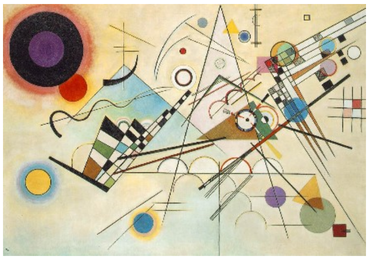


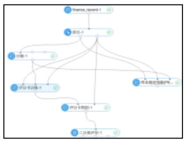
机器学习基础

主讲：王星
单位：中国人民大学统计学院
助教：杜露露
电话：**86-10-82500167**
上课时间：周一上午
上课地点：明法**0102**
Email: wangxingwisdom@126.com
办公地点: 明德主楼 **1019**

机器学习

- Image Classification
- Document Categorization
- Speech Recognition
- Protein Classification
- Spam Detection
- Branch Prediction
- Protein Classification
- Natural Language Processing—Knowledge Map
- Playing Games
- Computational Advertising

Images & Video
flickr
Google
YouTube
Text & Language
REUTERS
AP Associated Press
Wikipedia
Speech & Audio
Gene Expression
Product Recommendation
amazon
NETFLIX
ebY
Relational Data/ Social Network
facebook
twitter
Climate Change
Geological Data



Machine Learning is Changing the World

"Machine learning is the hot new thing"
(John Hennessy, President, Stanford)

"A breakthrough in machine learning would be worth ten Microsofts" (Bill Gates, Microsoft)

"Web rankings today are mostly a matter of machine learning"
(Prabhakar Raghavan, VP Engineering at Google)

SMARTER THAN YOU THINK

Aiming to Learn as We Do, a Machine Teaches Itself

Jeff Swensen for The New York Times

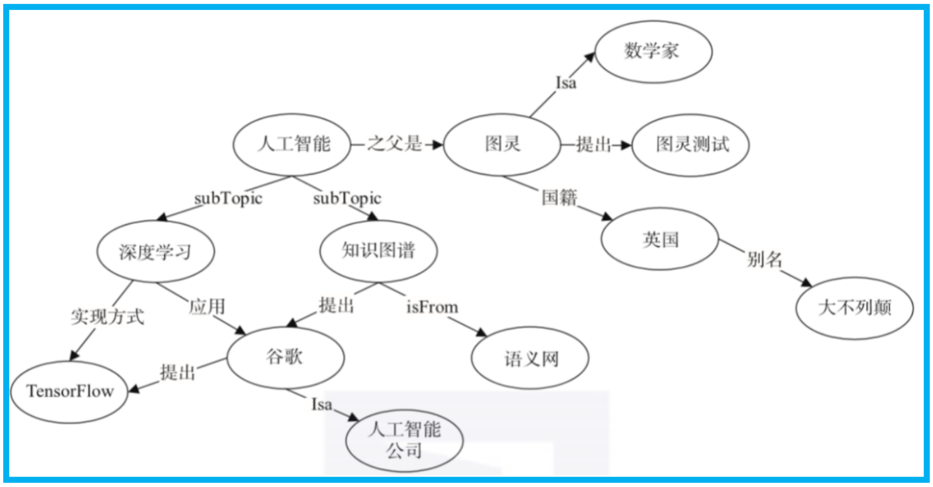
would sound like me.
So, what you see now is the result of that change.
I'm speaking in english and hopefully you'll hear me speaking chinese in my own voice.

The COOLEST TOPIC IN SCIENCE

- "A breakthrough in machine learning would be worthten Microsofts" (Bill Gates, Chairman, Microsoft)
- "Machine learning is the next Internet" (Tony Tether, Director, DARPA)
- Machine learning is the hot new thing" (John Hennessy, President, Stanford)
- "Web rankings today are mostly a matter of machine learning" (PrabhakarRaghavan, Dir. Research, Yahoo)
- "Machine learning is going to result in a real revolution" (Greg Papadopoulos, CTO, Sun)
- "Machine learning is today's discontinuity" (Jerry Yang, CEO, Yahoo)

机器学习的历史（外接视频）

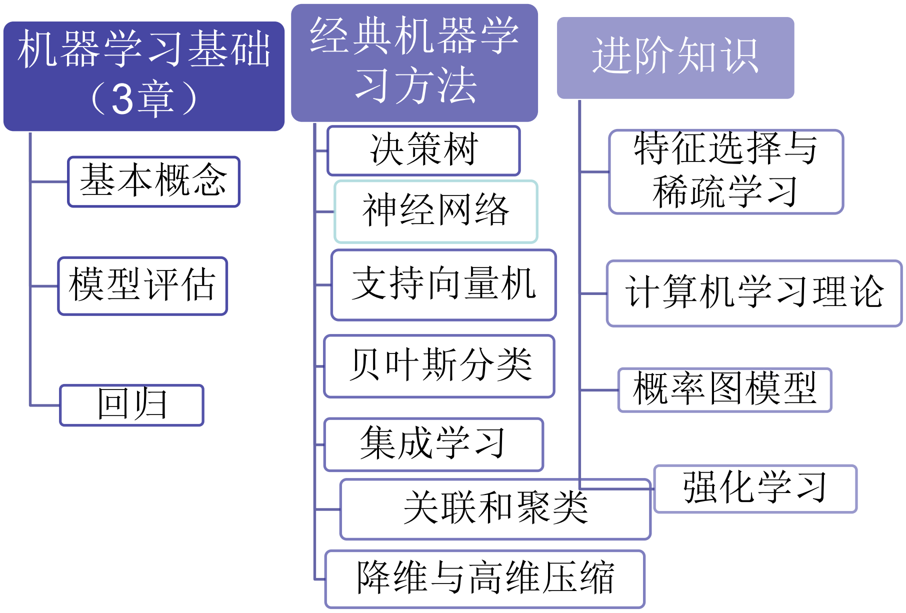
- https://www.iqiyi.com/w_19s5u1bb9p.html



学习目标

- 掌握机器学习算法基本概念和基本原理
- 针对任务选择合适算法；
- 对不同目的应用训练好的模型；
- 学习数据处理机制，准备数据；
- 评估模型性能以保证应用效果；
- 掌握Python 机器学习核心算法包；
- 使用示例代码设计和构建个性化模型；
- 构建实用的多功能预测模型。

该课程的定位：目录大纲 16章



机器学习基础（3章）

- 基本概念
- 模型评估
- 回归

经典机器学习方法

- 决策树
- 神经网络
- 支持向量机
- 贝叶斯分类
- 集成学习
- 关联和聚类
- 降维与高维压缩

进阶知识

- 特征选择与稀疏学习
- 计算机学习理论
- 概率图模型
- 强化学习

课堂里的机器学习案例

- 探索分类分析算法并将其应用于收入等级评估问题
- 使用预测建模并将其应用到实际问题中
- 了解如何使用无监督学习来实施市场细分
- 探索数据可视化技术以多种方式与数据进行交互
- 了解如何构建推荐引擎
- 理解如何与文本数据交互并构建模型来分析它
- 使用隐马尔科夫模型来研究语音数据并语音识别

基本要求

- 每周一次作业:There will be weekly homeworks；
- 1次习题课，1个 team project展示个 1次期末考试；
- 个人作业The weights are 平时20% for the personal homework, 期中20% for the first exam, 期末 60% for final exam.
- 作业不能拖延No late homework will be accepted: A,B,C
- Agree or other suggestions?
- 每周作业：编程、原理理解、数据分析，Jupyter Notebook

A类作业模板

[illegible]

3/1/2021

课本和参考书

教材：

1.周志华，机器学习，清华大学出版社，2016,01

2.Sebastian Raschka，Python机器学习，机械工业出版社，2017,03

3.李航，统计学习方法，清华大学出版社，2012,03

4. Christopher bishop, PRML,2007

5.T. M. Mitchell, Machine Learning, McGraw Hill, 1997

6.R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classificati
Wiley-Interscience, 2000

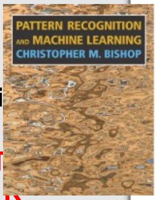
7.Gareth James, Daniela Witten, Trevor Hastie and Robert T
An Introduction to Statistical Learning with Applications in R

8. Hastie, Tibshirani and Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. The second Edition.

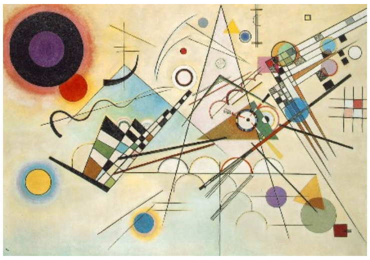
9.P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Dat
Wiley, 2005

10.Han Jiawei, Data mining Concepts and techniques, 机械工业

11.王星，大数据分析：方法与应用，清华大学出版社，201309



机器学习研究的对象？



机器学习里的经验

- 计算机中，“经验”通常是以“数据”形式储存下来，因此机器学习所研究的主要内容是关于如何从观测数据中不断学习和总结经验，从数据中产生“模型”（model）帮助计算机做出准确判断的自动化技术，这种技术称为算法，也称为“learning algorithm”.
- 两个注释：
  - 模型是算法学习的结果；
  - 算法将数据和经验打通。

问题与思考：经验预判

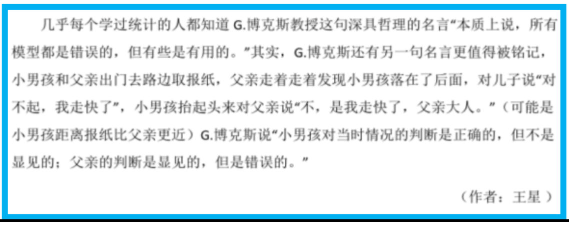
1. 天热去挑瓜，极大可能地挑出一个又甜又脆的好瓜?
2. 天上钩钩云,地上雨淋淋.天有城堡云,地上雷雨临；早晨浮云走,午后晒死狗.早雨一日晴,晚雨到天明；今晚花花云,明天晒死人.空中鱼鳞天,不雨也风颠。
3. 增加词汇量是增强英语理解能力的基本备考经验?
4. 便血可能是肠癌的前兆?
5. 新能源电动车电池寿命与经常使用快充还是慢充以及驾车员的驾车习惯很有关系；
6. 一位游泳10年的泳池达人可以告诉你抽筋溺水怎么办？一位瑜伽教练可以告诉你如何保护腰背不受伤?
7. 法院调解人对于分析案情找到当事人进行纠纷调解很有办法?

- 概念学习
- 经验分类
- 经验评估
- 新特征发现
- 风险因素
- 知识问答
- 信息抽取
- 经验排序

经验与数据

- 不确定性普遍存在，可以确定的是：人类可以依靠经验降低不确定问题中不确定的可能性；
- 经验：多次实践中得到的知识或技能（辞典）
- 经验经常以数据的形态存在；
- 有两种经验：显见的经验和不显见的经验

	显见	不显见
得当的经验		✓
不当的经验	✓	



几乎每个学过统计的人都知道G.博克斯教授这句深具哲理的名言“本质上说，所有模型都是错误的，但有些是有用的。”其实，G.博克斯还有另一句名言更值得被铭记，小男孩和父亲出门去路边取报纸，父亲走着走着发现小男孩落在了后面，对儿子说“对不起，我走快了”，小男孩抬起头来对父亲说“不，是我走快了，父亲大人。”（可能是小男孩距离报纸比父亲更近）G.博克斯说“小男孩对当时情况的判断是正确的，但不是显见的；父亲的判断是显见的，但是错误的。”

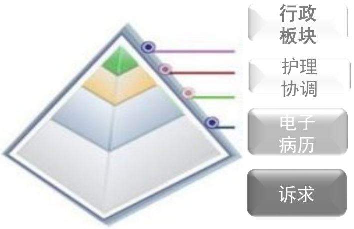
（作者：王星）

大数据带来的第一个好消息：不可能→可能→复现

提供了观察人类问题的新**视角：健康、教育**

的Holo Lens眼镜

HoloLens是一款现实增强耳机，它可以将虚拟世界与现实世界融合为一体。戴上耳机，镜片上的显示屏就会在物理世界上投射出一个虚拟世界层。



mHealth

MD2K Smoking Cessation Coach

- Wearable chest and wrist bands measure activity, stress, cigarette smoking…….
- Supportive stress-regulation interventions available on smartphone 24/7
- In which contexts should the wrist band provide supportive “cue” and smartphone activate to highlight associated support?

Today’s Challenge	New Data	What’s Possible
Healthcare Expensive office visits Hospital Dynamics	Remote patient monitoring, Hospital Sensors	Preventive care, reduced hospitalization, reduced human mistakes
Manufacturing In-person support	Product sensors	Automated diagnosis, customized support
Location-Based Services Based on home zip code	Real time location data	Geo-advertising, urban computing, mobile recommendation
Finance Fast-paced, Variety	Social Media, High-frequency Trading Data	Sentiment analysis Finance engineering
Retail One size fits all marketing	Market basket data, user behavior logs	Personalized Recommendation, Segmentation

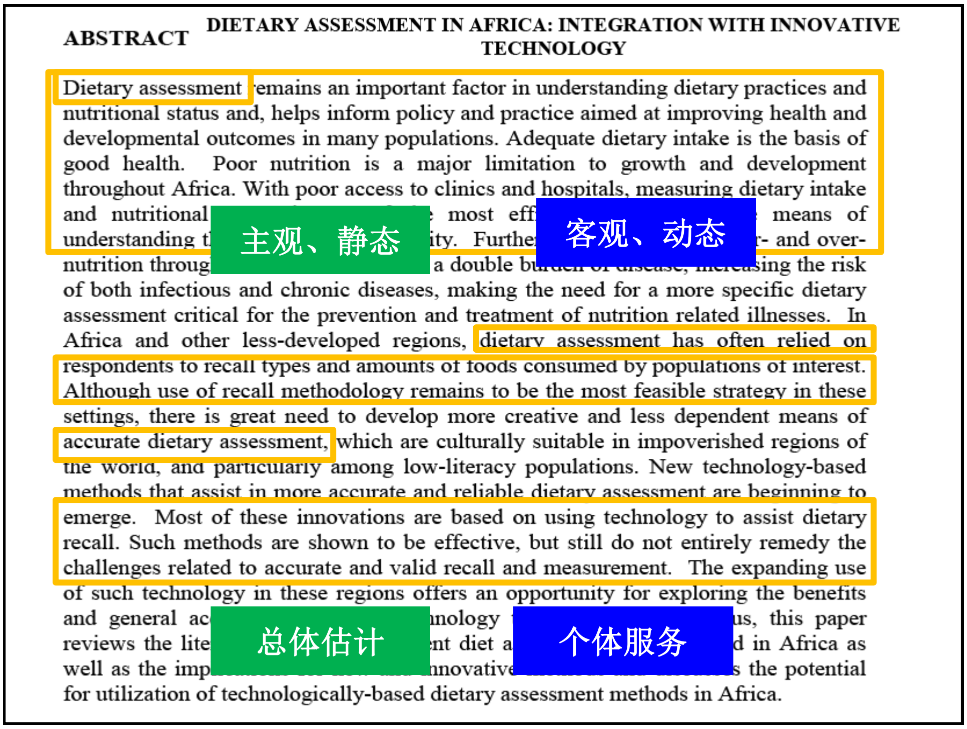
ABSTRACT **DIETARY ASSESSMENT IN AFRICA: INTEGRATION WITH INNOVATIVE TECHNOLOGY**

Dietary assessment remains an important factor in understanding dietary practices and [illegible]

南非人均国民收入 2005年超过 5 000美元，但是贫富悬殊，基尼系数高达 57~ 59。1994年非洲人国民大会 (下称 “非国大”) 执政以来，一个日渐增长的黑人中产阶层正在出现。根据南非贸工部的数字，黑人在企业主中已占 10%，在技术人员中已占 15%。黑人中最富有者的收入增长了 30%。[1] 南非百万美元富翁 2004年达 3. 7万人。根据 2005年《世界财富报告》南非占非洲 7. 5万个百万美元富翁群体的一半以上。一年内，新增百万美元富翁人数可与南非相比的只有新加坡、中国香港、澳大利亚。南非原来的矿业大亨（比如奥本海默家族和鲁珀尔斯家族），现与一批新生巨富共享超级富豪的地位。[2]

根据南非统计局的数据，1995年，南非 28%的家庭和 48%的人口生活在贫困线以下。1999年，南非全国 1 140万家庭中，有 370万处于贫困线以下，约占 33%。根据南非官方的解释，与 1995年相比，贫困家庭比例有明显增加的部分原因是贫困人口的传统大家庭分解为小家庭。贫苦家庭的收入平均低于贫困线 12%，与大多数发展中国家类似。” 由于长期种族隔离制度造成的广大黑人就业受限，失业率高，新政府在发展经济的同时，必须对缺乏收入来源的群体给与救助。同时大批流亡在国外的解放组织战士在 20世纪 90年代初期陆续回国，等待安置。

Africa and other less-developed regions, dietary assessment has often relied on respondents to recall types and amounts of foods consumed by populations of interest. Although use of recall methodology remains to be the most feasible strategy in these settings, there is great need to develop more creative and less dependent means of accurate dietary assessment, which are culturally suitable in impoverished regions of the world, and particularly among low-literacy populations. New technology-based methods that assist in more accurate and reliable dietary assessment are beginning to emerge. Most of these innovations are based on using technology to assist dietary recall. Such methods are shown to be effective, but still do not entirely remedy the challenges related to accurate and valid recall and measurement. The expanding use of such technology in these regions offers an opportunity for exploring the benefits and general acceptance of using technology to improve health. Thus, this paper reviews the literature concerning current diet assessment methods used in Africa as well as the implications for new and innovative methods and discusses the potential for utilization of technologically-based dietary assessment methods in Africa.


主观、静态
客观、动态
总体估计
个体服务

ABSTRACT

DIETARY ASSESSMENT IN AFRICA: INTEGRATION WITH INNOVATIVE TECHNOLOGY

Dietary assessment remains an important factor in understanding dietary practices and nutritional status and, helps inform policy and practice aimed at improving health and developmental outcomes in many populations. Adequate dietary intake is the basis of good health. Poor nutrition is a major limitation to growth and development throughout Africa. With poor access to clinics and hospitals, measuring dietary intake and nutritional most eff means of understanding t ity. Furthe r- and over-nutrition throug a double burden of disease, increasing the risk of both infectious and chronic diseases, making the need for a more specific dietary assessment critical for the prevention and treatment of nutrition related illnesses. In Africa and other less-developed regions, dietary assessment has often relied on respondents to recall types and amounts of foods consumed by populations of interest. Although use of recall methodology remains to be the most feasible strategy in these settings, there is great need to develop more creative and less dependent means of accurate dietary assessment, which are culturally suitable in impoverished regions of the world, and particularly among low-literacy populations. New technology-based methods that assist in more accurate and reliable dietary assessment are beginning to emerge. Most of these innovations are based on using technology to assist dietary recall. Such methods are shown to be effective, but still do not entirely remedy the challenges related to accurate and valid recall and measurement. The expanding use of such technology in these regions offers an opportunity for exploring the benefits and general ac nology us, this paper reviews the lite nt diet a d in Africa as well as the imp nnovative s the potential for utilization of technologically-based dietary assessment methods in Africa.

统计学成为数据科学的新宠

Tukey Leo Breiman、William Cleveland

A Data Scientist Is…

"A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician."

- Josh Blumenstock

"Data Scientist = statistician + programmer + coach + storyteller + artist"

- Shlomo Aragmon

大数据带来的第2个好消息：新模型和新算法

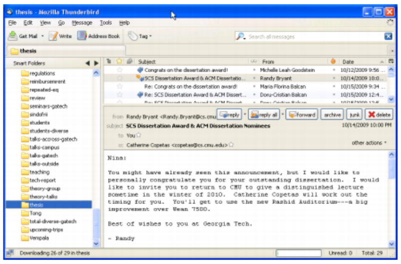
Supervised Classification. Example: Spam Detection

Decide which emails are spam and which are important.

Not spam

Supervised classification

spam



Subject	From	Date
Congrats on the dissertation award!	Michelle Leah Goodstein	10/12/2009 9:56 ...
SCS Dissertation Award & ACM Dissertatio...	Randy Bryant	10/14/2009 10:0...
Re: Congrats on the dissertation award!	Maria Florina Balcan	10/15/2009 9:34 ...
Re: SCS Dissertation Award & ACM Dissert...	Doru-Cristian Balcan	10/15/2009 12:4...

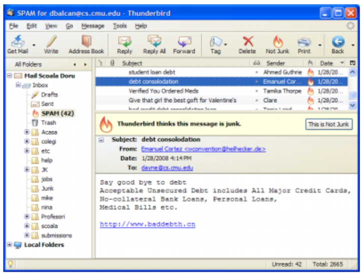
from Randy Bryant <Randy.Bryant@cs.cmu
subject **SCS Dissertation Award & ACM Dissertation Nominees** 10/14/2009 10:00 PM
to You
cc Catherine Copetas <copetas@cs.cmu.edu>

```
Nina:

You might have already seen this announcement, but I would like to
personally congratulate you for your outstanding dissertation.  I would
like to invite you to return to CMU to give a distinguished lecture
sometime in the winter of 2010.  Catherine Copetas will work out the
timing for you.  You'll get to use the new Rashid Auditorium---a big
improvement over Wean 7500.

Best of wishes to you at Georgia Tech.

- Randy
```



Subject	Sender	Date
student loan debt	Ahmed Guthrie	1/28/20...
debt consolodation	Emanuel Cor...	1/28/20...
Verified You Ordered Meds	Tamika Thorpe	1/28/20...
Give that girl the best goft for Valentine's	Clare	1/28/20...

Thunderbird thinks this message is junk.

Subject: **debt consolodation**
From: Emanuel Cortez <wconvention@hellhecker.de>
Date: 1/28/2008 4:14 PM
To: dayne@cs.cmu.edu

```
Say good bye to debt
Acceptable Unsecured Debt includes All Major Credit Cards,
No-collateral Bank Loans, Personal Loans,
Medical Bills etc.

http://www.baddebth.cn
```

Goal: use emails seen so far to produce good prediction rule for future data.

Supervised Classification. Example: Spam Detection

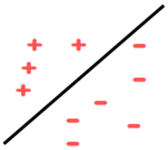
Represent each message by features. (e.g., keywords, spelling, etc.)

	"money"	"pills"	"Mr."	bad spelling	known-sender	spam?	
	Y	N	Y	Y	N	Y	
	N	N	N	Y	Y	N	
	N	Y	N	N	N	Y	
example	Y	N	N	N	Y	N	label
	N	N	Y	N	Y	N	
	Y	N	N	Y	N	Y	
	N	N	Y	N	N	N	

Reasonable RULES:

Predict SPAM if unknown AND (money OR pills)

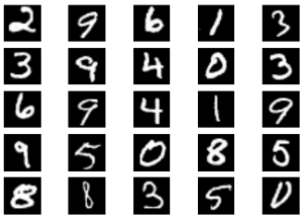
Predict SPAM if 2money + 3pills –5 known > 0



Supervised Classification. Example: Image classification

- Handwritten digit recognition (convert hand-written digits to characters 0..9)

Random Sampling of MNIST



- Face Detection and Recognition

Regression. Predicting a numeric value

Stock market

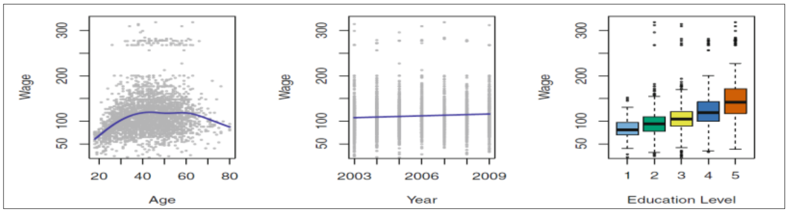
Weather prediction

Temperature
72° F

Predict the temperature at any given location

建模的例子

1. 影响劳动者劳动报酬高低的因素有哪些方面?
2. 在这些影响因素中，你认为什么的影响最大?



1. 受教育程度是影响劳动报酬分配的重要因素在其它条件相同情况下,受教育程度不同会引起劳动者劳动收入的差别是非常显著的。
2. 美国的教育学家布劳格说过：教育与收入之间的正相关的普遍性是现代社会最显著的发现质疑。也是少数几条适用于分析所有国家劳动力市场的准则之一、 在现在不同社会经济制度,不同经济发展水平和不同运行模式的国家都是普遍存在的.

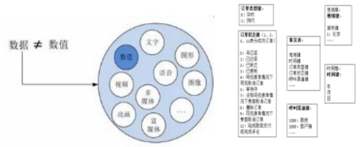
数据挑战：数据是什么(data as entity)

人大信息资源管理学院朝乐门《数据科学》

1.体（size）在数据科学中，具有物理属性byte的一切符号、编码、图像、语音、多媒体和富媒体等都称为数据；

2.历（resume）实体成长的记录，从个体数据来看，每条数据有创建、更新、使用等时间和空间地址等基本的自然属性。

3.通（sociability）从数据的价值来看，一个活跃的数据包含很多属性，数据分析也是从数据的很多属性挖掘出更多的属性之间的关系和模式为出发点开始的。



从分析的视角看，数据是可以形成问题的分析变量和变量的取值，但是这些变量有的是以实体的形式连接到主体，有的是以个体的形式互为连接，如何提取，这么提取之后是不是得到的结论稳定，都需要一个工艺设计和评估的过程。

Leo Brieman说：数据是实体，他说“如果生物学比物理学复杂，行为学比生物学复杂，那么数据分析则比行为学更复杂”。仅就关系理解的视角来看，物理学中物体的复杂性主要体现在其相对作用关系的不确定性；生物的多样性生物多样性是指一定范围内多种多样活的有机体(动物、植物、微生物) 有规律地结合所构成稳定的生态综合体，而行为学的不确定性体现在作为动态不确定性行为的个体和社群为适应内外环境变化 (刺激)所作出的反应。

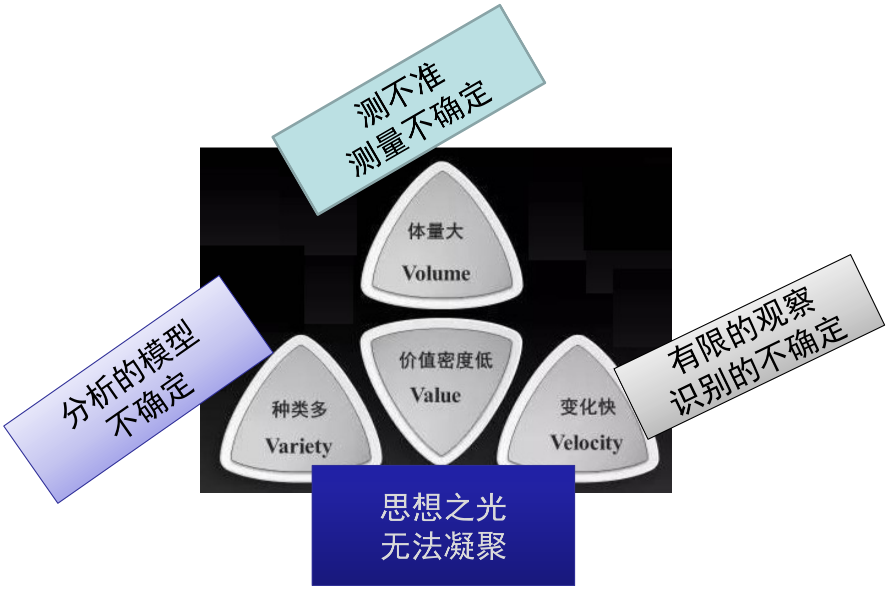
什么是大数据？

（观察、归纳和判断的模式化）

4V特性(Variable,Variety,Variance,Visulization)

体量Volume	*非结构化数据*的超大规模和增长 总数据量的80~90% 比结构化数据增长快10倍到50倍 是传统数据仓库的10倍到50倍
多样性Variety	大数据的异构和多样性 很多不同形式（文本、图像、视频、机器数据） 无模式或者模式不明显 不连贯的语法或句义
价值密度Value	大量的不相关信息 对未来趋势与模式的可预测分析 深度复杂分析（机器学习、人工智能Vs传统商务智能(咨询、报告等）
速度Velocity	*实时分析*而非批量式分析 数据输入、处理与丢弃 立竿见影而非事后见效

大数据及对应的科学问题



测不准
测量不确定

分析的模型
不确定

有限的观察
识别的不确定

思想之光
无法凝聚

不显见实体依赖的数据经验-COVID19-安全距离

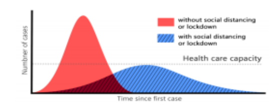
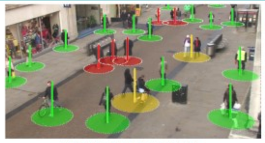


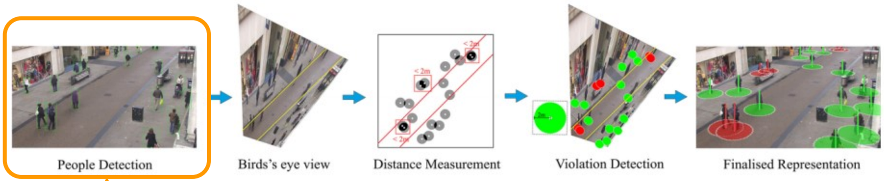
Figure 2. Gaussian distribution of infection transmission rate for a given population, with and without social distancing obligation.

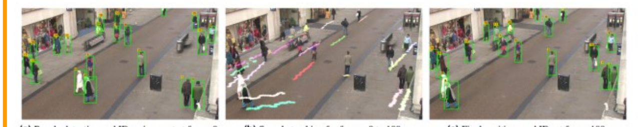


(a) Social distancing monitoring

(b) Accumulated infection risk (red-spots) due to multiple breaches of the social-distancing

- The statistics by WHO on 20 SEPT 2020 confirms 30.8 million infected people in more than 200 countries. The mortality rate of the infectious virus also shows a scary number of 1,005,000 people.
- No effective cure or available treatment for the virus.
- Precautions are taken by the whole world to limit the spread of infection
- Social distancing: refers to precaution actions to prevent the proliferation of the disease, by minimising the proximity of human physical contacts in covered or crowded public places (e.g. schools, workplaces, gyms, lecture theatres, etc.)
- Figure 2 demonstrates the effect of following appropriate social distancing guideline to reduce the rate of infection transmission among individuals. A wider Gaussian curve with a shorter spike within the range of the health system service capacity makes it easier for patients to fight to the virus by receiving continuous and timely support from the health care organizations. Any unexpected sharp spike and rapid infection rate (such as the red curve in Figure 2), will lead to the service failure, and consequently, exponential growth in the number of fatalities





(a) People detection and ID assignment at frame 0 **(b)** Sample tracking for frames 0 to 100 **(c)** Final positions and IDs at frame 100

Figure 7. People Detection, ID assignment, Tracking and moving trajectory representation.

不同类型和数据来源
的综合分析

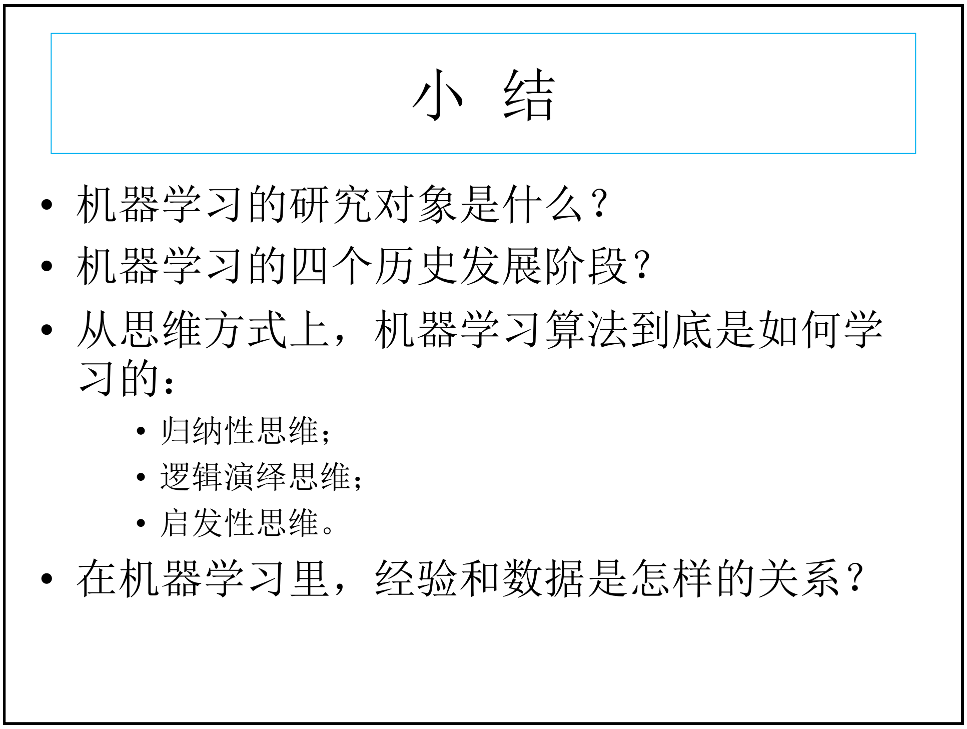
隐私数据的统计分析

重复性和可复制性的设计

新出现的网络、轨迹数据

Questions and considerations about designing a safe workplace should include:

- Can you redesign the workplace to maintain social distancing?你重新设计的工作场所以保持社交距离的效果如何
- Can you repurpose meeting rooms to spread employees out?改变会议室的用途，让员工分散开
- Can you reduce space pressure by reducing the number of employees required to work in an area (e.g. a proportion remains working from home)?减少在一个区域工作的员工数量
- In which places do people find it difficult to avoid one another (e.g. security points, lifts, stairs, lobbies, canteens, toilets, resource rooms, hot desks)?很难避开彼此(例如保安站、升降机、楼梯、大堂、食堂、厕所、资源室、热桌)?
- What can you do to smooth out their use and reduce this pressure (e.g. phased shift and break times, closure)?
- Can you provide more hand washing or sterilisation facilities around the workplace?在工作地点提供更多的洗手或消毒设施?
- Have you noted the places where most people commonly touch (e.g. equipment control panels, handles, handrails, kettles, hot desk surfaces)?大多数人经常接触的地方(例如设备控制面板，把手，扶手，水壶，办公桌热台面)?



小 结

- 机器学习的研究对象是什么？
- 机器学习的四个历史发展阶段？
- 从思维方式上，机器学习算法到底是如何学习的：
  - 归纳性思维；
  - 逻辑演绎思维；
  - 启发性思维。
- 在机器学习里，经验和数据是怎样的关系？