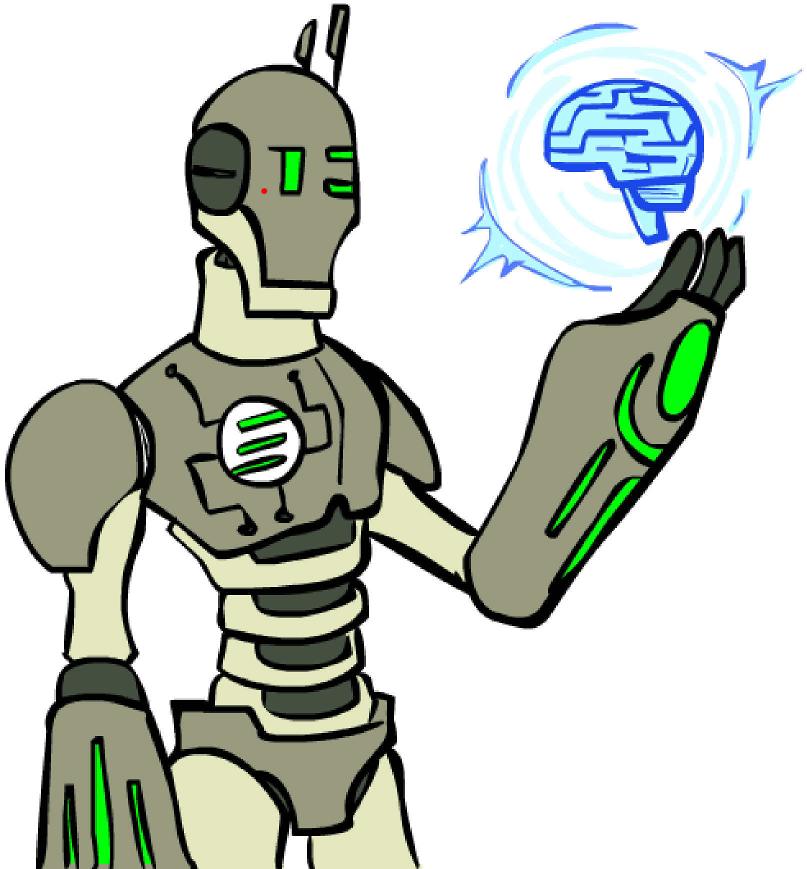
As you walk in

Welcome!

1) Sit at a table next to another student
2) Make name plate
   - Fold paper in half
   - Write preferred name
   - Below write you favorite fictional AI/robot



Images: ai.berkeley.edu

10-606 Mathematical Foundations for Machine Learning

Instructor: Pat Virtue

Today

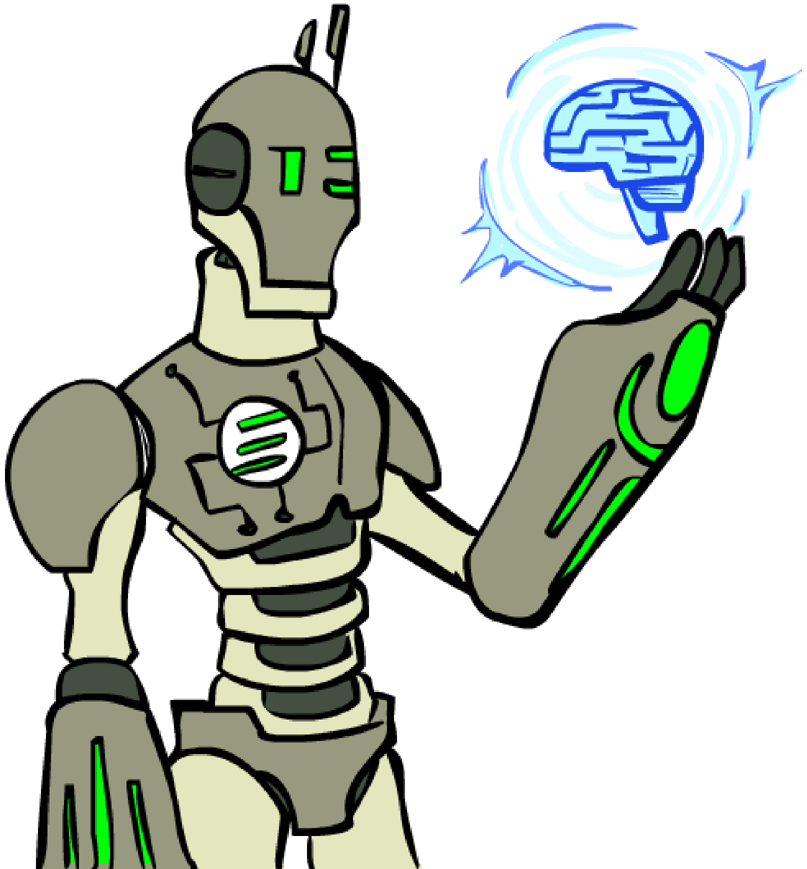
Course Info

Warm-up exercise

ML and Math Intro

Systems of equations

More Course Info



Images: ai.berkeley.edu

Course Team

Instructor

Pat
Virtue
pvirtue

Teaching Assistants

Kellen
Gibson
kagibson

Ian
Char
ichar

Course Team

Students!!

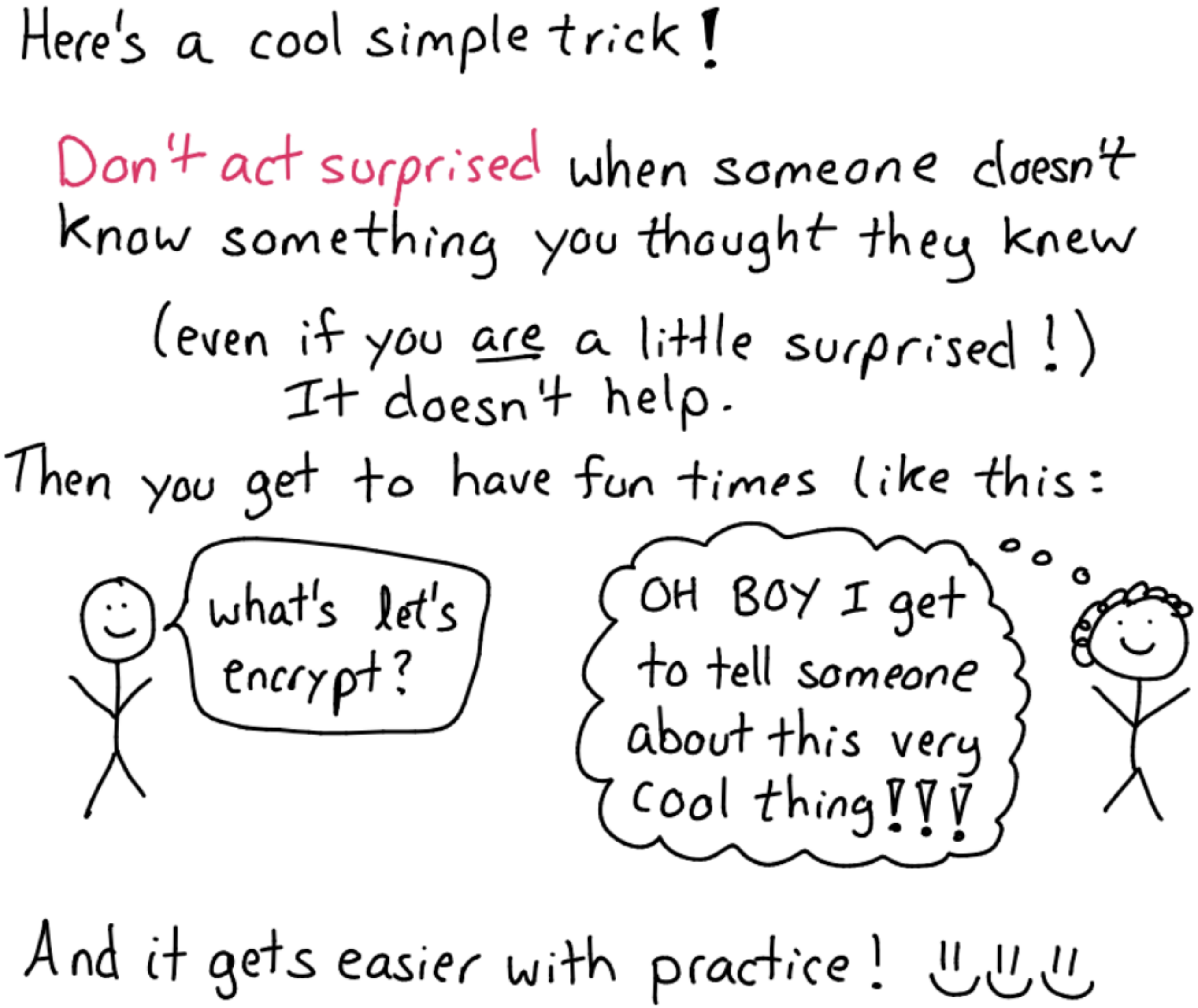
Team Tips

Try not to act surprised

https://jvns.ca/blog/2017/04/27/no-feigning-surprise/

Team Tips

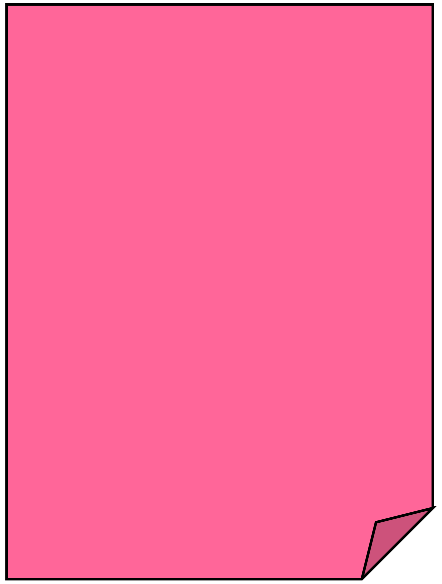
Try not to act surprised



https://jvns.ca/blog/2017/04/27/no-feigning-surprise/

Team Tips

Jargon card



Notation alert!

Warm-up Exercise

Suppose we want to find the parameter $\hat{\theta}$ that minimizes the function $J(\theta)$.

For three different sets of assumptions about $J(\theta)$, write down an algorithm (steps, pseudocode, etc.) that will try to find a good estimate $\hat{\theta}$.

1. Assumptions: $J: \mathbb{R} \to \mathbb{R}$. We are able to call $J(\theta)$ as much as we like. That's all we know.
   Algorithm: ?

   def J(theta):

Notation alert!

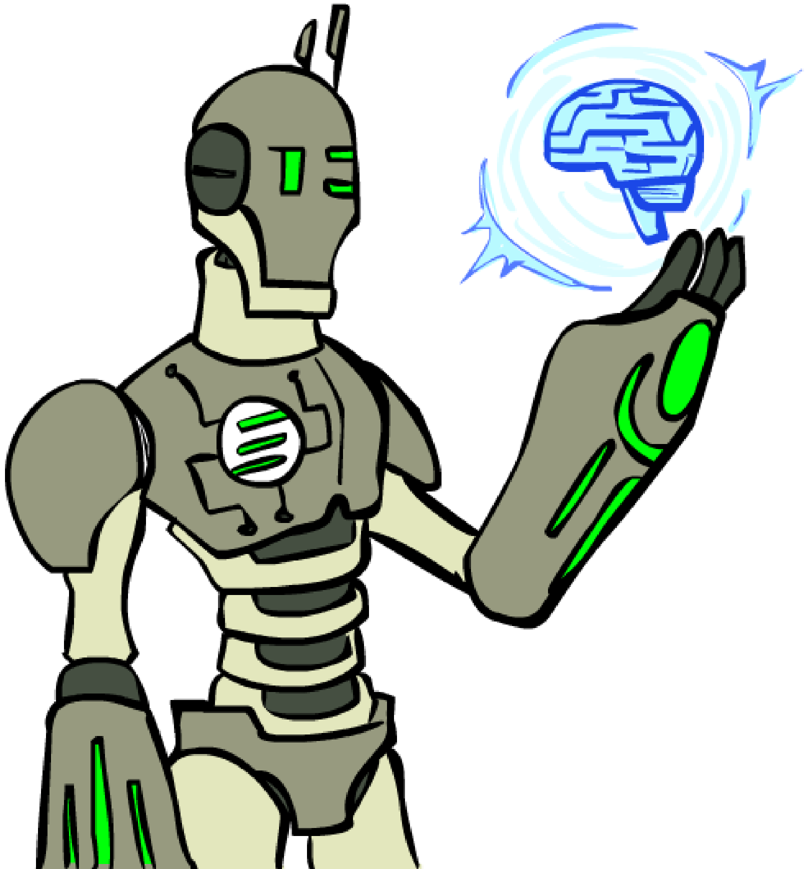
2. Assumptions: you choose
   Algorithm: ?

   convex, first deriv = 0
   $J'(\theta)$

3. Assumptions: you choose
   Algorithm: ?

Today

- Course Info
- Warm-up exercise
- ML and 606/607
- Systems of equations
- More Course Info



Images: ai.berkeley.edu

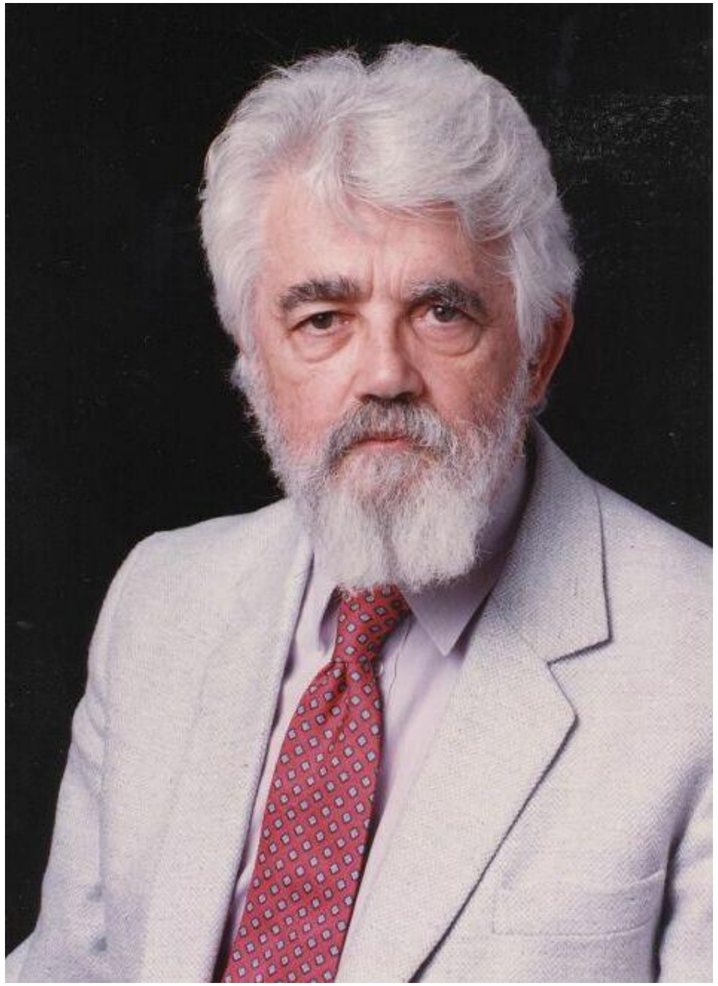
AI Definition by John McCarthy

What is artificial intelligence

- It is the science and engineering of making intelligent machines, especially intelligent computer programs

What is intelligence

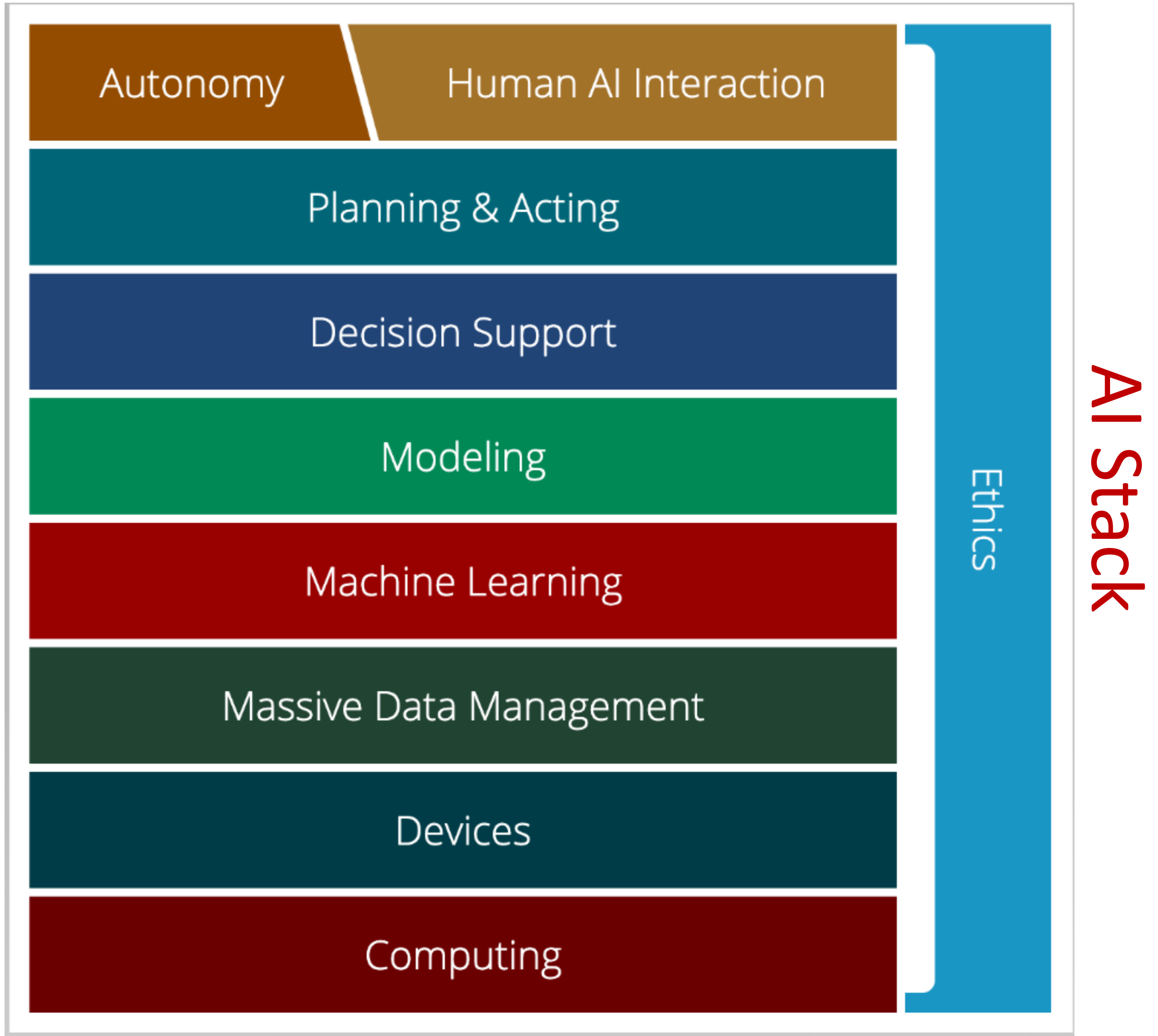
- Intelligence is the computational part of the ability to achieve goals in the world



http://www-formal.stanford.edu/jmc/whatisai/whatisai.html

AI Stack for CMU AI

"AI must understand the human needs and it must make smart design decisions based on that understanding"

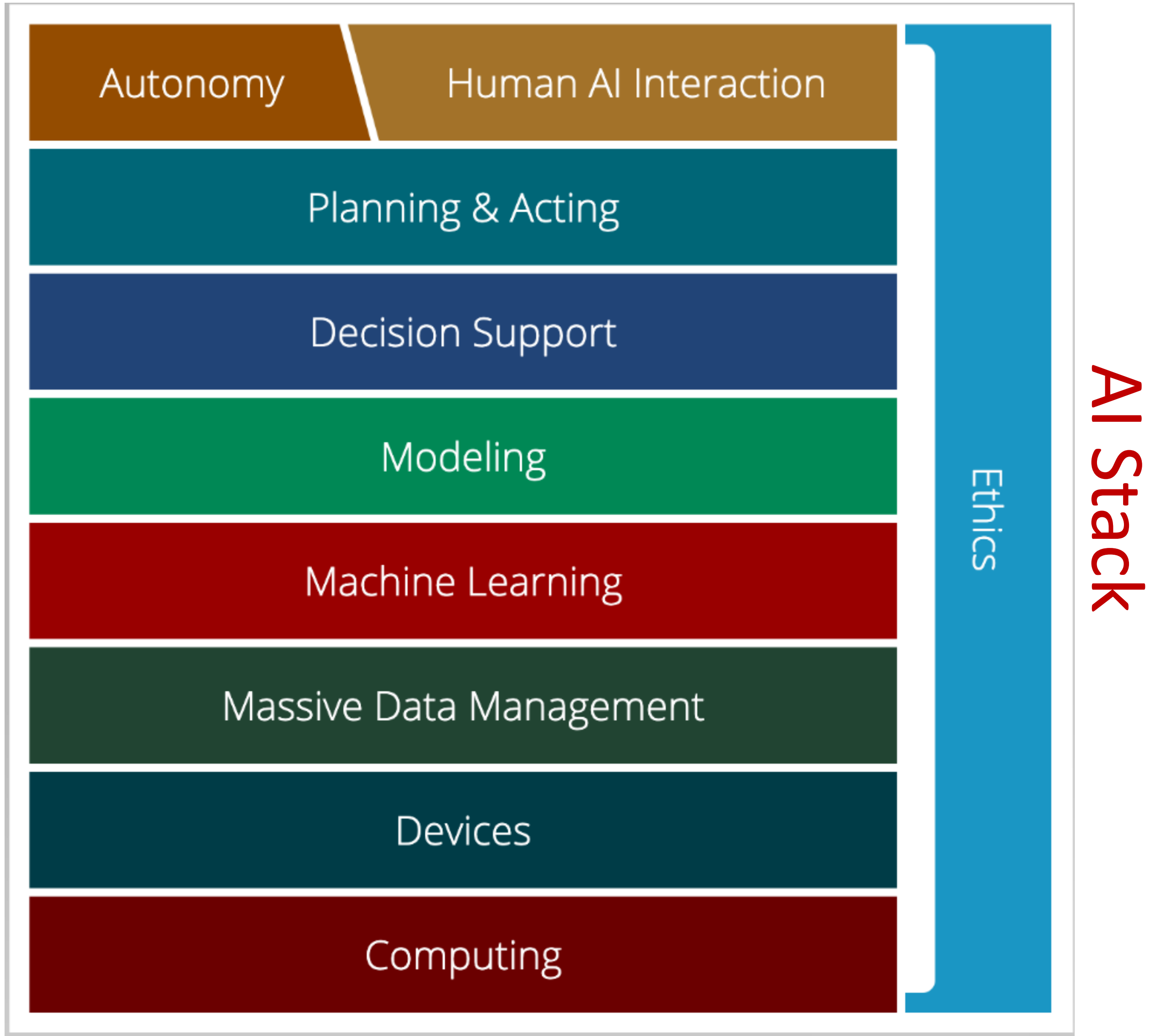


https://ai.cs.cmu.edu/about

AI Stack for CMU AI

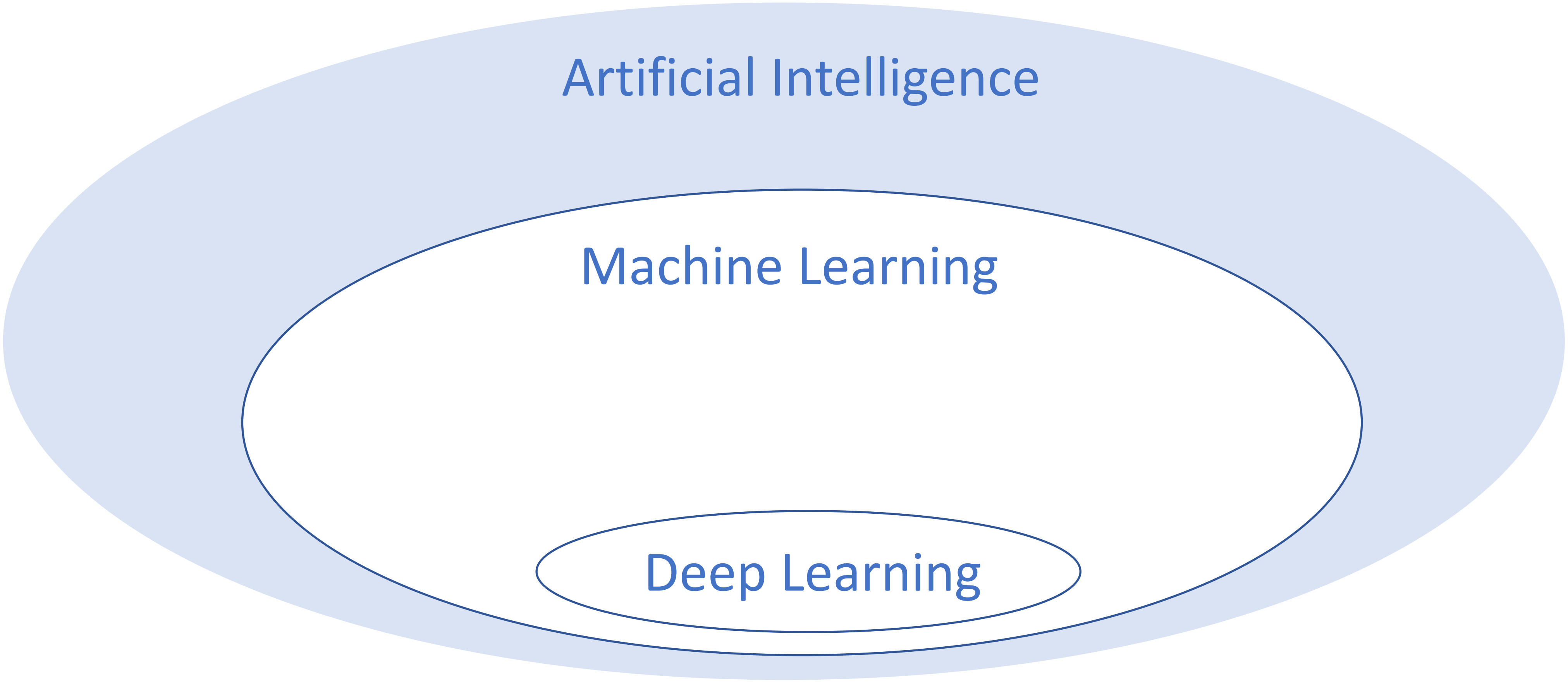
“Machine learning focuses on creating programs that learn from experience.”

“It advances computing through exposure to new scenarios, testing and adaptation, while using pattern- and trend-detection to help the computer make better decisions in similar, subsequent situations.”

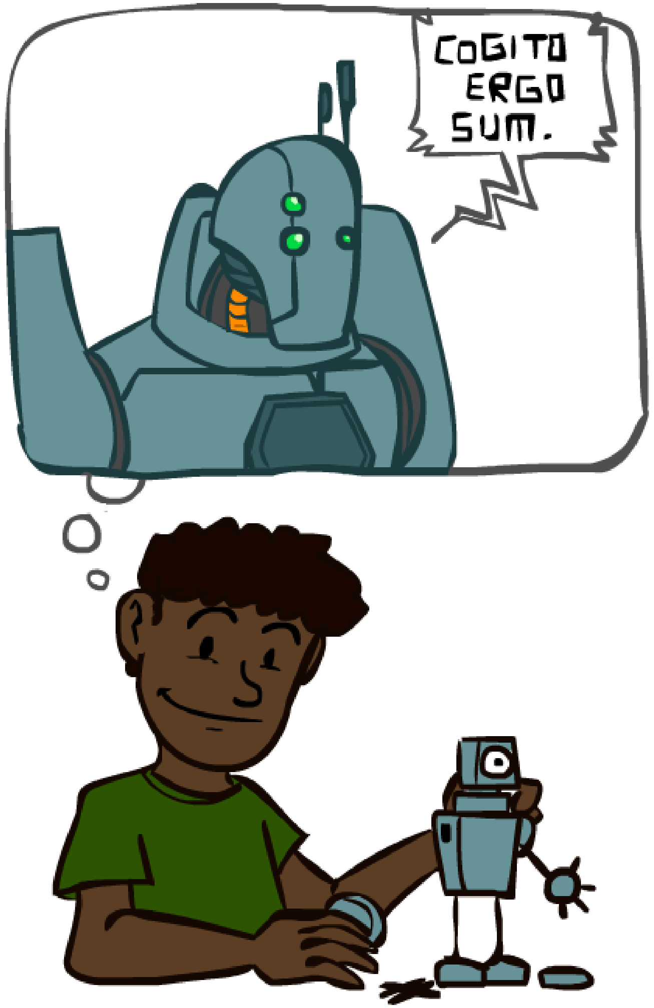


https://ai.cs.cmu.edu/about

Artificial Intelligence vs Machine Learning?

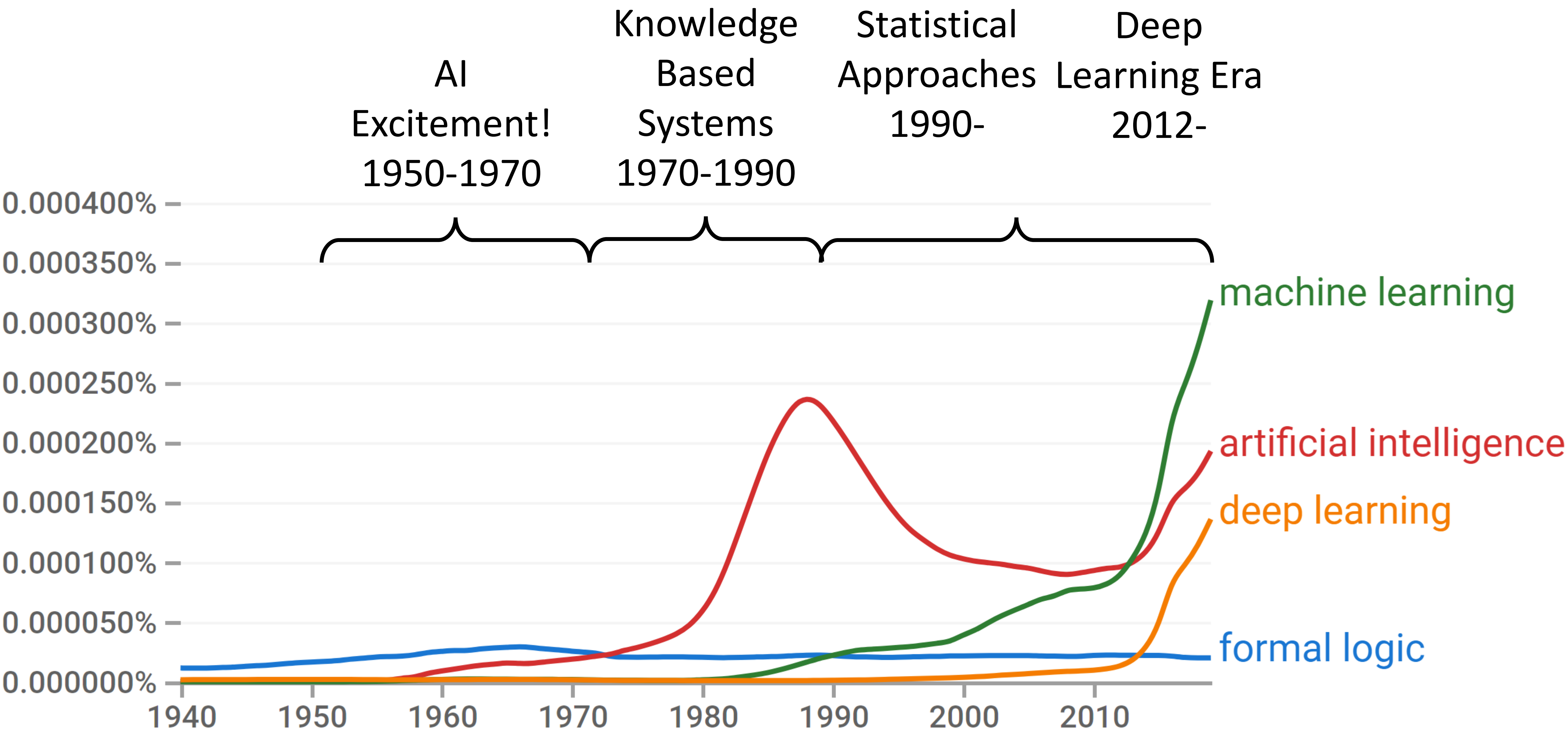


A Brief History of AI



Images: ai.berkeley.edu

A Brief History of AI



https://books.google.com/ngrams

A Brief History of AI

1940-1950: Early days

- 1943: McCulloch & Pitts: Boolean circuit model of brain
- 1950: Turing's “Computing Machinery and Intelligence”

1950—70: Excitement: Look, Ma, no hands!

- 1950s: Early AI programs, including Samuel's checkers program, Newell & Simon's Logic Theorist, Gelernter's Geometry Engine
- 1956: Dartmouth meeting: “Artificial Intelligence” adopted

1970—90: Knowledge-based approaches

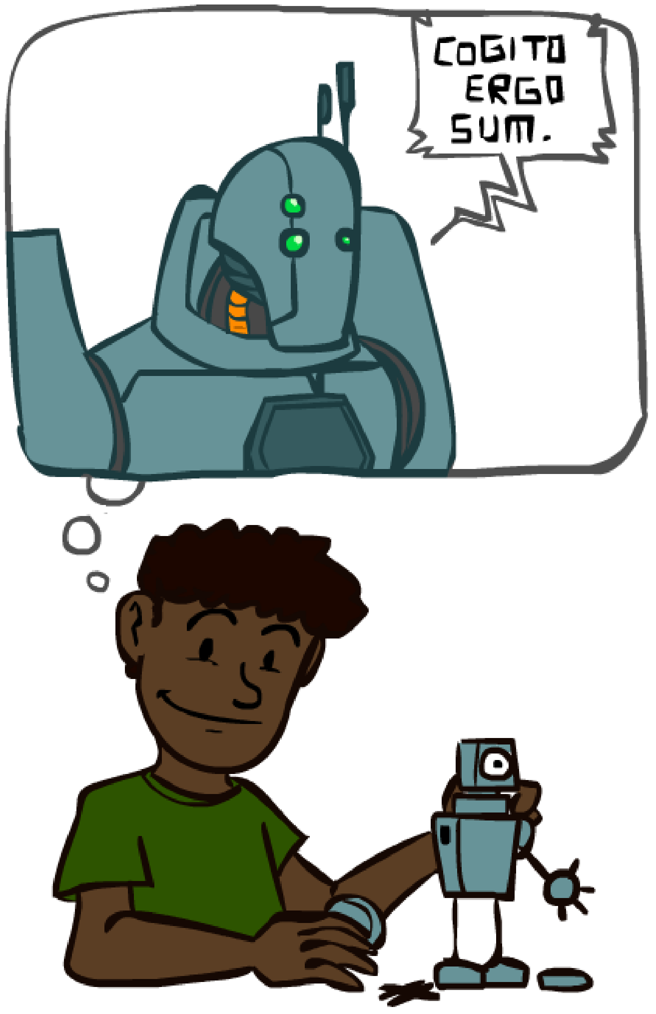
- 1969—79: Early development of knowledge-based systems
- 1980—88: Expert systems industry booms
- 1988—93: Expert systems industry busts: “AI Winter”

1990—: Statistical approaches

- Resurgence of probability, focus on uncertainty
- General increase in technical depth
- Agents and learning systems… “AI Spring”?

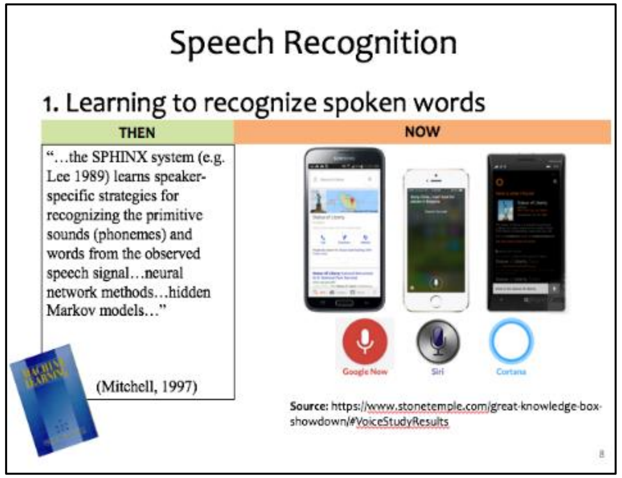
2012—: Deep learning

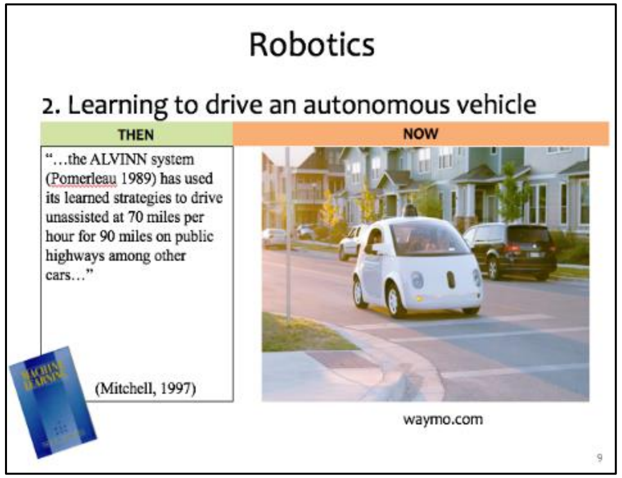
- 2012: ImageNet & AlexNet

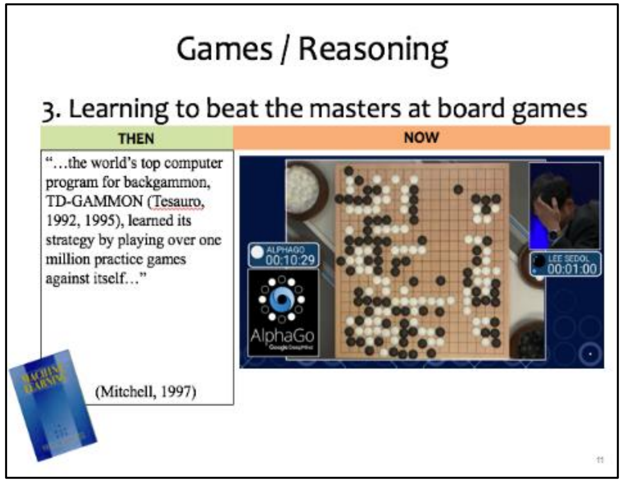


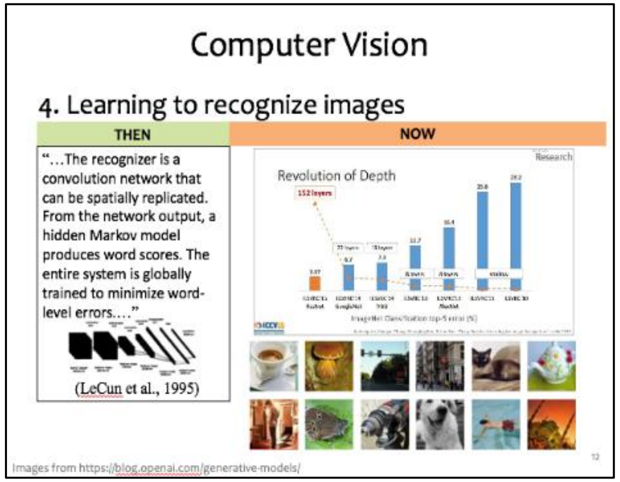
Images: ai.berkeley.edu

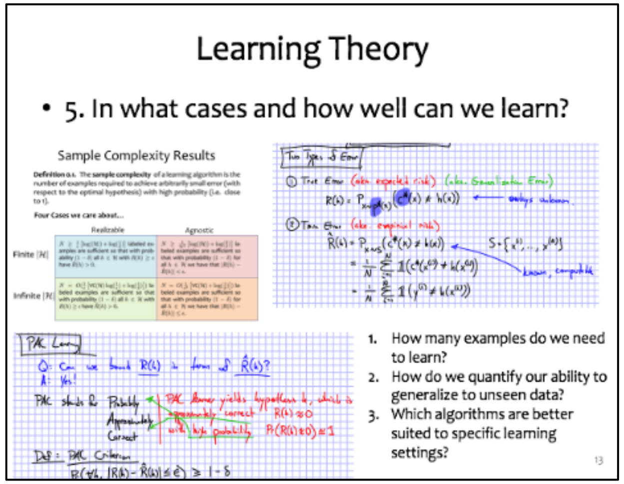
ML Applications?











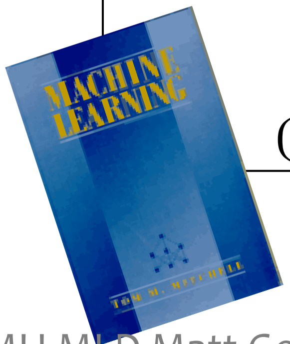
Speech Recognition

1. Learning to recognize spoken words

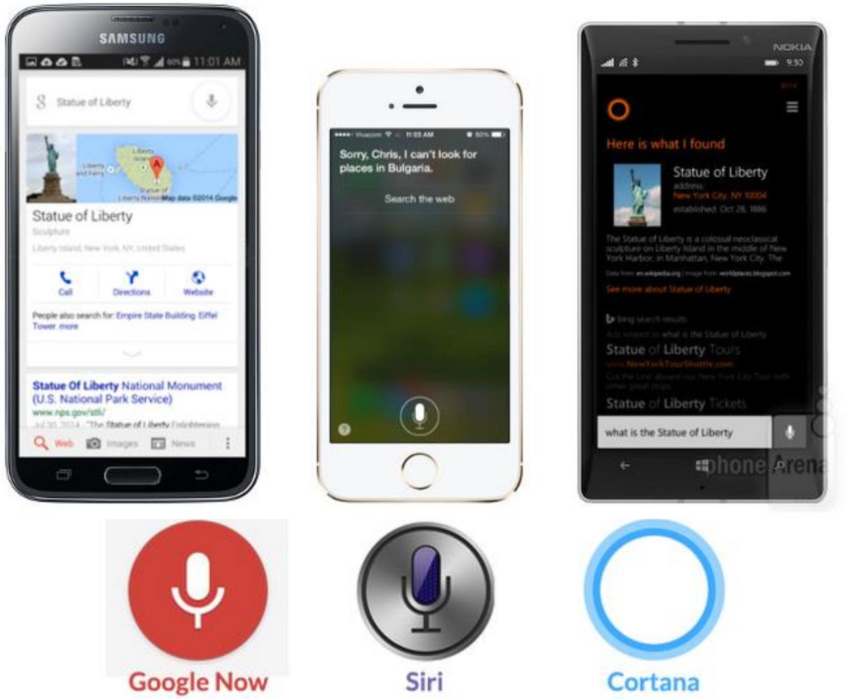
THEN

"…the SPHINX system (e.g. Lee 1989) learns speaker-specific strategies for recognizing the primitive sounds (phonemes) and words from the observed speech signal…neural network methods…hidden Markov models…"

(Mitchell, 1997)



NOW

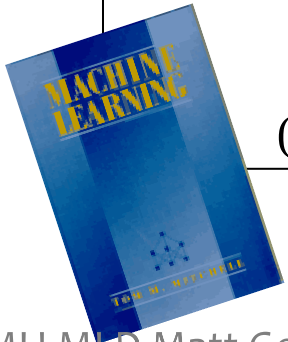


Source: https://www.stonetemple.com/great-knowledge-box-showdown/#VoiceStudyResults

Robotics

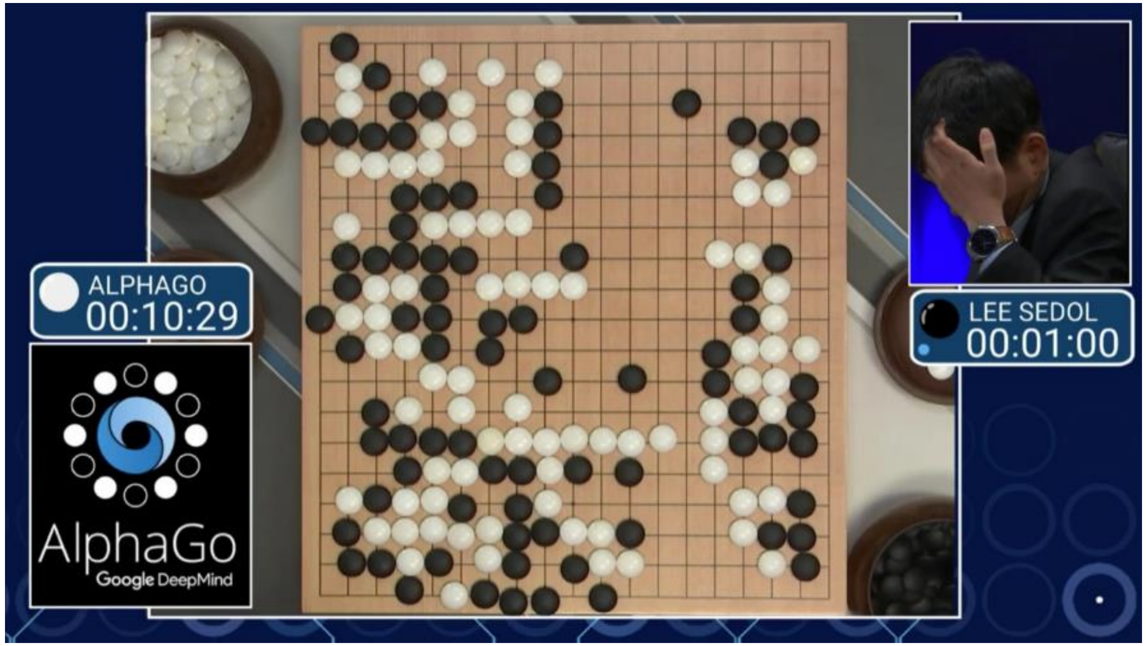
2. Learning to drive an autonomous vehicle

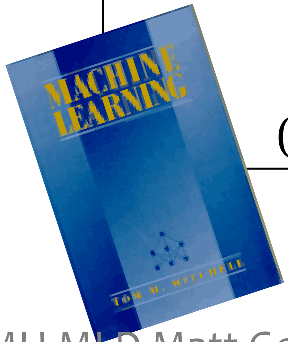
THEN	NOW
"…the ALVINN system (Pomerleau 1989) has used its learned strategies to drive unassisted at 70 miles per hour for 90 miles on public highways among other cars…" (Mitchell, 1997)	 https://www.geek.com/wp-content/uploads/2016/03/uber.jpg



Games / Reasoning

3. Learning to beat the masters at board games

THEN	NOW
"…the world's top computer program for backgammon, TD-GAMMON (Tesauro, 1992, 1995), learned its strategy by playing over one million practice games against itself…" (Mitchell, 1997)	



Computer Vision

4. Learning to recognize images

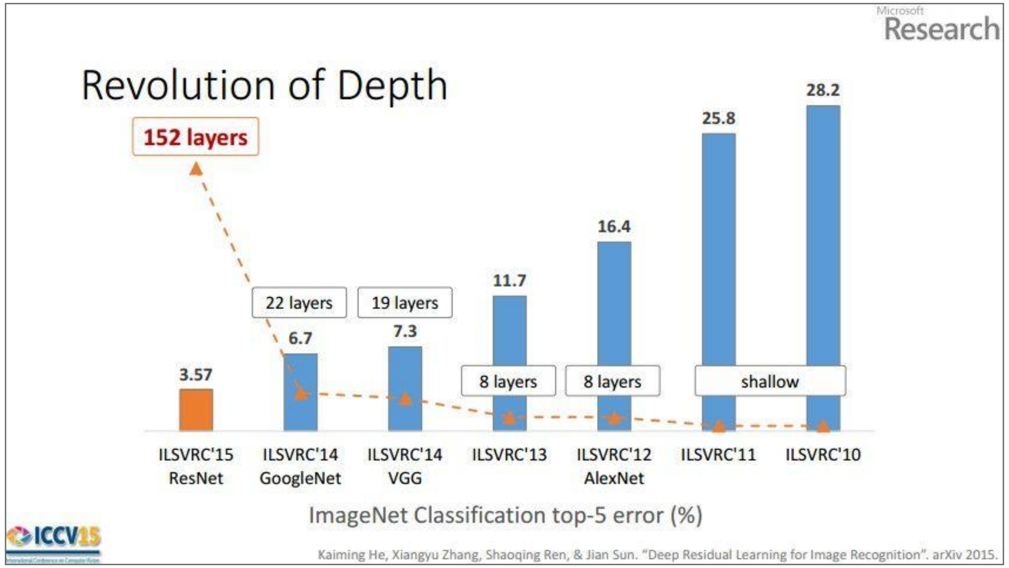
THEN	NOW

"…The recognizer is a convolution network that can be spatially replicated. From the network output, a hidden Markov model produces word scores. The entire system is globally trained to minimize word-level errors.…"

Figure 2: Convolutional neural network character recognizer. This architecture is robust to local translations and distortions, with subsampling, shared weights, and local receptive fields.

number of subsampling layers and the sizes of the kernels are chosen, the sizes of all the layers, including the input, are determined unambiguously. The only architectural parameters that remain to be selected are the number of feature maps in each layer, and the information as to what feature map is connected to what other feature map. In our case, the subsampling rates were chosen as small as possible (2 × 2), and the kernels as small as possible in the first layer (3 × 3) to limit the total number of connections. Kernel sizes in the upper layers are chosen to be as small as

(LeCun et al., 1995)



Slide credit: CMU MLD Matt Gormley

Images from https://blog.openai.com/generative-models/

Learning Theory

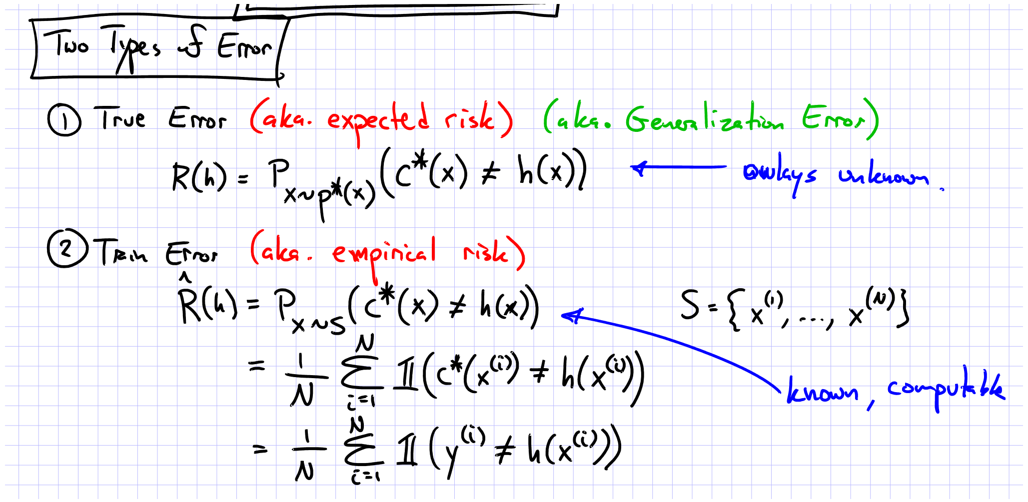
- 5. In what cases and how well can we learn?

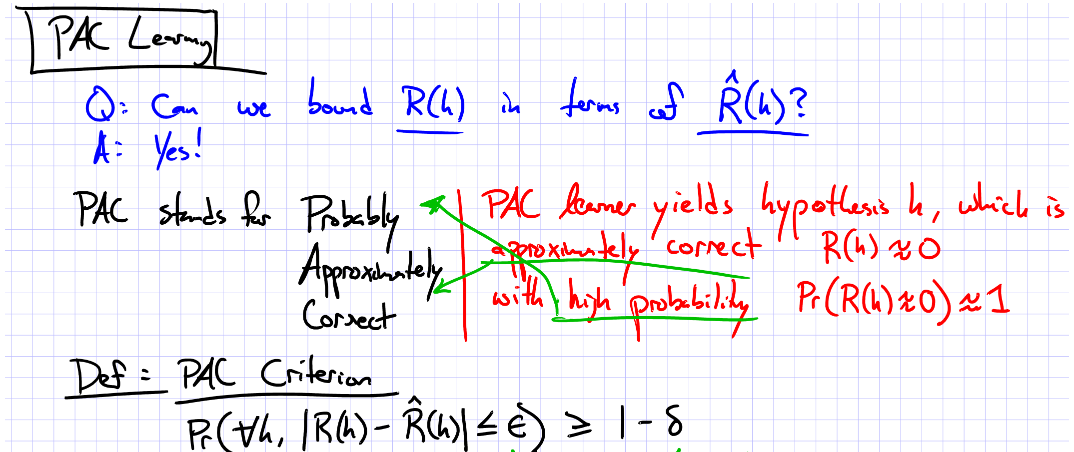
Sample%Complexity%Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four$Cases$we$care$about…

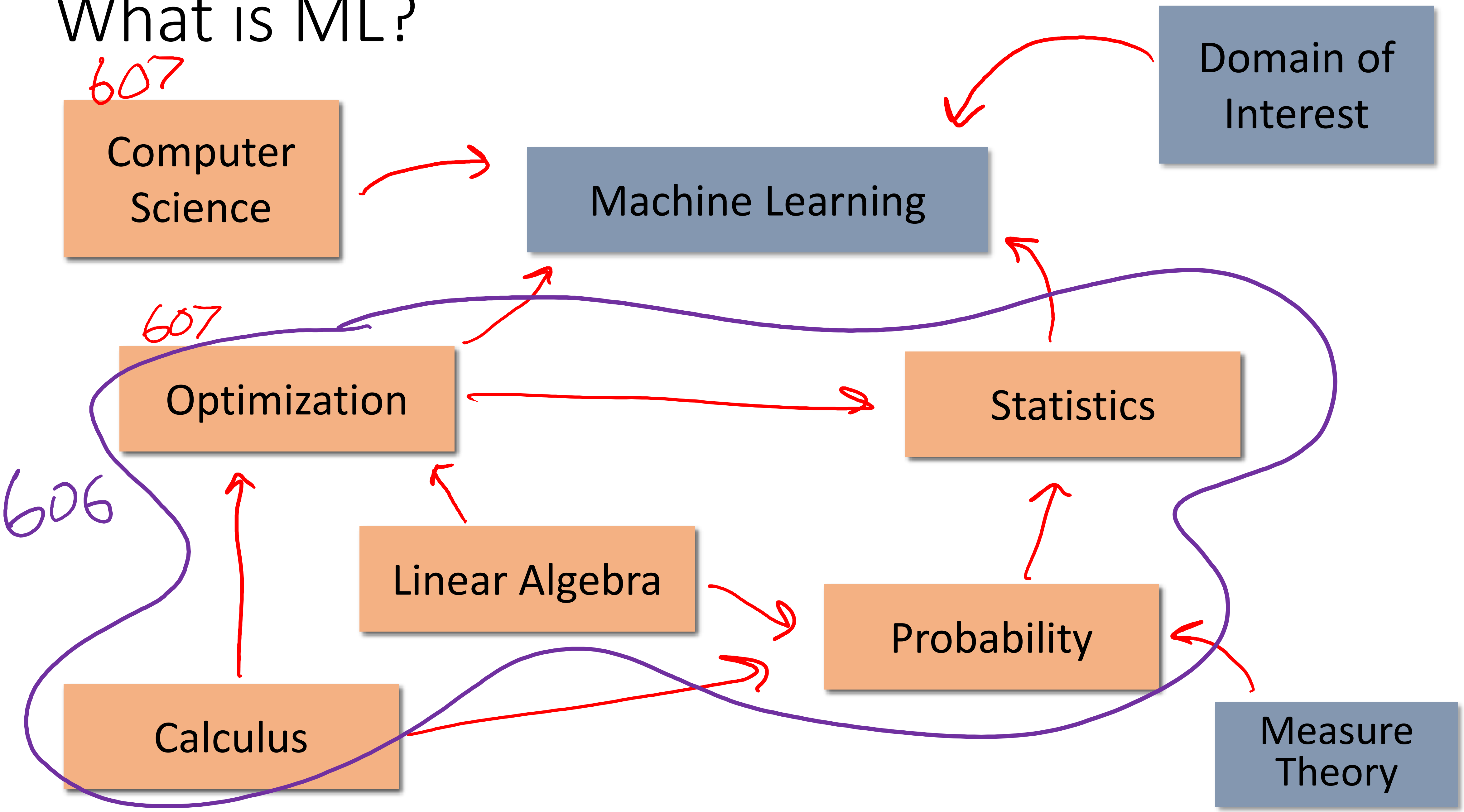
	Realizable	Agnostic
Finite $\lvert\mathcal{H}\rvert$	$N \geq \frac{1}{\epsilon}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $R(h) \geq \epsilon$ have $\hat{R}(h) > 0$.	$N \geq \frac{1}{2\epsilon^2}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert < \epsilon$.
Infinite $\lvert\mathcal{H}\rvert$	$N = O(\frac{1}{\epsilon}\left[\text{VC}(\mathcal{H})\log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $R(h) \geq \epsilon$ have $\hat{R}(h) > 0$.	$N = O(\frac{1}{\epsilon^2}\left[\text{VC}(\mathcal{H}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$.





1. How many examples do we need to learn?
2. How do we quantify our ability to generalize to unseen data?
3. Which algorithms are better suited to specific learning settings?

What is ML?


607
Computer Science
Machine Learning
Domain of Interest
607
Optimization
Statistics
606
Linear Algebra
Probability
Calculus
Measure Theory

Binary Logistic Regression

Gradient

$\vec{x} \in \mathbb{R}^M \quad \vec{\theta} \in \mathbb{R}^M$

$$J(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \underbrace{\log p(y^{(i)} \mid \vec{x}^{(i)}, \vec{\theta})}_{J^{(i)}(\theta)}$$

$\nabla_\theta J(\theta)$

$\mathbb{R}^M \rightarrow \mathbb{R}^M$

$$\nabla_\theta J^{(i)}(\theta) = \begin{bmatrix} \frac{\partial J^{(i)}}{\partial \theta_1} \\ \vdots \\ \frac{\partial J^{(i)}}{\partial \theta_m} \\ \vdots \end{bmatrix}$$

$$\frac{\partial J^{(i)}}{\partial \theta_m} = -\left(y^{(i)} - \sigma(\vec{\theta}^T \vec{x})\right) x_m^{(i)}$$

SGD

$$\nabla_\theta J^{(i)} = -\left(y^{(i)} - \sigma(\vec{\theta}^T \vec{x})\right) \vec{x}$$

Example slide from 10-601

Naïve Bayes MLE

$$\mathcal{D} = \{y^{(n)}, \boldsymbol{x}^{(n)}\}_{n=1}^{N}$$
$$y^{(n)} \in \{0,1\}$$
$$\boldsymbol{x}^{(n)} \in \{0,1\}^M$$
$$\phi \in [0,1]$$
$$\boldsymbol{\Theta} \in [0,1]^{Mx2}$$

Example slide from 10-601

$$L(\phi, \boldsymbol{\Theta}) = p(\mathcal{D} \mid \phi, \boldsymbol{\Theta})$$

$$= \prod_{n=1}^{N} p\left(\mathcal{D}^{(n)} \mid \phi, \boldsymbol{\Theta}\right)$$ i.i.d assumption

$$= \prod_{n=1}^{N} p\left(y^{(n)}, \boldsymbol{x}^{(n)} \mid \phi, \boldsymbol{\Theta}\right)$$

$$= \prod_{n=1}^{N} p\left(y^{(n)} \mid \phi\right) p\left(\boldsymbol{x}^{(n)} \mid y^{(n)}, \boldsymbol{\Theta}\right)$$ Generative model

$$= \prod_{n=1}^{N} p\left(y^{(n)} \mid \phi\right) p\left(x_1^{(n)}, x_2^{(n)}, \ldots, x_M^{(n)} \mid y^{(n)}, \boldsymbol{\Theta}\right)$$

$$= \prod_{n=1}^{N} p\left(y^{(n)} \mid \phi\right) \prod_{m=1}^{M} p\left(x_m^{(n)} \mid y^{(n)}, \theta_{m,y}\right)$$ Naïve Bayes

$$= \prod_{n=1}^{N} \phi^{y^{(n)}} (1-\phi)^{1-y^{(n)}} \prod_{m=1}^{M} \theta_{m,1}^{\mathbb{I}\left(y^{(n)}=1 \wedge x_m^{(n)}=1\right)} \left(1-\theta_{m,1}\right)^{\mathbb{I}\left(y^{(n)}=1 \wedge x_m^{(n)}=0\right)} \theta_{m,0}^{\mathbb{I}\left(y^{(n)}=0 \wedge x_m^{(n)}=1\right)} \left(1-\theta_{m,0}\right)^{\mathbb{I}\left(y^{(n)}=0 \wedge x_m^{(n)}=0\right)}$$

$$= \phi^{N_{y=1}} (1-\phi)^{N_{y=0}} \prod_{m=1}^{M} \theta_{m,1}^{N_{y=1,x_m=1}} \left(1-\theta_{m,1}\right)^{N_{y=1,x_m=0}} \theta_{m,0}^{N_{y=0,x_m=1}} \left(1-\theta_{m,0}\right)^{N_{y=0,x_m=0}}$$

Network Optimization: Layer Implementation

$J(\mathbf{w}) = z_3$

$z_3 = f_3(w_3, z_2)$

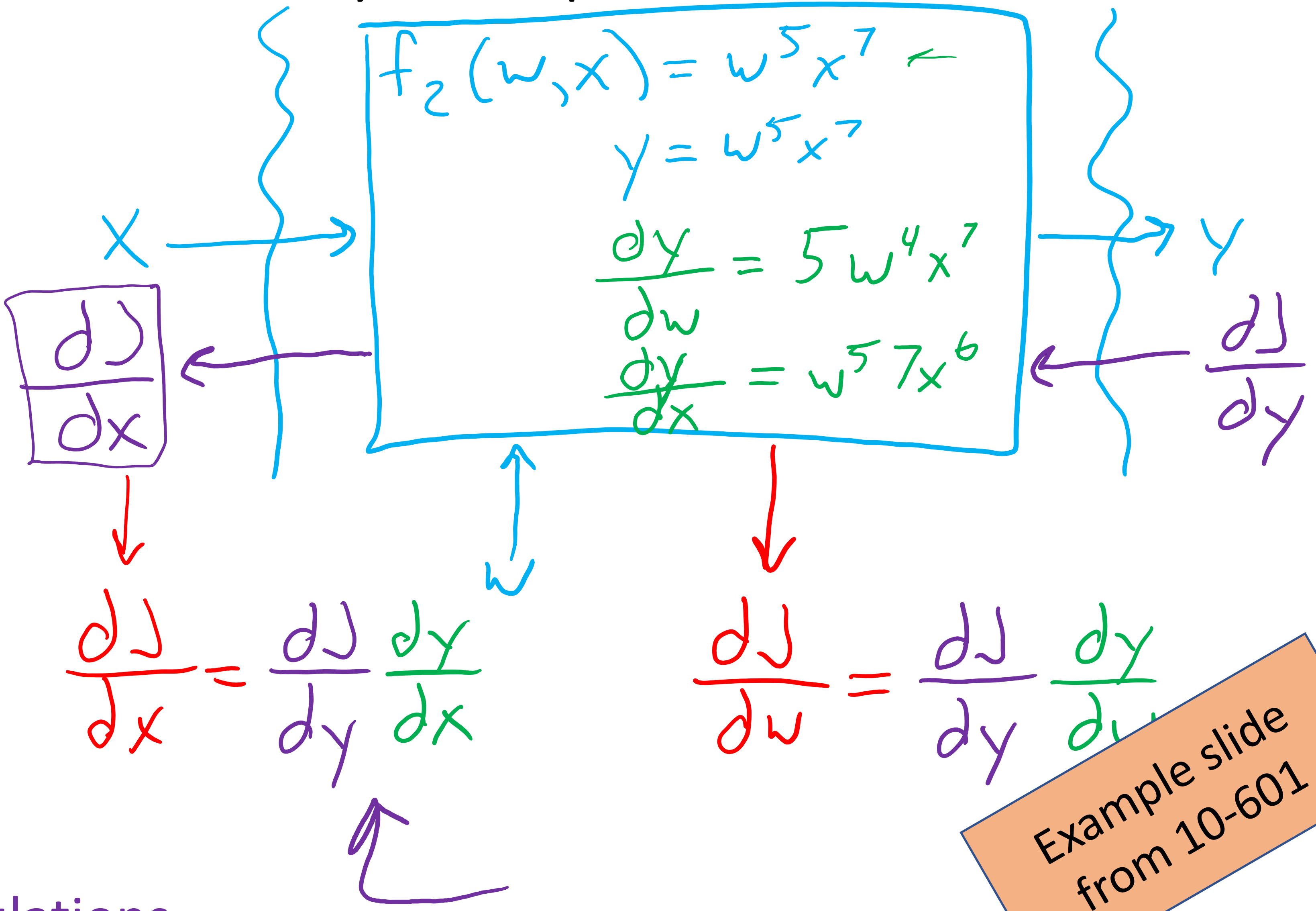
$z_2 = f_2(w_2, z_1)$

$z_1 = f_1(w_1, x)$

$$\frac{\partial J}{\partial w_3} = \frac{\partial J}{\partial z_3}\frac{\partial z_3}{\partial w_3}$$

$$\frac{\partial J}{\partial w_2} = \frac{\partial J}{\partial z_3}\frac{\partial z_3}{\partial z_2}\frac{\partial z_2}{\partial w_2}$$

$$\frac{\partial J}{\partial w_1} = \frac{\partial J}{\partial z_3}\frac{\partial z_3}{\partial z_2}\frac{\partial z_2}{\partial z_1}\frac{\partial z_1}{\partial w_1}$$



Lots of repeated calculations

Example slide from 10-601

PCA: the First Principal Component

To find the first principal component, we wish to solve the following constrained optimization problem (variance minimization).

$$\mathbf{v}_1 = \underset{\mathbf{v}:||\mathbf{v}||^2=1}{\text{argmax}} \ \mathbf{v}^T \mathbf{\Sigma} \mathbf{v} \tag{1}$$

So we turn to the method of Lagrange multipliers. The Lagrangian is:

$$\mathcal{L}(\mathbf{v}, \lambda) = \mathbf{v}^T \mathbf{\Sigma} \mathbf{v} - \lambda(\mathbf{v}^T \mathbf{v} - 1) \tag{2}$$

Taking the derivative of the Lagrangian and setting to zero gives:

$$\frac{d}{d\mathbf{v}} \left(\mathbf{v}^T \mathbf{\Sigma} \mathbf{v} - \lambda(\mathbf{v}^T \mathbf{v} - 1) \right) = 0 \tag{3}$$

$$\mathbf{\Sigma} \mathbf{v} - \lambda \mathbf{v} = 0 \tag{4}$$

$$\mathbf{\Sigma} \mathbf{v} = \lambda \mathbf{v} \tag{5}$$

Recall: For a square matrix $\mathbf{A}$, the vector $\mathbf{v}$ is an **eigenvector** iff there exists **eigenvalue** λ such that:

$$\mathbf{A} \mathbf{v} = \lambda \mathbf{v} \tag{6}$$

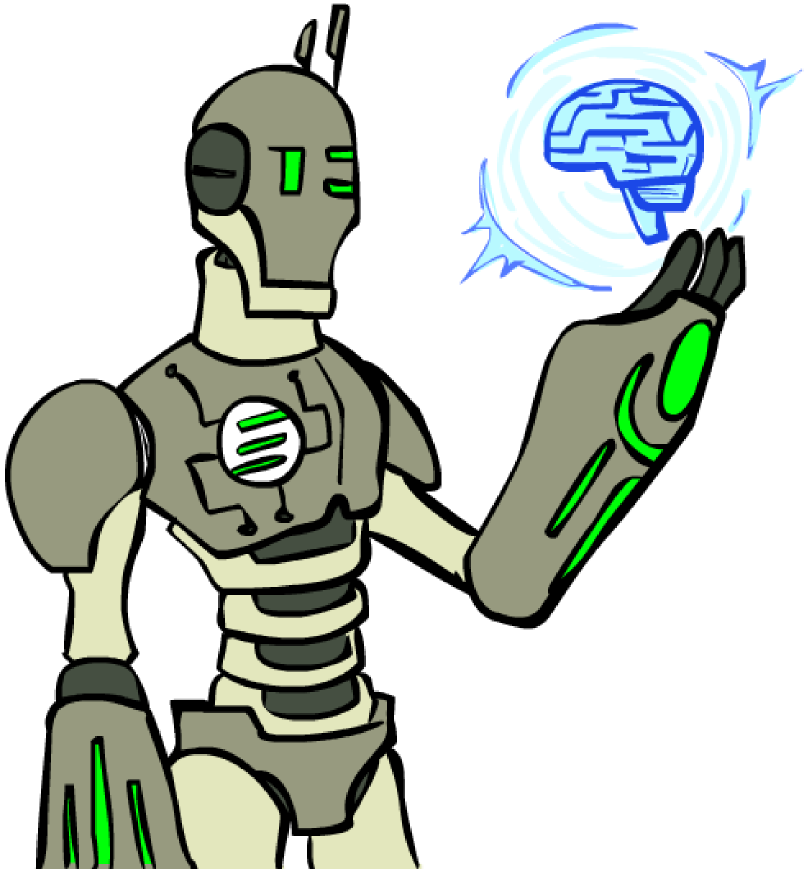
Example slide from 10-601

10-606 and 10-607

- Mini Courses
  - 10-606
  - 10-607
- Intro ML Courses
  - 10-315
  - 10-301/601
  - 10-701
  - 10-715
- Prerequisites

Today

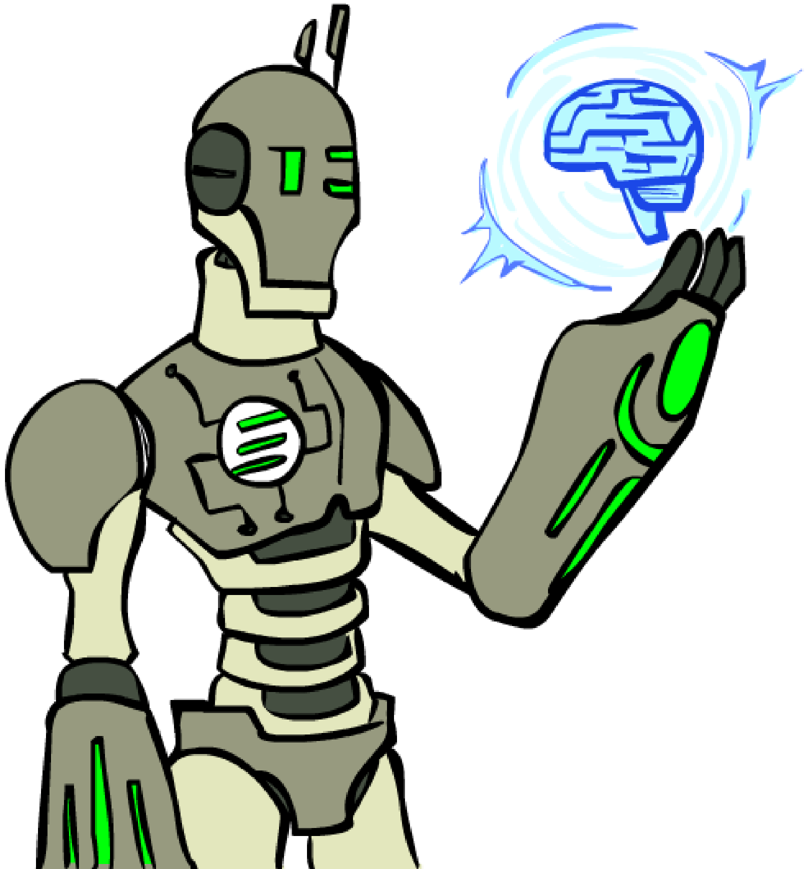
- Course Info
- Warm-up exercise
- ML and 606/607
- Systems of equations
- More Course Info



Images: ai.berkeley.edu

Today

- Course Info
- Warm-up exercise
- ML and 606/607
- Systems of equations
- More Course Info



Images: ai.berkeley.edu

Course Information

Website: https://www.cs.cmu.edu/~10606

Canvas: canvas.cmu.edu

Gradescope: gradescope.com

Communication:

piazza.com

E-mail (if piazza doesn't work):

pvirtue@andrew.cmu.edu

Course Information

Lectures

- Lectures are recorded
    - Shared with our course and ML course staff only
- Participation point earned by answering Piazza polls in lecture
- Quizzes will in lecture, announced two days ahead of time
- Slides will be posted

Recitations

- Recommended attendance
- No plans to record at this point
- No participation points in recitation
- Recitation materials are in-scope for quizzes and exams

Course Information

Office Hours

- OH calendar on course website
- OH-by-appointment requests are certainly welcome

Mental Health