中国科学院大学网络空间安全学院专业普及课

2020-2021学年秋季学期

数据科学导论

The Introduction of Data Science

授课团队：沙瀛　周川

助　　教：梁棋

数据科学导论

The Introduction of Data Science

[第　章] 总结与展望

授课教师：周川

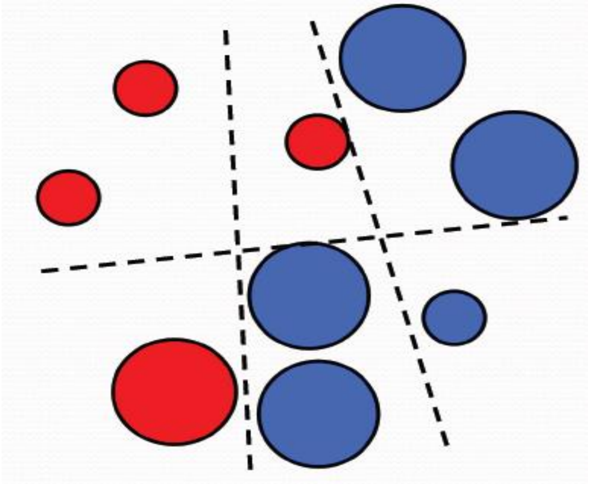
授课时间：2020年12月25日

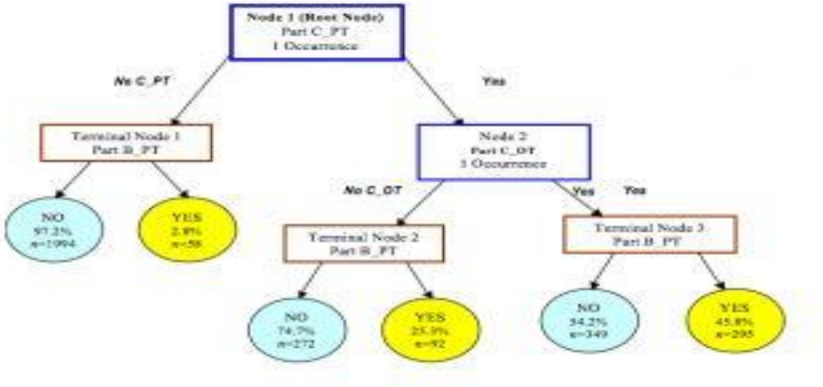
主要章节

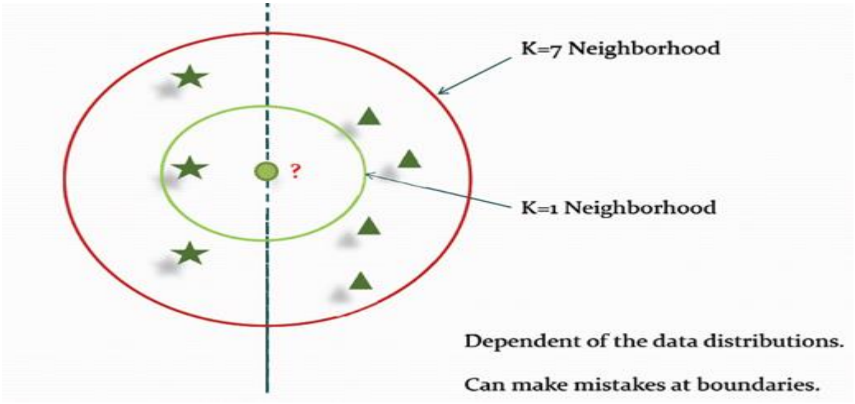
1. 引言
2. 数据科学生命周期
3. 相关与因果-批判性思维
4. 数据预处理
5. 数据分析与建模
6. 数据可视化
7. 项目实施与沟通
8. 分类
9. 聚类
10. 回归
11. 关联规则分析
12. 异常检测
13. 数据降维
14. 时间序列分析

第8章 分类

- 决策树算法
- 朴素贝叶斯分类器
- 最近邻分类器
- Logistics回归
- 提升方法（集成学习）

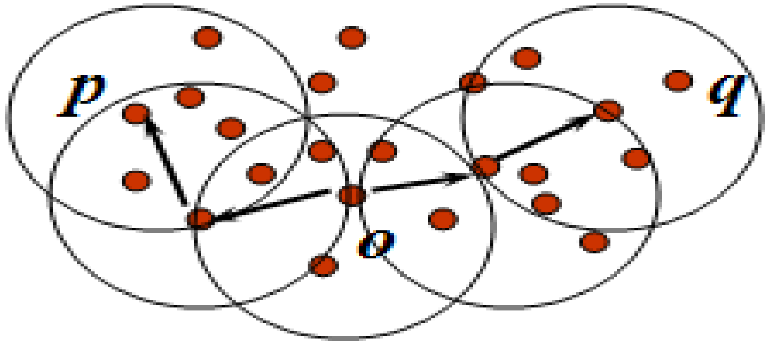


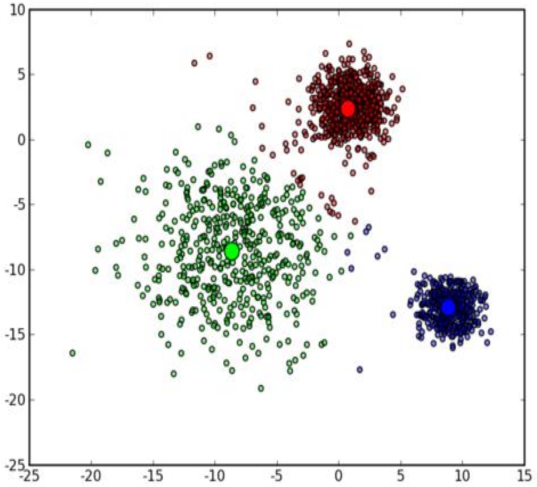


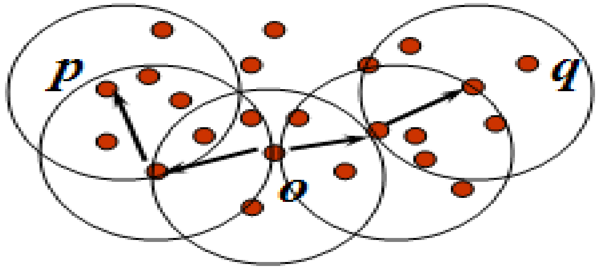

?
K=7 Neighborhood
K=1 Neighborhood
Dependent of the data distributions.
Can make mistakes at boundaries.

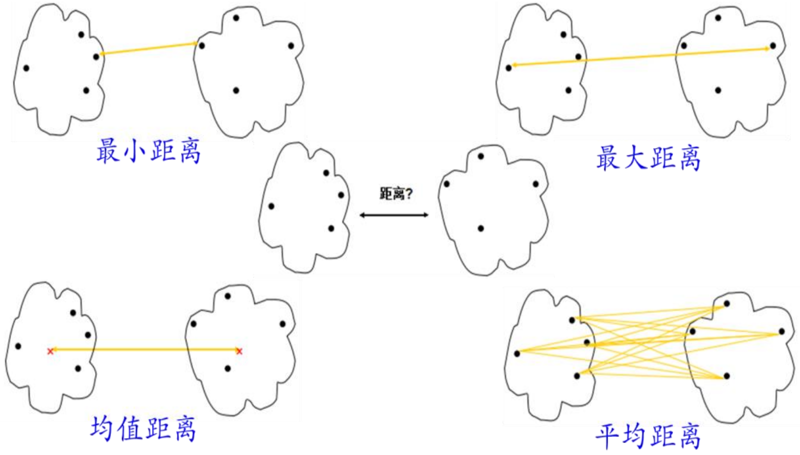
第9章 聚类

- 接近性度量/差异性度量
- 顺序聚类算法
- 划分聚类算法（K-means聚类算法）
- 层次聚类算法（ AGNES 、 DIANA ）
- 密度聚类算法（ DBSCAN ）



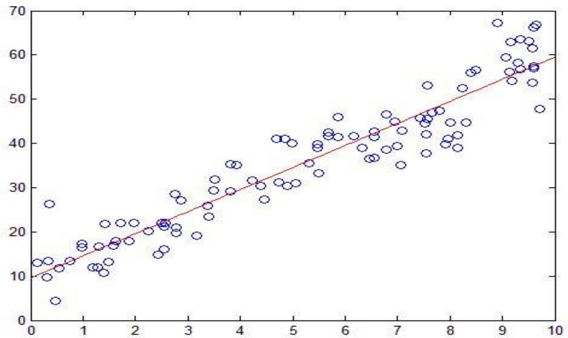


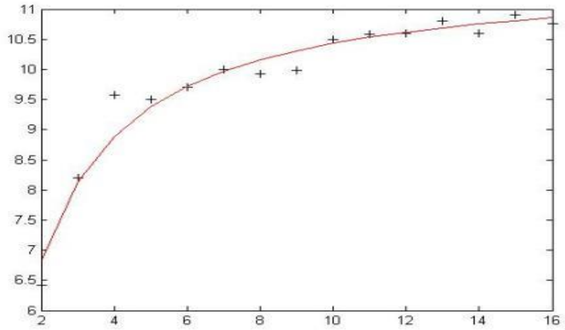




第10章 回归

- 一元线性回归
- 多元线性回归
- 非线性回归
- 附录：逐步回归、岭回归、Lasso回归



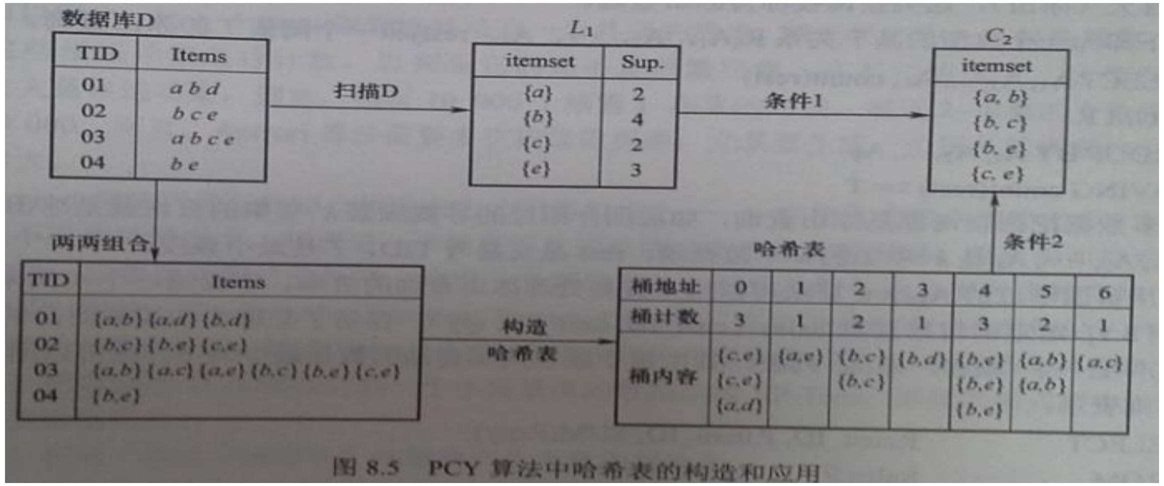


$$\hat{\beta} = (X^{\tau} X)^{-1} X^{\tau} Y$$

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \\ \varepsilon \sim N\left(0, \sigma^2\right) \end{cases}$$

第11章 关联规则分析

- 频繁相集/关联规则
- Apriori算法
- PCY算法
- FP-Growth算法
- 序列模式挖掘



数据库D

TID	Items
01	*a b d*
02	*b c e*
03	*a b c e*
04	*b e*

扫描D

L_1

itemset	Sup.
{*a*}	2
{*b*}	4
{*c*}	2
{*e*}	3

条件1

C_2

itemset
{*a*, *b*}
{*b*, *c*}
{*b*, *e*}
{*c*, *e*}

两两组合

TID	Items
01	{*a*,*b*} {*a*,*d*} {*b*,*d*}
02	{*b*,*c*} {*b*,*e*} {*c*,*e*}
03	{*a*,*b*} {*a*,*c*} {*a*,*e*} {*b*,*c*} {*b*,*e*} {*c*,*e*}
04	{*b*,*e*}

构造哈希表

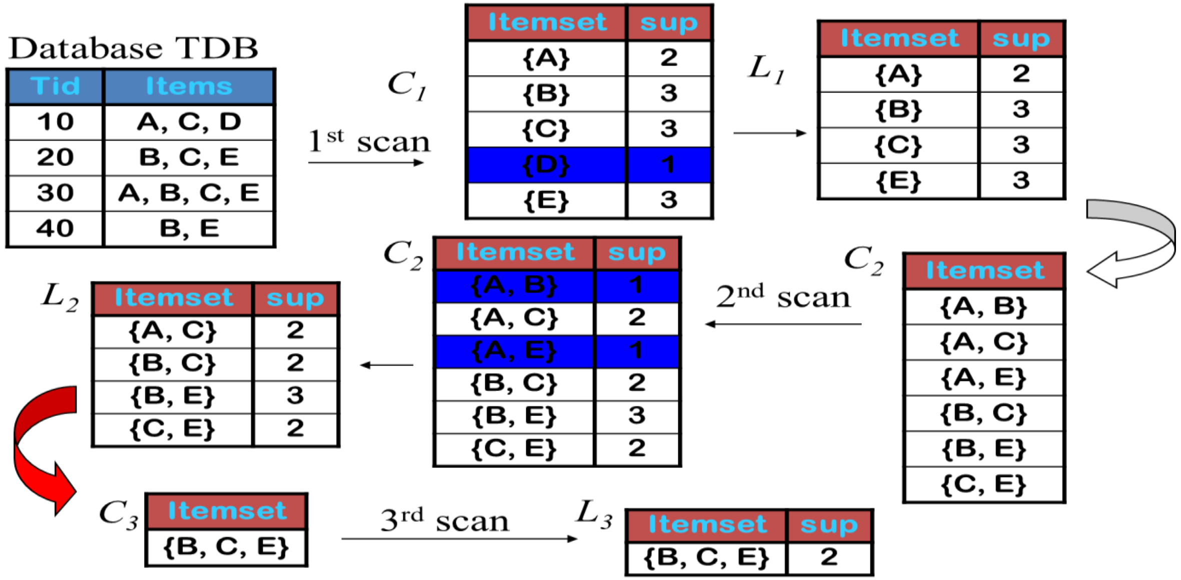
哈希表

桶地址	0	1	2	3	4	5	6
桶计数	3	1	2	1	3	2	1

桶内容	{*c*,*e*}
	{*c*,*e*}
	{*a*,*d*}
		{*a*,*e*}	{*b*,*c*}
			{*b*,*c*}
				{*b*,*d*}	{*b*,*e*}
					{*b*,*e*}
					{*b*,*e*}
						{*a*,*b*}
						{*a*,*b*}
							{*a*,*c*}

条件2

图 8.5 PCY 算法中哈希表的构造和应用



Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1st scan

C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan

C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

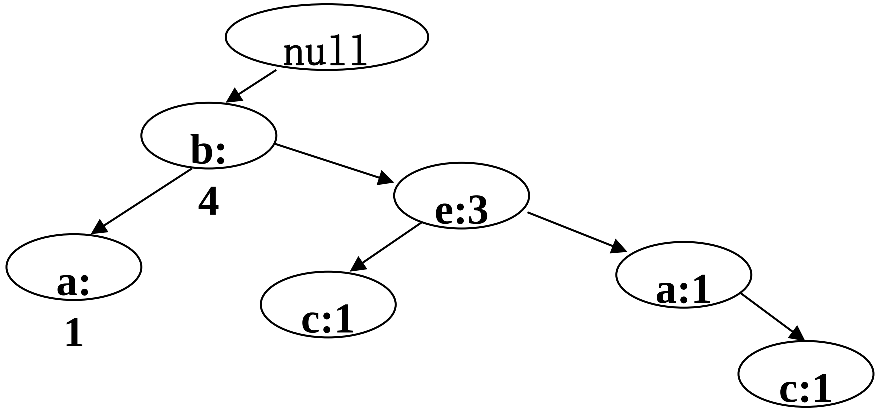
C_3

Itemset
{B, C, E}

3rd scan

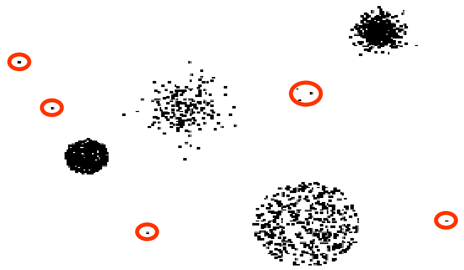
L_3

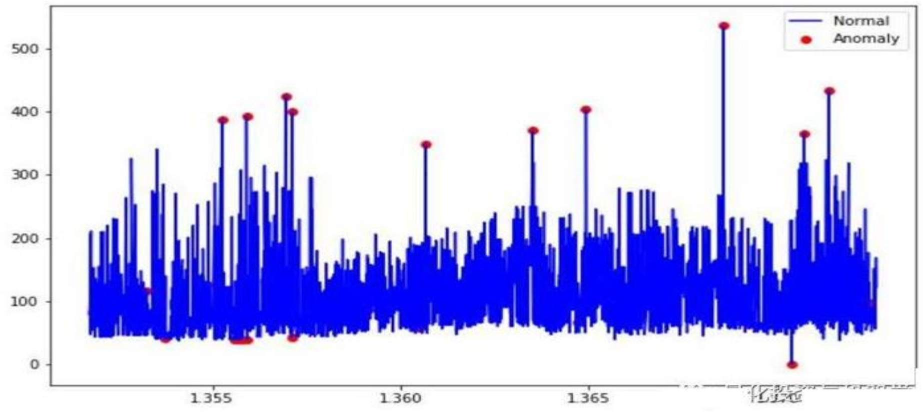
Itemset	sup
{B, C, E}	2

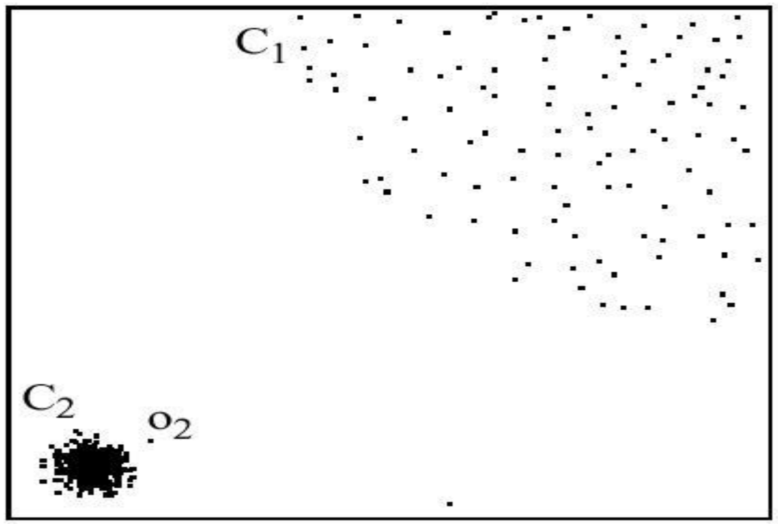


第12章 异常检测

- 定义与主要挑战
- 基于图形的方法
  - 箱型图
- 基于统计的方法
- 分类和聚类
- 基于距离和基于密度的方法
  - 基于K-means的方法
  - 局部异常因子（LOF）方法

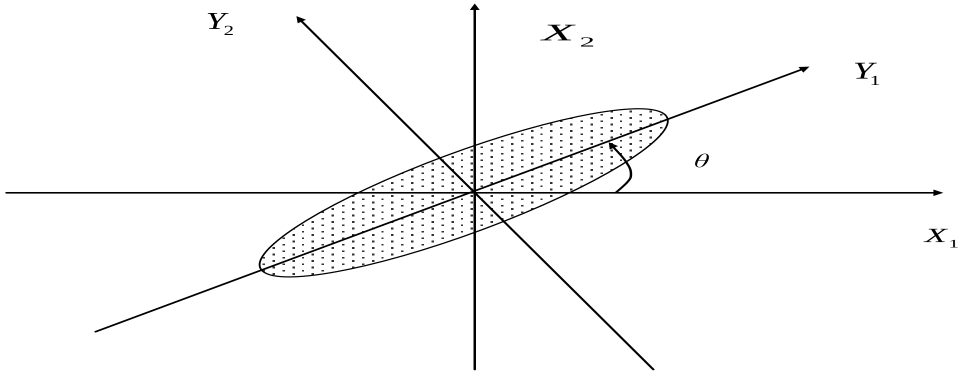






第13章 数据降维

- 主成分分析
  - 几何意义、总体主成分、样本主成分



- 因子分析
  - 因子载荷、因子旋转、因子得分

$$\boldsymbol{\Sigma} = \mathbf{AA}' + \mathbf{D} = \mathbf{U}\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p\end{bmatrix}\mathbf{U}'$$

- SVD分解

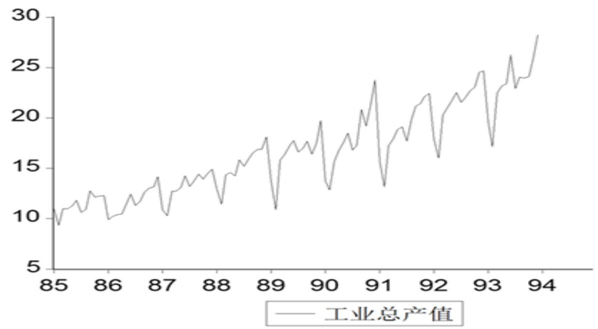
$$\boldsymbol{A}_{m\times n} = \boldsymbol{U}_{m\times m}\boldsymbol{\Sigma}_{m\times n}\boldsymbol{V}^{\mathrm{T}}_{n\times n}$$

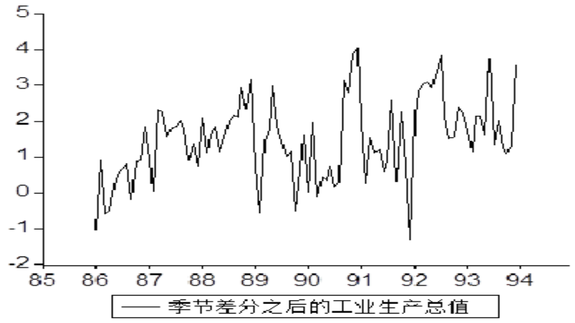
- 低维嵌入

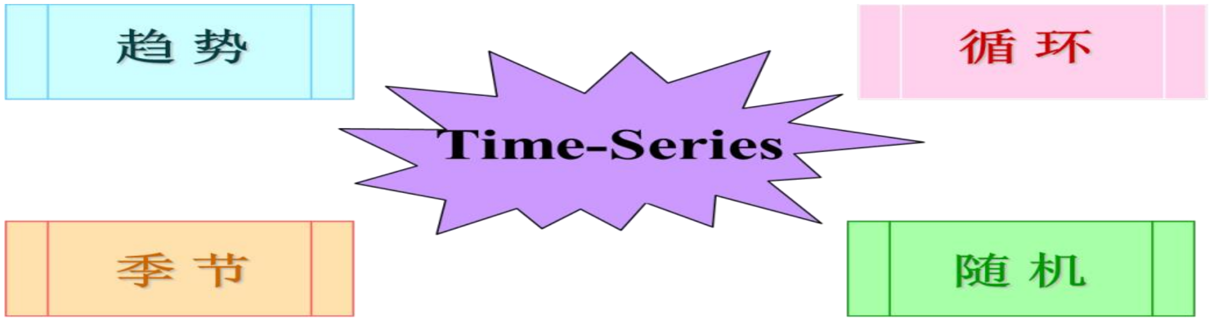
$$\varepsilon(\boldsymbol{W}) = \sum_{i=1}^{n}\left|\boldsymbol{x}_i - \sum_{j=1}^{k} w_{ij}\boldsymbol{x}_{i\_j}\right|^2 \qquad \Phi(\boldsymbol{W}) = \sum_{i=1}^{n}\left|\boldsymbol{y}_i - \sum_{j=1}^{k} w_{ij}\boldsymbol{y}_{i\_j}\right|^2$$

第14章 时间序列分析

- 时间序列分析的预处理
  - 平稳性检验
  - 随机性检验
- 平稳时间序列分析
  - ARMA模型
  - 平稳序列建模与学习
  - 序列预测
- 非平稳时间序列分析
  - 序列分解
  - ARIMA模型
  - 残差自回归模型
  - 指数平滑预测模型







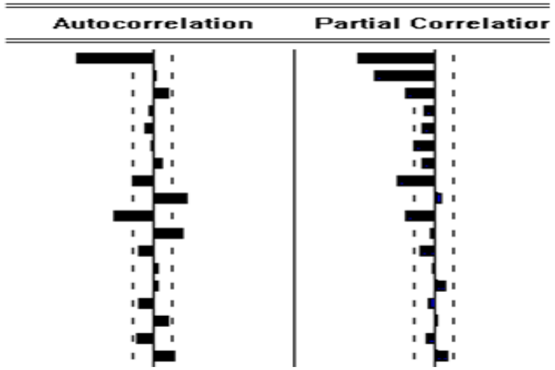
趋 势

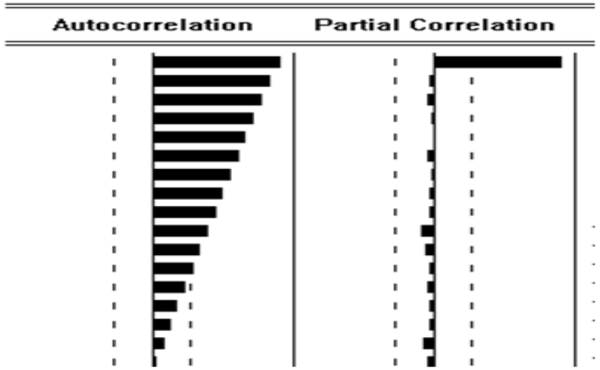
循 环

Time-Series

季 节

随 机

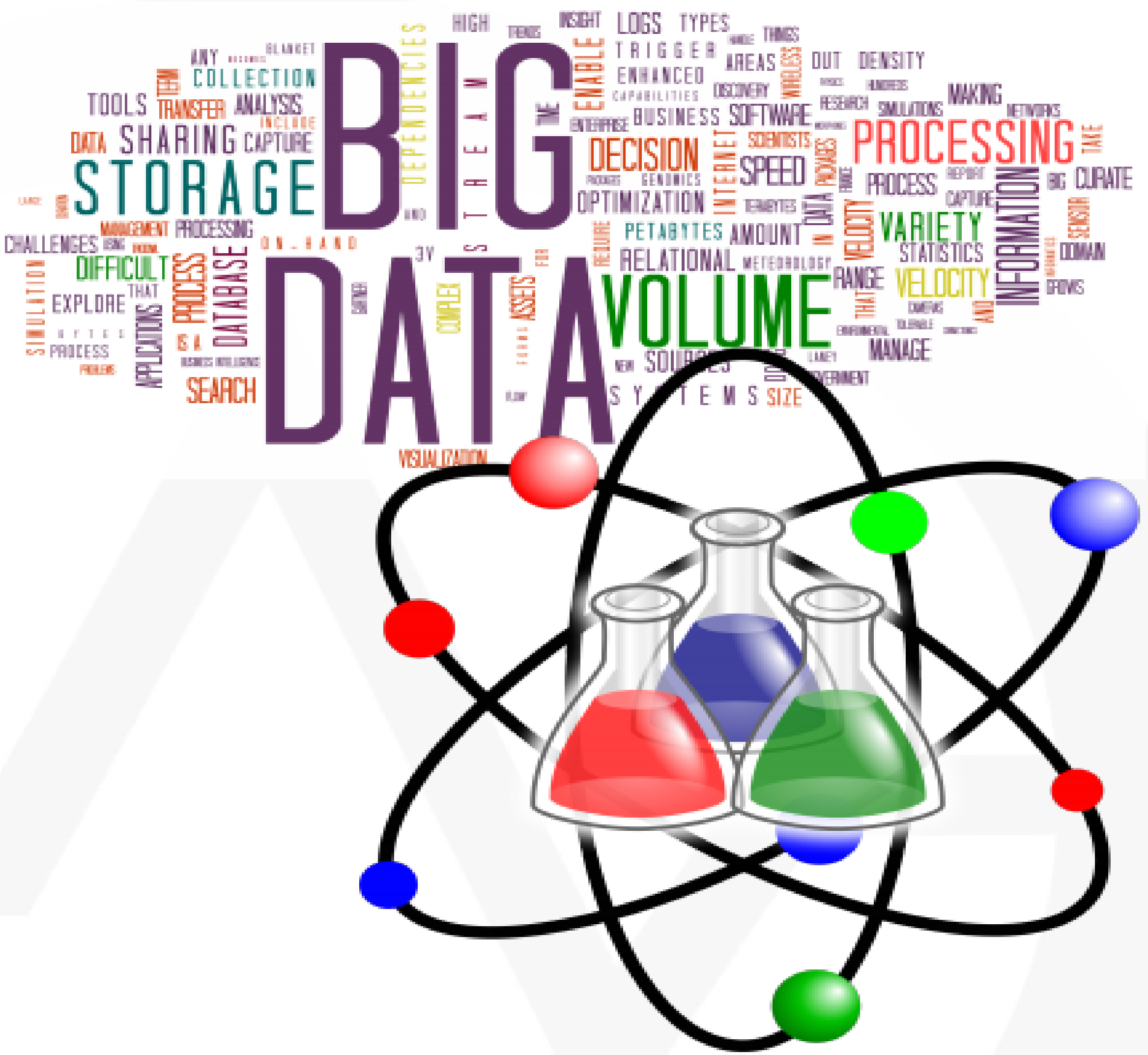




What is Data Science?

Extraction of knowledge from large volumes of data that are structured or unstructured.

It is a continuation of the fields **data mining** and **predictive analytics**



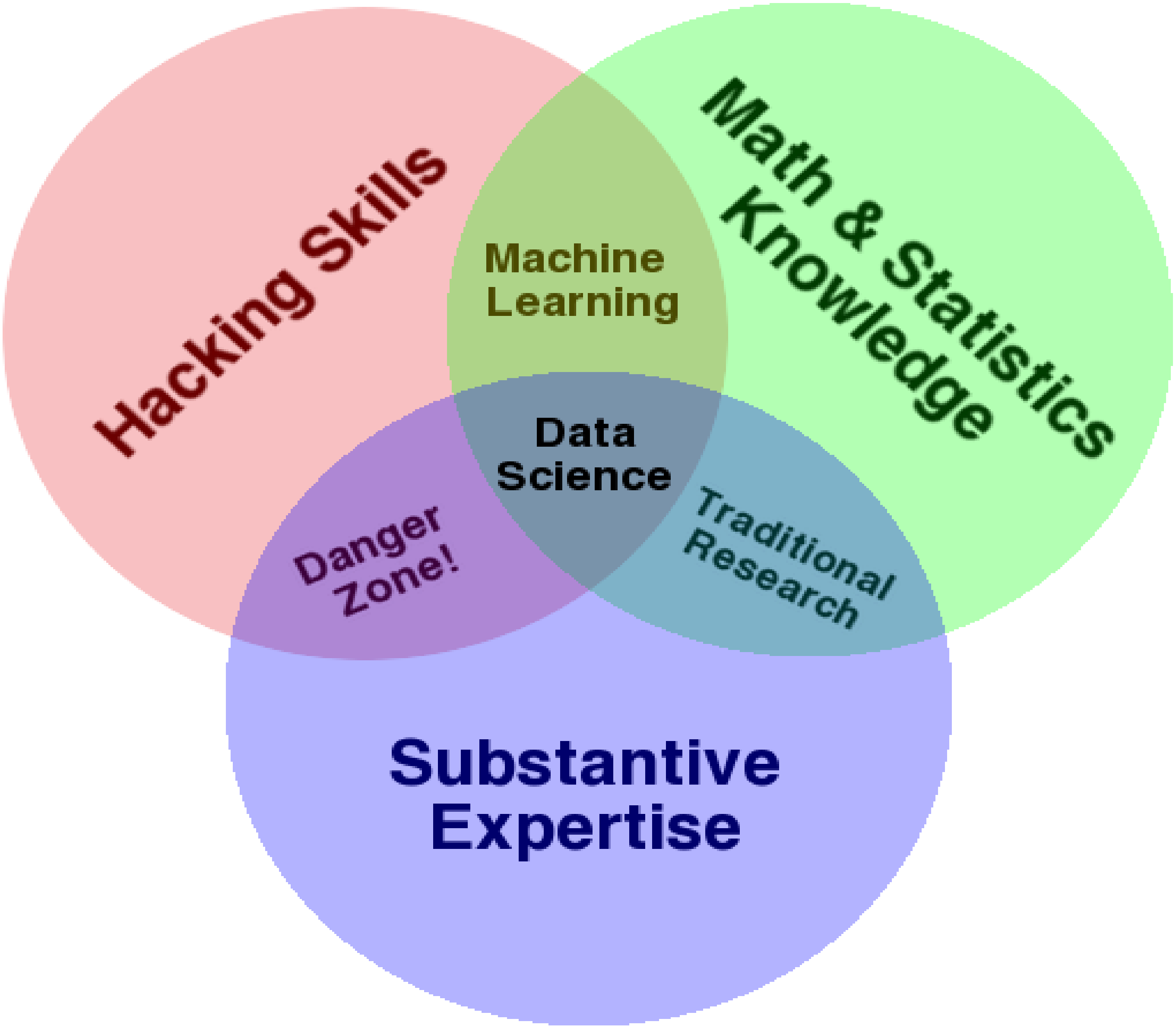
My Definition for Data Science

The application of **data centric**, **computational**, and **inferential thinking** to

understand the world	**&**	*solve problems*
Science		**Engineering**

- *Data science is fundamentally <u>interdisciplinary</u>*

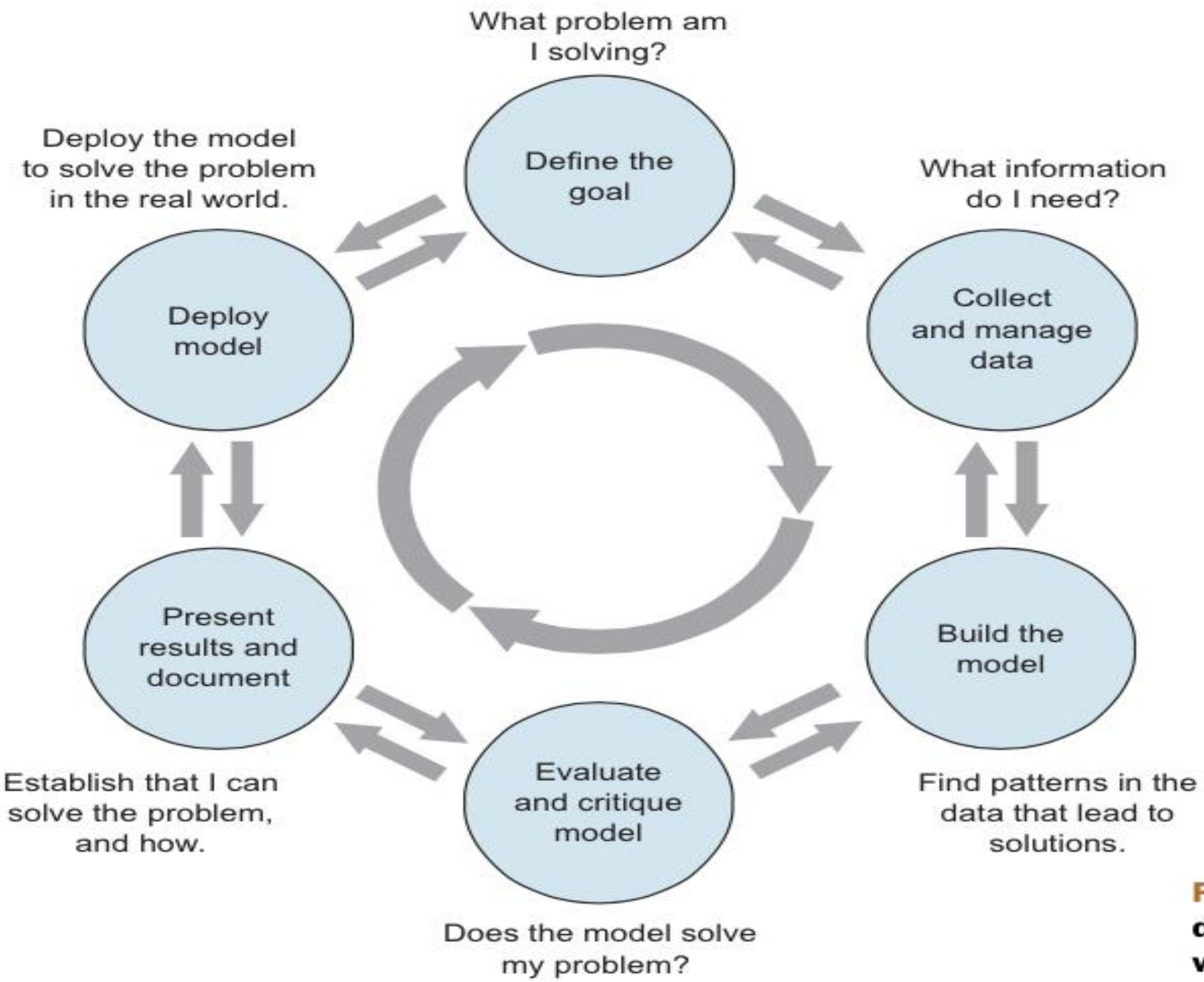
Data Science



Data Science Tools

python ™
ANACONDA®
Powered by Continuum Analytics
pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$
NumPy
SciPy.org
APACHE Spark™
matplotlib
Version 2.2.3
scikit learn

DATA SCIENCE LIFECYCLE: AN ALTERNATE VIEW



What problem am I solving?

Define the goal

What information do I need?

Collect and manage data

Build the model

Find patterns in the data that lead to solutions.

Evaluate and critique model

Does the model solve my problem?

Present results and document

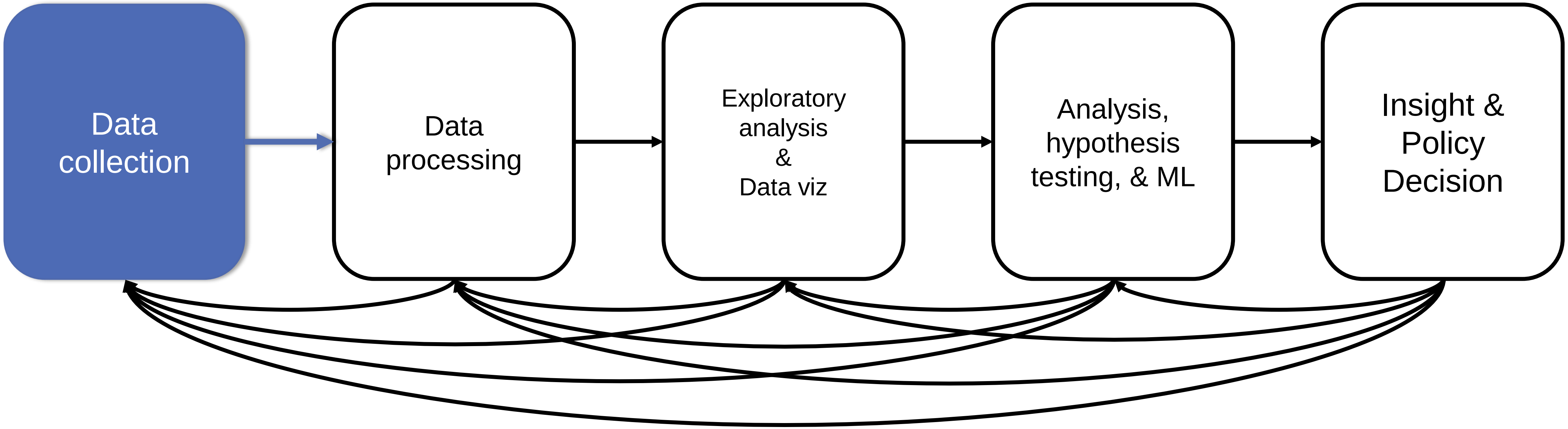
Establish that I can solve the problem, and how.

Deploy model

Deploy the model to solve the problem in the real world.

Figure 1.1 **The lifecycle of a data science project: loops within loops**

DATA SCIENCE LIFECYCLE



Data collection

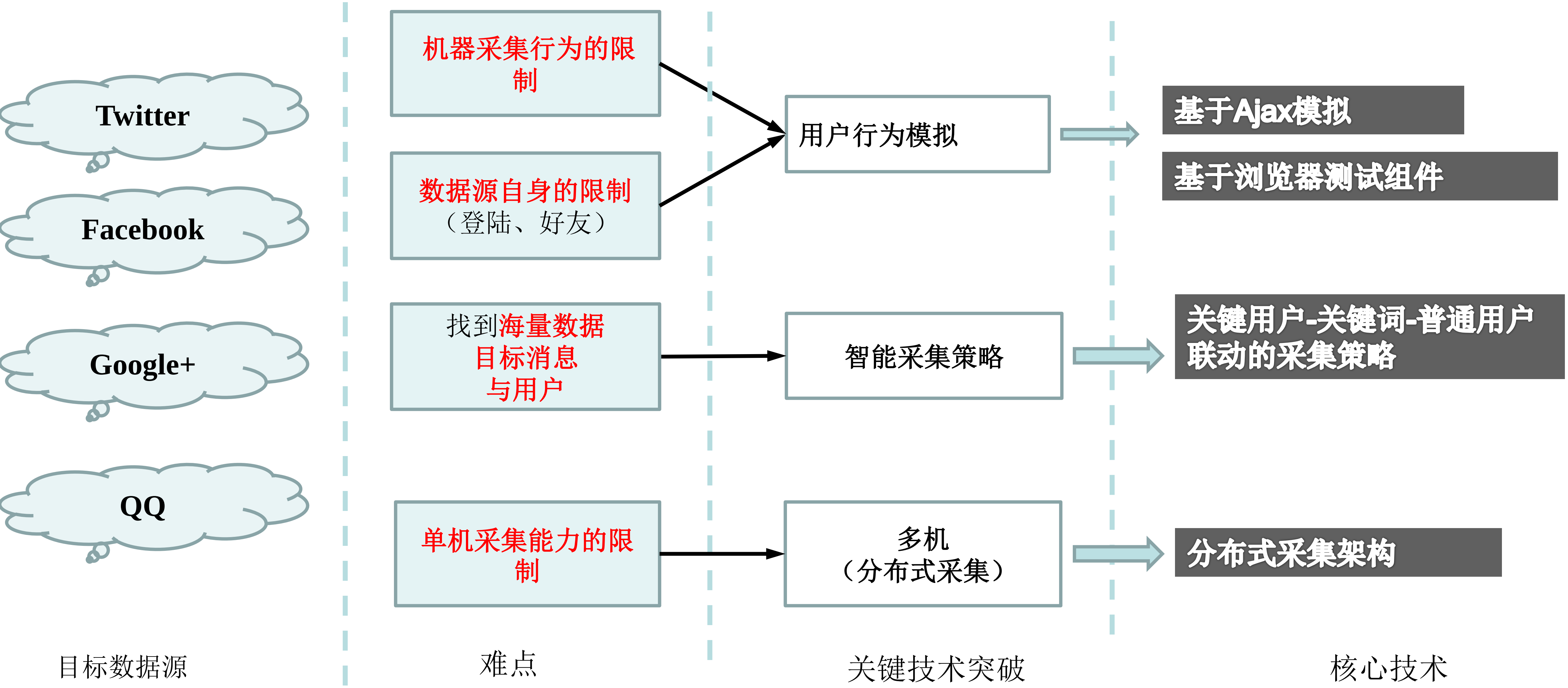
Data processing

Exploratory analysis & Data viz

Analysis, hypothesis testing, & ML

Insight & Policy Decision

采集的目标：快、准、全



Twitter

Facebook

Google+

QQ

机器采集行为的限制

数据源自身的限制（登陆、好友）

找到海量数据目标消息与用户

单机采集能力的限制

用户行为模拟

智能采集策略

多机（分布式采集）

基于Ajax模拟

基于浏览器测试组件

关键用户-关键词-普通用户联动的采集策略

分布式采集架构

目标数据源

难点

关键技术突破

核心技术

Data Cleanup

Real data is messay, often needs to cleaned up before useful.

- Bad forms
- Missing Data
- Useless Variables
- Wrong Data

Data Cleanup (Contd.)

- Transform variables (date formats, String to int)
- Create derived variables
  - Derive county from IP
  - Age from ID card number
- Normalize strings
  - Different spelling and nicknames (William -> Bill)
- Feature value rescaling
- Enrich

Data Exploration

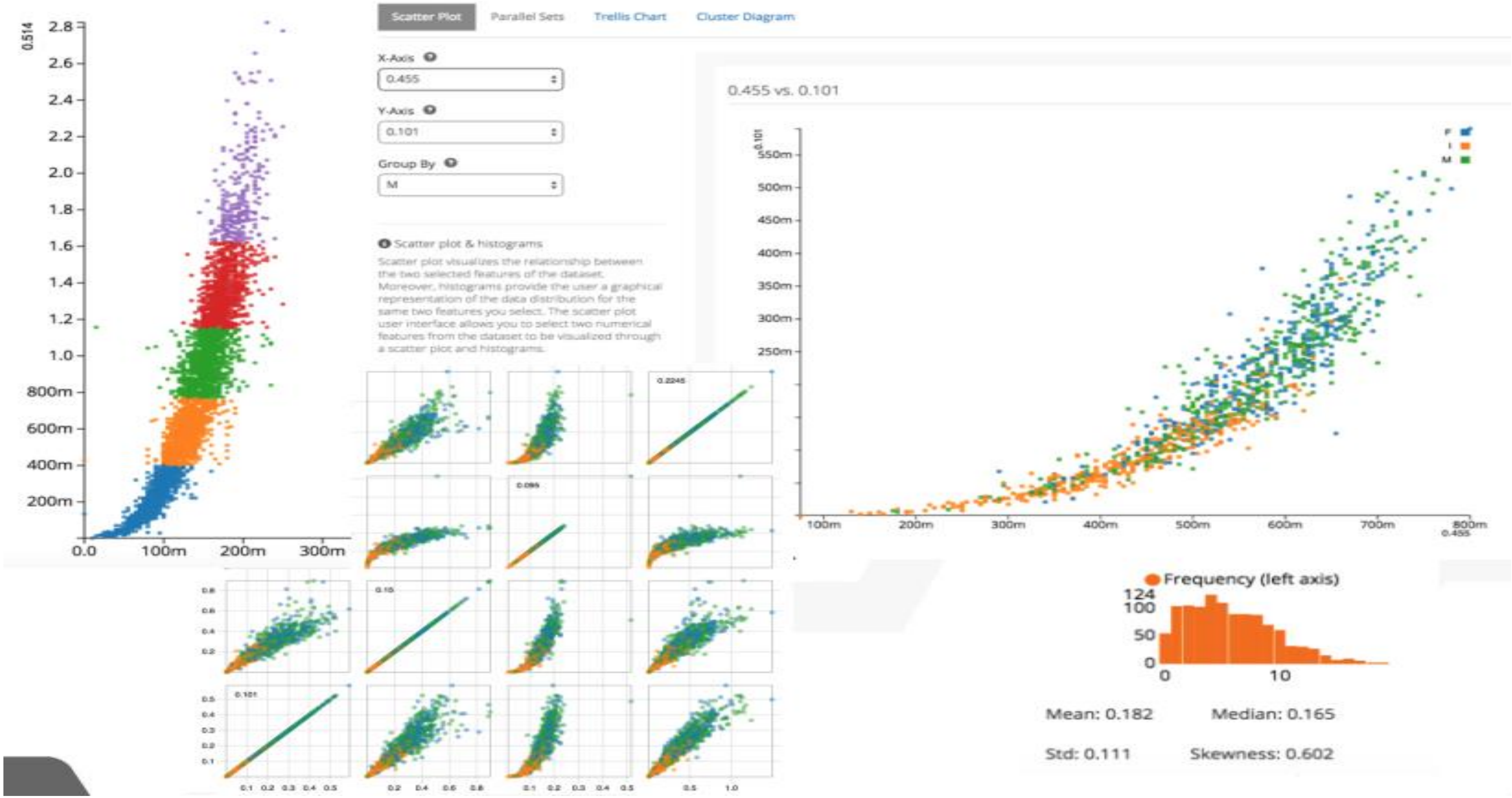
Understand, and get a feel for what is expected (models => densities, constraints) and unexpected/ residuals (errors, outliers)

- think what this is data about? domain, background, how it is collected, what each fields mean and range of values.
- head, tail, count, all descriptives (Mean, Max, median, percentiles ..) - Five number Summary. Min. 1st Qu. Median Mean 3rd Qu. Max.
- run a bunch of count/group-by statements to gauge if it's corrupt

Data Exploration (Contd.)

- Plot - take random sample and explore (scatter plot)
  - e.g. Draw scatter plot or Trellis Plot
- Find Dependencies between fields
  - Calculate Correlation
  - Dimensionality reduction
  - Cluster and look visualize clusters
- Look at frequency distribution of each field and try to find a known distribution if possible.

Data Exploration (Contd.)



X-Axis
0.455

Y-Axis
0.101

Group By
M

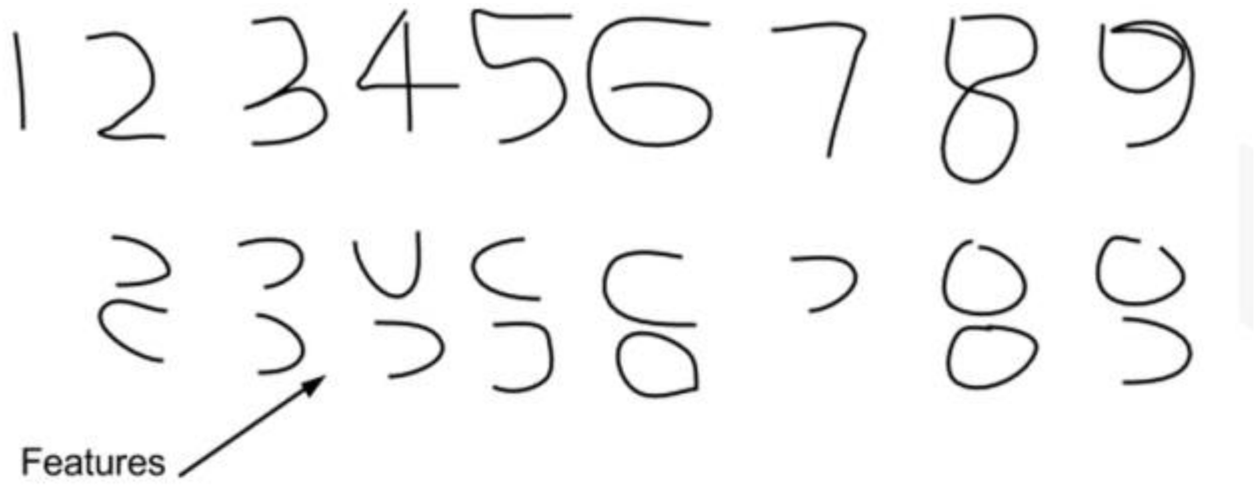
Scatter plot & histograms

Scatter plot visualizes the relationship between the two selected features of the dataset. Moreover, histograms provide the user a graphical representation of the data distribution for the same two features you select. The scatter plot user interface allows you to select two numerical features from the dataset to be visualized through a scatter plot and histograms.

Mean: 0.182 Median: 0.165

Std: 0.111 Skewness: 0.602

Feature Engineering



- Feature engineering is the art of finding feature that leads simplest decision algorithm. (Good features allow a simple model to beat a complex model.)
- Best features may be a subset, or a combination, or transformed version of the features.

How to do Feature Engineering?

- Manually pick by domain experts and trial and error.
- Search the possible combinations by training and combining subsets (e.g. Random Forest)
- Use statistical concepts like correlation and information criteria
- Reduce the features to a low dimension space using techniques like PCA.
- Automatic Feature Learning though Deep Learning

Analysis

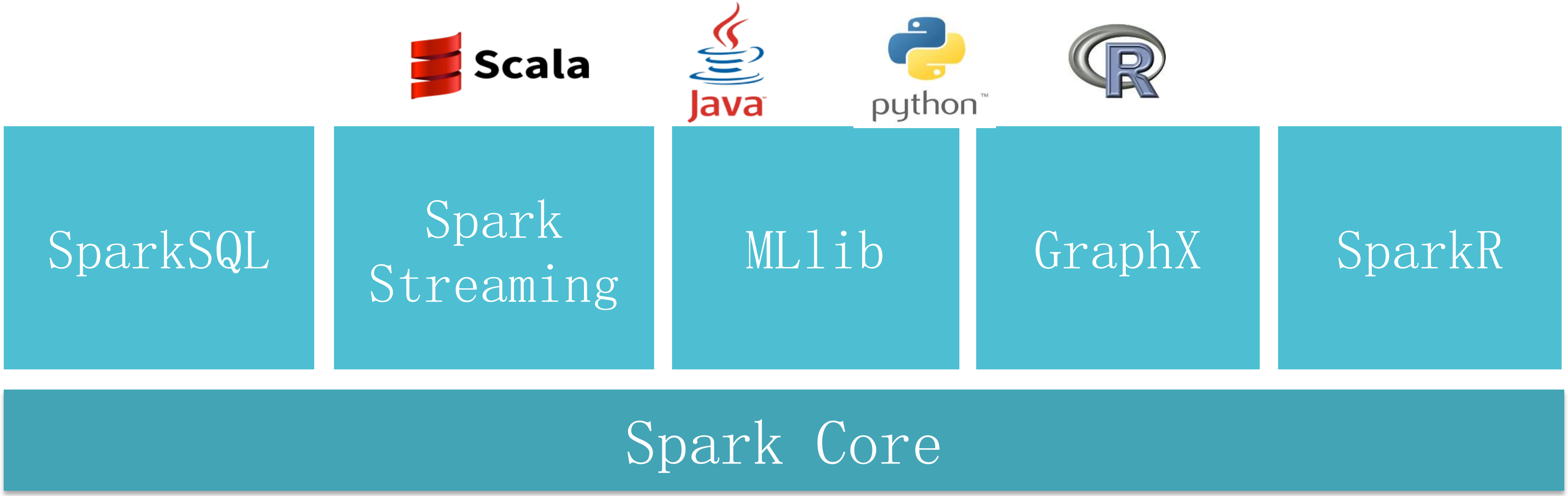
- Goal of analysis is to extract knowledge
- This knowledge usually come in one of the two forms
  - KPI (Key Performance Indicators)
    - Describe key measurement for what is being measured. (e.g. revenue per year, profit margin, revenue for sqft in retail, revenue per employer)
  - Models to describe or predict the data
    - e.g. Machine Learning models or Statistical models

4 Analysis types by time to decision

- Hindsight (what happened?)
  - Done using Batch Analytics like MapReduce
- Oversight (what is happening?)
  - Done using Realtime Analytics technologies like CEP
- Insight (why things happening?)
  - Done with Data Mining and Unsupervised learning algorithms like Clustering
- Foresight (what will happen?)
  - Done by building models using Machine learning or one of other techniques

Data Analytics Tools : Apache Spark

- Unifies ***batch, interactive, streaming*** workloads
- Easy to build sophisticated applications
  - Support iterative, graph-parallel algorithms
  - Powerful APIs in Scala, Python, Java, R



SparkSQL

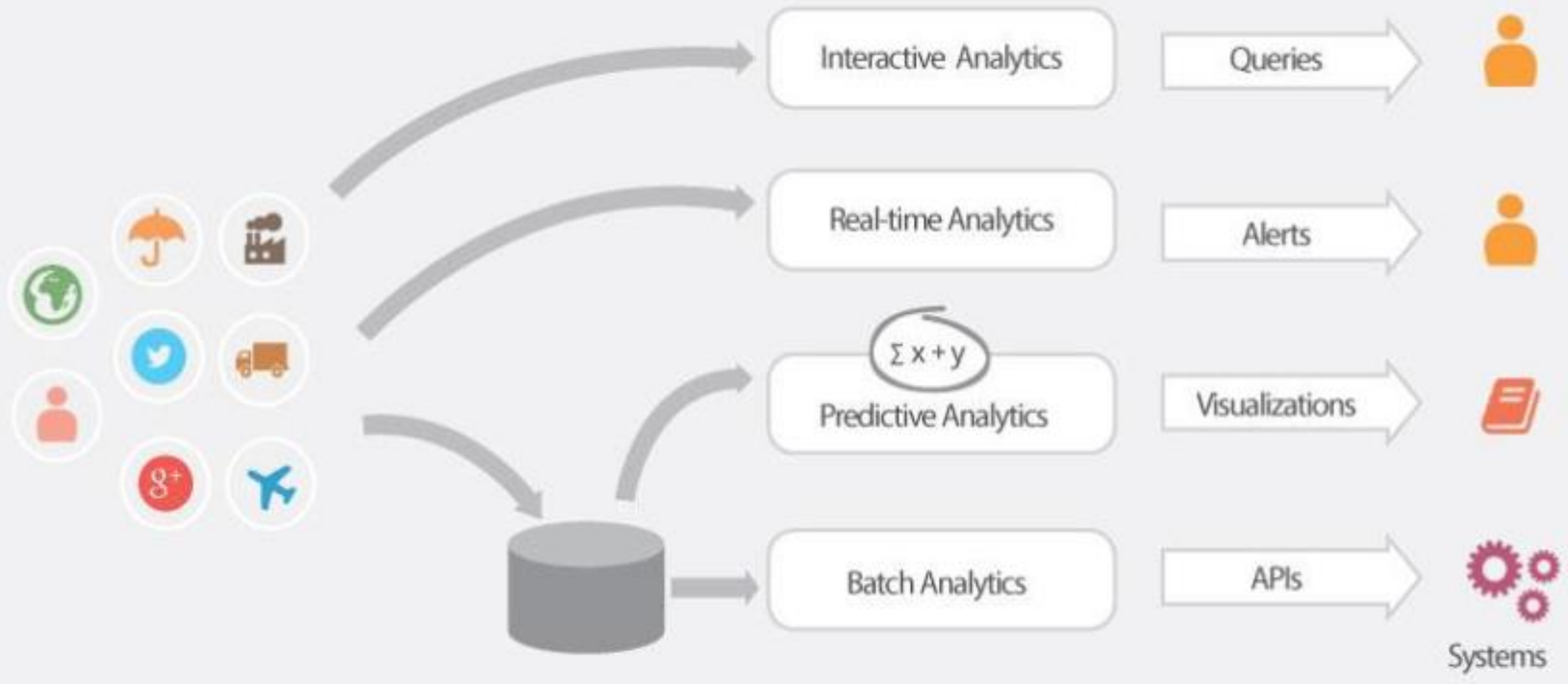
Spark Streaming

MLlib

GraphX

SparkR

Spark Core



Collect Data → Analyze & Make Decisions → Communicate

Inconvenient Truth About Data Science

- Data is **never clean**
- You will spend **most of your time** cleaning and preparing data.
- **95%** of tasks do not require deep learning
- In **90%** of cases generalized linear regression will do the trick
- Big Data is just a **tool**.
- You should embrace the **Bayesian** approach.
- No one cares **how you did it**.
- Academia and business are **two different** worlds.
- **Presentation is key** – be a master of Power Point.
- All models are false, but **some** are **useful**.
- There is no fully automated Data Science. You need to **get your hands dirty**.

Data Science is extracting knowledge by analyzing data

Hindsight

Oversight

Insight

Foresight

谢谢大家！