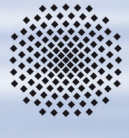


Universität Stuttgart

Institut für Maschinelle Sprachverarbeitung

Ethics and NLP

Sebastian Padó

SS 2021

First of all, welcome!

- This is an extraordinary situation for all of us
- First of all, some technical announcements
  - This seminar will happen **online, synchronously**
  - We will use WebEx meetings in the slot at https://unistuttgart.webex.com/join/sebastian.pado
- Please generally keep your microphones muted
  - Only open when you say something, use a headset
  - Use a “raise hand” option when possible (Desktop app)
  - You are invited to have your camera turned on ☺
- Important: communicate early, and openly!
  - ILIAS: Forum and Chat; Email

About me

- Sebastian Pado
  - Studied computational linguistics, computer science (Saarland University, Edinburgh)
  - PhD in computational linguistics (Saarland)
  - Postdoc in Stanford
  - Professor for computational linguistics in Heidelberg, Stuttgart
  - Main research: Semantics, Multilinguality, DH/CSS

Why Ethics and NLP?

- What is new?

Input: Language has gone digital
- Language input for programs (Siri, Google, …)
- Online communication (social networks, forums)

Processing: Big Data & DL
- Models can be trained on large datasets and filter out extremely subtle properties

Output: Language technology works
- More and more AI/NLP support in business: hiring, loans, business directions, customer interaction, …
- Suddenly, NLP shapes the way **society works**

”The AI winter”

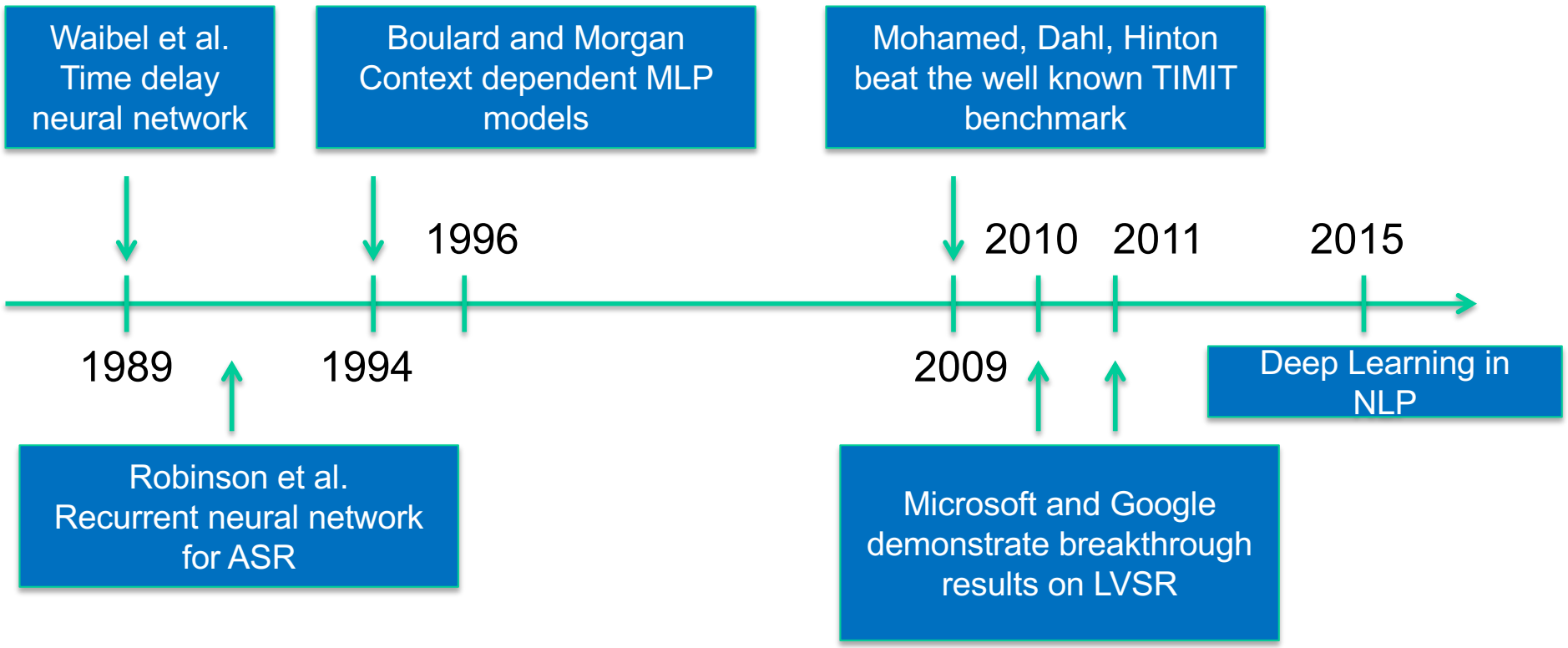
- Two major winters in 1974–80 and 1987–93 with some highlights:
  - 1966: the failure of machine translation
  - 1969: the abandonment of connectionism (first description of a neural network)
  - 1971–75: DARPA‘s frustration with the Speech Understanding research program at the Carnegie Mellon University
  - 1988: the cancellation of new spending on AI by the Strategic Computing Initiative

https://en.wikipedia.org/wiki/AI_winter

Computational Linguistics 1990-2010

- Empirical methods in 1990s: Statistical approach (Manning and Schütze, 1999)
  - Internet
  - Learning from data
- Machine learning methods:
  - Maximum Entropy, SVMs, Bayesian methods, etc. were successfully applied to NLP problems
- Common ground: supervised learning
  - Handcrafted **features**
  - Parameters (e.g. weights of the features) trained with decent amount of **data**

Timeline


Waibel et al.
Time delay
neural network
Boulard and Morgan
Context dependent MLP
models
Mohamed, Dahl, Hinton
beat the well known TIMIT
benchmark
1996
2010
2011
2015
1989
1994
2009
Deep Learning in
NLP
Robinson et al.
Recurrent neural network
for ASR
Microsoft and Google
demonstrate breakthrough
results on LVSR

Deep Learning

-
-
-

Current Developments

‘Data Collection’ Platforms:

Quantifying the tendencies

- Big Data – Unstructured Data
  - Google processes 3.5 trillion searches a day
  - YouTube users upload 48 hours of new video every minute of the day
  - Facebook has 1.8 billion daily users
  - Twitter sees roughly 500 million tweets every day …

Ethically relevant aspects of current NLP

Privacy

- Need for very large amounts of data to train models
  - Extreme incentive for companies to **collect (private) data**

- Feature developer as ‘middle man’ vanishes
  - Learning directly from (unfiltered) data
  - Learns everything in the data: potential for **biases**

Bias

- Model learns its own representations
  - **“Black box”** models

Explainability

Privacy

- AI can often add a lot to people's comfort
  - Siri: "Play the song 'Eye of the tiger'"
  - Google automatically extracts meeting dates from email and puts them in calendar

- But very often, there is a price to pay
  - *Either you pay with money or with your data*
  - Siri knows about your music taste
  - Google knows who you write to, what web pages you visit, where you are, whether you do that during work hours, …

NLP / Data Science is **dual use**
Are companies going to act responsibly with your data?

Privacy

- NLP is good enough to extract a lot of **personal attributes** from what people say
  - Most people are familiar with authorship attribution
  - In a similar vein:
    - Age
    - Gender
    - Native language (when speaking in 2nd language)
    - Extroverted/introverted, other personality traits
    - Early signs of mental health problems (dementia etc.)

Helpful in the right hands
But who would you trust with this information?

Non-NLP: Good & Bad

- Camera
  - Hold your beautiful moments
- Camera in public areas
  - Security, safety
  - But also privacy

There is awareness regarding picture taking (and laws)
What about speech / language?

https://www.lto.de/recht/nachrichten/n/ueberwachung-gesichtserkennung-software-polizei-kritik-datenschuetzer/

http://www.digicamhistory.com/1960s.html

Bias

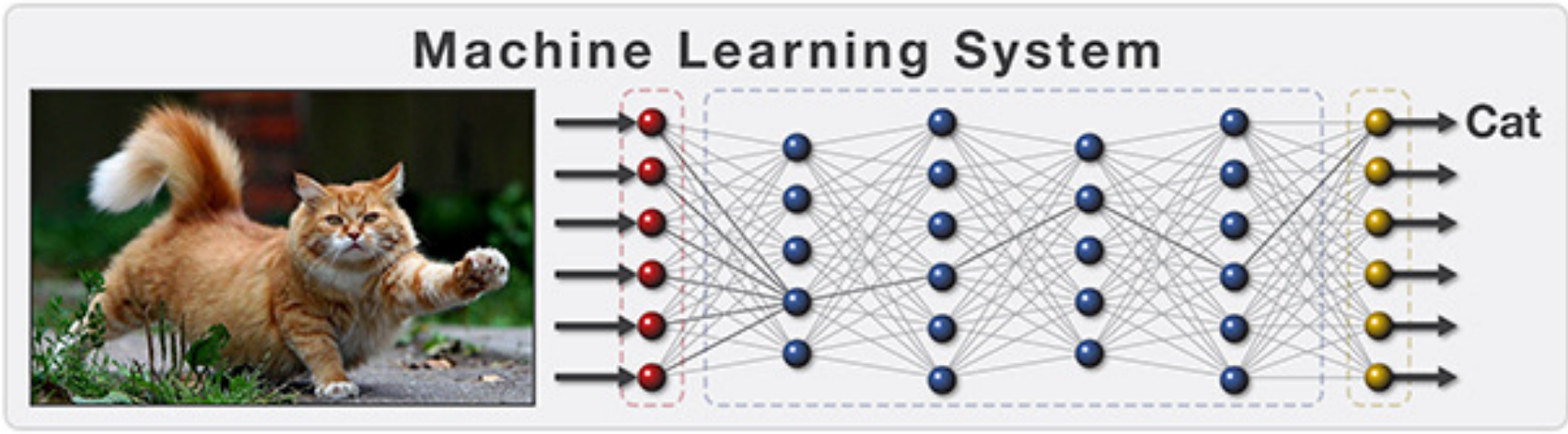
- "An unfair expression of prejudice for or against a person, a group, or an idea" (Bender and Friedman, 2018)
- Q: How can Bias turn up in NLP models?
  - Models pick up on properties of the training corpora
- Risk 1: Models internalize **facts about the world**
  - Analogy questions: King is to Queen as Doctor is to …?
  - Nurses tend to be female, doctors tend to be male
- AI systems making real decisions rely on such predictions
  - Perpetuate stereotypes / unfairness in the world
    - Black people / Immigrants are criminals / badly educated / …
    - Parole / credit application gets denied, ….

Bias

- Risk 2: Models internalize fact about **language**
  - Many standard models used to be trained on newspaper corpora (in the 2000s and beyond)
  - Written by a specific demographic: older, highly educated men
- Consequence: Speech technology works worse for input by the young, the less educated, women, …
  - Puts these groups at a disadvantage
  - Can be seen as overfitting

Explainable AI

- DARPA research focus



https://www.darpa.mil/program/explainable-artificial-intelligence

Explainable AI

- Interpretable ML Symposium at Neurips since 2017

http://interpretable.ml

The First Move in European Law

- European Union regulations on algorithmic decision-making and a “right to explanation” [Goodman, Flaxman, 2016]

Abstract

We summarize the potential impact that the European Union’s new General Data Protection Regulation will have on the routine use of machine learning algorithms. Slated to take effect as law across the EU in 2018, it will restrict automated individual decision-making (that is, algorithms that make decisions based on user-level predictors) which “significantly affect” users. The law will also effectively create a “right to explanation,” whereby a user can ask for an explanation of an algorithmic decision that was made about them. We argue that while this law will pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks which avoid discrimination and enable explanation.

Responsibility

Input: Language has gone digital
- Language input for programs (Siri, Google, …)
- Online communication (social networks, forums: “echo chambers”)

P

Arguably, Computational Linguistics / Machine Learning is the nuclear physics of the 2010s:

At the heart of scientific change and its societal consequences

Output: Langua
- More and more … loans, business direction…
- Suddenly, NLP shapes … **orks**

Role of a Computational Linguist Expert?

- *Code of Ethics for NLP and AI community? [Daumé, 2016]*
- IEEE Code:
  - to accept responsibility in **making decisions consistent with the safety, health, and welfare of the public**, and to disclose promptly factors that might endanger the public or the environment;
  - **to be honest and realistic in stating claims** or estimates based on available data;
  - **to improve the understanding of technology; its appropriate application, and potential consequences**;
  - …

https://nlpers.blogspot.de/2016/12/should-nlp-and-ml-communities-have-code.html

Role of a Computational Linguist Expert?

- *Code of Ethics for NLP and AI community? [Daumé, 2016]*
  - Responsibility in Research
    - Conduct research honestly, avoiding plagiarism and fabrication of results
  - Responsibility to Students, Colleagues, and other Researchers
    - Teach students ethical responsibilities
  - Compliance with the code
    - Uphold and promote the principles of this code
  - **Responsibility to the Public**
    - **Contribute to society and human well-being, and minimize negative consequences of computing systems**

https://nlpers.blogspot.de/2016/12/should-nlp-and-ml-communities-have-code.html

Organization

Goals of this Seminar

- Contents:
  - Explore current papers (**and other sources**) on different aspects connecting Ethics and NLP
- Skills:
  - Read and understand discussions in papers
  - Present a topic in front of audience with a diverse background
  - **Prepare discussion questions and lead discussion**
  - Develop own opinions/ideas
  - Documentation

More mixed sources than usual:
(Most) topics cover one NLP and one „context“ paper

Session Structure

To be modified if necessary to accommodate more presentations

- Presentation (20min)
  - Slide-based
  - Capture both technical level **and** ethical/social **context**
    - Less focused on individual papers
    - Larger degree of speculation than usual
    - I want to see your **personal assessment,** not just a summary
- Discussions (10-15min)
- Report (4 pages)
  - Coherent presentation/discussion of at least three sources
    - Should include outcome of class discussion

How to prepare slides

- Read material for topic
  - (Brownie points for reading beyond)
  - Check that you understand the papers
  - Get in touch with advisor if you have trouble with understanding
- Come up with a story that binds the papers together
- Create draft slides & questions
  - Submit **at least one week in advance** to me
  - I will do my best to provide feedback in a few days
  - I expect you to revise slides based on feedback
    - Send final slides to me 1 day before talk to be put online!

How to prepare the discussion

- Prepare **three questions about your topic** in writing
  - Should be discussion prompts
  - Please keep them concrete and controversial
    - Not: “What do you think about AI?” but: “Do you think this approach could be problematic from a privacy point of view”?
    - Not: “Do you think Google is a good company?” but: “What steps could be taken to make Google a better company?”
- Put them at the beginning of your slides to give people time to think about them -- also post them in the chat
  - We will attempt to do a “full interactive” discussion
  - If that doesn’t work (lag, turn taking, …), I will collect ideas from the chat and discuss with the presenter
  - **Chat** can be used for clarification questions all along

Modules

- Standard: Module “Ethics and NLP” (3 LP)
  - Cover one topic (presentation, questions, lead discussion, …)
  - Participate
  - Write short term paper on literature used for presentation
    - Include discussion based on class interactions
- Possible: Extension to Module “Project seminar CL” (6 LP)
  - Same as above
  - Additional effort possible through:
    - Long term paper about substantial body of literature
    - Practical project
  - Write me an email if you are interested

Grading Criteria

- Slides
  - Clear story? Factually accurate? Good grasp of relevance? Good use of presentation devices? Own assessment?
- Talk
  - Good speed? Comprehensible?
- Discussion
  - Questions well chosen? Good discussion lead?
- Term Paper
  - Well structured? (See slides)

Organization

- 19.04. : Introduction, Topics, Organization
- 26.04. : Discussion on J. Zittrain‘s talk (Inverted Classroom)
- 03.05. : Biases (Part 1)
- 10.05. : Biases (Part 2)
- 17.05. : Explainable AI
- 31.05. : Fake ‚Information‘
- 07.06. : Trolling, Hate Speech
- 14.06. : Privacy, User Profiling 1
- 21.06. : Privacy, User Profiling 2
- 28.06. : Generation
- 05.07. : Emotions and Chatbots
- 12.07. : Speech
- 19.07. : Speech & Wrap-up

Next Week: Zittrain

- Jonathan Zittrain: Professor of Internet Law at Harvard

- Great talk of his in the context of an opening session of his course on Ethics and Governance of AI at MIT
  - Please watch Zittrain's talk (1:00h to 2:30h) https://www.youtube.com/watch?v=MyW6eAGV-eM until next Monday
  - We will discuss the contents of the talk next week
    - Interactive whiteboard: https://cryptpad.fr/whiteboard/#/2/whiteboard/edit/d0ep-J4w9a+UcTMf2cVy3H1n/
  - **Please take notes for the questions on the next slides**

Questions for Zittrain

1. What is the one topic or statement in the talk that struck you most?

2. What is agenda setting?

3. Zittrain defines AI as 'arcane pervasive tightly coupled adaptive autonomous systems'. Please unpack that definition.

4. Zittrain talks about many problems that involve feedback loops (e.g. voter badge presentation in FB, rider rating in Uber). Are these really AI problems?

5. Two important concepts appears to be fairness and reponsibility. Please define them and outline whether you believe companies need to be fair and responsible.

Questions for Zittrain

6. Do you buy the dichotomy between "tools" and "friends“ that Zittrain claims, and the technological movement towards "friends"? What consequences does that have?

7. The topic of human involvement comes up several times. Can human involvement solve the problems that were sketched? (Which? How?)

8. At 1:57, Zittrain talks about 'overfitting'. Do you think that this the right term? Why / why not?

9. Why do justifications matter so much?

10. What do you think are reponsibilities of Data Scientists?

Paper Selection

- All papers I selected are on ILIAS
- You are welcome to find more papers!
  - See Literature folder
  - See http://www.cl.uni-heidelberg.de/courses/ws13/cl_gesellschaft/
  - See http://faculty.washington.edu/ebender/2017_575/
- Please send me until the end of this week (Fri April 23) by email to pado@ims:
  - Ranked list of 3 topics (=**presentations**) of your choice (send numbers!)
  - In which semester you are
- Selection process (if necessary):
  - Higher semester > lower semester; then remaining topics
- **Please only participate if you are serious about participating**

Session 1 – Biases [a]

- **Presentation 1: What is Bias (in NLP)?**
  - Blodgett, Barocas, Daume, Wallach 2020. Language (technology) is power: A critical survey of “bias” in NLP.
  - Bender and Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science.

- **Presentation 2: Debiasing**
  - Context: Hardt 2014. How big data is unfair -- Understanding unintended sources of unfairness in data driven decision making
  - Technology: Bolukbasi et al. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Session 2 – Biases [b]

- **Presentation 3: Bias in Sentiment Analysis**
  - Context: Yu et al. 2012. The impact of social and conventional media on firm equity value: A sentiment analysis approach
  - Technology: Kiritchenko and Mohammad 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems

- **Presentation 4: Bias in Computational Social Science**
  - Context: Haunss 2017. (De-)Legitimating Discourse Networks: Smoke without Fire?
  - Technology: Dayanik & Pado 2020. Masking Actor Names Leads to Fairer Claim Detection

Session 3 – Explanainable AI

- **Presentation 5: Integrated Explanations**
  - Context: Goodman and Flaxman 2017. European union regulations on algorithmic decision making
  - Technology: Liu et al. 2019. Towards Explainable NLP: A Generative Explanation Framework for Text Classification

- **Presentation 6: Post-hoc Explanations**
  - Context: Deeks 2019. The judicial demand for explainable artificial intelligence.
  - Technology: Ribeiro et al. 2018. Anchors: High-Precision Model-Agnostic Explanations

Session 4 – Fake Information

- **Presentation 7: Fake News**
  - Context: Allcott & Gentzkow 2017. Social Media and Fake News in the 2016 election
  - Technology: Rashkin et al. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking

- **Presentation 8**: **Fact Checking**
  - Context: Cohen et al. 2011, Computational Journalism: A Call to Arms to Database Researchers
  - Technology: Popat et al. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning

Session 5 – Trolling and Hate Speech

- **Presentation 9: Hate Speech**
  - Context: Bleich 2011. The Rise of Hate Speech and Hate Crime Laws in Liberal Democracies
  - Technology: Gröndahl et al. 2018, All You Need is “Love”: Evading Hate Speech Detection

- **Presentation 10: Trolling**
  - Context: Cheng et al. 2017. Anyone can become a troll: Causes of Trolling Behavior in Online Discussions
  - Technology: Samghabadi et al. 2017. Detecting Nastiness in Social Media

Session 6 – Privacy, User Profiling [a]

- **Presentation 11: Personal Information**
  - Context: Hovy & Spruit 2016, The Social Impact of Natural Language Processing
  - Technology: Benton et al. 2017. Multi-Task Learning for Mental Health using Social Media Text

- **Presentation 12: Privacy**
  - Context: Solove 2007. 'I've got nothing to hide' and other misunderstandings of privacy.
  - Technology: Zhang et al 2016. Safelog: Supporting web search and mining by differentially-private query logs

Session 7 – Privacy, User Profiling [b]

- **Presentation 13: Personalization**
  - Context: Bright 2008, Consumer control and customization in online environments
  - Technology: Grbovic 2018, Real-time Personalization using Embeddings for Search Ranking at Airbnb

- **Presentation 14: Privacy and Text**
  - Context: Garcia-Rivadulla 2016. Personalization vs. privacy: An inevitable trade-off?
  - Technology: Li et al 2018. Towards Robust and Privacy-preserving Text Representations

Session 8: Generation

- **Presentation 15: Overview, NL Generation**
  - Context: Dale 2020: NLG, the commercial state of the art
  - Technology: Smiley et al. 2017: Say the Right Thing Right: Ethics Issues in Natural Language Generation Systems.

- **Presentation 16: Chatbots**
  - Context: Schwartz 2019, In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation
  - Technology: Vinyals and Le 2015. A neural conversational model
  - (Bonus: Csáky 2017. Deep Learning based Chatbot Models.)

Session 9 – Emotions and Chatbots

- **Presentation 17: Emotion Generation**
  - Context: Wu 2019, Empathy in Artificial Intelligence
  - Technology**:** Zhou and Wang 2018, MOJITALK: Generating Emotional Responses at Scale

- **Presentation 18: Emotional Chatbots**
  - Context: Vincent 2016, Twitter users taught Microsoft Tay to be racist asshole in less than a day
  - Technology: Song et al. 2019, Generating Responses with a Specific Emotion in Dialog

Session 10 – Speech [a]

- **Presentation 19: Multispeaker TTS**
  - Context: UK Gov 2019, Snapshot Paper - Deepfakes and Audiovisual Disinformation
  - Technology: Jia et al. 2018, Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis

- **Presentation 20: Deception**
  - Context: Levitan et al. 2015, Individual Differences in Deception and Deception Detection
  - Technology: Levitan et al. 2016, Combining Acoustic-Prosodic, Lexical, and Phonotactic Features for Automatic Deception Detection

Session 11 – Speech [b]

- **Presentation 21: Multimodal information**
  - Context: Poropat 2009, A Meta-Analysis of the Five-Factor Model of Personality
  - Technology: Alam & Riccardi 2014, Predicting Personality Traits using Multimodal Information

- + Wrap-Up

"Fallback solution"

- I will put today's slides and the literature on my personal home page at https://www.nlpado.de/~sebastian/ to provide access to people who don't have ILIAS access yet
  - (still, please try to get access ASAP)

Thanks!
Questions?