

AI/大数据与 健康和医学的未来

演讲人：王大禹博士

跨界互联
数聚未来

第四届中国数据分析师行业峰会
CHINA DATA ANALYST SUMMIT

📍 北京 中国大饭店 2017.07

医生如何诊断疾病？

- 现在 vs 过去
 - Symptoms: 主观感受：头疼
 - Signs：客观事实：红斑，医学检测
 - 过去的sign：体检报告、医学检测
 - 过去的symptoms：靠回忆和病人的记录
- 生活
 - 生活习惯：吸烟
 - 工作性质：接触石棉、铅
- 家族
 - 家族遗传

医生与病人的信息不对称

- 医生：一般性的、统计意义的信息和知识
 - 几分钟到二十分钟诊断病人延续几天至几年的疾病
- 病人：对自己疾病的了解
 - 久病成医
 - 对历史的有意识、无意识隐瞒：House MD
- 信息时代病人对于诊断的参与的可能性
 - 准确、详尽的医疗信息
 - 丰富的医案
 - 数据处理技术
 - 医案的处理
 - 病人病史的处理



案例：维生素C每天应该吃多少？

- 40 mg/day : UK Food Standards Agency
- 45 mg/day : WHO
- 80 mg/day: the European Commission's Council on nutrition labeling
- 90 mg/day (males) and 75 mg/day (females): Health Canada 2007
- 60–95 mg per day: United States' National Academy of Sciences. The United States defined Tolerable Upper Intake Level for a 25-year-old male is 2,000 mg per day.
- 100 mg per day: Japan's National Institute of Health and Nutrition. The NIHN did not set a Tolerable Upper Intake Level.

案例：维生素C每天应该吃多少？

- 建议摄取量的简化
 - 吼猴吃多少？10-20倍
- 身体状况的需求：抑郁山羊主动多吃草
 - 压力大
 - 环境污染
 - 吸烟者
 - 偏食者
- 结论：发现维生素C也不意味着可以准确确定剂量
- 什么信息决定健康？决定诊断？决定治疗？决定治愈？

疾病的复杂性

社区民警高宝来 五年坚持接送学生 近期患肺癌病倒
孩子和家长很牵挂 警察爷爷 我们等你回来(组图)

2015-03-27 14:07:00 来源: 法制晚报(北京)

今年59岁的他当了35年警察，头发已白了大半。自确诊肺癌后，这是他第二次住院，马上要进行第二次化疗。老高的肺癌已是晚期，情况比较严重，肿瘤位置靠近心脏，手术风险大，医生们选择保守治疗。因肿瘤压到了声带，老高只能靠气嘘声说话，这些天也只能靠流食度日。

每天下午5点左右，老高都要经历一次剧烈咳嗽，持续一个小时候左右，让他一句完整的话都说不出来。

疾病的复杂性

- 徐灵胎医案

- 同里朱翁元亮，侨居郡城。岁初，其媳往郡拜贺其舅，舟过娄门，见城上蛇王庙，俗云烧香能免生疮肿，因往谒焉。归即狂言昏冒，舌动如蛇，称蛇王使二女仆一男仆来迎。延余诊视，以至宝丹一丸遣老岖灌之，病者言此系毒药，必不可服，含药喷姬，姬亦仆，不省人事，舌伸颈转，亦作蛇形。另易一人灌药讫，病者言一女使被烧死矣。凡鬼皆以朱砂为火也。次日煎药，内用鬼箭羽，病者又言一男使又被射死矣，鬼以鬼箭为矢也。从此渐安，调以消痰安神之品，月余而愈。此亦客许之类也，非金石及通灵之药，不能奏效。

基因组数据

- 存储：
 - CloudBurst, DistMap, SeqWare, Hydra
- 检错：
 - SAMQA, ART, CloudRS
- 数据分析
 - GATK, BlueSNP, Myrna, Eoulsan, SparkSeq
- 集成平台
 - CloVR

图像数据

- 存储和检索
 - PACS-DICOM
- 数据分享
- 数据分析
- 核心问题：
 1. 代替人工，看片快。e.g. 找问题-肠镜
 2. 接近事实：2D vs. 3D → 真实

临床信息数据

- 存储和检索
 - real-time是核心
e.g. EEG (脑电) : CloudWave 77G data, 5 sets in 1 min
 - 多模态的非结构化数据是难点：期待中文NLP的突破
 - 数据分享、数据安全
 - 数据分析 (预测)
 - E.g. PARAMO : 使用术语集(ICD, UMLS)
 - 机器学习算法 : LR , 随机森林 , $acc=77\%$, $recall=61\%$
- Zolfaghar K, et al. Big data solutions for predicting risk-of-readmission for congestive heart failure patients. In: 2013 IEEE International Conference on Big Data, PACS

健康信息

- 传染病监控
 - E.g. HIV prevalence analysis vs. tweet
- 人口健康管理
- 心理管理
 - E.g. NN预测心理疾病准确率82.35%
- 慢性病管理（中国的核心问题）
- 核心问题：生活、身体信息的高效、便捷获取

- Young SD, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. Prev Med. 2014;63(0):112–5.

- Dabek F, Caban JJ. Brain informatics and health. In: Guo Y, Friston K, Aldo F, Hill S, Peng H, eds. A Neural Network Based Model for Predicting Psychological Conditions. Cham: Springer International Publishing; 2015:252–61.

核心问题

- 自然语言处理
 - e.g. 症状识别
 - 术语和本体
- 信息检索
 - e.g. 时间识别
 - TREC
 - “右侧牙龈肿痛连及面颊2月”
 - “服清热泻火之剂10余天无效”

(1) 苔色分类器构建

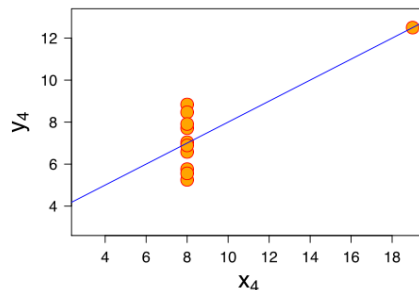
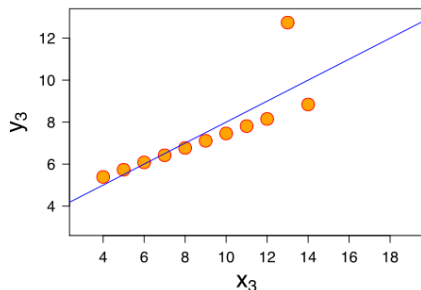
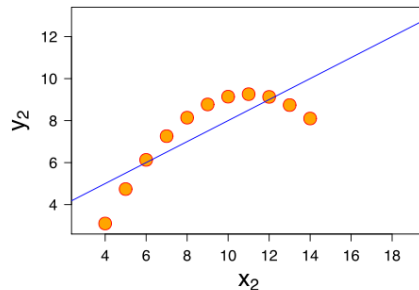
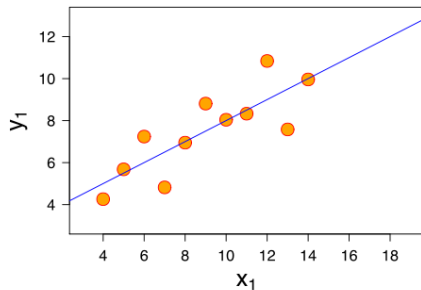
分类结果集合： $\langle \text{'白苔'}, \text{'黄苔'}, \text{'灰苔'}, \text{'黑苔'}, \text{'绿苔'}, \text{'锈酱苔'} \rangle - f(x)$

(2) 苔质分类器构建

分类结果集合： $\langle \text{'光剥舌'}, \text{'地图舌'}, \text{'中剥苔'}, \text{'前剥苔'}, \text{'花苔'}, \text{'舌面如镜'}, \text{'霉腐苔'}, \text{'舌干'}, \text{'脓腐苔'}, \text{'苔垢'}, \text{'舌苔腐垢'}, \text{'苔润'}, \text{'苔滑'}, \text{'苔燥'}, \text{'粉苔'}, \text{'粉苔'} \rangle$

数据的局限

- 描述性统计学的无力
- 测量的局限
- 人的自由意志



Property

Mean of x
Sample variance of x
Mean of y
Sample variance of y
Correlation between x and y
Linear regression line

Value

9 (exact)
11 (exact)
7.50
4.122 or 4.127
0.816
 $y = 3.00 + 0.500x$

医院面临的问题和其他领域的重叠

- 医院排班问题：每天放多少医生值班？v.s. 零售业
- 电子病历统一和普及（Electronic Health Records, EHRs）：
A [McKinsey report on big data in healthcare](#) states that “The integrated system has improved outcomes in cardiovascular disease and achieved an estimated \$1 billion in savings from reduced office visits and lab tests.”
- 实时提醒（real-time alerting）
- 防止药物滥用（Prevent Opioid Abuse）- 治疗方案规范化，尤其是肿瘤
- 战略计划与流行病
- 远程医疗-专家分身-医疗资源不平衡的解决方案
- 减少就诊和急诊-高质量的在线医疗常识信息

医学界、营养学界和人类面临的问题

- 未来人类的健康管理：主动管理医生和信息
- 医生的营养学修养：治未病和营养学方案
- 药理学的进展：MOA的突破
- 最有效 vs 最适合
 - 过度医疗
 - 副作用和统计学显著
- 名医的复制 vs 医疗记录检索
- 心理作用与安慰剂：主观和客观的治愈
- 对人和自然的不停歇的探索

王大禹

- Email : sywdy@qq.com
- 微信： 77839416
- @中国大禹



CDA 数据分析师
www.cda.cn

THANKS

跨界互联 数聚未来

第四届中国数据分析师行业峰会
CHINA DATA ANALYST SUMMIT