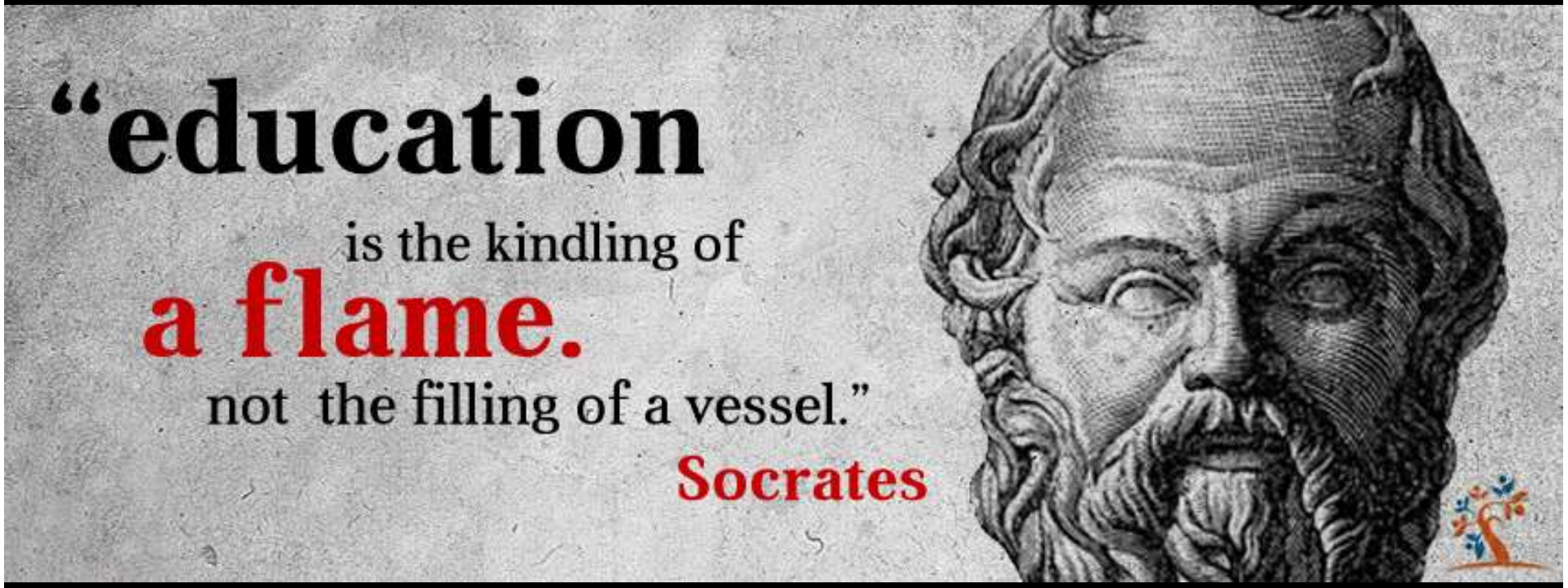
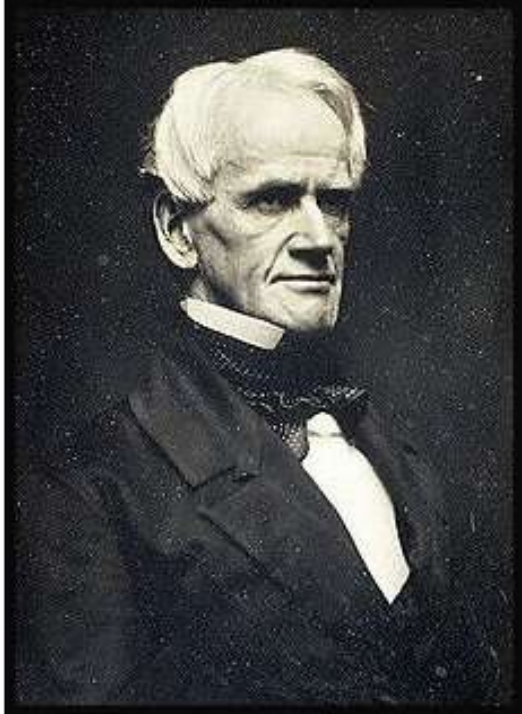
Data Mining:

Theory & Algorithms

Lecturer: Dr. Bo Yuan

E-mail: yuanb@sz.tsinghua.edu.cn


“education
is the kindling of
a flame.
not the filling of a vessel.”
Socrates



A teacher who is attempting to teach without inspiring the pupil with a desire to learn is hammering on cold iron.

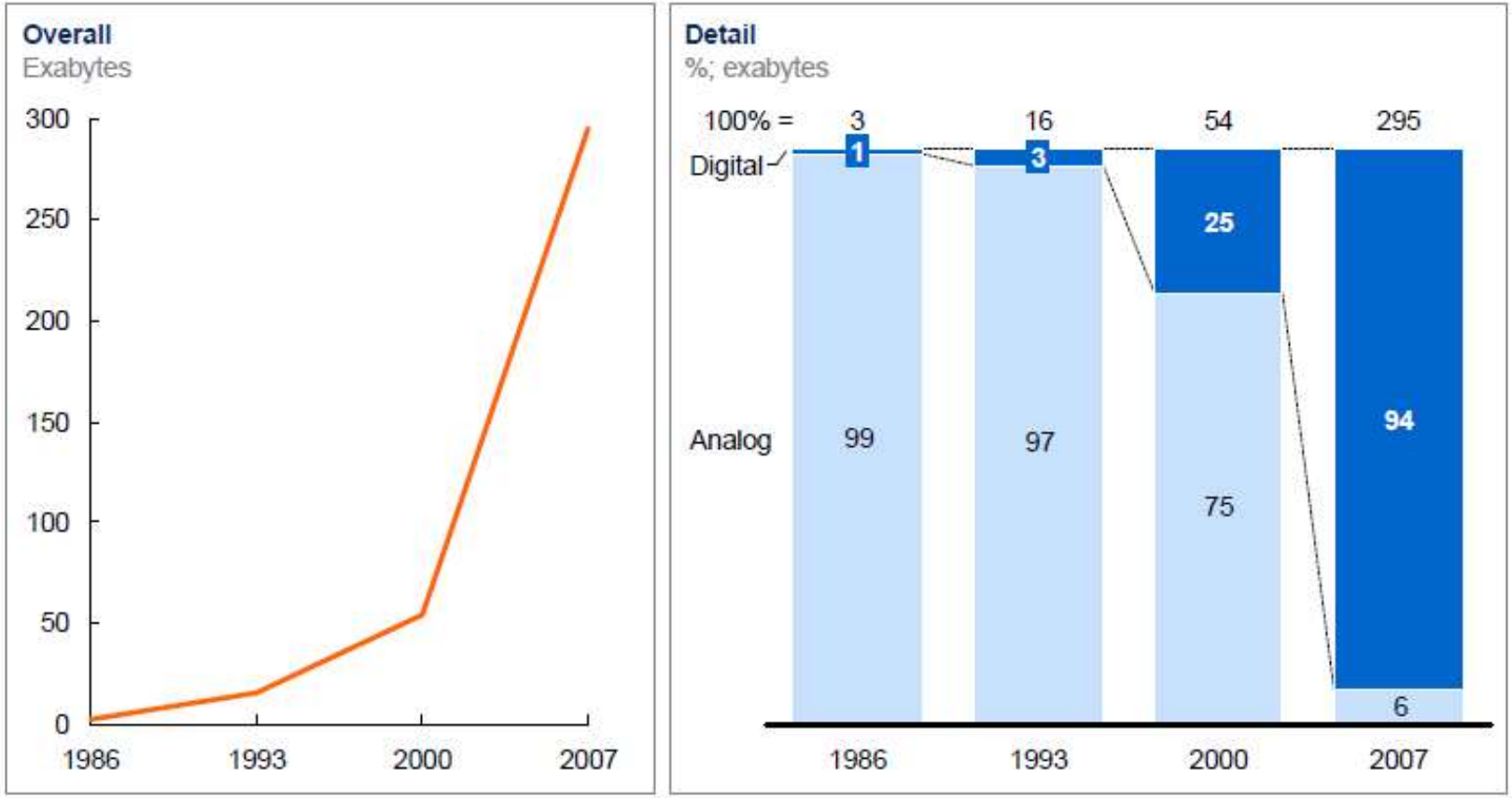
(Horace Mann)

Mining? Warehousing?

Technology Advancement

Data storage has grown significantly, shifting markedly from analog to digital after 2000

Global installed, optimally compressed, storage



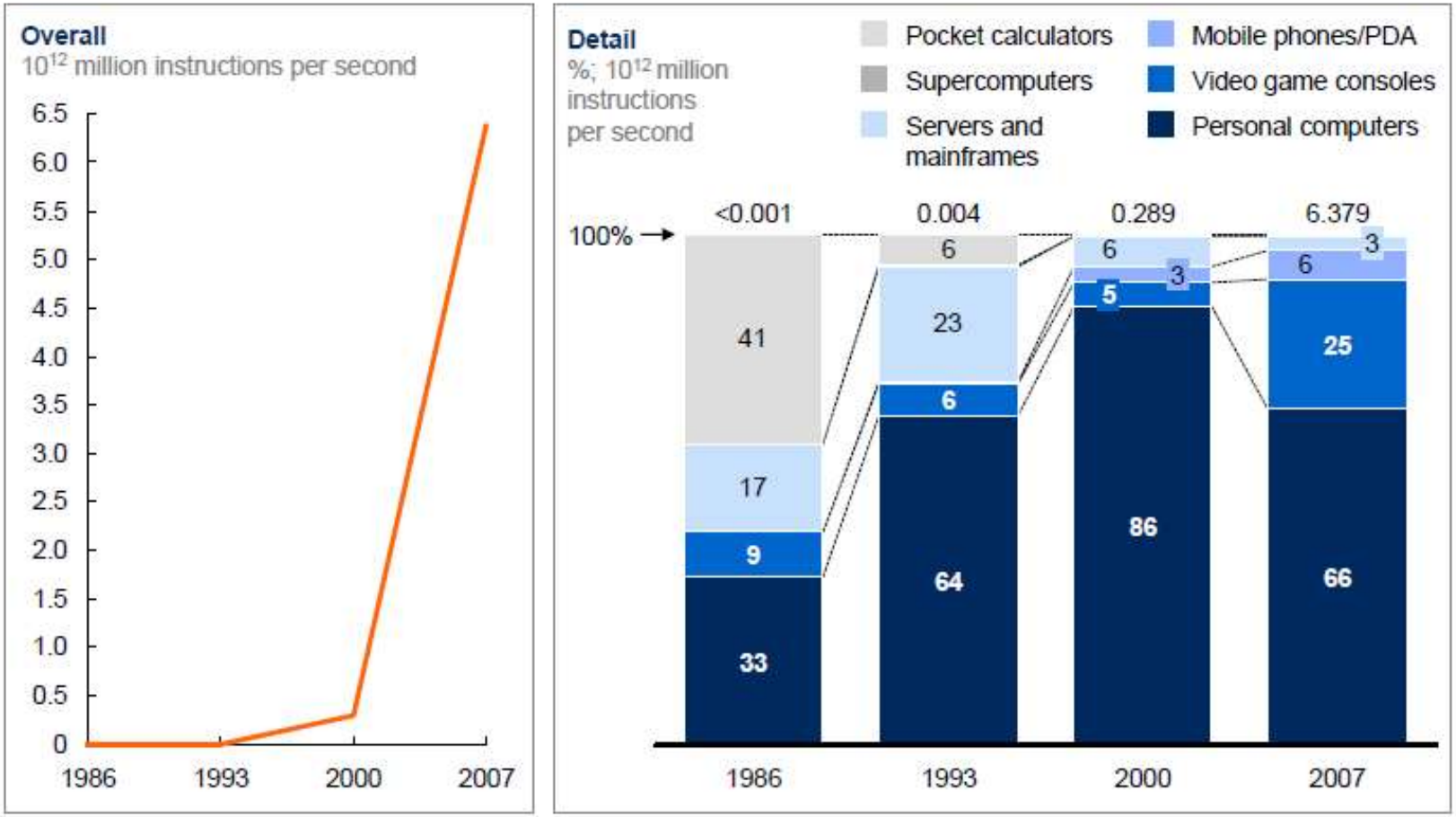
NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

Technology Advancement

Computation capacity has also risen sharply

Global installed computation to handle information



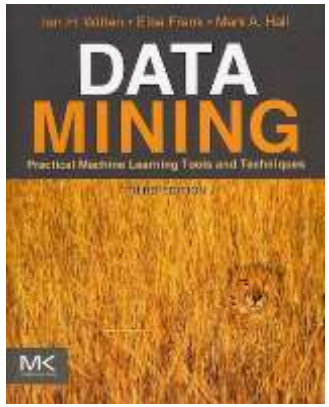
NOTE: Numbers may not sum due to rounding.

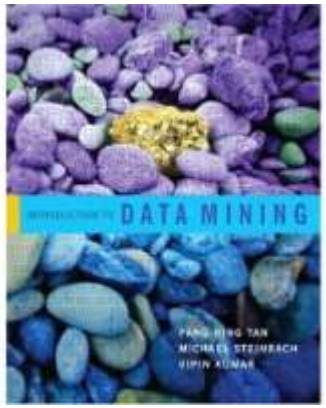
SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

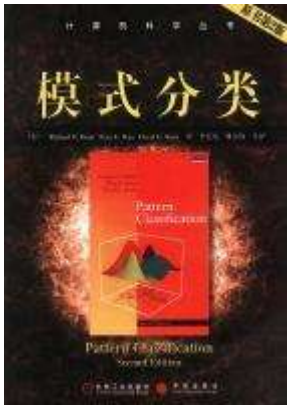
The World of Data

Data Rich, Information Poor

Learning Resources







International Conference on Data Mining

International Conference on Data Engineering

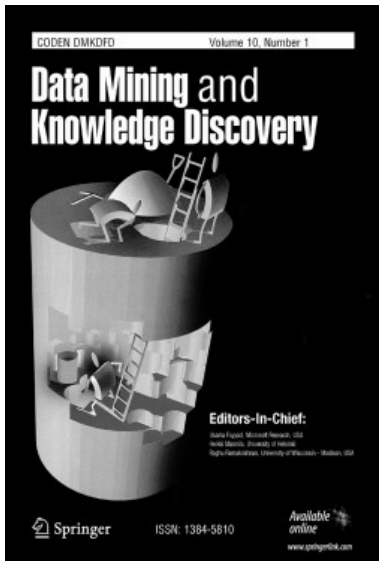
International Conference on Machine Learning

International Joint Conference on Artificial Intelligence

Pacific-Asia Conference on Knowledge Discovery and Data Mining

ACM SIGKDD Conference on Knowledge Discovery and Data Mining

Learning Resources



IEEE Computational Intelligence Society

IEEE computer society

Learning Resources

Xindong Wu

Zhihua Zhou

Jiawei Han

Jian Pei

Qiang Yang

Chih-Jen Lin

Philip S. Yu

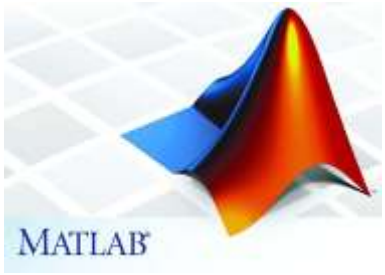
Changshui Zhang

Learning Resources

Google

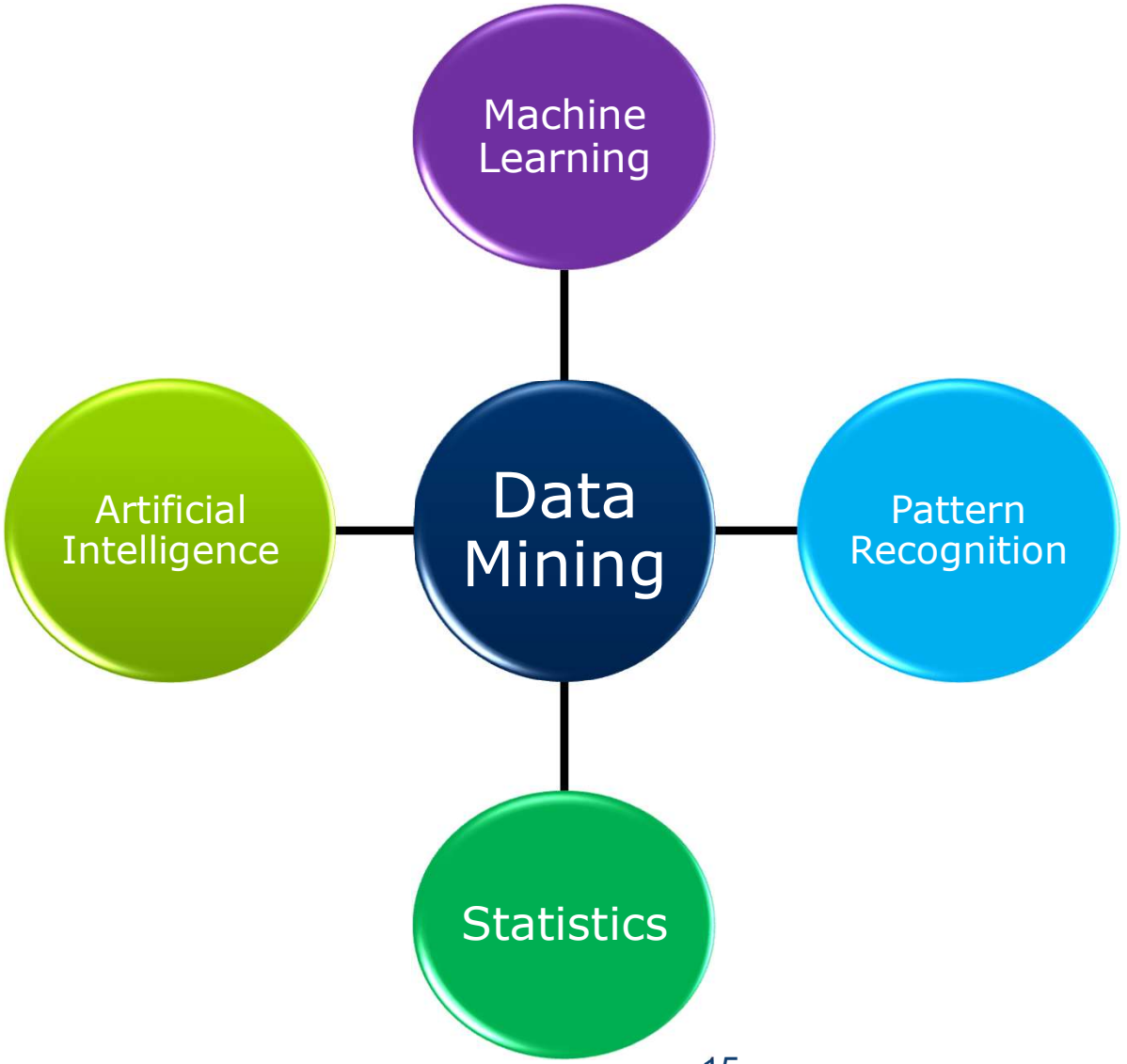
Google scholar

KDnuggets™

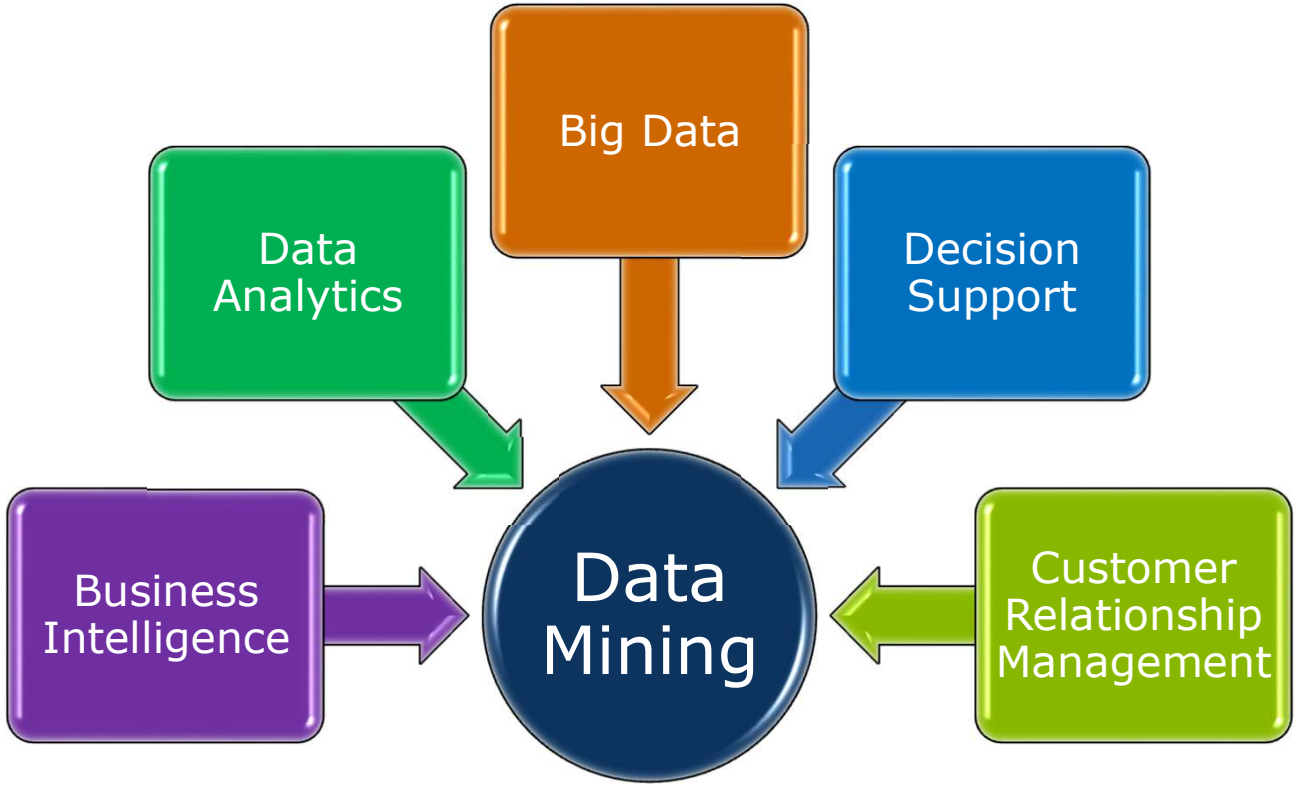


UCI Machine Learning Repository

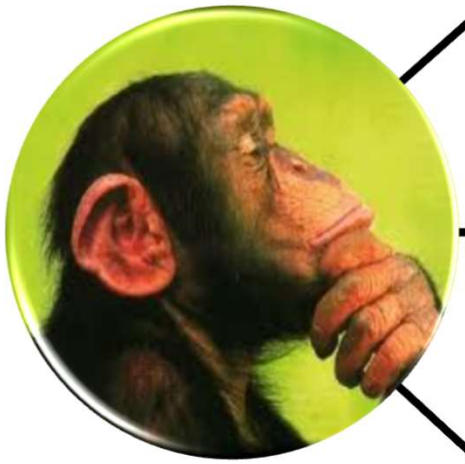
Interdisciplinary



Ubiquitous


Big Data
Data Analytics
Decision Support
Business Intelligence
Data Mining
Customer Relationship Management

Comprehensive Learning

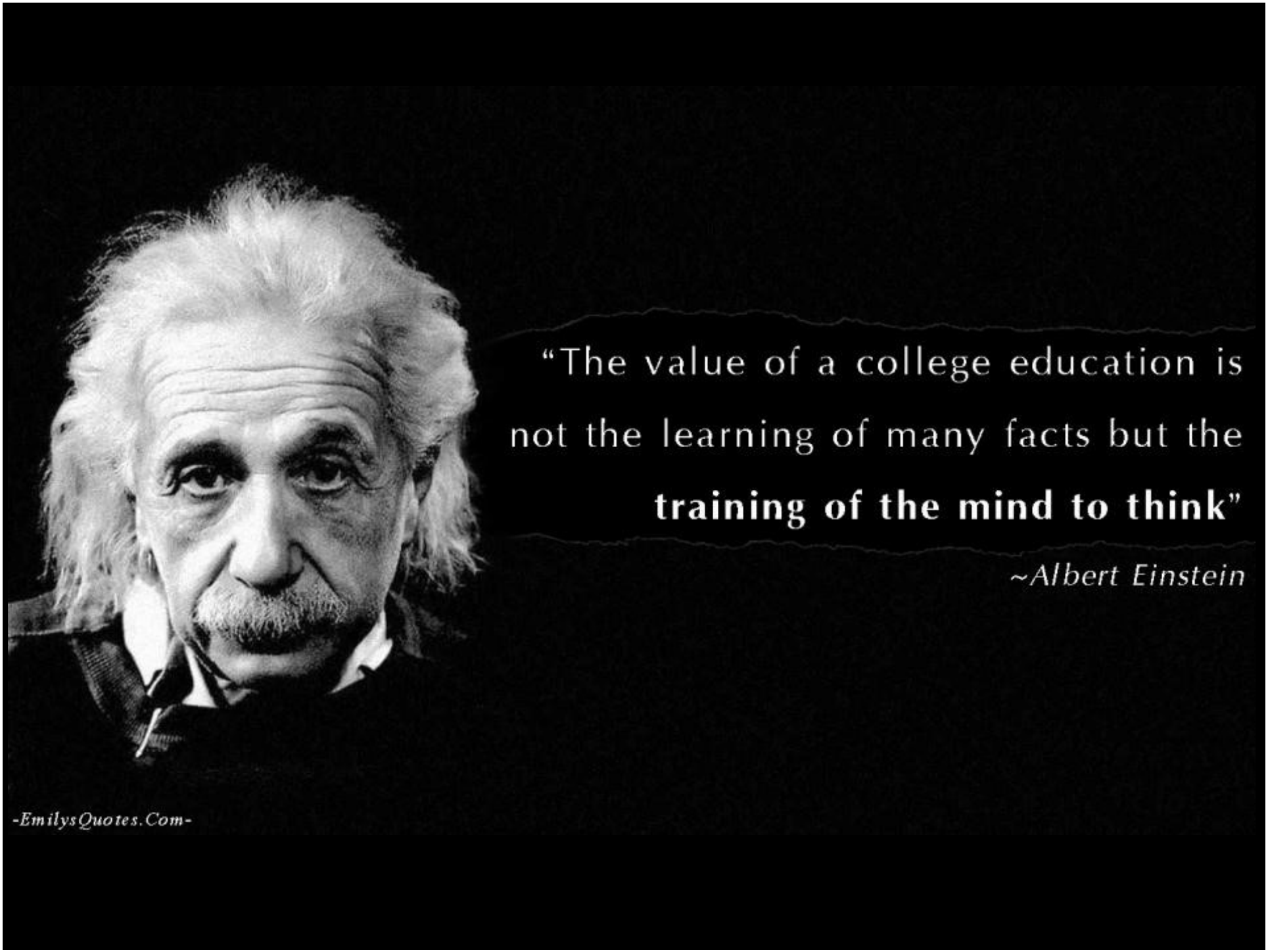


Class Teaching
- Thinking
- Discussion

Reading Materials
- Extension
- Inspiration

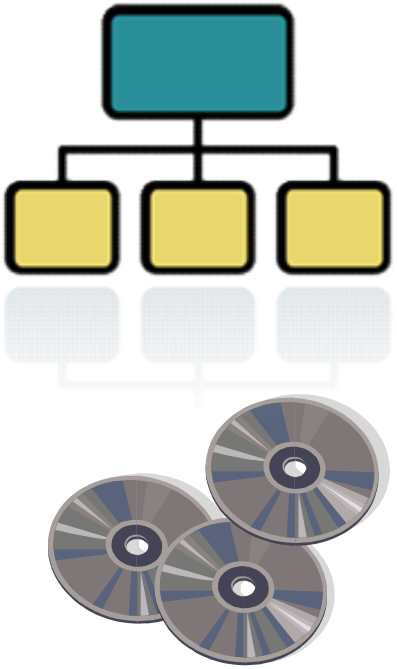
Practice
- Techniques
- Applications

Learning ≠ Listening


"The value of a college education is not the learning of many facts but the **training of the mind to think**"
~Albert Einstein
-EmilysQuotes.Com-

Data

- Definition
  - “Data are pieces of information that represent the qualitative or quantitative attributes of a variable or set of variables. Data are often viewed as the lowest level of abstraction from which information and knowledge are derived.”
- Data Types
  - Continuous, Binary
  - Discrete, String
  - Symbolic
- Storage
  - Physical
  - Logical
- Major Issues
  - Transformation
  - Errors and Corruption

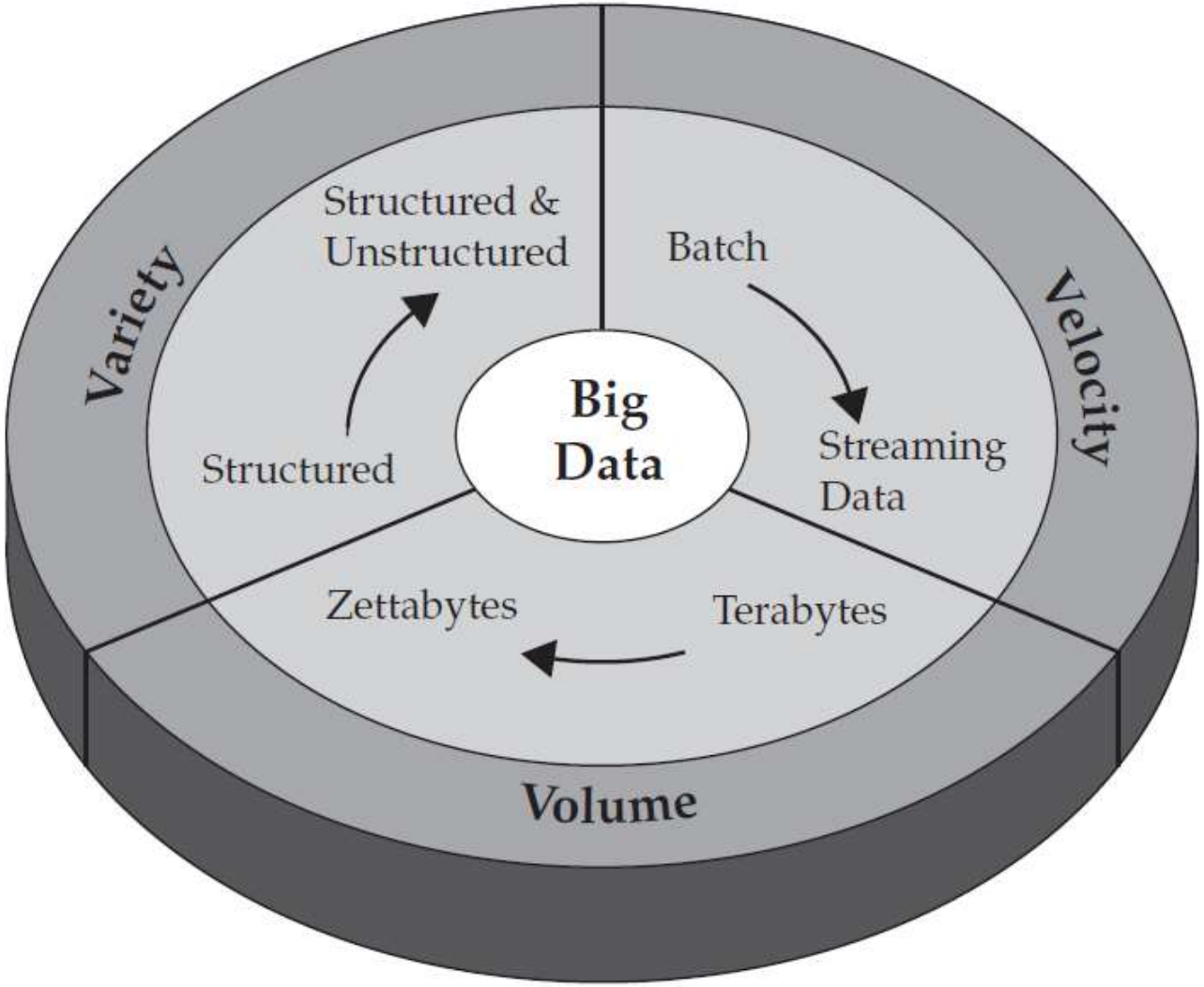


What is Big Data?

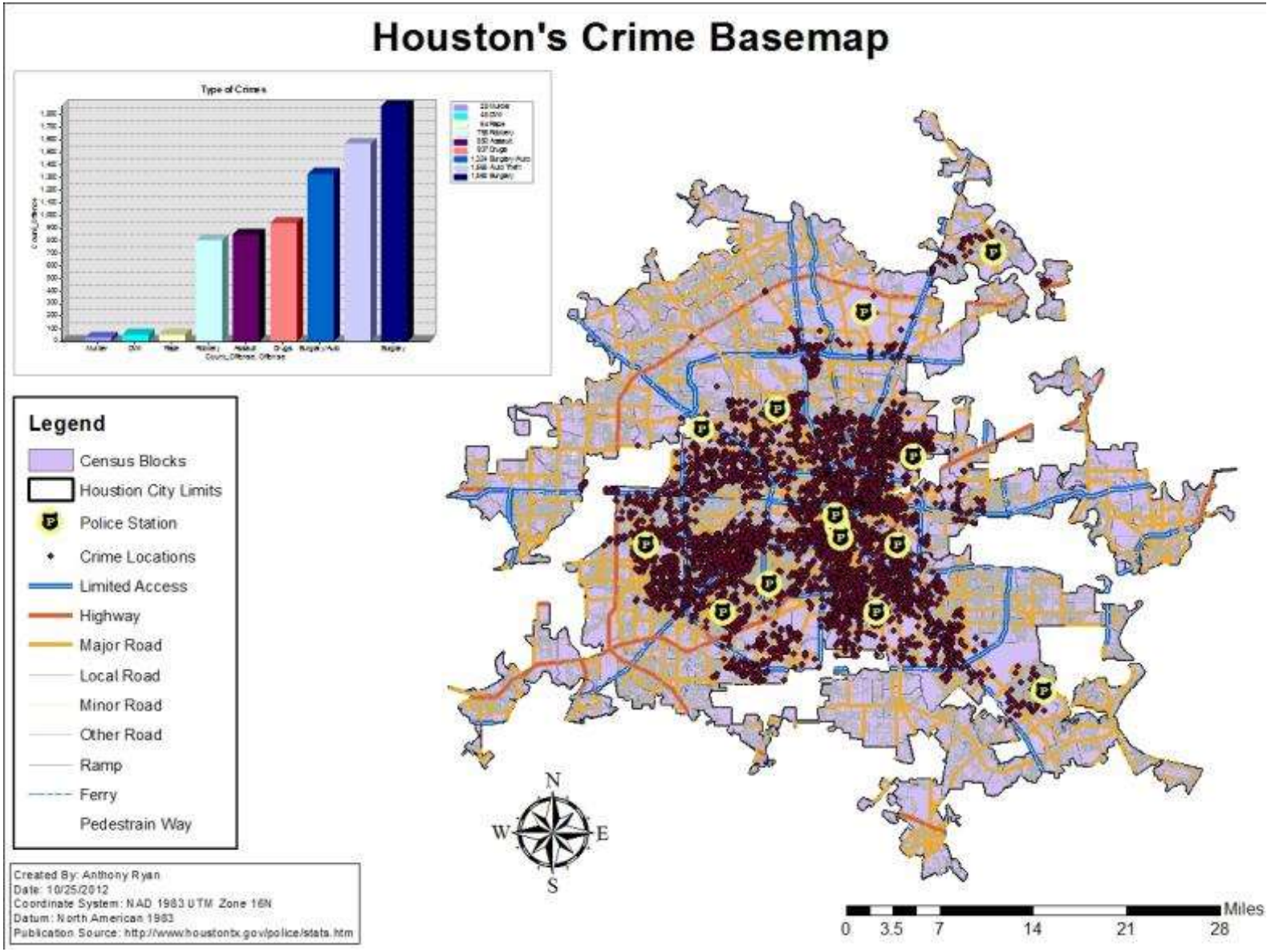
- **"Big data** is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." — ***Gartner***

- **"Big data** refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze." — ***Mckinsey & Company***

Big Data

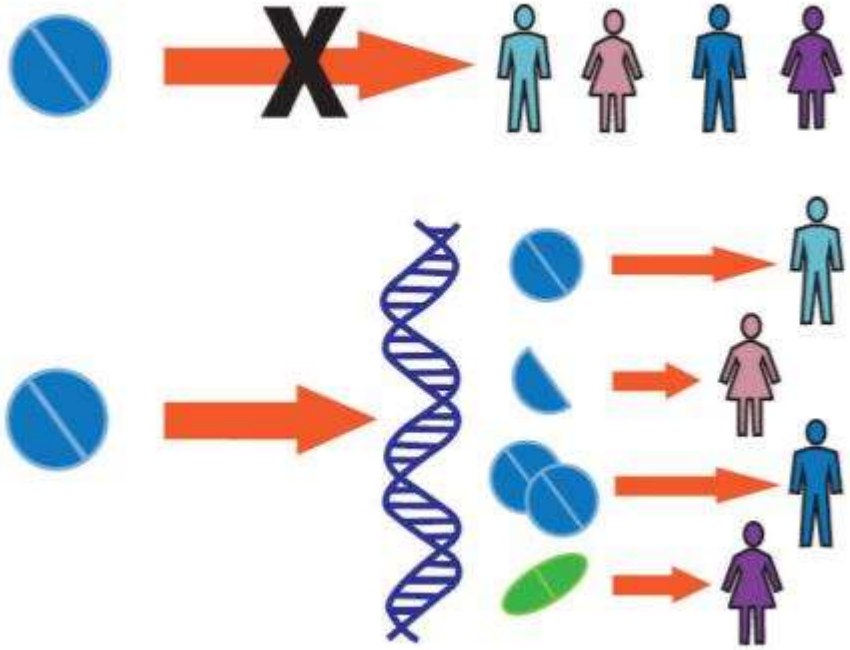

Variety
Structured &
Unstructured
Batch
Velocity
Big
Data
Structured
Streaming
Data
Zettabytes
Terabytes
Volume

Public Security



Health Care Application

Effectiveness Research



Personalized Medicine

Location Data: Urban Planning

Location Data: Mobile User

Mobile location-based services (LBS) and applications have proliferated

Mobile LBS applications continue to proliferate[1]

People locating (e.g., safety family/ child tracking, friend finder)

at&t FamilyMap

Pink Map

Location check-in/ sharing on social community applications

foursquare

loopt

Places

Who. What. When. And now **Where.**

City/regional guide, neighborhood service search

LON lonely planet

Location-enabled entertainment, e.g., mobile gaming, geo-tagged photo/travel

GEOCACHING.COM

Geo-tagged photo

Revenue through the "Freemium" model

Revenue model for these mobile LBS applications will be a mix of

- Free services/applications supported by advertising revenue
  - Sponsor links for mobile location-enabled (e.g., nearby point of interest) search
  - Advertising embedded in mobile applications
- Mobile apps requiring premiums for download or subscription
  - Onetime charge to download apps from mobile marketplaces
  - Recurring subscription fees for services/content
  - Add-on charges, e.g., purchase of virtual items in mobile games

Location Data: Shopper

How does location tracking work?

Input

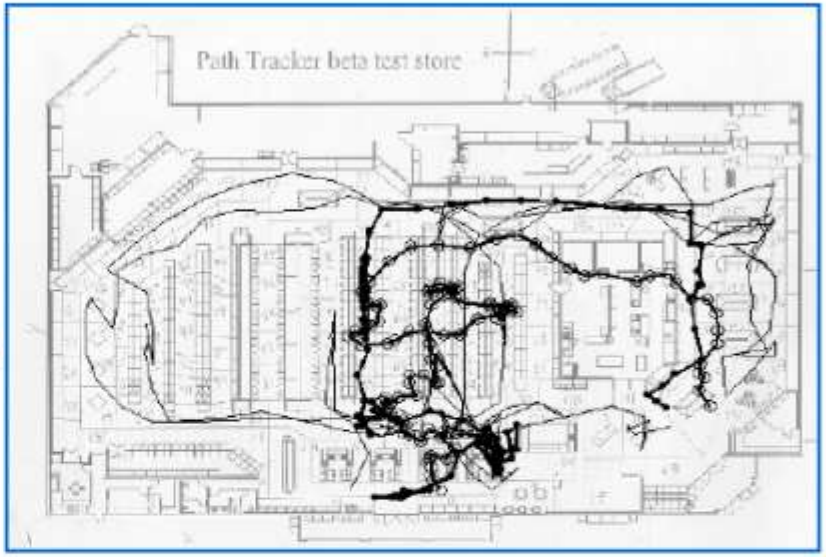
RFID Tag

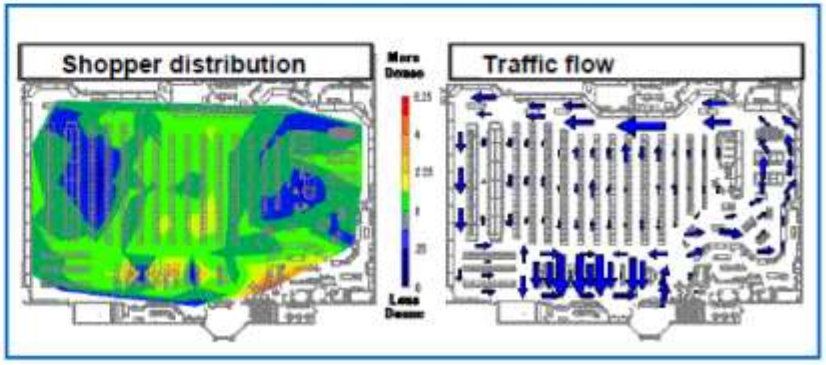
+ Mobile phones

Video

Personal tracking

Output



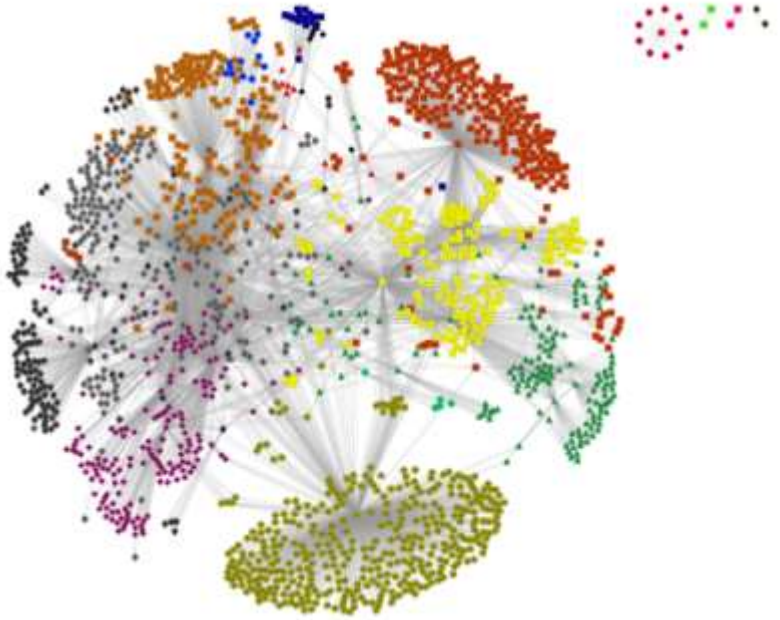


SOURCE: Press and literature search

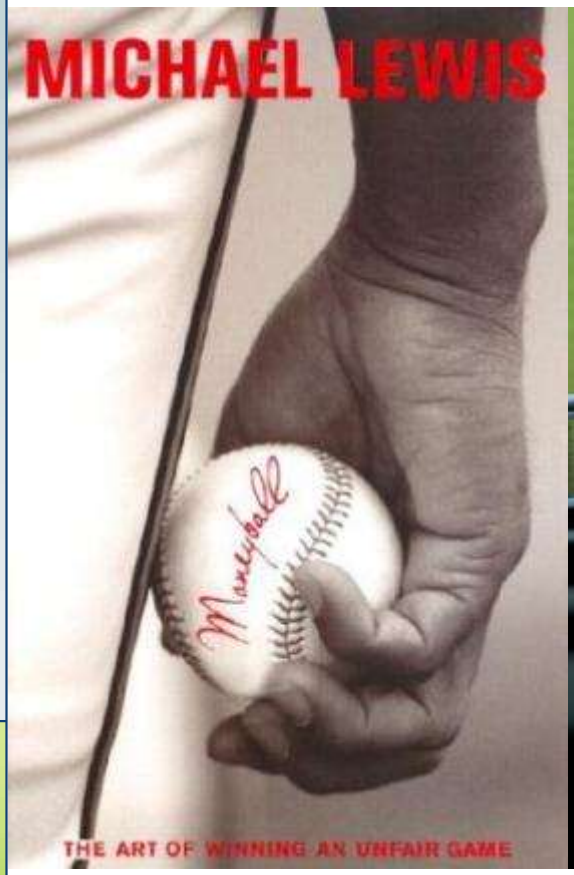
Retail Data: Targeted Marketing

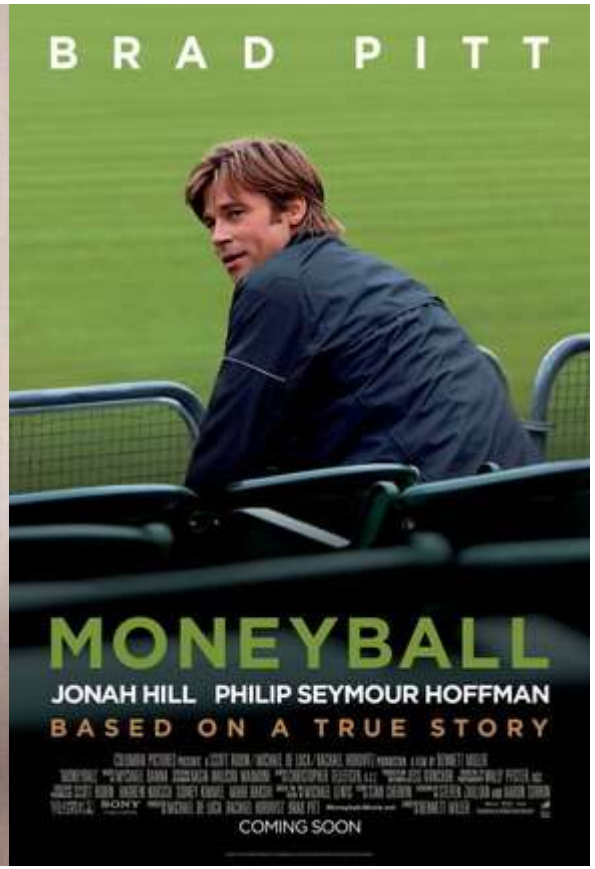
Retail Data: Sentiment Analysis

Social Networks



Sports





MAJOR LEAGUE BASEBALL

OAKLAND A's ATHLETICS ™

Attractiveness Mining

Attractiveness score:
- Dream Girl
- Number of Lights (first)
- Number of Lights (last)
- .....

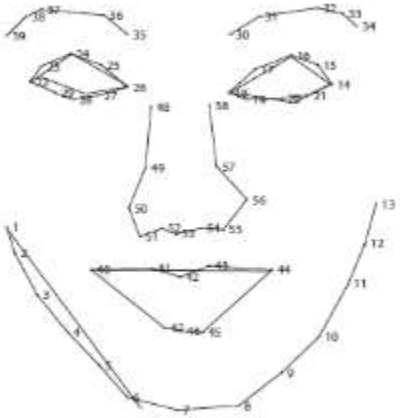
Image

Text

Children:
"I want two kids."
Perspective on Consumption:
"I do not buy luxury goods" .....

Numeric

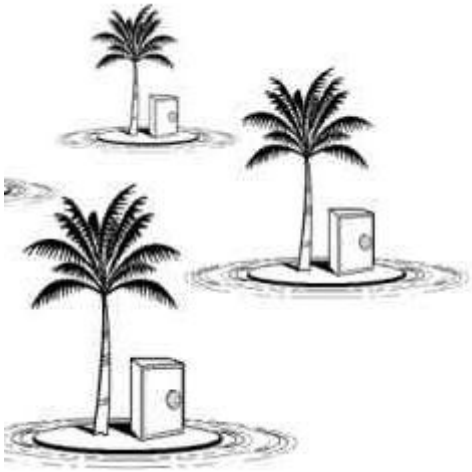
Age	City	Job
23	3	4
25	2	5
28	16	1
.....		

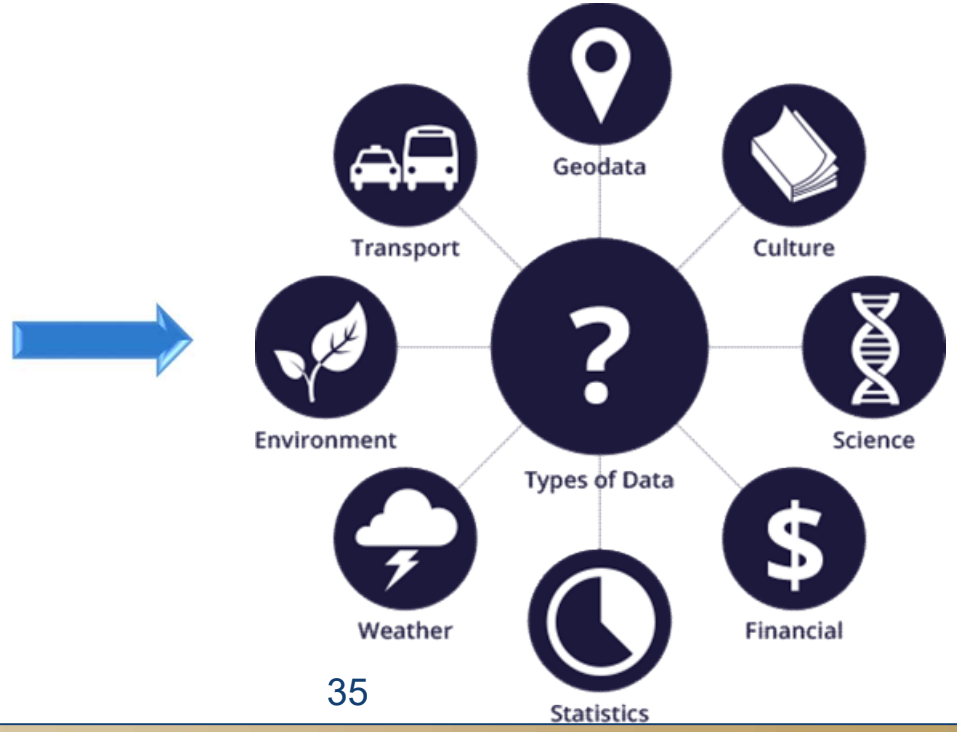


Open Data

- **Technically Open:** available in a machine-readable standard format, which means it can be retrieved and meaningfully processed by a computer application.

- **Legally Open:** explicitly licensed in a way that permits commercial and non-commercial use without restrictions.





Where to find data?

Data repositories

- AWS (Amazon Web Services) Public Data Sets, provides a centralized repository of public data sets that can be seamlessly integrated into AWS cloud-based applications.
- BigML big list of public data sources.
- Bioassay data, described in *Virtual screening of bioassay data*, by Amanda Schierz, J. of Cheminformatics, with 21 Bioassay datasets (Active / Inactive compounds) available for download.
- Bitly 1.usa.gov data, anonymized clicks on gov links.
- Canada Open Data, pilot project with many government and geospatial datasets.
- Causality Workbench data repository.
- Corral Big Data repository at Texas Advanced Computing Center, supporting data-centric science.
- Data Source Handbook, A Guide to Public Data, by Pete Warden, O'Reilly (Jan 2011).
- Datacatalogs.org, open government data from US, EU, Canada, CKAN, and more.
- Data.gov.uk, publicly available data from UK (also London datastore.)
- Data.gov/Education, central guide for education data resources including high-value data sets, data visualization tools, resources for the classroom, applications created from open data and more.
- DataMarket, visualize the world's economy, societies, nature, and industries, with 100 million time series from UN, World Bank, Eurostat and other important data providers.
- Datamob, public data put to good use.
- DataSF.org, a clearinghouse of datasets available from the City & County of San Francisco, CA.
- DataFerrett, a data mining tool that accesses and manipulates TheDataWeb, a collection of many on-line US Goverment datasets.
- Delve, Data for Evaluating Learning in Valid Experiments
- EconData, thousands of economic time series, produced by a number of US Government agencies.
- Enron Email Dataset, data from about 150 users, mostly senior management of Enron.
- Europeana Data, contains open metadata on 20 million texts, images, videos and sounds gathered by Europeana - the trusted and comprehensive resource for European cultural heritage content.
- FEDSTATS, a comprehensive source of US statistics and more
- FIMI repository for frequent itemset mining, implementations and datasets.
- NASDAQ Data Store, provides access to market data.
- National Government Statistical Web Sites, data, reports, statistical yearbooks, press releases, and more from about 70 web sites, including countries from Africa, Europe, Asia, and Latin America.
- National Space Science Data Center (NSSDC), NASA data sets from planetary exploration, space and solar physics, life sciences, astrophysics, and more.
- Open Data Census, assesses the state of open data around the world.
- OpenData from Socrata, access to over 10,000 datasets including business, education, government, and fun.
- Open Source Sports, many sports databases, including Baseball, Football, Basketball, and Hockey.
- Peter Skomoroch dataset Bookmarks
- PubGene(TM) Gene Database and Tools, genomic-related publications database
- Quandl, a collaboratively curated portal to millions of financial and economic time-series datasets.
- qunb, a platform to find and visualize quantitative data.
- Robert Schiller data on housing, stock market, and more from his book *Irrational Exuberance*.
- SMD: Stanford Microarray Database, stores raw and normalized data from microarray experiments.
- Jerry Smith dataset collection, with Finance, Government, Machine Learning, Science, and other data.
- SourceForge.net Research Data, includes historic and status statistics on approximately 100,000 projects and over 1 million registered users' activities at the project management web site.
- StatLib, CMU Datasets Archive.
- STATOO Datasets part 1 and STATOO Datasets part 2
- Time Series Data Library
- Visual Analytics Benchmark Repository.
- **UCI KDD Database Repository** for large datasets used in machine learning and knowledge discovery research.
- UCI Machine Learning Repository.
- UCR Time Series Data Archive, offering datasets, papers, links, and code.
- United States Census Bureau.
- Wikiposit, a (virtual) amalgamation of (mostly financial) data from many different sites, allowing users to merge data from different sources
- Wolfram Alpha disease and patient level dat.
- Yahoo Sandbox datasets, Language, Graph, Ratings, Advertising and Marketing, Competition
- Yelp Academic Dataset, all the data and reviews of the 250 closest businesses for 30 universities for students and academics to explore and research.

Open Government Data

Data Mining

- People have been analysing and investigating data for centuries.
- Statistics
  - Mean, Variance, Correlation, Distribution …
- In modern days, data are often far beyond human comprehension.
  - Diversity, Volume, Dimensionality
- Definition
  - Data Mining is the process of automatically extracting interesting and useful hidden patterns from usually massive, incomplete and noisy data.
- Not a fully automatic process
  - Human interventions are often inevitable.
  - Domain Knowledge
  - Data Collection and Pre-processing
- Synonym: Knowledge Discovery

Is DM really important?

- "If you are looking for a career where your services will be in high demand, you should find something where you provide a scarce, complementary service to something that is getting ubiquitous and cheap. So what's getting ubiquitous and cheap? Data. And what is complementary to data? Analysis. So my recommendation is to take lots of courses about how to manipulate and analyze data: databases, machine learning, econometrics, statistics, visualization, and so on."

An interview with Google Chief Economist
Hal Varian from the New York Times

职位名称

- 空间数据挖掘算法专家-数据平台 新
- 数据挖掘工程师 新
- 资深数据挖掘工程师/数据分析师（创新金融事业群.... 急 新
- 高端职位-高级数据挖掘专家-杭州 新
- 数据挖掘专家/工程师 急 新
- 资深数据挖掘工程师-数据平台 急 新
- 资深数据挖掘及统计算法研发（北京）-数据平台 急 新
- 数据挖掘工程师 新
- 数据挖掘专家-共享业务 新
- 资深数据挖掘工程师 急 新
- 高级数据开发工程师/高级数据挖掘工程师 急
- 数据挖掘专家-聚划算事业部-杭州
- 数据挖掘专家（挖掘应用方向）—天猫—杭州 急
- （集团安全）数据挖掘工程师-安全部-杭州 急
- 资深数据挖掘工程师-北京 急
- 高级数据挖掘工程师 急
- 资深数据挖掘工程师Data Mining Engineer 急
- 数据挖掘专家-商家业务事业部-数据部 急
- 高级数据开发工程师/高级数据挖掘工程师
- 数据挖掘专家 急

高级数据挖掘工程师

生效时间：2013-06-07

结束时间：2014-06-07

招聘人数：2人

工作性质：全职

学历要求：本科

应聘前,需要先创建您的个人简历，如果还没有简历请创建简历

立刻申请职位 加入收藏

所在地区：北京 浏览次数：3735 次

宝贝详情 | 部门介绍 | 应聘流程 | 留言簿

岗位描述：

1.根据数据产品的设计进行数据探索、选择和实现数据挖掘算法、构建数据挖掘模型，对算法和模型进行优化；
2.模型开发，与开发工程师一起完成算法和模型固化和并行等；
3.通过数据探索和模型的输出等发现业务问题，进行深入分析，规划数据挖掘模型和解决方案；

岗位要求：

1.统计学或数学专业，硕士及以上，具有扎实的统计和机器学习知识；
2.3年及以上海量数据挖掘和数据分析经验，能独立构建模型，完成数据分析等工作；
3.熟悉自然语言处理算法，有相关经验；有Deep learning 经验优先；
4.熟练使用c编程，具有丰富的独立实现算法和调优的经验，熟悉python等脚本语言优先；
5.熟练使用R/Matlab/Mahout/，对Hadoop、Map/reduce编程熟悉优先；
6. 对数据敏感，具有良好的逻辑思维能力、理解业务的能力、沟通能力和表达呈现能力；

岗位描述 - Business Intelligence

在这里，你将会接触各种各样的客户行为分析，通过海量的数据分析对客户做出相应的管理决策；

在这里，你将会学习到先进的统计建模方法，开发各类分析方法及模型工具，实现对客户的个性化管理；

在这里，你将会建立起对行业、产品、和客户的深度理解，能够不断开发新的客户经营机会并实施；

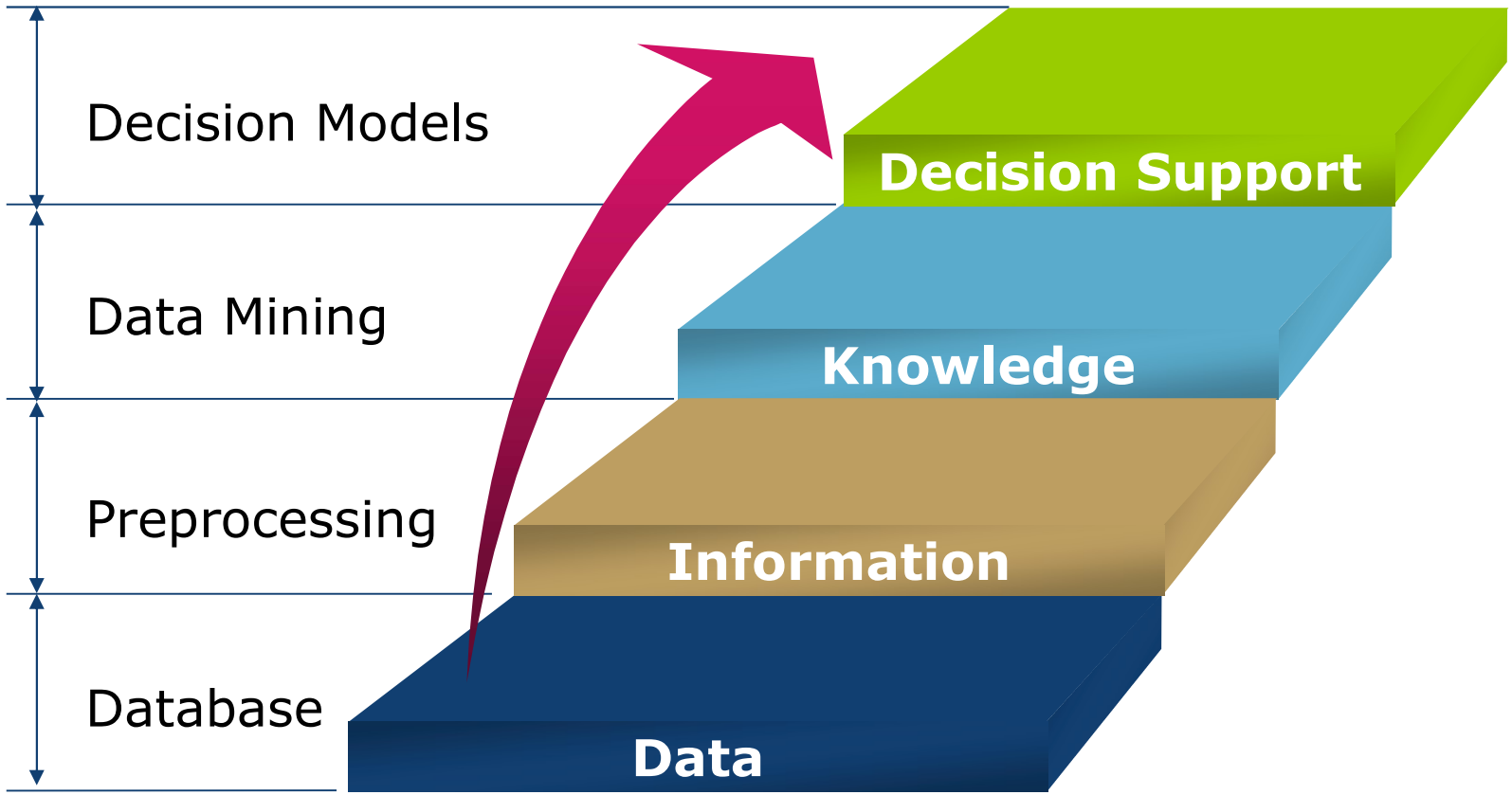
在这里，你将会开发并维护针对各客户群的商业分析，并根据分析报告做出相应决策；

在这里，你将会与运营部门、技术部门保持紧密的合作，完成数据的收集，促进分析决策的实施。

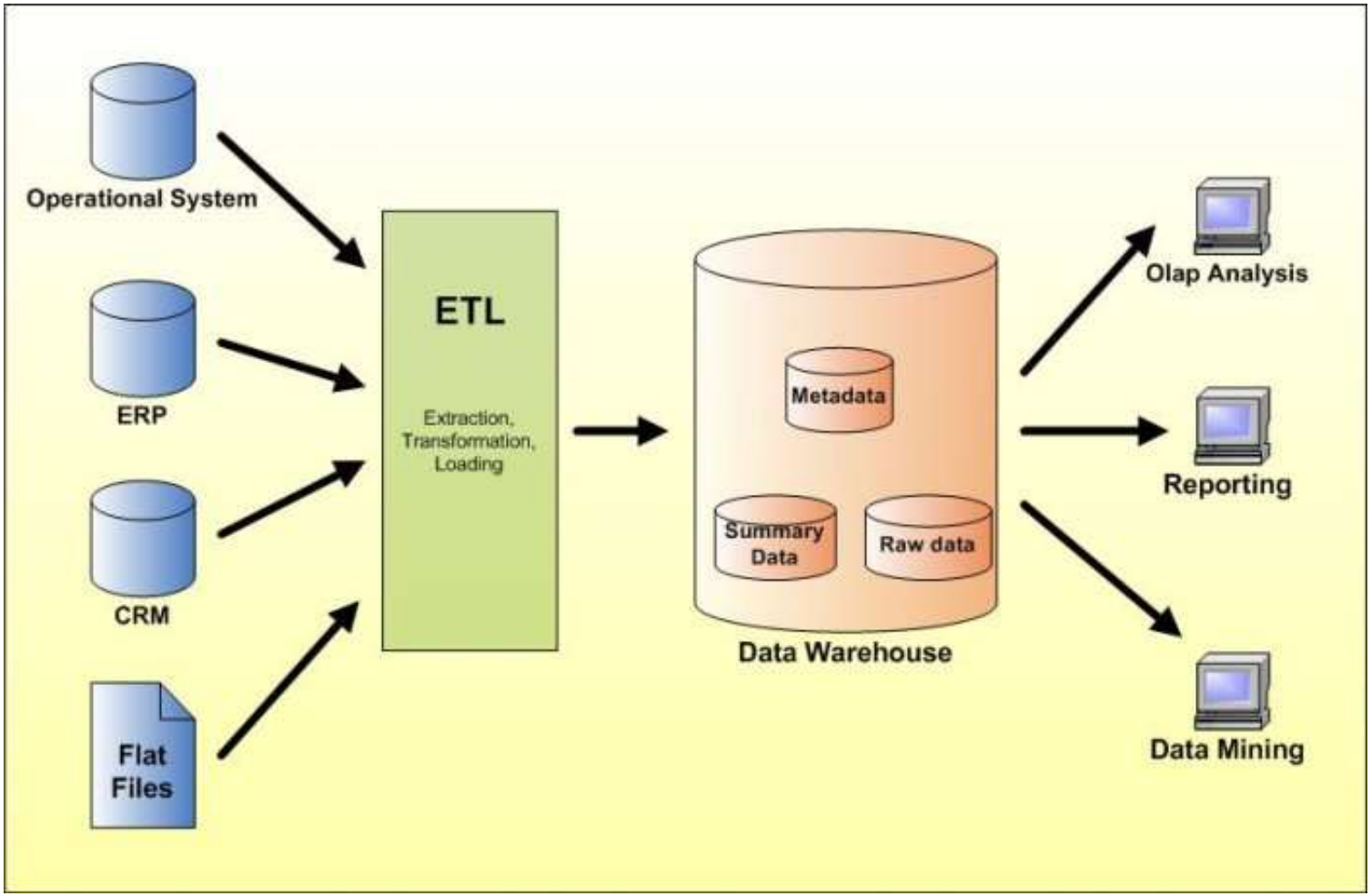
在这里，你将会得获得世界领先的互联网电子商务企业的管理理念，获得各类行业牛人的指导。

Business Intelligence

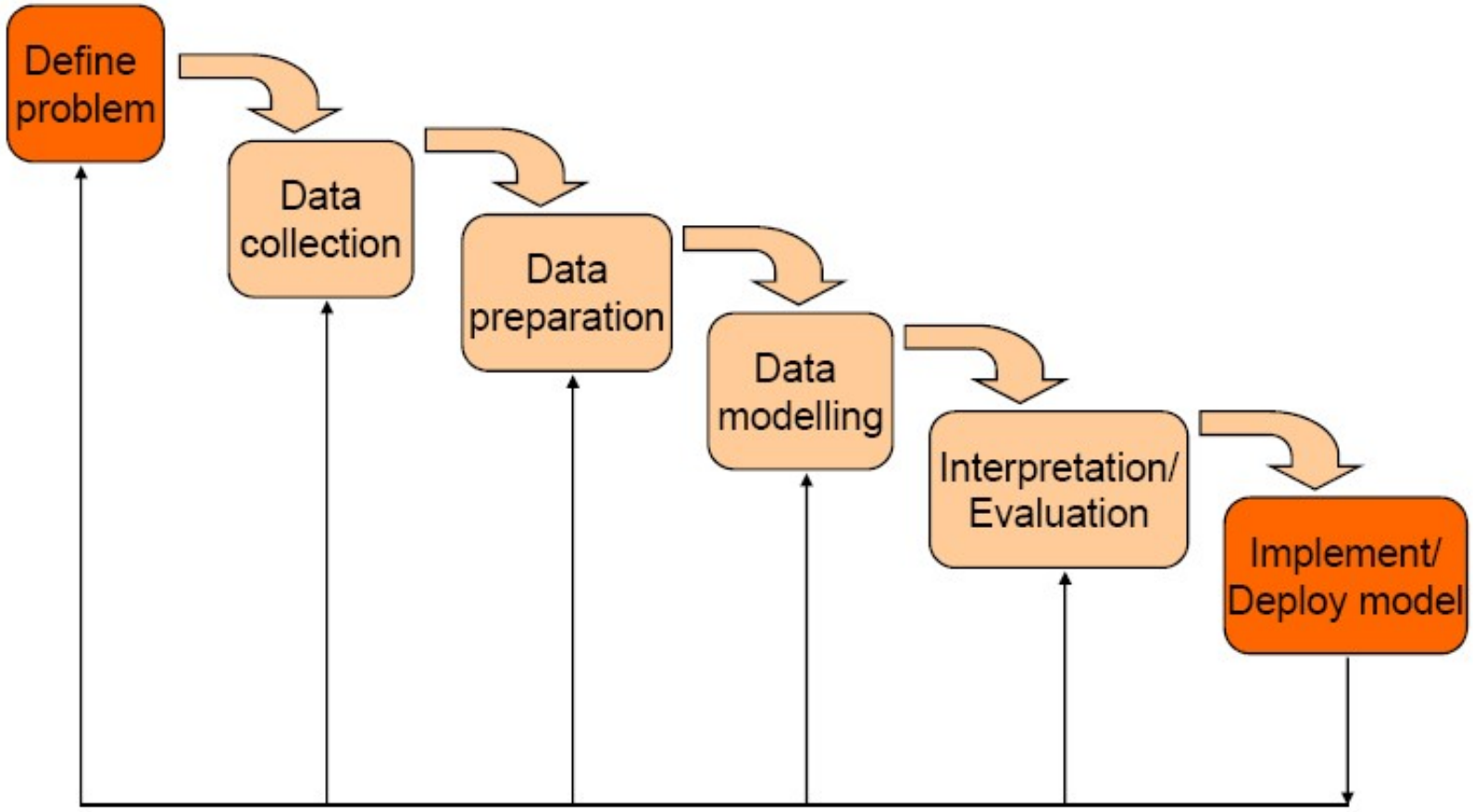
From Data To Intelligence

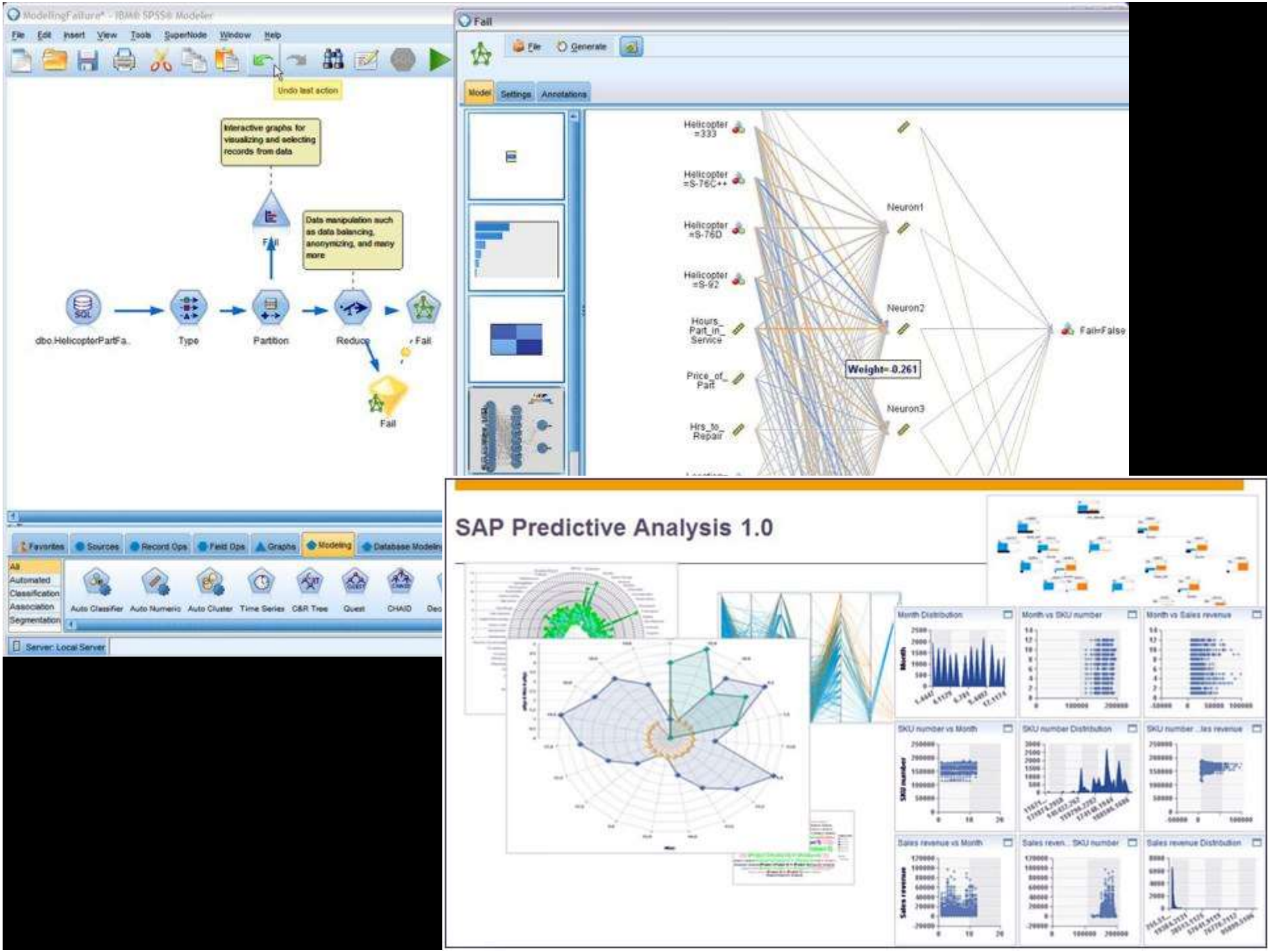


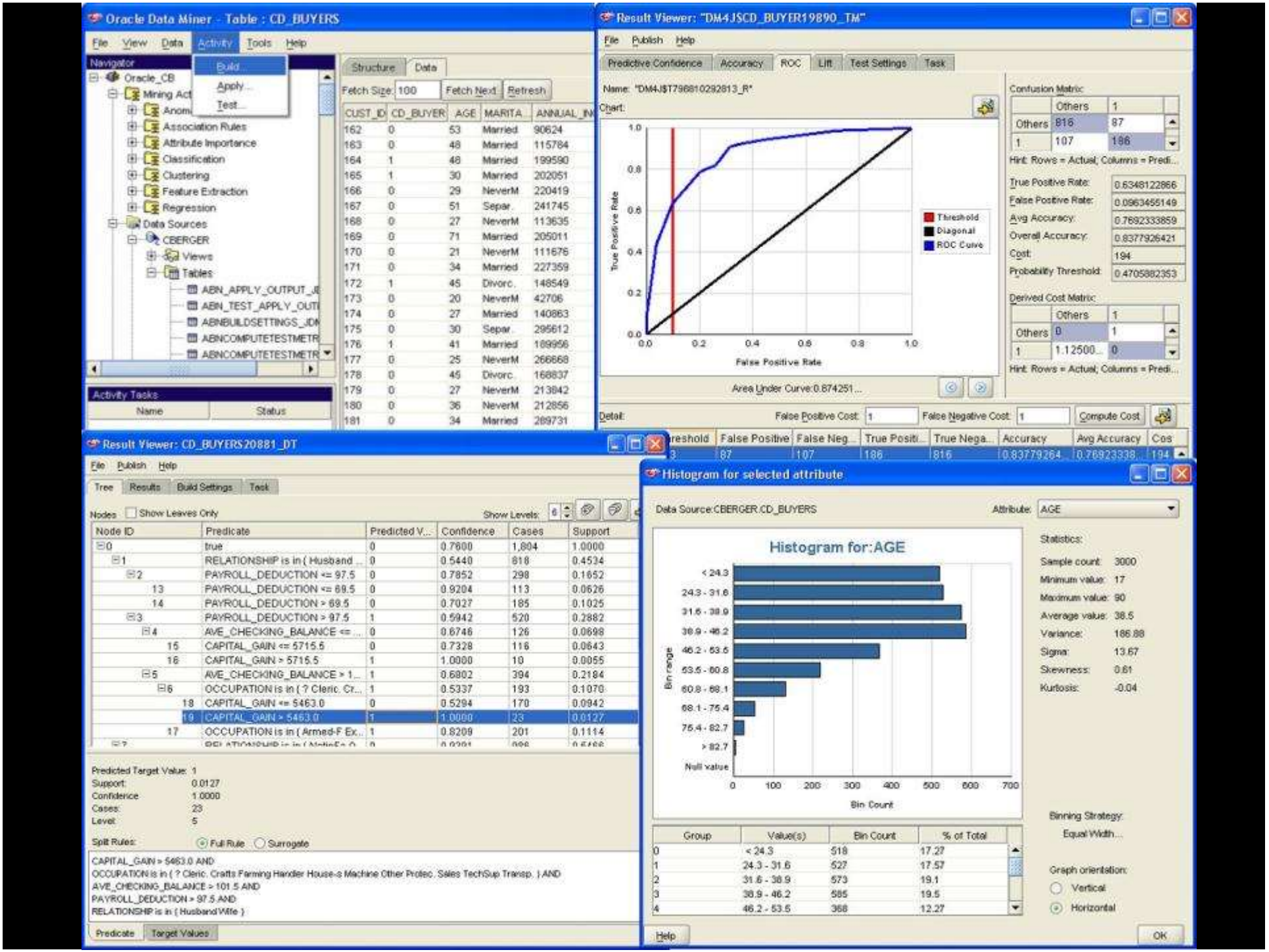
Data Integration & Analysis



The Process of Data Mining




ModelingFailure* - IBM® SPSS® Modeler
File Edit Insert View Tools SuperNode Window Help
Undo last action
Interactive graphs for visualizing and selecting records from data
Data manipulation such as data balancing, anonymizing, and many more
dbo.HelicopterPartFa...
Type
Partition
Reduce
Fail
Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
Database Modeling
All
Automated
Classification
Association
Segmentation
Auto Classifier
Auto Numeric
Auto Cluster
Time Series
C&R Tree
Quest
CHAID
Server: Local Server
Fail
File
Generate
Model
Settings
Annotations
Helicopter =333
Helicopter =S-76C++
Helicopter =S-76D
Helicopter =S-92
Hours_ Part_in_ Service
Price_of_ Part
Hrs_to_ Repair
Neuron1
Neuron2
Neuron3
Fail=False
Weight=-0.261
SAP Predictive Analysis 1.0
Month Distribution
Month vs SKU number
Month vs Sales revenue
SKU number vs Month
SKU number Distribution
SKU number ...les revenue
Sales revenue vs Month
Sales reven... SKU number
Sales revenue Distribution


Oracle Data Miner - Table : CD_BUYERS
File View Data Activity Tools Help
Navigator
Build...
Apply...
Test...
Oracle_CB
Mining Act
Anom
Association Rules
Attribute Importance
Classification
Clustering
Feature Extraction
Regression
Data Sources
CBERGER
Views
Tables
ABN_APPLY_OUTPUT_JI
ABN_TEST_APPLY_OUTI
ABNBUILDSETTINGS_JDM
ABNCOMPUTETESTMETR
ABNCOMPUTETESTMETR
Activity Tasks
Name
Status
Structure
Data
Fetch Size: 100
Fetch Next
Refresh
CUST_ID CD_BUYER AGE MARITA ANNUAL_IN
162 0 53 Married 90624
163 0 48 Married 115784
164 1 48 Married 199590
165 1 30 Married 202051
166 0 29 NeverM 220419
167 0 51 Separ. 241745
168 0 27 NeverM 113635
169 0 71 Married 205011
170 0 21 NeverM 111676
171 0 34 Married 227359
172 1 45 Divorc. 148549
173 0 20 NeverM 42706
174 0 27 Married 140863
175 0 30 Separ. 295612
176 1 41 Married 189956
177 0 25 NeverM 266868
178 0 45 Divorc. 166837
179 0 27 NeverM 213842
180 0 36 NeverM 212856
181 0 34 Married 289731
Result Viewer: "DM4JSCD_BUYER19890_TM"
File Publish Help
Predictive Confidence
Accuracy
ROC
Lift
Test Settings
Task
Name: "DM4J$T796810292813_R"
Chart:
True Positive Rate
False Positive Rate
Threshold
Diagonal
ROC Curve
Area Under Curve:0.874251...
Confusion Matrix:
Others 1
Others 816 87
1 107 186
Hint: Rows = Actual; Columns = Predi...
True Positive Rate: 0.6348122866
False Positive Rate: 0.0963455149
Avg Accuracy: 0.7892333859
Overall Accuracy: 0.8377926421
Cost: 194
Probability Threshold: 0.4705882353
Derived Cost Matrix:
Others 1
Others 0 1
1 1.12500... 0
Hint: Rows = Actual; Columns = Predi...
Detail:
False Positive Cost: 1
False Negative Cost: 1
Compute Cost
reshold False Positive False Neg... True Positi... True Nega... Accuracy Avg Accuracy Cos
3 87 107 186 816 0.83779264 0.76923338... 194
Result Viewer: CD_BUYERS20881_DT
File Publish Help
Tree
Results
Build Settings
Task
Nodes
Show Leaves Only
Show Levels: 6
Node ID Predicate Predicted V... Confidence Cases Support
0 true 0 0.7600 1,804 1.0000
1 RELATIONSHIP is in (Husband ... 0 0.5440 818 0.4534
2 PAYROLL_DEDUCTION <= 97.5 0 0.7852 298 0.1652
13 PAYROLL_DEDUCTION <= 69.5 0 0.9204 113 0.0626
14 PAYROLL_DEDUCTION > 69.5 0 0.7027 185 0.1025
3 PAYROLL_DEDUCTION > 97.5 1 0.5942 520 0.2882
4 AVE_CHECKING_BALANCE <= ... 0 0.6746 126 0.0698
15 CAPITAL_GAIN <= 5715.5 0 0.7328 116 0.0643
16 CAPITAL_GAIN > 5715.5 1 1.0000 10 0.0055
5 AVE_CHECKING_BALANCE > 1... 1 0.6802 394 0.2184
6 OCCUPATION is in (? Cleric. Cr... 1 0.5337 193 0.1070
18 CAPITAL_GAIN <= 5463.0 0 0.5294 170 0.0942
19 CAPITAL_GAIN > 5463.0 1 1.0000 23 0.0127
17 OCCUPATION is in (Armed-F Ex... 1 0.8209 201 0.1114
Predicted Target Value: 1
Support: 0.0127
Confidence 1.0000
Cases: 23
Level: 5
Split Rules:
Full Rule
Surrogate
CAPITAL_GAIN > 5463.0 AND
OCCUPATION is in (? Cleric. Craftsi Farming Handler House-s Machine Other Protec. Sales TechSup Transp.) AND
AVE_CHECKING_BALANCE > 101.5 AND
PAYROLL_DEDUCTION > 97.5 AND
RELATIONSHIP is in (Husband Wife)
Predicate
Target Values
Histogram for selected attribute
Data Source:CBERGER.CD_BUYERS
Attribute: AGE
Histogram for:AGE
Bin range
< 24.3
24.3 - 31.6
31.6 - 38.9
38.9 - 46.2
46.2 - 53.5
53.5 - 60.8
60.8 - 68.1
68.1 - 75.4
75.4 - 82.7
> 82.7
Null value
Bin Count
Statistics:
Sample count: 3000
Minimum value: 17
Maximum value: 90
Average value: 38.5
Variance: 186.88
Sigma: 13.67
Skewness: 0.61
Kurtosis: -0.04
Group Value(s) Bin Count % of Total
0 < 24.3 518 17.27
1 24.3 - 31.6 527 17.57
2 31.6 - 38.9 573 19.1
3 38.9 - 46.2 585 19.5
4 46.2 - 53.5 368 12.27
Binning Strategy:
Equal Width...
Graph orientation:
Vertical
Horizontal
Help
OK

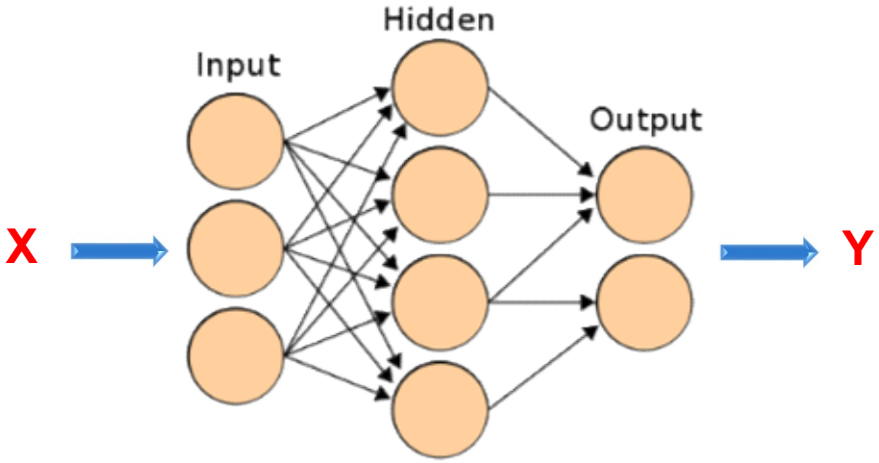
DM Techniques - Classification

- "Classification is a procedure in which individual items are placed into groups based on quantitative information on one or more characteristics (referred to as variables) and based on a training set of previously labeled items."

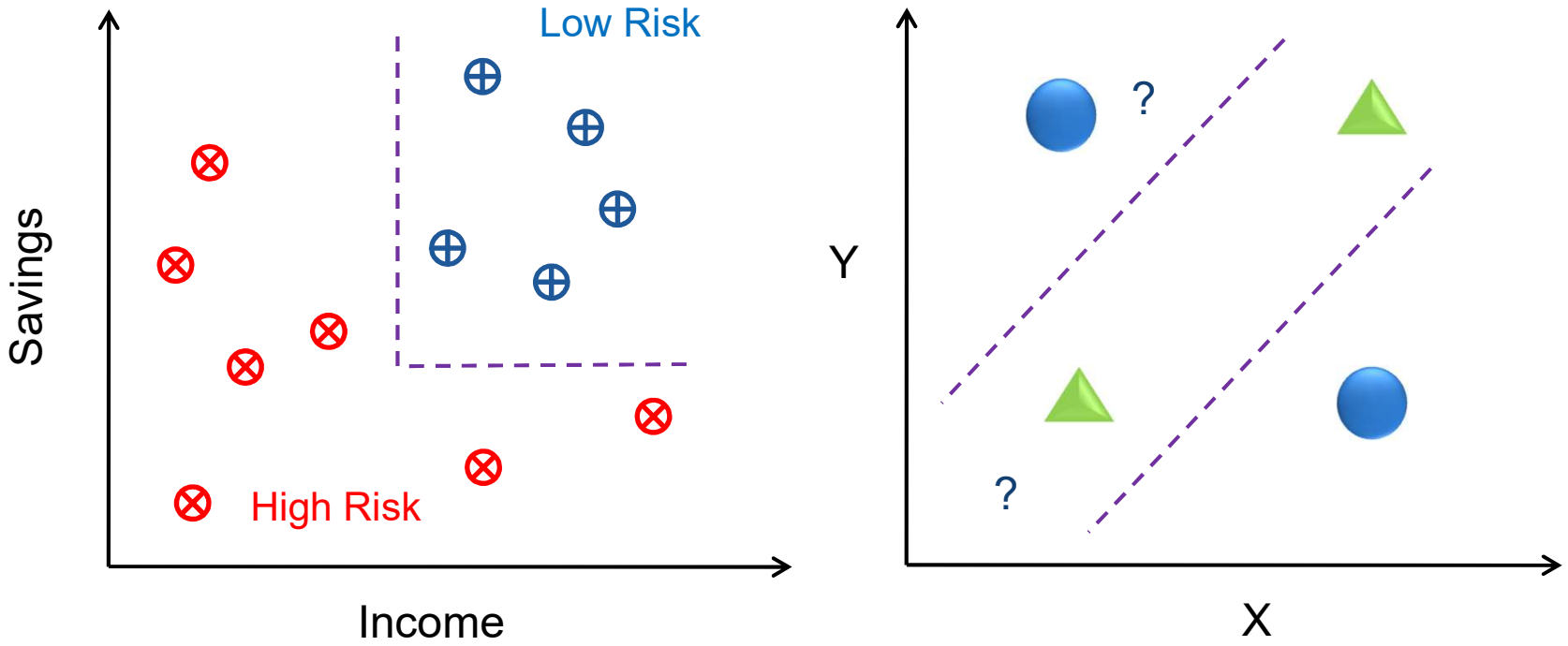
- Given a training set: $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, produce a classifier (function) that maps any unknown object x_i to its class label y_i.

- Algorithms
  - Decision Trees
  - K-Nearest Neighbours
  - Neural Networks
  - Support Vector Machines

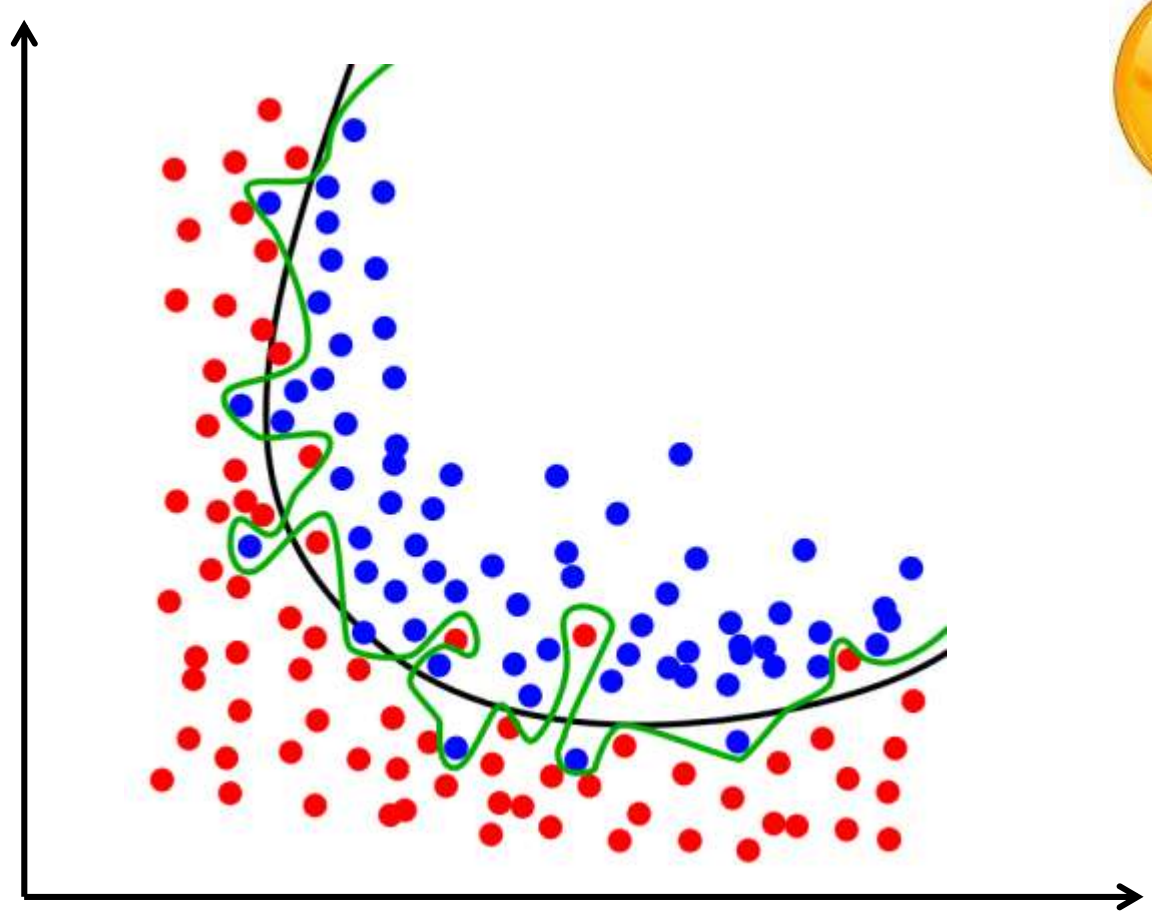
- Applications
  - Churn Prediction
  - Medical Diagnosis



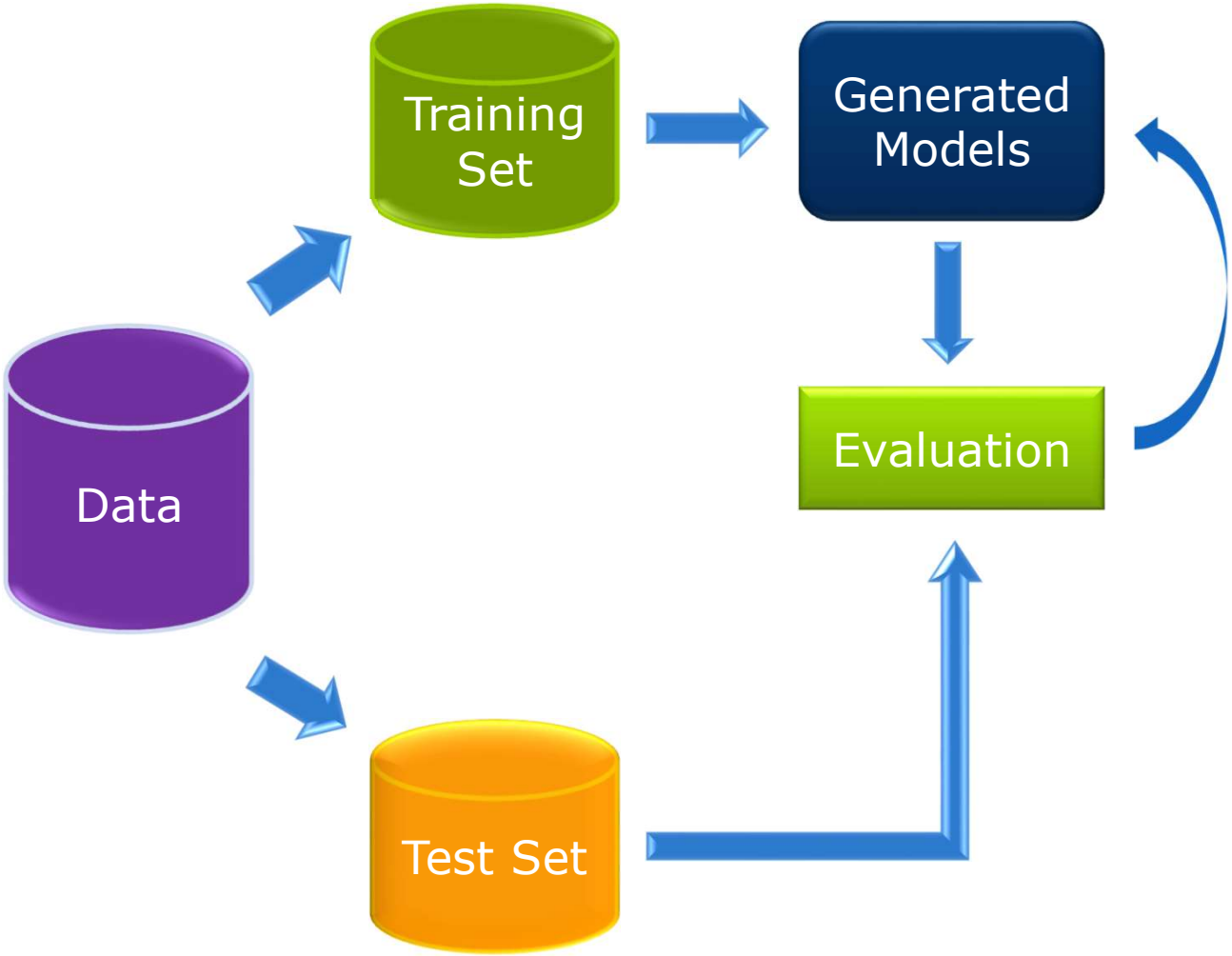
Classification Boundaries



Overfitting – Classification



Cross Validation



Confusion Matrix

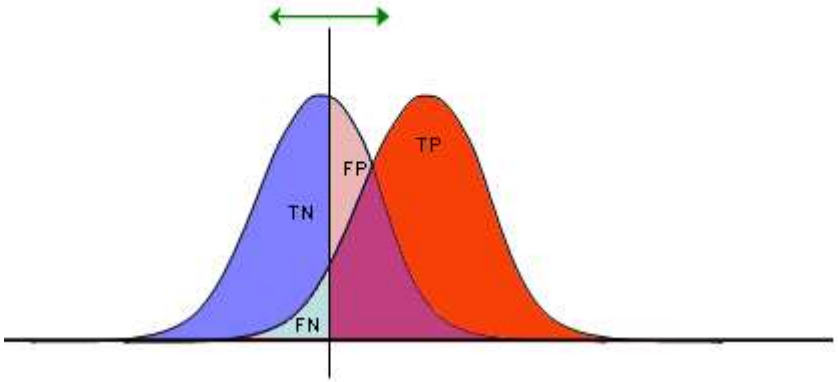
Confusion Matrix				
		Actual Value		
		Positive	Negative	Total
Predicted Value	Positive	True Positive	False Positive	P'
	Negative	False Negative	True Negative	N'
Total		P	N	

TPR=TP/(TP+FN)

TNR=TN/(TN+FP)

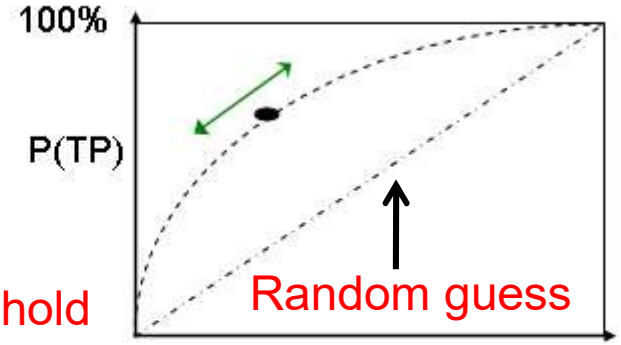
Accuracy=(TP+TN)/(P+N)

Receiver Operating Characteristic



TP	FP
FN	TN
1	1

Very small threshold

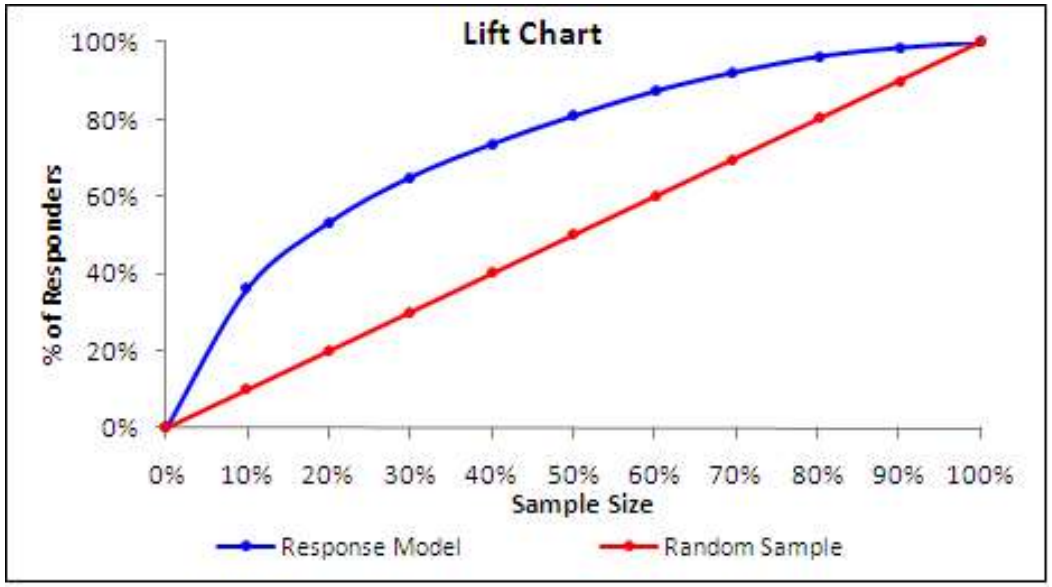


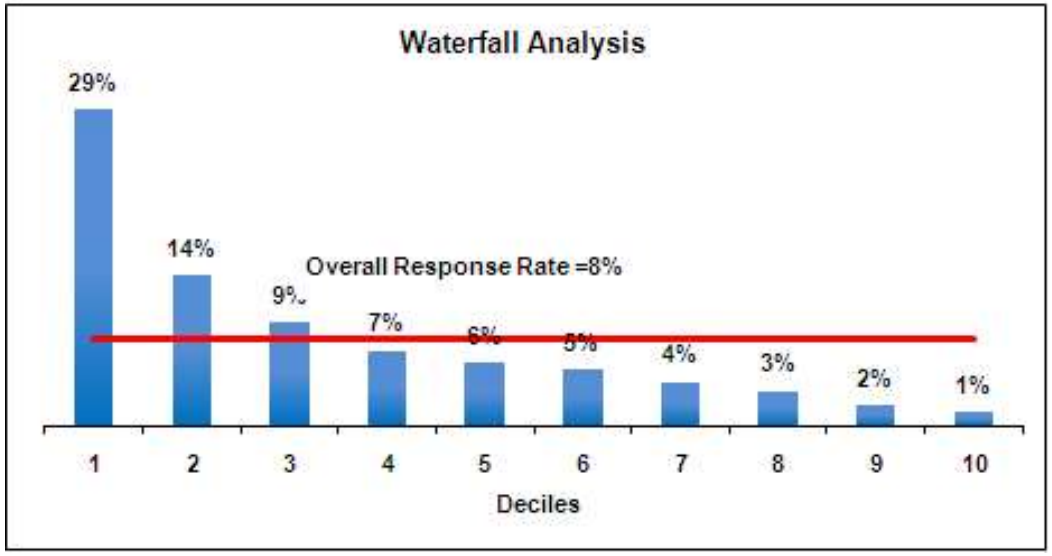
Very large threshold

Random guess

Cost Sensitive Learning

Lift Analysis





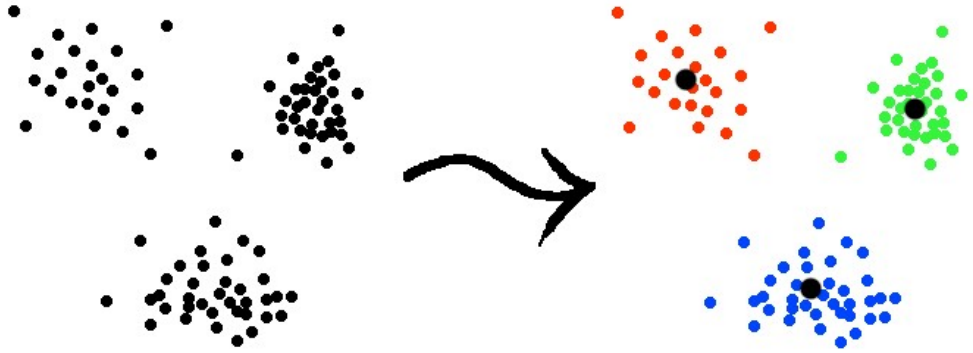
DM Techniques - Clustering

- "Clustering is the assignment of a set of observations into subsets (called *clusters*) so that observations in the same cluster are similar in some sense."

- Distance Metrics
  - Euclidean Distance
  - Manhattan Distance
  - Mahalanobis Distance

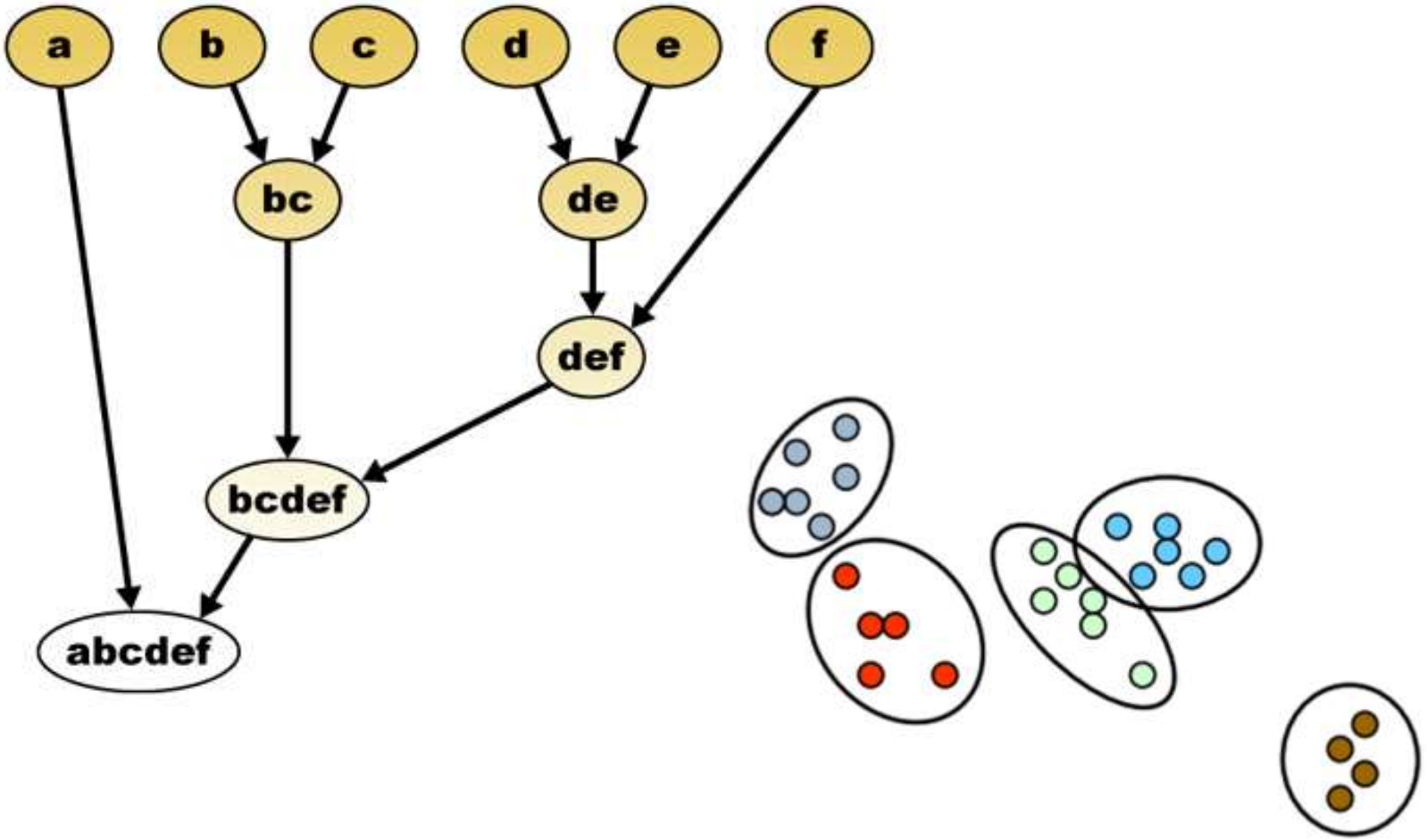
- Algorithms
  - K-Means
  - Sequential Leader
  - Affinity Propagation

- Applications
  - Market Research
  - Image Segmentation
  - Social Network Analysis



What is the difference between classification and clustering?

Hierarchical Clustering



DM Techniques – Association Rule

Association Rule

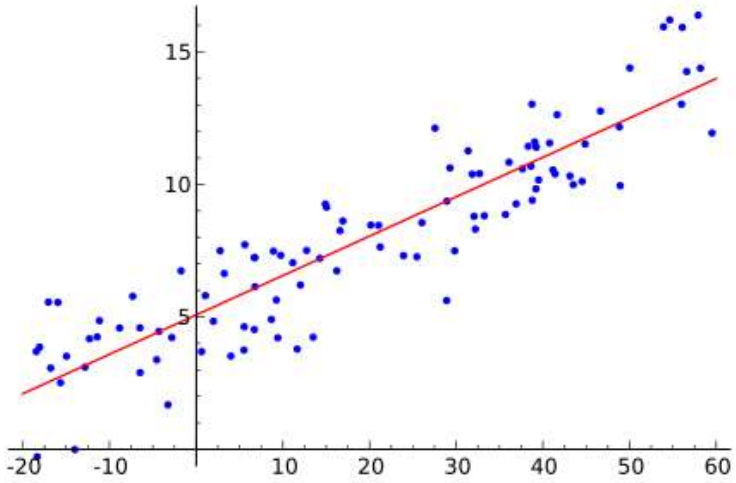
Transaction ID	Milk	Bread	Butter	Beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

$$\{\text{Milk, Bread}\} \Rightarrow \{\text{Butter}\}$$

DM Techniques – Regression

$$Y = f(X, \beta)$$

$$y = \beta_0 + \beta_1 x$$

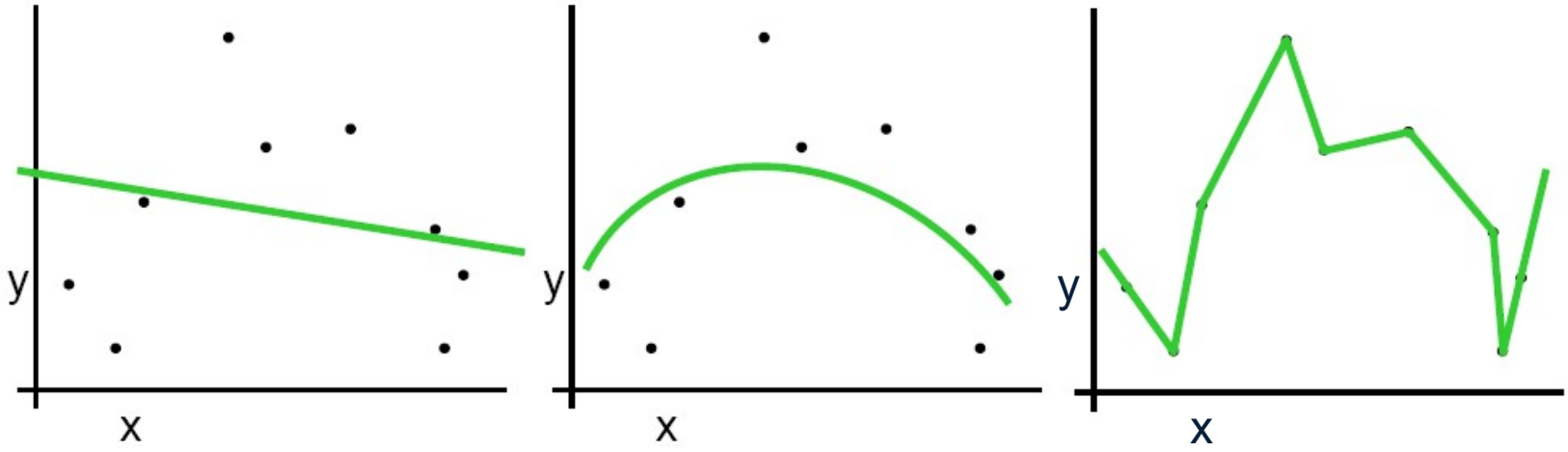


$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

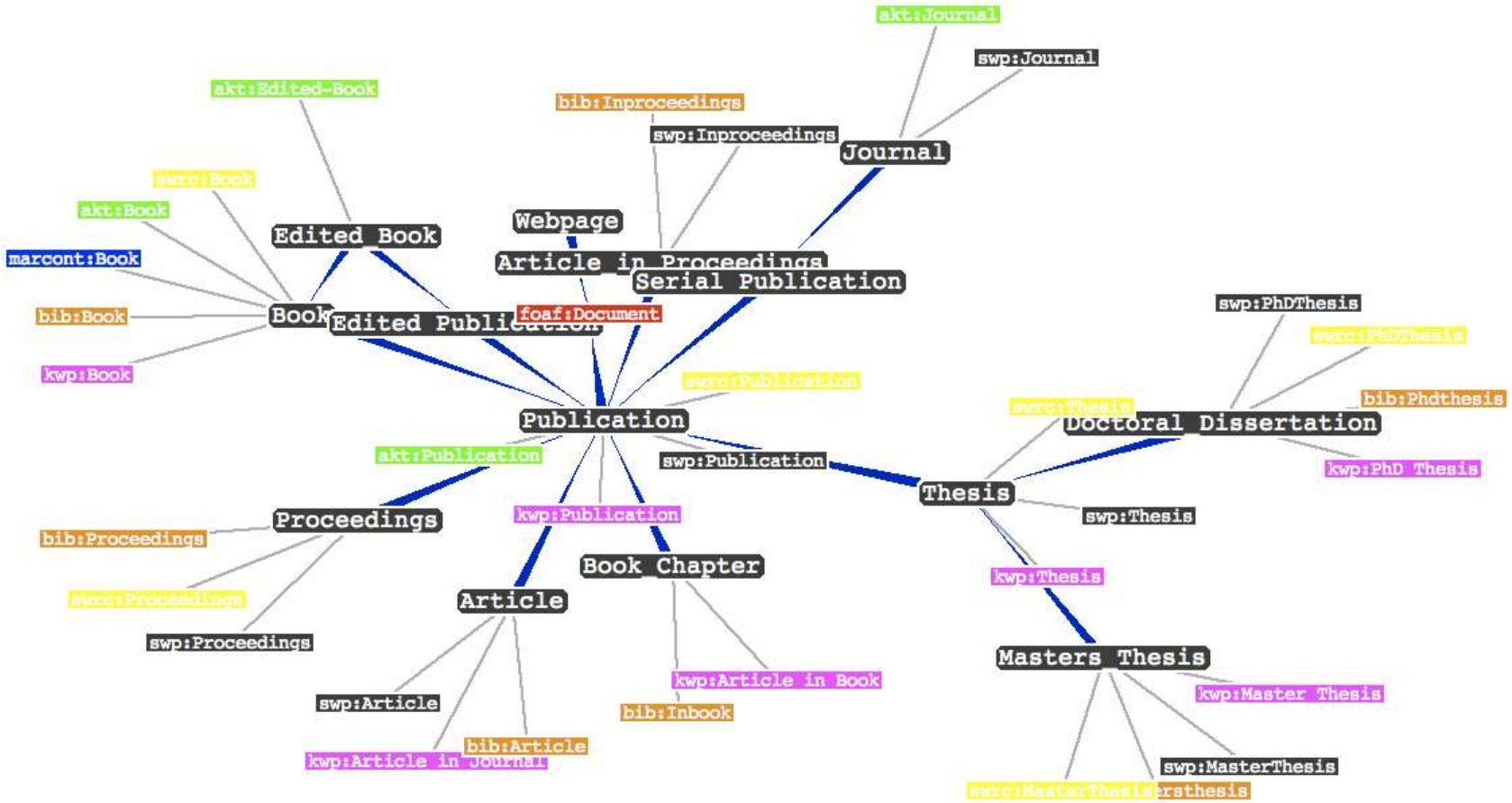
$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

$$y = \frac{1}{1 + e^{-z}}, \qquad z = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Overfitting – Regression



Seeing is Knowing



Performance Dashboard

Arbor.js

A library of force-directed layout algorithms plus abstractions for graph organization and refresh handling.

CartoDB

A web service for mapping, analyzing and building applications with data.

Chroma.js

Interactive color space explorer that allows to preview a set of linear interpolated equidistant colors.

Circos

A software package for visualizing data in a circular layout.

ColorBrewer

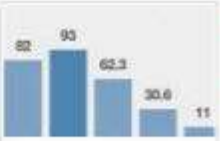
A web tool for selecting colors for maps.

Cubism.js

A library for creating interactive time series and horizon graphs based on D3.js

D3.js

An small, flexible and efficient library to create and manipulate interactive documents based on data.



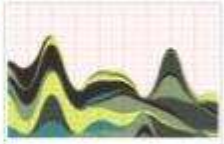
Dance.js

A simple data-driven visualization framework based on Data.js and Underscore.js

Data.js

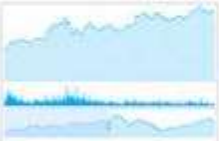
DataWrangler

An interactive web application for data cleaning and transformation.



Degrafa

A powerful declarative graphics framework for rich user interfaces, data visualizations and mapping.



Envision.js

A library for creating fast, dynamic and interactive time series visualizations.

Flare

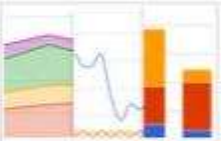
A set of software tools for creating rich interactive data visualizations in ActionScript.

GeoCommons

A public community and set of tools to access, visualize and analyze data with compelling map visualizations.

Gephi

A visualization and exploration platform for networks with dynamic and hierarchical graphs.



Google Chart Tools

A collection of simple to use, customizable and free to use interactive charts and data tools.

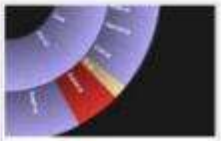
Google Fusion Tables

A web application that makes it easy to host, manage, collaborate on, visualize, and publish data tables.

Google

Impure / Quadrigram

A visual programming language aimed to gather, process and visualize information.

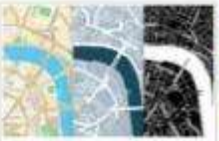


Java Script Infovis Toolkit

A Javascript library that provides tools for creating interactive data visualizations for the web.

Kartograph

A simple and lightweight framework for creating beautiful, interactive vector maps.



Leaflet

A light-weight JavaScript library for making tile-based interactive maps for desktop and mobile browsers.

Many Eyes

A web application to build, share and discuss graphic representation of user uploaded data.

MapBox

A web platform for hosting custom designed map tiles and a set of open source tools to produce them.

Miso Dataset

A client-side data transformation and management library to cleanse, parse, store, query & manipulate data.

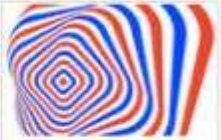
Modest Maps

A display and interaction library for tile-based maps in Flash, JavaScript and Python.

Mr. Da

NodeBox

A desktop application that lets you create generative, static, animated or interactive visuals.



Paper.js

A vector graphics scripting framework in a web designed, consistent and clean programming interface.

Peity

Peity is a simple jQuery plugin that converts an element's content into a simple mini pie, line or bar chart.

Polymaps

A library for making dynamic, interactive maps with image- and vector-based tiles.

Prefuse

A set of software tools for creating rich interactive data visualizations in Java.

Processing

An open source programming language and environment to create images, animations, and interactions.

Processing.js

The sister project of Processing that makes projects work using web standards and without any plug-ins.

Protovis

A library that composes custom views of data with simple marks such as bars and dots.

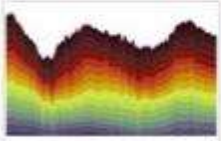
R

Raphaël

A small library that simplifies working with vector graphics on the web.

Reoline.js

A simple but powerful library for building data applications in pure JavaScript and HTML.



Rickshaw

A library for creating interactive time series graphs based on D3.js

Sigma.js

An open-source lightweight library to display interactively static and dynamic graphs

Tableau Public

A desktop application to build and post interactive graphs, dashboards, maps and tables to the web.

Tangle

A library that allow to interactively explore, play, and see the document update immediately.

Timeline

A tool to create timelines with data and media from different sources like Google Docs, Twitter, Flickr or Vimeo.

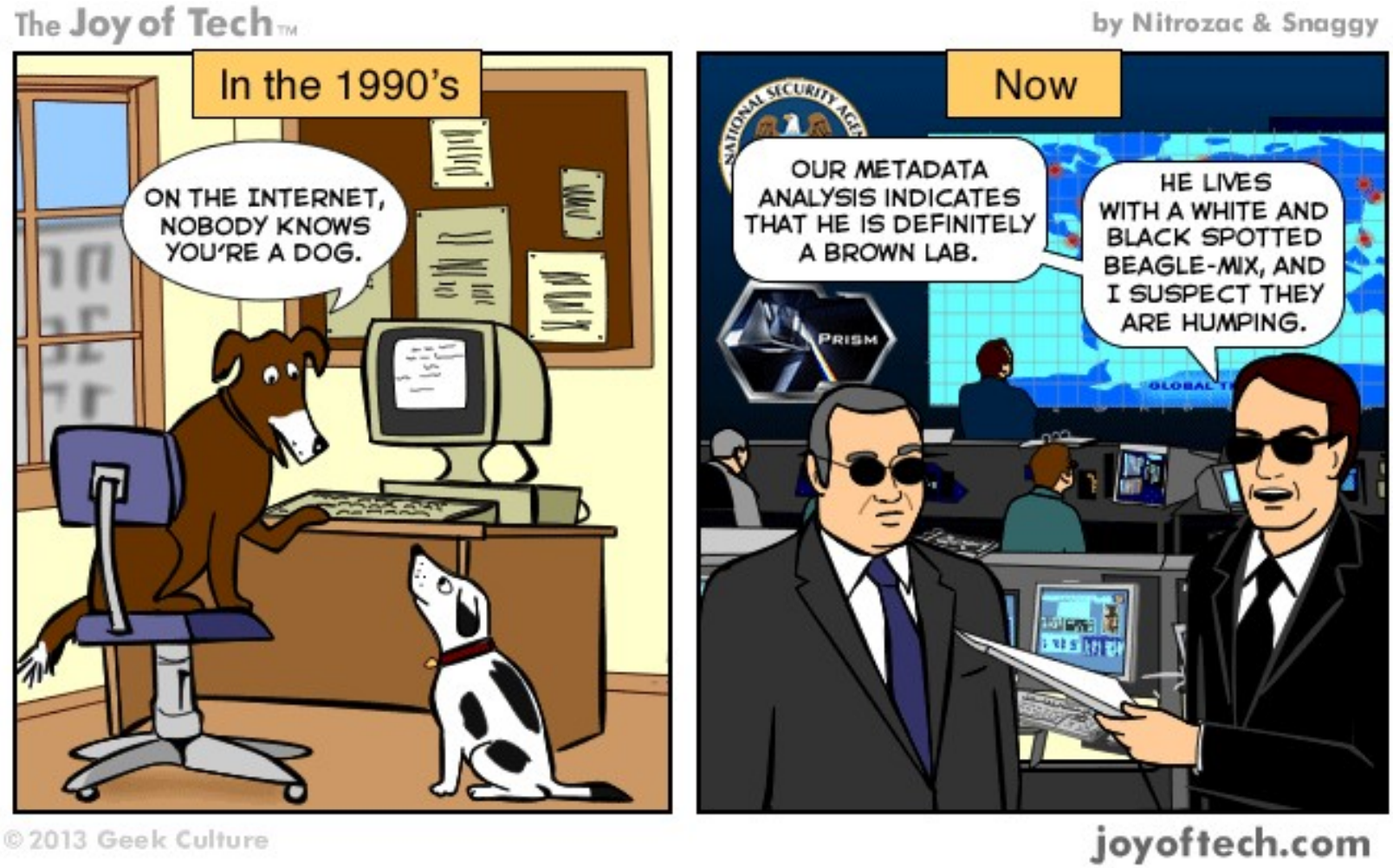
Data Preprocessing

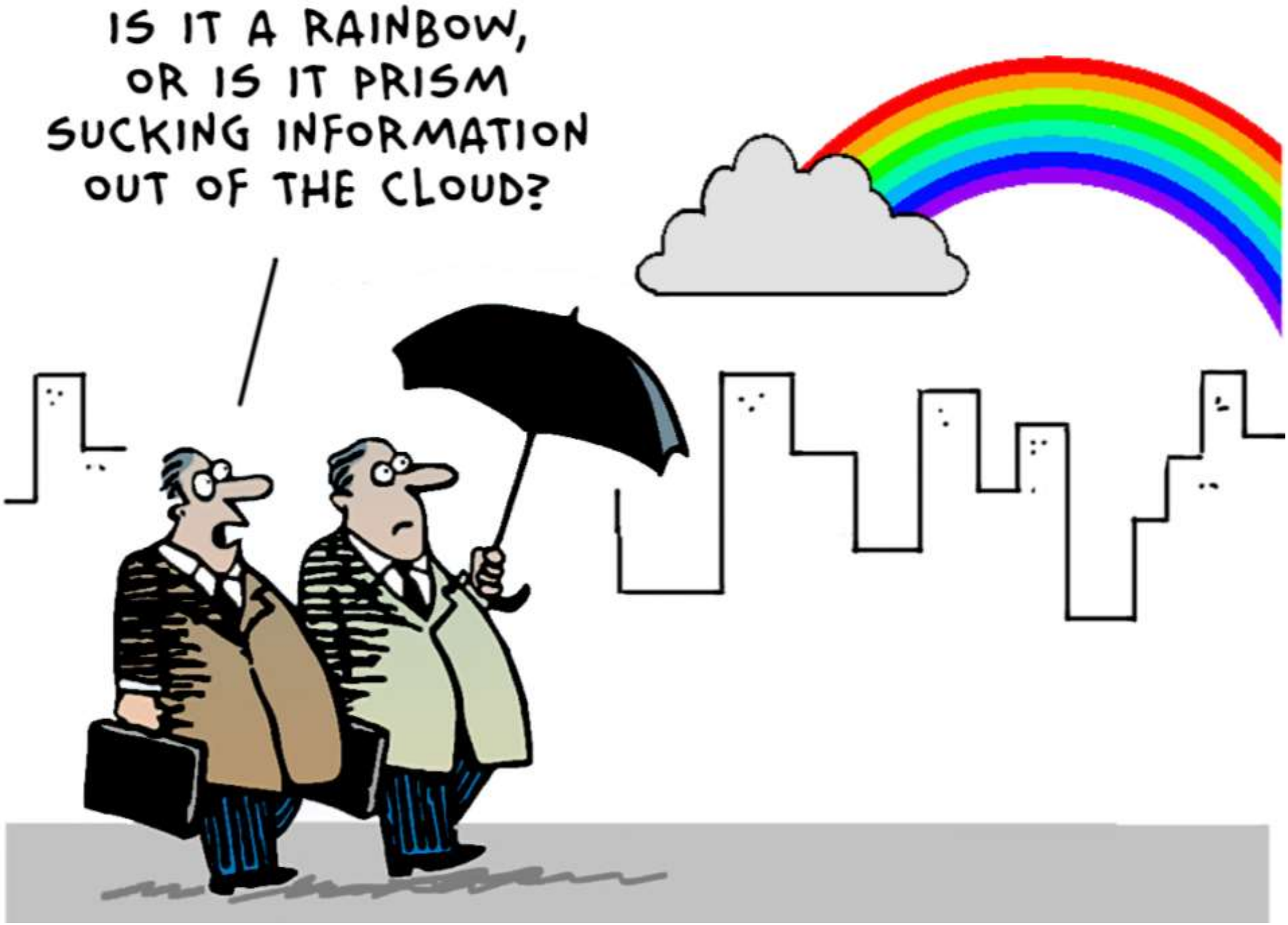
- Real data are often surprisingly dirty.
  - A Major Challenge for Data Mining
- Typical Issues
  - Missing Attribute Values
  - Different Coding/Naming Schemes
  - Infeasible Values
  - Inconsistent Data
  - Outliers
- Data Quality
  - Accuracy
  - Completeness
  - Consistency
  - Interpretability
  - Credibility
  - Timeliness

Data Preprocessing

- Data Cleaning
  - Fill in missing values.
  - Correct inconsistent data.
  - Identify outliers and noisy data.
- Data Integration
  - Combine data from different sources.
- Data Transformation
  - Normalization
  - Aggregation
  - Type Conversion
- Data Reduction
  - Feature Selection
  - Sampling

Internet Privacy




IS IT A RAINBOW, OR IS IT PRISM SUCKING INFORMATION OUT OF THE CLOUD?

Privacy Protection

- Data: A Double-Edged Sword
  - People can benefit greatly from data analysis.
  - The consequence of information leakage can be catastrophic.
- People may be reluctant to give sensitive information due to privacy concerns.
  - Drug, Tax, Sexuality …
- How to find out the percentage of people with a certain attribute?
  - The interviewer should not know the true answer of each respondent.
- Randomized Response
  - Used in structured survey research.
  - Can maintain the confidentiality of respondents.

Privacy Protection

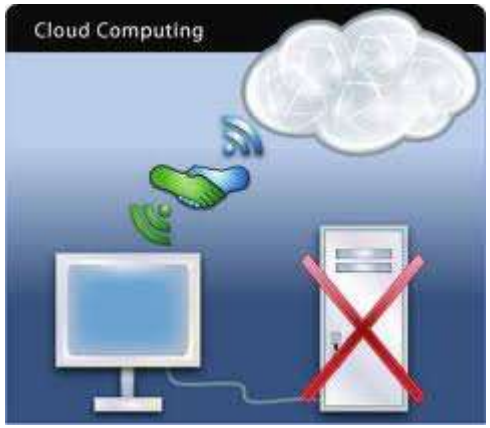
- Two questions are presented:
  - Q1: I have the attribute A.
  - Q2: I do not have the attribute A.

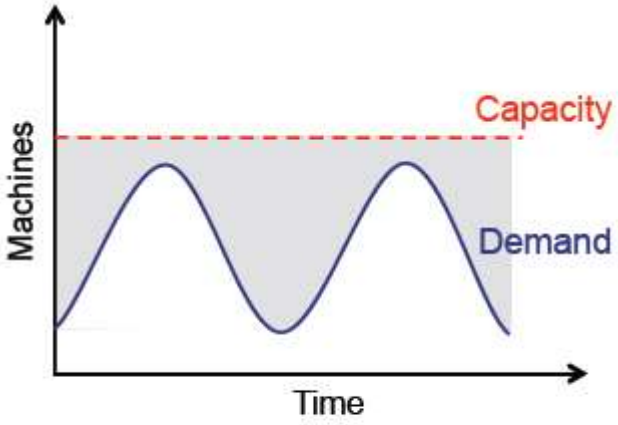
- The respondent uses a random device to:
  - Answer Q1 with probability *p*.
  - Answer Q2 with probability 1-*p*.
  - The interviewer has no idea about which question is answered.

$$P^*(True) = p \times P(True) + (1 - p) \times P(False)$$

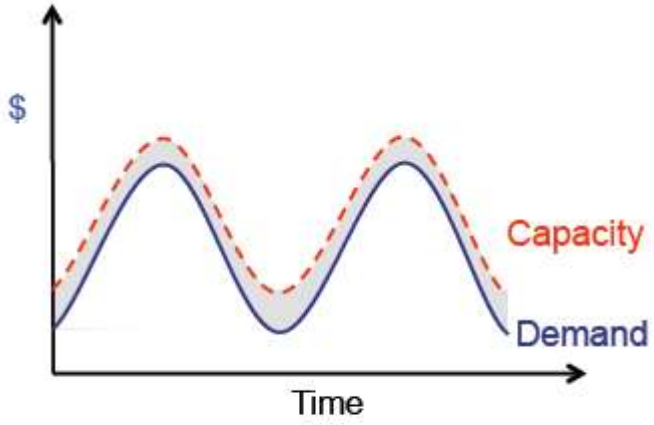
$$P(True) = \frac{P^*(True) + p - 1}{2p - 1} \qquad p \neq 0.5$$

Cloud Computing





"Statically provisioned" data center



"Virtual" data center in the cloud

Cloud Computing

- Pay As You Go

- Software as a Service

Google App Engine

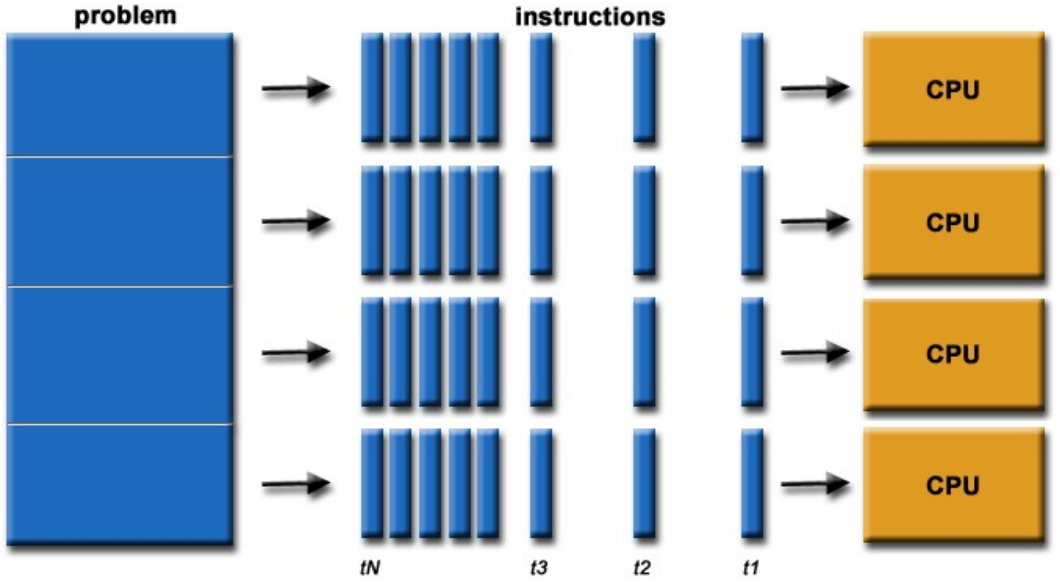
Google Docs

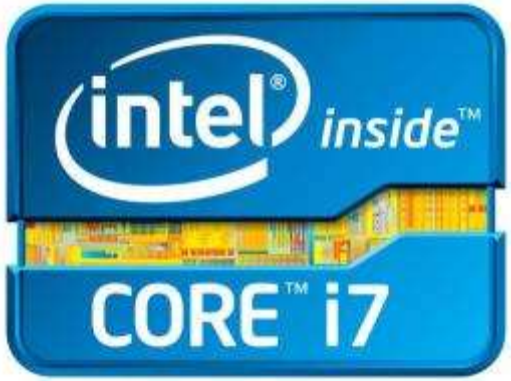
- Platform as a Service

- Infrastructure as a Service

amazon web services™

Parallel Computing

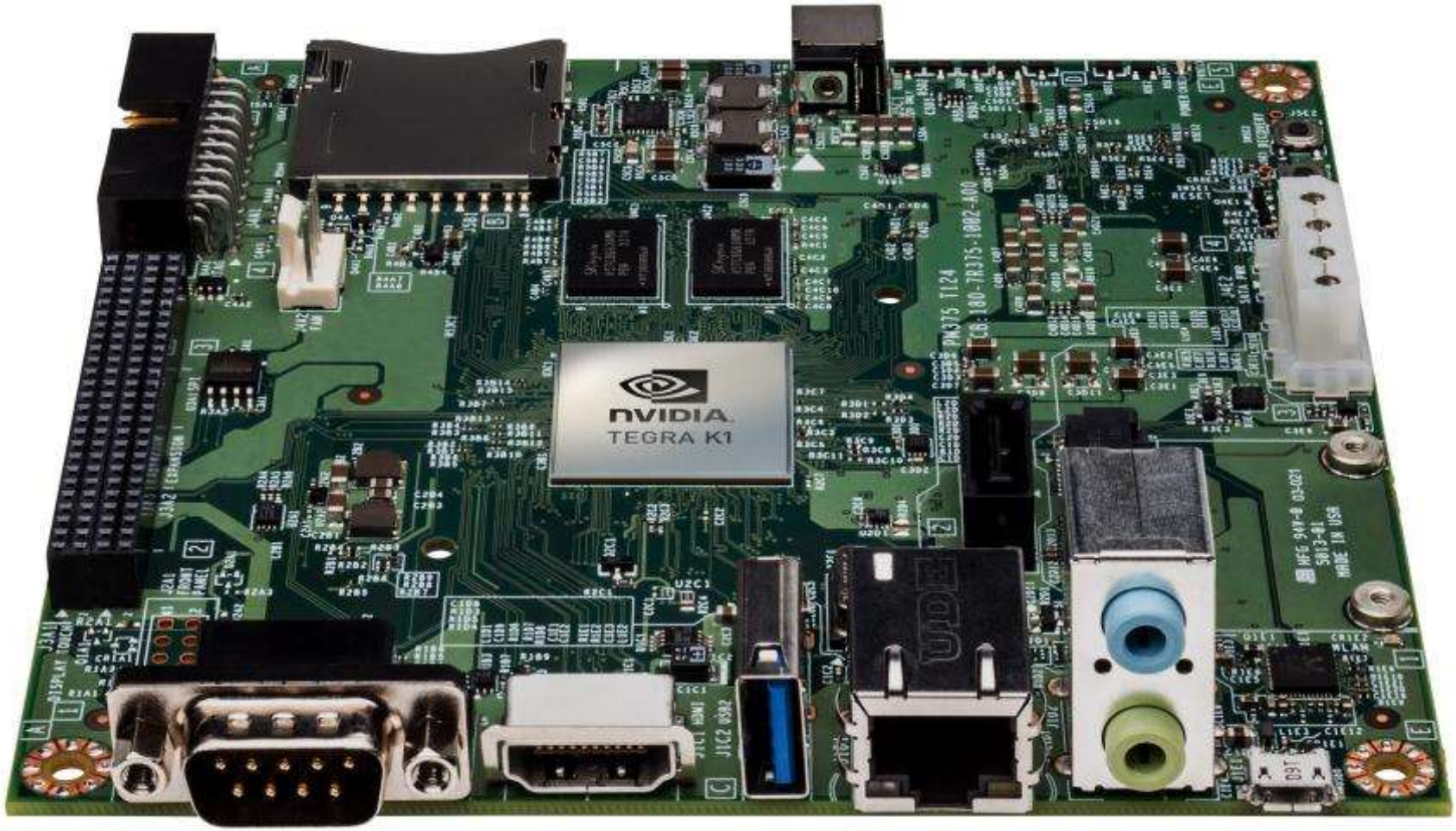




Parallel Computing


GEFORCE GTX
GEFORCE GTX
GEFORCE GTX
GEFORCE GTX
SAMSUNG
SAMSUNG
M2
1500W

Mobile Supercomputing



Intel MIC


intel
inside
XEON

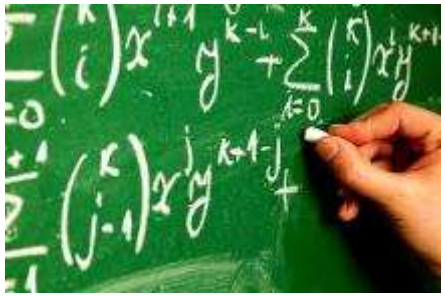
+


intel
inside
XEON PHI


intel
Xeon Phi Coprocessor


intel
inside
Xeon Phi

The Big Picture



No Free Lunch

- Why bother so many different algorithms?
- No algorithm is always superior to others.
- No parameter setting is optimal over all problems.
- Look for the best match between problem and algorithm.
  - Experience
  - Trial and Error
- Factors to consider:
  - Applicability
  - Computational Complexity
  - Interpretability
- Always start with simple ones.

Just in Case Someone Asks …

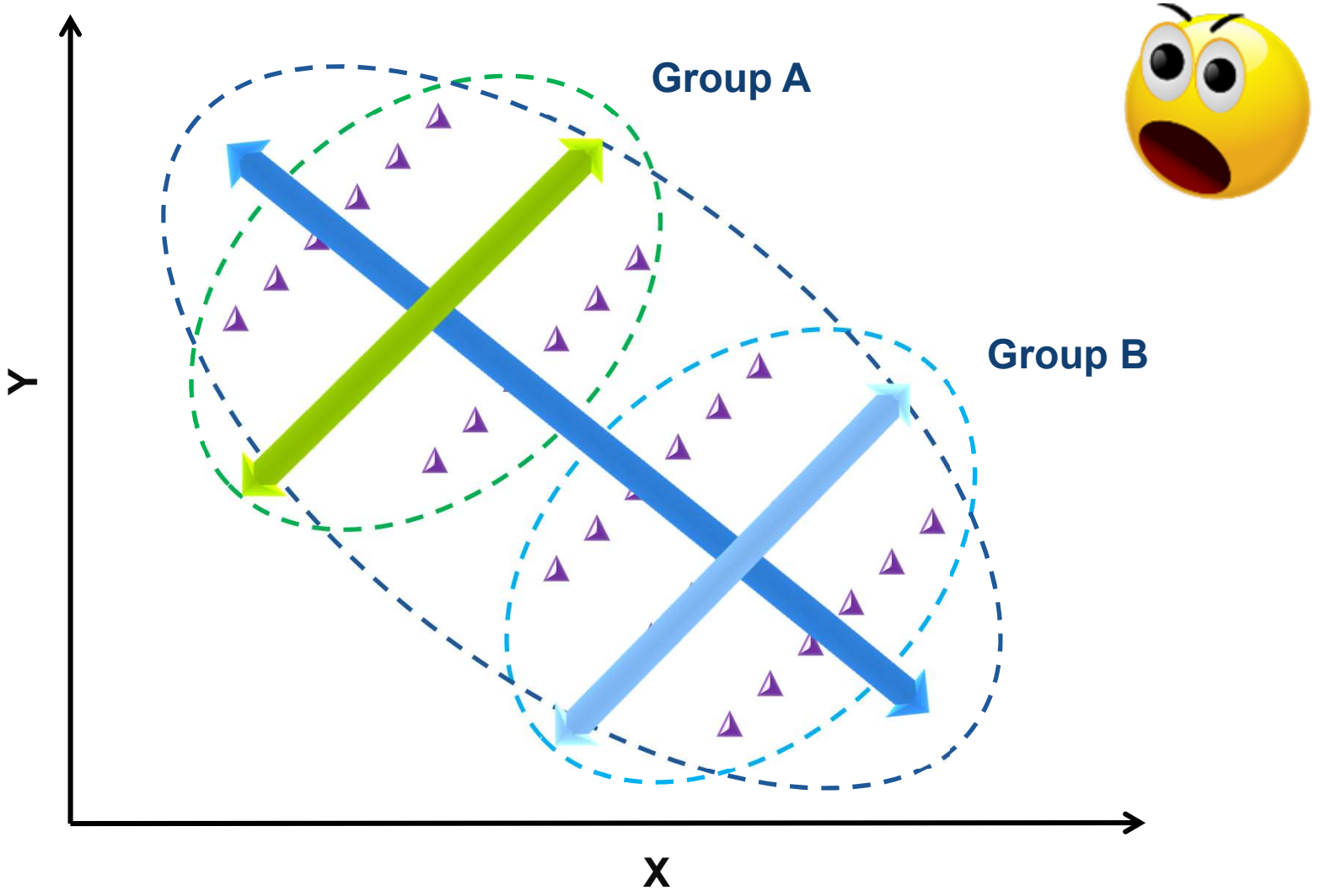
Just in Case Someone Asks ...

号码走势图 | 综合分析表

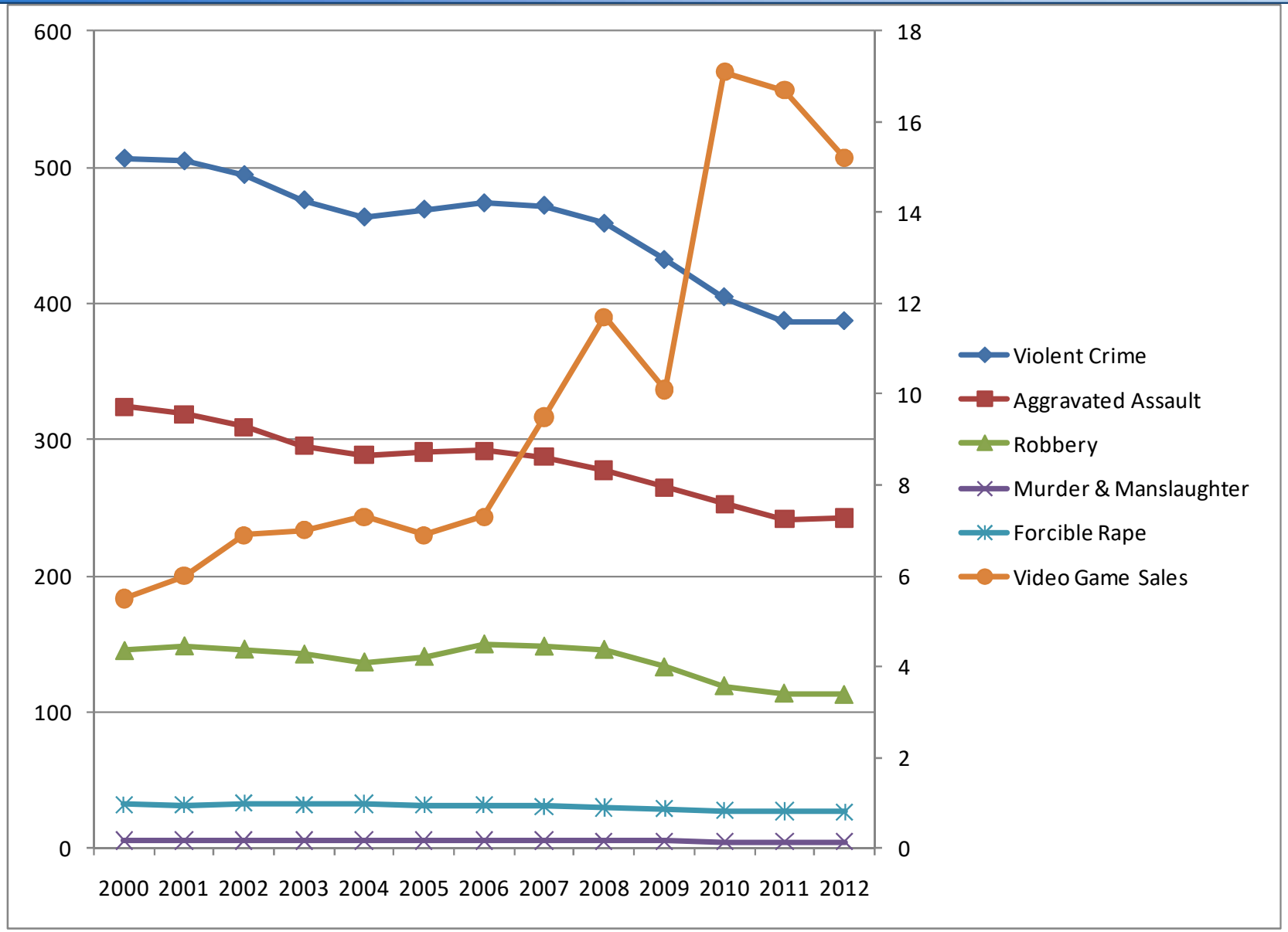
大乐透号码分布图 --请选择期号-- 至 --请选择期号-- 查看

期号	号码	中奖号码分布图																																												
		前区																																			后区											
		01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	01	02	03	04	05	06	07	08	09	10	11	12
09075	01 07 11 32 33 + 09 11	1						7				11																					32	33											9		11	
09076	03 09 11 18 24 + 04 05			3						9		11							18						24															4	5							
09077	07 20 28 32 34 + 06 08							7													20								28				32		34							6		8				
09078	11 23 26 31 34 + 05 11											11												23			26					31			34						5						11	
09079	01 02 07 13 34 + 04 10	1	2					7						13																					34					4						10		
09080	07 16 19 29 35 + 02 11							7									16			19										29						35		2									11	
09081	02 15 20 25 33 + 03 06		2													15					20					25								33					3			6						
09082	05 11 26 29 32 + 08 11					5						11															26			29			32											8			11	
09083	02 05 28 34 35 + 01 07		2			5																							28						34	35	1						7					
09084	05 09 19 26 35 + 02 10					5				9										19							26									35		2								10		
期号	号码	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	01	02	03	04	05	06	07	08	09	10	11	12

Grouping



Violent Crime vs. Video Game



Tricky?

- Is there correlation between height and business success?
- Average American Male is 5’9’’.
- Only 3.9% adult American men are taller than 6’2”.
- Around 30% CEOs of Fortune 500 are taller than 6’2’’.

Survivorship Bias

这是真的吗？

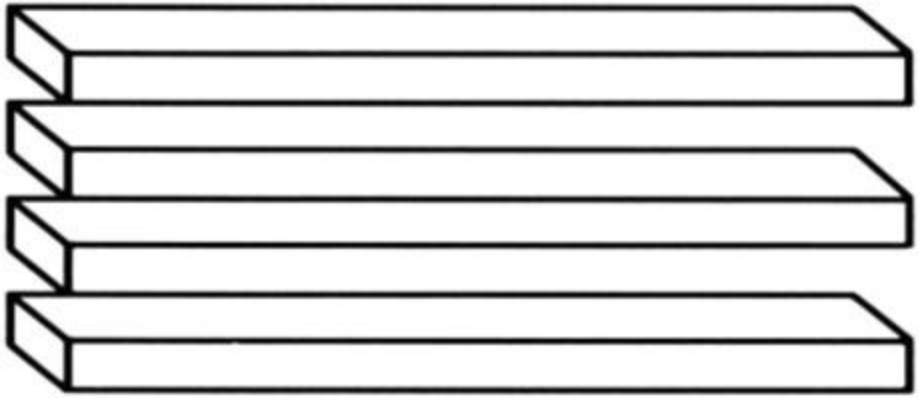
格力：市场起步阶段，竞争不激烈。经营格力品牌的商家，单店交易额随着商家数量的不断增长而增长，r 值呈现正相关关系，说明该品牌在天猫市场竞争中处于起步阶段，尚未白热化。

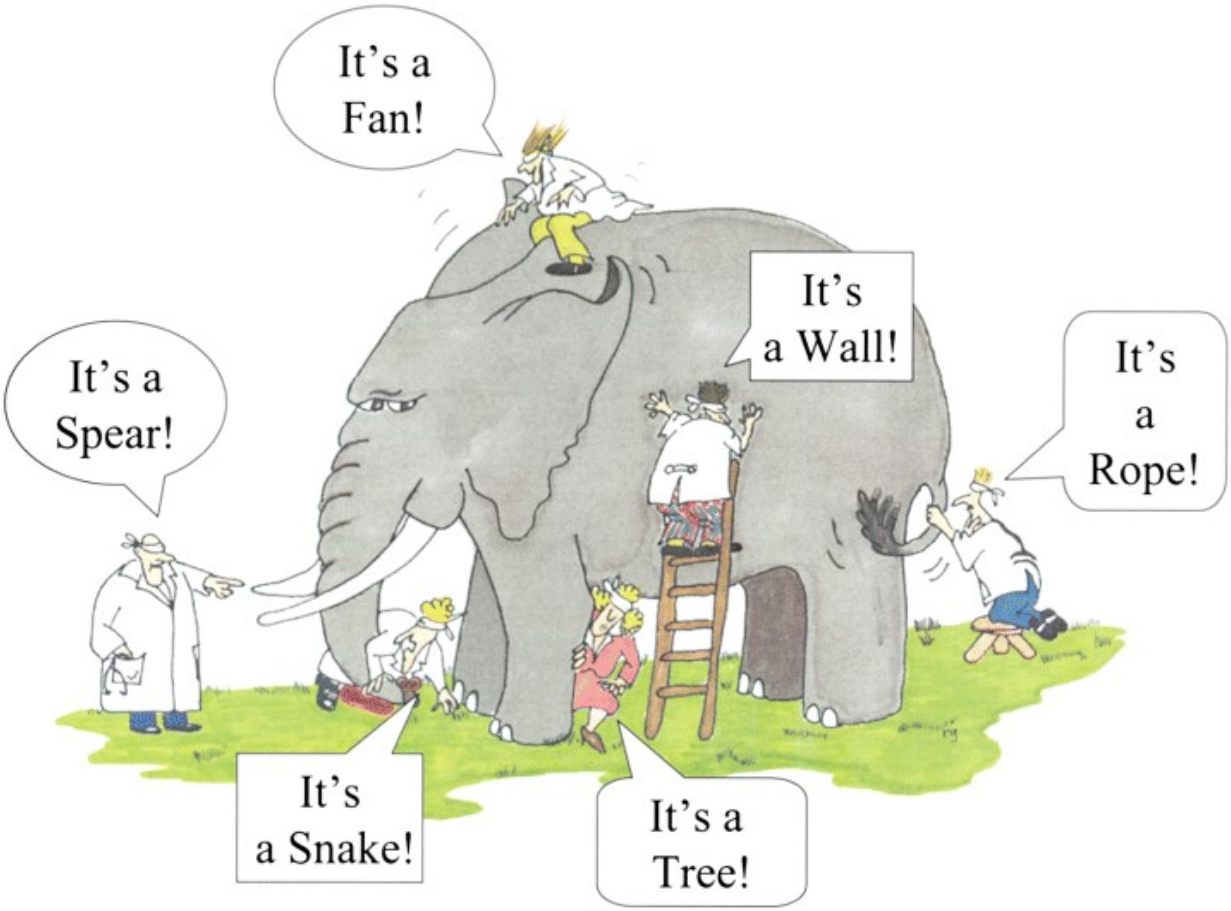
图 1-5　格力品牌商家单店支付宝金额和商家数量的关系

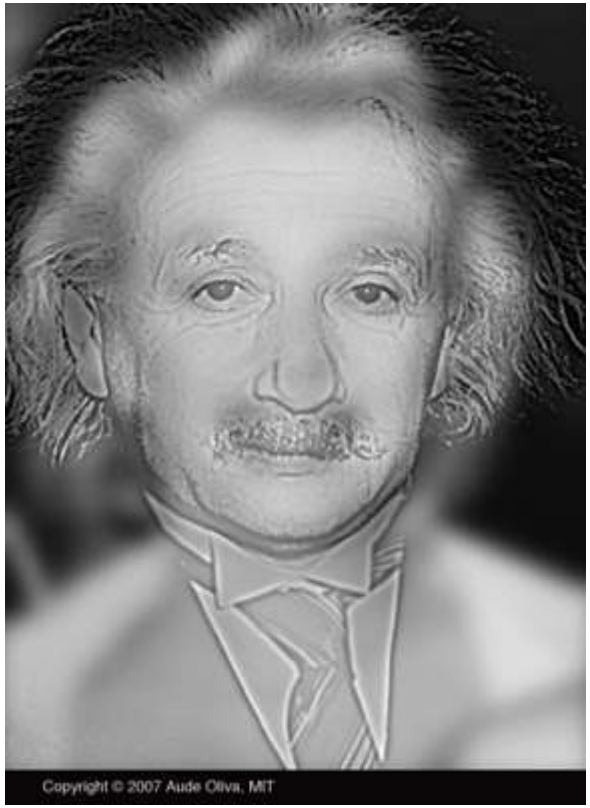
时间去哪儿了？

九阳：很饱和，竞争惨烈。经营九阳品牌的商家，单店交易额随着商家数量的不断增长而下降，r 值呈现负相关关系，说明该品牌在天猫市场中已

图 1-6　九阳品牌商家单店支付宝金额和商家数量的关系




It's a Fan!
It's a Spear!
It's a Wall!
It's a Rope!
It's a Snake!
It's a Tree!


Copyright © 2007 Aude Oliva, MIT