MIS 572:

Introduction to Big Data Analytics

Course Overview & Introduction

Yihuang K. Kang

"If you torture the data enough,

nature will always confess."

Ronald H.Coase

Instructor

- **Yihuang Kang (康藝晃), PhD**
  - Email: ykang@mis.nsysu.edu.tw
  - Office hour: Tue & Wed 10am-12pm (by appointment)
  - Phone: +886 7-5252000 ext. 4737

- **Shot me an email with "MIS 572 - "**
  - e.g. "MIS 572 - A quick question about Homework #1"
  - Please organize your email before you hit "Send"!
- **Talk is cheap, show me the data**
- **Contact me before "it's too late"**

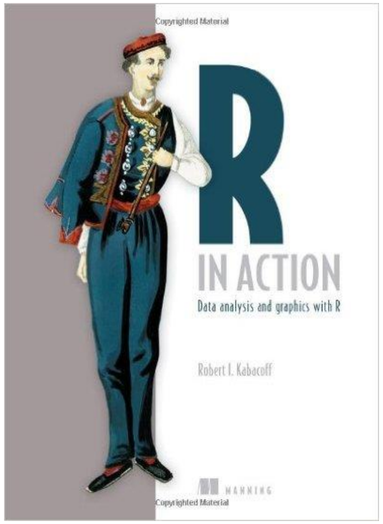
Syllabus

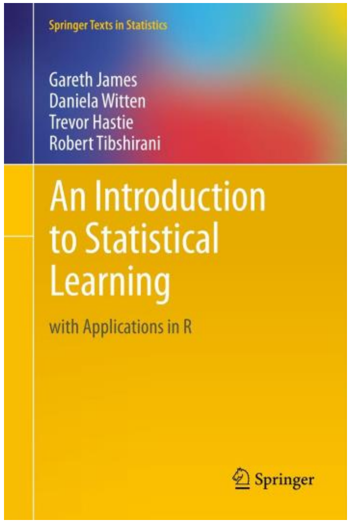
- **You will...**
  - exercise logical and computational thinking
  - sharpen your data analytics skills
  - learn how to use R to deal with "big data"

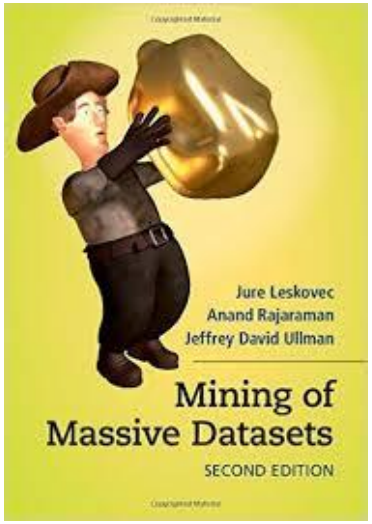
- **Time & Location**
  - Monday 3:10-6pm CM 3051

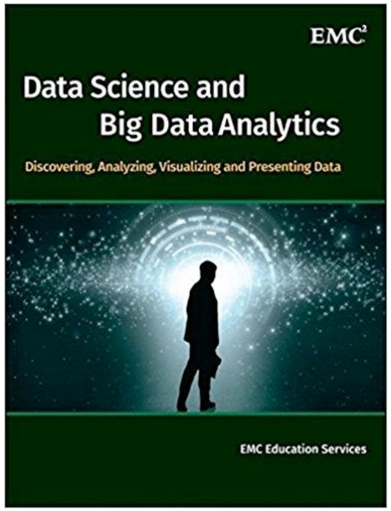
- **Teaching Assistant:**
  - 毛文瑞(rex850327@gmail.com)
  - 周詠捷(johnny83051202@gmail.com)
  - 楊宛蓁(w126251@gmail.com)
  - 郭博文(willy821002@gmail.com)

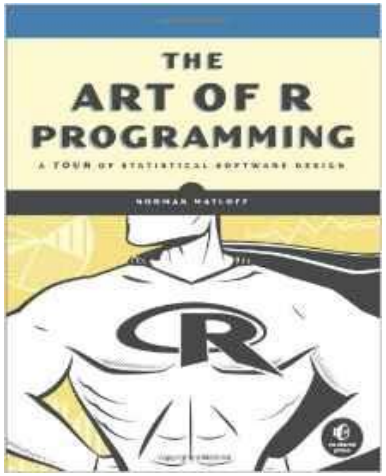
Syllabus (cont.)

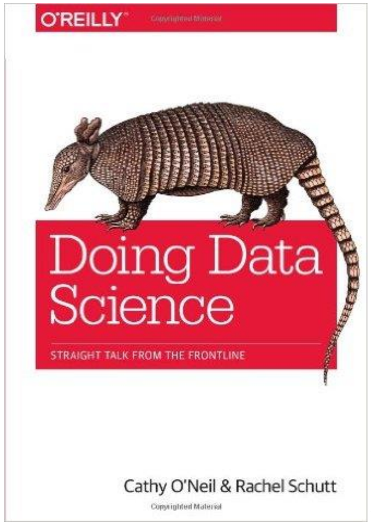




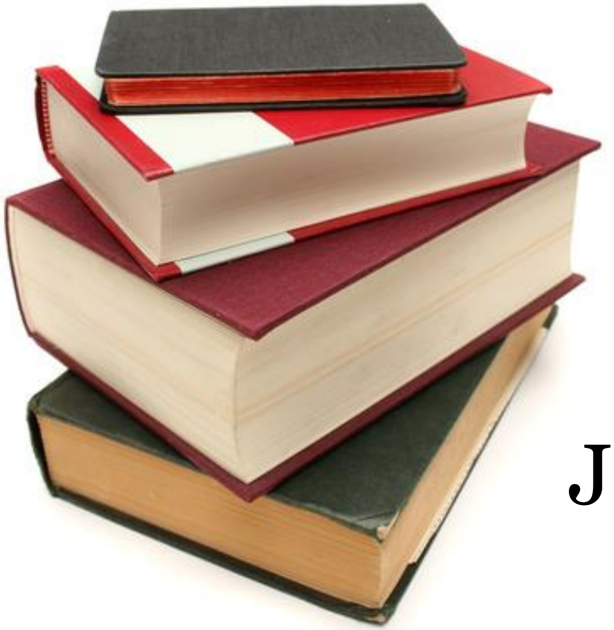



EMC²
Data Science and Big Data Analytics
Discovering, Analyzing, Visualizing and Presenting Data
EMC Education Services




O'REILLY
Doing Data Science
STRAIGHT TALK FROM THE FRONTLINE
Cathy O'Neil & Rachel Schutt

Syllabus (cont.)



As for the reading list…

Just read as many as you can!

Syllabus (cont.)

- **Grading**

  In-class group quiz: 30%   Homework: 30%

  Term project proposal: 20%   Term project defense: 20%

- **Term Project**

  ✓Organize your data science team

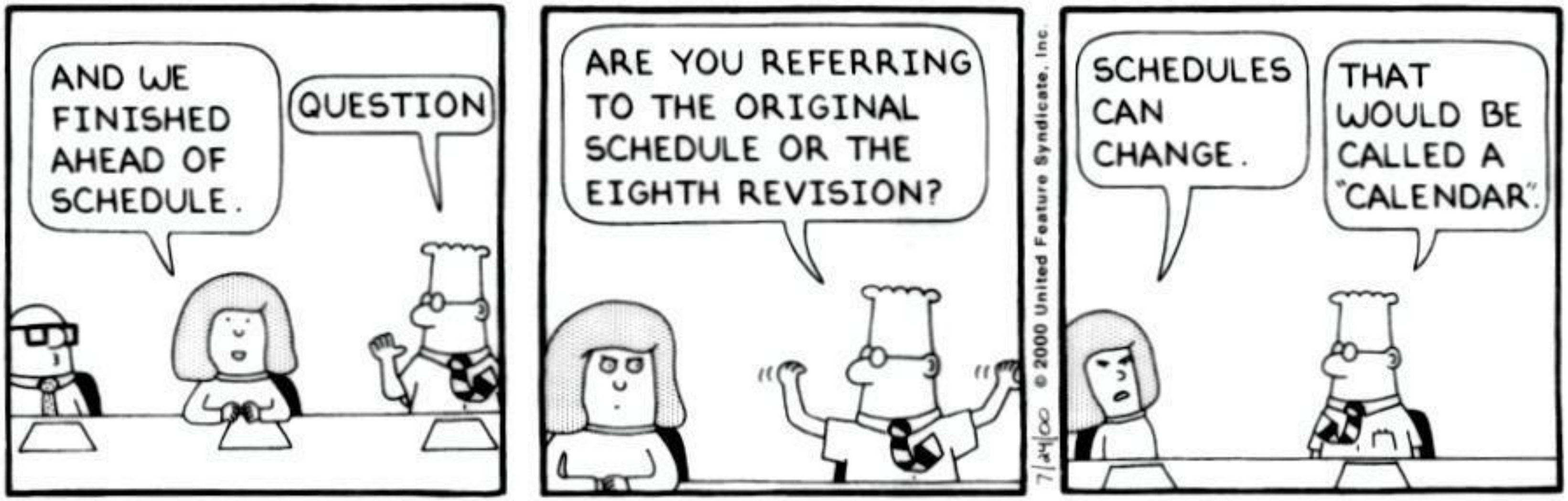
  ✓Group of **3-5** people

  ✓Get people with different backgrounds

Syllabus (cont.)

- **Schedule**

  Well, as always, it may change a bit later…


AND WE FINISHED AHEAD OF SCHEDULE.
QUESTION
ARE YOU REFERRING TO THE ORIGINAL SCHEDULE OR THE EIGHTH REVISION?
© 2000 United Feature Syndicate, Inc.
7/24/00
SCHEDULES CAN CHANGE.
THAT WOULD BE CALLED A "CALENDAR".

How to Survive

✓Practice makes perfect

✓Participate in class discussions

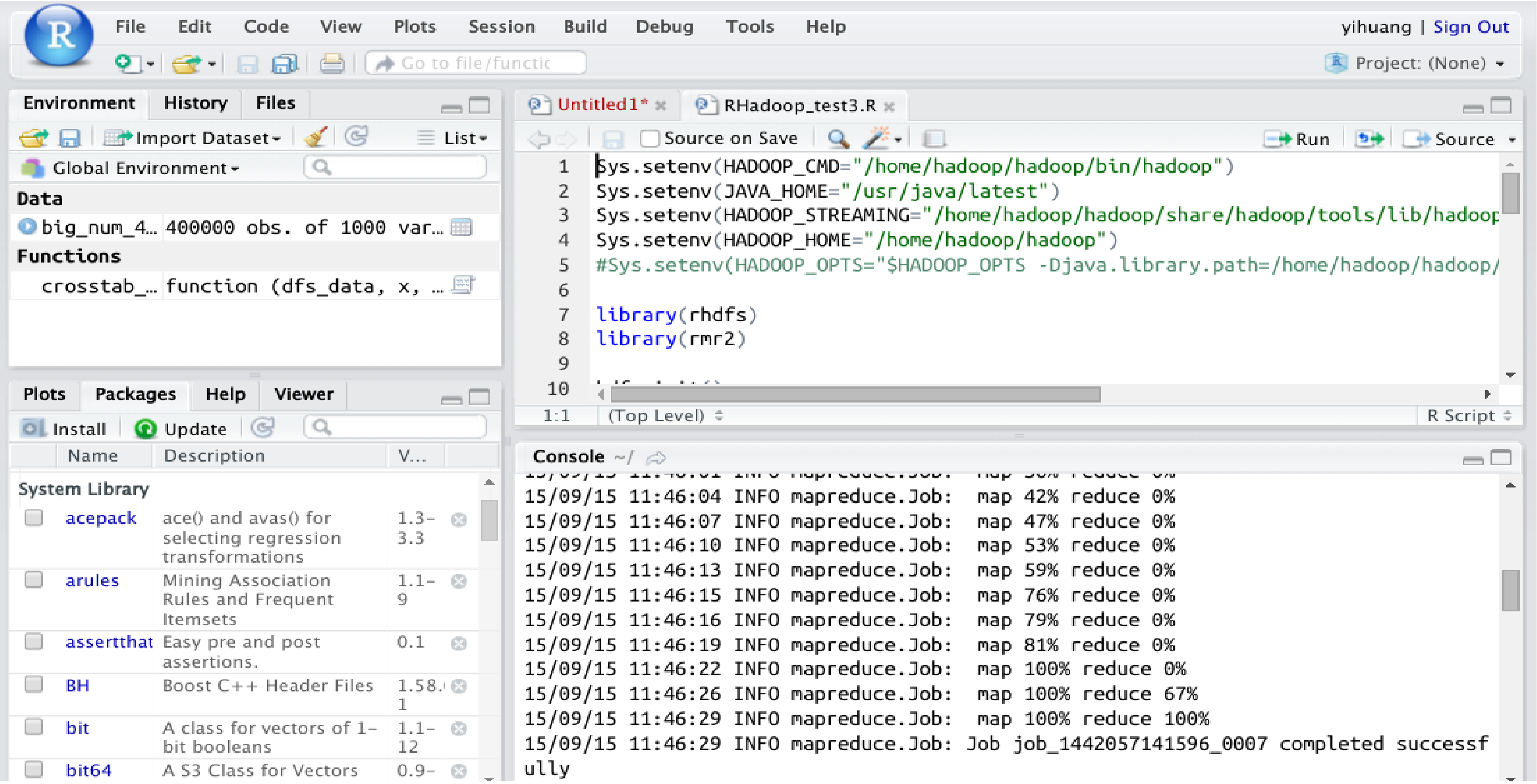
✓Love your data

✓Work ~~hard~~ ***smart!***

✓Ask geeks around you

CM Unified Analytics Platform

- RStudio Server: http://hdp.cm.nsysu.edu.tw:8787/



Data

big_num_4... 400000 obs. of 1000 var...

Functions

crosstab_... function (dfs_data, x, ...

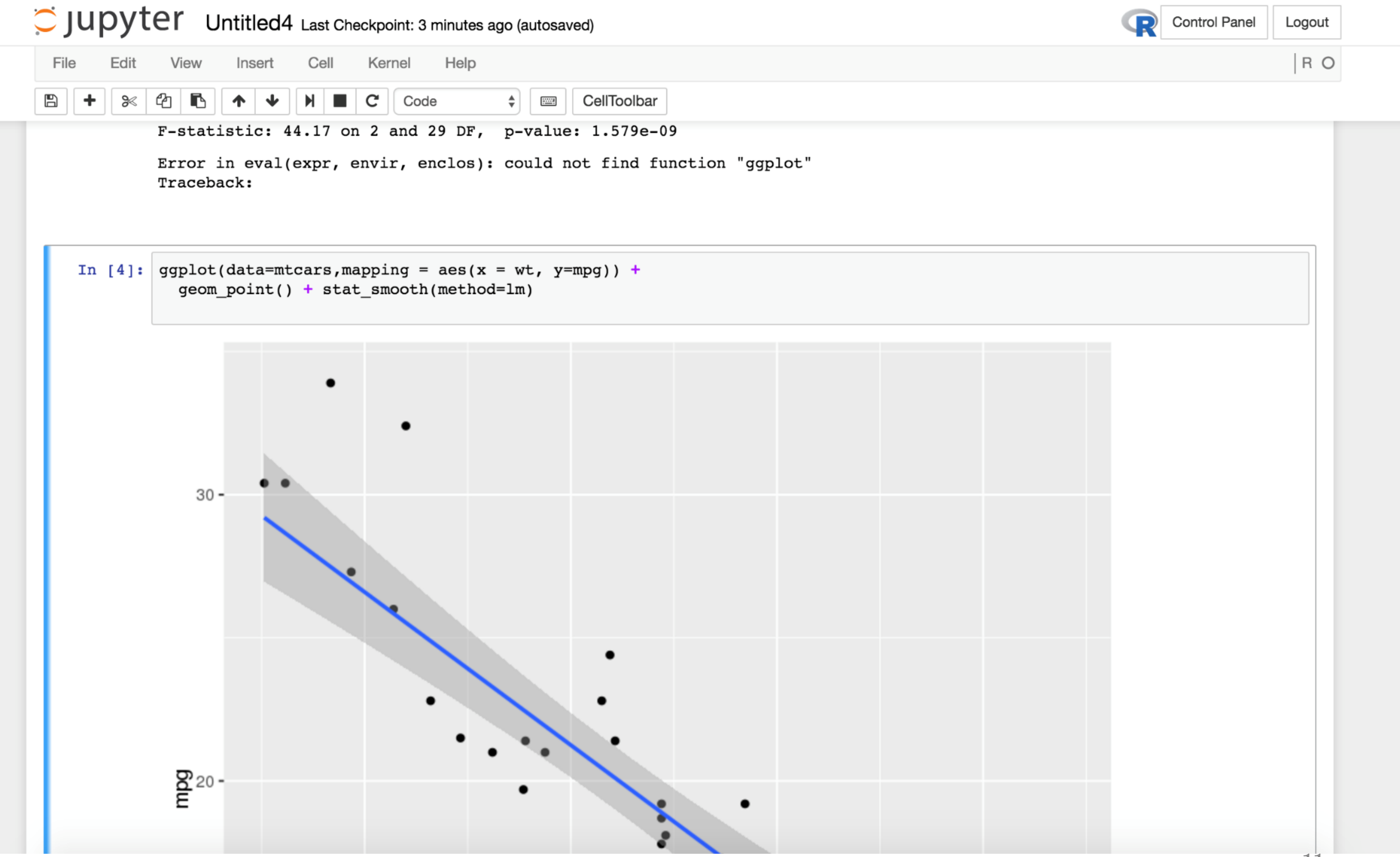
```
Sys.setenv(HADOOP_CMD="/home/hadoop/hadoop/bin/hadoop")
Sys.setenv(JAVA_HOME="/usr/java/latest")
Sys.setenv(HADOOP_STREAMING="/home/hadoop/hadoop/share/hadoop/tools/lib/hadoop
Sys.setenv(HADOOP_HOME="/home/hadoop/hadoop")
#Sys.setenv(HADOOP_OPTS="$HADOOP_OPTS -Djava.library.path=/home/hadoop/hadoop/

library(rhdfs)
library(rmr2)
```

	Name	Description	V...
	System Library		
□	acepack	ace() and avas() for selecting regression transformations	1.3-3.3
□	arules	Mining Association Rules and Frequent Itemsets	1.1-9
□	assertthat	Easy pre and post assertions.	0.1
□	BH	Boost C++ Header Files	1.58.1
□	bit	A class for vectors of 1-bit booleans	1.1-12
□	bit64	A S3 Class for Vectors	0.9-

```
15/09/15 11:46:04 INFO mapreduce.Job:  map 42% reduce 0%
15/09/15 11:46:07 INFO mapreduce.Job:  map 47% reduce 0%
15/09/15 11:46:10 INFO mapreduce.Job:  map 53% reduce 0%
15/09/15 11:46:13 INFO mapreduce.Job:  map 59% reduce 0%
15/09/15 11:46:15 INFO mapreduce.Job:  map 76% reduce 0%
15/09/15 11:46:16 INFO mapreduce.Job:  map 79% reduce 0%
15/09/15 11:46:19 INFO mapreduce.Job:  map 81% reduce 0%
15/09/15 11:46:22 INFO mapreduce.Job:  map 100% reduce 0%
15/09/15 11:46:26 INFO mapreduce.Job:  map 100% reduce 67%
15/09/15 11:46:29 INFO mapreduce.Job:  map 100% reduce 100%
15/09/15 11:46:29 INFO mapreduce.Job: Job job_1442057141596_0007 completed successf
ully
```

CM Unified Analytics Platform(cont.)



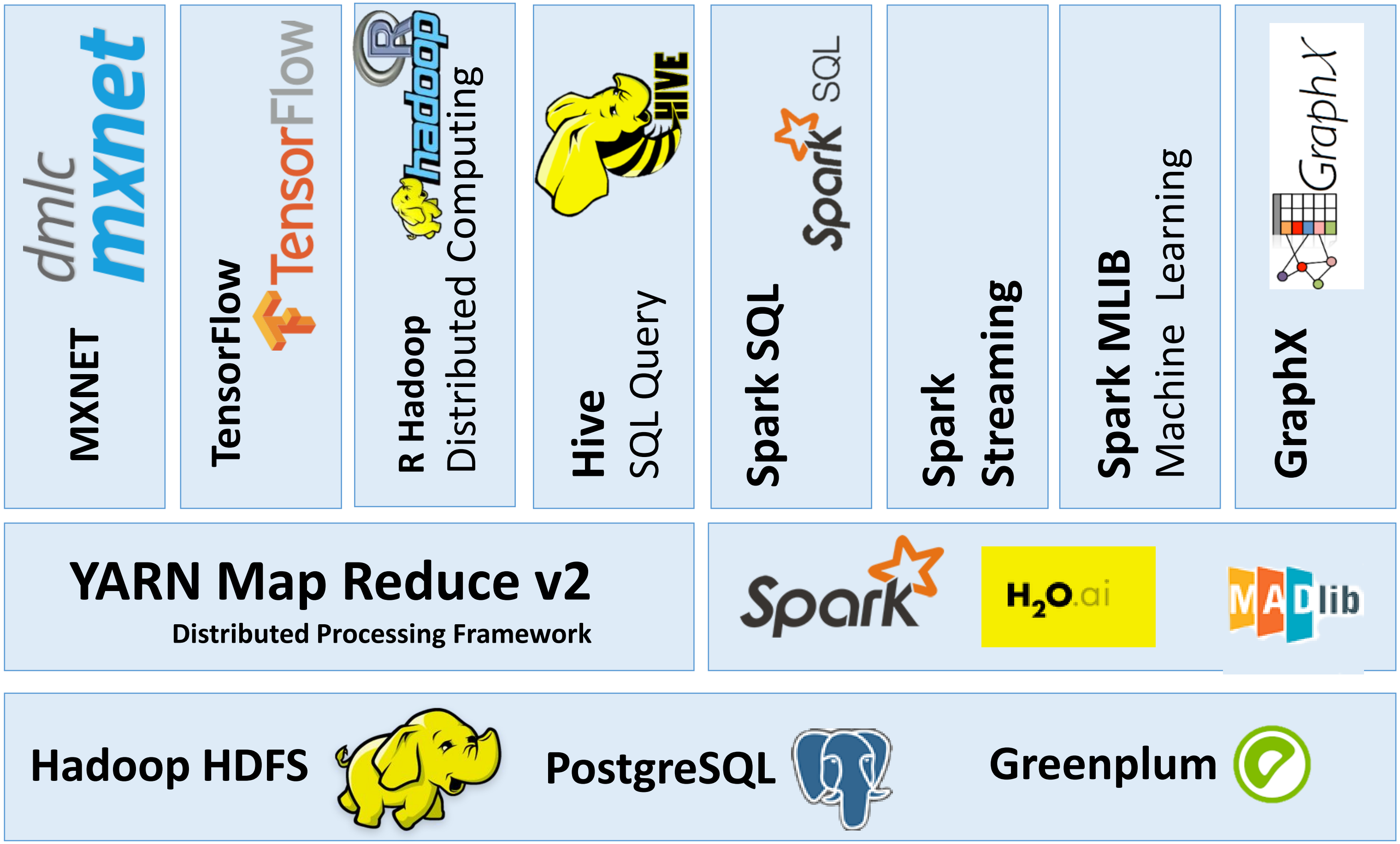
```
F-statistic: 44.17 on 2 and 29 DF,  p-value: 1.579e-09

Error in eval(expr, envir, enclos): could not find function "ggplot"
Traceback:
```

In [4]:
```
ggplot(data=mtcars,mapping = aes(x = wt, y=mpg)) +
  geom_point() + stat_smooth(method=lm)
```

CM Unified Analytics Platform(cont.)

CM Big Data Analytics Software Stack



MXNET

TensorFlow

R Hadoop
Distributed Computing

Hive
SQL Query

Spark SQL

Spark
Streaming

Spark MLIB
Machine Learning

GraphX

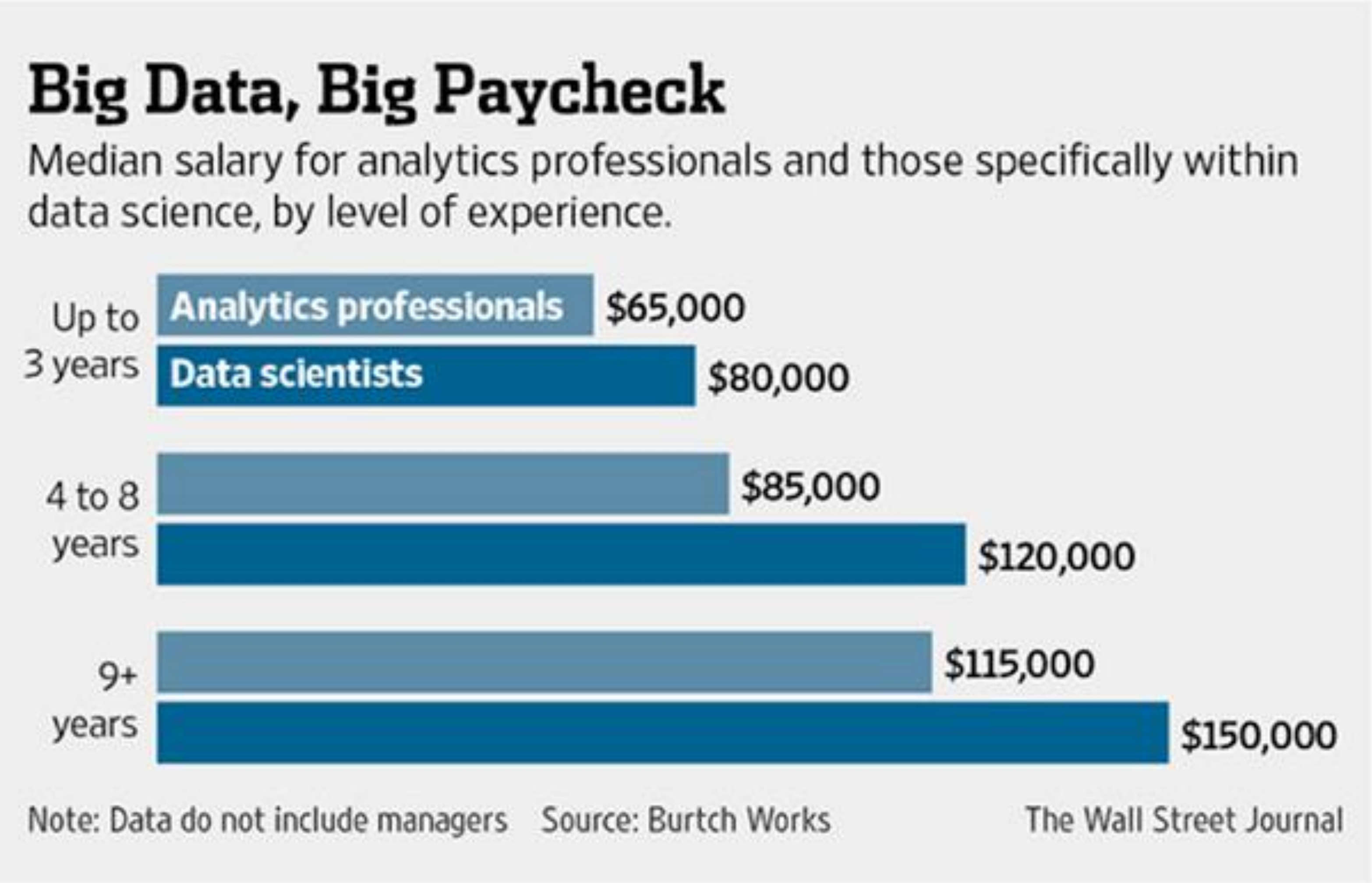
YARN Map Reduce v2
Distributed Processing Framework

Hadoop HDFS

PostgreSQL

Greenplum

> *"By 2018, the United States will experience a shortage of 190,000 skilled data scientists, and 1.5 million managers and analysts capable of reaping actionable insights from the big data deluge."*
>
> – *McKinsey Report, 2013*

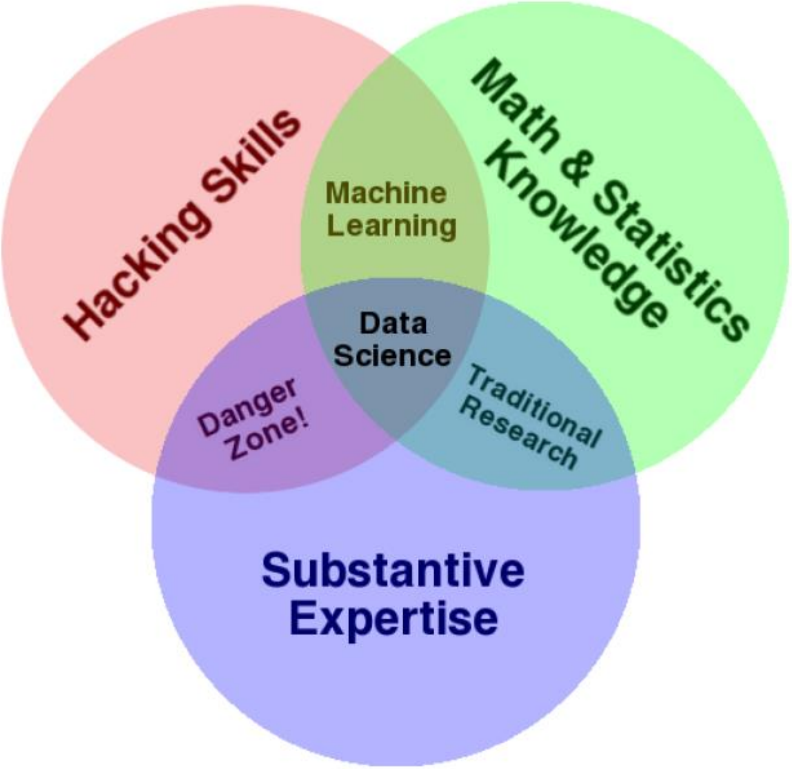


Big Data, Big Paycheck

Median salary for analytics professionals and those specifically within data science, by level of experience.

Note: Data do not include managers Source: Burtch Works The Wall Street Journal

What is "Data Science"

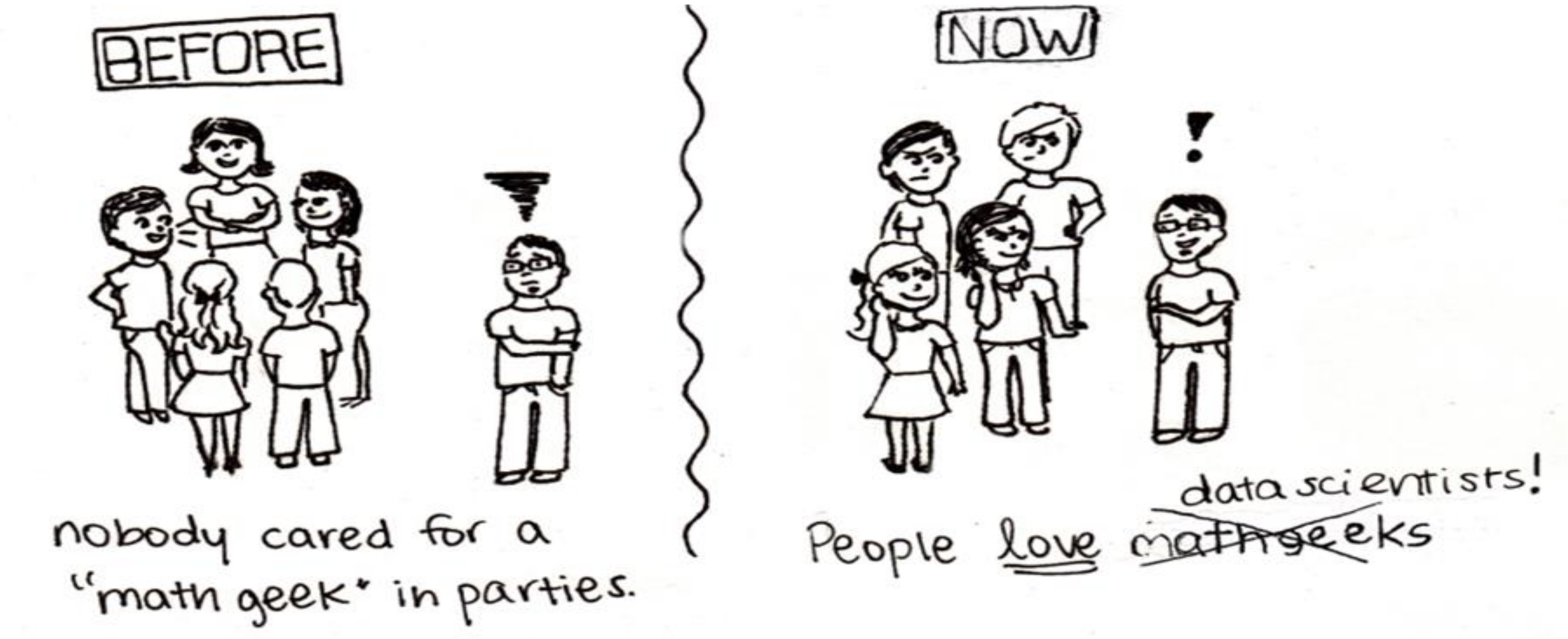


Drew Conway's Venn diagram of data science

The rise of "Data Scientists"

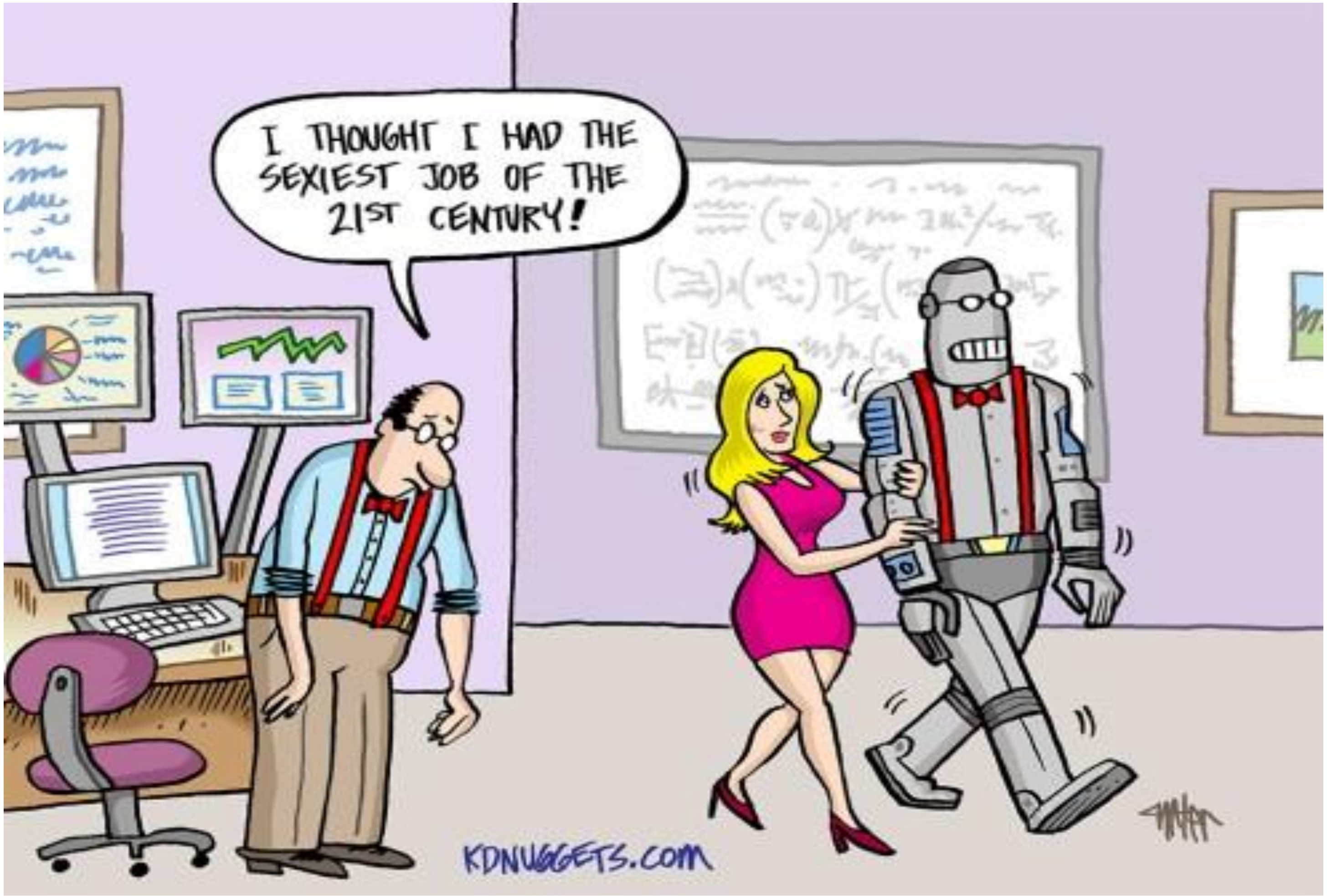
"Data Scientist is The Sexiest Job of the 21st Century"

—T. Davenport & D.J. Patil, Harvard Business Review


BEFORE
NOW
!
nobody cared for a "math geek" in parties.
People love ~~math geeks~~ data scientists!

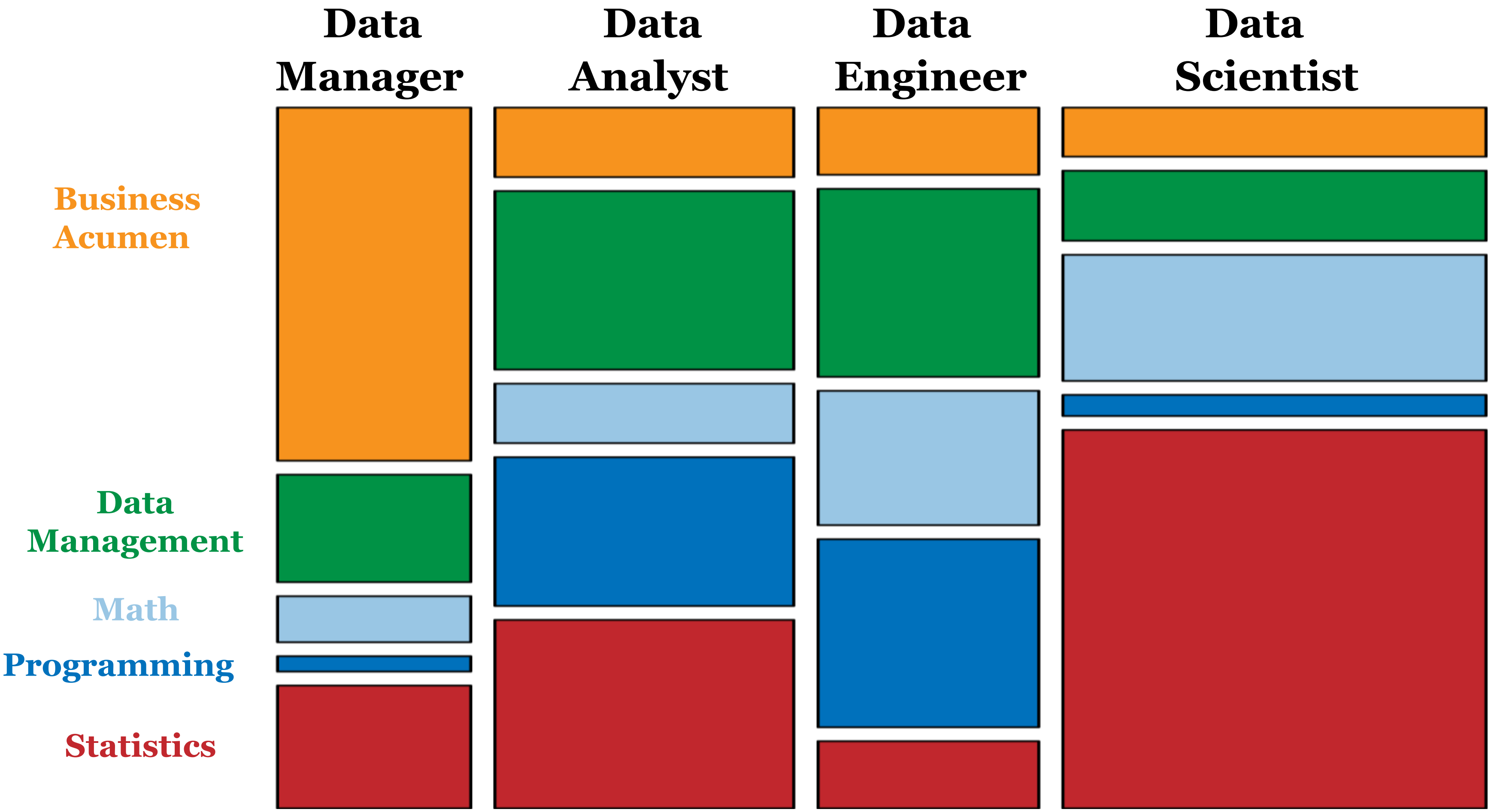
Source: http://www.techjuice.pk/how-to-become-a-data-scientist-for-free/

The "Dilemma" of the Data Scientists



Source: http://www.kdnuggets.com/2016/08/cartoon-data-scientist-sexiest-job-21st-century.html

Your data careers



What is "Big Data"

- Many people have defined "Big Data" with 3Vs, 4Vs, 5Vs…, many more Vs!

- My definition is: "*Too much and complicated data to be processed by a single machine with reasonable time or resources*".

Where does the big data come from?

- **Traditional Data**
    - Any digitized contents and/or archives acquired by traditional ways, e.g. survey data, interview records, and documents.

- **Machine Data**
    - Sensor data, web logs, any log data from monitoring information systems.

- **Network Data**
    - The network of computers (The Internet)
    - The network of people (Social Networks)
    - The network of things (Internet of Things)

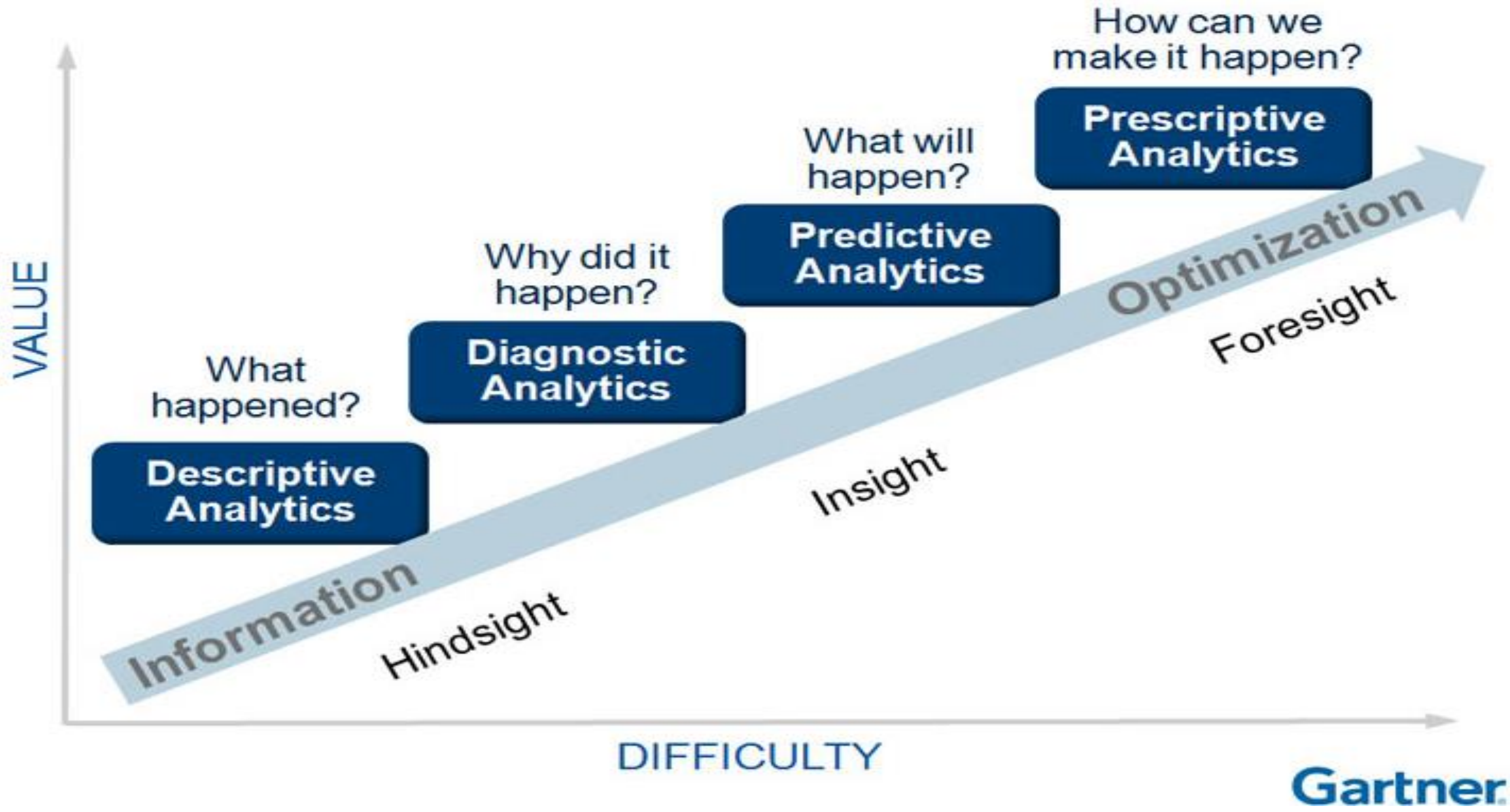
Types of Big Data

- **Structured data**
    - Data with clear schema/metadata/data model that describes & defines how the data elements relate to one another. E.g. relational databases, data cubes/warehouses.

- **Semi-Structured data**
    - Data with only tag/field definitions but without formal structures of data models to define relations. E.g. data used in information exchanges, such as XML & JSON. Emails/pictures/other files with tags/field definitions.

- **Unstructured data**
    - Unorganized data without any pre-defined schema. E.g. body of an e-mail message, pictures, audio, and video.

Wait.. we're talking about "Big Data Analytics"?



Data is NOT always the cure!

- The "Big Data" does eliminate intuition. However, our interpretations of it have great impact on the results. Let's check out this article in New York Time . It says "*Let's put everything in and let the data speak for itself.*" This is a bit horrible quote and don't let it mislead you.

"*...Data is just a quantitative,*

pale echo of the events of our society...".

–*O'Neil, "On Being a Data Skeptic"*

I’ve heard a ton about the “Big Data”.

So what?

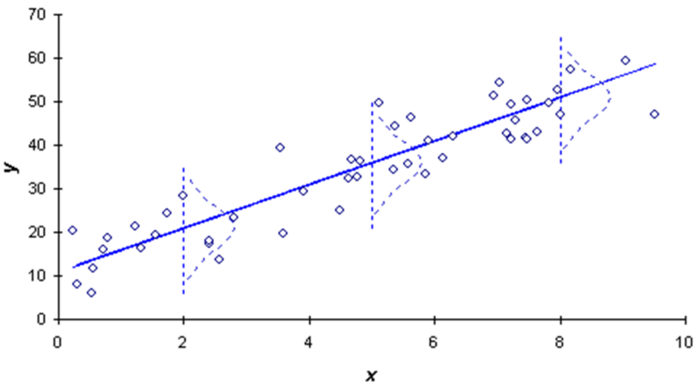
“Big Data is like teenage sex: everyone talks about it, nobody knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it.”

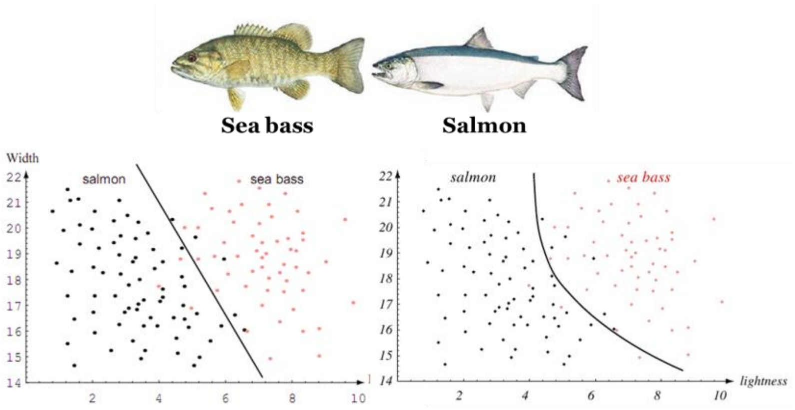
– *Dan Ariely*

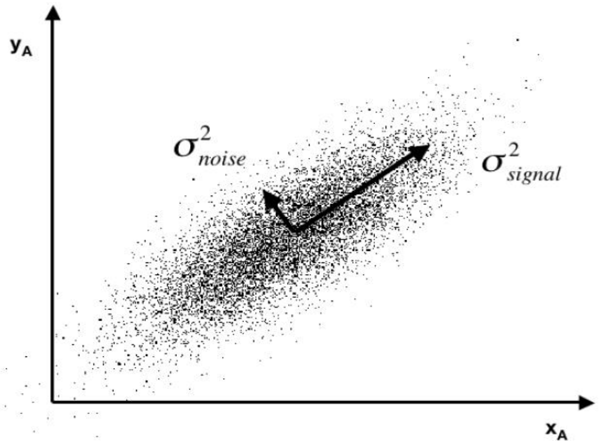
DO NOT "Talk" Big Data

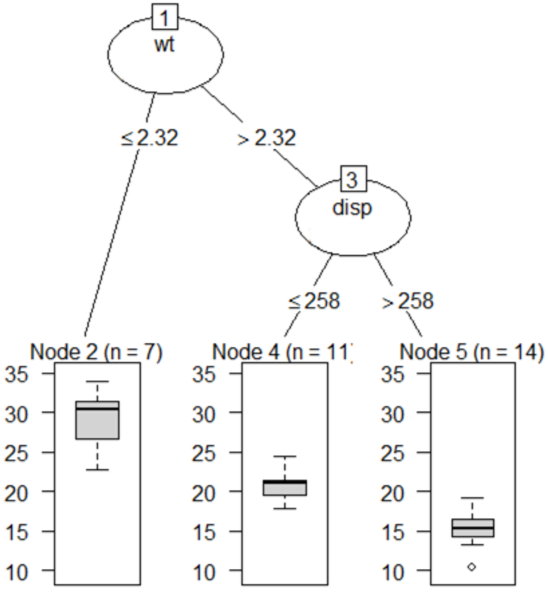
"Do" Big Data

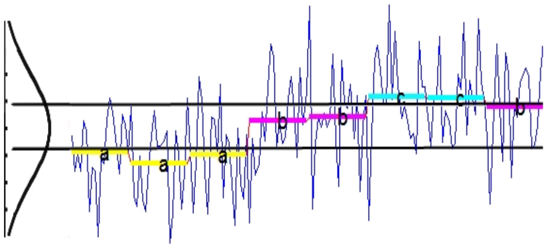
Statistical Learning







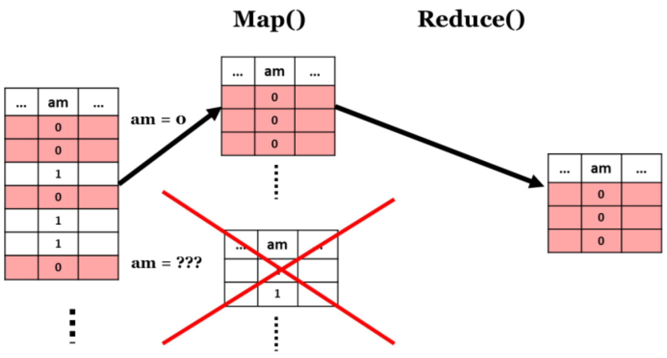




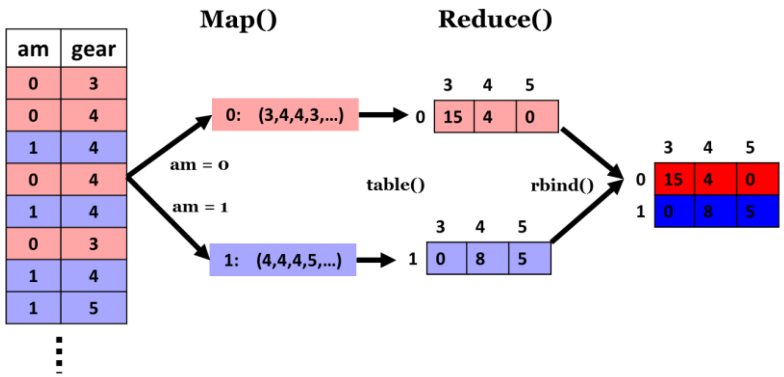
MapReduce Design Patterns

- We will surely do more than just the "word count"!

Filtering



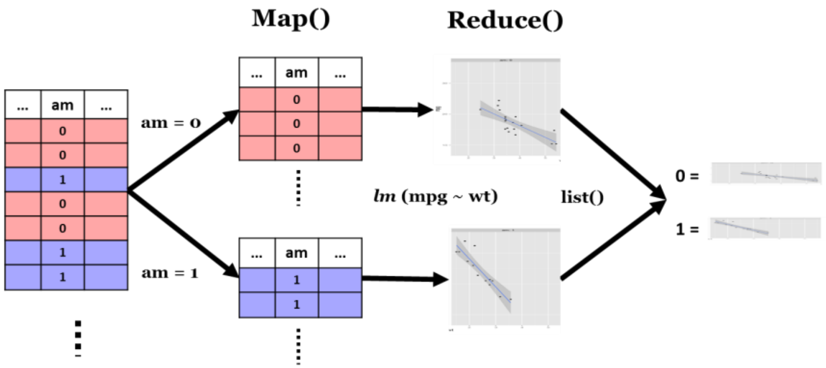
Aggregation (crosstab)



am	gear
0	3
0	4
1	4
0	4
1	4
0	3
1	4
1	5

	3	4	5
0	15	4	0
1	0	8	5

Split-Apply-Combine (e.g. model fitting)



Sorting patterns, Join patterns, ….

And many more!

High-performance R Programming

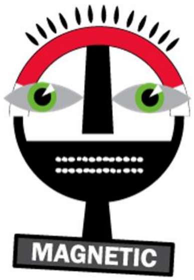
- We still don't know much about R's own limitations and capabilities when coping with Big Data. Why my R code is so slow? How to evaluate my R code?

- We will be discussing vectorized and functional programming, and why they matter in the age of Big Data.

- We will also be discussing how to tweak programs by writing more functional, primitive, and parallel R code, as well as how to use more CPU cores on both a single and a cluster of machines!

In-database Analytics

- Allow me to introduce different worlds of Big Data Analytics—how *in-database computing* solutions deal with massive dataset and how traditional database vendors (e.g Oracle, IBM,…etc) fight against this Big Data evolution.
- We will also be getting "MAD"—*Apache MADlib*, a popular in-database analytics package, along with *Greenplum database* and its procedure language.



Attract all data sources

Digest at a rapid pace

Go beyond rollups and drilldowns of BIs !

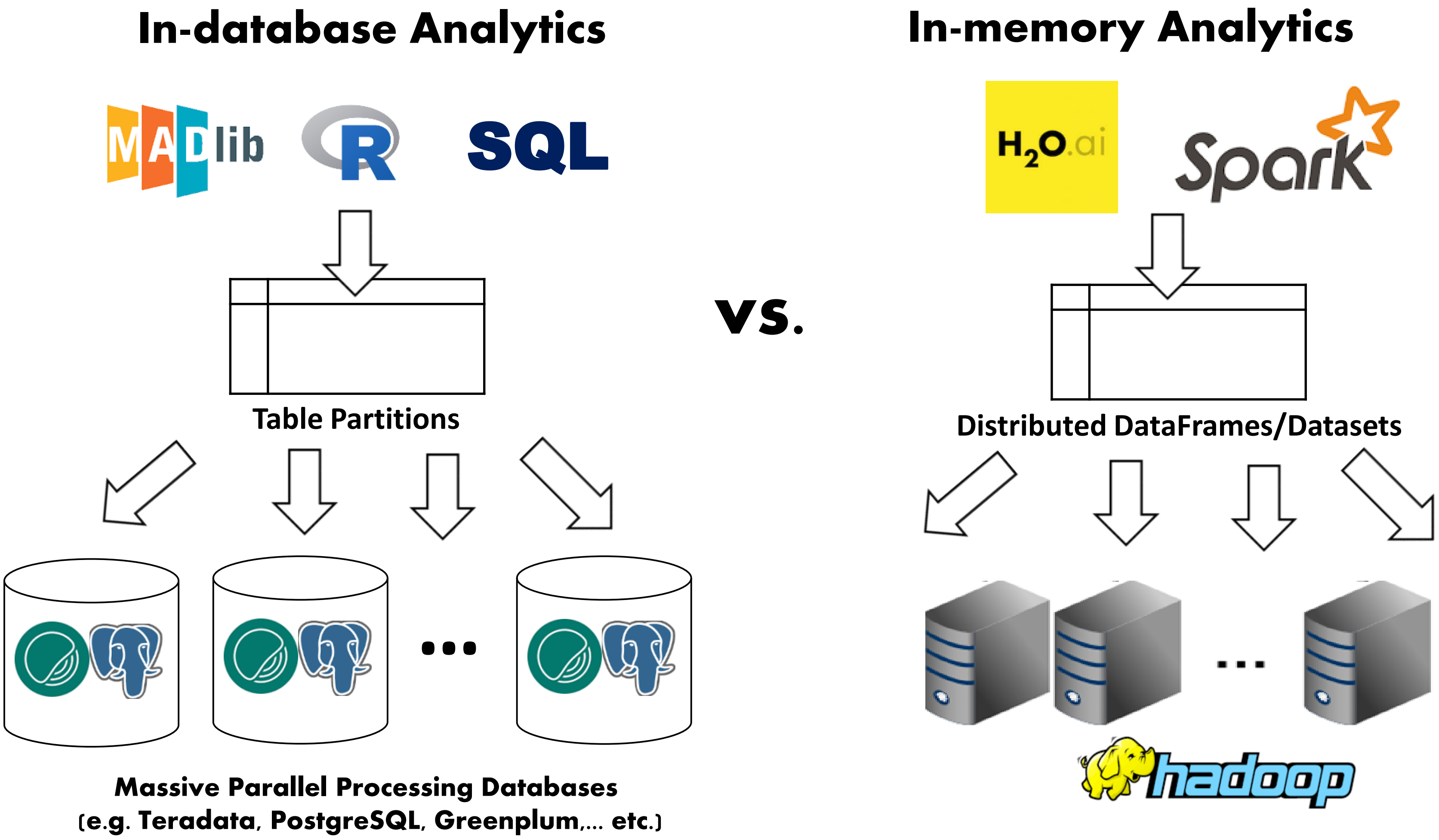
Fast Scalable Data Analytics

- Researchers have been proposing new data processing frameworks (e.g. Apache Hadoop) that incorporate and coordinate a cluster of computers to solve big data problems. However, we are still experiencing high latency when dealing with massive data or computation-intensive analysis tasks.

- Clumsy, centralized-managed, and batch data processing frameworks have hindered them from many data applications, which call for a **unified, fast, and scalable data analytics framework.**

- We need a distributed, peer-to-peer, and cross-platform computing architecture with rich data analysis libraries that allows us to easily set up a running-on-demand computing cluster capable of scaling out computations to multiple nodes.

Two Ecosystems of Scalable Data Analytics



In-database Analytics

Table Partitions

Massive Parallel Processing Databases
(e.g. Teradata, PostgreSQL, Greenplum,... etc.)

vs.

In-memory Analytics

Distributed DataFrames/Datasets

Trends in Big Data Analytics

✓The flood of "data lake".

✓ The rise of out-of-core learning algorithms.

✓ The dawn of *fast scalable data applications*.

✓ The use of *in-memory*, *in-database*, and *GPU* computing.

✓ The pursuit of interpretable analytics and explainable AI.

Your homework this week

- ❑ Make a choice. Quit or stay. We will be getting an account for the access to CM Big Data Analytics Platform next week.

- ❑ Review R programming. Especially those of you who are not familiar with any scientific computing languages.

- ❑ Get the textbooks & papers and start reading!

See you next week!