Thought leaders in data science and analytics:

Data Science

James G. Shanahan[1]

[1]***Independent Consultant***

EMAIL: James_DOT_Shanahan_AT_gmail_DOT_com

I 296A UC Berkeley

Lecture 1, Wednesday January 18, 2012

Brief Bio James G. Shanahan

- **20+ years in the field AI and information management**
  - Principal and Founder, Boutique Data Consultancy
    - Clients include: Adobe, Digg, SearchMe, AT&T, Ancestry, OfferPal,
  - Teach at University of California Santa Cruz (UCSC), ISM 209, 250, 251
  - Previously
    - Chief Scientist, Turn Inc. (A CPX ad network, DSP)
    - Principal Scientist, Clairvoyance Corp (CMU spinoff; sister lab to JRC)
    - Research Scientist, Xerox Research
    - Research Engineer, Mitsubishi Group
    - PhD in machine learning (1998), University of Bristol, UK; B.Sc. Comp. Scienc (1989), Uni. of Limerick, Ireland
- **Now: Machine Learning Consultant (San Francisco)**
  - IF *(you have large **data problems** and need a consultant)*
    THEN *{email me at James.Shanahan_AT_gmail.com}*
  - Where **problems** $\in$ *{web search, online advertising, machine learning, ranking, user modeling, statistics, social networks, operations research}*

Disclaimer

- The Authors retains all rights, including copyrights and distribution rights.
- No publication or further distribution in full or in part permitted without explicit written permission from the author
- Living vicariously!

Lecture Outline

- Course Background
- Advertising 101 and Digital advertising
- Predicting Click Through Rate
- Homework

This course is timely!

- **I 296 A core**
  - Look at how to leverage data modeling, machine learning, statistics, data mining for modern day problems?

- **…. with applications in digital advertising and marketing, healthcare, telecommunications, finance…**
- **Timely:**
  - Growing flood of online data, many budding industries (e.g., digital advertising, digital healthcare)
  - Computational power is available (PC, Cloud computing, Hadoop)
  - Progress in algorithms and theory and applications

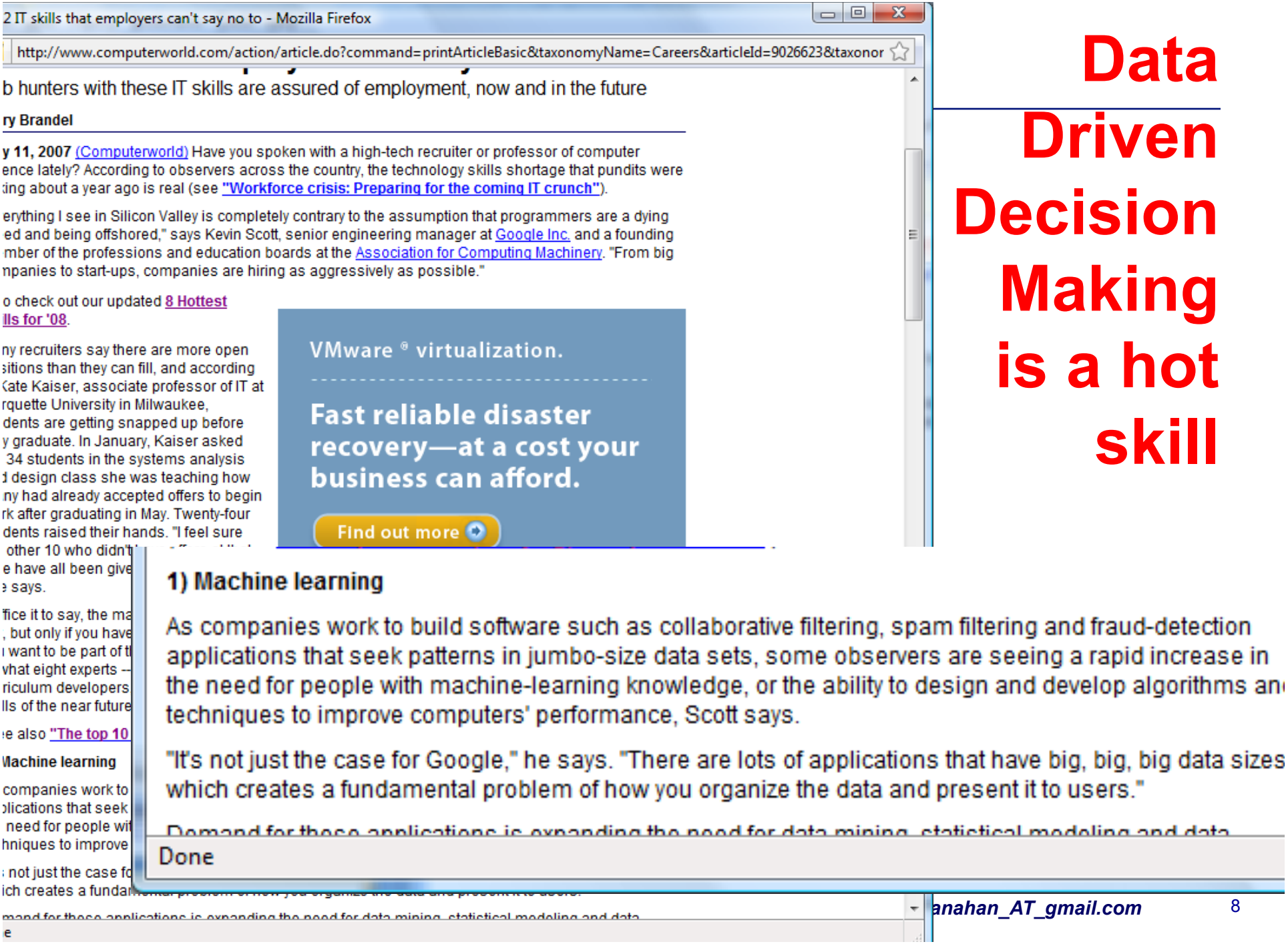
Summaries → Decisions

- **The old days were about asking, 'What is the biggest, smallest, and average?' " says Michael Olson, CEO of startup Cloudera. "Today it's, 'What do you like? Who do you know?' It's answering these complex questions."**

- **In the old days:**
  - A retailer such as Macy's (M) that once pored over last season's sales information could shift to looking instantly at how an e-mail coupon for women's shoes played out in different regions.

Data Driven Decision Making is hot skill

Data Driven Decision Making is a hot skill

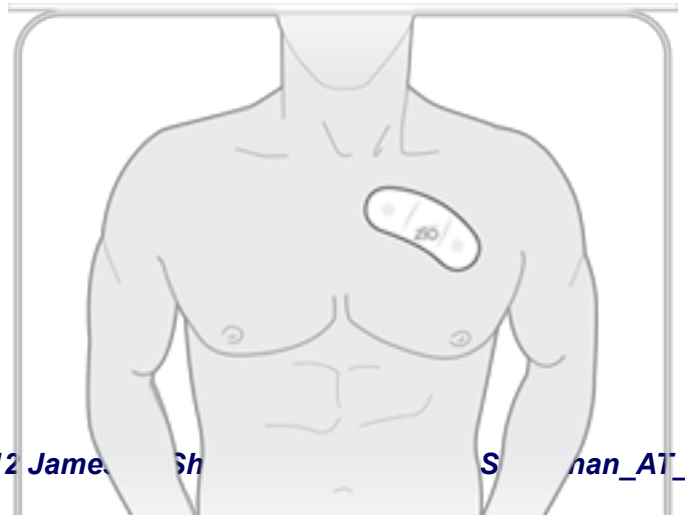


data analytics has even gotten hip!

- **It's not going too far to say that data analytics has even gotten hip.**
  - The San Francisco offices of startup Splunk have all the of-the-moment accoutrements you'd find at Twitter or Zynga.
- **The engineers work in what amounts to a *giant living room with pinball machines, foosball tables, and Hello Kitty-themed cubes*.**
- **Weekday parties often break out—during a recent visit, it was Mexican fiesta.**
  - Employees were wearing sombreros and fake moustaches while a dude near the tequila bar played the bongos.

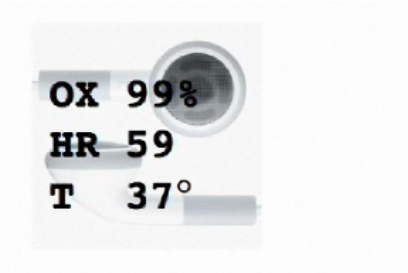
Irhythm: detect cardiac problems

- **IRhythm makes a type of oversize, plastic band-aid called the Zio Patch that helps doctors detect cardiac problems before they become fatal.**
  1. Patients affix the Zio Patch to their chests for two weeks to measure their heart activity.
  2. The patients then mail the devices back to IRhythm's offices, where a technician feeds the information into Amazon's cloud computing service.
- **Patients typically wear rivals' much chunkier devices for just a couple of days and remove them when they sleep or shower—which happen to be when heart abnormalities often manifest. The upside of the waterproof Zio Patch is the length of time that people wear it—but 14 days is a whole lot of data.**



Sensors + Services => Privacy Problem

- **Personal devices (with GPS' and accelerometers)**
  - Earphones; Nike+ (measures and records the distance and pace of a walk or run); asthma inhaler with built-in GPS tracking



- **Personal/social services**
  - Mint, Twitter, diets, health, exercise, FaceBook

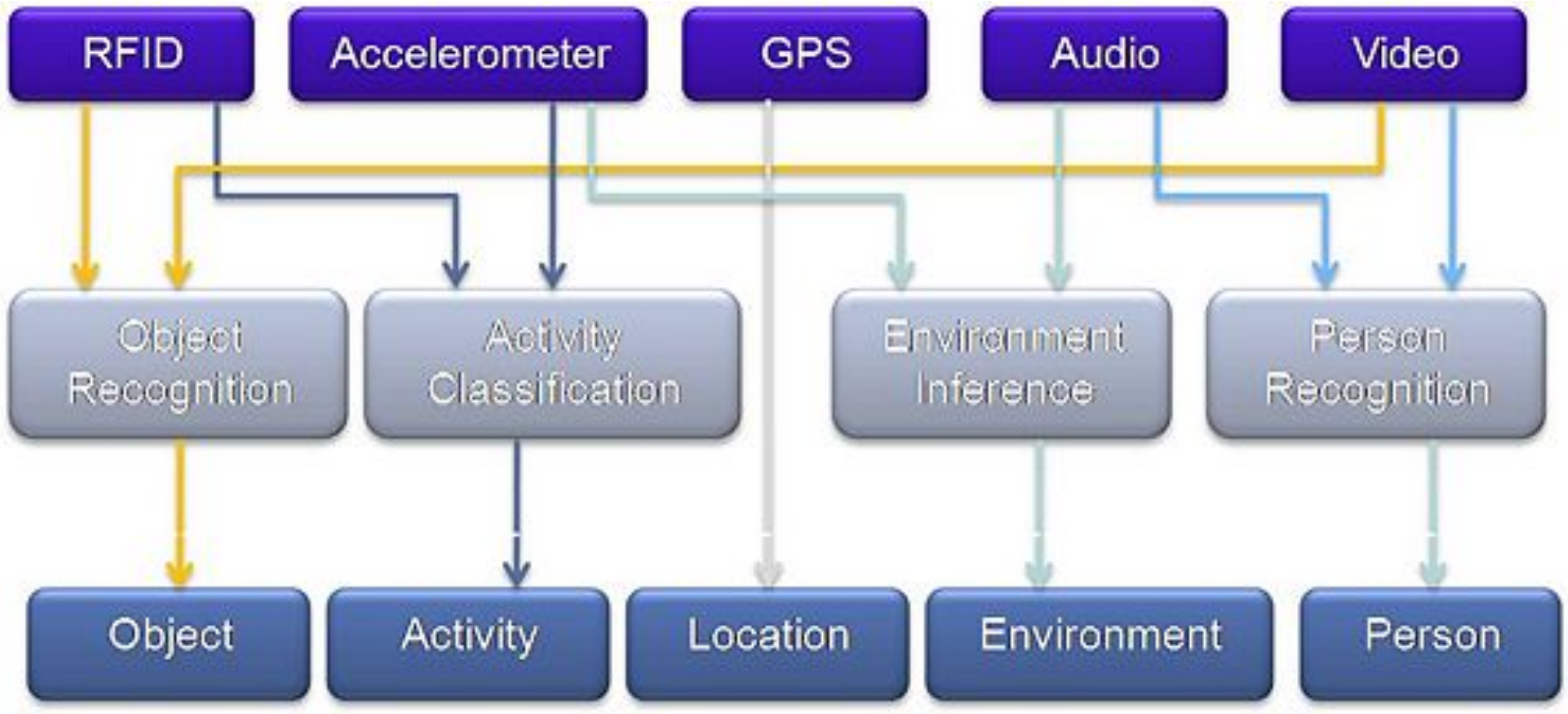
- **These data streams create a huge privacy problem**

Always connected at the extreme→ Lifelogging

Records events using multiple wearable sensors
Provides access to these data at multiple levels of granularity and abstraction, using access mechanism based on the episodic memory of human beings.

http://www.imrc.kist.re.kr/wiki/LifeLog

Backend Technology



3rdi Art Project

- **A New York University arts, Professor Bilal**
- **A surgically-implanted camera (12/15/2010)**
  - 3rdi Project, has already generated international media attention and anticipation. On Dec. 15 images from the "third eye" in the back of Bilal's head -- a surgically-impanted camera -- will be unveiled in Doha, Qatar as part of the Told/Untold/Retold exhibition that inaugurates the new Arab Museum of Modern Art near Education City, Doha's intellectual hub.
- **Transmits one image per minute to a website (www.3rdi.me), displayed a Doha gallery**
  - with the inaugural image
    designed room in the Do
    of the museum's new pe
    making, including more t
    from North Africa to the
    day.

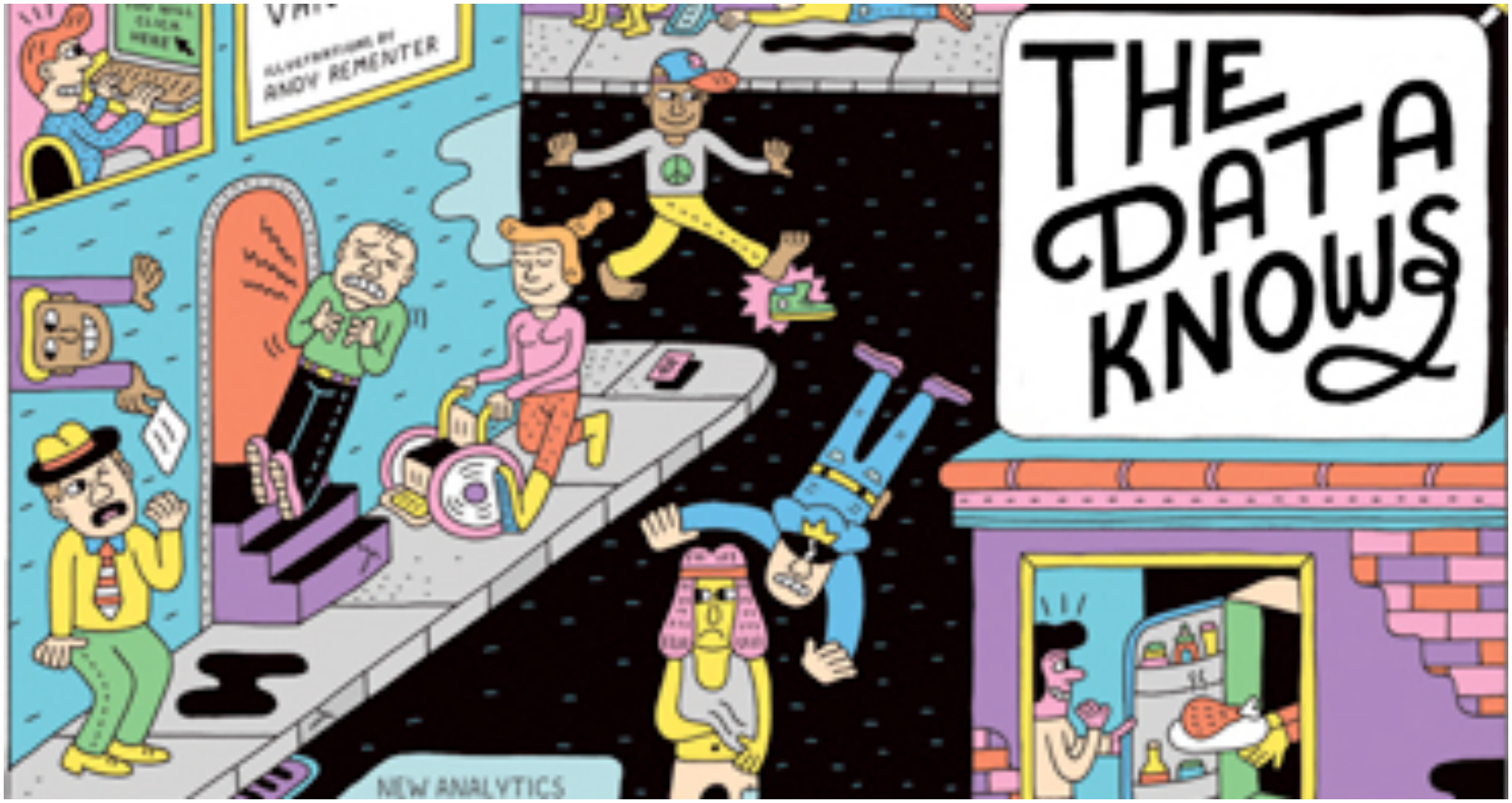
4 Screens: Mobile, Computer, TV, Theatre

- **Smartphones 50% share in mid2011 (US)**
- **Tablet computers**
  - Large Format Benefit
  - Enhanced mobile apps
  - Total media tablets device market
  - 28MM in 2011 (ABI, 2010; Barclays Capital, 2010)

- **IPTV**
  - Play IPTV digital content originating from the iTunes Store, Netflix, YouTube, Flickr, MobileMe or any Mac OS X or Windows computer running iTunes onto an enhanced-definition or high-definition widescreen television
  - Still early days but

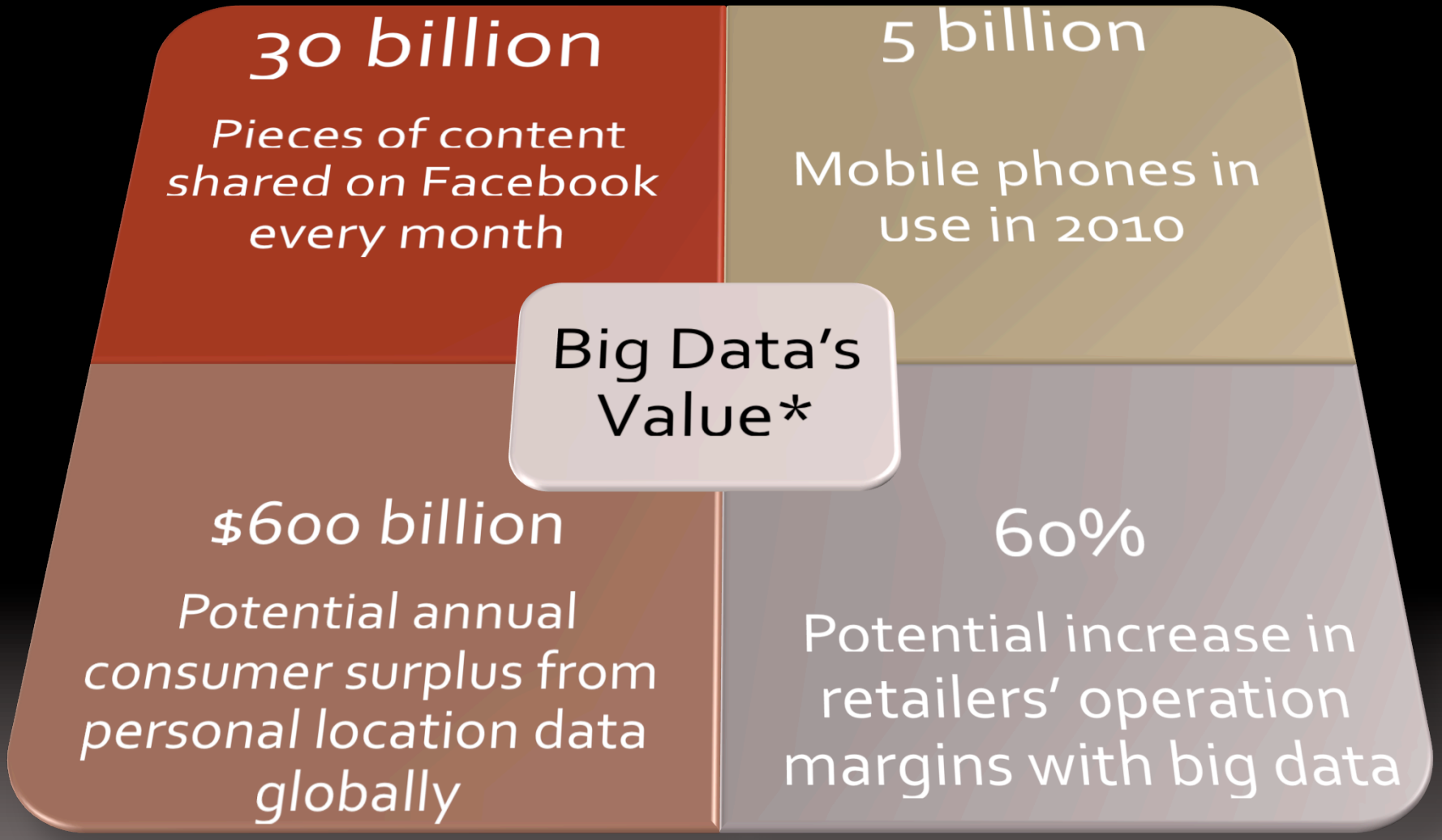
- **Theatre**

The Data Knows!



http://www.businessweek.com/magazine/data-analytics-crunching-the-future-09082011.html

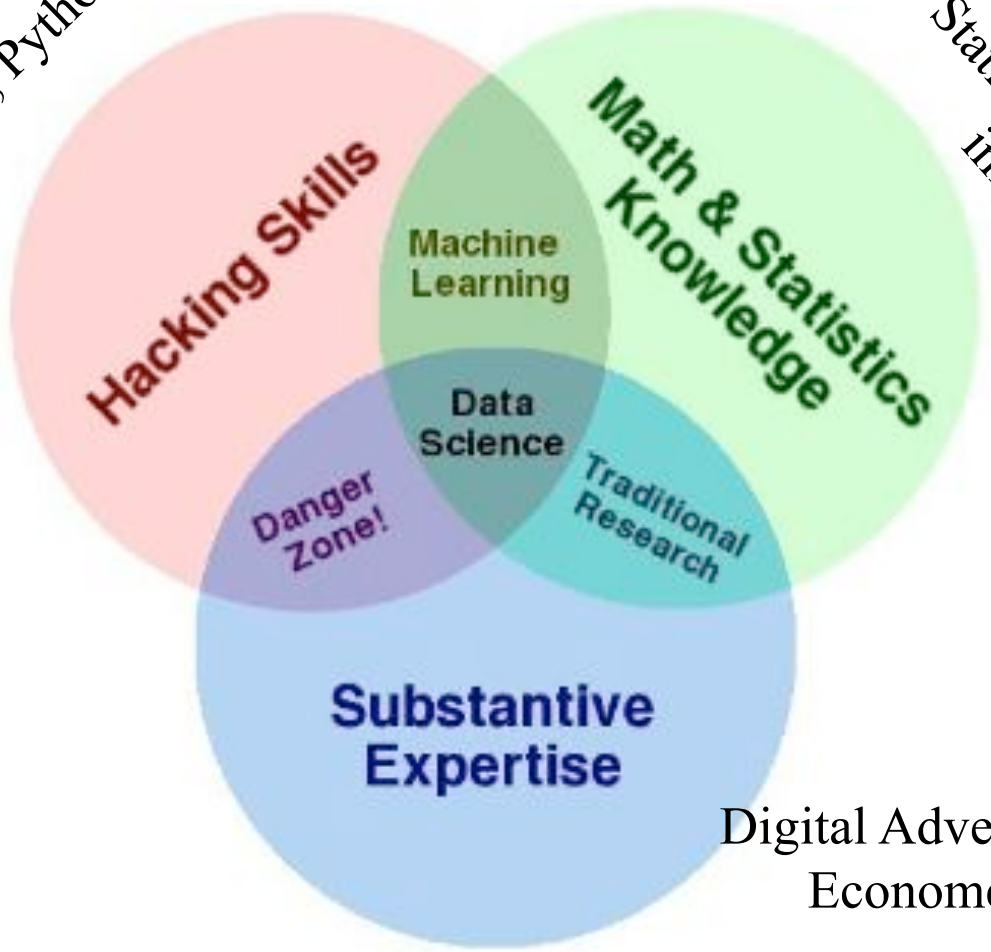
BIG DATA



* McKinsey Global Institute 2011

Wanted: Data Scientists

Hadoop, R, Python



Decision Trees, Gradient Descent, Statistics, Optimization theory, information science

Digital Advertising and Marketing, Econometrics, Web Search

150,000 Data Scientists needed in US

[McKinsey Report on Big Data]

More Data versus Rocket Science

Some simple math using a mountain of data can get you 80% of the way!

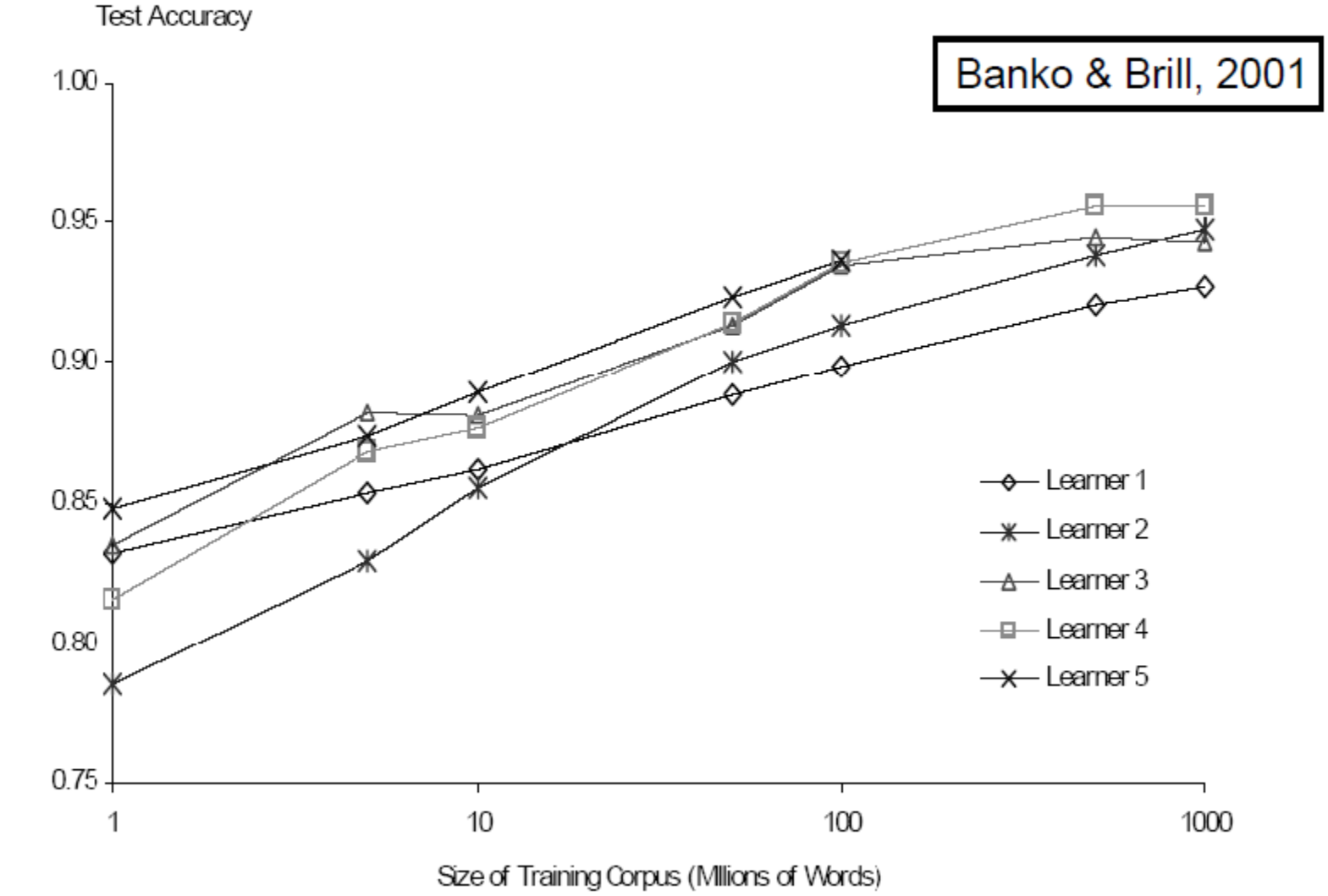


Figure 2. Learning Curves for Confusable Disambiguation

- **End of Lecture 1**