

# Lecture #1: Introduction and Overview\*

\* Slides partly based on Jiawei Han, and Jeff Ullman

# Data Mining: What?

- Automatic discovery of knowledge from data

# Data Mining: Why?

**\$600** to buy a disk drive that can store all of the world's music

**5 billion** mobile phones in use in 2010

**30 billion** pieces of content shared on Facebook every month

**40%** projected growth in global data generated per year vs.

**5%** growth in global IT spending

**\$5 million vs. \$400**

Price of the fastest supercomputer in 1975<sup>1</sup> and an iPhone 4 with equal performance

**235** terabytes data collected by the US Library of Congress by April 2011


**15 out of 17** sectors in the United States have more data stored per company than the US Library of Congress

# Data Mining: Why?

- Data contains value and knowledge



# Data Mining: How?

- To extract knowledge, data needs to be
  - ▲ stored
  - ▲ managed
  - ▲ analyzed  focus of this class

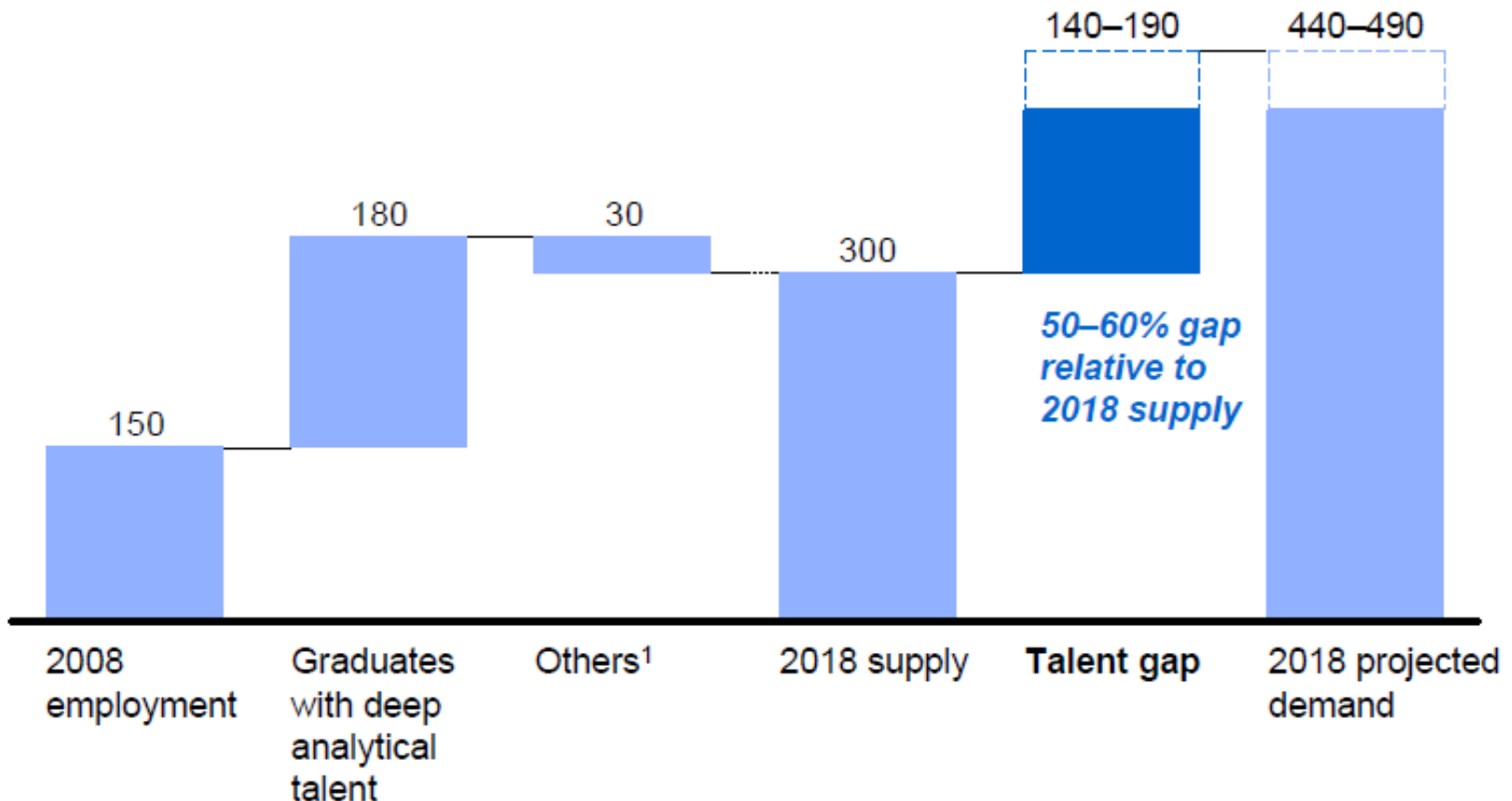
**Data Mining  $\approx$  Big Data  $\approx$   
Predictive Analytics  $\approx$  Data Science**

# Good News: Huge demand for data mining skills

**Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018**

Supply and demand of deep analytical talent by 2018

Thousand people



<sup>1</sup> Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

# What is Data Mining?

- Given lots of data
- Discover patterns and models that are
  - ▲ **Valid:** hold on new data with some certainty
  - ▲ **Useful:** should be possible to act on the item
  - ▲ **Unexpected:** non-obvious to the system
  - ▲ **Understandable:** humans should be able to interpret the pattern

# Frequent Itemset Mining Application



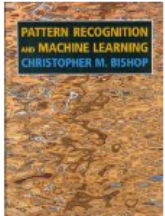
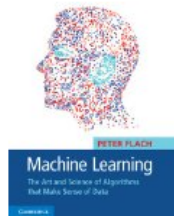

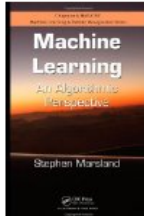
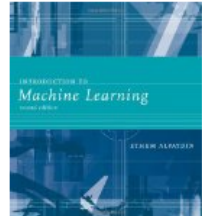
- Market-basket analysis to improve sales in stores such as Walmart
- “Classic” application was analyzing what people bought together in a brick-and-mortar store
  - ▲ Apocryphal story of “diapers and beer” discovery
  - ▲ Used to position potato chips between diapers and beer to enhance sales of potato chips



# Recommendation Systems: Example

- Amazon recommendation engine

## Related to Items You've Viewed










You viewed	Customers who viewed this also viewed					
						
Machine Learning ➤ Tom M. Mitchell Hardcover ★★★★☆ (48) \$217.87	Learning From Data ➤ Hsuan-Tien Lin, Yaser S. Abu-Mostafa, Malik Magdon-Ismail Hardcover ★★★★☆ (63)	Pattern Recognition and Machine Learning ➤ Christopher M. Bishop Hardcover ★★★★☆ (97) \$94.95 \$71.44	Machine Learning: The Art and Science... ➤ Peter A. Flach Paperback ★★★★☆ (11) \$64.00 \$57.60	The Elements of Statistical Learning... Trevor Hastie, Robert Tibshirani, ... Hardcover ★★★★☆ (36) \$89.95 \$71.96	Machine Learning: An Algorithmic... ➤ Stephen Marsland Hardcover ★★★★☆ (24) \$79.95 \$67.33	Introduction to Machine Learning ➤ Ethem Alpaydin Hardcover ★★★★☆ (26) \$60.00 \$48.12

➤ [View or edit your browsing history](#)

# Recommendation Systems: Example

- LinkedIn recommendation engine

People you may know

 <p><b>Ravi Madhira</b> Data Scientist at DXC technology (Formerly known as Oregon State University)</p> <p><a href="#">Connect</a></p>	 <p><b>Narsimha Rapaka</b> Postdoc at KAUST Indian Institute of Technology, Kanpur</p> <p><a href="#">Connect</a></p>	 <p><b>Emmanuel Sandeep Ganji</b> Senior Manager at SYNDICATE BANK Venkata Jamithireddy and 18 others</p> <p><a href="#">Connect</a></p>
 <p><b>Gurmeet Singh</b> Scientist at BARC Indian Institute of Technology, Kanpur</p> <p><a href="#">Connect</a></p>	 <p><b>Nilesh Mishra</b> Engineering Manager at LogMeIn Puneet Kaur and 25 others</p> <p><a href="#">Connect</a></p>	 <p><b>Sowjanya Addala</b> PreSales Lead at Tata Consultancy Services Andhra University</p> <p><a href="#">Connect</a></p>
 <p><b>Sudarshna Gangwar</b> Software Engineer at eKincare (Aayuv Technologies Private) Arwen Twinkle Griffioen, PhD and 1 other</p> <p><a href="#">Connect</a></p>	 <p><b>Svetlana Lockwood</b> Postdoc at University sun xueliang and 6 others</p> <p><a href="#">Connect</a></p>	 <p><b>Diwaker Tripathi</b> Research Associate at University of Washington, Washington State University</p> <p><a href="#">Connect</a></p>

# Spam Filtering



googleteam

GOOGLE LOTTERY WINNER! CONTACT

**From:** googleteam **To:**

**Subject:** GOOGLE LOTTERY WINNER! CONTACT YOUR AGENT TO CLAIM YOUR PRIZE.

GOOGLE LOTTERY INTERNATIONAL  
INTERNATIONAL PROMOTION / PRIZE AWARD .

(WE ENCOURAGE GLOBALIZATION)

FROM: THE LOTTERY COORDINATOR,  
GOOGLE B.V. 44 9459 PE.

RESULTS FOR CATEGORY "A" DRAWS

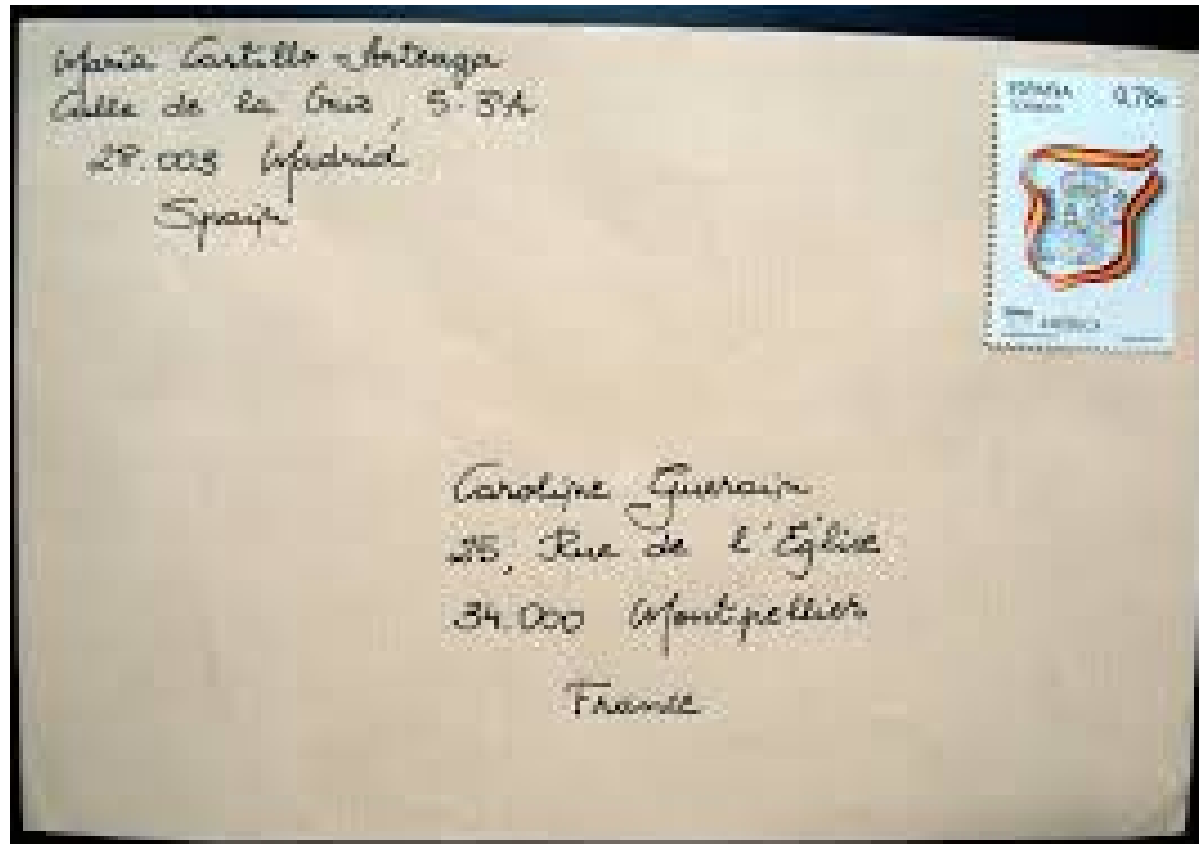
Congratulations to you as we bring to your notice, the results of the First Ca  
inform you that your email address have emerged a winner of One Million (1,  
money of Two Million (2,000,000.00) Euro shared among the 2 winners in this  
email addresses of individuals and companies from Africa, America, Asia, Au  
CONGRATULATIONS!

Your fund is now deposited with the paying Bank. In your best interest to avo  
award strictly from public notice until the process of transferring your claims |

NOTE: to file for your claim, please contact the claim department below on e

\*\*\*\*\*

# Optical Character Recognition



# Search Engines

The screenshot shows a Google search for "machine learning". The browser tabs include "Inbox - jana.iitk@gmail.co...", "Google Calendar", "CPT\_S 570: Machine Learning", "CPT\_S 570 (2 unread)", and "machine learning - Google...". The search bar shows "machine learning" with a magnifying glass icon. The results page shows "About 55,900,000 results (0.38 seconds)".

**Machine learning - Wikipedia, the free encyclopedia**  
[en.wikipedia.org/wiki/Machine\\_learning](https://en.wikipedia.org/wiki/Machine_learning) - Wikipedia  
Machine learning is a subfield of computer science (CS) and artificial intelligence (AI) that deals with the construction and study of systems that can learn from ...  
[List of machine learning ...](#) - [Computational learning theory](#) - [Unsupervised learning](#)

**Machine Learning | Coursera**  
<https://www.coursera.org/course/ml> - Coursera  
Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving ...

**ICML Beijing**  
[icml.cc/](http://icml.cc/)  
The 31st International Conference on Machine Learning (ICML 2014) will be held in ...  
ICML is the leading international machine learning conference and is ...  
You've visited this page many times. Last visit: 8/20/14

**Machine Learning Department - Carnegie Mellon University**  
[www.ml.cmu.edu/](http://www.ml.cmu.edu/) - Carnegie Mellon University  
The Machine Learning Department is an academic department within Carnegie Mellon University's School of Computer Science. We focus on research and ...

**Machine Learning - MIT OpenCourseWare**  
[ocw.mit.edu/.../6-867-machine-learning-fall-2006...](http://ocw.mit.edu/.../6-867-machine-learning-fall-2006...) - MIT OpenCourseWare  
6.867 is an introductory course on machine learning which gives an overview of many concepts, techniques, and algorithms in machine learning, beginning with ...

**CS 229: Machine Learning**  
[cs229.stanford.edu/](http://cs229.stanford.edu/)  
STANFORD. CS229 Machine Learning Autumn 2013. Announcements. new: We've just added extra office hours for PS4 and final project. Problem Set 4 has ...

**Machine Learning - Stanford School of Engineering ...**  
[see.stanford.edu/see/courseinfo.aspx?coll=348ca38a-3a6d-4052...](http://see.stanford.edu/see/courseinfo.aspx?coll=348ca38a-3a6d-4052...)  
Artificial Intelligence | Machine Learning ... This course provides a broad introduction to machine learning and statistical pattern recognition. Topics include: ...

**Machine Learning - Springer**  
[www.springer.com/computer/ai/.../109...](http://www.springer.com/computer/ai/.../109...) - Springer Science+Business Media

**Machine learning**  
Field of study  
Machine learning is a subfield of computer science and artificial intelligence that deals with the construction and study of systems that can learn from data, rather than follow only explicitly programmed instructions. [Wikipedia](#)

**Related topics**  
**Data mining** (the analysis step of the "Knowledge Discovery and Data Mining" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. [Wikipedia](#)  
**Explore:** [Data mining](#), [Computation](#)  
**Support vector machines** are a set of algorithms that learn from data by creating models that maximize their margin of error.  
[research.microsoft.com](http://research.microsoft.com)  
**Explore:** [Support vector machine](#)  
In machine learning, **pattern recognition** is the assignment of a label to a given input value. [Wikipedia](#)  
**Explore:** [Pattern recognition](#)

Feedback

# Machine Translation



Translate



English Spanish French English - detected ▼



English Spanish Arabic ▼

Translate

We are learning to learn

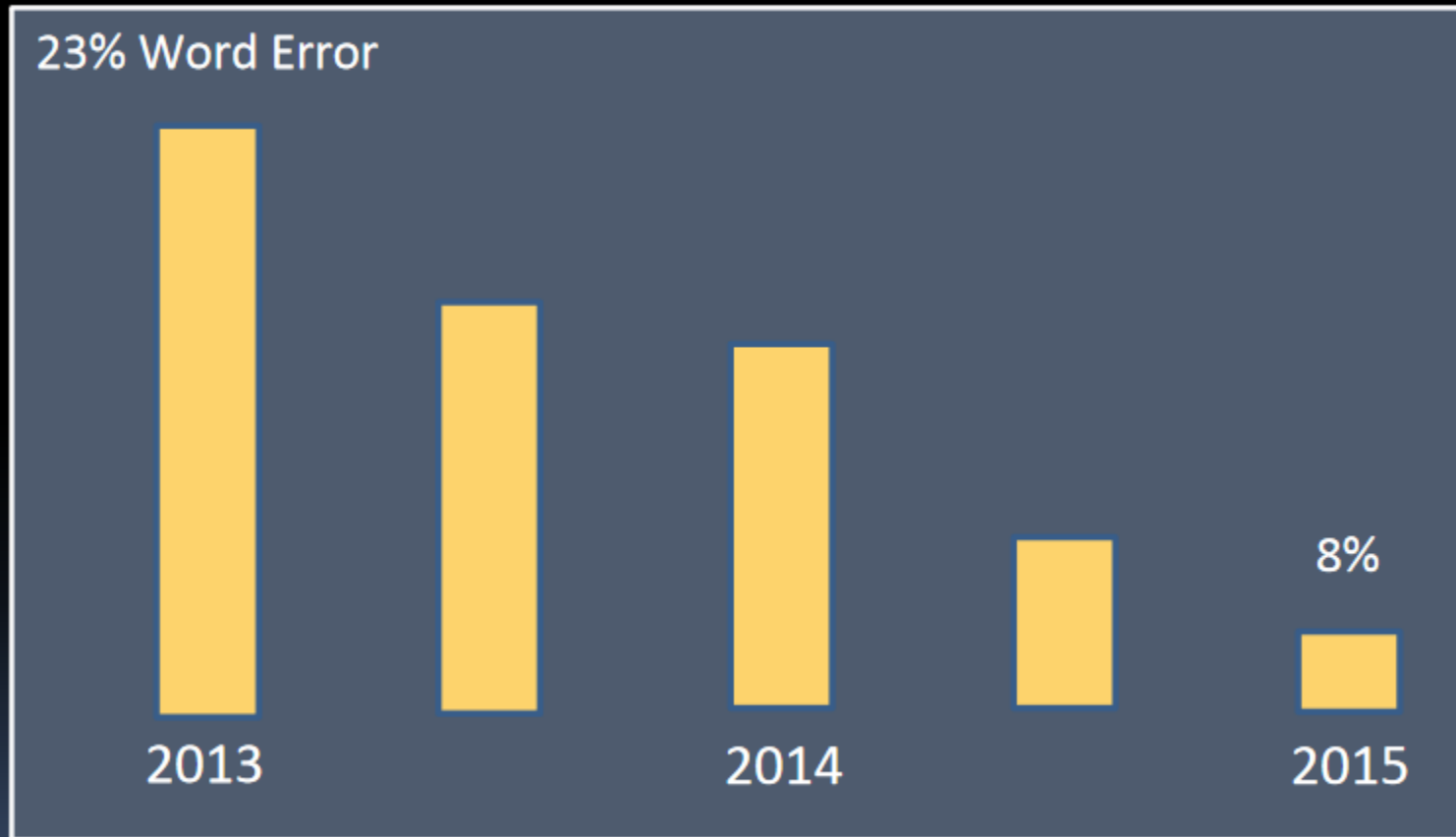


Estamos aprendiendo a aprender



# Speech Recognition

## Google Speech Recognition



Credit: Fernando Pereira & Matthew Firestone,  
Google

Credit: Tom Dietterich

# Image Captioning



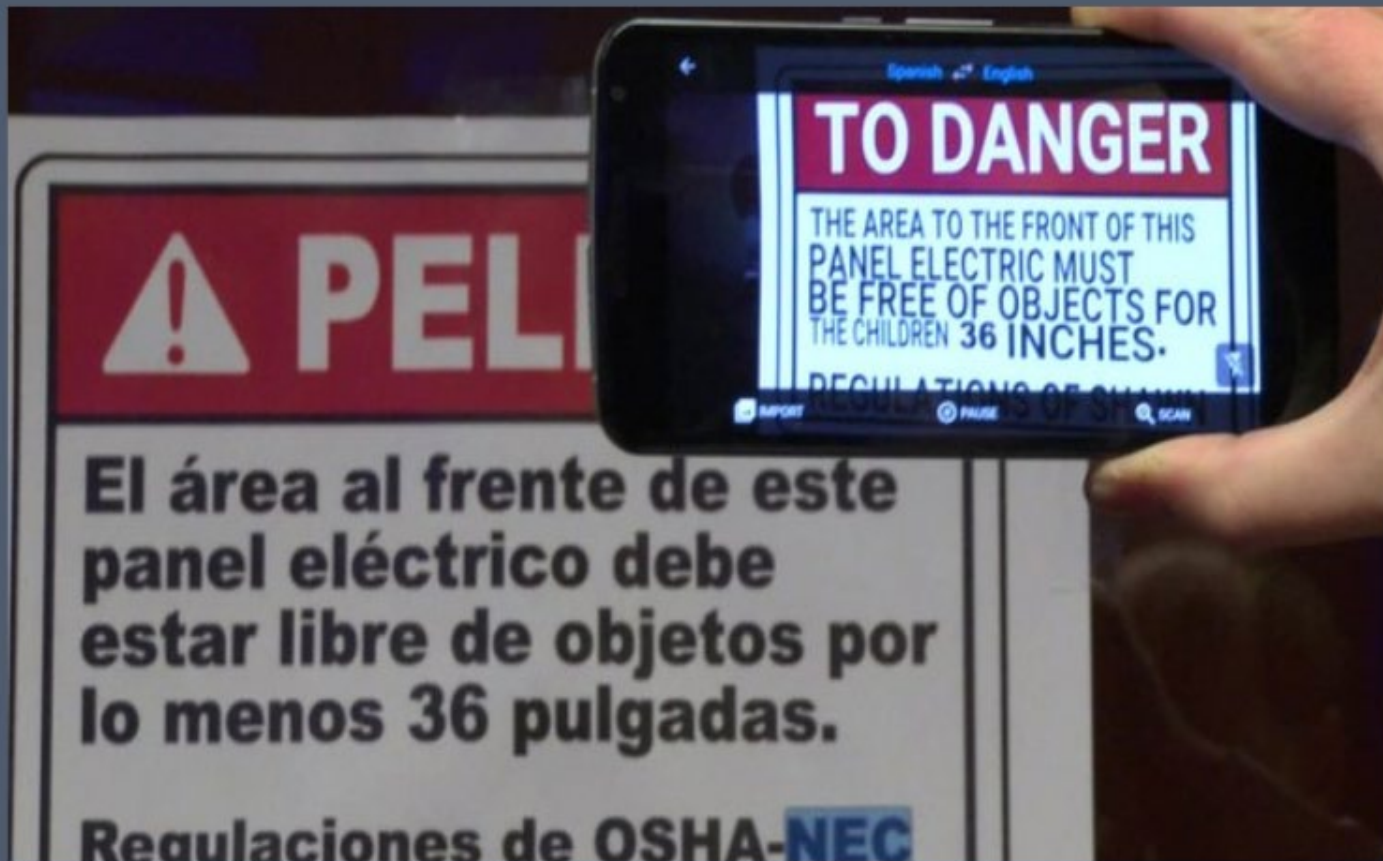
"a black and white cat is sitting  
on a chair."

Credit: Jeff Donahue, Trevor Darrell



# Perception + Translation

Google Translate from Images



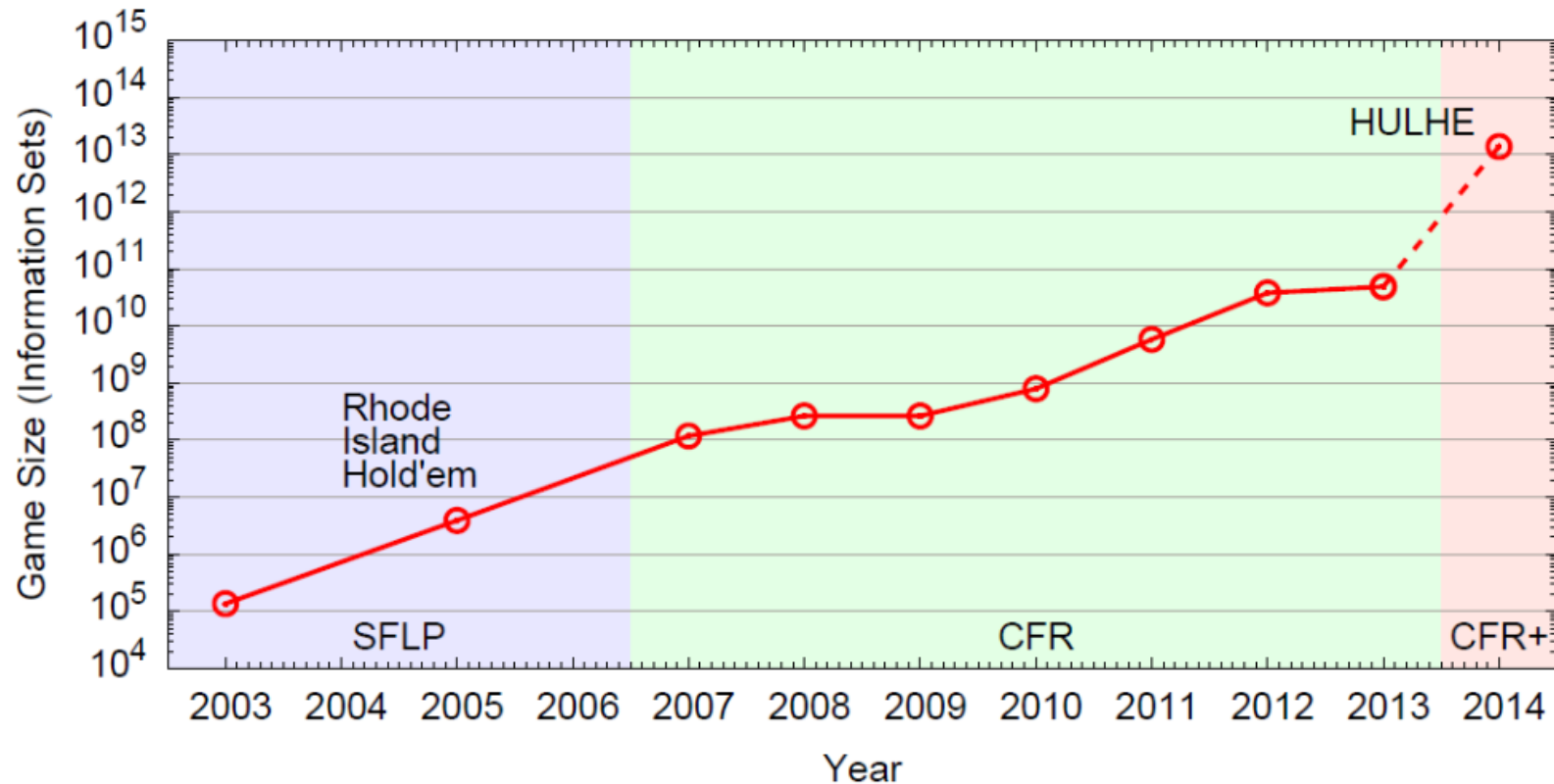
Credit: [www.bbc.com](http://www.bbc.com)

# Skype Translator



credit: Skype

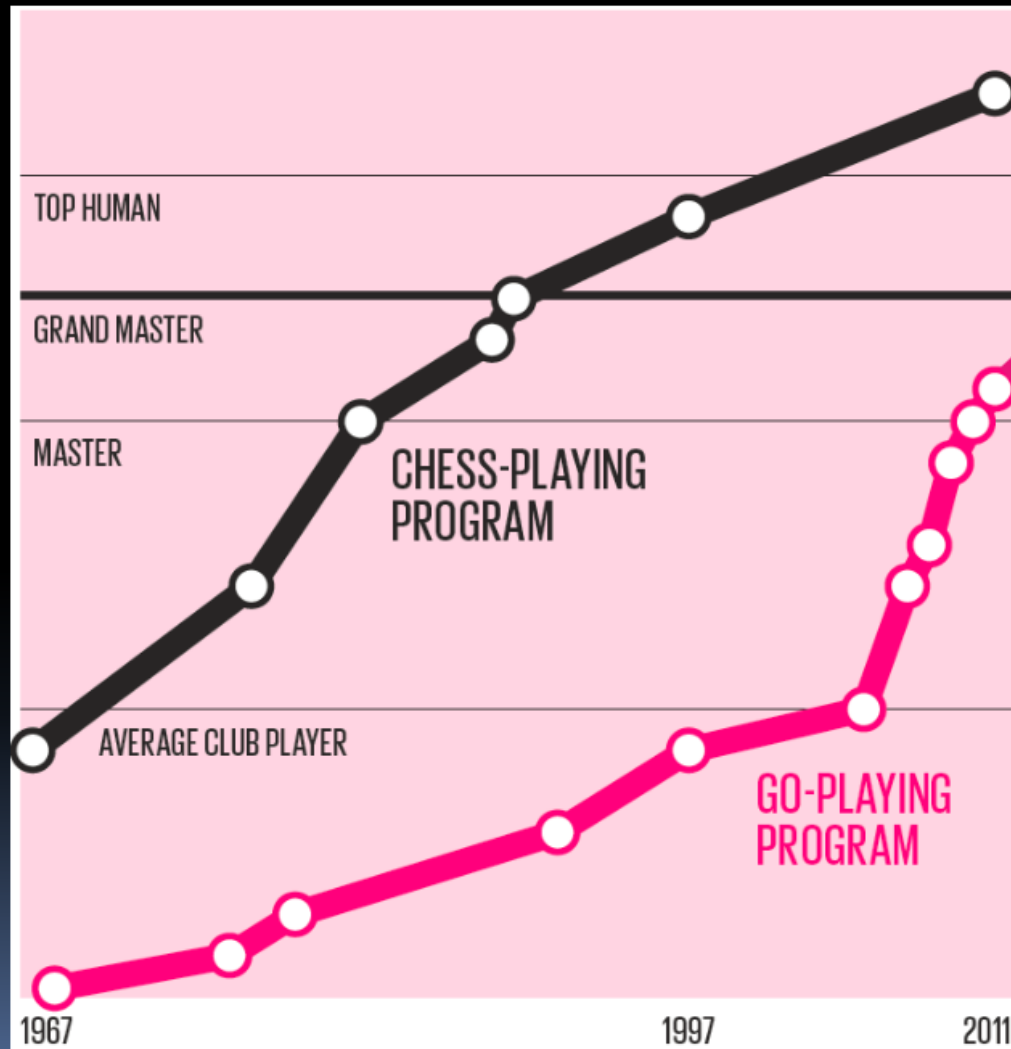
# Computers Playing Poker



Moore's Law

Credit: Michael Bowling

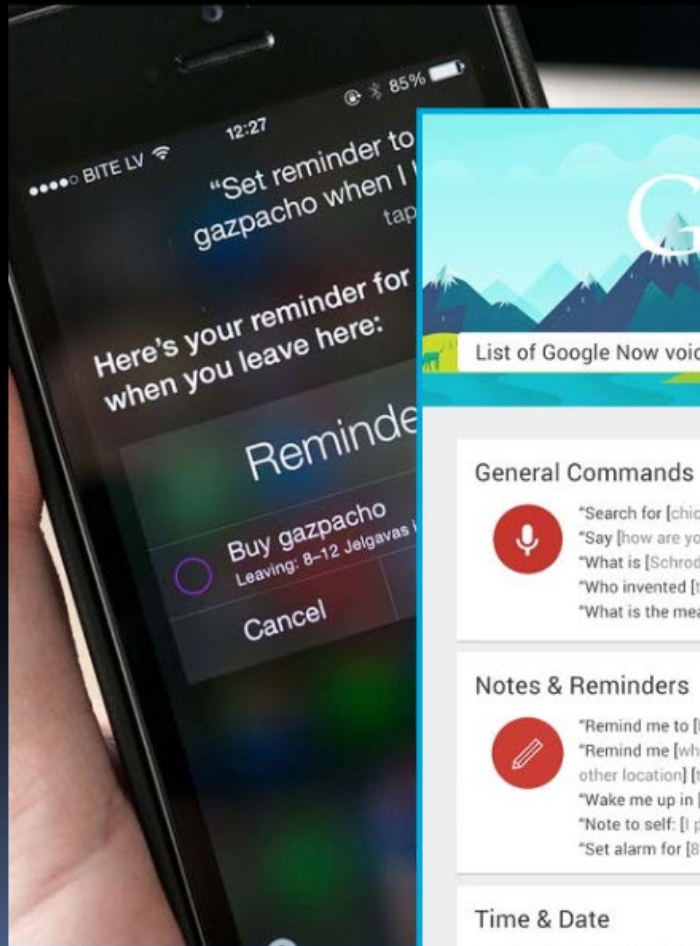
# Computers Playing Chess and Go



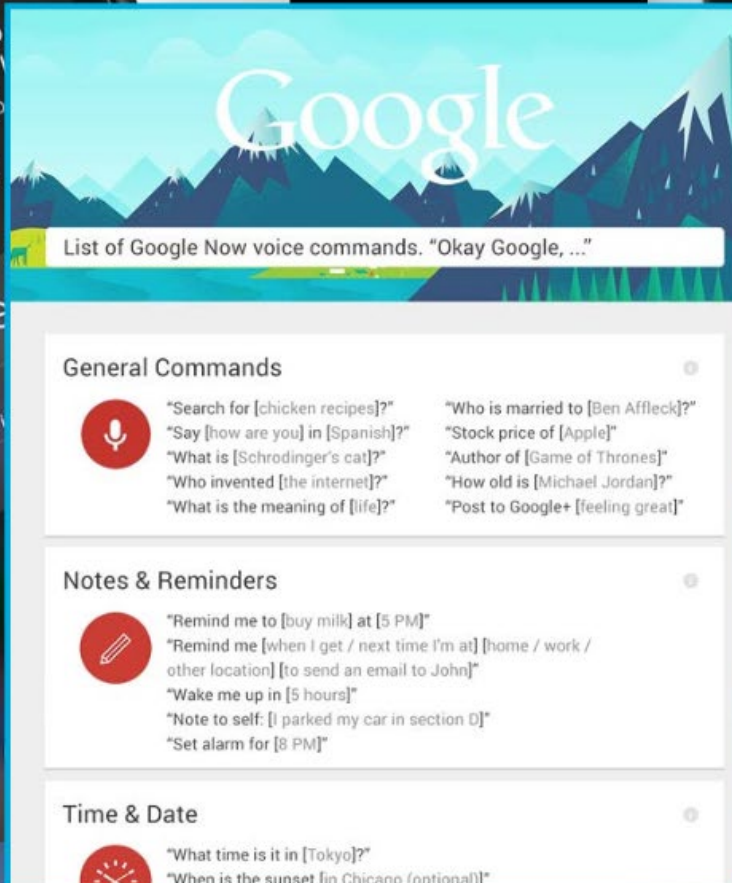
Silver, et al. (2016) *Nature*  
Deep Learning +  
Monte Carlo Tree Search

Credit: Martin Mueller

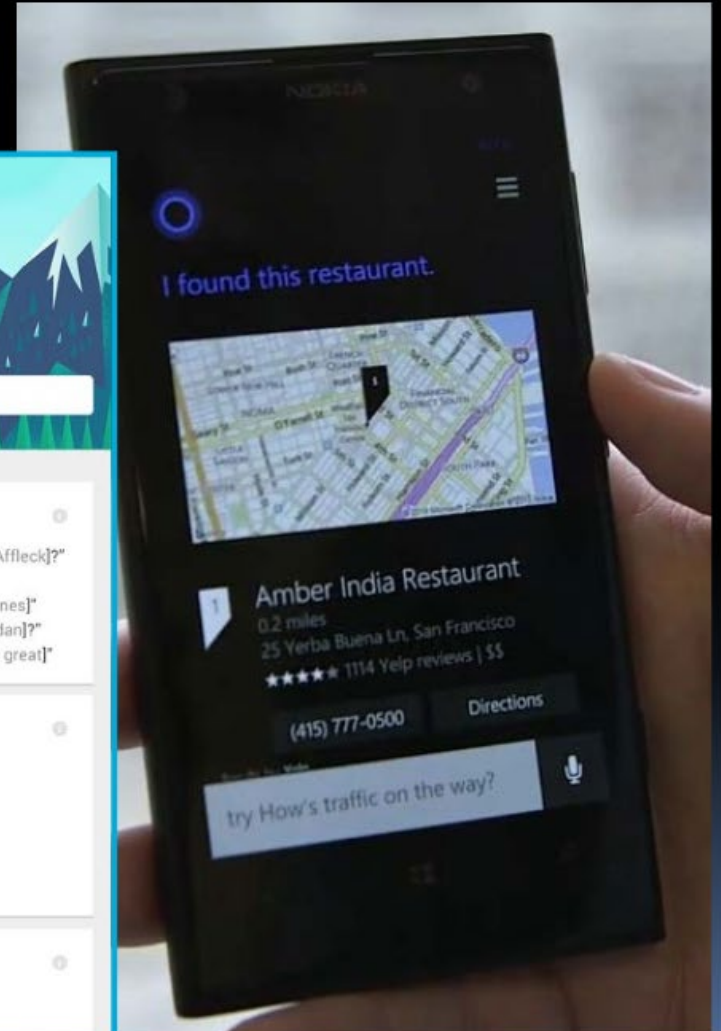
# Personal Assistants



Credit: mashable.com



Credit: trendblog.net



Credit: The Verge



# High-Stakes Applications: Self-Driving Cars



Credit: The Verge



Credit: delphi.com

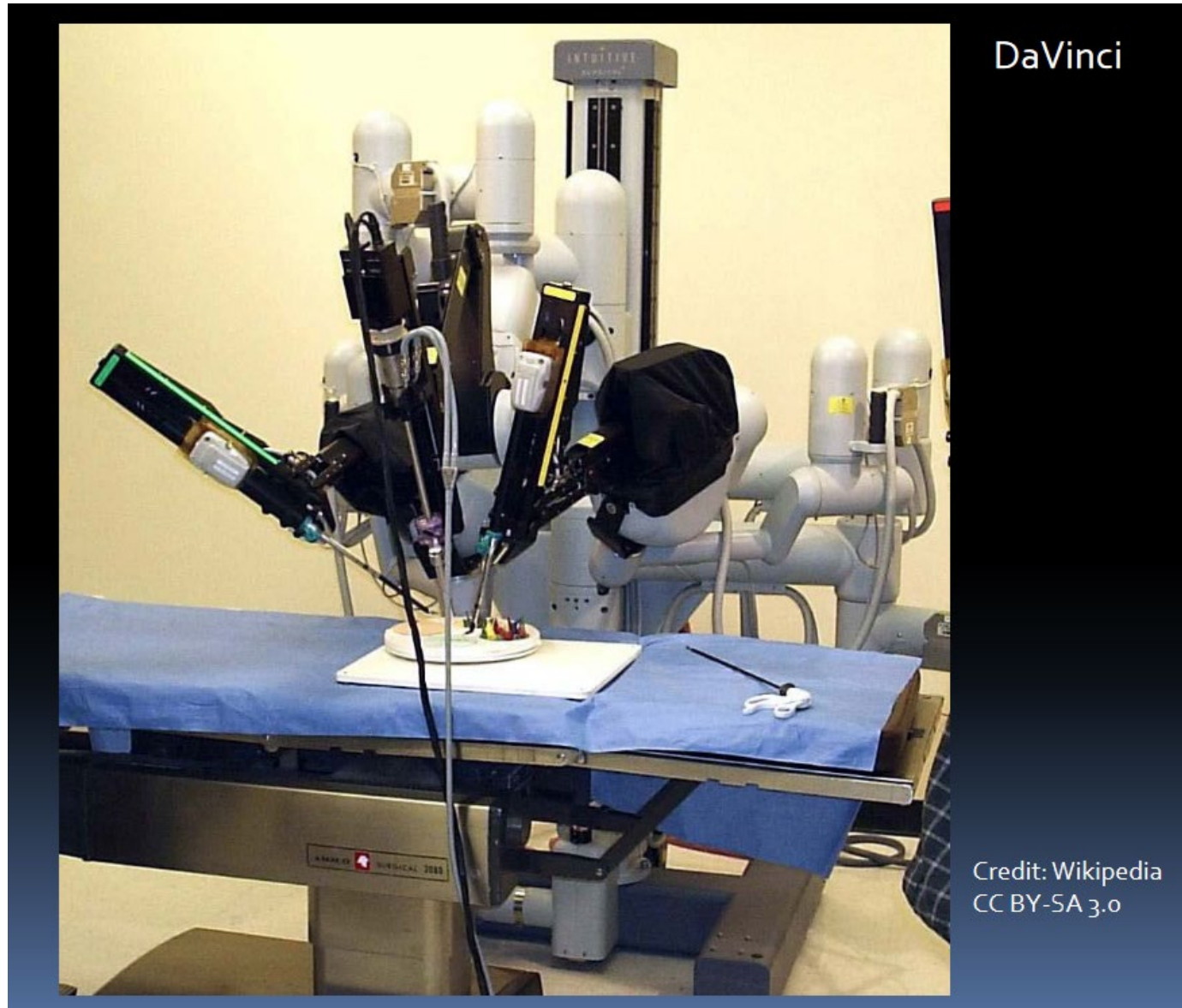
Tesla AutoSteer



Credit: Tesla Motors

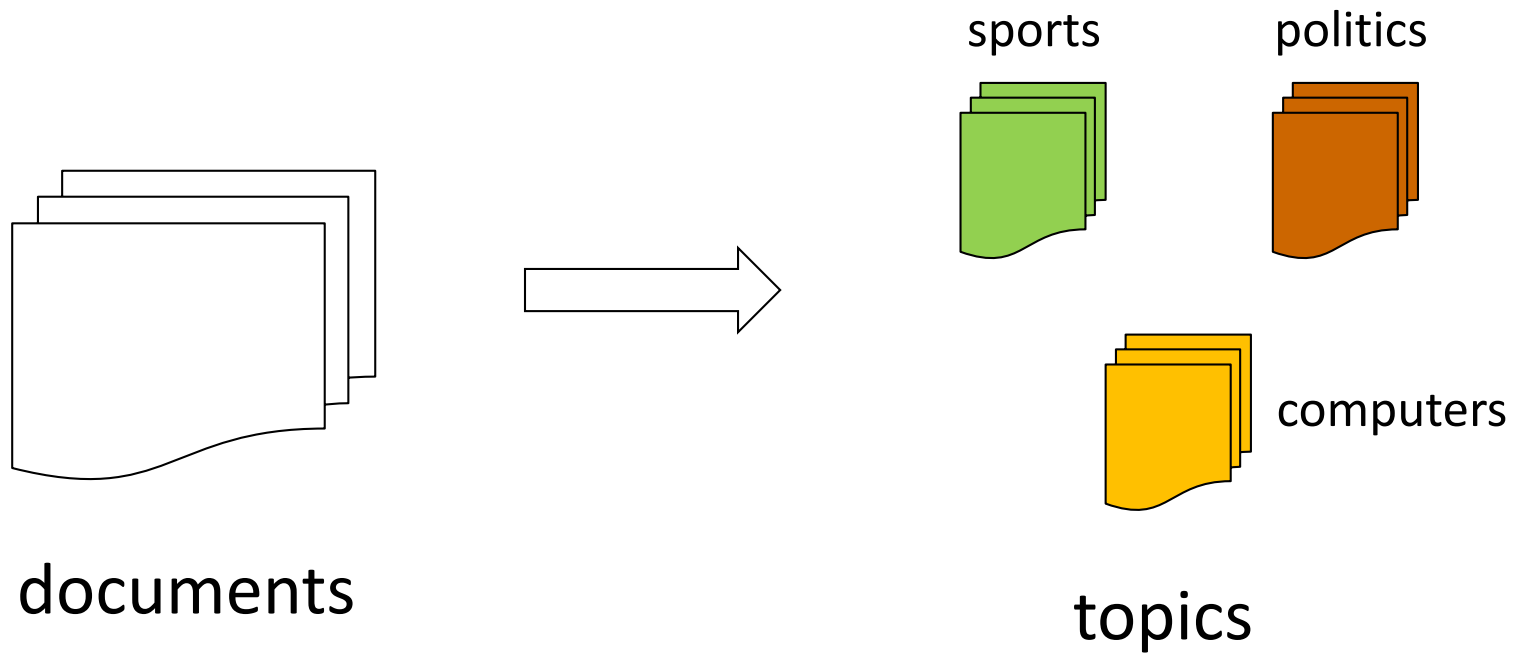
14

# High-Stakes Applications: Automated Surgical Assistants



# Clustering

- **Text Clustering**





# Anomaly Detection

- Anomaly is a pattern in the data that does not conform to the expected behavior
- Credit Card Fraud
  - ▶ An abnormally high purchase made on a credit card
- Cyber Intrusions
  - ▶ A web server involved in *ftp* traffic



# Computational Advertising: Ads vs. Search Results

## Web

Results 1 - 10 of about 2,230,000 for **geico**. (0.04 sec)

### [GEICO](#) Car Insurance. Get an auto insurance quote and save today ...

**GEICO** auto insurance, online car insurance quote, motorcycle insurance quote, online insurance sales and service from a leading insurance company.

[www.geico.com/](#) - 21k - Sep 22, 2005 - [Cached](#) - [Similar pages](#)

[Auto Insurance](#) - [Buy Auto Insurance](#)

[Contact Us](#) - [Make a Payment](#)

[More results from www.geico.com »](#)

### [Geico](#), Google Settle Trademark Dispute

The case was resolved out of court, so advertisers are still left without legal guidance on use of trademarks within ads or as keywords.

[www.clickz.com/news/article.php/3547356](#) - 44k - [Cached](#) - [Similar pages](#)

### Google and [GEICO](#) settle AdWords dispute | The Register

Google and car insurance firm **GEICO** have settled a trade mark dispute over ... Car insurance firm **GEICO** sued both Google and Yahoo! subsidiary Overture in ...

[www.theregister.co.uk/2005/09/09/google\\_geico\\_settlement/](#) - 21k - [Cached](#) - [Similar pages](#)

### [GEICO](#) v. Google

... involving a lawsuit filed by Government Employees Insurance Company (**GEICO**). **GEICO** has filed suit against two major Internet search engine operators, ...

[www.consumeraffairs.com/news04/geico\\_google.html](#) - 19k - [Cached](#) - [Similar pages](#)

## Sponsored Links

### [Great Car Insurance Rates](#)

Simplify Buying Insurance at Safeco

See Your Rate with an Instant Quote

[www.Safeco.com](#)

### [Free Insurance Quotes](#)

Fill out one simple form to get multiple quotes from local agents.

[www.HometownQuotes.com](#)

### [5 Free Quotes. 1 Form.](#)

Get 5 Free Quotes In Minutes!

You Have Nothing To Lose. It's Free

[sayyessoftware.com/Insurance](#)

Missouri

# Computational Advertising: Web 2.0

- **Performance-based advertising**

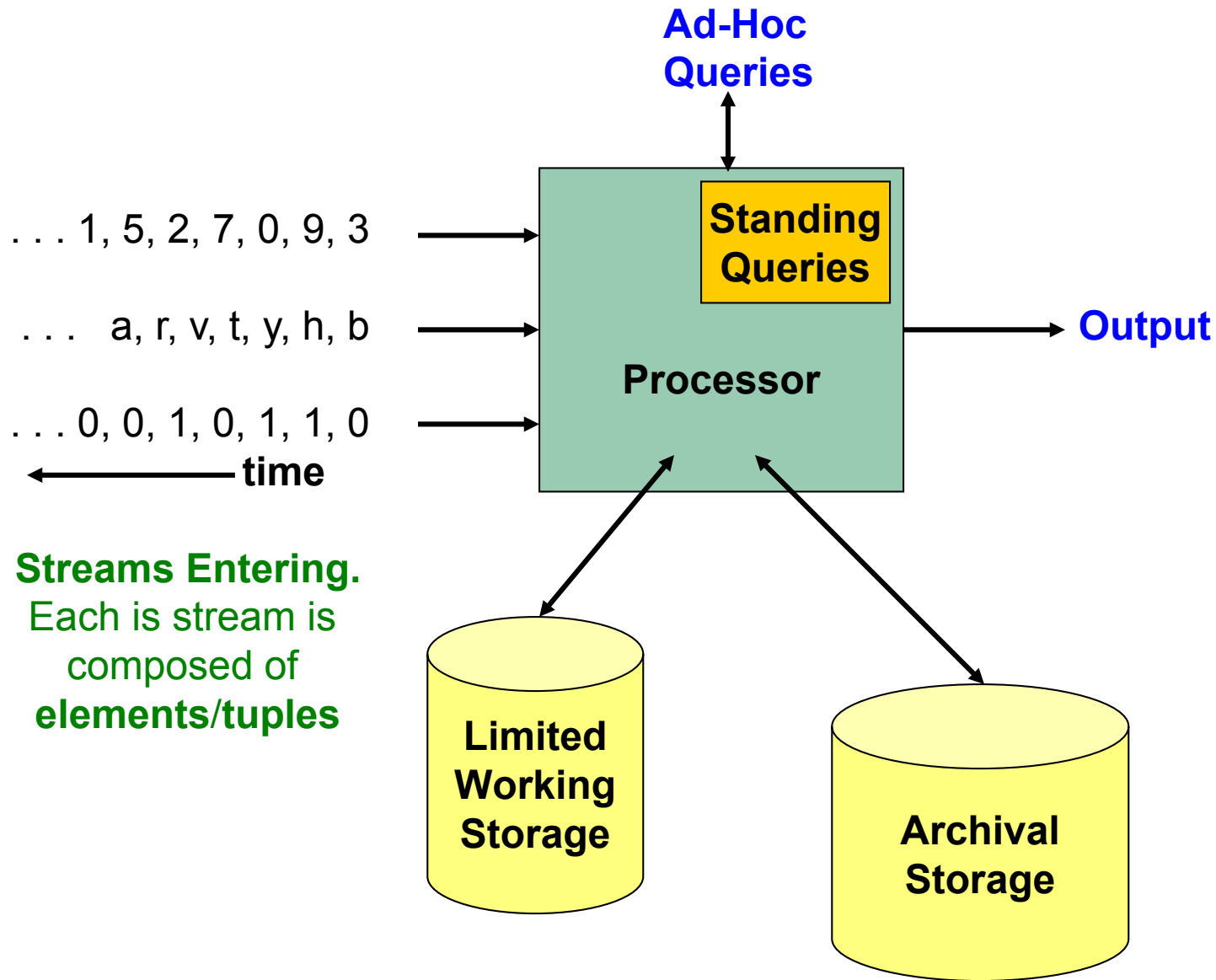
- ▲ Multi-billion-dollar industry

- **Interesting problem:**

- What ads to show for a given query?**

- **If I am an advertiser, which search terms should I bid on and how much should I bid?**

# Mining Data Streams



# Mining Data Streams: Applications

- **Mining query streams**

- ▲ Google wants to know what queries are more frequent today than yesterday

- **Mining click streams**

- ▲ Yahoo wants to know which of its pages are getting an unusual number of hits in the past hour

- **Mining social network news feeds**

- ▲ E.g., look for trending topics on Twitter, Facebook

# Course Contents

- **Introduction to the field of data mining**
    - ▲ Automatically analyze data using computers for discovering knowledge and insights
- ▲ **Computational problems** motivated from real-world applications
  - ▲ **Computational algorithms** to solve data analysis problems
  - ▲ **Real-world applications** for each of the data analysis problems

# Tentative Syllabus

- Mining frequent item sets and association rules
- Recommendation algorithms
- Supervised learning algorithms
  - ▲ Classification and regression tasks
- Clustering algorithms
- Outlier and anomaly detection algorithms
- Computational Advertising
- Mining Data Streams
- Responsible data mining