CS 335 Machine Learning

Dan Sheldon

Fall 2014

What is Machine Learning?

What is Machine Learning?

- How do you program a computer to
  - Recognize faces?
  - Recommend movies?
  - Decide which web pages are relevant to a Google search query?

A Simple Task: Recognize Obama

Page 2

A Simple Task: Recognize Obama

- Input: picture
- Output: yes/no

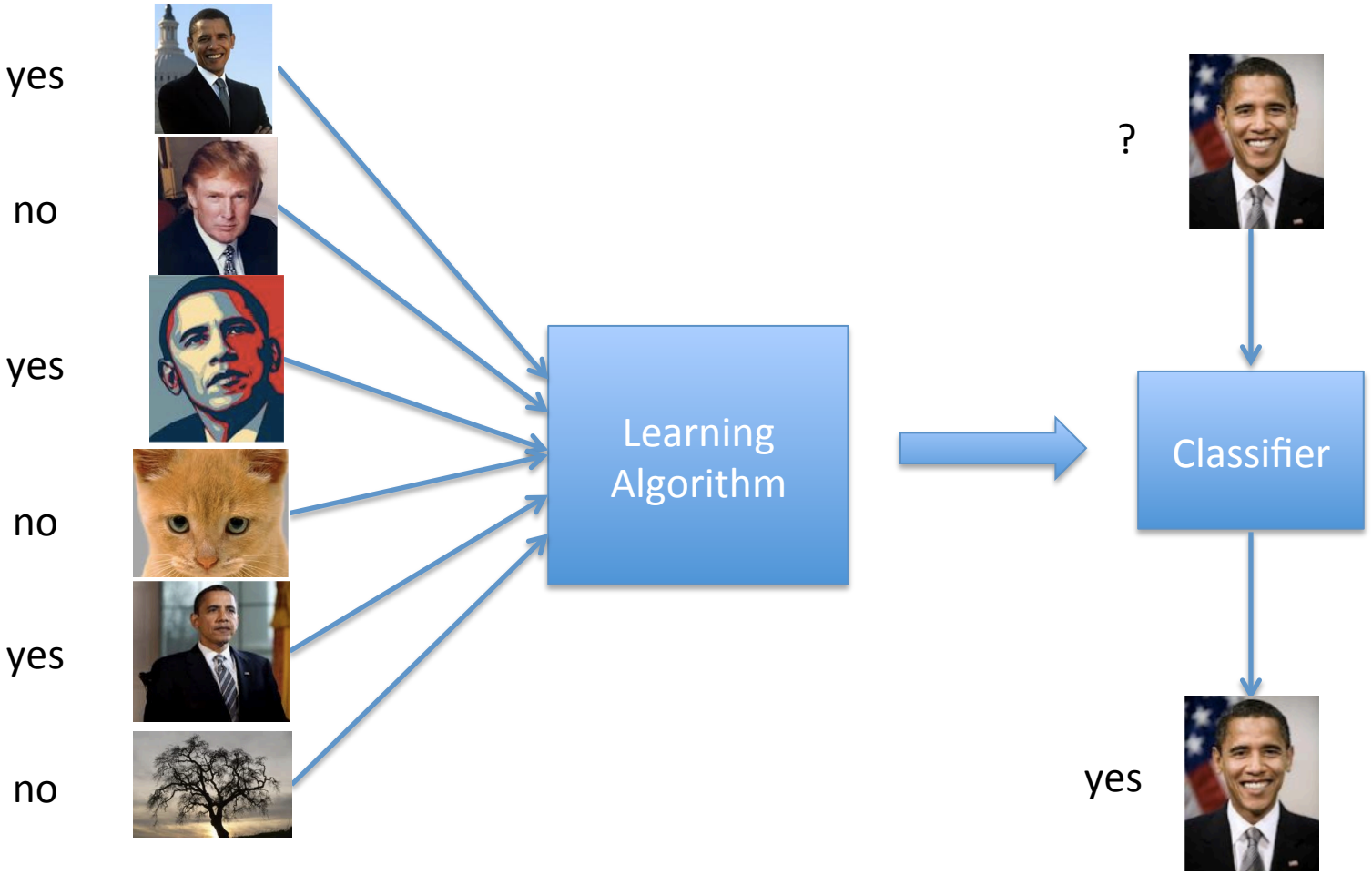
- Can you program this?
  - Probably not…
  - But you can *show* a computer how to solve this task

Examples

No

Yes

Learning from Examples



Discussion

- Is it easier to devise a learning algorithm than it is to program an Obama recognizer?
- Is it more useful to have a learning algorithm, or an Obama recognizer?

What is Machine Learning?

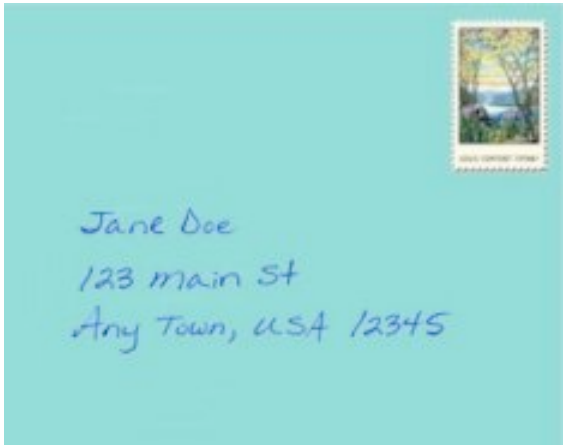
- *Machine learning is the practice of programming a computer to learn to solve a task through experience, rather than directly programming it to solve the task.*

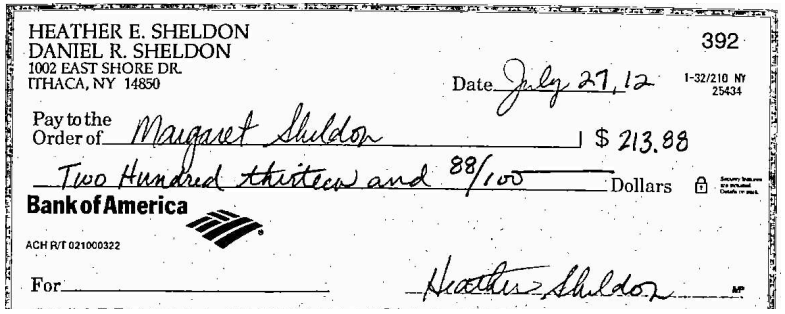
Why should I care about ML?

You tell me…

- What are some examples of ML in your day-to-day life?

ML makes the world go round.

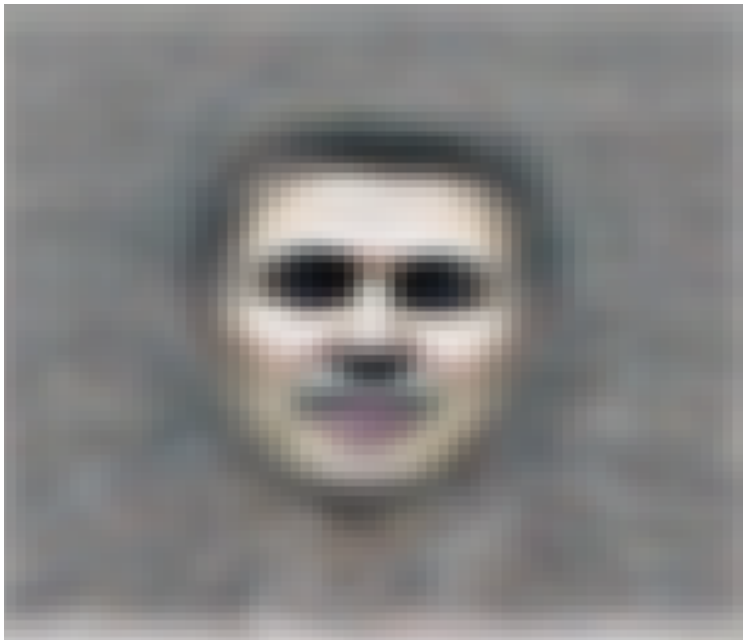




ML Achievements

ML wins Jeopardy!

ML Achievements



ML watches YouTube for three straight days!
(and learns to recognize cats)

http://www.npr.org/2012/06/26/155792609/a-massive-google-network-learns-to-identify

Building High-level Features Using Large Scale Unsupervised Learning

Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeffrey Dean, and Andrew Y. Ng

ML plays quiz bowl

QUESTION:
He left unfinished a novel whose title character forges his father's signature to get out of school and avoids the draft by feigning desire to join. A more famous work by this author tells of the rise and fall of the composer Adrian Leverkühn. Another of his novels features the jesuit Naptha and his opponent Settembrini, while his most famous work depicts the aging writer Gustav von Aschenbach. Name this German author of The Magic Mountain and Death in Venice.

ANSWER:
Thomas Mann

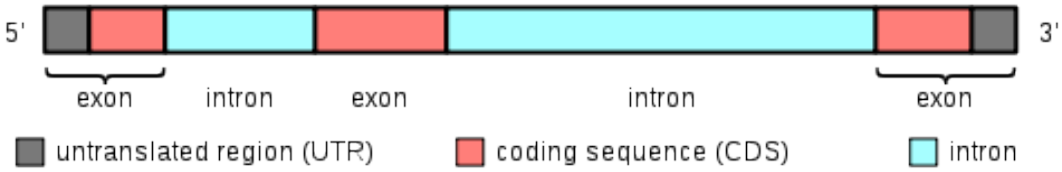
A NEURAL NETWORK FOR FACTOID QUESTION ANSWERING OVER PARAGRAPHS

Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III

In Proceedings of EMNLP 2014, download PDF here

ML in Science

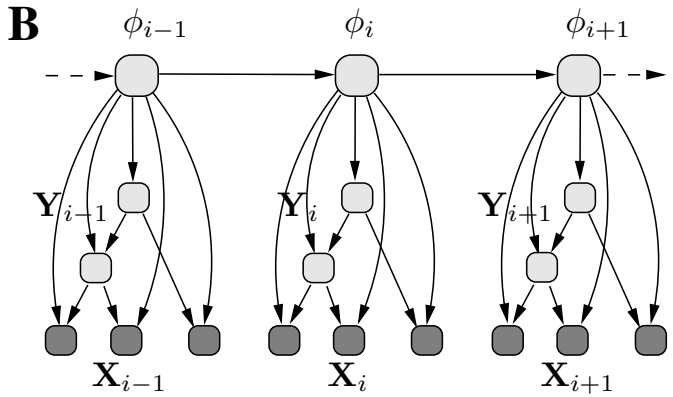
- Bioinformatics
  - Gene prediction



Computational Identification of Evolutionarily Conserved Exons

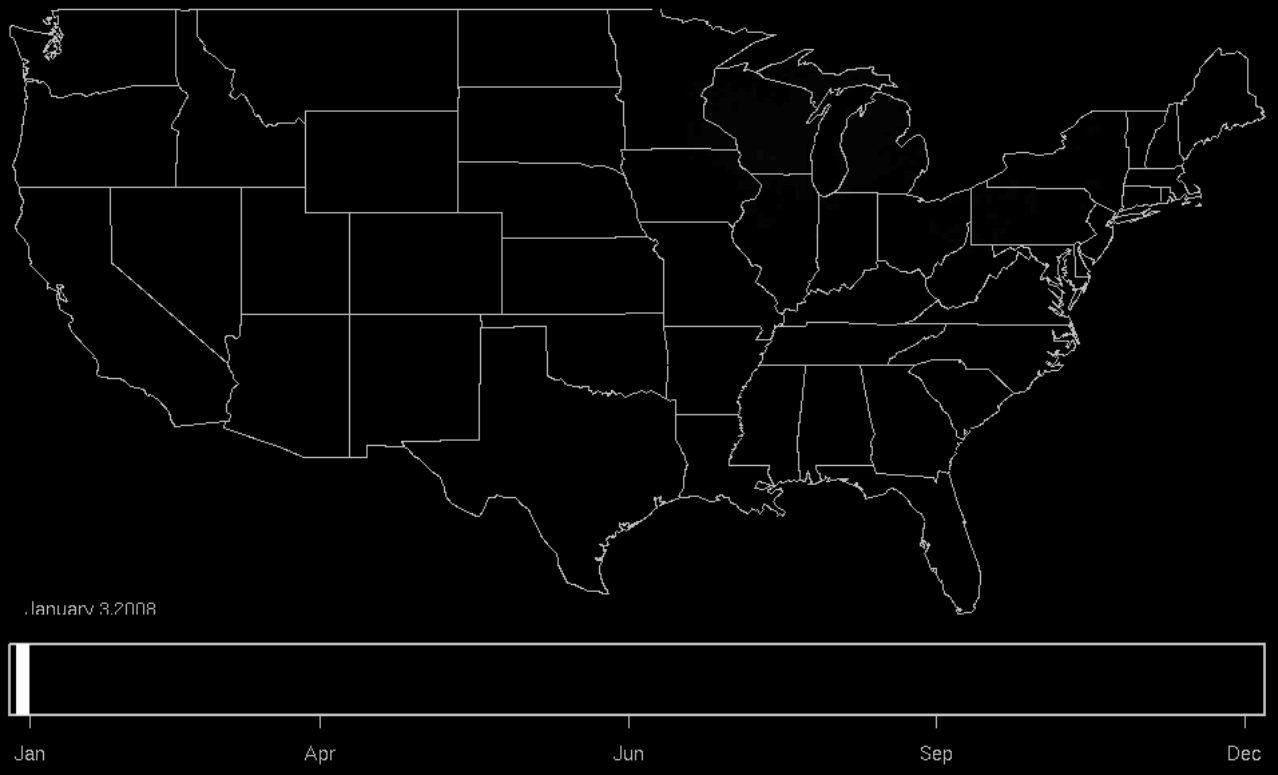
Adam Siepel
Center for Biomolecular Science and Engr.
University of California
Santa Cruz, CA 95064, USA
acs@soe.ucsc.edu

David Haussler
Howard Hughes Medical Institute and
Center for Biomolecular Science and Engr.
University of California
Santa Cruz, CA 95064, USA
haussler@soe.ucsc.edu



Hidden Markov model

Species Distribution Modeling



eBird

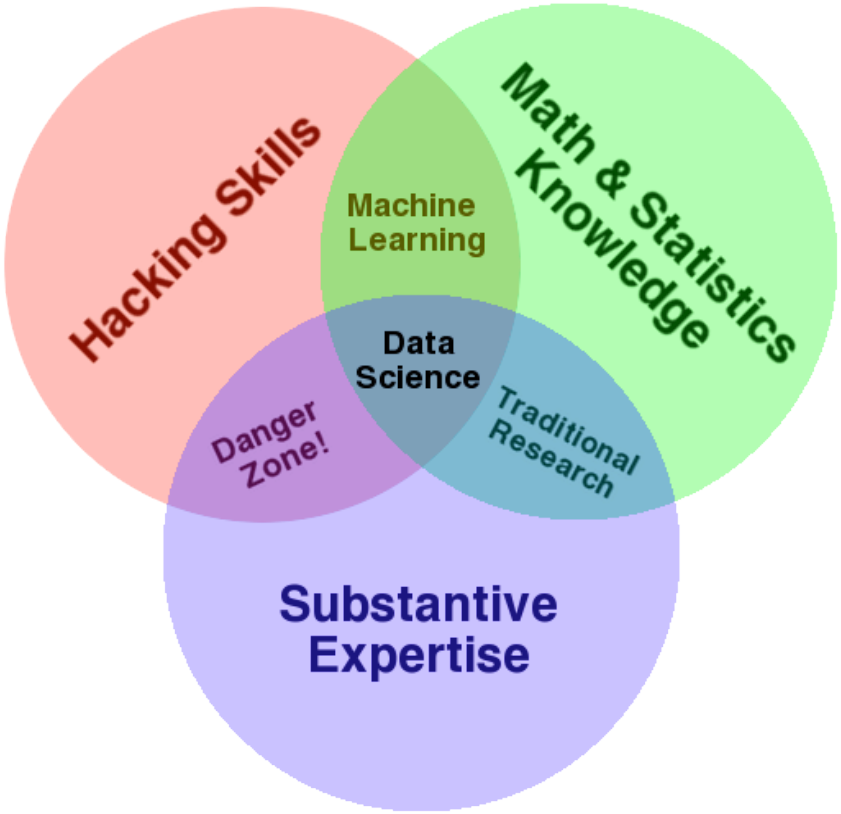
Big Data Revolution

- "Analyzing large data sets—so called big data—will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus…"

- US shortage of
  - 140,000 to 190,000 people with analytical expertise
  - 1.5 million managers and analysts with skills to understand and make decisions based on the analysis of big data

Source: Big data: The next frontier for innovation, competition, and productivity, McKinsey & Company, 2011

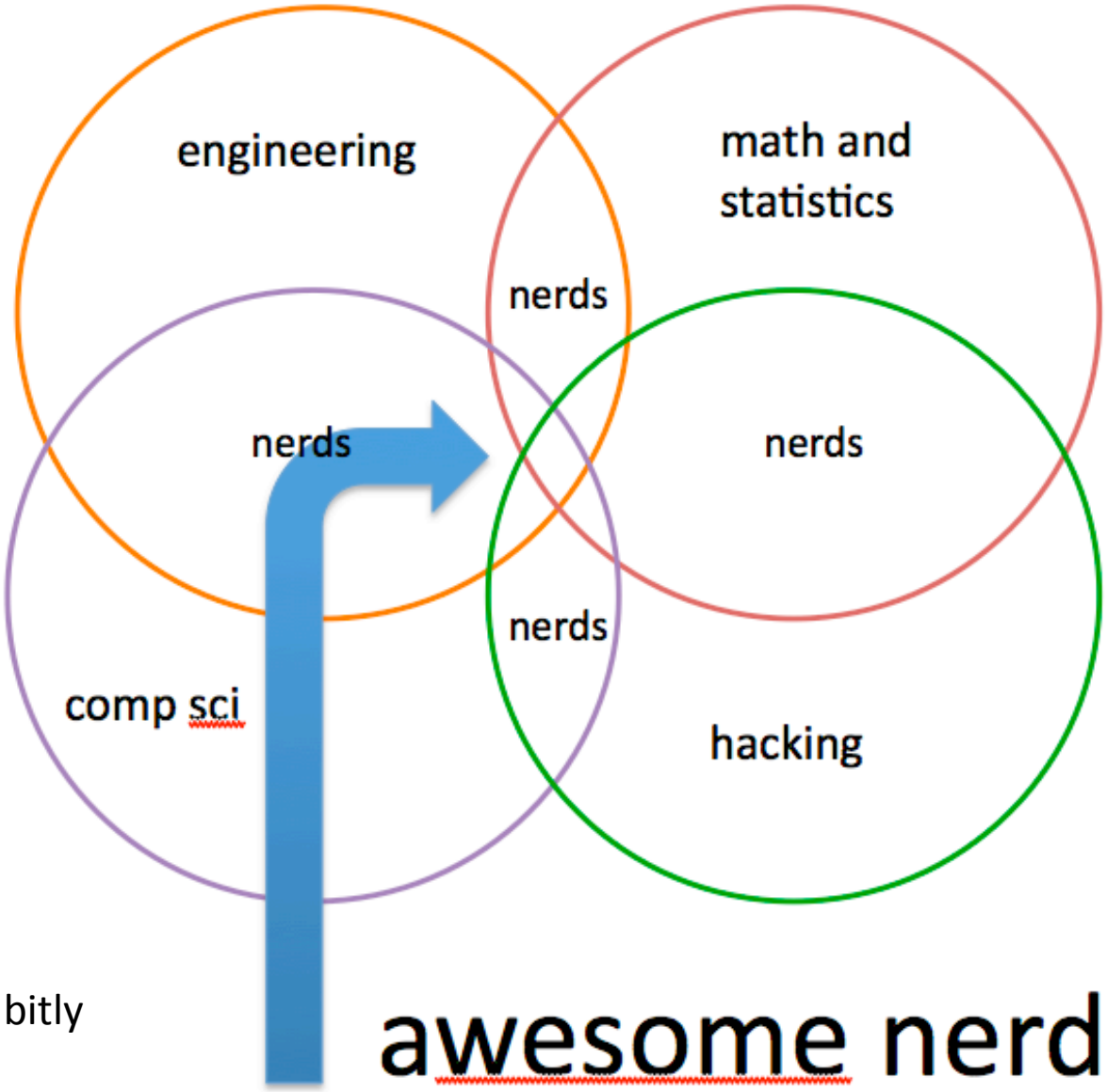
Data Science



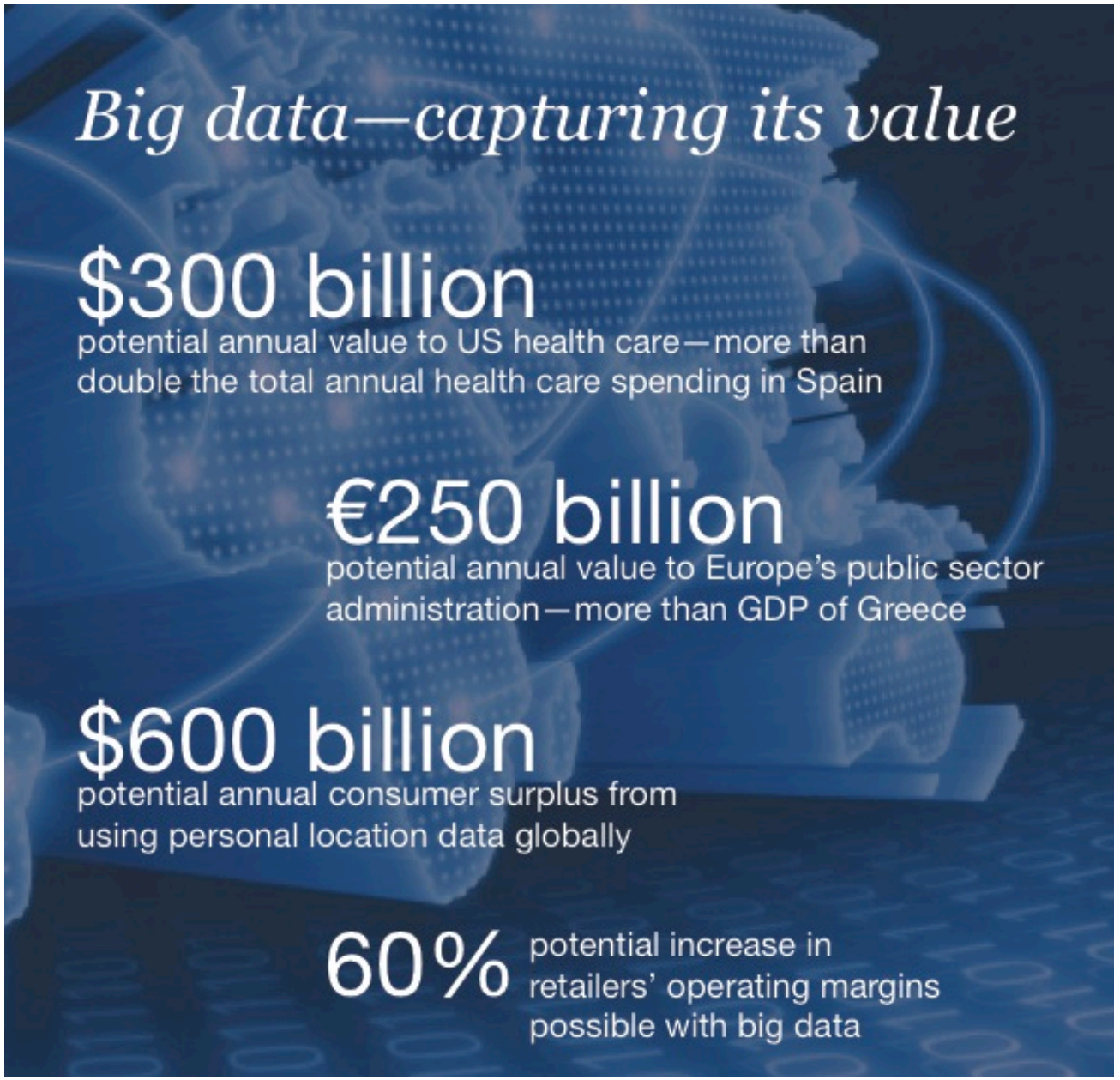
http://drewconway.com

Extracting insight and knowledge from (big) data

Data scientists?



By Hilary Mason, bitly
From forbes.com



Source: Big data: The next frontier for innovation, competition, and productivity, McKinsey & Company, 2011

OK, but what are we actually going to do in this class?

Course Goals

- Learn basic building blocks and general principles of designing ML algorithms
- Learn specific, widely used ML algorithms
- Learn methodology and tools to apply ML algorithms to real data and evaluate their performance
- Course is organized as a sequence of problems and algorithms

Course Outline

- Supervised learning
  - Learn from examples
- Unsupervised learning
  - Find patterns in data
- Probabilistic learning
  - Quantify uncertainty

"20% chance of rain"
"80% chance of survival"
"90% sure it is President Obama"

Logistics

Course Webpage

http://people.cs.umass.edu/~sheldon/teaching/cs335/index.html

- Entry point for all course information
  - Course policies (review)
  - Schedule
  - Slides
  - Homework

- Office hours:
  Tue 4-5pm, Thu 3:30-4:30 pm, Clapp 222B

Math

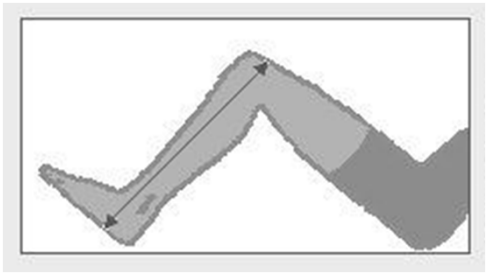
- Warning: there is math in this course
- What you should know
  - Calculus
    - Derivatives
    - (Partial derivatives)
  - Probability
    - Sample space, events, conditional probability, discrete random variables, expected value
    - Review later in semester
  - Linear algebra
    - Nothing (but hopefully seen matrices and vectors)
- Self-assessment in HW0

MATLAB

- All programming in this course done in MATLAB
  - Installed in Clapp 202
- MATLAB session: Fourth hour next Friday 9/12
  - Required

What's Next?

- Homework 0 posted
  - By Tuesday:
    - Math self-assessment
    - Read course policies
    - Post on Piazza
    - Help collect data (anonymous)
      - Height, knee height, arm span
      - List 5 movies
  - By next Friday:
    - Get started with MATLAB exercise



What's Next?

- Fourth Hour tomorrow (Optional)
  - 10:00–10:50 in Clapp 218
  - Calculus brush-up
    - Derivatives
    - Intuition and rules

(If time) First MATLAB Session

- Open MATLAB
- Change directories
- Edit file
- Run script

(If time) A First Example

- Polynomial fitting (on board)