Week 1

Agenda

1. Syllabus discussion
2. Introductions all around
3. What is machine learning?
4. 'Unreasonable Effectiveness' paper
5. Github Ids + Style Guide
6. Tutorial notebook

For next week:
- Read Domingos paper

Syllabus

Generally, the live sessions will consist of:
(1/3) A selective review of async
(1/3) A supplementary lecture
(1/3) Small group programming

Week 1: Welcome!
Week 2: Nearest Neighbors

Week 3: Naive bayes, Spam Classification
Week 4: Decision Trees, Bagging, Boosting
Week 5: Linear Regression, Logistic Regression.
Week 6: Gradient Descent, Regularization (Deep Learning)
Week 7: Neural Networks (Deep Learning)
Week 8: Algorithm Comparison (Deep Learning)

Week 9: K-Means
Week 10: Gaussian Mixture Models
Week 11: PCA

Week 12: Network Analysis
Week 13: Recommender Systems

Week 14: Class presentations

Due dates are always on Sunday nights.

Notes
- Grades are based on four projects. Each worth 25%.
- **The projects are very difficult and time consuming. Even for students with past ML experience.**
- The projects available now in Github. You can start any of them any time. **Start early!**
- Once a student gets behind, that student tends to stay behind.
- The projects are due the Sunday night after week 5, week 9, and week 12, and week 13.
- To turn in your work, just submit the completed ipython notebook to the ISVC.
- Projects are graded by me by hand. I read and run everyone's code.

Common Q's

- Office hours? Tuesdays 6:30-7:30 PST
- Can I get an extension? Yes, but 10% penalty
- How do I turn in work? Upload notebook to the ISVC
- Best place for q's? Wall is best, email second best
- Supplementary textbooks? I like…
  - https://web.stanford.edu/~hastie/Papers/ESLII.pdf
  - https://www.deeplearningbook.org/contents/TOC.html
- Others?

Introductions

1. Where are you located?
2. What brings you to MIDS?
3. When you think of machine learning, what do you think of?
4. One thing most people don't know about you?

What does it mean to learn a function?

Featured Prediction Competition

Zillow Prize: Zillow's Home Value Prediction (Zestimate)
Can you improve the algorithm that changed the world of real estate?

$1,200,000
Prize Money

Zillow · 3,779 teams · 8 months ago

"Zestimates" are estimated home values based on 7.5 million statistical and machine learning models that analyze hundreds of data points on each property. And, by continually improving the median margin of error (from 14% at the onset to 5% today), Zillow has since become established as one of the largest, most trusted marketplaces for real estate information in the U.S. and a leading example of impactful machine learning.

fireplace & Sound views. Spacious kitchen w/ slab granite surfaces & center island. Huge master suite with Jacuzzi tub & separate shower. Features; hdwd floors, all

Data Coverage and Zestimate Accuracy Table
Choose a location type below to change data:
Top Metro Areas
States/Counties*
National

	Zestimate Accuracy	Homes on Zillow	Homes With Zestimates	Within 5% of Sale Price	Within 10% of Sale Price	Within 20% of Sale Price	Median Error
Minneapolis-St Paul, MN	★★★★	1.2M					
New York, NY	★★★	5.3M	4.6M	49.1%	70.5%	84.8%	5.1%
Orlando, FL	★★★★	899.9K	791.5K	60.3%	79.8%	90.1%	3.7%
Philadelphia, PA	★★	2.1M	2.0M	48.1%	65.7%	77.7%	5.4%
Phoenix, AZ	★★★★	1.7M	1.5M	67.0%	84.8%	93.7%	3.1%
Pittsburgh, PA	★★	1.0M	867.7K	42.1%	61.9%	75.9%	6.5%
Portland, OR	★★★★	816.3K	704.0K	64.4%	84.4%	93.0%	3.4%

What does it mean to learn a function?

f(beds, baths, sqft) = alpha + Beta1 * beds + Beta2 * baths + Beta3 * sqft
f1(beds, baths) = 1 + .5 * beds + .5 * baths
f2(beds, baths) = 1 + .9 * beds + .1 * baths
f3(beds, baths) = 1 + .1 * beds + .9 * baths
Oracle | Always predicts correct answer

Hypothesis Space

f3

f1   f2

Oracle

F

All Possible Functions
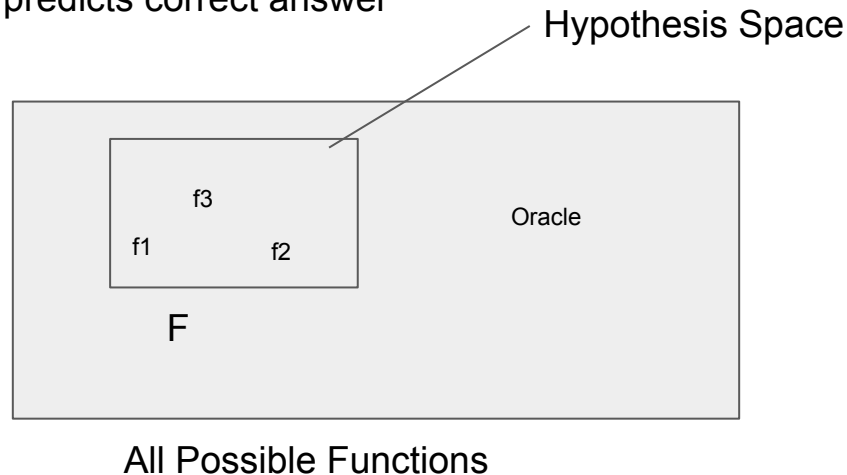
The Unreasonable Effectiveness of Data

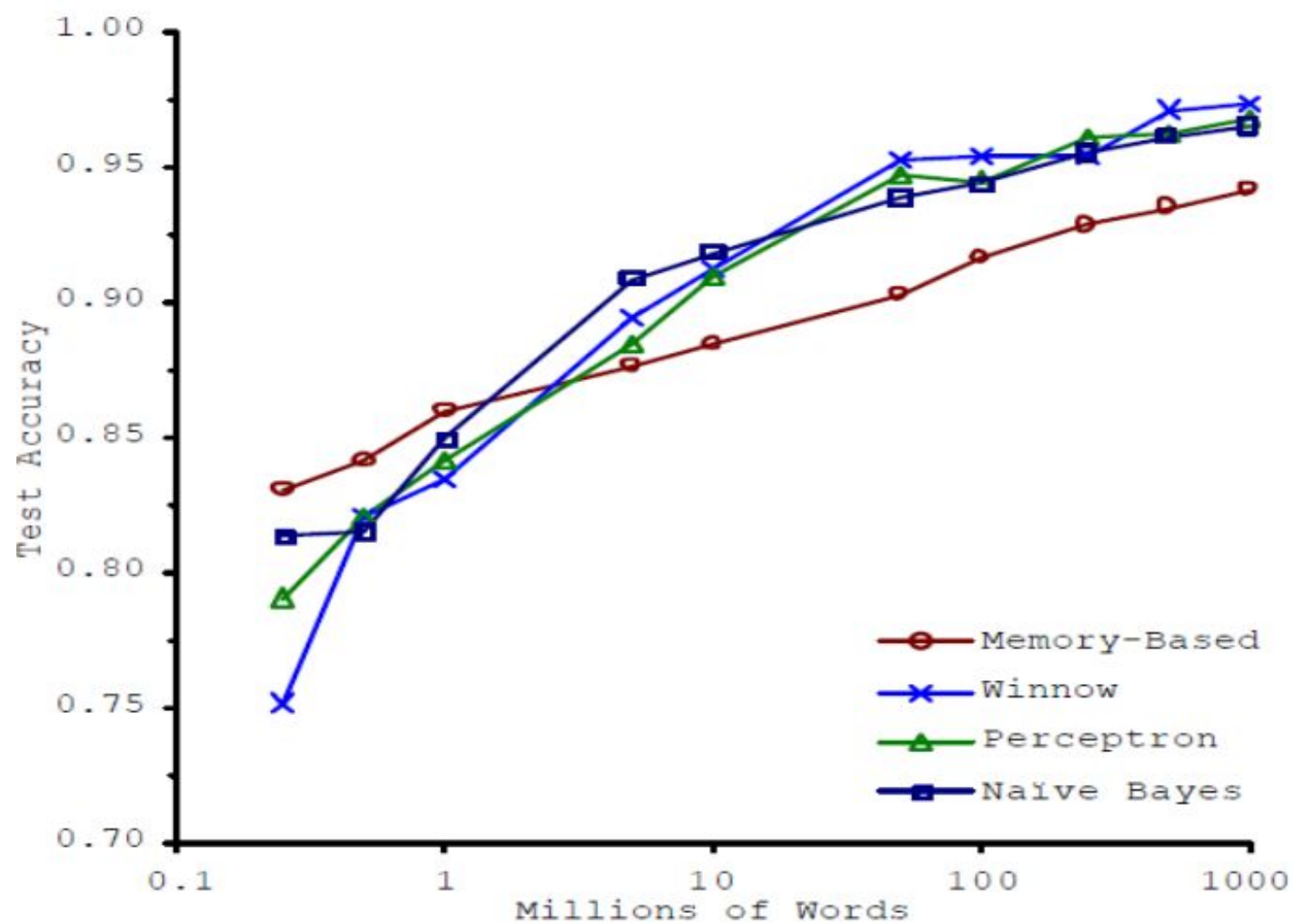**Alon Halevy, Peter Norvig, and Fernando Pereira,** *Google*

ugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"[1] examines why so much of physics can be neatly explained with simple mathematical formulas such as $f = ma$ or $e = mc^2$. Meanwhile, sciences that involve human beings rather than elementary particles have proven more resistant to elegant mathematics. Economists suffer from physics envy over their inability to neatly model human behavior. An informal, incomplete grammar of the English language runs over 1,700 pages.[2] Perhaps when it comes to natural language processing and related fields, we're doomed to complex theories that will never have the elegance of physics equations. But if that's so, we should stop acting as if our goal is to author extremely elegant theories, and instead embrace complexity and make use of the best ally we have: the unreasonable effectiveness of data.

behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

Learning from Text at Web Scale

The biggest successes in natural-language-related machine learning have been statistical speech recognition and statistical machine translation. The reason for these successes is not that these tasks are easier than other tasks; they are in fact much harder than tasks such as document classification that extract just a few bits of information from each document. The reason is that translation is a natural task routinely done every day for a real human need (think of the operations of the European Union or of news agencies). The same is true of speech transcription (think of closed-caption broadcasts). In other words, a large training set of the input-output behavior that we seek to automate is available to us *in the wild*. In contrast, traditional natural language

Final Thoughts?