中国科学院
自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

大数据机器学习

中国科学院自动化研究所

中科院大学人工智能学院

2019年12月19日

中国科学院
自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

课程提纲

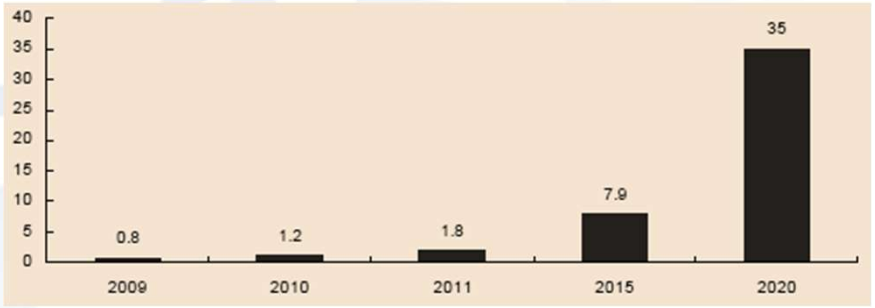
一、何谓大数据时代

二、大数据机器学习

三、典型应用举例

四、开源软件平台

数据量增加

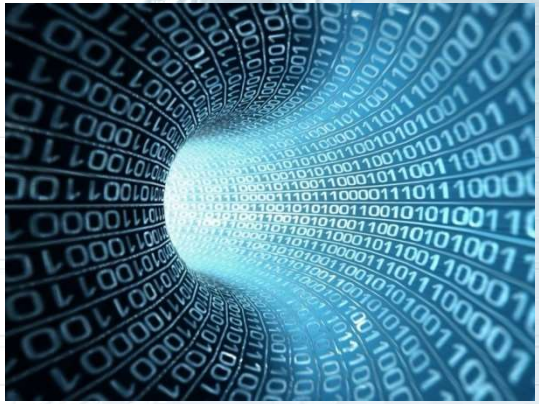


根据IDC 监测，人类产生的数据量正在呈指数级增长，大约每两年翻一番，这个速度在2020 年之前会继续保持下去。这意味着人类在最近两年产生的数据量相当于之前产生的全部数据量

TB ⇨ PB ⇨ EB ⇨ ZB

数据结构日趋复杂

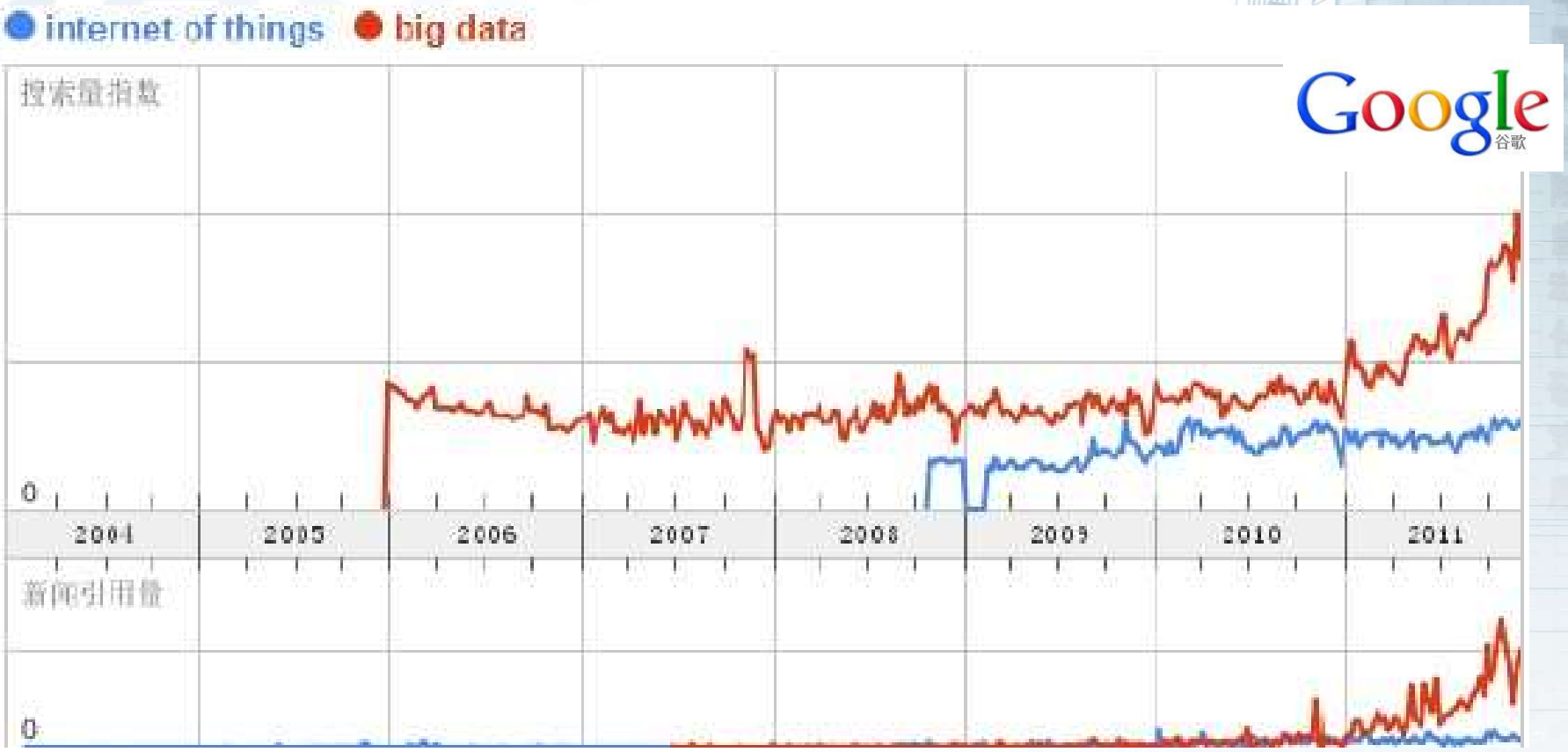
大量新数据源的出现则导致了非结构化、半结构化数据爆发式的增长



- 这些由我们创造的信息背后产生的这些数据早已经远远超越了目前人力所能处理的范畴
- 大数据时代正在来临…

EMC²

2011年5月，在“云计算相遇大数据”为主题的EMC World 2011会议中，EMC抛出了Big Data概念



“大数据”成为学术界关注焦点

中国科学院
自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

How to cope with the flood of data-

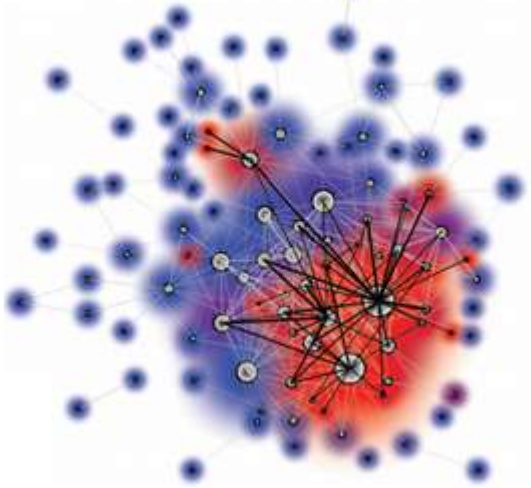
-- one of the **most daunting challenges** facing modern science.

Science
AAAS
Detecting Novel Associations in Large Data Set
David N. Reshef, *et al.*
Science **334**, 1518 (2011);
DOI: 10.1126/science.1205438

Detecting Novel Associations in Large Data Sets

David N. Reshef,[1,2,3*†] Yakir A. Reshef,[2,4*†] Hilary K. Finucane,[5] Sharon R. Grossman,[2,6] Gilean McVean,[3,7] Peter J. Turnbaugh,[6] Eric S. Lander,[2,8,9] Michael Mitzenmacher,[10‡] Pardis C. Sabeti[2,6‡]

Identifying interesting relationships between pairs of variables in large data sets is increasingly important. Here, we present a measure of dependence for two-variable relationships: the maximal information coefficient (MIC). MIC captures a wide range of associations both functional and not, and for functional relationships provides a score that roughly equals the coefficient of determination (R^2) of the data relative to the regression function. MIC belongs to a larger class of maximal information-based nonparametric exploration (MINE) statistics for identifying and classifying relationships. We apply MIC and MINE to data sets in global health, gene expression, major-league baseball, and the human gut microbiota and identify known and novel relationships.



大数据开启了一次重大的时代转型

- 望远镜让我们能够感受宇宙
- 显微镜让我们能够观测微生物
- 大数据正在改变我们的生活和理解世界的方式，成为新发明和新服务的源泉
- 更多的改变正蓄势待发！！！

中国科学院
自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

课程提纲

一、何谓大数据时代

二、大数据机器学习

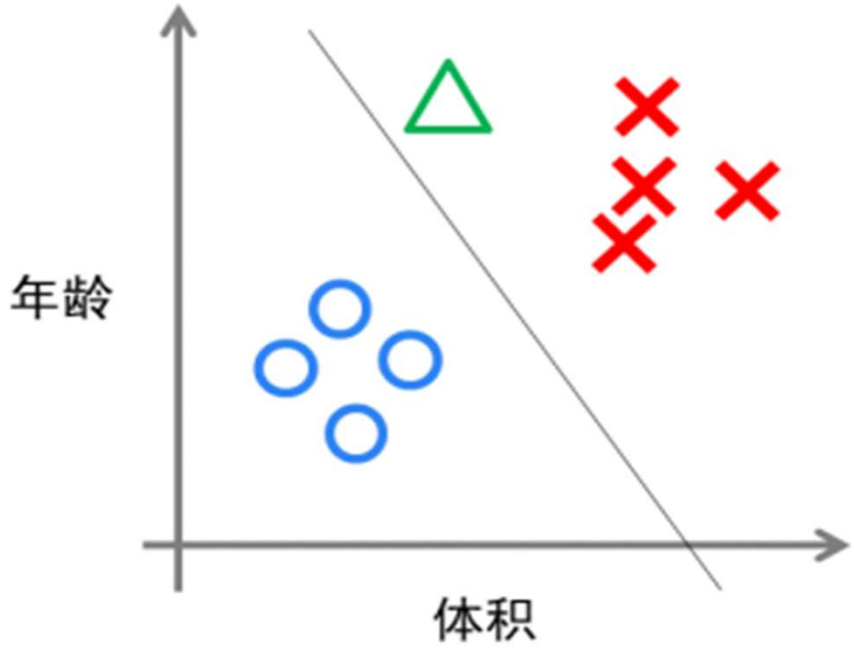
三、典型应用举例

四、开源软件平台

2014年5月16日，吴恩达加入百度，担任百度公司首席科学家，负责百度研究院、Baidu Brain计划

Geoffrey Hinton:Deep Learning祖师爷，多伦多大学教授，负责Google大脑

Facebook人工智能实验室主任、NYU数据科学中心创始人、深度学习界的泰斗Yann LeCun



目标：预测肿瘤的性质

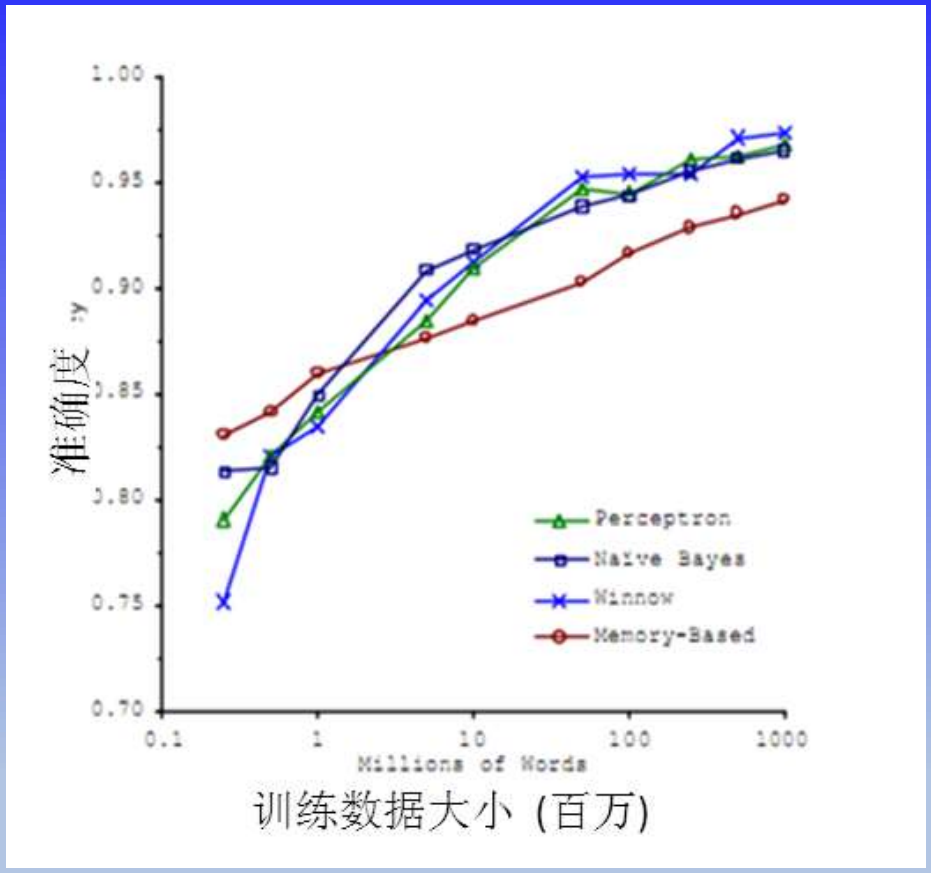
输入：肿瘤的体积，
患者的年龄

输出：良性或恶性


WARNING
WARNING
WARNING
WARNING
WARNING
WARNING
WARNING
H1N1
DANGER

Google


淘汰赛赛事预测
更新时间：07月11日
06.29 00:00
巴西 4
智利 3
06.29 04:00
哥伦比亚 2
乌拉圭 0
07.01 00:00
法国 2
尼日利亚 0
07.01 04:00
德国 2
阿尔及利亚 1
07.05 04:00
巴西 2
哥伦比亚 1
07.05 00:00
法国 0
德国 1
07.09 04:00
巴西 1
德国 7
冠军
德国
阿根廷
07.14 03:00
三四名决赛 07.13 04:00
巴西
荷兰
07.10 04:00
荷兰 2
阿根廷 4
07.06 04:00
荷兰 4
哥斯达黎加 3
07.06 00:00
阿根廷 1
比利时 0
06.30 00:00
荷兰 2
墨西哥 1
06.30 04:00
哥斯达黎加 6
希腊 4
07.02 00:00
阿根廷 1
瑞士 0
07.02 04:00
比利时 2
美国 1



It's not who has the best algorithm that wins .

It's who has the most data.

一种流行的大数据机器学习方法

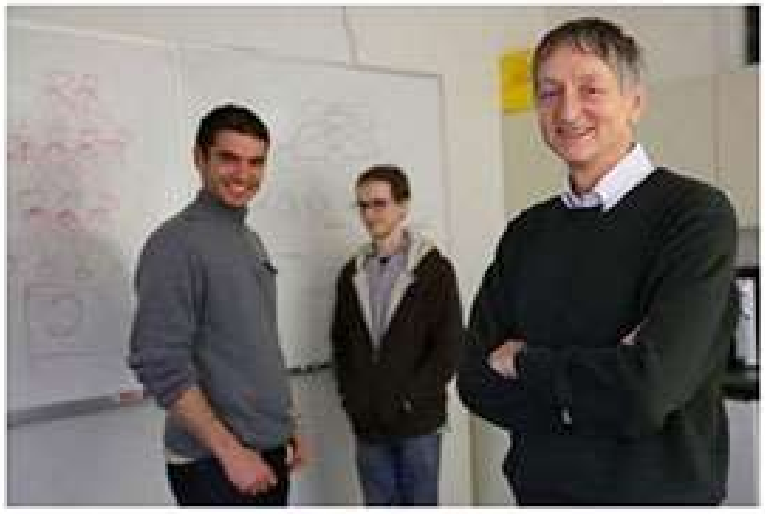
深度学习介绍【下次课详细讲】

2006年，Geoffrey Hinton在科学杂志《Science》上发表了一篇文章，论证了两个观点：

1. 多隐层的神经网络具有优异的特征学习能力，学习得到的特征对数据有更本质的刻画，从而有利于可视化或分类

2.深度神经网络在训练上的难度，可以通过“逐层初始化”来有效克服

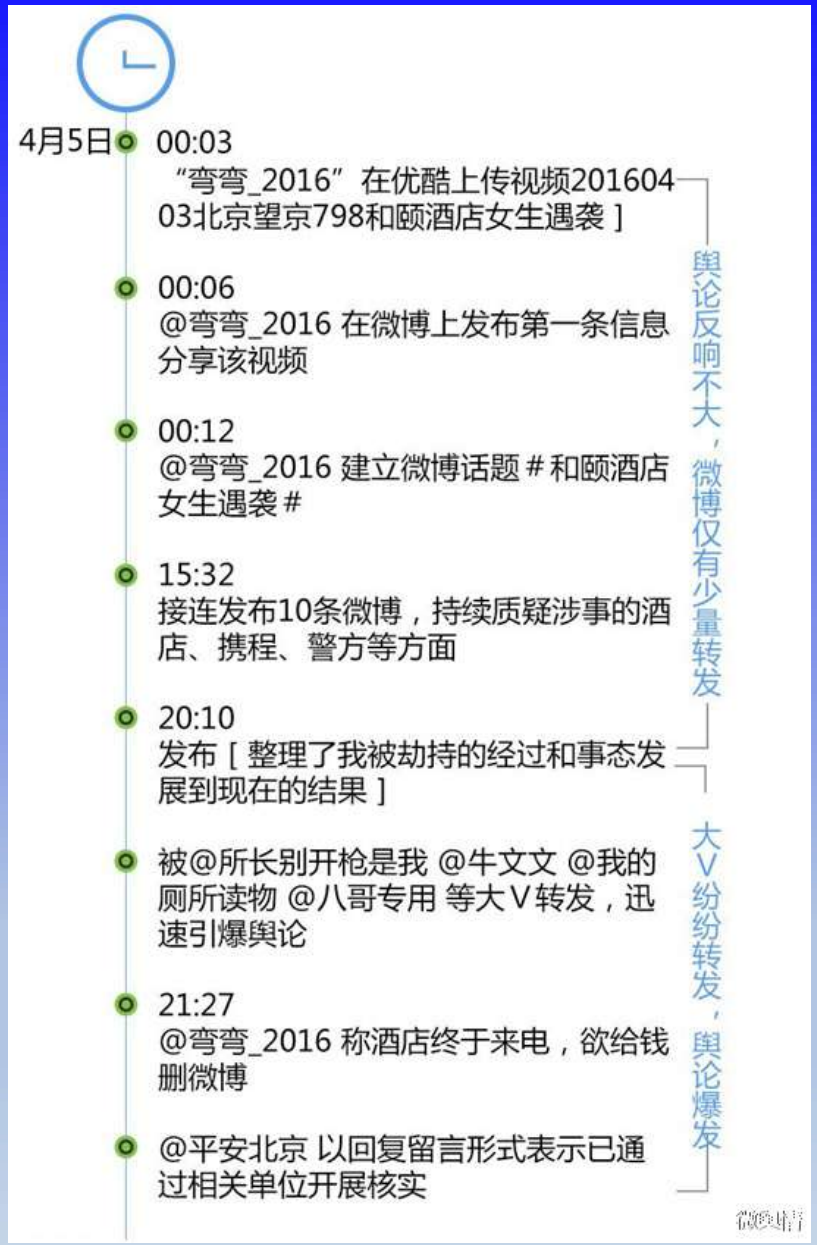
2006年，Geoffrey Hinton在科学杂志《Science》上发表了一篇文章，论证了两个观点：



2012年6月，《纽约时报》披露了Google Brain项目，这个项目是由Andrew Ng和Map-Reduce发明人Jeff Dean共同主导，用16000个CPU Core的并行计算平台训练一种称为“深层神经网络”的机器学习模型，在语音识别和图像识别等领域获得了巨大的成功，Andrew Ng就是文章开始所介绍的机器学习的大牛

2013年4月，《麻省理工学院技术评论》杂志将深度学习列为2013年十大突破性技术(Breakthrough Technology)之首

2015年10月阿尔法围棋以5：0完胜欧洲围棋冠军、职业二段选手樊麾。2016年3月挑战世界围棋冠军、职业九段选手李世石。根据日程安排，5盘棋将分别于3月9日、10日、12日、13日和15日举行，即使一方率先取得3胜，也会下满5盘。最后以4:1结束了这场“战争”。


4月5日
00:03
“弯弯_2016”在优酷上传视频20160403北京望京798和颐酒店女生遇袭]
00:06
@弯弯_2016 在微博上发布第一条信息分享该视频
00:12
@弯弯_2016 建立微博话题＃和颐酒店女生遇袭＃
15:32
接连发布10条微博，持续质疑涉事的酒店、携程、警方等方面
20:10
发布 [整理了我被劫持的经过和事态发展到现在的结果]
被@所长别开枪是我 @牛文文 @我的厕所读物 @八哥专用 等大Ｖ转发，迅速引爆舆论
21:27
@弯弯_2016 称酒店终于来电，欲给钱删微博
@平安北京 以回复留言形式表示已通过相关单位开展核实
舆论反响不大，微博仅有少量转发
大Ｖ纷纷转发，舆论爆发

中国科学院
自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

课程提纲

一、何谓大数据时代

二、大数据机器学习

三、典型应用举例

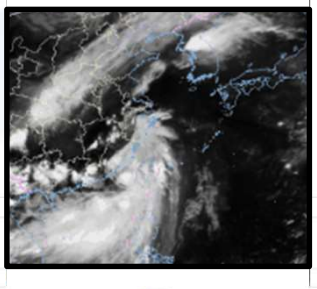
四、开源软件平台

- 气象数据真实存在的大数据
- 大数据颠覆了降水计算模式
- 精准降水估计高效学习算法

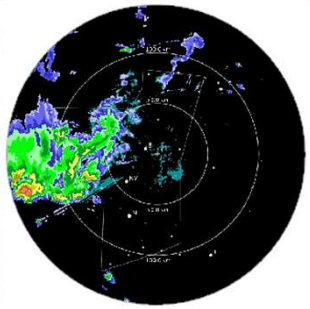
多源气象探测信息构成了名副其实的大数据

海量实时、多源异构、复杂多变

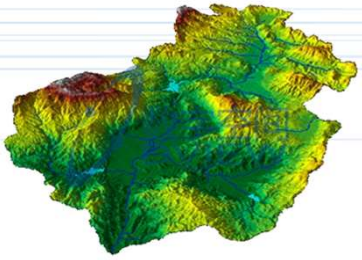
卫星云图



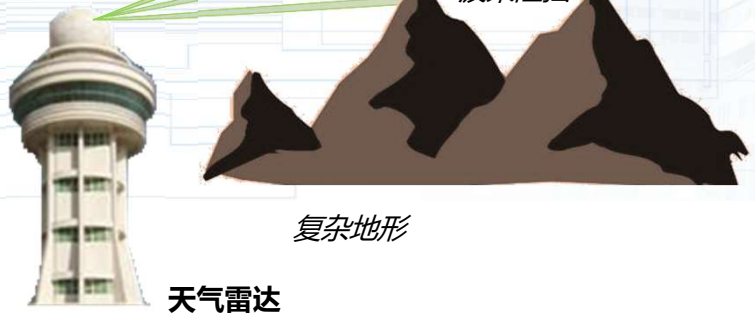
雷达回波

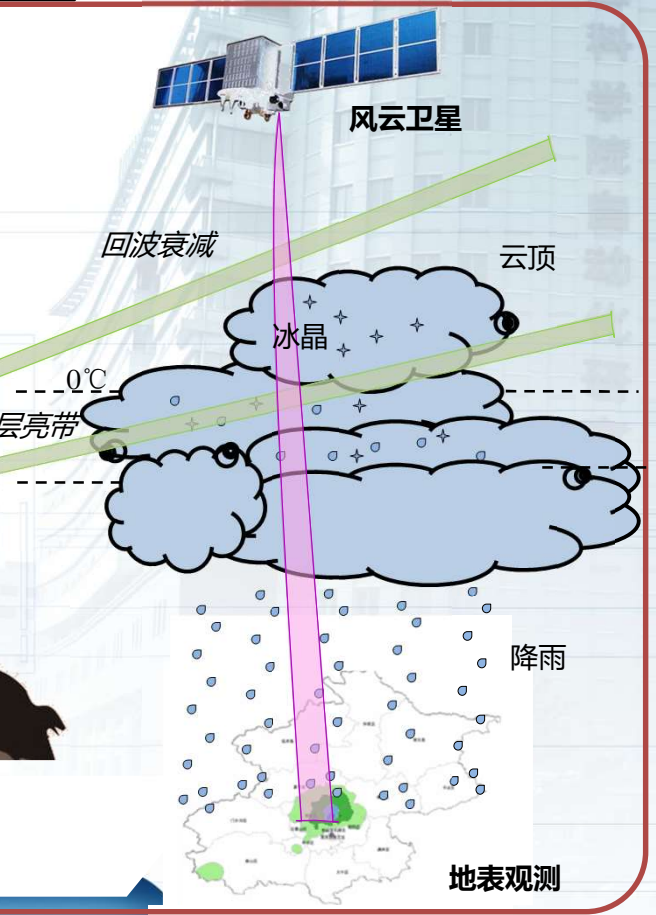


地理信息



- 23颗气象卫星92类卫星产品
  5000万文件3PB卫星数据
- 21类雷达产品超10PB数据
  600万观测/雷达/6分钟
- 6万自动气象站实时数据
  全国30m级DEM数据





- 气象数据真实存在的大数据
- 大数据颠覆了降水计算模式
- 精准降水估计高效学习算法

中国科学院
自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

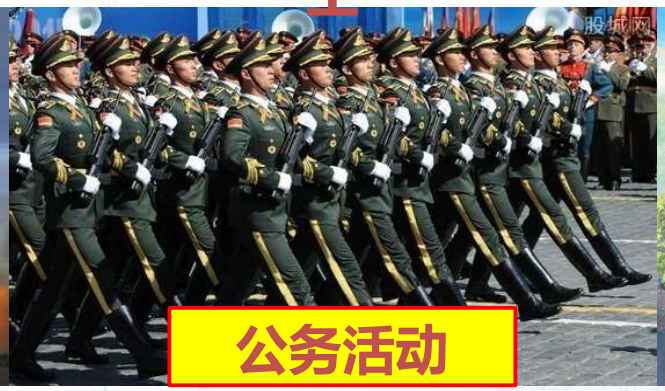
超短临降水预报服务迫切需求


抢险救灾


大型工程施工


高铁运行保障


航天器发射


公务活动


出行游乐

气象观测技术飞速发展，但我国目前还无法做到分钟级精确降水预报


兴 宁 路

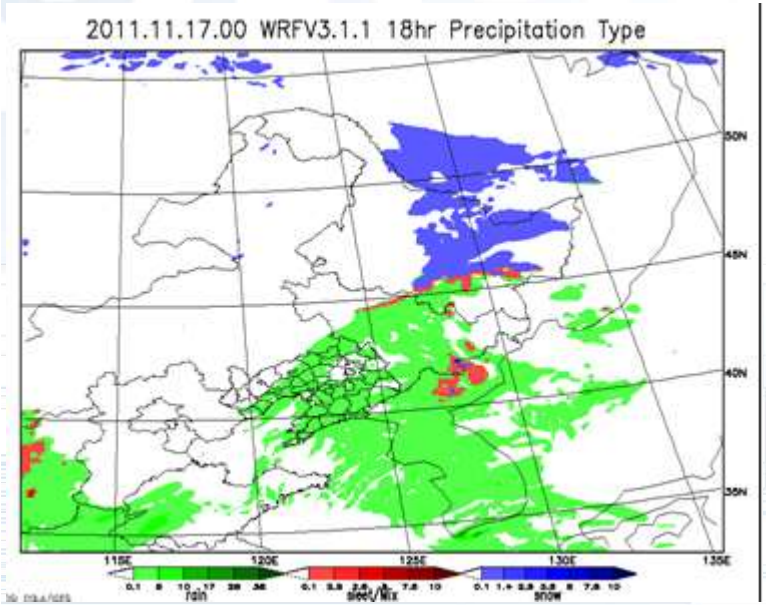
超短临降水是联合实验室业务需求

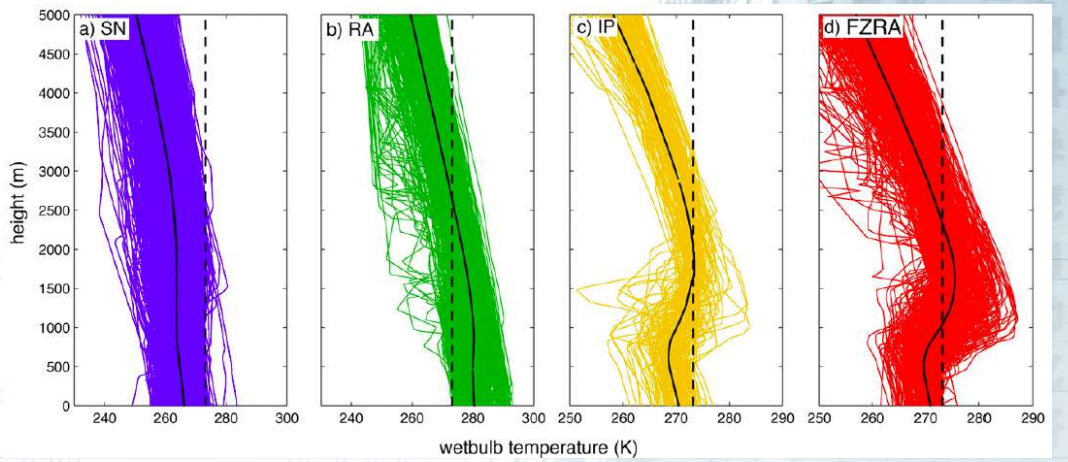
- 2015年6月，国家气象局公共气象服务中心和中国科学院自动化研究所和共同组建了“气象大数据与机器学习联合实验室”

- 超短临降水预报(分钟级)是联合实验室业务需求

降水相态识别：特别是春秋降水估计

如果将雨误判成雪，可能会将小雨级别的降水误报为大雪





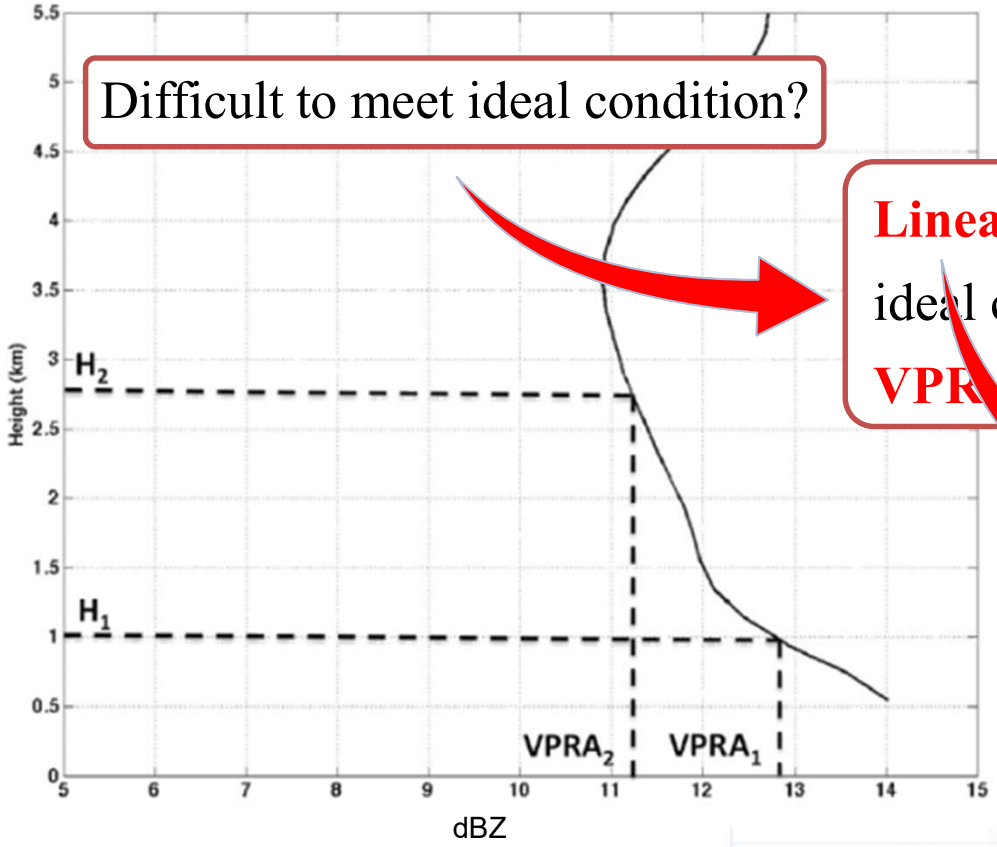
在我国探空气球非常少

降水相态主要与空间上不同高度的温度分布有关

- 气象数据真实存在的大数据
- 大数据颠覆了降水计算模式
- 精准降水估计高效学习算法

VPR Structure

$$\log(Z) = b\log(R) + \log(a) \quad \Rightarrow \quad R = f(\mathbf{Z})$$



Difficult to meet ideal condition?

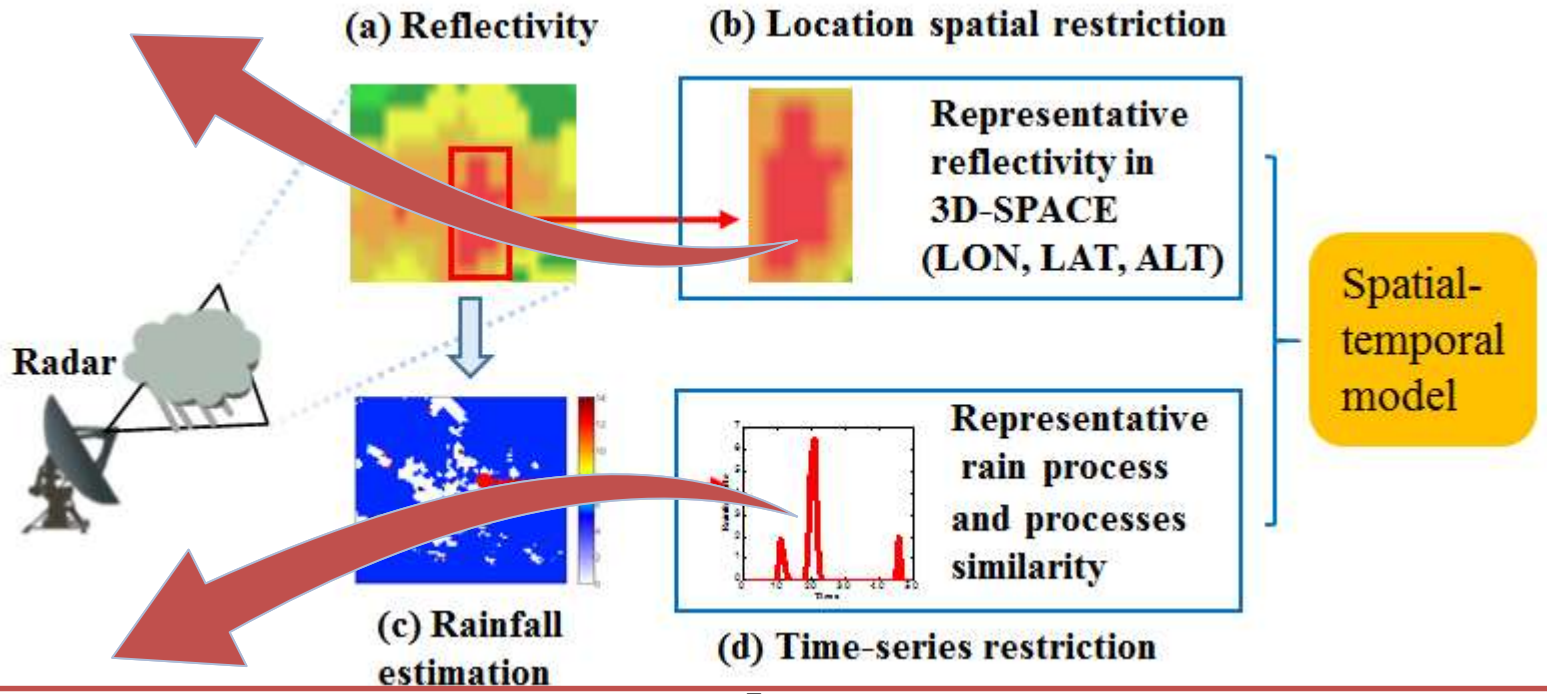
Linearity suits for measuring how the ideal condition is relaxed and capturing **VPR structure**

$$\rho_{\log(Z^h),\log(R)} = \frac{\mathrm{cov}\left(\log(Z^h),\log(R)\right)}{\sigma_{\log(Z^h)}\sigma_{\log(R)}}$$

Which level to select? → Consider VPR structure in statistical model

Spatial and temporal structure

spatial structure $\min_{\Theta} f(\Theta, x', y) = \sum \left(\sum_{v \in S_j} \left(\left[y(\Theta) - \bar{y}_j(\Theta) \right]^2 \right) - \sum_{i \in L,R} \sum_{v \in S_j} \left(\left[y(\Theta) - \bar{y}_j(\Theta) \right]^2 \right) \right)$



temporal structure $\min_{\omega,\lambda} g(\omega, \lambda) = \sum_{k=1}^{p} \left[\sum_{j=1}^{q(k)} \log\left(Z_k(x) \right) - \sum_{j=1}^{q(k)} \sum_{t=1}^{T} \omega_{kt} P_{kj}\left(y_{t-1}, y_t, x_t \right) + \lambda \left\| \omega_k \right\|_1 \right]$

Rainfall space-time characteristic?
→ Consider spatial and temporal structure in statistical model

Result

Comparison between radar rainfall estimation and raingauge measurement for Hangzhou radar coverage

	$RMSE_{mm}$	MAE_{mm}	CC
Z-R	3.05	1.64	0.659
SVR	2.69	1.51	0.717
RF	2.80	1.41	0.718
RANMP	2.72	1.36	0.737
LIST	2.51	1.14	0.763
RANLIST	**2.15**	**1.05**	**0.829**

1. Compared to conventional Z-R relationship:
Improvements of **30% in RMSE**, **36% in MAE** and **26% in CC** are obtained

2. Compared to random forest:
Improvements of **23% in RMSE**, **26% in MAE** and **15% in CC** are obtained

实验效果

Data set	P_{min} (%)						P_{avg} (%)					
	Base	SSMOTE	GCS	ABNC	OSMOTE	AMDO	Base	SSMOTE	GCS	ABNC	OSMOTE	AMDO
Contraceptive	$43.55_{6.04}$	$38.12_{6.78}$	$40.24_{5.81}$	$36.02_{4.51}$	**$47.10_{12.20}$**	*$45.36_{6.98}$*	$52.07_{2.15}$	$48.30_{3.56}$	$49.24_{0.74}$	$49.21_{2.19}$	$51.86_{3.04}$	**$52.52_{2.09}$**
Flare	$2.22_{4.97}$	$20.56_{14.65}$	**$30.28_{6.39}$**	$20.56_{12.36}$	$6.67_{9.94}$	*$25.56_{18.26}$*	$59.84_{1.51}$	$61.37_{3.40}$	*$63.04_{1.93}$*	$58.02_{1.58}$	$58.57_{1.80}$	**$63.45_{1.74}$**
Thyroid	$96.40_{2.49}$	$97.01_{2.99}$	*$97.61_{2.47}$*	$95.79_{2.68}$	**$98.81_{1.64}$**	$93.99_{2.97}$	$96.70_{1.58}$	$96.96_{0.49}$	*$97.34_{0.89}$*	$96.55_{0.74}$	**$97.72_{0.81}$**	$95.67_{0.98}$
Car	$63.08_{22.03}$	$61.54_{19.61}$	*$89.23_{6.88}$*	$67.69_{26.87}$	$61.54_{26.09}$	**$98.46_{3.44}$**	$77.26_{5.59}$	$76.81_{4.82}$	*$91.72_{4.12}$*	$78.54_{5.63}$	$80.37_{9.40}$	**$91.97_{1.31}$**
Nursery	$67.71_{3.70}$	$68.64_{6.64}$	*$77.45_{6.54}$*	$66.48_{3.31}$	$64.36_{6.46}$	**$93.90_{2.85}$**	$90.12_{1.19}$	$90.32_{1.94}$	*$93.45_{1.78}$*	$89.84_{1.05}$	$89.31_{1.86}$	**$95.75_{1.20}$**

□ **Xuebing Yang**, Qiuming Kuang, Wensheng Zhang, and Guoping Zhang, "AMDO: an Over-Sampling Technique for Multi-Class Imbalanced Problems," ***IEEE Transactions on Knowledge and Data Engineering***, 2017. (**SCI，CCF-A**)

TABLE 9
Means and Standard Deviations of AUCm (%) and MAUC (%) Results from Base, SSMOTE, GCS, ABNC, OSMOTE and AMDO for the Nominal/Mixed-type Data Sets

Data set	AUCm (%)						MAUC (%)					
	Base	SSMOTE	GCS	ABNC	OSMOTE	AMDO	Base	SSMOTE	GCS	ABNC	OSMOTE	AMDO
Contraceptive	*$64.23_{2.14}$*	$60.41_{3.20}$	$60.78_{1.44}$	$61.57_{1.32}$	$64.04_{3.10}$	**$64.63_{2.39}$**	*$64.05_{1.61}$*	$61.22_{2.67}$	$61.93_{0.55}$	$61.91_{1.64}$	$63.89_{2.28}$	**$64.39_{1.56}$**
Flare	$63.18_{0.96}$	$67.42_{4.40}$	**$70.28_{1.76}$**	$66.68_{3.30}$	$63.93_{2.43}$	*$69.53_{4.81}$*	$75.91_{0.90}$	$76.82_{2.04}$	*$77.82_{1.16}$*	$74.81_{0.95}$	$75.14_{1.08}$	**$78.07_{1.04}$**
Thyroid	$97.84_{1.11}$	$98.08_{0.86}$	*$98.37_{0.94}$*	$97.63_{0.87}$	**$98.82_{0.66}$**	$96.84_{1.07}$	$97.53_{1.19}$	$97.72_{0.37}$	*$98.00_{0.67}$*	$97.41_{0.55}$	**$98.29_{0.60}$**	$96.75_{0.73}$
Car	$81.64_{6.67}$	$81.08_{5.80}$	*$94.52_{3.03}$*	$82.99_{7.54}$	$82.94_{9.49}$	**$95.52_{1.08}$**	$84.84_{3.73}$	$84.54_{3.21}$	*$94.48_{2.75}$*	$85.69_{3.75}$	$86.92_{6.27}$	**$94.64_{0.87}$**
Nursery	$88.53_{1.33}$	$88.82_{2.32}$	*$92.12_{2.23}$*	$88.15_{1.18}$	$87.42_{2.24}$	**$96.65_{1.20}$**	$93.41_{0.79}$	$93.54_{1.29}$	*$95.63_{1.18}$*	$93.23_{0.70}$	$92.87_{1.24}$	**$97.16_{0.80}$**
Splice	*$95.54_{1.27}$*	$95.36_{1.21}$	$95.36_{1.29}$	$93.31_{2.51}$	$95.45_{1.33}$	**$95.99_{1.01}$**	$95.67_{0.96}$	$95.47_{0.73}$	$95.43_{0.92}$	$93.37_{2.00}$	*$95.69_{0.85}$*	**$96.09_{0.55}$**
PreND	$84.98_{0.00}$	$90.84_{3.25}$	$95.00_{0.34}$	*$95.12_{0.47}$*	$93.74_{1.99}$	**$95.15_{0.55}$**	$89.68_{0.00}$	$93.52_{2.53}$	$96.42_{0.23}$	*$96.63_{0.42}$*	$95.43_{0.78}$	**$96.70_{0.48}$**
Average	82.28	83.14	*86.63*	83.64	83.76	**87.76**	85.87	86.12	*88.53*	86.15	86.89	**89.11**
Mean rank	3.29	3.57	2.43	3.86	3.14	**1.71**	4.00	4.29	*2.86*	4.71	3.43	**1.71**

The best result is in bold face, and the second best result is in italics.

中国科学院
自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

课程提纲

一、何谓大数据时代

二、大数据机器学习

三、典型应用举例

四、开源软件平台

中国科学院
自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

AutoML开源软件平台

- 近年来，机器学习在各个领域都取得了重大突破。金融服务、医疗保健、零售、交通等领域一直在以某种方式使用机器学习系统，而且取得了很好的效果
- 传统的 ML 流程仍依赖于人力，但并非所有企业都有资源来投资经验丰富的数据科学团队，AutoML 可能正是这种困境的一个答案

- **自动机器学习**（AutoML）是将机器学习应用于现实问题的端到端流程自动化的过程
- AutoML 对于没有该领域专业知识的人使机器学习成为可能
- AutoML 通过使不同背景的人能够演进机器学习模型来解决复杂的场景，正在从根本上改变基于 ML 的解决方案给人们的印象

- Mercari 是一款在日本很受欢迎的购物应用程序，它一直使用 AutoML Vision（谷歌的 AutoML 解决方案）对图像进行分类
- Mercari 一直在“开发自己的 ML 模型，在照片上传的用户界面上推荐 12 个主要品牌的品牌名称”

- 在 TensorFlow 上训练的模型达到了 75% 的精度，AutoML Vision 的高级模式拥有 50,000 张训练图像，所以精度更是高达 91.3%，提升了竟有 15%
- 基于如此惊人的结果，Mercari 已经将 AutoML 集成到他们的系统中

AutoML 优 点

- 通过自动执行的重复性任务来 **提高工作效率** 。这使得数据科学家能够更多地关注问题而不是模型
- 自动化 ML 管道还有助于 **避免** 可能因手动引入的 **错误**
- AutoML 是向 **机器学习民主化** 迈出的一步，它使所有人都能使用 ML 的功能

中国科学院
自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

感谢同学们听课

欢迎讨论与交流