CS 124/LINGUIST 180
From Languages to Information

Dan Jurafsky

Stanford University

Introduction and Course Overview

From Languages to Information

Automatically extracting meaning and structure from:

- Human language text and speech (news, social media, etc.)
- Social networks
- Genome sequences

Interacting with humans via language

- Dialog systems/Chatbots
- Question Answering
- Recommendation Systems

Commercial World

Social World

Disaster Relief

Chatbots for Mental Health

Improve Police-Community relations via Body-Cameras

1. Extracting information from language

Information Retrieval

6,586,013,574 web searches every day (by one estimate)

Text-based information retrieval is thus likely the most frequently used piece of software in the world

How does it work? Can you build an IR engine?

Programming Assignment 4: Search!

Text Classification: Disaster Response

Haiti Earthquake 2010

Classifying SMS messages

Mwen thomassin 32 nan pyron mwen ta renmen jwen yon ti dlo gras a dieu bo lakay mwen anfom se sel dlo nou bezwen

I am in Thomassin number 32, in the area named Pyron. I would like to have some water. Thank God we are fine, but we desperately need water.

Programming Assignment 2: Triage!

Extracting Sentiment and Social Meaning

Lots of meaning is in **connotation**

> "connotation: an idea or feeling that a word invokes in addition to its literal or primary meaning."

Extracting connotation is generally called **sentiment analysis**

Programming Assignment 3: Thumbs up!

Sentiment Analysis

Extracting Social Meaning from Language

Uncertainty (students in tutoring)

Annoyance

- Talking to a computer travel agent:

Anger (police-community interaction)

Deception

Emotion

Intoxication

Sentiment in Restaurant Reviews

Dan Jurafsky, Victor Chahuneau, Bryan R. Routledge, and Noah A. Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. First Monday 19:4

900,000 Yelp reviews online

A very bad (one-star) review:

The bartender... absolutely horrible... we waited 10 min before we even got her attention... and then we had to wait 45 - FORTY FIVE! - minutes for our entrees... stalk the waitress to get the cheque... she didn't make eye contact or even break her stride to wait for a response ...

What is the language of bad reviews?

Negative sentiment language

horrible awful terrible bad disgusting

Past narratives about people

waited, didn’t, was

he, she, his, her,

manager, customer, waitress, waiter

Frequent mentions of we and us

... **we** were ignored until **we** flagged down a waiter to get **our** waitress ...

Other narratives with this language

A genre using:

Past tense, we/us, negative, people narratives

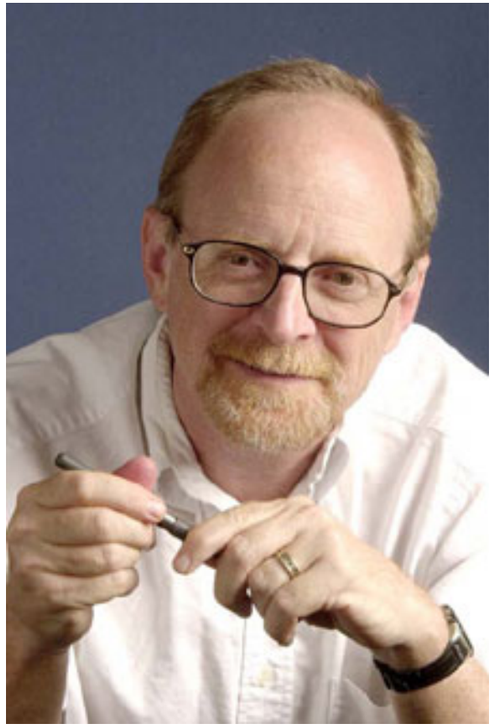
Texts written by **people suffering trauma**

- James Pennebaker lab at UT Austin
- Past tense is used for "distancing"
- Use of “we”: seeking solace in community

1-star reviews are trauma narratives!

The lesson of reviews:

It’s all about personal interaction



What about positive reviews?

Sex, Drugs, and Dessert

addicted to pepper shooters

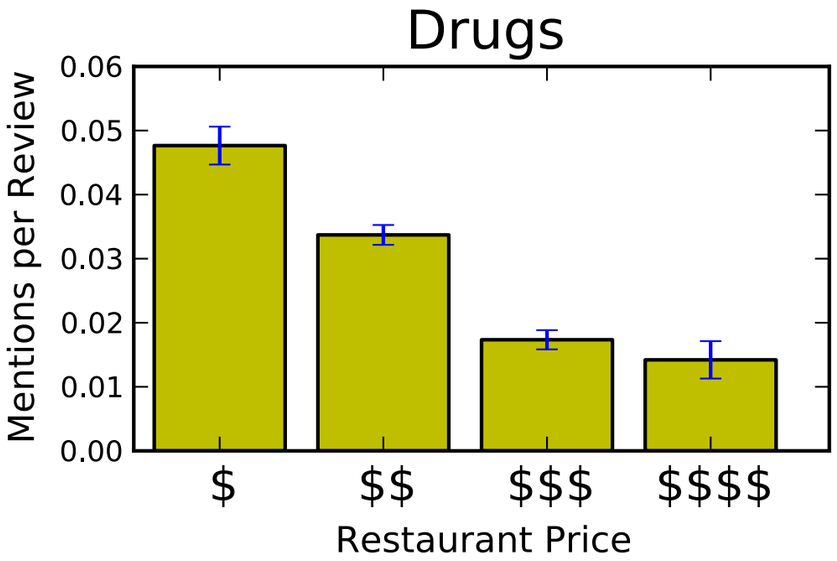
garlic noodles… my *drug of* choice

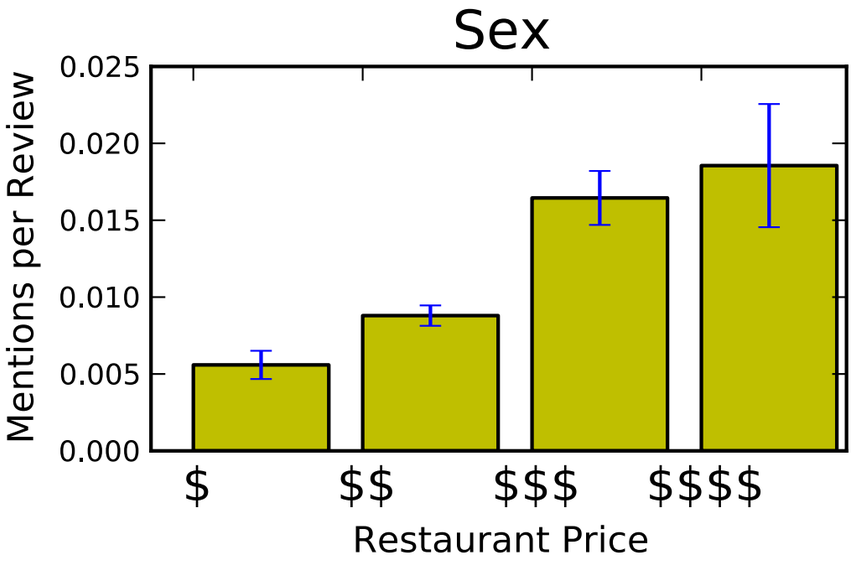
the fries are *like crack*

orgasmic pastry

sexy food

seductively seared fois gras





Computational Biology: Comparing Sequences

AGGCTATCACCTGACCTCCAGGCCGATGCCC

TAGCTATCACGACCGCGGTCGATTTGCCCGAC

```
-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
 | |  | | | | | | |  | | | | x |  | |  | | |  | | | | |
TAG-CTATCAC--GACCGC--GGTCGATTTGCCCGAC
```

Sequence comparison is key to

- Finding genes
- Determining function
- Uncovering evolutionary processes

This is also how spell checkers work!

We'll learn: edit distance algorithms (Quiz 1)

Social Networks

The network formed by your friends or other relations offline or online

- Can we compute properties of these networks?
- Extract information from them?

High school dating

What is the structure of social relations?

Imagine a graph of high school

- people are nodes
- links are romantic relationships

What will the shape of this graph be?

A densely connected graph?

A line?

A cycle?

Peter S. Bearman, James Moody and Katherine Stovel Chains of affection: The structure of adolescent romantic and sexual networks *American Journal of Sociology* 110 44-91 (2004)
Image drawn by Mark Newman

Help improve Police-Community Interaction (week 9)

Problems:

- A flood of viral videos show inappropriate officer use of force
- Black Americans report more negative interactions with police

Could natural language processing help?

- Quantify police-community interactions using body-worn cameras?
- Detect the potential for escalation?
- Help develop officer training?
- Reduce the chances of violence?

2. Interacting with humans via language

Personal Assistants

Siri

amazon alexa

Google Now

Listening..

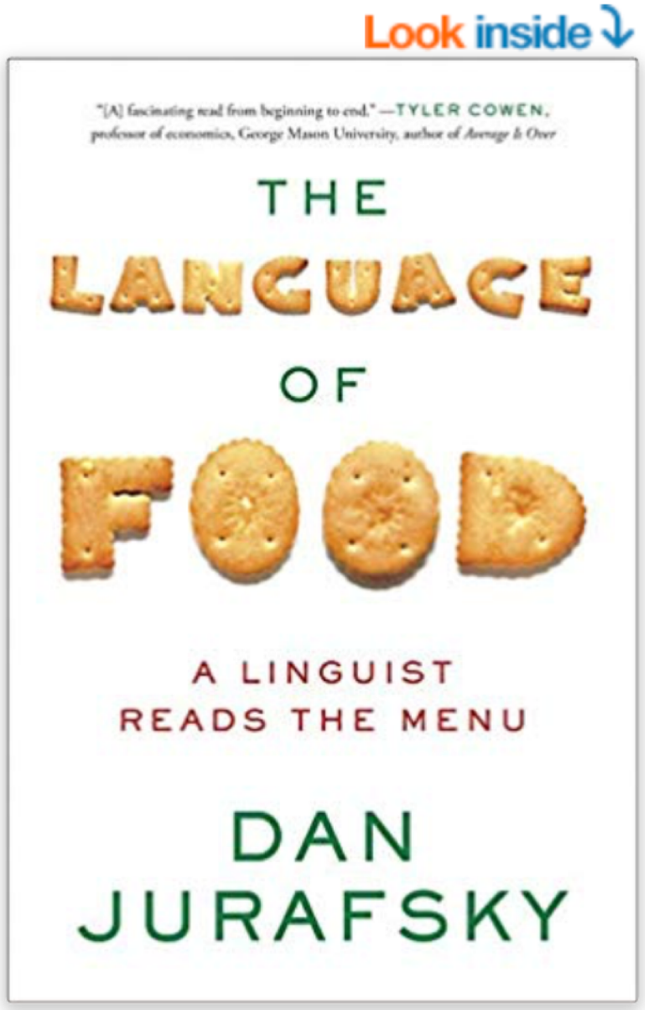
Facebook M

A personal assistant inside Messenger

Question Answering 10 years ago: IBM's Watson

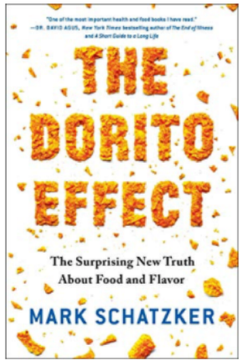
Recommendation Engines: The Good

If you bought….

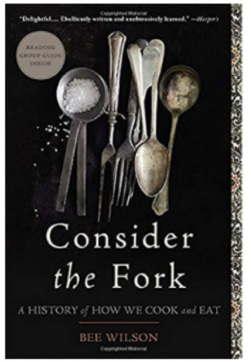


Customers who bought this item also bought

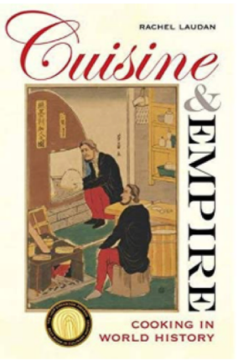
First Bite: How We Learn to Eat
› Bee Wilson
4.5 out of 5 stars 46
Paperback
$11.37 prime



The Dorito Effect: The Surprising New Truth About Food and Flavor
› Mark Schatzker
4.5 out of 5 stars 193
Paperback
$9.48 prime



Consider the Fork: A History of How We Cook and Eat
› Bee Wilson
4.0 out of 5 stars 253
Paperback
$15.65 prime



Cuisine and Empire: Cooking in World History (California Studies in…
› Rachel Laudan
4.0 out of 5 stars 35
Paperback
$16.20 prime

The dark side: YouTube Radicalization

Why is language interpretation hard?

Ambiguity

Resolving ambiguity is hard

Ambiguity

Find at least 6 meanings of this sentence:

```
I made her duck
```

Ambiguity

Find at least 6 meanings of this sentence:

```
I made her duck
```

I cooked waterfowl for her benefit (to eat)

I cooked waterfowl belonging to her

I created the (plaster?) waterfowl she owns

I caused her to quickly lower her head or body

I recognized the true identity of her spy waterfowl

I waved my magic wand and turned her into undifferentiated waterfowl

Ambiguity is Pervasive

I caused her to quickly lower her head or body

Part of speech: “duck” can be a Noun or Verb

I cooked waterfowl belonging to her.

Part of speech:

“her” is possive pronoun (“of her”)

“her” is dative pronoun (“for her”)

I made the (plaster) duck statue she owns

Word Meaning : “make” can mean “create” or “cook”

Ambiguity is Pervasive

Grammar: `make` can be:

Transitive: (verb has a noun direct object)

I cooked [waterfowl belonging to her]

Ditransitive: (verb has 2 noun objects)

I made [her] (into) [undifferentiated waterfowl]

Action-transitive (verb has a direct object + verb)

I caused [her] [to move her body]

Ambiguity is Pervasive: Phonetics!!!!!

Aye mate, her duck

I mate or duck

I'm eight or duck

Eye maid; her duck

I maid her duck

I'm aid her duck

I mate her duck

I'm ate her duck

I'm ate or duck

I mate or duck

More difficulties:
Non-standard language, emojis, hashtags, names

chowdownwithchan #crab and #pork #xiaolongbao at @dintaifungusa... where else? 😂🤷‍♀️ Note the cute little crab indicator in the 2nd pic 🦀💕

Making progress on this problem…

The task is difficult! What tools do we need?

- Knowledge about language and the world
- A way to combine knowledge sources

How we generally do this:

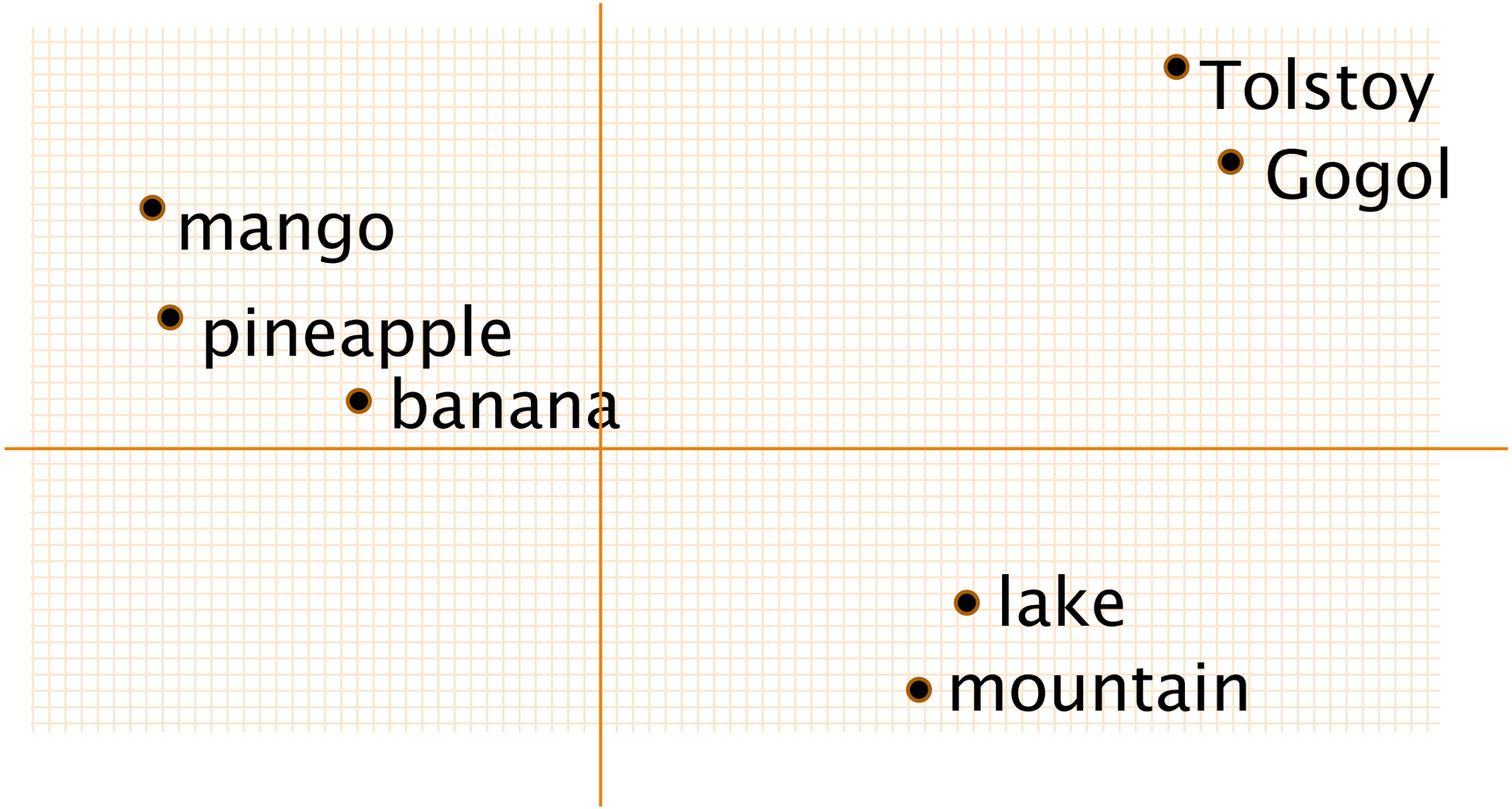
- neural and other machine learning models built from language data

Models and Tools

- Regular Expressions
- Edit distance and alignment
- Word embeddings
  - neural models of word meaning
- Language models
- Machine Learning classifiers
  - Naïve Bayes
  - Logistic Regression
  - Neural Networks
- Network algorithms
  - PageRank
- Recommendation algorithms
  - Collaborative filtering
- Linguistic tools
  - Sentiment lexicons
  - Emotion lexicons
- The GUS chatbot algorithm

Neural word embeddings

A word's meaning is a point in 300-dimensional space
A 2-D visualization:



Embeddings are the core of NLP

Core technology for any NLP task (question answering, machine translation, information retrieval, etc)

- Finding synonyms for words
- Deciding if two sentences have similar meaning

How to learn these "embeddings"?

Push words **together in space** they occur **together in text**

Read millions of words.
When you see:
Banana, mango, or pineapple are all delicious…

Move **banana** closer to **mango**
Move **banana** further from **Tolstoy**

Problem: Embeddings reflect cultural bias!

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings." In *NeurIPS 2016*, pp. 4349-4357.

Ask “Paris : France :: Tokyo : x”

- x = Japan

Ask “father : doctor :: mother : x”

- x = nurse

Ask “man : computer programmer :: woman : x”

- x = homemaker

What can we do about this problem? Week 5!

Logistics: Instructor

Instructor: Dan Jurafsky (he/him)

Professor in CS and Linguistics

Department Chair of Linguistics

Office Hours:

- Thurs after class 4:30-6:00
- Margaret Jacks Hall 117

*How to pronounce my name

Picture by Ross Petukhov

Course logistics

TAs: Urvashi Khandelwal (head TA)

Hancheng Cao	Chuma Kabaghe	Krishna Patel
Jennie Chen	Anjini Karthik	Andrew Tan
Cindy Jiang	Bryan Kim	Lauren Zhu

Time: TuTh 3:00-4:20, Hewlett 200

cs124.stanford.edu

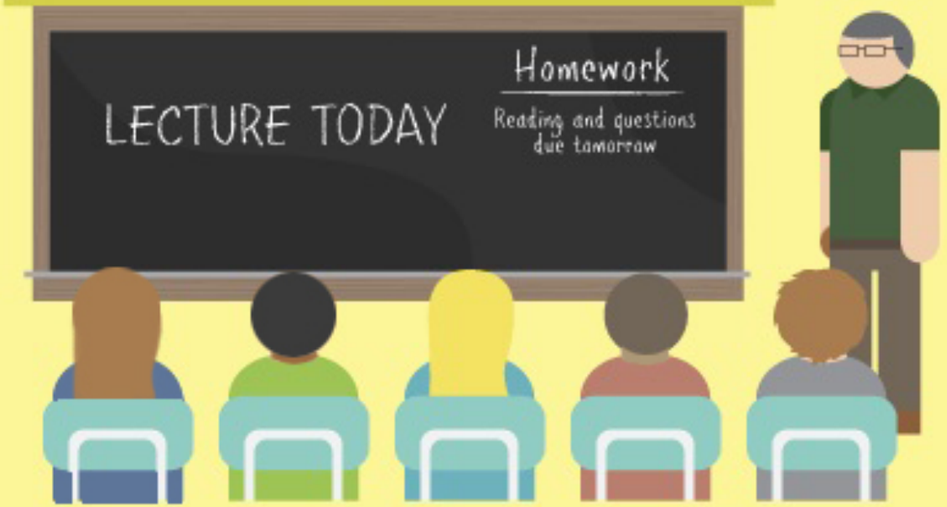
Evidence Based Pedagogy!

WHAT IS THE FLIPPED CLASSROOM?

The flipped classroom inverts traditional teaching methods, delivering instruction online outside of class and moving "homework" into the classroom.

THE INVERSION

The Traditional Classroom
Teacher's Role: Sage on the Stage



The Flipped Classroom
Teacher's Role: Guide on the Side

From (defunct) www.knewton.com/flipped-classroom/

Why the flipped classroom (1)

Mastery learning: Learn until you master

Benjamin Bloom, 1968

Bloom's mastery learning

Personalized, **goal-driven practice**, driven by **feedback**

1. Watch (and re-watch) lectures at your own pace and learn when it's best for you
2. Videos have embedded miniquizzes. If you get it wrong, it gives you feedback about why you misunderstood.
3. You have 2 chances at each weekly Tuesday Quiz, so you can go back to the lecture and retake them.
4. With programming assignments you can see your performance on the training and dev set to see what you're doing wrong!

Why the videos have embedded quizzes: “summative” vs “formative” assessment

Summative assessment

- Final exams: goal is grading

Formative assessment

- Along the way: goal is for **you** to find out what you don’t know so you can learn

Why the flipped classroom (2)

Attention span: everyone spaces out during long lectures

- Middendorf and Kalish, 1995, Johnstone and Percival 1976, Burns 1985

"the class started 1:00. The student sitting in front of me took copious notes until 1:20. Then he just nodded off… motionless, with eyes shut for about a minute and a half, pen still poised. Then he awoke and continued his rapid note-taking as if he hadn't missed a beat."

Student remembered only the first 15-20 minutes

Why the flipped classroom (3)

Active learning: Be in charge of your learning

- Obviously most important: programming assignments
- Active learning ("constructivism"), learning by doing

Collaborative learning: Learn from each other

- Use class time for group activities, worked problems
- "Small group active learning"

cs124: Semi-flipped classroom

1. **Lectures on video:** I expect you to:
   - Watch video lectures (~90 min/week)
   - Some people watch it speeded up
2. **Live lectures:**
   - 7 lectures and 1 group session are required
     - on final exam, no videos
3. **In-class group sessions (“active learning”)**
   - Optional but strongly recommended

Logistics More Specifically

Online Video Lectures w/embedded non-graded questions (before class)

20 pages of reading a week (up to you when to read)

Weekly online Review Quizzes (Tue of following week)

6 Python homeworks (Fri of following week)

Final Exam (Thursday March 19 12:15-3:15)

Class sessions: All encouraged; **8 live classes required**

Learning Goals

At the end of this course, you will be able to:

Learning goals

Write efficient regular expressions to solve any kind of text-based extraction task

Learning goals

Apply the edit distance algorithm to all sorts of text sequence problems

Learning goals

Build a supervised classifier to solve problems like sentiment classification

Learning goals

Build a search engine

Learning goals

Build a recommendation engine

Learning goals

Build a computational model of word meaning (using lexicons and neural word embeddings)

Learning goals

Build a chatbot

Learning goals

Understand and implement PageRank

This class is the undergrad intro to:

Win 2020: cs224N Natural Language Processing w/Deep Learning

Win 2020: cs246 Mining Massive Data Sets

Spr 2020: cs222U Natural Language Understanding

Spr 2020: cs224S Spoken Language Processing

Spr 2020: cs346 Ethical and Social Issues in NLP

Spr 2021: cs276 Information Retrieval and Web Search

Aut 2021: cs224W Machine Learning with Graphs

Aut 2021: cs221 Artificial Intelligence

Syllabus

http://web.stanford.edu/class/cs124

Coming up next class (Thursday)

Unix for poets

- grep
- sort

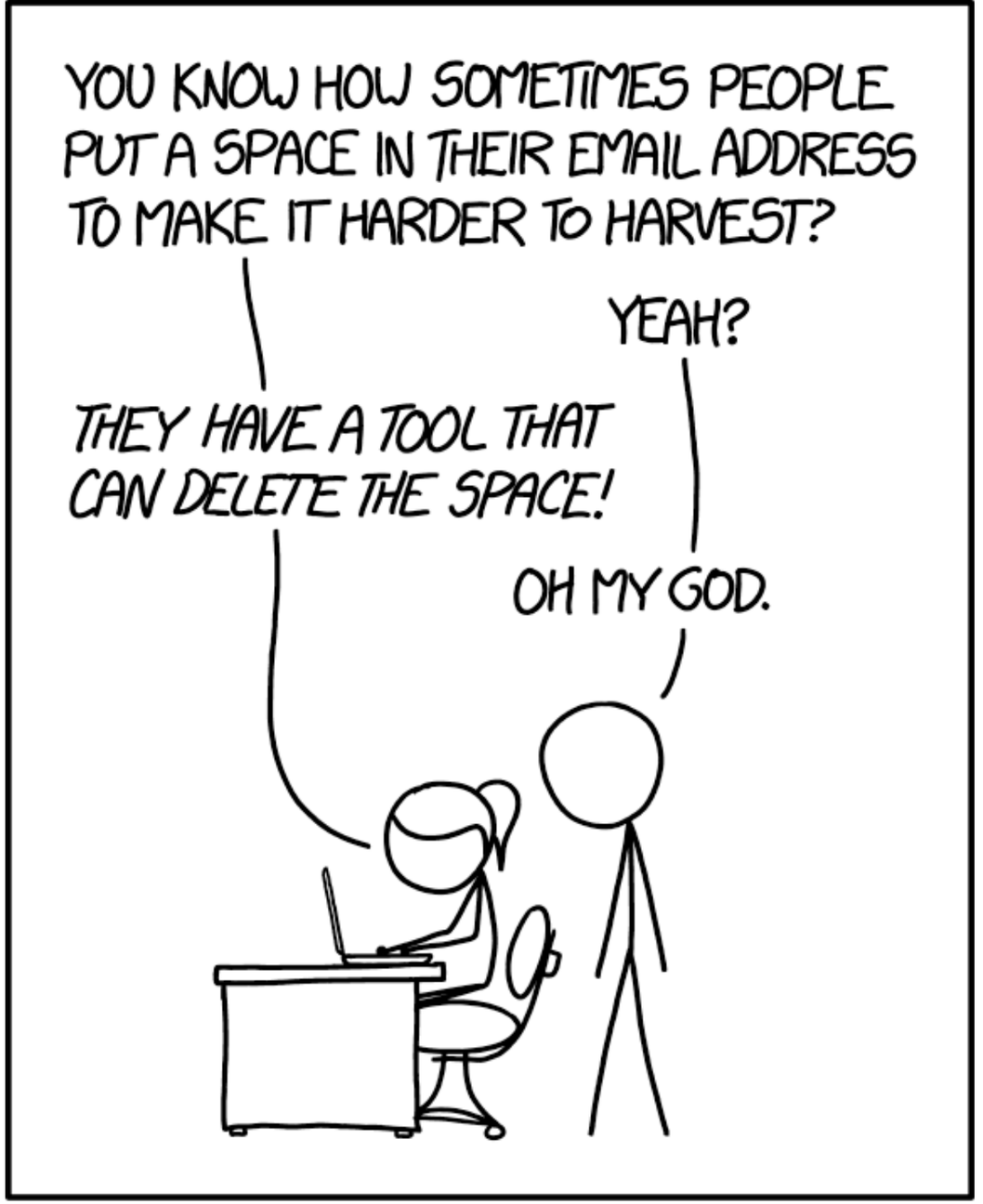
PA1: Spam Lord!

Write regular expressions to spread evil throughout the galaxy!

By extracting email addresses and phone numbers from the web!

```
jur a fs ky at st anford dot e d u
```

Goes live Friday!



LESS-DRAMATIC REVELATIONS FROM THE CIA HACKING DUMP

Action Items Before Thursday!

1) Read the syllabus webpage at cs124.stanford.edu

2) Sign up for piazza

3) Watch the first half of this week's videos ("Basic Text Processing") on Canvas before class!

4) Download this file to your laptop

http://cs124.stanford.edu/nyt_200811.txt.gz