QINGLU青鹿

数据分析那些事

The work of data analysis

分享人：杨毅

时间:2020/8/27

数据的获取1
The origin of data

通过**数据源网站**获取数据

1 数据报告

2 学术文献

3 统计信息

艾瑞咨询	镝数聚	QuestMobile
豆丁报告网	199IT	Mob研究院
谷歌学术	Sci-hub	百度学术
中国统计信息网	国家统计网	中国产业信息网

搜索引擎：百度、Google、Bing　　限制网站搜索site

数据的获取2
The origin of data

通过爬虫获取获取数据

爬取国家资源平台资源信息

网页的结构为树状结构

HtmlAgilityPack

利用Xpath爬取

GeckoFx

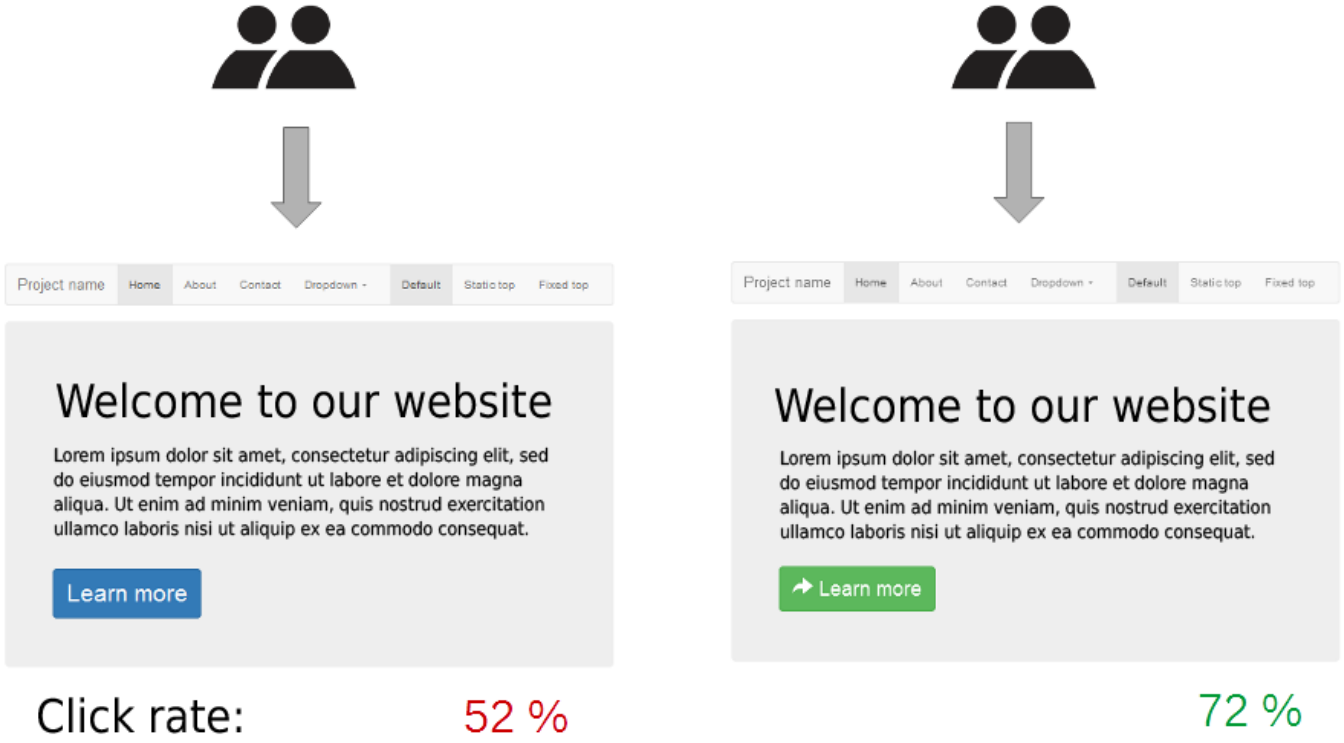
使用浏览器加载JS渲染后的网页爬取数据

数据的获取3
The origin of data

QINGLU青鹿

利用AB Test 获取数据

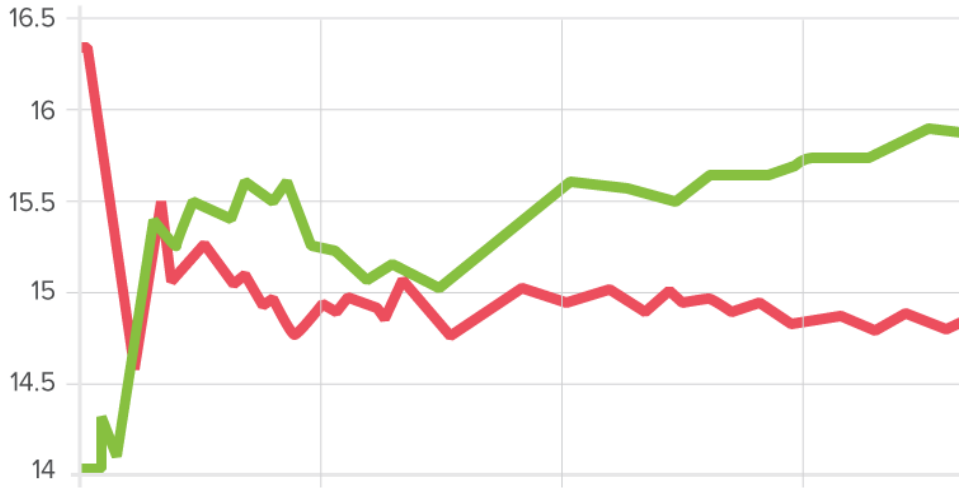
A/B Test 基本原理



Click rate: 52 %

72 %

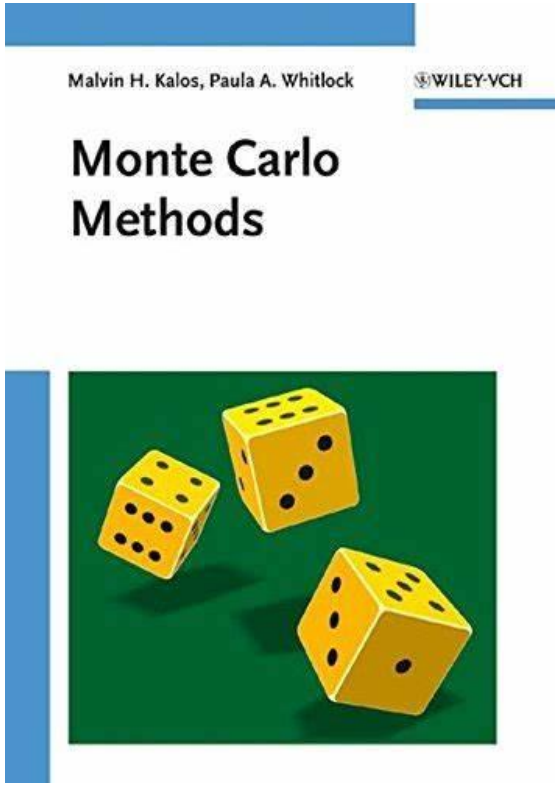
A/B Test 结果



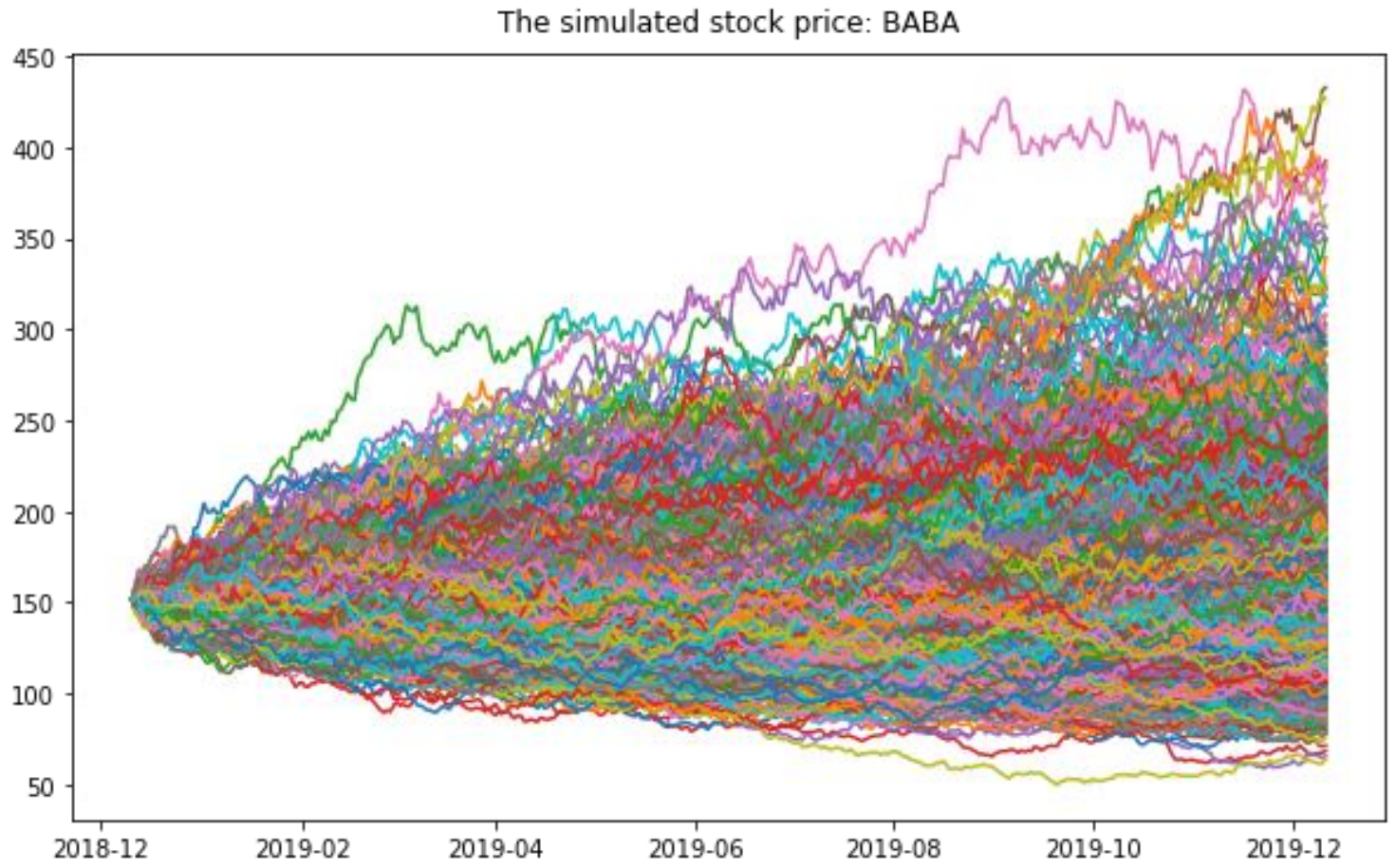
数据的获取4
The origin of data

QINGLU青鹿

利用Monte Carlo Method获取数据



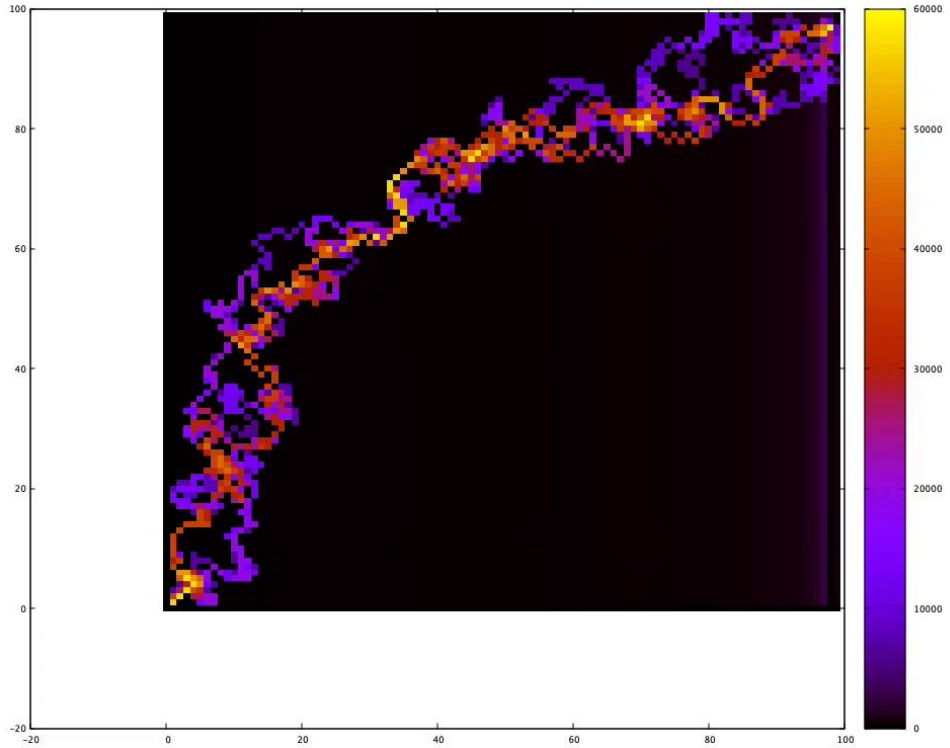
利用计算机产生随机，进行模拟



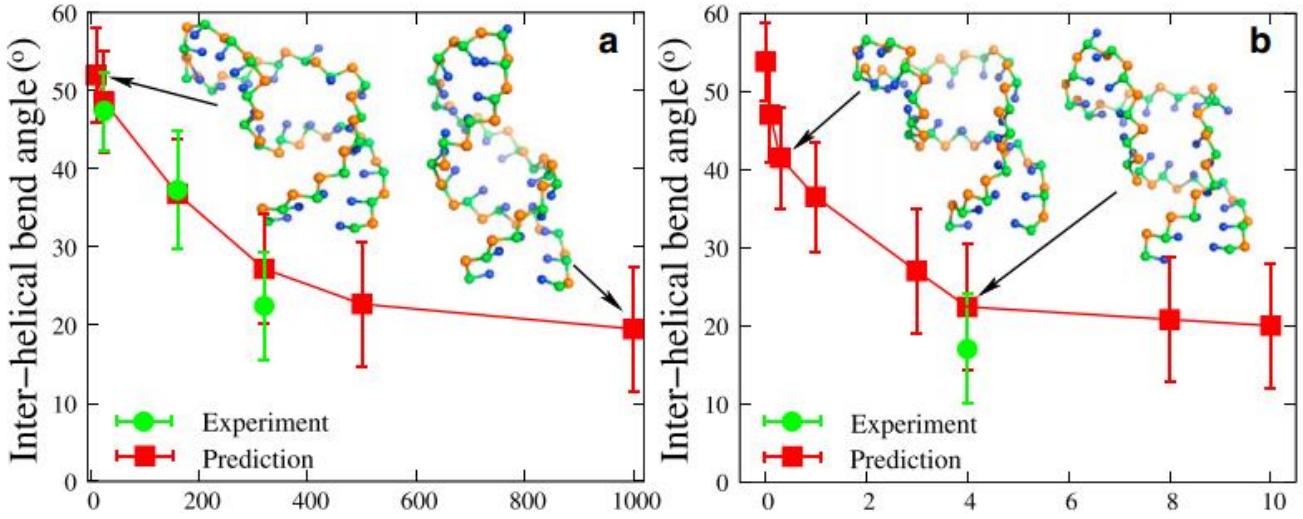
数据的获取4
The origin of data

QINGLU青鹿

利用Monte Carlo Method获取数据



仅考虑信息素的蚂蚁觅食行为模拟



不同浓度的Na与Mg离子溶液下RNA结构的预测

Predicting 3D structure, flexibility and stability of RNA hairpins in monovalent and divalent ion solutions[J]. Biophysical Journal, 2015, 109(12):2654-2665.

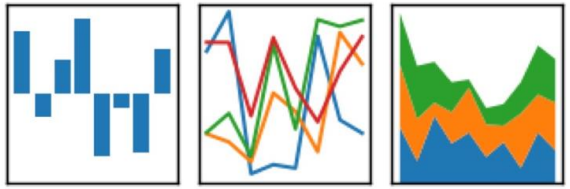
数据的清洗
The clean of data

QINGLU青鹿

利用Pandas进行数据的读取与清洗

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



数据透视表
Pivot Table

df

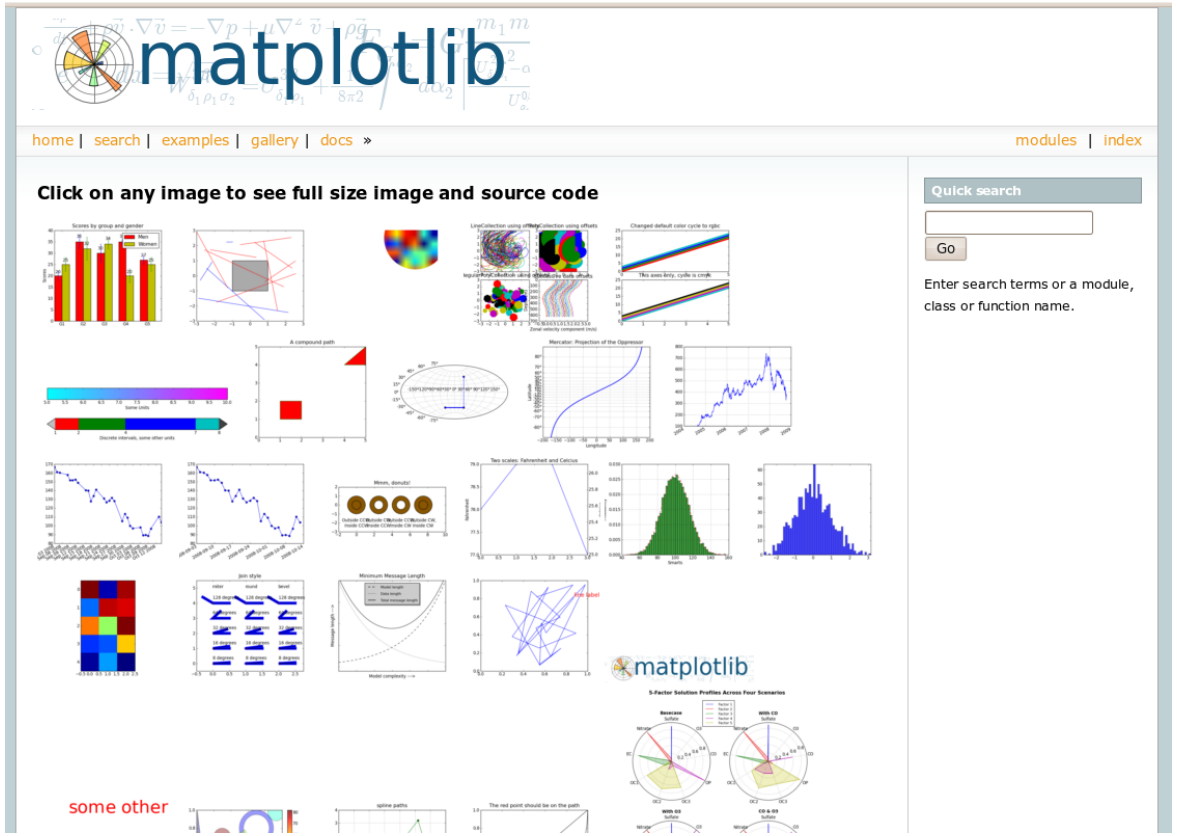
	foo	bar	baz	zoo
0	one	A	1	x
1	one	B	2	y
2	one	C	3	z
3	two	A	4	q
4	two	B	5	w
5	two	C	6	t

```
df.pivot(index='foo',
         columns='bar',
         values='baz')
```

bar	A	B	C
foo			
one	1	2	3
two	4	5	6

数据的观察
The exploration of data

通过作图进行初步的数据探索



https://dycharts.com/

数据的观察
The exploration of data

通过计算基本统计量来探索数据

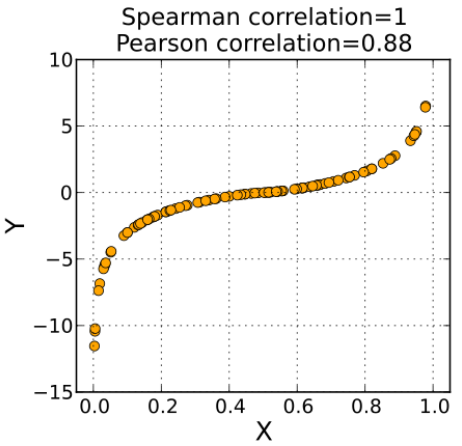
Sample mean: $\bar{x} = \frac{\sum x}{n}$	Sample median: Number that's in the middle once you order all the values.	Sample standard deviation: $s = \sqrt{\frac{(x-\bar{x})^2}{n-1}}$	Sample standard variance: $s^2 = \frac{(x-\bar{x})^2}{n-1}$
Z-value: $Z = \frac{x-\mu}{\sigma}$	Z-value based on the average: $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$	Confidence interval One population mean, σ known: $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$	Confidence interval One population mean, σ unknown: $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$
Hypothesis test One population mean, σ known: $Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	Hypothesis test One population mean, σ unknown: $t_{n-1} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	Normal approximation to the binomial $Z = \frac{x-np}{\sqrt{np(1-p)}}$	Paired t-test $t_{n-1} = \frac{\bar{d}-0}{s_d/\sqrt{n}}$
Confidence interval One population proportion: $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	Confidence interval Two population proportions: $(\hat{p}_1-\hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	Confidence interval Two population means: $(\bar{x}-\bar{y}) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	
Hypothesis test One population proportion: $Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	Hypothesis test Two population proportions: $Z = \frac{(\hat{p}_1-\hat{p}_2)-0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$	Hypothesis test Two population means: $Z = \frac{(\bar{x}-\bar{y})-0}{\sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}}$	

Pearson correlation coefficient

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var[X]\,Var[Y]}}$$

spearman correlation coefficient

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$

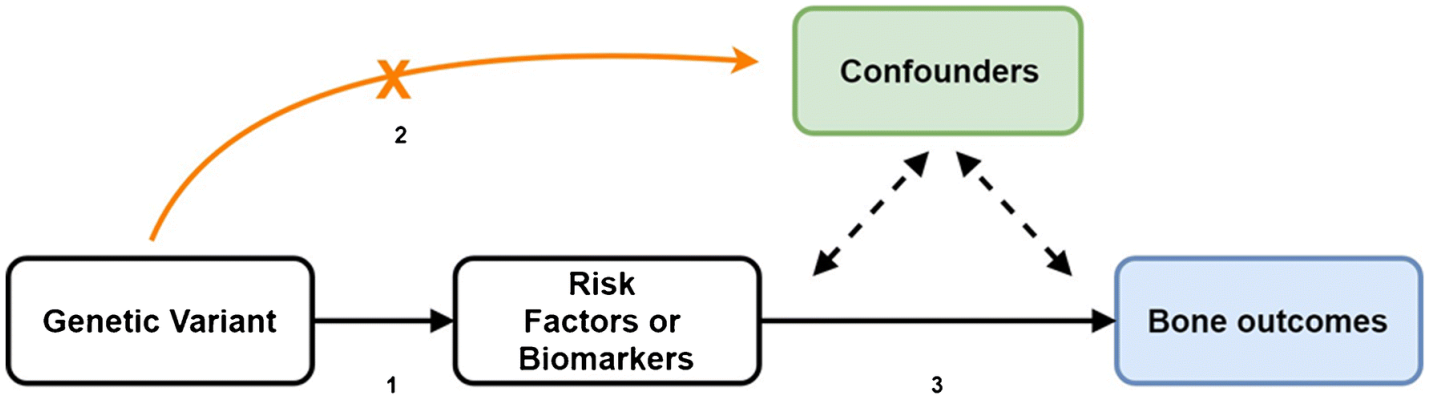


相关性并不等于因果性，但可以通过mendelian randomization分析因果性

QINGLU青鹿

孟德尔随机化分析

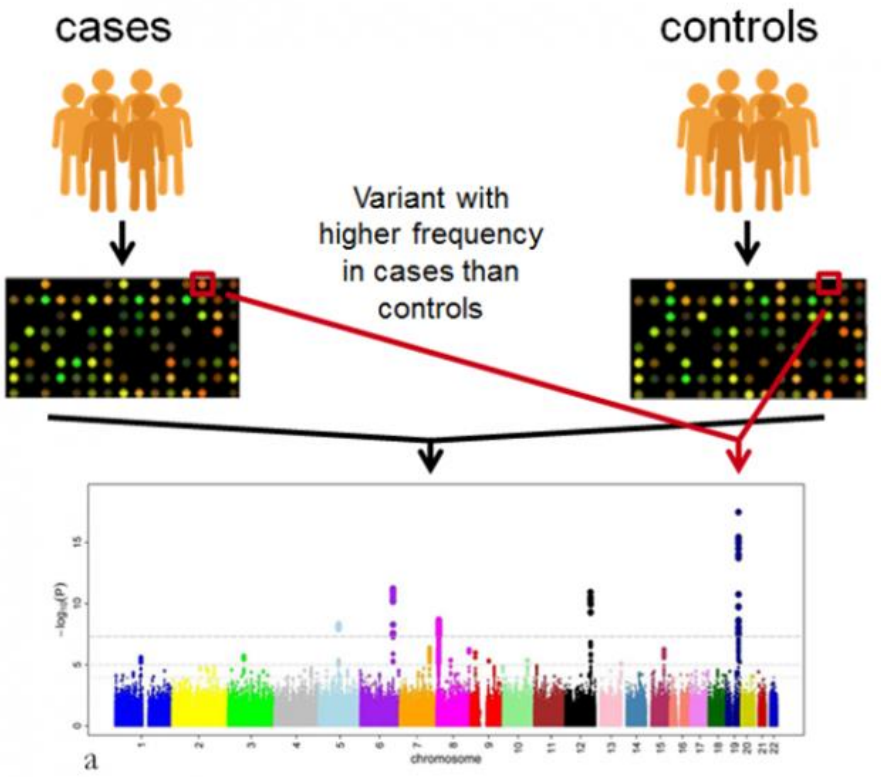
Mendelian randomization



Assumptions of Mendelian Randomization Study:

Genetic variants are associated with the risk factor
Genetic variants are not associated with confounders
Genetic variants influence bone outcomes only through the risk factor

GWAS



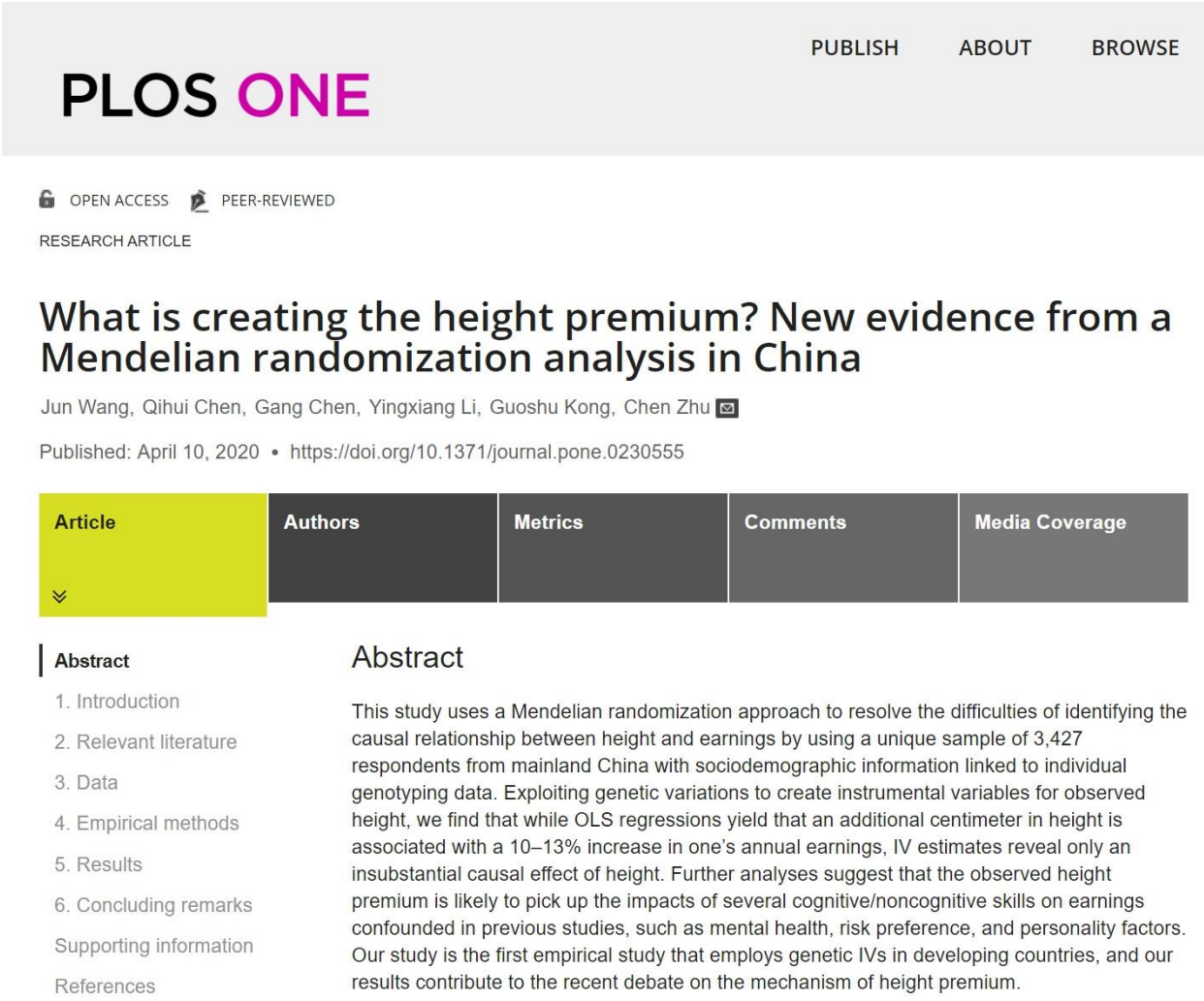
https://en.wikipedia.org/wiki/Mendelian_randomization

数据的观察
The explore of data

孟德尔随机化分析

为什么在劳动力市场中普遍存在身高溢价现象（即“身高越高收入越高”）？其背后机制是什么？

OLS回归结果表明身高每增加1厘米，一个人的年收入就会增加10-13%；但MR估计结果显示这并非实质性因果关系。身高溢价现象主要与多种认知/非认知技能对收入的影响有关



PLOS ONE

PUBLISH ABOUT BROWSE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

What is creating the height premium? New evidence from a Mendelian randomization analysis in China

Jun Wang, Qihui Chen, Gang Chen, Yingxiang Li, Guoshu Kong, Chen Zhu

Published: April 10, 2020 • https://doi.org/10.1371/journal.pone.0230555

Article | Authors | Metrics | Comments | Media Coverage

Abstract
1. Introduction
2. Relevant literature
3. Data
4. Empirical methods
5. Results
6. Concluding remarks
Supporting information
References

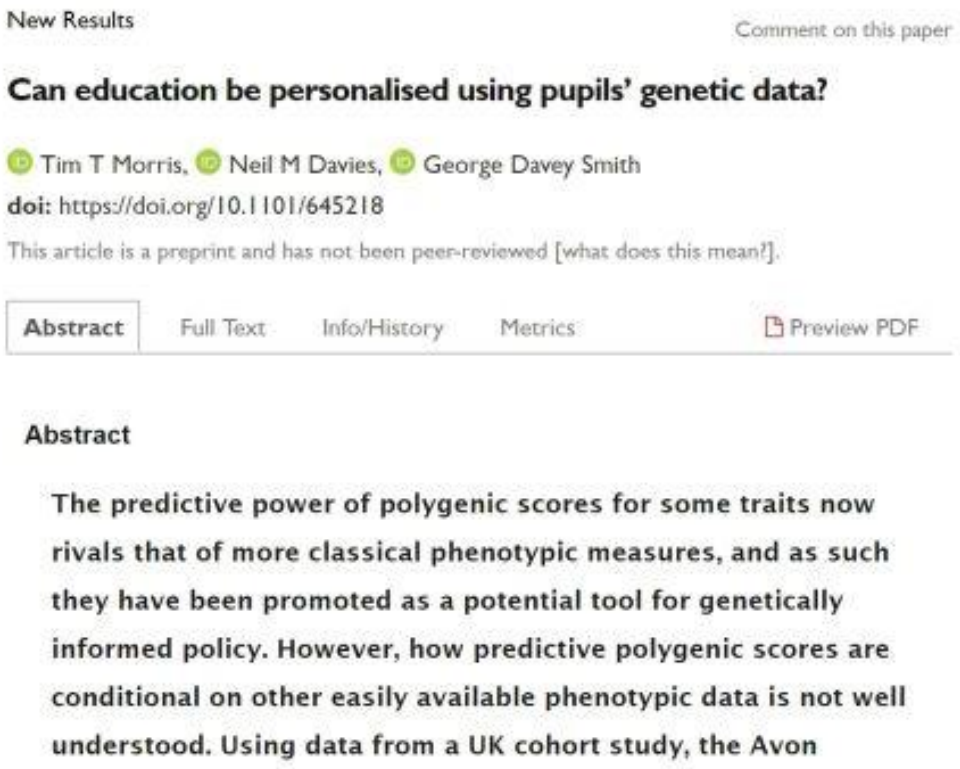
Abstract

This study uses a Mendelian randomization approach to resolve the difficulties of identifying the causal relationship between height and earnings by using a unique sample of 3,427 respondents from mainland China with sociodemographic information linked to individual genotyping data. Exploiting genetic variations to create instrumental variables for observed height, we find that while OLS regressions yield that an additional centimeter in height is associated with a 10–13% increase in one's annual earnings, IV estimates reveal only an insubstantial causal effect of height. Further analyses suggest that the observed height premium is likely to pick up the impacts of several cognitive/noncognitive skills on earnings confounded in previous studies, such as mental health, risk preference, and personality factors. Our study is the first empirical study that employs genetic IVs in developing countries, and our results contribute to the recent debate on the mechanism of height premium.

数据的观察
The explore of data

GWAS

教育成就多基因评分到底在何种程度上能够体现甚至是预测学生的实际学业表现？


New Results

Comment on this paper

Can education be personalised using pupils' genetic data?

Tim T Morris, Neil M Davies, George Davey Smith

doi: https://doi.org/10.1101/645218

This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract | Full Text | Info/History | Metrics | Preview PDF

Abstract

The predictive power of polygenic scores for some traits now rivals that of more classical phenotypic measures, and as such they have been promoted as a potential tool for genetically informed policy. However, how predictive polygenic scores are conditional on other easily available phenotypic data is not well understood. Using data from a UK cohort study, the Avon

基因影响我们的收入和财富积累吗？每一个人最终的财富水平（至少部分程度上）是与生俱来的吗？


Journal of Political Economy

Just Accepted

SUBSCRIBE/RENEW | BROWSE ISSUES | FORTHCOMING | CONTRIBUTORS

Previous Article | Next Article

Genetic Endowments and Wealth Inequality

Daniel Barth, **Nicholas W. Papageorge**, and **Kevin Thom**

Dr. Nicholas W. Papageorge (papageorge@jhu.edu)

ACCEPTED: July 11, 2019

数据的分析
The analysis of data

由样本数据估计总体情况

提出问题
Questions

1 2019年进行的**题库抽查数据**是否真的是我们的题库的真实情况？**学库宝题库的数据**是否真的是反映了其真实情况？

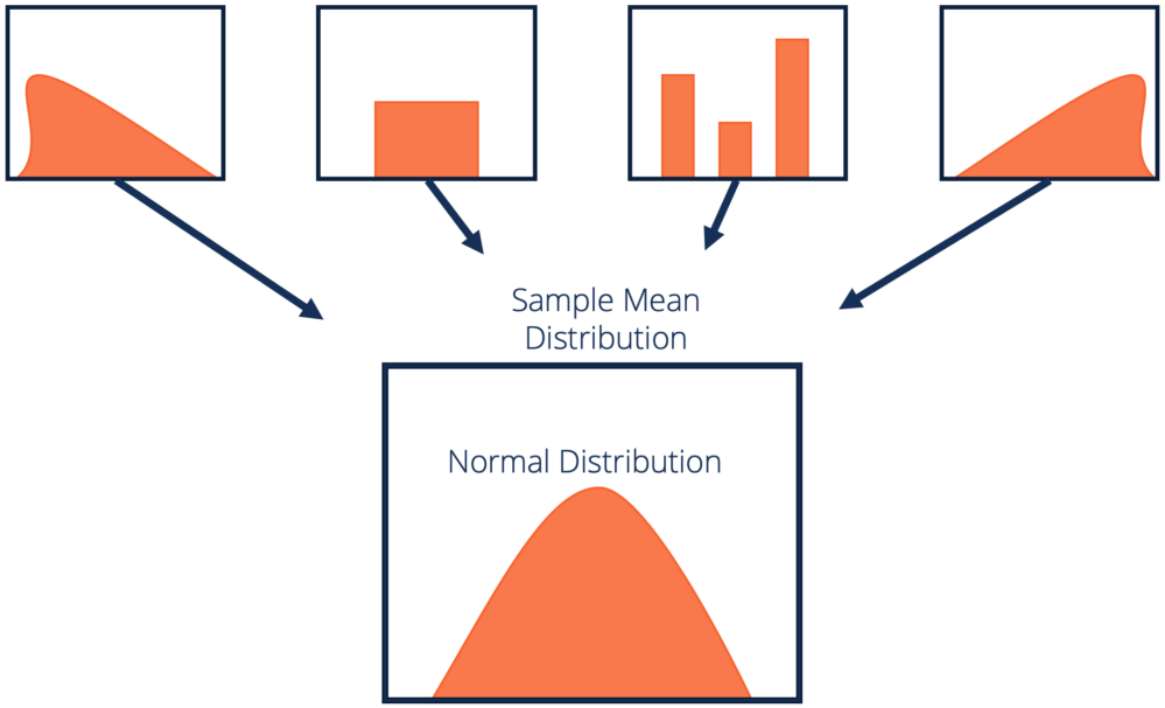
2 研究院进行的**教育研究实验**是否真实的反映了我们产品功能的特性？

3 技服同事记录的**产品的各种不稳定性**是否真实反映了我们的产品的问题？

4 我们在学校收集的**老师的意见和需求**是否真实的反应了绝大部分老师的需求？

中心极限定理

Central limit theorem



$$X_i \xrightarrow{i.i.d.} (\mu, \sigma^2)$$

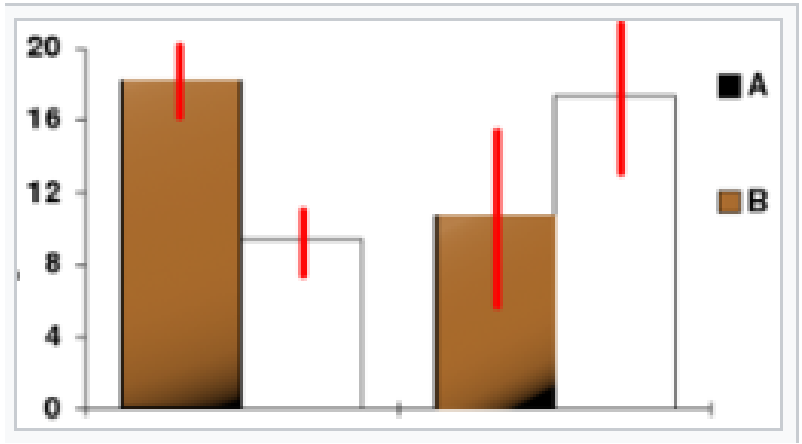
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

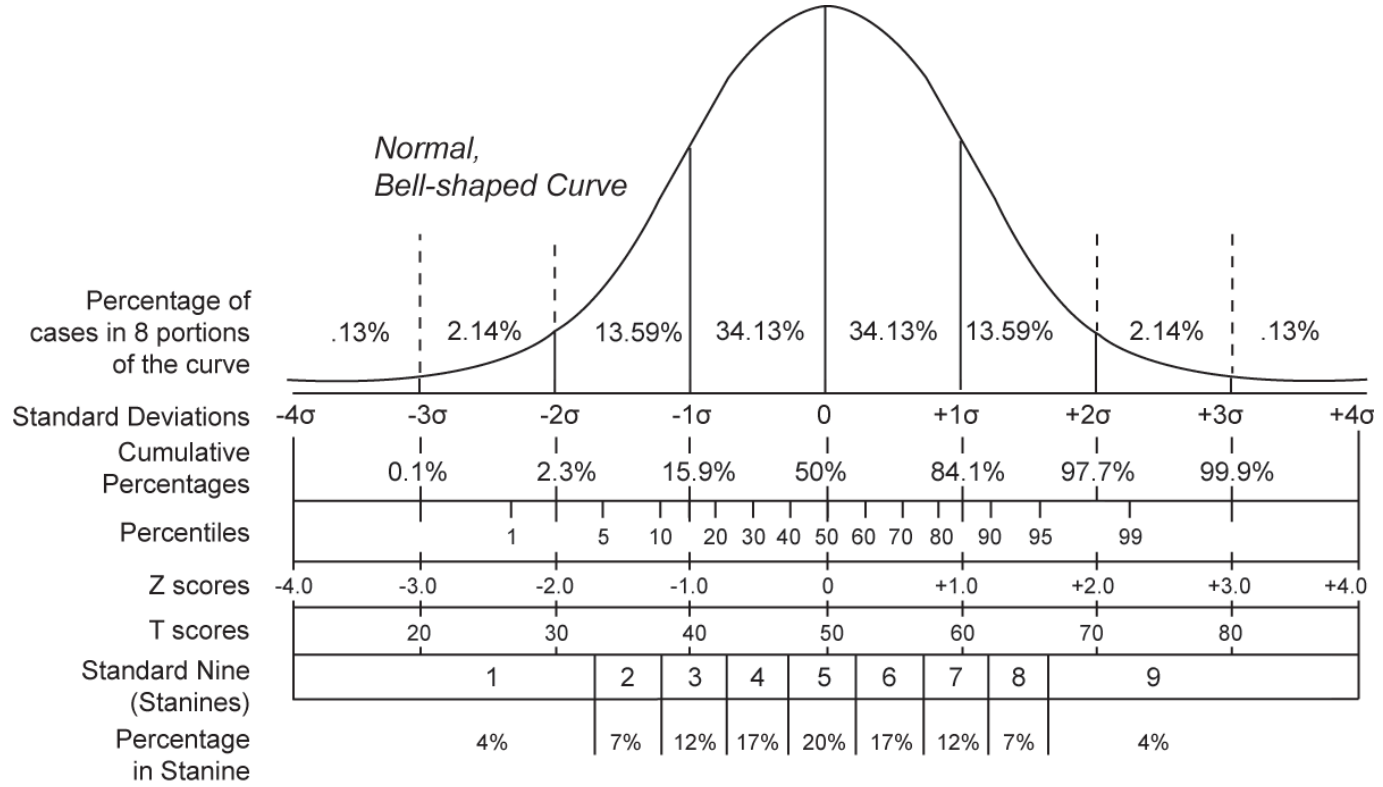
$$\bar{X} \to N\left(\mu, \frac{\sigma^2}{n}\right)$$

数据的分析
The analysis of data

QINGLU青鹿

置信区间



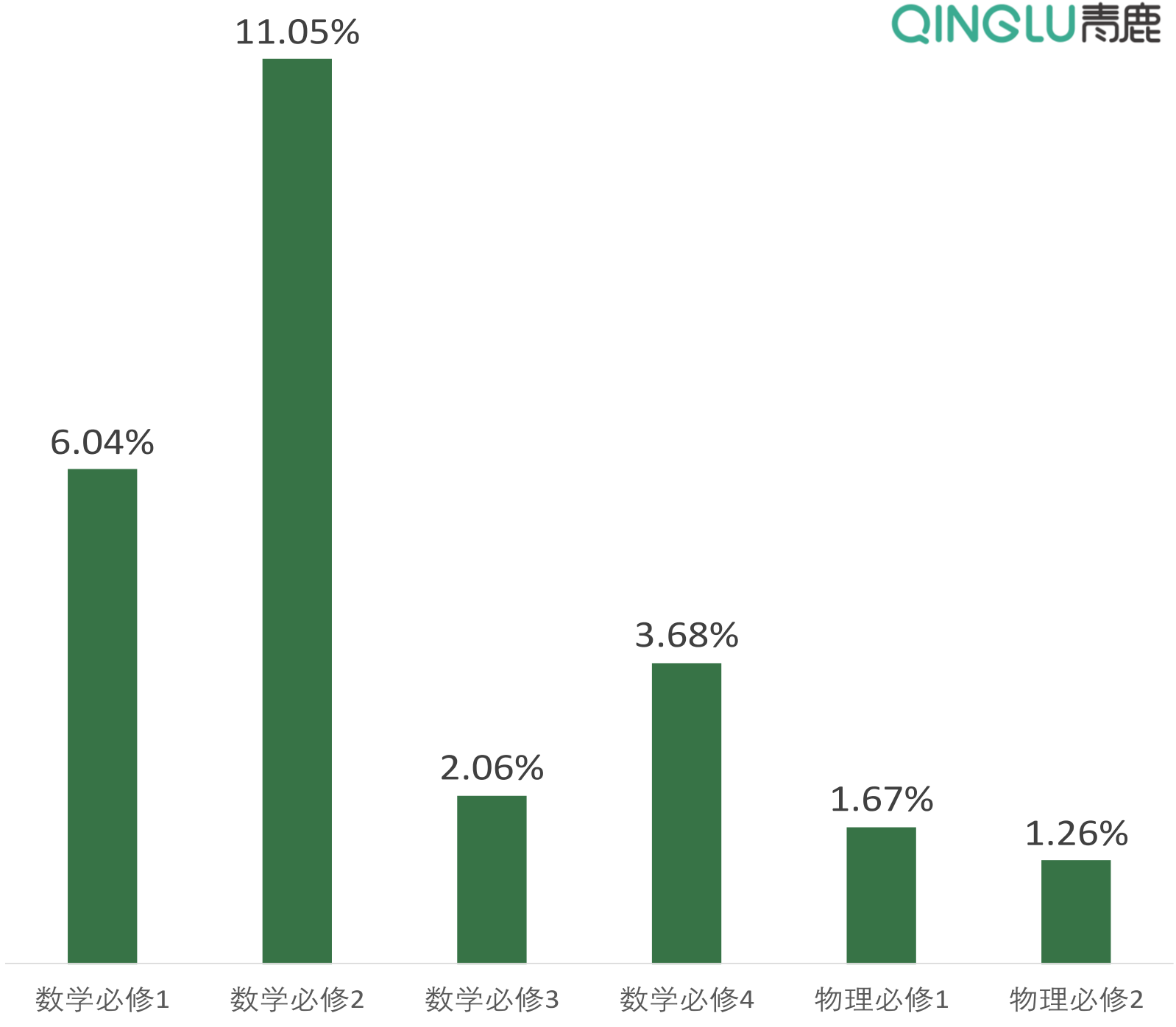


青鹿资源平台题库出错率

高中数学、高中物理题库平均错误率

3.58%

*数据来源与2019年度资源团队

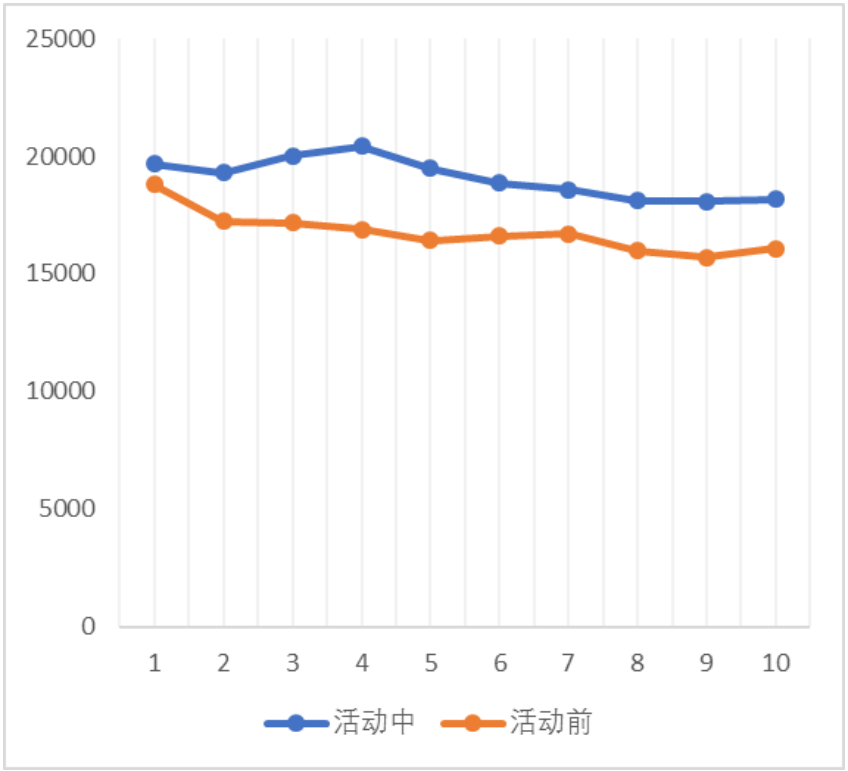


数据的分析
The analysis of data

QINGLU青鹿

假设检验

活动进程	日期	UV
活动中	7月20日	19684
	7月19日	19282
	7月18日	20017
	7月17日	20428
	7月16日	19476
	7月15日	18865
	7月14日	18568
	7月13日	18091
	7月12日	18060
	7月11日	18157
活动前	7月10日	18787
	7月9日	17240
	7月8日	17164
	7月7日	16901
	7月6日	16411
	7月5日	16617
	7月4日	16695
	7月3日	15967
	7月2日	15704
	7月1日	16089



16757.5

APP的UV从活动前10天

17351.4

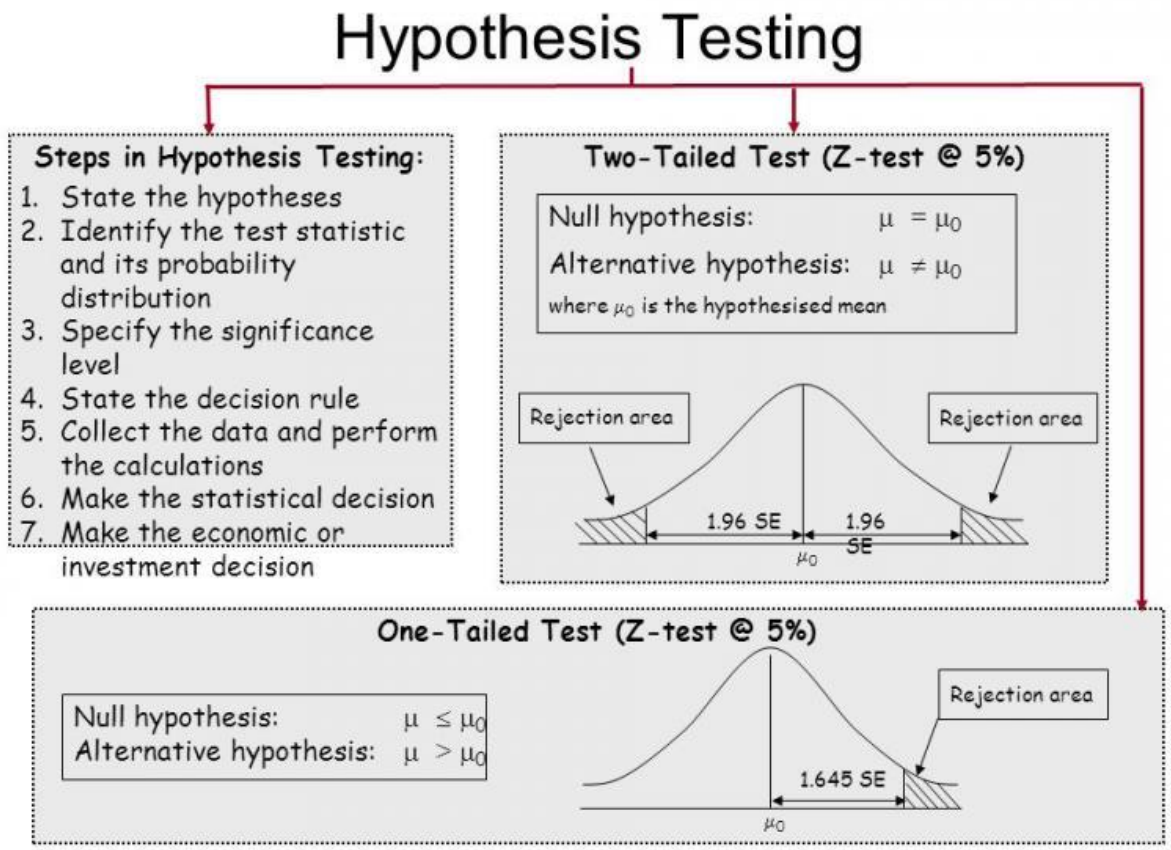
APP的UV活动中10天

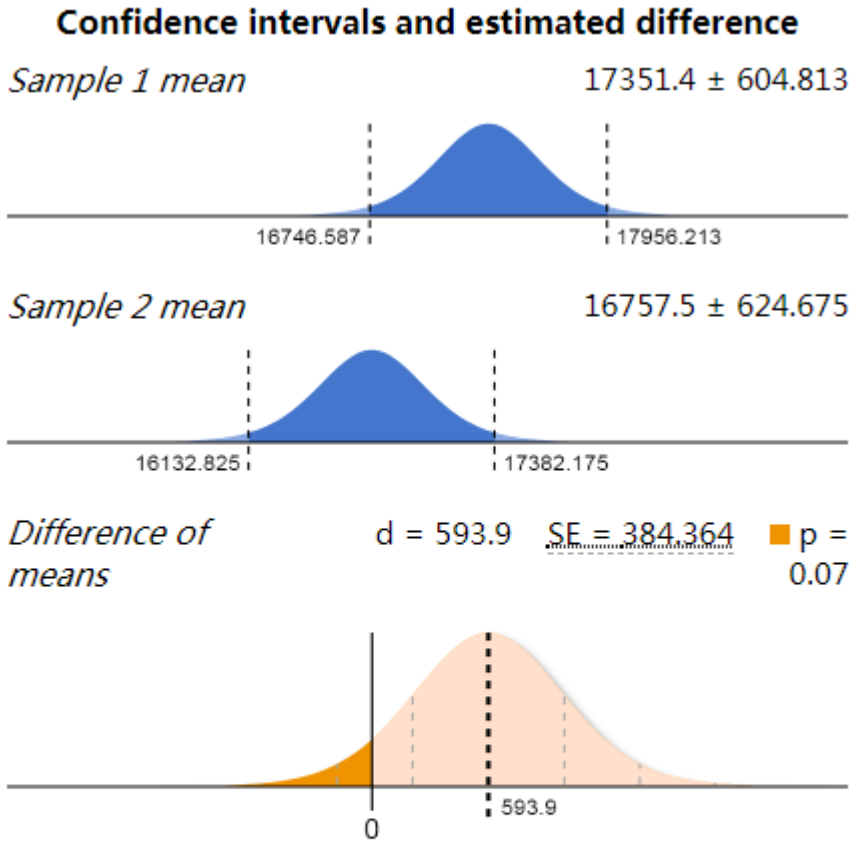
3.5%

活动办得很好

数据的分析
The analysis of data

假设检验



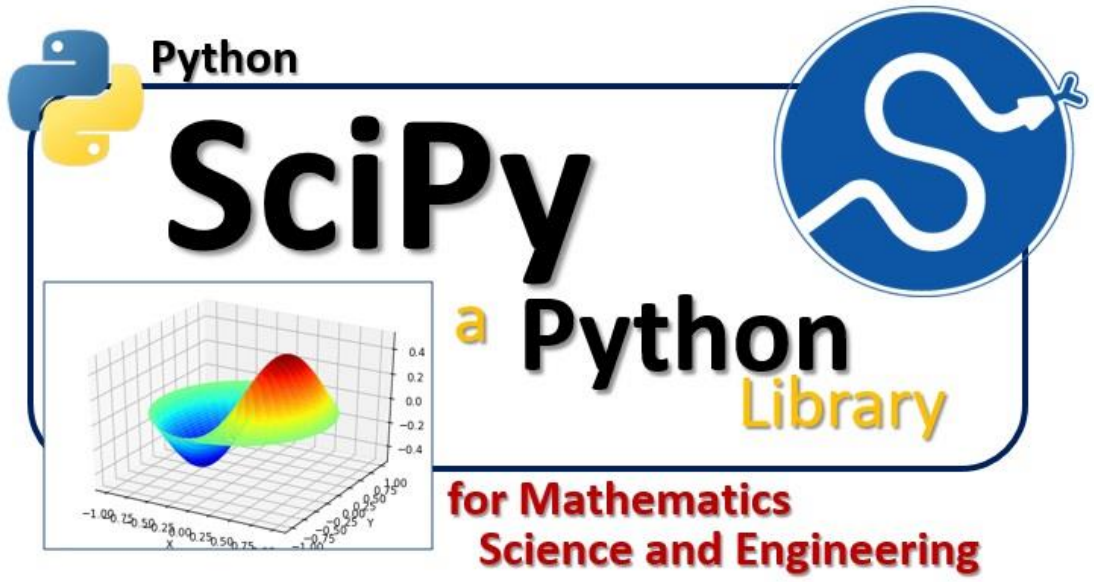


任何由样本估计总体的数据，都应该进行假设检验,计算P value

https://www.evanmiller.org/

统计分析计算工具


IBM SPSS
Statistics


Python
SciPy
a Python
Library
for Mathematics
Science and Engineering

Data CSV Analysis

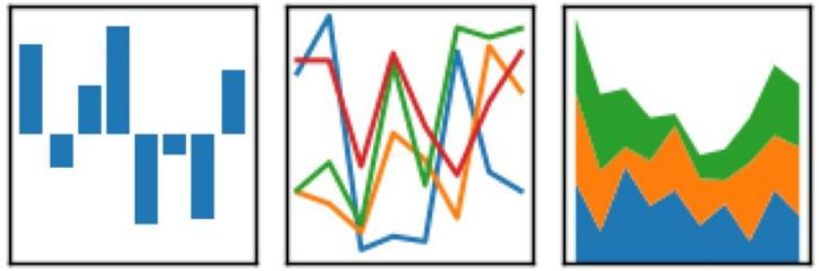
利用Pandas进行青鹿资源题库数据分析

Using Python library Pandas

QINGLU 青鹿

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



题目难度系数分布

题目数量分布

数据的预测

The prediction of data

Experience

以往经验

Machine Learning

机器学习

AlphaGo

Alpha Star

参数1750亿，并且使用45TB数据进行训

Machine Learning

机器学习

什么是机器学习

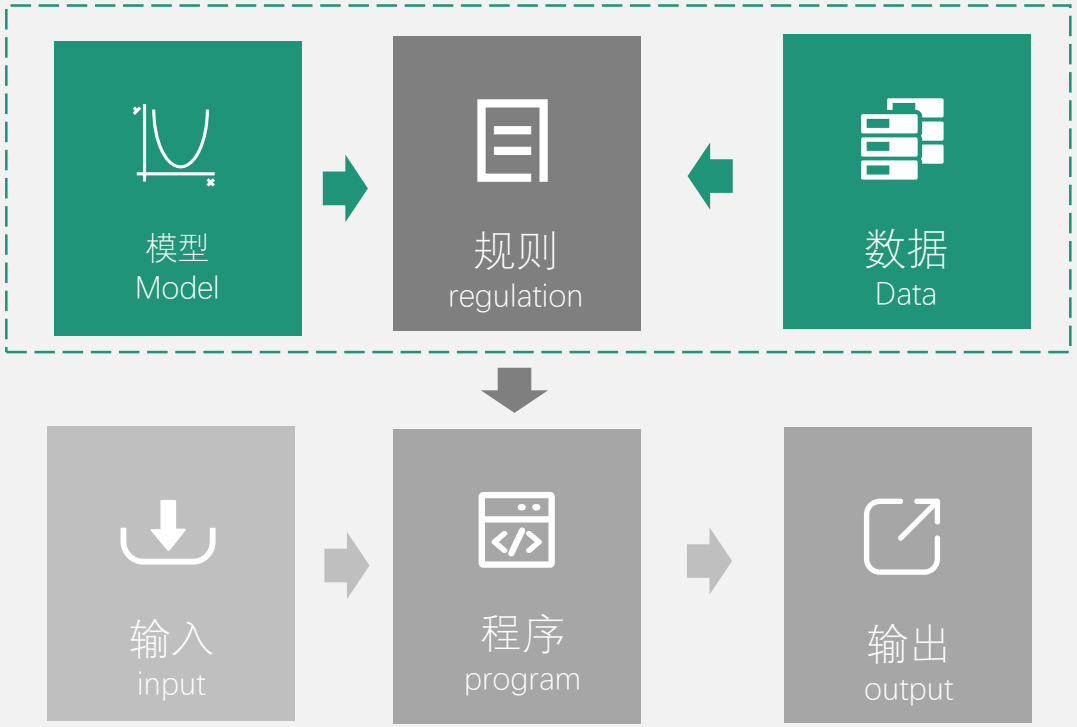
What is machine learning?

"Machine Learning is the study of computer algorithms that improve automatically through experience.

——Wikipedia

机器学习方法与传统程序

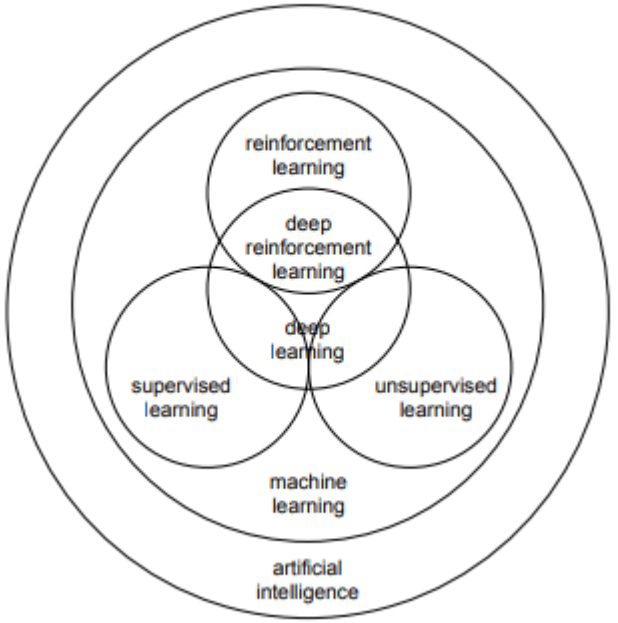
Machine learning and the normal programming



Nils J. Nilsson, Introduction to Machine Learning

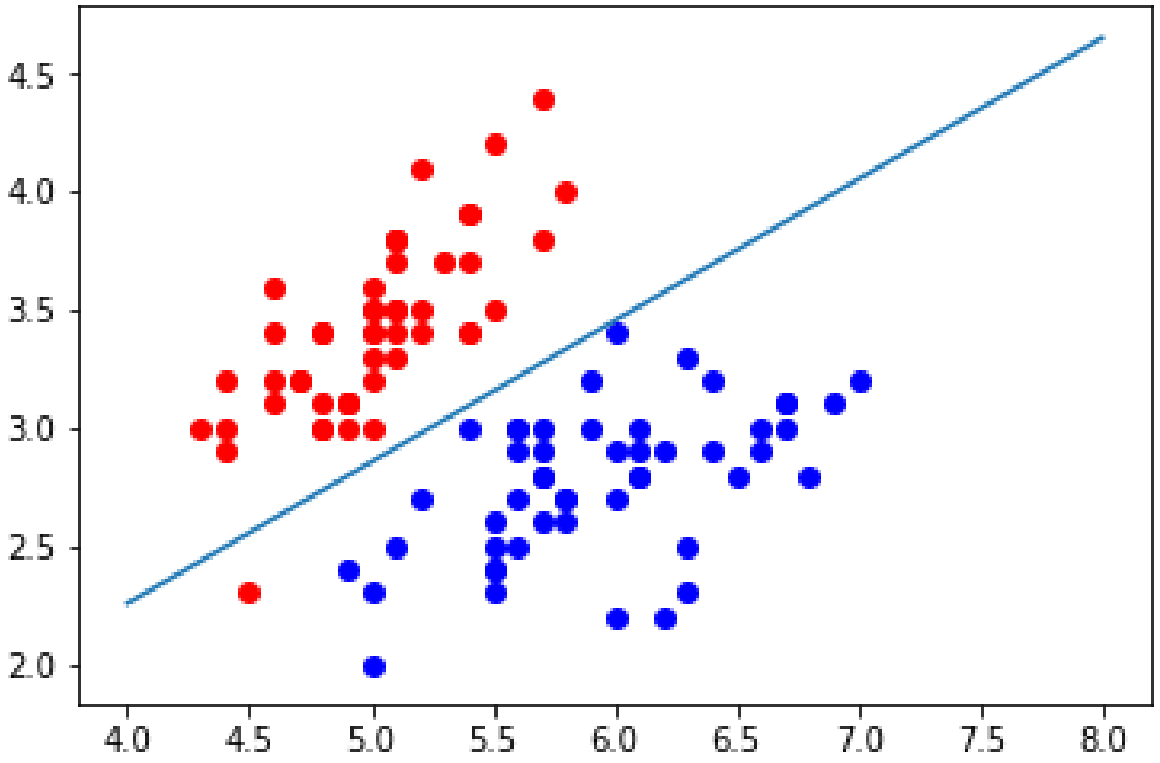
人工智能、机器学习、深度学习、大数据

Artificial Intelligence, Machine learning, Deep Learning, Bigdata

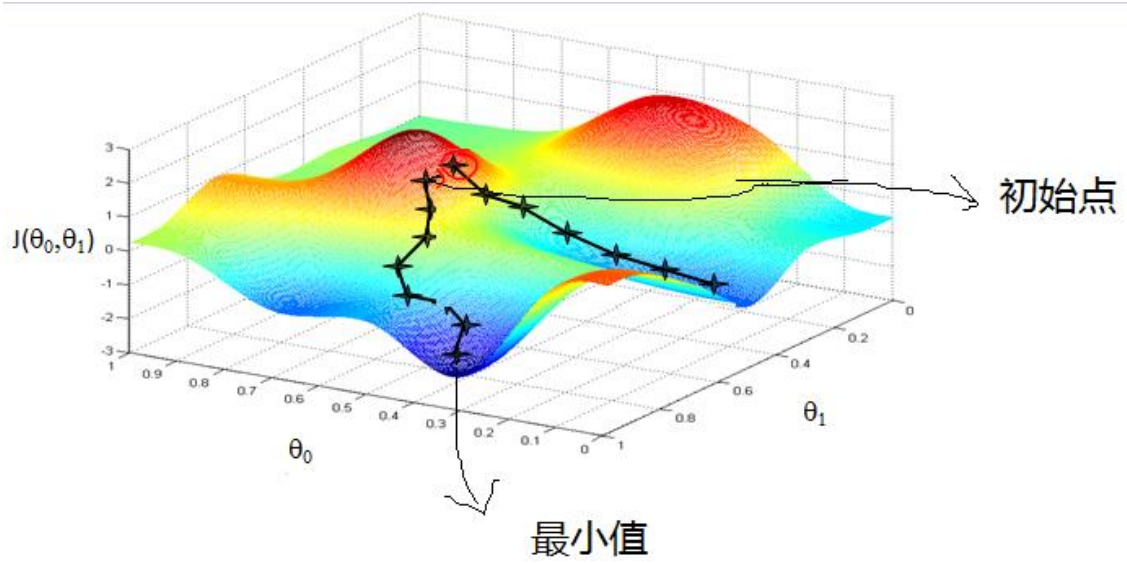


数据的预测

The prediction of data



Logistic Regression

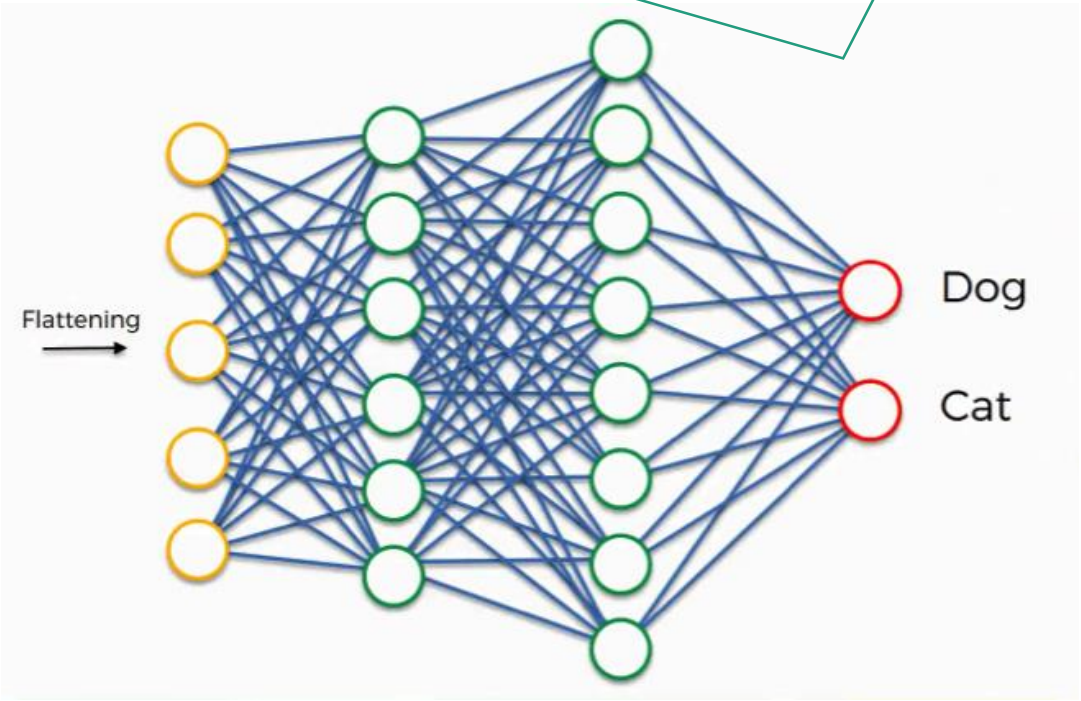


Gradient Descent

机器学习中的深度学习

Deep Learning

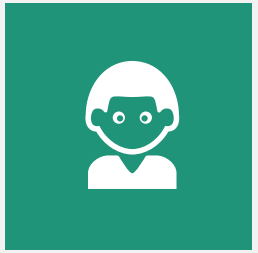
“利用神经网络模型进行学习过程的机器学习为深度学习“



Full connected neural network

全连接神经网络

QINGLU青鹿



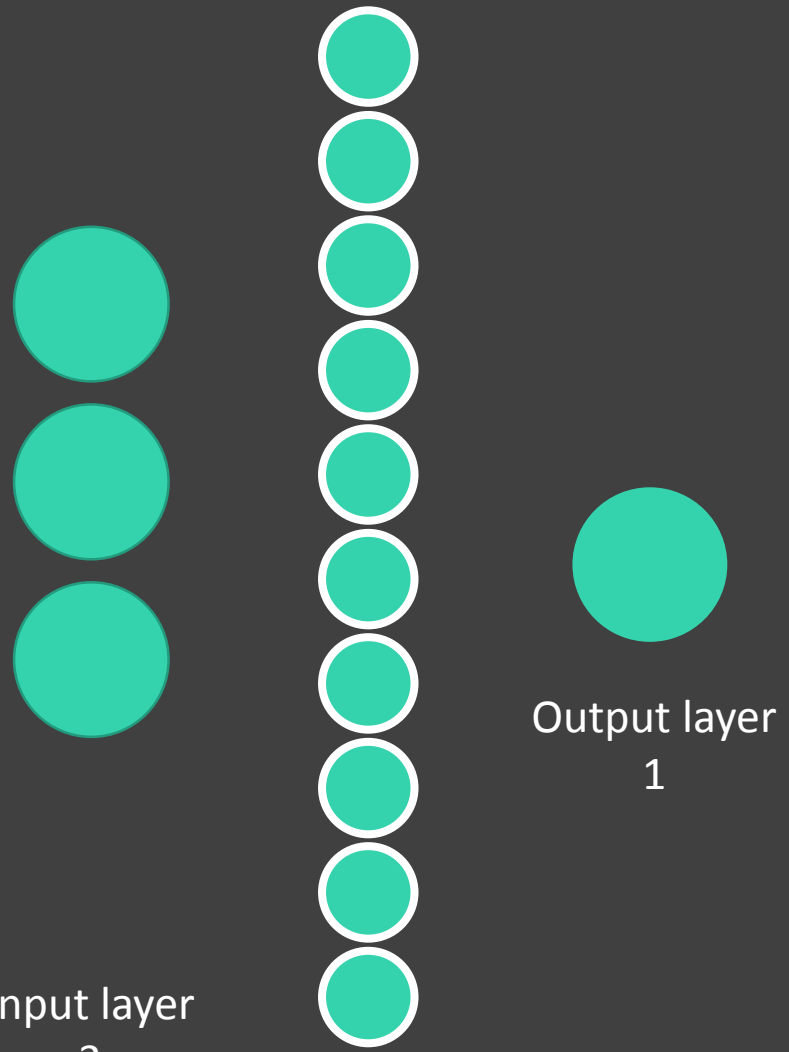
预测小强去不去看电影

Predict the results

如花	小倩	小明	小强
0	0	1	0
1	1	1	1
1	0	1	1
0	1	1	0
1	1	0	?

利用Deep Learning预测小强是否去看电影

Using Deep Learning to predict the results



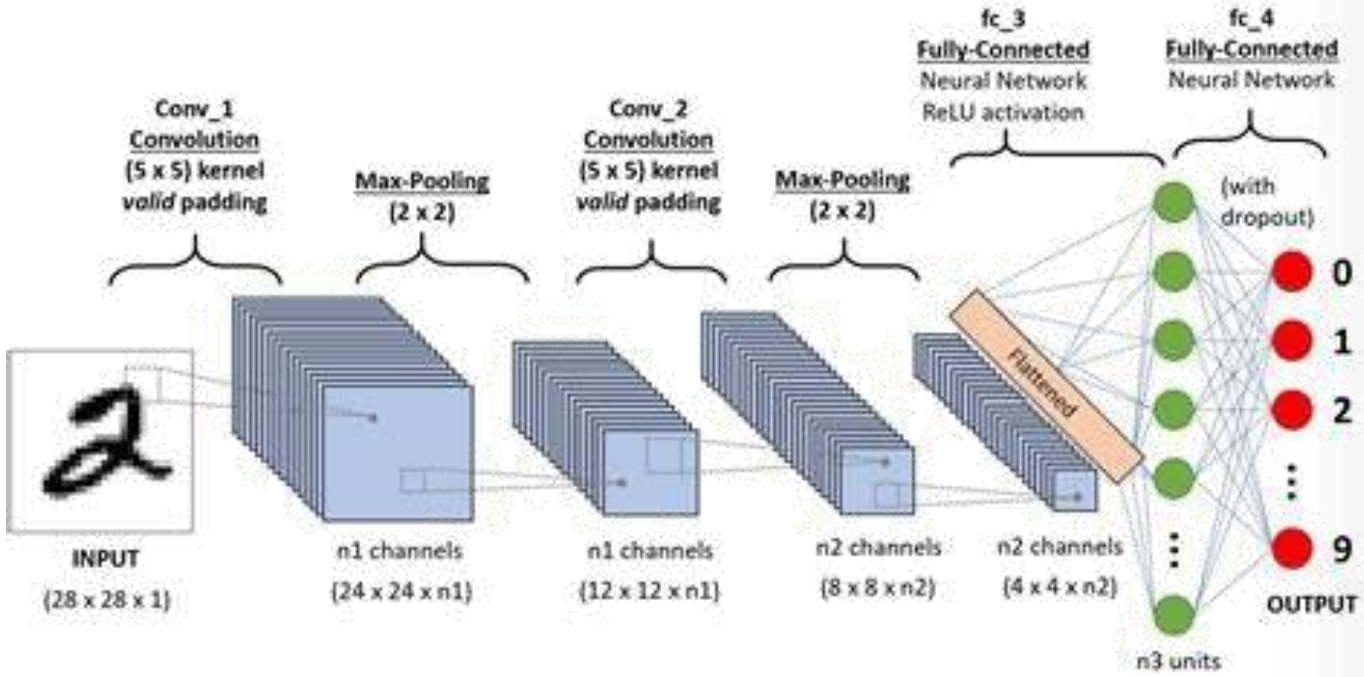

```
from numpy import random,
dot, exp, array

#正向推导：根据输入和权重，
算出结果
def fp(input):
    l1 = 1/(1+exp(-dot(input,
w0)))
    l2 = 1/(1+exp(-dot(l1, w1)))
    return l1, l2

#反向转播：用计算结果和实
际结果的误差，反向推算权
重的调整量
def bp(l1, l2, y):
    error = y - l2
    slope = l2 * (1-l2)
    l1_delta = error * slope

    l0_slope = l1 * (1-l1)
    l0_error = l1_delta.dot(w1.T)
    l0_delta = l0_slope * l0_error

    return l0_delta, l1_delta
```

```
#准备数据: X是输入参数, y是正确结果
X = array([[0,0,1],[0,1,1],[1,0,1],[1,1,1]])
y = array([[0,1,1,0]]).T

#设置随机的权重
random.seed(1)
w0 = random.random((3,10)) * 2 - 1
w1 = random.random((10,1)) * 2 - 1

for it in range(10000):   #迭代循环

    l0 = X
    l1, l2 = fp(l0)    #正向传播计算


  l0_delta, l1_delta = bp(l1, l2, y)     #反向传播计算


    w1 = w1 + dot(l1.T, l1_delta)     #更新权重
    w0 = w0 + dot(l0.T, l0_delta)

    Print(fp([[1,0,1]])[1])   #输出结果
```

结果：0.9914

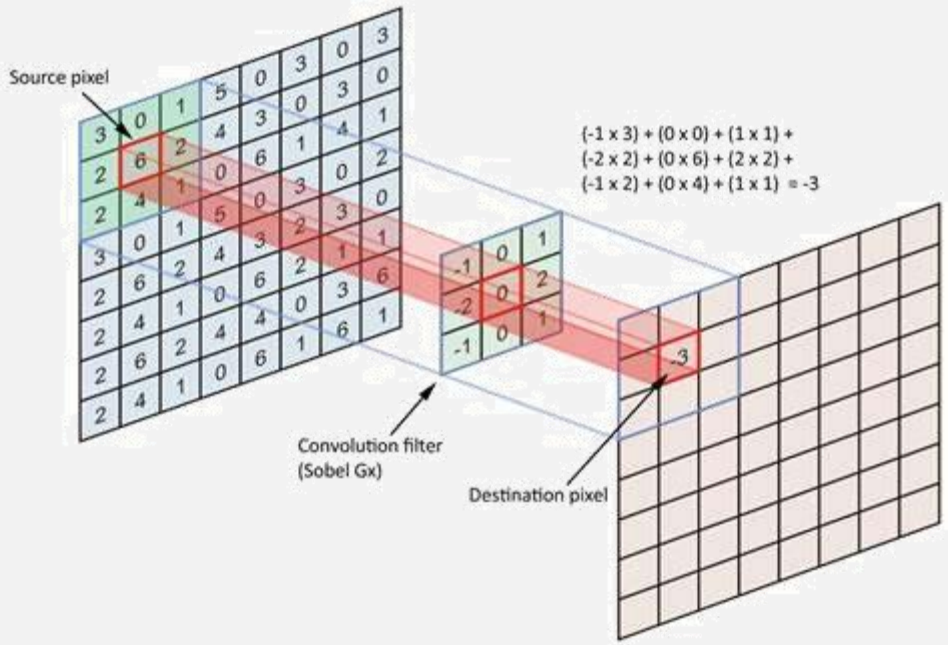
卷积神经网络

Convolution Neural Network



卷积核

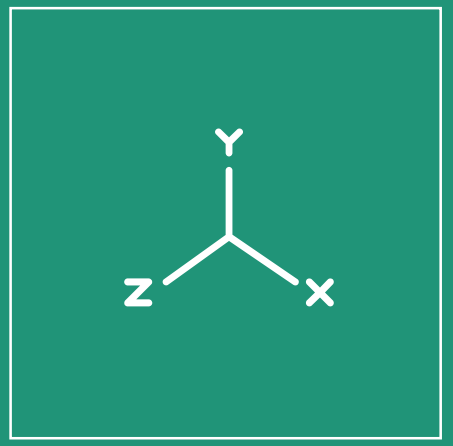
Convolution Kernel



Simon Haykin, Bart Kosko. GradientBased Learning Applied to Document Recognition[M]. Wiley-IEEE Press, 2009.

深度学习
资源类产品的应用

Deeping Learning in K12education



利用各种卷积神经网络,如TextCNN等

Using various CNN

智能标注

OCR识别

智能推送

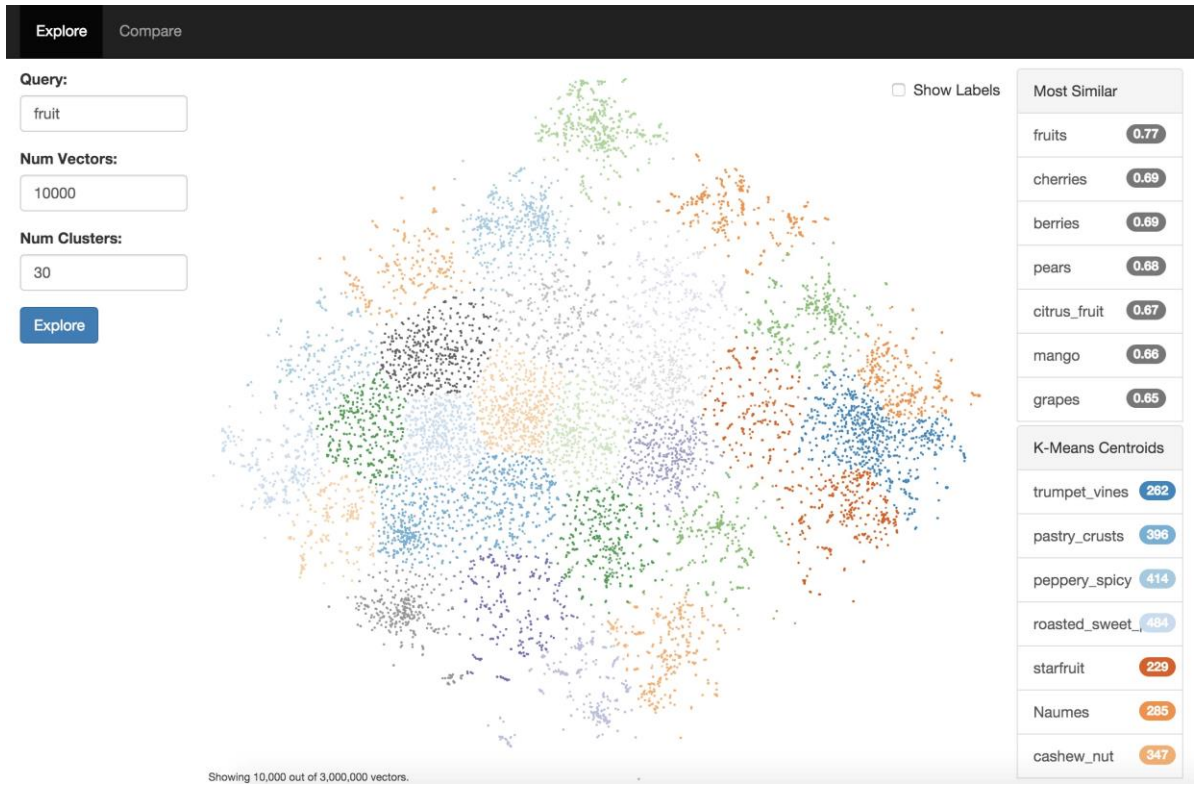
自动批改

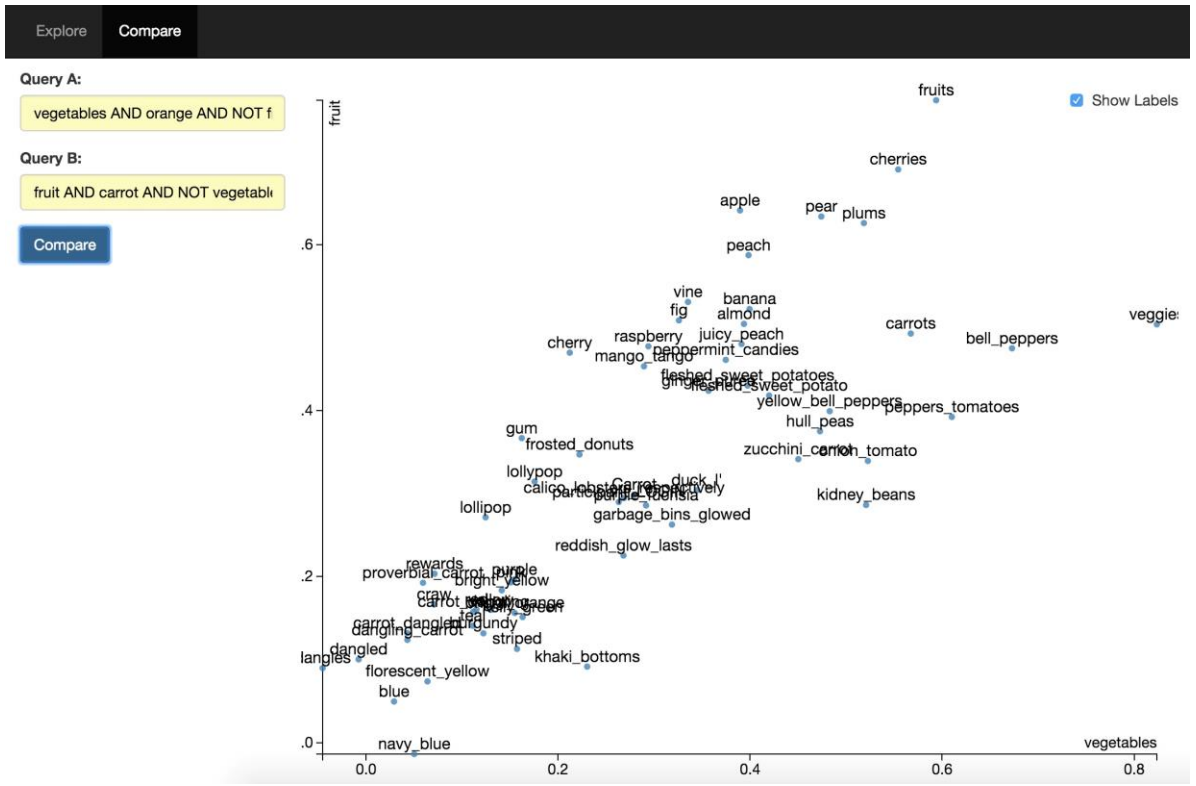
知识点空间

词向量空间
Word2Vec

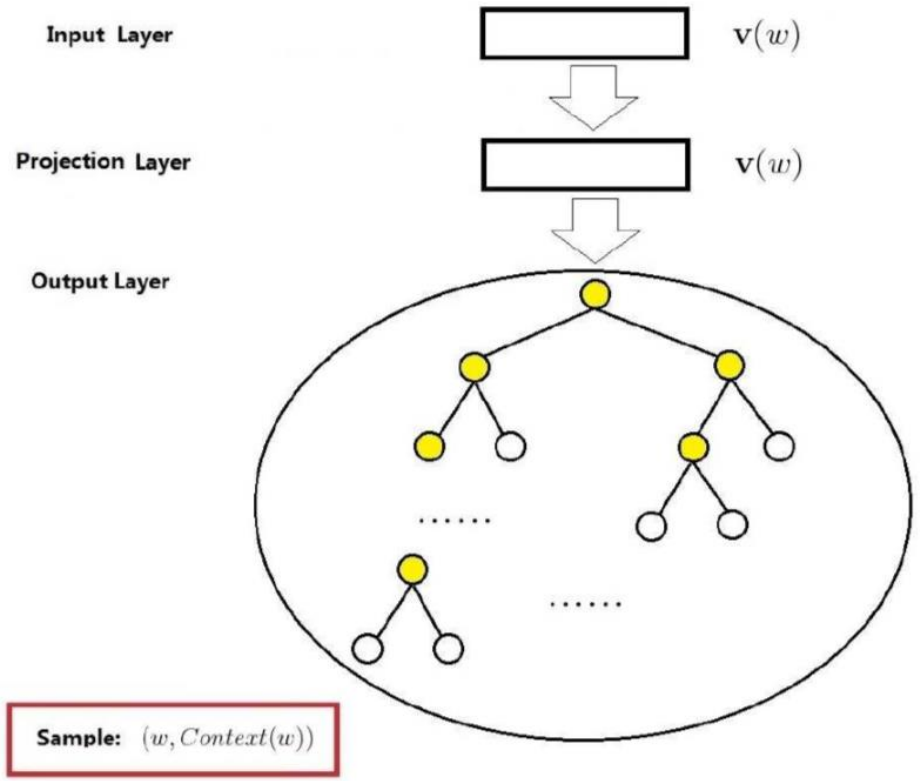
QINGLU青鹿

词（words） → 向量（Vector）





词向量空间
Word2Vec



$$\mathcal{L} = \sum_{w \in \mathcal{C}} \log \prod_{u \in Context(w)} \prod_{j=2}^{l^u} \{[\sigma(\mathbf{v}(w)^\top \theta_{j-1}^u)]^{1-d_j^u} \cdot [1 - \sigma(\mathbf{v}(w)^\top \theta_{j-1}^u)]^{d_j^u}\}$$

$$= \sum_{w \in \mathcal{C}} \sum_{u \in Context(w)} \sum_{j=2}^{l^u} \{(1 - d_j^u) \cdot \log[\sigma(\mathbf{v}(w)^\top \theta_{j-1}^u)] + d_j^u \cdot \log[1 - \sigma(\mathbf{v}(w)^\top \theta_{j-1}^u)]\}.$$

$$\mathbf{v}(w) := \mathbf{v}(w) + \eta \sum_{u \in Context(w)} \sum_{j=2}^{l^u} \frac{\partial \mathcal{L}(w, u, j)}{\partial \mathbf{v}(w)}.$$

基于自主构建的学科词向量库，可以实现智能推送、自主批改、知识点章节分析等功能

QINGLU青鹿

“数据分析与预测的未来大概率的是基于机器学习（深度学习）”

Machine Learning

机器学习

PHYSICAL REVIEW LETTERS

Highlights Recent Accepted Collections Authors Referees Search Press About

Featured in Physics Editors' Suggestion

Discovering Physical Concepts with Neural Networks

Raban Iten, Tony Metger, Henrik Wilming, Lídia del Rio, and Renato Renner

Phys. Rev. Lett. **124**, 010508 – Published 8 January 2020

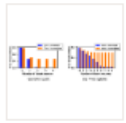
Physics See Viewpoint: Physics Insights from Neural Networks

More

Article References Citing Articles (12) Supplemental Material PDF HTML Export Citation

ABSTRACT

Despite the success of neural networks at solving concrete physics problems, their use as a general-purpose tool for scientific discovery is still in its infancy. Here, we approach this problem by modeling a neural network architecture after the human physical reasoning process, which has similarities to representation learning. This allows us to make progress towards the long-term goal of machine-assisted scientific discovery from experimental data without making prior assumptions about the system. We apply this method to toy examples and show that the network finds the physically relevant parameters, exploits conservation laws to make predictions, and can help to gain conceptual insights, e.g., Copernicus' conclusion that the solar system is heliocentric.


Received 17 July 2019

DOI: https://doi.org/10.1103/PhysRevLett.124.010508

Physics Subject Headings (PhySH)

Issue

Vol. 124, Iss. 1 — 10 January 2020