wray Buntine

Outline

- Unit motivation (Why data science?)
- About this Unit
  - Teaching team
  - Unit logistics
- Overview of data science
  - Data science definition
  - Data science process
  - Relevant fields to data science

Motivation for the Unit

from *The Digital Universe of Opportunities*: Rich Data and the Increasing Value of the Internet of Things, April 2014

Data is everywhere!

- Google: processes 24 peta bytes of data per day.
- Facebook: 10 million photos uploaded every hour.
- YouTube: 1 hour of video uploaded every second.
- Astronomy: Satellite data is in hundreds of PB.
- Twitter: 400 million tweets per day.
- In 2020, every person generated 1.7 megabytes of data in just a second.
- Cloud data storage around the world: 200+ Zettabytes by 2025 (*Cybercrime Magazine*).

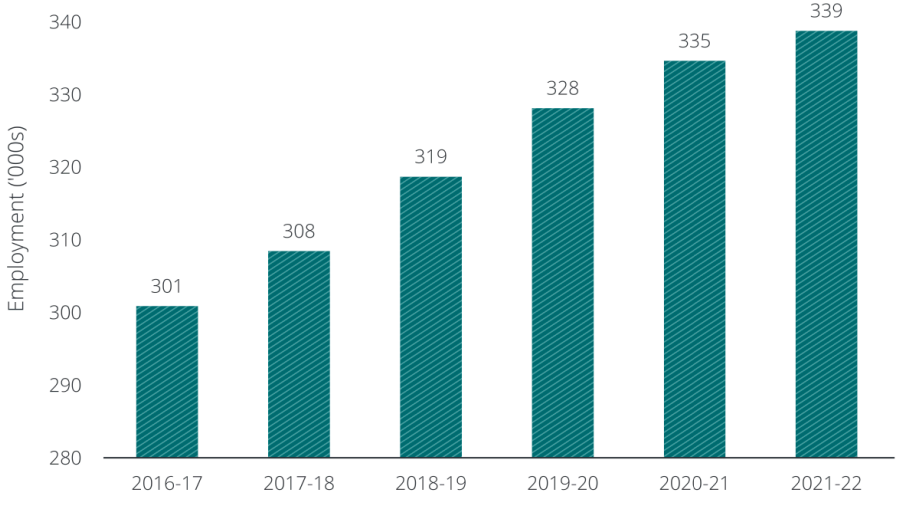
image src: corporatecomplianceinsights.com

Motivation for the Unit

Data Science is in its growth phase:

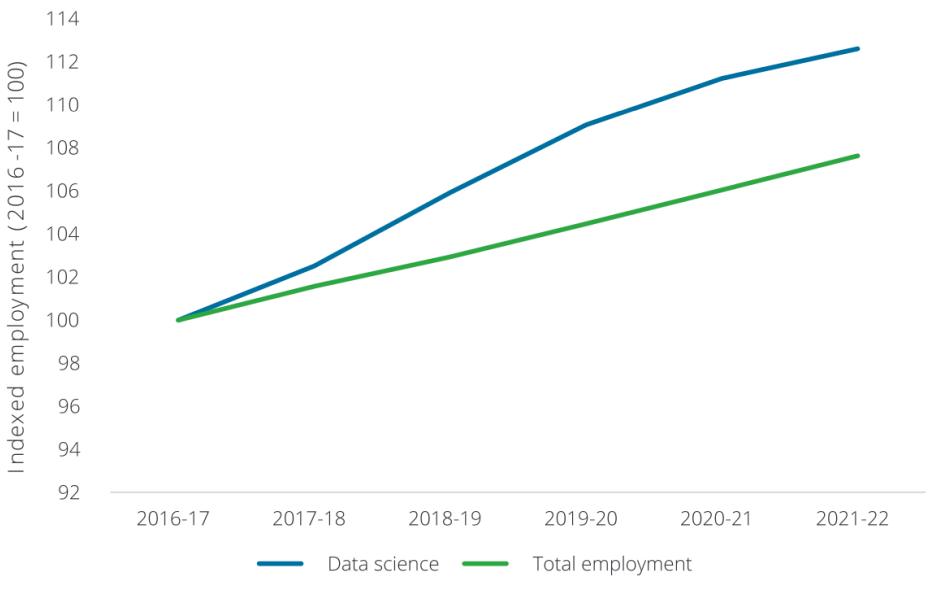
- The demand for skilled data science practitioners in industry, academia, and government is rapidly growing.

Chart 1: Data science employment forecasts, 2016-17 to 2021-22



Source: Deloitte Access Economics (2017)

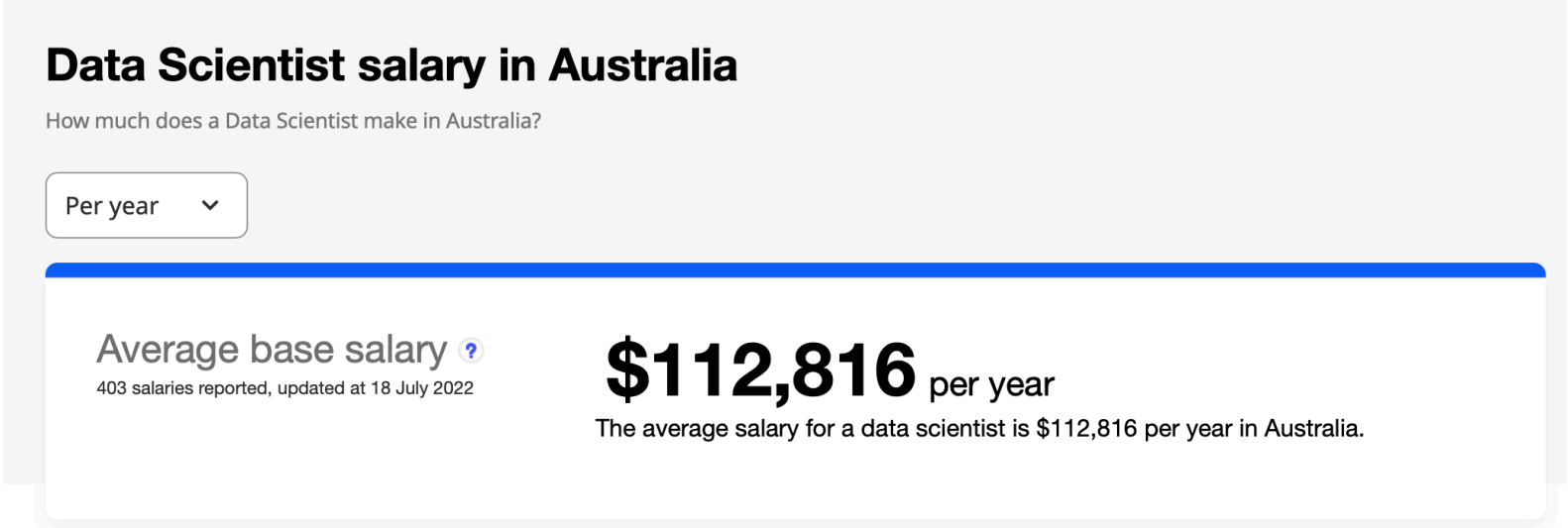
Chart 2: Data science employment and total employment, 2016-17 to 2021-22



Motivation for the Unit

Data Science is very well paying field

- Average base salary of almost $113,000 per year in Australia in 2022



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

- background and context
- leading review articles, lectures, introductions

Teaching Team

Unit email address: fit1043.clayton-x@monash.edu
Lecturer: Mahsa Salehi
Consultation: 2:30-3:30pm Mondays
Office 264, 20 Exhibition Walk, Clayton (Woodside building)

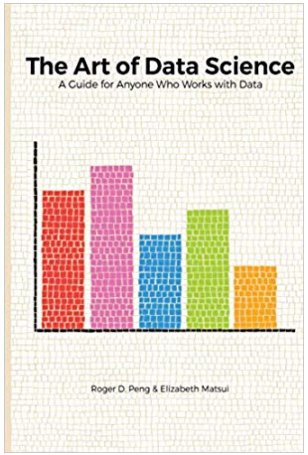
Dr. Mahsa Salehi	Lecturer	mahsa.salehi@monash.edu
Dr. Chris Yun	Admin TA	chris.yun@monash.edu
Dr. Heshan Kumarage	Admin TA	heshan.kumarage@monash.edu
Dr. Dilini Rajapaksha	TA	dilini.rajapakshahewaranasingh@monash.edu
Dr. Jesmin Naher	TA	jesmin.nahar@monash.edu
Md Mohaimen	TA	md.mohaimen@monash.edu
Mike Wang	TA	qizhou.wang@monash.edu
Navid Foumani	TA	navid.MohammadiFoumani@monash.edu
Sehrish Iqbal	TA	Sehrish.Iqbal@monash.edu

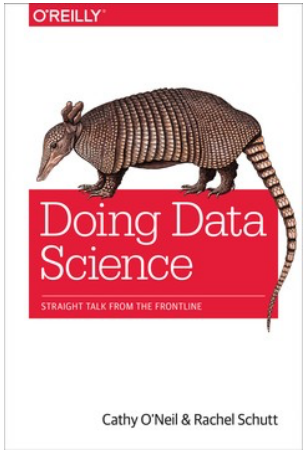
Note: Consultation times will be posted on Moodle, no consultation on week 1

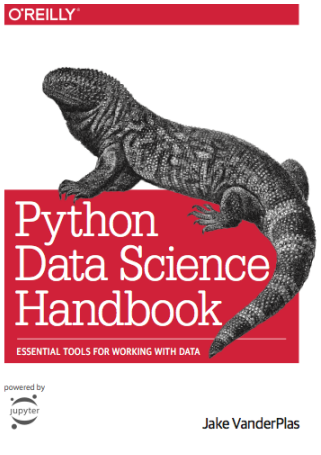
Resources

1. Moodle contains
   - Unit Information, Assessments and Discussion Forums
   - Lecture Notes: contain active links to recommended videos & readings

2. 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
   - Doing Data Science by Rachel Schutt and Cathy O'Neil
   - Python Data Science Handbook by Jake vanderPlas







Resources

3. review of Ed Lessons
   - LOTS of additional resources and exercises
   - get the big picture from articles/videos

4. be aware also of the:
   - library services available
   - special consideration policies
   - disability support available

Prerequisites

You will need:

- high school level of mathematics and statistics
- basic programming
- 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
  - Amazon, Google, Twitter, ...

Getting Started

How these classes are run

- 2 hour lecture, Monday 16:00 – 18:00
- 2 hour tutorial (laboratory), check Allocate+
- watch videos & read background material between classes
- bring a device to lectures to participate in class activities
- prepare for tutorials (labs)

Unit Schedule

Week	Activities	Assignments
1	Overview of data science	Weekly quiz
2	Introduction to Python for data science	
3	Data visualisation and descriptive statistics	
4	Data sources and data wrangling	
5	Data analysis theory	Assignment 1
6	Regression analysis	
7	Classification and clustering	
8	Introduction to R for data science	
9	Characterising data and "big" data	Assignment 2
10	Big data processing	
11	Issues in data management	
12	Industry guest lecture	Assignment 3

Assessment

Assessment task	Value	Due date
Assignment 1	10%	Week 5
Assignment 2	20%	Week 9
Assignment 3	15%	Week 12
Weekly quiz	5%	Weeks 2-11
Examination 1	50%	To be advised

- Assignments 1-3 coding tasks based on Python/R/bash subsets covered in lectures and tutorials
- Active participation:
  - All onshore students are expected to attend the on-campus lectures and tutorials (highly recommended).
  - The weekly quizzes examine the material covered within the lectures and tutorials. They will be open after the lectures on Mondays 6pm and will close the following Wednesdays 6pm (after 48 hours).
  - Bring your device, participate using FLUX, instructions next
- Exam based on material covered in lectures and tutorials

Instructions to participate in the poll (using FLUX)

- Visit https://flux.qa on your phone, tablet or laptop
- Enter your email address
- Log in using your Monash account details
- Click the + symbol in the top right hand corner
- Enter the code (**5NKWED**)
- Answer questions when they pop up
- That's it ☺
- Download a copy of instructions

To participate, go to

flux.qa/5NKWED

FLUX Question: Your Background

1. What programming language are you most experienced in?

2. What kinds of data are you familiar with?

Learning Outcomes (Week 1)

By the end of this week you should be able to:

- Explain what is data science and Drew Conway's Venn diagram
- Comprehend the usefulness of machine learning
- Explain different components of a data science process
- Differentiate data science from other related disciplines
- Learn how to install and start coding in Python with Jupyter Notebook

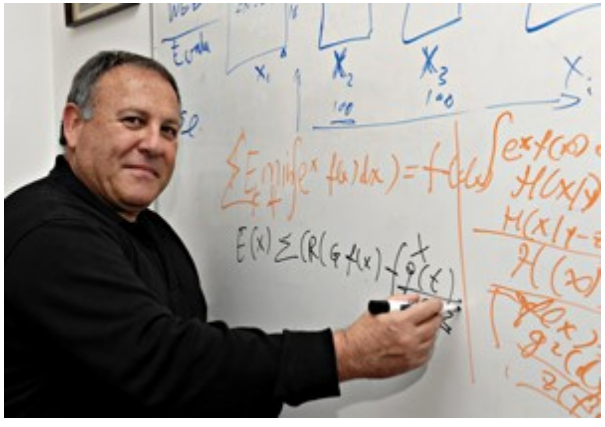
Overview of Data Science

a quick overview of the context

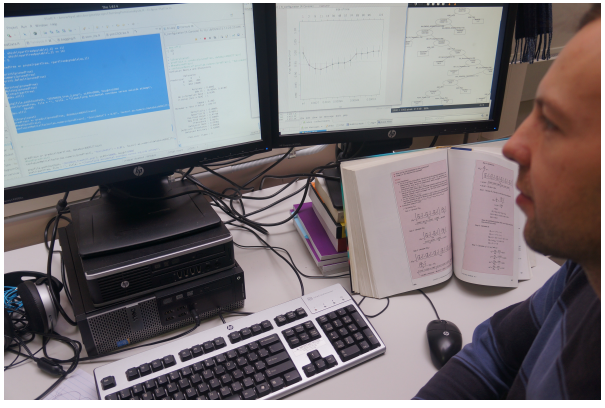
FLUX Question : Who are the Data Scientists?

person A

person B



person C



person D

What is Data Science?

how can we define data science?

Defining Data Science

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!*

What is Data Science?

Definitions: from Wikipedia

Data Science is the extraction of knowledge from data, which is a continuation of the field 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.

What is Data Science?

A quote from *Hal Varian* (From What is Data Science?)

> *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.*

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

FLUX Question

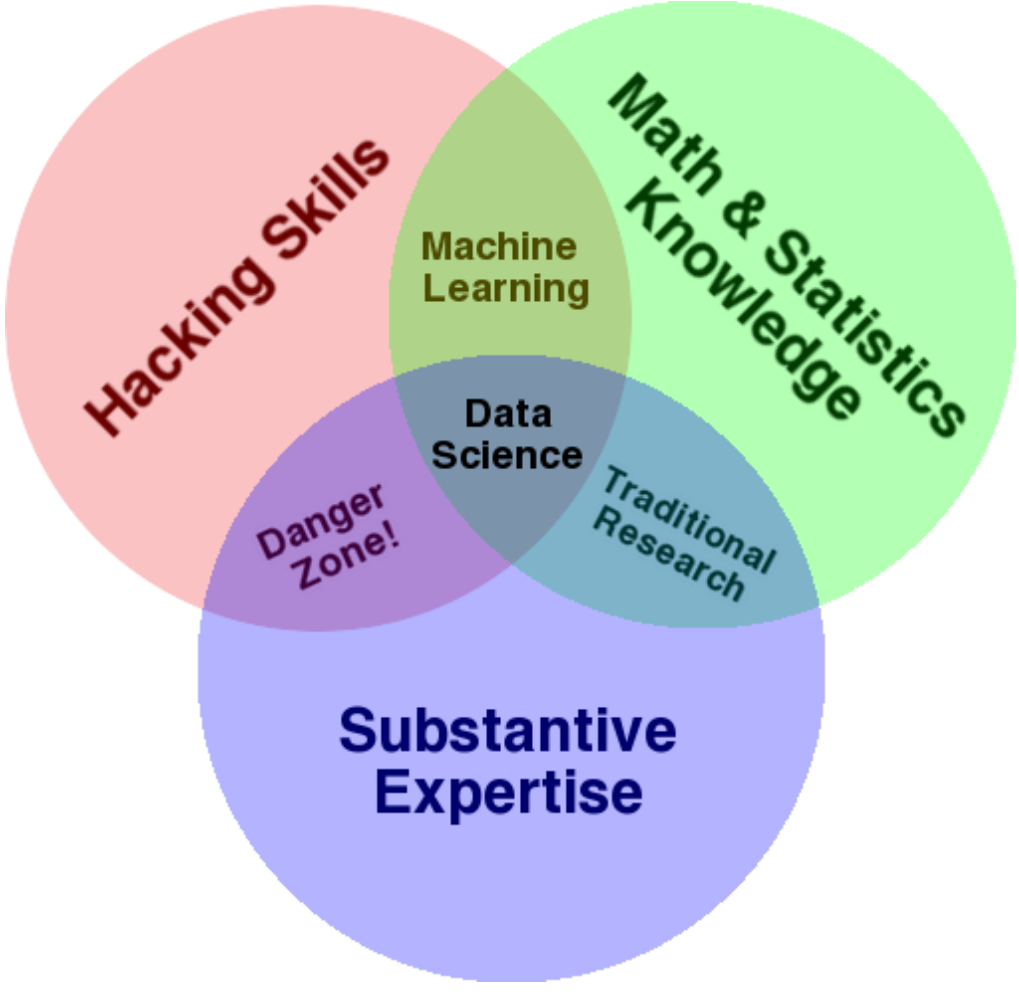
Which of the following data science definition you like most?

Data Science is

A. machine learning on big data
B. extraction of knowledge/value from data through the complete data lifecycle process
C. almost everything that has something to do with data: collecting, analyzing, modeling, etc, yet the most important part is its applications — all sorts of applications

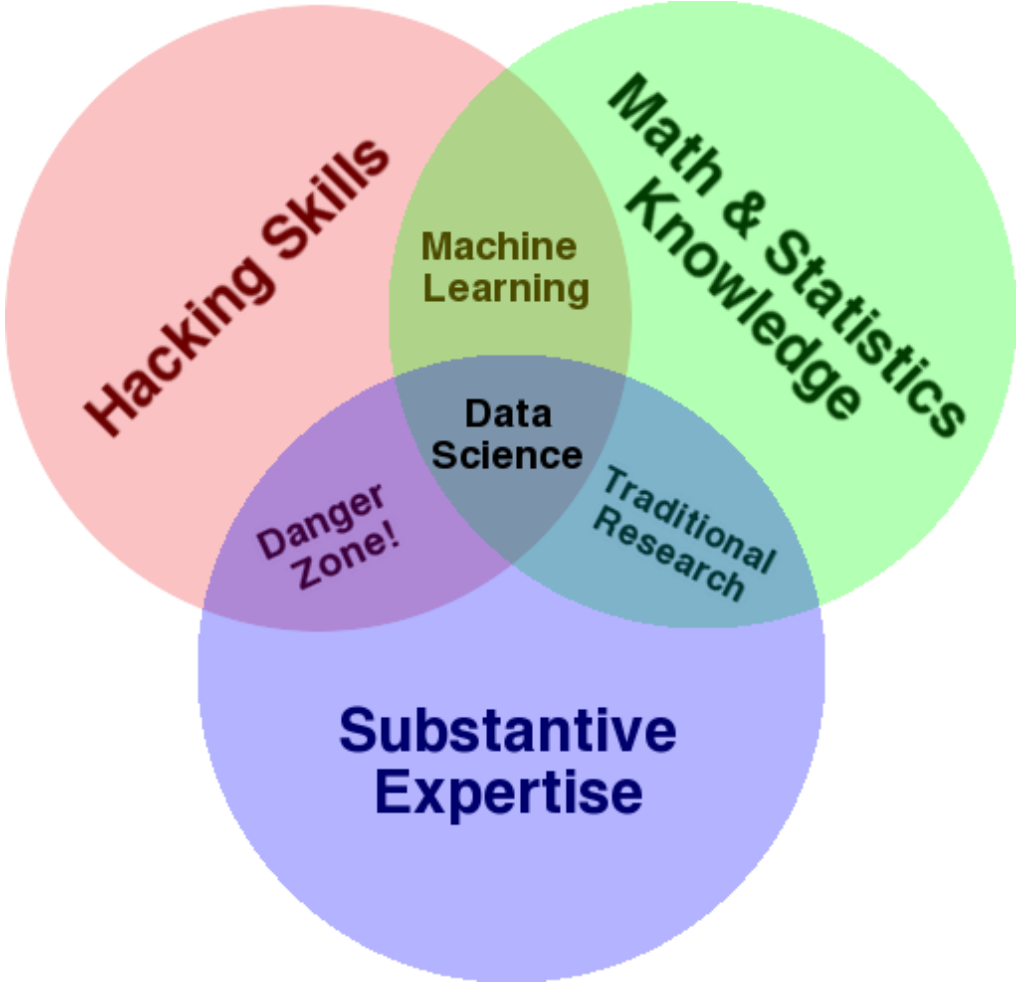
Data Science Venn Diagram

Drew Conway's Venn diagram of data science



Data Science Venn Diagram

Drew Conway's Venn diagram of data science



- Combination of different skill sets
- Diverse skills are needed

Data Science Examples

Some famous data science projects and investigations:

- Google's spell checker and *translation engine*

FLUX Question

Provide another data science example.

Data Science Examples

Some famous data science projects and investigations:

- Google's spell checker and *translation engine*

Other examples to explore:

- Amazon.com's *recommendation engine*
- Public health: *"saturated fat is not bad for you after all"*
- Microsoft's *predictive analytics for traffic*

Defining Machine Learning

Unlike Data Science, the definition for Machine Learning is better understood and more agreed upon:
Machine Learning is concerned with the development of algorithms and techniques that allow computers to *learn*.

- concerned with building computer programs that can learn, oftentimes with computational output
- but the underlying theory is statistics

see *A Gentle Guide to Machine Learning*

Why use Machine Learning?

Machine learning is useful when:

- Human expertise is not available
  e.g. Martian exploration

- Many solutions need to be adapted automatically
  e.g. user personalisation

- Humans are expensive to use for the work
  e.g. handwritten zipcode recognition

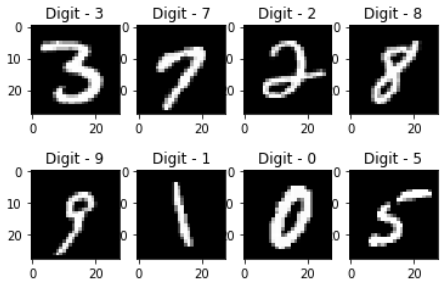


image src: theconversation.com, meduim.com, blog.prioridata.com

Why use Machine Learning?

Machine learning is useful when:

- Situation changes over time
  e.g. junk email

- There are large amounts of data
  e.g. discover astronomical objects

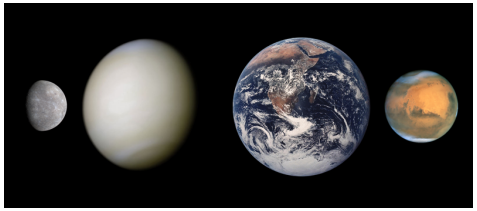


image src: lifewire.com, clrealyexplained.com

Why use Machine Learning?

- because you do not want to be this poor guy!
- sifting through all the data by hand

FLUX Question

Which of the following is real world applications of Machine Learning?

A. Video Games
B. Self-driving cars
C. Spam filtering
D. Predictions
E. All of the options

The Data Science Process

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

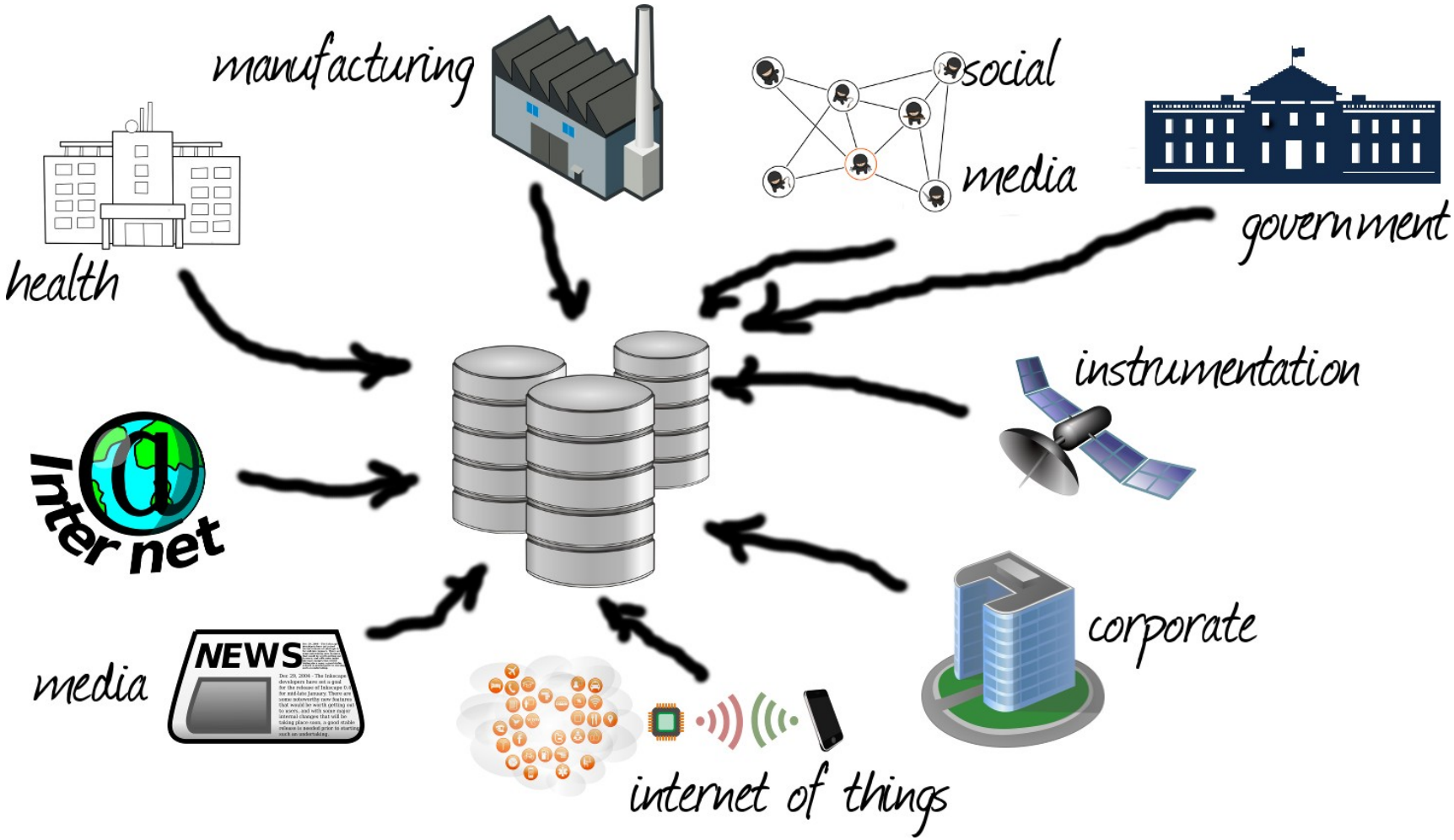
- Many different tasks come together to complete a Data Science project
  - a data scientist should be familiar with most, but doesn't need to be an expert in all
- Not all are labelled as Data Science
  - some from other field such as computer engineering, business, ...

1. Pitching ideas for data science projects to investors/managers.

image src: "Young Business Man Holding a Tablet" by Pic Basement

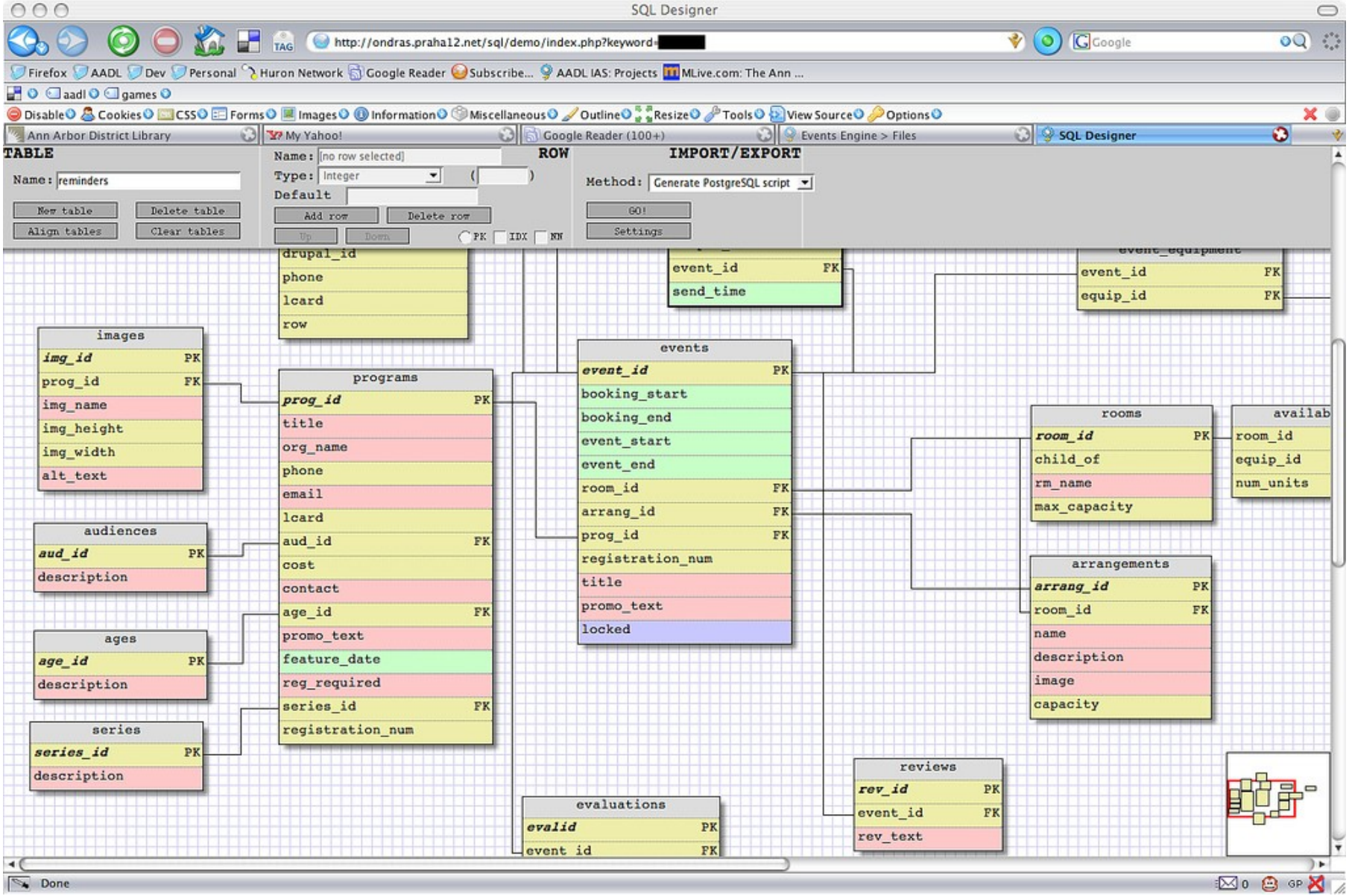
2. Collecting data: researchers preparing to x-ray a patient.

image src: Stephen Ausmus acquired from USDA ARS, public domain.



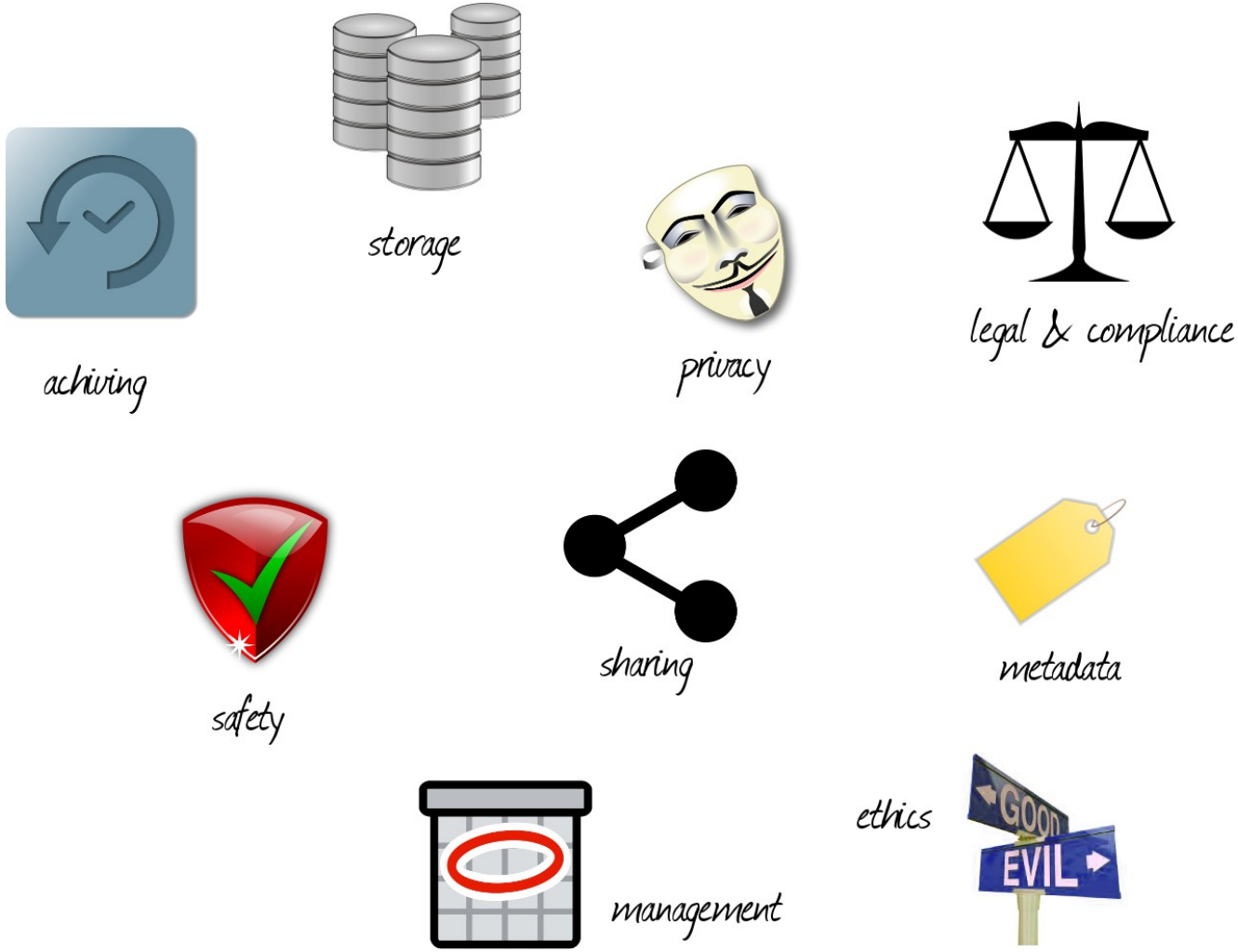
3. Integration: Data can come from many different sources.

icons from by Openclipart.org, public domain



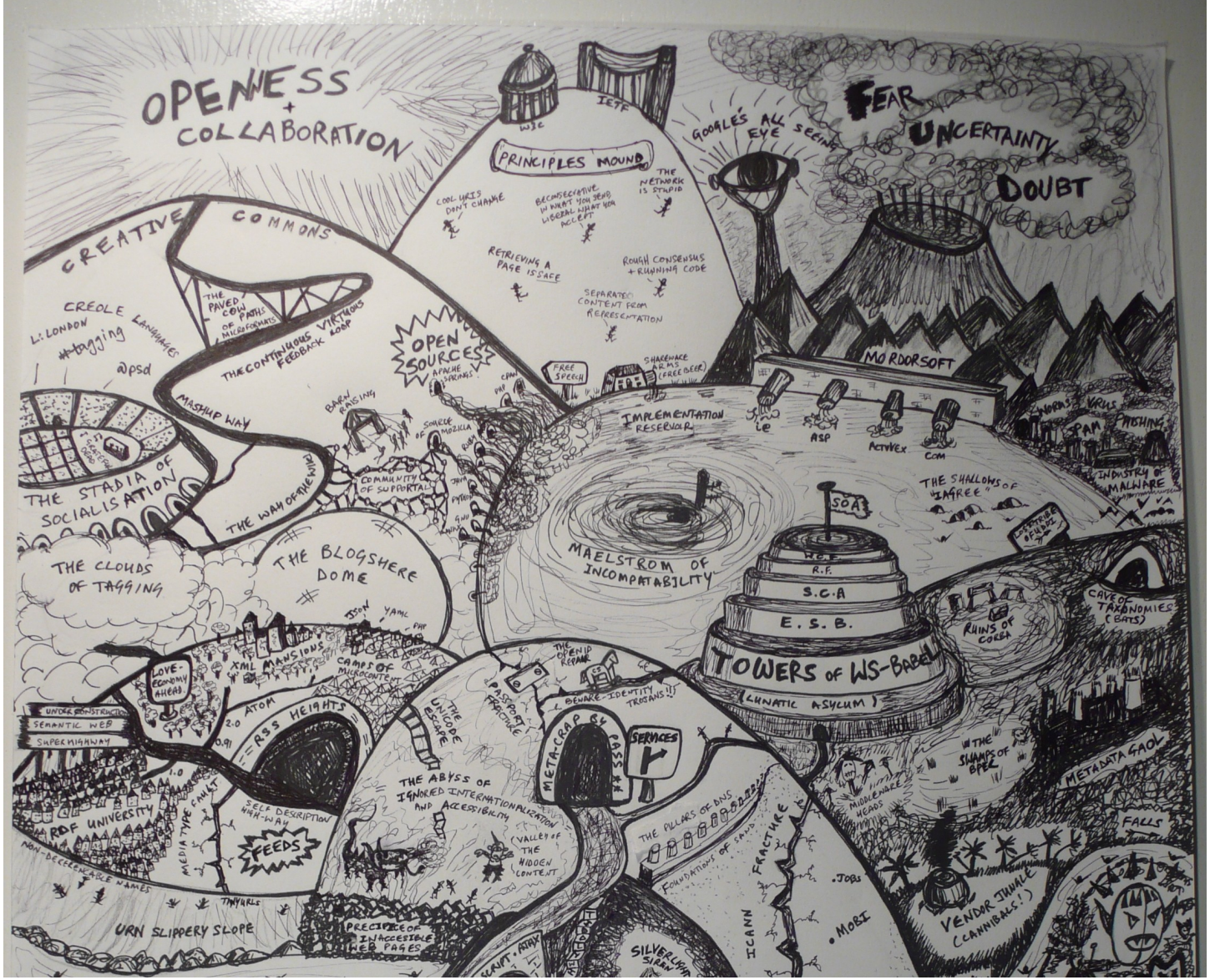
4. Interpretation: e.g. data can be described using a database schema.

image src: Eric, Sql Designer



5. Governance: caring for the data and its subjects.

icons from by Openclipart.org, public domain; Good and Evil by AJC ajcann.wordpress.com

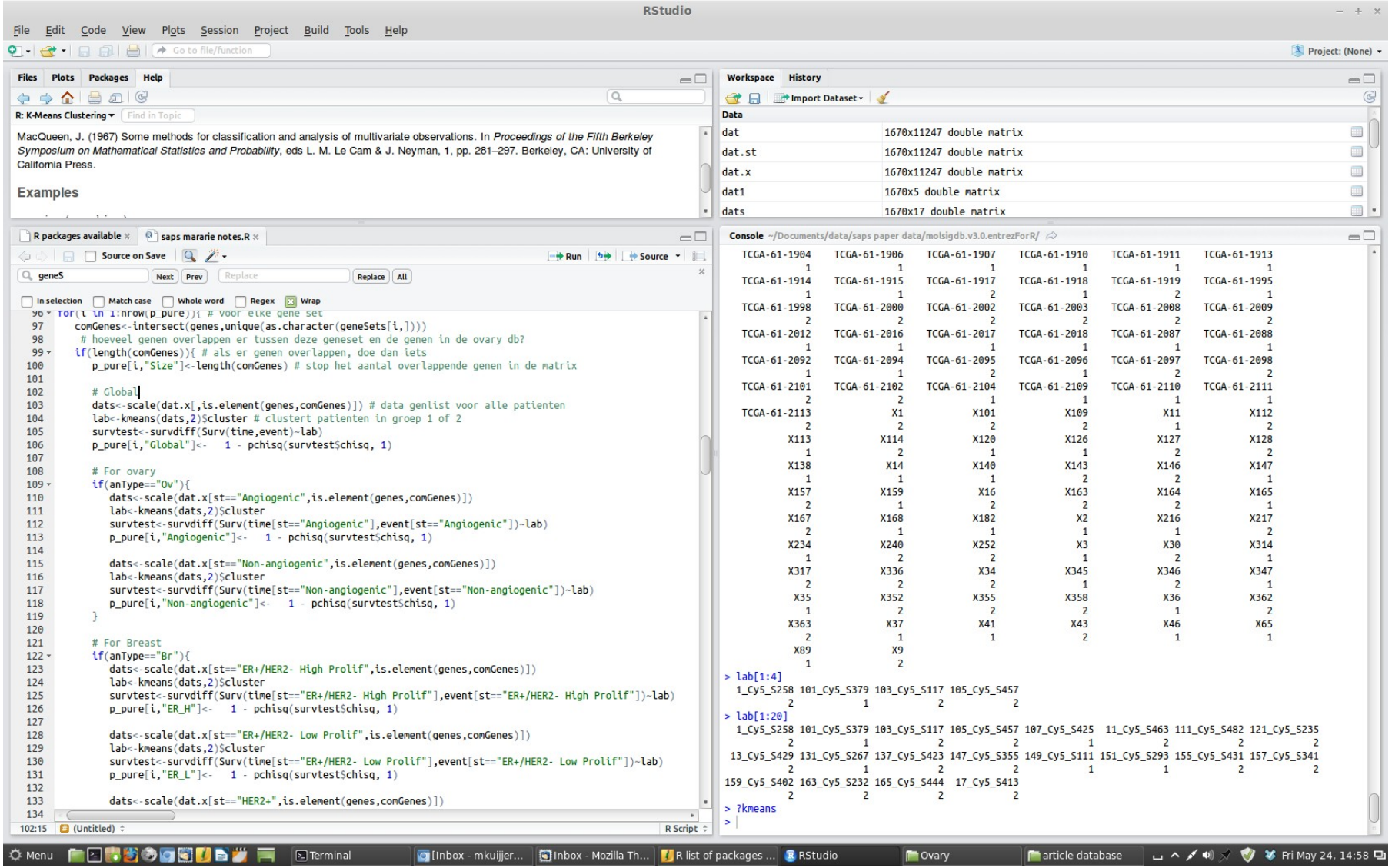


5. Governance: managing data standards and formats

"The Web is Agreement" cropped, by Paul Downey

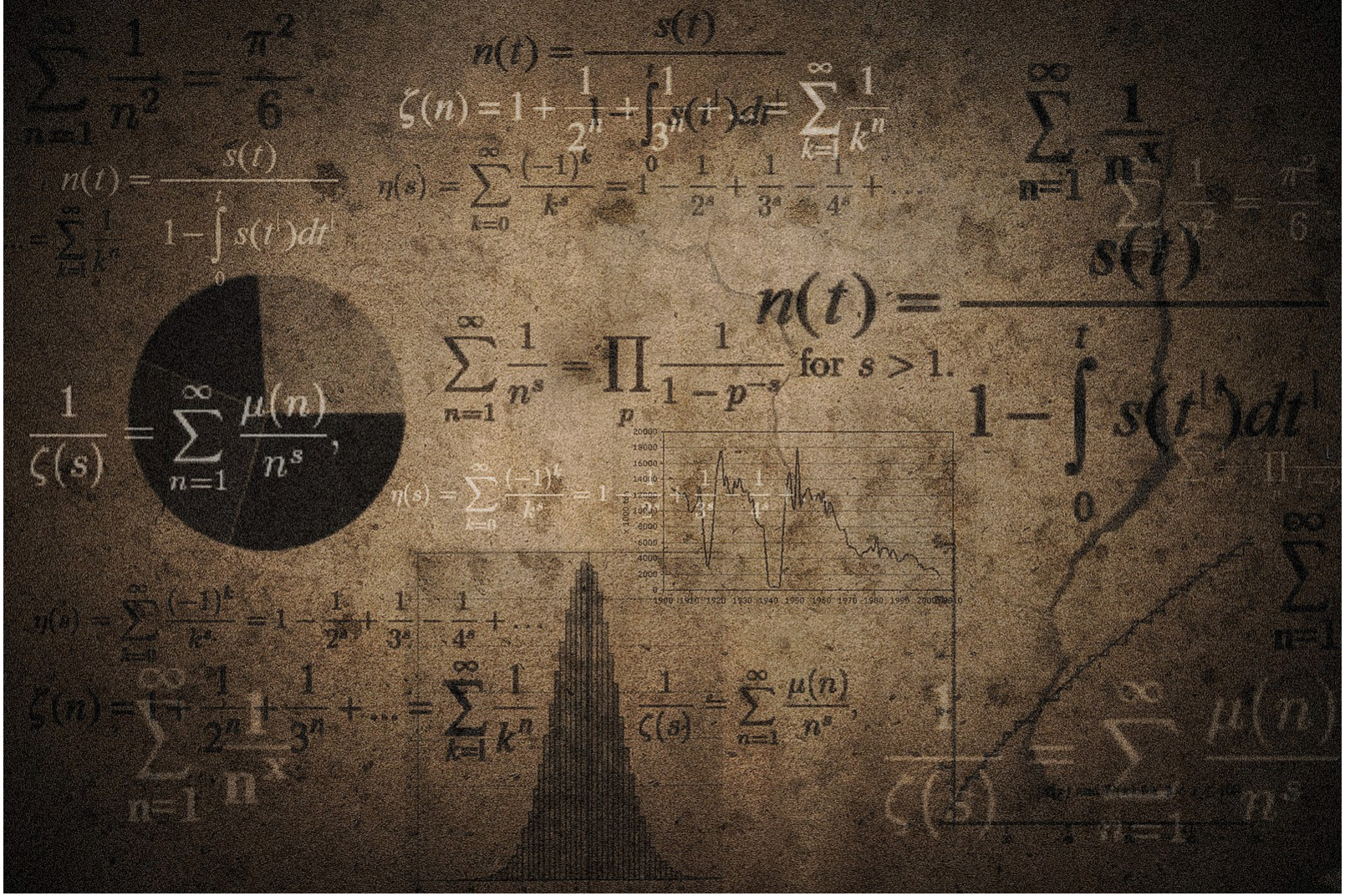
6. Engineering: Data engineers make the back-end work

image src: by Intel Free Press



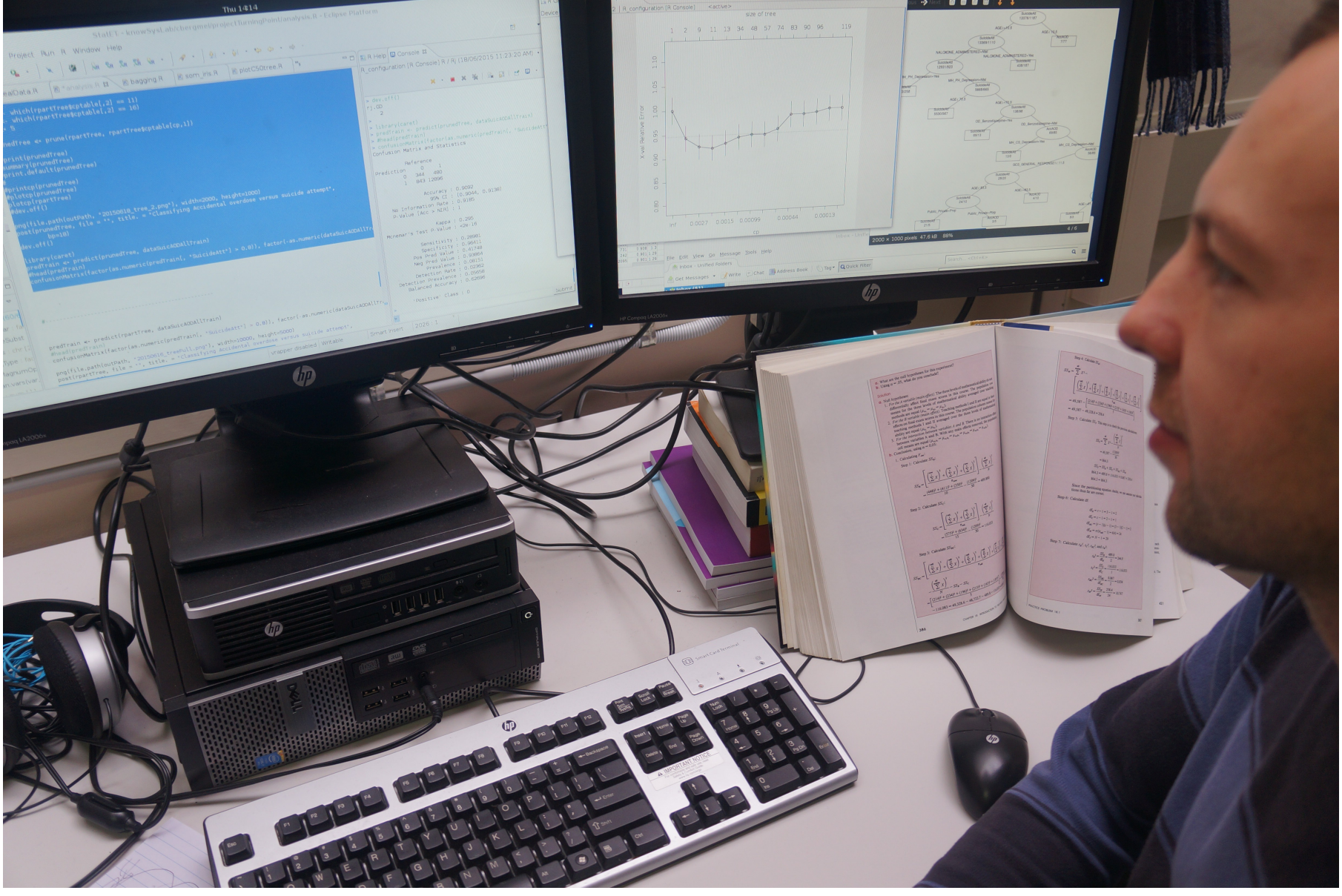
7. Wrangling: Inspecting and cleaning the data.

image src: "rstudio" by marari

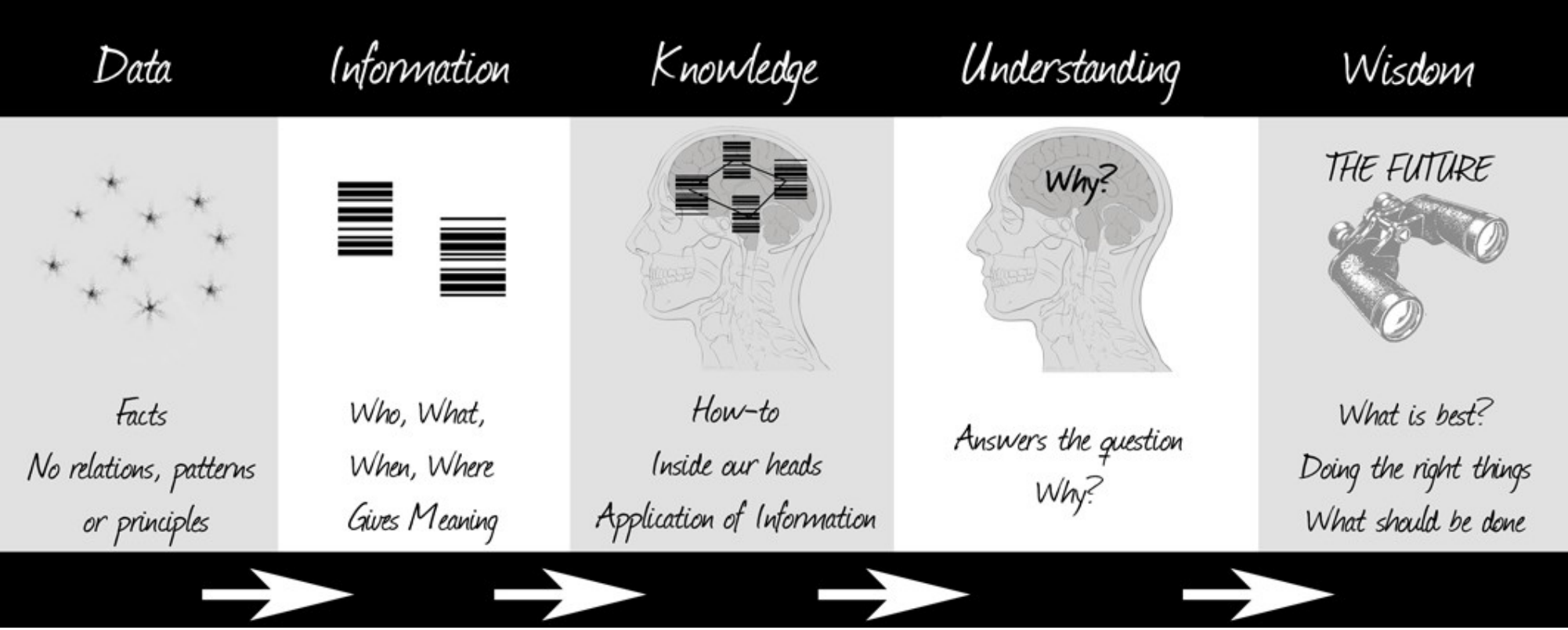


8. Modelling: Proposing a conceptual / mathematical / functional model.

image src: "Mathematics" by Tom Brown



8. Modelling: Analyst building models with his favourite tool.



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

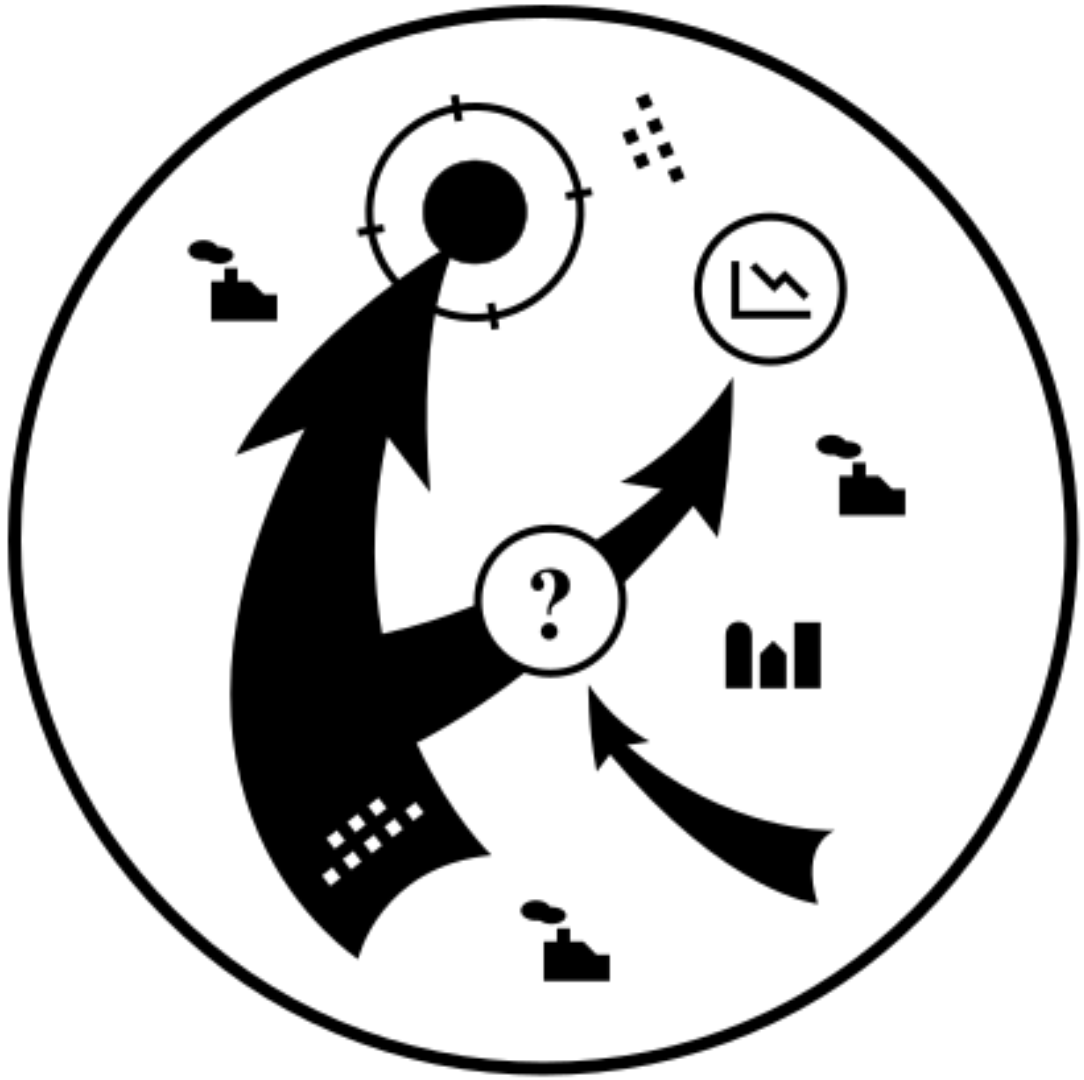
image src: "From Data to Wisdom" by Nick Webb

9. Visualisation: Visualising data to interpret it and present results.

image src: Stephen Ausmus acquired from USDA ARS, public domain

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

image src: "Visualization Matrix" cropped, by Lauren Manning



10. Operationalization: putting the results to work.

image src: "Illustration of Strategy“ by Denis Fadeev

FLUX Question

Using a short phrase or word, which activity in data science process is the most interesting to you.

The Data Science Process:
Our Standard Value Chain

our model of the process

Data Science Project Tasks

Collection: getting 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

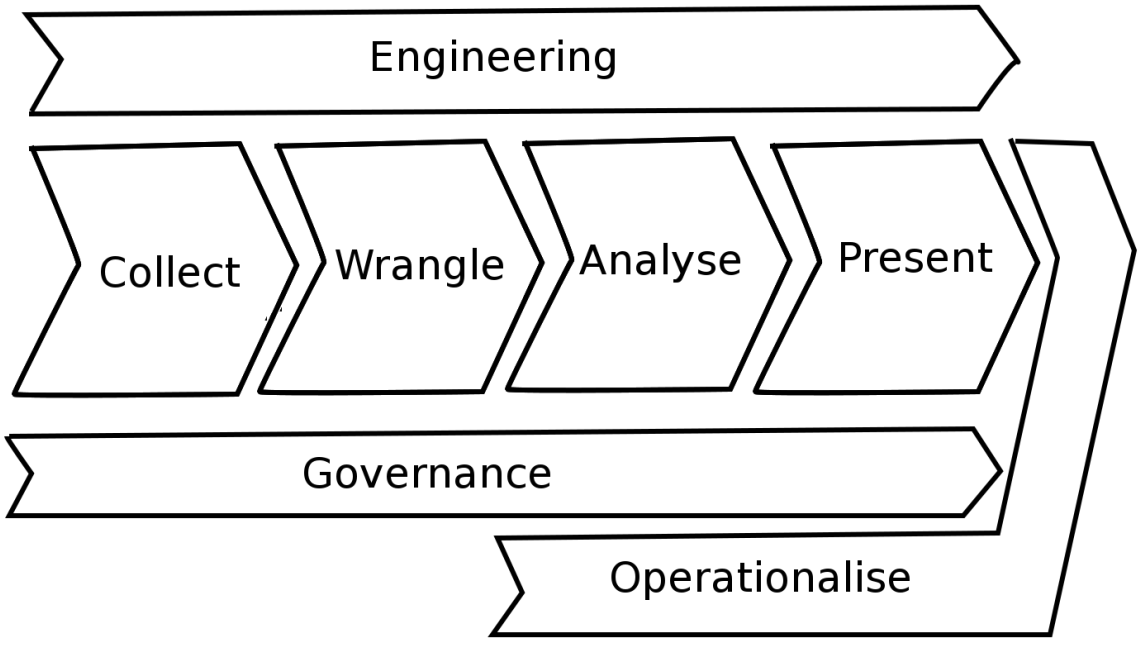
Operationalisation: putting the results to work, so as to gain benefits or value

We call this the Standard Value Chain.

Data Science Process

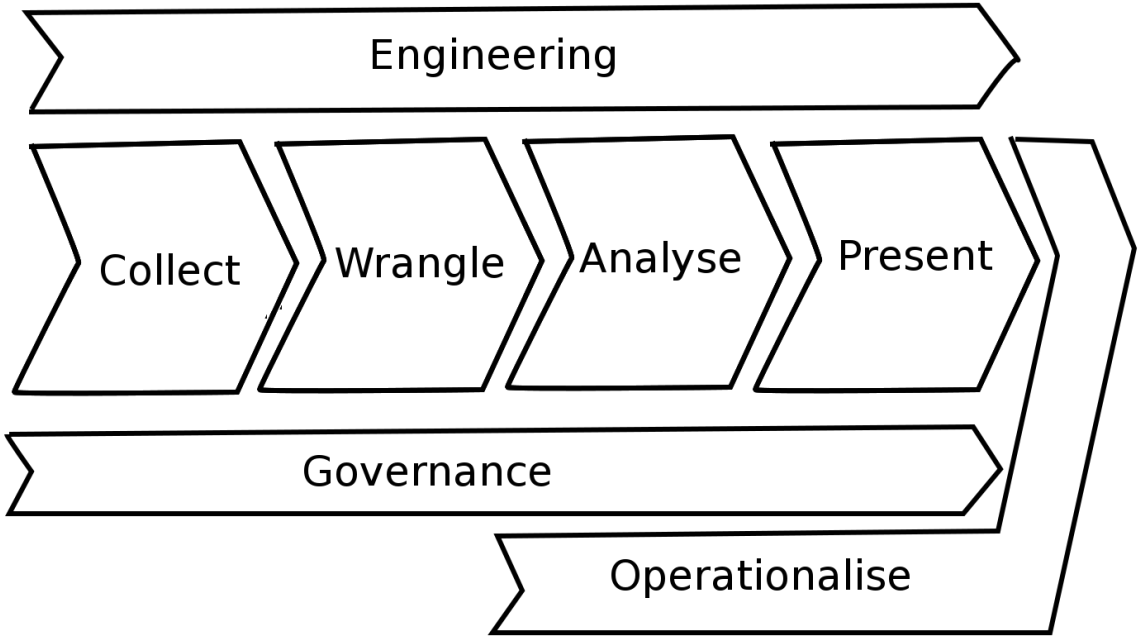
from *Doing Data Science* by Schutt and O'Neil, 2013, (available digitally through library)

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

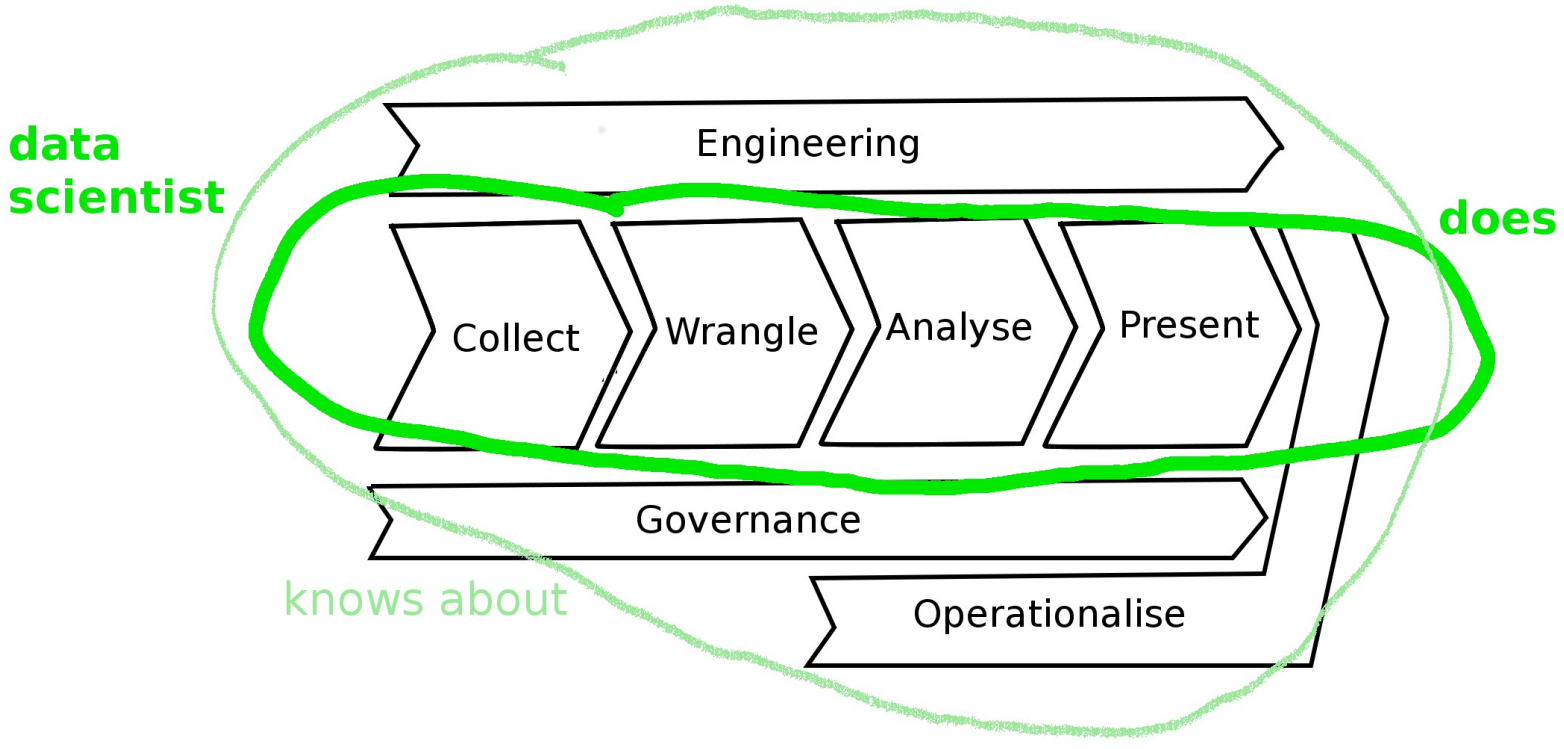


Data Science Process

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

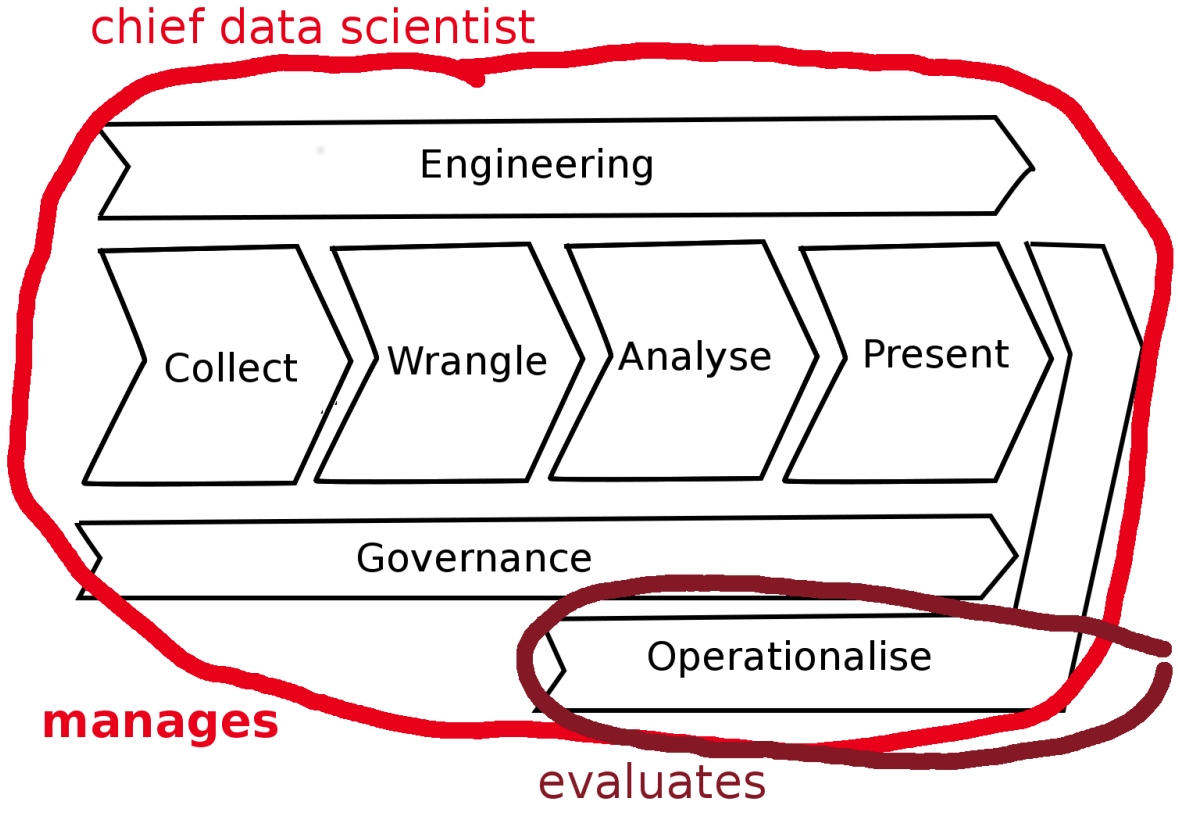


Data Science Process



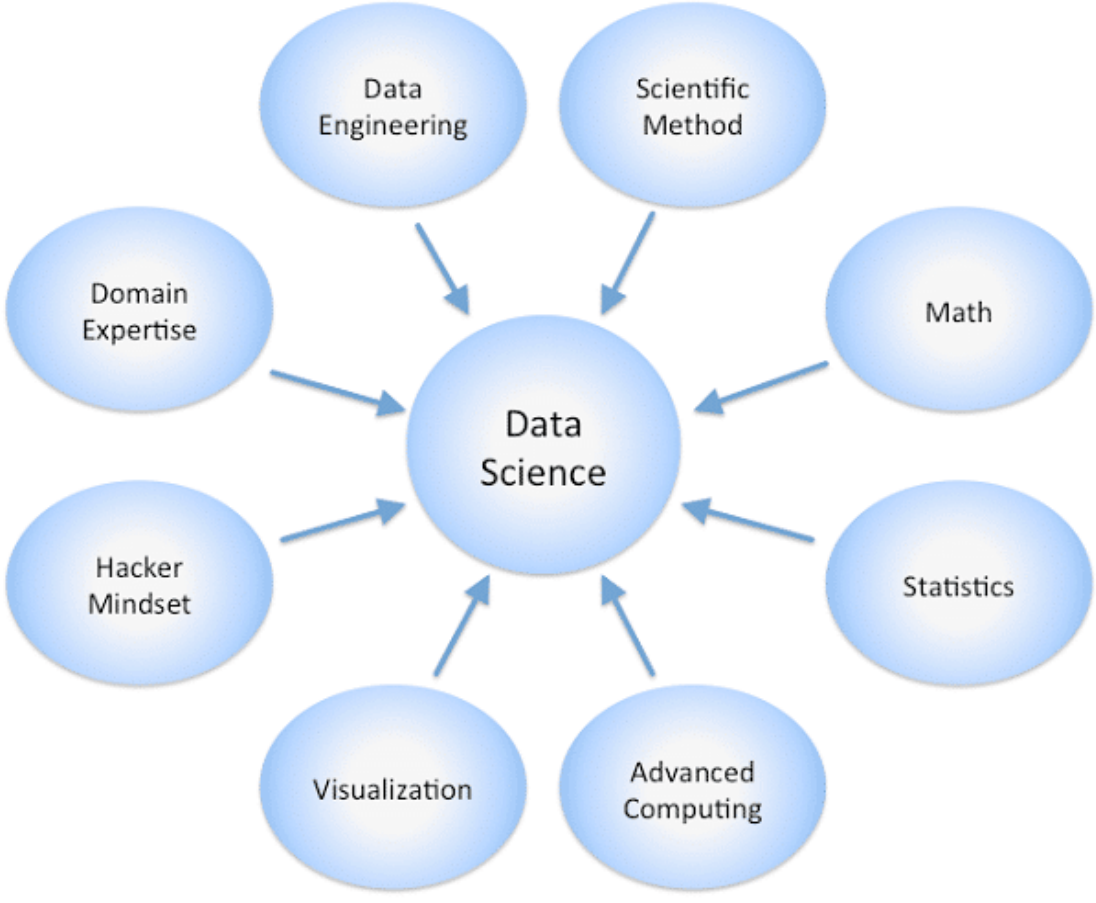
Data scientist ::= addresses the data science process to extract meaning/value from data

Data Science Process



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

Relationship of Data Science to Other Disciplines



Related: Data Engineering

building scalable systems for storage, processing data

- *e.g.* Hadoop
- databases, distributed processing, datalakes, cloud computing, GPUs, wrangling, ...

Related: Data Analysis

performing analysis and understanding results

- *e.g.* R and Microsoft Azure Machine Learning
- machine learning, computational statistics, visualisation, ...

Related: Data Management

managing data through its lifecycle

- ethics, privacy, curation, backup, governance, ...

Tutorial/Lab week 1

- Guide to install Anaconda and Python

- Be prepared for Lecture 2, bring your device (Introduction to Python)

Home Activities

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

Home Activities

- 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*