Responsible Machine Learning (Part I)

A/Prof. Hongzhi Yin

https://sites.google.com/view/hongzhi-yin/home

About Me

- ARC Future Fellowship 2021, https://itee.uq.edu.au/article/2021/09/data-scientist-wins-prestigious-arc-future-fellowship
- Field Leader of Data Mining & Analysis in Australia, The Australian's Research 2020 magazine, https://specialreports.theaustralian.com.au/1540291/15/
- UQ Foundation Research Excellence Award 2019, https://www.uq.edu.au/news/article/2019/09/uq%E2%80%99s-finest-researchers-awarded

Responsible Big Data Intelligence Group

- To develop **decentralized (cloud-free), on-device, and trustworthy** (e.g., privacy-preserving, secure, robust, explainable and fair) data mining and machine learning techniques with theoretical backbones to better discover actionable patterns and intelligence from **large-scale, heterogeneous, networked, dynamic and sparse data**

- To integrate these **actionable patterns** with **domain knowledge** from environment, urban transportation, healthcare, smart grid, E-commerce and marketing to help solve societal, environmental and economical challenges facing humanity, in pursuit of **a sustainable future**

A variety of real-life applications

- mobile recommender systems
- chatbot
- sales prediction for supply-chain optimisation
- user account linkage and multi-view user profiling
- event detection and tracking
- rumour detection and argument discovery
- rare disease prediction
- on-device disease detection and prediction
- real-time traffic speed prediction
- passenger demand prediction and passenger-driver matching
- anomaly detection in decentralized smart grid

What will be covered in week 10's and week 11's lectures

- Motivation for Ethics research in machine learning
- How and why machine learning models are unfair
- Various types of machine learning fairness issues and mitigation approaches

What will not be covered

- Definitive answers to fairness/ethical questions
- Prescriptive solutions to fix ML (un)fairness

Introduction of Ethics in Machine Learning

Ethical Considerations in ML (AI) Research

What is Machine Learning

As with any concept, machine learning may have a slightly different definition, depending on whom you ask. We combed the Internet to find five practical definitions from reputable sources:

1. "Machine Learning at its most basic is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world." – Nvidia
2. "Machine learning is the science of getting computers to act without being explicitly programmed." – Stanford
3. "Machine learning is based on algorithms that can learn from data without relying on rules-based programming."- McKinsey & Co.
4. "Machine learning algorithms can figure out how to perform important tasks by generalizing from examples." – University of Washington
5. "The field of Machine Learning seeks to answer the question "How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?" – Carnegie Mellon University

What is Ethics?

“Ethics is a study of what are **good and bad** ends to pursue in life and what it is **right and wrong** to do in the conduct of life.

It is therefore, above all, a **practical discipline**.

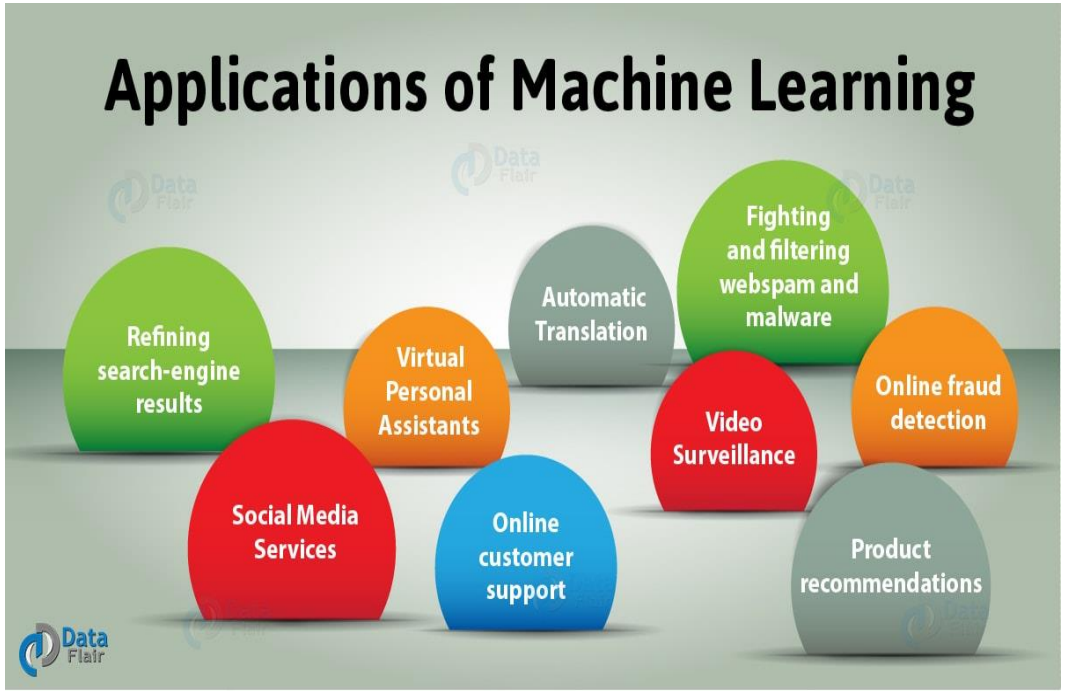
Its primary aim is to determine how one ought to live and what actions one ought to do in the conduct of one’s life.”

-- Introduction to Ethics, John Deigh

ML (AI) and People

- ML (AI) systems are pervasive in our world.
- Questions of ethics are raised in human-centered ML (AI) systems.
- Various ML systems designed for:
  - helping people (improving service like credit card check, self-driving vehicles)
  - or interacting with people (conversational agents/virtual assistants)
  - or reasoning about people (profiling for recommendation systems)
  - or they affecting people in another way (parole/visa/job decisions)

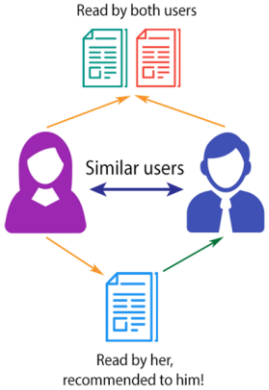
- Human-centered Applications



ML (AI) and People

- Automatic decisions about people
  - Parole decisions, employment, immigration

- People profiling applications
  - Demographic and personality profiling
  - Targeted content: ads, propaganda, agenda-setting, fake news



- Applications that interact with people
  - Conversational agents, personal assistants

- NLP Applications
  - Machine Translation
  - Information Retrieval
  - Question Answering
  - Dialogue Systems
  - Sentiment Analysis
  - ...

All NLP applications are human-centered

All NLP applications are human-centered

The common misconception is that language has to do with **words** and what they mean.

It doesn't.

It has to do with **people** and what ***they*** mean.

Herbert H. Clark & Michael F. Schober, 1992

Working on Ethical Issues in ML (AI)

- **Ethics is hard even to define, it is subjective and it changes over time**: should we be then trying to quantify and evaluate ethics in ML(AI)?
  - It is another problem with an ill-defined answer
    - It still has some definition of good and bad
    - Not everyone agrees on all examples
    - But they do agree on some examples
    - They do have some correlation between people
  - Complex ML problems are also hard to quantify and evaluate
    - Summarization, QA, dialogue (chatbot), speech synthesis

Our Goals In Teaching This Lecture

Identify a range of problems where ethical issues emerge, particularly focusing on **ML Technologies such as NLP and CV** that interact with **People**

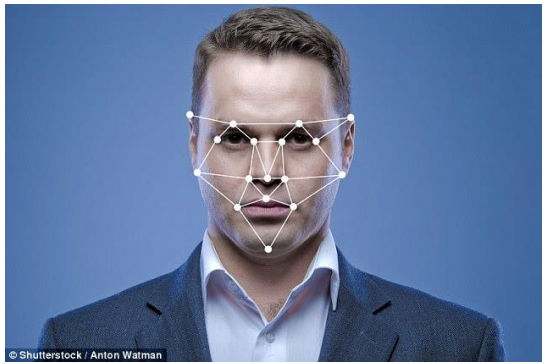
Identify a range of questions that we should be asking ourselves when working with these problems

Discuss ample examples.

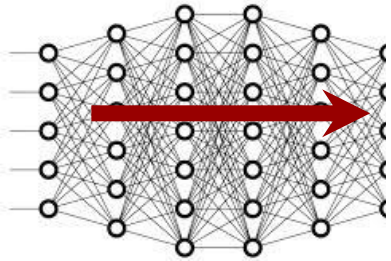
Explore socially responsible ML techniques.

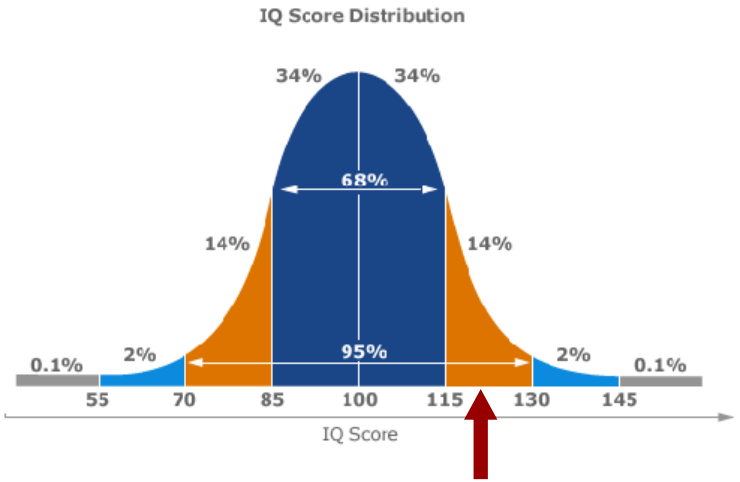
Let’s Train an IQ Classifier

Let’s train a DNN model to predict people’s IQ from their photos.



DNN Model





- **I**ntelligence **Q**uotient: a number used to express the apparent relative intelligence of a person

An IQ Classifier

Let’s train a classifier to predict people’s IQ from their photos.

- Who could benefit from such a classifier?
  - University?
  - Army?
  - Immigration offices?
  - Employers?
  - Plastic Surgery Hospitals?

An IQ Classifier

Let's train a classifier to predict people's IQ from their photos.

- Assume the classifier is 100% accurate. Who can be harmed from such a classifier?
  - Hard working people with a lot of knowledge, good soft skills, but non-genius IQ
- How it can be misused?
  - -- if a university decides: "We'll use this classifier to accept only smart people" -- does it actually identify smart people?
  - -- IQ is not an absolute measure of intelligence. Just one metric. Other variables (knowledge, skills, motivation) can be ignored if we trust this classifier too much.

An IQ Classifier

Let’s train a classifier to predict people’s IQ from their photos.

- Who could benefit from such a classifier?
- Who can be harmed by such a classifier?
- We know in a real life classifiers always make mistakes. Suppose, our test results show 90% accuracy
    - What happens if the classifier makes a mistake? The cost of misclassification is expensive.
    - Are there any minority groups for which the accuracy is much lower?

An IQ Classifier

Let’s train a classifier to predict people’s IQ from their photos.

- Who could benefit from such a classifier?
- Who can be harmed by such a classifier?
- Suppose, our test results show 90% accuracy (average accuracy)
  - Evaluation reveals that white females have 95% accuracy
  - People with blond hair under age of 25 have only 60% accuracy
  - How about the accuracy for white females under age of 25 with blond hair?
    - Unknown due to the lack of such test examples in the test set.
  - Should it be reported in the paper?

An IQ Classifier

Let’s train a classifier to predict people’s IQ from their photos.

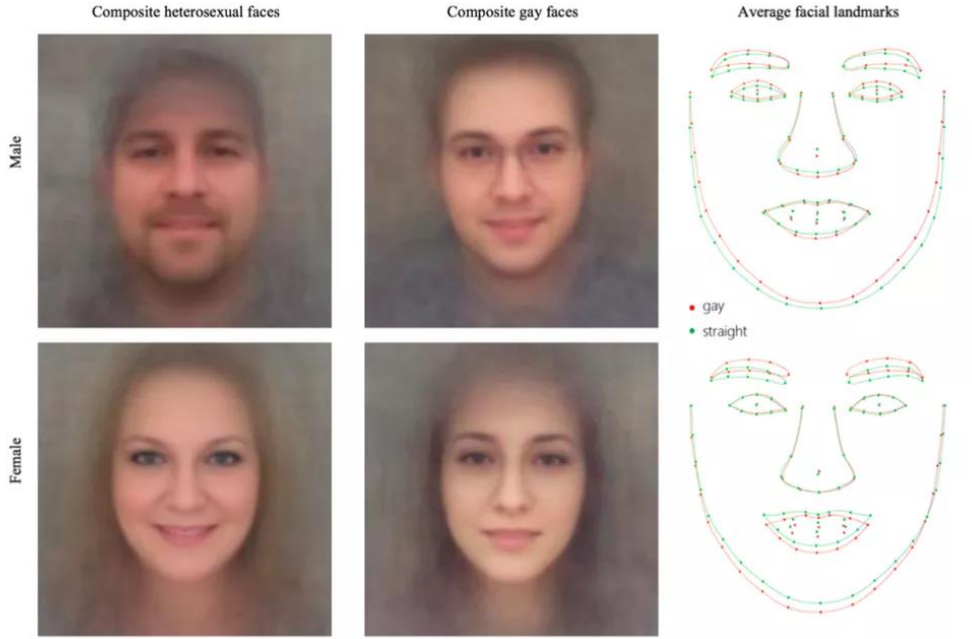
- Who could benefit from such a classifier?
- Who can be harmed by such a classifier?
- Are there biases in training data?
  - More white females than black females in the training data
  - More white females than white males in the training data
  - What will happen?

An IQ Classifier

Let’s train a classifier to predict people’s IQ from their photos.

- Who could benefit from such a classifier?
- Who can be harmed by such a classifier?
- Are there biases in the training data?
- Is the test set big enough to cover all possible cases?
- What personal data was used as training data? Privacy concerns?
- Who is responsible?
  - Researcher/developer? Reviewer? University? Society?

A Recent Study: the “A.I. Gaydar”



“GAYDAR” refers to the ability to accurately identify others’ sexual orientation from mere observation.

Wang & Kosinski. **Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.** *Journal of Personality and Social Psychology (in press).* September 7, 2017.

A Case Study: the “A.I. Gaydar”

Abstract. We show that faces contain much more information about sexual orientation than can be perceived and interpreted by the human brain. We used deep neural networks to extract features from 35,326 facial images. These features were entered into a logistic regression aimed at classifying sexual orientation. Given a single facial image, a classifier could correctly distinguish between gay and heterosexual men in 81% of cases, and in 74% of cases for women. Human judges achieved much lower accuracy: 61% for men and 54% for women. The accuracy of the algorithm increased to 91% and 83%, respectively, given five facial images per person. Facial features employed by the classifier included both fixed (e.g., nose shape) and transient facial features (e.g., grooming style). Consistent with the prenatal hormone theory of sexual orientation, gay men and women tended to have gender-atypical facial morphology, expression, and grooming styles. Those findings advance our understanding of the origins of sexual orientation and the limits of human perception. Additionally, given that companies and governments are increasingly using computer vision algorithms to detect people’s intimate traits, our findings expose a threat to the privacy and safety of gay men and women.

Wang & Kosinski. **Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.** *Journal of Personality and Social Psychology (in press).* September 7, 2017.

A Case Study: the “A.I. Gaydar”

- Research question
  - Identification of sexual orientation from visual features
- Data collection
  - Photos downloaded from a popular American dating website
  - 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly
- Method
  - A deep learning model CNN was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification
- Accuracy
  - 81% for men, 74% for women

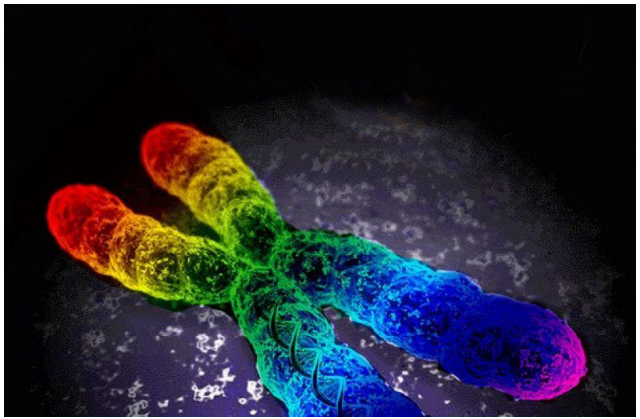
Let’s Discuss...

- Research question
  - Identification of sexual orientation from visual features
- Data collection
  - Photos downloaded from a popular American dating website
  - 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly
- Method
  - A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification
- Accuracy
  - 81% for men, 74% for women

What ethical questions could be asked here?

Research Question

- Identification of sexual orientation from visual features



Studies ranged from 19th century's measurements of homosexual men's hips,
to late 20th century's so-called "gay genes," "gay brains," "gay ring fingers", "lesbian ears,"
or other physical differences between homosexual and heterosexual bodies.

Is it ok to do research in this area?

Research Question

- Identification of sexual orientation from visual features

How can people be harmed by this research?

- In many countries being gay person is not acceptable (by law or by society) and in some places there is even death penalty for it
- It might affect people's employment; family relationships; health care opportunities;
- Personal attributes, e.g. gender, race, sexual orientation, religion are social constructs. They can change over time. They can be non-binary. They are private, intimate, often not visible publicly.
- Importantly, these are properties for which people are often discriminated against.

Research Question

- Identification of sexual orientation from visual features

“... Additionally, given that companies and governments are increasingly using computer vision algorithms to detect people’s intimate traits, our findings expose a threat to the privacy and safety of gay men and women.”

→ your thoughts on this? Should we do this kind of reseach?

There is an increasing divide between what is technically possible and what is ethical and allowed.

Data

- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly

Data – Problem 1 : Privacy

- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly

Public ≠ Publicized

Public is fine: they want to be found by their target audience. But there are different social circles and people can decide to reveal different information at different circles. What they want to put on dating websites might not be what they want to discuss in family or at work.

Did these people agree to participate in the study?

According to GDPR, people has the right to know how their data is used and the right to restrict processing/using of their data. In this case, the purpose of using the data has changed from dating to the research on sexual orientation.

Data - Problem 2: Bias

- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly

Data-Problem 2: Bias

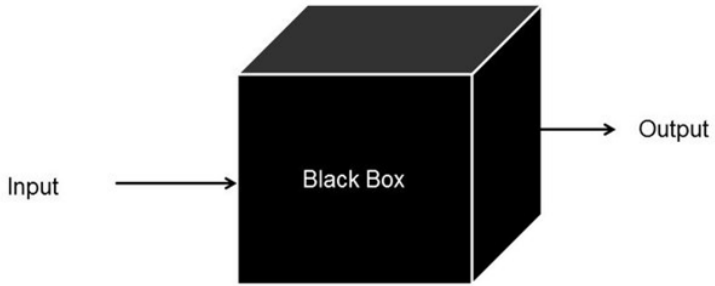
- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly

Only white people, who self-disclose their orientation, certain social groups, certain age groups, certain time range/fashion;
the photos were carefully selected by subjects to be attractive so there is even self-selection bias…
The dataset is balanced, which does not represent true class distribution.

(Race bias, self-selection bias, sampling bias)

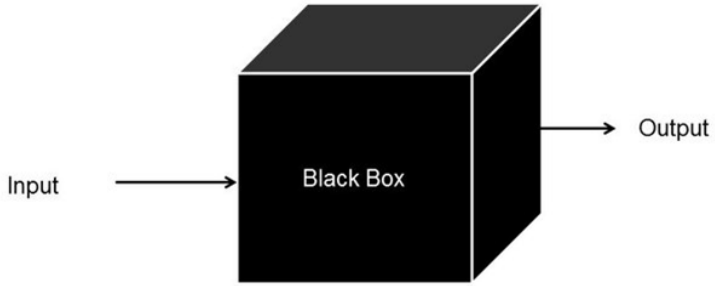
Method – Lack of interpretablity

- A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification



Method-Lack of interpretablity

- A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification



- Can we use non-interpretable models when we make predictions about sensitive attributes, about complex experimental conditions that require broader world knowledge?
- **It is important to have classification or decisions be explainable/interpretable in the sensitive areas; settings in which this interpretability is crucial today: medical diagnostics, social science research that explores sensitive questions**

Evaluation – Huge Cost of Misclassfication

- Accuracy: 81% for men, 74% for women
- No ML model is 100% accurate, and all are biased to training data.
- In these settings, misclassification is expensive to individuals misclassified, and it can affect their lives.
- Note that since the social contract was violated at the stage of research question definition, fixing the problem of bias and other problems would not help making this study acceptable.

Learn to Assess ML/AI Systems Adversarially

- Who could benefit from such a technology?
- Who can be harmed by such a technology?

- Representativeness of training data? Any biases in the training data?
- Collected training data reflecting the true distribution in the real world?
- Could sharing the data have major effect on people's lives?

- Does the system optimize for the "right" objective?
- Could prediction errors have major effect on people's lives?

Learn to Assess AI Systems Adversarially

- Who could benefit from **your** technology?
- Who can be harmed by **your** technology?

- Representativeness of **your** training data
- Could **you** by sharing this data have negative effect on people's lives?

- Does **your** system optimize for the "right" objective?
- Could prediction errors of **your** technology have major effect on people's lives?

Topics in the Intersection of Ethics and ML (AI)

Fairness

Privacy

Responsible Machine Learning Beyond Accuracy

Transparency

Explainability

Research Topics in the Intersection of Ethics and ML

- Fairness-aware ML
- Privacy-preserving ML
- Explainable and Transparent ML

Next Lecture in Week 11

- Fairness-aware ML
- Privacy-preserving ML
- Explainable and Transparent ML

References

- Yulia Tsvetkov and Alan W Black, CMU Computational Ethics for NLP http://demo.clab.cs.cmu.edu/ethical_nlp/

- Yulia Tsvetkov, Vinodkumar Prabhakaran and Rob Voigt, Socially Responsible Natural Language Processing, The Web Conference 2019

- Emily M. Bender, Ethics in NLP, http://faculty.washington.edu/ebender/2017_575/

- Graeme Hirst, Social Impact of Information Technology http://www.cs.utoronto.ca/~gh/cscD03/lectures.shtml