Truth or Lie? Spoken Indicators of Deception in Speech

Julia Hirschberg

Columbia University

EMNLP 2018

Collaborators

- ***Sarah Ita Levitan***
- ***Michelle Levine***
- ***Guozhen An (CUNY)***
- ***Andrew Rosenberg (CUNY, now Google)***
- Gideon Mendels
- Nishmar Cestero
- Rivka Levitan
- Xi Chen
- Kara Schechtman
- Angel Maredia
- Jessica Xiang
- Bingyan Hu
- William Wang
- James Shin
- Yocheved Levitan
- Mandi Wang
- Kai-Zhan Lee
- Zoe Baker-Peng
- Ivy Chen
- Meredith Cox
- Leighanne Hsu
- Yvonne Missry
- Gauri Narayan
- Molly Scott
- Jennifer Senior
- Grace Ulinksi
- Mukund Yelahanka Raghuprasad

Columbia Speech Lab

Deceptive Speech

- ***Deliberate choice to mislead***
  - ***Without*** prior notification
  - To gain some ***advantage*** or to avoid some ***penalty***
- ***Deception does not include***:
  - Self-deception, delusion, pathological behavior
  - Theater
  - Falsehoods due to ignorance/error
- ***Everyday (White) Lies*** very hard to detect
- But ***Serious Lies may*** be easier…

Why might Serious Lies be easier to identify?

- ***Hypotheses*** in research and among practitioners:
  - Our ***cognitive load*** is increased when we lie because…
    - We must keep our story straight
    - We must remember what we ***have*** and ***have not*** said
  - Our ***fear of detection*** is increased if…
    - We believe our target is difficult to fool
    - Stakes are high: serious rewards and/or punishments
- ***All this makes it hard for us to control potential indicators of deception***

Humans are very poor at Recognizing these Cues (Aamodt & Mitchell 2004 Meta-Study)

()

Group	#Studies	#Subjects	Accuracy %
Criminals	*1*	*52*	*65.40*
Secret service	**1**	**34**	**64.12**
Psychologists	4	508	61.56
Judges	**2**	**194**	**59.01**
Cops	**8**	**511**	**55.16**
Federal officers	**4**	**341**	**54.54**
Students	122	8,876	54.20
Detectives	**5**	**341**	**51.16**
Parole officers	**1**	**32**	**40.42**

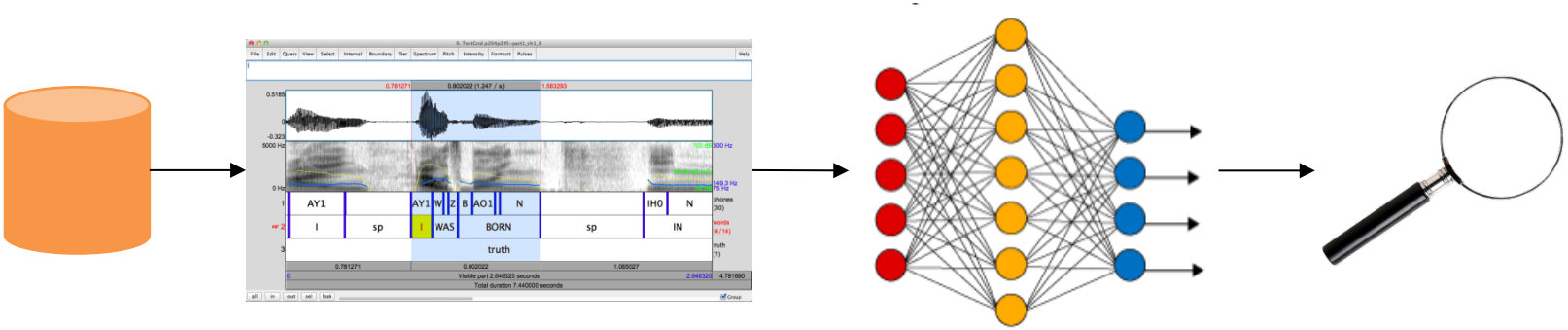
Current Approaches to Deception Detection

- ***'Automatic' methods*** (polygraph, commercial products) no better than chance
- ***Human training***: e.g. John Reid & Associates
  - Behavioral Analysis: Interview/Interrogation no empirical support, e.g.
  - Truth: *I didn't take the money* vs. Lie: *I did not take the money* **(but non-native speakers use contractions less….)**
- ***Laboratory studies***: Production and perception (facial expression, body posture/gesture, statement analysis, brain activation, odor,…)

Our Goal

- ***Conduct objective experiments on human subjects*** to identify ***spoken language*** cues to deception
- Collect speech data and extract ***acoustic-prosodic***, and ***lexical cues*** automatically
- Examine ***Individual Differences***: Take ***gender, ethnicity, culture, and personality factors*** into account as features in classification
- Use ***Machine Learning*** techniques to train models to classify deceptive vs. non-deceptive speech and ***use these to improve deception detection*** by humans by creating better methods of identifying the subtle cues humans may miss: ***Collaborative AI***

Deception Detection from Spoken Language



Corpus collection

Annotation, feature extraction

Classification

Analysis

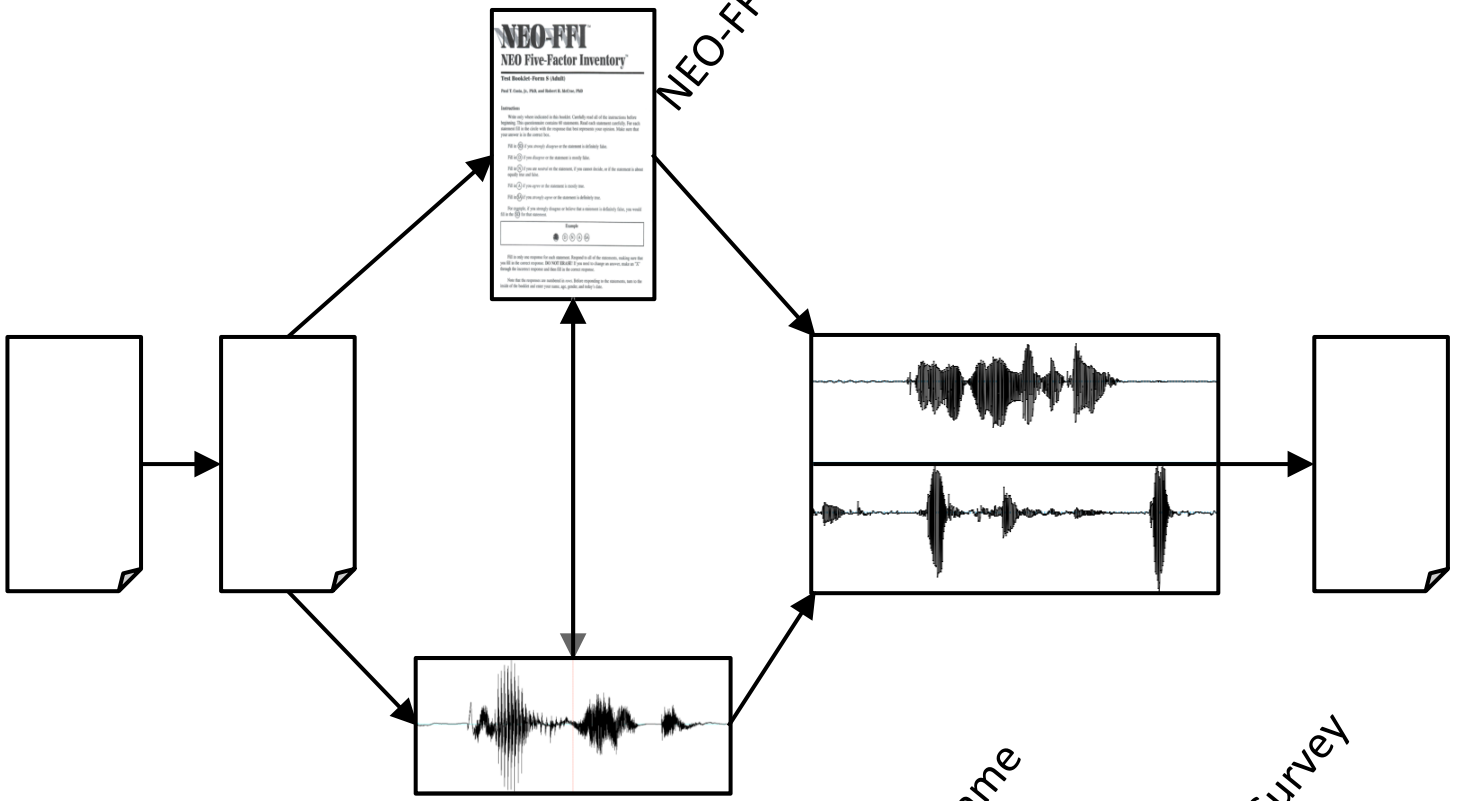
Columbia SRI Colorado Deception Corpus (‘03--)

- ***Corpus***:
  - ***7h of speech*** from 32 Standard American English-speaking subjects performing tasks and asked to lie about half
  - Examined ***lexical and acoustic-prosodic features***
  - Obtained classification *accuracy* of (70.67%) ***significantly better than the*** corpus baseline (59.93%)
  - Humans judging the same data performed worse than both (58.2%)
- ***Other findings:***
  - Our classifiers ***performed better on male subjects than females***
  - Our human ***judges who were high in certain personality features*** (openness-to-experience, agreeableness) performed much better than others on judging deception

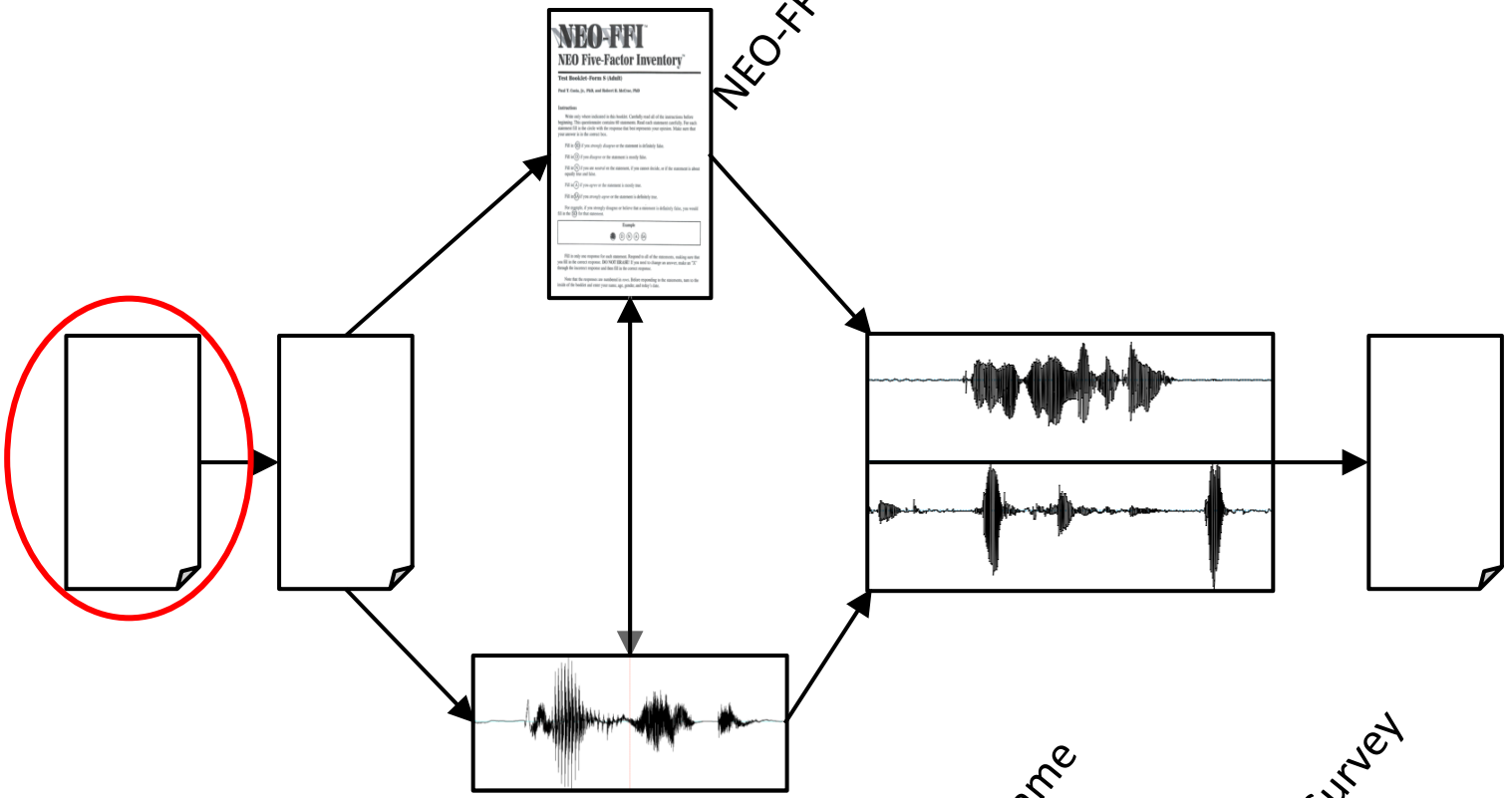
Columbia Cross-Cultural Deception Corpus (CXD)

- ***Include***
  - ***Gender and personality information for all subjects***
  - ***Compare subjects with different cultural and language backgrounds***
- ***CXD:*** Pair native speakers of SAE with native speakers of Mandarin Chinese, all speaking English, interviewing each other

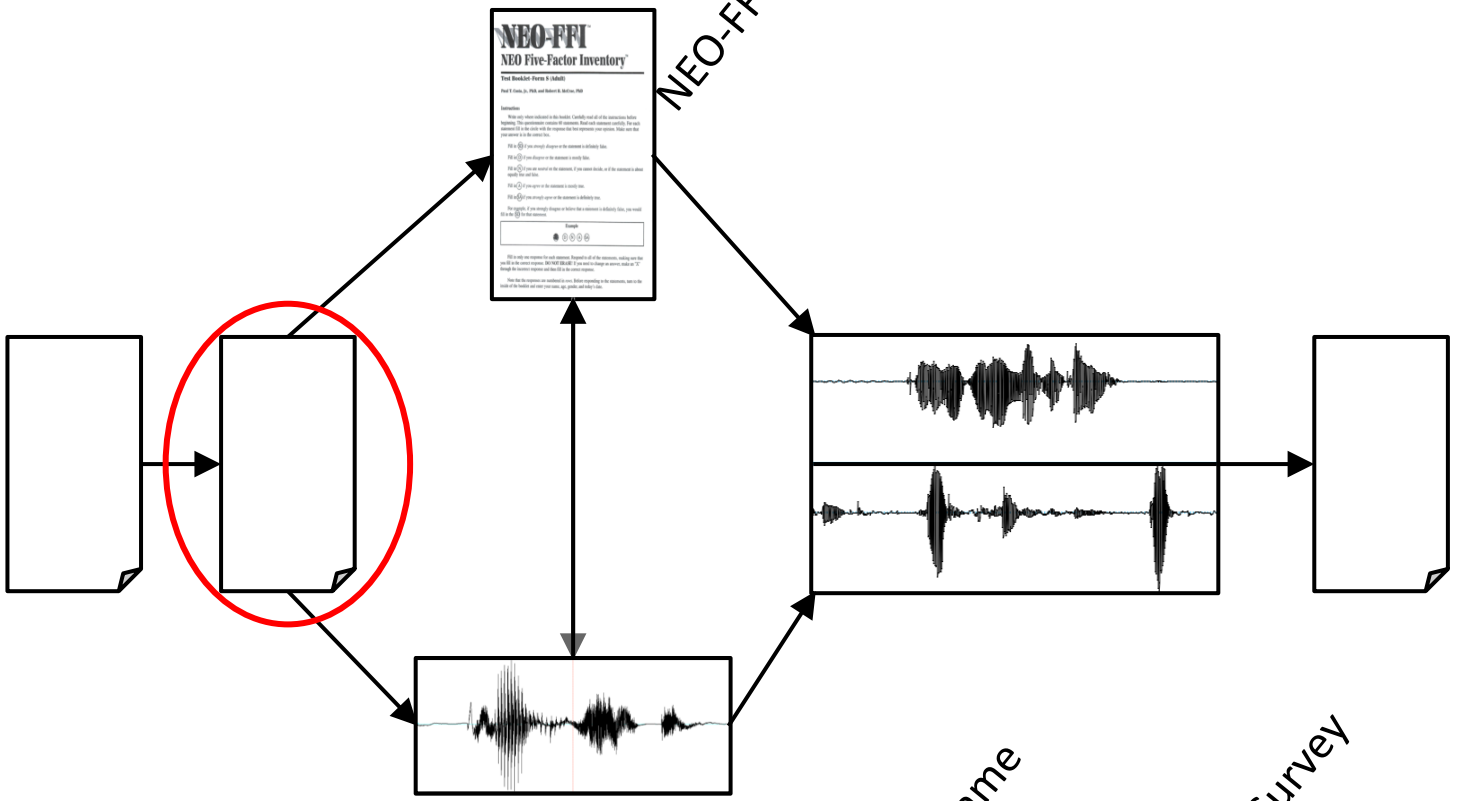
Our CXD Experiment



Our CXD Experiment



Our CXD Experiment



Biographical Questionnaire

Participant No. _____ Date _____

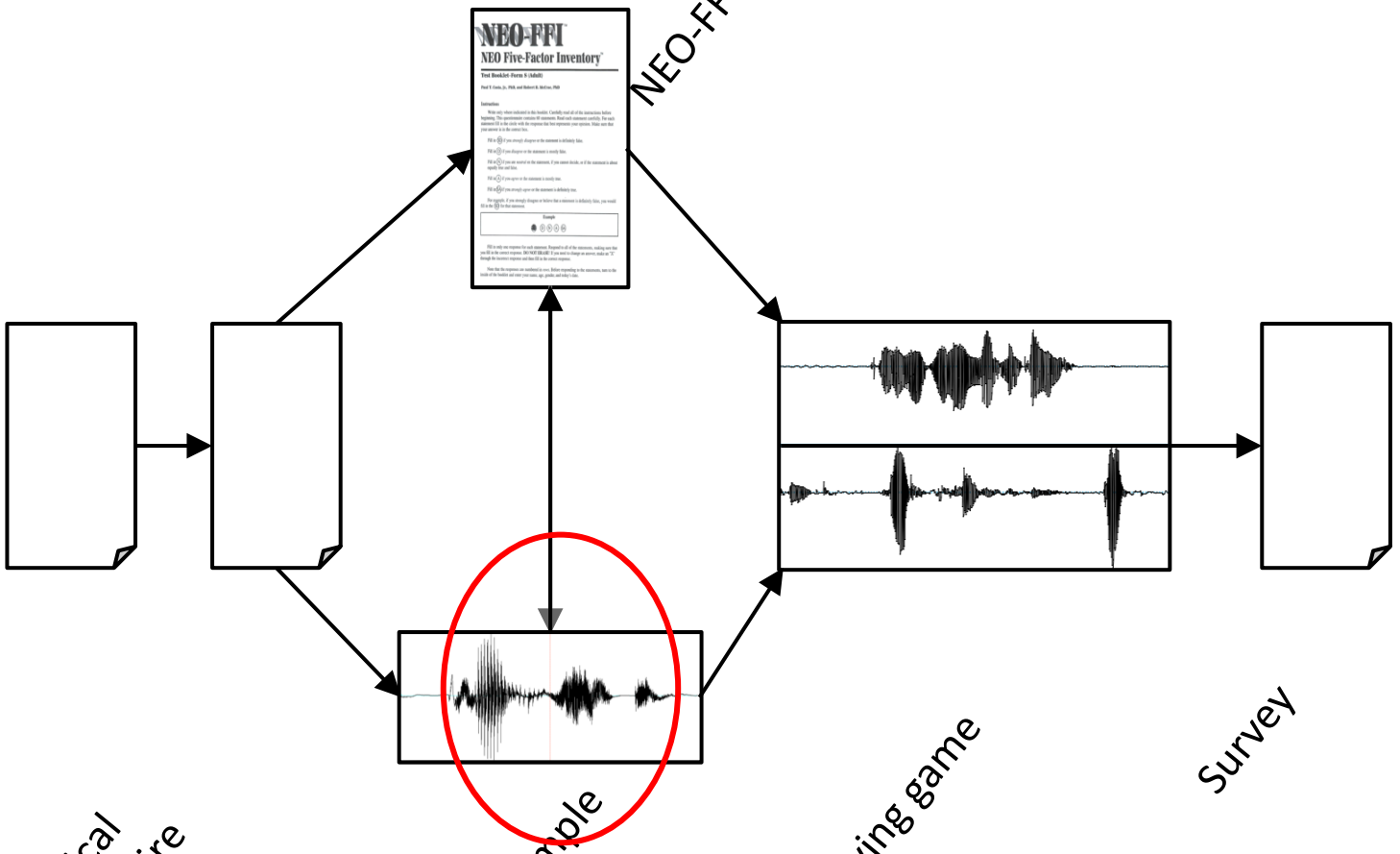
Instructions

Please carefully look through the questions. Write down the true answer to each question in the "True Answer" column. When you have finished that, for all the questions that have don't have "X"s in the "False Answer" column, make up an answer. Consult the additional sheet you have been given. You want to choose a lie that you are not as familiar with as the true answer.

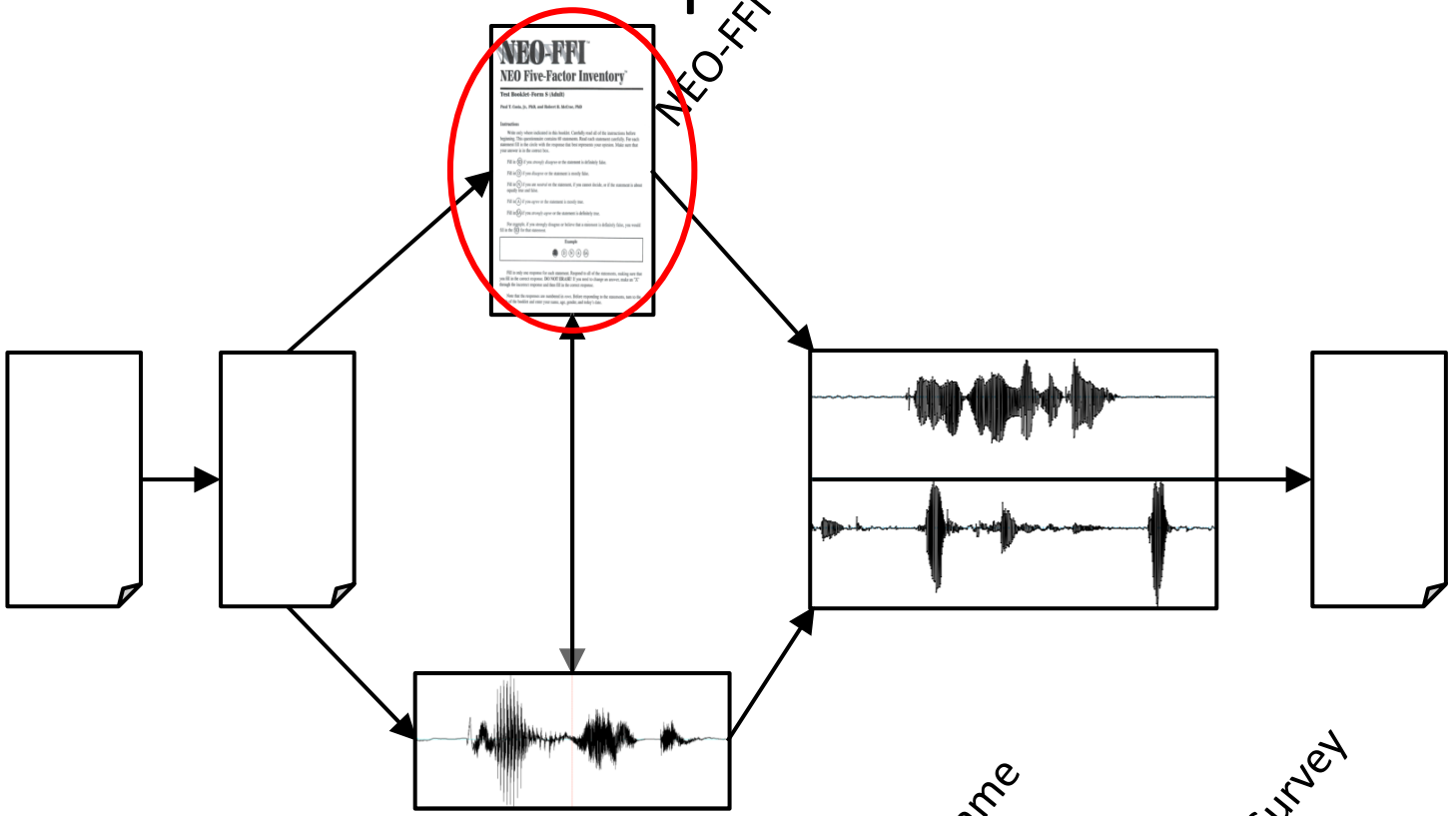
This experiment is completely anonymous- your name will never be linked to the data.

No.	Questions	True Answer	False Answer
1	Where were you born?		
2	How many years did you live in your first home?		
3	What is your mother's job?		
4	What is your father's job?		[redacted]
5	Have your parents divorced?		
6	Have you ever broken a bone?		[redacted]
7	Do you have allergies to any foods?		
8	Have you ever stayed overnight in a hospital as a patient?		
9	Have you ever tweeted? (posted a message on twitter)		
10	Have you ever bought anything on eBay?		
11	Do you own an e-reader of any kind?		
12	Who was the last person you were in a physical fight with?		[redacted]
13	Have you ever gotten into trouble with the police?		
14	Who ended your last romantic relationship?		
15	Whom do you love more, your mother or father?		
16	What is the most you have ever spent on a pair of shoes?		[redacted]
17	What is the last movie you saw that you really hated?		

Our CxD Experiment



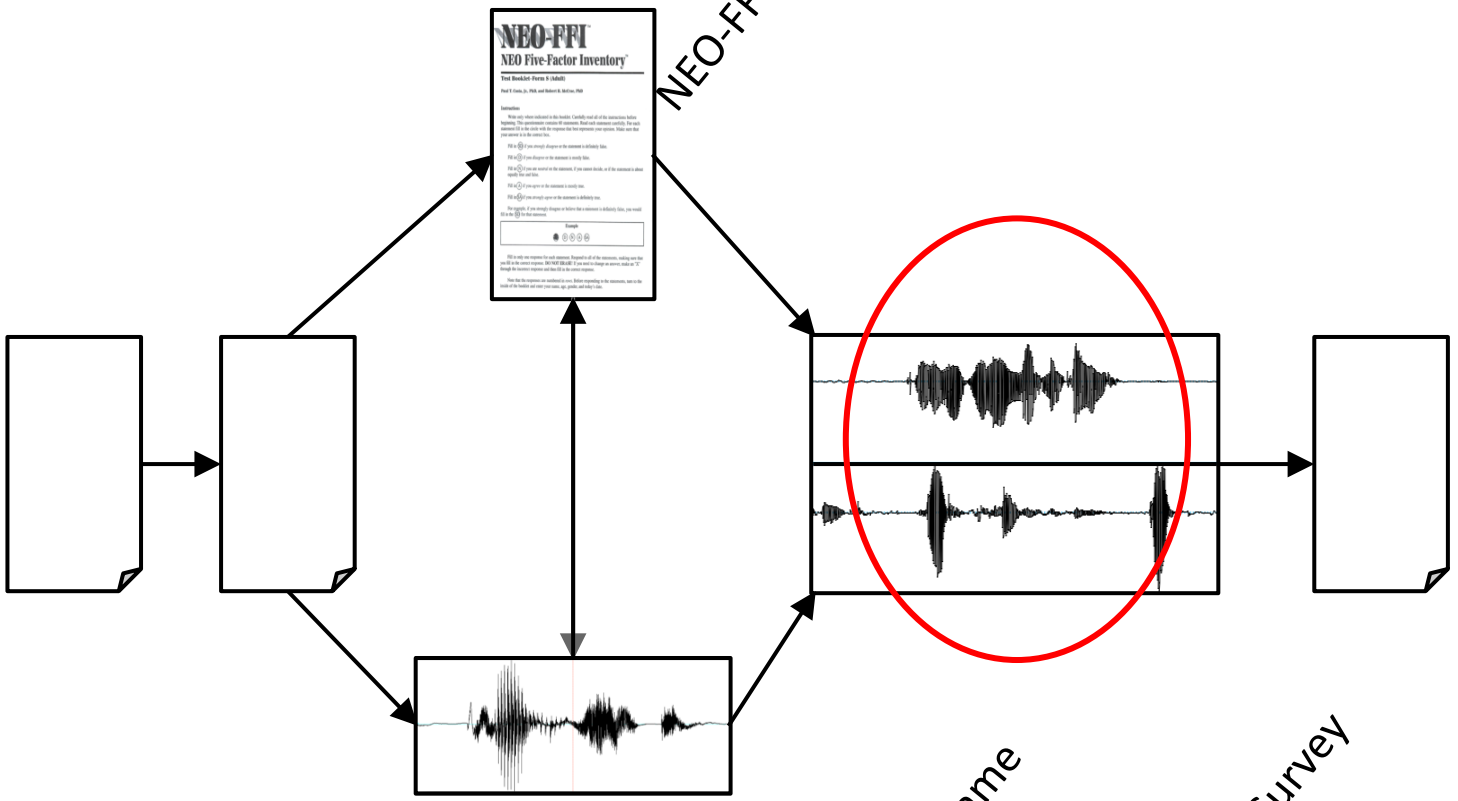
Our CXD Experiment



The Big Five NEO-FFI (Costa & McCrae, 1992)

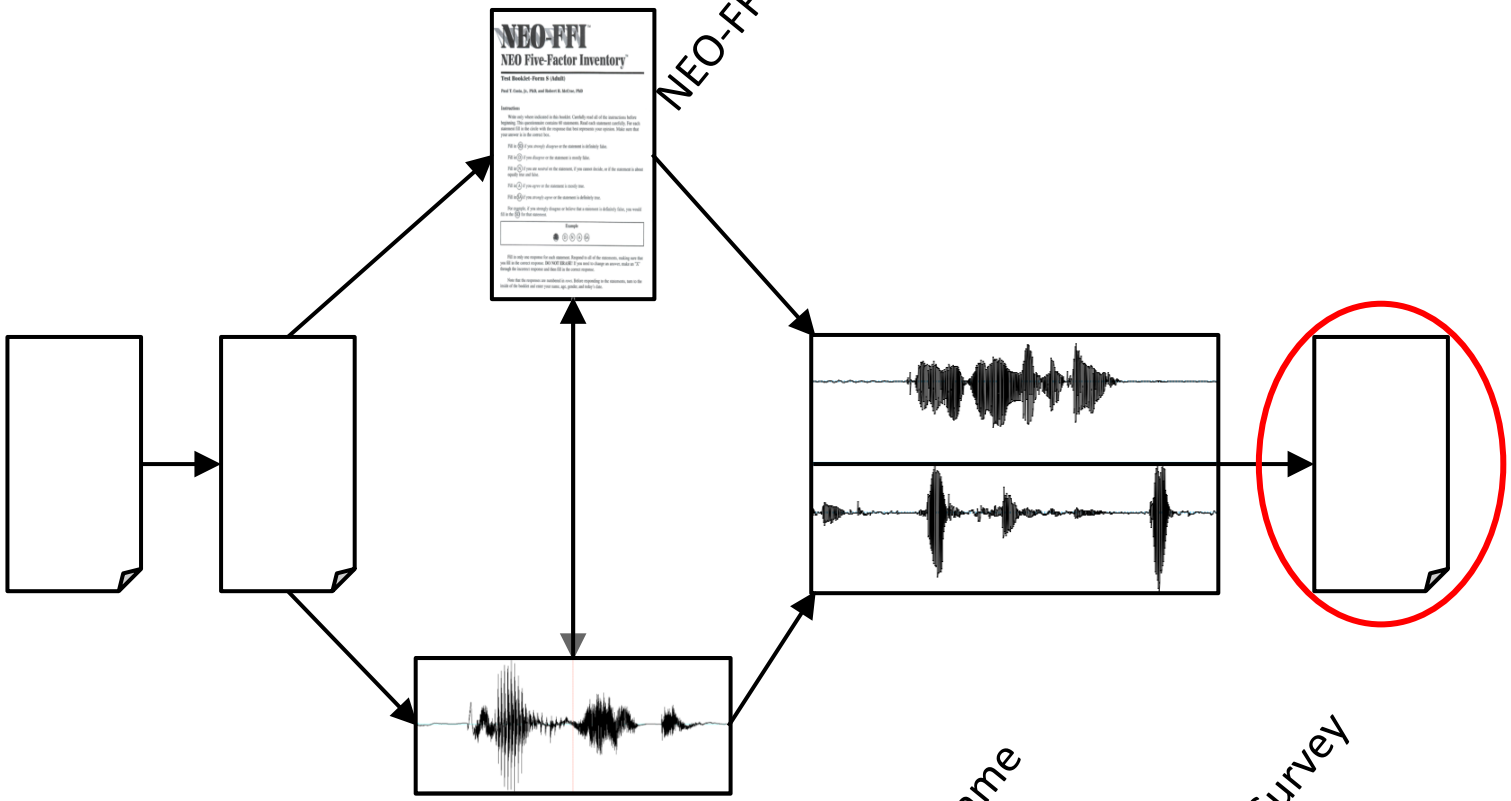
- **Openness to Experience:** "I have a lot of intellectual curiosity."
- **Conscientiousness:** "I strive for excellence in everything I do."
- **Extraversion:** "I like to have a lot of people around me."
- **Neuroticism:** "I often feel inferior to others."
- **Agreeableness:** "I would rather cooperate with others than compete with them."

Our CXD Experiment



Our CXD Experiment

Our CXD Experiment



Motivation and Scoring

- ***Monetary motivation***
  - **Success for interviewer**:
    - Add $1 for every correct judgment, truth or lie
    - Lose $1 for every incorrect judgement
  - **Success for interviewee:**
    - Add $1 for every lie interviewer thinks is true
    - Lose $1 for every lie interviewers thinks is a lie
- ***Good liars tell the truth as much as possible*** when lying, so how do we know what's true or false for follow-up questions?
  - **Interviewees press T/F keys after every phrase**

Columbia X-Cultural Deception Corpus

- ***340 subjects, balanced by gender and native language (American English, Mandarin Chinese)***: 122 hours of speech
- ***Crowdsourced transcription***, automatic speech alignment (hand-corrected)
- Interviewee speech segmented into
  - **Inter-pausal units (IPUs):** 111,479
  - **Speaker turns:** 43,706
  - **Question/answer sequences** (Q/1st Response and Q/Resp+follow-up): 7,418

Deception Annotation

- ***Deception annotation***
  - ***Local deception:*** T/F keypresses
  - ***Global deception:*** biographical questionnaire, automatically "chunked" (Maredia et al 2017)
- ***Example***
  - Interviewer: *"What is your mother's job?"*
  - Interviewee: *"My mother is a doctor [F]. She has always worked very late hours and I felt neglected as a child [T]."*
    - Global lie with local truth…

“Did you ever cheat on a test in high school?”

TRUE or FALSE?

“Did you ever cheat on a test in high school?”

“Did you ever cheat on a test in high school?”

TRUE or FALSE?

“Did you ever cheat on a test in high school?”

Features Extracted

- **Text-based**: n-grams, psycholinguistic, Linguistic Inquiry and Word Count (LIWC) (Pennybaker et al), word embeddings (GloVe trained on 2B tweets)
- **Speech-based**: openSMILE IS09 (e.g. f0, intensity, speaking rate, voice quality)(386)
- **Gender, native language, NEO-FFIs personality scores and clusters**
- **Syntactic features** (complexity), **entrainment, regional origin**

Segmentations

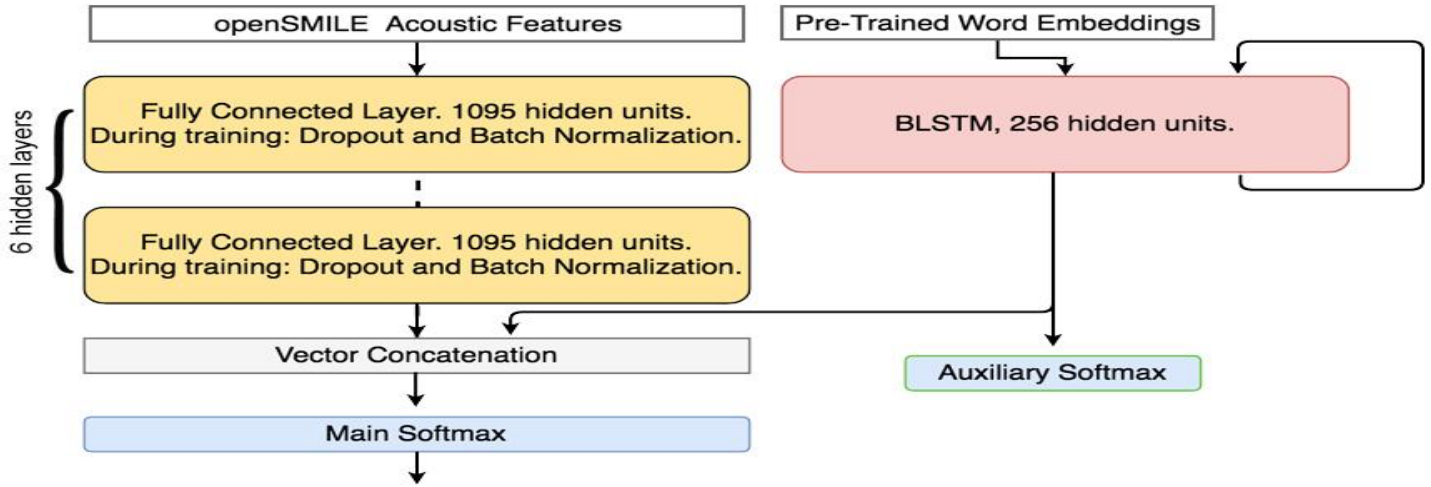
- ***IPUs (Inter-Pausal Units):*** single-speaker phrases separated by at least 50ms
- ***Speaker turns***
- ***First response*** to questions: First Turn
- ***Entire set of responses*** to a question: ”Chunks”

Classifiers

- ***Random Forest***
- ***SVMs***
- ***Deep Learning***
  - ***DNNs***
  - ***BLSTMs***
  - ***Hybrid models:*** BLSTM-GloVe embeddings + DNN-openSMILE

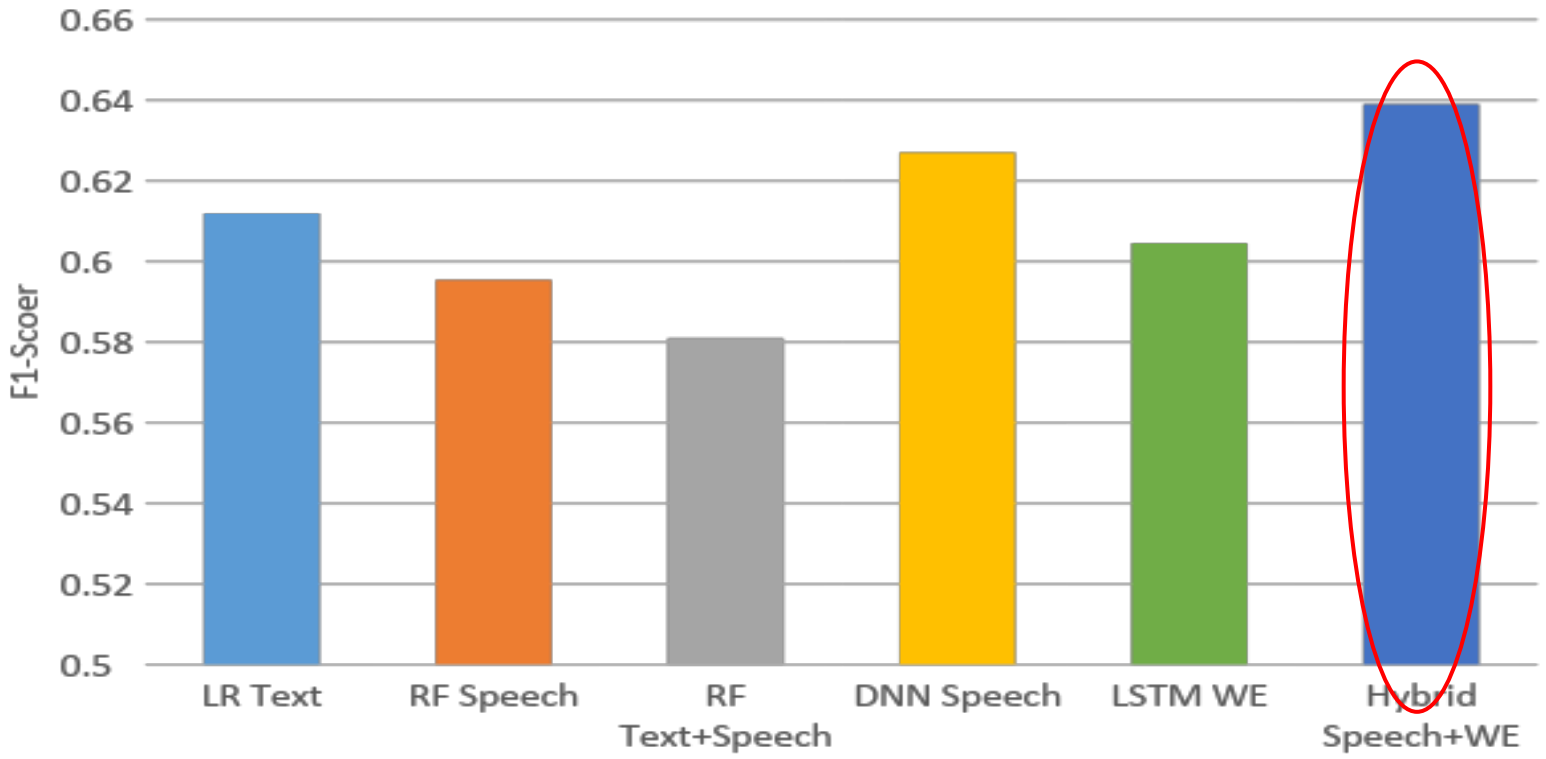
Deep Learning on Word Embeddings and openSmile Acoustic Features

- BLSTM-word embeddings
- DNN-openSMILE
- ***Hybrid: BLSTM-lexical + DNN-openSMILE***



Mendels, Levitan et al. 2017, "Hybrid acoustic lexical deep learning approach for deception detection," Interspeech, Stockholm.

IPU Classification: Hybrid Achieves Best F1



Mendels, Levitan et al. 2017, “Hybrid acoustic lexical deep learning approach for deception detection ”

Improving Deception Detection with Personality Features

- Ahn et al '18 trained deception classifiers on ***speaker turns (not IPUs)*** using acoustic/ prosodic and lexical information (LIWC, DAL, LLDs, pre-trained word embeddings (GloVe, Google)
  - ***Baseline models:*** Multilayer perceptron, LSTM, and Hybrid model combining both
  - ***Adding personality*** through multi-task learning or adding personality scores as features improved F1 from .68 to .744 for the MLP model and slighted less for the Hybrid model (no improvement in the LSTM)

Best MLP Models Adding Personality

Figure 1: *Diagram of multi-task learning MLP model (variant i).*

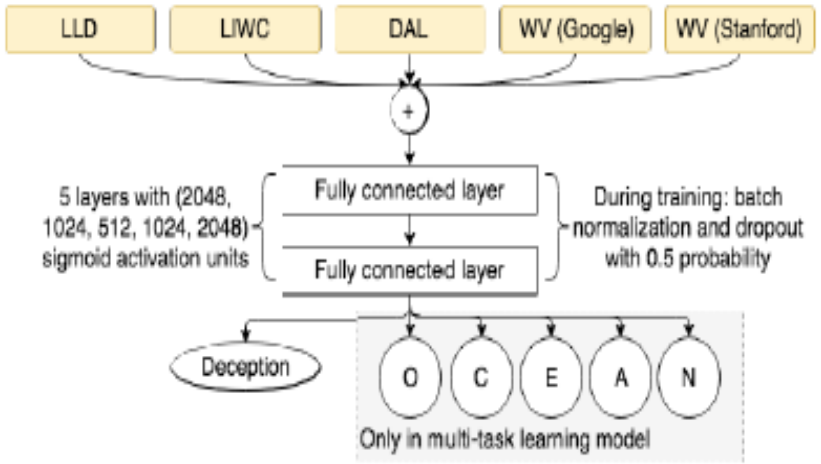
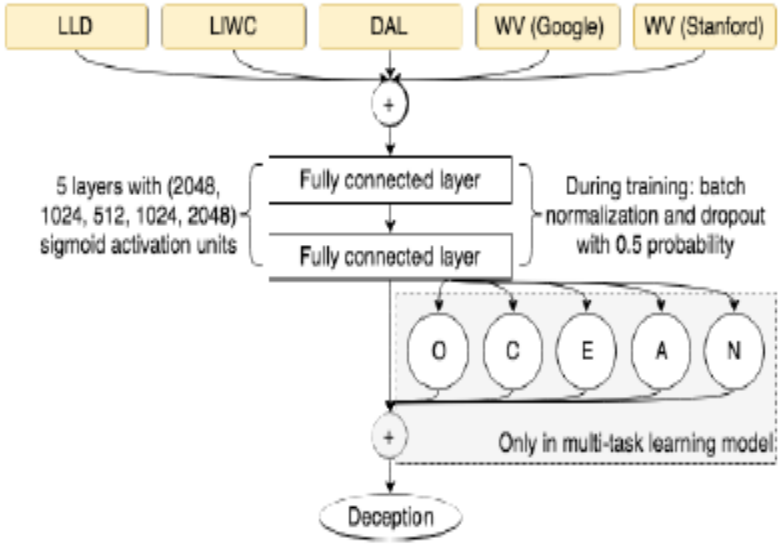


Figure 2: *Diagram of multi-task learning MLP model (variant ii).*



An, Levitan et al. 2018, "Deep Personality Recognition for Deception Detection," Interspeech, Hyderabad.

Results for Speaker Turns

Baseline w/out Personality

Model	Prec	Recall	F1
MLP	68.08	67.95	68.01
LSTM	65.64	66.08	65.78
Hybrid	69.43	69.46	***69.45***

Multi-task Learning (1)

Model	Prec	Recall	F1
MLP	74.33	74.51	***74.39***
LSTM	64.61	65.56	64.40
Hybrid	69.42	69.76	69.51

Multi-task Learning (2)

Model	Prec	Recall	F1
MLP	74.37	74.67	***74.38***
LSTM	66.13	67.03	65.89
Hybrid	72.58	72.98	72.70

What Next Can We Learn from Gender and Native Language?

- Extract ***simple acoustic/prosodic features*** from question responses
- ***Compare distributions of features*** over all and by gender and native language
  - When interview***ees*** lie vs. tell the truth
  - When interview***ees*** are trusted (believed) or are not
  - When interview***ers*** trust (believe) an interview***ee*** or do not
- Perform ***paired t-tests*** to compare feature means
  - Tests for significance correct for family-wise Type I error by controlling the false discovery rate at α=0.05. (Parentheses indicate an uncorrected p <=0.05.)

Individual Differences in Deceptive vs. Truthful Speech by Gender and Native Language

Feature	Male	Female	English	Chinese	All
Pitch Max					✔
Pitch Mean					
Intensity Max					✔
Intensity Mean					
Speaking Rate					
Jitter					
Shimmer					
NHR					

Deceptive True

Individual Differences in Deceptive vs. Truthful Speech by Gender and Native Language

Feature	Male	Female	English	Chinese	All
Pitch Max	✔				
Pitch Mean					
Intensity Max	✔	(✔)			
Intensity Mean					
Speaking Rate					
Jitter		(✔)			
Shimmer					
NHR					

Deceptive True

Individual Differences in Deceptive vs. Truthful Speech by Gender and Native Language

Feature	Male	Female	English	Chinese	All
Pitch Max				✔	
Pitch Mean					
Intensity Max			✔		
Intensity Mean			(✔)		
Speaking Rate				✔	
Jitter					
Shimmer					
NHR					

Deceptive True

Individual Differences in Deceptive vs. Truthful Speech by Gender and Native Language

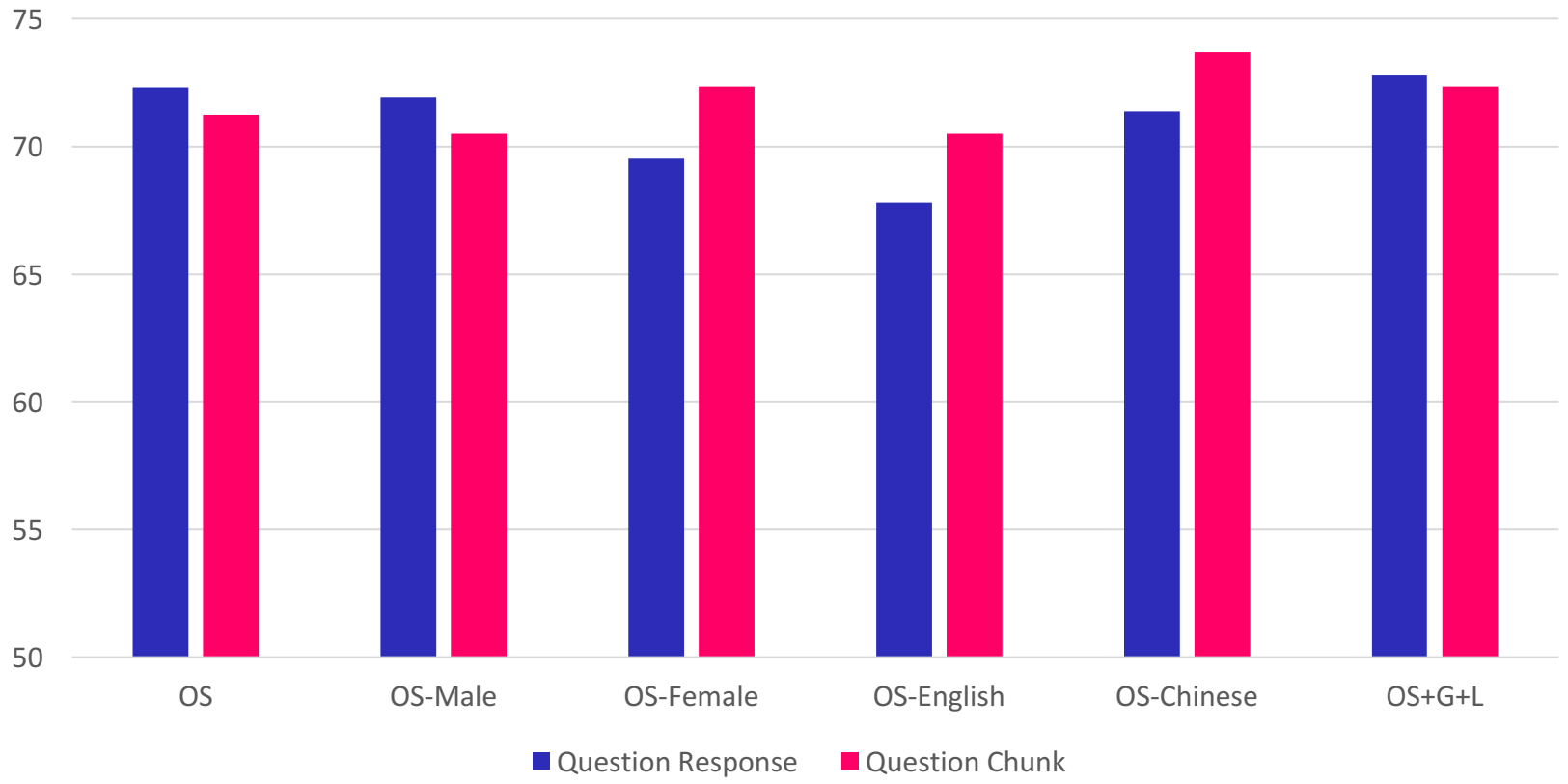
Feature	Male	Female	English	Chinese	All
Pitch Max	✔			✔	✔
Pitch Mean					
Intensity Max	✔	(✔)	✔		✔
Intensity Mean			(✔)		
Speaking Rate				✔	
Jitter		(✔)			
Shimmer					
NHR					

Deceptive True

Do Gender and Native Language Help in Classification?

- ***Features:*** openSMILE (384 acoustic/prosodic features), Gender, Native Language
- ***Classifier:*** Random Forest – fewer inputs
- Data: 7,878 question responses (first response and "chunks") and subsets for Gender and Native Language
- ***Baselines:*** random, 50.0 F1 (data is balanced for T and F labels); human baseline for this data is 46.0 F1

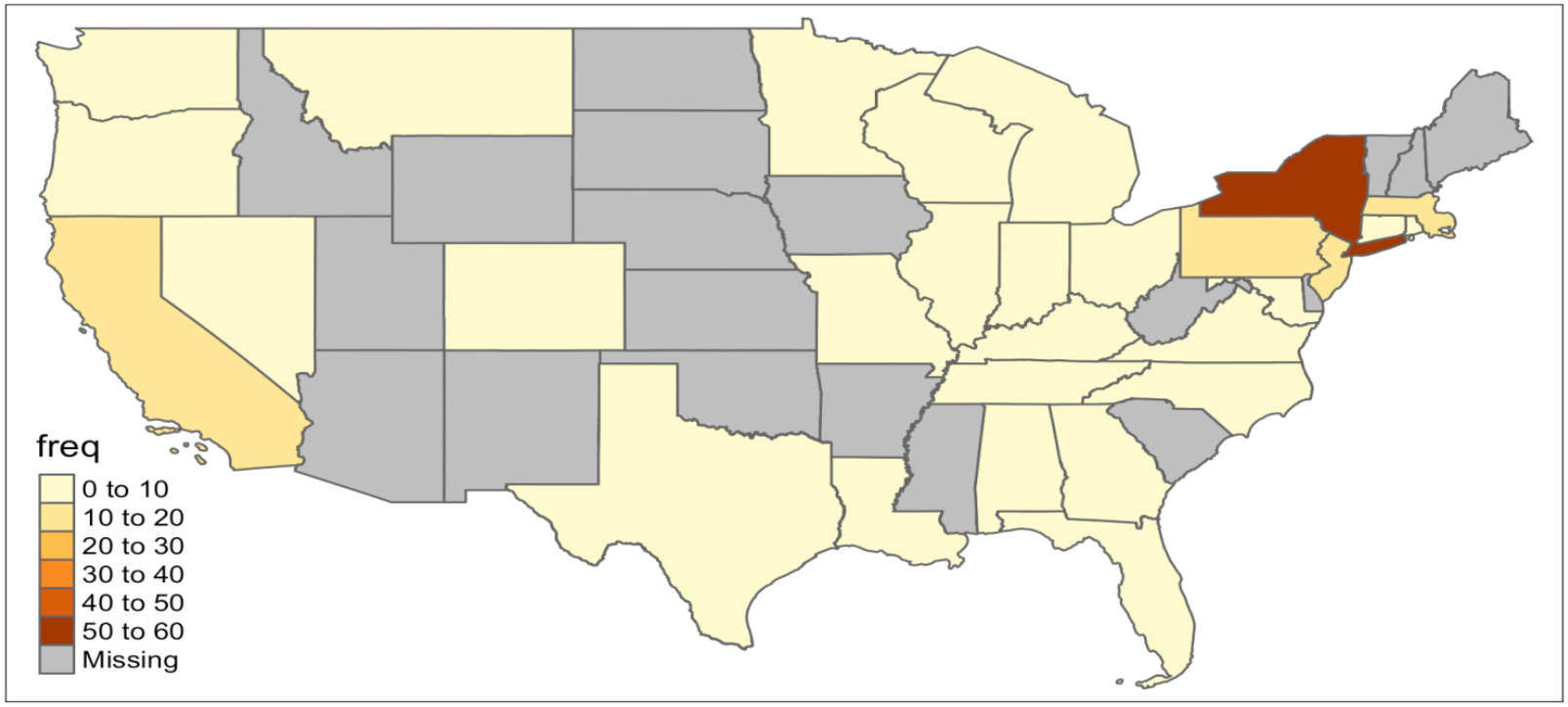
Train/Test on Gender and Native Language Groups and Adding Features to a General Classifier



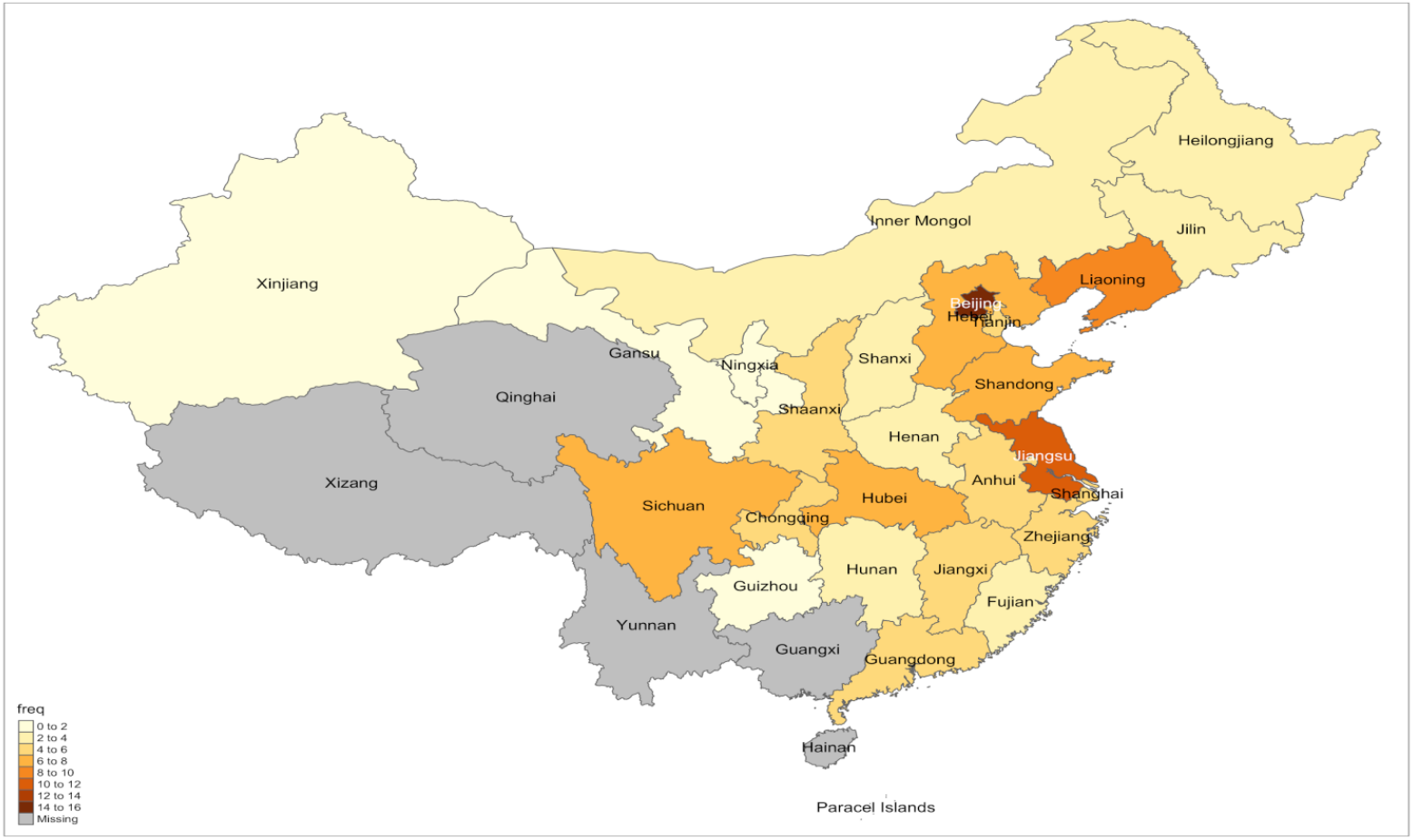
More Individual Differences

- We have also found individual differences in ***interviewee responses that interviewers believe true***, based both on the gender and native language of the (a) ***interviewee*** and that of the (b) ***interviewer***
- We have also found differences in ***entrainment*** in deceptive vs. truthful speech which may also serve as useful deception indicators
- Our next look at individual differences is looking more specifically at ***regional background***

Native American English Speaker States in USA



Native Mandarin Speaker Provinces in China



Recall that Machine Learning Models Perform much Better than Humans in Deception Detection

- Corpus Baseline: 50.00 (balanced)
- Human Performance:
  - Accuracy: 56.75
  - Precision: 56.50
  - Recall: 40.00
  - F1: 46.50

How Do Our Classifiers Compare to Humans?

- Corpus Baseline: 50.00 (balanced)
- Human vs. ML Performance:
  - Accuracy: 56.75/ 75.21
  - Precision: 56.50/ 79.30
  - Recall: 40.00/ 67.23
  - F1: 46.50/ 72.77
- ***But where do the differences lie?***

In our AI-Centric World: When are Machines Better (or Worse) than Humans?

- Are ***certain speakers*** more difficult to judge than others?
- Are ***certain groups*** more difficult to judge than others?
- Are ***particular question types*** more difficult to judge than others?
- ***Does this differ for humans and machines?***

Method

- Data
  - Naïve Bayes classifier
  - 5000 lexical + syntactic features
  - Question chunk segmentation (same as humans)
- Procedure
  - Aggregate all question chunk segments by speaker
  - Compute $F1_C$: average classifier F1 per speaker
  - Compute $F1_{human}$: average human F1 per speaker
  - 340 speakers x 24 questions = 8160 aggregated segments

Findings

- ***No correlation*** between humans and classifiers in ***which speakers are easy or difficult to judge*** – here classifiers and humans differ – how?
- ***No difference between humans and classifiers in judging gender or native language groups*** although classifier much better at judging speakers who scored ***low in Conscientiousness (NEO-FFI)***
- ***Both humans and classifiers judged longer responses as lies and shorter ones as truthful*** and ***both found most of the same questions easy or hard to judge***

Some “Easy” Questions for Both

- Question #5: Have your parents divorced?
- Question #13: Have you ever gotten into trouble with the police?
- Question # 16: What is the most you have ever spent on a pair of shoes?
- ***What makes these easier to classify?***

Some “Easy” Questions for Both

- Question #5: Have your parents divorced?
- Question #13: Have you ever gotten into trouble with the police?
- Question # 16: What is the most you have ever spent on a pair of shoes?
- What makes these easier to classify?
  - ~80% of interviewee parents were not divorced
  - ~80% of interviewees had never gotten into trouble with the police
  - Most spent on pair of shoes: median(T)=$150; median(F)=$350

A Hard Question for Both

- Question #8: Have you ever stayed overnight in the hospital as a patient?
  - Human_{F1}: 50 F1 (-6.37 from mean)
  - Classifier_{F1}: 64.59 F1 (-5.22 from mean)
- What makes this harder to classify?

A Hard Question for Both

- Question #8: Have you ever stayed overnight in the hospital as a patient?
  - $Human_{F1}$: 50 F1 (-6.37 from mean)
  - $CLassifier_{F1}$: 64.59 F1 (-5.22 from mean)
- What makes this harder to classify?
  - ~60% of interviewees never stayed overnight in the hospital as a patient

Hard for Humans, Easy for Classifier

- Question #6: Have you ever broken a bone?
  - $Human_{F1}$: 51.55 F1 (-4.82 from mean)
  - $Classifier_{F1}$: 72.61 F1 (+2.85 from mean)
- What percentage of interviewees do you think had ever broken a bone?

Hard for Humans, Easy for Classfiers

- Question #6: Have you ever broken a bone?
  - $Human_{F1}$: 51.55 F1 (-4.82 from mean)
  - $Machine_{F1}$: 72.61 F1 (+2.85 from mean)
- What percentage of interviewees do you think had ever broken a bone?
  - ~75% had never broken a bone

Performance by Question Type

- Yes-no (13) vs. open-ended (11) questions
- Sensitive (8) vs. non-sensitive (16) questions
- Sensitive: related to money, parental or romantic relationships, mortality, socially undesirable behaviors or experiences (Tourangeau & Yan, 2007)
  - e.g. "Who ended your last romantic relationship?" "Who do love more, your mother or your father?"

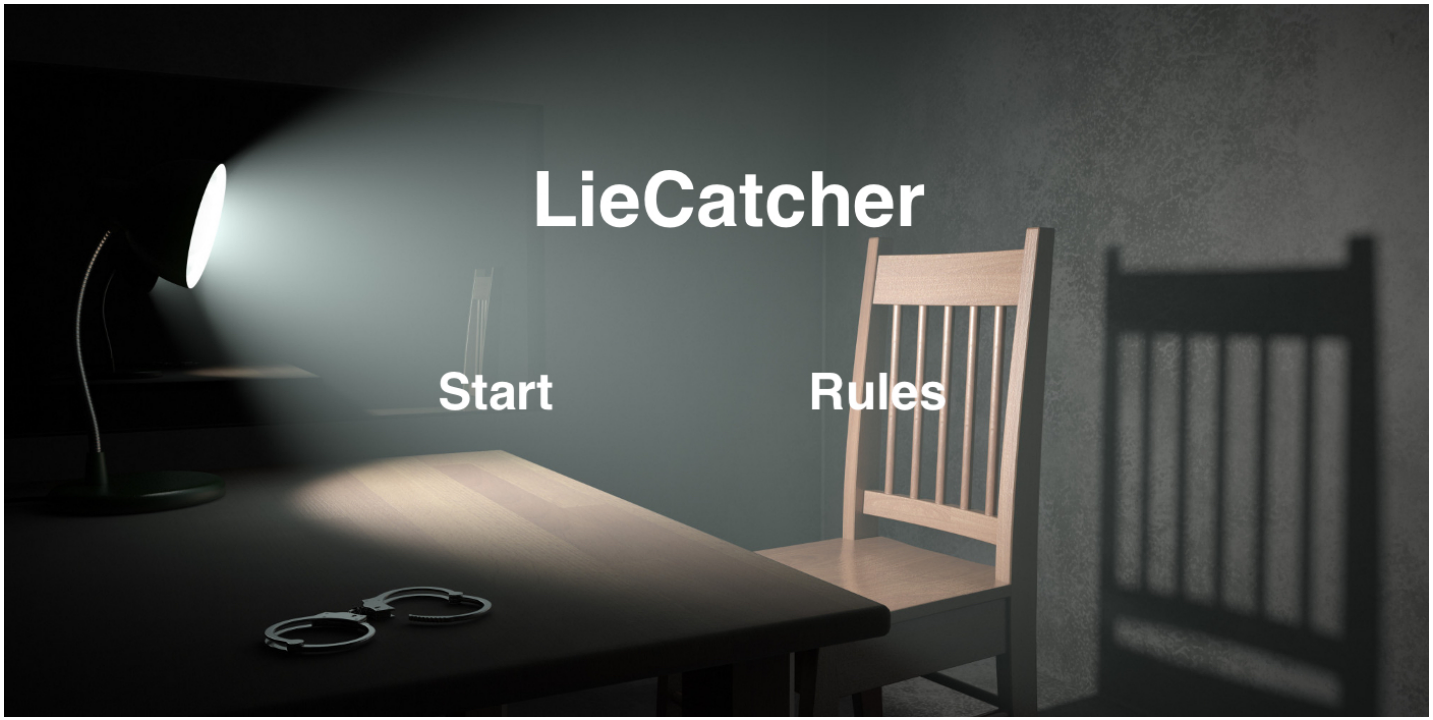
Performance by Question Type

- Yes-no (13) vs. open-ended (11) questions
- Sensitive (8) vs. non-sensitive (16) questions
- Sensitive: related to money, parental or romantic relationships, mortality, socially undesirable behaviors or experiences (Tourangeau & Yan, 2007)
  - e.g. “Who ended your last romantic relationship?” “Who do love more, your mother or your father?”
- Human were ***significantly better at judging sensitive questions than non-sensitive***, while Classifiers showed no differences across question types

Current and Future Research

- ***Many individual differences to be considered: Incorporating all features*** into our deep learning classifiers: ***gender, native language, personality, entrainment, regional origin as well as acoustic, lexical and syntactic features***
- ***Creating "trusted" and "mistrusted" synthetic voices*** based on our findings (What are the acoustic-prosodic features of voices that hearers believe or do not believe?) for robots, avatars, chatbots
- ***Obtain more human judgments*** on our data through crowd-sourcing: ***The Lying Game***

Games with a Purpose



Levitan et al. 2018, "LieCatcher: Game framework for collecting human judgments of deceptive speech," LREC 2018, Miyazaki.

“Who was the last person you had a physical fight with?”

True or False?

“Who was the last person you had a physical fight with?”

“Who was the last person you had a physical fight with?”

True or False?

“Who was the last person you had a physical fight with?”

“Who was the last person you had a physical fight with?”

True or False?

“Who was the last person you had a physical fight with?”

“Who was the last person you had a physical fight with?”

True or False?

“Who was the last person you had a physical fight with?”

Thank you!


DEPARTMENT OF
U.S.
HOMELAND SECURITY


NSF


AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
UNITED STATES AIR FORCE