Machine Learning I 60629A

Week #1

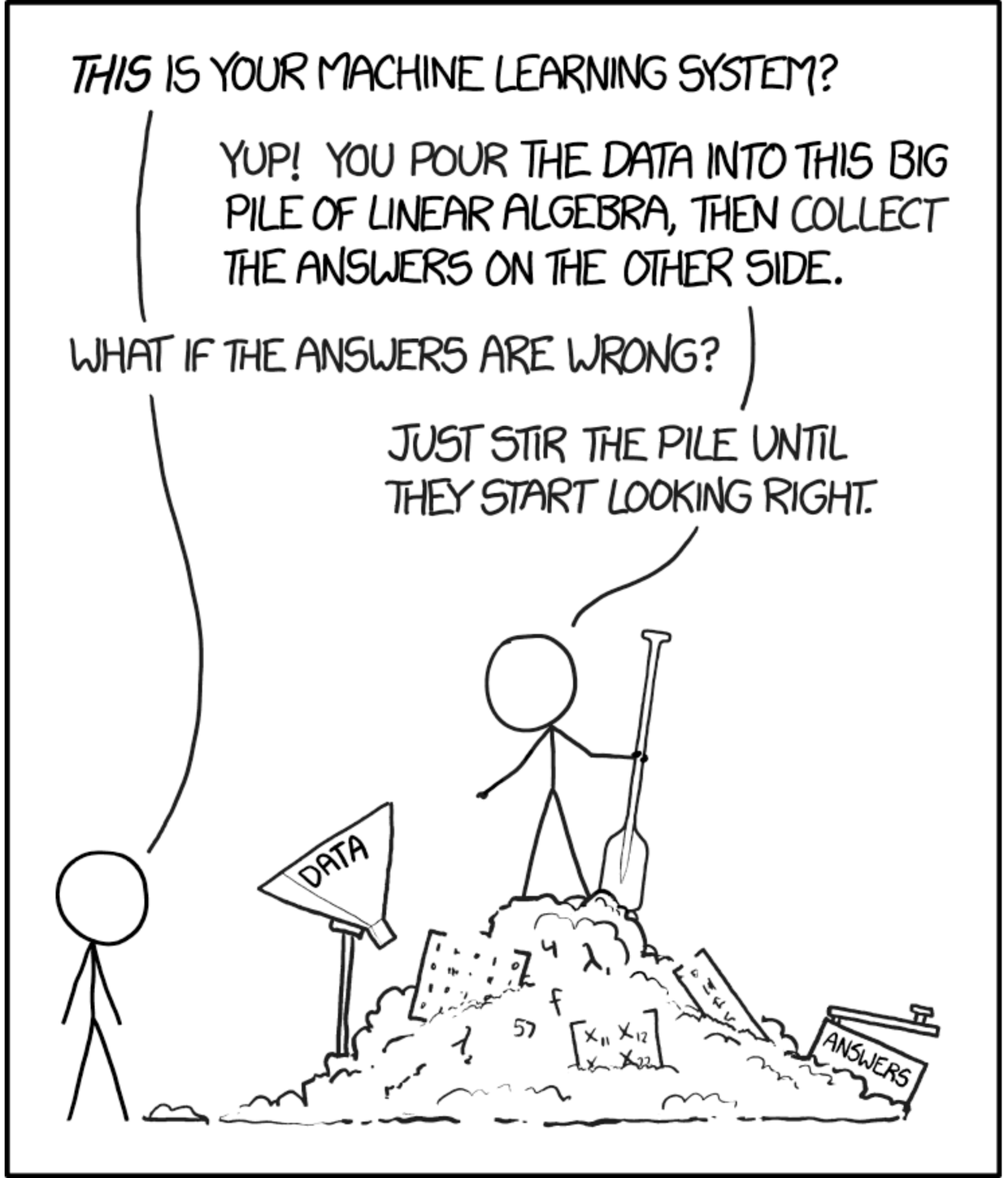
Today

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

Machine Learning (ML)

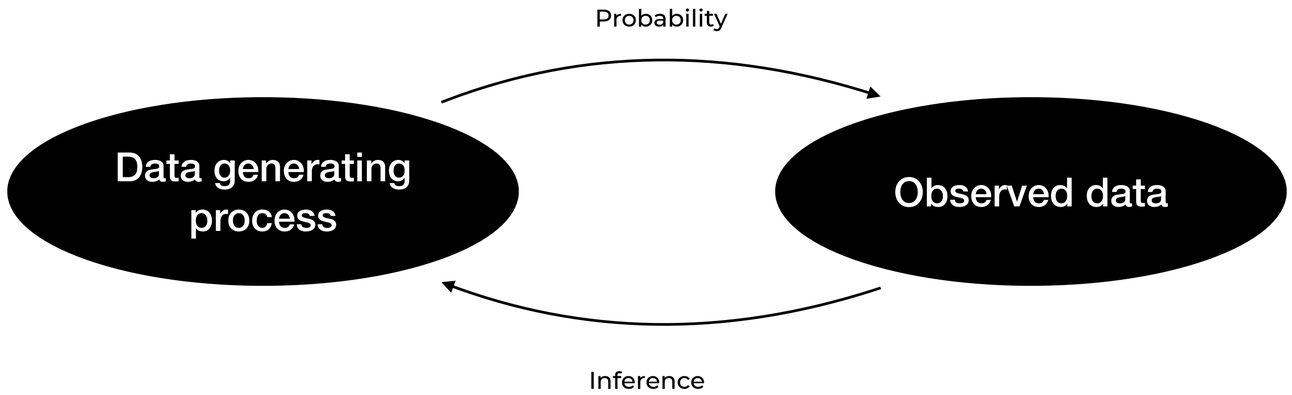
- Science that studies statistical and computational aspects of modeling data for predictive purposes
  - (Mostly) Empirical science

- Task: Predict whether an image contains a malignant tumor
- Task: Predict the next movie a person should watch
- Task: Answer a question, summarize a long text, provide an analysis of a video...



https://xkcd.com/1838/

“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: dialogue (think ChatGPT), image recognition, image generation
    - In that context: create a machine with human-like capacities? Or a machine that can help humans?

- **For this course:**

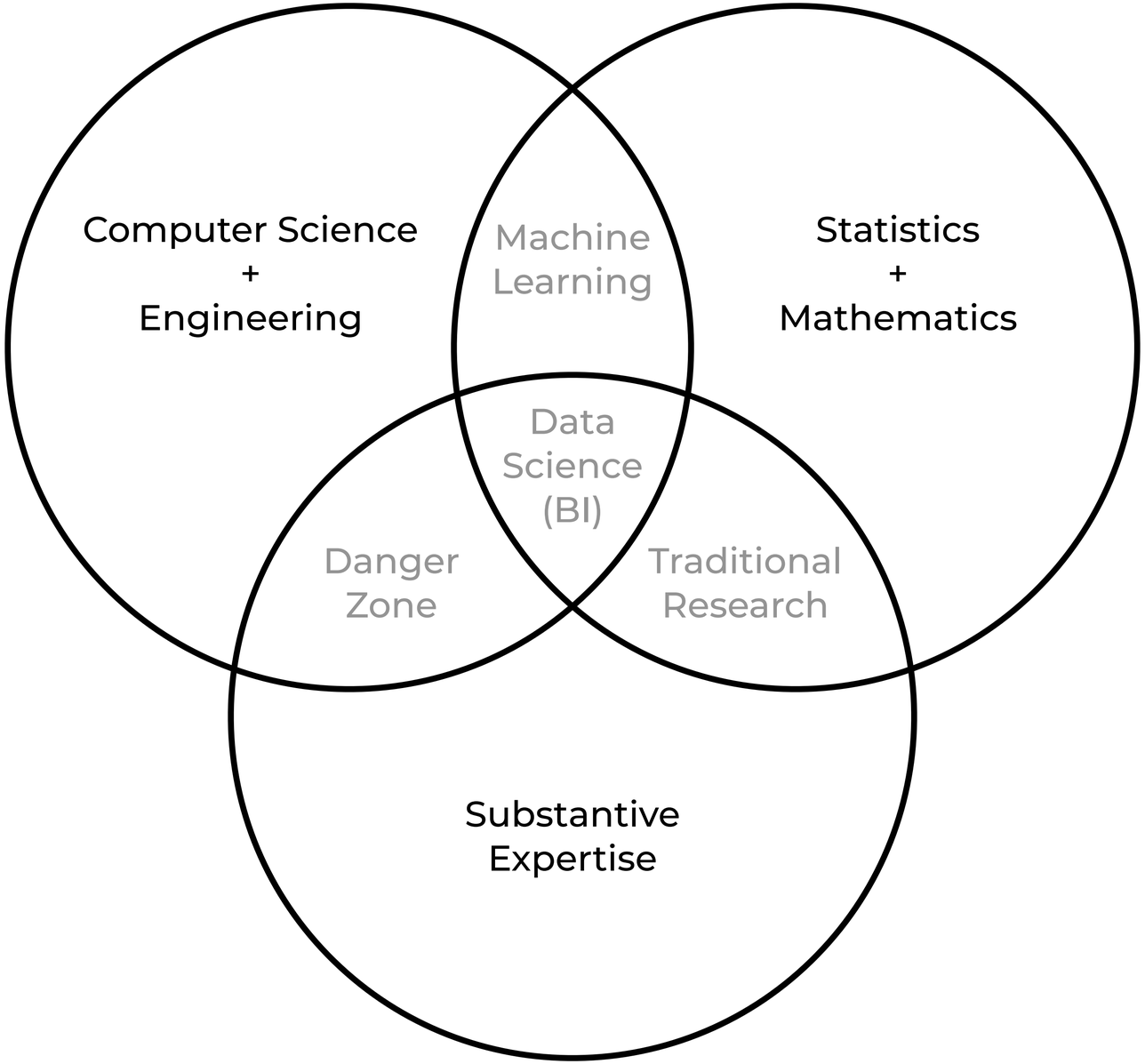
  Understand data through predictive models

  Understand the world through predictive models

How does ML relate to other fields?

Historical View

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



[Inspired by Drew Conway]

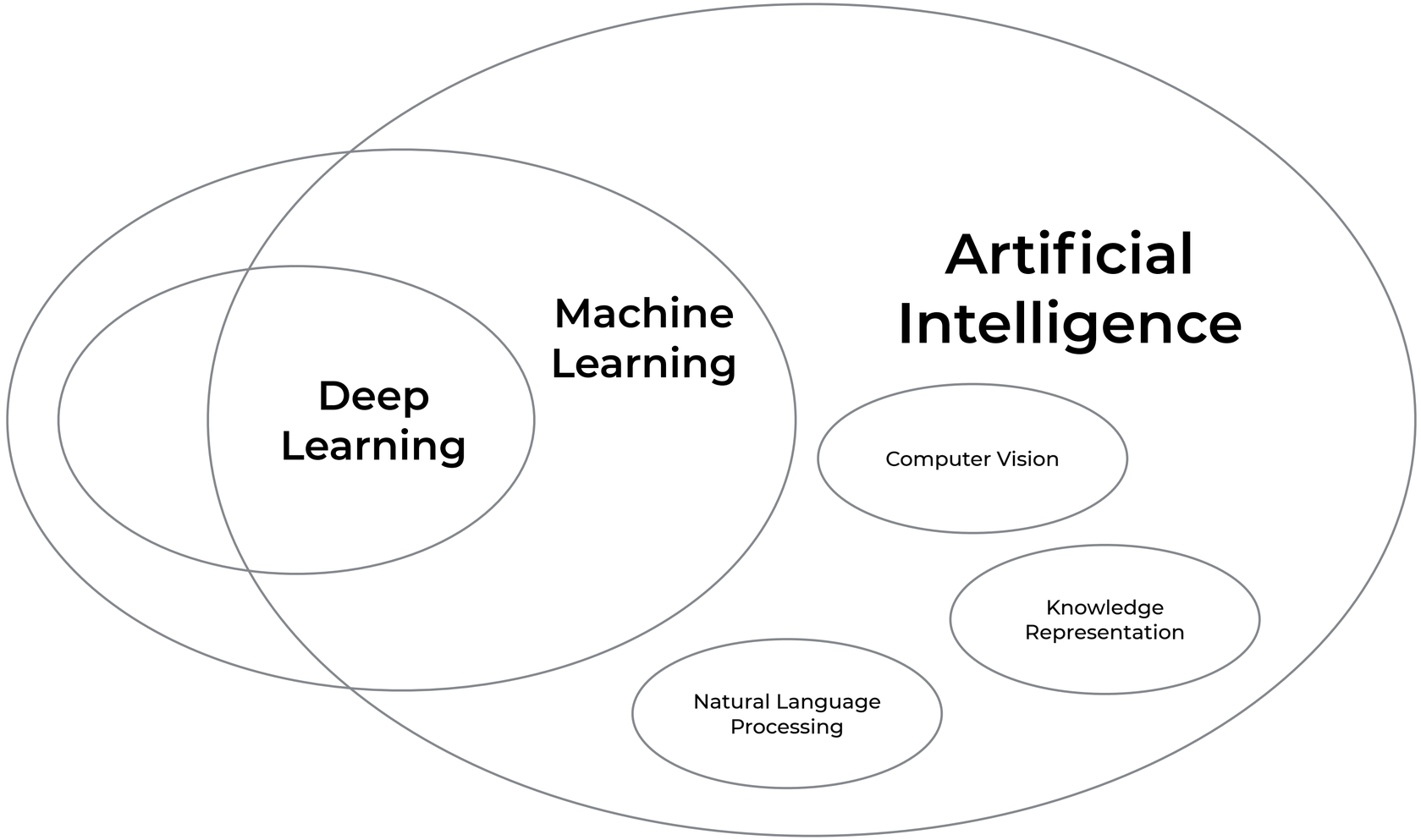
Attitudes in Machine Learning and Data Mining Versus Attitudes in Traditional Statistics

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

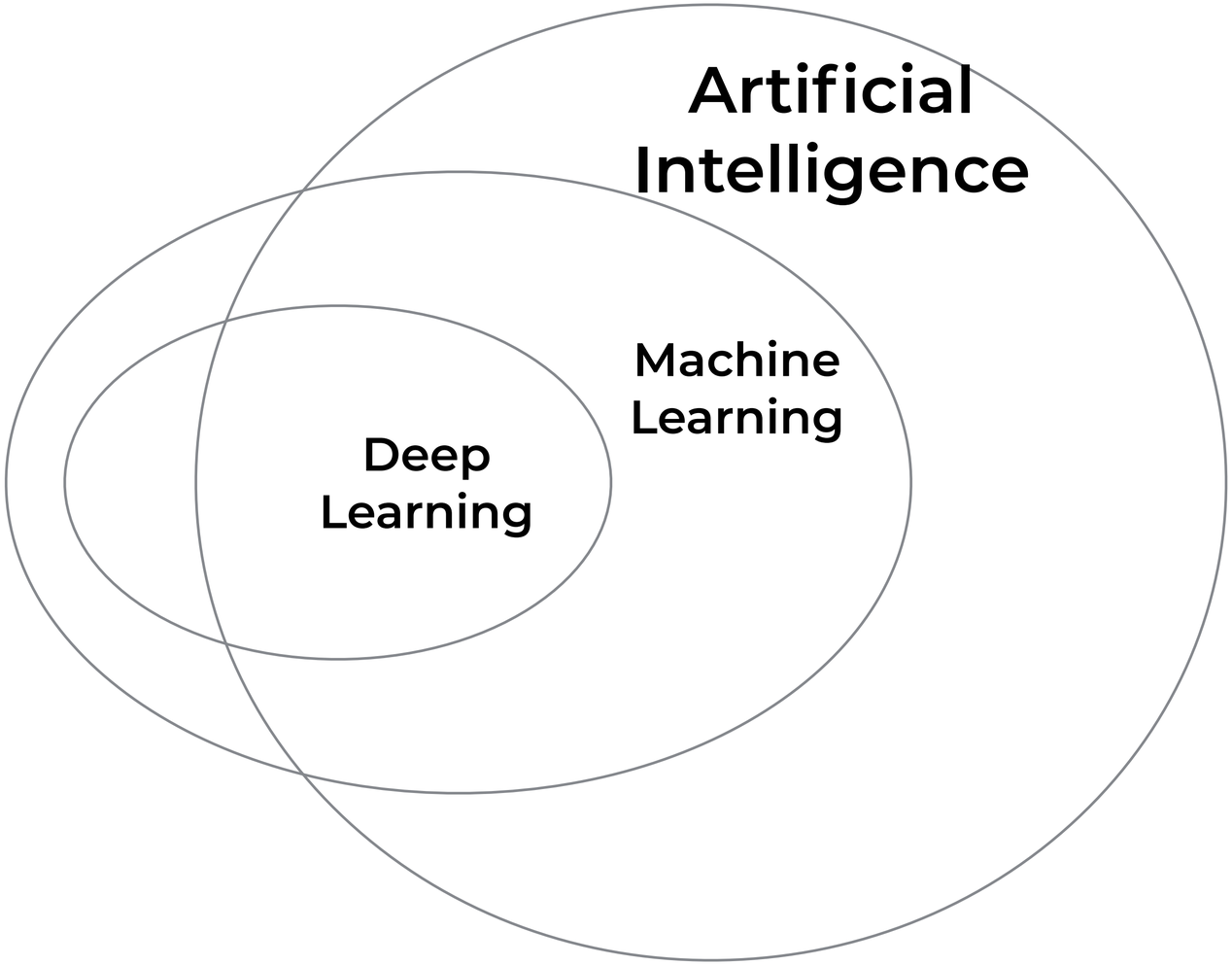
Machine learning	**Traditional statistics**
No settled philosophy or widely accepted theoretical framework.	Classical (frequentist) and Bayesian philosophies compete.
Willing to use *ad hoc* methods if they seem to work well (though appearances may be misleading).	Reluctant to use methods without some theoretical justification (even if the justification is actually meaningless).
Emphasis on automatic methods with little or no human intervention.	Emphasis on use of human judgement assisted by plots and diagnostics.
Methods suitable for many problems.	Models based on scientific knowledge.
Heavy use of computing.	Originally designed for hand-calculation, but computing is now very important.

[Reproduced from R. Neal: http://www.utstat.utoronto.ca/~radford/sta414.S14/week1.pdf]

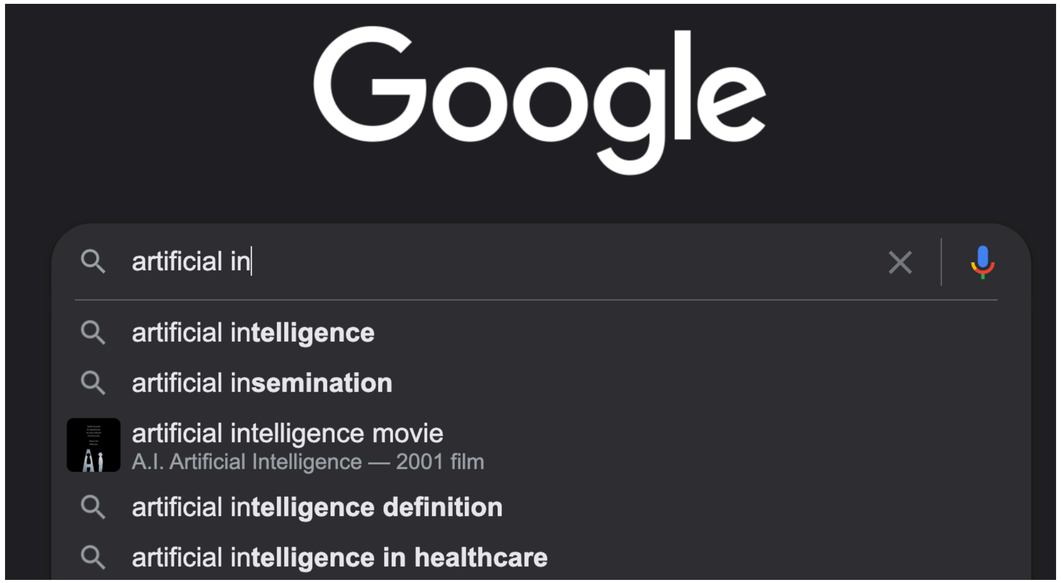
2000

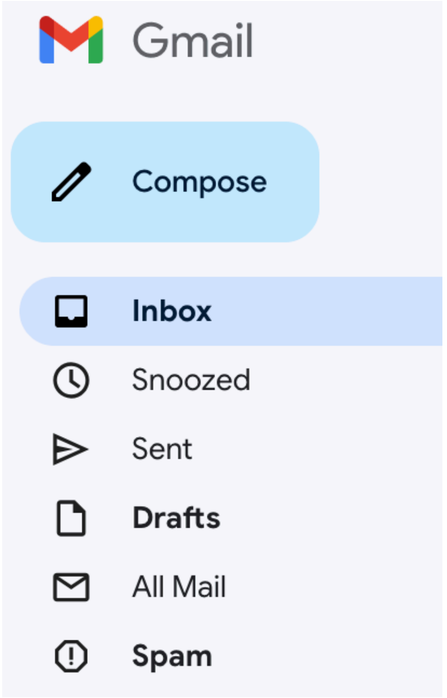

Artificial Intelligence
Machine Learning
Deep Learning
Computer Vision
Knowledge Representation
Natural Language Processing

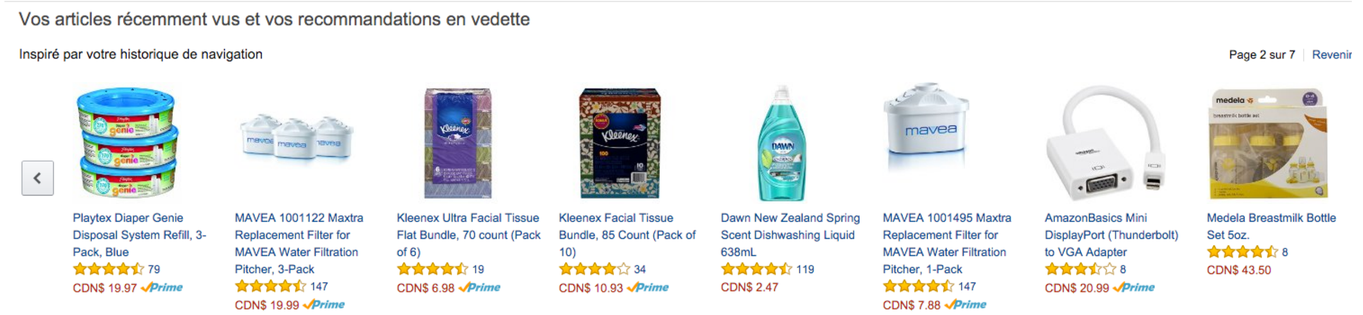
2024



Applications of ML

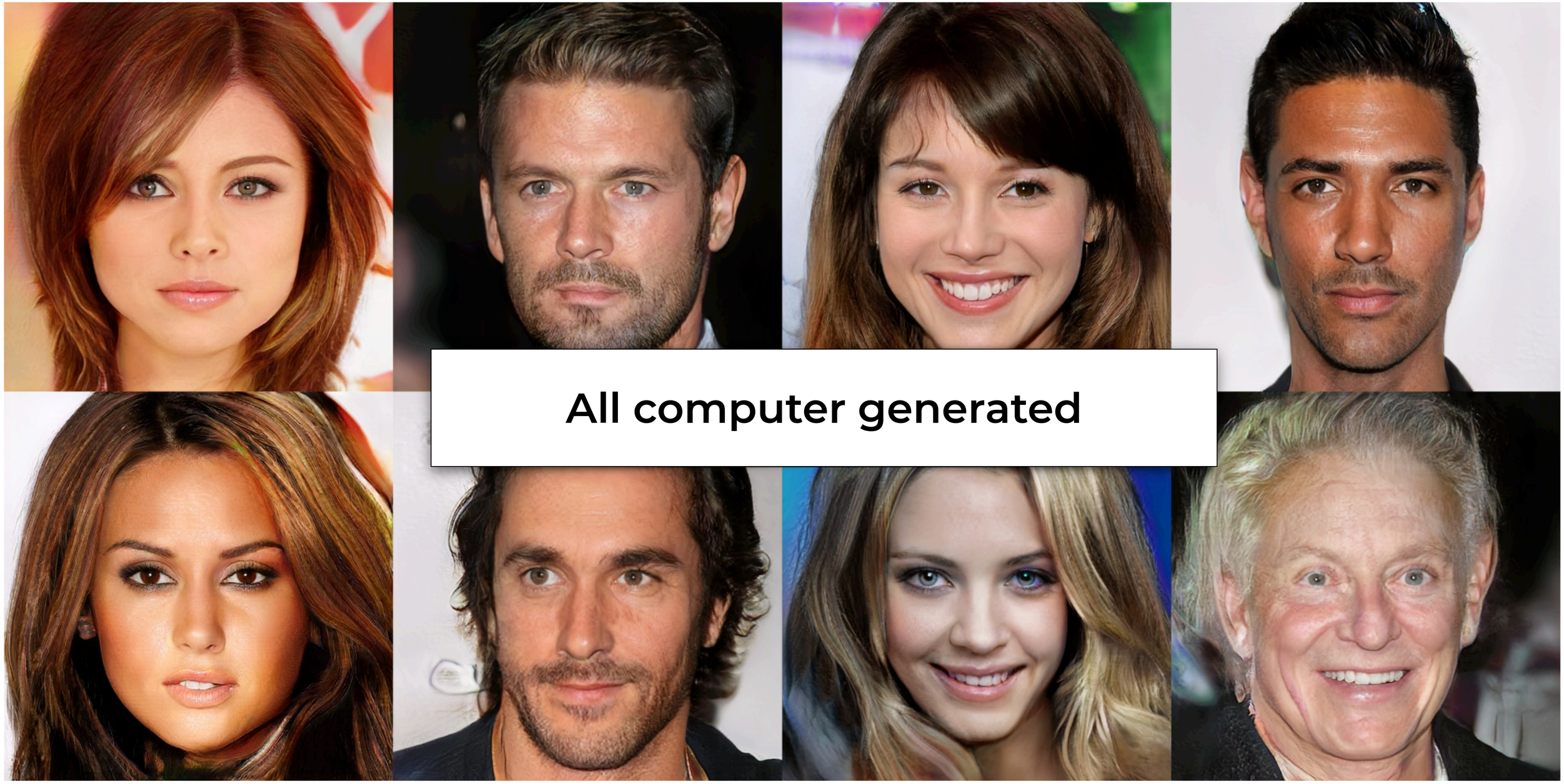

Google
artificial in
artificial intelligence
artificial insemination
artificial intelligence movie
A.I. Artificial Intelligence — 2001 film
artificial intelligence definition
artificial intelligence in healthcare


Gmail
Compose
Inbox
Snoozed
Sent
Drafts
All Mail
Spam


Vos articles récemment vus et vos recommandations en vedette
Inspiré par votre historique de navigation
Page 2 sur 7
Revenir
Playtex Diaper Genie Disposal System Refill, 3-Pack, Blue
79
CDN$ 19.97
MAVEA 1001122 Maxtra Replacement Filter for MAVEA Water Filtration Pitcher, 3-Pack
147
CDN$ 19.99
Kleenex Ultra Facial Tissue Flat Bundle, 70 count (Pack of 6)
19
CDN$ 6.98
Kleenex Facial Tissue Bundle, 85 Count (Pack of 10)
34
CDN$ 10.93
Dawn New Zealand Spring Scent Dishwashing Liquid 638mL
119
CDN$ 2.47
MAVEA 1001495 Maxtra Replacement Filter for MAVEA Water Filtration Pitcher, 1-Pack
147
CDN$ 7.88
AmazonBasics Mini DisplayPort (Thunderbolt) to VGA Adapter
8
CDN$ 20.99
Medela Breastmilk Bottle Set 5oz.
8
CDN$ 43.50

NVIDIA

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



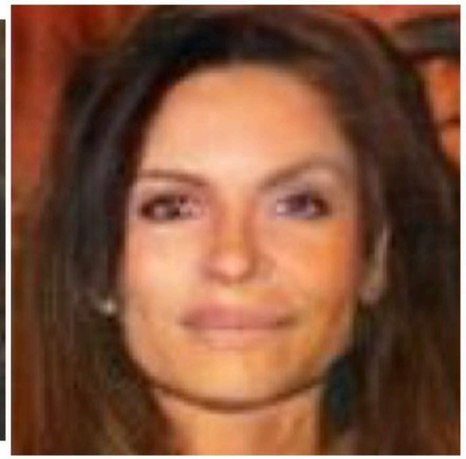
All computer generated

NVIDIA

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

2014

2015



2016

2017

2018

https://twitter.com/goodfellow_ian/status/1084973596236144640/photo/1

https://twitter.com/StelfieTT

https://twitter.com/StelfieTT

https://twitter.com/StelfieTT

Laurent Charlin — 60629

Laurent Charlin — 60629

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

AI-GENERATED IMAGES

https://openai.com/blog/dall-e/

- 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...

Risks

- Powerful technology that continues to improve
- Dual-use, like most technologies
- Biases, high-energy costs, truth/fiction, difficult to evaluate

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.**

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

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

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

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

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

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}

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

a_{12}

Simple operations

- **Transpose**

$$\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}$$

$$\mathbf{a}^\top = \begin{bmatrix} a_0 & a_1 & a_2 \end{bmatrix}$$

$$(\mathbf{A}_{ij})^\top = \mathbf{A}_{ji}$$

- **Addition**
  - **Vectors and matrices w. the same shape**

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} \qquad \mathbf{a} + \mathbf{b} = \begin{bmatrix} a_0 + b_0 \\ a_1 + b_1 \\ a_2 + b_2 \end{bmatrix}$$

$$(\mathbf{A} + \mathbf{B})_{ij} = \mathbf{A}_{ij} + \mathbf{B}_{ij}$$

Simple operations

- Multiply by a scalar $\alpha \mathbf{a} = \begin{bmatrix} \alpha\, a_0 \\ \alpha\, a_1 \\ \alpha\, a_2 \end{bmatrix}$
- Vector product.
  - The dot product $\mathbf{a}^\top \mathbf{a} = \sum_i a_i a_i$
    - Note: it yields a scalar.
  - Element-wise product: $\mathbf{a} \odot \mathbf{a} = \begin{bmatrix} a_0 a_0 \\ a_1 a_1 \\ a_2 a_2 \end{bmatrix}$
    - 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)**

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

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

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

A square matrix not invertible is *singular*

- Inverses are useful to solve systems of equations:

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

Norms

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

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

- Standard norms in ML:
  - Euclidean norm (p=2) $$\| a \|_2 = \sqrt{\left(\sum_i |a_i|^2 \right)}$$
  - Dot product w. 2-norm: $a^\top b = ||a||_2 ||b||_2 \cos \theta_{ab}$
  - Frobenius norm (matrix): $$\| A \|_2 = \sqrt{\left(\sum_i \sum_j |a_{ij}|^2 \right)}$$

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.

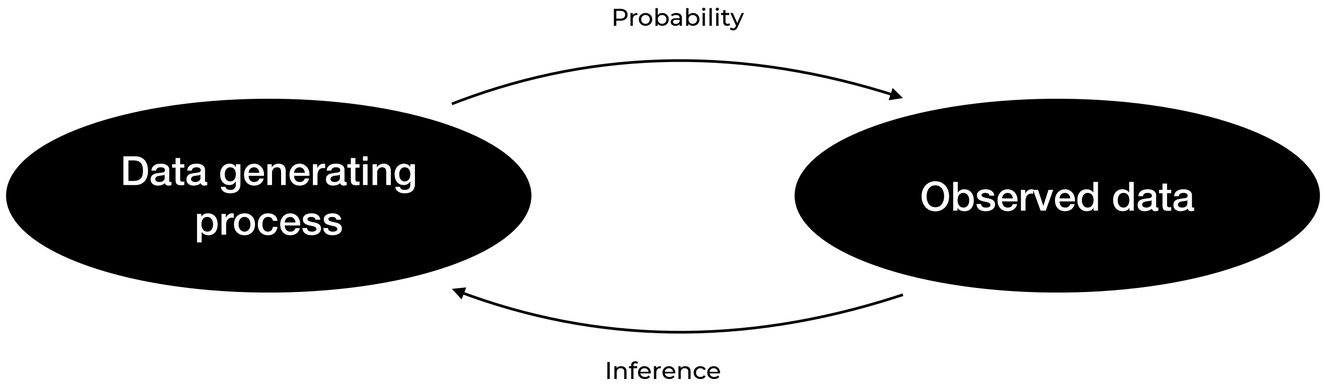
Why probabilities?

- 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: $P(x) \geq 0, \forall x \in X$

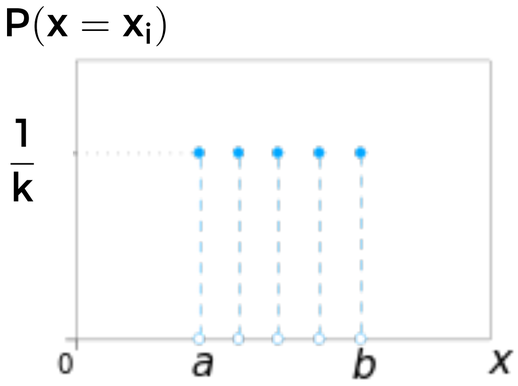
$$P(\bigcup x) = 1$$

Discrete RVs

- An RV is discrete if it takes a finite number of values[1]

$$P(x = x_i) \geq 0, \forall i$$

$$\sum_i P(x = x_i) = 1$$

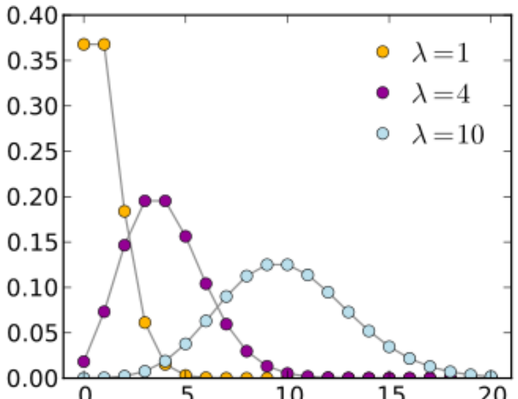


- E.g., uniform distribution:

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

- E.g., Poisson distribution:

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



both images are from: wikipedia.org

1. technically: it must be countable

Continuous RVs

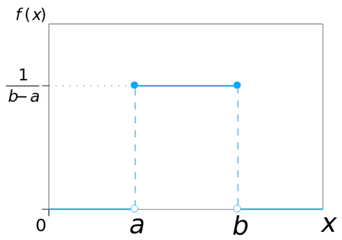
- An RV is continuous if $f(x) \geq 0, \forall x \in X$

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

$$P(a < x < b) = \int_{a}^{b} f(x)dx$$

- f(x) is a probability density function (PDF)
  - E.g., (continuous) uniform distribution:

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



from: wikipedia.org

  - E.g., Gaussian distribution

A few useful properties
(shown for discrete variables for simplicity)

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

Moments

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

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