MONASH University

FIT1043 Introduction to Data Science

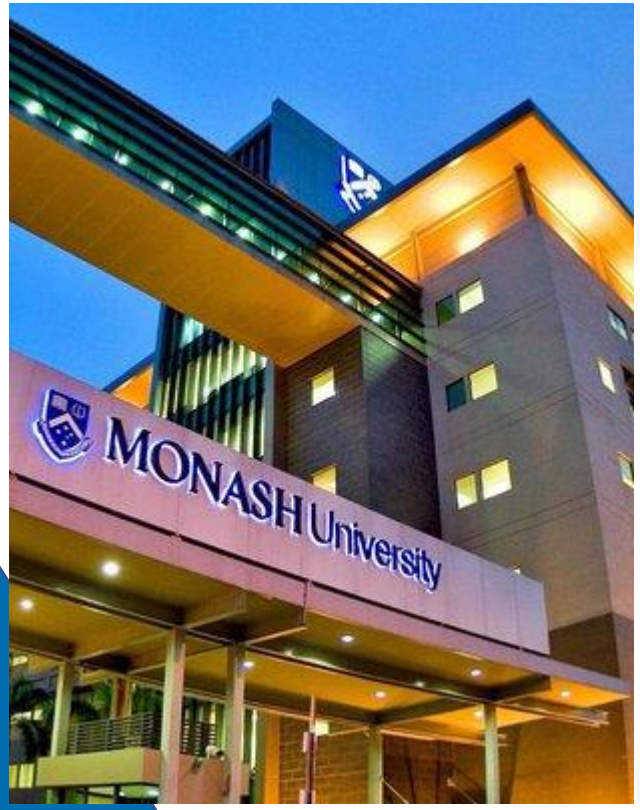
Week 1, Data Science Process

Dr. Sicily Ting Fung Fung

School of Information Technology
Monash University Malaysia

Data Science Process

ePub Section 1.3



MONASH University

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
    - To be achieved in your tutorial / laboratory session

The Data Science Process

ePub Section 1.3

What happens in a Data Science project?

- Illustrating the process
  - A quick walkthrough illustrating the s
- The standard value chain
  - Our model of the process

What is Data Science?

Quote from Hal Varian

The ability to **take data** and;

- 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.

MONASH University

The Data Science Process

Illustrating the 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, ...

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

1. Pitching ideas:

For data science projects to investors / managers

MONASH University

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

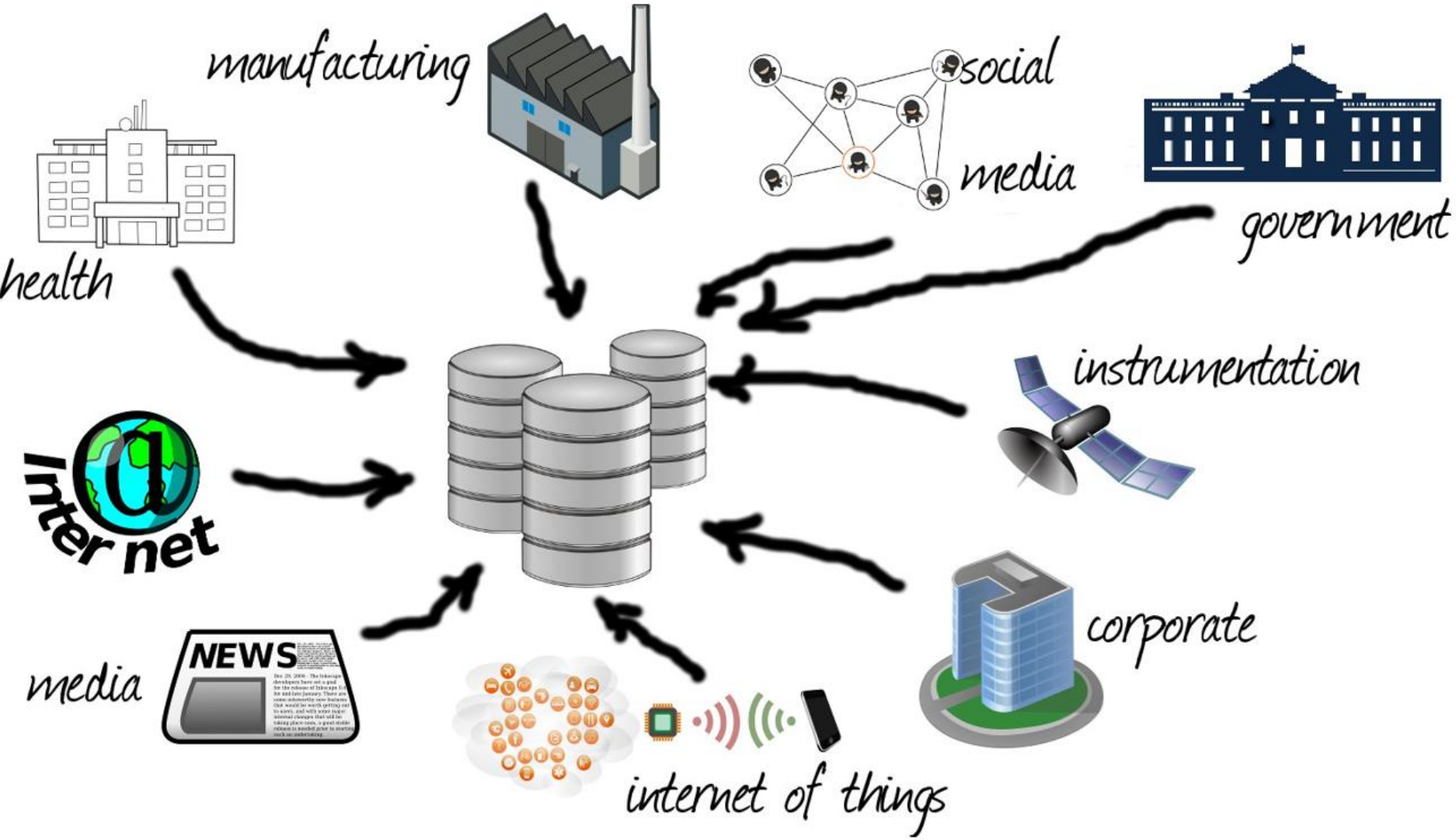
2. Collecting data:

Researchers preparing to x-ray a patient.

MONASH University

3. Integration:

Data can come from many different sources.



icons from by Openclipart.org, public domain

4. Interpretation:

Data can be described using a database schema.

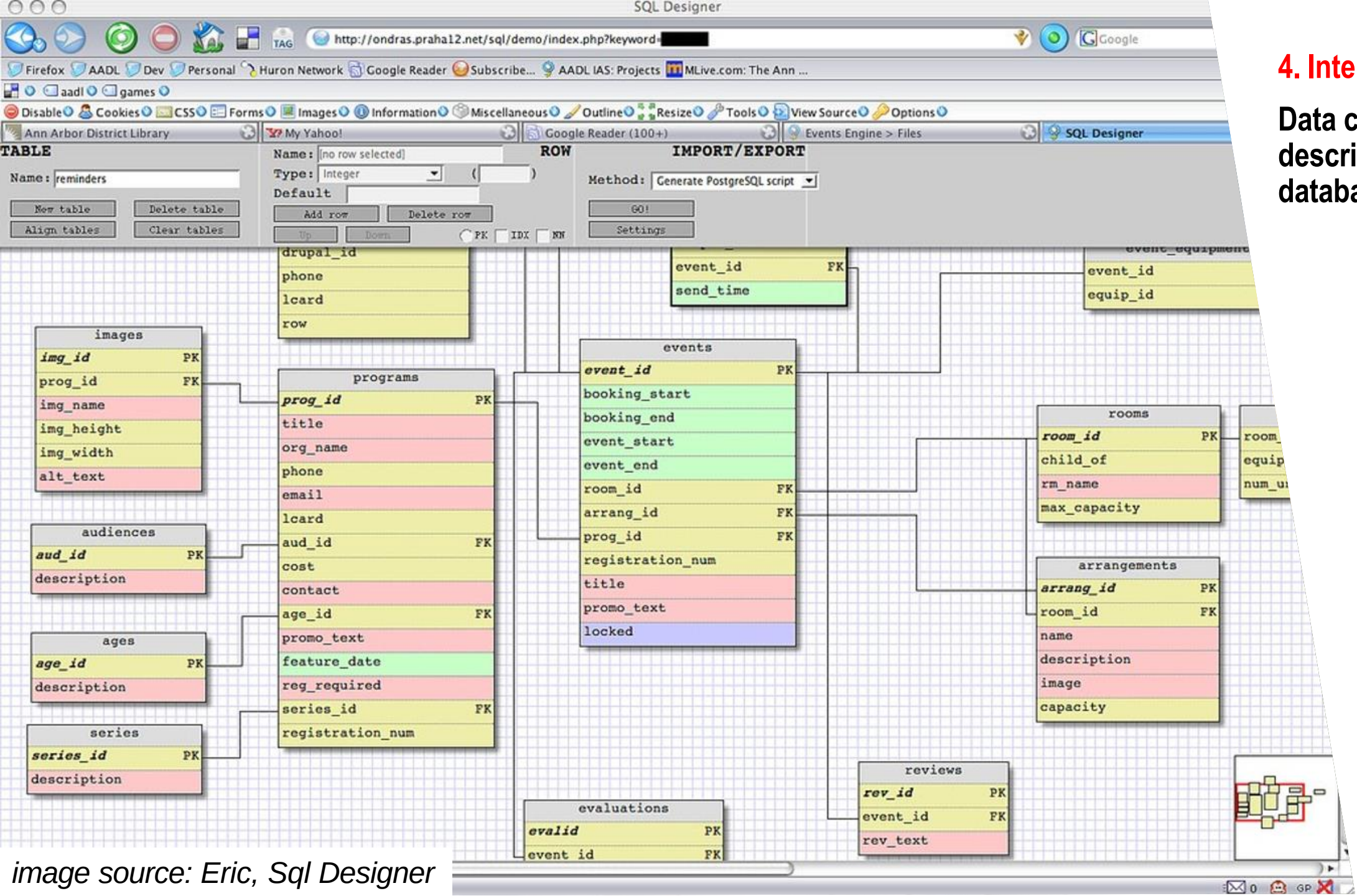


image source: Eric, Sql Designer

MONASH University

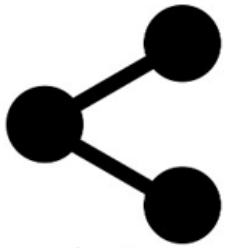
achiving

storage

privacy

legal & compliance

safety



sharing

metadata

management

ethics

5. Governance:

(i) caring for the data and its subjects.

(ii) managing data standards and formats.

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

MONASH University

6. Engineering:

Data engineers make the back-end work

image source: by Intel Free Press

MONASH University

7. Wrangling:

Inspecting and cleaning the data.

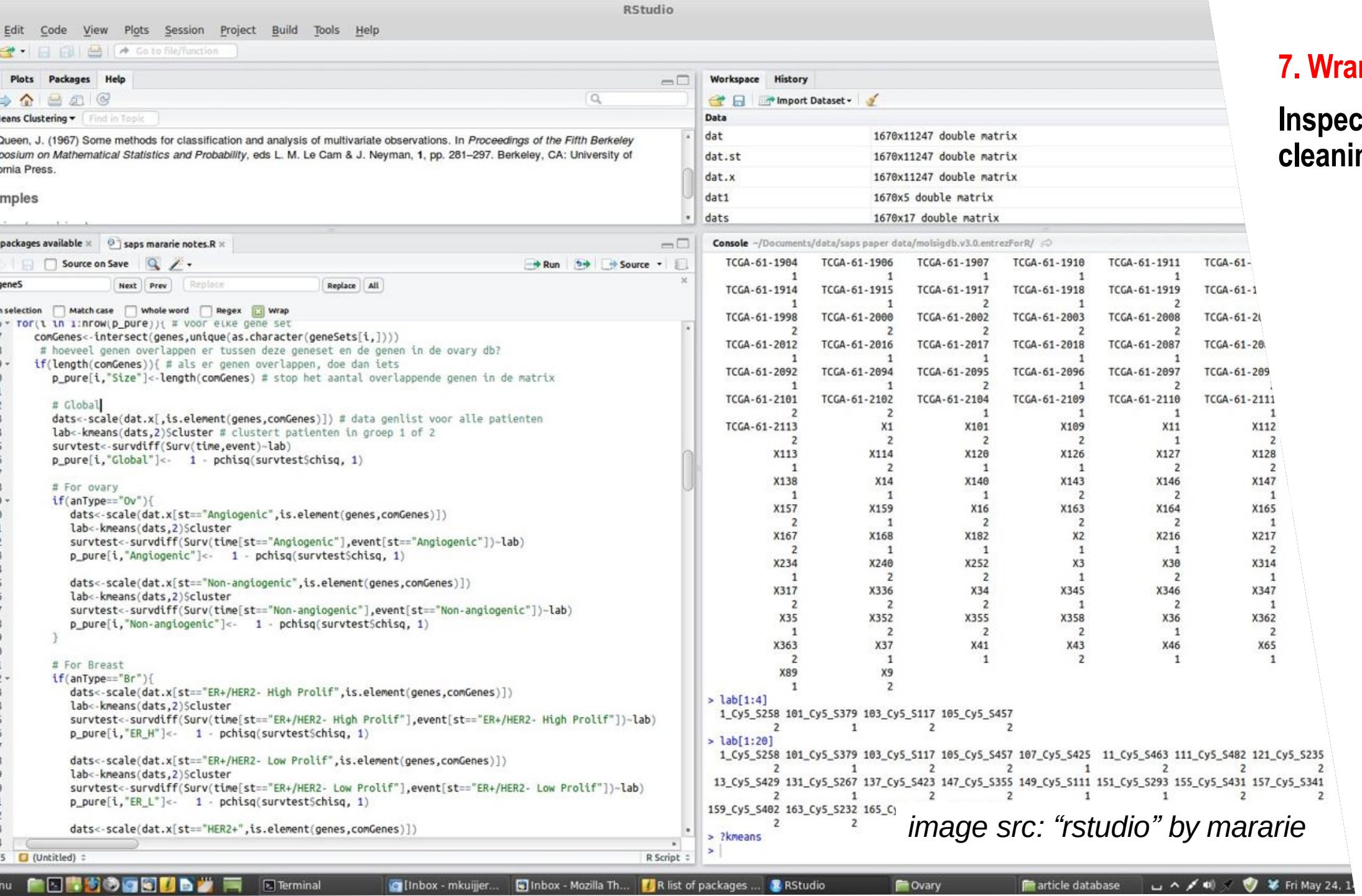


image src: "rstudio" by marariе

MONASH University

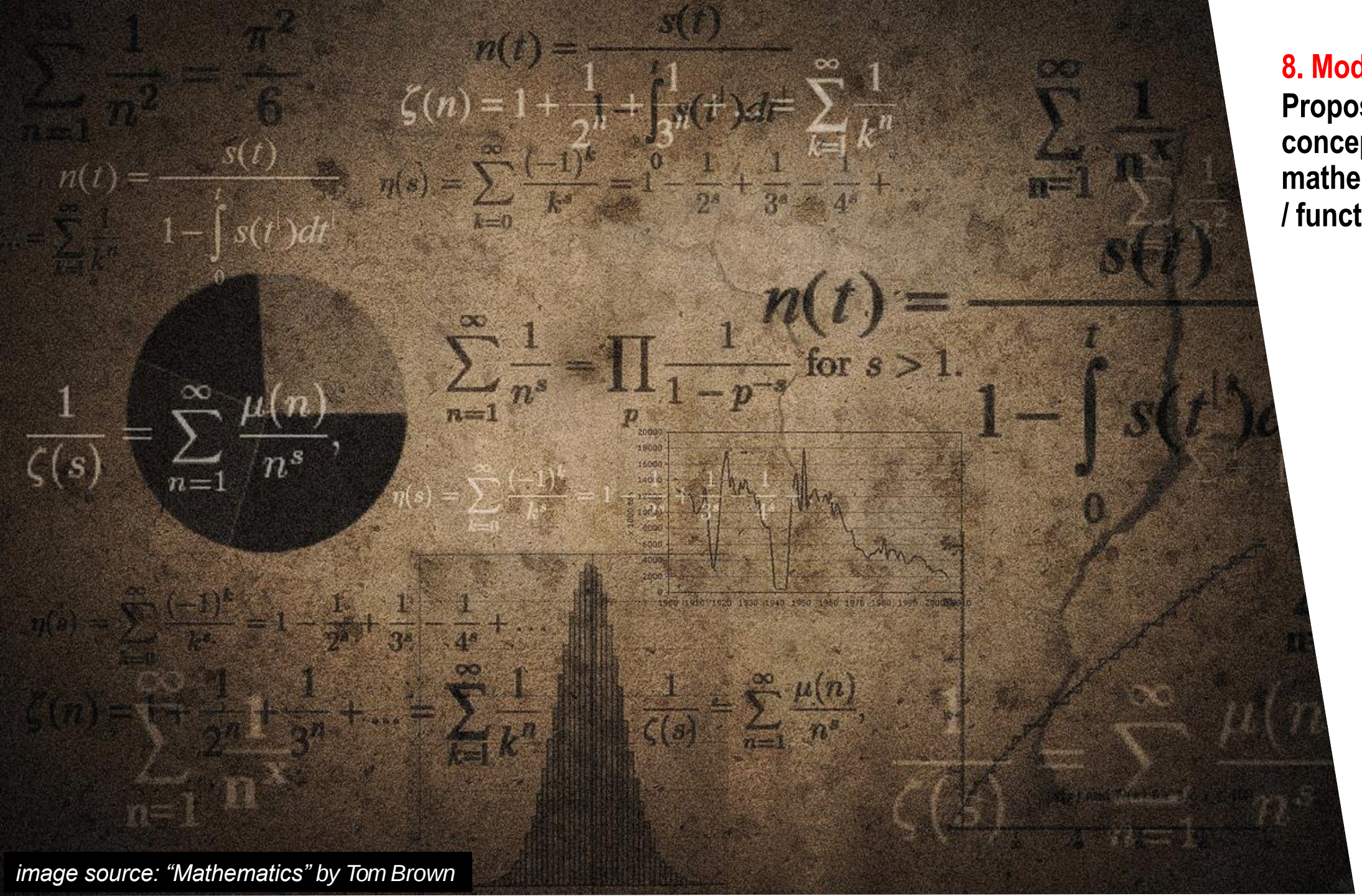


image source: "Mathematics" by Tom Brown

8. Modelling:

Proposing a conceptual / mathematical / functional model.

MONASH University

8. Modelling:

Analyst building models with his favourite tool.

MONASH University

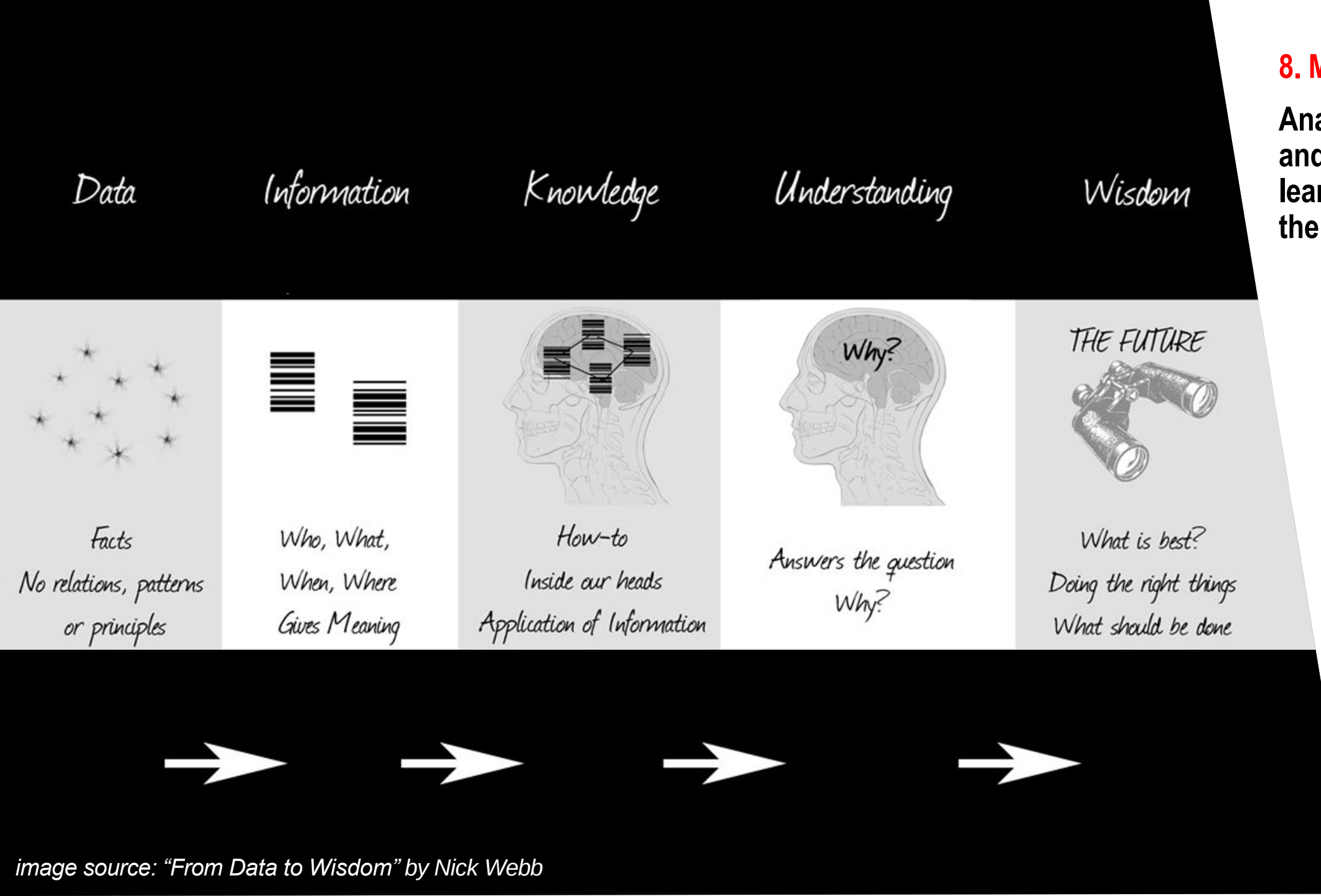


image source: "From Data to Wisdom" by Nick Webb

8. Modelling:

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

MONASH University

9. Visualisation:

Visualising data to interpret it and present results.

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

MONASH University

9. Visualisation:

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

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

MONASH University

10. Operationalize:

Putting the results to work.

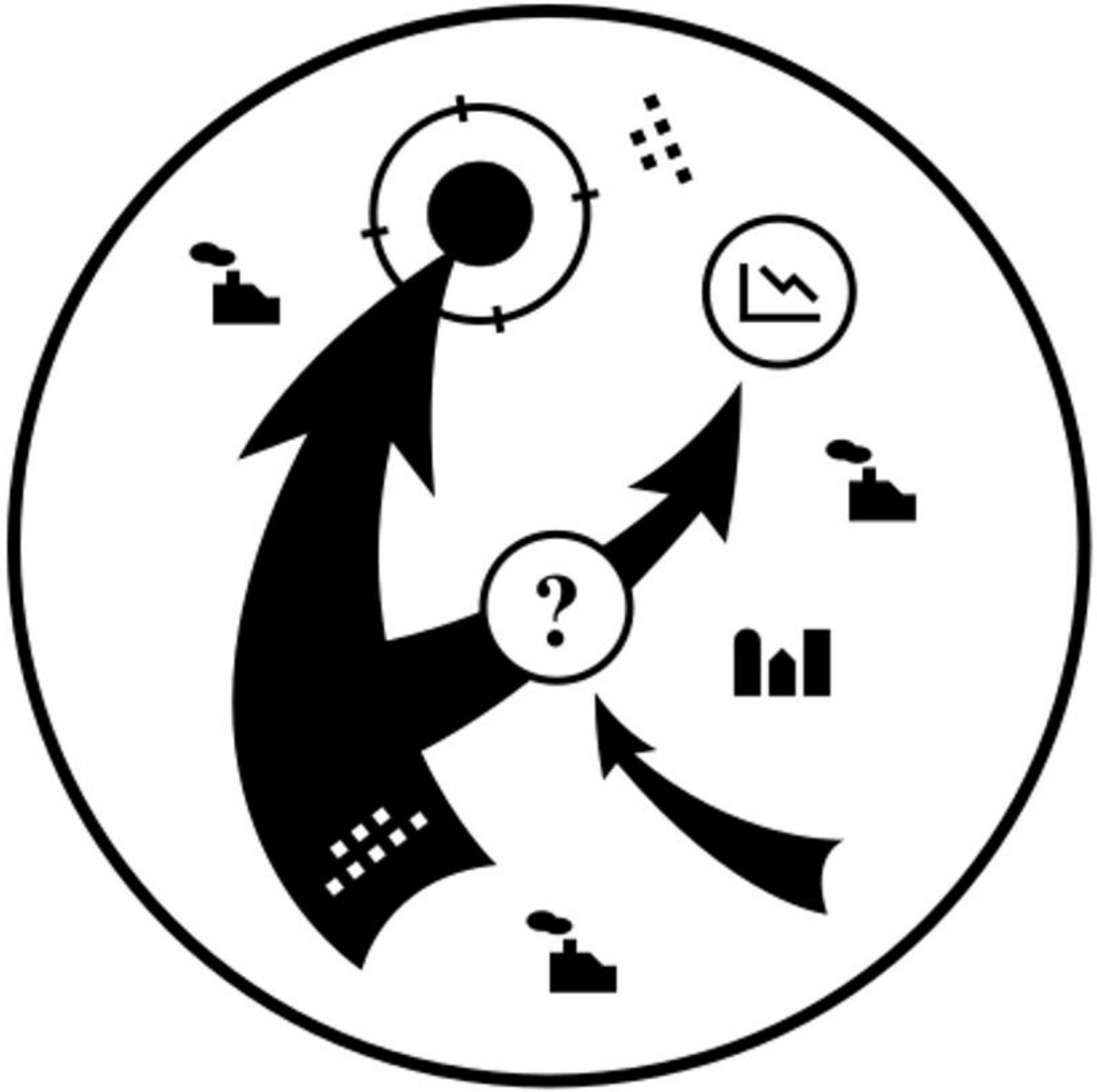


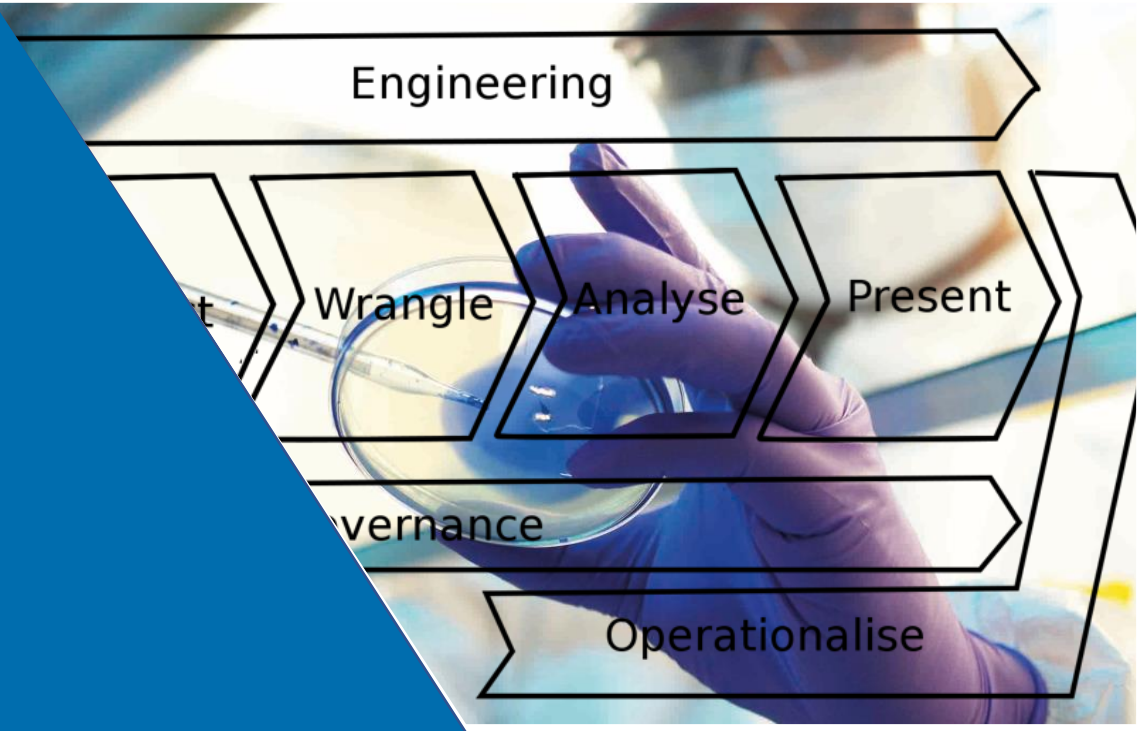
image source "Illustration of Strategy“ by Denis Fadeev

MONASH University

Data Science Process

Standard Value Chain:

Our Mode of the Process



Parts of a Data Science Project

Collection

- Getting the data

Engineering

- Storage and computational resources across full lifecycle

Governance

- Overall management of data across full lifecycle

Wrangling

- Data pre-processing, cleaning

Analysis

- Discovery (learning, visualisation, etc.)

Visualization

- Arguing the case that the results are significant and useful

Operationalize

- Putting the results to work, so as to gain benefits or value

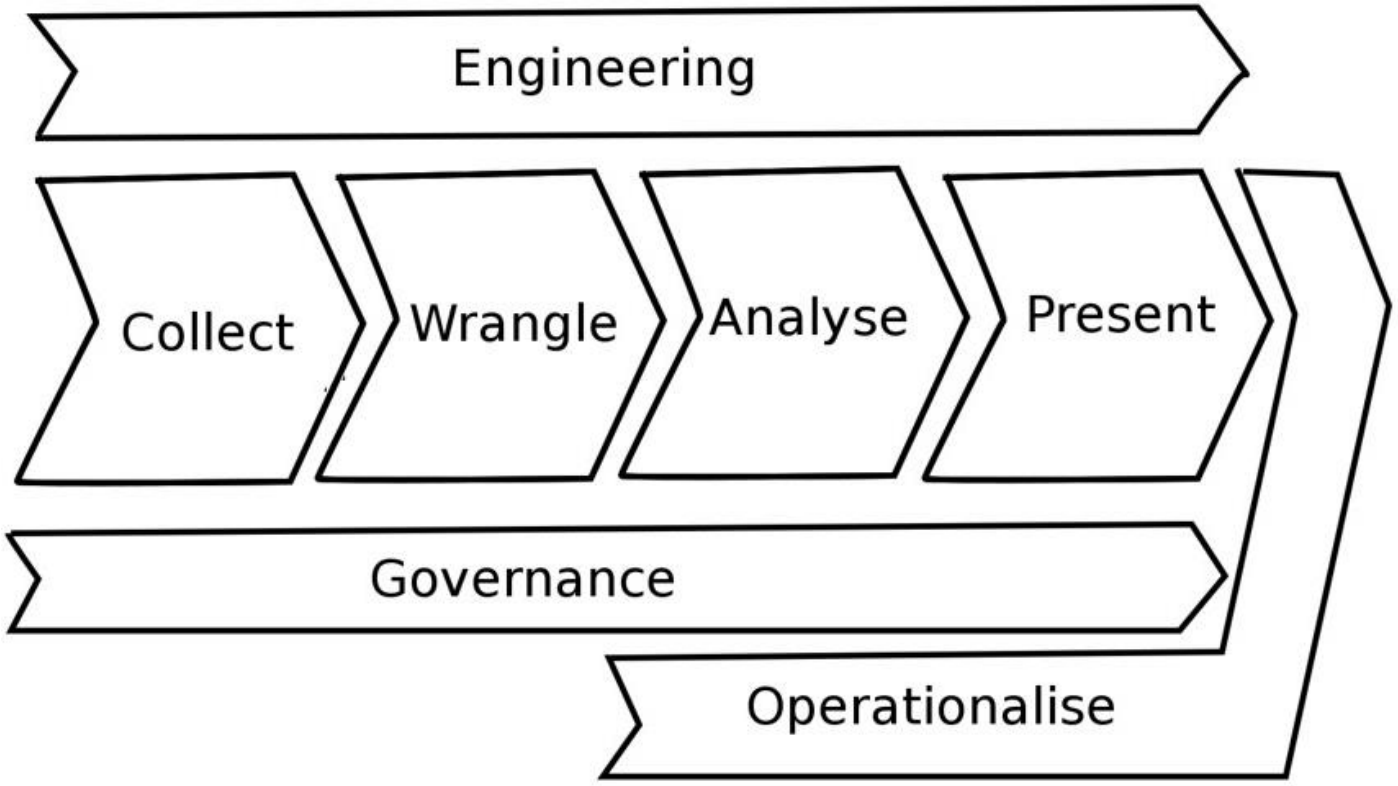
We call this the Standard Value Chain.

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.

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

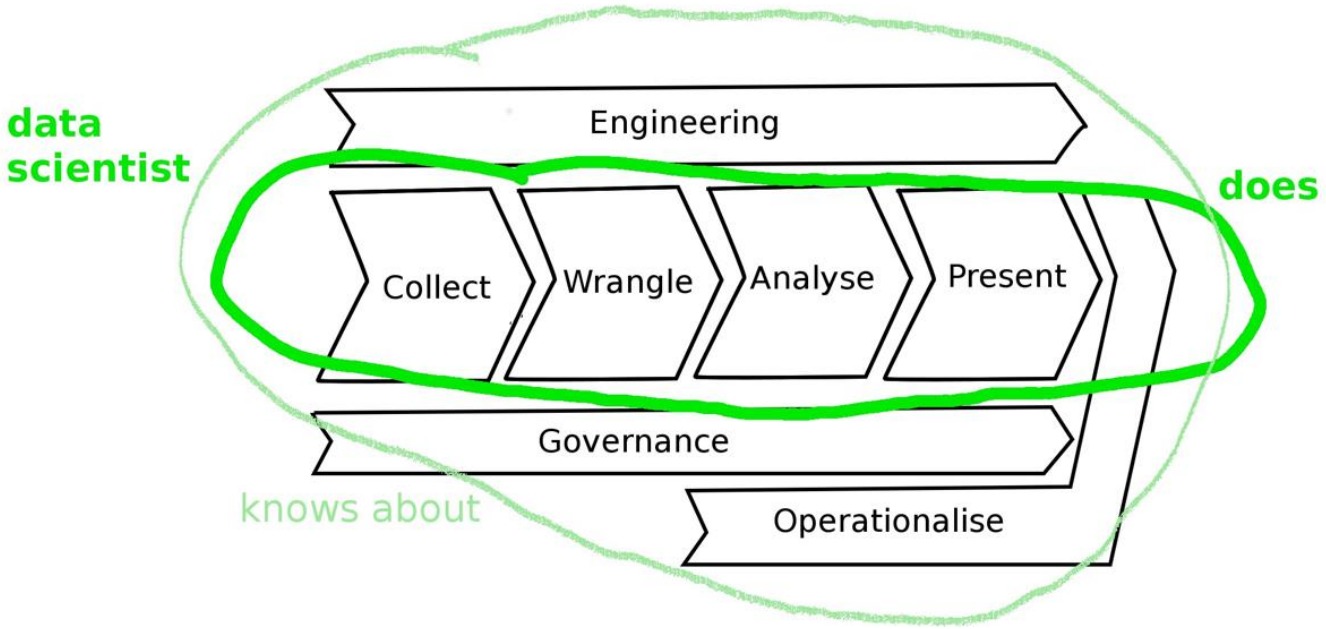


Data Science Process

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

Data Scientist

Addresses the data science process to extract meaning / value from data



MONASH University

Data Science Process

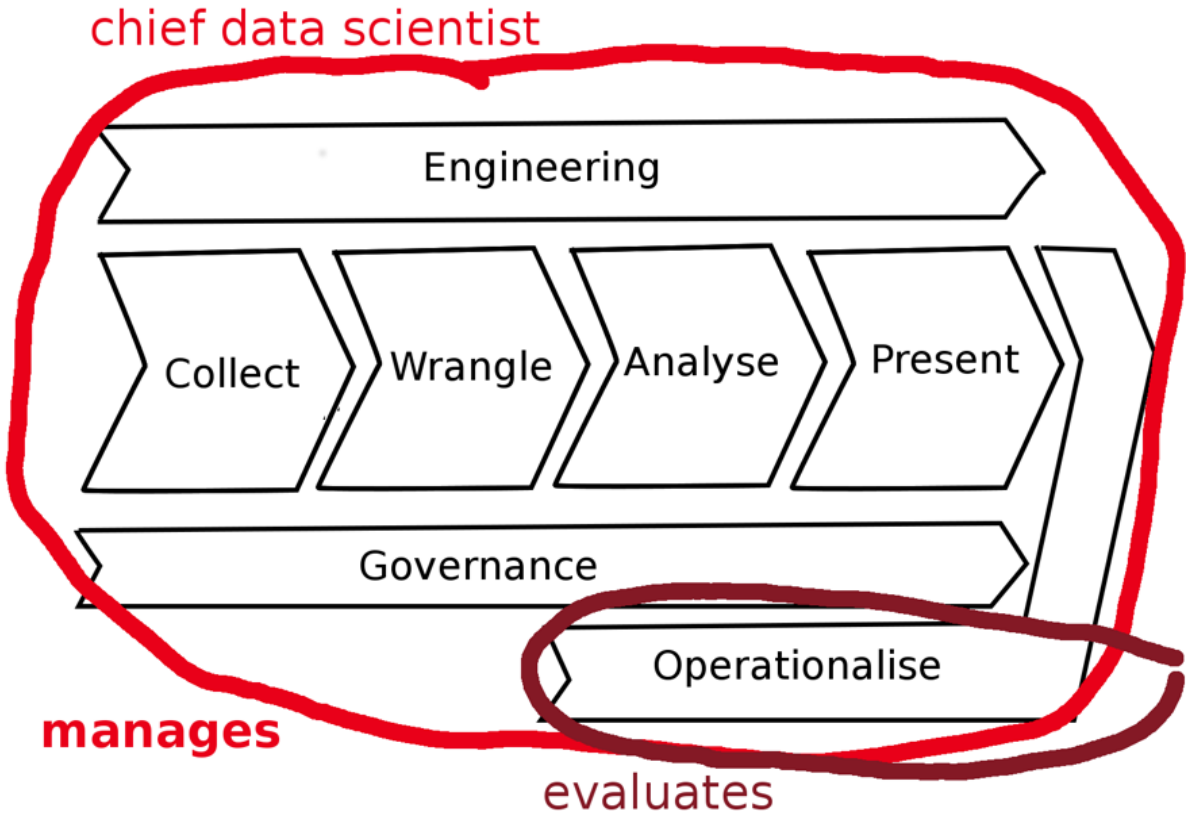
from *Doing Data Science* by Schutt and O'Neil, 2013, (available digitally through library)

Chief Data Scientist

A form of chief scientist who addresses data management, data engineering and data science goals.

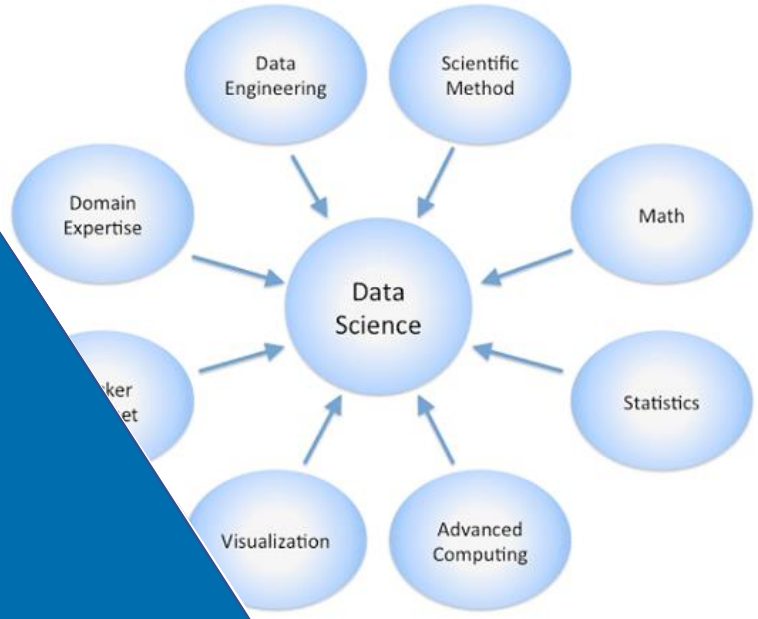
Chief Scientist

corporate position, responsible for science related aspects of a company/organisation



Relationship of Data Science to Other Disciplines

http://growtrue.net/coursera-data-science-courses/



Related: Data Engineering

Building scalable systems for storage, processing data

Hadoop,

Databases,

Distributed processing,

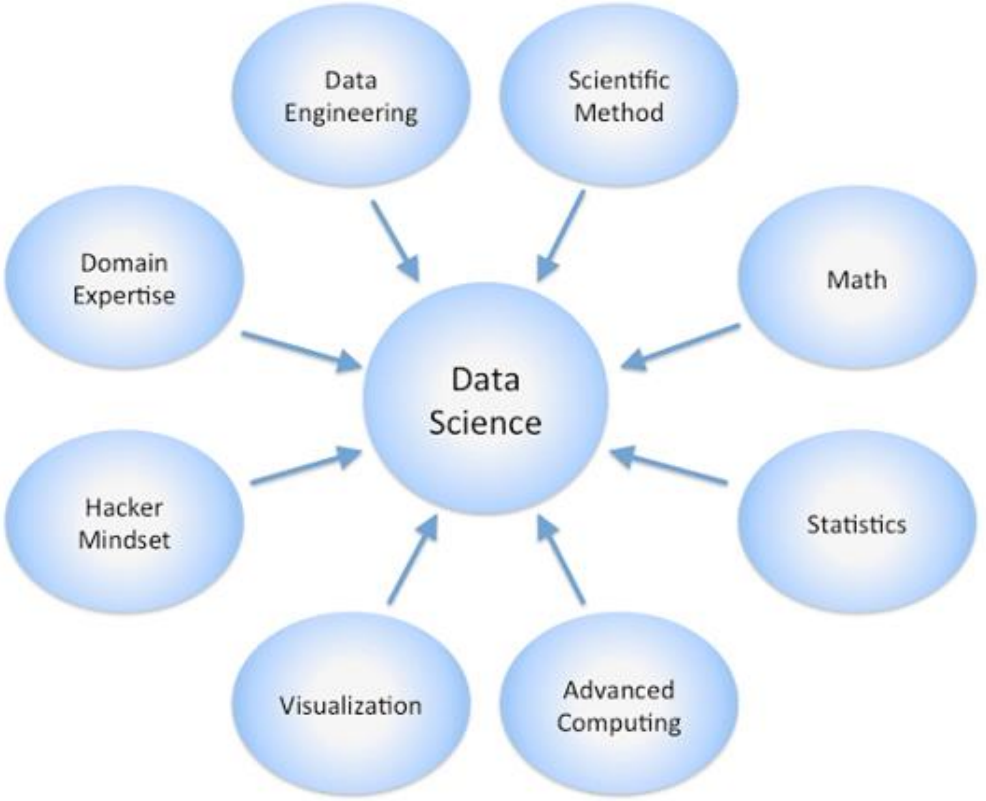
Datalakes,

Cloud computing,

GPUs,

Data wrangling, …

Huge, continuous improvement ....



MONASH University

Related: Data Analyst

Performing analysis and understanding results

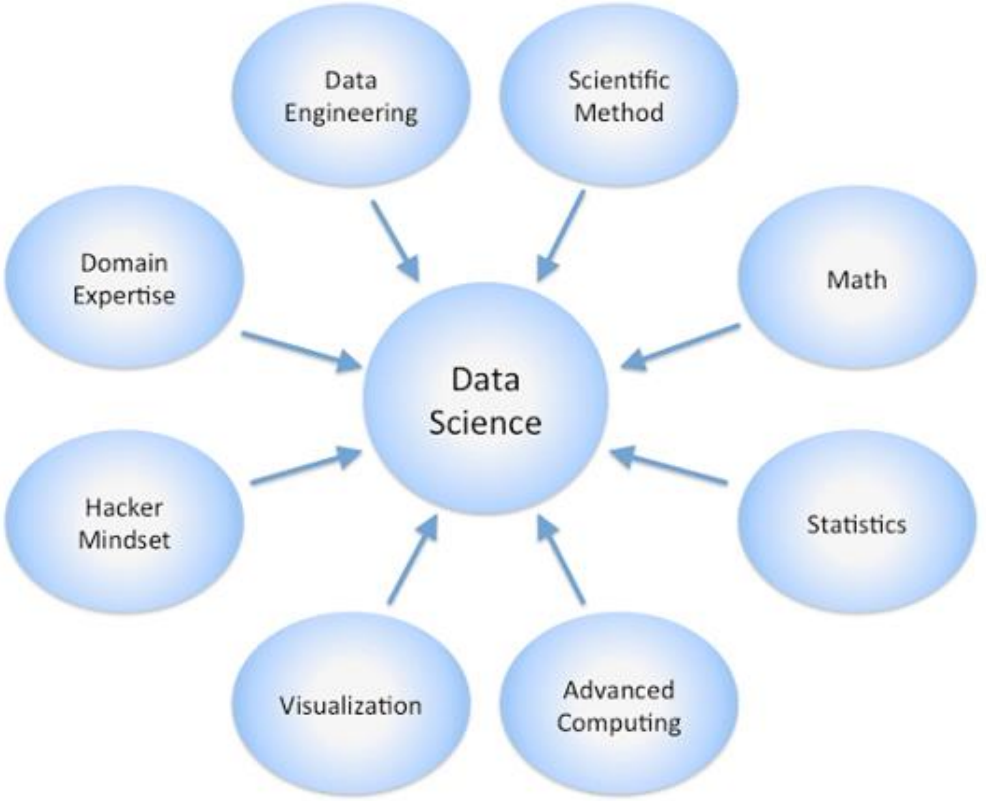
R and Microsoft Azure Machine Learning

Machine learning,

Computational statistics,

Visualisation, ...

Huge, continuous improvement ....



MONASH University

Related: Data Management

Managing data through its lifecycle

ANDS

Ethics,

Privacy,

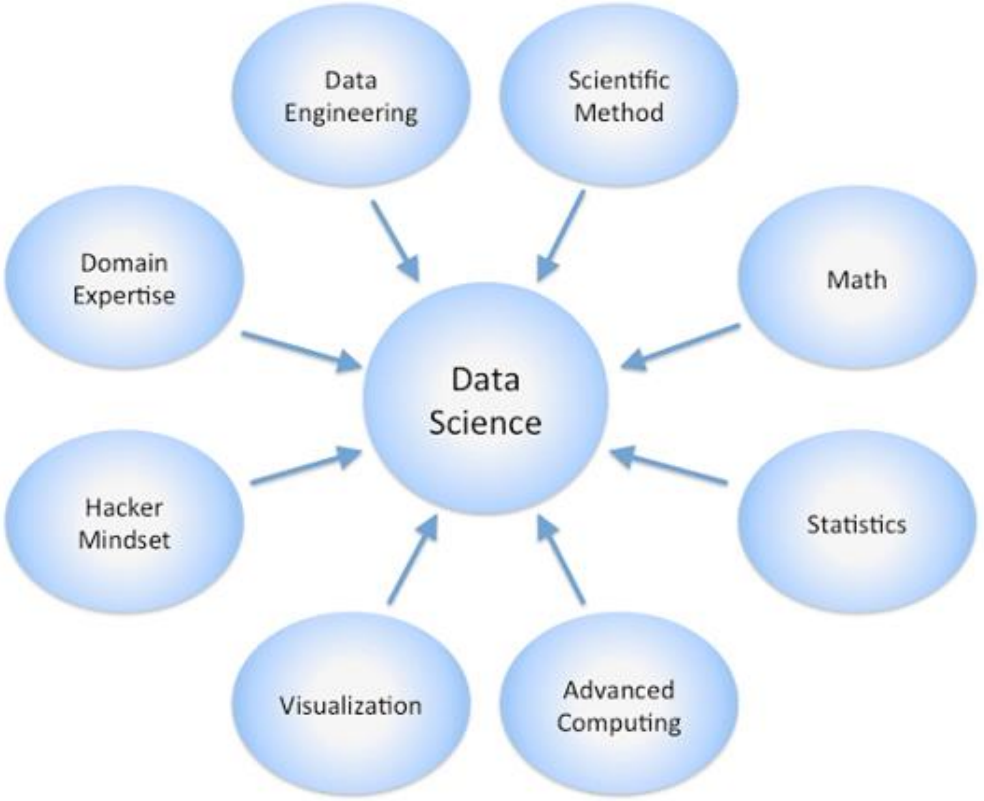
Providence,

Curation,

Backup,

Governance, ...

Huge, continuous improvement ....



MONASH University

Home Activities

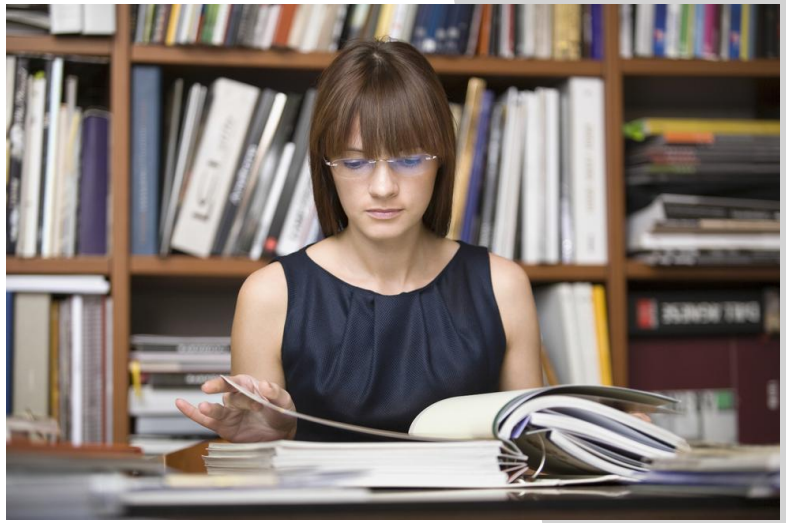
Suggested Activities for the week end

Videos

Watch *Cukier's TED talk on "Big Data"*

Watch the CERN video, *"Big Data" from Tim Smith*

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

What is Data Science?

MONASH University

Recap: 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
    - To be achieved in your tutorial / laboratory session