CS 6604
Data Challenges in Machine Learning

Instructor: Ismini Lourentzou

Assistant Professor

Computer Science, Virginia Tech

https://isminoula.github.io/

Logistics

- Class: Tuesdays and Thursdays, 9:30-10:45 am EST
- Office hours: Tuesdays and Thursdays 11:30-12:30 am EST
  (starting next week)

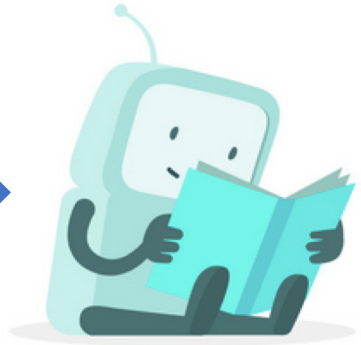
- **Web:** https://isminoula.github.io/cs6604SP21/
- **Piazza:** https://piazza.com/vt/spring2021/cs6604
- **Slack**: cs-vt.slack.com → **cs6604dcml** channel
- **Instructor Email:** ilourentzou@vt.edu **Title: [CS6604]**

Student ordering during office hours: type your name in the chat as soon as you enter the Zoom room.
For one-on-one interactions with the instructor, please post a private note on Piazza or use Slack.

Machine Learning

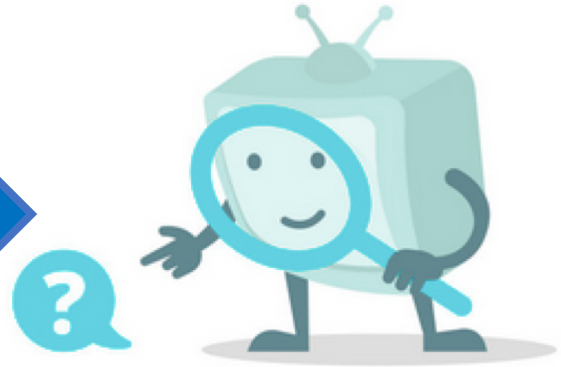
Designing systems that can learn from data.

Training Data

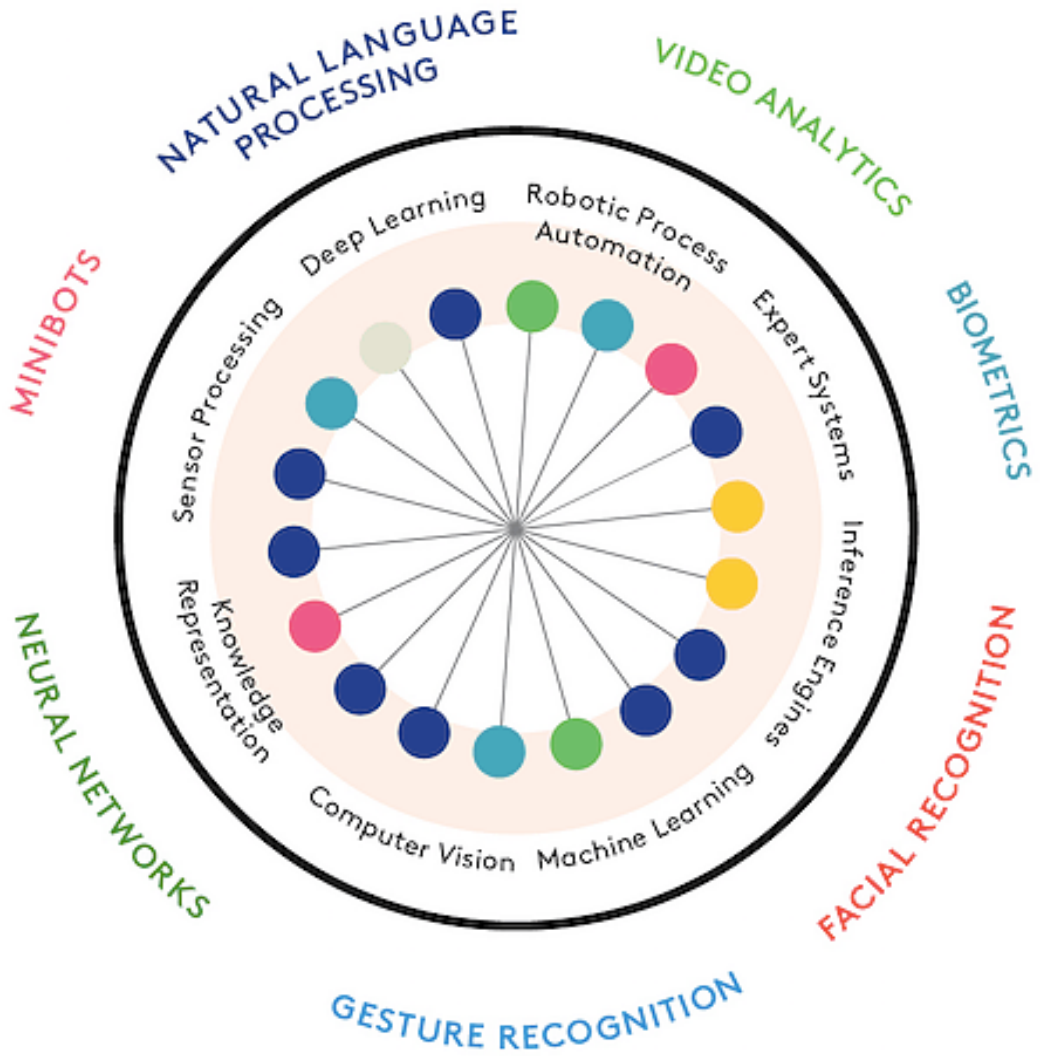


model/ learn

Testing Data



model/ predict

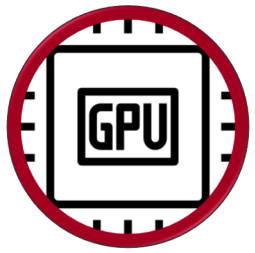


https://www.weforum.org/agenda/2017/12/why-ai-is-entering-its-golden-age

Deep Learning trends in research

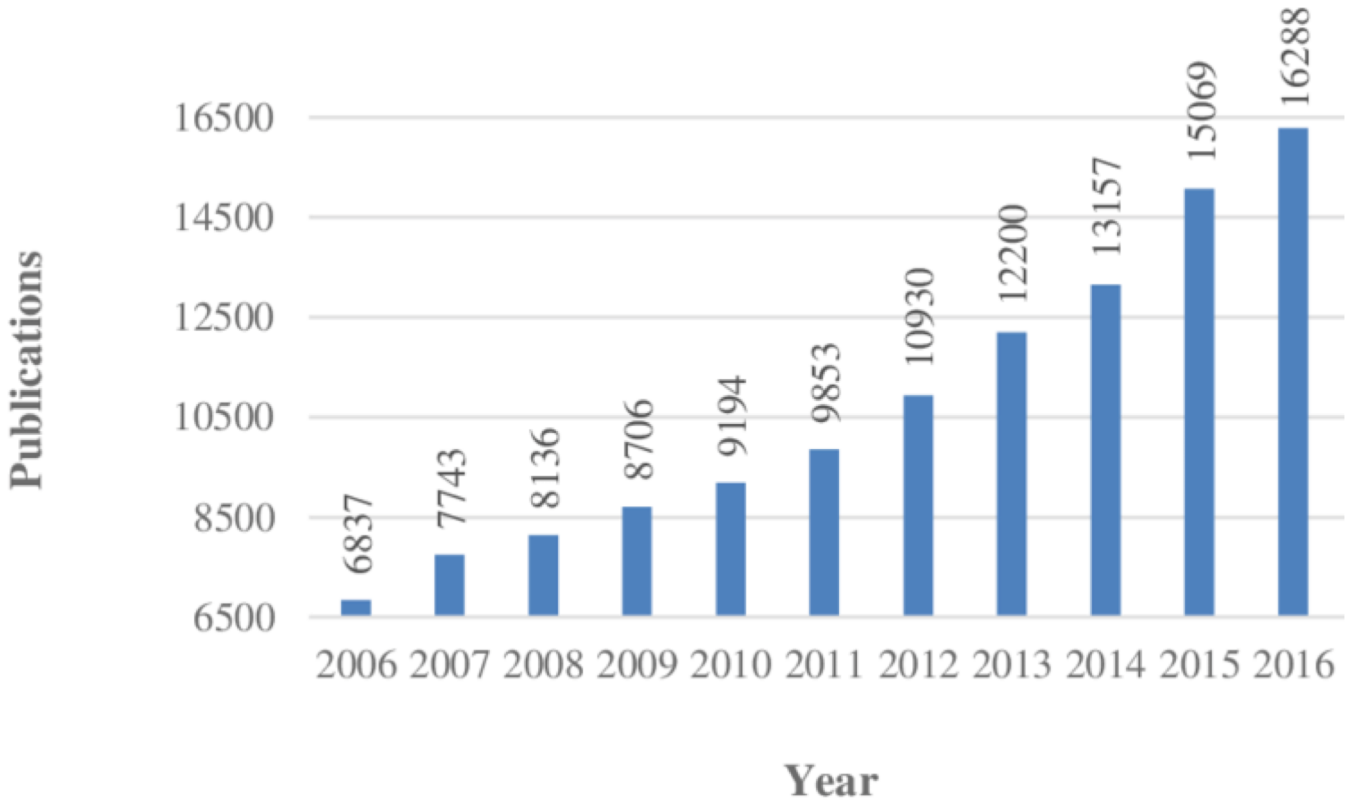
Data

large quantities
of academic training data



Computing Power

better infrastructure



Growth of the number of publications in Deep Learning [1]

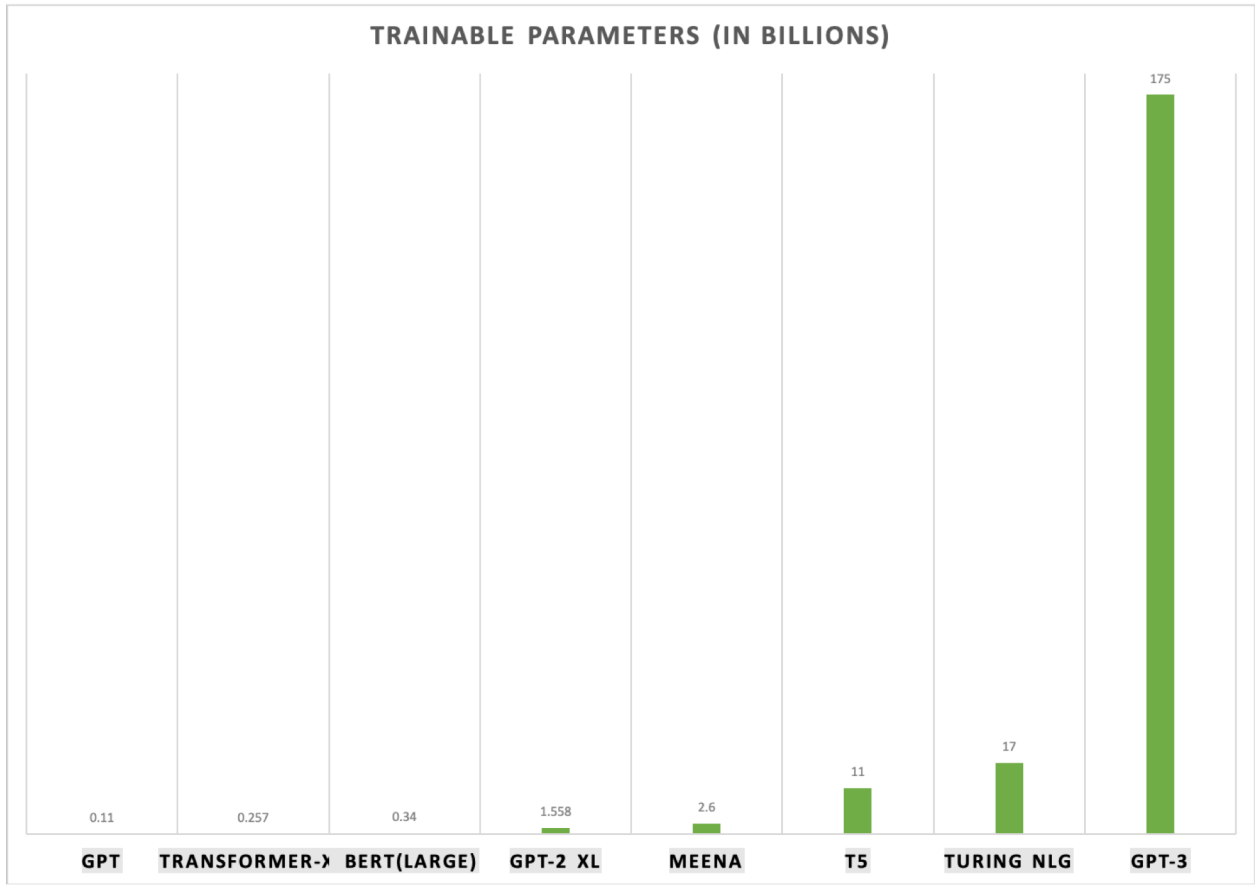
[1] Vargas, Rocio, Amir Mosavi, and Ramon Ruiz. "Deep learning: a review", Advances in Intelligent Systems and Computing 2017

A fun analogy (gardening)

- **Seeds** = Algorithms
- **Nutrients** = Data
- **Gardener** = You
- **Plants** = Programs

Source: Pedro Domingo's CS446 ML

Why Data Challenges? (Scalability)



Language Model hyper-parameters

GTP-3 training data
45 TB text data

Dataset	**Quantity (tokens)**	**Weight in training mix**	**Epochs elapsed when training for 300B tokens**
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

https://in.springboard.com/blog/openai-gpt-3/

Why Data Challenges? (Bias)

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Figure from [1]: Occupations as projected on to the she−he gender direction
Occupations such as businesswoman, where gender is suggested by the orthography, were excluded

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

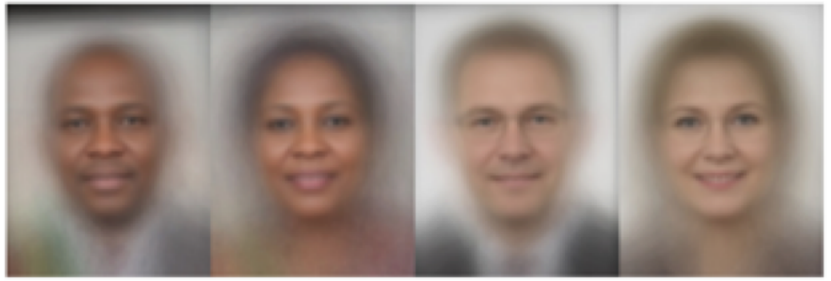


Figure from [2]:
Least accuracy classifying **darker females**
Highest accuracy classifying lighter males
Discrepancy as large as **34.4%**

[1] Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *Advances in neural information processing systems* 29 (2016): 4349-4357.
[2] Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Conference on fairness, accountability and transparency*. 2018.

Data in the real world (More challenges!)

Incomplete

Missing
Aggregated
Private

Linguistic variations
Outliers
Faulty sensors
Biases

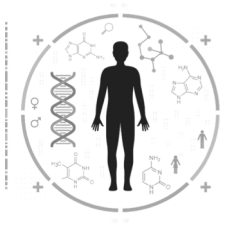
Noisy

Inconsistent

Duplicates
Discrepancies
Heterogeneous

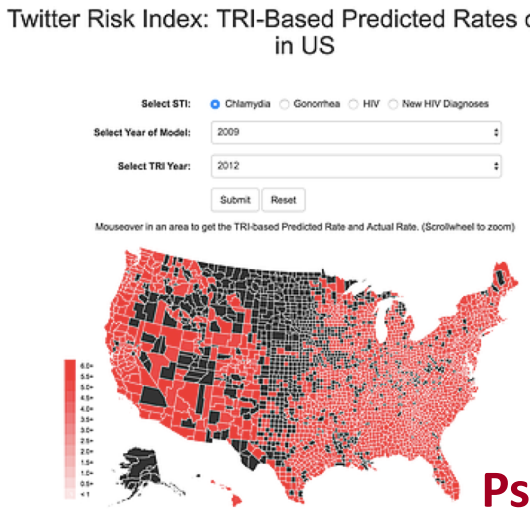
https://www.ab2l.org.br/wp-content/uploads/2018/01/machine_learning.png

Data is a critical bottleneck

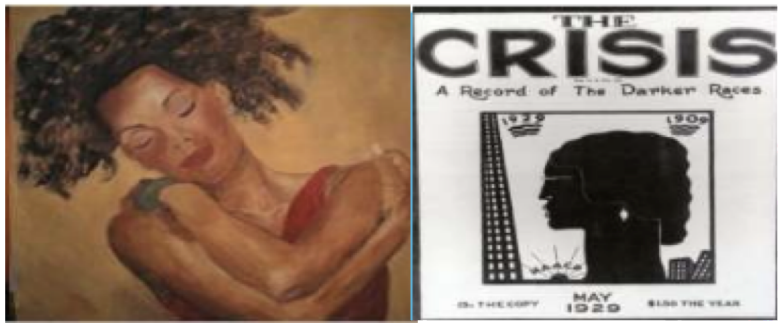


Not enough training data in **new domains**

- Example: new viruses or deceases



Psychology



https://www.hpcuserforum.com/presentations/austin2016/BlackLivesShort.pdf

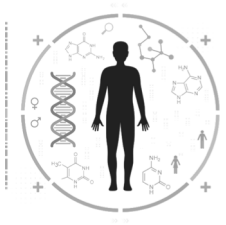
Social Sciences

https://pages.semanticscholar.org/coronavirus-research

Public Health

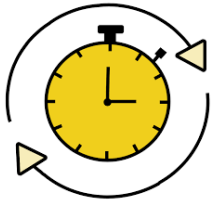
https://static.wixstatic.com/media/e3eff4_1b3129903e0d4a2494ad645a38a7b2e7~mv2_d_1788_1482_s_2.png/v1/fill/w_414,h_343,al_c,q_90,usm_0.66_1.00_0.01/e3eff4_1b3129903e0d4a2494ad645a38a7b2e7~mv2_d_1788_1482_s_2.webp

Data is a critical bottleneck



Not enough training data in **new domains**

- Example: new viruses or deceases

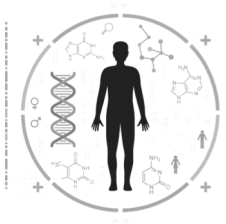


Preparing data for ML is **resource-demanding** and **expensive**

- Collecting, cleaning, feature engineering

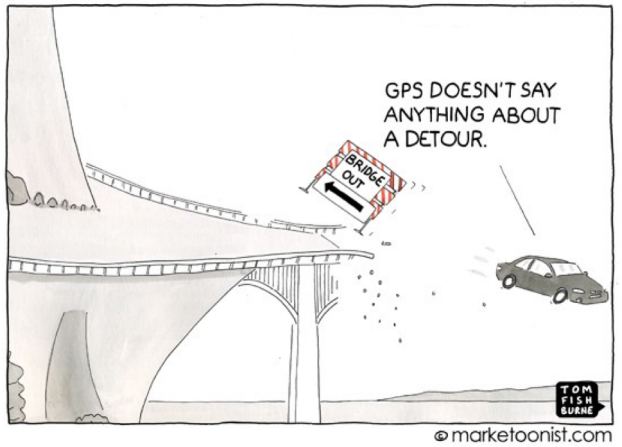
https://medium.com/@SeattleDataGuy/good-data-quality-is-key-for-great-data-science-and-analytics-ccfa18d0fff8

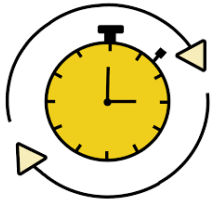
Data is a critical bottleneck



Not enough training data in **new domains**

- Example: new viruses or deceases





Preparing data for ML is **resource-demanding** and **expensive**

- Collecting, cleaning, feature engineering

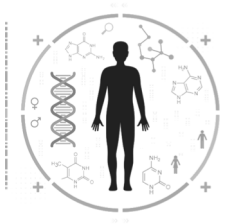
Even with massive data, **quality** is necessary

Low quality data → Poor decisions

- Incorrect models can mislead to incorrect decisions
- Potential critical issues in industrial applications

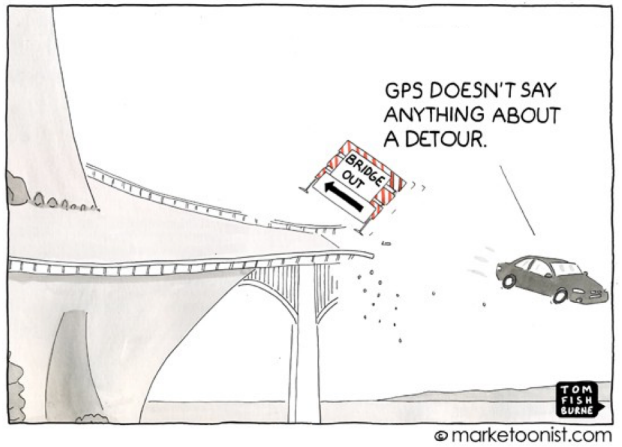
https://medium.com/@SeattleDataGuy/good-data-quality-is-key-for-great-data-science-and-analytics-ccfa18d0fff8

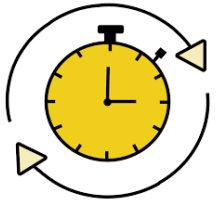
Data is a critical bottleneck



Not enough training data in **new domains**

- Example: new viruses or deceases





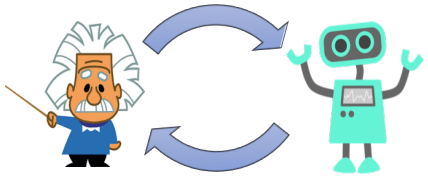
Preparing data for ML is **resource-demanding** and **expensive**

- Collecting, cleaning, feature engineering

Even with massive data, **quality** is necessary

Low quality data → Poor decisions

- Incorrect models can mislead to incorrect decisions
- Potential critical issues in industrial applications

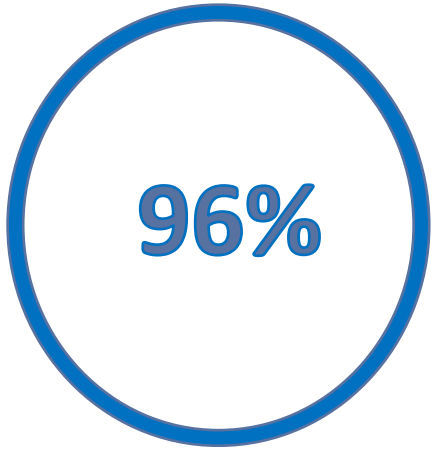


Learning is **continual** and **interactive**

Data Challenges in Industry

Cost of poor data quality US business, each year

Lack of data is the 2nd highest barrier

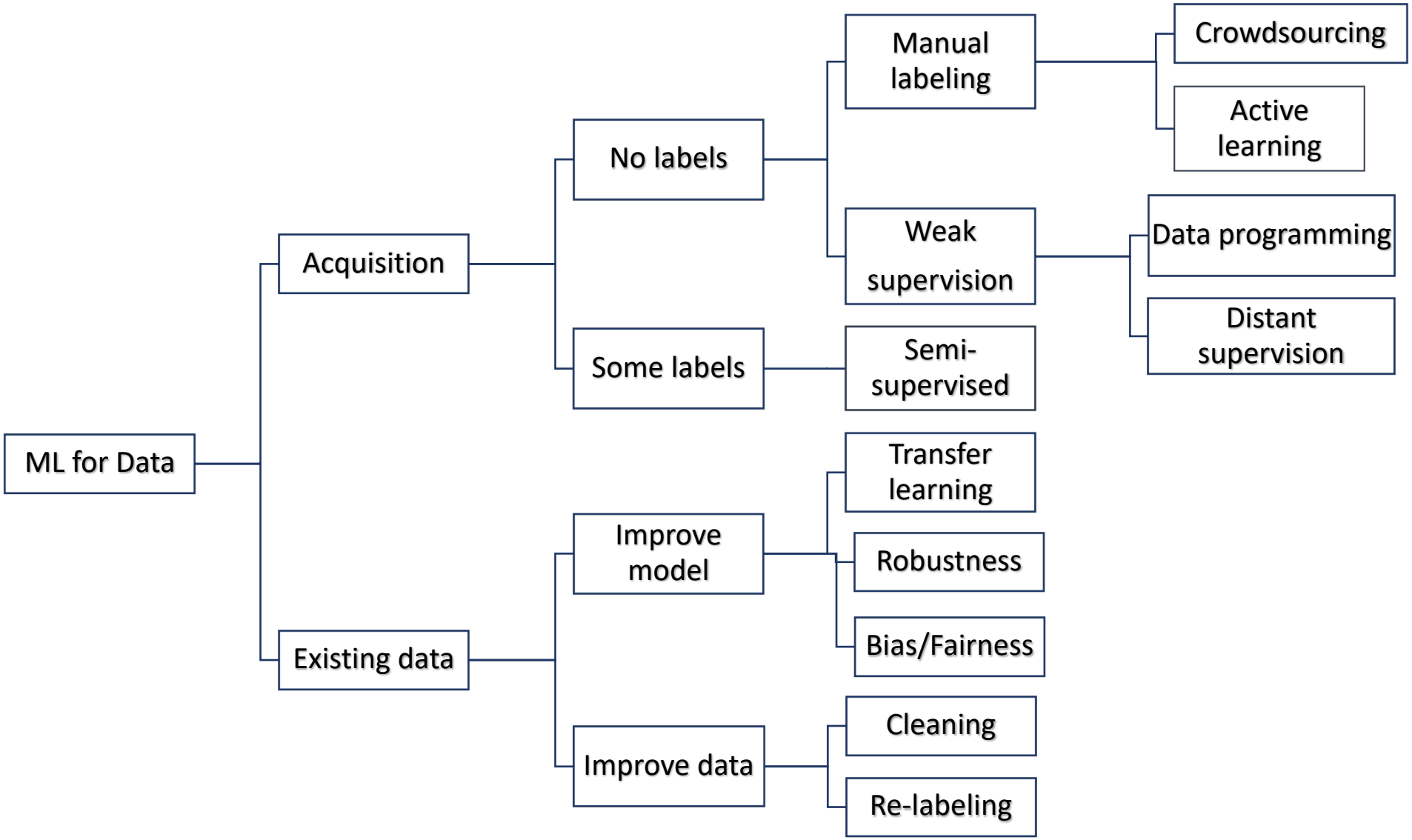


Companies run into data **quality** problems

*data scientists and technology experts in financial institutions with more than $1bn in revenue

https://www.refinitiv.com/perspectives/wp-content/uploads/2019/06/2-Machine-learning-for-risk-and-returns.png
https://www.celsiusinternational.com/need-customer-engagement-data-partner/
https://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data
https://www.roboticsbusinessreview.com/ai/almost-80-of-ai-and-ml-projects-have-stalled-survey-says/

ML areas on data challenges



What is this course about?

- Explore recent advances that address data challenges
  - Active Learning, Semi-supervised Learning, annotation noise
  - Weak Supervision and Self-supervision
  - Data Augmentation and Adversarial Training
  - Data bias and fairness (e.g., selection, confirmation and confounding biases)
  - Class imbalance, skewed distributions and class miss-match
  - Outliers, out-of-distribution instances
  - Missing attributes/values
  - Robustness, generalization and interpretability

- Obtain thorough understanding of these methods
  - Advance research on ML
  - Apply methods to other areas

Prerequisites

- Experience with Machine Learning, Data Analytics or Deep Learning
- Familiarity with Linear Algebra, Statistics and Probability
- Design and implementation of ML models
  - *Ideally PyTorch, Tensorflow, Keras, etc.*
- **Extract key concepts and ideas from reading ML conference papers**

Not sure? Contact me!
Piazza → Slack → Email

Course Structure

- Reading, presenting and discussing weekly papers
- **Everyone** expected to have read the paper prior to class
- Two groups, one on Tuesdays and the other one on Thursdays.
- Each student in the **presenting group** will be given a rotating **role**
- Formal presentation (**slides ~10 mins**): your assigned **role** determines what you should include in the slides.

lens through which student reads paper

https://ipt.imgix.net/205419/x/0/best-lenses-for-landscape-photography-in-2020-3.jpg

Reading Roles (presenting)

Two (2) students will team up
Pairings can change in each class

- **Presenter:** Create the main presentation
  - Motivation, problem definition, method, experimental findings

- **Archaeologist:** Previous and subsequent work report
  - Older paper that substantially influenced current paper
  - Newer paper citing current paper

- **Industry Expert:** Propose new application or company product based on paper
  - Discuss positive and negative impact of this application.
  - Convince your industry boss that it's worth investing time and money to implement this paper.
  - With arguments particularly applicable to the chosen industry market.

- **Hacker:** Implement a small part of the paper
  - On a small dataset or toy problem or
  - Any other simplified version of the paper.
  - Share a Jupyter Notebook with code
  - **DO NOT** simply download and run an existing implementation
  - You can use existing implementations for "backbone" code (build model, load data, train, etc.) **CITE**

Jupyter Notebook
No slides

Reading Roles (presenting)

Designed for solo work, 1 student

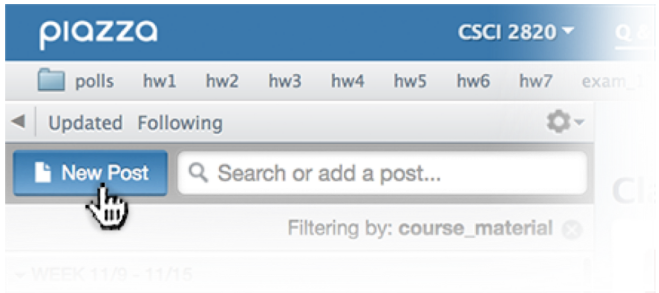
- **Reviewer:** Complete review of the paper.
  - Follow NeurIPS review questions 1-6 under "Review Content"
  - Assign Overall score (question 9) + Confidence score (question 10)

- **Researcher:** Propose follow-up project that has become possible due to the existence and success of the current paper

- **Ethicist:** You are an ethical impact assessor from 2021 (or even 2051). What has been the impact (good or bad) of this paper on the economy, society, and/or the environment?

*Depending on changes in course enrollment, the roles might **change**.*
Remove roles or make roles optional in case enrollment decreases.
Allow groups of two students for all roles in the event of enrollment increase.

Everyone, every week



- Post your thoughts on **Piazza**, e.g.:
  - Which parts did you enjoy reading?
  - What results and insights did you find interesting?
  - Can you propose a missing result the paper could have included?

- **Important**: **"Weekly assignment: Paper title"**
  - Assignment completion checks will be done automatically

- Like it? Thumps up! Endorse student's posts
  - You can also post a reply with your additional thoughts.

- By **9 pm** on the **same day** of class session

Final projects

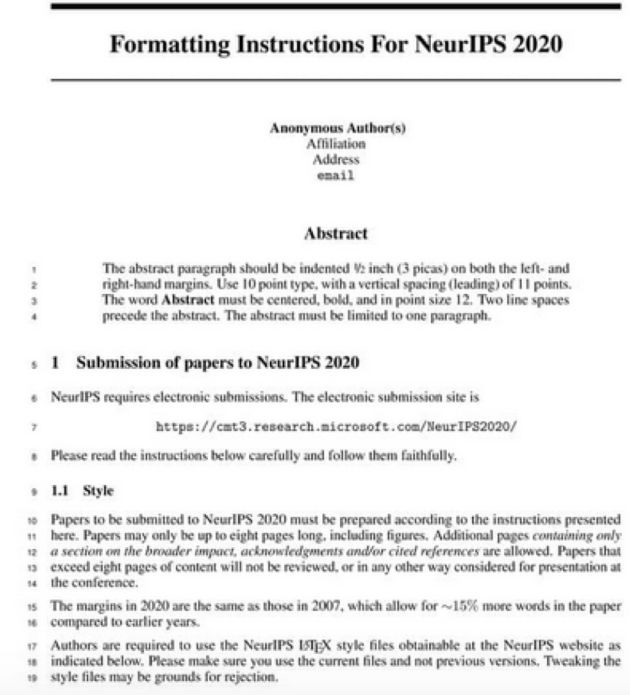
1. Extend papers from topics covered in class
2. Experimentally demonstrate any limitations of related work
3. Suggest improvements by applying the methods to public datasets

Work in groups <= 3 members

- Work produced *proportional* to number of team members
- Include "*contributions*" section in final project report

Report: research paper in a standard conference paper format
https://www.overleaf.com/latex/templates/neurips-2020/mnshsmqkjsqz

*Please familiarize yourself with **GitHub**, **LaTex** and paper **writing***



Formatting Instructions For NeurIPS 2020

Anonymous Author(s)
Affiliation
Address
email

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Submission of papers to NeurIPS 2020

NeurIPS requires electronic submissions. The electronic submission site is

`https://cmt3.research.microsoft.com/NeurIPS2020/`

Please read the instructions below carefully and follow them faithfully.

1.1 Style

Papers to be submitted to NeurIPS 2020 must be prepared according to the instructions presented here. Papers may only be up to eight pages long, including figures. Additional pages *containing only a section on the broader impact, acknowledgments and/or cited references* are allowed. Papers that exceed eight pages of content will not be reviewed, or in any other way considered for presentation at the conference.

The margins in 2020 are the same as those in 2007, which allow for ~15% more words in the paper compared to earlier years.

Authors are required to use the NeurIPS LaTeX style files obtainable at the NeurIPS website as indicated below. Please make sure you use the current files and not previous versions. Tweaking the style files may be grounds for rejection.

Final projects

1. Extend papers from topics covered in class
2. Experimentally demonstrate any limitations of related work
3. Suggest improvements by applying the methods to public datasets

Projects hosted on GitHub https://github.com/CS6604VT

Written report (research paper) + Jupyter Notebook

Example: http://nlp.seas.harvard.edu/2018/04/03/attention.html

Final presentations during the last two class sessions (PowerPoint or LaTex slides)

List of suggested topics

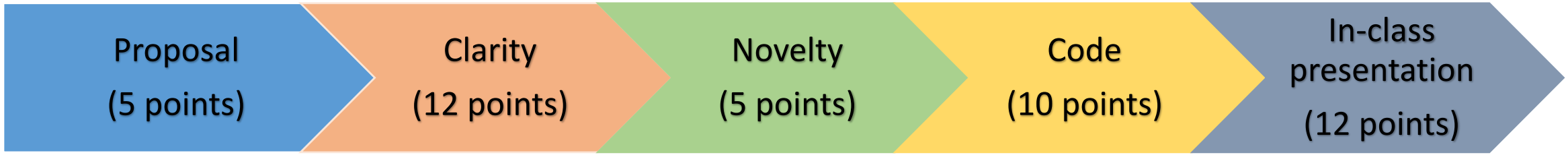
Details about the project proposal

Evaluation

Readings: 56 points

12 sessions x 3.5 points each time you present (all presenting roles considered equal)
12 sessions x 1 point Piazza assignment + Class participation
+ 2 points peer-review

Final Project: 44 points



Extra credit: up to 3 points to the most well-made presentation and notebook

Attendance & Late work

- "Presenting" role:
  1. Create presentation for your assigned session
  2. Find someone else to present for you before the day of the presentation

Please make arrangements to avoid disrupting the class

12 points penalty otherwise

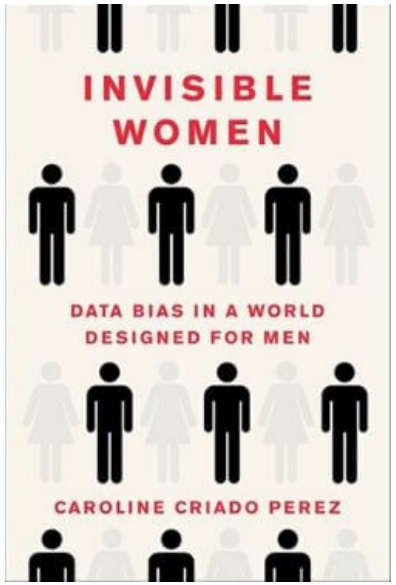
- Miss non-presenting assignment: 0 points for the assignment ☹
- Cannot postpone **final project presentations** (switch your timeslot with another group)
- Other material are negotiable (final project submission, project report, etc.)
  - Based on the severity of the request, e.g., medical reasons.

✓ Anonymous feedback: https://virginiatech.qualtrics.com/jfe/form/SV_6u4OIe19cosSYzc

✓ Students with disabilities: contact me + Services for Students with Disabilities office

What is Coming Up in Next Class?

Ethicist: You are an ethical impact assessor from 2021 (or even 2051). What has been the impact (good or bad) of this paper on the economy, society, and/or the environment?



Diversity & inclusion in ML

https://www.sundance.org/projects/code-for-bias

https://www.goodreads.com/book/show/41104077-invisible-women

LXAI

Black in AI

A place for **sharing** ideas, **fostering** collaborations and **discussing** initiatives to increase the presence of **Black people** in the field of **Artificial Intelligence**.

blackinai.org/

WiML
Women in Machine Learning

{DIS}ABILITY IN AI

What is Coming Up in Next Class?

Ethicist: You are an ethical impact assessor from 2021 (or even 2051). What has been the impact (good or bad) of this paper on the economy, society, and/or the environment?

This Thursday 01/21:

Bring a relevant blog post, news article or paper about diversity & inclusion and bias in ML/NLP/CV/DS to discuss in class

Next class: Survey of Active Learning + some of my work (presented by me)

Action items (for you)

1) Divide into groups:

- Tuesdays: 11 students
- Thursdays: 12 students

https://tinyurl.com/cs6604groups

- By tomorrow night

Wednesday 01/20/2021

B	C
Tuesdays	Thursdays
Name LastName	Ismini Lourentzou

2) Background survey (with an option for force-add requests)
To be completed by 01/31
https://tinyurl.com/cs6604survey

Action items (for me)

Paper list for each sub-topic (link in course web page)

- Active Learning, Semi-supervised Learning, annotation noise
- Weak Supervision and Self-supervision
- ….

Updated by this Friday

Date	Reading
Tuesday, 01/19/2021	Course introduction & logistics
Thursday, 01/21/2021	Active Learning
Tuesday, 01/26/2021	Active Learning
Thursday, 01/28/2021	Semi-supervised Learning
Tuesday, 02/02/2021	Semi-supervised Learning
Thursday, 02/04/2021	Self-supervision
Tuesday, 02/09/2021	Self-supervision
Thursday, 02/11/2021	Weak supervision
Tuesday, 02/16/2021	Weak supervision
Thursday, 02/18/2021	Data Augmentation
Tuesday, 02/23/2021	
Thursday, 02/25/2021	Adversarial Training