AutoStat@Big²Data.IoT

(Statistics for Internet of Things)

物联网时代的统计思维

Dennis Lin（林共进）
University Distinguished Professor
Department of Statistics
The Pennsylvania State University
(美国宾州州立大学统计系杰出讲座教授)

at
2015 JMP Discovery Summit数据分析峰会
15 May, 2015

统计的未来在哪里？

What is Future Statistics?

今日报告主要议题 (Main issues)

- 什么是物联网 What is IoT?
- 物联网下的统计思维 Statistics for IoT?
- 自动化统计 (Automatic Statistics)
- 个性化统计 (Personalization Statistics)
- *追求系统化的知识，而非支离破碎的常识*
- *要花钱买体验，而不是买物品*

Did you know???

微信：十大热门学科

Top Ten (10) Professions in China

统计到底是干什么的？

What's
Statistics
All About?

镜子与窗子

镜子 (统计描述 Descriptive Statistics)
窗子 (统计推断 Statistical Inference)

镜子(统计描述 Descriptive)

白的不会是黑的，
红的不会是蓝的，
该什么就是什么，
我不在其中

原原本本反映出数据的真实面貌

窗子(统计推断 Inference)

- 推断
- 统计推断
- 基本上为演绎与归纳

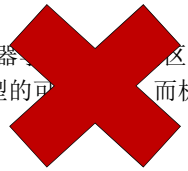
科学探索

- 演绎（Deduction）
  - 物理
  - 数学
- 归纳（Induction）
  - 生物
  - 化学
- 统计分析比较偏重于观察与实验的积累（也就是“归纳”类型的研究）。

推断（预测），各个学科都有
描述（建模），只有少数学科 有（统计）

Ryan Adams:

“我认为统计学和机器[illegible]区别在于根本目标不同。统计学家更关心模型的可[illegible]而机器学习专家更关心模型的预测能力。”



万事万物有其相同性，
这使得统计变成可能；
Things are similar
which makes Statistics possible.

万事万物有其相异性，
这使得统计变成必须。
Things are different
which makes Statistics necessary.

Dennis Lin

理论与实践

- 理论（传递性）
  - 数学

    A=B, B=C, 那么 A=C
- 真实（实用性）
  - 统计 H_1: $\mu_A=\mu_B$

    H_2: $\mu_B=\mu_C$ then H_3: $\mu_A\neq\mu_C$
  - 社交
    - 甲爱乙， 乙爱丙， 那么甲会爱丙?

Statistics for Big Data: Are Statisticians Ready?

John Jordan and Dennis K.J. Lin
(ICSA-Bulletine 2014)

With Considerable Input from Roger Hoerl

什么是大数据?
What is BIGdata?

谁在搞 大数据?
Who is working on BIGdata?

大数据需要什么样的统计?
What kinds of Statistics are needed?

Website

大数据就像是年轻人谈论性爱
Big data is like teenagers' sex…

- Everyone talk about it…
  Nobody really knows how to do it.
  所有人都在说，没人真正知道怎么做
- Everyone thinks everyone else is doing it…
  So everyone claims they are doing it.
  所有人都觉得其他人都在做，
  所以 所有人都声称 自己也在做

谁在搞大数据
Who is working on BIGdata?

- 计算机科学 Computer Science
- 信息科学 Information Technology
- 应用数学 Applied Mathematics
- 计算数学 Computational Math/Science
- 运筹学 Operation Research
- 工业工程 Industrial Engineering Informatics
- 数据科学 Science/Data Scientist
- 电子工程 Electronic Engineering
- 政治科学 Political Science
- *所有需要钱的人 Anyone who needs funding $$$*

BIG Data大数据

这是统计的黄金时代，
但不一定是统计家的黄金时代
This is a golden age for Statistics,
But not necessary for statisticians.

--Hahn and Hoerl

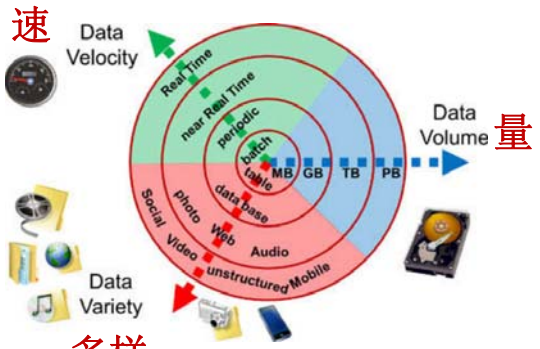
怎么会这样 How come?????

当 数据流行时，统计出现了！

当大数据流行时，统计消失了！

When Data is popular, Statistics appears!
When Big Data is popular, Statistics disappear!!!

大数据特征
Characteristics of Big Data



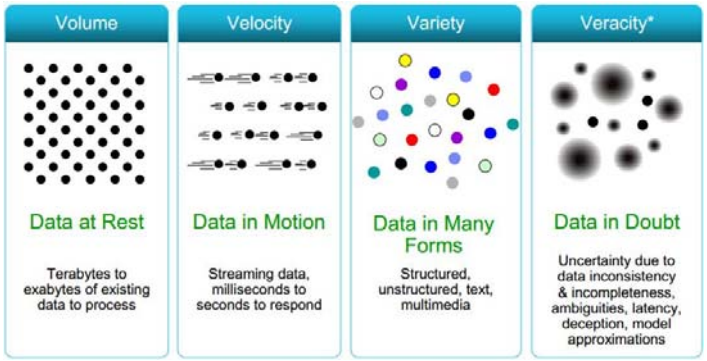
19

谨记 (Keep in mind)

- 大数据单是大，但未必完整，准确，真实
  Big data is simple big, but not necessarily complete, accurate, or true!
- 价值由评价的人决定，而不是处理数据的人
  Value is in the eye of beholder, not the person crunching the numbers!
- 更大未必更好...如果有初始偏差
  Bigger does not always implies better

大数据4V (Big Data: The 4V's)

体量， 速度， 多样， 真实



理想中的统计：愿景清单

What kind of Statistics: Wish-List

- 探索有意义的问题 *Seek out high-impact problems*
  - Do Something Matter
- 为未定义问题提供结构 Provide structure for poorly defined problems
- 为实际的大数据世界推出新理论 Develop new theories for new (reality) BIG data world.

Some examples 一些例子

- 描述性的统计量和图表 Statistics and plots for (many descriptive) statistics
  - 怎么综合上千的p值？ How to summarize thousands of p-values?
  - 相关性？方差分析？直方图？ Correlations? ANOVA tables? Histograms?
- 低维表现 Low-dimension behavior
- 正态还是极端 Norm or Extreme
  - 模式识别和特征提取 Pattern Recognition and Feature Extraction
- 新型数据的方法 Methodology for new type of data
  - 网络数据的回归模型 Regression model for network data?
- 预测与估计 Prediction vs Estimation

统计真正的挑战和机会

第三个V：多样

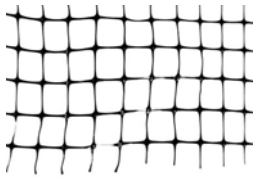
*The 3rd V: **Variety***

新型数据 New Types of Data

先来上一点英语课
English -101

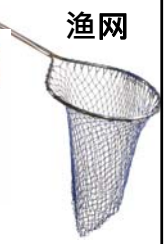
- Net 网
- Internet 互联网
- Internet of thing 物联网
- Internet of everything 万物联网

Nets (网：能困住你的东西)




格子点网

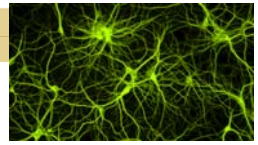


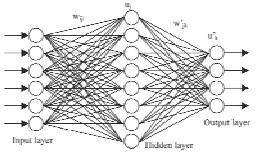
渔网

情网　　蜘蛛网　　篮网

Network 网路



网络是一个由线相互连接组成的系统




Internet 因特网
—互相联系的 网



互联网（Internet），又称网际网路，是网路与网路之间所串连成的庞大网路

Internet of Thing (IoT) 物联网

物联网（Internet of Things，IOT）
是一个基于互联网，
让所有能够被独立寻址的普通物理对象
实现互联互通的网络。

新常态：从信息技术到数据技术

New Normal: From IT to DT



Data Technology

物联网 ***Internet of Things***

在虚拟网络中联结客观事物

connecting physical things in a virtual networks

- 网络不是目的导向 Network is not purpose-oriented.
- 系统完全自制化 System is fully customized.
- 人工干预最小化，但可完全控制 Human intervention is minimized, but dominating.

物联网 ***Internet of Things***

结合 Integration of
快 Too Fast,
大 Too Large, and
复杂 Too Complicate.

模型包含的越广，就有越多的噪声。
你需要统计！
The more inclusive the model, the more noise is introduced. You will need Statistics!

物联网：

什么Internet和什么Thing联结？

（即使他们都存在）

- 如何思考？
- 懂得别人会怎样思考？
- 然后想想自己该怎么思考（逆向思维）？

自动化 Automatic

自动化诗歌 Automatic Poem
自动化报告 Automatic Report
自动化统计 Automatic Statistics

个性化 Personalization

个性化统计 Personalized Statistics
个性化数据 Personalized Data
个性化JMP Personalized JMP

写首诗送给你 (Auto-Poem)

1空	21一笑	41深处	61一片	81不是
2东风	22黄昏	42时节	62桃李	82时候
3何处	23当年	43平生	63人生	83肠断
4人间	24天涯	44凄凉	64十分	84富贵
5风流	25相逢	45春色	65心事	85蓬莱
6归去	26芳草	46匆匆	66黄花	86昨夜
7春风	27尊前	47功名	67一声	87行人
8西风	28一枝	48一点	68佳人	88今夜
9归来	29风雨	49无限	69长安	89谁知
10江南	30流水	50今日	70东君	90不似
11相思	31依旧	51天上	71断肠	91江上
12梅花	32风吹	52杨柳	72而今	92悠悠
13千里	33风月	53西湖	73鸳鸯	93几度
14回首	34多情	54桃花	74为谁	94青山
15明月	35故人	55扁舟	75十年	95何时
16多少	36当时	56消息	76去年	96天气
17如今	37无人	57憔悴	77少年	97惟有
18阑干	38斜阳	58何事	78海棠	98一曲
19年年	39不知	59芙蓉	79寂寞	99月明
20万里	40不见	60神仙	80无情	100往事

π= 3.1415926535

3 14 15 92 65 35

3 何处
14 回首
15 明月
92 悠悠
65 心事
35 故人

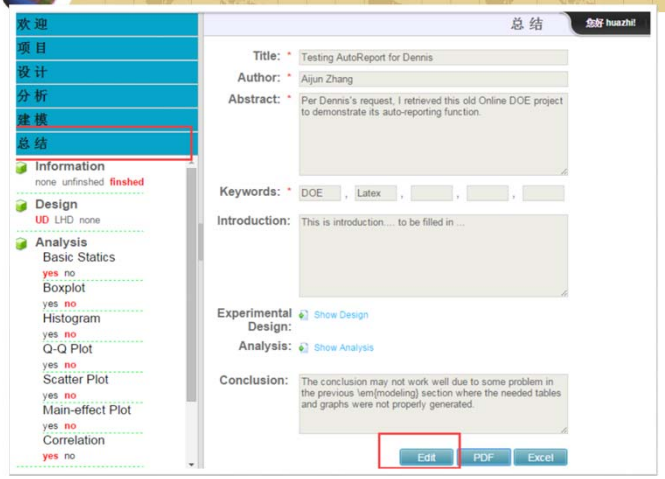
何处回首明月，
悠悠心事故人

自动化报告 Auto Report

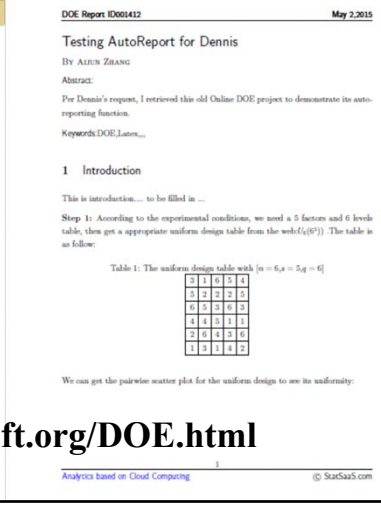
http://www.statsoft.org/DOE.html

Visit http://www.statsoft.org/DOE.html
Use AJ's demo account: huazhi/huazhi12

Go to "Summary", Click "Edit" to edit the inputs



Finally,
Click "PDF"
or "Excel"
to generate reports


DOE Report ID001412 May 2,2015

Testing AutoReport for Dennis

By Aijun Zhang

Abstract:

Per Dennis's request, I retrieved this old Online DOE project to demonstrate its auto-reporting function.

Keywords: DOE,Latex,...

1 Introduction

This is introduction.... to be filled in ...

Step 1: According to the experimental conditions, we need a 5 factors and 6 levels table, then get a appropriate uniform design table from the web:$U_6(6^5)$. The table is as follow:

Table 1: The uniform design table with $[n = 6, s = 5, q = 6]$

3	1	6	5	4
5	2	2	2	5
6	3	3	6	3
4	4	5	1	1
2	6	4	3	6
1	3	1	4	2

We can get the pairwise scatter plot for the uniform design to see its uniformity:

Analytics based on Cloud Computing © StatSaaS.com

http://www.statsoft.org/DOE.html

自动化统计 Automatic Statistics

- 自动化统计学家 (Automatic Statistician)
- 统计的人工智能 (Artificial Intelligent for Statistics)
- 自动化过程，包括：
  - 统计质量控制 (Statistical Process Control)
  - 统计模型选择 (Model Selection)
  - 数据分析 (Data Analysis & Prediction)
  - 报告 (Report)
  - etc

个性化统计
Personalization Statistics

- 个性化电脑 Personal Computer
- 个性化手机 Personal Phone
- 个性化网站 Personal website
- 个性化医疗 Personal medicine
- 个性化统计 Personal Statistics
- 个性化**JMP** Personal JMP

个性化统计（利用物联网）
Personal Statistics (with IoT)

- 健康数据 Health Data
- 财政数据 Financial Data
- 社交数据 Social Network Data
- 旅行、食品、兴趣数据 Travel/Food/Hobby Data
- 物联网—连接性 Internet of Things—Connectivity
- 统计—描述和推断 Statistics—Descriptive & Inference

(a) 追求系统化的知识，而非支离破碎的常识
Looking for systematic knowledge, not piecewise information.

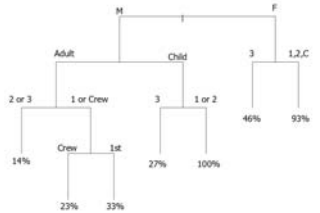
(b) 花钱买体验，而不是买物品
Buying experience, not the good!

追求系统化的知识，而非支离破碎的常识。

破碎的知识

- 船是哪一年造的？
- 船的造价是多少？
- 船长叫什么名字？
- 船的负载能力如何？
- 船主是从事什么行业？
- …

系统化的学问



闻道有先后（知识），术业有专攻（学问）。

谷歌/百度 (Google/Baidu)

我们培养了很多高学历的野蛮人

文 | 鲍鹏山

- 知识就是力量，良知才是方向。
- 没有知识可以被宽容，没有良知不可以被宽容。
- 生活中有太多这样无用的知识—无趣、无聊。
  - 很多人关心某个明星喜欢的颜色是什么，星座是什么，结了几次婚，又离了几次婚。
  - 把精力花在这些地方时，他可能获得了知识，并且在饭桌上能与人聊天，但他会变得特别琐碎。
- 用琐碎的知识，把自己的人生切割成碎片。
- 追求系统化的知识，而非支离破碎的常识。

*要花钱买**体验** (experience)，*
*而不是买**物品** (goods)。*

当基本需求都已经满足之后...

花钱买体验，而不是买物品

iPhone/BMW 是物品
跳伞/音乐是体验

物件都是生不带来死不带走的独立存在。
你的人生体验才是你真正的自己。

Send $500 to

- Dennis Lin

University Distinguished Professor
317 Thomas Building
Department of Statistics
Penn State University

- +1 814 865-0377 (phone)
- +1 814 863-7114 (fax)
- DennisLin@psu.edu

(Customer Satisfaction or your money back!)