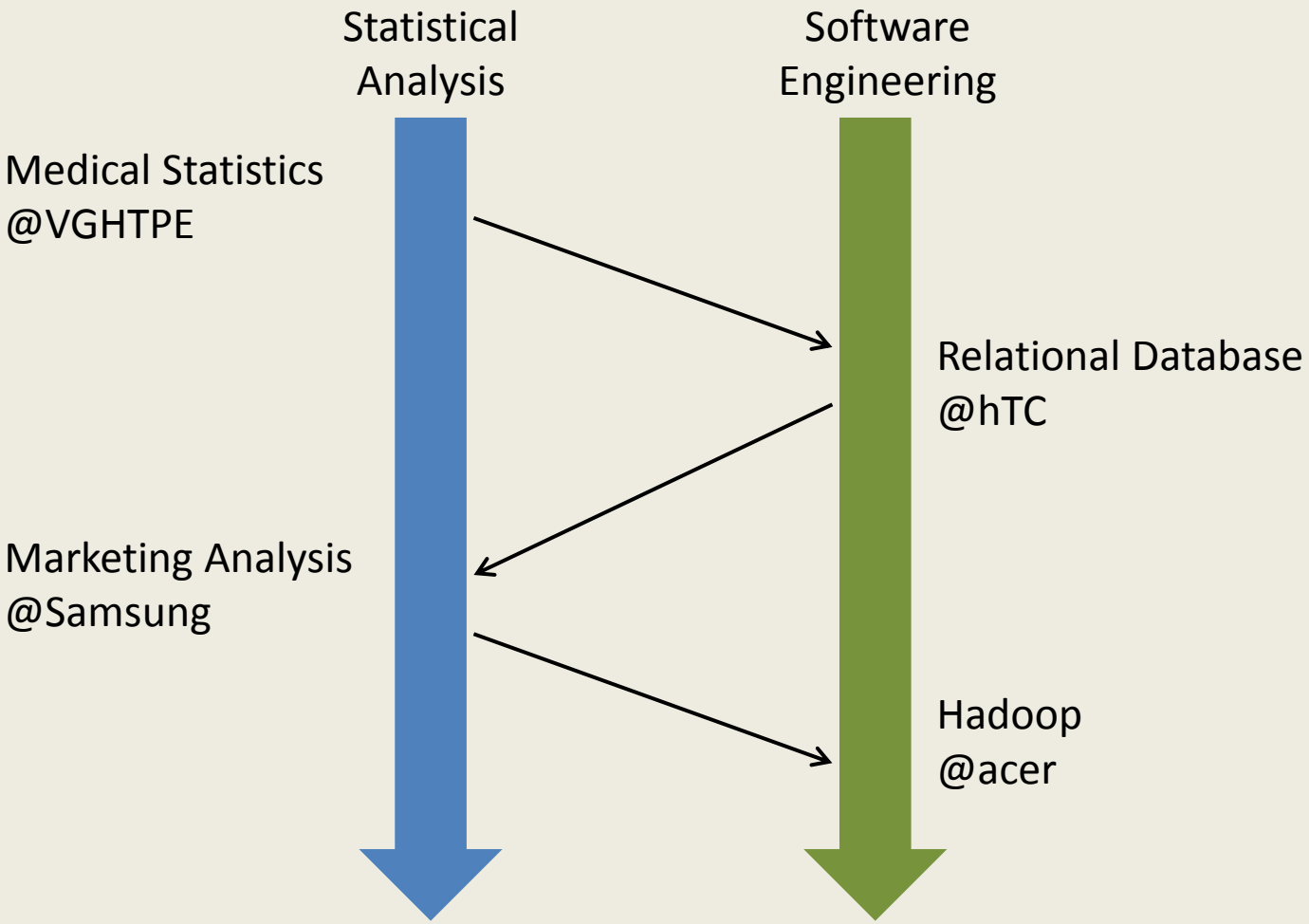
Between Statistics and IT



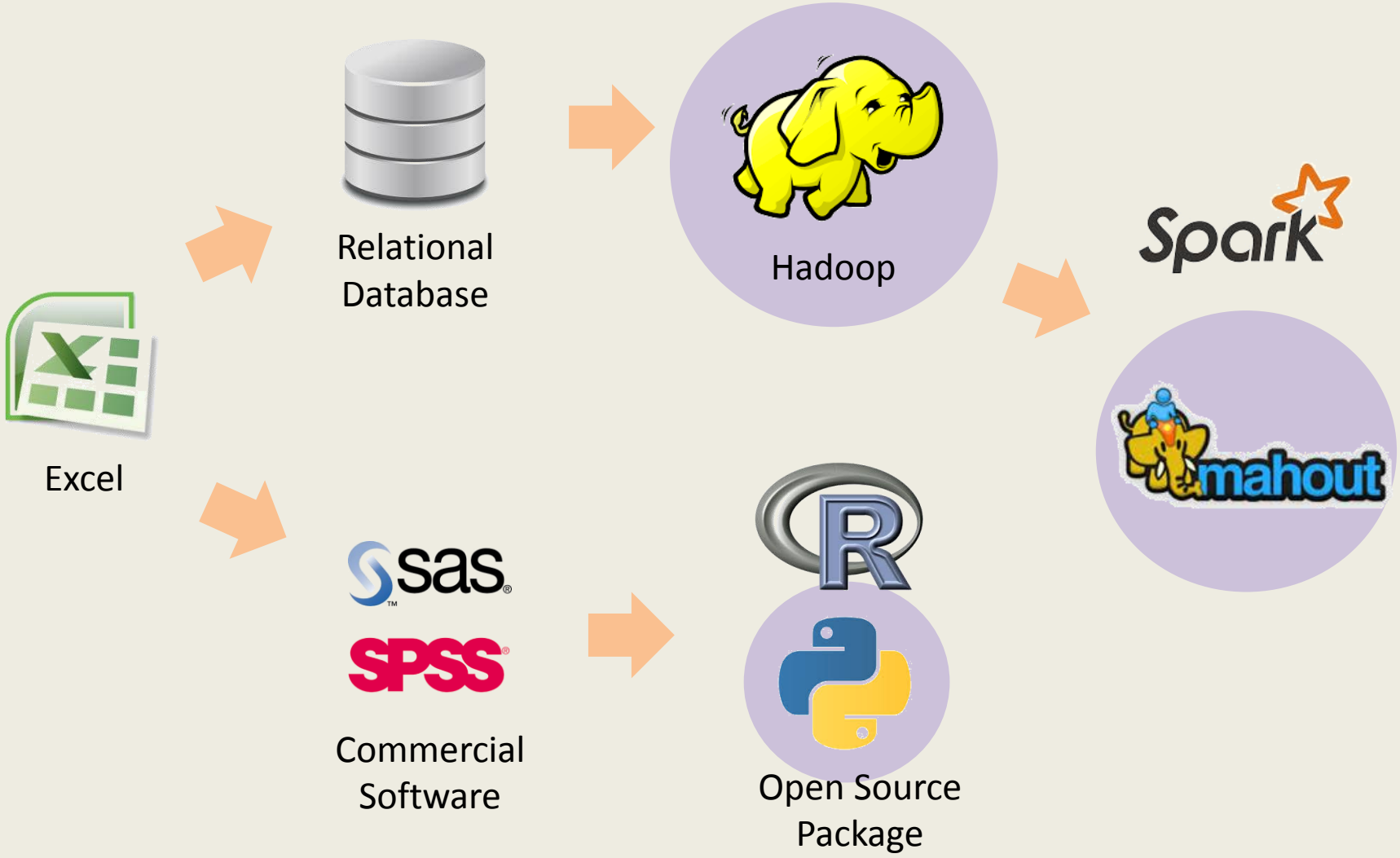
Definition

- Data Scientist (n.):

Person who is better at statistics than any software engineer and better at software engineering than any statistician.

-- Josh Wills @Cloudera

Useful Tools



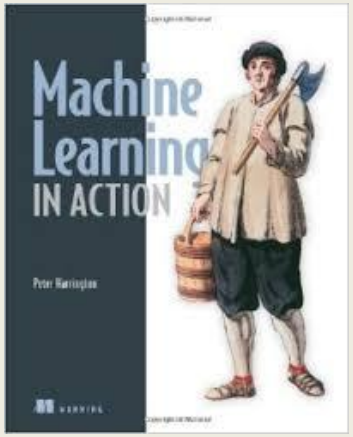
Python Packages

- NumPy – 向量運算
- SciPy – 科學計算
- Matplotlib – 繪圖
- Pandas – 資料分析
- StatsModels – 統計模型
- scikit.learn – 機器學習

- Integrated Package: Anaconda of Python(X,Y)

Machine Learning in Action

- Machine Learning Algorithms with Python
  - Classification: KNN, decision tree, naïve Bayes, logistic regression, SVM, AdaBoost
  - Predictive: regression, tree-based regression
  - Unsupervised: K-mean, Apriori, FP-growth
  - Others: PCA, SVD



Famous Cases

- Walmart - 啤酒與尿布
  - Walmart運用購物籃分析, 發現每週四啤酒與尿布一起銷售的機率非常高, 調查後發現男性常在週四採購周末Party需要的啤酒, 妻子則會順便提醒先生順便買小孩的尿布. 得知此事實後Walmart調整這兩項商品貨架位置, 使得獲利上升.

    -- 1998年哈佛商業評論

- Target - 預測少女懷孕
  - 一位父親怒氣沖沖的到Target百貨向經理控訴: 你們怎麼可以寄送嬰兒用品折價券給未成年高中生女兒? 一個月後, 父親又回去跟經理道歉, 原來女兒真的懷孕了. Target百貨依據消費者個人資料以及購物行爲改變(無香味乳液與葉酸營養品), 成功預測消費者已經懷孕

    -- 2012年紐約時報

Hadoop Streaming with Python

- hadoop jar $HADOOP_HOME/hadoop-streaming.jar \
  -D mapred.job.name=‘job001' \
  -input /data \
  -output /output \
  -file knn_map.py \
  -mapper knn_map.py \
  -file knn_reduce.py \
  -reducer knn_reduce.py

Hive with Python

```
from (
        select transform(a)
        using '${env:HOME}/josh/script/s1.py'
        as a1
        from table1
) t1
select transform(a1)
using '${env:HOME}/josh/script/s2.py'
as b1
;
```

Mahout

- 機器學習函式庫
- MapReduce為主, 目前轉向Spark
- 演算法三大類別:
  - 推薦(recommendation)
  - 群集(clustering)
  - 分類(classification)

Recommendation (User-Based)

假設有顧客1~5,
對商品編號101~107
的評價如下表
現在要向顧客1號推薦商品

item \ user	1	2	3	4	5
101	5	2	2	5	4
102	3	2.5			3
103	2.5	5		3	2
104		2	4	4.5	4
105			4.5		3.5
106				4	4
107			5		

依據共同評價過的商品
計算顧客之間的相似度
發現顧客1號與4, 5號正相關
與2號負相關
3號則無法計算(樣本不夠)

user \ user	1	2	3	4	5
1	1.00	-0.76	-	1.00	0.94
2	-0.76	1.00	-	-0.97	-0.94
3	-	-	1.00	-1.00	-0.65
4	1.00	-0.97	-1.00	1.00	0.88
5	0.94	-0.94	-0.65	0.88	1.00

經計算後, 將4, 5號顧客評價高
分的商品104, 106推薦給顧客1

item \ user	1	2	3	4	5
101	5	2	2	5	4
102	3	2.5			3
103	2.5	5		3	2
104	5	2	4	4.5	4
105			4.5		3.5
106	4			4	4
107			5		

Recommendation Options

- 推薦方法
  - User-based, Item-based, Slope-One, SVD, Knn Item-Based, Tree Cluster
- 相似度估計
  - Pearson Correlation, Euclidean Distance, Consine, Spearman Correlation, Tanimoto Coefficient, LogLokelihood
- 鄰居選擇方法
  - Nearest-N, Threshold
- 其他因素
  - Boolean? Weight?

Run MapReduce

```
hadoop jar $MAHOUT_HOME/mahout-core-0.9-job.jar \
org.apache.mahout.cf.taste.hadoop.item.RecommenderJob \
-Dmapred.input.dir=/data \
-Dmapred.output.dir=/output \
-s SIMILARITY_PEARSON_CORRELATION
```

Clustering

- Google 新聞群組

Twitter Text Mining

- The Acer engineers are smoking crack again ?

Twitter Text Mining

Aspire R7 的四種使用模式

筆電模式　觸控模式　分享模式　平板模式

Twitter Text Mining

- 文字分析發現Meme已經開始傳播並衍生出三種版本
- 找出病毒傳播的網站並消毒澄清
- 將”acer engineers”設為關鍵字並持續追蹤

Total			**495**	
	the acer engineers are smoking crack again		**381**	
		i.imgur.com	240	
		9gag.com	52	
		_invalid	12	
		bestof4chan.org	9	
		www.bestof9gag.com	7	
		www.reddit.com	7	
		zumlerr.com	4	
		adf.ly	2	
		imgur.com	2	
		limk.com	2	
		q.gs	2	
		www.facebook.com	2	
		anotherboringblog.com	1	
		geekstumbles.com	1	
		hanyokasha.tumblr.com	1	
		kartelnometry.tumblr.com	1	
		latestfunnystuff.com	1	
		latestvcnews.wordpress.com	1	
		lazeblog.com	1	
		m.vk.com	1	
		omg-pictures.tumblr.com	1	
		tumblokami.tumblr.com	1	
		uberhumor.com	1	
		www.awwomg.com	1	
		www.dailylmao.net	1	
	the acer engineers are doing drugs again		**103**	
		twitter.com	103	@BestProHumor
	Acer engineers want to see the world burn		**11**	
		www.quiterly.com	1	

Classification

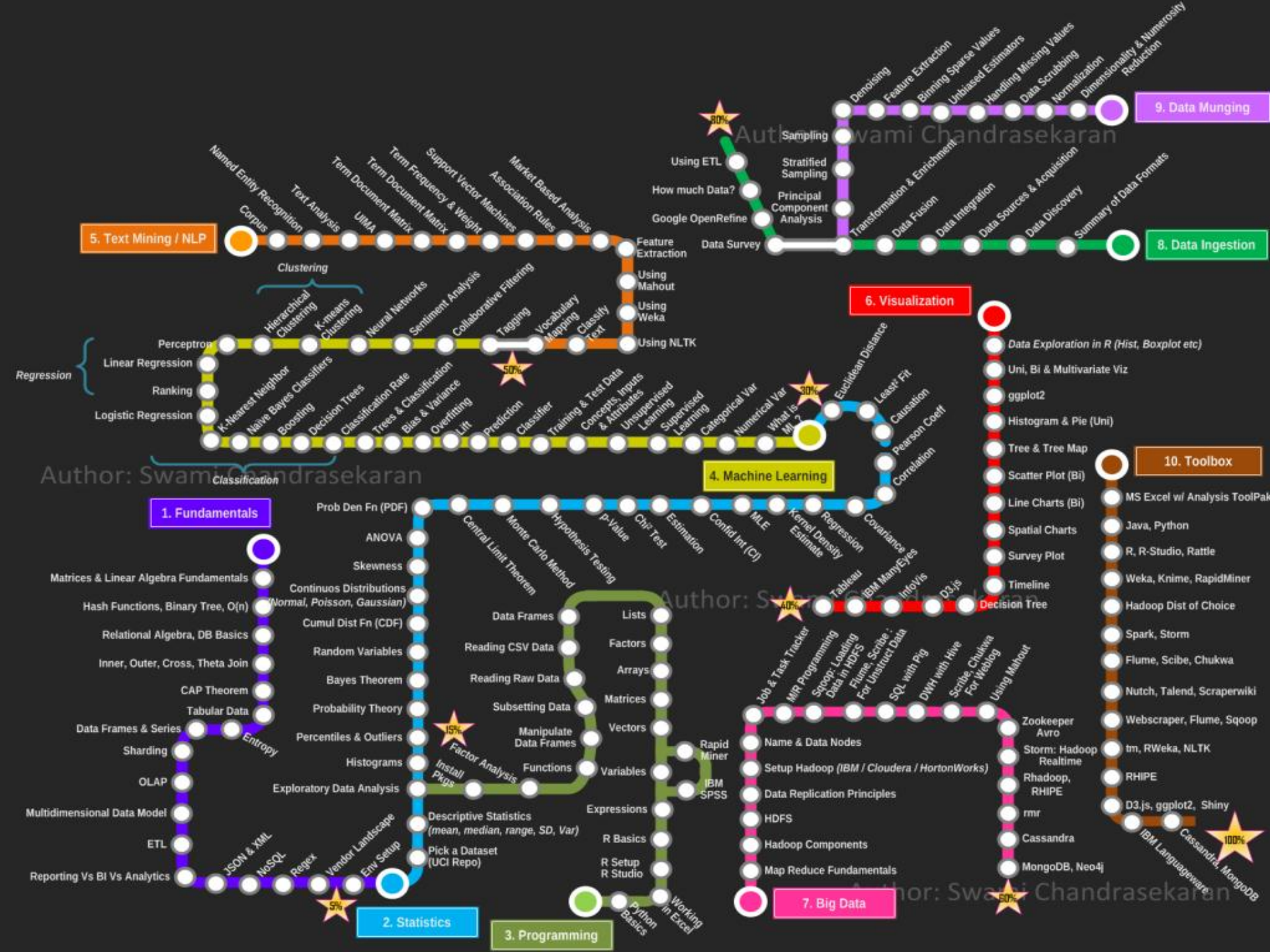
- Gmail垃圾郵件分類

Classification Algorithm

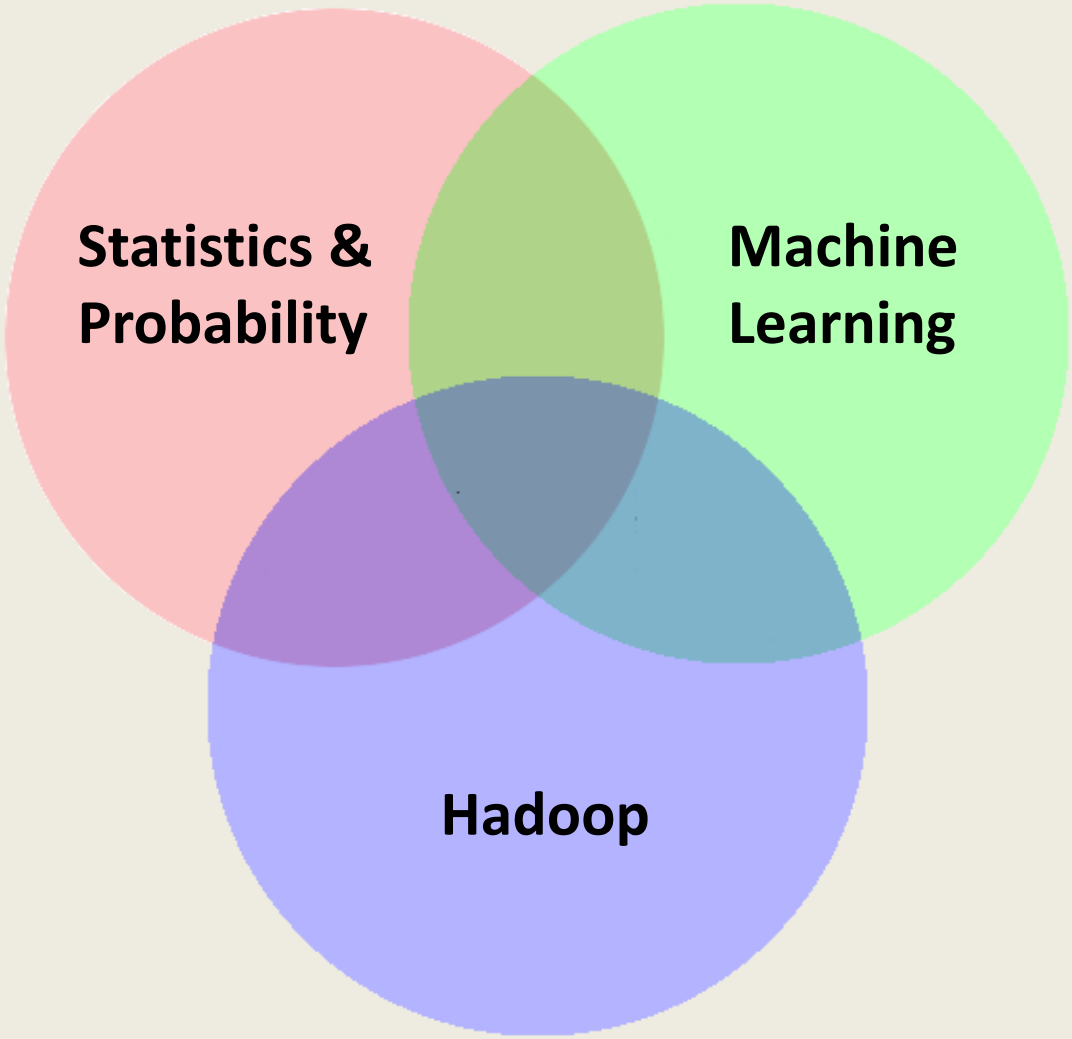
- Logistic Regression
- Naïve Bayesian
- Random Forests
- Hidden Markov Models

Cloudera Certification

- Hadoop Developer CCDH
- Hadoop Admin CCAH
- HBase Specialist CCSHB
- CCP: Data Scientist
  - Data Science Essentials (DS-200)
  - Data Science Challenge


1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Entropy
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Prob Den Fn (PDF)
ANOVA
Skewness
Continuos Distributions
(Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Descriptive Statistics
(mean, median, range, SD, Var)
Pick a Dataset
(UCI Repo)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
15%
Factor Analysis
Install Pkgs
3. Programming
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup
R Studio
Rapid Miner
IBM SPSS
Python Basics
Working in Excel
4. Machine Learning
Regression
Perceptron
Linear Regression
Ranking
Logistic Regression
Classification
K-Nearest Neighbor
Naive Bayes Classifiers
Boosting
Decision Trees
Classification Rate
Trees & Classification
Bias & Variance
Overfitting
Lift
Prediction
Classifier
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
What is ML?
50%
Clustering
Hierarchical Clustering
K-means Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
5. Text Mining / NLP
Corpus
Text Analysis
Named Entity Recognition
UIMA
Term Document Matrix
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Feature Extraction
Using Mahout
Using Weka
Using NLTK
30%
Euclidean Distance
Least² Fit
Causation
Pearson Coeff
Correlation
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
40%
Tableau
IBM ManyEyes
InfoVis
D3.js
7. Big Data
Job & Task Tracker
M/R Programming
Sqoop: Loading Data in HDFS
Flume, Scribe : For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper
Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%
8. Data Ingestion
Data Survey
Google OpenRefine
How much Data?
Using ETL
80%
Sampling
Stratified Sampling
Principal Component Analysis
Transformation & Enrichment
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
9. Data Munging
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%
Author: Swami Chandrasekaran

Data Science Essentials (DS-200)



Statistics & Probability

- Statistics Concepts
- Descriptive Statistics
- Probability Distributions
- Bayes Probability
- Likelihood Function
- Bagging and Boosting
- Other Statistics Concepts

Statistics & Probability

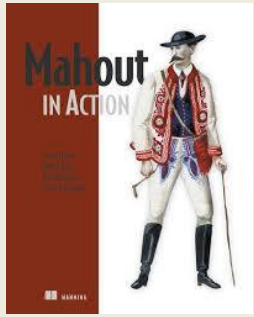
- 參考資料: 各種機率統計書籍
- 注重機率論, 統計模擬及相關線性代數
- 不重一般統計學: 假設檢定, 參數估計, ...

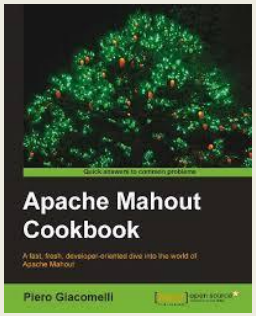
Machine Learning

- Algorithms Concept
  - Naïve Bayes
  - SVM
  - K-Mean
  - Collaborative Filtering
- Over/Under Fitting
- VC dimension
- Stochastic Gradient Descent

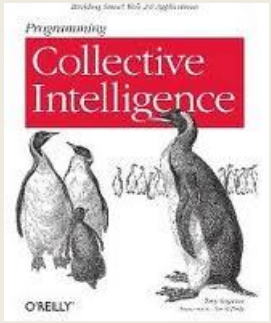
Machine Learning

- References:
  - Machine Learning in Action
  - Mahout in Action
  - Programming Collective Intelligence





- Coursera:
  - Machine Learning, Andrew Ng
  - 機器學習基石, 林軒田

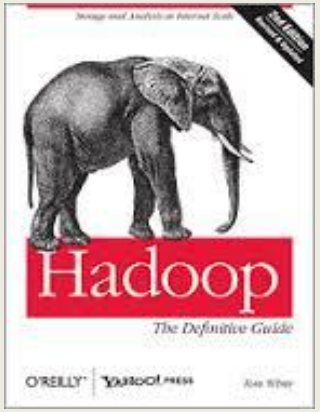


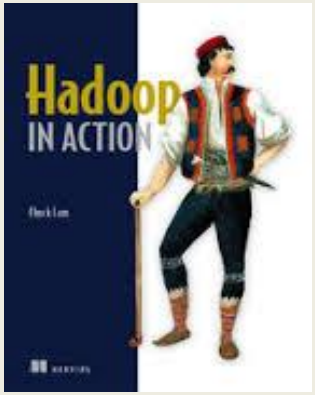
Hadoop

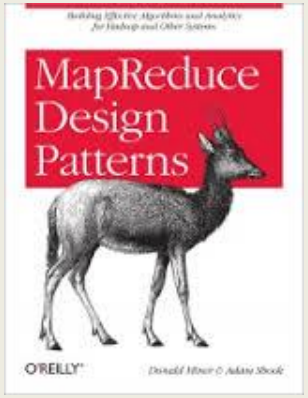
- MapReduce Concepts
- Combiner and Partitioner
- Distributed Cache
- Pseudo Code Analysis
- Hive, Pig , Sqoop and Flume Use Case

Hadoop

- Reference:
  - Hadoop Definitive Guide
  - Hadoop In Action
  - MapReduce Design Pattern







Data Science Challenge

- Web Analytics Challenge: Classification, Clustering, and Collaborative Filtering
  - Explore and Summary Data
  - Clean Data
  - Classify Adult or Kid
  - Cluster Users
  - Recommend File to User

Solution Kit

CCP: Data Scientist Solution Kit

- Hadoop Streaming + Python to Clean and Summary JSON data
- Python + bash shell to classify with SimRank Algorithm
- Cluster users with K-Mean Algorithm using Cloudera KM
- Recommend Film with Mahout

The End