
Thought leaders in data science and analytics: Data Science

James G. Shanahan¹

¹Independent 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. Science (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.

2 IT skills that employers can't say no to - Mozilla Firefox

http://www.computerworld.com/action/article.do?command=printArticleBasic&taxonomyName=Careers&articleId=9026623&taxonor

b hunters with these IT skills are assured of employment, now and in the future

ry Brandel

y 11, 2007 (Computerworld) Have you spoken with a high-tech recruiter or professor of computer ence lately? According to observers across the country, the technology skills shortage that pundits were ing about a year ago is real (see "[Workforce crisis: Preparing for the coming IT crunch](#)").

erything I see in Silicon Valley is completely contrary to the assumption that programmers are a dying ed and being offshored," says Kevin Scott, senior engineering manager at [Google Inc.](#) and a founding mber of the professions and education boards at the [Association for Computing Machinery](#). "From big npanies to start-ups, companies are hiring as aggressively as possible."

o check out our updated [8 Hottest lls for '08](#).

ny recruiters say there are more open sitions than they can fill, and according Gate Kaiser, associate professor of IT at rquette University in Milwaukee, dents are getting snapped up before y graduate. In January, Kaiser asked 34 students in the systems analysis 1 design class she was teaching how ny had already accepted offers to begin rk after graduating in May. Twenty-four dents raised their hands. "I feel sure other 10 who didn't have offers at that e have all been given an offer by now," e says.

vice it to say, the market for IT talent is , but only if you have the right skills. If i want to be part of the wave, take a look what eight experts -- including recruiters, riculum developers, computer science professors and other industry observers -- say are the hottest lls of the near future.

re also "[The top 10 dead \(or dying\) computer skills](#)".)

Machine learning

companies work to build software such as collaborative filtering, spam filtering and fraud-detection lications that seek patterns in jumbo-size data sets, some observers are seeing a rapid increase in need for people with machine-learning knowledge, or the ability to design and develop algorithms and hniques to improve computers' performance, Scott says.

: not just the case for Google," he says. "There are lots of applications that have big, big, big data sizes, ich creates a fundamental problem of how you organize the data and present it to users."

mand for these applications is expanding the need for data mining, statistical modeling and data

e

VMware® virtualization.

Fast reliable disaster recovery—at a cost your business can afford.

Find out more



Data Driven Decision Making is hot skill

2 IT skills that employers can't say no to - Mozilla Firefox

http://www.computerworld.com/action/article.do?command=printArticleBasic&taxonomyName=Careers&articleId=9026623&taxonor

b hunters with these IT skills are assured of employment, now and in the future

ry Brandel

y 11, 2007 (Computerworld) Have you spoken with a high-tech recruiter or professor of computer ence lately? According to observers across the country, the technology skills shortage that pundits were ing about a year ago is real (see "[Workforce crisis: Preparing for the coming IT crunch](#)").

erthing I see in Silicon Valley is completely contrary to the assumption that programmers are a dying ed and being offshored," says Kevin Scott, senior engineering manager at [Google Inc.](#) and a founding mber of the professions and education boards at the [Association for Computing Machinery](#). "From big npanies to start-ups, companies are hiring as aggressively as possible."

o check out our updated [8 Hottest lls for '08](#).

ny recruiters say there are more open sitions than they can fill, and according Gate Kaiser, associate professor of IT at rquette University in Milwaukee, dents are getting snapped up before y graduate. In January, Kaiser asked 34 students in the systems analysis 1 design class she was teaching how ny had already accepted offers to begin rk after graduating in May. Twenty-four dents raised their hands. "I feel sure other 10 who didn't e have all been give e says.

nce it to say, the ma , but only if you have i want to be part of t what eight experts -- riculum developers lls of the near future e also "[The top 10 Machine learning](#)

companies work to olications that seek need for people with hniques to improve : not just the case fo ich creates a fundam

mand for these applications is expanding the need for data mining, statistical modeling and data

e

VMware® virtualization.

Fast reliable disaster recovery—at a cost your business can afford.

Find out more

1) Machine learning

As companies work to build software such as collaborative filtering, spam filtering and fraud-detection applications that seek patterns in jumbo-size data sets, some observers are seeing a rapid increase in the need for people with machine-learning knowledge, or the ability to design and develop algorithms and techniques to improve computers' performance, Scott says.

"It's not just the case for Google," he says. "There are lots of applications that have big, big, big data sizes which creates a fundamental problem of how you organize the data and present it to users."

Demand for these applications is expanding the need for data mining, statistical modeling and data

Done

anahan_AT_gmail.com

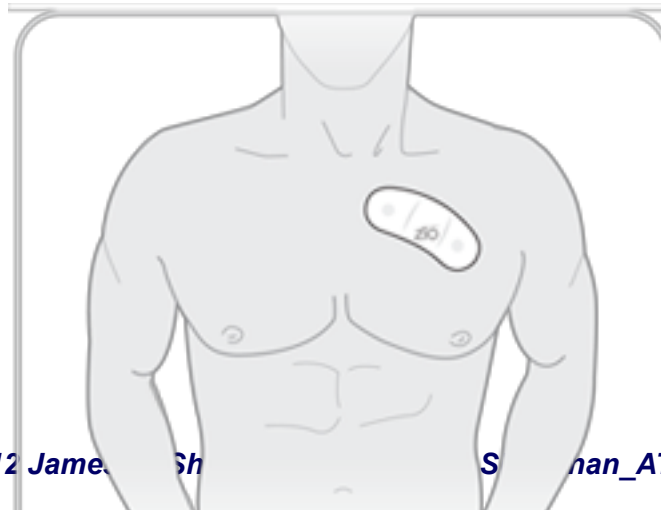
Data Driven Decision Making is a hot skill

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.

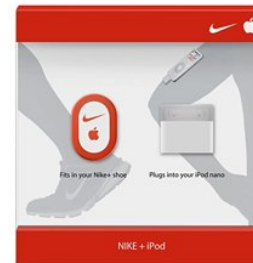
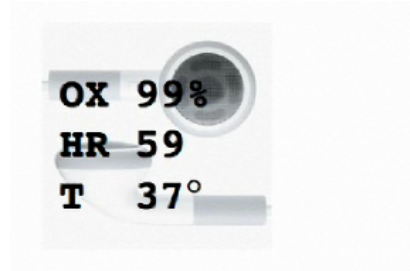
Irhythm: detect cardiac problems

- **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**

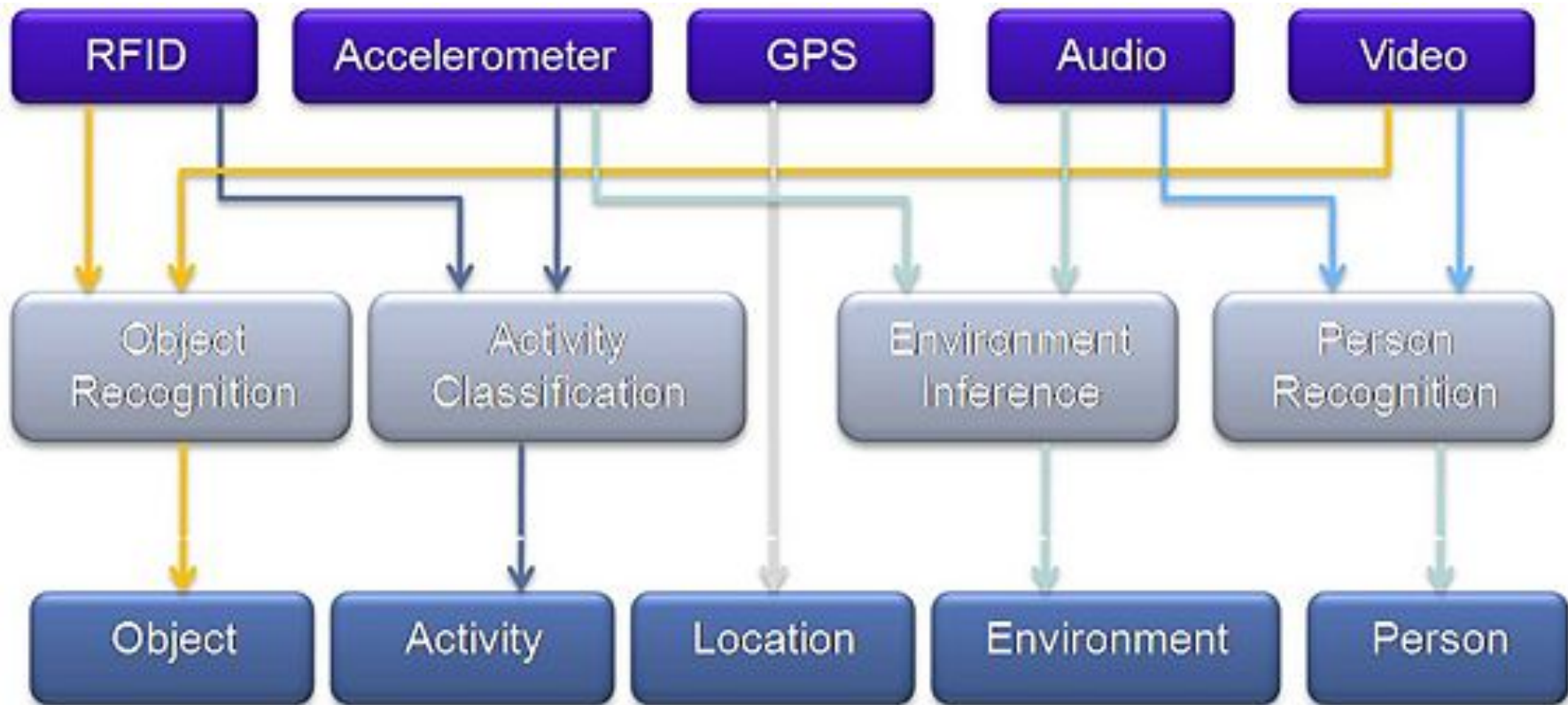
Always connected at the extreme → Lifelogging



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



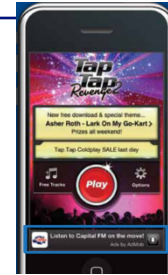
imscshankar

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 images
- designed room in the Doha
of the museum's new per
making, including more th
from North Africa to the C
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

30 billion

Pieces of content shared on Facebook every month

5 billion

Mobile phones in use in 2010

Big Data's Value*

\$600 billion

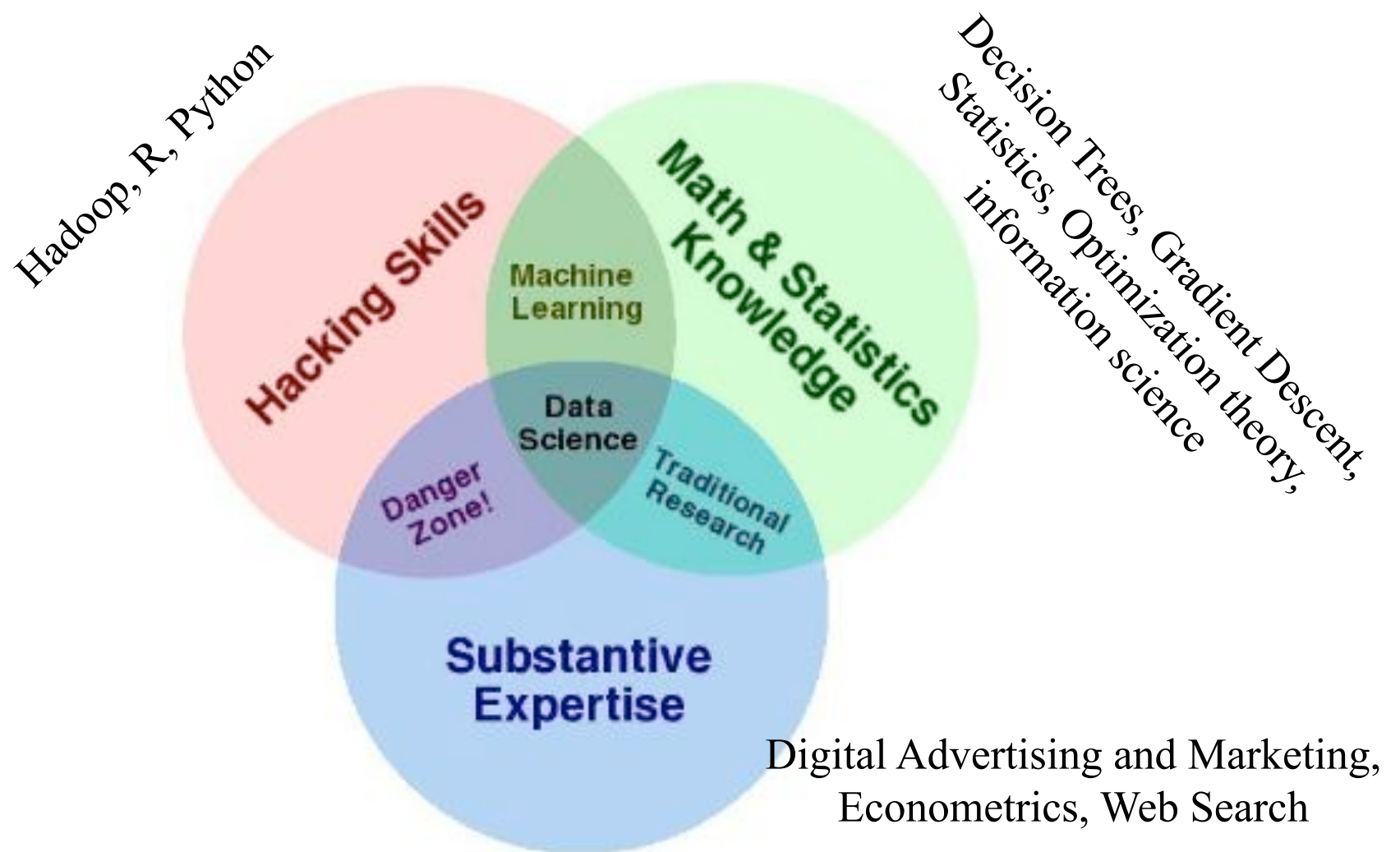
Potential annual consumer surplus from personal location data globally

60%

Potential increase in retailers' operation margins with big data

* McKinsey Global Institute 2011

Wanted: Data Scientists



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!

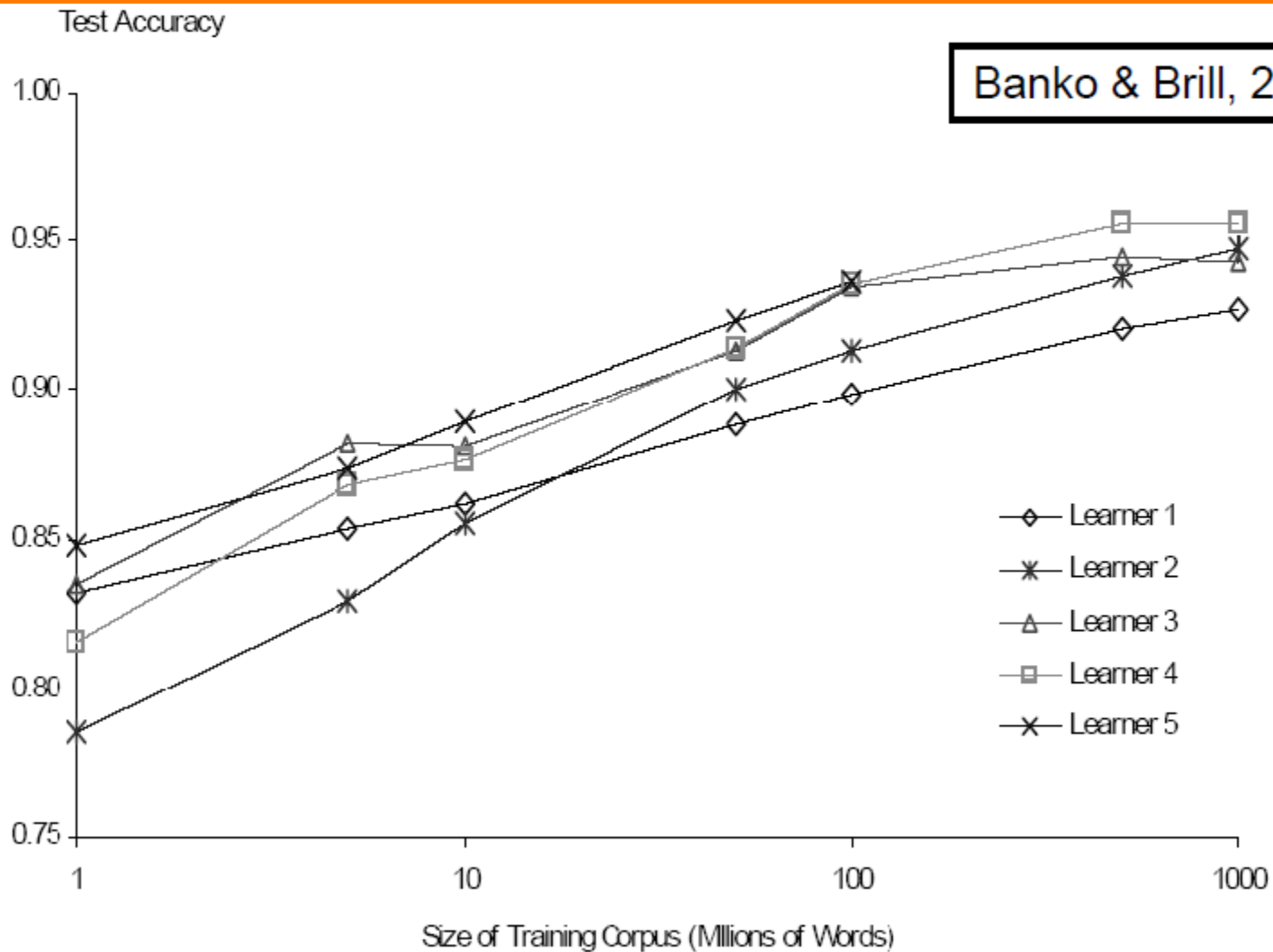


Figure 2. Learning Curves for Confusable Disambiguation

-
- **End of Lecture 1**