上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

New Media Data Analytics and Application

Lecture 1: A Brief Introduction

Ting Wang

Outlines

1. Significance of Data Analysis
2. Definition of Data Analysis
3. History of Computer Data Analytics
4. Domains of New Media Data Analytics

the significance of data analysis

Why Data Analysis

Why Data Analysis

EXAMPLE 1: Money Exchange

Why Data Analysis

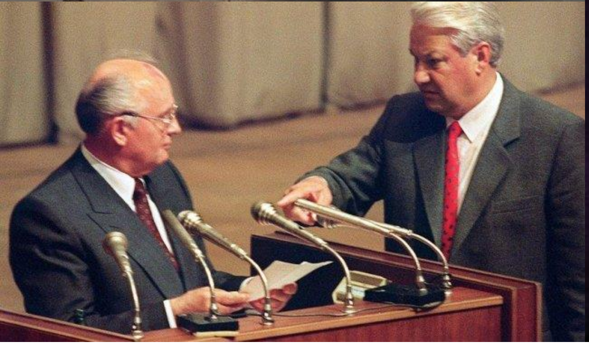
US Dollars,

The King of the World

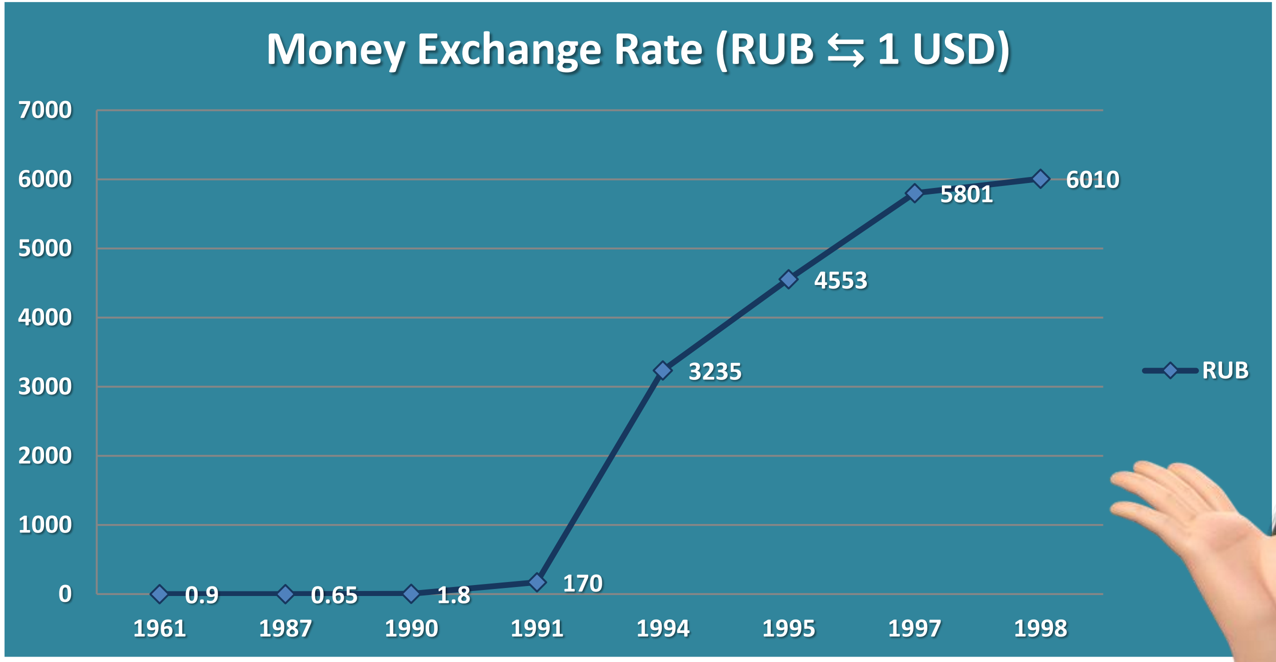
Bretton Woods, 1944

Why Data Analysis

Political events impact on economic trends



Why Data Analysis

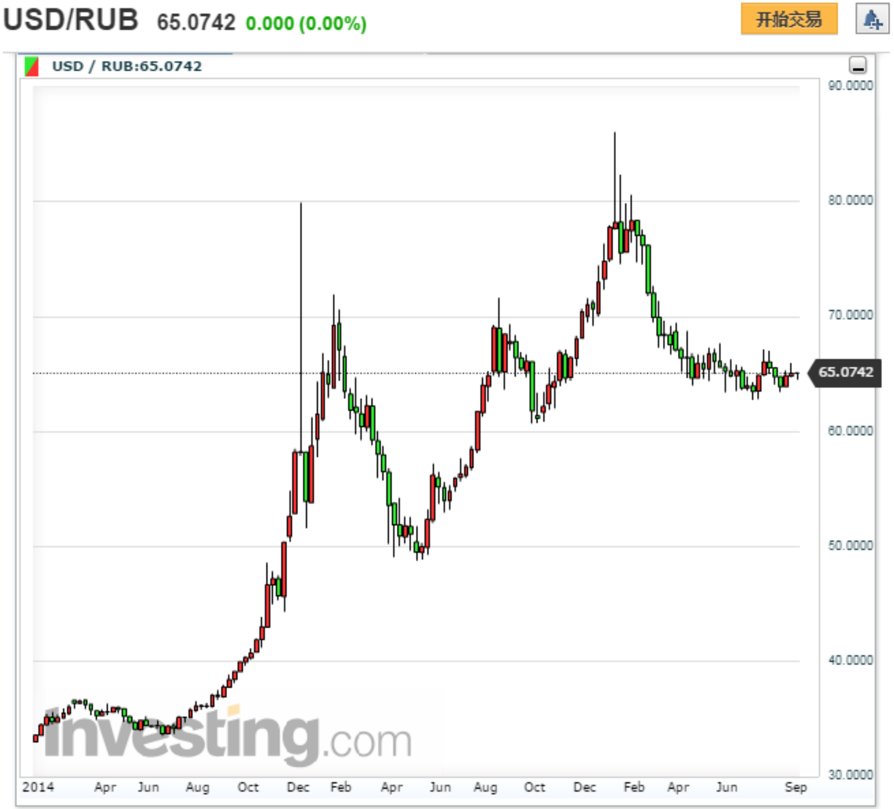


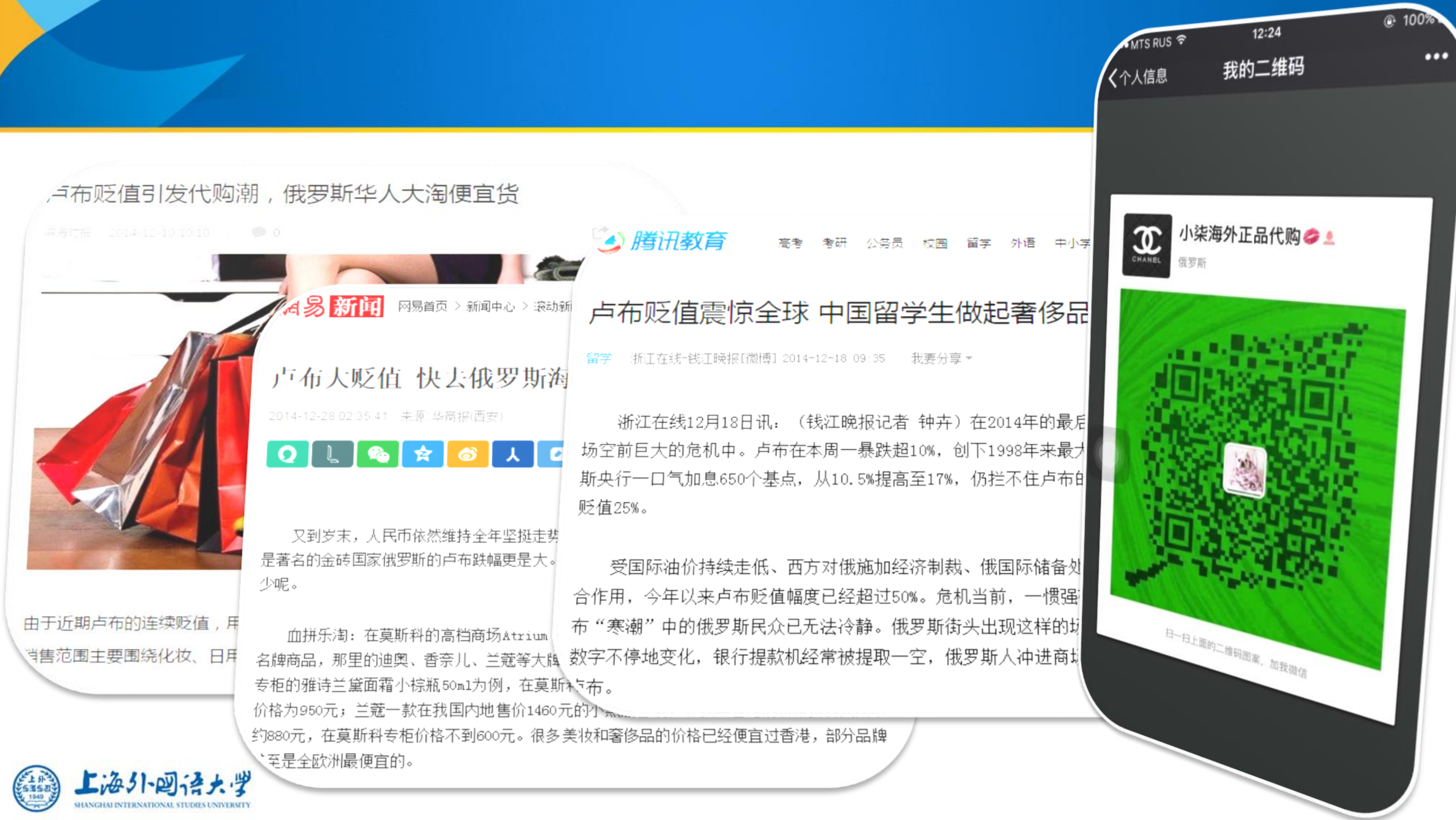
Money Exchange Rate (RUB ⇆ 1 USD)

Why Data Analysis

http://cn.investing.com/

KHERSON OBLAST
KHERSON OBLAST
ZAPORIZHIA OBLAST
Gulf of Karkinit
Sea of Azov
Armyansk
Krasnoperekopsk
Dzhankoy
AUTONOMOUS REPUBLIC OF CRIMEA
Kerch
RUSSIA
Yevpatoriya
Saky
Simferopol
Feodosiya
Black Sea
Bakhchysaray
Alushta
Sevastopol
SEVASTOPOL
Yalta
Black Sea





卢布贬值引发代购潮，俄罗斯华人大淘便宜货

2014-12-10 10:10 0

由于近期卢布的连续贬值，用

销售范围主要围绕化妆、日用

卢布大贬值 快去俄罗斯淘

2014-12-28 02:35:41 来源: 华商报(西安)

又到岁末，人民币依然维持全年坚挺走势
是著名的金砖国家俄罗斯的卢布跌幅更是大。
少呢。

血拼乐淘：在莫斯科的高档商场Atrium
名牌商品，那里的迪奥、香奈儿、兰蔻等大牌
专柜的雅诗兰黛面霜小棕瓶50ml为例，在莫斯
价格为950元；兰蔻一款在我国内地售价1460元的小
约880元，在莫斯科专柜价格不到600元。很多美妆和奢侈品的价格已经便宜过香港，部分品牌
甚至是全欧洲最便宜的。

腾讯教育
高考 考研 公务员 校园 留学 外语 中小学

卢布贬值震惊全球 中国留学生做起奢侈品

留学 浙江在线-钱江晚报[微博] 2014-12-18 09:35 我要分享

浙江在线12月18日讯：（钱江晚报记者 钟卉）在2014年的最后
场空前巨大的危机中。卢布在本周一暴跌超10%，创下1998年来最大
斯央行一口气加息650个基点，从10.5%提高至17%，仍拦不住卢布的
贬值25%。

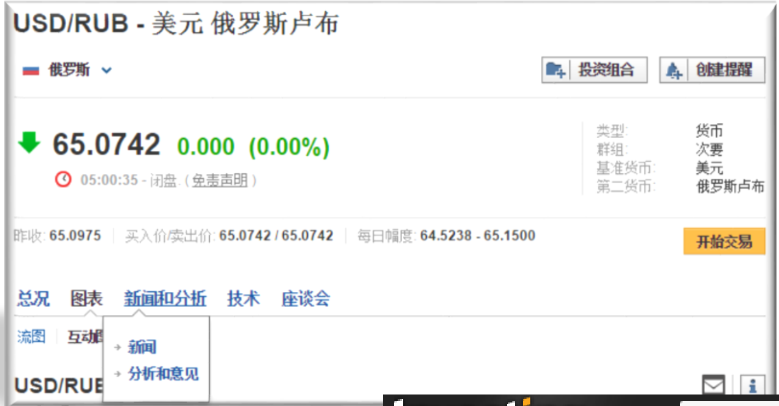
受国际油价持续走低、西方对俄施加经济制裁、俄国际储备处
合作用，今年以来卢布贬值幅度已经超过50%。危机当前，一惯强
布“寒潮”中的俄罗斯民众已无法冷静。俄罗斯街头出现这样的场
数字不停地变化，银行提款机经常被提取一空，俄罗斯人冲进商场
布。

小柒海外正品代购
俄罗斯

扫一扫上面的二维码图案，加我微信

Why Data Analysis

News and New Media are crucial to the investment.



USD/RUB - 美元 俄罗斯卢布

俄罗斯

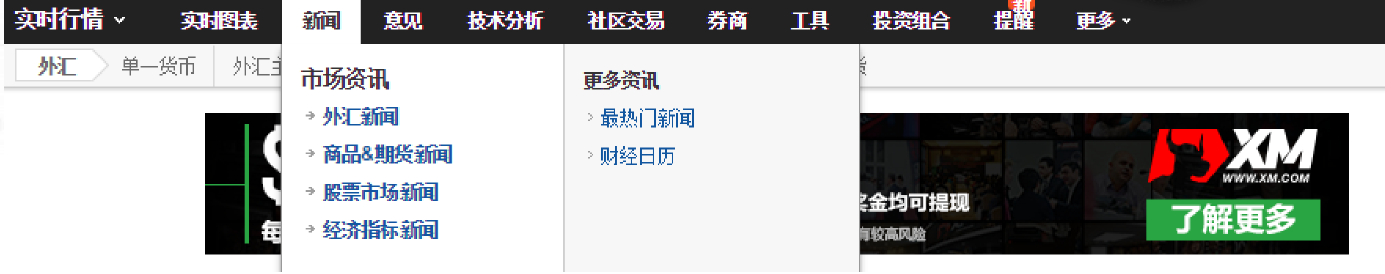
65.0742 0.000 (0.00%)

05:00:35 - 闭盘. (免责声明)

类型:	货币
群组:	次要
基准货币:	美元
第二货币:	俄罗斯卢布

昨收: 65.0975 | 买入价/卖出价: 65.0742 / 65.0742 | 每日幅度: 64.5238 - 65.1500

USD/RUB



Investing.com

欧元/美元 或者 0941

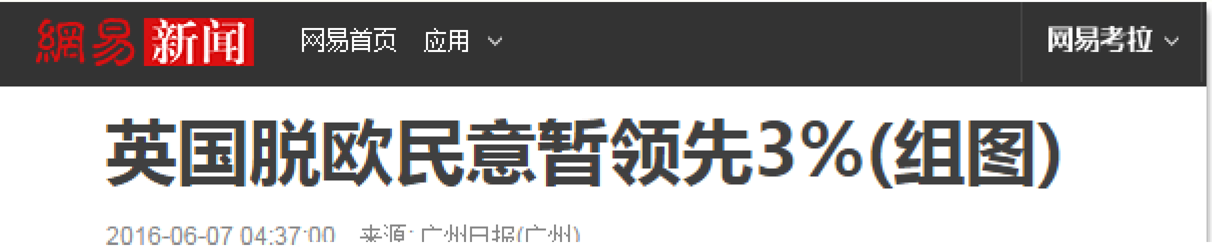
市场资讯
- 外汇新闻
- 商品&期货新闻
- 股票市场新闻
- 经济指标新闻

更多资讯
- 最热门新闻
- 财经日历

上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

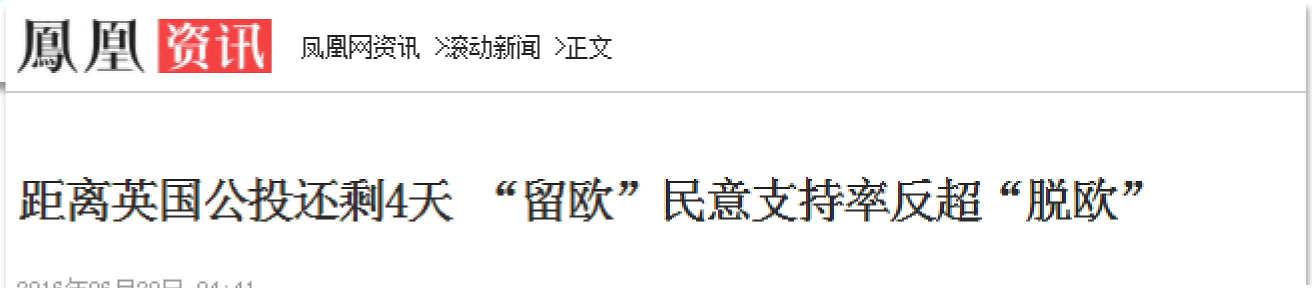
Why Data Analysis

Why Data Analysis



英国脱欧民意暂领先3%(组图)

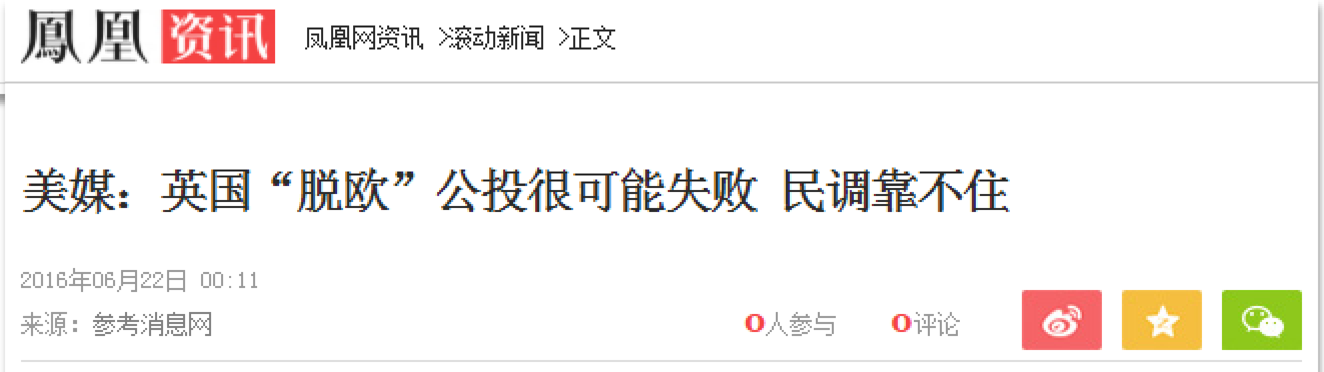
2016-06-07 04:37:00 来源: 广州日报(广州)



距离英国公投还剩4天 “留欧”民意支持率反超“脱欧”

2016年06月20日 04:41

来源：新快报



美媒：英国“脱欧”公投很可能失败 民调靠不住

2016年06月22日 00:11

来源：参考消息网

Why Data Analysis

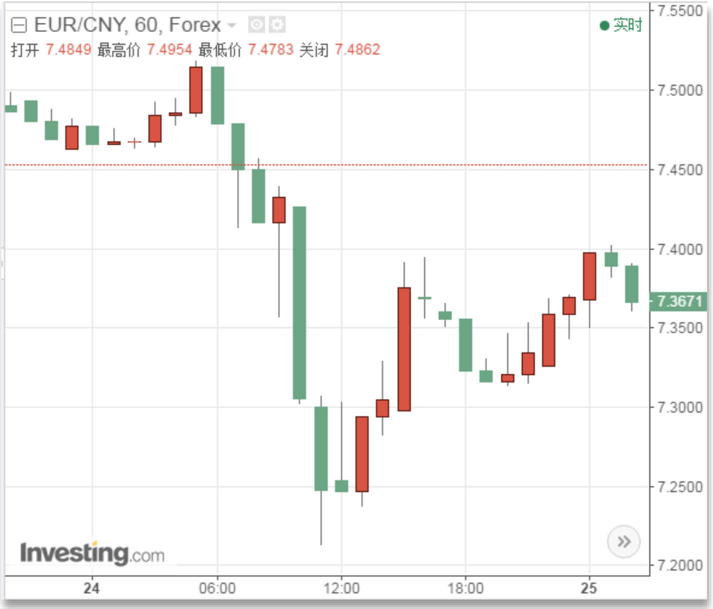
June 23, 2016

Why Data Analysis

Real time data analysis based on news and my EURO currency exchange

http://finance.qq.com/zt2016/gongtou/index.htm

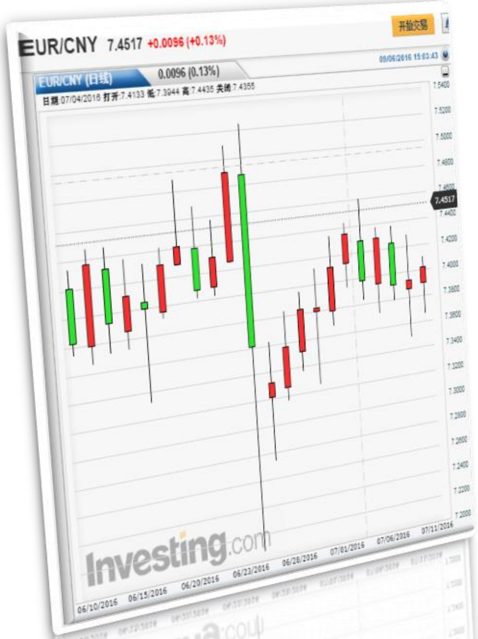
Time	News	Rate
06-24 06:17	Gibraltar (IN)	1:7.52
06-24 06:31	New Castle (IN)	1:7.51
06-24 07:28	Sunderland (OUT)	1:7.47
06-24 09:23	Oxford (IN), North Ireland (IN)	1:7.43
06-24 10:01	49.79% (IN), 50.21% (OUT)	1:7.30
06-24 11:07	48.95% (IN), 51.05% (OUT)	1:7.25
06-24 11:40	Wales (OUT)	1:7.22
06-24 12:13	48.28% (IN), 51.72% (OUT), 339/382 Regions	1:7.24



Why Data Analysis

Final Currency Exchange Results:

(7.52-7.24)*2200=616 RMB


EUR/CNY 7.4517 +0.0096 (+0.13%)

日期:07/04/2016 打开:7.4133 低:7.3944 高:7.4435 关闭:7.4355

Why Data Analysis

Do you have a similar experience using data analysis based on news?

Why Data Analysis

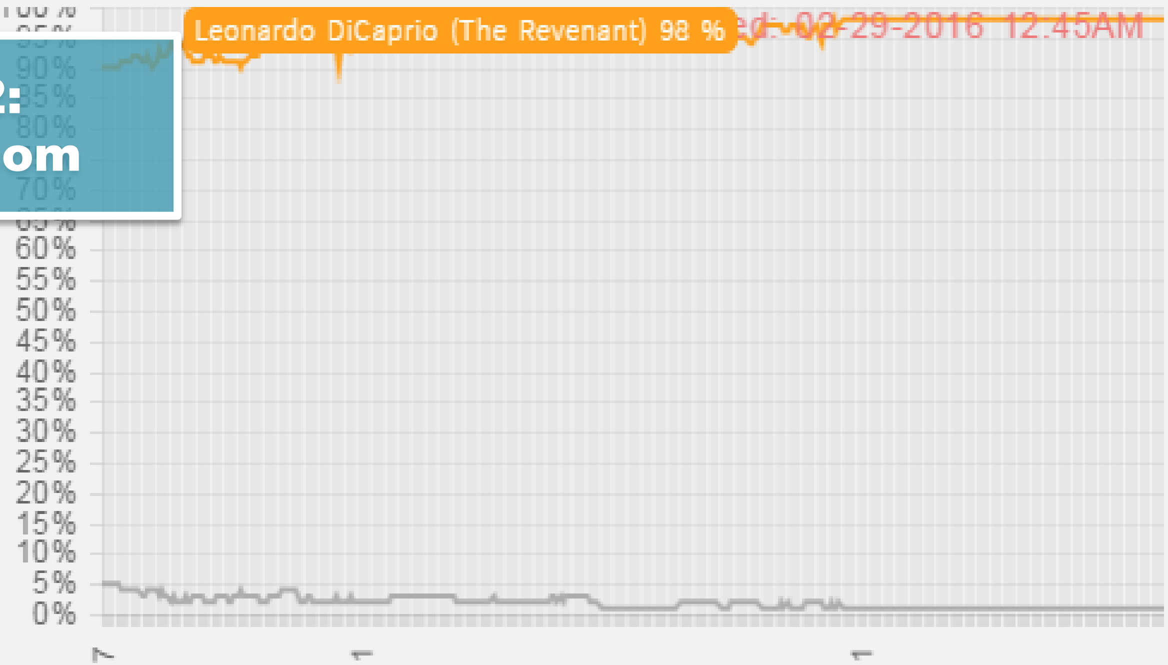
Leading Actor

Leonardo DiCaprio (The Revenant) 98 %

Eddie Redmayne (The D

1 %

EXAMPLE 2: predictwise.com



Why Data Analysis

David Rothschild,

PhD of Wharton School of Business at the University of Pennsylvania

Microsoft researcher at Microsoft Research in New York City

He correctly predicted 50 of 51 Electoral College outcomes in February of 2012, average of 20 of 24 Oscars from 2013-5, and 15 of 15 knockout games in the 2014 World Cup.

- POLITICS
- SPORTS
- ENTERTAINMENT
- ECONOMIC/FINANCIAL

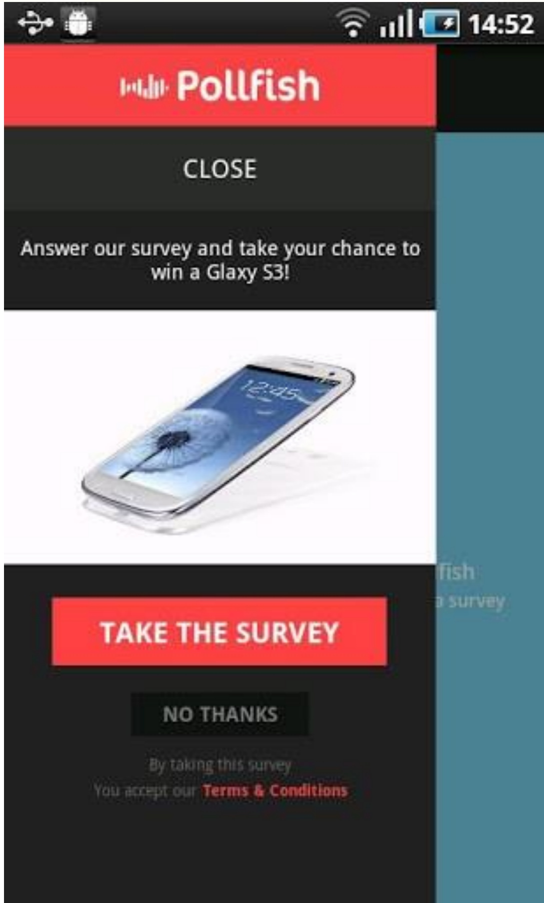
Why Data Analysis

Approaches

- – Data Collection:
  Pollfish, MSN, Xbox

- – Data Analysis:
  Statistical Analysis based on Historical Data

http://predictwise.com/

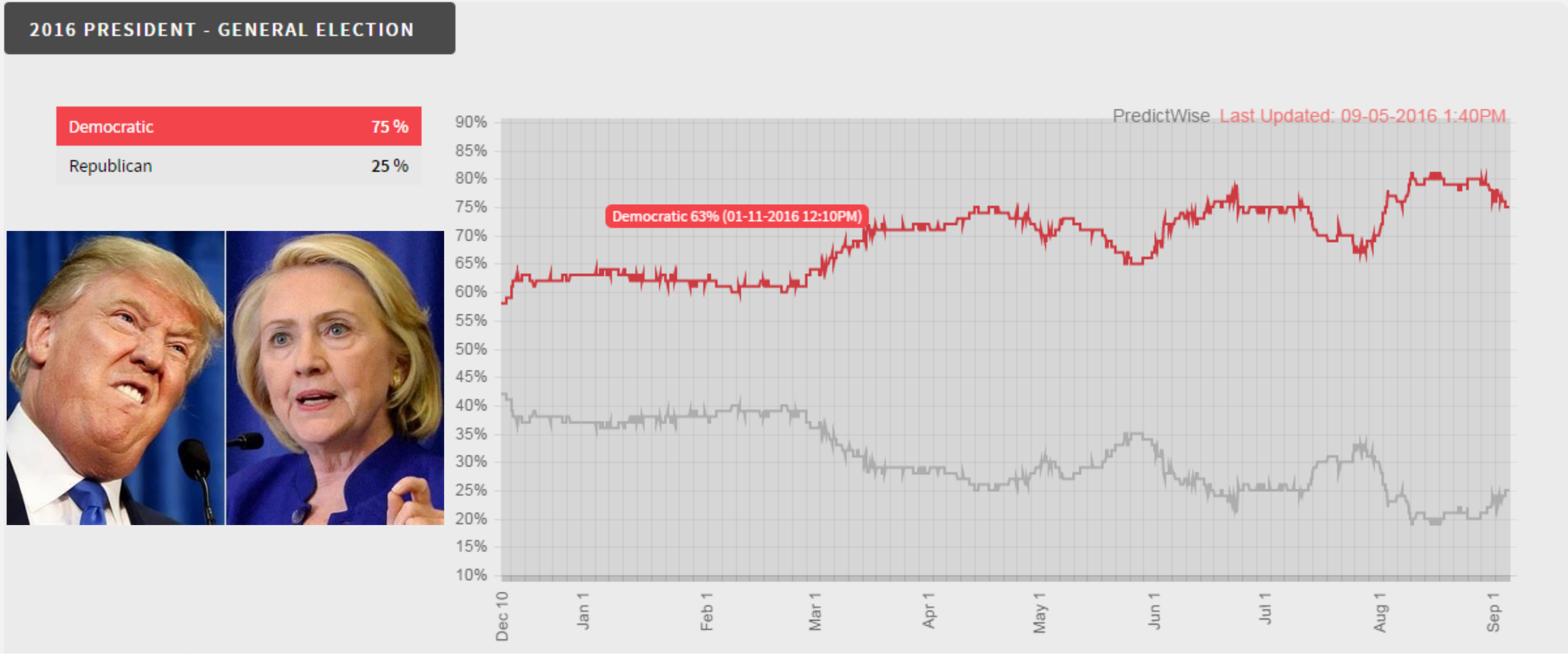


Answer our survey and take your chance to win a Glaxy S3!

By taking this survey
You accept our Terms & Conditions

Why Data Analysis

Politics

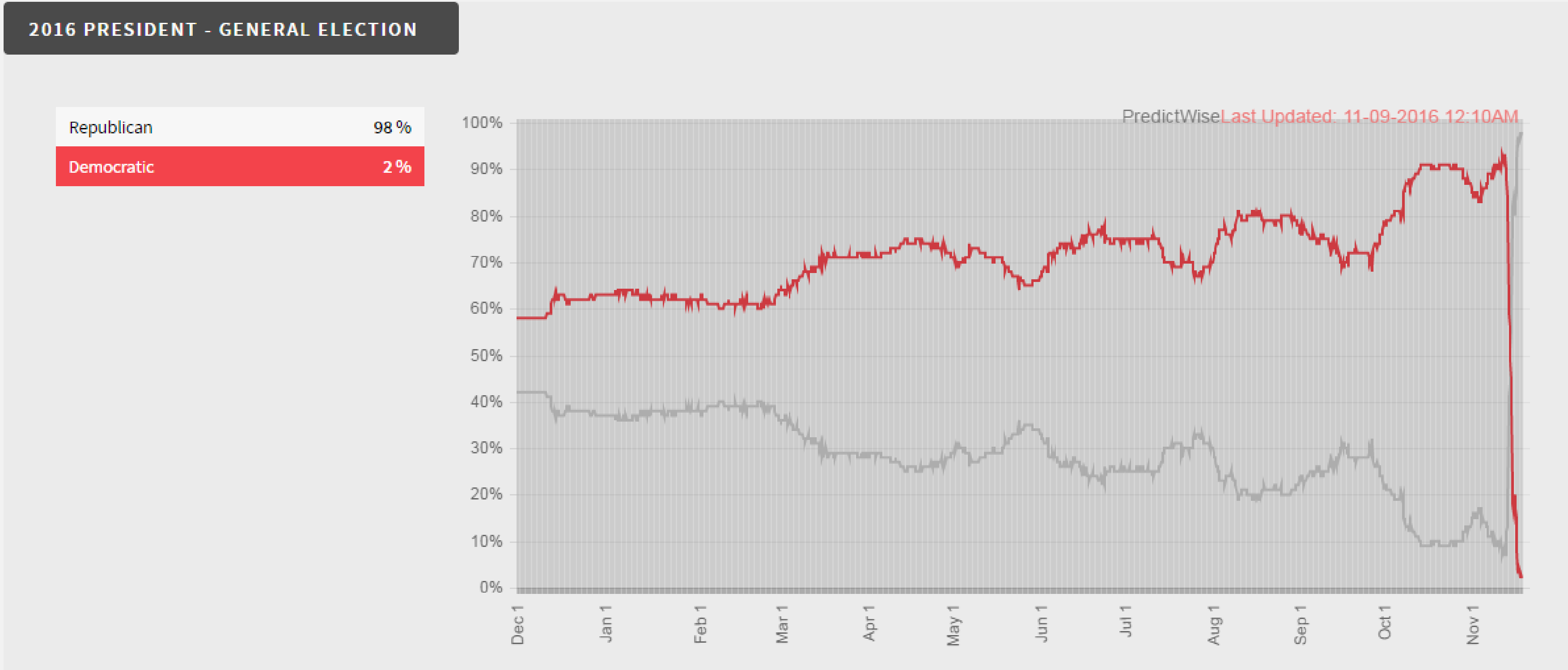


2016 PRESIDENT - GENERAL ELECTION

Democratic	75 %
Republican	25 %

Why Data Analysis

Politics



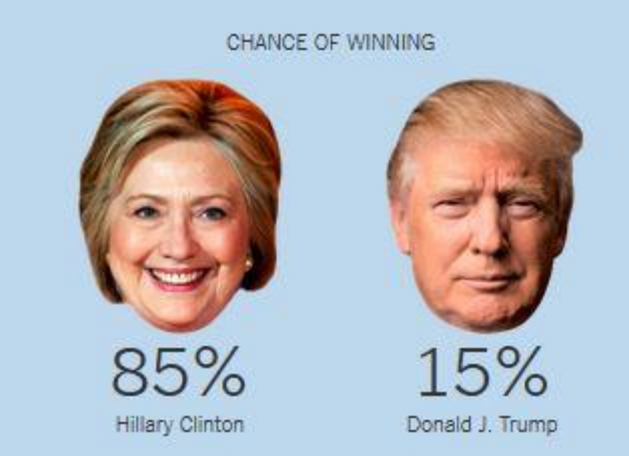
2016 PRESIDENT - GENERAL ELECTION

Republican	98 %
Democratic	2 %

Why Data Analysis

1. Examine data quality - in this election polls were not reaching all likely voters

2. Beware of your own biases: many pollsters were likely Clinton supporters and did not want to question the results that favored their candidate. For example, Huffington Post had forecast 98% chance of Clinton Victory.



http://www.kdnuggets.com/2016/11/trump-shows-limits-prediction.html the definition of data analysis for journalism

What is Data Analysis

What is Data Analysis

The significance of Data Analysis

1. *To obtain new information*
2. *To enlarge the benefits*
3. *To avoid the risks*

What is Data Analysis

INFORMATION DISCOVERY

CONCLUSION SUGGESTING

DECISION SUPPORT

are three objectives of data analysis

What is Data Analysis

Definition:

https://en.wikipedia.org/wiki/Data_analysis

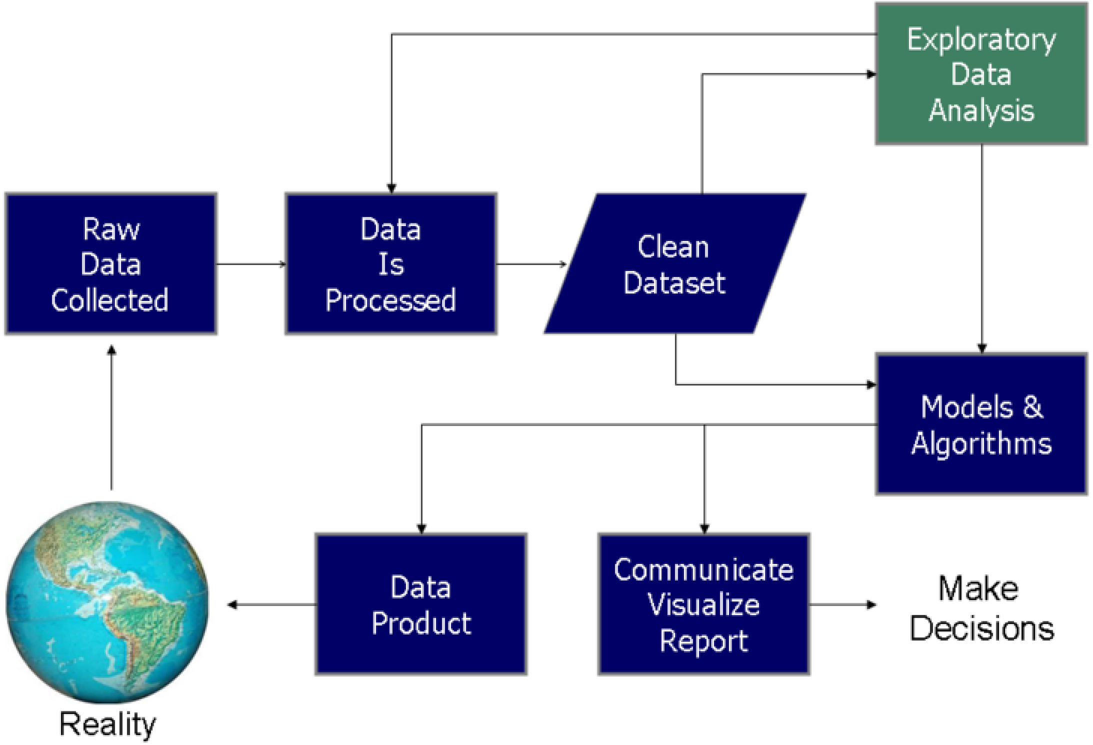
Analysis of data *is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.*

What is Data Analysis

- Two types of decisions:
  - Quantitative Decision with a value
    - Prediction, Regression
  - Qualitative Decision with a label
    - Classification, Clustering

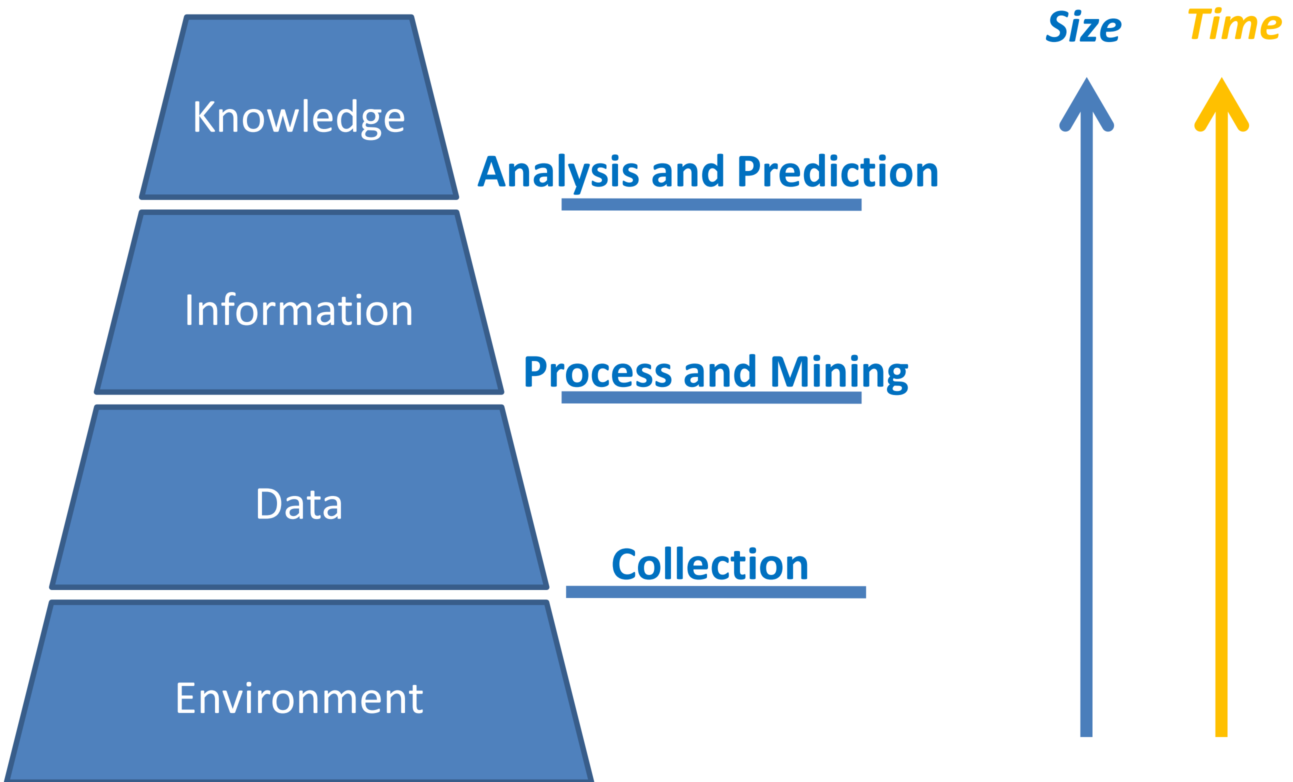
What is Data Analysis

Data Science Process



What is Data Analysis

Relationship between data, information and knowledge



What is Data Analysis

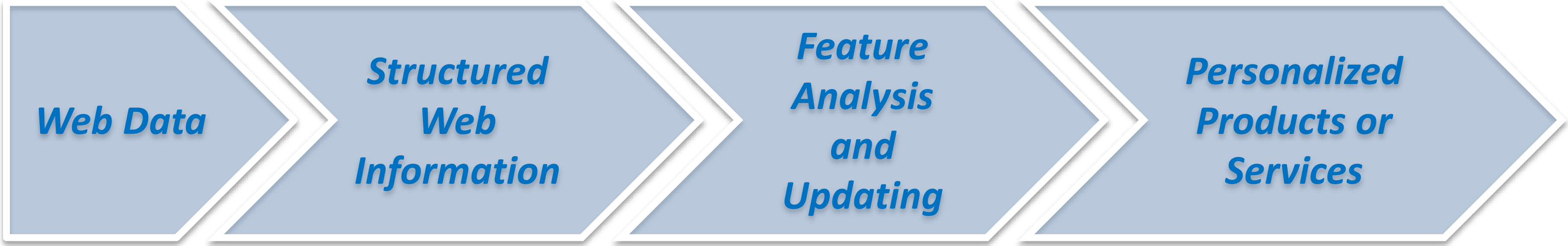
Methodology

1 *Information Acquisition*

2 *Data Cleaning and Information Retrieval*

3 *Knowledge Fusion and Information Updating*

4 *Prediction or Classification*



Web Data

Structured Web Information

Feature Analysis and Updating

Personalized Products or Services

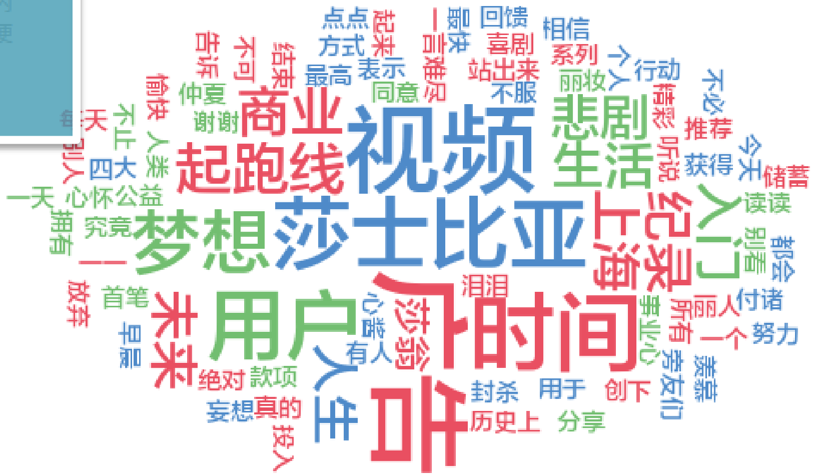
What is Data Analysis

围脖关键词

http://app.thunlp.org/

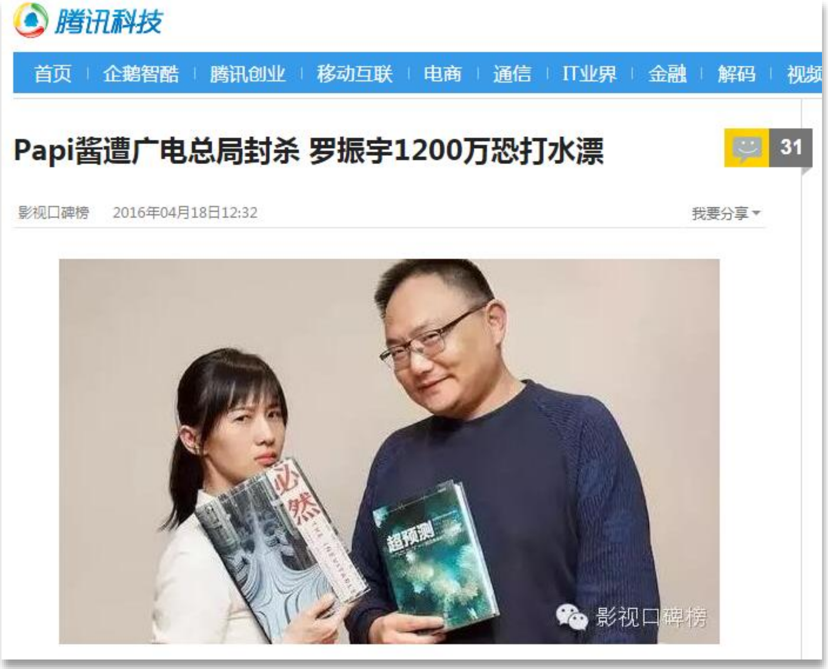
围脖关键词利用自然语言处理的关键词抽取技术，分析用户近期发表微博内容，提取代表用户兴趣的关键词，并采用标签云可视化技术对关键词进行可视化，便于用户快速了解自己、好友、主题等的关键词。

EXAMPLE 3:
papi酱



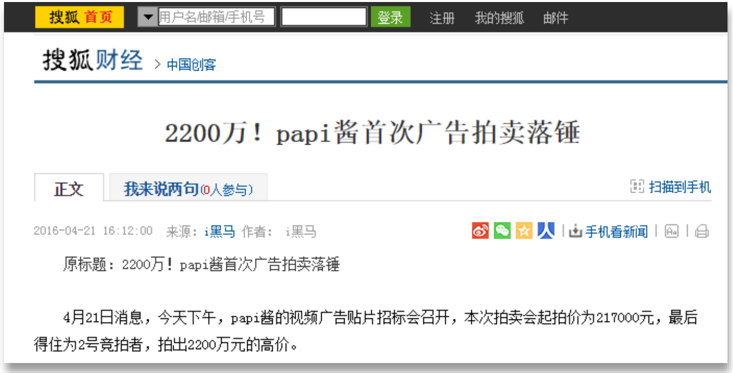
papi酱美女

What is Data Analysis


腾讯科技
首页 | 企鹅智酷 | 腾讯创业 | 移动互联 | 电商 | 通信 | IT业界 | 金融 | 解码 | 视频
31
我要分享
必然
超预测
影视口碑榜

Papi酱遭广电总局封杀 罗振宇1200万恐打水漂

影视口碑榜 2016年04月18日12:32

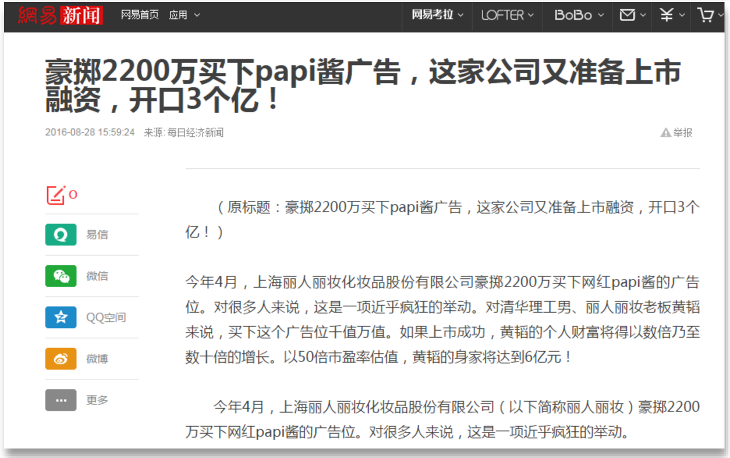


2200万！papi酱首次广告拍卖落锤

2016-04-21 16:12:00 来源：i黑马 作者： i黑马

原标题：2200万！papi酱首次广告拍卖落锤

4月21日消息，今天下午，papi酱的视频广告贴片招标会召开，本次拍卖会起拍价为217000元，最后得住为2号竞拍者，拍出2200万元的高价。



豪掷2200万买下papi酱广告，这家公司又准备上市融资，开口3个亿！

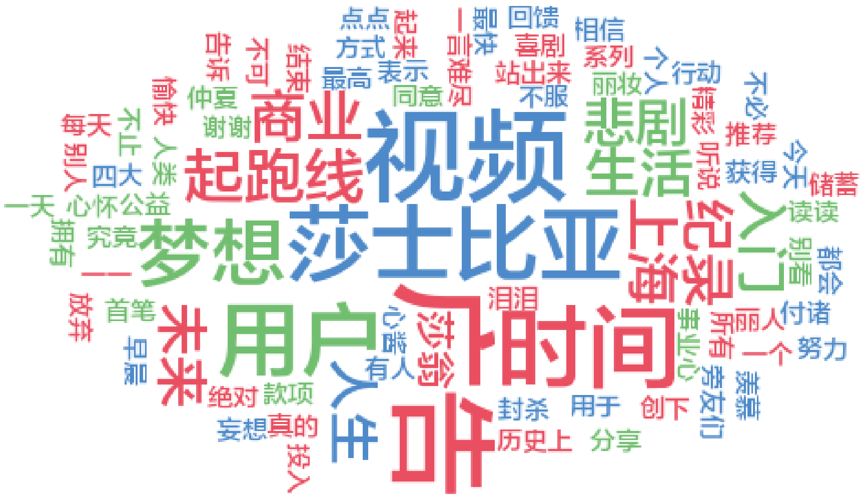
2016-08-28 15:59:24 来源: 每日经济新闻

（原标题：豪掷2200万买下papi酱广告，这家公司又准备上市融资，开口3个亿！）

今年4月，上海丽人丽妆化妆品股份有限公司豪掷2200万买下网红papi酱的广告位。对很多人来说，这是一项近乎疯狂的举动。对清华理工男、丽人丽妆老板黄韬来说，买下这个广告位千值万值。如果上市成功，黄韬的个人财富将得以数倍乃至数十倍的增长。以50倍市盈率估值，黄韬的身家将达到6亿元！

今年4月，上海丽人丽妆化妆品股份有限公司（以下简称丽人丽妆）豪掷2200万买下网红papi酱的广告位。对很多人来说，这是一项近乎疯狂的举动。

What is Data Analysis



papi酱美女

What is Data Analysis

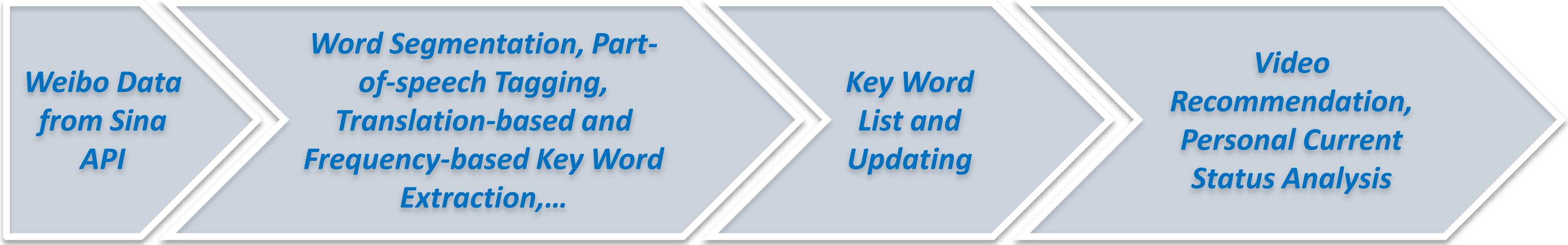
Weibo Key Words for papi酱

1 Information Acquisition

2 Data Cleaning and Information Retrieval

3 Knowledge Fusion and Information Updating

4 Prediction or Classification



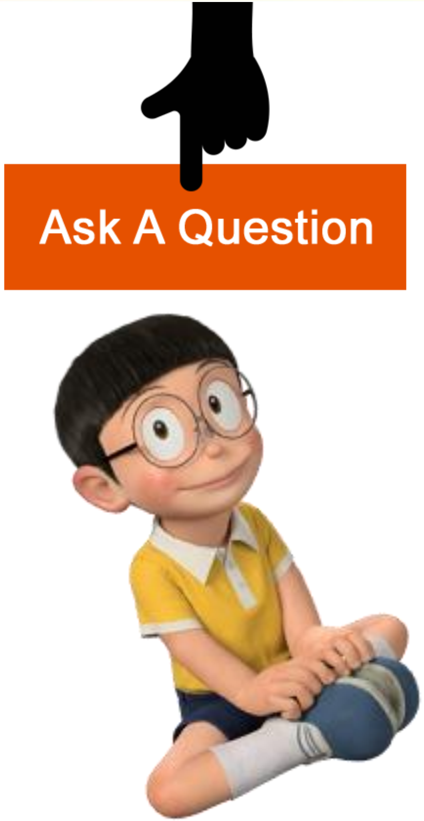
Weibo Data from Sina API

Word Segmentation, Part-of-speech Tagging, Translation-based and Frequency-based Key Word Extraction,…

Key Word List and Updating

Video Recommendation, Personal Current Status Analysis

What is Data Analysis

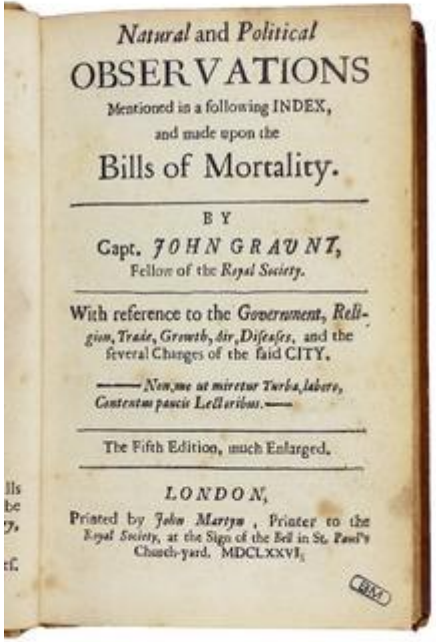


Do you have some methods to forecast the auto sale market in the next several months?

computational journalism will be everywhere in the future

The History of Computer Data Analytics

The History of Computer Data Analytics



Natural and *Political*
OBSERVATIONS
Mentioned in a following INDEX,
and made upon the
Bills of Mortality.

BY
Capt. *JOHN GRAUNT*,
Fellow of the *Royal Society*.

With reference to the *Government*, *Reli-gion*, *Trade*, *Growth*, *Air*, *Diſeaſes*, and the ſeveral Changes of the ſaid CITY.

——*Non, me ut miretur Turba, laboro,*
Contentus paucis Lectoribus.——

The Fifth Edition, much Enlarged.

LONDON,
Printed by *John Martyn*, Printer to the
Royal Society, at the Sign of the Bell in St. *Paul's*
Church-yard. MDCLXXVI.

社会统计 1662

John Graunt (24 April 1620 – 18 April 1674) used statistical analysis to predict the onset and spread of bubonic plague in London, which led him to the Royal Society.

The History of Computer Data Analytics

Thomas Bayes (1701-1761)

贝叶斯决策理论 1763

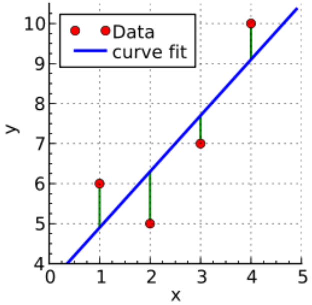
Bayes, Thomas; Price, Mr. (1763). "An Essay towards solving a Problem in the Doctrine of Chances.《机会问题的解法》". Philosophical Transactions of the Royal Society of London. 53 (0): 370–418. doi:10.1098/rstl.1763.0053

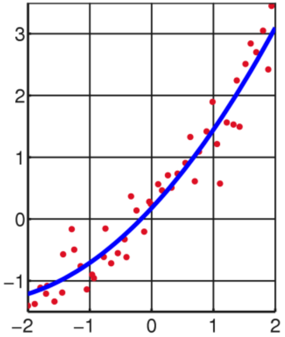
The History of Computer Data Analytics

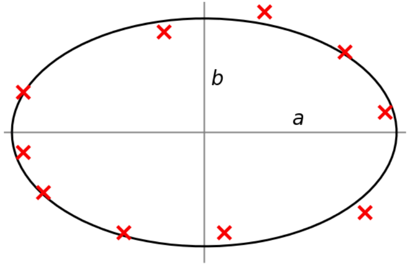
Carl Friedrich Gauss (1777–1855)

最小二乘法1805

Least squares for data fitting and regression





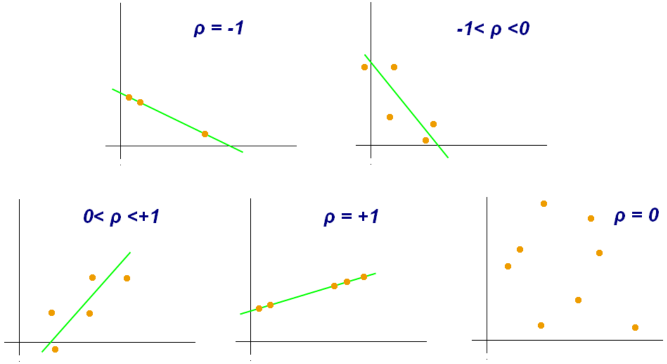


The History of Computer Data Analytics

Karl Pearson
(27 March 1857 – 27 April 1936)

相关系数1880

Pearson Correlation Coefficient

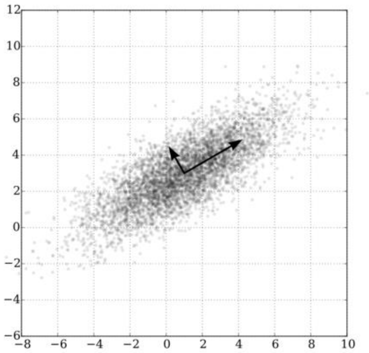


The History of Computer Data Analytics

Karl Pearson
(27 March 1857 – 27 April 1936)

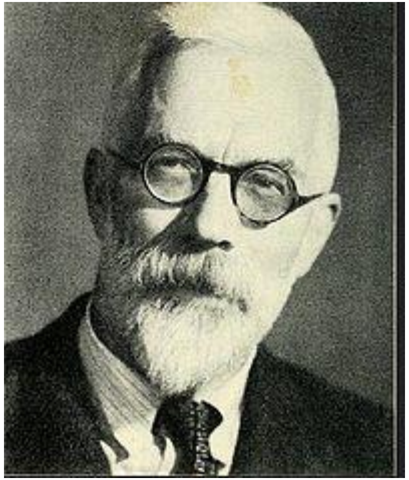
主成分分析1901

Principal Component Analysis



上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

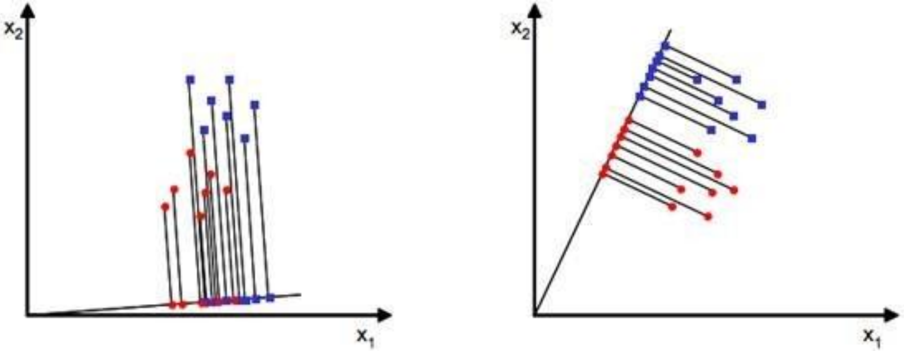
The History of Computer Data Analytics



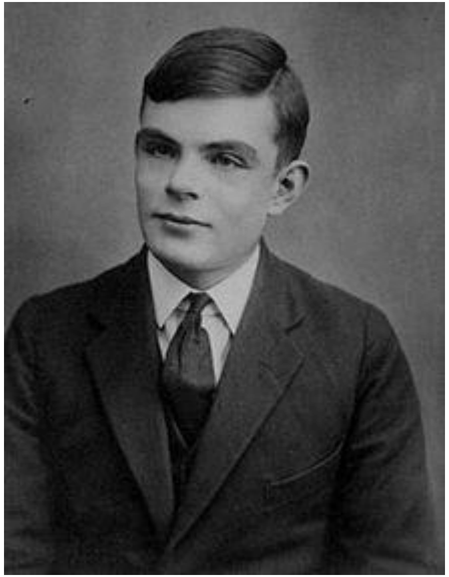
R. A. Fisher
(17 February 1890 – 29 July 1962)

线性判别 1936

Linear Discriminant Aanalysis



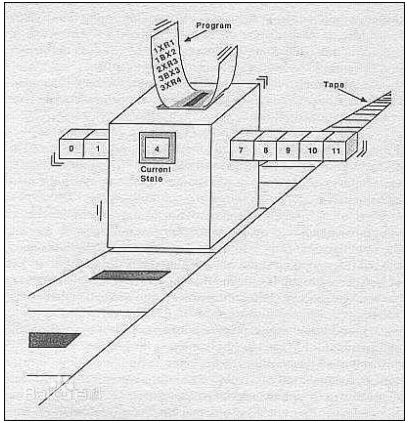
The History of Computer Data Analytics



Alan Turing
(23 June 1912 – 7 June 1954)

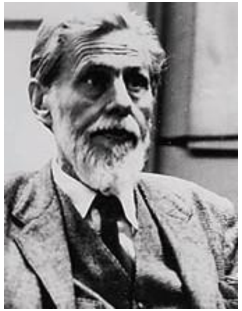
图灵机 1936

Turing Machine



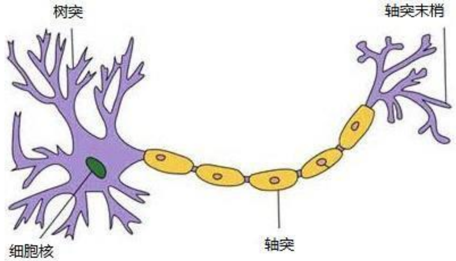
The History of Computer Data Analytics

Warren McCulloch



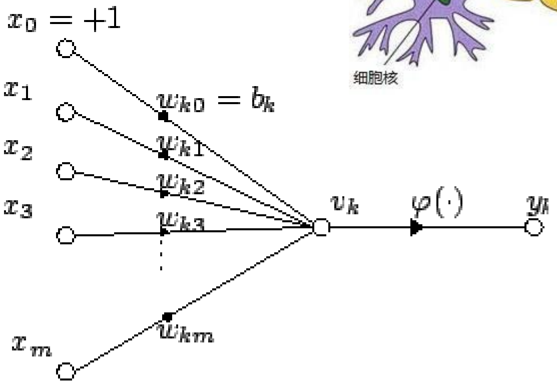
Walter Pitts

人工神经元 1943

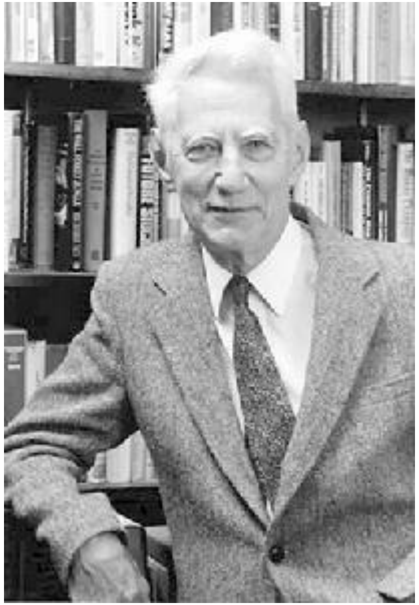


Artificial Neuron

$$y_k = \varphi\left(\sum_{j=0}^{m} w_{kj}x_j\right)$$



The History of Computer Data Analytics



Claude Shannon
(April 30, 1916 – February 24, 2001)

信息论 1948

Information theory

- –Entropy （信息熵）
- –Mutual Information （互信息）

The History of Computer Data Analytics

2006, 50 Years Anniversary

人工智能 1956

Summer Research Project on Artificial Intelligence, Dartmouth College

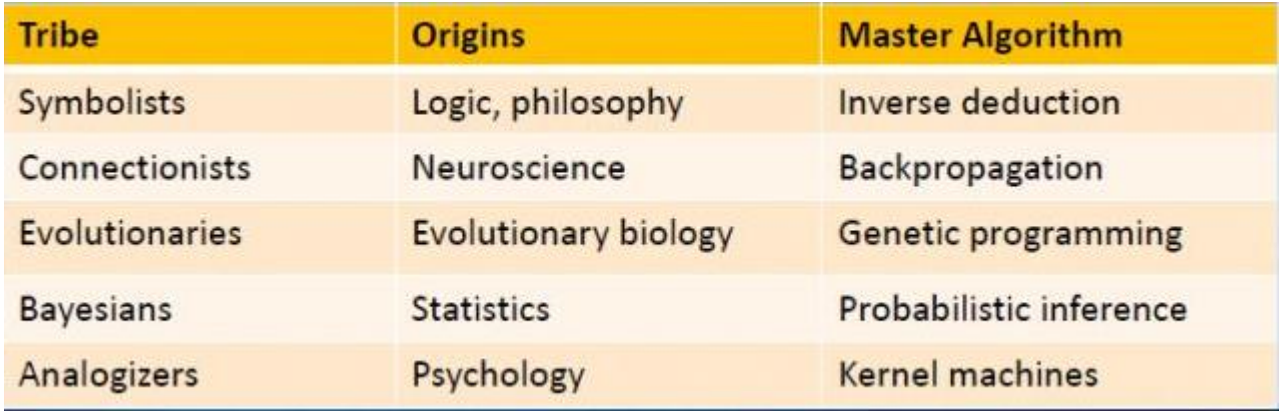
The History of Computer Data Analytics

Prof. Pedro Domingos
University of Washington

2015, ACM

1. 符号主义
2. 联结主义
3. 进化主义
4. 贝叶斯主义
5. 类推主义

Five Tribes in AI



Tribe	Origins	Master Algorithm
Symbolists	Logic, philosophy	Inverse deduction
Connectionists	Neuroscience	Backpropagation
Evolutionaries	Evolutionary biology	Genetic programming
Bayesians	Statistics	Probabilistic inference
Analogizers	Psychology	Kernel machines

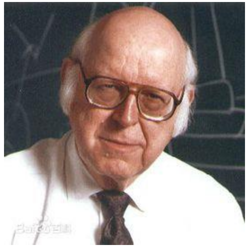
The History of Computer Data Analytics

Avram Noam Chomsky

符号主义 1957
Symbolism

1. Plato is a man.
2. Man will die.
3. Plato will die.

Herbert Simon

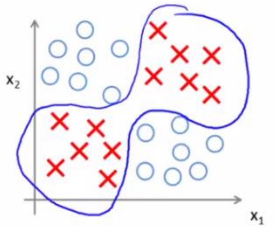


Allen Newell

Expert System
Universal Grammar and Chomsky Hierarchy

The History of Computer Data Analytics

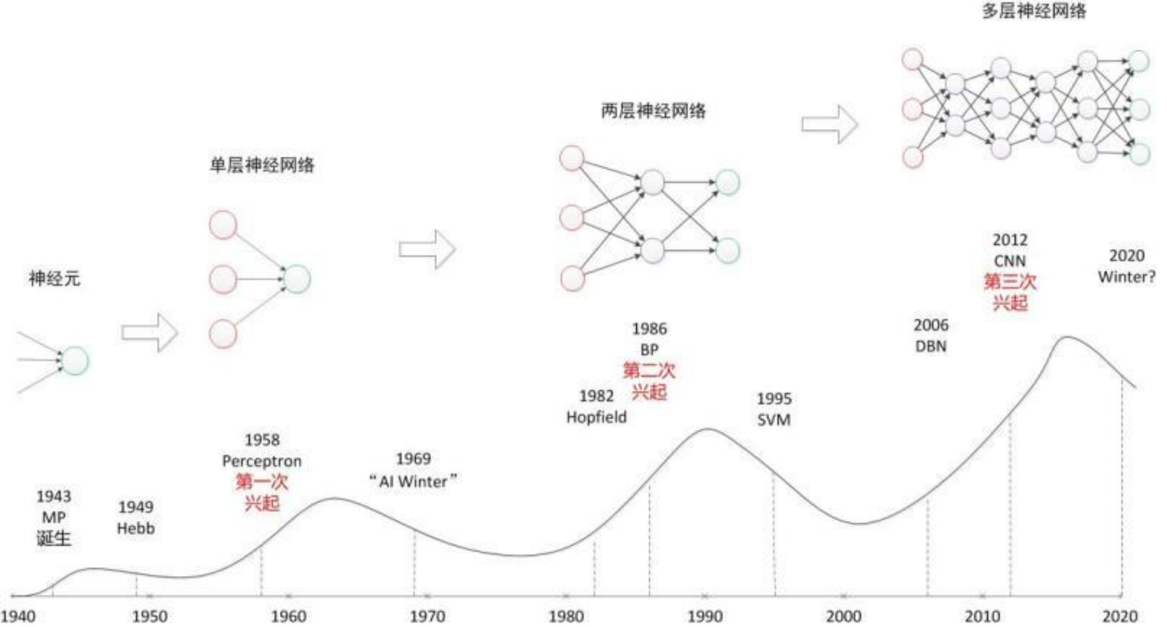
Donald Olding Hebb



联结主义 1943 connectionism

Frank Rosenblatt

Marvin Lee Minsky



上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

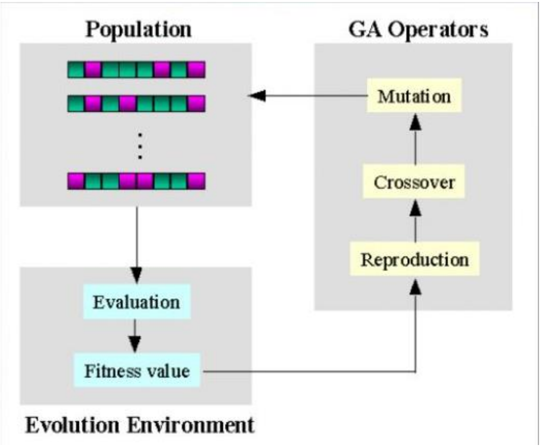
The History of Computer Data Analytics

进化主义 1970’s Evolutionism

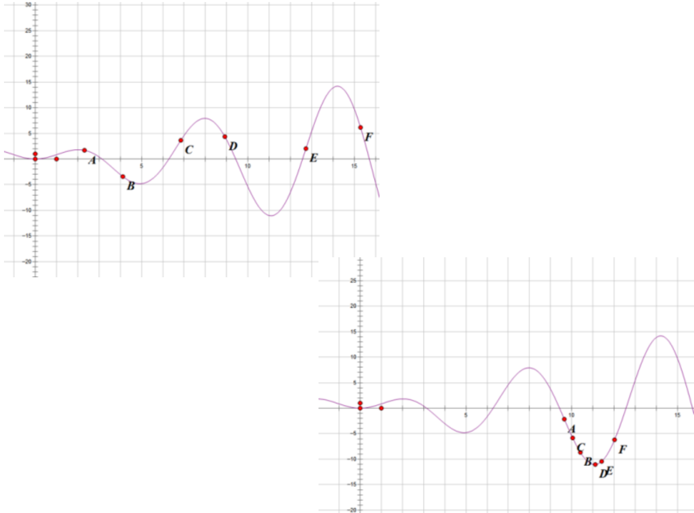
John Henry Holland

Yuhui Shi

Genetic Algorithm



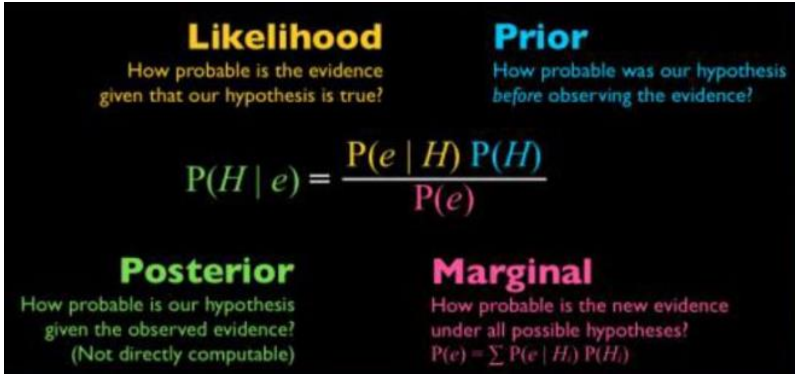
Particle Swarm Optimization



The History of Computer Data Analytics

Judea Pearl

贝叶斯主义 1763 Bayesianism



Likelihood
How probable is the evidence given that our hypothesis is true?

Prior
How probable was our hypothesis *before* observing the evidence?

$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)}$$

Posterior
How probable is our hypothesis given the observed evidence?
(Not directly computable)

Marginal
How probable is the new evidence under all possible hypotheses?
$P(e) = \sum P(e \mid H_i)\, P(H_i)$

The History of Computer Data Analytics

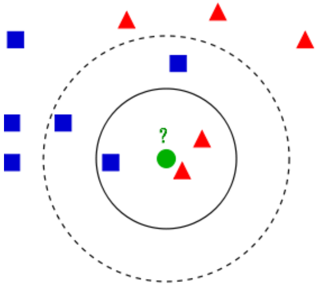
类推主义 1951

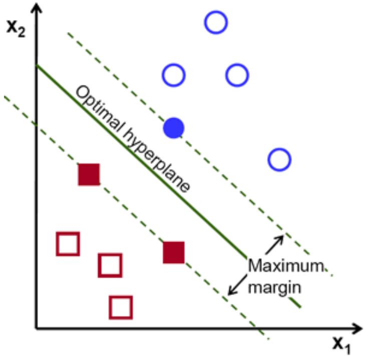
Analogism

Vladimir Vapnik

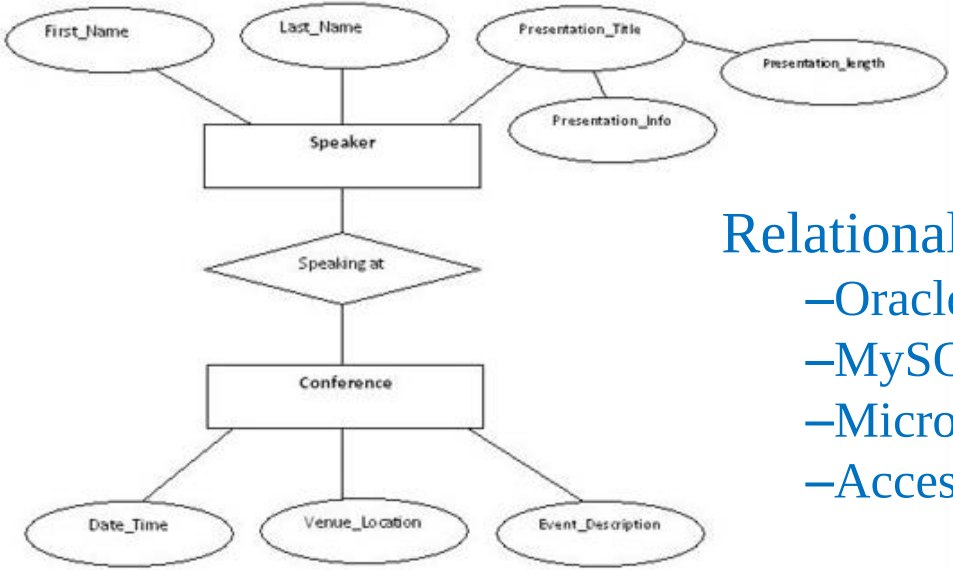
K-Nearest Neighbour

Support Vector Machine





The History of Computer Data Analytics



Entity Relationship Diagram

数据库 1951

Relational Database:

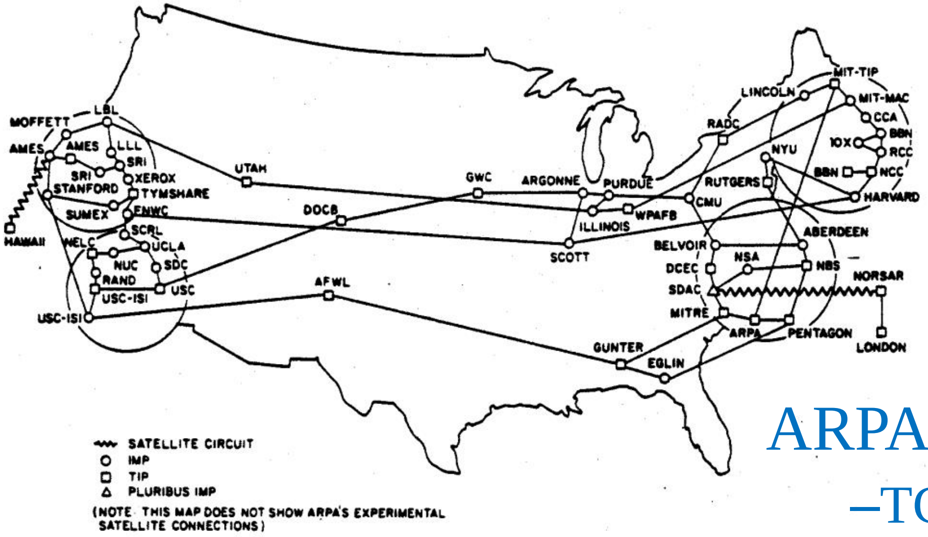
- –Oracle
- –MySQL
- –Microsoft SQL Server
- –Access

NoSQL:

- – MongoDb

上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

The History of Computer Data Analytics



互联网1969

ARPA Net:

- –TCP/IP

The Internet:

- –World Wide Web
- –Email

The History of Computer Data Analytics

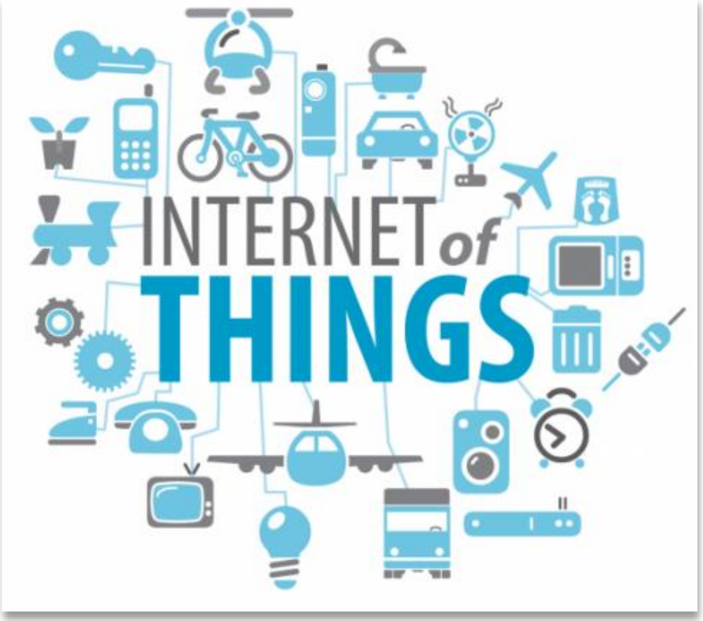
Web2.0 2004

Web 1.0 Contents made by Providers

Web 2.0 Contents made by Customers

Semantic Web: Web 3.0?

The History of Computer Data Analytics



物联网 2005

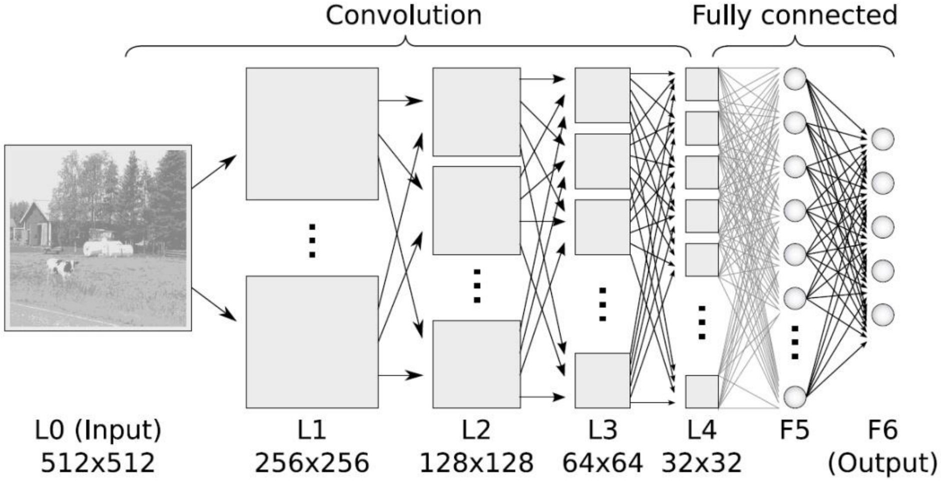
The Internet of Things

The History of Computer Data Analytics

Geoffery Hinton

深度学习 2006

Deep Learning



The History of Computer Data Analytics

云计算 2006

Cloud Computing

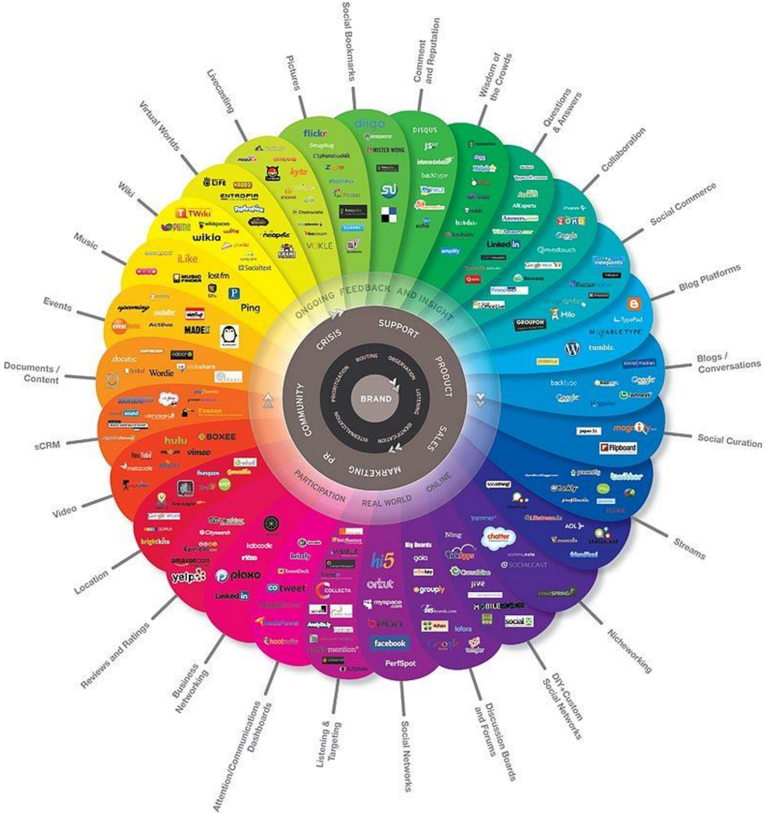
- –Grid Computing
- –Distributed Computing
- –Parallel Computing
- –Utility Computing
- –…

Google

The History of Computer Data Analytics

大数据2008

The History of Computer Data Analytics

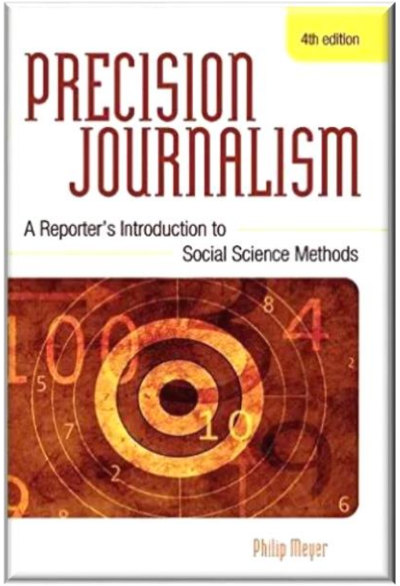


社交媒体的繁荣

2000’s

Social Media

The History of Computer Data Analytics



数据新闻1967

Data Analysis in Journalism

Philip Meye, Precision Journalism, 1970’s

The National Institute for Computer-Assisted Reporting – NICAR, 1994

Computational Journalism, Georgia Institute of Technology, 2006

Frontiers of Computational Journalism, Columbia Journalism School, 2012

Masters in Computational Journalism, Syracuse University, 2015

Computational Journalism Lab, Stanford University, 2015

上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

knowledge domains of computational journalism

Domains of New Media Data Analytics

Domains of New Media Data Analytics

Relevant Disciplines

- – Journalism
- – Computer Science
- – Mathematics
- – Psychology
- – Economics
- – Politics
- – Linguistics
- – ….

Domains of New Media Data Analytics

Corresponding Technologies

- – Computer Science
- – Artificial Intelligence
- – Machine Learning
- – Statistical Analysis
- – Natural Language Processing
- – Pattern Recognition
- – …

Home Work

Home Work

1. Identify at least three major side effects of information sharing on social media.

2. Rumors spread rapidly on social media. Can you think of some methods to block the spread of rumors on social media?

The End of Lecture 1

Thank You

http://www.wangting.ac.cn