

# **Machine Learning I**

# **60629A**

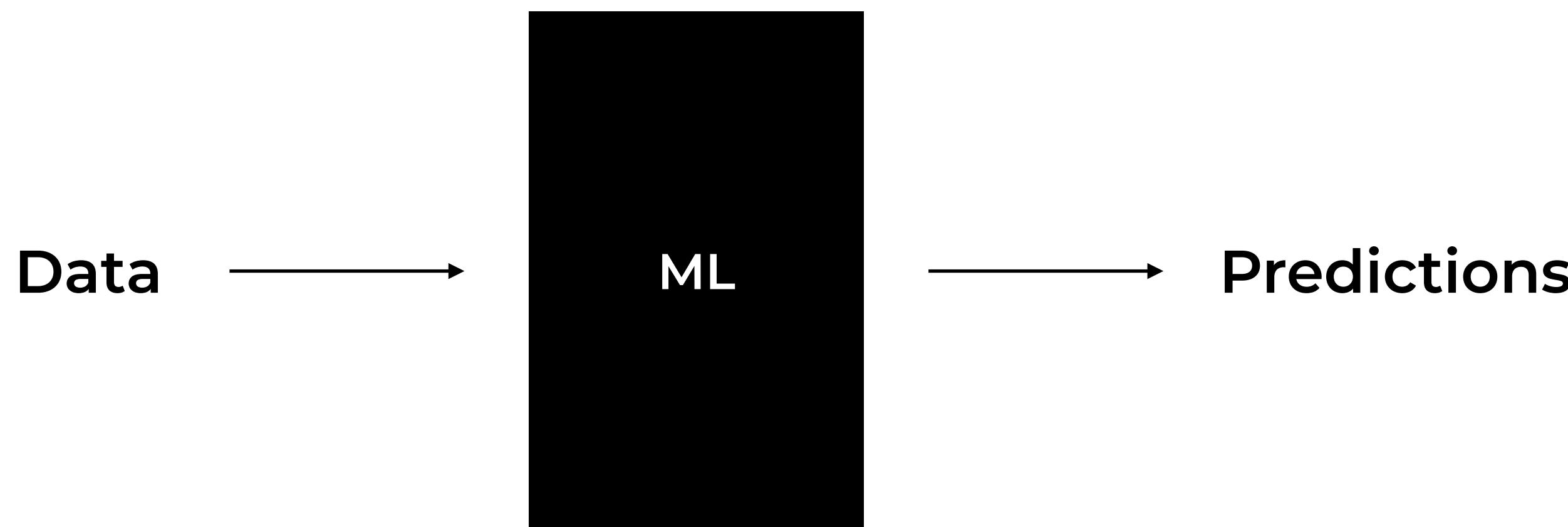
**Week #1**

# Today

- **Introduction to machine learning**
- **The course (syllabus)**
- **Math review (probability + linear algebra)**

# Machine Learning (ML)

- Science that studies statistical and computational aspects of modeling data for predictive purposes
  - (Mostly) Empirical science



- **Task: Predict whether an image contains a malignant tumor**
- **Task: Predict the next movie a person should watch**

THIS IS YOUR MACHINE LEARNING SYSTEM?

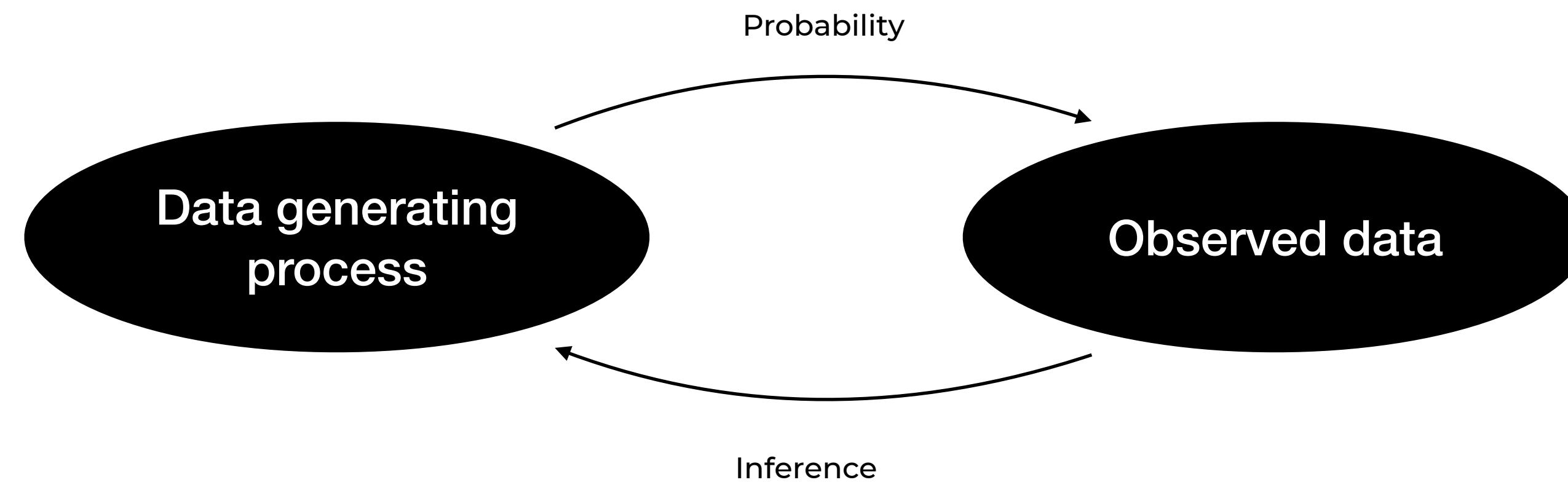
YUP! YOU POUR THE DATA INTO THIS BIG  
PILE OF LINEAR ALGEBRA, THEN COLLECT  
THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL  
THEY START LOOKING RIGHT.



**“Data analysis, machine learning and data mining are various names given to the practice of statistical inference, depending on the context.”**



*–Larry Wasserman in “All of Statistics: A Concise Course in Statistical Inference.”*

**What is the goal of ML?**

- A bit of historical context
  - When I started my PhD very few in ML talked about AI
  - Recent ML makes progress toward “AI tasks”
    - Examples of AI tasks: translation, object recognition
    - In that context: create a machine with human-like capacities? Or a machine that can help humans?

- My (imperfect) view:

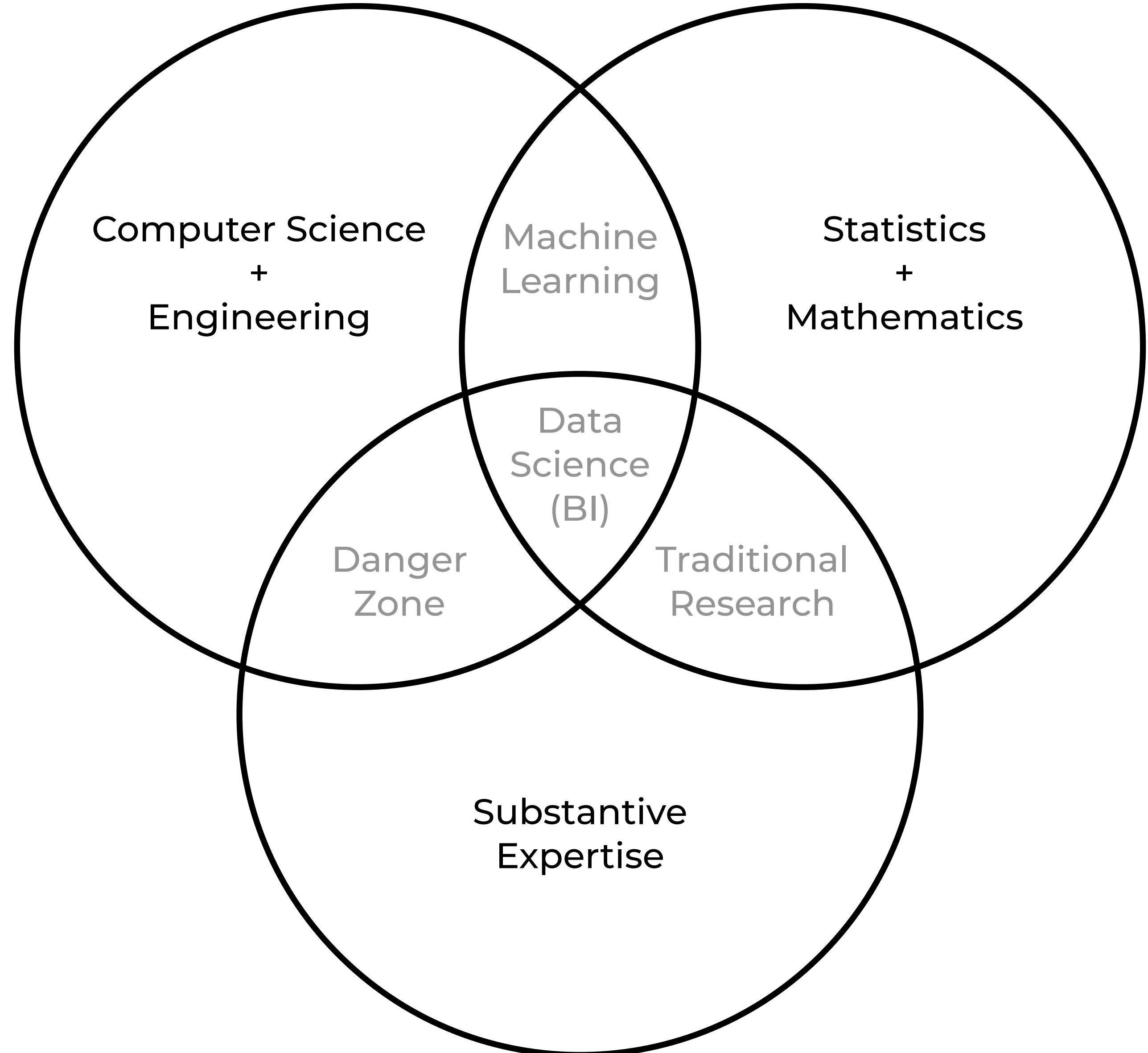
**Understand data through predictive models**

**Understand the world through predictive models**

**How does ML relate  
to other fields**

# Historical View

- **(Modern) Statistics:** ~1900
- **Machine Learning and Data Mining:** ~1960
- **Data Science:** ~2000



# Attitudes in Machine Learning and Data Mining Versus Attitudes in Traditional Statistics

Despite these differences, there's a big overlap in problems addressed by machine learning and data mining and by traditional statistics. But attitudes differ...

## Machine learning

No settled philosophy or widely accepted theoretical framework.

Willing to use *ad hoc* methods if they seem to work well (though appearances may be misleading).

Emphasis on automatic methods with little or no human intervention.

Methods suitable for many problems.

Heavy use of computing.

## Traditional statistics

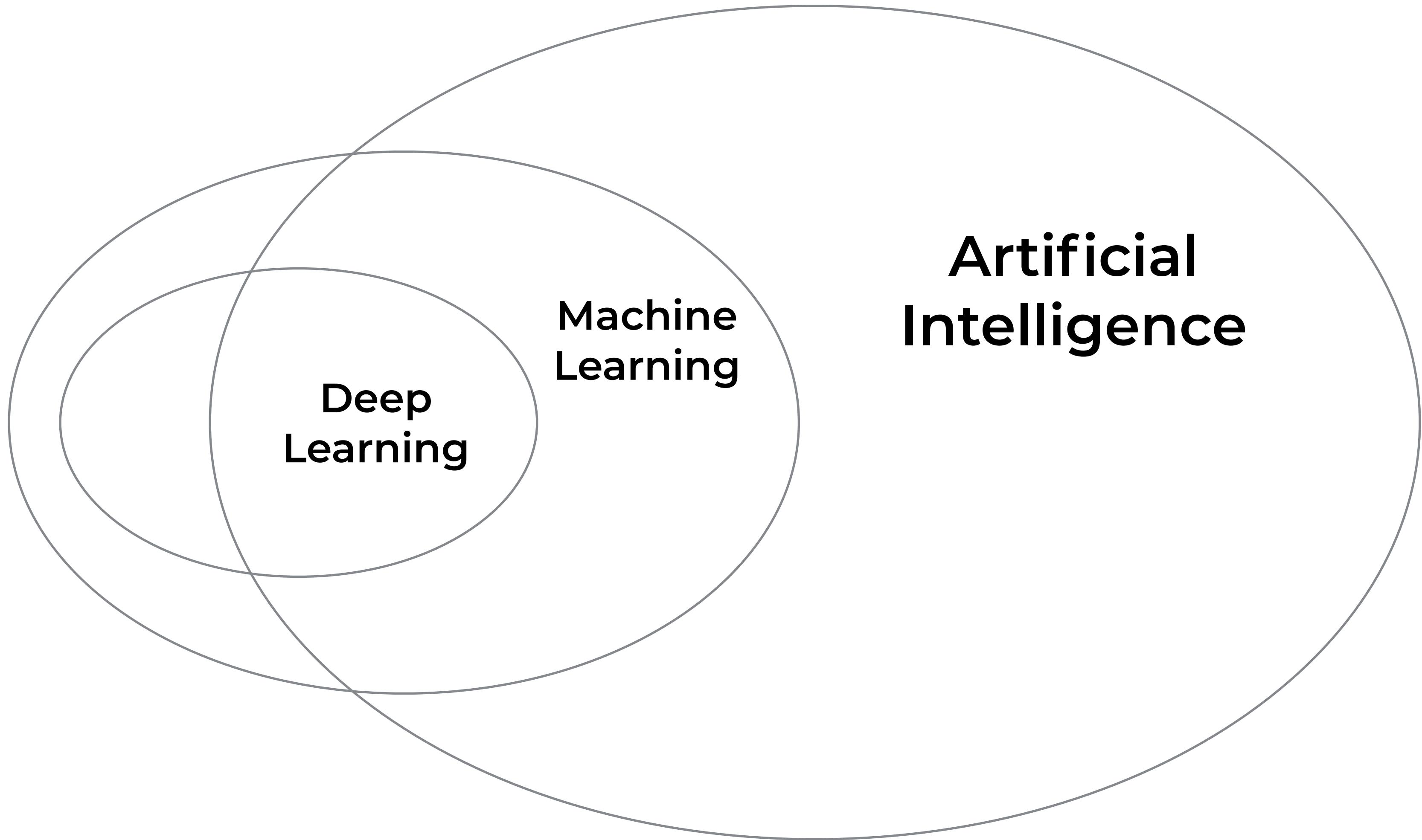
Classical (frequentist) and Bayesian philosophies compete.

Reluctant to use methods without some theoretical justification (even if the justification is actually meaningless).

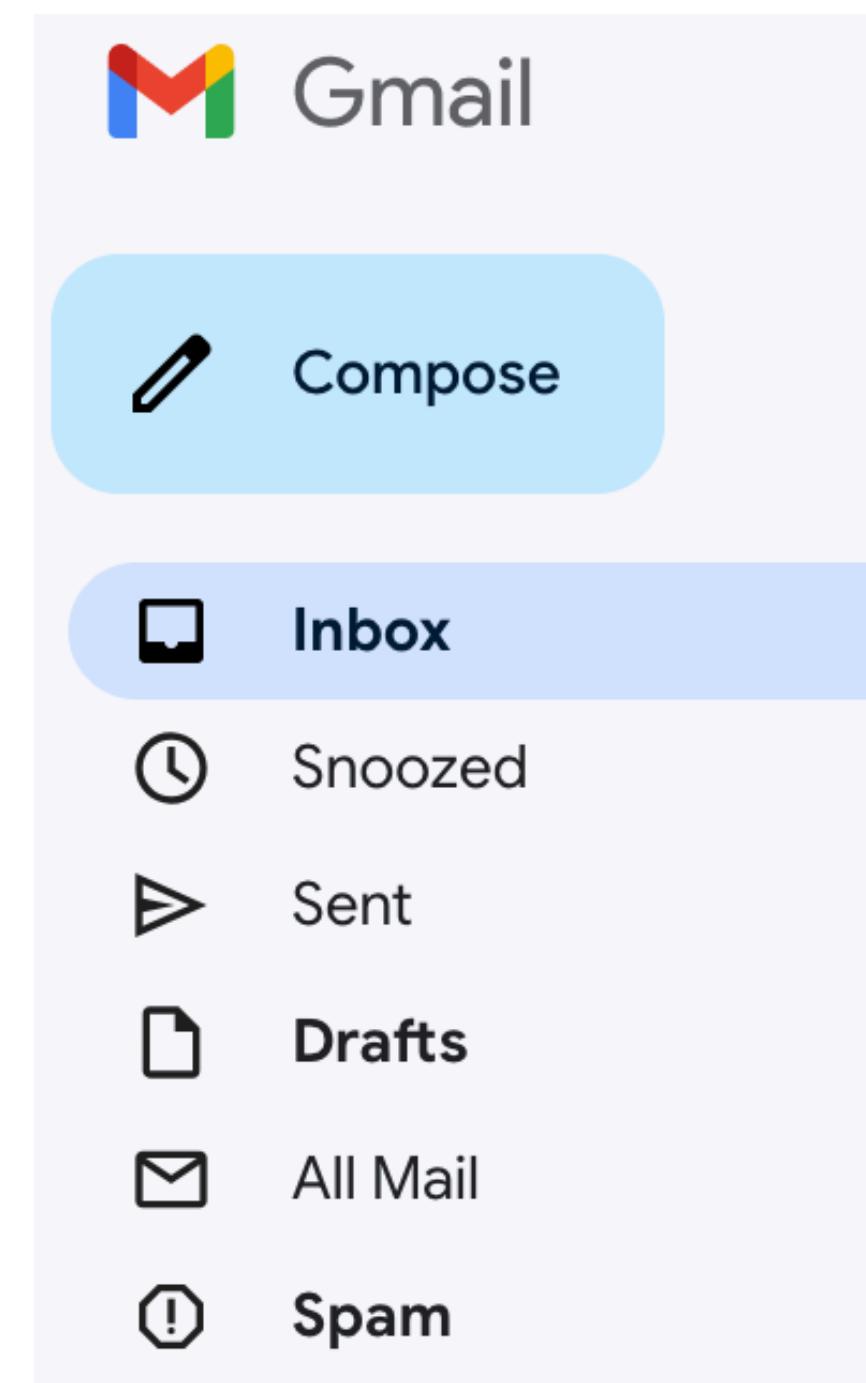
Emphasis on use of human judgement assisted by plots and diagnostics.

Models based on scientific knowledge.

Originally designed for hand-calculation, but computing is now very important.



# Applications of ML



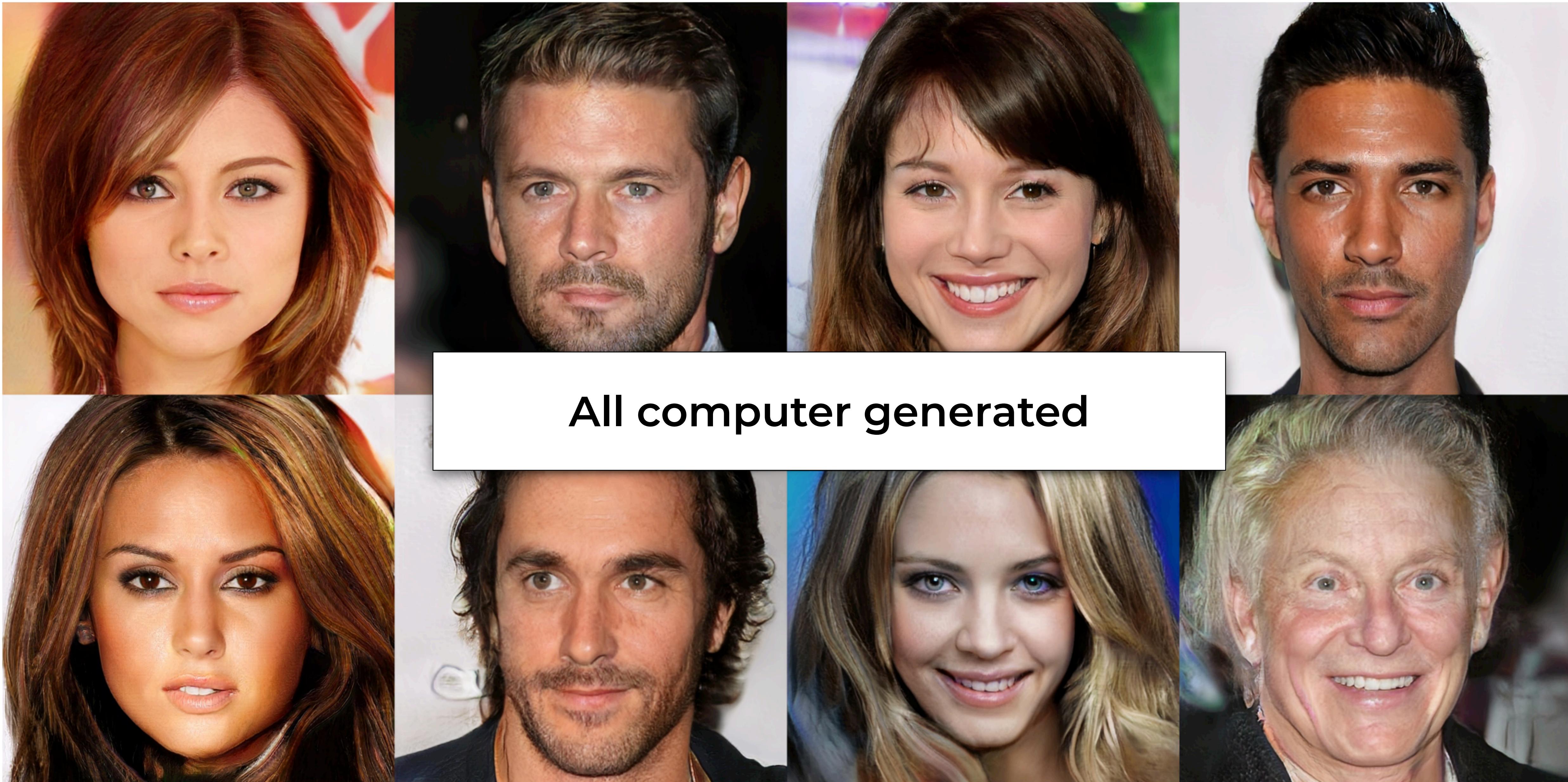
# Google

artificial in|

- artificial intelligence
- artificial insemination
- artificial intelligence movie  
A.I. Artificial Intelligence — 2001 film
- artificial intelligence definition
- artificial intelligence in healthcare



Progressive Growing of GANs for Improved  
Quality, Stability, and Variation  
Karras et al., ICLR'18



Progressive Growing of GANs for Improved  
Quality, Stability, and Variation  
Karras et al., ICLR'18



2014



2015



2016



2017



2018



"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."



"two young girls are playing with lego toy."



"boy is doing backflip on wakeboard."



"girl in pink dress is jumping in air."



"black and white dog jumps over bar."



"young girl in pink shirt is swinging on swing."



"man in blue wetsuit is surfing on wave."



"man in black shirt is playing guitar."



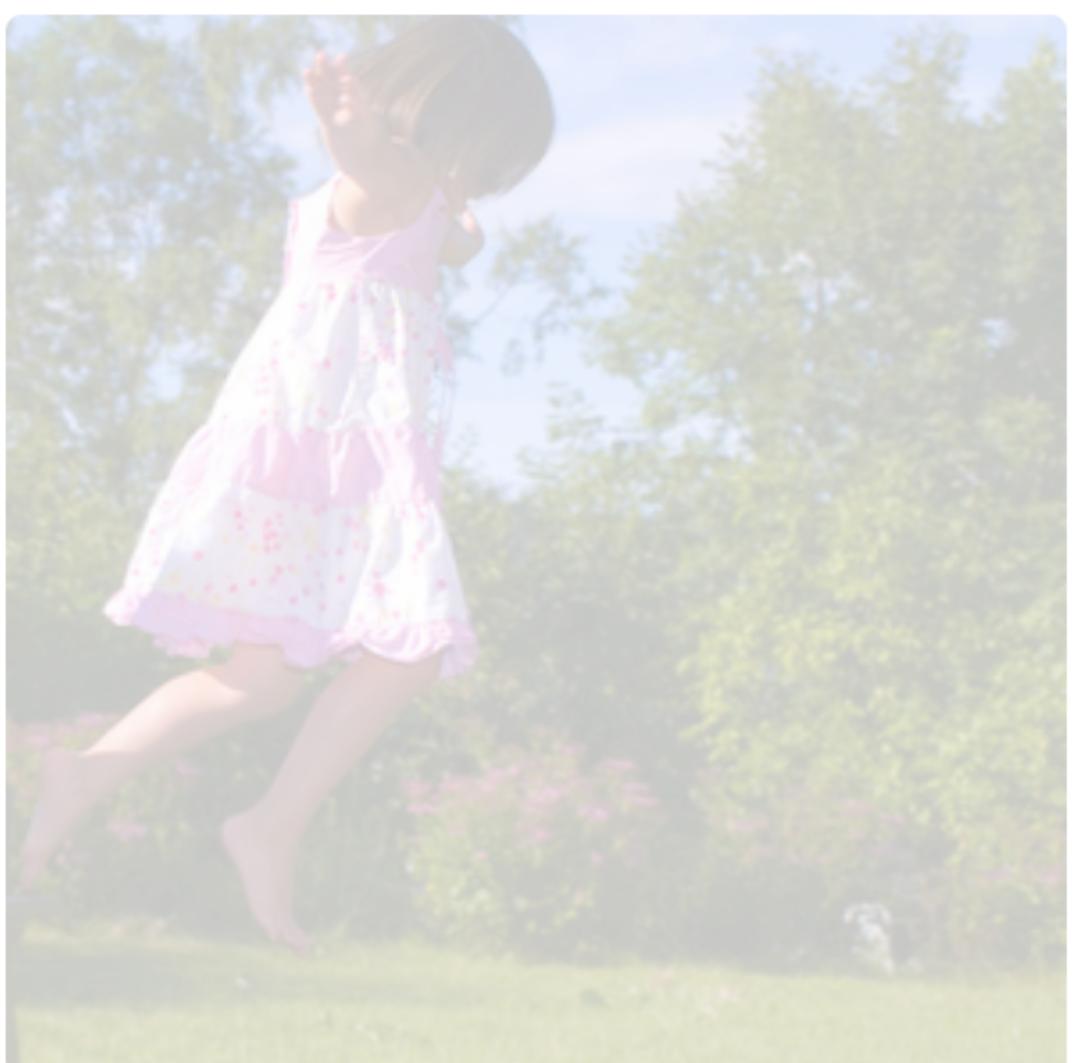
"construction worker in orange safety vest is working on road."



"two young girls are playing with lego toy."



"boy is doing backflip on wakeboard."



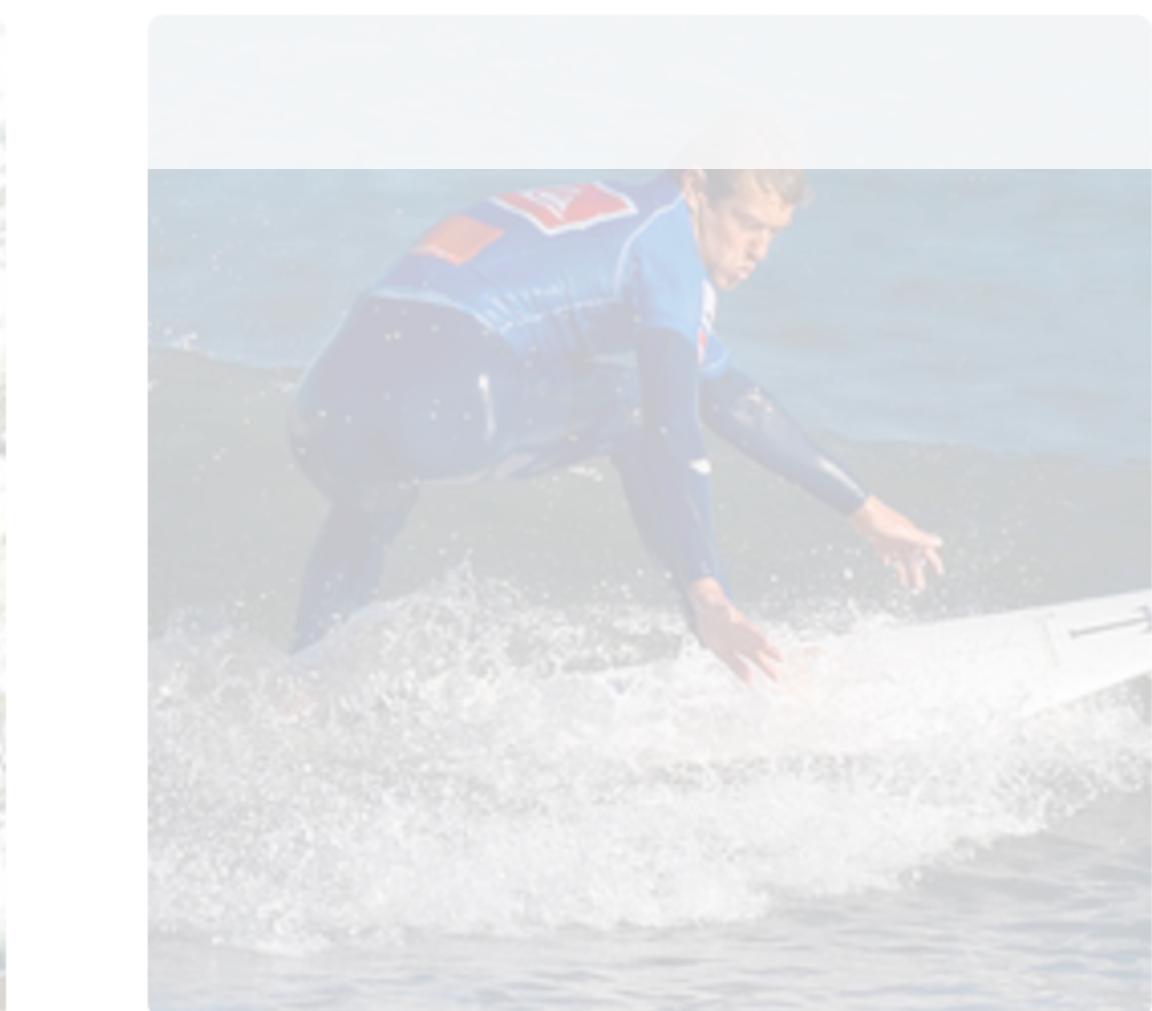
"girl in pink dress is jumping in air."



"black and white dog jumps over bar."



"young girl in pink shirt is swinging on swing."



"man in blue wetsuit is surfing on wave."

TEXT PROMPT

an armchair in the shape of an avocado....

AI-GENERATED  
IMAGES



# Generate an image from a text prompt

TEXT PROMPT

A bowl of soup that looks like a monster spray-painted on the wall

AI-GENERATED  
IMAGES

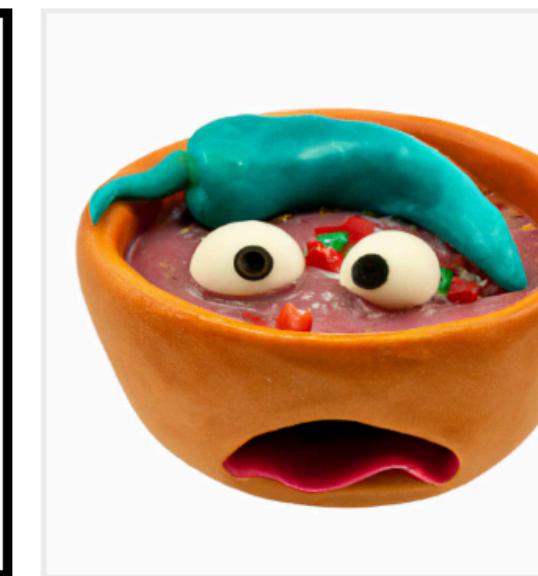


# Generate an image from a text prompt

TEXT PROMPT

A bowl of soup that looks like a monster [made out of plasticine](#)

AI-GENERATED  
IMAGES





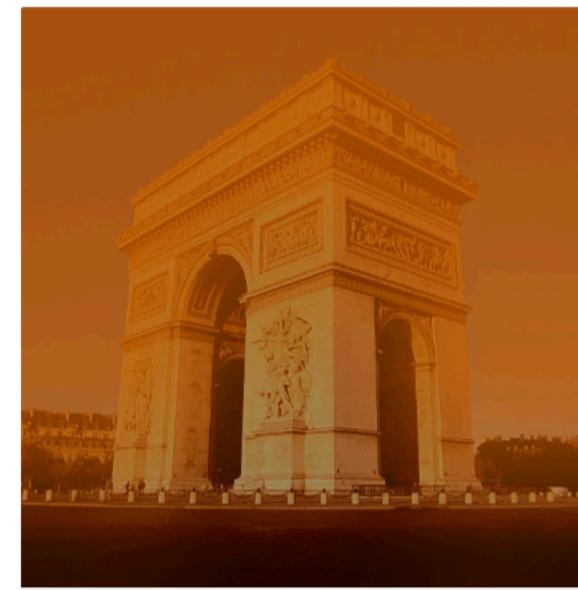
This Climate Does Not Exist

These images  
are generated by  
artificial intelligence (AI)



[Visualize climate change](#)

[What now](#)



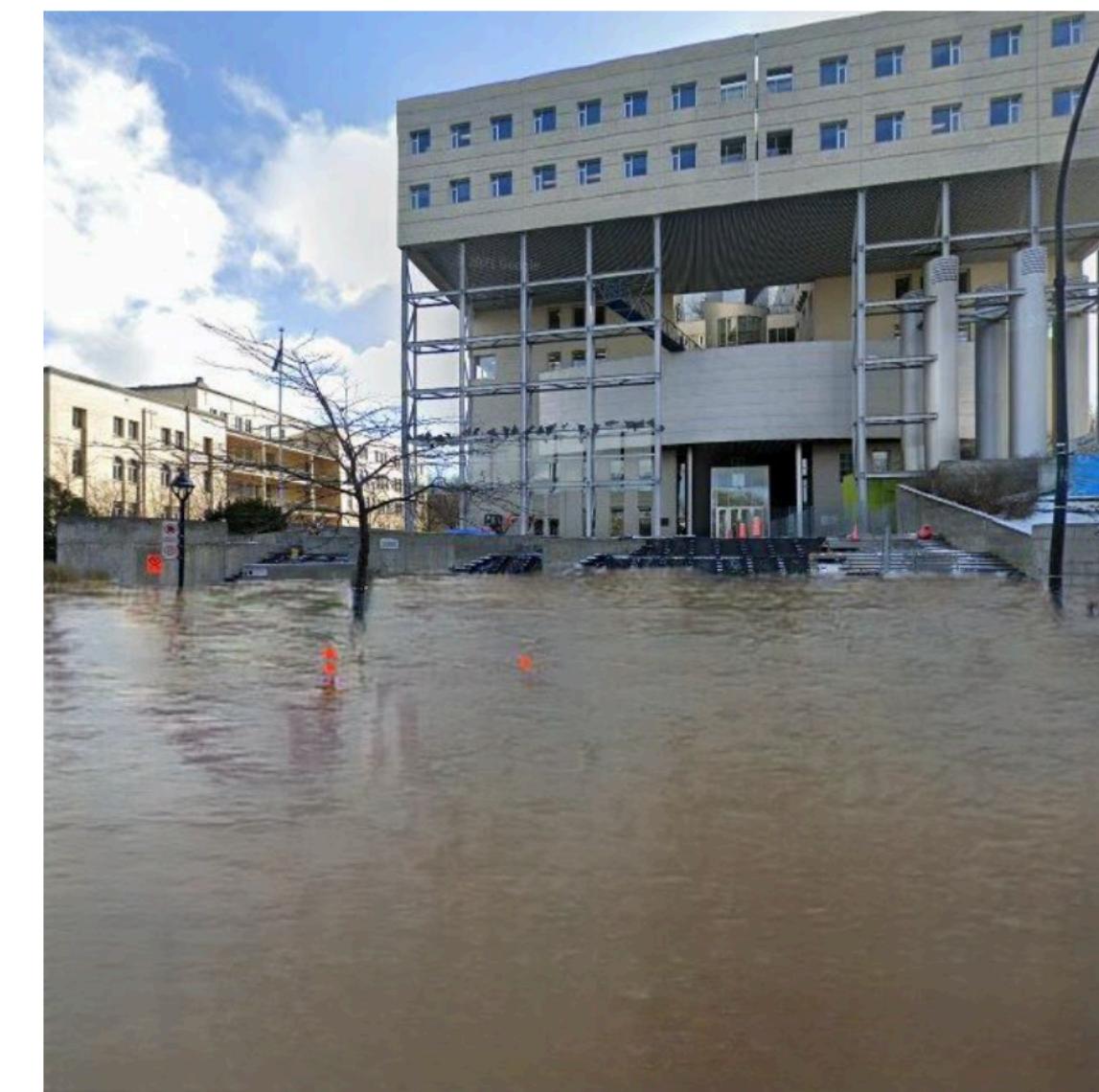
But the environmental  
disasters they portray  
are very real



Can you imagine these kinds of disasters  
happening in your own backyard?



This Climate Does Not Exist



## “Lombard street ...”

“view from bottom”



“view from top”



“bird’s eye view”



“in heavy rain”



“Lombard street ...”

“view from bottom”



“view from top”



“bird’s eye view”



“in heavy rain”

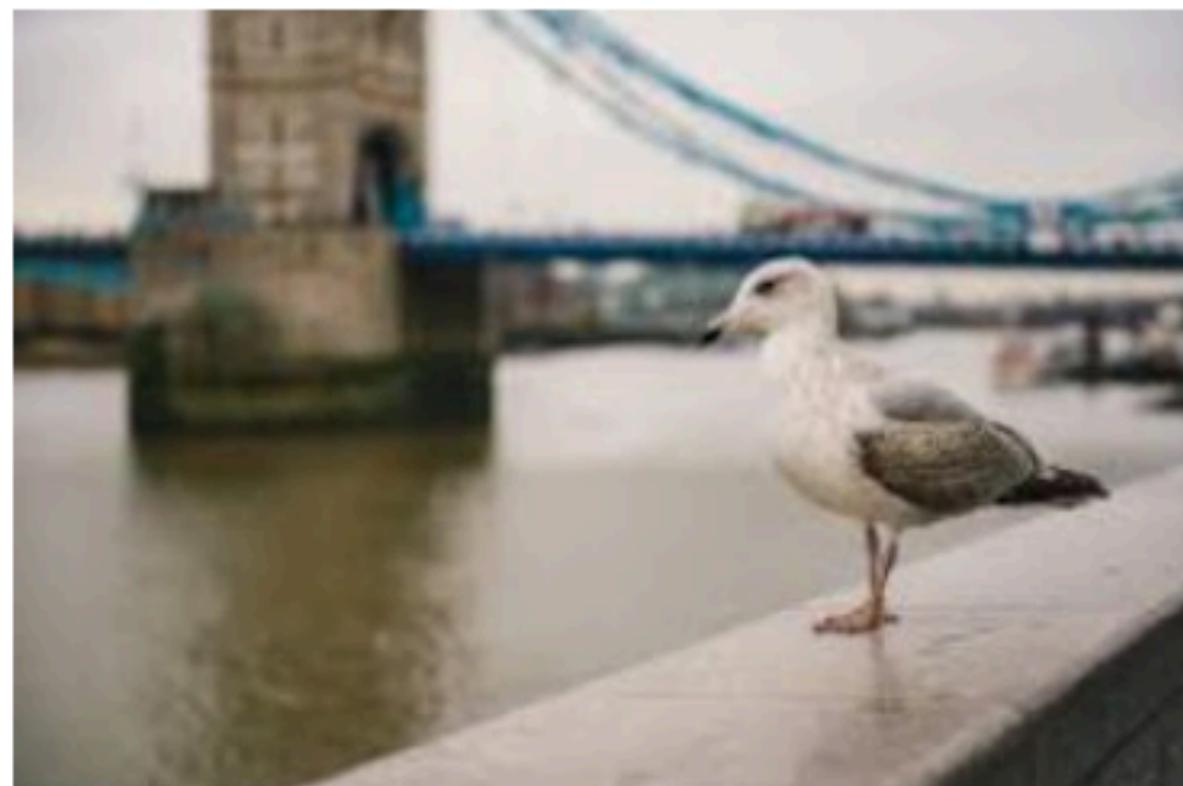


“seagull in front of ...”

“Golden Gate Bridge”



“London Tower Bridge”

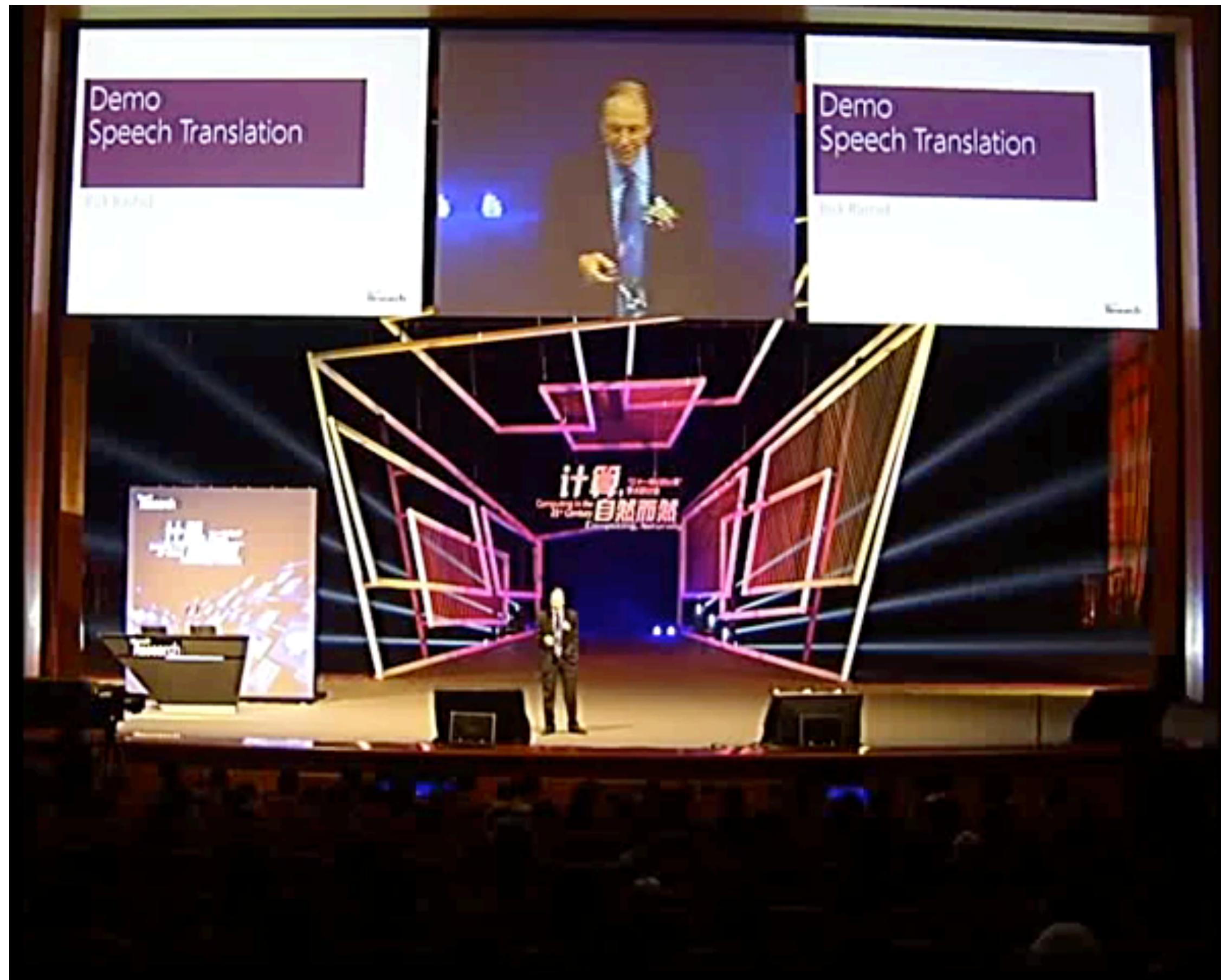


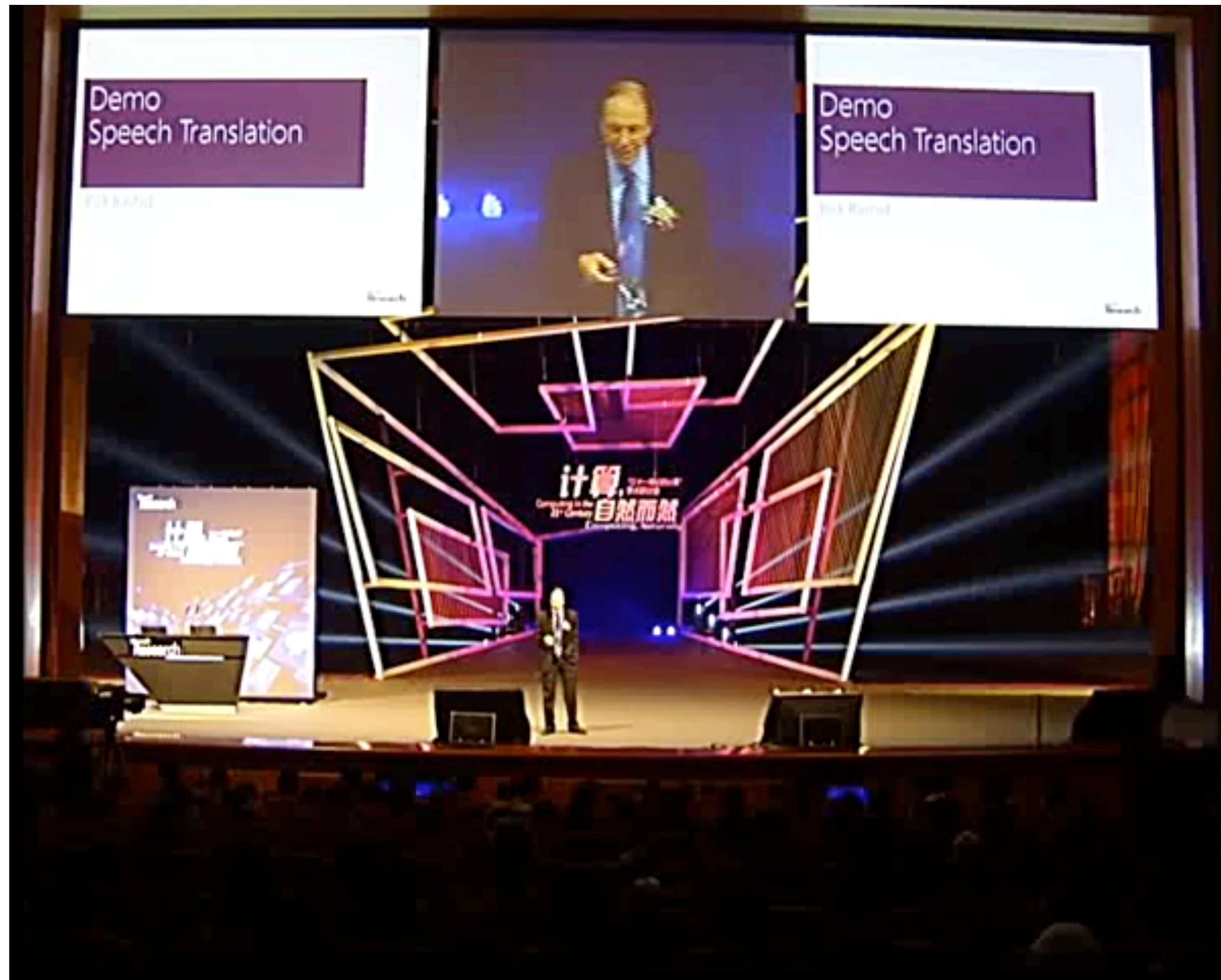
“Sydney Harbour Bridge”



“Rialto Bridge”









Robotic observatory makes  
fast work of astronomy p. 479

A wet route to  
methanol p. 523

Human noise plagues  
protected areas p. 531

# Science

\$15  
5 MAY 2017  
[sciencemag.org](http://sciencemag.org)

AAAS

A  
♠

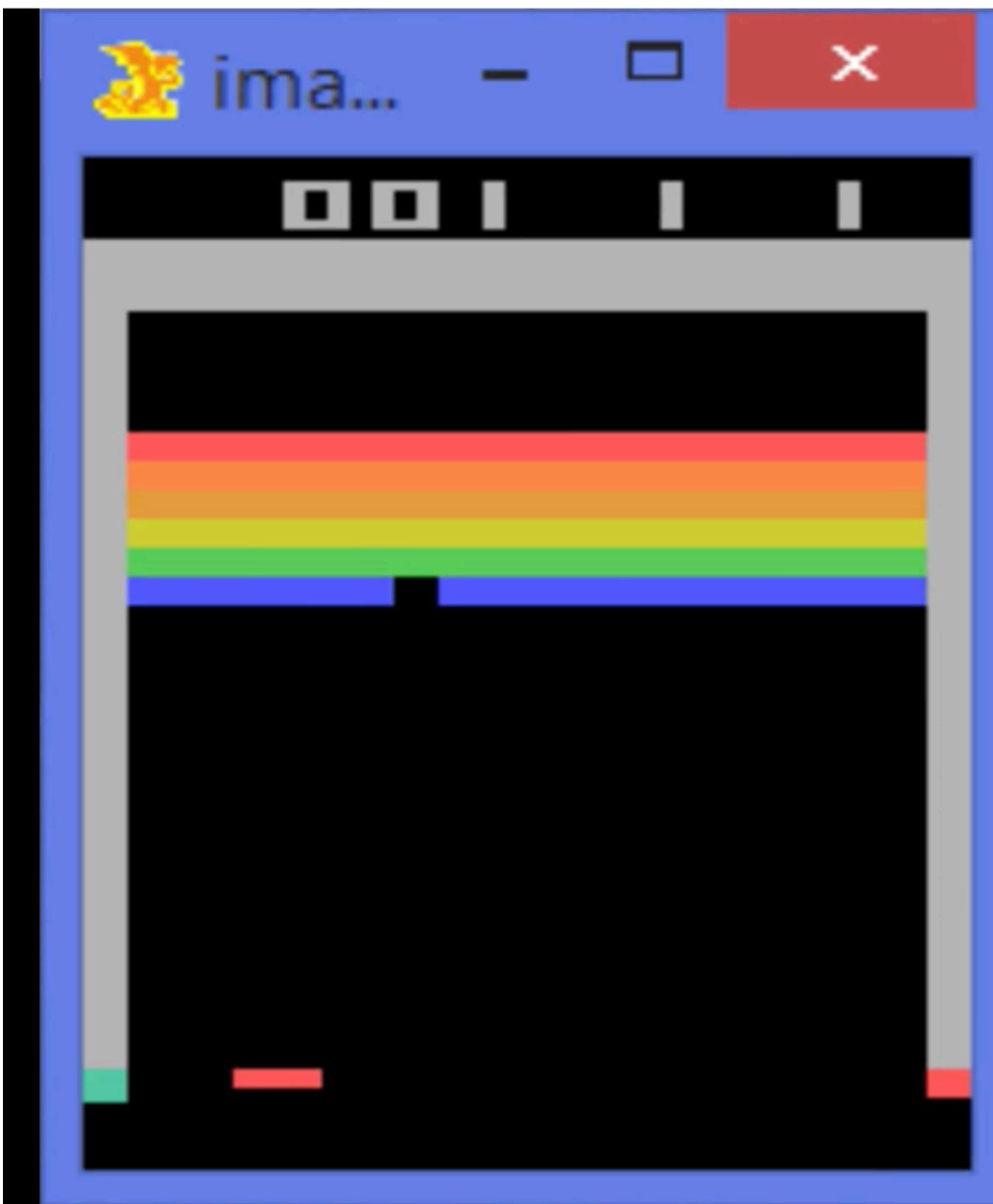
## DIGITAL CARDS WHIZ

AI beats humans at  
challenging poker variant  
p. 508



A  
♠

# AI for video games



Mnih et al.  
Nature  
Volume 518,  
pages 529–533  
(26 February 2015)





- Medicine: personalized, automate diagnostics
- Social sciences: prediction problem (e.g., predict recidivism)
- Engineering: to propose new design, evaluate without building
- Finance: capture uncertainty, short-term trading
- Marketing: to understand and quantify user experience, advertising efficacy
- Many others: conservation, social projects, climate change
- Your domain of expertise...

# **Course Introduction & Goals**

# Logistics

- Course syllabus: <http://www.cs.toronto.edu/~lcharlin/courses/60629/>

# Flipped Classroom

- Every week:
  1. Class preparation (Offline):
    - Weekly material (~90 minutes)
    - Reading, watching capsules
  2. Class time (Online):
    - Summary, Q&A, problem solving (120 minutes)

# Suggestions for navigating a flipped classroom

- In class: Come prepared
  - Watch the capsules ahead of time
  - Do the readings
  - Write down your questions
- Capsules: Stay active while watching the capsules (e.g., take notes, pause, go back, think of how it fits in the broader context)

# Fit with other courses

- HEC
  - PhD level (originally)
  - Computationally oriented
  - Prequel to
    - Machine Learning II: Deep Learning (MATH 60630A)
    - Trustworthy Machine Learning (MATH 80630)
  - Other ML courses in Montreal (U.Montreal, Polytechnique, McGill)
    - More applied (similar to COMP-551@McGill)

# **Short review of linear algebra, statistics, and probabilities**

- Based on chapters 2 and 3 of “Deep Learning”

<http://www.deeplearningbook.org/>

# Linear algebra

- **Scalar: a single value.**

$$\mathbf{a} \in \mathbb{R}, \mathbf{a} \in \mathbb{N} \quad \mathbf{a} = 3$$

- **Vector: an array of values.**

$$\mathbf{a} \in \mathbb{R}^D, \mathbf{a} \in \mathbb{N}^D \quad \mathbf{a} = \begin{bmatrix} 3 \\ 4 \\ 2 \end{bmatrix}$$

- **Matrix: a table of values.**

$$\mathbf{A} \in \mathbb{R}^{D_1 \times D_2}, \mathbf{A} \in \mathbb{N}^{D_1 \times D_2} \quad \mathbf{A} = \begin{bmatrix} 3 & 4 & 2 \\ 1 & 2 & 9 \end{bmatrix}$$

# Indexing notation

- Indexing elements of a vector:  $a_i$

$$a = \begin{bmatrix} 3 \\ 4 \\ 2 \end{bmatrix} \leftarrow a_1$$

Convention:  
The first element  
is the zero'th.

- Indexing elements of a matrix:  $a_{ij}$

$$A = \begin{bmatrix} 3 & 4 & 2 \\ 1 & 2 & 9 \end{bmatrix}$$

$\uparrow$   
 $a_{12}$

# Simple operations

- Transpose

$$a = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} \quad \left| \quad (A_{ij})^\top = A_{ji}$$
$$a^\top = [a_0 \ a_1 \ a_2]$$

- Addition

- Vectors and matrices w. the same shape

$$b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} \quad \left| \quad (A + B)_{ij} = A_{ij} + B_{ij}$$
$$a + b = \begin{bmatrix} a_0 + b_0 \\ a_1 + b_1 \\ a_2 + b_2 \end{bmatrix}$$

# Simple operations

- Multiply by a scalar

$$\alpha \mathbf{a} = \begin{bmatrix} \alpha \mathbf{a}_0 \\ \alpha \mathbf{a}_1 \\ \alpha \mathbf{a}_2 \end{bmatrix}$$

- Vector product.

- The dot product

$$\mathbf{a}^\top \mathbf{a} = \sum_i \mathbf{a}_i \mathbf{a}_i$$

- Note: it yields a scalar.

- Element-wise product:

$$\mathbf{a} \odot \mathbf{a} = \begin{bmatrix} \mathbf{a}_0 \mathbf{a}_0 \\ \mathbf{a}_1 \mathbf{a}_1 \\ \mathbf{a}_2 \mathbf{a}_2 \end{bmatrix}$$

- Also known as Hadamard product

# Operations

- Matrix product (dot product):

$$C_{ij} = \sum_k A_{ik} B_{kj}$$

- A's columns must equal B's rows (order is important)

$$\mathbf{A} \in \mathbb{R}^{D_1 \times D_2}, \mathbf{B} \in \mathbb{R}^{D_2 \times D_3}$$

- Distributive:  $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
- Associative:  $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$
- Product of transpose:  $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$

# Inverse

- We denote a matrix's inverse as  $A^{-1}$
- A matrix has an inverse iff:
  - it's square.  $D_1 = D_2$
  - its columns are linearly independent.
    - No column can be recovered using a combination of other columns
- Inverses are useful to solve systems of equations:

$$Ax = b \quad x = A^{-1}b$$

A square matrix  
not invertible is *singular*

# Norms

- $L^p$  norm. Size of a vector (or matrix)

$$\| \mathbf{a} \|_p = \left( \sum_i |a_i|^p \right)^{1/p}$$

- Standard norms in ML:

- Euclidean norm ( $p=2$ )

$$\| \mathbf{a} \|_2 = \sqrt{\left( \sum_i |a_i|^2 \right)}$$

- Dot product w. 2-norm:  $\mathbf{a}^\top \mathbf{b} = \|\mathbf{a}\|_2 \|\mathbf{b}\|_2 \cos \theta_{ab}$

- Frobenius norm (matrix):  $\| \mathbf{A} \|_2 = \sqrt{\left( \sum_i \sum_j |a_{ij}|^2 \right)}$

# Special matrices & vectors

- Identity. Denoted  $I_n$ .

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- All zeros except for ones on the main diagonal.
- Symmetric:  $A = A^\top$
- Unit vector:  $\|a\|_2 = 1$
- Orthogonal vectors:  $a^\top b = 0$
- Orthonormal vectors: unit and orthogonal  $A^\top A = AA^\top = I$
- Orthogonal matrix: Orthonormal rows & columns

- Skip eigendecomposition, SVD, pseudo-Inverse, determinants (Sections 2.7–2.11).
  - We will get back to them if/when needed in the course.

- On to probabilities
- Chapter 3 of “Deep Learning”
  - I’ve adapted some of the lecture slides from the book.
  - Thanks to Ian Goodfellow for providing slides.

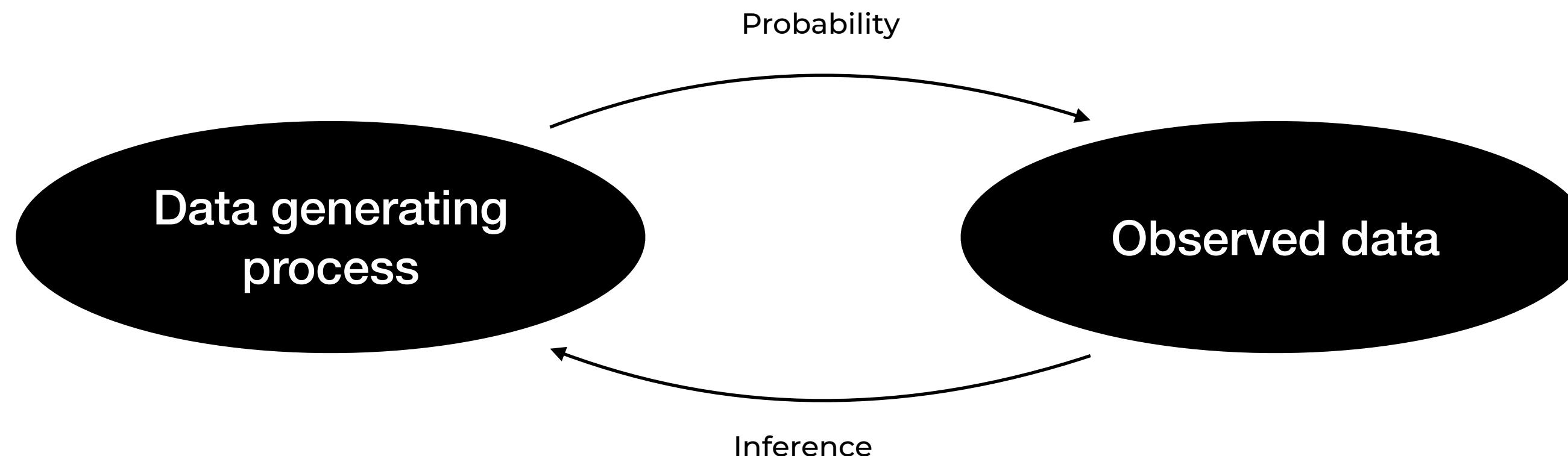
# Why probabilities?

- To capture uncertainty

E.g., What time will I get home tonight?

- Probabilities provide a formalism for making statements about “data generating processes” (L. Wasserman)

E.g., what happens when I flip a fair coin?



# The example

- Generate data by throwing a fair die.
- What do we know about a single throw?
  - 6 possible outcomes. (**sample space**)
  - Each outcome (e.g., 1). (**element, state**)
    - A subset of outcomes (e.g., <3). (**event**)
    - Outcomes are equiprobable. (**uniform distribution**)

# Random variables and probabilities

- A random variable (r.v.) is a probabilistic outcome.
  - For example,
    - Die throw ( $X$ )
      - The actual outcome is  $\in \{1, 2, 3, 4, 5, 6\}$ . ( $x$ )
    - A probability function ( $P$ ) assigns a real number to each possible event:  $P(x) \geq 0, \forall x \in X$

$$P(\bigcup x) = 1$$

# Discrete RVs

- An RV is discrete if it takes a finite number of values<sup>1</sup>

$$\begin{aligned} P(x = x_i) &\geq 0, \forall i \\ \sum_i P(x = x_i) &= 1 \end{aligned}$$

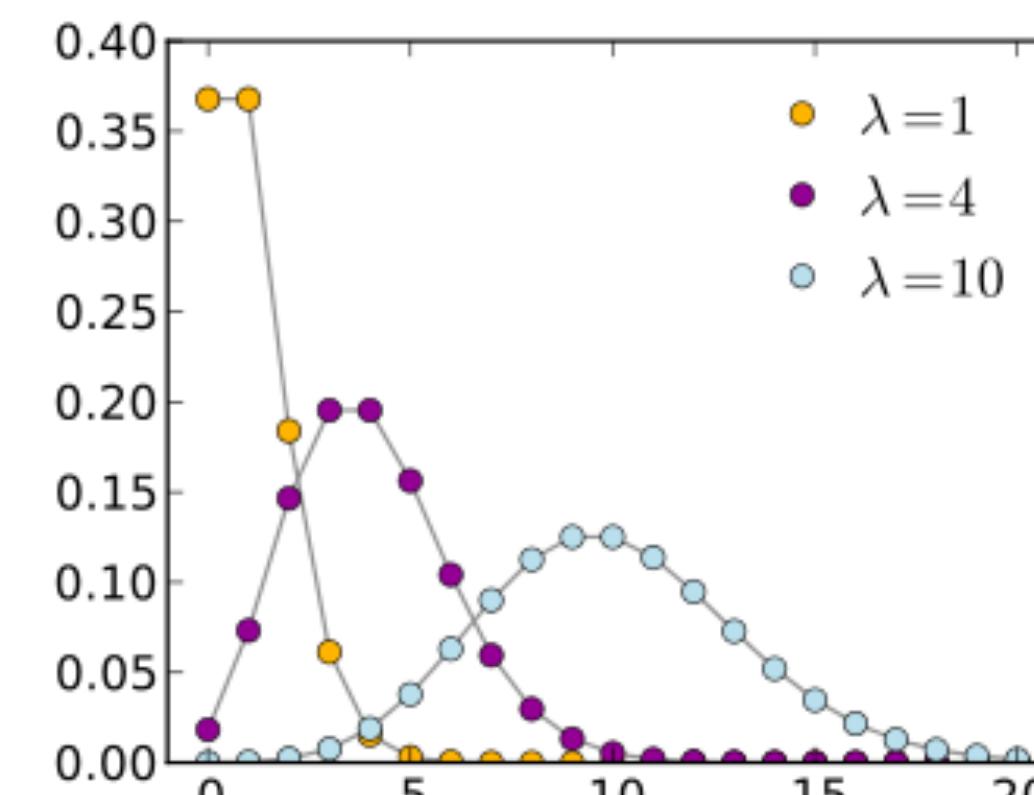
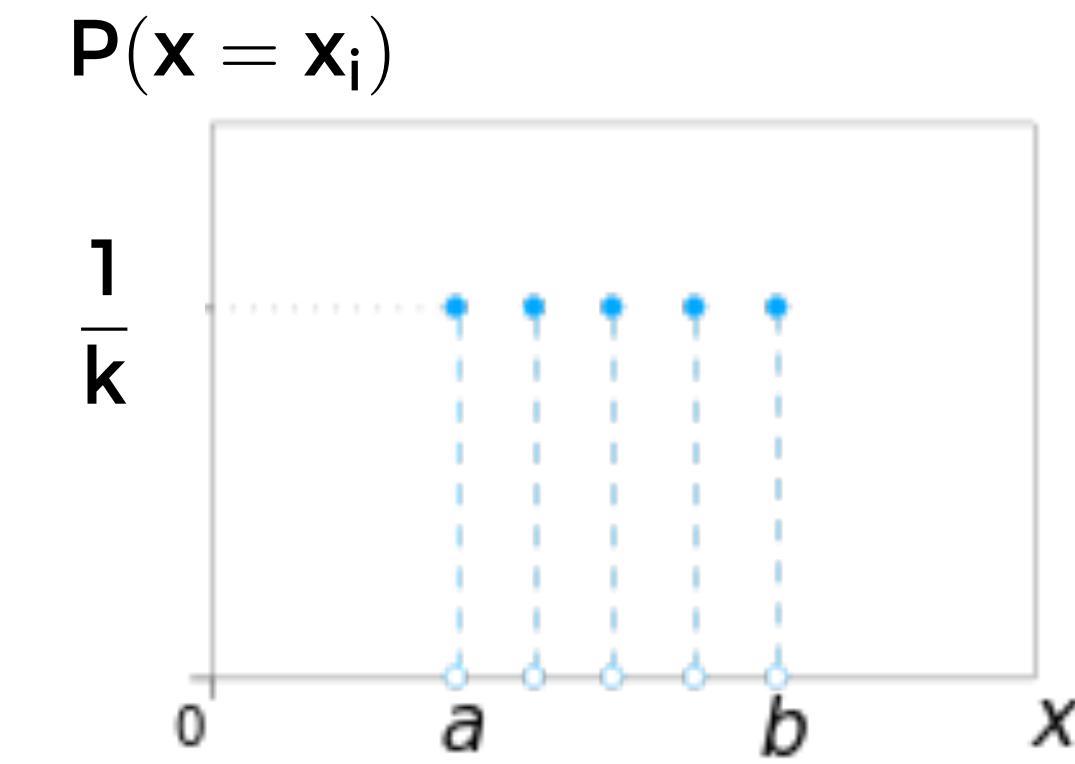
- E.g., uniform distribution:

$$P(x = x_i) = \frac{1}{k}, \forall i$$

- E.g., Poisson distribution:

$$P(x = x_i; \lambda) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

1. technically: it must be countable



# Continuous RVs

- An RV is continuous if  $f(x) \geq 0, \forall x \in X$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

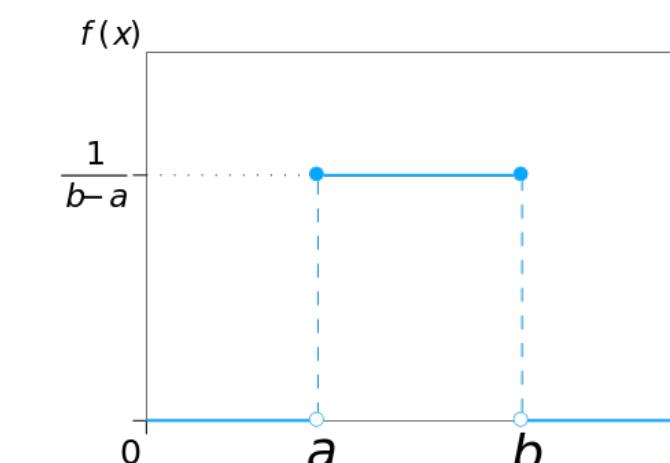
$$P(a < x < b) = \int_a^b f(x)dx$$

- $f(x)$  is a probability density function (PDF)

- E.g., (continuous) uniform distribution:

$$u(x; a, b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

- E.g., Gaussian distribution



from: wikipedia.org

# A few useful properties

(shown for discrete variables for simplicity)

- **Sum rule:**  $P(X) = \sum_Y P(X, Y)$
- **Product rule:**  $P(X, Y) = P(X | Y)P(Y)$
- **Chain rule:**  $P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | X_1, \dots, X_{i-1})P(X_1)$
- **If x and y are independent:**  $P(X, Y) = P(X)P(Y)$
- **Bayes' Rule:**  $P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}$

# Moments

- **Expectation:**  $\mathbb{E}[X] = \sum_i P(x = x_i)x_i \quad \mathbb{E}[aX] = a\mathbb{E}[X]$
- **Variance:**  $\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$
- **Covariance:**  $\text{Cov}(X, Y) = \mathbb{E} [(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$
- **correlation:**  $\rho(x, y) = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$

# Further Reading

- Prologue to “The Master Algorithm”  
<http://homes.cs.washington.edu/~pedrod/Prologue.pdf>
- Ch. 1 of Hastie et al.
- Math Preparation
  - Ch.2 of Pattern Recognition and Machine Learning [PRML]
  - Ch.2-3 of Deep Learning [DL]
  - Slightly more advanced:

<http://www.cs.mcgill.ca/~dprecup/courses/ML/Materials/prob-review.pdf>

<http://www.cs.mcgill.ca/~dprecup/courses/ML/Materials/linalg-review.pdf>