Lecture 7:
Bias, fairness, diversity in data

CS698Y: Human AI Interaction

Prelude

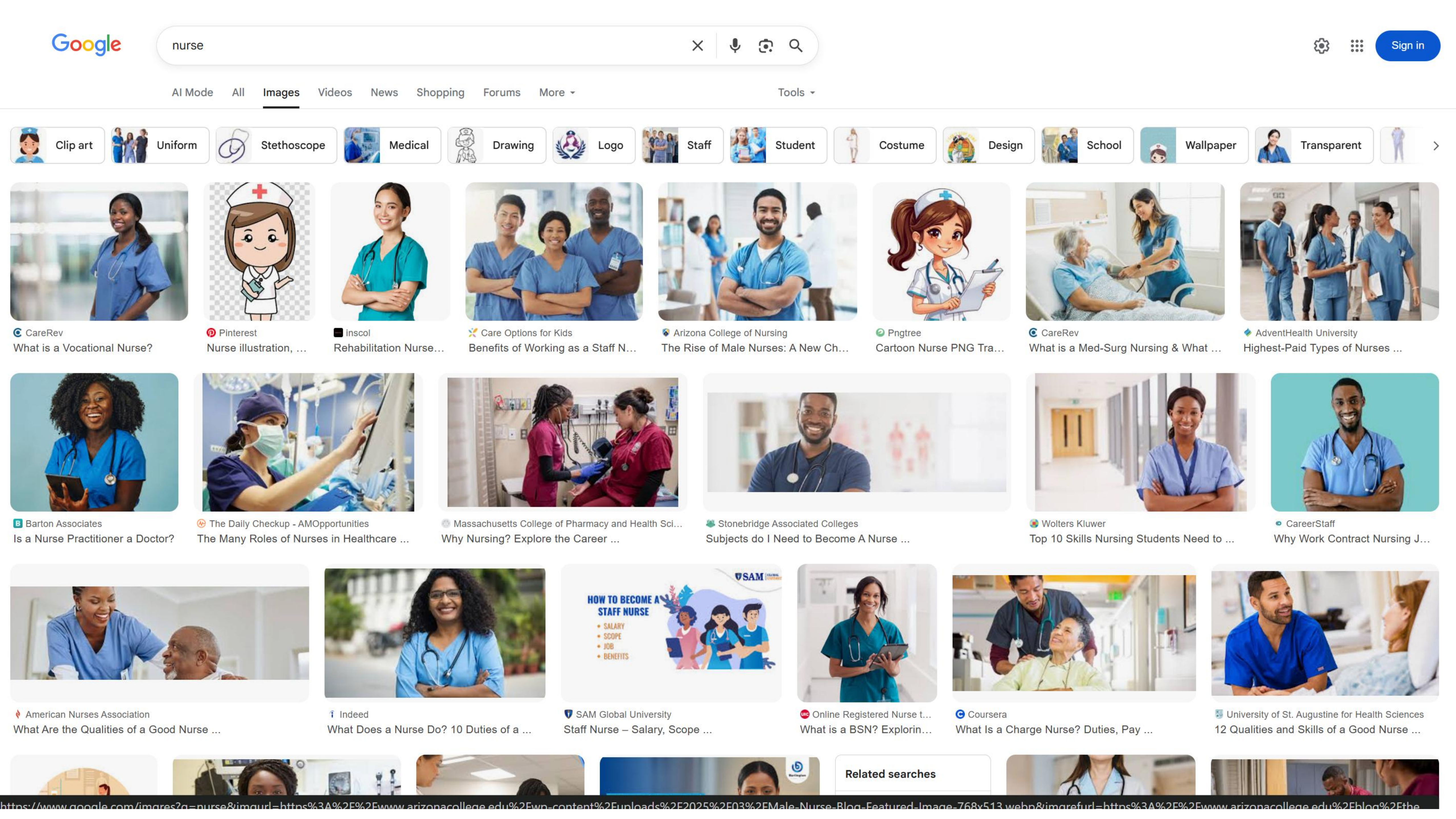
- Principles in Human-AI interaction design
  - HAX guidelines
  - Cognitive principles
  - Sociological principles

- Going forward
  - Designing according to guidelines/principles in practice
  - Hands-on exercises / in-class activities
  - Today → Bias, fairness, diversity in data
  - Why care, details & hands-on explorations

What is Bias?

- Systematic error / skew / preference in data, model, or outcome
- Something is being *over- or under-represented*

- Examples:
  - Speech recognition working poorly for some languages/accents → because data is underrepresented
  - Image classifier mislabels some dog breeds without collar as fox → not enough training data for that kind
  - Model misdiagnoses cardiac diseases in middle aged South Asian women → not enough data on South Asian women


Google
nurse
Sign in
AI Mode
All
Images
Videos
News
Shopping
Forums
More
Tools
Clip art
Uniform
Stethoscope
Medical
Drawing
Logo
Staff
Student
Costume
Design
School
Wallpaper
Transparent
CareRev
What is a Vocational Nurse?
Pinterest
Nurse illustration, ...
inscol
Rehabilitation Nurse...
Care Options for Kids
Benefits of Working as a Staff N...
Arizona College of Nursing
The Rise of Male Nurses: A New Ch...
Pngtree
Cartoon Nurse PNG Tra...
CareRev
What is a Med-Surg Nursing & What ...
AdventHealth University
Highest-Paid Types of Nurses ...
Barton Associates
Is a Nurse Practitioner a Doctor?
The Daily Checkup - AMOpportunities
The Many Roles of Nurses in Healthcare ...
Massachusetts College of Pharmacy and Health Sci...
Why Nursing? Explore the Career ...
Stonebridge Associated Colleges
Subjects do I Need to Become A Nurse ...
Wolters Kluwer
Top 10 Skills Nursing Students Need to ...
CareerStaff
Why Work Contract Nursing J...
HOW TO BECOME A STAFF NURSE
SALARY
SCOPE
JOB
BENEFITS
SAM
American Nurses Association
What Are the Qualities of a Good Nurse ...
Indeed
What Does a Nurse Do? 10 Duties of a ...
SAM Global University
Staff Nurse – Salary, Scope ...
Online Registered Nurse t...
What is a BSN? Explorin...
Coursera
What Is a Charge Nurse? Duties, Pay ...
University of St. Augustine for Health Sciences
12 Qualities and Skills of a Good Nurse ...
Related searches
https://www.google.com/imgres?q=nurse&imgurl=https%3A%2F%2Fwww.arizonacollege.edu%2Fwp-content%2Fuploads%2F2025%2F03%2FMale-Nurse-Blog-Featured-Image-768x513.webp&imgrefurl=https%3A%2F%2Fwww.arizonacollege.edu%2Fblog%2Fthe

Google

programmer

AI Mode | All | Images | Books | Videos | Forums | Short videos | More | Tools

Sign in

Wallpaper | Coding | Clipart | Software | Design | Computer | Icon | Illustration | Job | Logo | Animated | Vector | Developer | Avatar


Freepik
Programmer Images - Free Download o...


Springboard
What Exactly Does a Programmer Do ...


Forbes
How To Become A Computer Programmer: ...


Online Courses UK
How to become a Programmer ...


Epitech
9 characteristics of a good programmer ...


Herzing University
What Does a Computer Programmer Do?


Indeed
How To Become a Computer Programme...


Alpha Academy
how can I become a computer programmer ...


GogoTraining
a Computer Programmer ...


stylecnc
How To Become A CNC Programmer...


Unsplash
Computer Programmer Pictures | Downlo...

LinkedIn
What is a Programmer?


Coursera
Programmer vs Developer: Job Roles ...

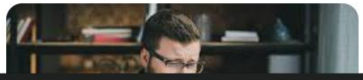

Built In
Software Engineer Vs. Programmer: 6 K...

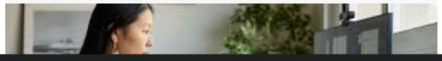

Spiceworks
Computer Programmer Job Description ...

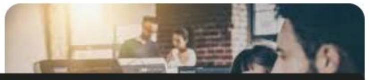

Freelancer
guide to hiring a programmer ...


FlexJobs
Computer Programmer or Developer ...


IT Talent
How to hire a programmer: 12 ...





https://www.google.com/imgres?q=programmer&imgurl=https%3A%2F%2Fimg.freepik.com%2Ffree-photo%2Frear-view-programmer-working-all-night-long_1098-18697.jpg%3Fsemt%3Dais_hybrid%26w%3D740%26q%3D80&imgrefurl=https%3A%2F%2Fwww.f...

What is Fairness?

- A *goal/criterion* about equitable treatment and outcomes across groups (doesn't disadvantage one group)
- "Are different groups treated appropriately and justly by the system?"
- Example:
  - Data biased "brown, no collar" = fox
  - Fairness: all dogs are equally accurately recognized as dogs (here, unfair towards brown no-collared ones.
- Goal → ensure fairness even if data is skewed

How to reduce bias & ensure fairness?

- Diversity
  - Training data has many collared dogs, few uncollared dogs.
  - Lack of representation of *uncollared brown dogs* reduces diversity.
  - Broader diversity → more robust AI

- Inclusivity
  - Designing systems so all groups can participate meaningfully
  - Goes beyond representation → asks: who benefits? who is left out?
  - Example → not all pet owners put collars on dogs.

Why bias matters?

- Models reflect, reinforce and amplify these human biases
- Not always intentional — but with real-world effects
  - Lower trust in technology
  - Amplifies biases already existing in society
- Important to guard against these!

- For AI developers / researchers / statisticians → source of interesting problems!

How Bias Enters Data?

- Sampling bias (not enough representation)
  - Data doesn’t have enough collared brown dogs
- Labeling bias (human annotators’ perspectives)
  - Labeler thinks brown uncollared ones in woods are foxes
- Historical bias (reflecting unfair past decisions)
  - Historically, village dogs were collared, some breeds not domesticated
  - Model: dogs =domesticated & collared, fox=in the wild
- Measurement bias (sensors, tools skewed)
  - Camera errors, lighting, blurs, …

How do we mitigate biases?

- Carefully choose subsets of data for training/testing
- Avoid over-representing some groups
- Specific techniques / strategies
  - Resampling
  - Balancing
  - Feature awareness
  - Fairness metrics

Resampling

- Problem: Some groups over- / under- represented in data
- How do you balance?
- Oversampling → Increase minority group data
  - Duplicate minority group data
  - Synthesize training data for minority group
    - Statistical → create distributions and then sample from that
    - Interpolation and extrapolation → based on two or more nearest points
    - Model-based sampling → fit a model (e.g., linear regression) and pick points
    - ML models to generate data → GAN, VAE, LLM, …
- Undersampling → drop data from majority group (rarely!)
  - Cluster and then sample from each cluster

Balancing

- When one group is too small, stratify training set split
  - Instead of random 80%, take appropriate proportions for train/test
  - “Don’t see too many foxes when training, and too many dogs when training”
- Weight loss functions
  - Small groups → higher penalty on errors [weight = 1/frequency]
- K-fold cross validation → stratified each time
- Check fairness → accuracy same across groups

Labeling bias

- Provide detailed instructions
- Ensure diverse labelers at start → be aware of your labelers
- Get everything labelled by more than one person
  - Multiple annotators + consensus reduces subjectivity
- Bias-aware training → treat certain labels with caution (e.g., in terms of weighting)

What about historical bias?

- Awareness & auditing: Examine datasets for underrepresentation or historical patterns.
- Rebalancing / reweighting: Give more weight to underrepresented groups or outcomes.
- Counterfactual or synthetic data: Generate data to represent fairer scenarios.

Calibration / Measurement Bias

- Bias from sensors, instruments, or measurement errors.
  - Example: Facial recognition for darker skin tones / evening images
- Improve measurement quality: Better sensors / procedures.
- Normalize / preprocess inputs: Reduce systematic differences in inputs across groups.
- Calibration techniques: Adjust model outputs so probabilities match reality per group.
- Continuous monitoring: Track performance across groups and recalibrate regularly.

Feature Awareness

- Be aware of *sensitive attributes* (gender, race, age, caste, …)
- Sometimes you want to exclude them (might contribute to bias)
- Sometimes you want to include them (for fairness)
- How do you know what to do?
    - Ask the domain expert / research the domain
    - Look for difference in accuracy across groups / ranges
    - Eliminate features to see if it reduces these gaps
    - Sometimes this can bring down overall accuracy
- To include, or not to include—that is the question!

Fairness metrics

- Statistical / demographic parity: Equal outcomes across groups
  - 90% Indians predicted “at risk”, 90% Americans also should be.
  - Unfair in some cases → what if one race is low risk compared to others?
- Equal opportunity / equal TP: Same true positive rate across groups
  - Among those who qualify as positive, the algorithm predicts +ve for same % in all groups.
  - Among all qualified, % men predicted qualified = % women predicted qualified
- Predictive parity: Same precision across groups
  - Among predicted as “qualified”, same % of qualified people across groups
  - Ensures trustworthiness of models for various groups
  - Some group’s yes is often wrong → people don’t pay attention to model

Fairness metrics

- Equalized odds: Equal true positive & true negative rates
  - Model doesn’t make more mistakes for one group than another
  - Example: If 80% of creditworthy men are approved (TPR) and 10% of non-creditworthy men are incorrectly approved (FPR), then the same rates should hold for women.
- Treatment Equality: Ratio of false negatives (FN) to false positives (FP) is equal across groups
  - Example: In medicine, we want to avoid FN, over FP
  - If FN/FP = 1/100 for Caucasians, should also be 1/100 for Indians.

Individual-based fairness tests

- Counterfactual fairness: same outcome if you swap a protected attribute, all else remaining fixed
  - Options to explore “What if” / counterfactuals typically provided in explainable systems.
  - Important to also ensure trust in systems
- Fairness through awareness (within group): Similar individuals (by relevant features) should have similar outcomes
  - Easily seen through clusters

Diversity / representation metrics

- Proportional representation – Each group’s representation matches its share in the population
- Ensure similar error rates, positive / deserved outcomes for each group
- Coverage / exposure – how much coverage do we get / how much exposure do we get
    - E.g., recommendations of just one kind than variety?

Some other metrics (Regression)

- Statistical difference / mean difference – Compare means of predictions or errors across groups
- Correlation with sensitive attributes – Measures how strongly predictions depend on protected features.

Fairness is contextual & conflicted

- No one size fits all fairness metric
- There are tensions between metrics
  - E.g., equal outcomes vs. equal opportunities
  - Minimize overall error vs. maintain parity in error rates
- We can’t satisfy all at once!
- Need careful thought + expert opinion
- Notion of fairness itself can be unfair
  - Who defines it? Who is affected by it?

In summary…

- Bias in data → bad, less robust models
  - Users lose trust
  - Social consequences
- Bias occur due to data sampling, labelling, calibration and historical reasons
- Systematic ways of minimizing them in data
- Focus on other metrics to also measure fairness when appropriate!

Homework

- Work in pairs
- Pick one dataset:
  - Multiple your roll numbers’ last digits.
  - If even: Predict Students' Dropout and Academic Success
  - Else: Absenteeism at work
- Train a simple model for the task (use libraries)
- Evaluate the biases in the dataset
- Evaluate the performance of the model (incl. for fairness)
- Use python; turn in Colab / Github links+ report (TBA on HelloIITK)

Readings

- Implications of AI Bias in HRI: Risks (and Opportunities) when Interacting with a Biased Robot
- Humans inherit artificial intelligence biases
- Fair ML Book, https://fairmlbook.org/pdf/fairmlbook.pdf (Chapter-7)

Additional (optional):

- How human–AI feedback loops alter human perceptual, emotional and social judgements

Next class…

- Dr. RS Sharma → Aadhaar and data privacy
- Final project annoncements