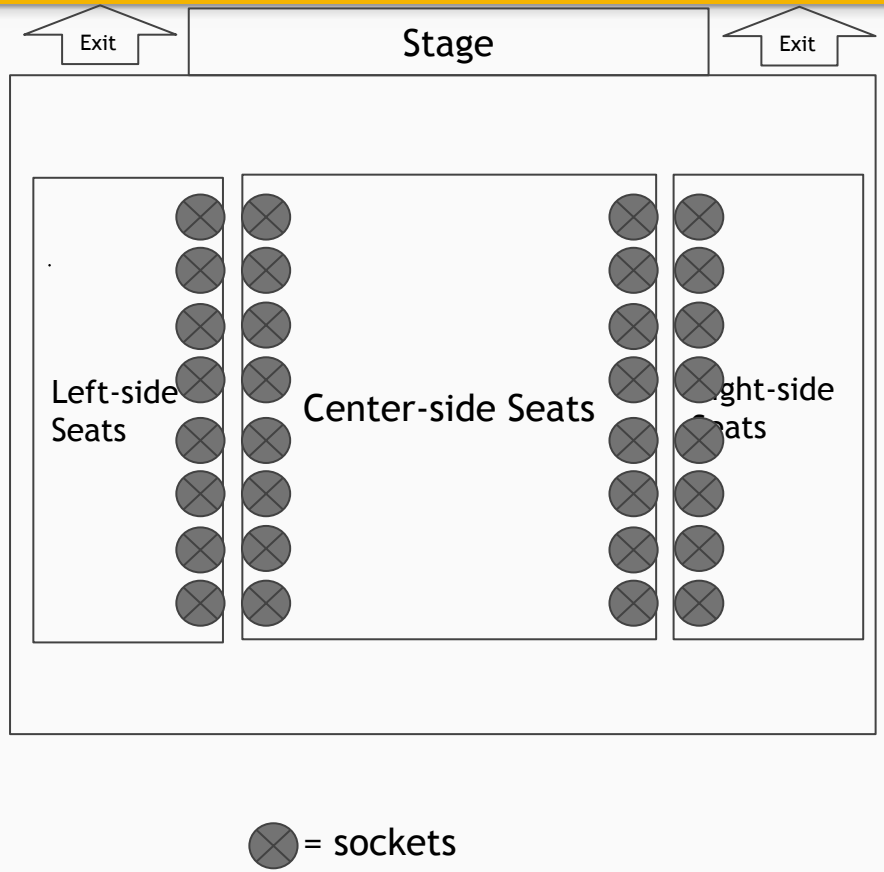
Yasumoto International Academic Park - YIA LT6

Limited Mobile Signal. Please use on-campus wifi.
32 Sockets. But please bring your own charger.



Python Pandas, Numpy Basics Introduction to Data Science

CUHK MSc Data Science & Biz Stat. Program
STAT5106 - Programming Techniques for Data Science
Week 4 @ 3 Oct 2024

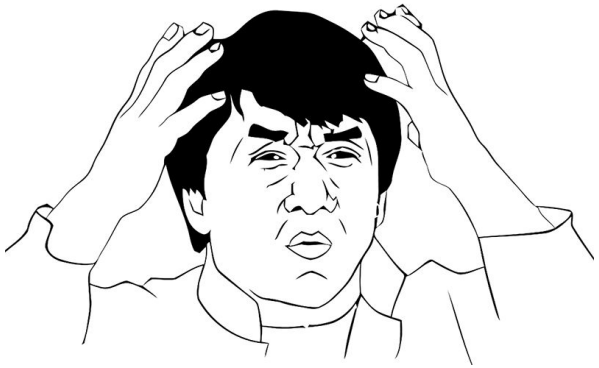
What is Data Science ?

- 1960 - Computer Scientist Peter Naur introduced the term “datalogy”
- 1962 - Statistician John Tukey wrote a paper titled “The Future of Data Analysis”, referring to the merging of statistics and computer
- 1974 - Naur published Concise Survey of Computer Methods, which freely used the term **data science** in its survey of the contemporary data processing methods.
- 1977 - Tukey wrote a second paper, titled Exploratory Data Analysis, arguing the importance of using data in selecting “which” hypotheses to test, and that confirmatory data analysis and exploratory data analysis should work hand-in-hand
- 1997 - Jeff Wu , the father of EM, resampling methods, presented his lecture entitled "Statistics = Data Science?". He characterized statistical work as a trilogy of **data collection, data modeling and analysis, and decision making**, and advocated that **statistics be renamed as data science**; and statisticians as data scientists.
- 2001 - William S. Cleveland , American Computer Scientist and Statistician, introduced data science as an independent discipline, extending the field of statistics to incorporate "**advances in computing with data**"
- 2002 - Data Science Journal started by CODATA; 2003 - The Journal of Data Science started by Columbia U
- 2012 - Harvard Business Review article "Data Scientist: The Sexiest Job of the 21st Century"
- UP TO NOW - THE INDUSTRY FLIES !!!

What is Data Science ?

- 1960 - Computer Scientist Peter Naur introduced the term “datalogy”
- 1962 - Statistician John Tukey wrote a paper titled “The Future of Data Analysis”, referring to the m
- 1974 - **ata scien**
- 1977 - ce of using d explo
- 1997 - tistics = Data **mode as data s**
- 2001 - ta science as an independent discipline, extending the field of statistics to incorporate "**advances in computing with data**"
- 2002 - Data Science Journal started by CODATA; 2003 - The Journal of Data Science started by Columbia U
- 2012 - Harvard Business Review article "Data Scientist: The Sexiest Job of the 21st Century"
- UP TO NOW - THE INDUSTRY FLIES !!!

WAIT...Then **what is the definition of DATA SCIENCE**... ?

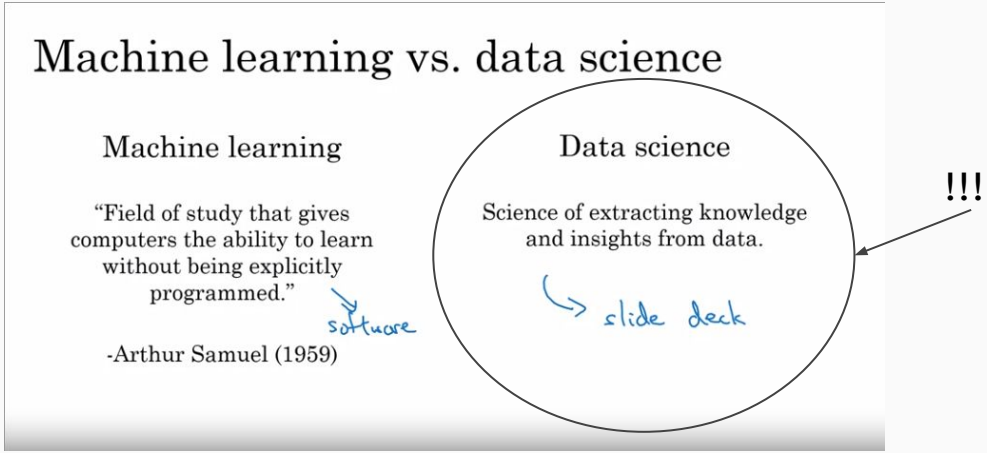


What is Data Science ?

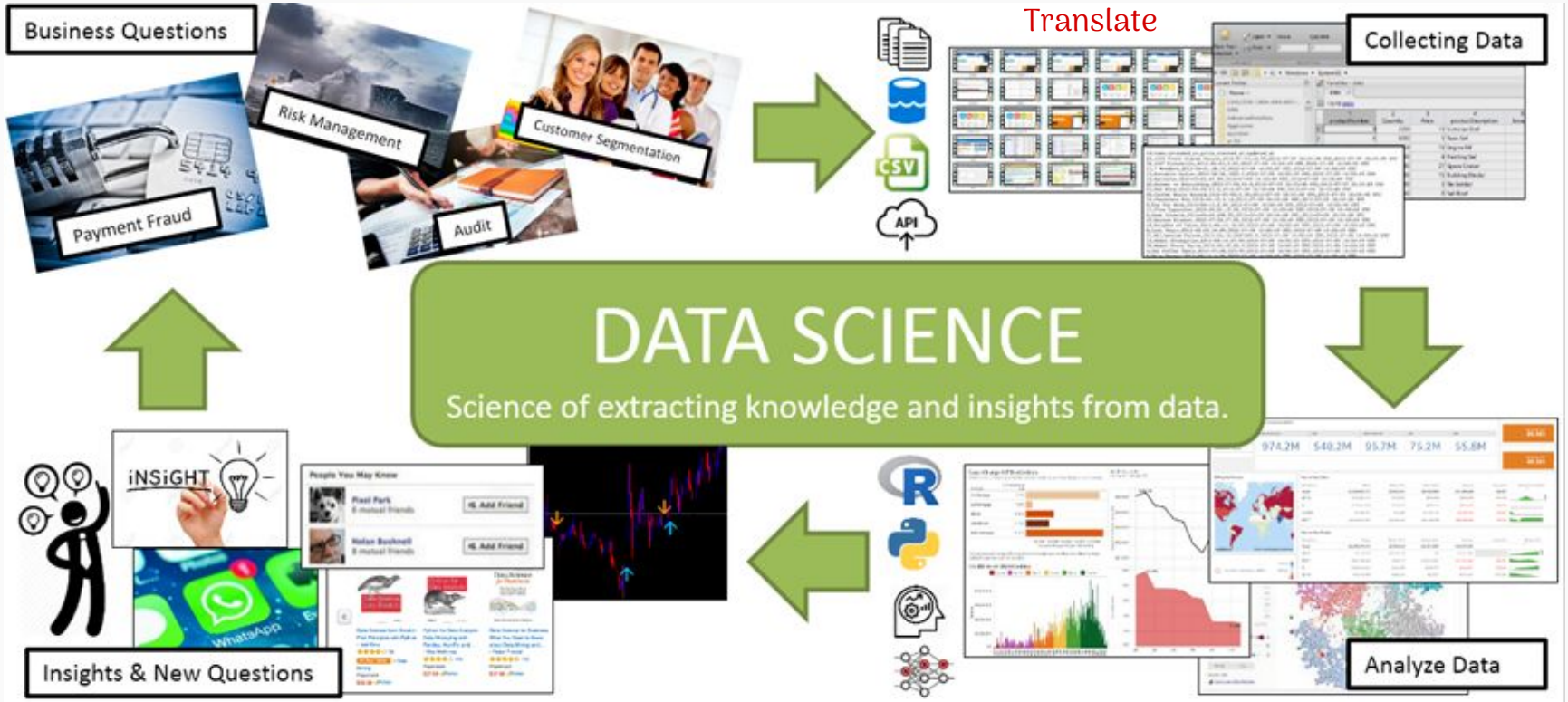
The WAR, still continues…

- Statistician View:
  John Chambers urges statisticians to adopt an **inclusive concept of learning from data**
- Computer Scientist View:
  William Cleveland urges to prioritize extracting **from data applicable predictive tools over explanatory theories**

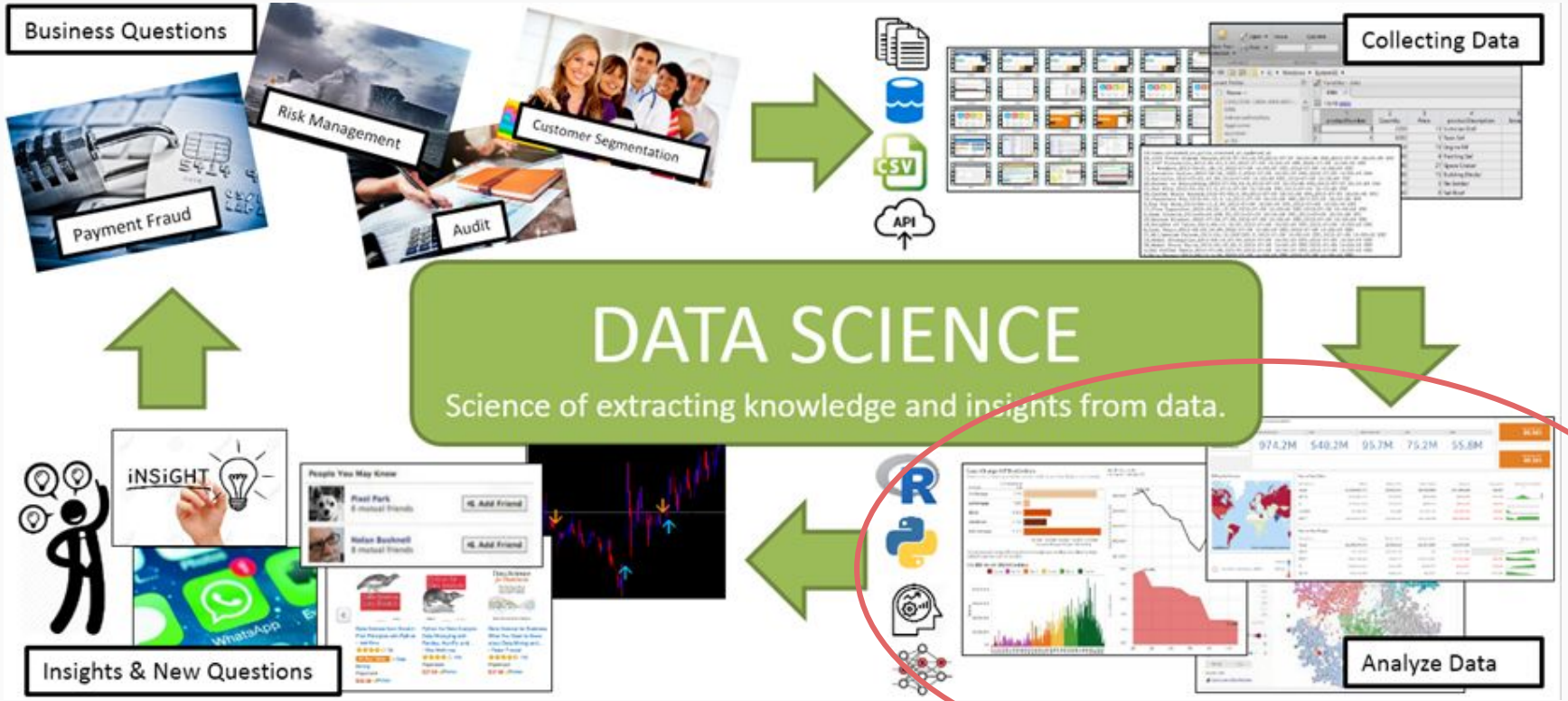
For me, Andrew Ng , co-founder of Coursera , said in the course AI for Everyone .



What is Data Science ?



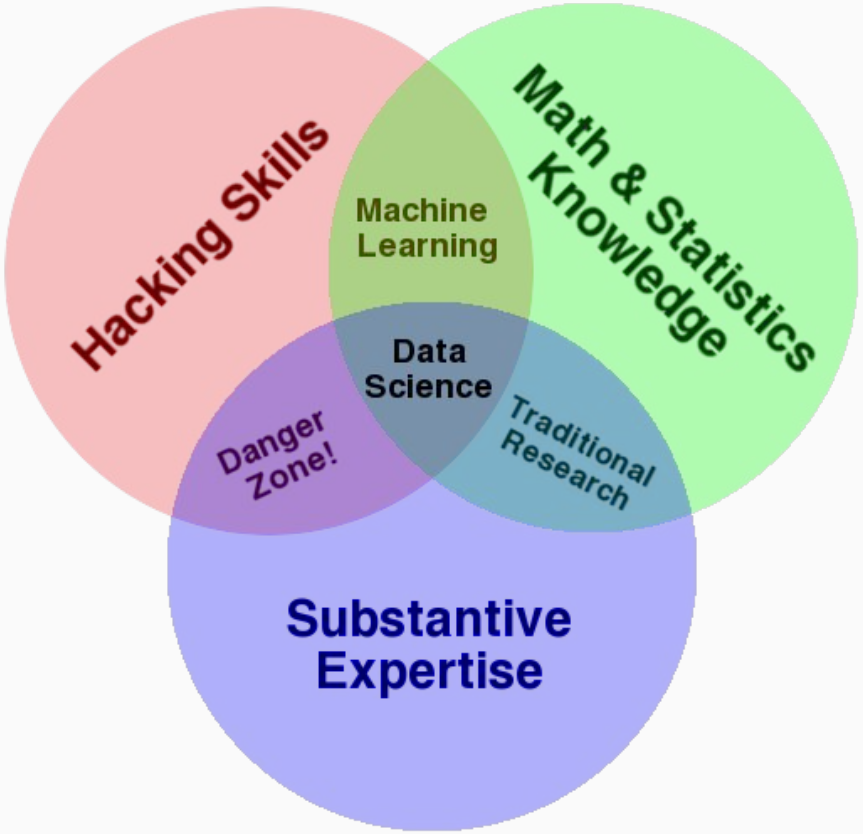
What is Data Science ?



In this course, most of the time on here...

Data Science Venn's Diagram

- Purposed by Drew Conway (Homepage)
- Details
- Misleading ?
  - Hard Skills Only ?
- How to utilize the venn's diagram
  - Balance for each dimension in your career



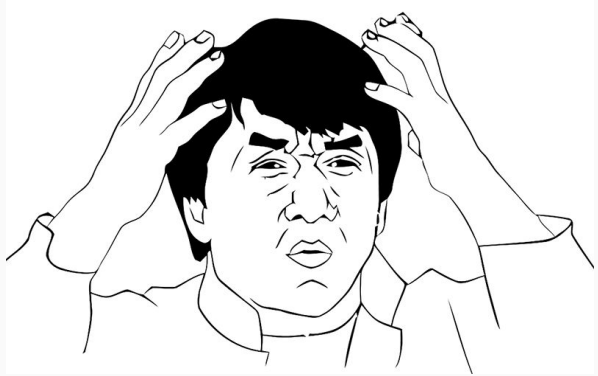
Okay. This is Data Science.
Science of Extracting Knowledge and Insights from data.

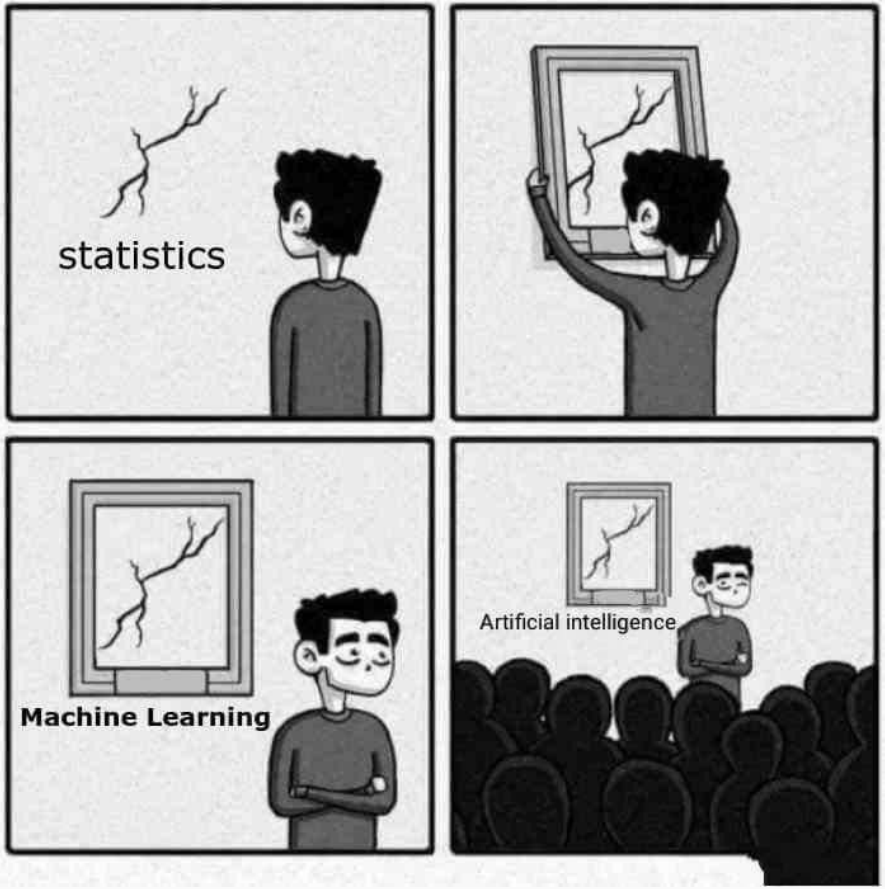
Then what is AI ?

What is AI ?

This brings out the following

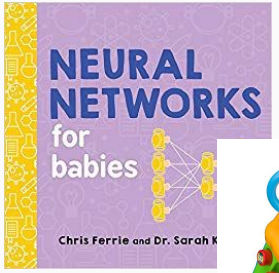
- What is Statistics ? (You are studying)
- What is Machine Learning ? (Slide 6)
- What is Big Data ? (Week 3)
- What is AI ?

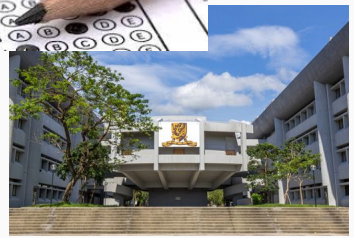




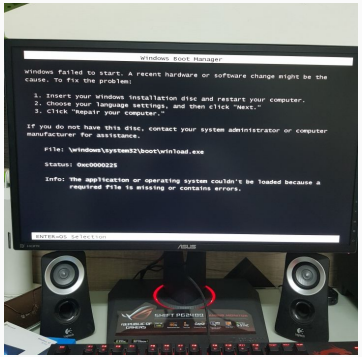
What is AI ?

Natural Intelligence

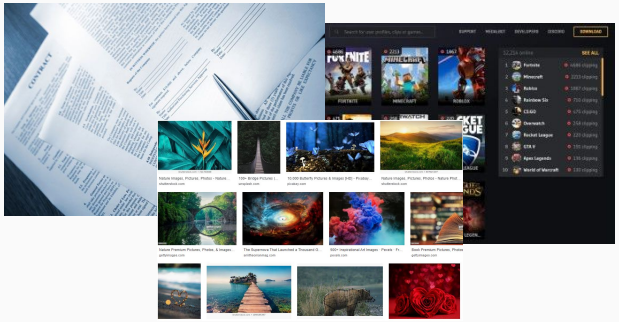




Artificial Intelligence

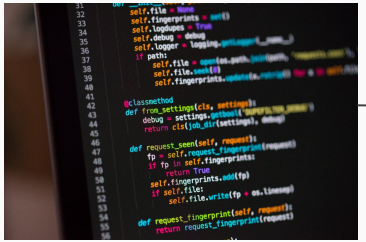


(Big) Data



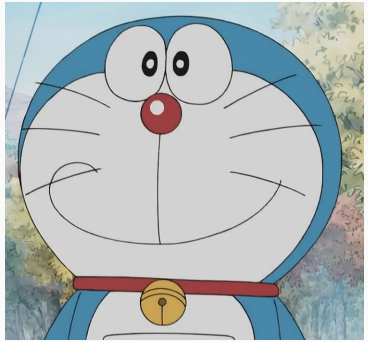
+

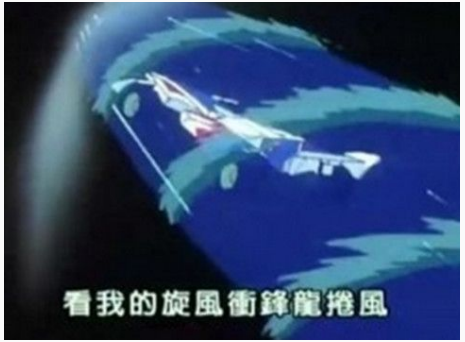
Algorithm

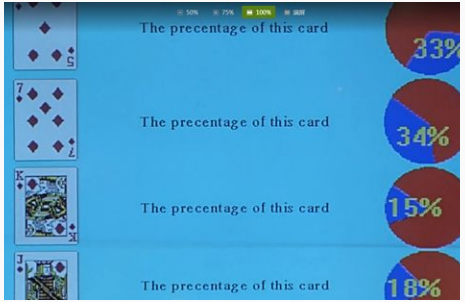


What is AI ?

Select all images about AI.







VERIFY

Okay. This is Data Science.
Science of Extracting Knowledge and Insights from data.

Then what is AI ?

<u>Realization of Data Science.</u>

Start Coding…

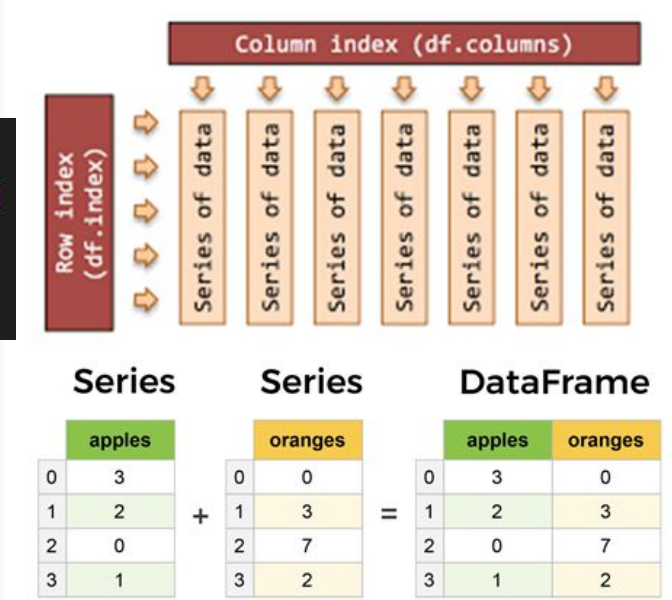
Please access…Week 4 pandas ; Week 4 numpy

Introduction to pandas

- Initial release: 11 January 2008
- Author: Wes McKinney – Developer
- Need for Quantitative Analysis on financial data
- Pandas = The short from PANel DAta

Series / DataFrame

```
flights_col = {
    'data no': [1, 2, 3]
    , 'date': ['2022-08-08 07:50', '2022-08-08 08:10', '2022-08-08 09:30']
    , 'flight no': ['CX 958', '5J 111', 'NH 812']
}
df_flights_col = pd.DataFrame(flights_col)
print(df_flights_col)
```



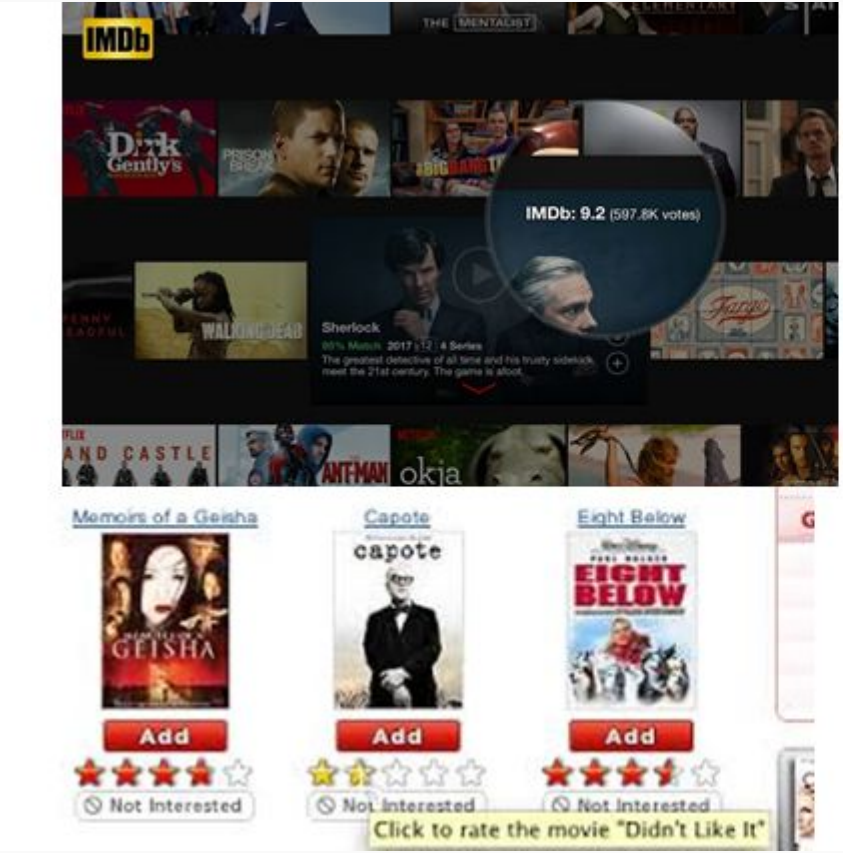
Data basics, Python pandas with R

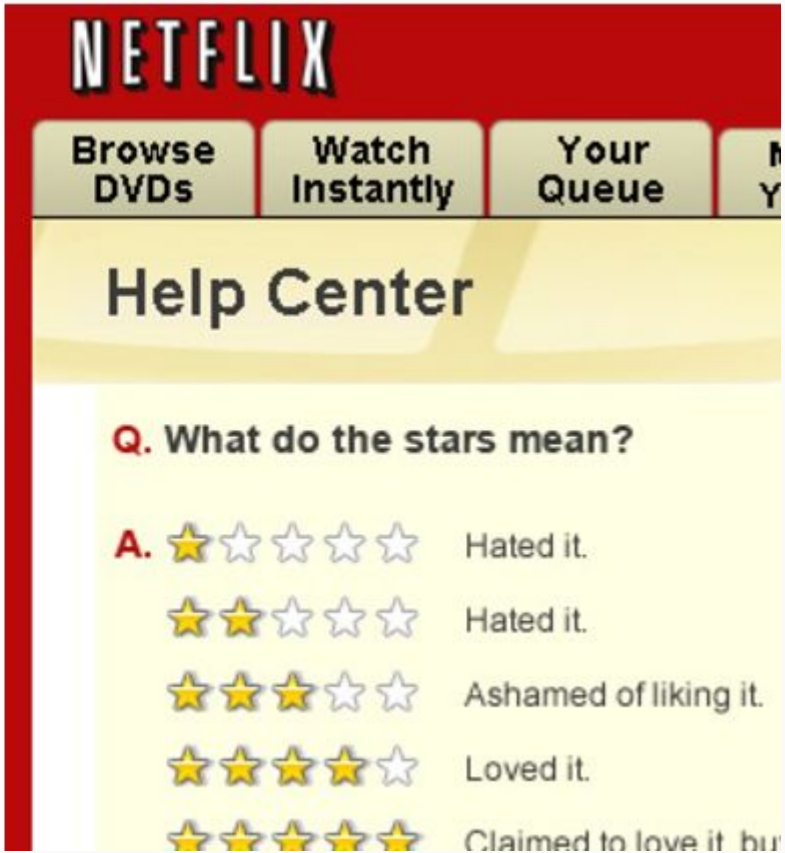
	In Python pandas	In R
create dataset	`df = pd.DataFrame({'a': [1,2,3], 'b': [4,5,6]`	`d <- data.frame(a=c(1,2,3), b=c(4,5,6))`
select columns	`df[ [List of column names] ]`	`d %>% select(c( colnames )) / d[, colnames]`
select rows	`df.loc[ [List of index names] / conditions ]` `df.iloc[ no. of rows ]`	`d %>% filter( conditions ) /` `d[c(...)/conditions, ]`
Select elements	`df.loc[ index names , column names ]` `df.iloc[ no. of rows , no. of columns ]`	`d[ rownames / no. of rows , colnames / no. of cols ]`
Quick overview	`df.info() , df.describe() , df.head() , df.tail()`	`str(d), summary(d), head(d), tail(d)`
read table / csv read xlsx	`df.read_table(...) / df.read_csv(...)` `df.read_excel(...)`	`read.csv() / read_csv() / fread()` `library(openxlsx / xlsx / readxl)`
write table / csv write xlsx	`df.to_csv(...)` `df.to_excel(...)`	`write.csv() / write_csv() / fwrite()` `library(openxlsx / xlsx / writexl)`

Data Pre-processing, Python pandas vs R

	In Python pandas	In R
Sorting	`df.sort_values([cols], ascending=[TF vectors])`	`d %>% arrange(col1, -col2)`
Merging	`df1.merge(df2, how, left_on, right_on)`	`d %>% xxxx_join(d2, by=(c(col1=col2)))`
Transforming	`df['new_col'] = any_transformations…`	`d %>% mutate(new_col = any_trans…)` `d$new_col = any_trans…`
Aggregating	`df.groupby(grouped_col).agg(dictionary…)`	`d %>% group_by(grouped_col) %>% summarise(...)`
delete rows delete columns	`df.drop(row_nos)` `del df[ [the_cols] ]`	`d[(-row_nos), ]` `d %>% select(-the_cols)`
drop NA	`df.dropna()`	`d %>% na.omit()`
drop duplicates	`df.drop_duplicates()`	`d %>% unique()`

Pandas Example - Movielens





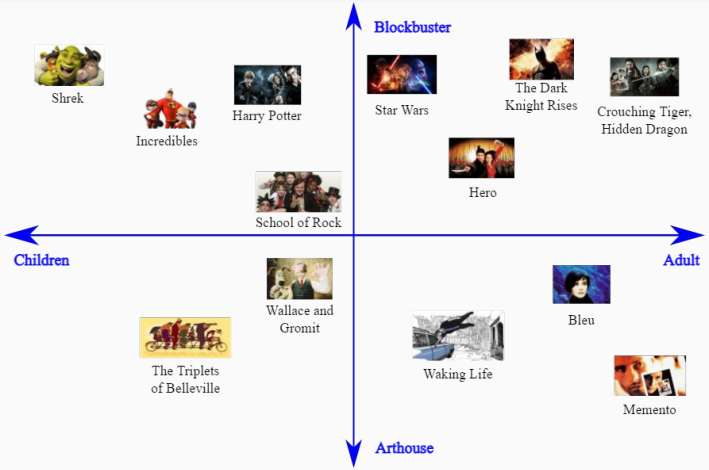
Grouplens, Netflix Price, and hence Recommender System

History

- 1992: Grouplens founded for collaborative filtering research, as well as the core studies of recommender system
- 2006 - 2009: Netflix Prize (should be the first concept of Hackathon), $1M prize returning $1.4B revenue
- and hence application exploding…
- More reference on this

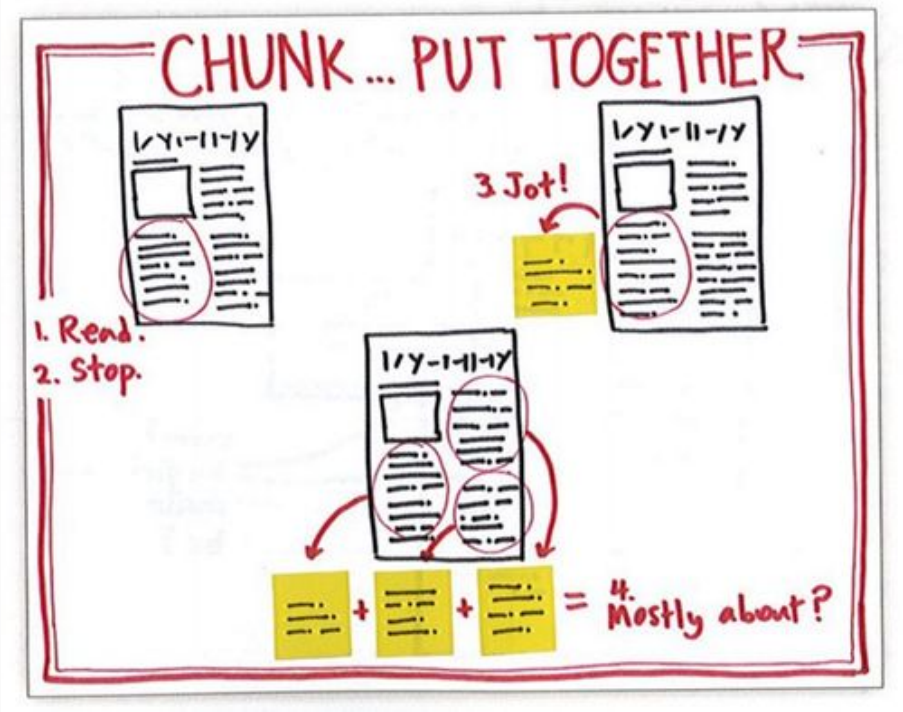
Tools needed

- PATTERN RECOGNITION
- Dimension Reduction / Embedding / Vectorization, such as PCA
  → The n nearest neighbourhood
- Python package surprise



Chunk in pd.read_csv

In case: a 100GB csv file? How to read?



- df_chunk = read_csv(filename, chunksize = 1024)
- Chunksize best choice = 2^N (1024, 2048, …) , depends to you PC RAM

Generators

```python
def main():
    with open("outfile.txt", "w") as outfile:
        for outrow in lowercase_rows("chromosome_y.fa"):
            outfile.write(outrow)
        outfile.close()

def lowercase_rows(filename):
    with open(filename) as infile:
        for row in infile:
            yield (row.lower())
        infile.close()

if __name__ == '__main__':
    main()
```

A yield ("return") for every iteration, so it can be used immediately without temp data structure == GOOD!

```python
# Initiate the generator
rows_gen = lowercase_rows('bigfile.txt')

print( next(rows_gen) )
# print the 1st line with lowercase
print( next(rows_gen) )
# print the 2nd line with lowercase
print( next(rows_gen) )
# print the 3rd line with lowercase
print( next(rows_gen) )
# print the 4th line with lowercase
...
```

Introduction to NumPy

- support for large, multi-dimensional arrays and matrices, along with mathematical functions to operate on them
- Older Array Packages
  - In 1995, package called “Numeric” released.
  - Somewhile, another package called “numarray” released.
  - In 2006, “Numeric” and “numarray” are integrated to the new package “NumPy”, by Travis Oliphant (After that he founded Anaconda)

Matrix Computation, Comparing with R

	In Python numPy	In R
array creation	`m = np.array([row1, row2])`	`m <- matrix(c(row1, row2), nrow, ncol)`
array dimension	`m.shape`	`dim(m)`
Sequence Interpolation	`np.arange(from, to, by)` `np.linspace(from, to, length)`	`seq(from, to, by)` `seq(from, to, length=...)`

1-matrix
0-matrix
identity matrix
	`np.ones([nrow, ncol])`
	`np.zeros([nrow, ncol])`
	`np.eye(n)`
		`matrix(0, nrow, ncol)`
		`matrix(1, nrow, ncol)`
		`diag(rep(1, n))`

diagonalization / get diagonal	`np.diag(list/m)`	`diag(vector/m)`
repeat arrays	`np.array(list * n) / np.repeat(m, n)`	`rep(vector, n)`
combining arrays	`np.vstack([m1, m2]) / np.hstack([m1, m2])`	`rbind(m1, m2) / cbind(m1, m2)`
Matrix Product	`m1.dot(m2)`	`m1 %*% m2`
transpose	`m.T`	`t(m)`
Inverse	`np.linalg.inv(m)`	`solve(m)`

About Mid-Term

Mid-Term Take Home Exam Instructions.ipynb

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

CUHK STAT5106 Mid-Term Take Home Exam

25 Oct 2024 (Fri), 2200 - 27 Oct 2024 (Sun), 2159

48 hours are given, but 9 hours are expected to be finished.
This is open-book; open-notes; and open-web exam. Welcome to utilize any resources.
Please follow the instructions step-by-step:

- Please access into Blackboard > STAT5106 > Course Content > Mid-Term Take Home Exam
- Access the link below, copy the ipynb to anywhere such as local computer, or personal cloud drive, so this becomes savable.
  For the best experience - using **colab and your google drive account**.
- Code and complete all questions in this ipynb.
- Please keep on saving for protecting what you have input.
  Finally you need to upload this to Blackboard account.
- Upload this in Mid-Term Take Home Exame page, liked how you have submitted assignments.

For any queries, please submit your questions on BLACKBOARD called **MID-TERM discussion room**. During the mid-term exam period - in the 48 hours. For fair the publicity to all students.

Alan will answer most questions around the following time point:

- 1000, 26 Oct 2024
- 2200, 26 Oct 2024
- 1000, 27 Oct 2024

Some of them I would not answer if your question is not suitable - such as asking the full answers.

DON'T email to tutors or me - we won't answer in the period.

Thanks and good luck,
Alan

Just FYI: Python vs R …

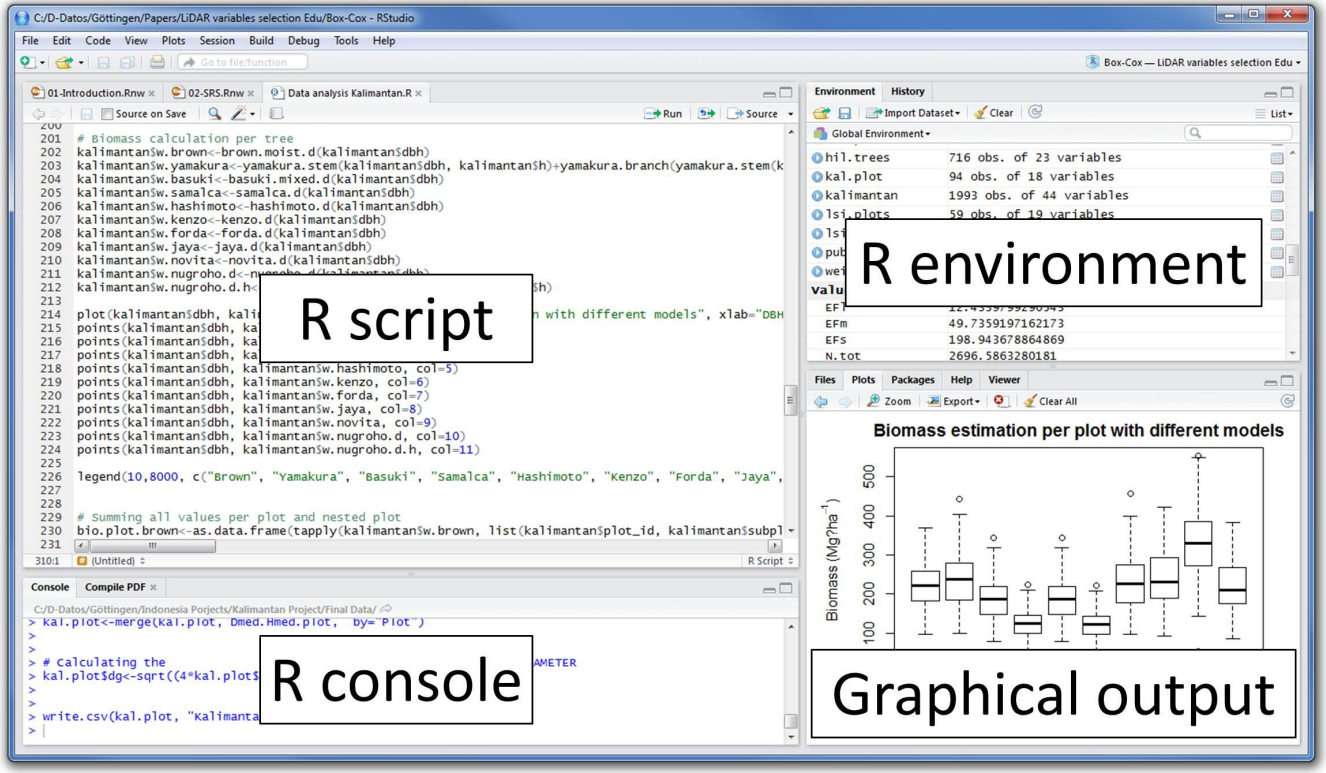
R History

- initial version released in 1995
- by Ross Ihaka and Robert Gentleman, Statisticians at the Auckland U
- named partly after the first names of the first two R authors
- Reference is here

- Version
  - R 1.0 - 2000-02-29
  - R 2.0 - 2004-10-04
  - R 3.0 - 2013-04-03
  - R 4.0 - 2020-04-24
- Recent – R 4.3 - 2023-06-16

Install and Quick-Start on R Studio

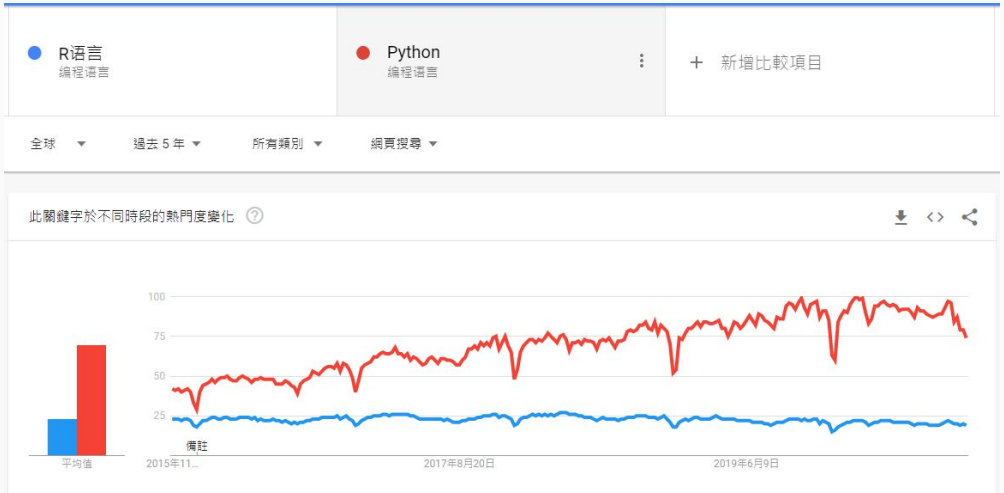
1. Install R gui at https://cloud.r-project.org
2. Install R Studio at https://www.rstudio.com/products/rstudio/download/

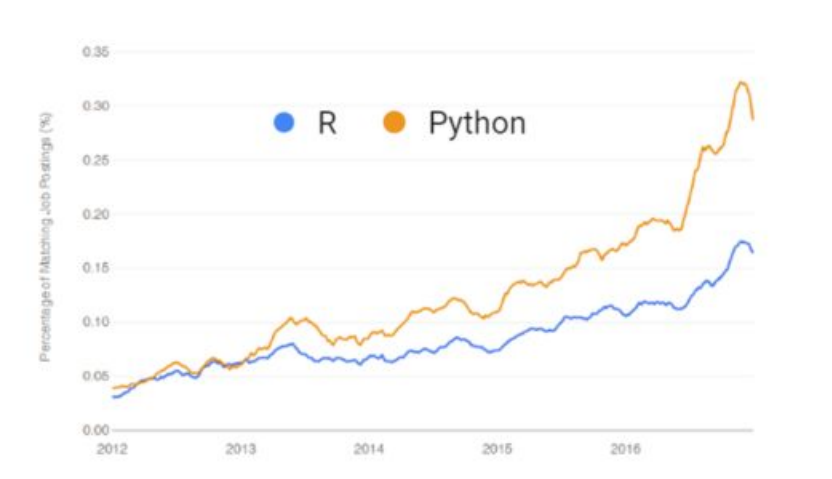


R vs Python

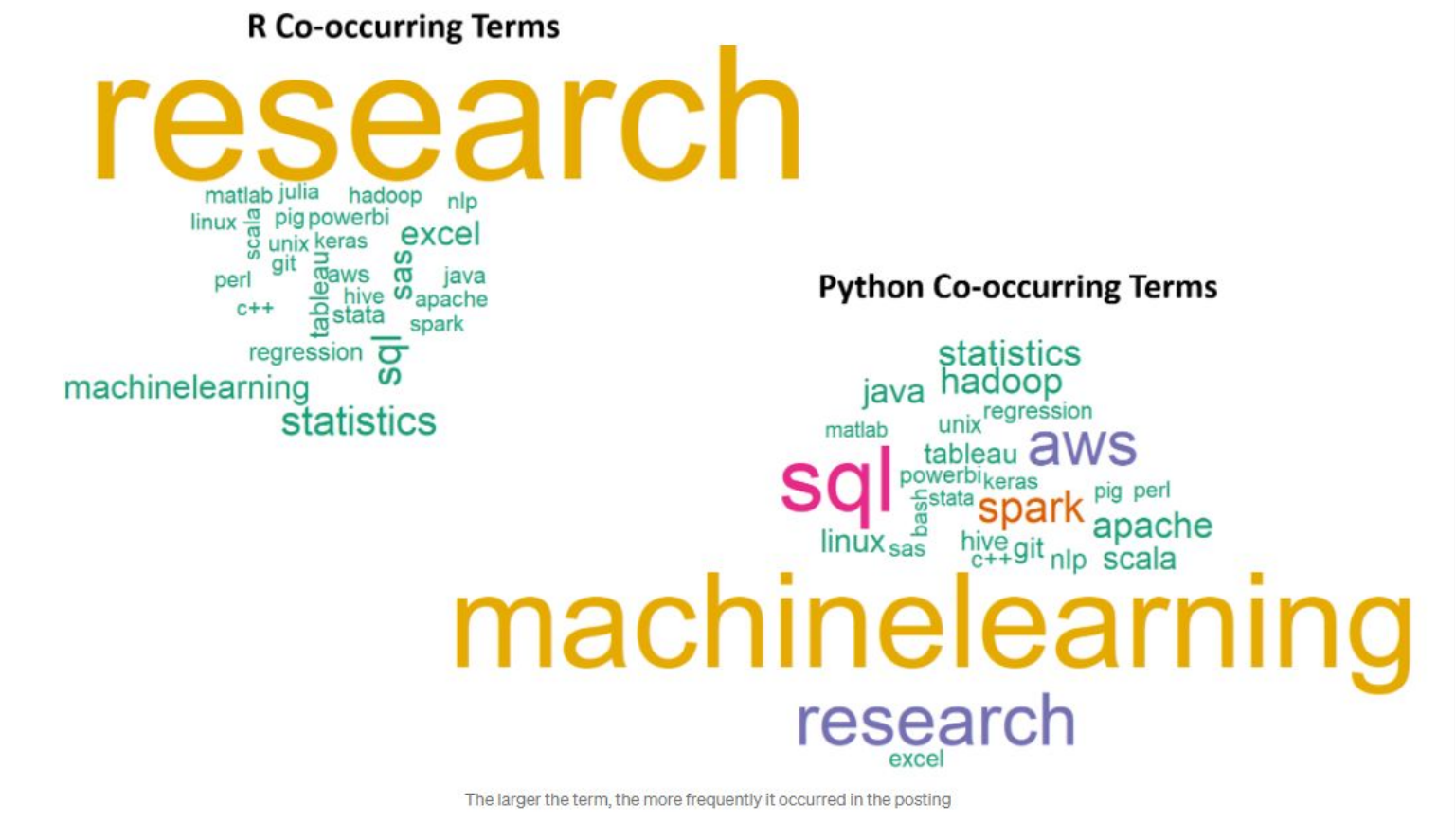
Must Knows	Must Knows
1. R is an implementation of S programming language (Bell Labs). 2. R's design and evolution is handled by the R-core group and R foundation. 3. R's software environment was written primarily in C, Fortran and R.	1. Python was inspired by C, Modula-3, and particularly ABC. 2. Python gets its name from the "Monty Python's Flying Circus" comedy series. 3. Python Software Foundation (PSF) takes care of Python's advances.
Purpose	
R focuses on better, user friendly data analysis, statistics and graphical models.	Python emphasizes productivity and code readability.
Used By?	
R has been used primarily in academics and research. However, R is rapidly expanding into the enterprise market. *"The closer you are to statistics, research and data science, the more you might prefer R."*	Python is used by programmers that want to delve into data analysis or apply statistical techniques, and by developers that turn to data science. *"The closer you are to working in an engineering environment, the more you might prefer Python."*

R vs Python


R语言
編程語言
Python
編程語言
新增比較項目
全球
過去 5 年
所有類別
網頁搜尋
此關鍵字於不同時段的熱門度變化
100
75
50
25
備註
平均值
2015年11...
2017年8月20日
2019年6月9日


R
Python
Percentage of Matching Job Postings (%)
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
2012
2013
2014
2015
2016

R vs Python

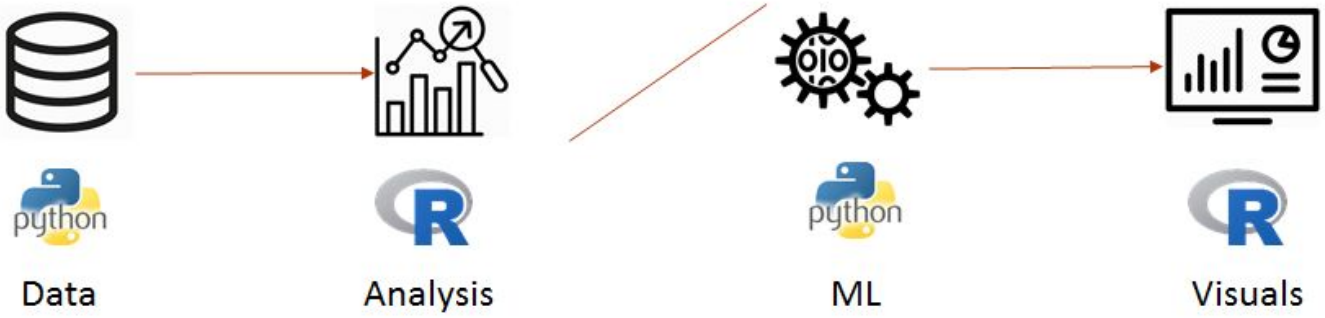

R Co-occurring Terms
research
matlab
julia
hadoop
nlp
linux
pig
powerbi
scala
unix
keras
excel
git
perl
aws
tableau
sas
java
hive
apache
c++
stata
spark
regression
sql
machinelearning
statistics
Python Co-occurring Terms
statistics
hadoop
java
regression
matlab
unix
aws
tableau
sql
powerbi
keras
bash
stata
spark
pig
perl
apache
linux
sas
hive
c++
git
nlp
scala
machinelearning
research
excel
The larger the term, the more frequently it occurred in the posting

R vs Python – My Practice

- R
  - For data analysis / mining / visualization
  - For data analytics perspective
  - For STATISTICAL Modelling use
- Python
  - For large scale data flow / engineering
  - For web scraping / getting web data
  - For Machine Learning / Deep Learning Modelling use
  - For productionalizing DS projects

"The closer you are to statistics, research and data science, the more you might prefer R."

"The closer you are to working in an engineering environment, the more you might prefer Python."



To be continue…