Project presentations and deadlines

- Project presentations. In class on 3/15 and 3/17
  - Upload 5 minute video by Sunday 3/14 11:59pm. No late days.
  - In-class we play your video, allow 5 minutes for questions
  - 2 rooms in parallel, Expect 5:30 – 7:15pm. Please attend as much as you can to see classmates’ projects
  - Sign up for your presentation time on Gdrive sheet
- Public YouTube video of all presentations when final papers are ready. You can opt out.
- Final project paper *deadline extended*. No late days Monday 3/22 11:59pm
- HW4 *deadline extended*. Friday 3/12 11:59pm
Supervised ML as an analysis tool

1. Define your task / hypothesis (e.g. detecting alcohol intoxication in speech)

2. Collect/access data and annotations (y) for your task. Ensure data is representative for your problem

3. Define and test your modeling approach and inputs (x)
   1. Optimize to find x,f() to improve f(x)=y for your data

4. Analyze results, feature importance, model weights etc.
Supervised ML as an analysis tool

- Examples we will explore today:
  - Flirtation / interpersonal stance
  - Intoxication
- Framework is highly flexible. Requires careful decisions about datasets and what conclusions to draw
- Use cases:
  - Accurately recognize $y$ in a live/ongoing system
    - Validate $x$ is reasonable/ethical. Build best possible $f(x)$. Favor accuracy
  - Infer findings/relationships predictive of $y$
    - Careful analysis of $x, f(x)$. Favor interpretability
Interpersonal Stance: Our Goals

- Friendliness
- Assertiveness
- Flirtation
- Awkwardness
Methodology

- Given speech and text from a conversation
- Can we tell if a speaker is
  - Awkward?
  - Flirtatious?
  - Friendly?
- Dataset:
  - 1000 4-minute “speed-dates”
  - Each subject rated their partner for these styles
  - The following segment has been lightly signal-processed:

(Jurafsky, Ranganath & McFarland. 2009)
(Ranganath, Jurafsky & McFarland. 2009)
**Speed dating** *noun*

*speed dating [uncountable]*

an event at which you meet and talk to a lot of different people for only a few minutes at a time. People do this in order to try to meet someone and have a romantic relationship.


Extracting social meaning

- **Stance**
  - Friendly, flirt, awkward, assertive

- **Social Bond**
  - Clicking or Connection
  - Romantic Interest

- **946 4-minute dates**
  - ~800K words, hand-transcribed
  - ~60 hours, from shoulder sash recorders
  - 3 events, 20x20=400 dates x 3
  - Date perceptions, demographics, preferences
Data annotation

- Each speaker wore a microphone
- So each date had two recordings
- The wavefile from each speaker was manually segmented into 4-minute dates
- Professional transcription service produced:
  - words, laughter, disfluencies
  - timestamps for turn beginning and end (1 second)
    - for 10% of the dates, timestamp at 0.1 second granularity
  - using both recordings
Study 1:
What we attempted to predict

• Conversational style:
  • How often did they behave in the following ways on this date?
    • On a scale of 1-10 (1=never, 10=constantly)

awkward
friendly
flirtatious
assertive
Features

• **Prosodic**
  - pitch (min, mean, max, std)
  - intensity (min, max, mean, std)
  - duration of turn
  - rate of speech (words per second)

• **Lexical**
  - negation words (don’t, didn’t, won’t, can’t, not, never)
  - hedges (kind of, sort of, probably, I don’t know)
  - personal pronouns (I, you, we, us)

• **Dialog**
  - questions
  - backchannels (“uh-huh”, “yeah”)
  - appreciations (“Wow!”, “That’s great!”)
  - sympathy (“That’s awful!” “Oh, that sucks!”)
Engineering studies
16 binary classifiers

- Female ± Awkward, Male ± Awkward,
- Female ± Friendly, Male ± Friendly,
- Female ± Flirtatious, Male ± Flirtatious,
- Female ± Assertive, Male ± Assertive

- Each study run twice, on:
  - self-assessed
  - alter-assessed

- Multiple classifier experiments
  - L1-regularized logistic regression
  - SVM w/RBF kernel
Test set

- For each of the 16 experiments
  - Sort all 946 dates
  - Choose top 10% as positive class
  - Choose bottom 10% as negative class
  - Ignore 80% of dates in the middle!
- 5-fold cross-validation within this small training and test set

- Goal: distinguishing social interactants who are reported to exhibit (or not exhibit) clear social intentions or styles
Results using SVM Classifier

Using my speech to predict what my date says about me

<table>
<thead>
<tr>
<th></th>
<th>Male speaker</th>
<th>Female speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flirting</td>
<td>65%</td>
<td>78%</td>
</tr>
<tr>
<td>Friendly</td>
<td>71</td>
<td>64</td>
</tr>
<tr>
<td>Awkward</td>
<td>67</td>
<td>67</td>
</tr>
<tr>
<td>Assertive</td>
<td>65</td>
<td>69</td>
</tr>
</tbody>
</table>
Results using SVM Classifier

- Using my speech to predict what I say about myself

<table>
<thead>
<tr>
<th></th>
<th>Male speaker</th>
<th>Female speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flirting</td>
<td>66%</td>
<td>74%</td>
</tr>
<tr>
<td>Friendly</td>
<td>76</td>
<td>71</td>
</tr>
<tr>
<td>Awkward</td>
<td>63</td>
<td>67</td>
</tr>
<tr>
<td>Assertive</td>
<td>73</td>
<td>64</td>
</tr>
</tbody>
</table>
What do flirters do?

- Women when flirting:
  - raise pitch ceiling
  - talk faster
  - say “I” and “like”, use more hedges
  - laugh at themselves

- Men when flirting:
  - raise their pitch floor
  - laugh at their date (teasing?)
  - say “you”
  - don’t use words related to academics
  - say “um”, “I mean”, “you know”
Unlikely words for male flirting

- academia
- interview
- teacher
- phd
- advisor
- lab
- research
- management
- finish
Assertive

• Assertive men
  • talk more
  • use more negative emotion
  • lower their pitch floor
  • use more agreements and appreciations
  • use more “um”, “you”
  • use less negation

• Assertive women:
  • use more negation ("no", “didn’t”, “don’t”)
  • talk about academics
  • are less sympathetic
  • accommodate more (content words)
  • use more “I” and “I mean”
  • use less negative emotion
What makes an awkward conversationalist?

• Awkward people:
  • use more hedges
  • ask more questions

• Awkward men
  • don’t talk about academics
  • do swear or use negative emotion

• Awkward women:
  • do talk about academics
  • talk more, and talk faster
  • don’t laugh at their date
  • don’t use “I”
(Actionable) conclusions

- How to date:
- Don’t talk about your advisor
- Focus on the empowered party
- Flirting women raise pitch ceiling – flirting men raise pitch floor
What makes someone seem friendly? “Collaborative conversational style”


- **Friendly people:**
  - laugh at themselves
  - don’t use negative emotions

- **Friendly men**
  - are sympathetic and agree more often
  - don’t interrupt
  - don’t use hedges

- **Friendly women:**
  - higher max pitch
  - laugh at their date
Intoxication
Hollien et al 2001

- **Methods:**
  - 35 young adults, 19 males, 16 females
  - given series of doses of alcohol
  - speech collected at 4 BAC stages
    - Rainbow passage
    - difficult words (buttercup, shapupie)
    - extemp speech (“Tell us about your favorite TV program”)
    - head-mounted mics
- **Investigated:**
  - F0 mean and variance
  - duration/rate of speech
  - intensity
  - disfluencies
Hollien et al 2001 Results: Duration

<table>
<thead>
<tr>
<th>Group</th>
<th>Level of intoxication (BrAC)</th>
<th>Shift (0.00–0.12)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td><strong>Men</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (s)</td>
<td>25.3</td>
<td>25.8</td>
</tr>
<tr>
<td>S.D. (s)</td>
<td>2.9</td>
<td>2.5</td>
</tr>
<tr>
<td><strong>Women</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (s)</td>
<td>25.1</td>
<td>25.5</td>
</tr>
<tr>
<td>S.D. (s)</td>
<td>2.2</td>
<td>2.2</td>
</tr>
</tbody>
</table>
Hollien et al 2001 Results: Disfluencies

<table>
<thead>
<tr>
<th>Subjects</th>
<th>N</th>
<th>Experimental condition (BrAC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>Males</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>3.2</td>
</tr>
<tr>
<td>SD</td>
<td></td>
<td>2.0</td>
</tr>
<tr>
<td>Females</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>2.2</td>
</tr>
<tr>
<td>SD</td>
<td></td>
<td>1.7</td>
</tr>
<tr>
<td>Mean</td>
<td>35</td>
<td>2.7</td>
</tr>
</tbody>
</table>
Hollien et al
2001 Results: Magnitudes
A famous case study

The Exxon Valdez oil spill occurred in Prince William Sound, Alaska, on March 24, 1989, when the Exxon Valdez, an oil tanker bound for Long Beach, California, struck Prince William Sound's Bligh Reef and spilled 260,000 to 750,000 barrels (41,000 to 119,000 m³) of crude oil.[1][2] It is considered to be one of the most devastating human-caused environmental disasters.[3] As
Was Captain Hazelwood drunk?

- Not clear if this is relevant, since seems like other questionable corporate things were going on:
  - he was asleep below deck
  - The third mate was in charge of the wheelhouse
  - the ship’s radar was broken
- But is a well-studied case
Johnson et al examined 3 kinds of cues

- Segmental Effects (phoneme, syllable, word level)
- Disfluencies
- Suprasegmental Effects (stress, intonation, etc.)
Keith Johnsons /s/ and /ʃ/

Fig. 1. Power spectra of /s/ (a) and /ʃ/ (b) produced by K. J. in a quiet recording booth with recording equipment responsive up to 5,000 Hz.

e.g. “sun” vs “shun”
"she's"

"shout"

Fig. 2. Power spectra of /ʃ/ produced by Captain Hazelwood in the words she's and shout recorded 33 h before the accident. Each spectrum is paired with a spectrum of the background noise from a nearby open-mike pause.
33 Hrs before

1 Hr before

Immediately after

1 Hr after

9 Hrs after
Fig. 2. Power spectra of /ʃ/ produced by Captain Hazelwood in the words she's and shout recorded 33 h before the accident. Each spectrum is paired with a spectrum of the background noise from a nearby open-mike pause.
Duration

Segment Durations of "Exxon Valdez"

33 Hrs before
1 Hr before
Immediately after
1 Hr after
9 Hrs after

Duration (ms)
Table 3. Summary of phenomena found in the analysis of the NTSB tape (numbers in parentheses indicate the time of recording)

<table>
<thead>
<tr>
<th>Gross effects</th>
<th>revisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-1) Exxon Ba, uh Exxon Valdez</td>
<td>(-1) departed disembarked</td>
</tr>
<tr>
<td>(-1) I, we’ll</td>
<td>(-1) columbia gla, columbia bay</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Segmental effects</th>
<th>misarticulation of /r/ and /l/</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0) northerly, little, drizzle,</td>
<td>visibility</td>
</tr>
<tr>
<td>visibility</td>
<td>(/s/ becomes /ʃ/ (fig. 3)</td>
</tr>
<tr>
<td>final devoicing (e.g. /z/ → /s/)</td>
<td>(-1,0,+1) Valdez → Valdes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Suprasegmental effects</th>
<th>reduced speaking rate (fig. 4, 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean change in pitch range</td>
<td>(talker-dependent, fig. 6)</td>
</tr>
<tr>
<td>(talker-dependent, fig. 6)</td>
<td>increased F₀ jitter (fig. 6)</td>
</tr>
</tbody>
</table>
Problems

- If intoxicated speech, why wasn’t s pronounced as sh 1 hour before?
- Other kinds of speaker state could cause drop in F0, slower speech, and disfluencies?
  - Stress, just having woken up, trauma....
Automatic Classification

- Use of prosodic speech characteristics for automated detection of alcohol intoxication
  Michael Levit, Richard Huber, Anton Batliner, Elmar Noeth

- Break utterance into phrases automatically, based on
  - fundamental frequency (where possible);
  - zero-crossing rate

![Waveform with labeled phrases]

- PhU1
- PhU2
- PhU3
Then use 4 classes of features

- **Prosodic**
  - F0 max, F0 min, energy max, energy min, pause length
- **Duration of voiced regions, unvoiced regions, etc.**
- **Jitter and shimmer**
  - jitter is variation in pitch
  - shimmer is variation in energy
- **Average cepstrum and cepstral slope**
Methods

- Alcoholized speech samples collected at the Police Academy of Hessen, Germany
  - 120 readings (87 minutes) of a fable
  - 33 male speakers
  - BAC between 0 and .24/mille

<table>
<thead>
<tr>
<th>Alcohol Blood Level</th>
<th>0.0</th>
<th>&lt; 0.4</th>
<th>&lt; 0.8</th>
<th>&lt; 1.2</th>
<th>&lt; 1.6</th>
<th>&lt; 2.0</th>
<th>&lt; 2.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recordings</td>
<td>32</td>
<td>20</td>
<td>20</td>
<td>18</td>
<td>20</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>

- Binary task: above or below 0.8/mille
- leave-one-out cross-validation
- neural net classifier
Results of Levit et al.

- Used dev set to find best classifier
- This suggested two feature classes:
  - Prosodic features
  - Jitter/shimmer
- Results with this classifier
  - 62% phrase-accuracy
  - 69% for the whole speech sample
    - voting of the phrases
Alcohol Language Corpus

Florian Schiel et al 2009, 2010

- [http://www.bas.uni-muenchen.de/forschung/Bas/BasALCeng.html](http://www.bas.uni-muenchen.de/forschung/Bas/BasALCeng.html)

- 162 speakers (77 female, 85 male)
  - recorded in a car (sometimes with engine running)
  - command and control speech (“turn off the radio”)
  - spontaneous dialogue, monologue, question answering
  - read speech
  - counts of disfluencies, etc

- sample, drunk:
- sample, sober:
Automatic detection in ALC: Paralinguistic Challenge 2011

- Human: 66-72%  (Schiel 2011, Ultes, Schmitt, Minker 2011)
- Machine: roughly 65%-70%
- Example features from winning system:


- Prosody (f0, duration, energy, jitter, shimmer)
- Spectral (MFCC, MFB log-energy, formants)
- Computed over whole utterance and small windows
- normalized phoneme duration
- iterative speaker normalizeaton
Supervised ML as an analysis tool

1. Define your task / hypothesis (e.g. detecting alcohol intoxication in speech)

2. Collect/access data and annotations (y) for your task. Ensure data is representative for your problem

3. Define and test your modeling approach and inputs (x)
   1. Optimize to find $x, f()$ to improve $f(x) = y$ for your data

4. Analyze results, feature importance, model weights etc.
Social Signal Processing
= Affect/Emotion Detection

- Detecting frustration of callers to a help line
- Detecting stress in drivers or pilots
- Detecting depression, intoxication
- Detecting interest, certainty, confusion in on-line tutors
  - Pacing/Positive feedback
- Hot spots in meeting summarizers/browsers
Questions in Emotion Recognition

- How do we know what emotional speech is?
  - Acted speech vs. natural (hand labeled) corpora
- What can we classify?
  - Distinguish among multiple ‘classic’ emotions
  - Distinguish
    - Valence: is it positive or negative?
    - Activation: how strongly is it felt? (sad/despair)
- What features best predict emotions?
- What techniques best to use in classification?

Slide from Julia Hirschberg