Data Science in Action

Peerapon Vateekul, Ph.D.

Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University

Chula Data Science

Outlines

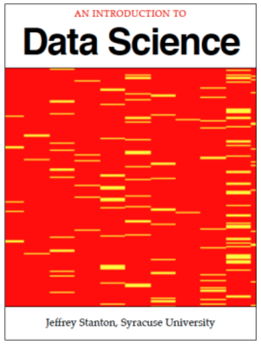
- Data Science & Data Scientist
- Data Mining
- Analytics with R
- A Framework for Big Data Analytics

+

Data Science & Data Scientist

What is Data Science?

Hortonworks


AN INTRODUCTION TO

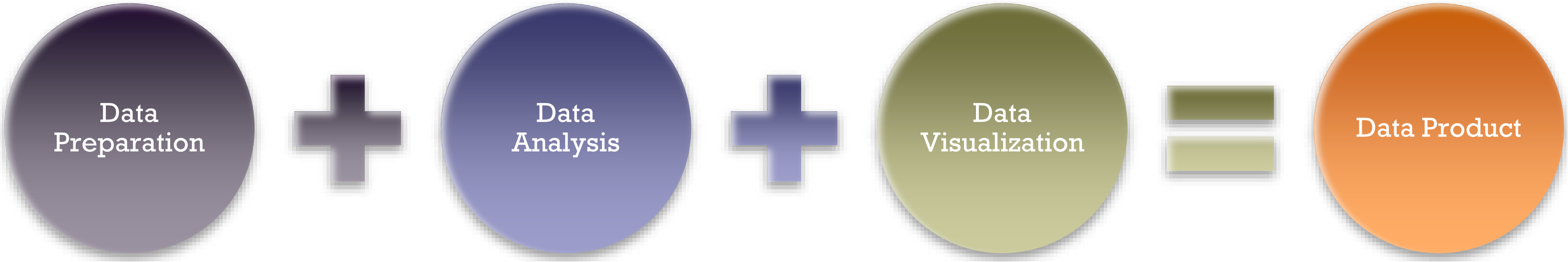
Data Science

Jeffrey Stanton, Syracuse University

What is Data Science?

The future belongs to the companies and people that turn data into products

- Data
  - Facts and statistics collected together for reference or analysis
- Science
  - A systematic study through observation and experiment
- Data Science
  - The scientific exploration of data to extract meaning or insight
  - , and the construction of software to utilize such insight in a business context.



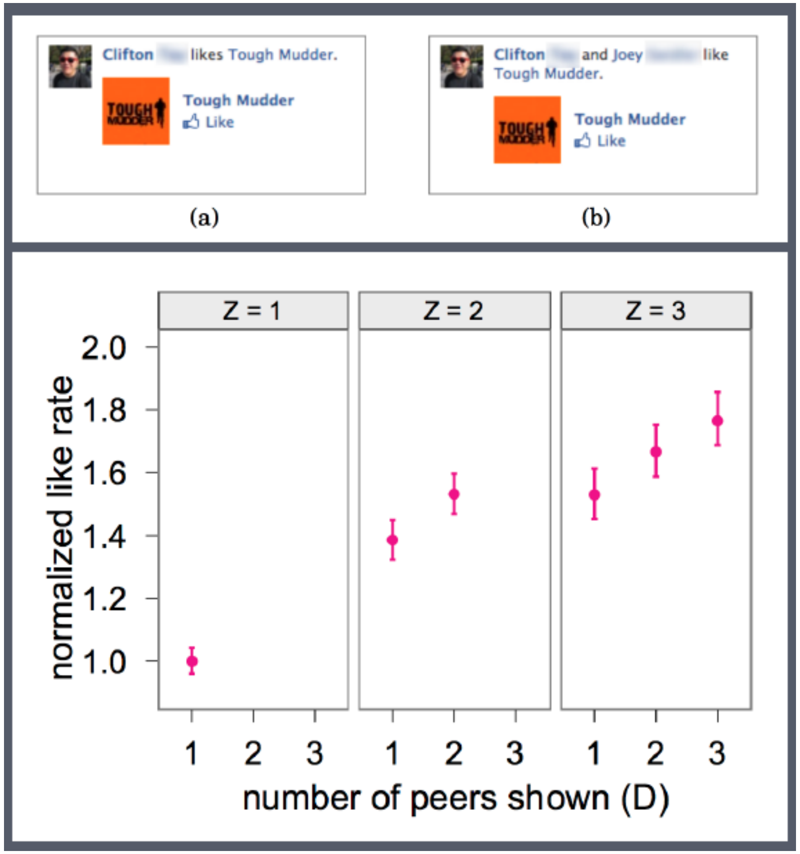
What is Data Science? (cont.)

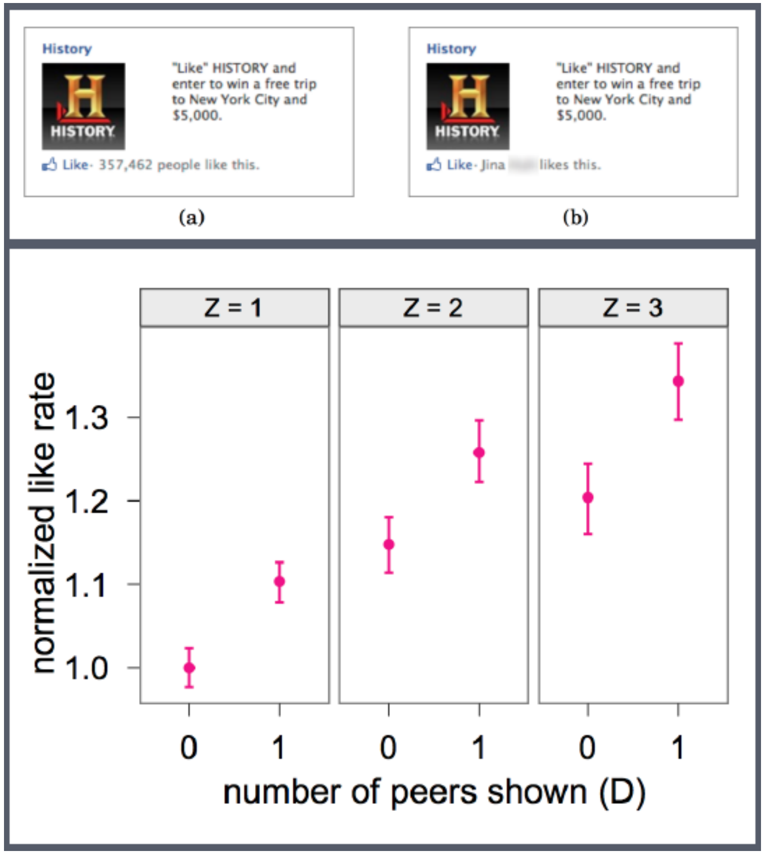
- Transform data into valuable insights
- Transform data into data products
- Transform data into interesting stories

Ta Virot Chiraphadhanakul
Data Scientist, Facebook

What is Data Science? (cont.)
Transform data into valuable insights

Ta Virot Chiraphadhanakul
Data Scientist, Facebook

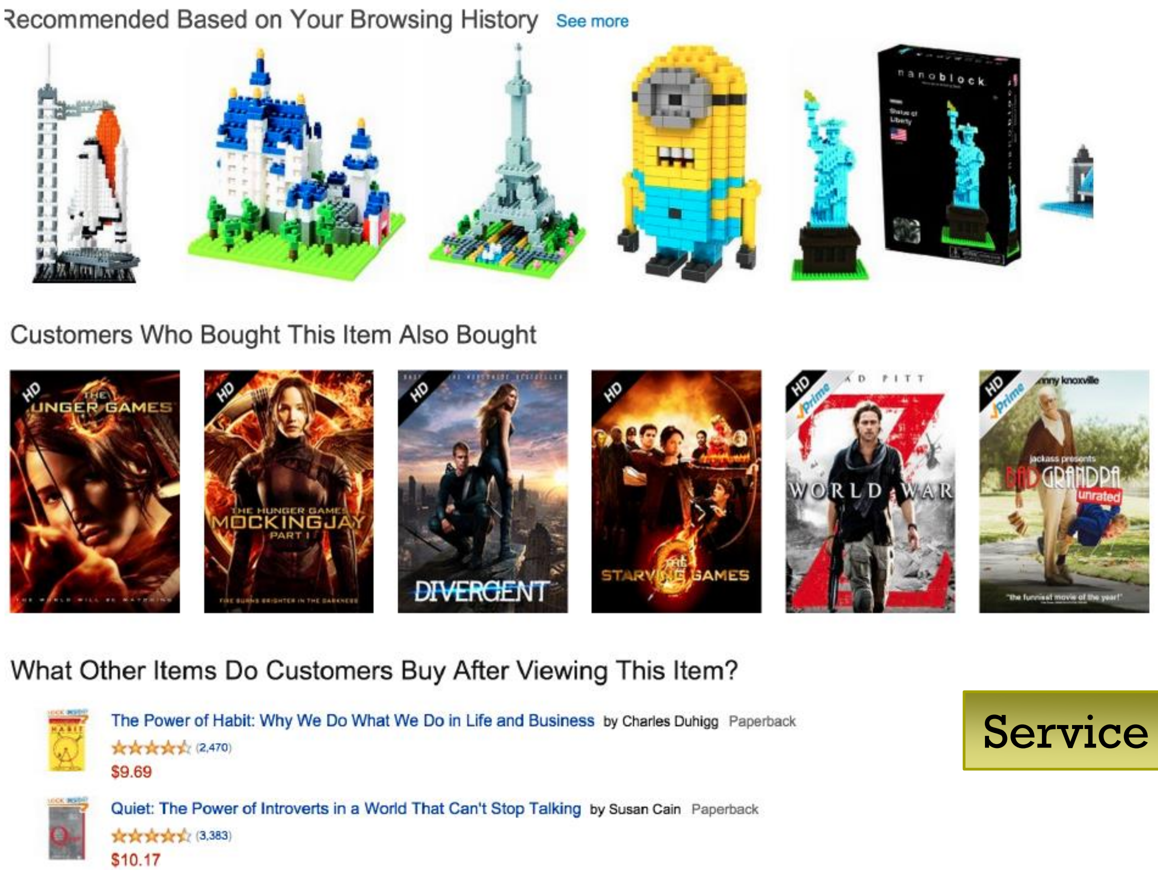




Social Influence in Social Advertising: Evidence from Field Experiments (Bakshy et al. 2012)

What is Data Science? (cont.) Transform data into data products

Ta Virot Chiraphadhanakul
Data Scientist, Facebook



Recommended Based on Your Browsing History See more

Customers Who Bought This Item Also Bought

What Other Items Do Customers Buy After Viewing This Item?

The Power of Habit: Why We Do What We Do in Life and Business by Charles Duhigg Paperback
★★★★☆ (2,470)
$9.69

Quiet: The Power of Introverts in a World That Can't Stop Talking by Susan Cain Paperback
★★★★☆ (3,383)
$10.17

amazon

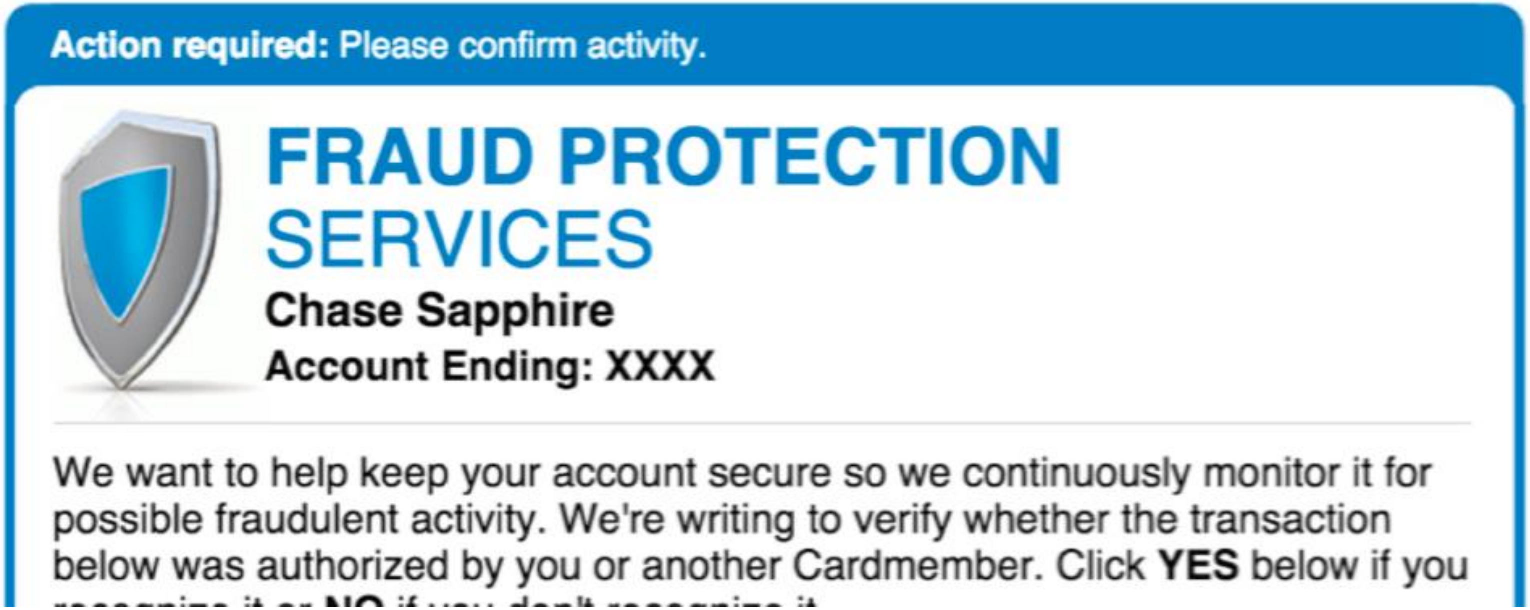
Service Recommendation

+ What is Data Science? (cont.) Transform data into data products

Ta Virot Chiraphadhanakul
Data Scientist, Facebook

CHASE

Fraud Detection



Action required: Please confirm activity.

FRAUD PROTECTION SERVICES

Chase Sapphire

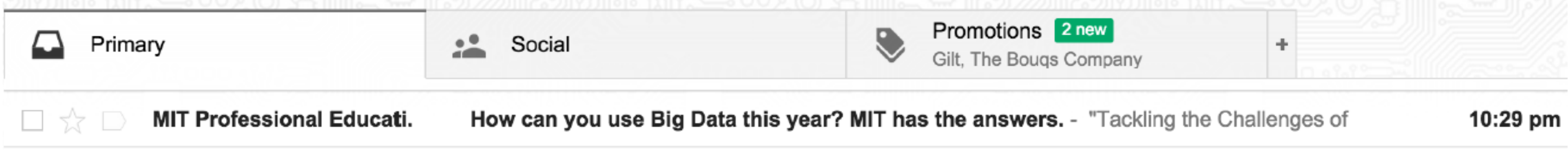
Account Ending: XXXX

We want to help keep your account secure so we continuously monitor it for possible fraudulent activity. We're writing to verify whether the transaction below was authorized by you or another Cardmember. Click **YES** below if you recognize it or **NO** if you don't recognize it.

+ What is Data Science? (cont.) Transform data into data products

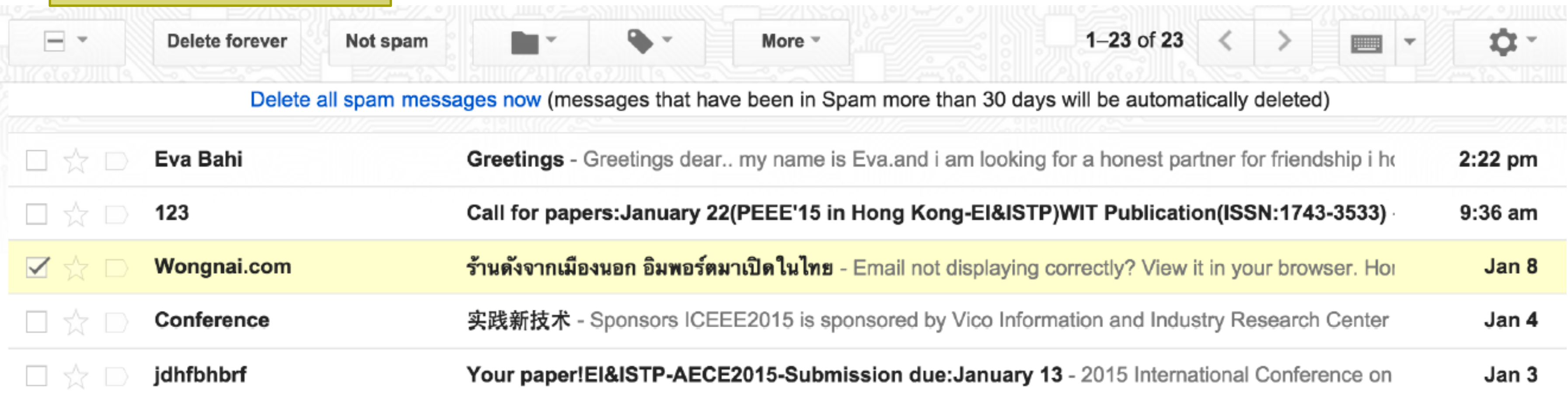
Ta Virot Chiraphadhanakul
Data Scientist, Facebook

Email Classification



☐ ☆	MIT Professional Educati.	**How can you use Big Data this year? MIT has the answers.** - "Tackling the Challenges of	10:29 pm

Spam Detection



Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

☐ ☆	Eva Bahi	**Greetings** - Greetings dear.. my name is Eva.and i am looking for a honest partner for friendship i hc	2:22 pm
☐ ☆	123	**Call for papers:January 22(PEEE'15 in Hong Kong-EI&ISTP)WIT Publication(ISSN:1743-3533)** -	9:36 am
☑ ☆	Wongnai.com	**ร้านดังจากเมืองนอก อิมพอร์ตมาเปิดในไทย** - Email not displaying correctly? View it in your browser. Ho	Jan 8
☐ ☆	Conference	**实践新技术** - Sponsors ICEEE2015 is sponsored by Vico Information and Industry Research Center	Jan 4
☐ ☆	jdhfbhbrf	**Your paper!EI&ISTP-AECE2015-Submission due:January 13** - 2015 International Conference on	Jan 3

+ What is Data Science? (cont.)

Ta Virot Chiraphadhanakul
Data Scientist, Facebook

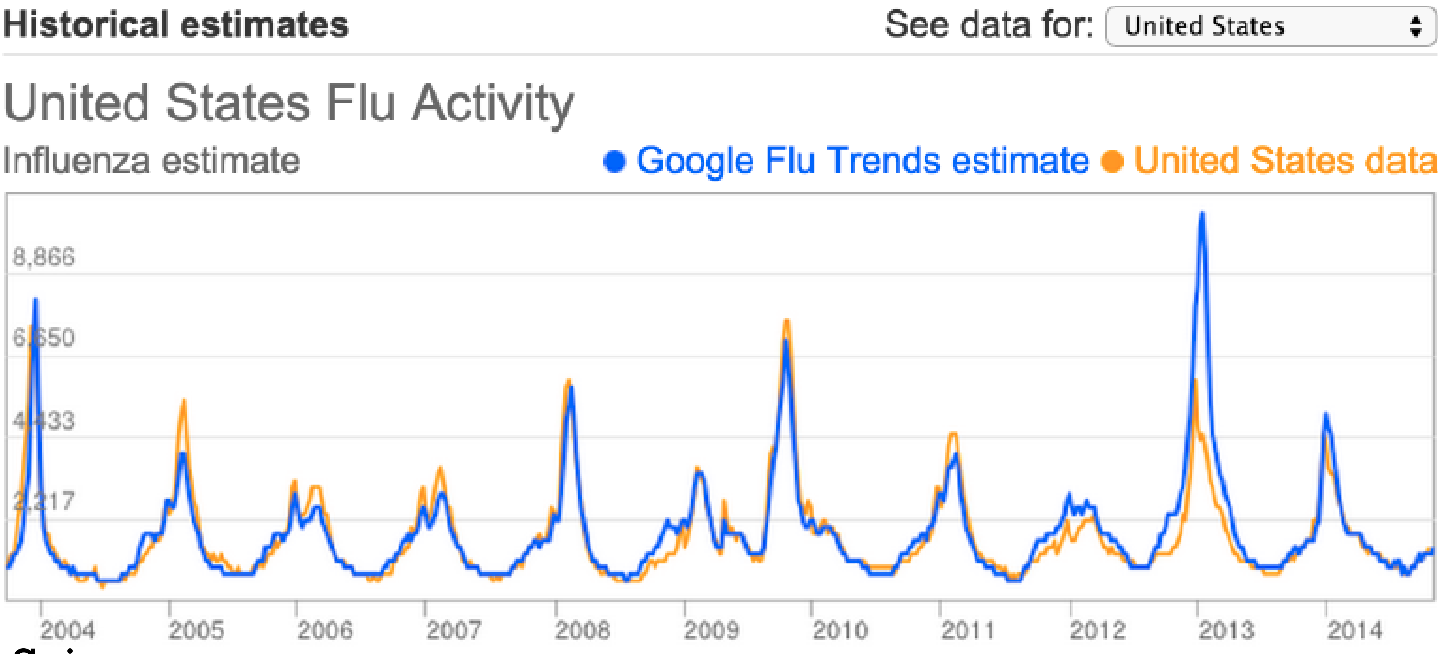
Transform data into interesting stories

Google

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

[1]Google Inc. [2]Centers for Disease Control and Prevention



Historical estimates

See data for: United States

United States Flu Activity

Influenza estimate

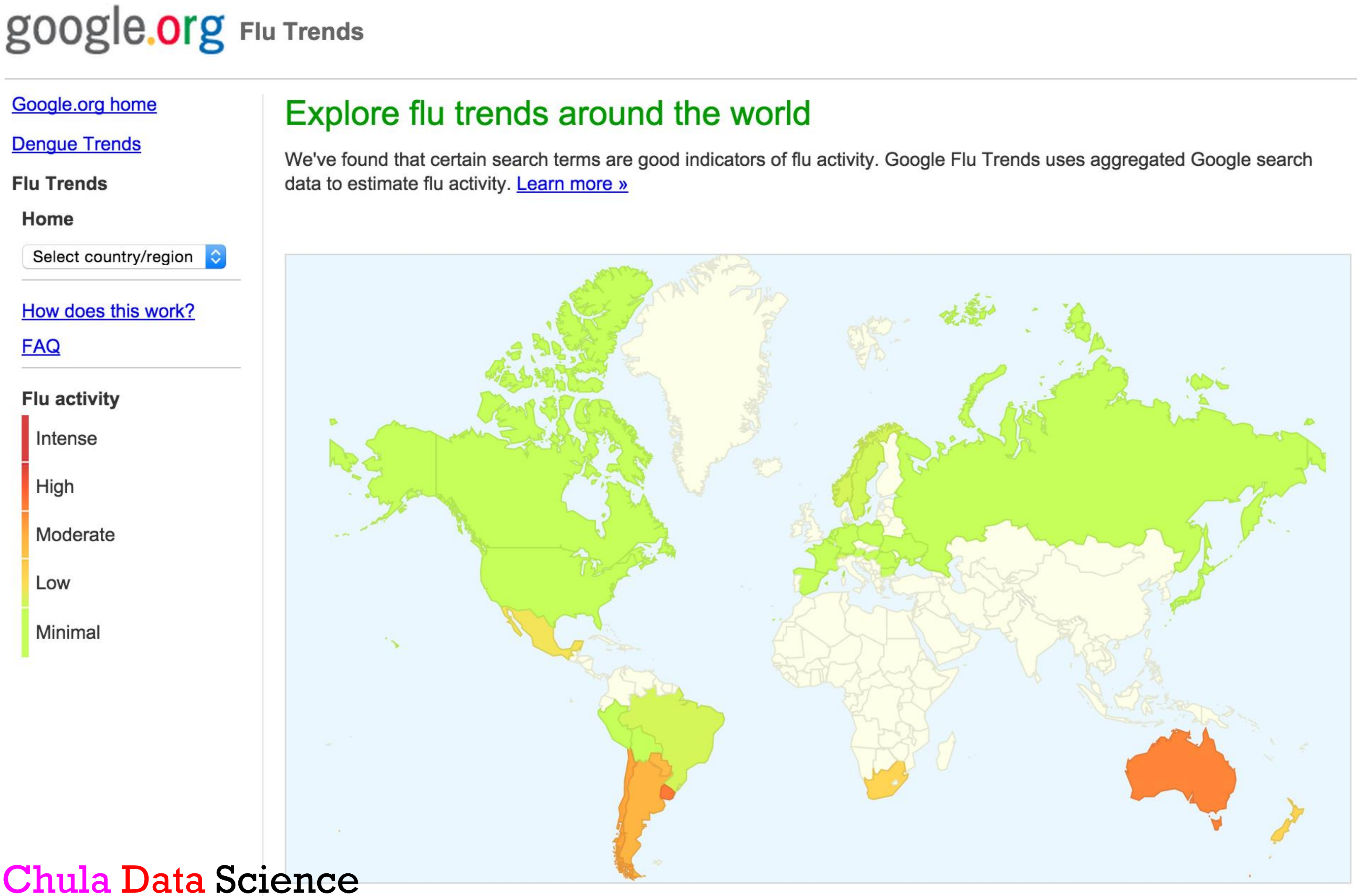
● Google Flu Trends estimate ● United States data

United States: Influenza-like illness (ILI) data provided publicly by the U.S. Centers for Disease Control.

What is Data Science? (cont.)

Ta Virot Chiraphadhanakul
Data Scientist, Facebook

Transform data into interesting stories



google.org Flu Trends

Google.org home

Dengue Trends

Flu Trends

Home

Select country/region

How does this work?

FAQ

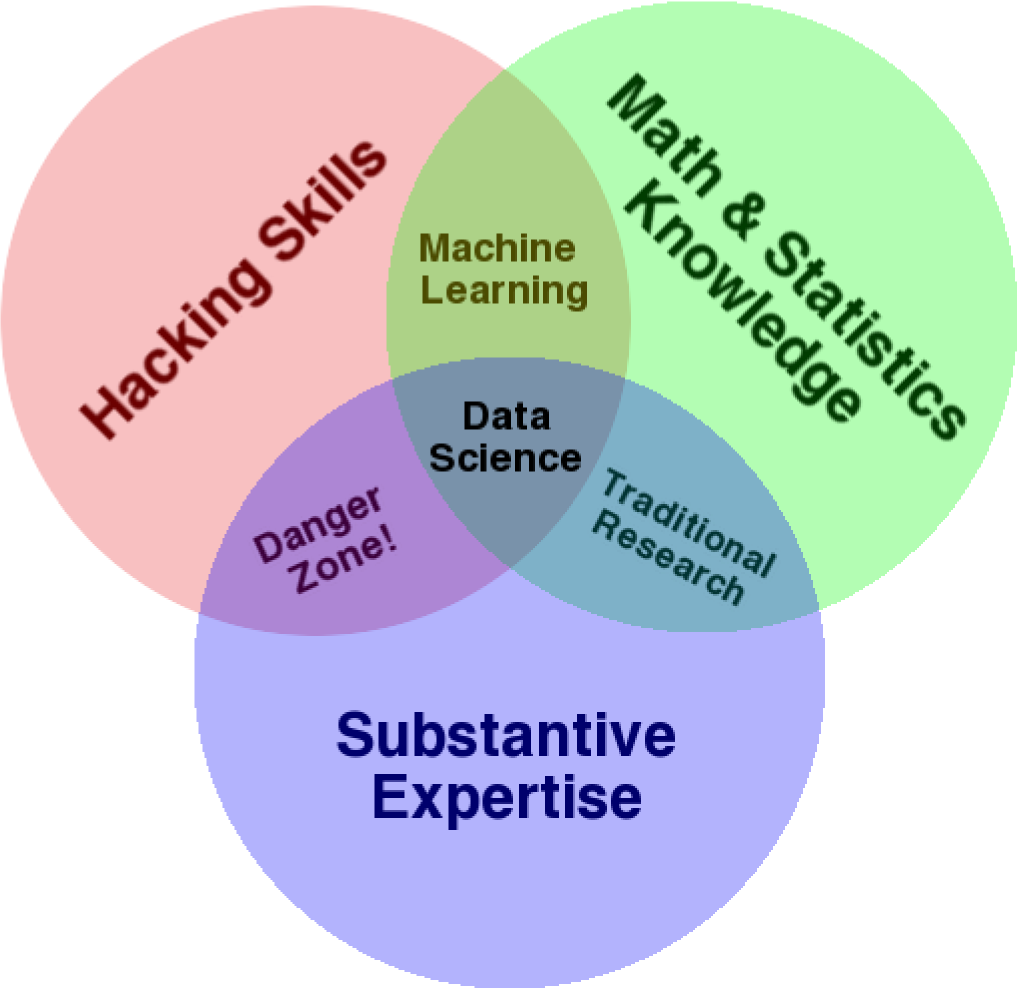
Flu activity

- Intense
- High
- Moderate
- Low
- Minimal

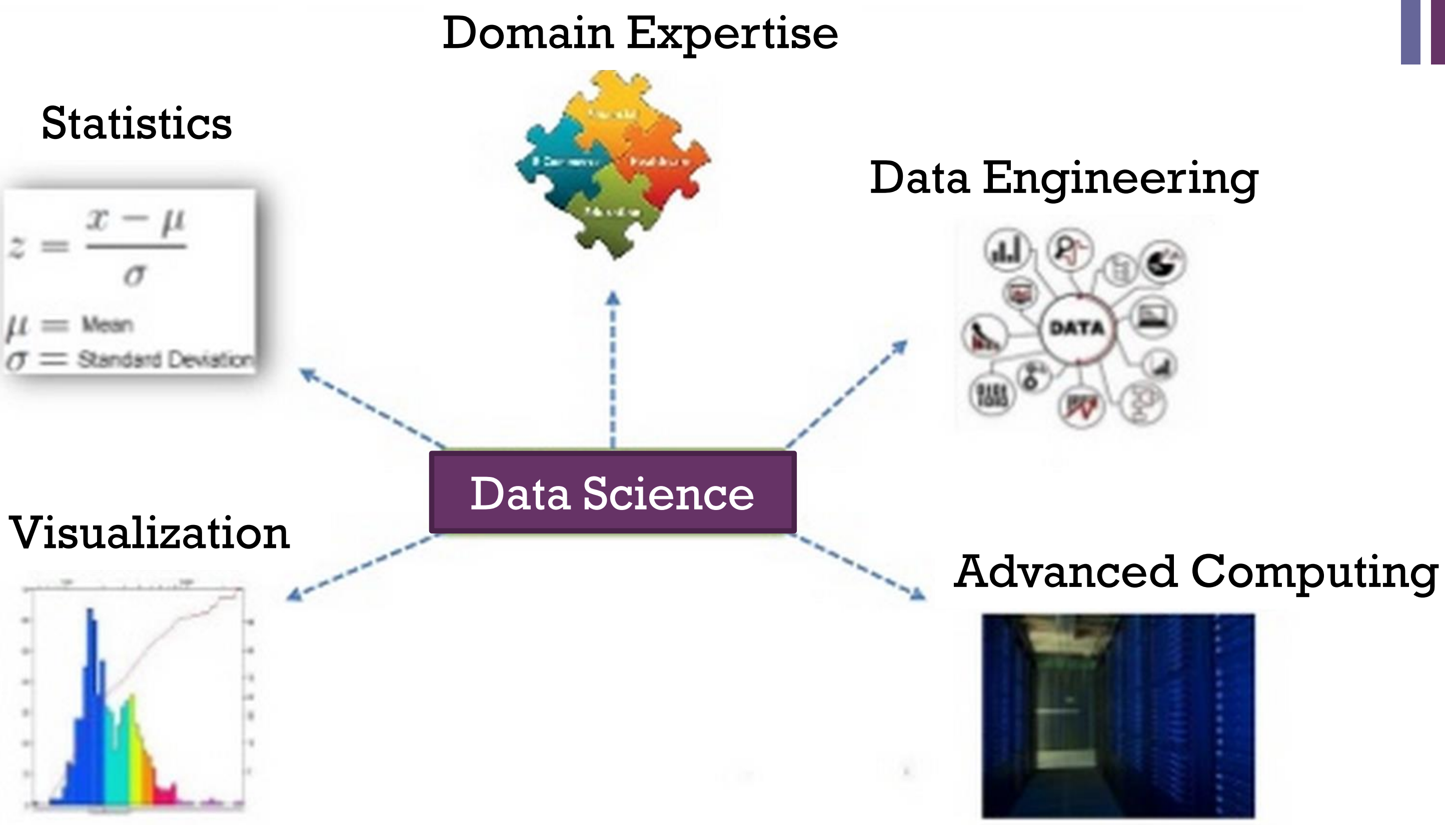
Explore flu trends around the world

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

Data Science: Famous Definition



Data Science: Components



Domain Expertise

Statistics

Data Engineering

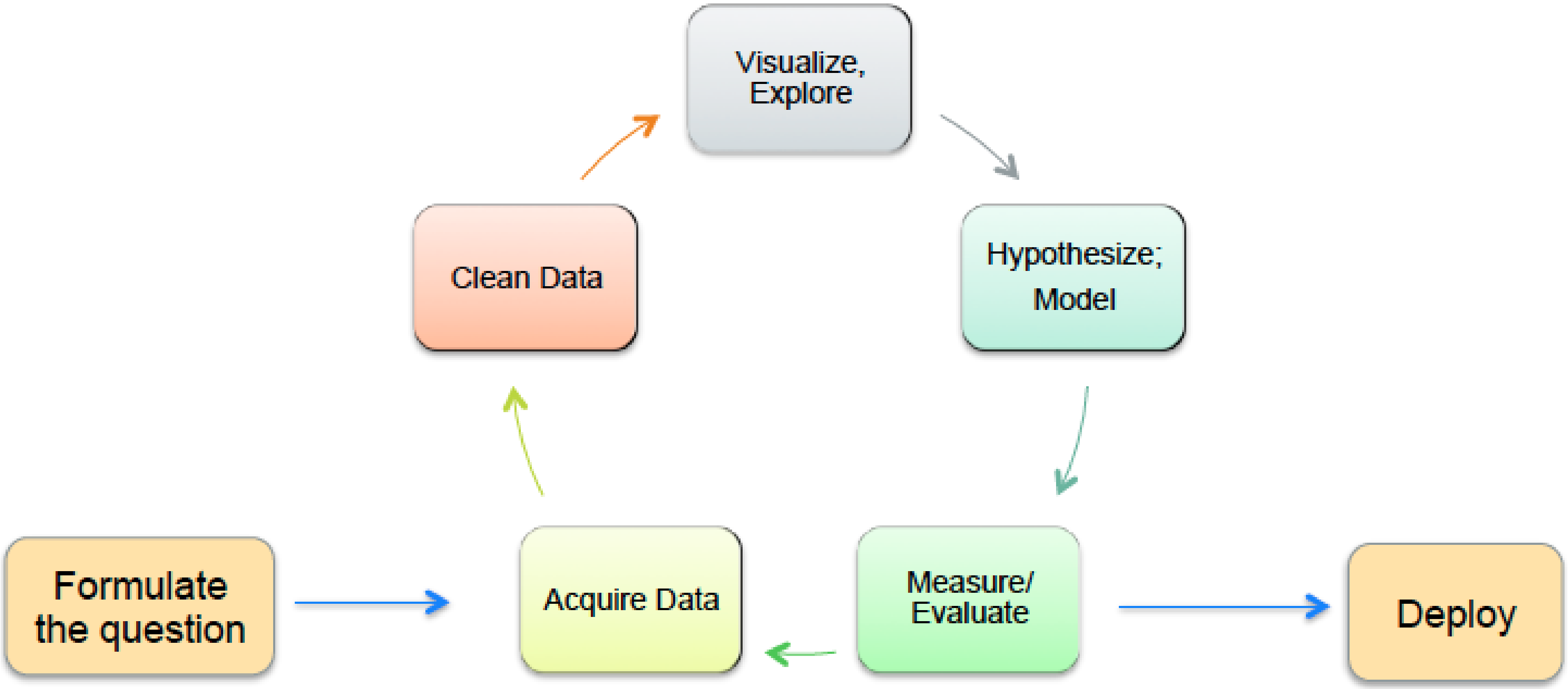
Data Science

Visualization

Advanced Computing

Data Science Process: Iterative Activity

Hortonworks



Formulate the question

Acquire Data

Clean Data

Visualize, Explore

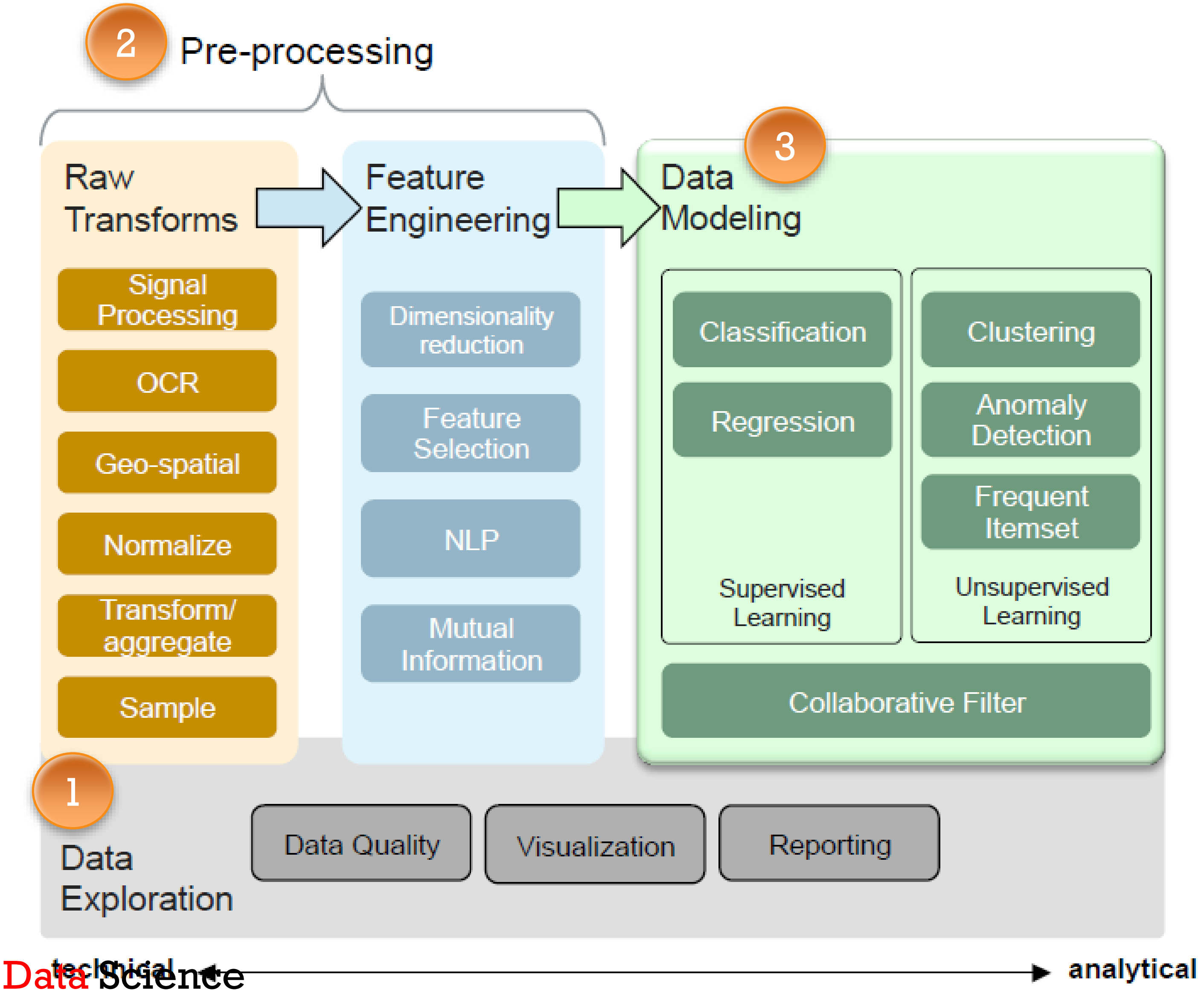
Hypothesize; Model

Measure/ Evaluate

Deploy

Data Science Tasks

Hortonworks



Data Science with Big Data

Hortonworks

- Very large raw data sets are now available:
  - Log files
  - Sensor data
  - Sentiment information
- With more raw data, we can build better models with improved predictive performance.
- To handle the larger datasets we need a scalable processing platform like Hadoop and YARN

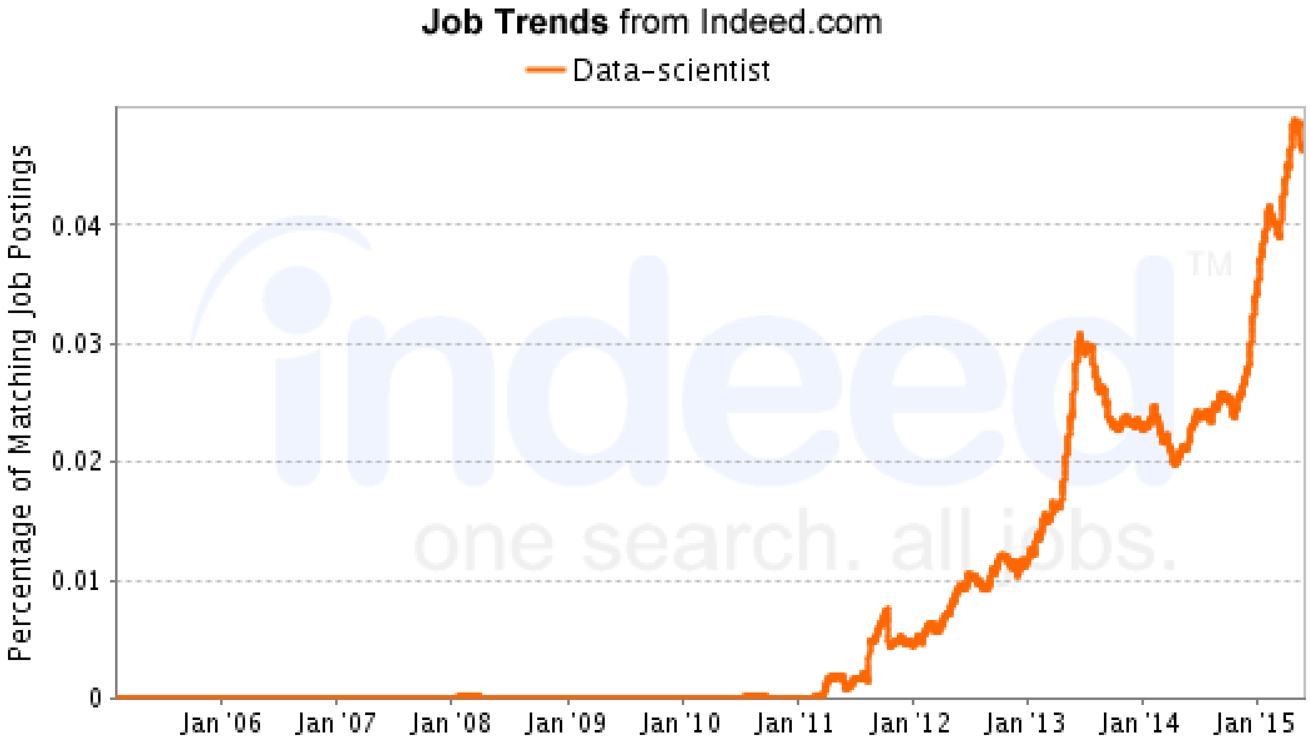
Who builds these systems?

Ta Virot Chiraphadhanakul
Data Scientist, Facebook

Harvard Business Review

Data Scientist: The Sexiest Job of 21st Century

By Thomas H. Davenport and D.J. Patil
From the October 2012 issue



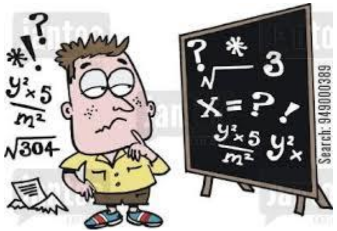
Data science jobs pay an average of $118,000/year

It is estimated that by 2018, US could have a shortage of 140,000+ people with advanced analytical skills!

Definition

Computer Scientist

- Data collection systems
- Machine learning algorithms
- Interface design
- Design/manage/query database
- Data aggregation
- Data mining



Mathematician

- Statistical models
- Evaluation metrics
- Predictive analytics
- Data visualization

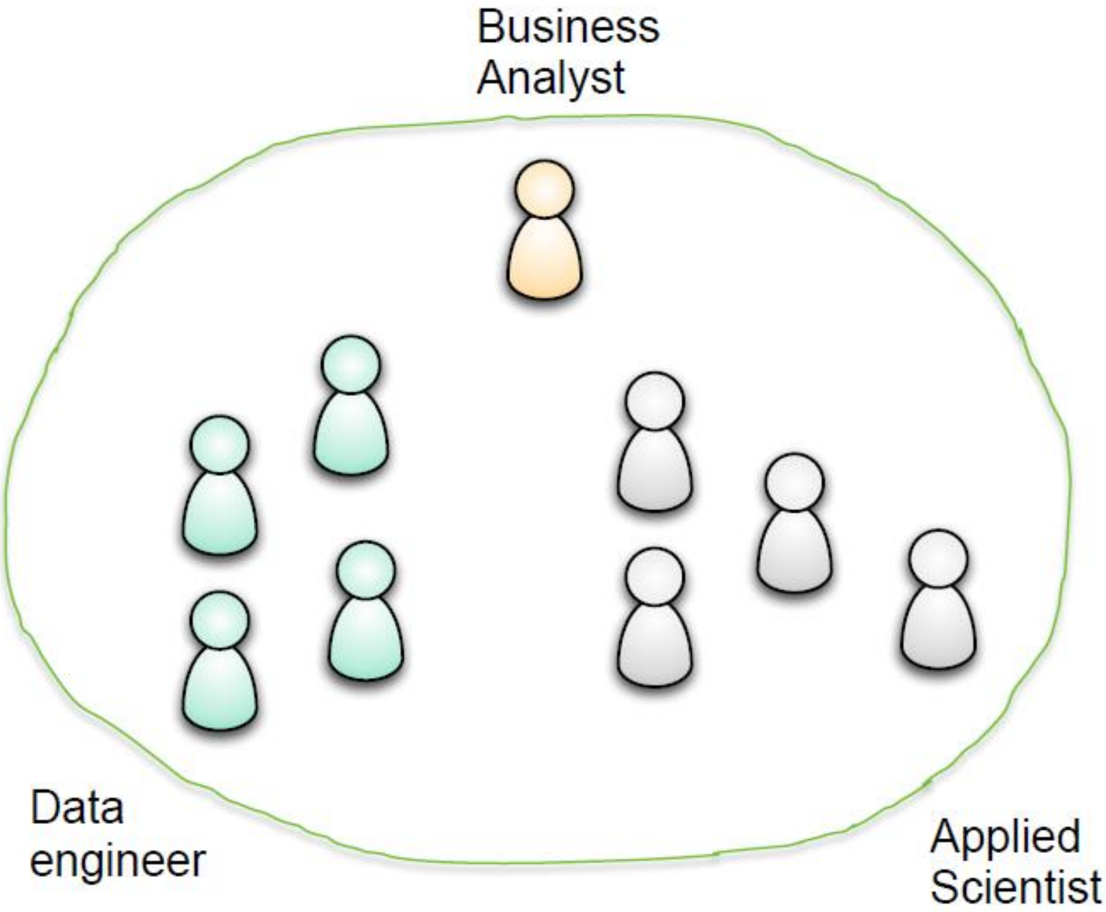
Business Person

- Domain expertise
- Knowing what questions to ask
- Interpreting results for business decisions
- Presenting outcomes

Needed Skills

- **Applied Science**
  - Statistics, applied math
  - Machine Learning, Data Mining
  - Tools: Python, R, SAS, SPSS
- **Business Analysis**
  - Data Analysis, BI
  - Business/domain expertise
  - Tools: SQL, Excel, EDW
- **Data engineering**
  - Database technologies
  - Computer science
  - Tools: Java, Scala, Python, C++
- Big data engineering
  - Big data technologies
  - Statistics and machine learning over large datasets
  - Tools: Hadoop, PIG, HIVE, Cascading, SOLR, etc.

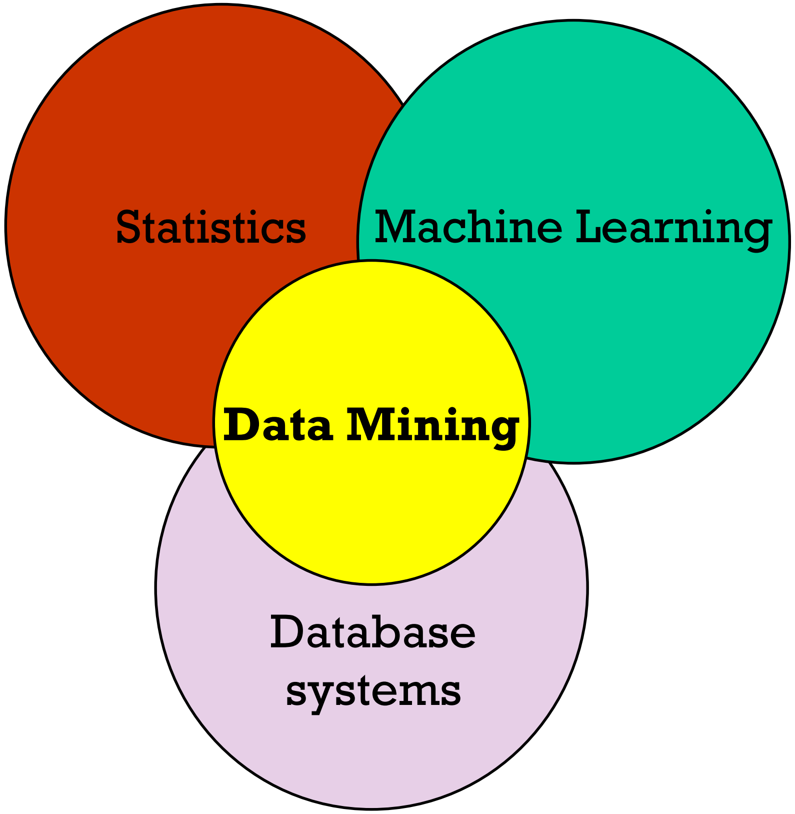
The Data Science Team



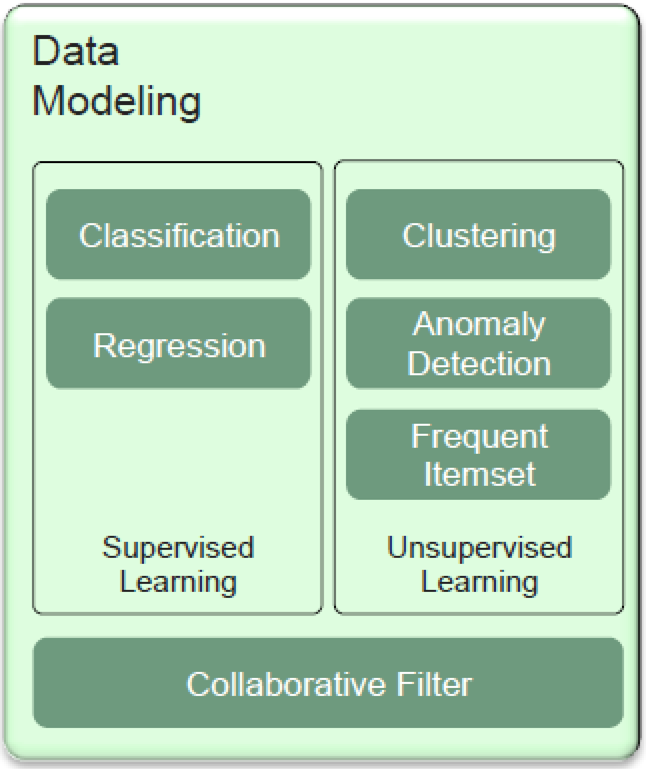
Data Mining

What is Data Mining (DM)?

- An automatic process of
- discovering useful information
- in large data repositories
- with sophisticated algorithm



Data Mining Tasks



- Predictive Task *(Supervised Learning)*
  - Classification
  - Regression
- Descriptive Task *(Unsupervised Learning)*
  - Clustering
  - Association Rules Mining
  - Sequence Analysis
- Other:
  - Collaborative filtering: (recommendations engine) uses techniques from both supervised and unsupervised world.

Supervised Learning: learning from target

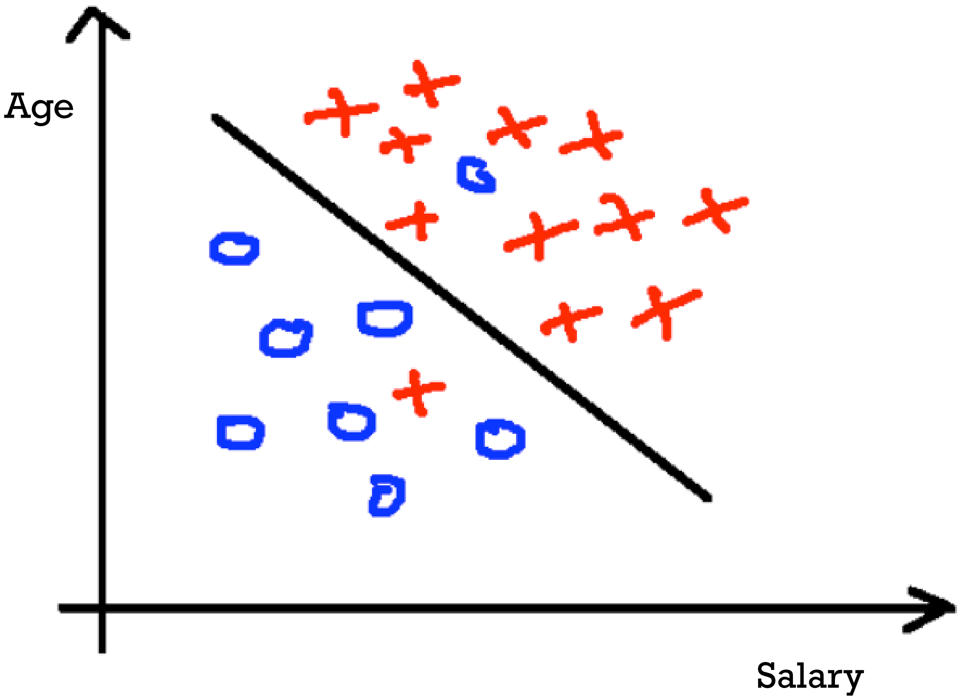
Training dataset:

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,1,1,0,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0	1
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0	0
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0	1
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0	0
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0	1
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,1	0

Test dataset:

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0 ?

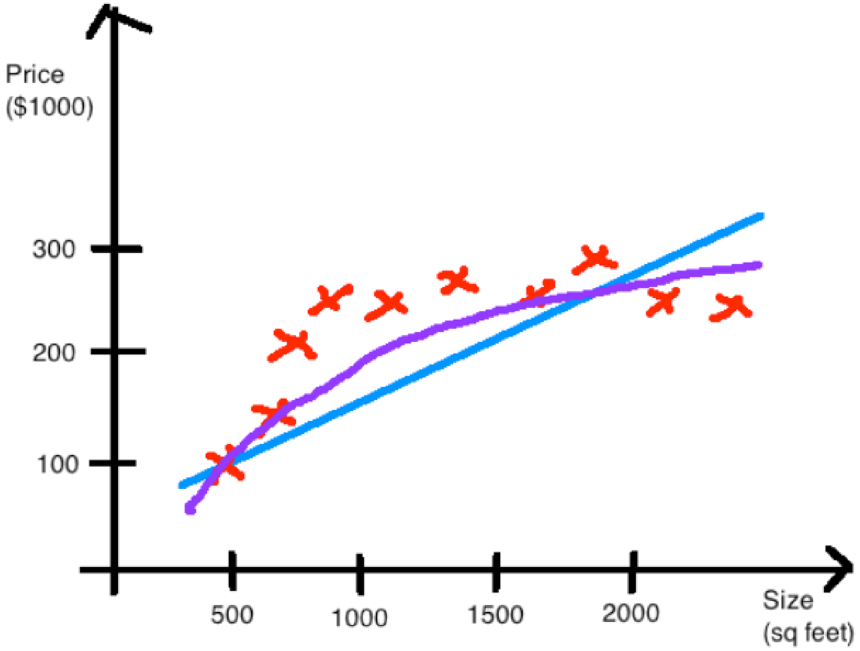
+ Classification: predicting a category



Predict targeted customers who tend to buy our product (yes/no)

- **Some techniques:**
  - Naïve Bayes
  - Decision Tree
  - Logistic Regression
  - Support Vector Machines
  - Neural Network
  - Ensembles

Regression: predict a continuous value



Predict a sale price of each house

- **Some techniques:**
  - Linear Regression / GLM
  - Decision Trees
  - Support vector regression
  - Neural Network
  - Ensembles

Predictive Modeling Applications

- Database marketing
- Financial risk management
- Fraud detection
- Pattern detection

+ Unsupervised Learning: detect natural patterns

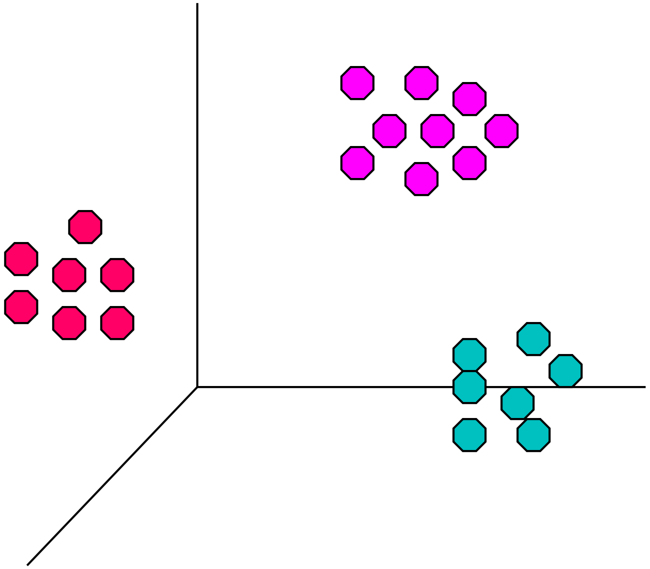
Training dataset:

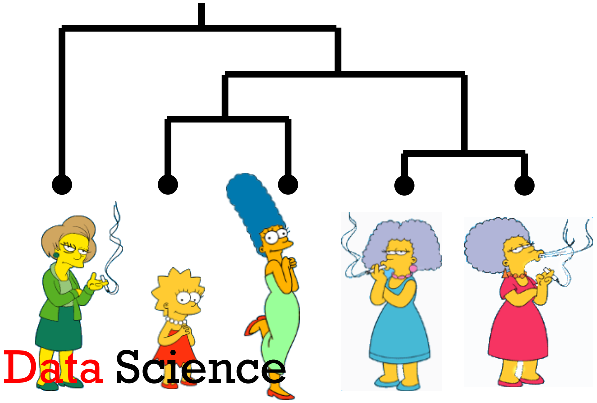
57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0	~~0~~
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0	~~1~~
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0	~~0~~
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	~~1~~
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0	~~1~~
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0	~~0~~
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0	~~1~~
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0	~~0~~
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	~~0~~
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0	~~1~~
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1	~~0~~

Test dataset:

~~71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0~~ ?

Clustering: detect similar instance groupings





- **Some techniques:**
  - k-means
  - Spectral clustering
  - DB-scan
  - Hierarchical clustering

Example: Customer Segmentation

Association Rule Discovery

Store layout design/promotion

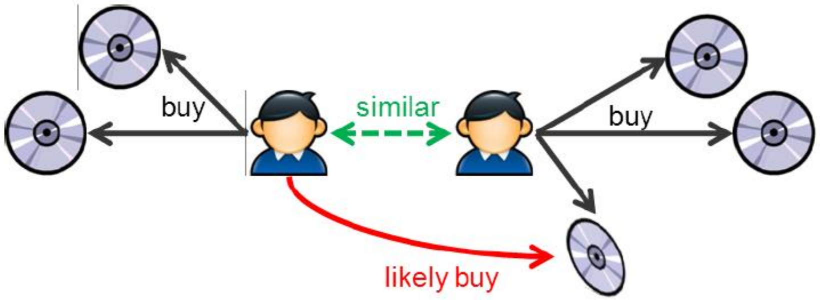
TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

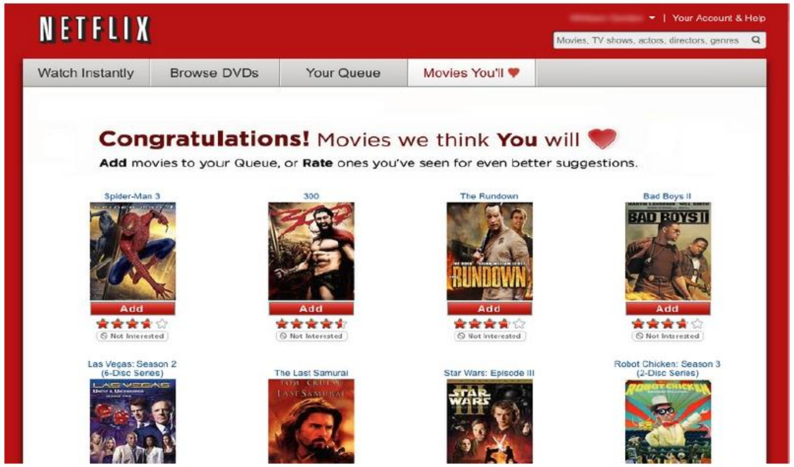
Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Product recommendation: predicting "preference"




NETFLIX

Your Account & Help

Movies, TV shows, actors, directors, genres

Watch Instantly | Browse DVDs | Your Queue | Movies You'll ♥

Congratulations! Movies we think **You** will ♥

Add movies to your Queue, or **Rate** ones you've seen for even better suggestions.

Spider-Man 3 — Add — Not Interested

300 — Add — Not Interested

The Rundown — Add — Not Interested

Bad Boys II — Add — Not Interested

Las Vegas: Season 2 (6-Disc Series)

The Last Samurai

Star Wars: Episode III

Robot Chicken: Season 3 (2-Disc Series)

	Harry potter	X-Men	Hobbit	Argo	Pirates
101	5	2	4	?	?
102	?	?	5	2	?
103	1	2	?	?	3
104					
105					
…					

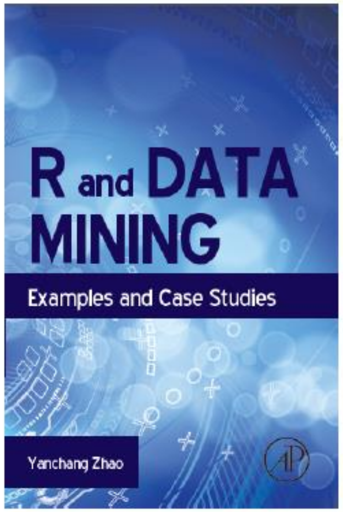
	Harry potte	X-Men	Hobbit	Argo	Pirates
101	5	2	4	1	3
102	4	1	5	2	3
103	1	2	4	1	3
104					
105					
…					

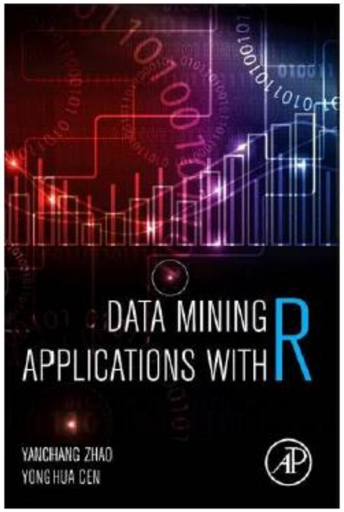
+

Analytics with R

What is R?

- R is a free software environment for statistical computing and graphics.
- R can be easily extended with 5,800+ packages available on CRAN (as of 13 Sept 2014).
- Many other packages provided on Bioconductor, R-Forge, GitHub, etc.
- R manuals on CRAN

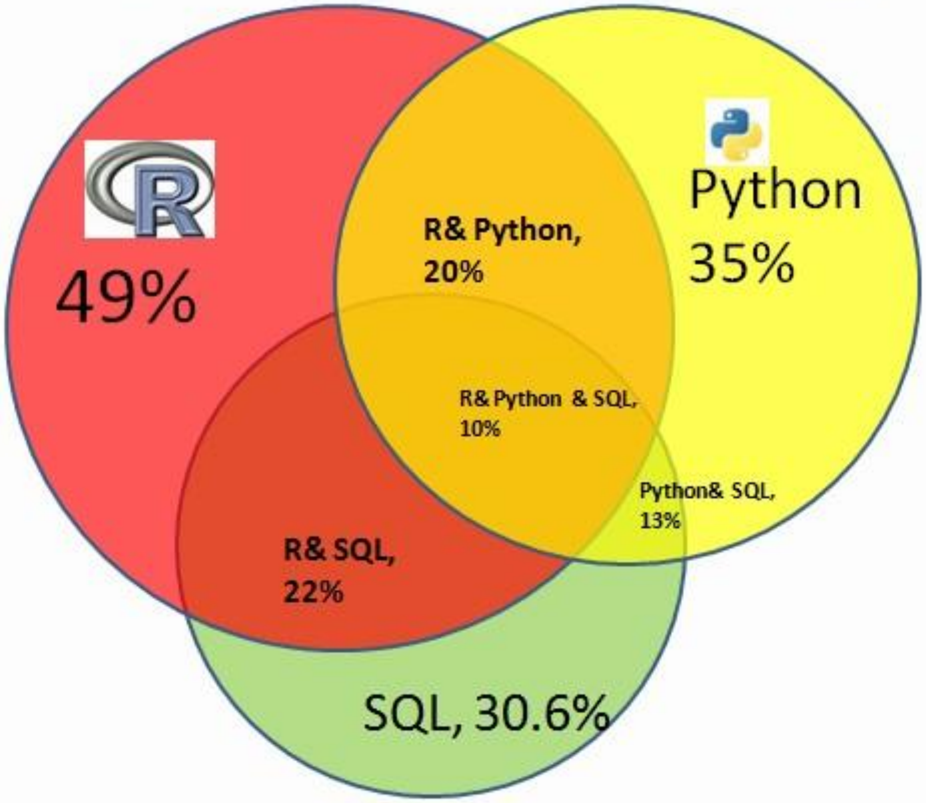




Why R?

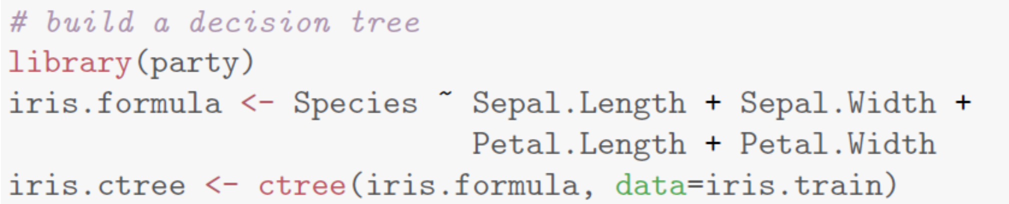
- R is widely used in both academia and industry.
- R was ranked no. 1 in the KDnuggets 2014 poll on Top Languages for analytics, data mining, data science (actually, no. 1 in 2011, 2012 & 2013!).
- The CRAN Task Views 9 provide collections of packages for different tasks.

KDnuggets 2014 Poll: Languages used for Analytics/Data Mining

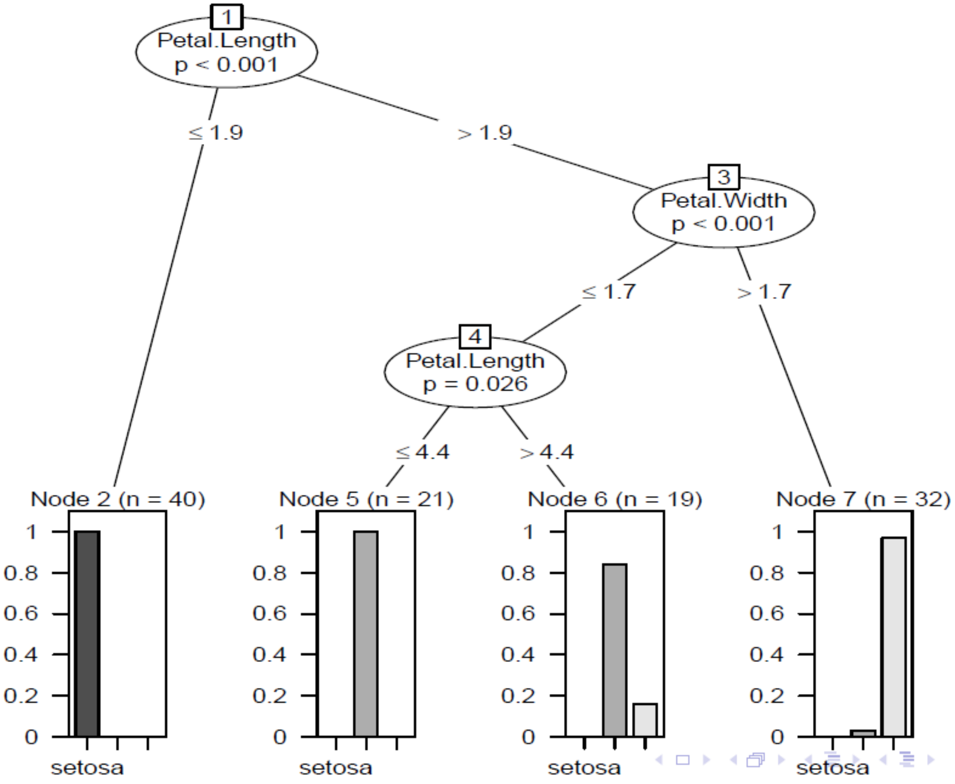


Classification with R

- Decision trees: *rpart*, *party*
- Random forest: *randomForest*, *party*
- SVM: *e1071*, *kernlab*
- Neural networks: *nnet*, *neuralnet*, *RSNNS*
- Performance evaluation: *ROCR*

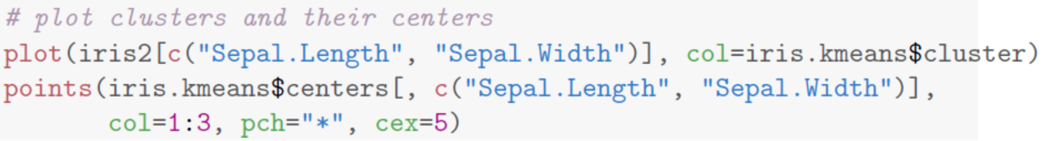


```
# build a decision tree
library(party)
iris.formula <- Species ~ Sepal.Length + Sepal.Width +
                          Petal.Length + Petal.Width
iris.ctree <- ctree(iris.formula, data=iris.train)
```

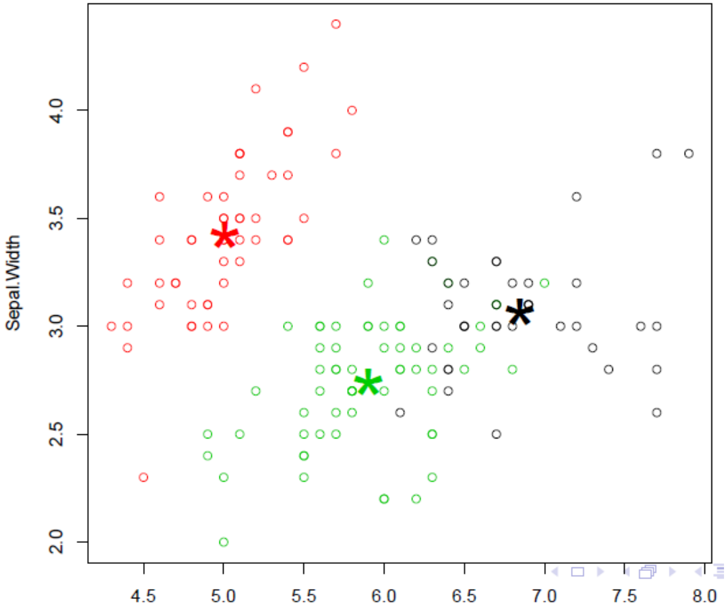


Clustering with R

- k-means: *kmeans(), kmeansruns()*
- k-medoids: *pam(), pamk()*
- Hierarchical clustering: *hclust(), agnes(), diana()*
- DBSCAN: *fpc*
- BIRCH: *birch*

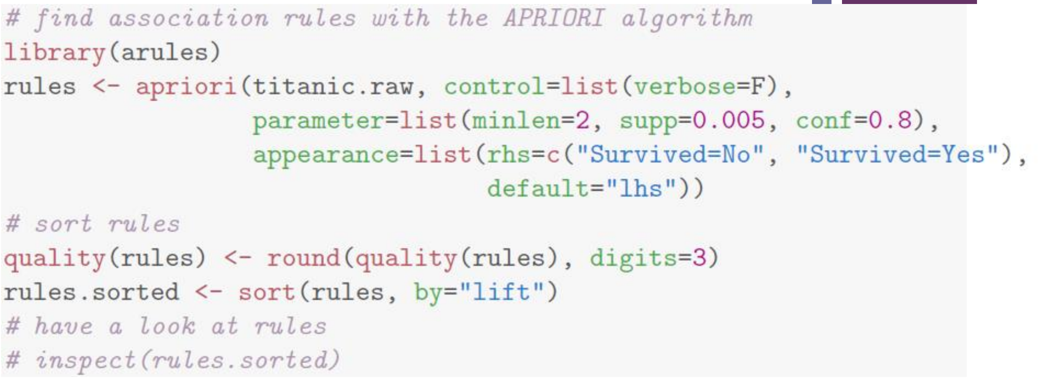


```
# plot clusters and their centers
plot(iris2[c("Sepal.Length", "Sepal.Width")], col=iris.kmeans$cluster)
points(iris.kmeans$centers[, c("Sepal.Length", "Sepal.Width")],
       col=1:3, pch="*", cex=5)
```

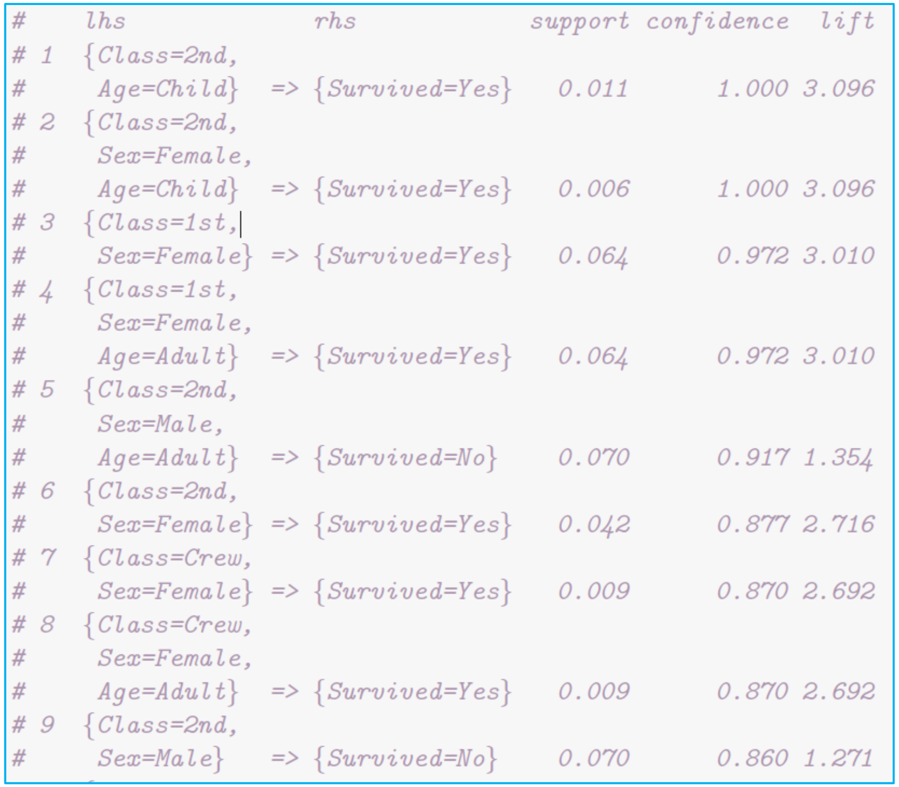


+ Association Rule Mining with R

- Association rules: *apriori(), eclat()* in package *arules*
- Sequential patterns: *arulesSequence*
- Visualization of associations: *arulesViz*



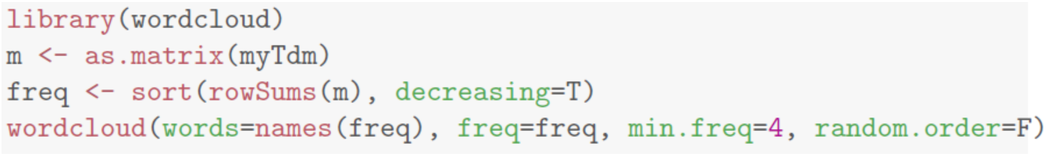
```
# find association rules with the APRIORI algorithm
library(arules)
rules <- apriori(titanic.raw, control=list(verbose=F),
                 parameter=list(minlen=2, supp=0.005, conf=0.8),
                 appearance=list(rhs=c("Survived=No", "Survived=Yes"),
                                 default="lhs"))
# sort rules
quality(rules) <- round(quality(rules), digits=3)
rules.sorted <- sort(rules, by="lift")
# have a look at rules
# inspect(rules.sorted)
```



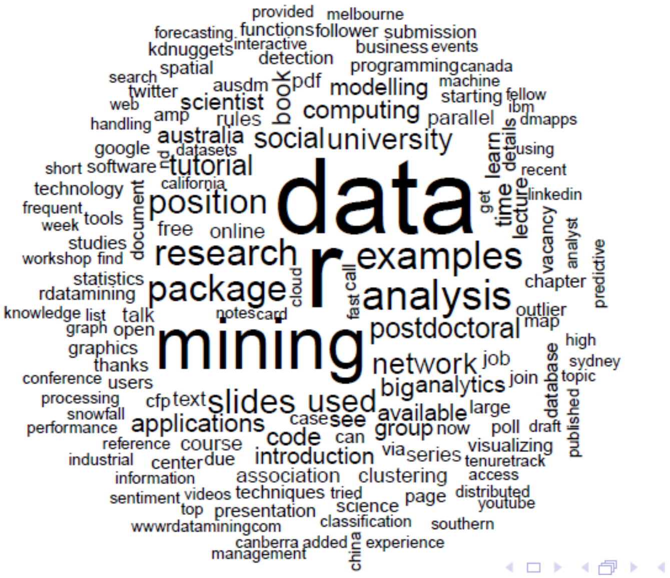
```
#     lhs                rhs            support confidence  lift
# 1  {Class=2nd,
#     Age=Child}     => {Survived=Yes}   0.011      1.000 3.096
# 2  {Class=2nd,
#     Sex=Female,
#     Age=Child}     => {Survived=Yes}   0.006      1.000 3.096
# 3  {Class=1st,
#     Sex=Female}    => {Survived=Yes}   0.064      0.972 3.010
# 4  {Class=1st,
#     Sex=Female,
#     Age=Adult}     => {Survived=Yes}   0.064      0.972 3.010
# 5  {Class=2nd,
#     Sex=Male,
#     Age=Adult}     => {Survived=No}    0.070      0.917 1.354
# 6  {Class=2nd,
#     Sex=Female}    => {Survived=Yes}   0.042      0.877 2.716
# 7  {Class=Crew,
#     Sex=Female}    => {Survived=Yes}   0.009      0.870 2.692
# 8  {Class=Crew,
#     Sex=Female,
#     Age=Adult}     => {Survived=Yes}   0.009      0.870 2.692
# 9  {Class=2nd,
#     Sex=Male}      => {Survived=No}    0.070      0.860 1.271
```

Text Mining with R

- Text mining: *tm*
- Topic modelling: *topicmodels, lda*
- Word cloud: *wordcloud*
- Twitter data access: *twitteR*



```
library(wordcloud)
m <- as.matrix(myTdm)
freq <- sort(rowSums(m), decreasing=T)
wordcloud(words=names(freq), freq=freq, min.freq=4, random.order=F)
```



Time Series Analysis with R

- Time series decomposition: *decomp(), decompose(), arima(), stl()*
- Time series forecasting: *forecast*
- Time Series Clustering: *TSclust*
- Dynamic Time Warping (DTW): *dtw*

Social Network Analysis with R

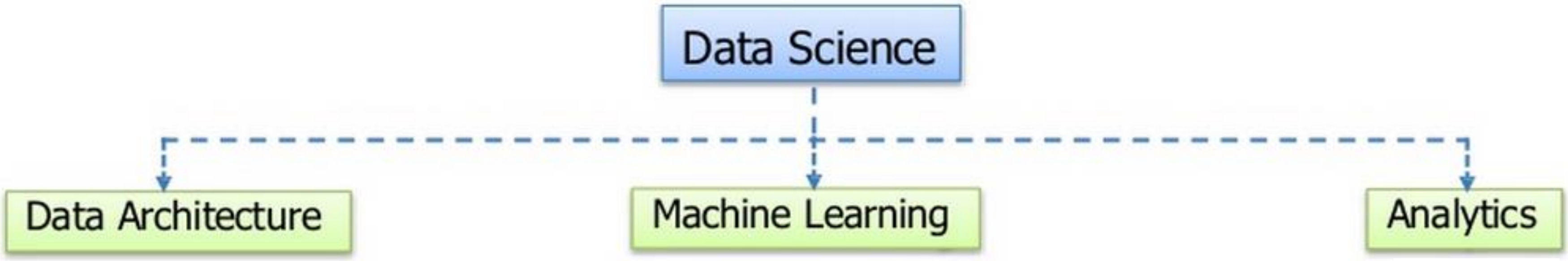
- Packages: *igraph, sna*
- Centrality measures: *degree(), betweenness(), closeness(), transitivity()*
- Clusters: *clusters(), no.clusters()*
- Cliques: *cliques(), largest.cliques(), maximal.cliques(), clique.number()*
- Community detection: *fastgreedy.community(), spinglass.community()*

R and Big Data

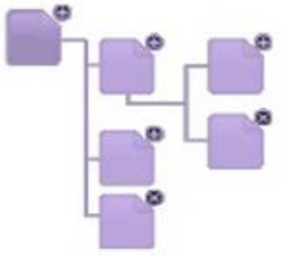
- Hadoop
  - Hadoop (or YARN) - a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models
  - R Packages: RHadoop, RHIPE
- Spark
  - Spark - a fast and general engine for large-scale data processing, which can be 100 times faster than Hadoop
  - SparkR - R frontend for Spark
- H2O
  - H2O - an open source in-memory prediction engine for big data science
  - R Package: h2o
- MongoDB
  - MongoDB - an open-source document database
  - R packages: rmongodb, RMongo

A Framework for Big Data Analytics

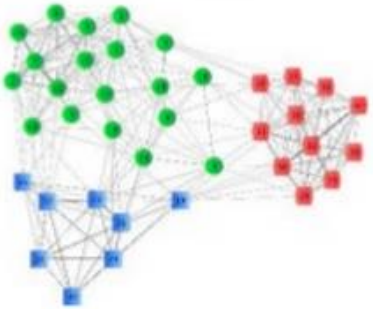
Big Data Analytics: Components



Tool: Hadoop

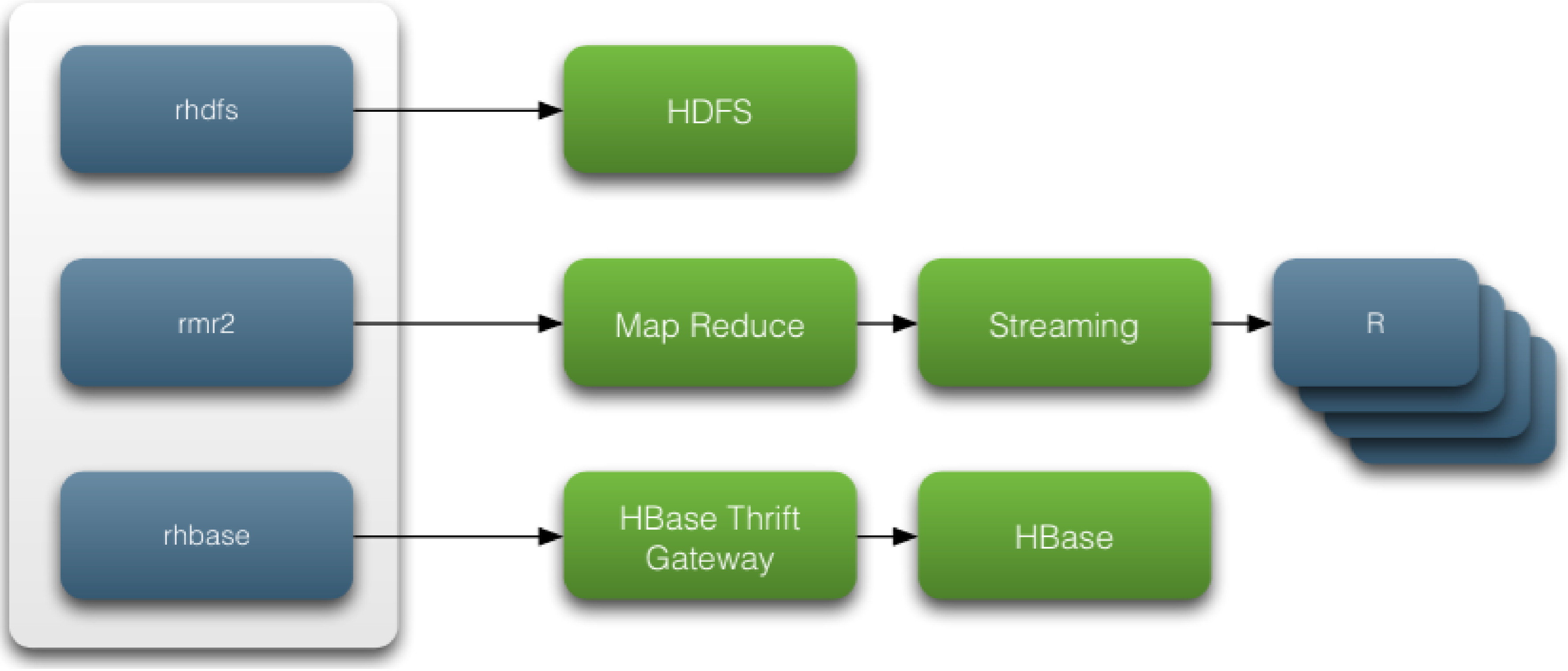


R Tools: RHadoop, H2O

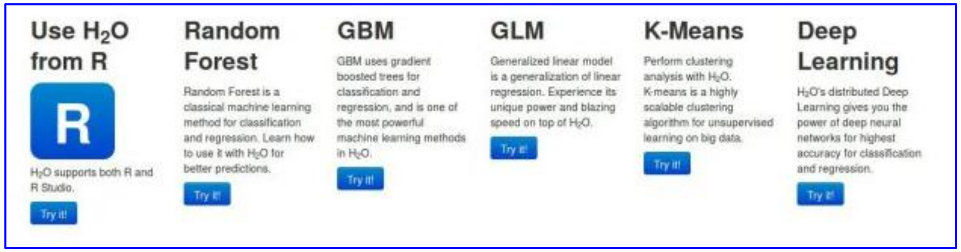


Tool: R

RHadoop


rhdfs
rmr2
rhbase
HDFS
Map Reduce
Streaming
R
HBase Thrift Gateway
HBase

H2O



Use H_2O from R

H_2O supports both R and R Studio.

Random Forest

Random Forest is a classical machine learning method for classification and regression. Learn how to use it with H_2O for better predictions.

GBM

GBM uses gradient boosted trees for classification and regression, and is one of the most powerful machine learning methods in H_2O.

GLM

Generalized linear model is a generalization of linear regression. Experience its unique power and blazing speed on top of H_2O.

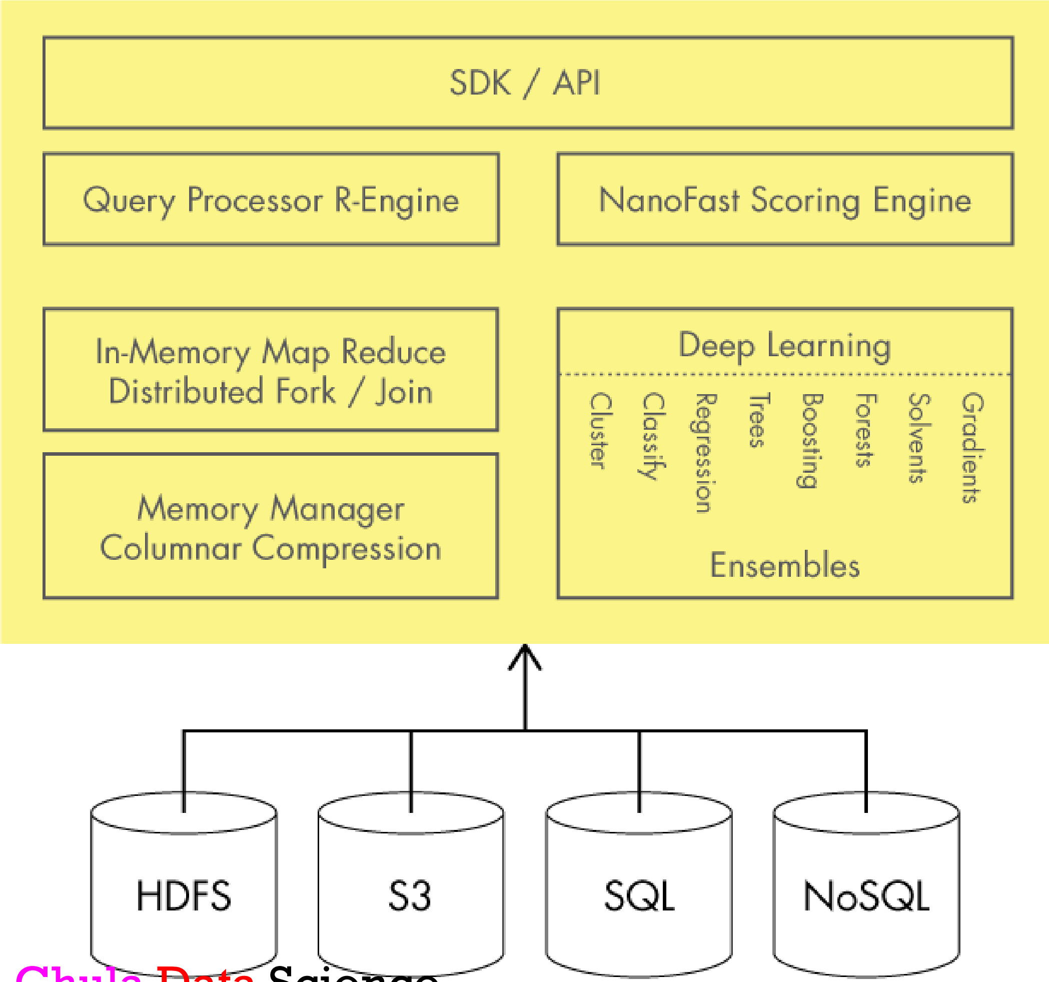
K-Means

Perform clustering analysis with H_2O. K-means is a highly scalable clustering algorithm for unsupervised learning on big data.

Deep Learning

H_2O's distributed Deep Learning gives you the power of deep neural networks for highest accuracy for classification and regression.

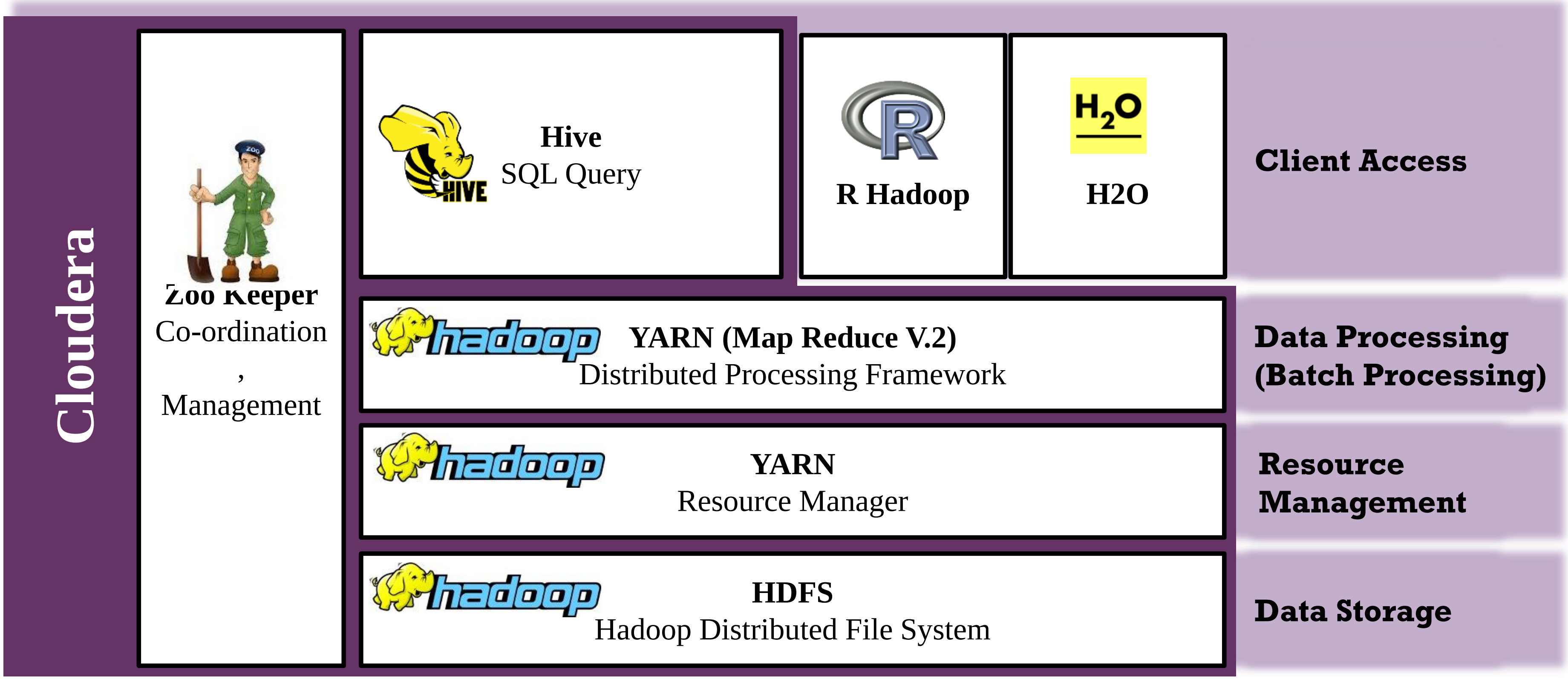
Python JSON R Scala Tableau Excel



1. Regression
2. Classification
3. Clustering
4. Others: Recommendation, Time Series

Chula Data Science

Big data & Analytic Architecture



Cloudera

Zoo Keeper
Co-ordination
,
Management

Hive
SQL Query

R Hadoop

H2O

Client Access

YARN (Map Reduce V.2)
Distributed Processing Framework

Data Processing (Batch Processing)

YARN
Resource Manager

Resource Management

HDFS
Hadoop Distributed File System

Data Storage

YARN enables multiple processing applications

Program List

Language	Management	Hadoop Ecosystem	Analytic
JAVA R	Cloudera	HDFS YARN HIVE Zoo Keeper	RHadoop RStudio Server H2O

Use Case: Predict Airline Delays

- Every year approximately 20% of airline flights are delayed or cancelled, resulting in significant costs to both travelers and airlines.

- Datasets:
  - Airline delay data (1987-2008)
  - http://stat-computing.org/dataexpo/2009/
  - **12 GB!**

- Goal:
  - Predict delay (delayTime >= 15 mins) in flights

+

Thank you & Any questions?