

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



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

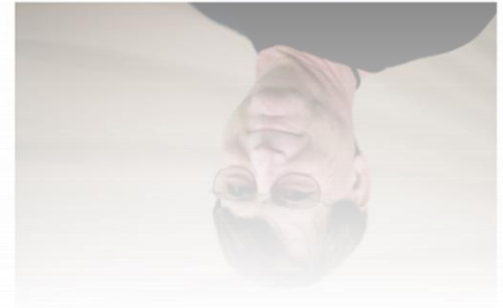
That's going to be a hugely important skill in the next decades.



important skill in the next decades.

That's going to be a hugely

- to **communicate** it
- to **visualize** it
- to **extract value** from it
- to **process** it
- to be able to **understand** it



The Data Science Process

Illustrating the Process

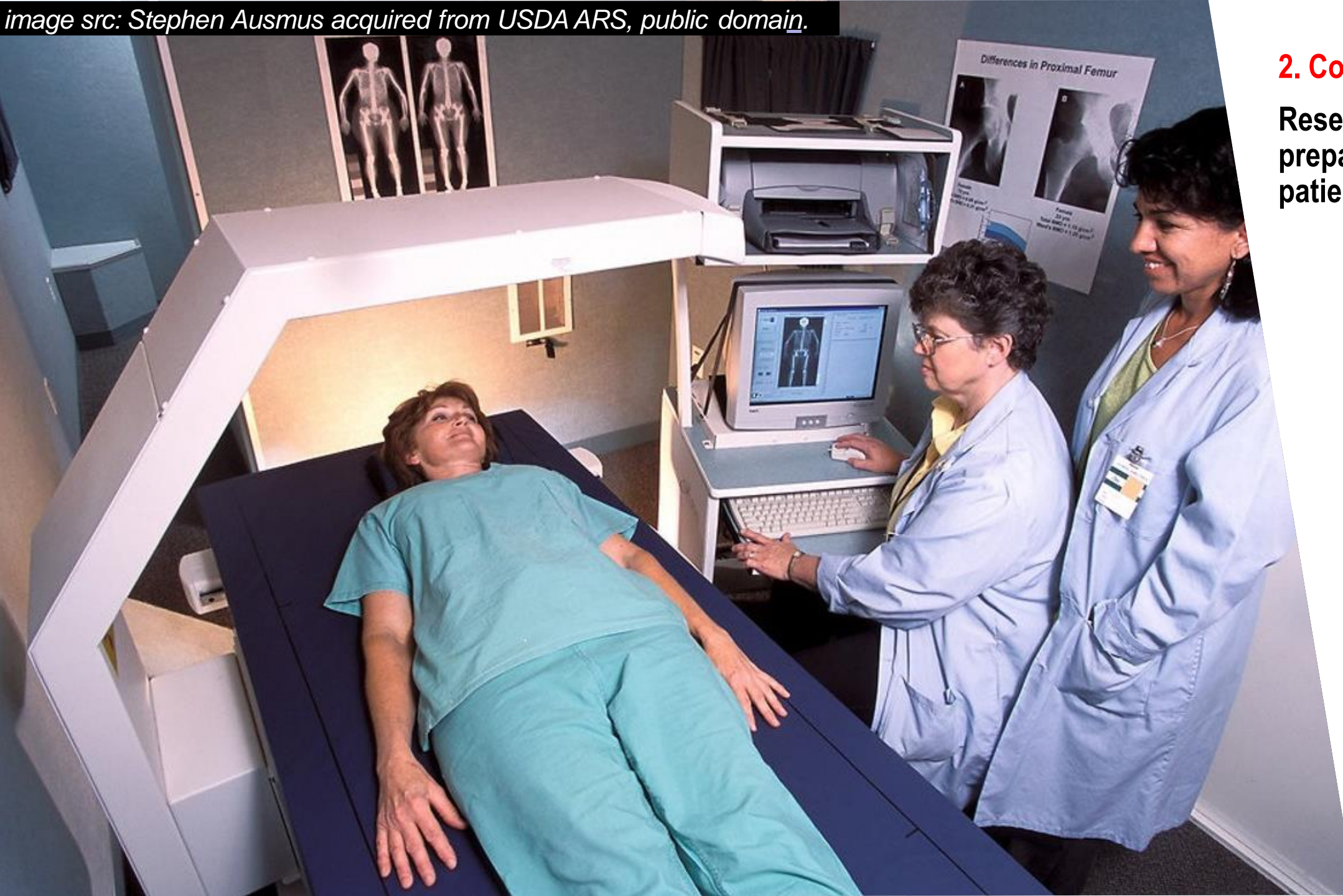
- Many different tasks come together to complete a Data Science project
 - A data scientist should be familiar with most, but doesn't need to be an expert in all
- Not all are labelled as Data Science
 - Some from other field such as computer engineering, business, ...

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



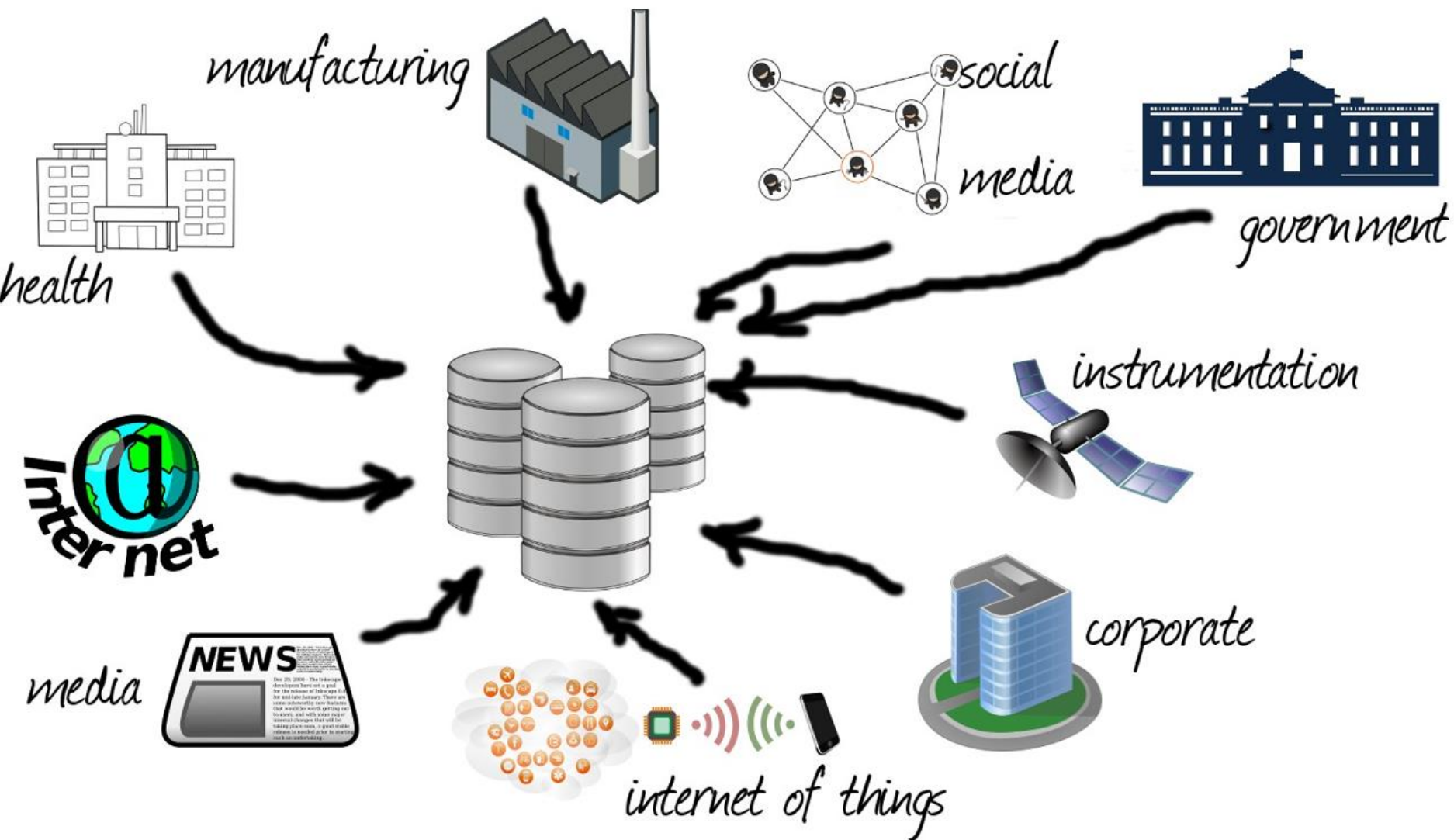
1. Pitching ideas:

For data science
projects to investors
/ managers



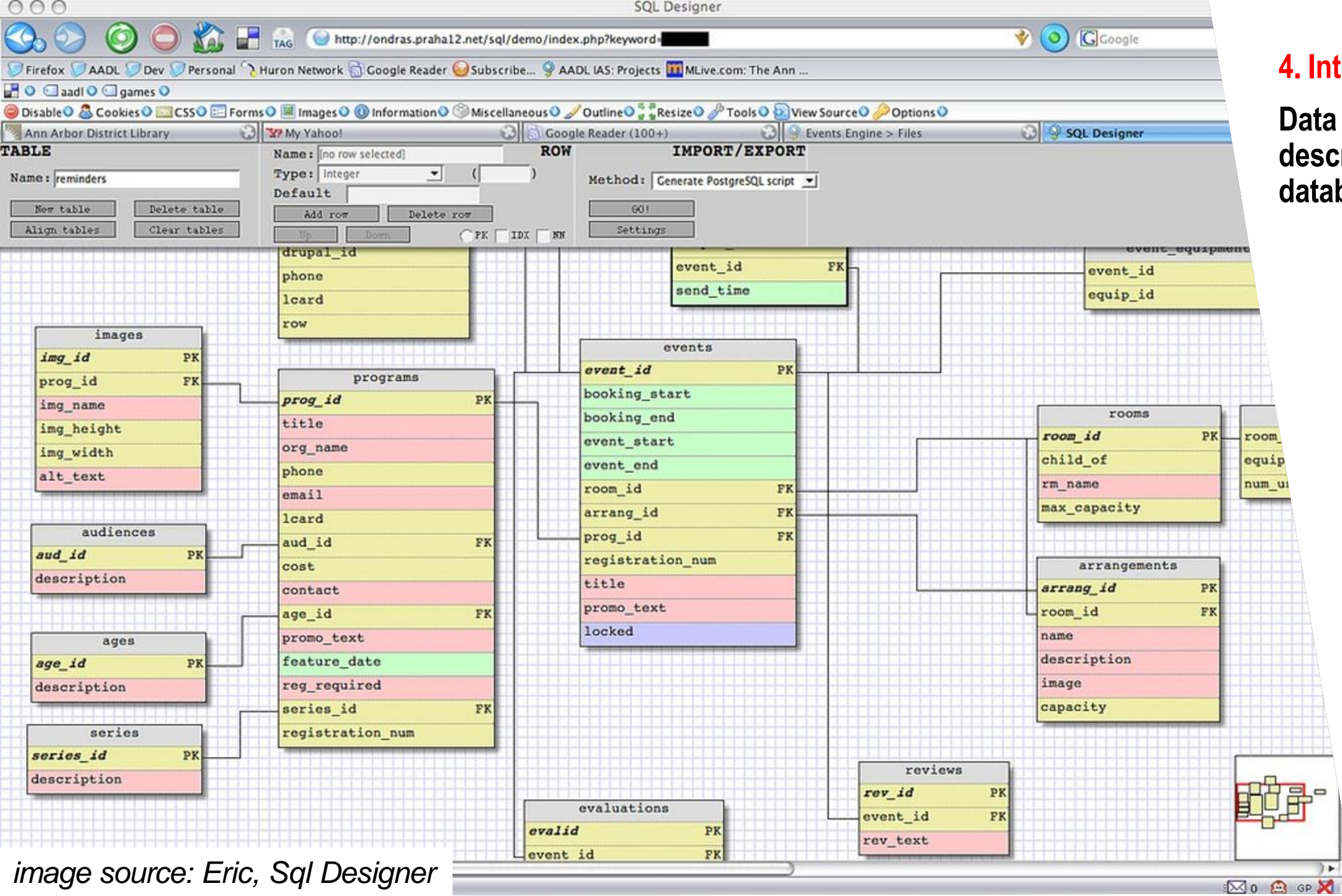
2. Collecting data:

Researchers preparing to x-ray a patient.



3. Integration:

Data can come from many different sources.



4. Interpretation:

Data can be described using a database schema.



archiving



storage



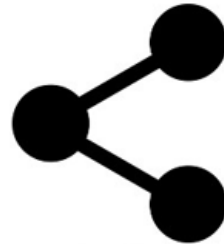
privacy



legal & compliance



safety



sharing



metadata



management

ethics



5. Governance:

(i) caring for the data and its subjects.

(ii) managing data standards and formats.

6. Engineering:

Data engineers make the back-end work



image source: by Intel Free Press

RStudio

Edit Code View Plots Session Project Build Tools Help

Go to file/function

Plots Packages Help

Means Clustering Find in Topic

Queen, J. (1967) Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, eds L. M. Le Cam & J. Neyman, 1, pp. 281–297. Berkeley, CA: University of California Press.

Examples

packages available saps mararie notes.R

Source on Save

genes

Match case Whole word Regex Wrap

```
for(i in 1:nrow(p_pure)){ # voor elke gene set
  conGenes<-intersect(genes,unique(as.character(geneSets[i])))
  # hoeveel genen overlappen er tussen deze geneset en de genen in de ovary db?
  if(length(conGenes)){ # als er genen overlappen, doe dan iets
    p_pure[i,"Size"]<-length(conGenes) # stop het aantal overlappende genen in de matrix

    # Global
    dats<-scale(dat.x[,is.element(genes,conGenes)]) # data genlist voor alle patienten
    lab<-kmeans(dats,2)$cluster # cluster patienten in groep 1 of 2
    survtest<-survdiff(Surv(time,event)~lab)
    p_pure[i,"Global"]<- 1 - pchisq(survtest$chisq, 1)

    # For ovary
    if(anType=="Ov"){
      dats<-scale(dat.x[st=="Angiogenic",is.element(genes,conGenes)])
      lab<-kmeans(dats,2)$cluster
      survtest<-survdiff(Surv(time[st=="Angiogenic"],event[st=="Angiogenic"])-lab)
      p_pure[i,"Angiogenic"]<- 1 - pchisq(survtest$chisq, 1)

      dats<-scale(dat.x[st=="Non-angiogenic",is.element(genes,conGenes)])
      lab<-kmeans(dats,2)$cluster
      survtest<-survdiff(Surv(time[st=="Non-angiogenic"],event[st=="Non-angiogenic"])-lab)
      p_pure[i,"Non-angiogenic"]<- 1 - pchisq(survtest$chisq, 1)
    }

    # For Breast
    if(anType=="Br"){
      dats<-scale(dat.x[st=="ER+/HER2- High Prolif",is.element(genes,conGenes)])
      lab<-kmeans(dats,2)$cluster
      survtest<-survdiff(Surv(time[st=="ER+/HER2- High Prolif"],event[st=="ER+/HER2- High Prolif"])-lab)
      p_pure[i,"ER_H"]<- 1 - pchisq(survtest$chisq, 1)

      dats<-scale(dat.x[st=="ER+/HER2- Low Prolif",is.element(genes,conGenes)])
      lab<-kmeans(dats,2)$cluster
      survtest<-survdiff(Surv(time[st=="ER+/HER2- Low Prolif"],event[st=="ER+/HER2- Low Prolif"])-lab)
      p_pure[i,"ER_L"]<- 1 - pchisq(survtest$chisq, 1)

      dats<-scale(dat.x[st=="HER2+",is.element(genes,conGenes)])
```

Workspace History

Import Dataset

Data

dat	1670x11247 double matrix
dat.st	1670x11247 double matrix
dat.x	1670x11247 double matrix
dat1	1670x5 double matrix
dats	1670x17 double matrix

Console ~/Documents/data/saps paper data/moisigdb.v3.0.entrezForR/

```
TCGA-61-1904 TCGA-61-1906 TCGA-61-1907 TCGA-61-1910 TCGA-61-1911 TCGA-61-1912
1 1 1 1 1 1
TCGA-61-1914 TCGA-61-1915 TCGA-61-1917 TCGA-61-1918 TCGA-61-1919 TCGA-61-1920
1 1 2 1 2 2
TCGA-61-1998 TCGA-61-2000 TCGA-61-2002 TCGA-61-2003 TCGA-61-2008 TCGA-61-2009
2 2 2 2 2 2
TCGA-61-2012 TCGA-61-2016 TCGA-61-2017 TCGA-61-2018 TCGA-61-2087 TCGA-61-2090
1 1 1 1 1 1
TCGA-61-2092 TCGA-61-2094 TCGA-61-2095 TCGA-61-2096 TCGA-61-2097 TCGA-61-2099
1 1 2 1 2 2
TCGA-61-2101 TCGA-61-2102 TCGA-61-2104 TCGA-61-2109 TCGA-61-2110 TCGA-61-2111
2 2 1 1 1 1
TCGA-61-2113 X1 X101 X109 X11 X112
2 2 2 1 2 2
X113 X114 X120 X126 X127 X128
1 2 1 1 2 2
X138 X14 X140 X143 X146 X147
1 1 1 2 2 1
X157 X159 X16 X163 X164 X165
2 1 2 2 2 1
X167 X168 X182 X2 X216 X217
2 1 1 1 1 2
X234 X240 X252 X3 X30 X314
1 2 2 1 2 1
X317 X336 X34 X345 X346 X347
2 2 2 1 2 1
X35 X352 X355 X358 X36 X362
1 2 2 2 1 2
X363 X37 X41 X43 X46 X65
2 1 1 2 1 1
X89 X9
1 2
```

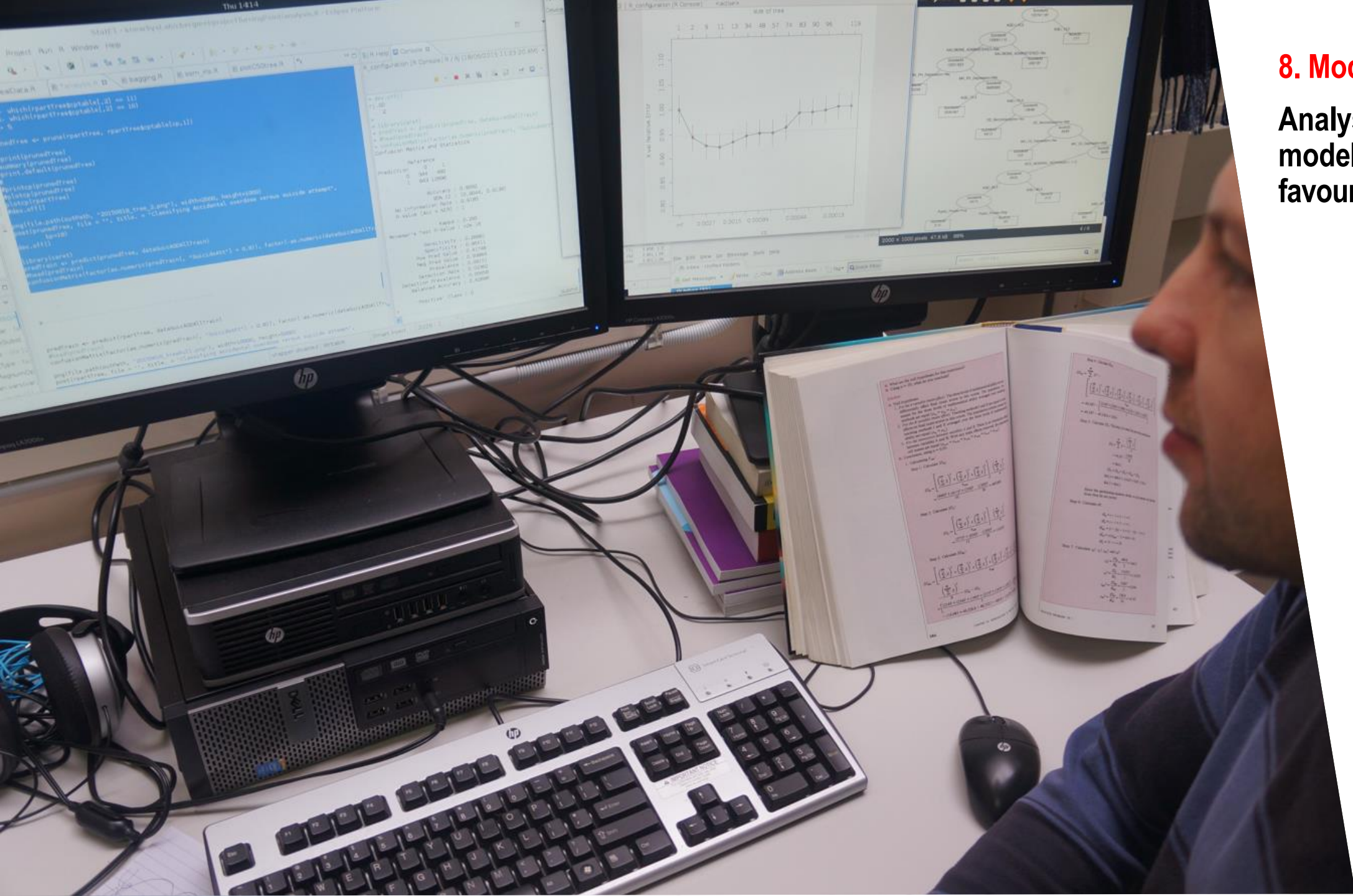
```
> lab[1:4]
1_Cy5_S258 101_Cy5_S379 103_Cy5_S117 105_Cy5_S457
2 1 2 2
> lab[1:20]
1_Cy5_S258 101_Cy5_S379 103_Cy5_S117 105_Cy5_S457 107_Cy5_S425 11_Cy5_S463 111_Cy5_S482 121_Cy5_S235
2 1 2 2 1 2 2 2
13_Cy5_S429 131_Cy5_S267 137_Cy5_S423 147_Cy5_S355 149_Cy5_S111 151_Cy5_S293 155_Cy5_S431 157_Cy5_S341
2 1 2 2 1 1 2 2
159_Cy5_S402 163_Cy5_S232 165_Cy5_S232
2 2
```

```
> ?kmeans
> |
```

image src: "rstudio" by mararie

7. Wrangling: Inspecting and cleaning the data.

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



8. Modelling:

Analyst building models with his favourite tool.

Data

Information

Knowledge

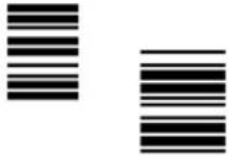
Understanding

Wisdom



Facts

No relations, patterns
or principles



Who, What,
When, Where
Gives Meaning



How-to
Inside our heads
Application of Information



Answers the question
Why?

THE FUTURE



What is best?
Doing the right things
What should be done



8. Modelling:

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

9. Visualisation:

Visualising data to interpret it and present results.



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

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

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

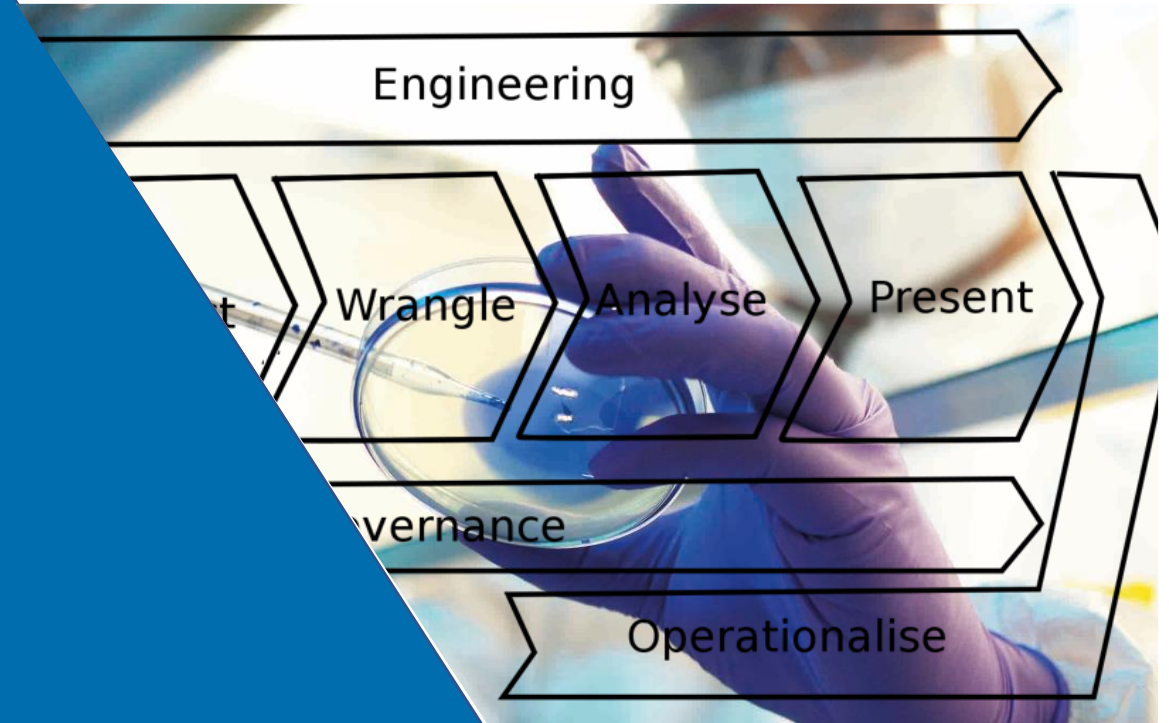


10. Operationalize:

Putting the results to work.

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

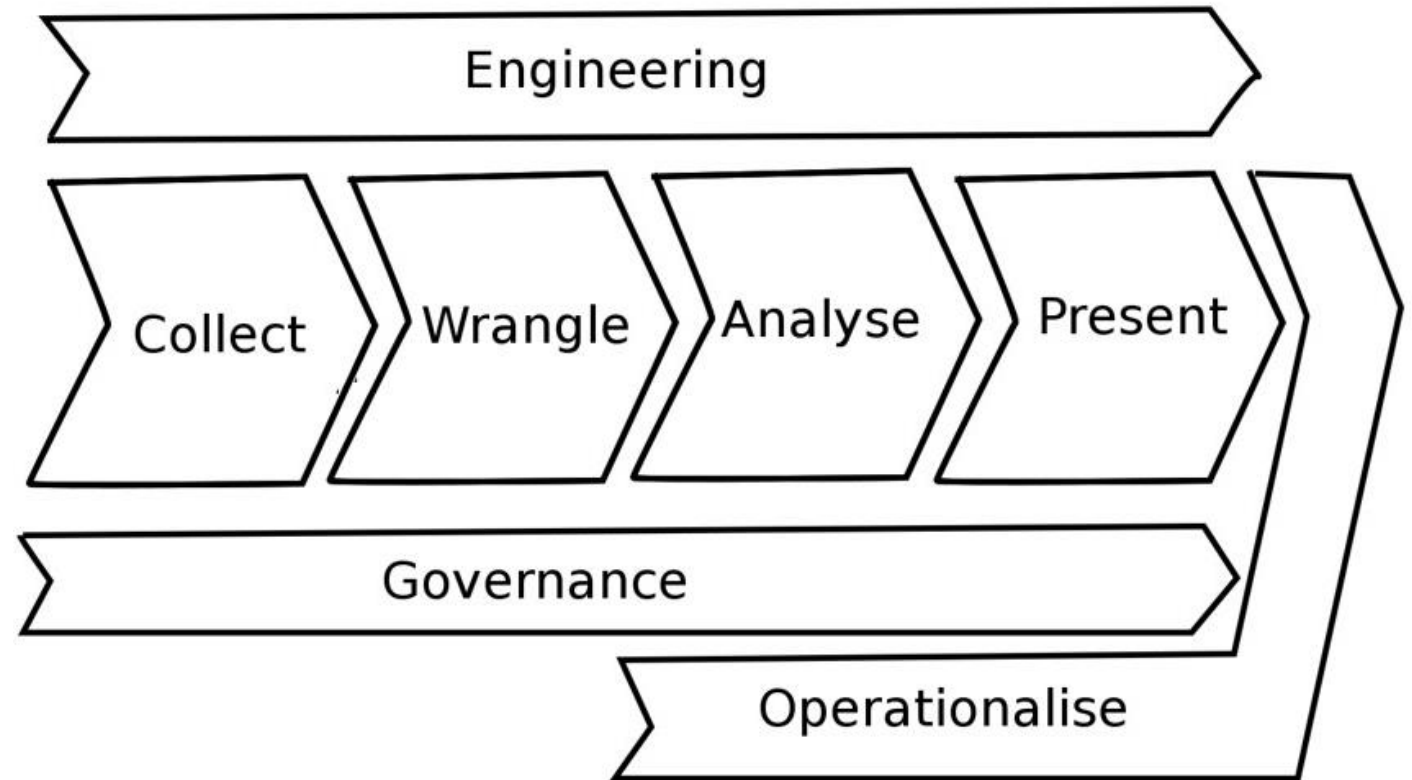
We call this the
**Standard Value
Chain.**

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

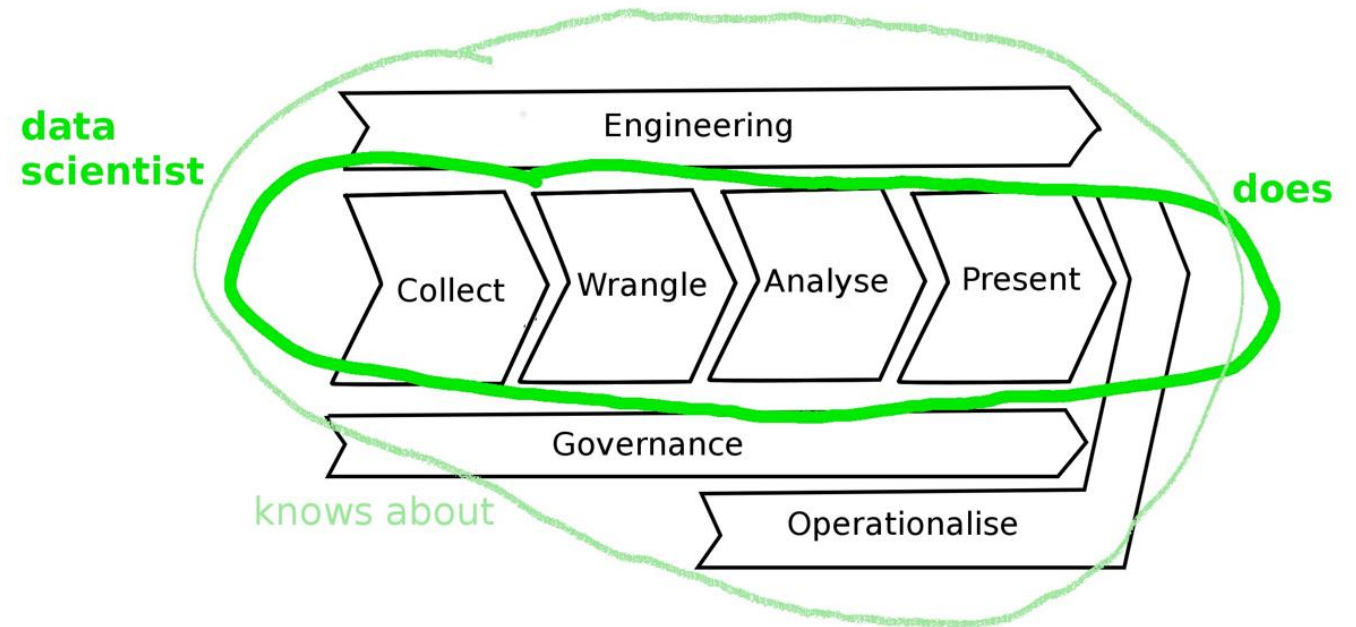


Data Science Process

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



Data Science Process

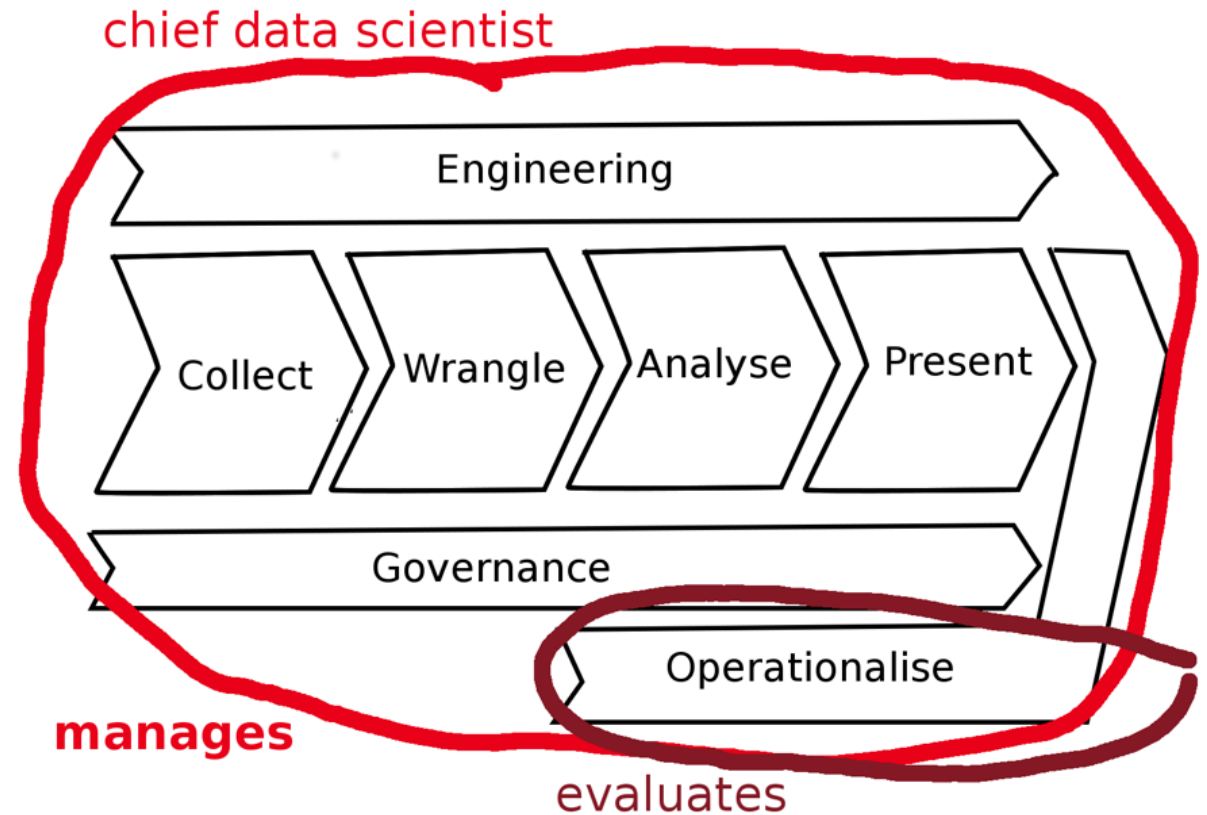
from [Doing Data Science](#) by Schutt and O'Neil, 2013, (available digitally through library)

Chief Data Scientist

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

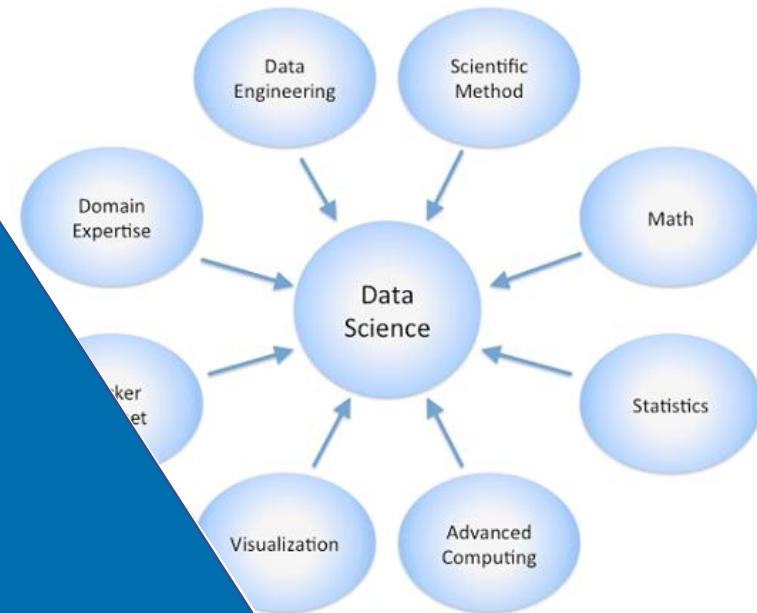
Chief Scientist

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,

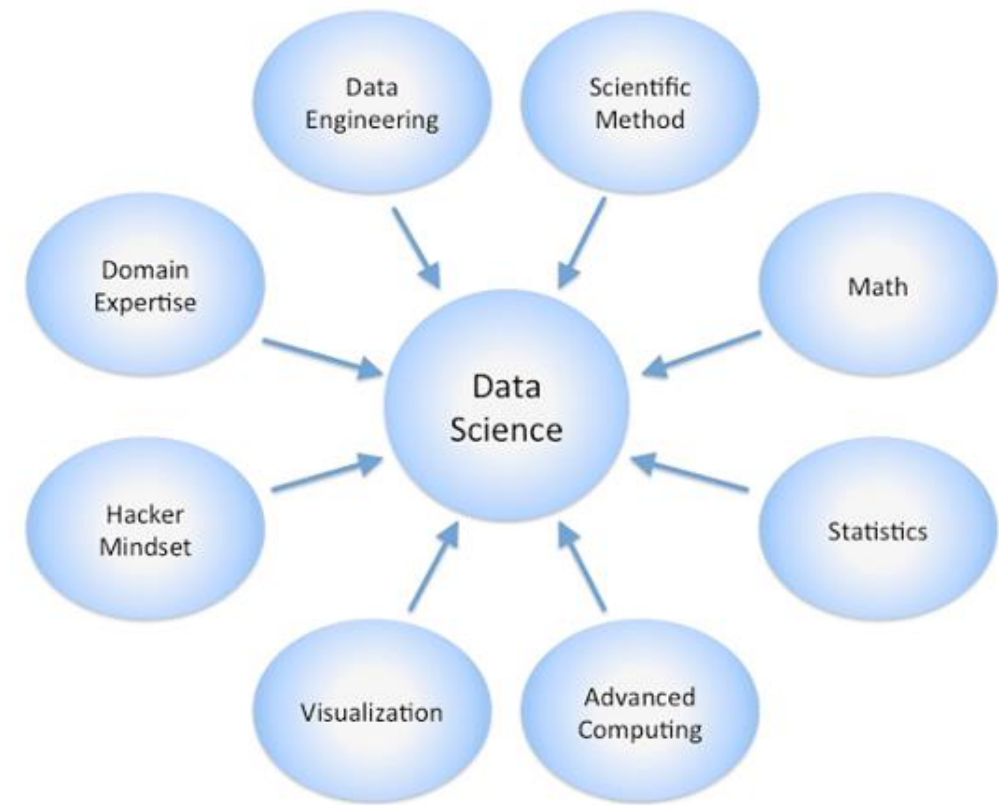
Datalakes,

Cloud computing,

GPUs,

Data wrangling, ...

Huge, continuous improvement



Related: Data Analyst

Performing analysis and understanding results

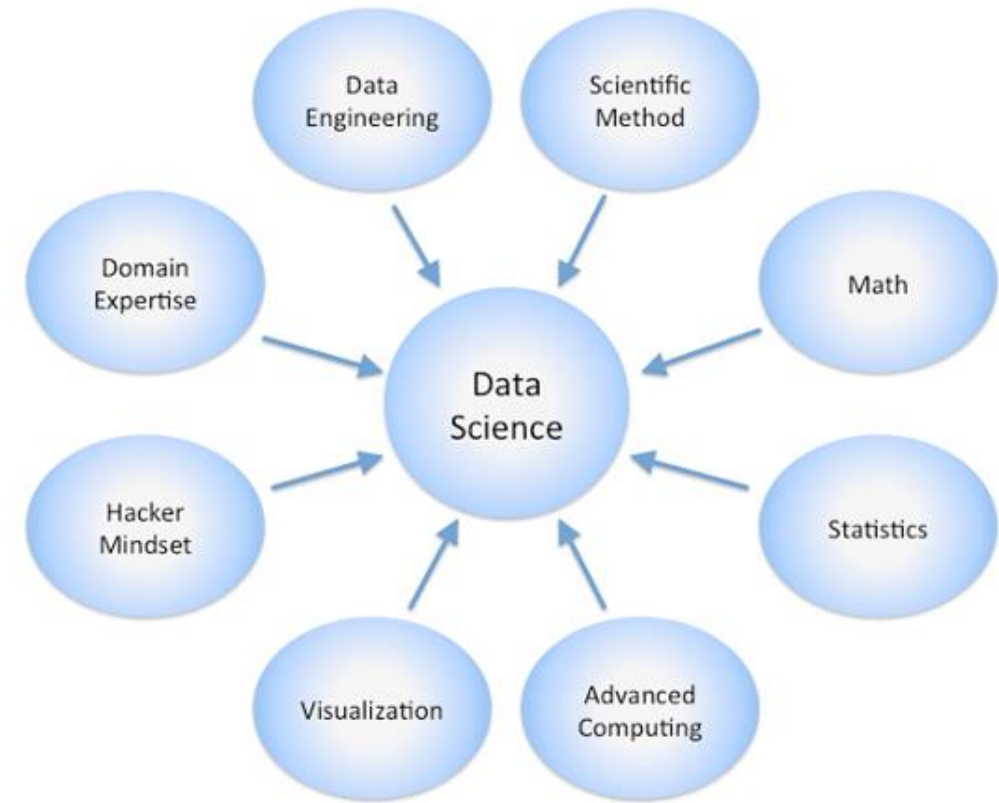
[R](#) and [Microsoft Azure Machine Learning](#)

Machine learning,

Computational statistics,

Visualisation, ...

Huge, continuous improvement



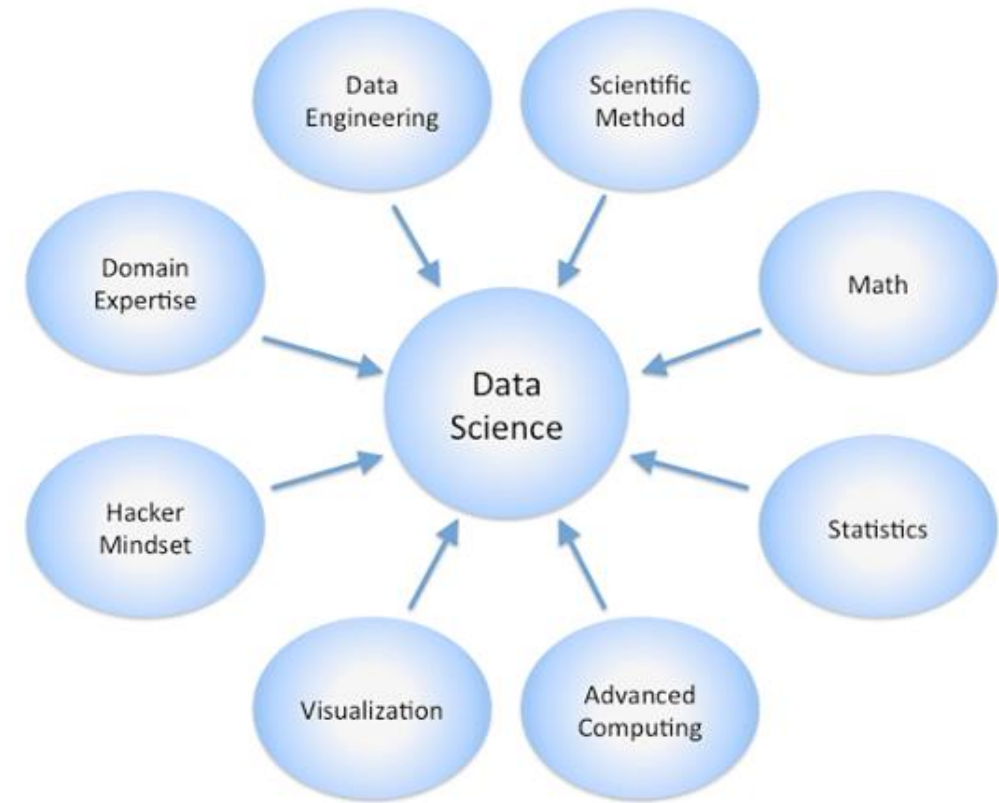
Related: Data Management

Managing data through its lifecycle

ANDS

Ethics,
Privacy,
Providence,
Curation,
Backup,
Governance, ...

Huge, continuous improvement



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?](#)



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