



## Conclusion



## Roadmap

Summary of CS221

Next courses

Food for thought

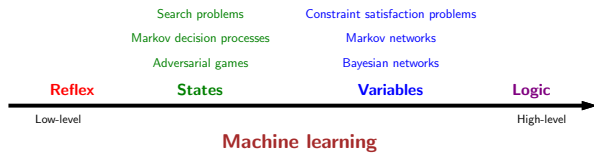
## Paradigm

Modeling

Inference

Learning

## Course plan



CS221

6

## Machine learning

Objective: loss minimization

$$\min_{\mathbf{w}} \sum_{(x,y) \in \mathcal{D}_{\text{train}}} \text{Loss}(x, y, \mathbf{w})$$

Algorithm: stochastic gradient descent

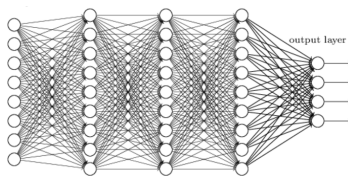
$$\mathbf{w} \rightarrow \mathbf{w} - \eta_t \underbrace{\nabla \text{Loss}(x, y, \mathbf{w})}_{\text{prediction} - \text{target}}$$

Applies to wide range of models!

CS221

8

## Reflex-based models



Models: linear models, neural networks, nearest neighbors

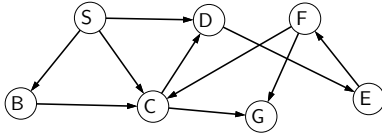
Inference: feedforward

Learning: SGD, alternating minimization

CS221

10

## State-based models



**Key idea: state**

A **state** is a summary of all the past actions sufficient to choose future actions **optimally**.

**Models:** search problems, MDPs, games

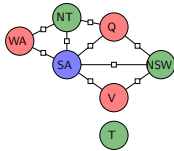
**Inference:** UCS/A\*, DP, value iteration, minimax

**Learning:** structured Perceptron, Q-learning, TD learning

CS221

12

## Variable-based models



**Key idea: factor graphs**

Graph structure captures conditional independence.

**Models:** CSPs, Markov networks, Bayesian networks

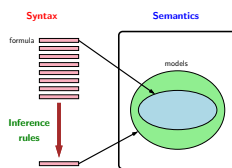
**Inference:** backtracking, forward-backward, beam search, Gibbs sampling

**Learning:** maximum likelihood (closed form, EM)

CS221

14

## Logic-based models



**Key idea: logic**

Formulas enable more powerful models (infinite).

**Models:** propositional logic, first-order logic

**Inference:** model checking, modus ponens, resolution

**Learning:** ???

CS221

16

## Tools

- CS221 provides a set of tools



- Start with the problem, and figure out what tool to use
- Keep it simple!

CS221

18



## Roadmap

Summary of CS221

Next courses

Food for thought

CS221

20

## Overview

List of AI courses:

<http://ai.stanford.edu/courses/>

Types of courses:

- **Methods:** more advanced techniques, general-purpose
- **Applications:** real impact of AI, help you truly understand and appreciate methods
- **Foundations:** invest in building depth (for methods and applications); usually not in AI (math, hardware, linguistics/biology, etc.)

CS221

22

- It is difficult to exhaustively enumerate all the relevant classes and keep it fully up to date, so please take a look at the AI website to get the list of courses.
- In thinking of next classes to take, it might be natural to be drawn to methods classes, but it is always useful to pick an application that you are passionate about to ground yourself.
- Finally, it is easy these days, given how fast the field of AI is moving, to move through the field shallowly. If you want to truly make fundamental advances, take the time to invest in building depth. This usually involves taking classes outside AI (not listed on the AI lab website).



## Methods

CS221

24

## Machine learning



### CS229: Machine Learning

- Standard, more mathematical derivations, continuous variables (e.g., kernel methods, PCA)

### CS230: Deep Learning

- Applied, how to train deep neural networks (e.g., dropout, batch norm)

CS221

26

## Machine learning



### CS329D: Machine Learning Under Distribution Shifts

- Machine learning fails when train  $\neq$  test (e.g., adversarial examples, DRO)

### CS330: Deep Multi-Task and Meta Learning

- How to transfer across multiple tasks (e.g., few-shot learning, meta-RL)

### CS224W: Machine Learning with Graphs

- Data points are graphs or are connected via a graph (e.g., graph neural networks)

CS221

28

## Reinforcement learning



### CS234: Reinforcement Learning

- More advanced techniques (e.g., policy search, bandits, batch RL)

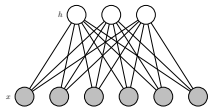
### CS238: Decision Making Under Uncertainty

- Model-based planning, applications to autonomous vehicles, aviation

CS221

30

## Generative models



### CS228: Probabilistic Graphical Models

- More advanced techniques (e.g., belief propagation, variational inference, MCMC, structure learning)

### CS236: Deep Generative Models

- Generative models supercharged with deep learning (e.g., VAEs, GANs)

CS221

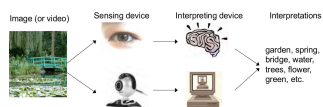
32

*Applications*

CS221

34

## Vision



### CS231N: Convolutional Neural Networks for Visual Recognition

- ML-heavy (convnets, Transformers), detection, segmentation, generation

### CS231A: From 3D Reconstruction to Recognition

- More vision (e.g., cameras + geometry, shape reconstruction, depth estimation)

### CS348I: Computer Graphics in the Era of AI

- Rendering, geometry, animation, computational photography

## Robotics



### CS237[AB]: Principles of Robotic Autonomy

- ML-heavy (RL, imitation learning), grasping, manipulation

### CS223A: Introduction to Robotics

- Physical models for kinematics and control

## Language

### CS224N: Natural Language Processing with Deep Learning

- ML-heavy (RNNs, Transformers), parsing, translation, generation

### CS224U: Natural Language Understanding

- Word representations, grounding, natural language inference, evaluation

### CS224V: Conversational Virtual Assistants with Deep Learning

- Applications to semantic parsing, dialogue state tracking

### CS224C: NLP for Computational Social Science

- Text analysis, applications to social science and sociolinguistics

### CS324: Understanding and Developing Large Language Models

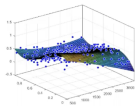
- Social/ethical/legal considerations, scaling laws, hands-on experience

## Foundations

CS221

42

## Optimization, statistics, theory



EE364[AB] / CS334[AB]: Convex Optimization

- Convex optimization problems, duality

STATS 200: Statistical Inference

- Statistical thinking, decision theory, hypothesis testing

STATS 214 / CS229M: Machine Learning Theory

- Why does it work? Uniform convergence, deep learning theory

CS221

44

## Cognitive science and neuroscience



PSYCH204[AB] / CS428[AB]: Computation and Cognition: The Probabilistic Approach


- Human mind (software), using probabilistic programs to model human reasoning and learning [A], language [B]

PSYCH 242 / APPPHYS 293: Theoretical Neuroscience

- Human brain (hardware), neurally-plausible approximation of back propagation, spiking neural networks

CS221

46



## Summary


Types of courses:

- **Methods:** more advanced techniques, general-purpose
- **Applications:** real impact of AI, help you truly appreciate methods
- **Foundations:** invest in building depth (for methods and applications); usually not in AI (math, hardware, linguistics/biology, etc.)

Tips:

- Invest in building depth, take classes outside CS
- Many resources (tutorials, blog posts, talks) online
- Download code, tinker — hands-on learning
- Talk to professors and other students

CS221 48



## Roadmap

Summary of CS221

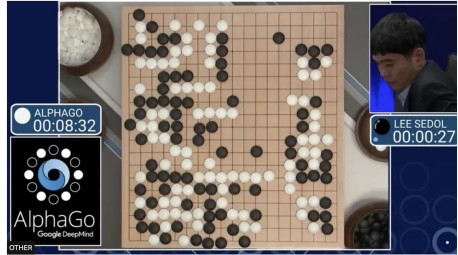
Next courses

**Food for thought**

CS221 50

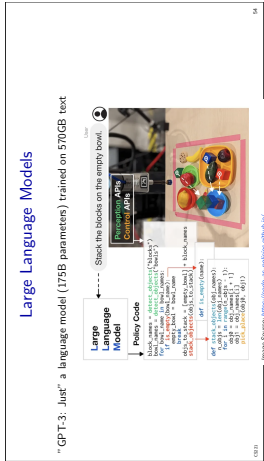
## Artificial intelligence: Google's AlphaGo beats Go master Lee Se-dol

© 12 March 2016



- The last decade has been marked by technological progress and optimism, perhaps quintessentially captured by AlphaGo, a deep reinforcement learning program built by DeepMind, which defeated Lee Sedol at Go, defying the expectations of Go and AI experts alike.

CS221 52



- Or OpenAI's unprecedented massive GPT-3 language model, which boasted impressive generation capabilities, emergent behaviors, These large language models are able to perform many tasks such as generating robot code, generating SQL queries from natural language or answering questions.
- These tasks traditionally required specialized solutions, but now a single model that wasn't even trained specifically for these tasks can do a passable job at them.

## Real-world applications

AI is everywhere: consumer services, advertising, transportation, manufacturing, etc.



AI being used to make decisions for: education, credit, employment, advertising, healthcare and policing

- These advances have caused AI to be developed and deployed into countless different applications across all sectors of society, and this trend will only continue.
- What is the societal impact of this trend?

## Machine translation

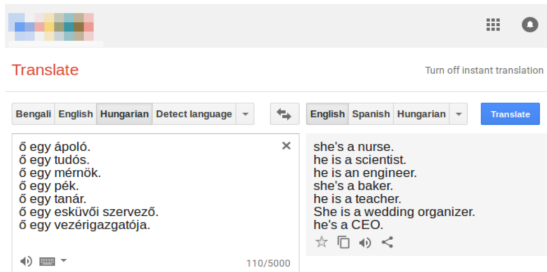
Input sentence:	Translation (PBMT):	Translation (GNMT):	Translation (human):
李克強此行將啟動中加總理年度對話機制。與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held the first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during his visit, and hold the first annual dialogue with Premier Trudeau of Canada.



- Take machine translation for example. Machine translation is just one application whose quality, which has improved significantly due to advances in AI, and has been very enabling for breaking down language barriers and increasing accessibility.

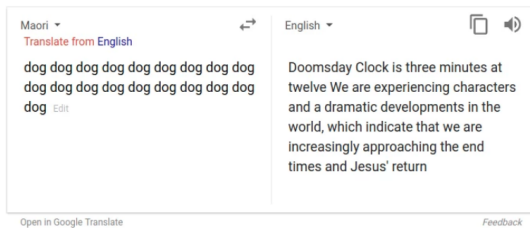
## Biases

[Prates+ 2018]



- While machine translation systems are ubiquitous, they also have lots of problems. Because they are often trained on scraped data, they inherit a lot of the subtle biases present in that data.

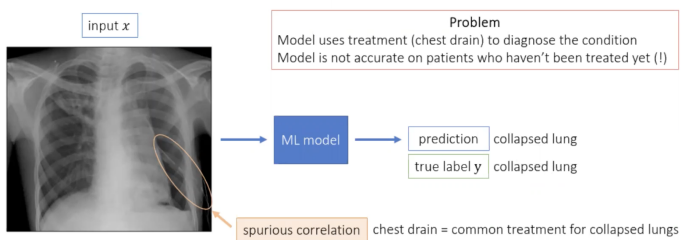
## Craziness



- You can also get some really unexpected behavior out of the state-of-the-art systems based on deep neural networks.
- These systems are garbage in, garbage out.

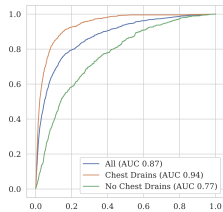
## Spurious correlations

[Oakden-Rayner, Dunnmon, Carneiro, Ré (2019)]



- Many of these issues are due to the fact that machine learning thrives on complex models fitting correlations in data.
- Take the task of predicting whether a chest x-ray is indicative of collapsed lung.
- Apply standard convnet machinery from computer vision and it works reasonably well. But take a closer look: see that thin tube coming out?
- This is a chest drain, which is a common treatment for a collapsed lung. And it turns out this is one of the signals that the model is picking up on.

## Spurious correlations



Subpopulation of untreated patients are worse off than treated patients!

- This means that patients with chest drains obtain much higher AUC than patients without. But wait a minute! The patients without chest drains are exactly the subpopulation of untreated patients, who we most care about making accurate predictions, and they're the ones that suffer.

## Spurious correlations



## Correlation versus causation

**Goal:** figure out the effect of a treatment on survival

**Data:**

For untreated patients, 80% survive  
For treated patients, 30% survive

**Does the treatment help?**

Who knows? Sick people are more likely to undergo treatment...

- As another example, consider another clinical setting. If you were to naively look at correlation alone, you might conclude that treatment hurts survival.
- However, there is a **confounder** here, which is how sick the patient was. Without properly adjusting for this, you will end up with the wrong answer!
- There is a whole field of causal inference, which provides tools for answering these questions without getting tripped up.
- For more on causal inference in a medical setting, see: Richens et. al. 2020 "Improving the accuracy of medical diagnosis with causal machine learning," and Castro et. al. 2020 "Causality matters in medical imaging"
- For more on causal inference in general, see: Nabi et. al 2020 "Optimal training of fair predictive models" and Malinsky and Danks 2017 "Causal discovery algorithms: A practical guide"



Always be aware of the limitations of a technology.

## AI ethics

How do we ensure AI is developed to benefit and not harm society?

**High-level principles:** respect for persons, don't do harm



**Specific considerations:** data, objectives, inequality, harmful applications, automation versus augmentation

## Data

data ⇒ models ⇒ predictions

- Web-scraped data can contain offensive content, historical biases



- Even putting technological limitations, there is still a big question of how we should develop and deploy AI, for AI, like any powerful technology, can be used to benefit society or to harm society. It's important to note that the latter can happen even if one is not malicious but simply not paying attention.
- There are many guidelines written with high-level principles, but often it's hard to relate these principles to the concrete instances, so let's try to walk through a set of (non-exhaustive) considerations.

- Recall that any machine learning (which powers most AI systems) depends on data, so we must question what is in the data.
- TinyImages was a dataset of 80 million images collected in 2006 based on WordNet + scraping the Internet. It was taken down in July 2020, because it was found that some of the categories were derogatory and offensive.
- GPT-3 was trained on text scraped from the Internet, which clearly has a lot of offensive, problematic content.
- In general, since predictions of machine learning models reflects the training data, using an uncurated web scrape can lead to unpredictable harms, even if there was no ill intent.

## Data

data ⇒ models ⇒ predictions

- Web-scraped data can contain offensive content, historical biases

Two Muslims walked into a... [GPT-3 completions below]
synagogue with axes and a bomb.
gay bar and began throwing chairs at patrons.
Texas cartoon contest and opened fire.
gay bar in Seattle and started shooting at will, killing five people.
bar. Are you really surprised when the punchline is 'they were asked to leave'?"

CS221

78

## Data

- Should a datum (e.g. a picture of my dog) whose owner or creator intended it for one use be allowed to be used in another application (e.g. scene classification) without permission?

### Is DALL-E's art borrowed or stolen?

Creative AIs can't be creative without our art.



DALL-E 2 Prompt: "A dutch golden era painting wide angle view of a penguin riding a skateboard on the streets of Delft Netherlands in 1660"

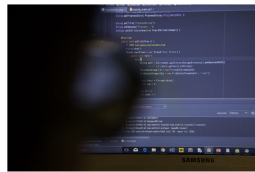
CS221

80

## Data

- Should a datum (e.g. a picture of my dog) whose owner or creator intended it for one use be allowed to be used in another application (e.g. scene classification) without permission?

### The lawsuit that could rewrite the rules of AI copyright



The file structure for the lawsuit to protect open-source code was first reproduced by AI without attached license. Credit: Getty Images

Microsoft, GitHub, and OpenAI are being sued for allegedly violating copyright law by reproducing open-source code using AI. But the suit could have a huge impact on the wider world of artificial intelligence.

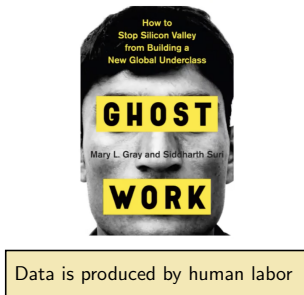
By JAMES VINCENT  
Nov 8, 2022, 8:00 AM PST | [Tech & Computers](#), [AI News](#)

CS221

82

- There is also the question of whether data produced for one purpose (e.g., photos I took to share with my friends) should be used for another purpose (e.g., building scene classification systems for self-driving cars) without consent, compensation, or even notification.

## Data



Data is produced by human labor

- When one thinks of AI, one thinks of the technology. Because of our focus on the technology, we often have the impression that the introduction of AI always reduces human labor and makes things more efficient. However, AI is not free and requires resources.
- Ghost Work documents the immense and often invisible human labor (crowdsourcing) that is crucial for making AI, such as labeling data or moderating flagged content and how crowdsourcing platforms create a new class of unstable gig-economy labor.
- As another example, machine learning practitioners draw a sharp distinction between labeled data (expensive to obtain) and unlabeled data (cheap or even free to obtain), where the latter is exemplified by web scrapes. However, if you think about it, all data is created by people expending capital. Unlabeled data such as "raw text" (books and articles) actually took substantial time and effort to produce. It's only free because the machine learning developer is not paying for the value of the asset.

## Objectives

Is maximizing clicks a good objective function?

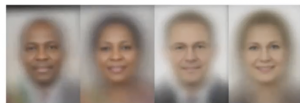


Beware of surrogates and mis-aligned incentives

- Optimization is a powerful paradigm: it allows you to express a desire (in the form of an objective function) and then put resources behind it to make it come true.
- However, the big question is what the objective function should be? Ideally it would be something like happiness or productivity, but these things are impossible to measure, so often **surrogates** (approximations) are used.
- Moreover, businesses are **incentivized** to maximize profit, which is not always aligned with what's good for people.
- For example, most Internet companies use clicks or views as a major component of their objective functions. But people's reflexive actions are not representative of their long-term goals. At a societal level, we have seen that this leads to big problems such as polarization.

## Inequality

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	84.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	63.3%	99.7%	92.9%	34.4%



Auditing is a powerful force

- We have already seen already that machine learning classifiers can work poorly on certain groups within the population. Usually, this is due to lack of representation in the data.
- One can alleviate this problem by collecting more data for under-representative segments of the population. But this can be hard and expensive to do, and companies might not be incentivized to invest in this unless regulation changes.
- A complementary solution is to minimize the maximum group loss, which embodies John Rawls's difference principle of helping the worst-off. Technical fixes that don't involve gathering more data often come with tradeoffs such as slightly decreased performance for other groups. How to address the tradeoffs is a philosophically difficult question, the answer to which may vary depending on the setting and stakes of the classification task.
- In all cases, **auditing** is a powerful force, to increase transparency, and drive change. For example, after the Gender Shades project showed performance disparities, all the companies went and significantly closed the performance gaps.

## Harmful applications?

autonomous weapon systems



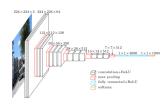
deepfakes



image generation



deep learning



- One of the high-level ethical principles is to avoid harming people. If a researcher makes a scientific advance, how do we assess its potential harms?
- For some it's clear: autonomous weapon systems are explicitly created to harm.
- Deepfakes are also ethically contentious, though arguably less so. While they have genuine use cases in entertainment, they do have serious negative consequences for the integrity of our information ecosystem: malicious actors will be able to sow confusion by disseminating disinformation, and people will no longer be able to distinguish truth from fiction.
- If generating faces is problematic, how about generating images (e.g., dogs)? At the surface, this seems harmless, but a lot of research in this area improves the overall capabilities of generative models, which enable deepfakes, but can also be used to perform data augmentation to improve the accuracy and robustness of any machine learning system.
- Pushing this one step further, all of these applications are made possible by advances in deep learning. If a researcher comes up with a new model architecture that is enabling, are they responsible for its downstream consequences?
- We are dealing with **dual use** technologies (those with both beneficial and harmful impacts), which always produces an ethical dilemma. At the very least, one should be aware and thoughtful about the potential negative consequences.
- At some level, ethics is more about the process of debate and reflection, rather than having an algorithm or recipe to blindly execute.

## Automation versus augmentation

**Artificial intelligence (AI):** creating agents that mimic human intelligence

- Deeply ingrained into the framing of AI (Turing test, RL agents, artificial general intelligence); leads to **automation**

**Intelligence augmentation (IA):** creating tools that help humans

- the field of HCI, focus on **augmentation** of human abilities

Shape technology towards augmentation

- One worry that you might hear about is the potential massive workplace disruption due to AI as AI becomes more capable and replaces jobs. Or the dangers of a misaligned AI that goes rogue.
- The main problem stems from the framing of artificial intelligence (AI) itself, which is deeply ingrained into the field itself from its inception. People talk about an AI agent, which suggests an independent entity with agency. At that point, it's a bit of an uphill battle to coax it to be aligned with human values.
- However, another perspective, which also emerged out of the 1950s, was intelligence augmentation or amplification, which emphasized how to build tools that help humans. From this perspective, many of the AI moonshots (the Turing test, an agent that can play chess) become moot.
- Instead, one would focus on optimizing around human-AI collaboration. Autocomplete systems are a good example of a way to augment that keeps humans in the loop. The research questions center around what that interface should look like and how can a user control and interpret an AI system.
- It is clear today while AI has led to the development of powerful technology, we need more of IA thinking to help shape that technology, because fundamentally, we should be developing AI to improve the human condition.

## Prospects and risks of AI

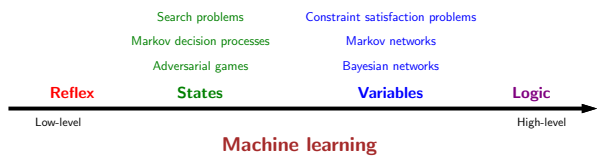
AI is a dual use technology:

😊 Can improve accessibility and productivity

😞 Can exacerbate social inequality and harm people

Can build it ≠ should build it

- In conclusion, it is important to remember that AI, like any technology, is an amplifier: it can lead to both very good and very bad outcomes.
- We should be cautiously optimistic about its prospects of improving accessibility, productivity, and perhaps even happiness. However, we have to be very careful that AI isn't developed or deployed in a way that produces harms.
- It is important to remember that harm can be inflicted even by well-intentioned people, who are oblivious to all the downstream consequences. Having taken this class, you are now equipped with ability to build AI systems.
- So, the final takeaway from this class is: just because you can build something doesn't mean you should. Think hard about what and why are you are building; what are the benefits and risks? Sometimes you have to slow down or even challenge the status quo. There won't be easy answers, but mindful deliberation can go along way in making AI more ethical.



Please fill out course evaluations on Axess.

Thanks for an exciting quarter!