

化智能问题为数据问题

吴军

人工智能六十年 (1956-2016)



机器产生智能的方式是否和人一样



机器智能中的空气动力学原理



数据量的作用

从中文到英语的翻译

封闭测试集

Google	35.31%
南加州大学	30.73%
马里兰大学	30.00%
德国亚琛工学院	29.37%
约翰·霍普金斯 - 剑桥大学	28.27%
IBM	25.71%
.....	
SYSTRAN 公司	14.71%

开放测试集

Google	35.16%
中国科学院	12.93%
哈工大	7.97%

从阿拉伯语到英语的翻译

封闭测试集

Google	51.31%
南加州大学	46.57%
IBM 华生实验室	46.46%
马里兰大学	44.97%
约翰·霍普金斯大学	43.48%
.....	
SYSTRAN 公司	10.79%

开放测试集

Google	51.37%
SAKHR 公司	34.03%
美军 ARL 研究所	22.57%

计算机回答问题

- 图灵测试：让计算机回答问题
 - 语音识别
 - 语法和语音分析
 - 机器翻译 (Alignment)
 - 写作
- 但是，.....，复杂的问题 (Why, How) 还不能回答 (2012)

天为什么是蓝色的 (2014)



why is the sky blue



Web

Videos

Images

Shopping

Books

More ▾

Search tools

About 309,000,000 results (0.42 seconds)

A clear cloudless day-time **sky** is **blue** because molecules in the air scatter **blue** light from the sun more than they scatter red light. When we look towards the sun at sunset, we see red and orange colours because the **blue** light has been scattered out and away from the line of sight.

[Why is the sky Blue?](#)

math.ucr.edu/.../BlueSky/blue_sky.html

University of California, Riverside ▾

Feedback

夏天为什么比冬天热



why is summer hotter than winter



Web

News

Videos

Images

Shopping

More ▾

Search tools

About 1,240,000 results (0.48 seconds)

During the **summer**, the sun's rays hit the Earth at a steep angle. The light does not spread out as much, thus increasing the amount of energy hitting any given spot. Also, the long daylight hours allow the Earth plenty of time to reach warm temperatures.

[Why is it hot in summer and cold in winter? \(Everyday ...](#)

www.loc.gov/rr/scitech/mysteries/seasons.html Library of Congress ▾

Feedback

将Q&A问题变成数据问题

- 一系列通信问题 (Fred Jelinek)
 - 语音识别: $P(S | A)$
 - 机器翻译: $P(E | C)$
- 当数据量大了以后
 - 写作: $P(S | S_1, S_2, S_3, \dots, S_N)$

机器回答问题

- 传统的做法：阅读题考试
 - 给一篇文章，题一个问题，寻找答案
- 换个角度想问题
 - 给一篇文章，看看能回答什么问题
- 大数据的完备性

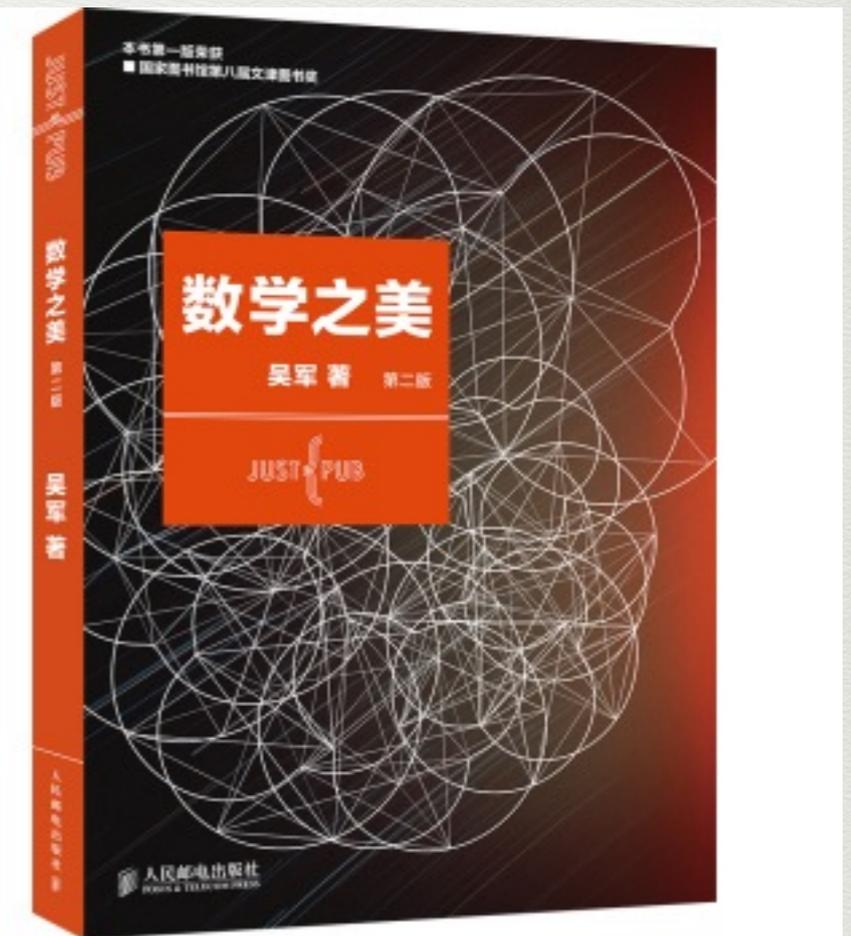
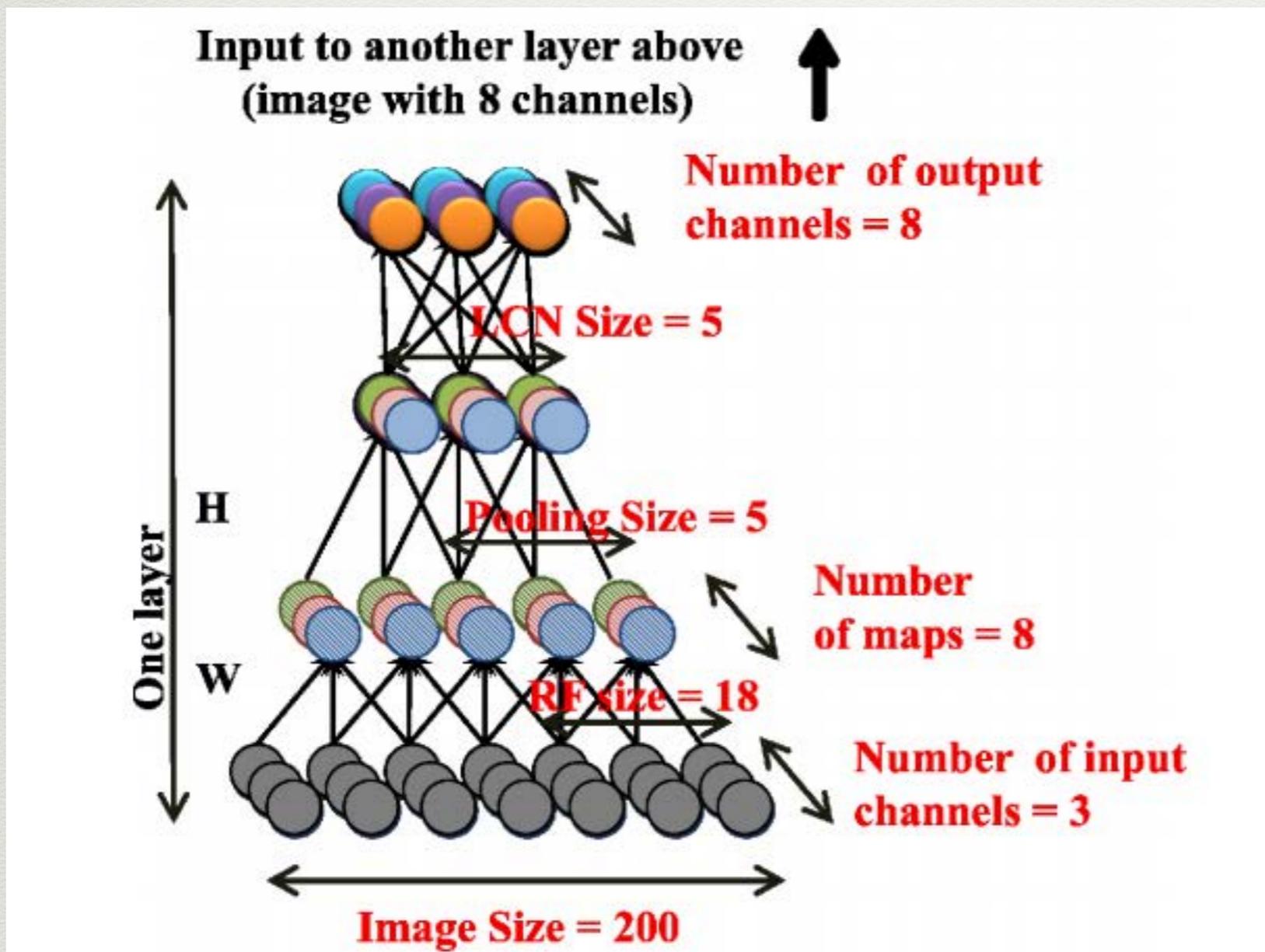
将Q&A问题变成机器翻译问题

- 机器翻译
 - $P(E|C) = P(C|E) \cdot P(E) / P(C)$
- Q&A
 - $P(A|Q) = P(Q|A) \cdot P(A) / P(Q)$
 - $P(Q|A)$: Alignment
 - $P(A)$: 写作的模型

步骤

- Crawl all web pages (~10B, 40% English)
- Parse all sentences
- Find top 50 documents per question
- Align sentences to questions
- Find top candidate sentences
- Recompose sentences and paragraphs

Deep Learning: Google Brain



自动驾驶的问题也是一个数据问题

- 2004 DARPA拉力赛 8英里
- 2010 Google 14万英里无交通事故
- 2016 200万+英里，两次主动交通事故

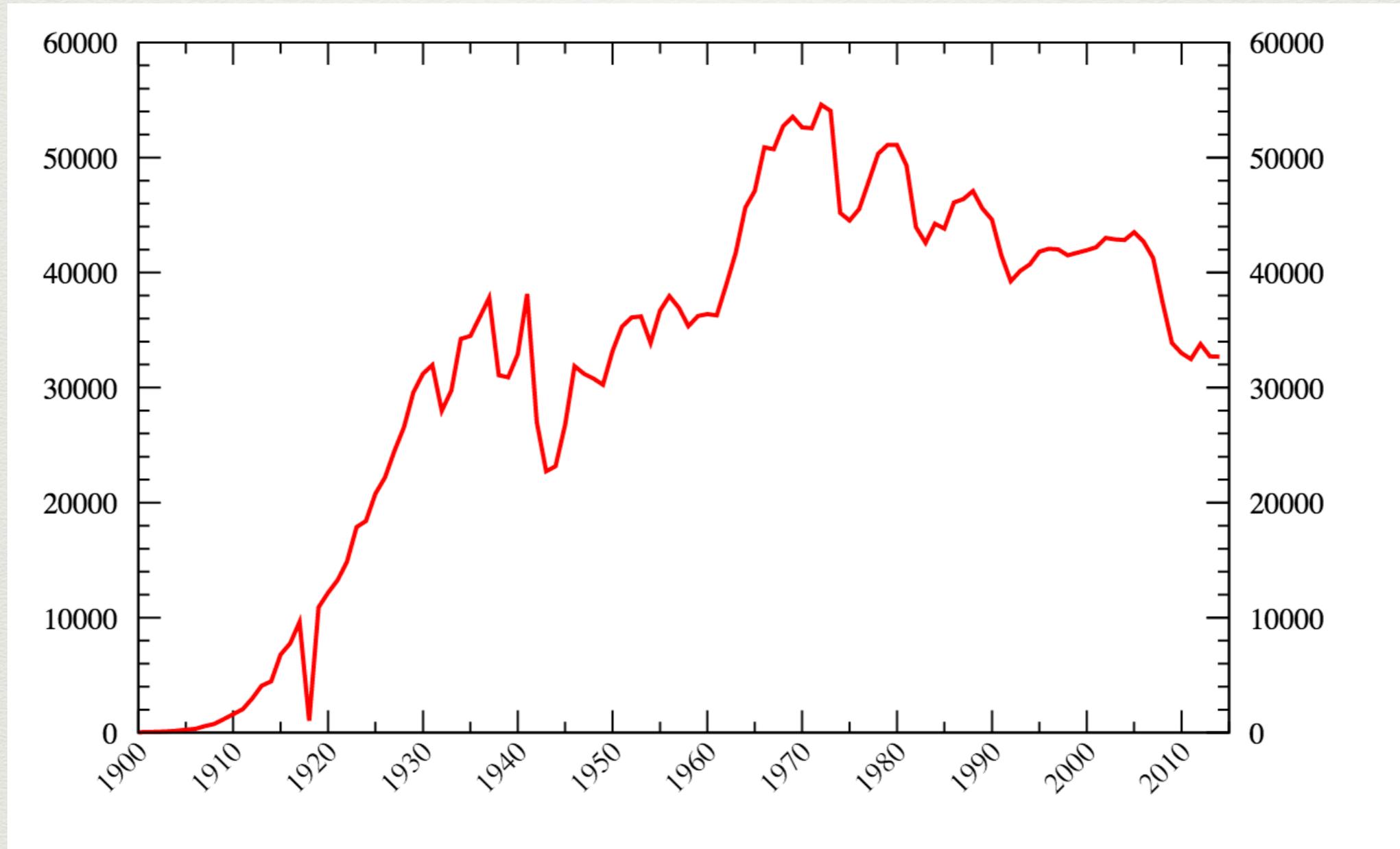


最好的时代

- 未来交通
 - 智能交通（美国大城市，省20分钟）
 - 无人驾驶



今天美国每年汽车交通事故死亡3万人



未来的产业离不开机器智能

- NBA金州勇士队（Golden State）的奇迹

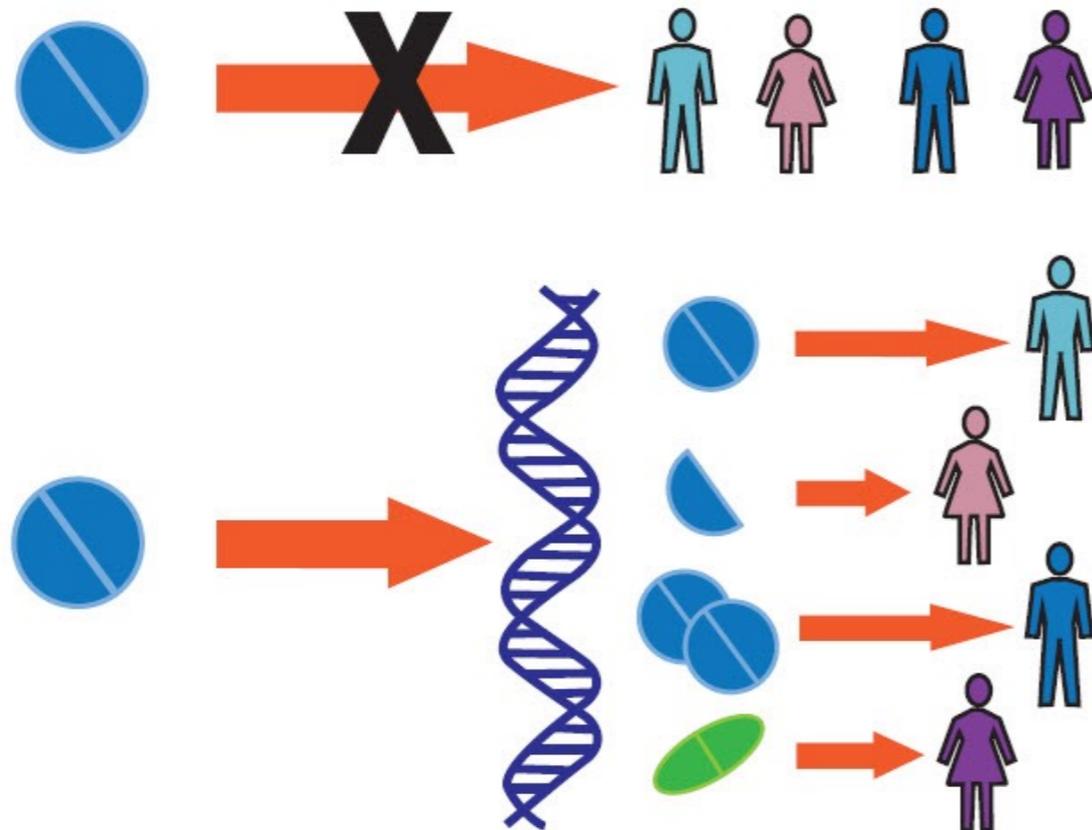


治愈癌症

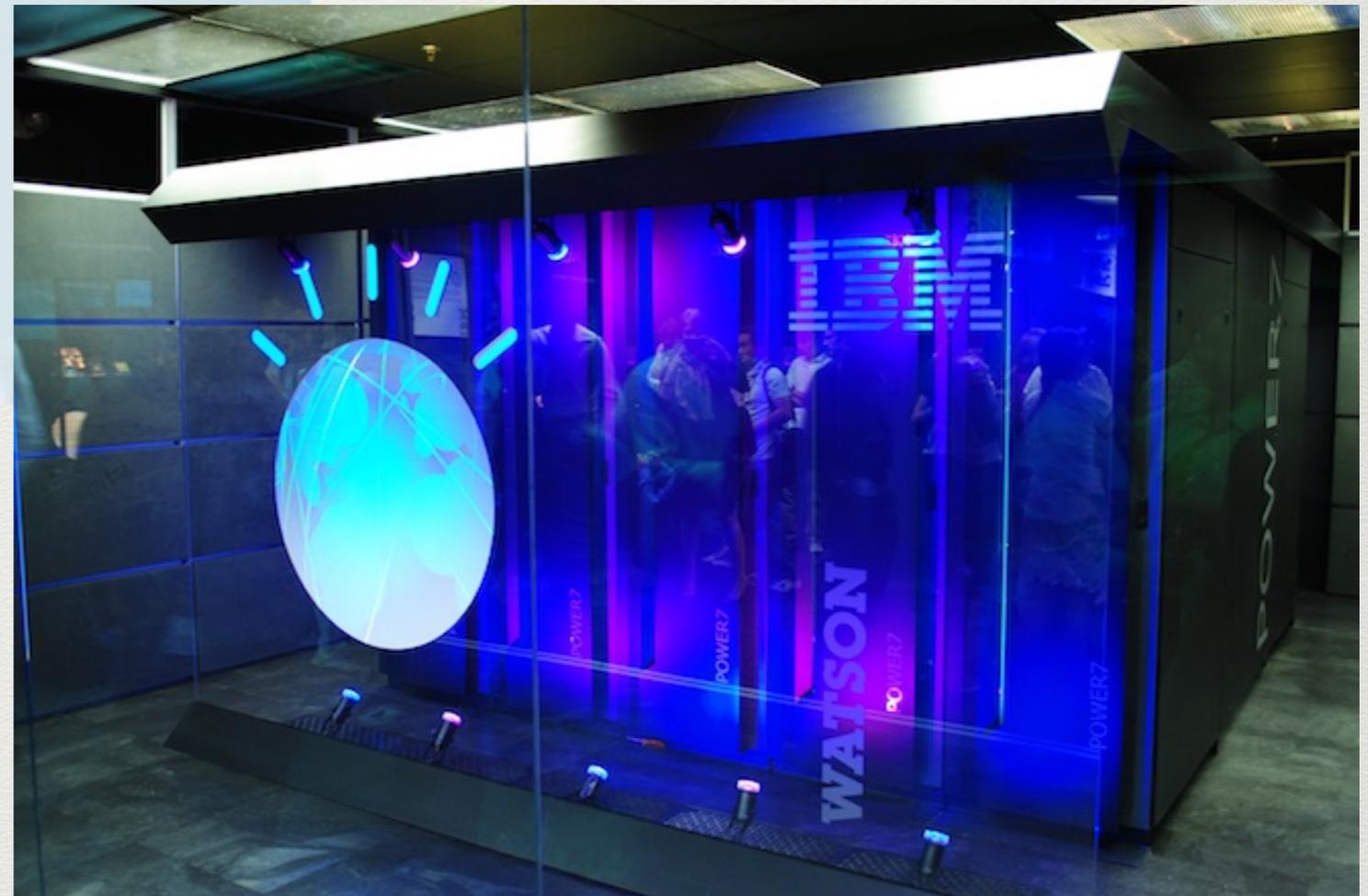
- 为什么癌症难以医治
 - 和人有关
 - 不断变化
- 为每一个人研制一种专卖的抗癌药
 - 10亿美元

个性化制药

- 癌症：几百种癌症，5000种基因变异
- 能否通过几百种药全面覆盖所有的情况？

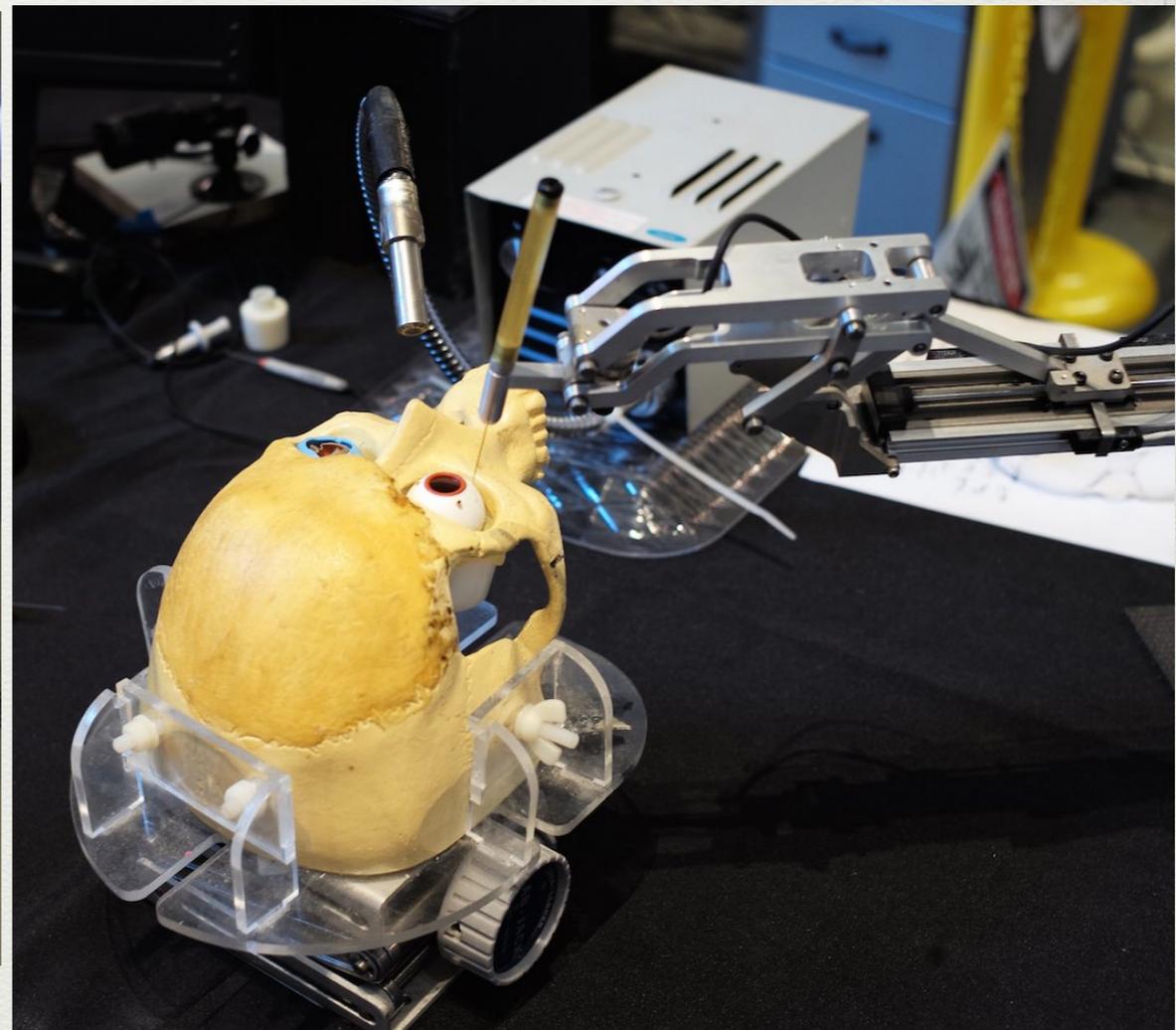


未来医疗



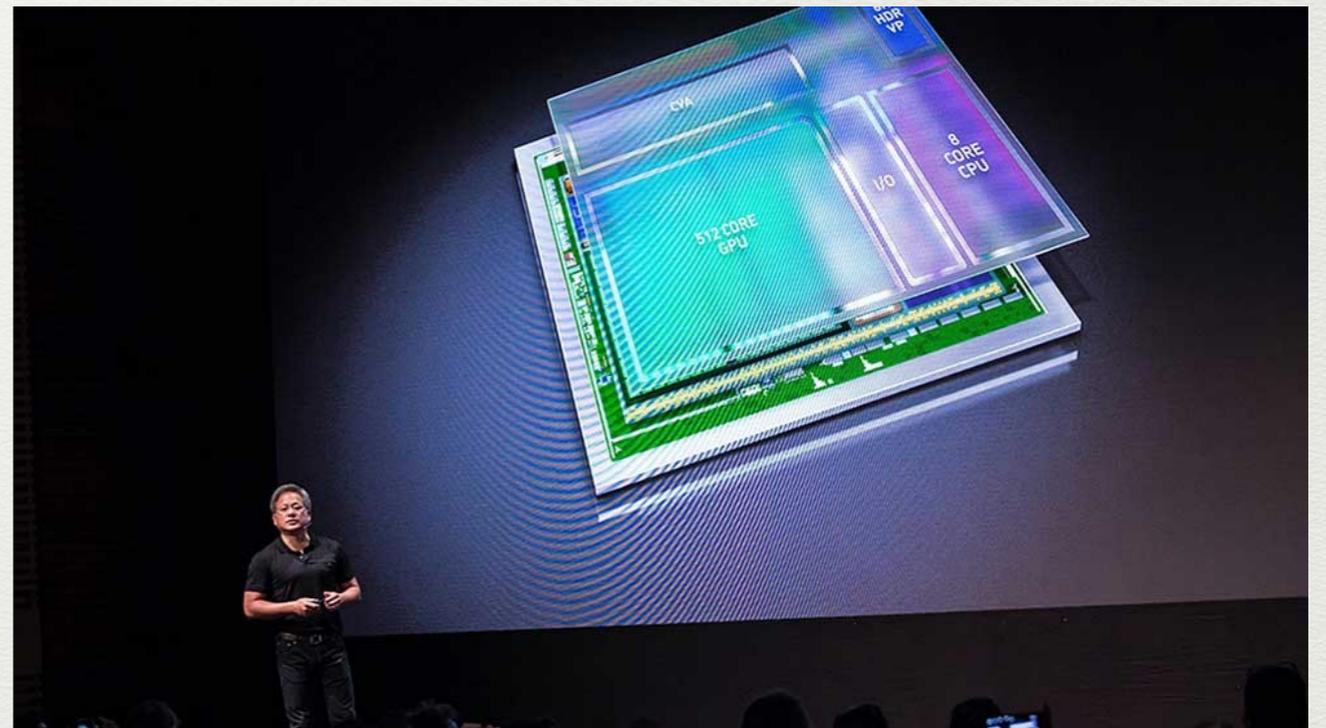
手术机器人

- 精确度要比人高得多



专用芯片

- 摩尔定律遇到瓶颈，但是真的机器智能的专用芯片依然有机会
- 嵌入机器学习算法特定功能
 - 计算机视觉
 - Google Brain芯片版
- 计算存储一体化
 - 手机大小的数据中心



小结

- 智能问题能够转化成数据问题
- 机器智能 = 摩尔定律 + 大数据 + 深度学习
- 未来的20-50年，是智能时代