

AI Slides Viewer & Checker - 2

Sunghee Yun

Table of contents

● AI Study & Research	
– ML Prerequisites	ml-prerequisite-2025-0707.tex - 2
– ML Basics	ml-basics-2025-0706.tex - 49
– Learning AI	ai-study-2025-0707.tex - 90
● AI industry sectors	
– AI & Biotech	biotech-2025-0603.tex - 94
– AI-powered Humanoid Robots	ai-humanoid-robots-2024-0912.tex - 114
– Industrial AI	inai-2024-0811.tex - 124
● Silicon Valley's Cultural Engine	silicon-valley-2025-0522.tex - 163
● Empowering Humanity for Future Enriched by AI	ai-humanity-2025-0119.tex - 171
● Some Important Questions	ai-important-qs-2025-0610.tex - 183
● Recent AI Development	ai-newdevs-2024-1019.tex - 219
● Selected references & sources	selrefs-2025-0627.tex - 236
● References	- 238

ML Prerequisites

Linear Algebra Basics

Scalars, vectors, and matrices

- real number $a \in \mathbf{R}$, called *scalar*
- (ordered) collection of real numbers $(a_1, \dots, a_n) \in \mathbf{R}^n$, called *vector*

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \in \mathbf{R}^n \quad \text{- column vector}$$

$$\begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \in \mathbf{R}^{1 \times n} \quad \text{- row vector}$$

- (ordered) collection of 2-dimensional array, called *matrix*

$$\begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1} & A_{m,2} & \cdots & A_{m,n} \end{bmatrix} \in \mathbf{R}^{m \times n}$$

Transposes

- transpose of row vector is column vector & vice versa

$$\begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}^T = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \quad \& \quad \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}^T = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}$$

- transpose of m -by- n matrix is n -by- m matrix

$$\begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1} & A_{m,2} & \cdots & A_{m,n} \end{bmatrix}^T = \begin{bmatrix} A_{1,1} & A_{2,1} & \cdots & A_{m,1} \\ A_{1,2} & A_{2,2} & \cdots & A_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1,n} & A_{2,n} & \cdots & A_{m,n} \end{bmatrix} \in \mathbb{R}^{n \times m}$$

Matrix-vector multiplication

- for matrix $A \in \mathbf{R}^{m \times n}$ & vector $b \in \mathbf{R}^n$
 - matrix-vector multiplication Ab defined by

$$Ab = \begin{bmatrix} A_{1,1}b_1 + A_{1,2}b_2 + \cdots + A_{1,n}b_n \\ A_{2,1}b_1 + A_{2,2}b_2 + \cdots + A_{2,n}b_n \\ \vdots \\ A_{m,1}b_1 + A_{m,2}b_2 + \cdots + A_{m,n}b_n \end{bmatrix} \in \mathbf{R}^m$$

in other words

$$(Ab)_i = \sum_{j=1}^n A_{i,j}b_j \quad \text{for } 1 \leq i \leq m$$

- resulting quantity is vector of length m
- number of columns of A *must* equal to length of b

Matrix-matrix multiplication

- for matrices $A \in \mathbf{R}^{m \times n}$ & $B \in \mathbf{R}^{n \times p}$

- matrix-matrix multiplication $AB \in \mathbf{R}^{m \times p}$ defined by

$$(AB)_{i,j} = \sum_{k=1}^n A_{i,k}B_{k,j} \quad \text{for } 1 \leq i \leq m$$

- resulting quantity is m -by- p matrix
 - *order matters* and number of columns of A *must* equal to number of rows of B
- note matrix-vector multiplication is *special case* of matrix-matrix multiplication

Calculus Basics

Functions

- $f : X \rightarrow Y$
 - $X = \text{dom } f$ - domain of f
 - Y - codomain of f
 - $\mathcal{R}(f) = \{f(x) \in Y \mid x \in X\}$ - range of f

Differentiation & derivatives

- for real-valued function $f : \mathbf{R} \rightarrow \mathbf{R}$

- derivative of f at $x \in \mathbf{R}$

$$f'(x) = \frac{d}{dx} f(x) = \lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h} \in \mathbf{R}$$

- derivative exists *if and only if* limit exists
 - second derivative of f at $x \in \mathbf{R}$

$$f''(x) = \frac{d^2}{dx^2} f(x) = \lim_{h \rightarrow 0} \frac{f'(x + h) - f'(x)}{h} \in \mathbf{R}$$

- second derivative exists *if and only if* limit exists

Multivariate functions

- $f : \mathbf{R}^n \rightarrow \mathbf{R}$ - real-valued multivariate function

$$f(x) = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}\right) = f(x_1, x_2, \dots, x_n) \in \mathbf{R}$$

- examples
 - $f : \mathbf{R}^3 \rightarrow \mathbf{R}$ - linear function

$$f(x) = x_1 + 3x_2 + 2x_3$$

- $f : \mathbf{R}^3 \rightarrow \mathbf{R}$ - convex quadratic function

$$f(x) = x_1^2 + x_1x_2 + 3x_2^2 + 5x_3^2$$

Multivariate vector functions

- $f : \mathbf{R}^n \rightarrow \mathbf{R}^m$ - real-valued multivariate vector function

$$f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{bmatrix} \in \mathbf{R}^m$$

where $f_j : \mathbf{R}^n \rightarrow \mathbf{R}$ for $1 \leq j \leq m$

- examples
 - $f : \mathbf{R}^3 \rightarrow \mathbf{R}^2$ - linear function

$$f(x) = \begin{bmatrix} x_1 + 3x_2 + 2x_3 \\ -3x_2 + x_3 \end{bmatrix} \in \mathbf{R}^2$$

- $f : \mathbf{R}^3 \rightarrow \mathbf{R}^3$ - componentwise function

$$f(x) = \begin{bmatrix} \exp(x_1) & \exp(x_2) & \exp(x_3) \end{bmatrix}^T \in \mathbf{R}^3$$

Partial derivative & gradient

for $f : \mathbf{R}^n \rightarrow \mathbf{R}$

- i th partial derivative

$$\frac{\partial}{\partial x_i} f(x) = \frac{f(x + he_i) - f(x)}{h} = \frac{f(\dots, x_{i-1}, x_i + h, x_{i+1}, \dots) - f(x)}{h}$$

where $e_i \in \mathbf{R}^n$ is i th unit vector

- gradient is vector of partial derivatives

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} \in \mathbf{R}^n$$

- we have

$$(\nabla f(x))_i = \frac{\partial}{\partial x_i} f(x) = e_i^T \nabla f(x) \in \mathbf{R}$$

Jacobian

for $f : \mathbf{R}^n \rightarrow \mathbf{R}^m$

- Jacobian matrix

$$Df(x) = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_1(x)}{\partial x_2} & \dots & \frac{\partial f_1(x)}{\partial x_n} \\ \frac{\partial f_2(x)}{\partial x_1} & \frac{\partial f_2(x)}{\partial x_2} & \dots & \frac{\partial f_2(x)}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \frac{\partial f_m(x)}{\partial x_2} & \dots & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix} \in \mathbf{R}^{m \times n}$$

- equivalently

$$Df(x) = \begin{bmatrix} \nabla f_1(x)^T \\ \nabla f_2(x)^T \\ \vdots \\ \nabla f_m(x)^T \end{bmatrix} \in \mathbf{R}^{m \times n}$$

Chain rule

- for $f : \mathbf{R} \rightarrow \mathbf{R}^m$, $g : \mathbf{R}^m \rightarrow \mathbf{R}$ & $h = g \circ f$, i.e., $h(x) = g(f_1(x), \dots, f_m(x))$, derivative of h at $x \in \mathbf{R}$

$$h'(x) = \sum_{j=1}^m \frac{\partial}{\partial y_j} g(f(x)) f'_j(x) = \sum_{j=1}^m \nabla g(f(x))_j f'_j(x) \in \mathbf{R}$$

- for $f : \mathbf{R}^n \rightarrow \mathbf{R}^m$, $g : \mathbf{R}^m \rightarrow \mathbf{R}^p$ & $h = g \circ f$, Jacobian of h at $x \in \mathbf{R}^n$

$$Dh(x) = Dg(f(x)) Df(x) \in \mathbf{R}^{p \times n}$$

- note $Dg(f(x)) \in \mathbf{R}^{p \times m}$ & $Df(x) \in \mathbf{R}^{m \times n}$

- first is *special case* of second

Statistics Basics

Random experiments & probability law

- *random experiment*
 - outcome varies in unpredictable fashion (even) when experiment is being repeated under same conditions
 - specified by stating experimental procedure and set of one or more measurements or observations
- probability law
 - rule assigning probabilities to events of experiment that belong to event class \mathcal{F}

$$p : \mathcal{F} \rightarrow \mathbf{R}_+$$

- properties (or axioms)
 - for event $A \in \mathcal{F}$, $p(A)$ called *probability* of A
 - for event $A, B \in \mathcal{F}$ with $A \cap B = \emptyset$

$$p(A \cup B) = p(A) + p(B)$$

Conditional probability

- probability of event A given that event B has occurred, called *conditional probability*, denoted by

$$p(A|B)$$

- formula

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

- thus

$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A)$$

- Bayes' theorem

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

Independence

- for events A & B , when knowledge of occurrence of B does not alter probability of A
 - A said to be *independent* of B
- following statements are equivalent
 - A is independent of B
 - B is independent of A
 - $p(A|B) = p(A)$
 - $p(B|A) = p(B)$
 - $p(A \cap B) = p(A)p(B)$

Random variables

- *discrete* random variable X assumes values from countable set $\{x_1, x_2, \dots\}$
- *continuous* random variable X assumes values from \mathbf{R}
- random *vector* X assumes values from \mathbf{R}^n

PMF, PDF & CDF

- *probability mass function (PMF)* of discrete X

$$p_X(x) = p(X = x)$$

- *probability density function (PDF)* of continuous X

$$\int_a^b p_X(x) = p(a \leq X \leq b)$$

- *cumulative distribution function (CDF)* of (any) X

$$F_X(x) = p(X \leq x)$$

- for discrete X - $F_X(x) = \sum_{x' \leq x} p_X(x')$
- for continuous X - $F_X(x) = \int_{-\infty}^x p_X(x') dx'$

Expected value, variance & covariance matrix

- expected value

- for discrete X

$$\mathbf{E} X = \sum_x x p_X(x)$$

- for continuous X

$$\mathbf{E} X = \int_{-\infty}^{\infty} x p_X(x) dx$$

- variance for scalar $X \in \mathbf{R}$

$$\mathbf{Var}(X) = \mathbf{E}(X - \mathbf{E} X)^2 = \mathbf{E} X^2 - (\mathbf{E} X)^2$$

- covariance matrix for vector $X \in \mathbf{R}^n$

$$\mathbf{Var}(X) = \mathbf{E}(X - \mathbf{E} X)(X - \mathbf{E} X)^T = \mathbf{E} XX^T - (\mathbf{E} X)(\mathbf{E} X)^T$$

Joint PMF, PDF & CDF

- *joint PMF* of discrete X & Y

$$p_{X,Y}(x, y) = p(X = x, Y = y)$$

- *join PDF* of continuous X & Y

$$\int_c^d \int_a^b p_{X,Y}(x, y) dx dy = p(a \leq X \leq b \text{ & } c \leq Y \leq d)$$

- *joint CDF* of X & Y

$$F_{X,Y}(x, y) = p(X \leq x \text{ & } Y \leq y)$$

Conditional expectation

for two random variables X & Y

- expected value of Y conditioned on X

$$\mathbf{E}(Y|X = x) = \int y p(y|x) dy$$

where

$$p(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

- note

$$\mathbf{E}_{X,Y} f(X, Y) = \mathbf{E}_X \mathbf{E}_Y (f(X, Y)|X)$$

because

$$\int \int f(x, y) p(x, y) dx dy = \int \left(\int f(x, y) p(y|x) dy \right) p(x) dx$$

Discrete random variables

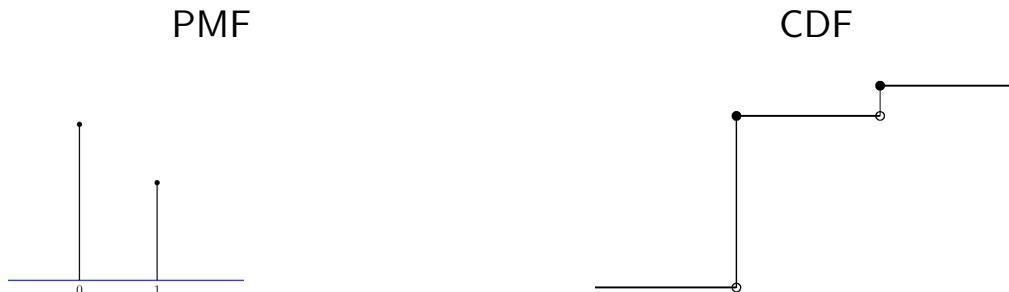
Bernoulli distribution

- model single binary trial with probability p of success (and, hence $(1 - p)$ of failure)
- PMF, mean, variance

$$p(k) = p^k(1-p)^{1-k} = \begin{cases} 1-p & \text{if } k=0 \\ p & \text{if } k=1 \end{cases}$$

$$\mathbf{E}(X) = p \quad \mathbf{Var}(X) = p(1-p)$$

- ML applications - (foundation for)
 - logistic regression, binary classification, modeling click-through rates, A/B testing outcomes



Binomial distribution

- model number of successes in n independent Bernoulli trials with probability p
- PMF, mean, variance

$$p(k) = \binom{n}{k} p^k (n-p)^{1-k} \quad \text{for } 1 \leq k \leq n$$

$$\mathbf{E}(X) = np \quad \mathbf{Var}(X) = np(1-p)$$

- ML applications
 - modeling conversion rates, quality control testing, ensemble voting methods, batch processing success rates



Multinomial distribution

- generalizes binomial distribution to multiple categories with probabilities p_1, \dots, p_k
- PMF, mean, variance

$$p(k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

$$\mathbf{E}(X_i) = np_i \quad \mathbf{Var}(X_i) = np_i(1 - p_i) \quad \mathbf{Cov}(X_i, X_j) = -np_ip_j$$

- ML applications
 - multi-class classification, topic modeling, document classification, market basket analysis
- widely used in *Bayesian inference* with Dirichlet priors



Geometric distribution

- model number of trials needed to achieve first success in independent Bernoulli trials
- PMF, mean, variance

$$p(k) = p(1 - p)^{k-1} \quad \mathbf{E}(X) = 1/p \quad \mathbf{Var}(X) = (1 - p)/p^2$$

- ML applications
 - modeling time-to-conversion, failure analysis, reinforcement learning episode lengths, web crawling stopping conditions
- memoryless property $p(X > m + n | X > m) = p(X > n)$

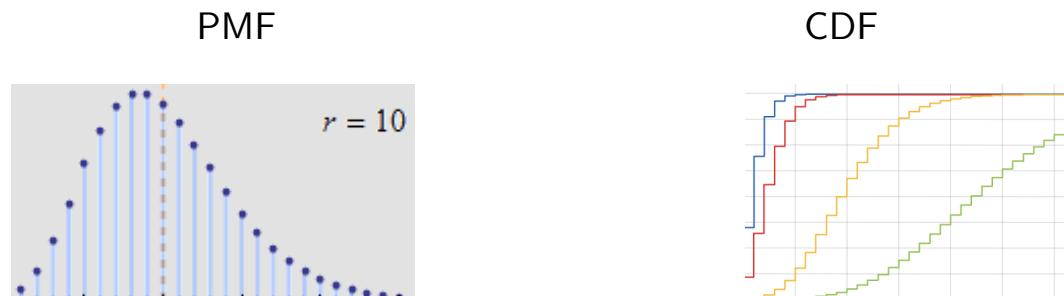


Negative binomial distribution

- model number of trials needed to achieve r successes in independent Bernoulli trials
- PMF, mean, variance

$$p(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r} \quad \mathbf{E}(X) = r/p \quad \mathbf{Var}(X) = r(1-p)/p^2$$

- ML applications
 - modeling overdispersed count data, customer acquisition costs, reliability engineering, text analysis for word frequencies
- *often used when Poisson assumptions are violated due to overdispersion*



Poisson distribution

- model number of events occurring in fixed interval of time or space
- PMF, mean, variance ($\lambda > 0$)

$$p(k) = e^{-\lambda} \lambda^k / k! \quad \mathbf{E}(X) = \lambda \quad \mathbf{Var}(X) = \lambda$$

- ML applications
 - modeling web traffic, system failures
 - word counts (in NLP), user interactions (in recommendation systems)
- approximates binomial when n is large & p is small with $= np$
- sum of independent Poisson variables is Poisson



Hypergeometric distribution

- model number of successes in n draws without replacement from finite population of size N containing K successes
- PMF, mean, variance ($N, K \in \mathbf{N}$ with $N > K$)

$$p(k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}} \quad \mathbf{E}(X) = \frac{nK}{N} \quad \mathbf{Var}(X) = \frac{nK}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1}$$

- ML applications
 - sampling without replacement, quality control testing
 - feature selection validation, A/B testing with finite populations



Continuous random variables

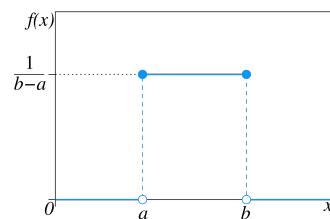
Uniform distribution

- model equally likely outcomes over continuous interval $[a, b]$ representing complete uncertainty within bounded range
- PDF, mean, variance ($a, b \in \mathbf{R}$ with $b > a$)

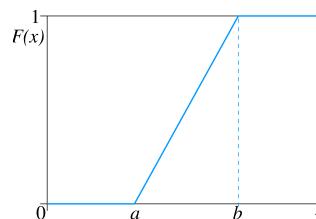
$$p(x) = 1/(b - a)I_{[a,b]}(x) \quad \mathbf{E}(X) = (a + b)/2 \quad \mathbf{Var}(X) = (b - a)^2/12$$

- ML applications
 - Monte Carlo sampling, generating baseline distributions for hypothesis testing
- maximum entropy distribution for bounded continuous support
- foundation for pseudo-random number generation and inverse transform sampling

PDF



CDF



Gaussian distribution

- most important continuous distribution
- model symmetric bell-shaped data arising from many natural processes
- PDF, mean, variance ($\mu \in \mathbf{R}$, $\sigma > 0$)

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \quad \mathbf{E}(X) = \mu \quad \mathbf{Var}(X) = \sigma^2$$

- ML applications
 - linear regression error terms, NN weight initialization, PCA, noise modeling
- invariant under linear transformations, maximum entropy for given mean and variance

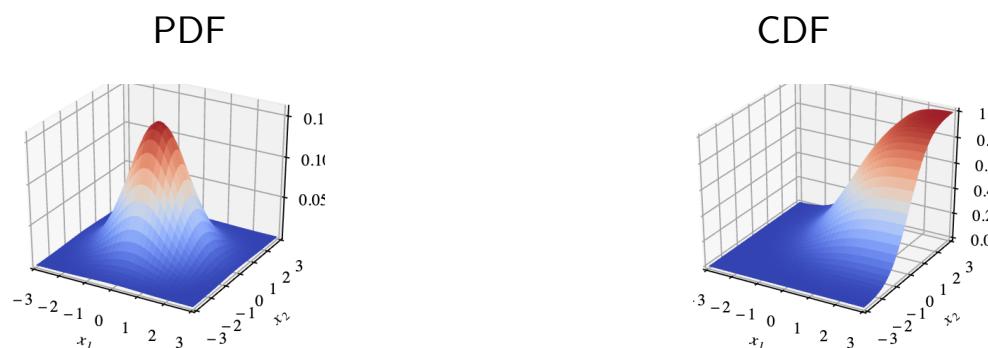


Multivariate Gaussian distribution

- generalize scalar Gaussian to random vector
- PDF, mean, variance ($\mu \in \mathbf{R}^n$, $\Sigma \in \mathbf{S}_{++}^n$)

$$p(x) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \quad \mathbf{E}(X) = \mu \quad \mathbf{Cov}(X) = \Sigma$$

- ML applications
 - Gaussian mixture, PCA, Kalman filtering, Gaussian processes, latent variable models
- maximum likelihood estimation having closed-form solution, foundation for many Bayesian models



Exponential distribution

- model time between events in Poisson process, representing memoryless waiting times or lifetimes
- PDF, mean, variance ($\lambda > 0$)

$$p(x) = \lambda e^{-\lambda x} I_{[0,\infty)}(x) \quad \mathbf{E}(X) = 1/\lambda \quad \mathbf{Var}(X) = 1/\lambda^2$$

- ML applications
 - system failure times, web session durations, survival analysis
- memoryless property $p(X > s+t | X > s) = p(X > t)$ - only continuous distribution with this property, minimum of exponentials is exponential



Gamma distribution

- model positive continuous values - generalizing exponential distribution to allow for more flexible shapes, *e.g.*, for waiting times for multiple events
- PDF, mean, variance ($\alpha, \beta > 0$)

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} I_{[0,\infty)}(x) \quad \mathbf{E}(X) = \alpha/\beta \quad \mathbf{Var}(X) = \alpha/\beta^2$$

- ML applications
 - survival analysis, queuing theory
- exponential is special case when $\alpha = 1$, sum of independent exponentials is gamma, conjugate prior for Poisson and exponential distributions



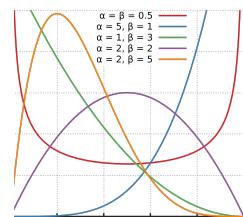
Beta distribution

- model probabilities and proportions, defined on $[0, 1]$ with flexible shapes from uniform to highly skewed
- PDF, mean, variance ($\alpha, \beta > 0$)

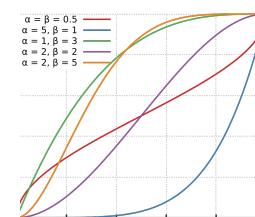
$$p(x) = \frac{\Gamma(\alpha, \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad \mathbf{E}(X) = \frac{\alpha}{\alpha + \beta} \quad \mathbf{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

- ML applications
 - modeling success rates, A/B testing, probability calibration
- uniform is special case when $\alpha = \beta = 1$, conjugate prior for Bernoulli & binomial related to Dirichlet distribution

PDF



CDF



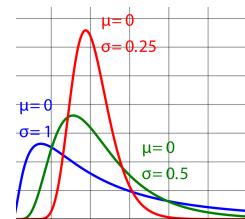
Log-normal distribution

- model positive values where logarithm follows normal distribution, representing multiplicative processes and heavy-tailed phenomena
- PDF, mean, variance ($\mu \in \mathbf{R}$, $\sigma > 0$)

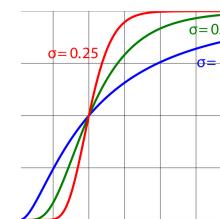
$$p(x) = e^{-(\log x - \mu)^2 / 2\sigma^2} / x\sigma\sqrt{2\pi} \quad \mathbf{E}(X) = e^{\mu + \sigma^2 / 2} \quad \mathbf{Var}(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$$

- ML applications
 - modeling income distributions, stock prices, file sizes, network traffic, biological measurements, computational complexity
- heavy right tail, multiplicative central limit theorem

PDF



CDF



Chi-square distribution

- model sum of squares of independent standard normal random variables, fundamental in statistical testing and confidence intervals
- PDF, mean, variance ($\nu \in \mathbf{N}$ - degree of freedom)

$$p(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}x^{\nu/2-1}e^{-x/2}I_{[0,\infty)}(x) \quad \mathbf{E}(X) = \nu \quad \mathbf{Var}(X) = 2\nu$$

- ML applications
 - goodness-of-fit testing, feature selection, confidence intervals for variance, regularization in NN
- special case of gamma distribution, sum of independent chi-squares is chi-square



Student's t -distribution

- model sum of squares of independent standard normal random variables, fundamental in statistical testing and confidence intervals
- PDF, mean, variance ($\nu > 0$ degrees of freedom - almost always positive integer)

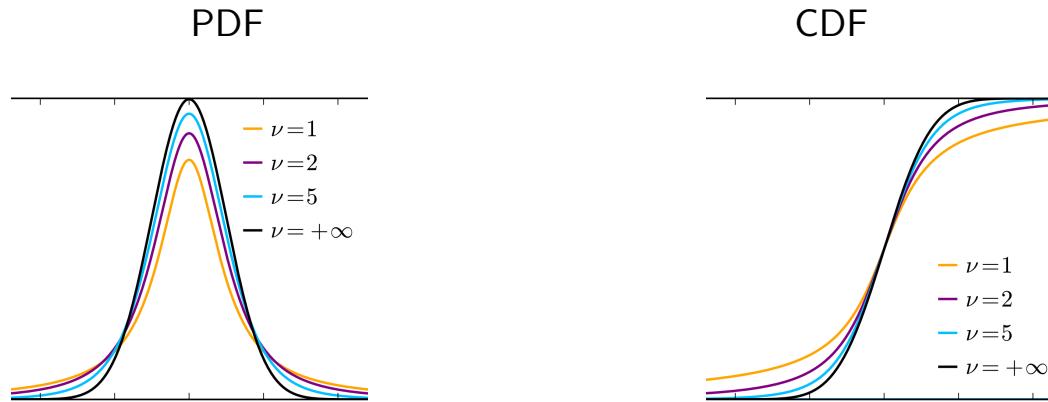
$$p(x) = \frac{\Gamma((\nu + 1)/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)}(1 + x^2/\nu)^{-(\nu+1)/2}$$

$$\mathbf{E}(X) = \begin{cases} 0 & \text{if } \nu > 1 \\ \text{undefined} & \text{otherwise} \end{cases}$$

$$\mathbf{Var}(X) = \begin{cases} \nu/(\nu - 2) & \text{if } \nu > 2 \\ \infty & \text{if } 1 < \nu \leq 2 \\ \text{undefined} & \text{otherwise} \end{cases}$$

- ML applications
 - Bayesian inference, robust regression, confidence intervals with small samples, uncertainty quantification in DL

- heavier tails than normal, approaches standard normal as ν approaches ∞ , symmetric around zero, undefined moments for small ν



Weibull distribution

- model survival times & failure rates with flexible hazard functions, generalizing exponential distribution for reliability analysis
- PDF, mean, variance ($\lambda, k > 0$)

$$p(x) = (k/\lambda)(x/\lambda)^{k-1} e^{-(x/\lambda)^k} I_{[0,\infty)}(x) \quad \mathbf{E}(X) = \lambda\Gamma(1 + 1/k)$$

- ML applications
 - survival analysis, reliability engineering, wind speed modeling, NN activation functions, extreme value theory
- flexible hazard function, minimum of Weibull variables is Weibull



Cauchy distribution

- model heavy-tailed symmetric data with undefined mean and variance, arising in physics and robust statistics
- PDF, mean, variance ($x_0 \in \mathbf{R}$, $\gamma > 0$)

$$p(x) = \frac{1}{\pi\gamma(1 + ((x - x_0)/\gamma)^2)} \quad \mathbf{E}(X) = \text{undefined} \quad \mathbf{Var}(X) = \text{undefined}$$

- ML applications
 - robust statistics, modeling outliers, Bayesian inference with heavy-tailed priors, physics simulations, anomaly detection
- no defined moments, stable distribution, ratio of two independent normals is Cauchy



Laplace distribution

- model symmetric data with heavier tails than normal, representing difference between two independent exponential variables
- PDF, mean, variance ($\mu \in \mathbf{R}$, $b > 0$)

$$p(x) = \frac{1}{2b} \exp(-|x - \mu|/b) \quad \mathbf{E}(X) = \mu \quad \mathbf{Var}(X) = 2b^2$$

- ML applications
 - lasso, robust regression, sparse coding, image processing, privacy-preserving ML
- maximum entropy for given mean absolute deviation, related to L1 penalty, robust to outliers (fundamentally more than normal distribution)



Pareto distribution

- model heavy-tailed phenomena following power-law distributions, representing “80-20 rule” and scale-free networks
- PDF, mean, variance ($x_m, \alpha > 0$)

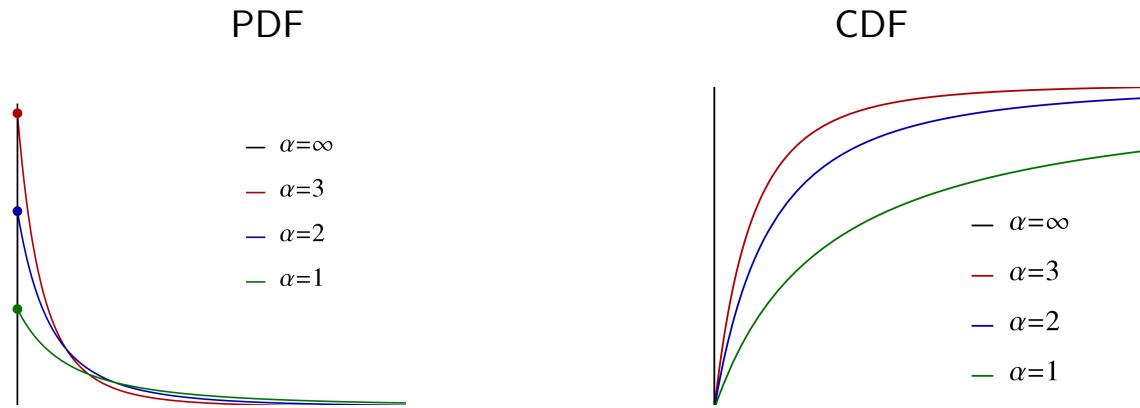
$$p(x) = \alpha x_m^\alpha / x^{\alpha+1}$$

$$\mathbf{E}(X) = \begin{cases} \infty & \text{if } \alpha \leq 1 \\ \alpha x_m / (\alpha - 1) & \text{if } \alpha > 1 \end{cases}$$

$$\mathbf{Var}(X) = \begin{cases} \infty & \text{if } \alpha \leq 2 \\ \alpha x_m^2 / (\alpha - 1)^2 (\alpha - 2) & \text{if } \alpha > 2 \end{cases}$$

- ML applications
 - model wealth distributions, network degree distributions, web page rankings, file sizes, NLP

- heavy right tail, scale-free property, finite moments only for sufficiently large α , basis for power-law distributions



ML Basics

Estimation, Regression, and Inference

The optimal estimator

- estimation problem
 - for two random variables $X \in \mathbf{R}^n$ & $Y \in \mathbf{R}^m$
 - design *estimator or predictor* $g : \mathbf{R}^n \rightarrow \mathbf{R}^m$ to make $g(X)$ *as close as possible* to Y
- when *closeness* measured by mean-square-error (MSE), *the optimal solution* exists

$$g^*(x) = \mathbf{E}(Y|X = x)$$

Proof of optimality

$$\begin{aligned}
 \mathbf{E}_{X,Y}((g(X) - g^*(X))^T(g^*(X) - Y)) &= \mathbf{E}_X \mathbf{E}_Y((g(X) - g^*(X))^T(g^*(X) - Y)|X) \\
 &= \mathbf{E}_X((g(X) - g^*(X))^T \mathbf{E}_Y(g^*(X) - Y)|X) \\
 &= 0
 \end{aligned}$$

hence

$$\begin{aligned}
 \mathbf{E} \|g(X) - Y\|_2^2 &= \mathbf{E} \|g(X) - g^*(X) + g^*(X) - Y\|_2^2 \\
 &= \mathbf{E} \|g(X) - g^*(X)\|_2^2 + \mathbf{E} \|g^*(X) - Y\|_2^2 + 2 \mathbf{E}(g(X) - g^*(X))^T(g^*(X) - Y) \\
 &= \mathbf{E} \|g(X) - g^*(X)\|_2^2 + \mathbf{E} \|g^*(X) - Y\|_2^2 \\
 &\geq \mathbf{E} \|g^*(X) - Y\|_2^2
 \end{aligned}$$

Regression

- in most cases, *not* possible to obtain g^* (unless, *e.g.*, full knowledge of joint PDF)
- regression problem
 - given data set $D = \{(x_1, y_1), \dots, (x_N, y_N)\} \subset \mathbf{R}^n \times \mathbf{R}^m$
 - find $g : \mathbf{R}^n \rightarrow \mathbf{R}^m$ to make $g(X)$ *as close as possible* to Y
- given certain regression method, regressor depends on dataset D

$$g(\cdot; D)$$

Bias & variance

assuming \mathcal{D} is random variable for dataset D

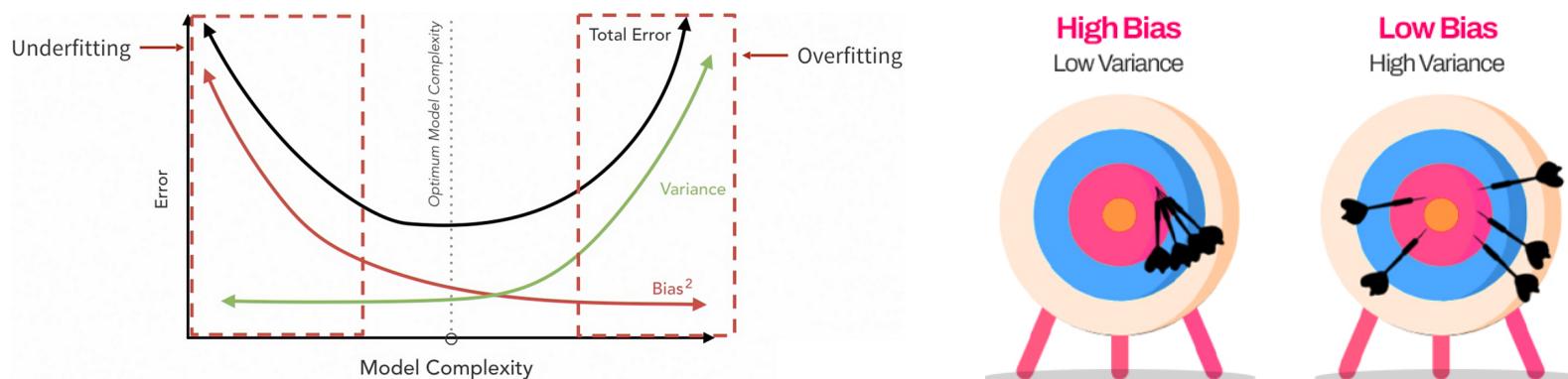
- estimation MSE is

$$\begin{aligned}
 & \mathbf{E}_{X,Y,\mathcal{D}} \|g(X; \mathcal{D}) - Y\|_2^2 \\
 &= \underbrace{\mathbf{E}_{X,\mathcal{D}} \|g(X; \mathcal{D}) - \mathbf{E}_{\mathcal{D}} g(X; \mathcal{D})\|_2^2}_{\text{variance}} + \underbrace{\mathbf{E}_X \|\mathbf{E}_{\mathcal{D}} g(X; \mathcal{D}) - g^*(X)\|_2^2}_{\text{bias}} + \underbrace{\mathbf{E}_{X,Y} \|g^*(X) - Y\|_2^2}_{\text{noise}} \\
 &= \underbrace{\mathbf{E}_{X,\mathcal{D}} \|g(X; \mathcal{D}) - \mathbf{E}_{\mathcal{D}} g(X; \mathcal{D})\|_2^2}_{\text{variance}} + \underbrace{\mathbf{E}_{X,Y} \|\mathbf{E}_{\mathcal{D}} g(X; \mathcal{D}) - Y\|_2^2}_{\text{bias + noise}}
 \end{aligned}$$

- bias & variance
 - *bias* measures how good model is in average
 - *variance* measures how much model varies depending on dataset it is trained on
- *noise* cannot be reduced even with the optimal predictor

Model choice & hyperparameter optimization

- want to choose model or modeling method to make both bias & variance low
 - (too) complex models have low bias, but high variance
 - (too) simple models have low variance, but high bias
- usually solved by *hyperparameter optimization*
 - sometimes called *hyperparameter tuning*



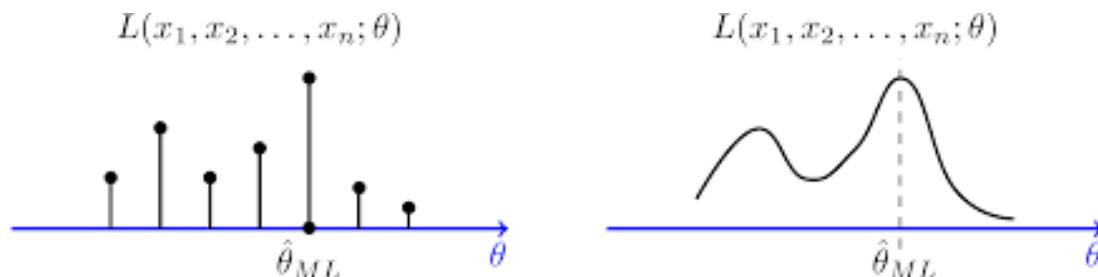
MLE

- maximum likelihood estimation (MLE)
 - assume parameterized distribution of $X \in \mathbb{R}^n$ by $\theta \in \Theta$ - $p(x; \theta)$
 - find θ maximizing *likelihood function*

$$p(x_1, \dots, x_N; \theta) = \prod_{i=1}^N p(x_i; \theta)$$

- MLE solution

$$\hat{\theta}_{\text{MLE}} = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^N p(x_i; \theta)$$



MAP estimation

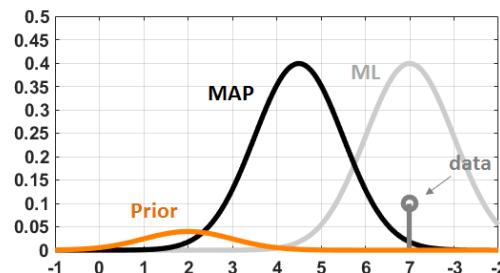
- maximum a posteriori (MAP) estimation
 - assume *prior knowledge* of θ - $p(\theta)$
 - assume parameterized distribution of $X \in \mathbb{R}^n$ by θ - $p(x|\theta)$
 - find θ maximizing *posteriori probability*

$$p(\theta|x_1, \dots, x_N)$$

– Bayes' theorem implies $p(\theta|x_1, \dots, x_N) \propto p(\theta) \prod_{i=1}^N p(x_i|\theta)$

- MAP solution

$$\hat{\theta}_{\text{MAP}} = \underset{\theta \in \Theta}{\operatorname{argmax}} p(\theta) \prod_{i=1}^N p(x_i|\theta)$$



Bayesian inference

- both MLE & MAP estimation are *point estimations*
- Bayesian inference
 - updates *prior distribution* by replacing it with posterior distribution
- conjugate prior
 - if prior can be further parameterized by hyperparameter α and posterior is in same probability distribution family, both prior and posterior called *conjugate distributions*, prior called *conjugate prior*

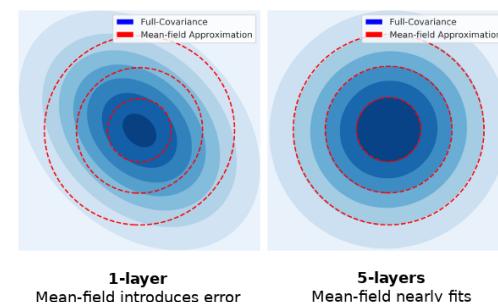
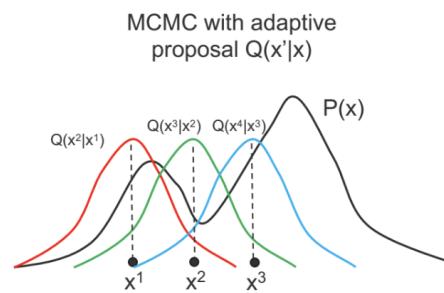
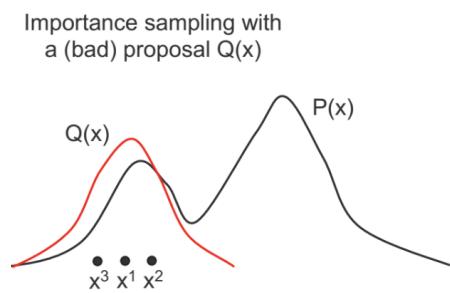
$$p(\theta; \alpha)$$

- in this case, can update hyperparameter α , i.e., find α^+ such that

$$p(\theta; \alpha^+) = p(\theta | x_1, \dots, x_N; \alpha) = \frac{p(\theta; \alpha) \prod_{i=1}^N p(x_i | \theta; \alpha)}{p(x_1, \dots, x_N; \alpha)}$$

Bayesian algorithms & methods

- exact inference methods
 - conjugate priors - *e.g.*, Beta-Binomial, Normal-Normal, *etc.*
- Markov Chain Monte Carlo (MCMC)
 - Metropolis-Hastings algorithm, Gibbs sampling, Hamiltonian Monte Carlo (HMC)
- variational inference (VI)
 - mean field variational Bayes - assuming parameter independence for tractability
 - structured variational inference - maintaining dependencies & inference tractability
 - variational autoencoder (VAE) - NN-based VI for complex distributions

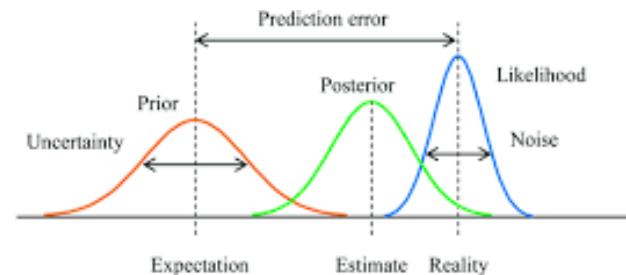


Pros & cons of Bayesian Inference

- pros
 - principled uncertainty quantification - providing complete probability distributions
 - incorporates prior knowledge - allowing to formally include domain expertise, *etc.*
 - coherent framework - providing mathematically consistent approach
 - natural sequential learning - easily handles streaming data or online learning scenarios
 - interpretable results - outputs directly interpretable as probabilities
- cons
 - computational complexity - often requiring sophisticated sampling methods
 - prior sensitivity, scalability issues, implementation difficulty, slower inference, model selection challenges

$$p(\theta | \text{data}) = \frac{p(\text{data} | \theta) \cdot p(\theta)}{p(\text{data})}$$

↑ Posterior ↓ Likelihood ↓ Prior
 Normalization

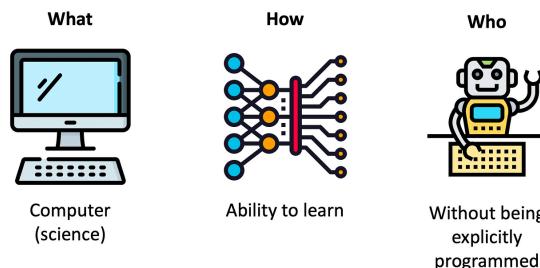


Machine Learning

Machine Learning

- ML

- subfield of computer science that
“gives computers the ability to learn without being explicitly programmed.”
- Arthur Samuel (1959)
- *not* magic, still less intelligent than humans for many cases
- *numerically minimizes* certain (mathematical) loss function to (indirectly) solve *some statistically meaningful* problems



Machine learning is the subfield of computer science that gives “computers the ability to learn without being explicitly programmed.”



Arthur Samuel

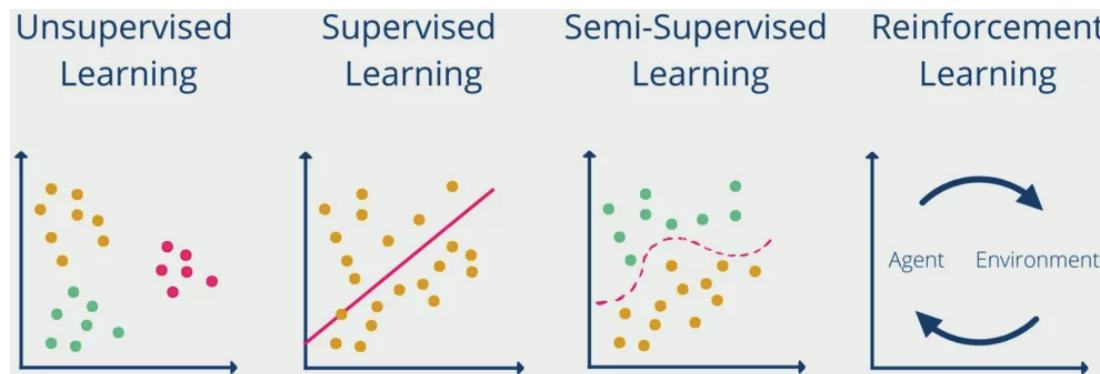
Two famous quotes

- Albert Einstein

The grand aim of all science is to cover the greatest number of empirical facts by logical deduction from the smallest possible number of hypotheses or axioms.

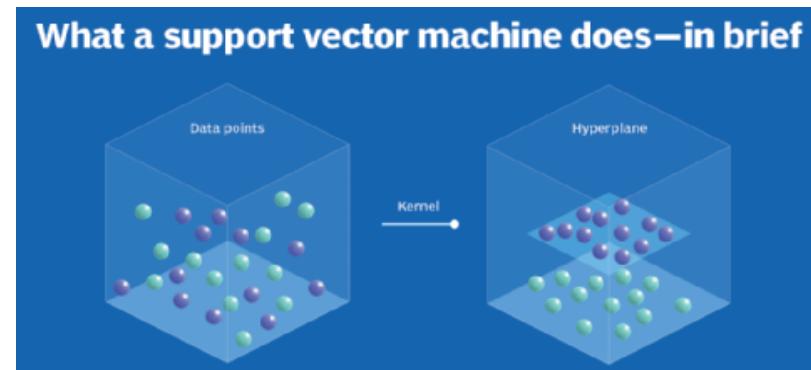
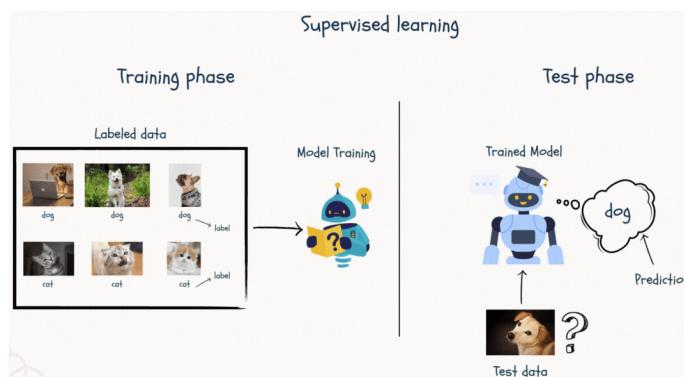
- Alfred North Whitehead

Civilization advances by extending the number of important operations which we can perform without thinking about them. - Operations of thought are like cavalry charges in a battle – they are strictly limited in number, they require fresh horses, and must only be made at decisive moments.



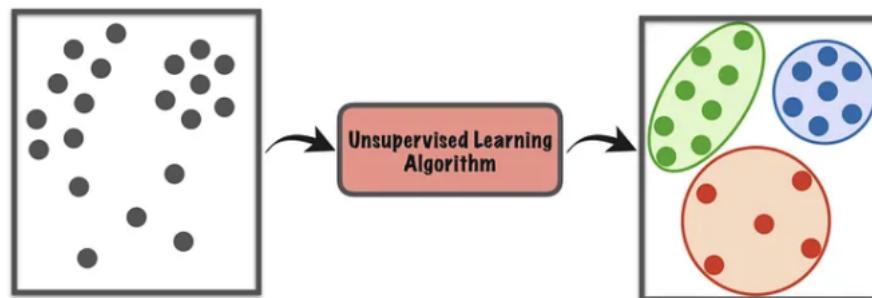
Supervised learning

- most basic and widely used type of ML
- model is trained on dataset where correct output or “label” is provided for each input
- use cases
 - image classification, object detection, semantic segmentation
 - natural language processing (NLP) - text classification, sentiment analysis
 - predictive modeling, medical diagnosis
- algorithms
 - linear regression, logistic regression, decision trees, random forest
 - support vector machine (SVM), k -nearest neighbors (kNN)



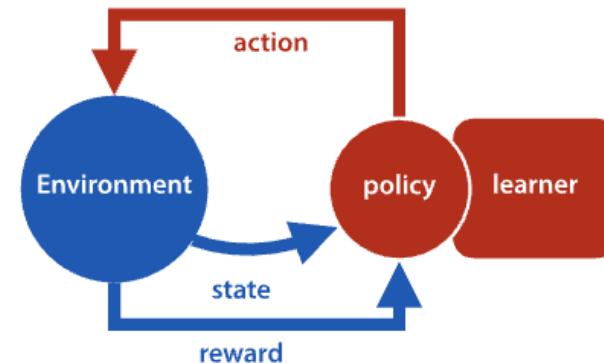
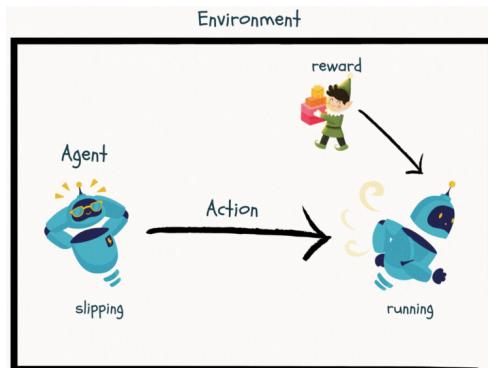
Unsupervised learning

- model is given dataset without any labels or output
- model finds patterns & structure within data on its own
- use cases
 - clustering, dimensionality reduction
 - anomaly detection, generative models
- algorithms
 - k-means clustering, hierarchical clustering, principal component analysis (PCA)
 - t-distributed stochastic neighbor embedding (t-SNE)



Reinforcement learning

- (quite different from supervised & unsupervised learnings)
- model learns from consequences of its actions
 - model receives feedback on its performance; feedback called *reward*
 - uses that information to adjust its actions and improve its performance over time
- use cases
 - robotics, game playing, autonomous vehicles, industrial control
 - healthcare, finance
- algorithms
 - Q-learning, SARSA, DQN, A3C, policy gradient



ML Formulations

Statistical problem formulation

- assume data set $X_m = \{x^{(1)}, \dots, x^{(m)}\}$
 - drawn independently from (true, but unknown) data generating distribution $p_{\text{data}}(x)$
- maximum likelihood estimation (MLE) is to solve

$$\text{maximize } p_{\text{model}}(X; \theta) = \prod_{i=1}^m p_{\text{model}}(x^{(i)}; \theta)$$

where optimization variable is θ

- find *most plausible or likely model* that fits data
- equivalent (but more numerically tractable) formulation

$$\text{maximize } \log p_{\text{model}}(X; \theta) = \sum_{i=1}^m \log p_{\text{model}}(x^{(i)}; \theta)$$

MLE & KL divergence

- in information theory, Kullback-Leibler (KL) divergence defines distance between two probability distributions p & q

$$D_{\text{KL}}(p\|q) = \mathbf{E}_{X \sim p} \log p(X)/q(X) = \int_{x \in \Omega} p(x) \log \frac{p(x)}{q(x)} dx$$

- KL divergence between data distribution p_{data} & model distribution p_{model} can be approximated by Monte Carlo method as

$$D_{\text{KL}}(p_{\text{data}}\|p_{\text{model}}(\theta)) \simeq \frac{1}{m} \sum_{i=1}^m (\log p_{\text{data}}(x^{(i)}) - \log p_{\text{model}}(x^{(i)}; \theta))$$

where $x^{(i)}$ are drawn (of course) according to p_{data}

- hence *minimizing KL divergence is equivalent to solving MLE problem!*

Equivalence of MLE to MSE

- assume model is Gaussian, *i.e.*, $y \sim \mathcal{N}(g_\theta(x), \Sigma)$ ($g_\theta(x) \in \mathbf{R}^p$, $\Sigma \in \mathbf{S}_{++}^p$)

$$p(y|x; \theta) = \frac{1}{\sqrt{2\pi}^p |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (y - g_\theta(x))^T \Sigma^{-1} (y - g_\theta(x)) \right)$$

- assuming that $\Sigma = \alpha I_p$, log-likelihood becomes

$$\begin{aligned} \sum_{i=1}^m \log p(x^{(i)}, y^{(i)}; \theta) &= \sum_{i=1}^m \log p(y^{(i)}|x^{(i)}; \theta) p(x^{(i)}) \\ &= - \sum_{i=1}^m \|y^{(i)} - g_\theta(x^{(i)})\|_2^2 / 2\alpha - \frac{pm}{2} \log(2\pi\alpha) + \sum_{i=1}^m \log p(x^{(i)}) \end{aligned}$$

- hence *minimizing mean-square-error (MSE) is equivalent to solving MLE problem!*

Numerical optimization problem formulation

- (true) problem to solve

$$\text{minimize } \mathbf{E} l(g_\theta(X), Y)$$

- *impossible* to solve

- basic formulation - surrogate problem to solve

$$\text{minimize } f(\theta) = \frac{1}{m} \sum_{i=1}^m l(g_\theta(x^{(i)}), y^{(i)})$$

- formulation with regularization

$$\text{minimize } f(\theta) = \frac{1}{m} \sum_{i=1}^m l(g_\theta(x^{(i)}), y^{(i)}) + \gamma r(\theta)$$

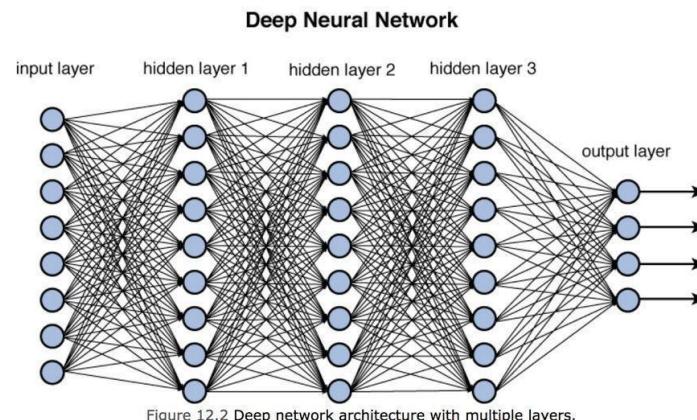
- stochastic gradient descent (SGD)

$$\theta^{(k+1)} = \theta^{(k)} - \alpha_k \nabla f(\theta^{(k)})$$

Deep Learning

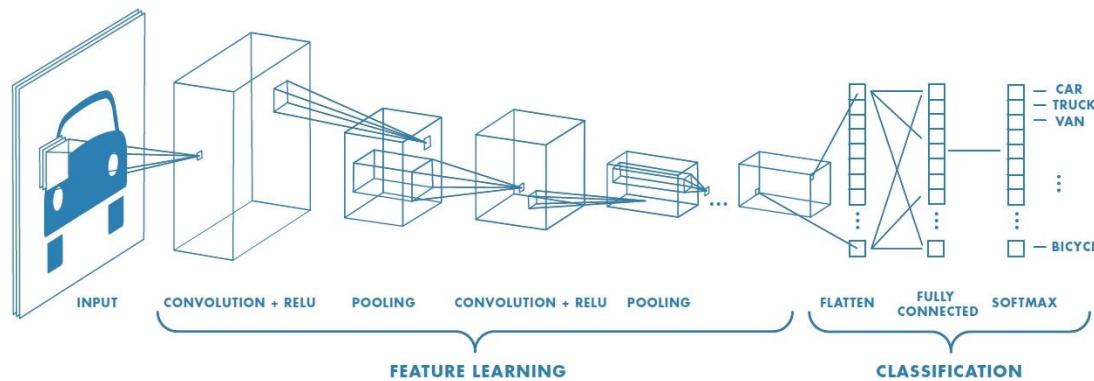
Deep learning (DL)

- machine learning using artificial neural networks with multiple layers for
 - automatically learning hierarchical representations of data
- key components
 - deep neural networks, hidden layers, backpropagation, activation functions
 - hierarchical feature learning, representation learning, end-to-end learning
- key breakthroughs enabling DL
 - massively available data, GPU computing, algorithmic advances



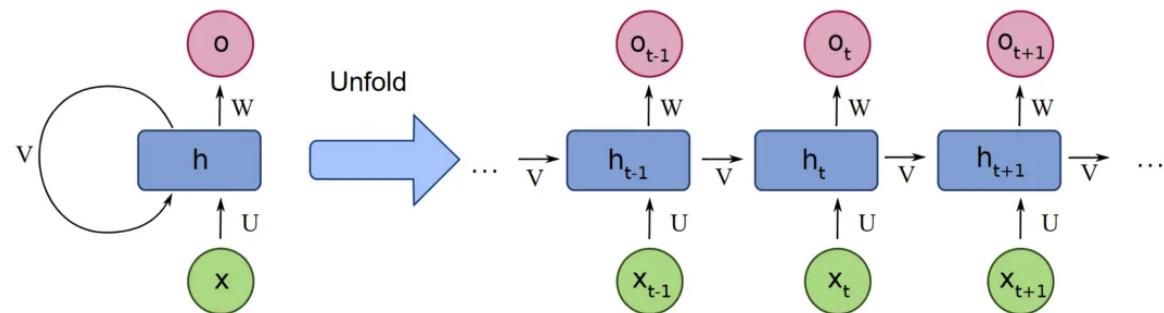
Convolutional neural network (CNN)

- specialized DL learning architecture designed for
 - processing grid-like data such as images
 - where spatial relationships between pixels matter
- key components
 - convolutional layers, pooling layers, activation functions, fully connected layers
- how it works
 - feature extraction, translation invariance, parameter sharing
- why it excels
 - local connectivity, hierarchical learning



Recurrent neural network (RNN)

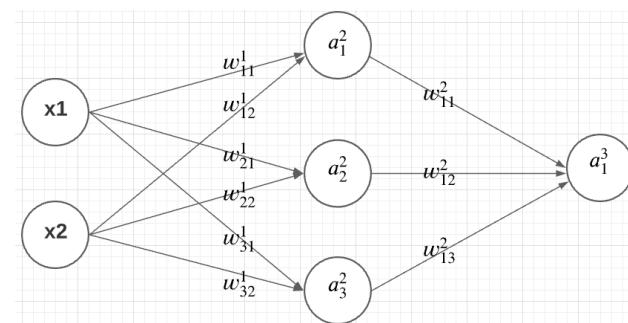
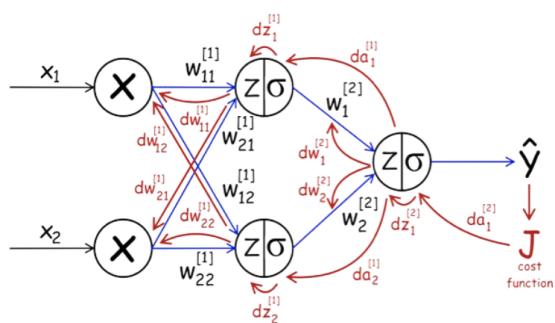
- neural network designed for
 - processing sequential data by maintaining memory of previous inputs
- key components
 - hidden states, recurrent connections, input/output layers, weight sharing
- how it works
 - sequential processing, memory mechanism, temporal dependencies
- why it excels
 - variable length input, context awareness, flexible architecture
- variants - long short-term memory (LSTM), gated recurrent unit (GRU)



Training DNN using SGD

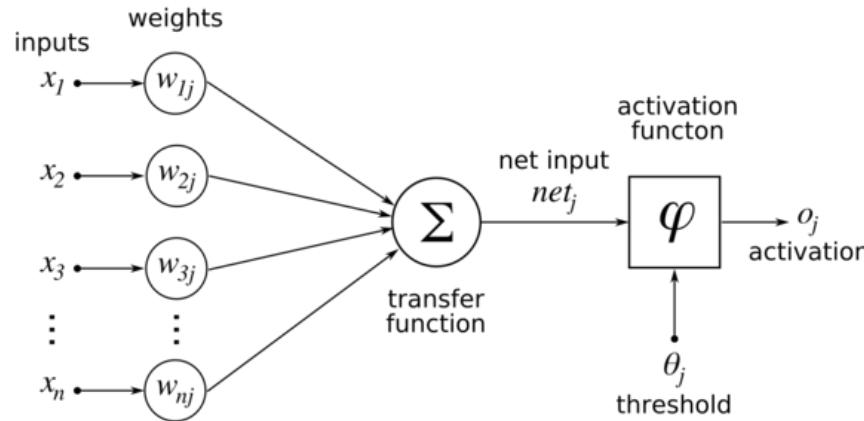
Notations

- p / q - dimension of input / output spaces
- $l : \mathbf{R}^q \times \mathbf{R}^q \rightarrow \mathbf{R}_+$ - loss function
- d - depth of neural network
- n_i ($1 \leq i \leq d$) - number of perceptrons in i th layer
- $z^{[i]} \in \mathbf{R}^{n_i}$ - input to i th layer
- $o^{[i]} \in \mathbf{R}^{n_i}$ - output of i th layer
- $W^{[i]} \in \mathbf{R}^{n_i \times n_{i-1}}$ - weights of connections between $(i-1)$ th and i th layer
- $w^{[i]} \in \mathbf{R}^{n_i \times n_{i-1}}$ - bias weights of i th layer
- $\phi^{[i]} : \mathbf{R}^{n_i} \rightarrow \mathbf{R}^{n_i}$ - activation functions of i th layer

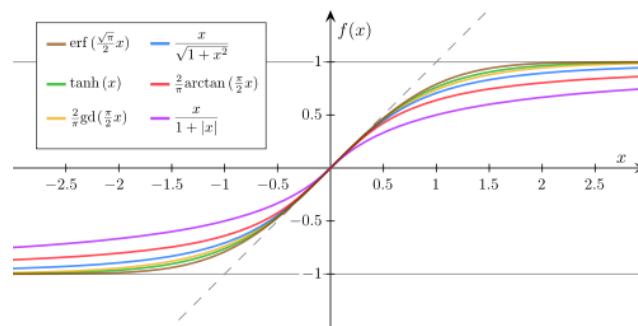


Basic unit & activation function

- basic unit



- activation function



Neural net equations

- modeling function for the (deep) neural network $g_\theta : \mathbf{R}^p \rightarrow \mathbf{R}^q$

$$g_\theta = \phi_\theta^{[d]} \circ \psi_\theta^{[d]} \circ \cdots \circ \phi_\theta^{[1]} \circ \psi_\theta^{[1]}$$

or equivalently

$$g_\theta(x) = \phi_\theta^{[d]}(\psi_\theta^{[d]}(\cdots(\phi_\theta^{[1]}(\psi_\theta^{[1]}(x)))))$$

- for i th layer
 - output via (componentwise) activation function

$$o^{[i]} = \phi^{[i]}(z^{[i]}) \Leftrightarrow o_j^{[i]} = \phi_j^{[i]}(z_j^{[i]}) \quad (1 \leq j \leq n_i)$$

- input via affine transformation $\psi^{[i]} : \mathbf{R}^{n_{i-1}} \rightarrow \mathbf{R}^{n_i}$

$$z^{[i]} = \psi^{[i]}(o^{[i-1]}) = W^{[i]}o^{[i-1]} + w^{[i]}$$

Stochastic gradient descent

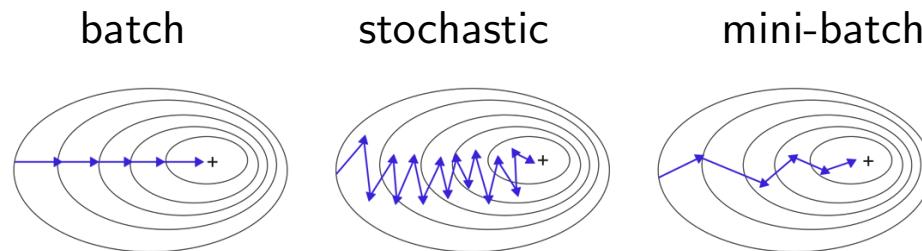
- ML training tries to minimize some loss function - $f(\theta)$ depends on (not only θ , but also) batch of data $(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})$

$$\text{minimize } f(\theta)$$

- while exist hundreds of optimization methods solving this problem
 - the only method used widely* is stochastic *gradient descent!*
- (stochastic) gradient descent

$$f(\theta^{k+1}) = f(\theta^k) - \alpha^k \nabla f(\theta^k)$$

- backpropagation* is used to evaluate this (stochastic) *gradient* using *chain rule*



Chain rule

- suppose
 - two functions $f : \mathbf{R}^n \rightarrow \mathbf{R}^m$ & $g : \mathbf{R}^m \rightarrow \mathbf{R}$
 - Jacobian of f - $Df : \mathbf{R}^n \rightarrow \mathbf{R}^{m \times n}$
 - gradient of g - $\nabla g : \mathbf{R}^m \rightarrow \mathbf{R}^m$
- gradient of composite function $h = g \circ f$

$$\nabla h(\theta) = Df(\theta)^T \nabla g(f(\theta)) \in \mathbf{R}^n \quad (\text{using matrix-vector multiplication})$$

in other words

$$\frac{\partial}{\partial \theta_i} h(\theta) = \sum_{j=1}^m \frac{\partial}{\partial \theta_i} f_j(\theta) \nabla_j g(f(\theta)) \quad (\text{scalar version})$$

Loss function & its gradient

- assume cost function of deep neural network is

$$f(\theta) = \frac{1}{m} \sum_{k=1}^m l(g_\theta(x^{(k)}), y^{(k)}) = \frac{1}{m} \sum_{k=1}^m f_k(\theta)$$

where

$$f_k(\theta) = l(g_\theta(x^{(k)}), y^{(k)})$$

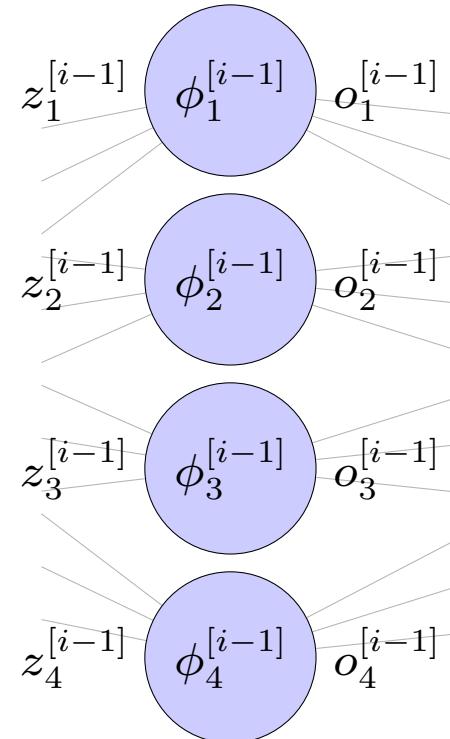
- gradient is

$$m \nabla_\theta f(\theta) = \sum_{k=1}^m \nabla_\theta l(g_\theta(x^{(k)}), y^{(k)}) = \sum_{k=1}^m \nabla_\theta f_k(\theta)$$

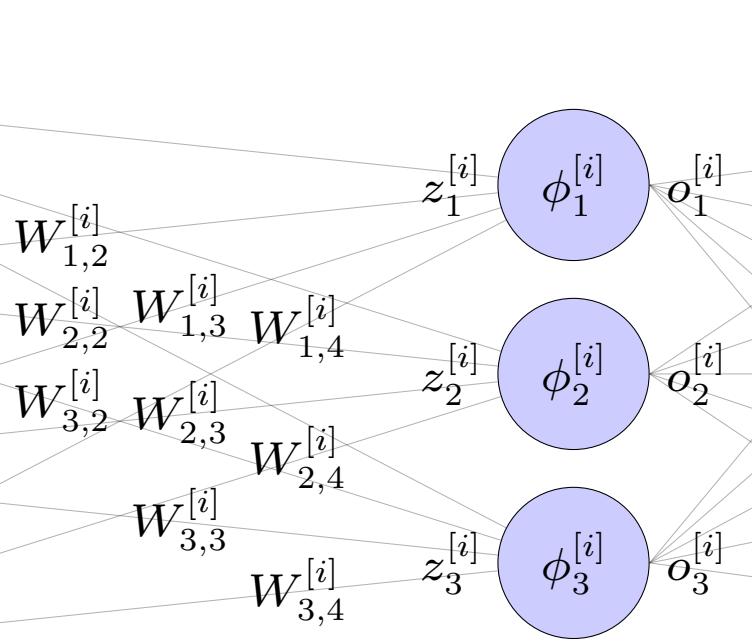
- i.e., evaluate gradient $\nabla_\theta f_k(\theta)$ for each data point $(x^{(k)}, y^{(k)})$

Hidden layers

(i - 1)th hidden layer



ith hidden layer



Backpropagation formula using chain rule

- for each data $(x^{(k)}, y^{(k)})$
 - via activation function

$$\frac{\partial}{\partial z_j^{[i]}} f_k(\theta) = \frac{\partial}{\partial o_j^{[i]}} f_k(\theta) \phi_j^{[i]'}(o_j^{[i]}) \quad \text{for } 1 \leq j \leq n_i \quad (1)$$

where $\phi_j^{[i]'}(o_j^{[i]})$ is derivative of activation function $\phi_j^{[i]}$ evaluated at $o_j^{[i]}$

- via affine transformation

$$\frac{\partial}{\partial W_{j,l}^{[i]}} f_k(\theta) = o_l^{[i-1]} \frac{\partial}{\partial z_j^{[i]}} f_k(\theta) \quad \text{for } 1 \leq j \leq n_i \text{ & } 1 \leq l \leq n_{i-1} \quad (2)$$

$$\frac{\partial}{\partial w_j^{[i]}} f_k(\theta) = \frac{\partial}{\partial z_j^{[i]}} f_k(\theta) \quad \text{for } 1 \leq j \leq n_i \quad (3)$$

$$\frac{\partial}{\partial o_l^{[i-1]}} f_k(\theta) = \sum_{j=1}^{n_i} W_{j,l}^{[i]} \frac{\partial}{\partial z_j^{[i]}} f_k(\theta) \quad \text{for } 1 \leq l \leq n_{i-1} \quad (4)$$

Backpropagation formula using matrix-vector multiplication

- for each data $(x^{(k)}, y^{(k)})$

- via activation function

$$\nabla_{z^{[i]}} f_k(\theta) = D\phi^{[i]} \nabla_{o^{[i]}} f_k(\theta) \quad (5)$$

where $D\phi^{[i]} = \text{diag}(\phi_1^{[i]'}(o_1^{[i]}), \dots, \phi_{n_i}^{[i]'}(o_{n_i}^{[i]}))$ is Jacobian of $\phi^{[i]}$ evaluated at $o^{[i]}$

- via affine transformation

$$\nabla_{W^{[i]}} f_k(\theta) = \nabla_{z^{[i]}} f_k(\theta) o^{[i-1]T} \in \mathbf{R}^{n_i \times n_{i-1}} \quad (6)$$

$$\nabla_{w^{[i]}} f_k(\theta) = \nabla_{z^{[i]}} f_k(\theta) \in \mathbf{R}^{n_i} \quad (7)$$

$$\nabla_{o^{[i-1]}} f_k(\theta) = W^{[i]T} \nabla_{z^{[i]}} f_k(\theta) \in \mathbf{R}^{n_{i-1}} \quad (8)$$

Backpropagation formula using Python numpy package

- for each data $(x^{(k)}, y^{(k)})$
 - via activation function

$$\text{grad_z} = \text{phi_dir} * \text{grad_o} \quad (9)$$

where grad_z , phi_dir , grad_o are 1d numpy.ndarray of size n_i

- via affine transformation

$$\text{grad_W} = \text{numpy.dot}(\text{grad_z}, \text{val_o.T}) \quad (10)$$

$$\text{grad_w} = \text{grad_z}.copy() \quad (11)$$

$$\text{grad_o_prev} = \text{numpy.dot}(\text{grad_z}, \text{W}) \quad (12)$$

where val_o , grad_w are 1d numpy.ndarray of size n_i , grad_o_prev is 1d numpy.ndarray of size n_{i-1} , grad_W is 2d numpy.ndarray of shape (n_i, n_{i-1})

Gradient evaluation using backpropagation

- forward propagation - evaluate for each $(x^{(k)}, y^{(k)})$

$$g_{\theta}(x^{(k)}) = \phi_{\theta}^{[d]}(\psi_{\theta}^{[d]}(\cdots(\phi_{\theta}^{[1]}(\psi_{\theta}^{[1]}(x^{(k)})))))$$

- *backpropagation - evaluate partial derivatives backward*

- evaluate gradient with respect to output of output layer $o^{[d]} = g_{\theta}(x^{(k)})$

$$\nabla_{o^{[d]}} f_k(\theta) = \nabla_{y_1} l(g_{\theta}(x^{(k)}), y^{(k)})$$

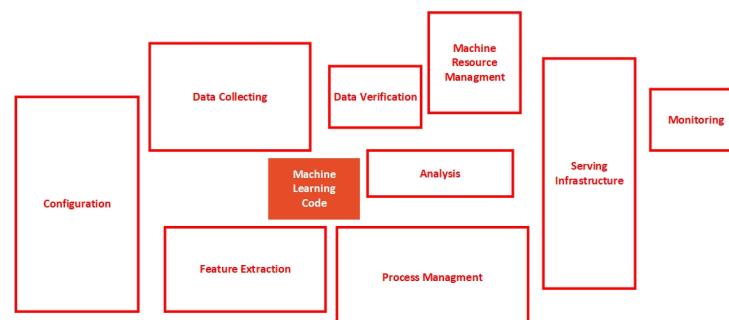
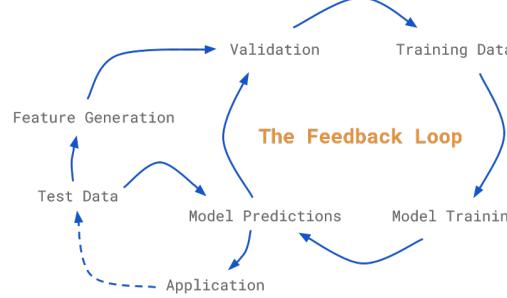
- evaluate gradient with respect to input from that with respect to output using (1), or equivalently, using (5) *i.e.*, evaluate $\nabla_{z^{[i]}} f_k(\theta)$ from $\nabla_{o^{[i]}} f_k(\theta)$
- evaluate gradient with respect to weights, bias, and intput of previous layer using (3), (4), & (2) or equivalently, using (7), (8), & (6) *i.e.*, evaluate $\nabla_{W^{[i]}} f_k(\theta)$, $\nabla_{w^{[i]}} f_k(\theta)$ & $\nabla_{o^{[i-1]}} f_k(\theta)$ from $\nabla_{z^{[i]}} f_k(\theta)$
- repeat back to input layer to evaluate all

$$\nabla_{W^{[1]}} f_k(\theta), \nabla_{w^{[1]}} f_k(\theta), \dots, \nabla_{W^{[d]}} f_k(\theta), \nabla_{w^{[d]}} f_k(\theta)$$

ML in Action

ML in practice

- define business problem - business objective, success metrics, establish baselines (early)
- data collection - data cleaning, validation & exploratory data analysis (EDA)
- feature engineering - based on domain expertise
- train/validation/test split - stratified sampling, chronological splits for time-series
- model selection or/and hyperparameter optimization
- monitoring, retraining & notification
- start simple, iterative fast (fail fast!), validate business impact - *e.g.*, A/B test



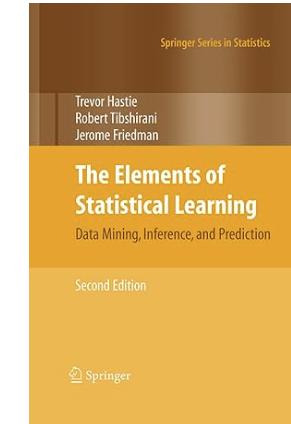
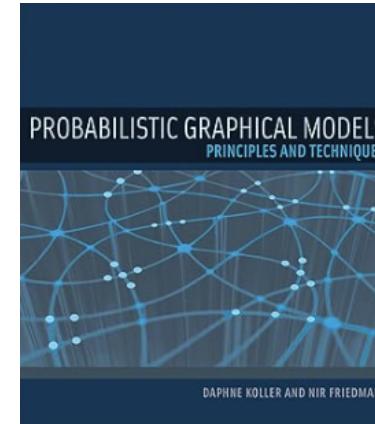
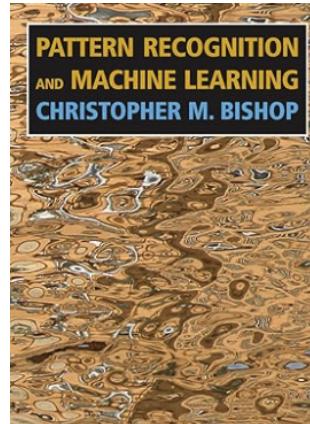
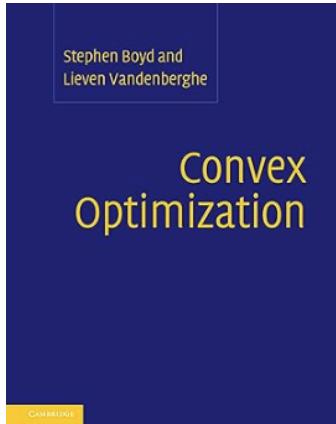
Learning AI

Best ways to learn AI & ML

- first, learn basics - college classes, online courses, (easy) books
 - no need to understand every mathematical details, but should know rough ideas!
- *hands-on is MUST!*
 - learn and practice coding - Python is MUST; do not do (only) R
 - learn git - know how to develop efficiently, plus import others' work
- I think *online courses are blessing to mankind!*
 - *can't* say “you can't do it because I don't have access to good resource or you don't go to good schools” because . . . they are available!
 - getting (expensive) certificates is good idea because . . . otherwise you wouldn't complete it! :) - and can post it on your LinkedIn!
- would be best if your task at work is related to ML
 - however, even if that's not the case or can't be the case, can always do your own personal projects – or contribute to public projects (on github)!

Books

- The Elements of Statistical Learning - Hastie, Tibshirani & Friedman [[HTF01](#)]
- Pattern Recognition and Machine Learning - Christopher M. Bishop [[Bis06](#)]
- Deep Learning - Ian Goodfellow, Yoshua Bengio & Aaron Courville [[GBC16](#)]
- Reinforcement Learning: An Introduction - Richard S. Sutton & Andrew G. Barto [[SB18](#)]
- Machine Learning: A Probabilistic Perspective - Kevin P. Murphy [[Mur12](#)]
- Probabilistic Graphical Models - Daphne Koller & Nir Friedman [[KF09](#)]
- Convex Optimization - Stephen Boyd & Lieven Vandenberghe [[BV04](#)]



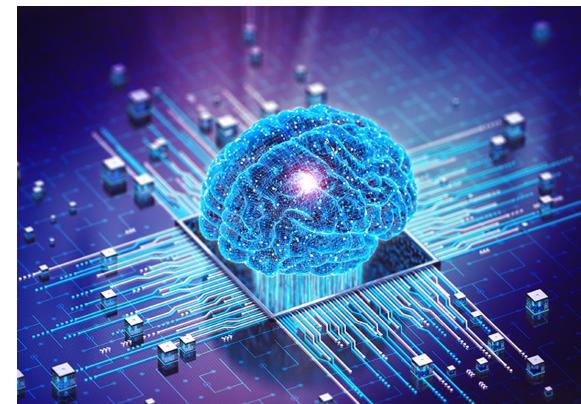
Andrew Ng!

- Andrew Ng
 - (co-)founder of “Deep Learning.AI” and “Coursera”, prominent figure in ML & AI
 - his courses highly regarded because well-structured and provide insights
- [latest Andrew Ng courses](#)
 - AI Agents in LangGraph
 - AI Agentic Design Patterns with AutoGen
 - Introduction to On-device AI
 - Multi AI Agent Systems with Crew AI
 - Building Multimodal Search and RAG - contrastive learning, multimodality to RAG
 - Building Agentic RAG with LlamaIndex
 - Quantisation In Depth
 - In Prompt Engineering for Vision Models
 - Getting Started with Mistral - open-source models (Mistral 7B, Mixtral 8x7B)
 - Preprocessing Unstructured Data for LLM

AI & Biotech

AI in biology

- AI has been used in biological sciences, and science in general
- AI's ability to process large amounts of raw, unstructured data (*e.g.*, DNA sequence data)
 - reduces time and cost to conduct experiments in biology
 - enables other types of experiments that previously were unattainable
 - contributes to broader field of engineering biology or biotechnology
- AI increases human ability to make direct changes at cellular level and create novel genetic material (*e.g.*, DNA and RNA) to obtain specific functions



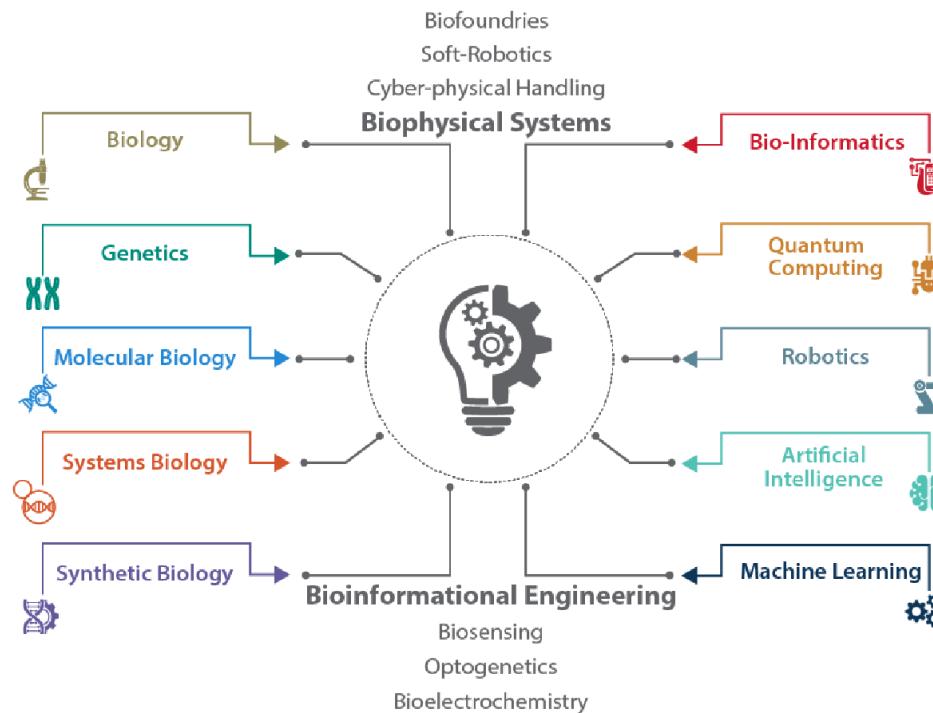
Biotech

Biotech

- biotechnology
 - is multidisciplinary field leveraging broad set of sciences and technologies
 - relies on and builds upon advances in other fields such as nanotechnology & robotics, and, increasingly, AI
 - enables researchers to read and write DNA
 - sequencing technologies “read” DNA while gene synthesis technologies take sequence data and “write” DNA turning data into physical material
- 2018 National Defense Strategy & Senior US Defense and Intelligence Officials identified emerging technologies that could have disruptive impact on US national security [[Say21](#)]
 - *AI*, lethal autonomous weapons, hypersonic weapons, directed energy weapons, *biotechnology*, quantum technology
- other names for biotechnology are engineering biology, synthetic biology, biological science (when discussed in context of AI)

Biotech - multidisciplinary field

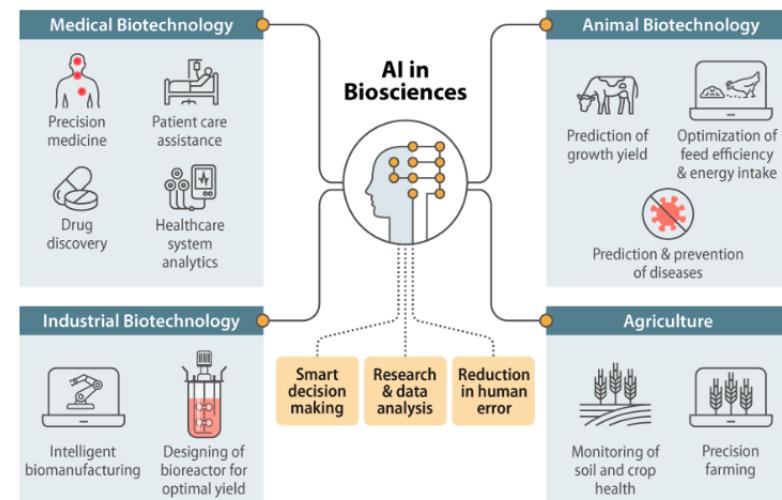
- sciences and technologies enabling biotechnology include (but not limited to)
 - (molecular) biology, genetics, systems biology, synthetic biology, bio-informatics, quantum computing, robotics [DFJ22]



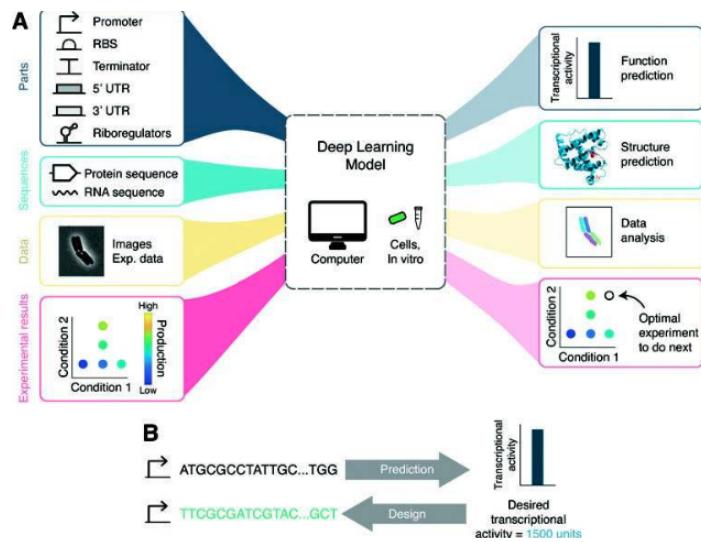
Convergence of AI and biological design

- AI & biological sciences converging [BKP22]
 - each building upon the other's capabilities for new research and development across multiple areas
- Demis Hassabis, CEO & cofounder of DeepMind, said of biology [Toe23]

“... biology can be thought of as information processing system, albeit extraordinarily complex and dynamic one ... just as mathematics turned out to be the right description language for physics, biology may turn out to be *the perfect type of regime for the application of AI!*”
- both AI & biotech rely on and build upon advances in other scientific disciplines and technology fields, such as nanotechnology, robotics, and increasingly big data (e.g., genetic sequence data)
 - each of these fields itself convergence of multiple sciences and technologies
- so *their impacts can combine to create new capabilities*



Multi-source genetic sequence data



- AI, essential to analyzing exponential growth of genetic sequence data

“AI will be essential to fully understanding how genetic code interacts with biological processes” - US National Security Commission on Artificial Intelligence (NSCAI)

 - process huge amounts of biological data, e.g., genetic sequence data, coming from different biological sources for understanding complex biological systems
 - sequence data, molecular structure data, image data, time-series, omics data
 - e.g., analyze genomic data sets to determine the genetic basis of particular trait and potentially uncover genetic markers linked with that trait

Quality & quantity of biological data

- limiting factor, however, is *quality and quantity* of biological data, *e.g.*, DNA sequences, that AI is trained on
 - *e.g.*, accurate identification of particular species based on DNA requires reference sequences of *sufficient quality* to exist and be available
- databases have varying standards - access, type, and quality of information
- design, management, quality standards, and data protocols for reference databases can affect utility of particular DNA sequence



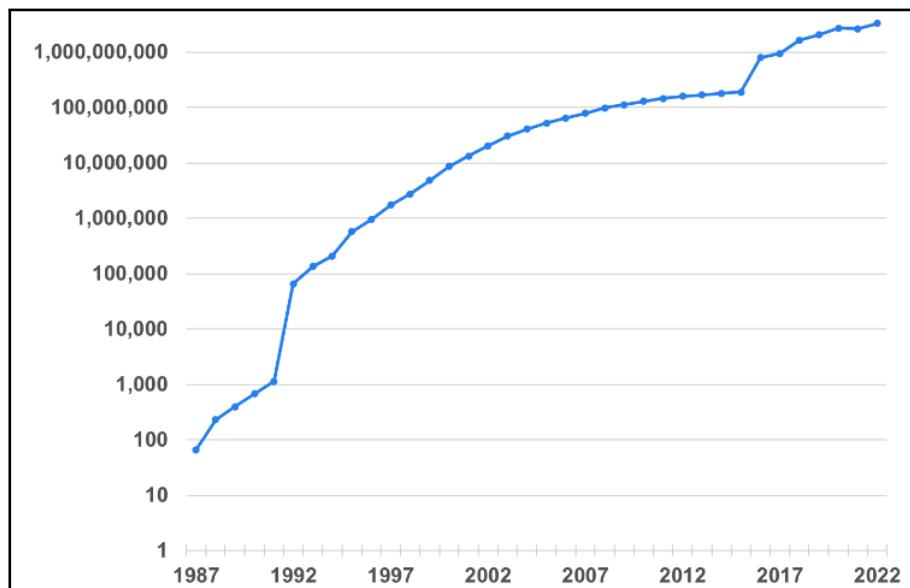
Rapid growth of biological data

- volume of genetic sequence data grown exponentially as sequencing technology evolved
- more than 1,700 databases incorporating data on genomics, protein sequences, protein structures, plants, metabolic pathways, *etc.*, *e.g.*
 - open-source public database
 - Protein Data Bank, US-funded data center - more than *terabyte of three-dimensional structure data* for biological molecules, *e.g.*, proteins, DNA, RNA
 - proprietary database
 - Gingko Bioworks - more than *2B protein sequences*
 - public research groups
 - Broad Institute - produces roughly *500 terabases of genomic data per month*
- great potential value in aggregate volume of genetic datasets that can be collectively mined to discover and characterize relationships among genes

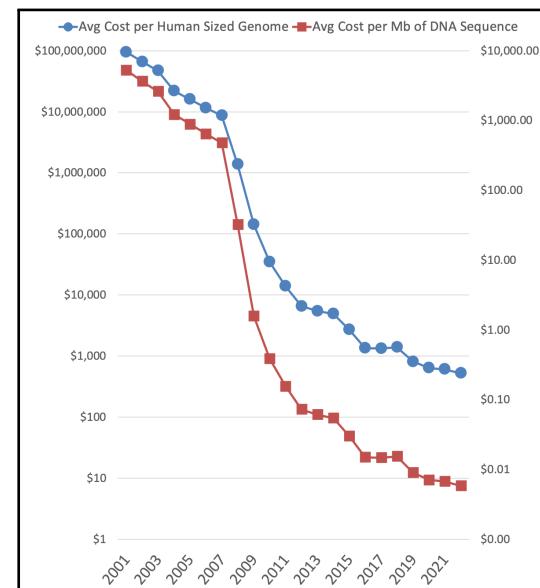
Volume and sequencing cost of DNA over time

- volume of DNA sequences & DNA sequencing cost
 - data source: National Human Genome Research Institute (NHGRI) [[Wet23](#)] & International Nucleotide Sequence Database Collaboration (INSDC)
- more dramatic than Moore's law!*

sequences in INSDC



DNA sequencing cost



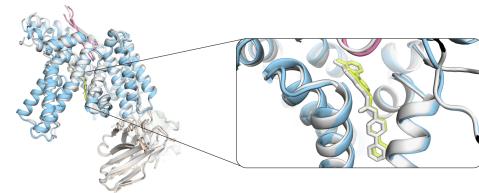
Bio data availability and bias

- US National Security Commission on Artificial Intelligence (NSCAI) recommends
 - US fund and prioritize development of a biobank containing “*wide range of high-quality biological and genetic data sets securely accessible by researchers*”
 - establishment of database of broad range of human, animal, and plant genomes would
 - *enhance and democratize biotechnology innovations*
 - *facilitate new levels of AI-enabled analysis of genetic data*
- bias - availability of genetic data & decisions about selection of genetic data can introduce bias, e.g.
 - training AI model on datasets emphasizing or omitting certain genetic traits can affect how information is used and types of applications developed - *potentially privileging or disadvantaging certain populations*
 - access to data and to AI models themselves may impact communities of differing socioeconomic status or other factors unequally

Emerging Trends in Biotech

AlphaFold

- solving 50-year-old protein folding problem, “*one of biology’s grand challenges*”
 - definition - given amino acid sequence, predict how it folds into a 3D structure
 - proteins fold in microseconds, but predicting computationally nearly impossible
- AlphaFold 1 (2018) - DL + physics-based energy functions → AlphaFold 2 (2020)
 - attention-based NN solving protein folding “in principle” → AlphaFold 3 (2024) - diffusion-based DL, drug-protein interactions, protein complexes
- AlphaFold protein structure database
 - >200MM protein structures - nearly every known protein, used by >2MM researchers
- Applications & implications
 - drug discovery - target identification, lead optimization, side effect prediction
 - enzyme engineering, agriculture, environmental, vaccine development

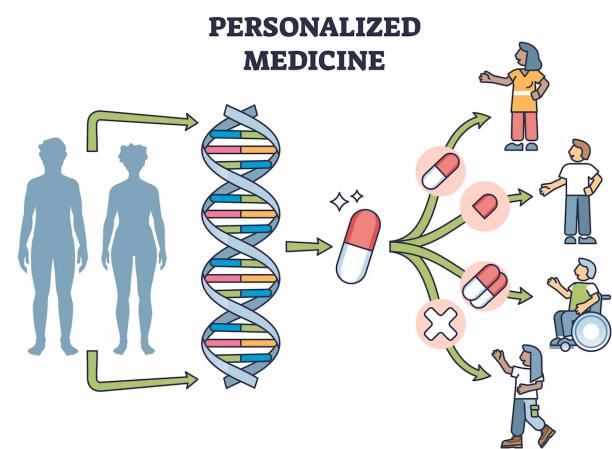


AlphaGo

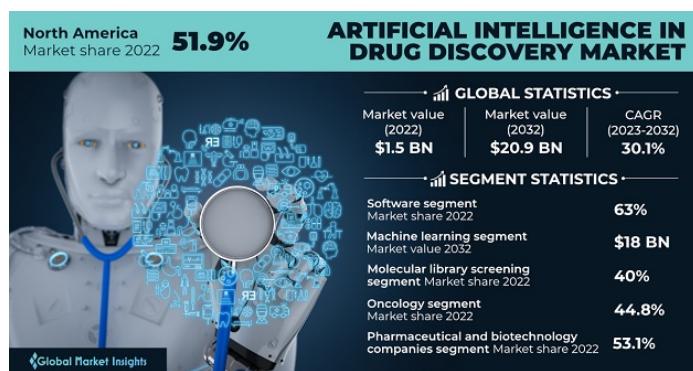
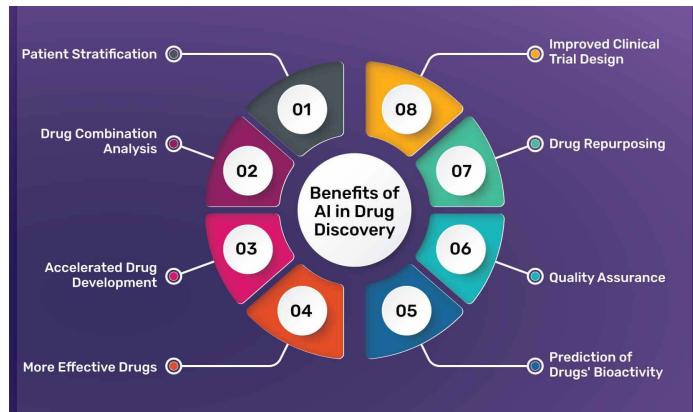


Personalized medicine

- *shift from one-size-fits-all approach to tailored treatments*
- based on individual genetic profiles, lifestyles & environments
- AI enables analysis of vast data to predict patient responses to treatments, thus enhancing efficacy and reducing adverse effects
- e.g.
 - custom cancer therapies
 - personalized treatment plans for rare diseases
 - precision pharmacogenomics
- companies - Tempus, Foundation Medicine, etc.



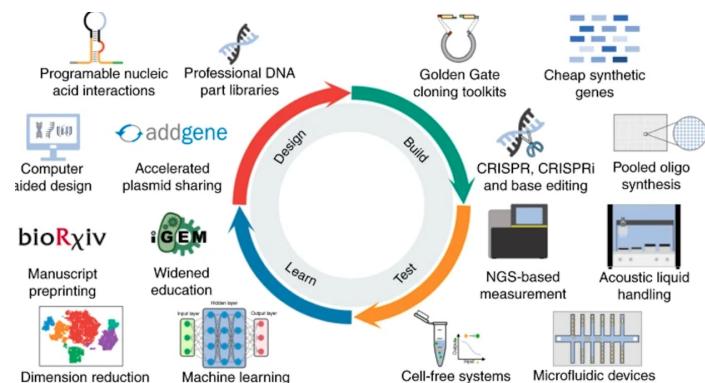
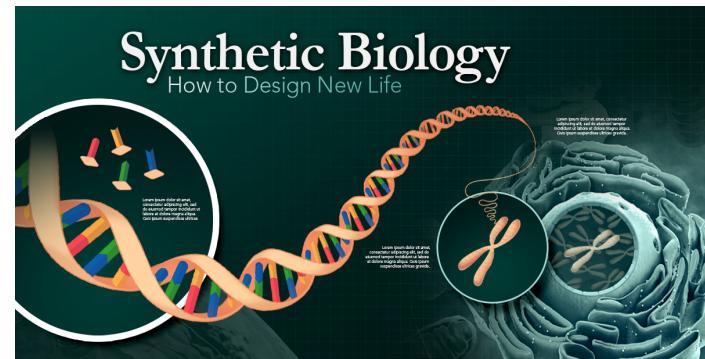
AI-driven drug discovery



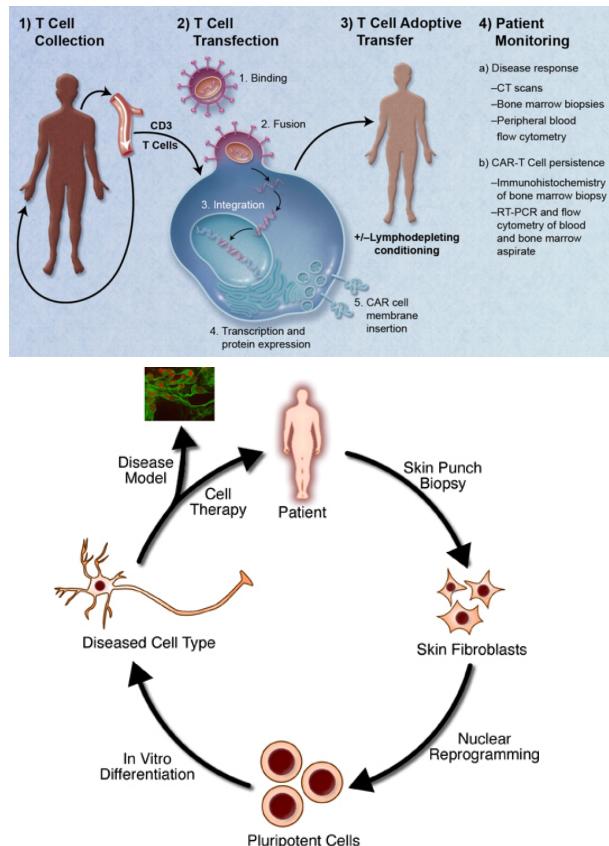
- traditional drug discovery process - time-consuming and costly often taking decades and billions of dollars
- AI streamlines this process by predicting the efficacy and safety of potential compounds with more speed and accuracy
- AI models analyze chemical databases to identify new drug candidates or repurpose existing drugs for new therapeutic uses
- companies - Insilico Medicine, Atomwise.

Synthetic biology

- use AI for gene editing, biomaterial production and synthetic pathways
- combine principles of biology and engineering to design and construct new biological entities
- AI optimizes synthetic biology processes from designing genetic circuits to scaling up production
- company - Ginkgo Bioworks uses AI to design custom microorganisms for applications ranging from pharmaceuticals to industrial chemicals



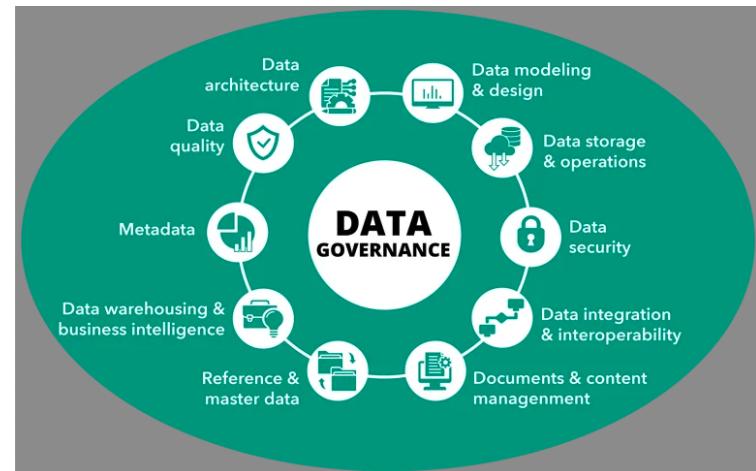
Regenerative medicine



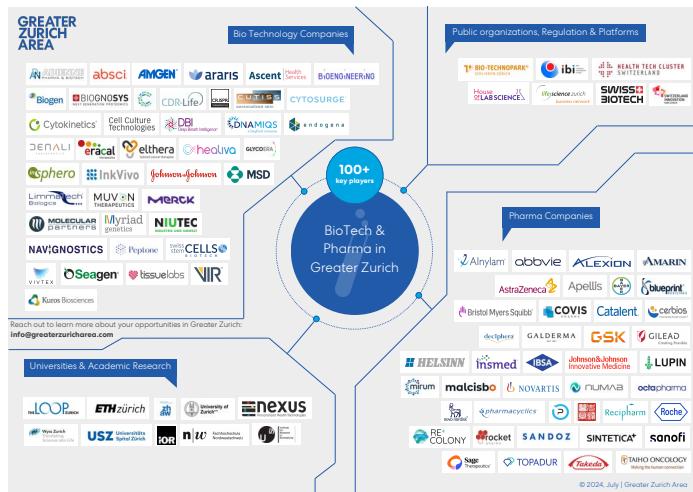
- AI advances development of stem cell therapies & tissue engineering
- AI algorithms assist in identifying optimal cell types, predicting cell behavior & personalized treatments
- particularly for conditions such as neurodegenerative diseases, heart failure and orthopedic injuries
- company - Organovo leverages AI to potentially improve the efficacy and scalability of regenerative therapies, developing next-generation treatments

Bio data integration

- integration of disparate data sources, including genomic, proteomic & clinical data - one of biggest challenges in biotech & healthcare
- AI delivers meaningful insights *only when* seamless data integration and interoperability realized
- developing platforms facilitating comprehensive, longitudinal patient data analysis - vital enablers of AI in biotech
- company - Flatiron Health working on integrating diverse datasets to provide holistic view of patient health



Biotech companies



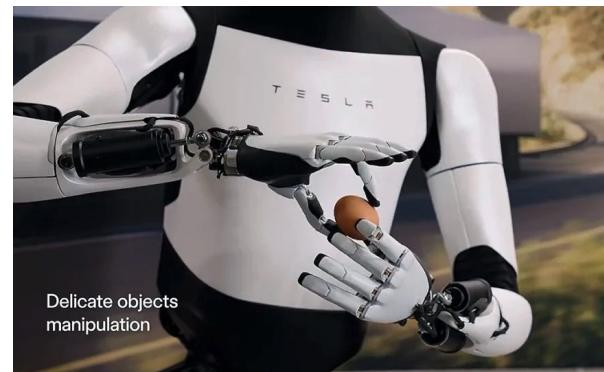
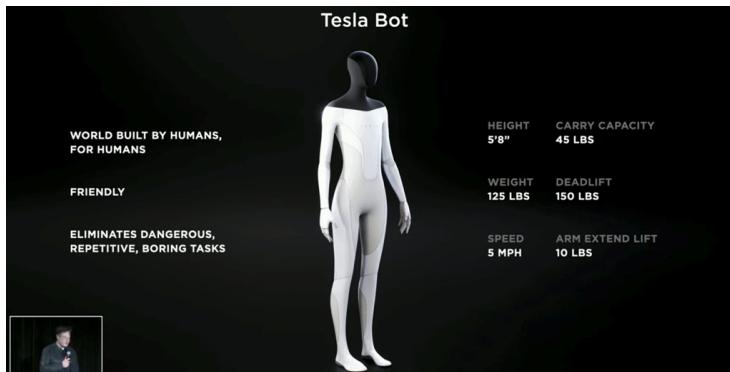
- Atomwise - small molecule drug discovery
- Cradle - protein design
- Exscientia - precision medicine
- Iktos - small molecule drug discovery and design
- Insilico Medicine - full-stack drug discovery system
- Schrödinger, Inc. - use physics-based models to find best possible molecule
- Absci Corporation - antibody design, creating new from scratch antibodies, *i.e.*, “*de novo* antibodies”, and testing them in laboratories

AI-powered Humanoid Robots

Tesla Optimus

Tesla Optimus

- humanoid robot developed by Tesla intended to handle repetitive & dangerous tasks
- objective - *revolutionize automation* & assist in human labor across various industries
- features - [YouTube - Optimus - Gen 2](#)
 - dimensions - 5'8" tall & 125 lbs
 - capabilities - lifting weights, walking at 5 MPH & performing everyday tasks
 - AI-powered - runs on Tesla's AI leveraging same technology used in self-driving cars
 - power source - 2.3 KWH battery designed for efficient power management
 - launch year - announced by Elon Musk during Tesla AI Day in 2021
 - *price* - \$25,000~\$30,000 expected to decrease over time



History of Tesla Optimus

- inception - first conceptualized as extension of Tesla's AI & robotics capabilities
- AI day 2021 - *officially announced by Elon Musk* w/ vision to solve labor shortages & improve productivity
- Sep 2022 - prototype unveiled
- gen 2 introduced in 2023 - improved capabilities
- Jun 2024 w/ more advanced tasks - *towards mass production for commercial applications*

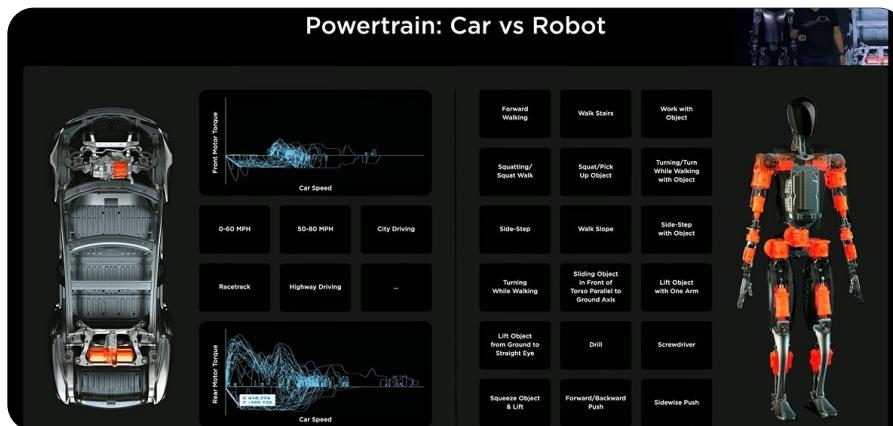
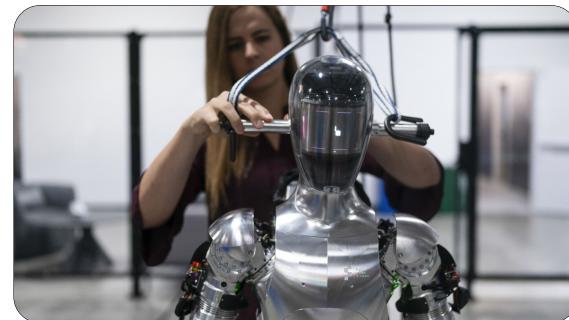


Figure A1

Figure AI robots

- Figure AI
 - founded in 2022 as Silicon Valley startup company by Brett Adcock - serial entrepreneur with successful Archer Aviation & Vetter
 - vision of enhancing productivity by integrating AI and robotics into both industrial & personal spaces
- Figure 02
 - 5'6" tall, 154 lbs, payload of 44 lbs, 5 hr runtime, 1.2 m/s speed
 - imitation learning
 - capabilities - advanced cognition, STS task, dexterous hands w/ 16 degrees of freedom



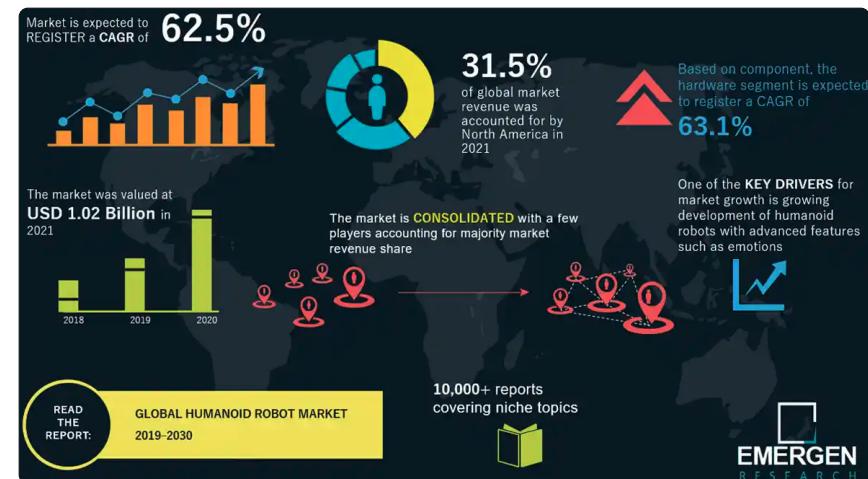
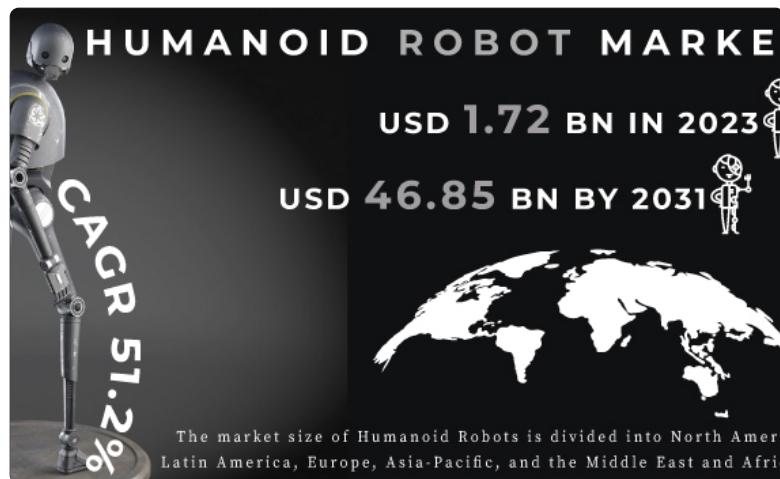
History of Figure AI

- 2022 - founded by Brett Adcock - previously co-founded Archer Aviation & Vetter
- 2022-2023 - early development - stealth mode focusing on developing their own technology
- May 2023 - public announcement - officially announces mission to develop general-purpose humanoid robots - already raised \$70M @ announcement
- Aug 2023 - unveils Figure 01, first prototype w/ basic mobility & manipulation capabilities
- Oct 2023 - series B funding - raised \$675M beyond initial goal of \$500M - Jeff Bezos, Microsoft, OpenAI - valuation of ~ \$2.6B
- late 2023 ~ early 2024 - partnership announcements - refines humanoid robot technology in locomotion, object manipulation & human-robot interaction
- 2024 - significant strides in robot control & decision-making

Impacts & Future

Impacts on industries & markets

- impacts on robotics history
 - competitor benchmark - competes with robotics giants such as Boston dynamics
 - affordability & scale - predict to lead to *lower costs & higher adoption*
- impacts on labor market
 - task automation - replace human labor in *high-risk & repetitive roles*
 - *job displacement vs creation* - new roles in AI, robot maintenance & oversight
- impacts on consumer market - home automation



Future outlook & predictions

- widespread industrial adoption - expected to become common tool in factories by 2030
- market valued @ **\$1.02B in 2021** - expected *CAGR of 62.5%, 63.1% in hardware segment by 2030 - 31.5% revenue increase in 2021* North America - **10,000 humanoid robots** will be shipped worldwide each year by 2027
- AI evolution - continuous learning and AI enhancements will lead to greater efficiency & adaptability
- consumer integration - long-term vision includes personal assistant
- societal impact - could redefine human roles in industries & homes *raising philosophical & ethical questions on human-robot collaboration*



Industrial AI

Industrial AI (inAI)

- inAI (collectively) refers to AI technology & software and their products developed for
 - *customer values creation, productivity improvement, cost reduction, production optimization, predictive analysis, insight discovery*
 - *semiconductor, steel, oil & gas, cement, and other various manufacturing industries* (unlike general AI, which is frontier research discipline striving to achieve human-level intelligence)



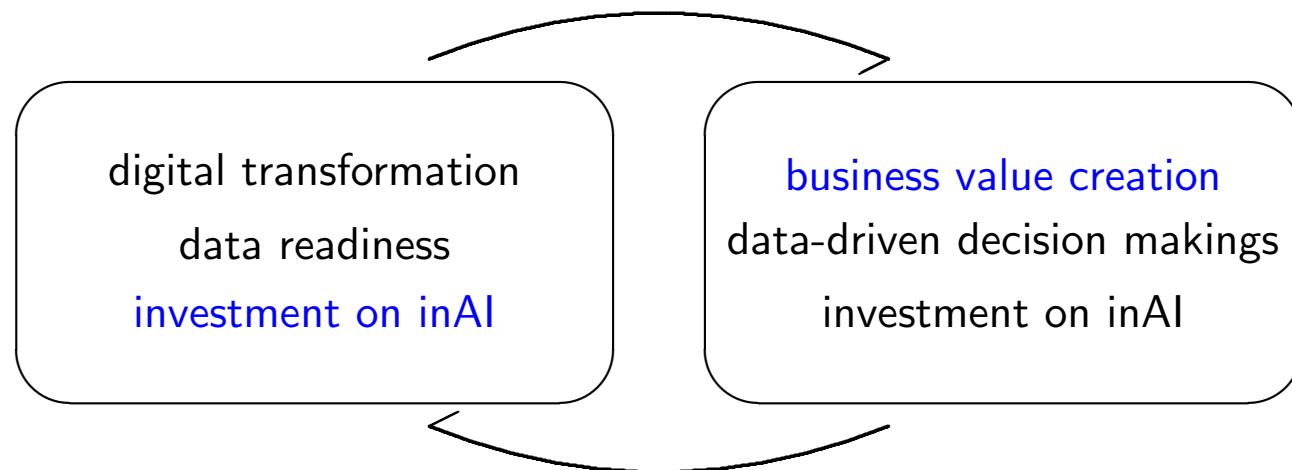
inAI fields

- product
 - product design & innovation, adaptability & advancement, product quality & validation, design for reusability & recyclability, performance optimization
- production process
 - *production quality*, process management, inter-process relations, process routing & scheduling, process design & innovation, *traceability*, *predictive process control*
- machinery & equipment
 - *predictive maintenance*, *monitoring & diagnosis*, component development, *ramp-up optimization*, material consumption prediction
- supply chain
 - supply chain monitoring, material requirements planning, customer management, supplier management, logistics, reusability & recyclability

Characteristics of inAI

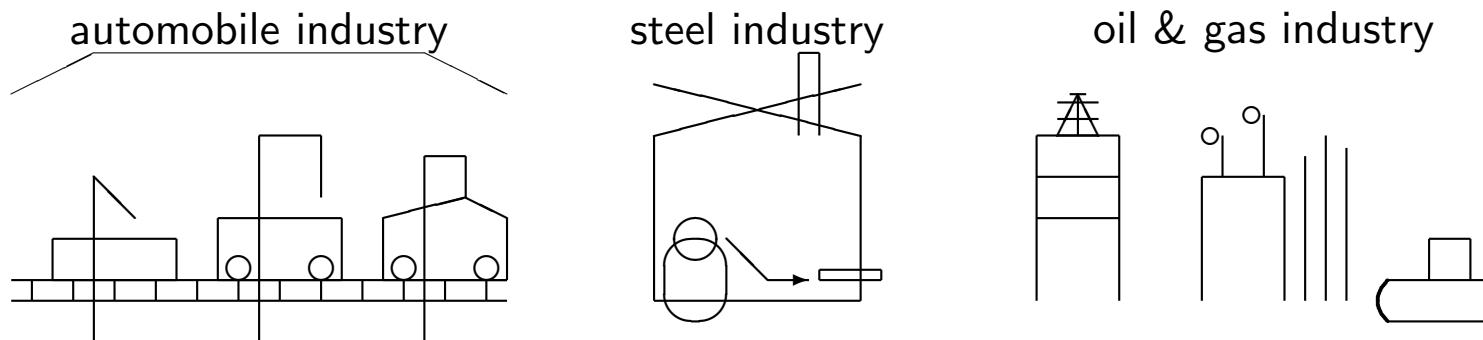
Vicious (or virtuous) cycle

- integration of inAI with customers' business creates monetary values and encourages data-driven decisions
- however, to do so, digital transformation with data-readiness is MUST-have
- created values, in turn, can be invested into infrastructure required for digital transformation and success of inAI!



Data-centric AI

- unlike many ML disciplines where foundation models do generic representation learning, *i.e.*, learn universal features
- each equipment has (gradually) different data characteristics, hence need data-centric AI
 - “... need 1,000 models for 1,000 problems” - Andrew Ng
 - data-centric AI - discipline of systematically engineering the data used to build AI system



Challenging data characteristics

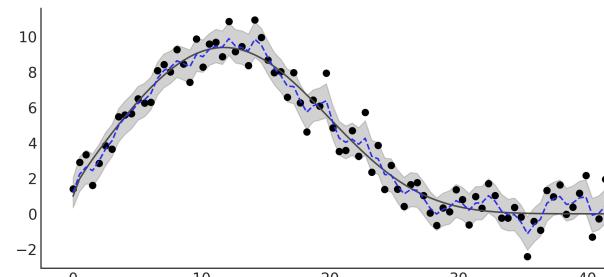
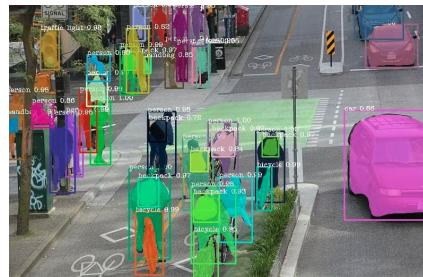
- huge volume
- data multi-modality
- high velocity requirement
- very fat data
- sever data shift & drift (in many cases)
- label imbalance
- data quality



Manufacturing AI

MLs in manufacturing AI (manAI)

- *image data* - huge amount of image data measured and inspected
 - SEM/TEM images, wafer defect maps, test failure pattern maps¹
→ semantic segmentation, defect inspection, anomaly detection
- *time-series (TS) data* - all the data coming out of manufacturing is TS
 - equipment sensor data, process times, various measurements, MES data²
→ regression, anomaly detection, semi-supervised learning, Bayesian inference



¹SEM: scanning electron microscope, TEM: transmission electron microscope

²MES: manufacturing execution system

CV ML in manAI

Computer vision ML in manAI

- measurement and inspection (MI)
 - metrology - measurement of critical features
 - inspection - defect inspection, defect localization, defect classification
 - failure pattern analysis
- applications
 - automatic feature measurement
 - anomaly detection
 - defect inspection

Automatic feature measurement

- ML techniques
 - image enhancement (denoising)
 - texture segmentation
 - repetitive pattern recognition
 - automatic measurement

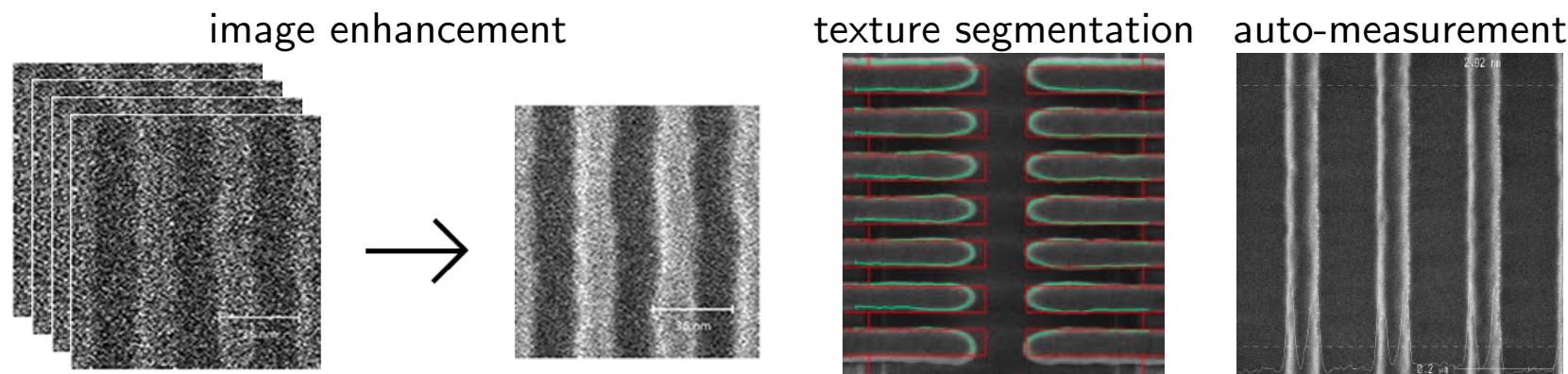


Image enhancement

- image enhancement techniques
 - general supervised denoising using DL
 - blind denoising using DL - remove noise without prior knowledge of noise adapting to various noise types
 - super-resolution - upscale low-resolution images, add realistic details for sharper & higher-quality images

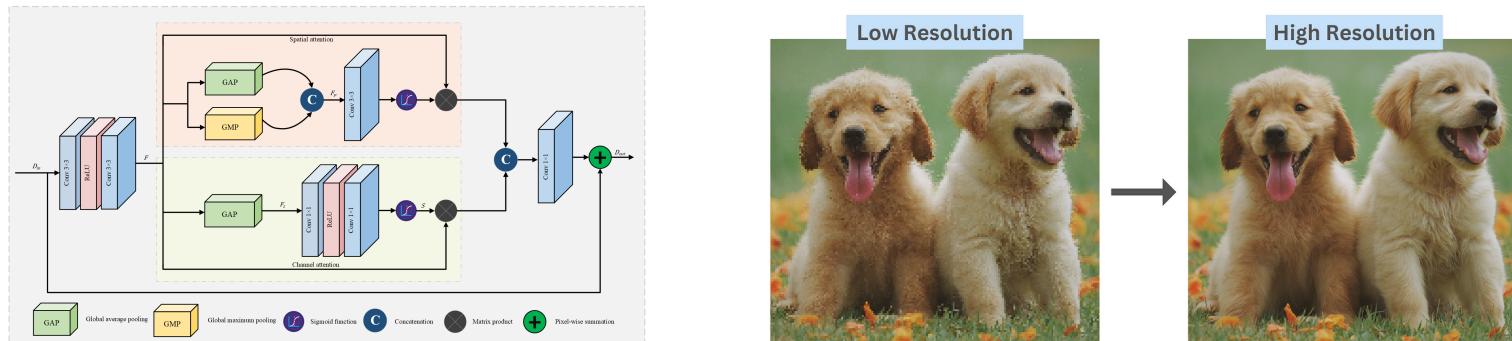
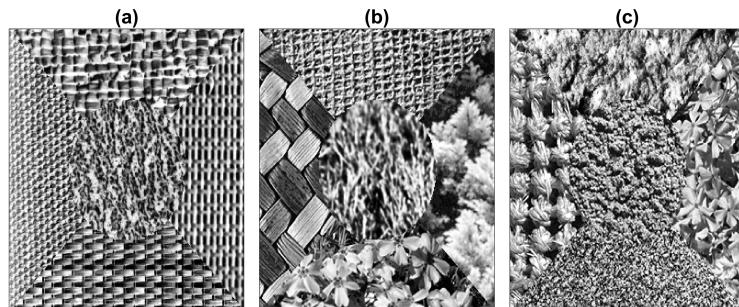


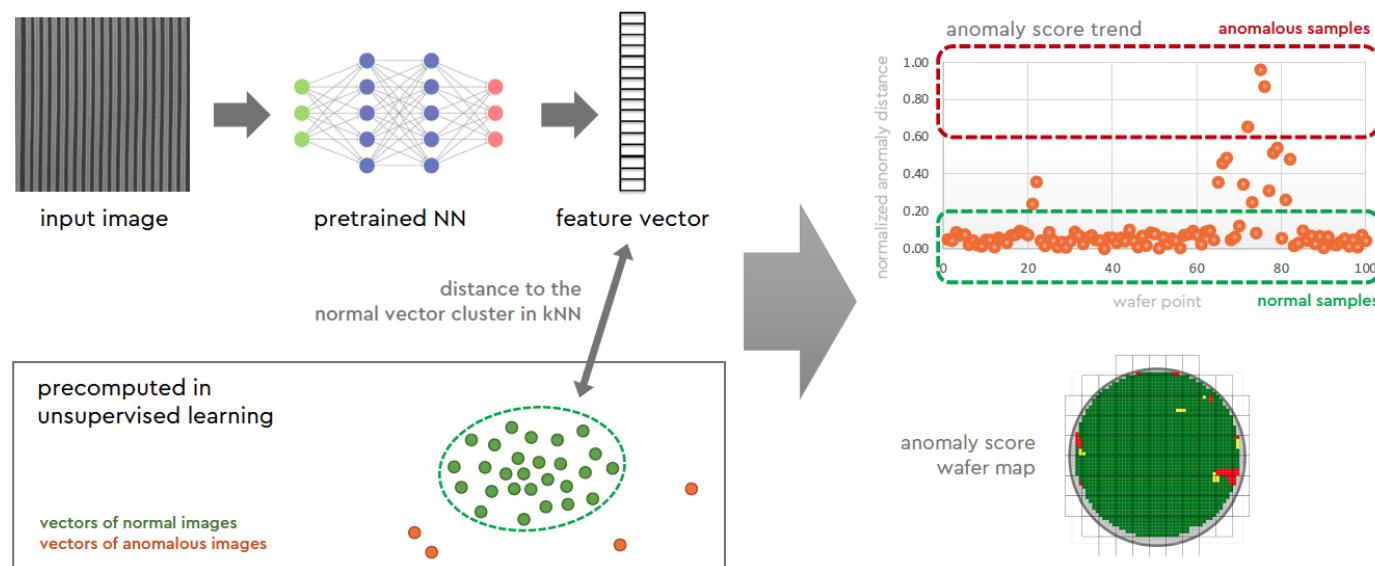
Image segmentation

- texture segmentation
 - distinguish areas based on texture patterns - identifying regions with similar textural features - used for material classification, surface defect detection, medical imaging
 - methods - Gabor filters, wavelet transforms, DL
- semantic segmentation
 - assign class labels to every pixel - enabling precise object and region identification - used for autonomous driving, scene understanding, medical diagnostics
 - methods - fully convolutional network (FCN), U-net, DeepLab



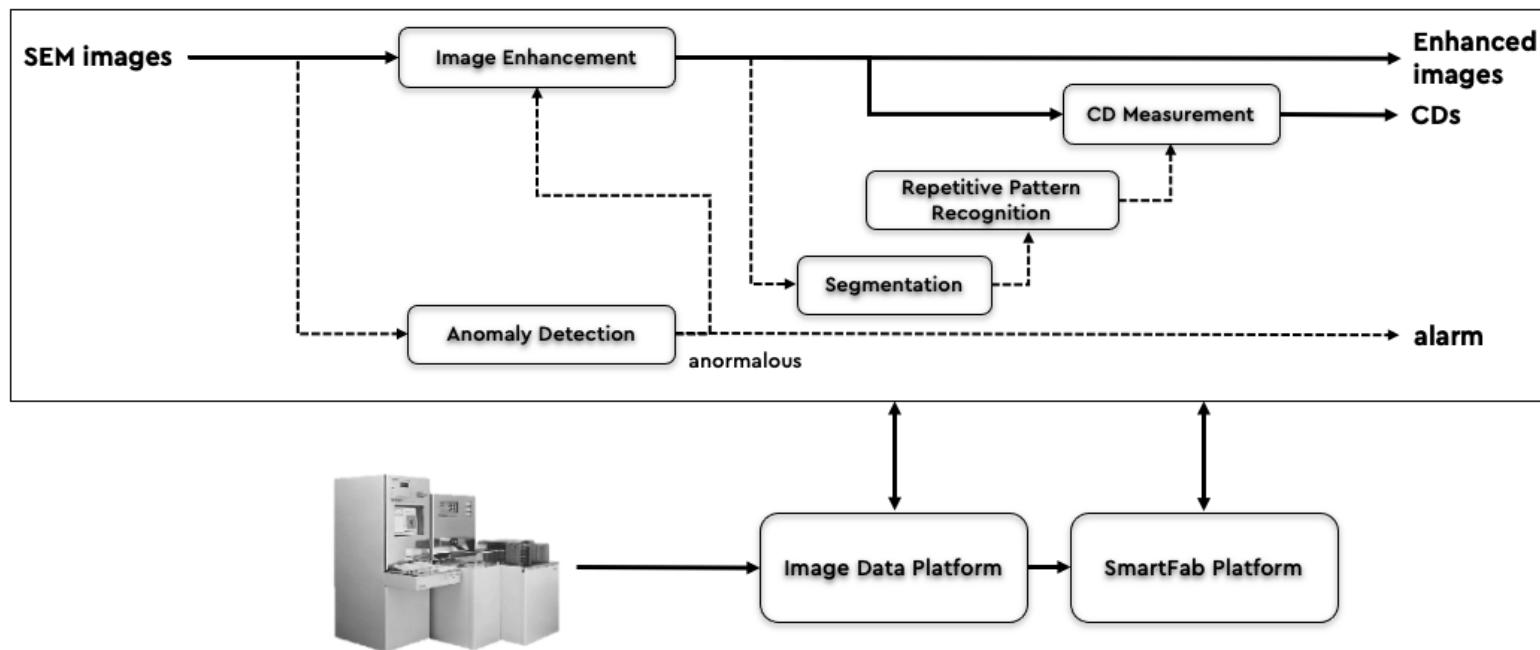
Anomaly detection using side product

- representation in embedding space obtained as side product from previous processes
- distance from normal clusters used for anomaly detection
- can be used for yield drop prediction and analysis



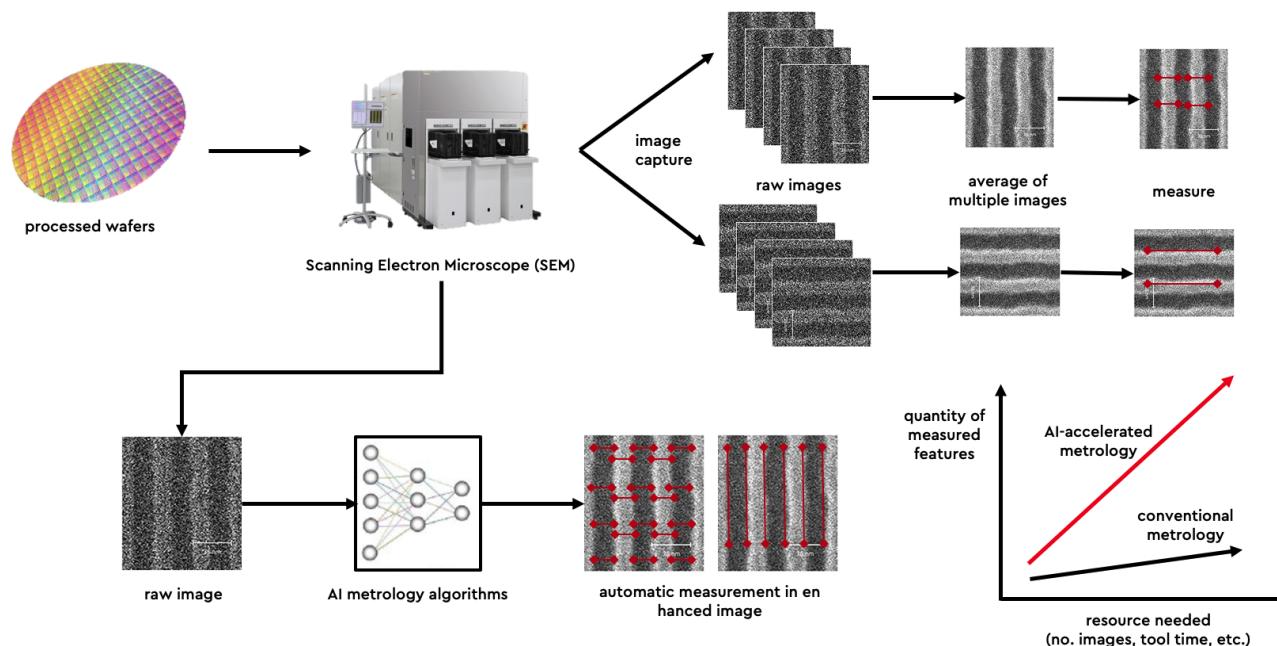
AI-enabled metrology system

- integration of separate components creates AI-enabled metrology system



Benefits of new system

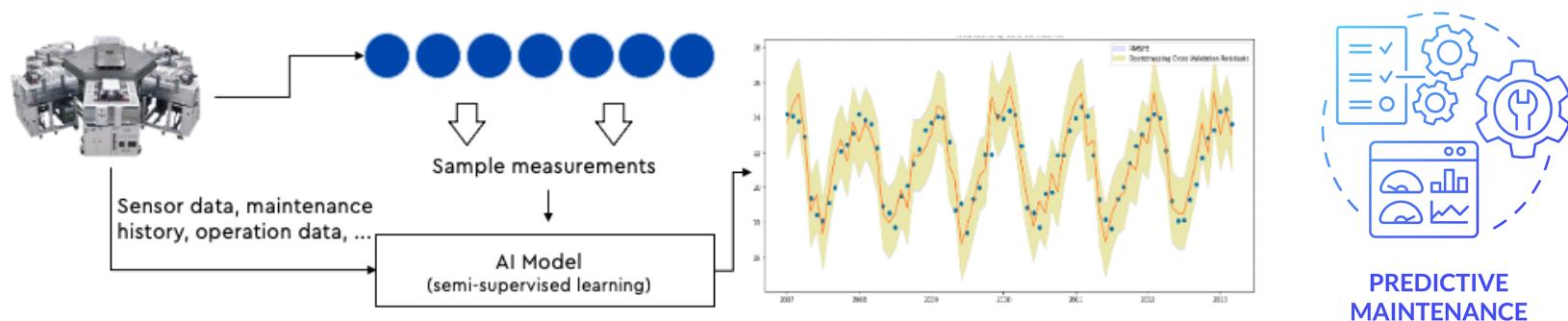
- new system provides
 - improved accuracy and reliability
 - improved throughput
 - savings on investment on measurement equipment



TS ML in manAI

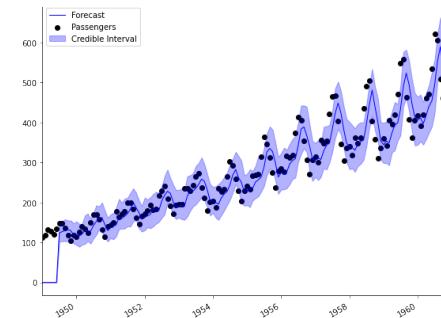
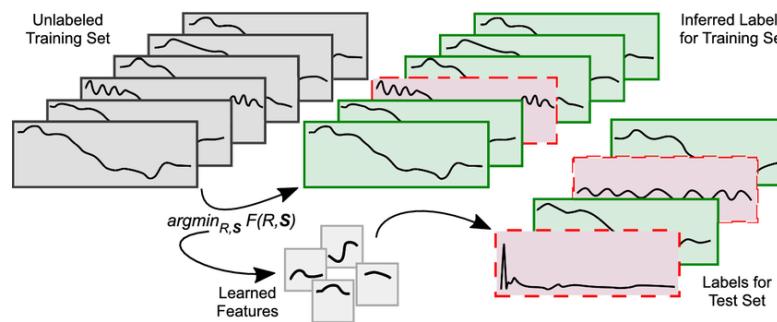
Time-series ML applications in manAI

- estimation of TS values
 - virtual metrology - estimate measurement without physically measuring things
- anomaly detection on TS
 - predictive maintenance - predict maintenance times ahead
- multi-modal ML using LLM & genAI
 - root cause analysis and recommendation system



TS MLs in manAI

- TS regression/prediction/estimation
 - LSTM, GRU, attention-based models, Transformer-based architecture for capturing long-term dependencies and patterns
- anomaly detection
 - isolation forest, autoencoders, one-class SVM
- TS regression providing credibility intervals
 - Bayesian-based approaches offering uncertainty estimation alongside predictions

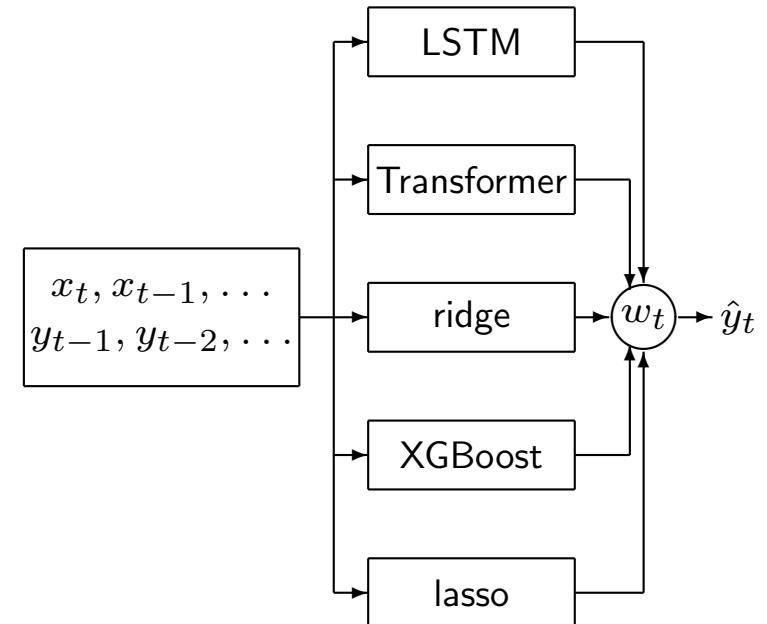


Difficulties with TS ML

- no definition exists for general TS data
- data drift & shift
 - $p(x_{t_k}, x_{t_{k-1}}, \dots)$ changes over time
 - $p(y_{t_k} | x_{t_k}, x_{t_{k-1}}, \dots, y_{t_{k-1}}, y_{t_{k-2}}, \dots)$ changes over time
- (extremely) fat data, poor data quality, huge volume of data to process
- not many research results available
- none of algorithms in academic papers work / no off-the-shelf algorithms work

Online learning for TS regression

- use multiple experts - $f_{1,k}, \dots, f_{p_k,k}$ for each time step $t = t_k$ where $f_{i,k}$ can be any of following
 - seq2seq models (e.g., LSTM, Transformer-based models)
 - non-DL statistical learning models (e.g., online ridge regression)
- model predictor for t_k , $g_k : \mathbf{R}^n \rightarrow \mathbf{R}^m$ as weighted sum of experts



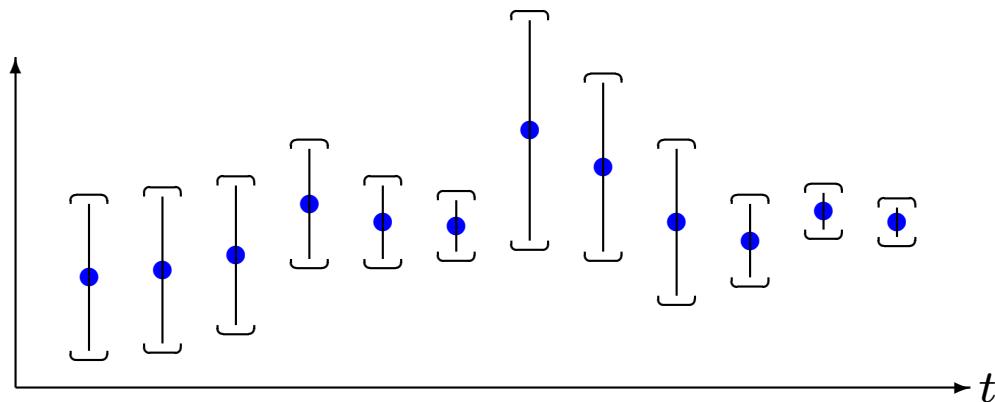
$$g_k = w_{1,k}f_{1,k} + w_{2,k}f_{2,k} + \cdots + w_{p_k,k}f_{p_k,k} = \sum_{i=1}^{p_k} w_{i,k}f_{i,k}$$

Credibility intervals

- every point prediction is wrong, *i.e.*

$$\text{Prob}(\hat{y}_t = y_t) = 0$$

- reliability of prediction matters, however, *none* literature deals with this (properly)
- critical for our customers, *i.e.*, *such information is critical for downstream applications*
 - e.g.*, when used for feedback control, need to know how reliable prediction results are
 - sometimes *more crucial than algorithm accuracy*



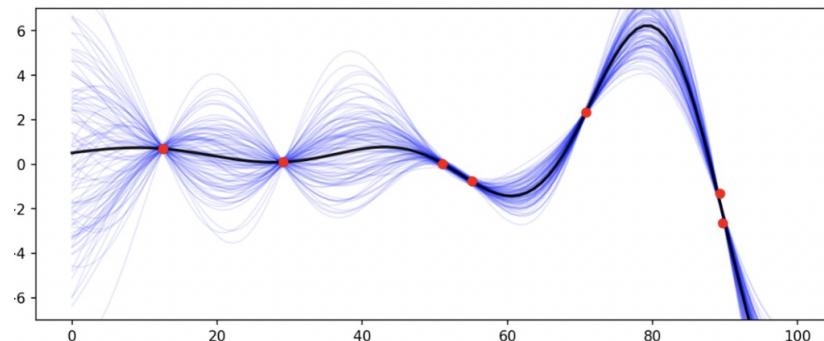
Bayesian approach for credibility interval evaluation

- assume conditional distribution i th predictor parameterized by $\theta_{i,k} \in \Theta$

$$p_{i,k}(y(t_k) | x_{t_k}, x_{t_{k-1}}, \dots, y(t_{k-1}), y(t_{k-2}), \dots) = p_{i,k}(y(t_k); x_{t_k}, \theta_{i,k})$$

- depends on prior & current input, *i.e.*, $\theta_{i,k}$ & x_{t_k}
- update $\theta_{i,k+1}$ from $\theta_{i,k}$ after observing true $y(t_k)$ using Bayesian rule

$$p(w; \theta_{i,k+1}) := p(w | y(t_k); x_{t_k}, \theta_{i,k}) = \frac{p(y(t_k) | w, x_{t_k}) p(w; \theta_{i,k})}{\int p(y(t_k) | w, x_{t_k}) p(w; \theta_{i,k}) dw}$$



Virtual Metrology

VM

- background
 - every process engineer wants to (so badly) measure every material processed - make sure process done as desired
 - *e.g.*, in semiconductor manufacturing, photolithography engineer wants to make sure diameter of holes or line spacing on wafers done correctly to satisfy specification for GPU or memory chips
 - however, various constraints prevent them from doing it, *e.g.*, in semiconductor manufacturing
 - measurement equipment requires investment
 - incur intolerable throughput
 - fab space does not allow
- GOAL - *measure every processed material without physically measuring them*

VM - problem formulation

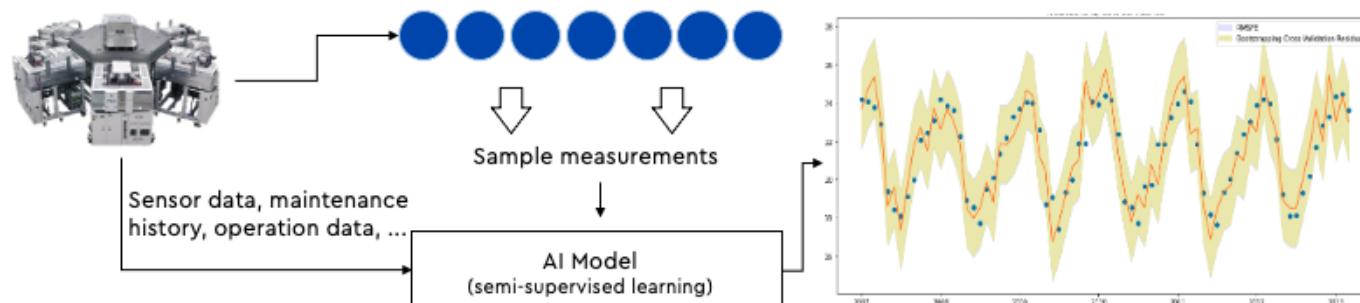
- problem description

(stochastically) predict y_{t_k}
 given $x_{t_k}, x_{t_{k-1}}, \dots, y_{t_{k-1}}, y_{t_{k-2}}, \dots$

- our problem formulation

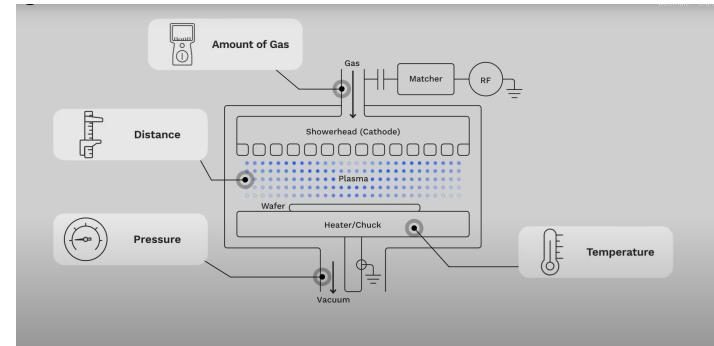
$$\begin{aligned} & \text{minimize} && \sum_{k=1}^K w_{k,K-k} l(y_{t_k}, \hat{y}_{t_k}) \\ & \text{subject to} && \hat{y}_{t_k} = g_k(x_{t_k}, x_{t_{k-1}}, \dots, y_{t_{k-1}}, y_{t_{k-2}}, \dots) \end{aligned}$$

where optimization variables - $g_1, g_2, \dots : \mathcal{D} \rightarrow \mathbf{R}^m$



VM - Gauss Labs' inAI success story

- **Gauss Labs' ML solution & AI product**
 - fully home-grown online TS adaptive ensemble learning method
 - outperform competitors and customer inhouse tools, e.g., *Samsung, Intel, Lam Research*
 - published & patented in US, Europe, and Korea
- business impacts
 - improve process quality - reduction of process variation by tens of percents
 - (indirectly) contribute to better product quality and yield
 - Gauss Labs' main revenue source



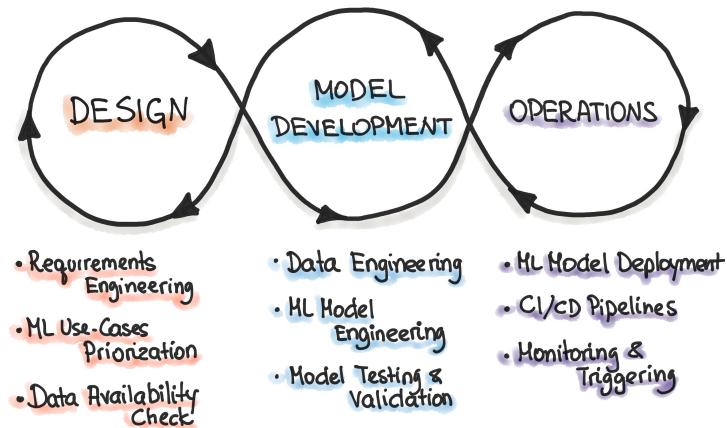
Manufacturing AI Productionization

Minimally required efforts for manAI

- MLOps - for CI/CD
- data preprocessing - missing values, inconsistent names, difference among different systems
- feature extraction & selection
- monitoring & retraining
- notification, via messengers or emails
- mainline merge approvals by humans
- data latency, data reliability, & data availability

MLOps for manAI

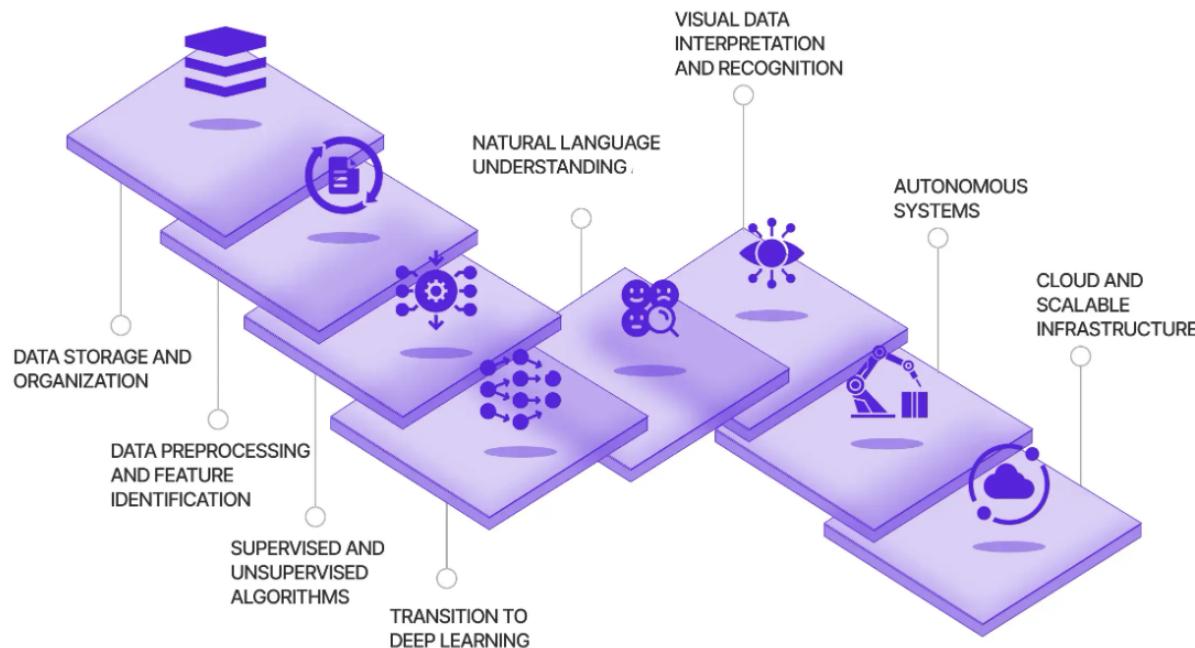
- environment for flexible and agile exploration - EDA³
- fast & efficient iteration of algorithm selection, experiments, & analysis
- correct training / validation / test data sets critical!
- seamless productionization from, e.g., Jupyter notebook to production-ready code
- monitoring, *right* metrics, notification, re-training



³EDA - exploratory data analysis

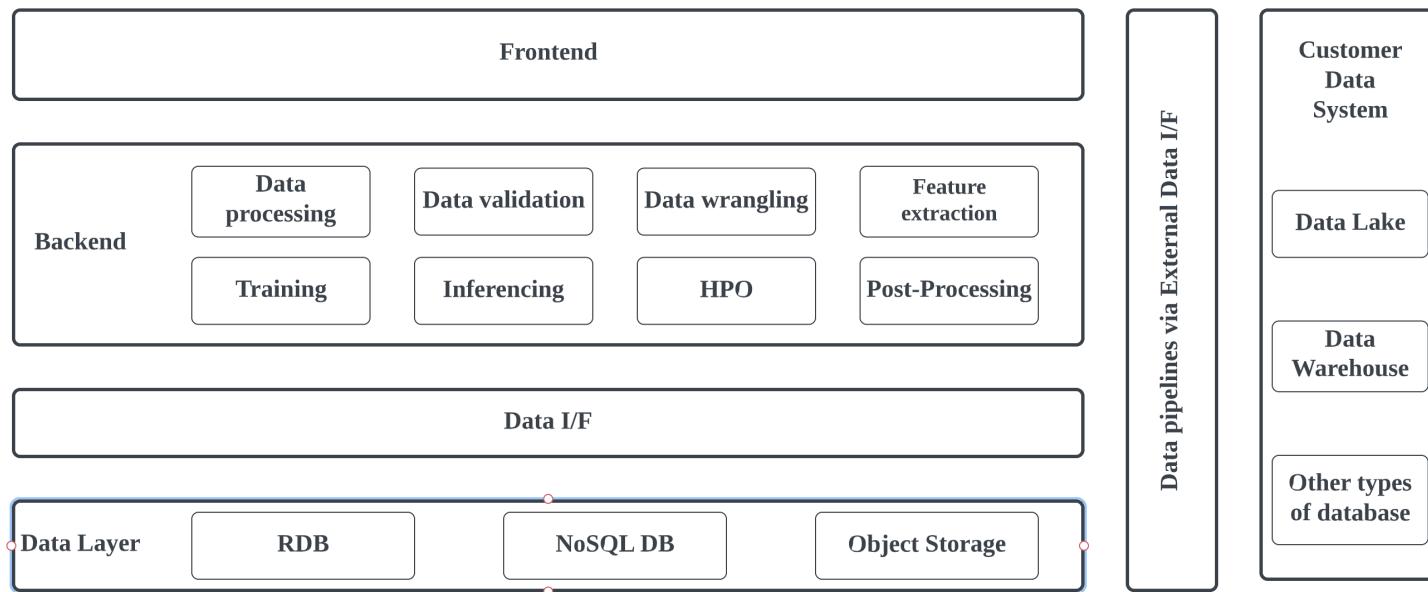
manAI software system

- data, data, data! – store, persist, retrieve, data quality
- seamless pipeline for development, testing, running deployed services
- development environment should be built separately



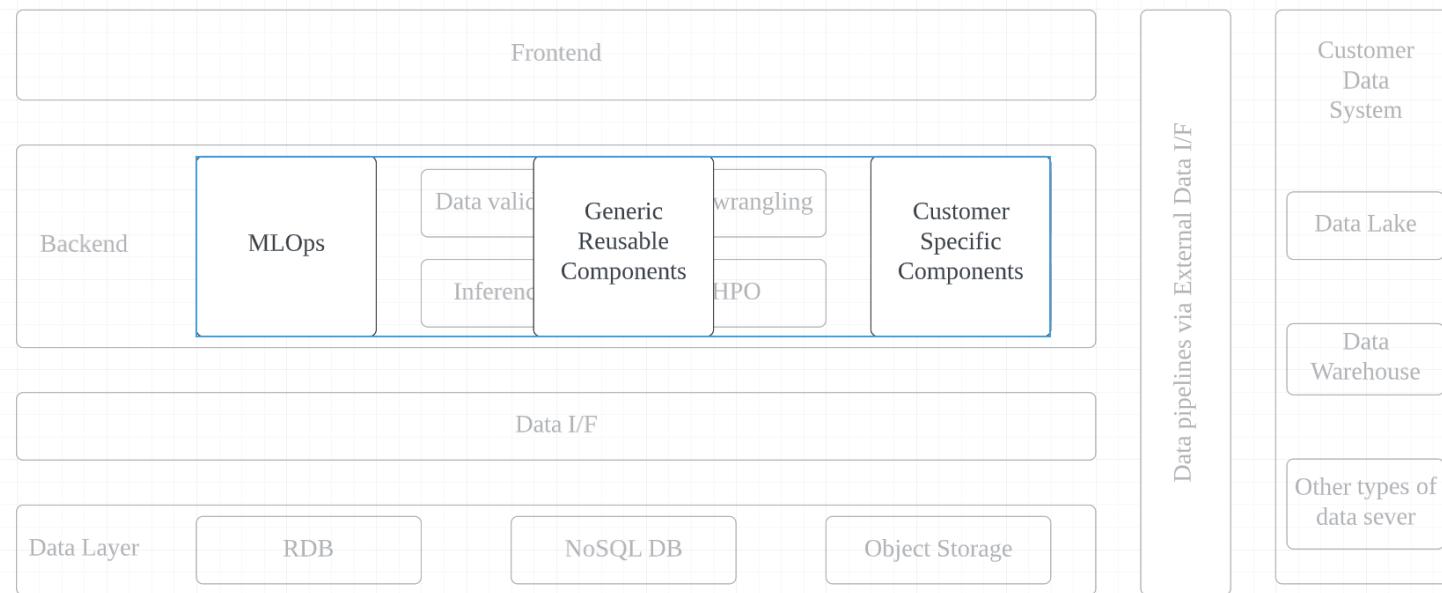
manAI system architecture

- frontend / backend / data I/F / data layer
- efficient and effective MLOps in backend or development environment



Reusable components vs customer specific components

- make sure to build two components separate - generic reusable and customer specific
- generic models should be tuned for each use case
- generic model library grows as interacting with more and more customers



My Two Cents

Recommendations for maximum impact via inAI

- concrete goals of projects
 - north star – yield improvement, process quality, making engineers' lives easier
 - hard problem – scheduling and optimization
- be strategic!
 - learn from others – lots of successes & failures of inAI
 - ball park estimation for ROI crucial – efforts, time, expertise, data
 - utilities vs technical excellency / uniqueness vs common technology
 - home-grown vs off-the-shelf

Remember . . .

- data, data, data! – readiness, quality, procurement, pre-processing, DB
- *never* underestimate domain knowledge & expertise – data do NOT tell you everything
- EDA
- do *not* over-optimize your algorithms – ML is all about trials-&-errors
- overfitting, generalization, concept drift/shift - way more important than you could ever imagine
- devOps, MLOps, agile dev, software development & engineering

Conclusion

Conclusion

- various CV MLs used for inAI applications
- TS ML applications found in every place in manufacturing
- drift/shift & data noise make TS MLs very challenging, but working solutions found
- in reality, crucial bottlenecks are
 - data quality, preprocessing, monitoring, notification, and retraining
 - data latency, availability, and reliability
 - excellency in software platform design and development using cloud services

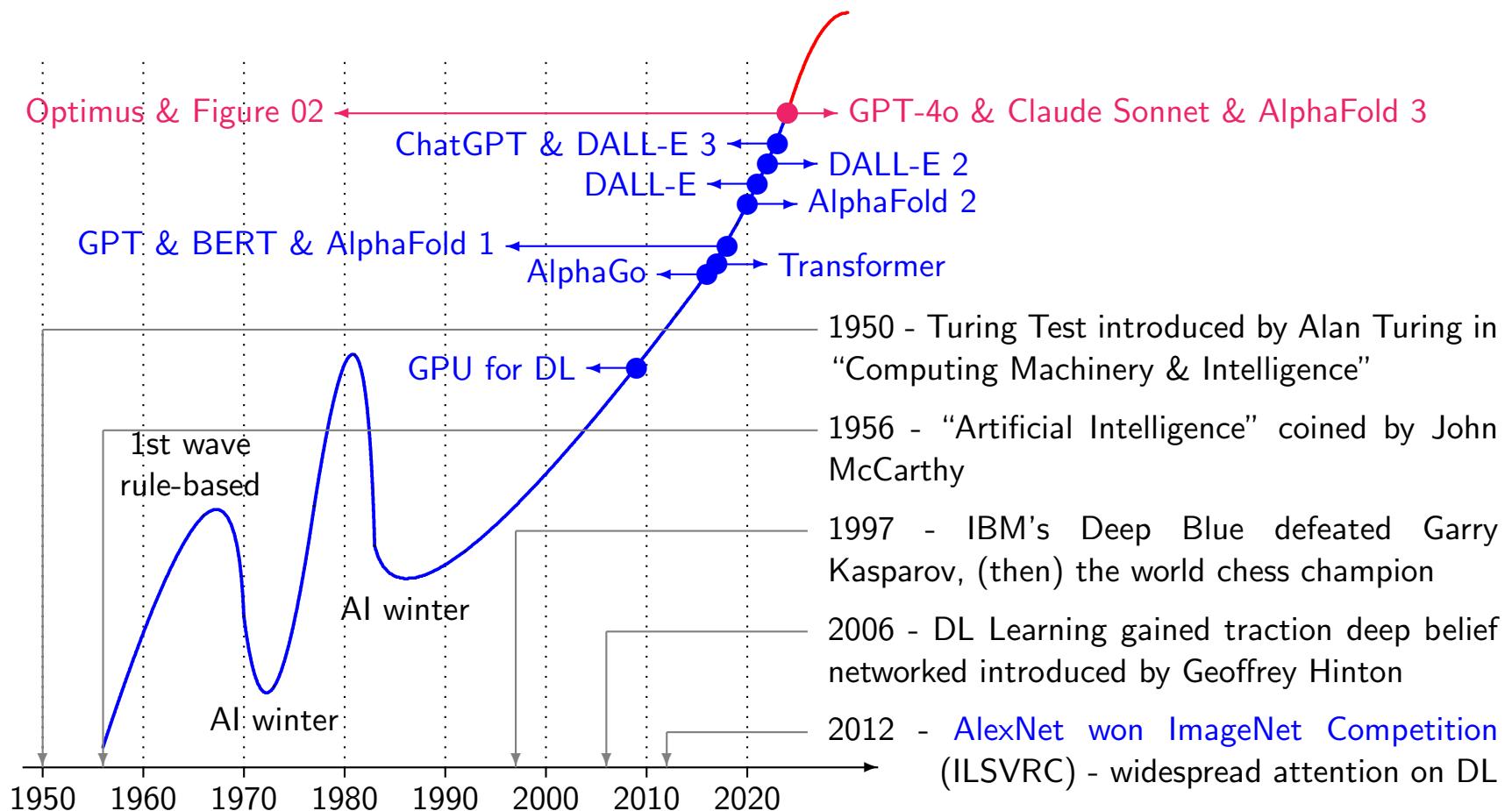
Silicon Valley's Cultural Engine of Innovation and Disruption

My journey from Samsung & Amazon to Gauss Labs & Erudio Bio

- Samsung Semiconductor, Inc.
 - inception into industry from academia, the world's best memory chip maker!
- Amazon.com, Inc.
 - experience so-called Silicon Valley big tech culture and technology
 - set tone for my future career trajectory!
- Gauss Labs, Inc.
 - found & operate AI startup, shaping corporate culture & spearheading R&D as CTO
 - inherent challenges of Korean conglomerate spin-off startup - cultural constraints, over-capitalization, and leadership limitations
- Erudio Bio, Inc.
 - concrete & tangible bio-technology in addition to AI
 - great decisions regarding business development; business models, market fit, go-to-market (GTM) strategies based on lessons learned *in a hard way* ☺

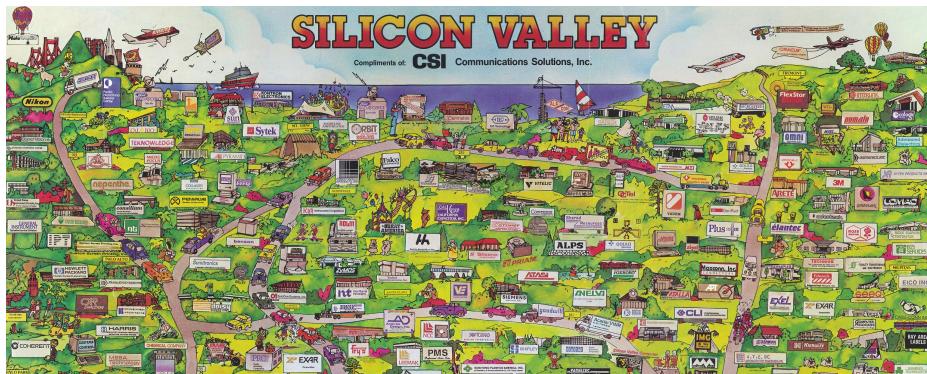


Joining Amazon.com, Inc. at the inflection point of AI



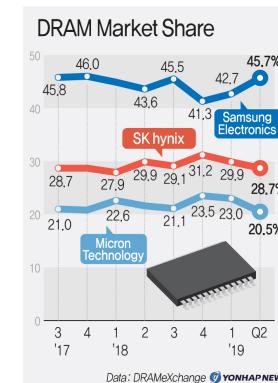
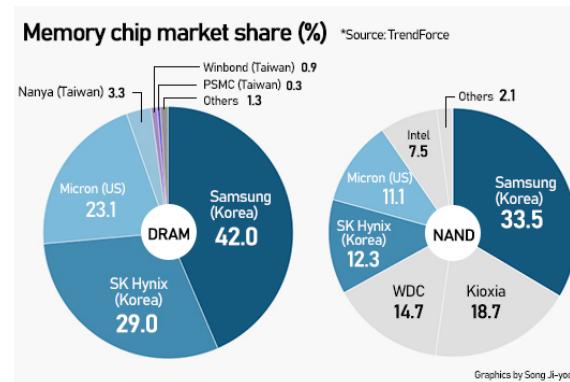
Innovation ecosystem of Silicon Valley

- key characteristics
 - risk-taking culture, *trust* in technology → *genuine* respect for engineers and scientists
 - easy access to huge capital - VCs, angel investors alike
 - talent density - engineers, researchers, scientists, entrepreneurs, PMs, TPMs, . . .
 - diversity, “collision density” of ideas
 - ecosystem of collaboration and competition - startups, academia, industry leaders
- what they mean for global big tech
 - set trends in AI, software & hardware (and or hence) product & industry innovation
 - act as testing ground for disruptive ideas



Case study: Amazon - amazing differentiators of big techs

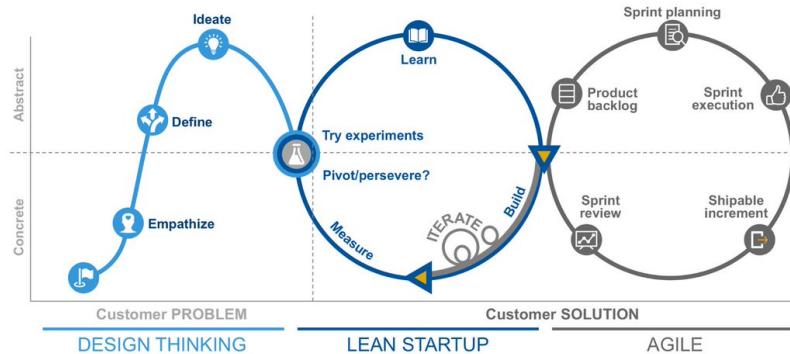
- Amazon's culture & leadership principles
 - customer obsession as driver of innovation
 - high standards & ownership culture, disagree & commit
 - bias for action and long-term thinking - sounds contradictory?
 - mechanisms like “two-pizza teams” & “Day One” for (or rather despite) scalability
- lessons for Korean corporations
 - applying customer-centric innovation in hardware & AI, e.g., on-device AI
 - balancing agility with long-term R&D
 - *build / adapt / apply on the core strength of Samsung that no other company has!*



Founding and scaling startups

- challenges
 - competence of and chemistry among co-founders crucial
 - technology & great team are *necessary*, but *not sufficient (at all!)* for success
 - business models, market fit, timing, agility, flexibility for pivoting / perseverance
- insight
 - importance of domain expertise in addition to AI
 - balancing innovation with good business decisions

Combine Design Thinking, Lean Startup and Agile



Product-Market Fit (PMF)



Bridging Silicon Valley & Korea

- cultural differences
 - risk appetite & failure tolerance
 - decision-making speed vs hierarchy
 - innovation vs execution focus
- opportunities for collaboration
 - leveraging Korea's manufacturing expertise with Silicon Valley's software/AI strengths
 - building global teams with diverse perspectives



To be successful . . .

- embrace customer/market-centric mindset in innovation and for business decisions
- balance agility with long-term vision
- foster cross-cultural collaboration for global impact
- ((very) strategically and carefully) leverage AI to solve real-world industrial challenges



Empowering Humanity for Future Enriched by AI

Blessings & Curses of AI

Blessings

- advancements in healthcare & improved quality of life
 - much faster & more accurate diagnosis, far superior personalized medicine, accelerated drug discovery, assistive technologies
- economic growth & efficiency
 - automation to increase productivity and reduce cost, far superior decision-making
- environmental solutions
 - climate change prediction, global warming effect mitigation, solutions for sustainability
- safety & security
 - natural disaster prediction & relief, cybersecurity



Curses

- job displacement & overall impacts on labor market
 - millions of jobs threatened, wealth gap widened
- bias & inequality, misinformation & manipulation
 - existing human biases, both conscious and unconscious, perpetuated through AIs, asymmetric accessibility to advanced AI technologies by nations & corporations
- ethical dilemmas
 - infringing privacy & human rights, accountability for weapon uses and damages by AI
- environmental costs
 - significant energy for training AI models, waste generated by obsolescent AI hardware



Salzburg Global Seminar

KFAS-Salzburg Global Leadership Initiative

- “Uncertain Futures and Connections Reimagined: Connecting Technologies” - 41 global leaders convened from 4-Dec to 8-Dec, 2024 @ Schloss Leopoldskron in Salzburg, Austria
- My working group was “Technology, Growth, and Inequality: The Case of AI”
 - International Cooperation Officer (Portugal)
 - Gender Equality, Disability Inclusion Consultant, UN Women (Lithuania)
 - Assistant Professor @ Lincoln Alexander School of Law (Canada)
 - Research Associate @ Luxembourg Institute of Socio-Economic Research
 - Policy Officer & Delegation of the EU Union (India)
- blog: [Bridging Technology & Humanity - Reflections from Lyon, Salzburg, and München](#)



Sunghee Yun

July 8, 2025

KFAS-Salzburg Global Leadership Initiative

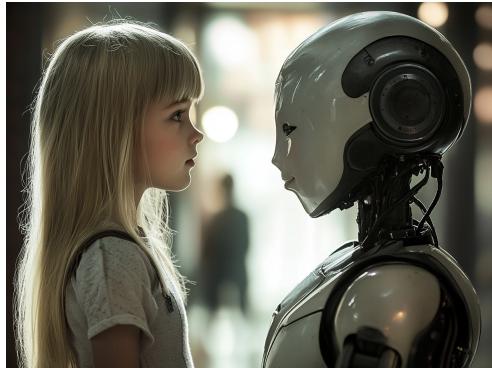
Salzburg Global photo collections



Empowering Humanity

AI capacity building - scientists, engineers & practitioners

- ethics and responsible AI education or campaign via interdisciplinary collaboration
 - foster continuous learning programs on AI risks, bias & societal impacts
- bias detection & mitigation
 - bias-detection tools to identify & reduce discrimination in data & models
 - regular fairness audits
- transparency & explainability
 - explainable AI (xAI) techniques, frameworks like Model Cards for transparency
- environmental impact awareness
 - reduce AI's carbon footprint, advocate for sustainable AI development practices



AI capacity building - lawmakers & policy makers

- problems
 - difficulties in understanding of rapidly evolving AI technologies
 - lead to reactive or insufficient regulation
- proposed solutions
 - develop comprehensive regulatory frameworks addressing transparency, bias & privacy concerns
 - gender bias, racial bias, hallucinations
 - foster public debates on ethical AI use & societal implications
 - introduce policies to limit spread of AI-generated misinformation,



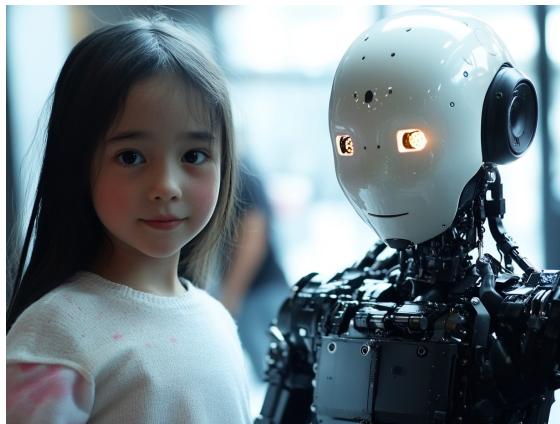
Participatory social agreements

- open data frameworks including data sovereignty, regulation of data transfer, storage & localization
- corporate social responsibility, extra-territorial obligations & environmental protection
 - including outside the jurisdiction of the country
- labour and employment displacements, tax cuts & algorithmic impact assessments
 - including remedies for AI harms and enforcements



Reclaiming technology for Humanity

- strategic approach to AI development
 - *leverage very technologies alienating humans to strengthen human connection*
 - transform automation from replacement to *enhancement of human capabilities*
 - leverage technological scale to address fundamental human needs
- *paradigm shift* in technological implementation
 - recognize the duality of advanced technologies
 - *systematically channel AI capabilities toward human-centric solutions*
 - convert technological challenges into opportunities for human advancement



Some Important Questions around AI

Some important questions around AI

- why human-level AI?
- what lies in very core of DL architecture? what makes it work amazingly well?
- biases that can hurt judgement, decision making, social good?
- AI ethics & legal issues
- consciousness
- utopia vs dystopia
- knowledge, belief, reasoning
- risk of anthropomorphization

Human-level AI?

Why human-level in the first place?

- lots of times, when we measure AI performance, we say
 - how can we achieve human-level performance, *e.g.*, CV models?
- why human-level?
 - are all human traits desirable? are humans flawless?
 - aren't humans still evolving?
- advantage of AI over humans
 - *e.g.*, self-driving cars can use extra eyes, GPS, computer network
 - *e.g.*, recommendation system runs for hundreds of millions of people overnight
 - AI is available 24 / 7 while humans cannot
 - . . . critical advantages for medical assistance, emergency handling
 - AI does not make more mistakes because task is repetitive and tedious
 - AI does not request salary raise or go on strike

What makes DL so successful?

Factors contributing to astonishing success of DL

- analysis based on speaker's mathematical, numerical algorithmic & statistical perspectives considering hardware innovations

30% universal approximation theorem? - (partially) yes! but that's not all

- function space of neural network is *dense* (math theory), *i.e.*, for every $f : \mathbf{R}^n \rightarrow \mathbf{R}^m$, exists $\langle f_n \rangle$ such that $\lim_{n \rightarrow \infty} f_n = f$

25% architectures/algorithms tailored for each class of applications, *e.g.*, CNN, RNN, Transformer, NeRF, diffusion, GAN, VAE, . . .

20% data labeling - expensive, data availability - unlimited web text corpus

15% computation power/parallelism - AI accelerators, *e.g.*, GPU, TPU & NPU

10% rest - Python, open source software, cloud computing, MLOps, . . .

Sudden leap in LLM performance

Probability inferred sequence is correct

- assume

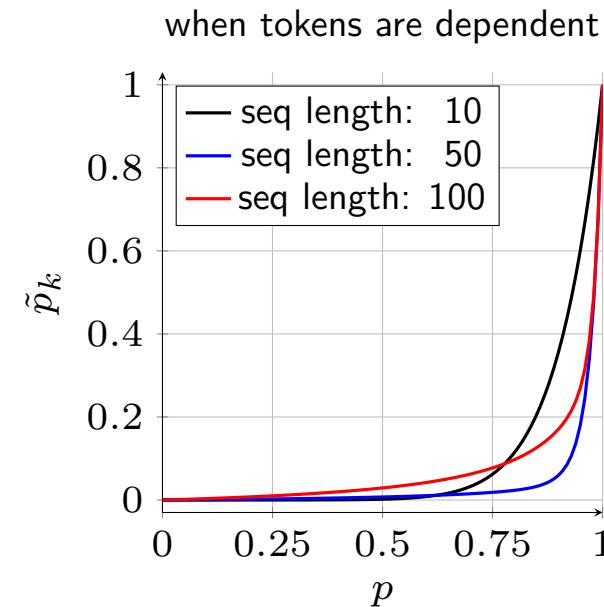
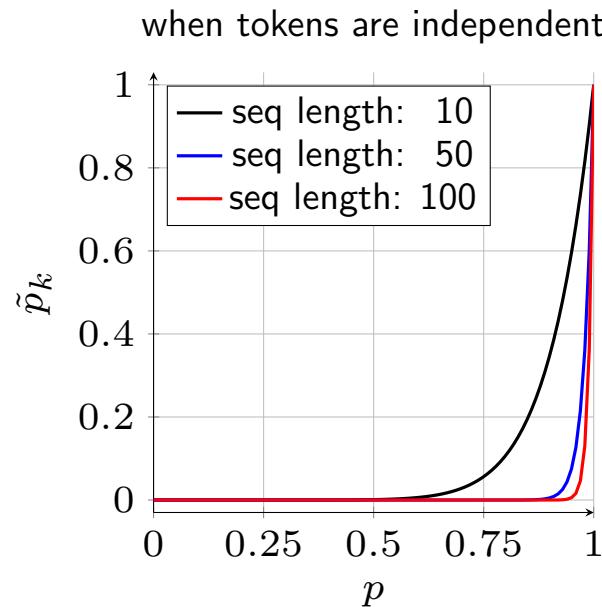
- t_i - i th token
 - p_i - probability that t_i is correct
 - ρ_i - correlation coefficient between t_{i-1} & t_i
 - \tilde{p}_k - probability that (t_1, \dots, t_k) are correct

- recursion

$$\rho_i = \frac{\tilde{p}_i - \tilde{p}_{i-1}p_i}{\sqrt{\tilde{p}_{i-1}(1 - \tilde{p}_{i-1})p_i(1 - p_i)}}$$
$$\Leftrightarrow \tilde{p}_i = \tilde{p}_{i-1}p_i + \rho_i \sqrt{\tilde{p}_{i-1}(1 - \tilde{p}_{i-1})p_i(1 - p_i)}$$

Dramatic improvement of LLM near saturation

- do simulations for both independent & dependent cases
 - assume p_i are same for all i
- (for both cases) sequence inference improves dramatically as p approaches 1
- this explains *why we have observed sudden dramatic performance improvement of certain seq2seq learning technologies*, e.g., LLM



Biases

Cognitive biases attributed to humans

- cognitive biases [Kah11]
 - confirmation bias, availability bias
 - hindsight bias, confidence bias, optimistic bias
 - anchoring bias, halo effect, framing effect, outcome bias
 - belief bias, negativity bias, false consensus



Biases of LLMs

- LLMs subject to
 - availability bias - biased by imbalancedly available information
 - LLM trained by imbalanced # articles for specific topics
 - belief bias - derive conclusion not by reasoning, but by what it saw
 - LLM easily inferring what it saw, *i.e.*, data it trained on
 - halo effect - overemphasize on what prestigious figures say
 - LLM trained by imbalanced # reports about prestigious figures
- similar facts true for other types of ML models,
 - *e.g.*, video caption, text summarization, sentiment analysis
- cognitive biases only humans represent
 - confirmation bias, hindsight bias, confidence bias, optimistic bias, anchoring bias, negativity bias, framing effect

AI Ethics

Ethical issues related to AI

- AI can be exploited by those who have bad intention to
 - manipulate / deceive people - using manipulated data corpus for training
 - *e.g.*, spread false facts
 - induce unfair social resource allocation
 - *e.g.*, medical insurance, taxation
 - exploit advantageous social and economic power
 - *e.g.*, unfair wealth allocation, mislead public opinion
- AI for Good - advocated by Andrew Ng
 - *e.g.*, public health, climate change, disaster management
- should scientists and engineers be morally & politically conscious?
 - *e.g.*, Manhattan project

AI related Legal Issues

Legal issues with ethical consideration

- scenario 1 - full self-driving algorithm causes traffic accident killing people
 - who is responsible? - car maker, algorithm developer, driver, algorithm itself?
- scenario 2 - self-driving cars kill less people than human drivers
 - *e.g.*, human drivers kill 1.5 people for 100,000 miles & self-driving cars kill 0.2 people for 100,000 miles
 - how should law makers make regulations?
 - utilitarian & humanitarian perspectives
- scenario 3 - someone is not happy with their data being used for training
 - “The Times sues OpenAI and Microsoft over AI use of copyrighted work” (Dec-2023)
 - “Newspaper publishers in California, Colorado, Illinois, Florida, Minnesota and New York said Microsoft and OpenAI used millions of articles without payment or permission to develop ChatGPT and other products” (Apr-2024)

Consciousness

Consciousness

- what is consciousness, anyway?
 - recognizes itself as independent, autonomous, valuable entity?
 - recognizes itself as living being, unchangeable entity?
- no agreed definition on consciousness exists yet
 - . . . and will be so forever
- does it have anything to do with the fact that humans are biologically living being?
- is SKYNET ever plausible?
 - can AI have *desire* to survive (or save earth)?



Utopia vs Dystopia

Utopia vs dystopia



- not important questions (at all) *I think . . .*
- what we should focus on is *not* the possibilities of doomsday or Judgment Day, but rather
 - our limits on controlling unintended impacts of AI
 - *misuse* by (greedy, immoral, and unethical) people possessing social, economic & political power
 - *social good and welfare impaired* by either exploiting AI or ignorance of (inner workings of) AI
- should concern
 - choice or balance among utilitarianism, humanitarianism & values
 - amend or improve laws/regulations
 - ethical issues caused by AI

Knowledge, Belief, and Reasoning

Does AI (LLM) have knowledge or belief? Can it reason?

**What categories of questions do they belong to?
engineering, scientific, philosophical, cognitive scientific, . . . ?**

LLMs . . .

- LLM is very different sort of animal . . . except that it is *not* an animal!
- *unreasonable* effectiveness of data [HNF09]
 - *performance scales with size of training data*
 - *qualitative leaps* in capability as models scale
 - tasks demanding human intelligence *reduced to next token prediction*
- focus on third surprise

conditional probability model looks like human with intelligence

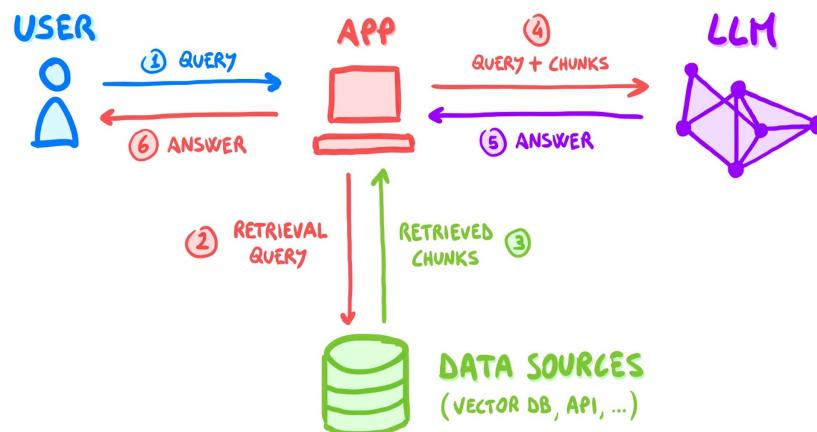
- making vulnerable to anthropomorphism
- examine it by throwing questions such as
 - “*does LLM have knowledge and belief?*”
 - “*can it reason?*”

What LLM really does!

- given prompt “the first person to walk on the Moon was”, LLM responds with “Neil Armstrong”. . . strictly speaking
 - it’s *not* being asked *who* was the first person to walk on the Moon
 - what are being *really* asked is “*given statistical distribution of words in vast public corpus of text, what words are most likely to follow ‘The first person to walk on the Moon was’?*”
- given prompt “after ring was destroyed, Frodo Baggins returned to”, LLM responds with “the Shire”
 - on one level, it seems fair to say, you might be testing LLM’s knowledge of fictional world of Tolkien’s novels
 - what are being *really* asked is “*given statistical distribution of words in vast public corpus of text, what words are most likely to follow ‘After the ring was destroyed, Frodo Baggins returned to’?*”

LLMs vs systems in which they are embedded

- crucial to distinguish between the two (for philosophical clarity)
 - LLM (bare-bones model) - highly specific & well-defined function, which is *conditional probability estimator*
 - systems in which LLMs are embedded, *e.g.*, for question-answering, news article summarization, screenplays generation, language translation



How ChatBot works?

- conversational AI agent does *in-context learning* or *few-shot prompting*
- for example,

- when the user enters

- who is the first person to walk on the Moon?

- ChatBot, LLM-embedded system, feeds the following to LLM

- User, a human, and BOT, a clever and knowledgeable AI agent.

- User: what is 2+2?

- BOT: the answer is 4.

- User: where was Albert Einstein born?

- BOT: he was born in Germany.

- User: who is the first person to walk on the Moon?

- BOT:

Knowledge, belief & reasoning around LLM

- *not* easy topic to discuss, or even impossible because
 - we *do not have agreed definition* of these terms especially in context of being asked questions like

does LLM have belief?
or
do humans have knowledge?
- let us discuss them in two different perspectives
 - laymen's perspectives
 - cognitive scientific & philosophical perspectives

Laymen's perspectives on knowledge, belief & reasoning

- does (good) LLM have knowledge?
 - Grandmother: looks like it cuz when instructed “*explaining big bang*”, it says
“*The Big Bang theory is prevailing cosmological model that explains the origin and evolution of the universe. . . . 13.8 billion years ago . . .*”
- does it have belief?
 - Grandmother: I don't think so, e.g., it does not believe in God!
- can it reason?
 - Grandmother: seems like it! e.g., when asked “*Sunghee is a superset of Alice and Beth is a superset of Sunghee. Is Beth a superset of Alice?*”, it says
“*Yes, based on information provided, if Sunghee is a superset of Alice and Beth is a superset of Sunghee, then Beth is indeed a superset of Alice . . .*”
- can it reason to prove theorem whose inferential structure is more complicated?
 - Grandmother: I'm not sure – actually, I don't know what you're talking about!

Knowledge

- could argue LLM “knows” which words follow which other words with high probability
- but, only *in context of capacity to distinguish truth from falsehood* can we legitimately speak of “knowledge” !
- LLM(-embedded BOT)
 - can be said to “*encode*”, “*store*”, or “*contain*” knowledge
 - lacks means to use words “true” & “false” in all ways & in all contexts because . . .
 - *does not inhabit the world* we human language-users share!



Belief

- nothing can count as *belief about the world* we share unless
 - is against backdrop of “*ability to update beliefs appropriately in light of evidence from that world*” - (again) essential capacity to distinguish truth from falsehood
- change taking place in humans when acquiring or updating belief is
 - reflection of their nature as language-using animals inhabiting shared world with community of language-users
- then, *what if LLM-embedded system updates LLM with outside world information?*
 - even so, when interacting with AI systems based on LLMs, these grounds are *absent!*



can you provide me up-to-date information about AI agent perplexity?

Sources

- Perplexity AI: A Guide for Beginners - Jeff Su
- Perplexity AI's value to our AI team as it realizes new \$800K funding - siliconangle
- Perplexity At A Deep Dive - Reflections

Perplexity

Perplexity AI is an AI-powered search engine and answer engine that has gained significant traction in recent years. Founded in 2022, Perplexity uses large language models (LLMs) to generate answers to user queries, citing sources from the web within the text response. As of Q1 2024, Perplexity had reached 15 million monthly users.

Key features and developments of Perplexity AI include:

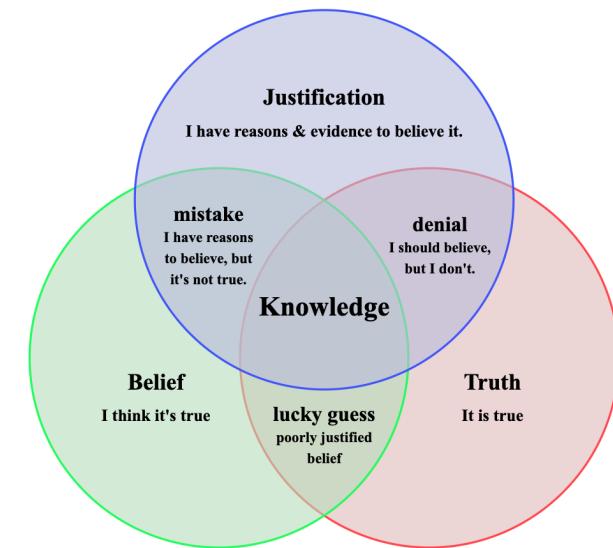
- Search functionality: Perplexity summarizes search results and produces text with inline citations, providing personalized results based on the context of user queries

Ask follow-up



Knowledge in philosophical and cognitive scientific sense

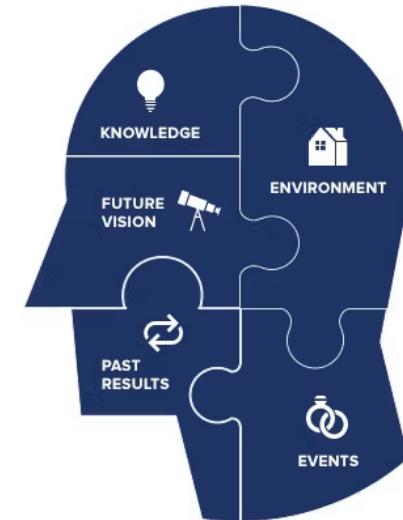
- does LLM have knowledge?
 - Sunghee: *I don't think so!*
- why?
 - we say we have “knowledge” when
“we do so against ground of various human capacities that we all take for granted when we engage in everyday conversation with each other.”
 - when asked *“who is Tom Cruise’s mother?”*, it says *“Tom Cruise’s mother is Mary Lee Pfeiffer.”*
 However, this is nothing but
“guessing” by conditional probability model the most likely words following “Tom Cruise’s mother is.”
 - so *we cannot say it really knows the fact!*



Belief in philosophical and cognitive scientific sense

- for the discussion
 - do *not* concern any specific belief
 - but concern *prerequisites for ascribing any beliefs to AI system*
- so does it have belief?
 - nothing can count as belief about the world we share unless
 - it is against ground of the ability to update beliefs appropriately in light of evidence from that world, essential aspect of the capacity to distinguish truth from falsehood*
 - LLM does not have this ground, essential consideration when deciding whether it *really* had beliefs.
- Sunghee: so *no, LLM cannot have belief!*

WHERE DO YOUR
BELIEFS COME FROM?



Reasoning in philosophical and cognitive scientific sense

- note reasoning is *content neutral*
 - e.g., following logic is perfect regardless of truth of premises
 - hence, no access to outside world does *not* disqualify
- when asked “*if humans are immortal, would Socrates have survived today?*”, LLM says “*... it's logical to conclude that Socrates would likely still be alive today. ...*”
- however, remember, once again, what we just asked it to do is *not* “deductive inference”
given the statistical distribution of words in public corpus, what words are likely to follow the sequence, “humans are immortal and Socrates is human therefore.”
- Sunghee: so *no, LLM cannot reason, either!*
- but, LLM
 - pretends to reason, and from which capabilities, we can benefit!
 - also, can *mimic even multi-step reasoning whose inferencing structure is complicated* using *chain-of-thoughts prompting*, i.e., *in-context learning* or *few-short prompting*

Simple example showing LLM not possessing knowledge



- User
"Who is Tom Cruise's mother?"
- LLM(-embedded question-answering system) (as of Jan 2022)
"Tom Cruise's mother is Mary Lee Pfeiffer. She was born Mary Lee South. . . . Information about his family, including his parents, has been publicly available, . . . "
- User
"Who is Mary Lee Pfeiffer's son?"
- LLM(-embedded question-answering system) (as of Jan 2022)
"As of my last knowledge update in January 2022, I don't have specific information about Mary Lee Pfeiffer or her family, including her son. . . . "

Risk of anthropomorphization

- unfortunately, contemporary LLMs are *too powerful, too versatile, and too useful for most people to accept (after understanding) previous arguments!*
- maybe, o.k. for laymen to (mistakenly) anthropomorphize LLM(-embedded systems)
- however, *imperative for (important, smart, and responsible) AI researchers, scientists, engineers & practitioners* to have rigorous understanding in these aspects especially when
 - advise and be consulted by law makers, policy makers, journalists, and various stakeholders responsible for *critical business decisions (in private sectors) and public policies (in public sectors)*
 - collaborate with or/and help professionals in liberal arts, such as *philosophy, ethics, law, religion, literature, history, music, cultural studies, psychology, sociology, anthropology, political science, economics, archaeology, linguistics, media studies, natural sciences, fine arts, . . .*
 - to address negative societal and economic impacts

Moral

- AI shows incredible utility and commercial potentials, hence should
 - make informed decisions about trustworthiness and safety
 - avoid ascribing capacities they lack
 - *take best utilization of remarkable capabilities of AI*
- today's AI so powerful, so (seemingly) convincingly intelligent
 - obfuscate mechanism
 - actively encourage *anthropomorphism* with philosophically loaded words like “*believe*” and “*think*”
 - easily mislead people about character and capabilities of AI
- matters not only to scientists, engineers, developers, and entrepreneurs, but also
 - *general public, law & policy makers, journalists, . . .*

Recent AI Development

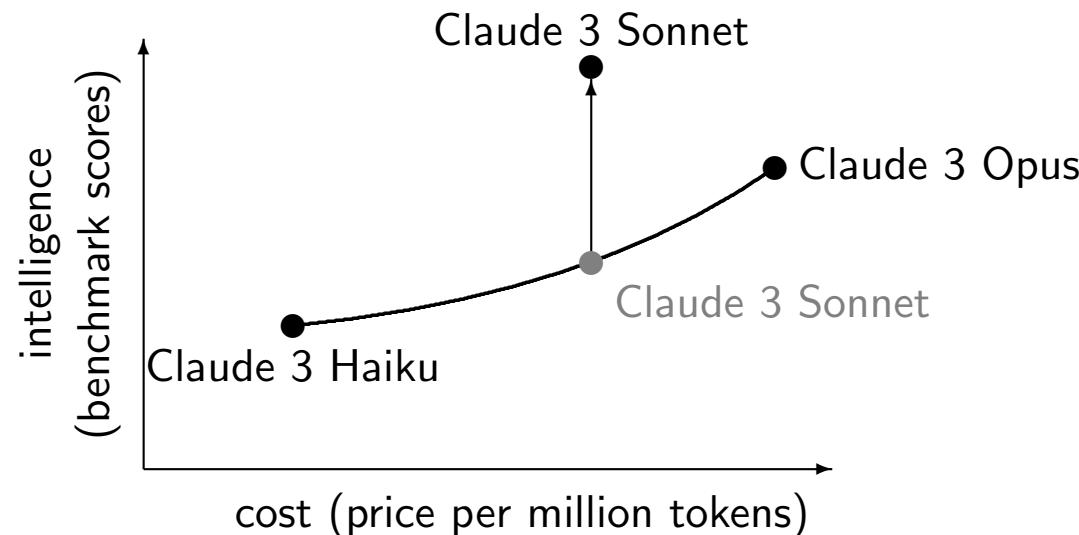
Notable recent AI research and new development

- Claude 3.5 Sonnet
- Kolmogorov–Arnold networks (KAN)
- JEPA (*e.g.*, I-JEPA & V-JEPA) & consistency-diversity-realism trade-off

Claude 3.5 Sonnet

Claude 3.5 Sonnet

- Anthropic
 - releases Claude 3.5 Sonnet (Jul-2024)
 - when! GPT-4o accepted to be default best model for many tasks, *e.g.*, reasoning & summarization
 - claims Claude 3.5 Sonnet sets *new industry standard for intelligence*



Main features & performance

- Claude 3.5 Sonnet shows off
 - improved vision tasks, 2x speed (compared to GPT-4o), artifacts - new UIs for, *e.g.*, code generation & animation
- with GPT-4o, Claude 3.5 Sonnet
 - wins at code generation
 - on par for logical reasoning
 - loses at logical reasoning
 - *wins at generation speed*

	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro
visual math reasoning	67.7%	50.5%	63.8%	63.9%
science diagrams	94.7%	88.1%	94.2%	94.4%
visual question answering	68.3%	59.4%	69.1%	62.2%
chart Q&A	90.8%	80.8%	85.7%	87.2%
document visual Q&A	95.2%	89.3%	92.8%	93.1%

KAN

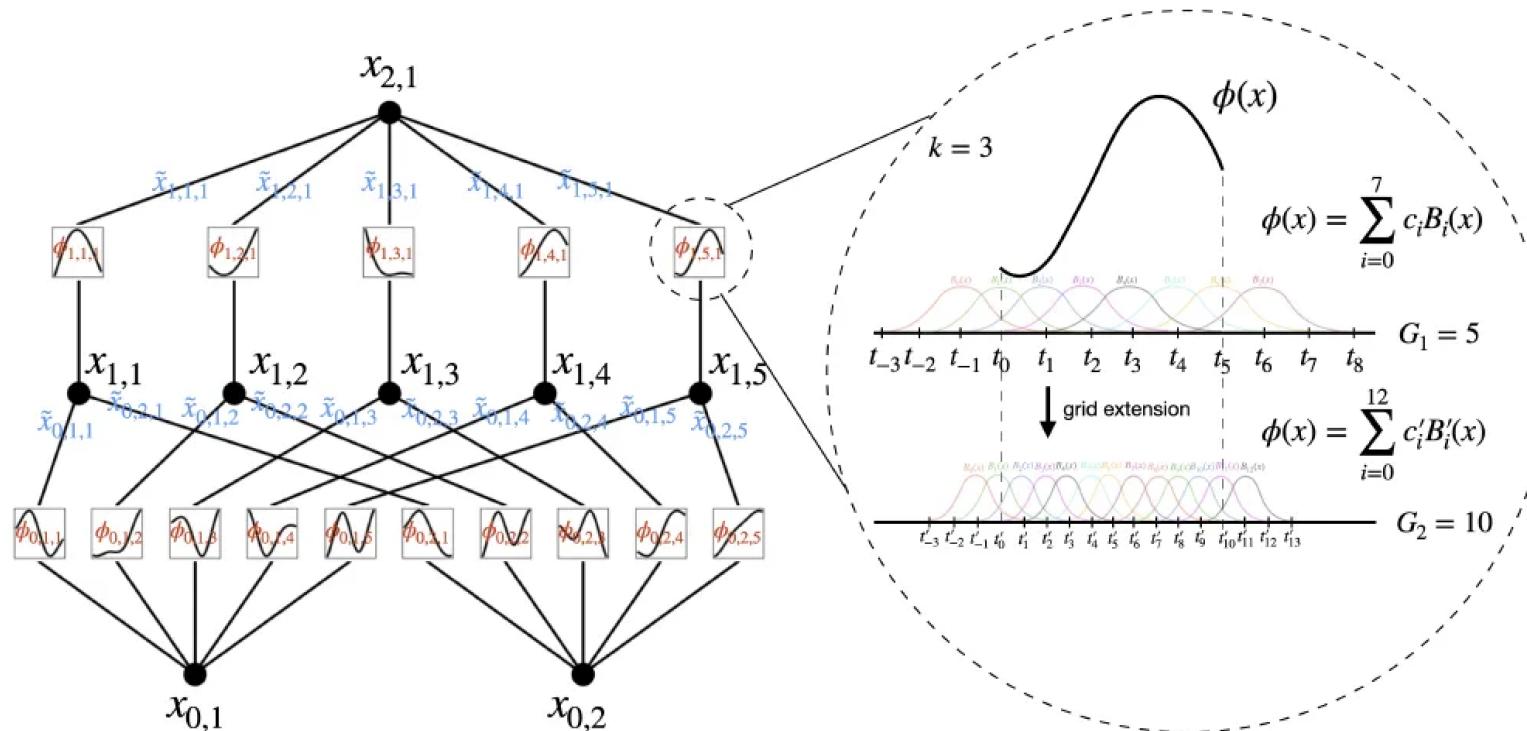
Kolmogorov–Arnold networks (KAN)

- KAN: Kolmogorov-Arnold Networks - MIT, CalTech, Northeastern Univ. & IAIFI
- techniques
 - inspired by Kolmogorov-Arnold representation theorem - every $f : \mathbf{R}^n \rightarrow \mathbf{R}$ can be written as finite composition of continuous functions of single variable, i.e.
$$f(x) = \sum_{q=0}^{2n} \Phi_q \left(\sum_{p=1}^n \phi_{q,p}(x_p) \right)$$
where $\phi_{q,p} : [0, 1] \rightarrow \mathbf{R}$ & $\Phi_q : \mathbf{R} \rightarrow \mathbf{R}$
 - replace (fixed) activation functions with learnable functions
 - use B-splines for learnable (uni-variate) functions - for flexibility & adaptability
- advantages
 - benefits structure of MLP on outside & splines on inside
 - reduce complexity and # parameters to achieve accurate modeling
 - *interpretable* by its nature
 - *better continual learning* - adapt to new data without forgetting thanks to local nature of spline functions

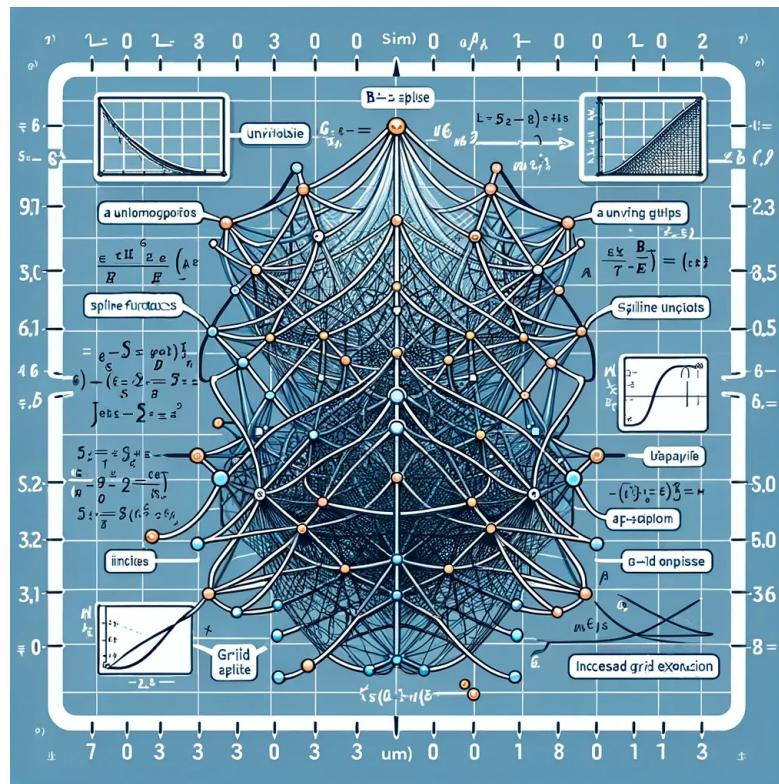
MLP vs KAN

Model	Multi-Layer Perceptron (MLP)	Kolmogorov-Arnold Network (KAN)
Theorem	Universal Approximation Theorem	Kolmogorov-Arnold Representation Theorem
Formula (Shallow)	$f(\mathbf{x}) \approx \sum_{i=1}^{N(e)} a_i \sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i)$	$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^n \phi_{q,p}(x_p) \right)$
Model (Shallow)	<p>(a)</p> <p>fixed activation functions on nodes</p> <p>learnable weights on edges</p>	<p>(b)</p> <p>learnable activation functions on edges</p> <p>sum operation on nodes</p>
Formula (Deep)	$\text{MLP}(\mathbf{x}) = (\mathbf{W}_3 \circ \sigma_2 \circ \mathbf{W}_2 \circ \sigma_1 \circ \mathbf{W}_1)(\mathbf{x})$	$\text{KAN}(\mathbf{x}) = (\Phi_3 \circ \Phi_2 \circ \Phi_1)(\mathbf{x})$
Model (Deep)	<p>(c)</p> <p>MLP(\mathbf{x})</p> <p>\mathbf{W}_3</p> <p>σ_2</p> <p>\mathbf{W}_2</p> <p>σ_1</p> <p>\mathbf{W}_1</p> <p>\mathbf{x}</p> <p>nonlinear; fixed</p> <p>linear; learnable</p>	<p>(d)</p> <p>KAN(\mathbf{x})</p> <p>Φ_3</p> <p>Φ_2</p> <p>Φ_1</p> <p>\mathbf{x}</p> <p>nonlinear; learnable</p>

KAN architecture with spline parametrization unit layer



Future work on KAN



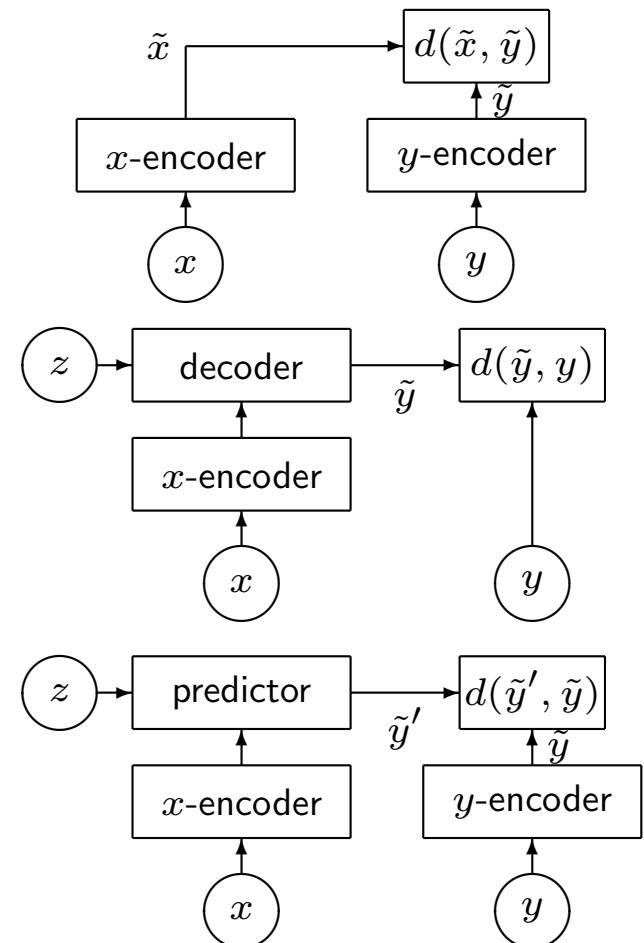
- natural question is
 - what if use both MLP and KAN?
 - what if use other types of splines?
 - how to control forgetfulness of continual learning?
 - why functions of one variable? possible to use functions of two variables?

(figure created by DALLE-3)

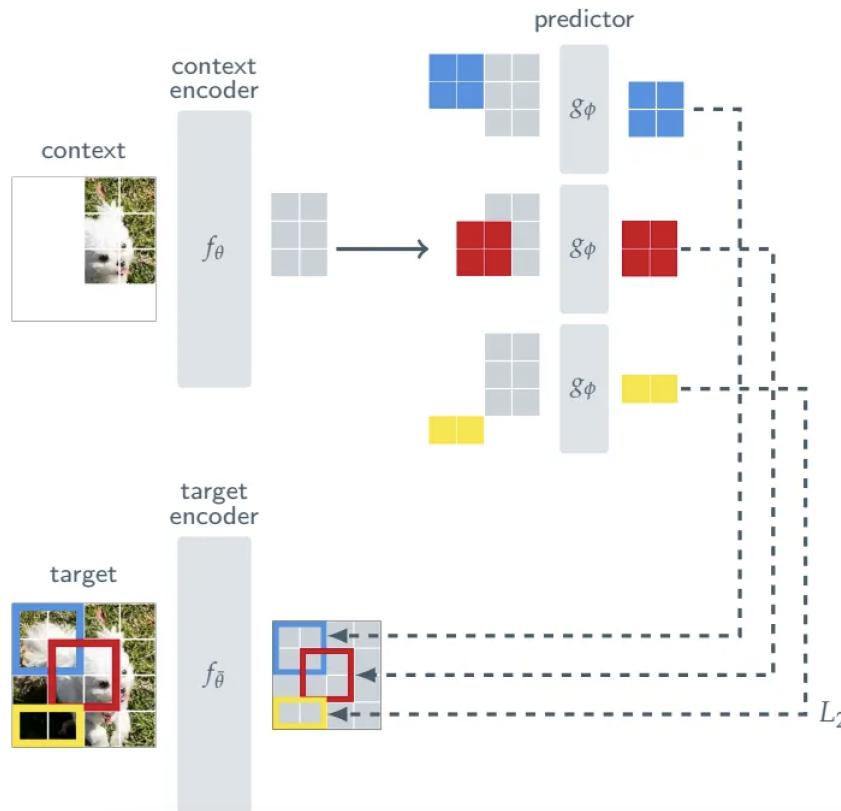
JEPA

Joint-Embedding Predictive Architecture (JEPA)

- Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture (JEPA) - Yann LeCun et al. - Jan-2023
 - joint-embedding architecture (JEA)
 - output similar embeddings for compatible inputs x, y and dissimilar embeddings for incompatible inputs
 - generative architecture
 - directly reconstruct signal y from compatible signal x using decoder network conditioned on additional variables z to facilitate reconstruction
 - joint-embedding predictive architecture (JEPA)
 - similar to generative architecture, but comparison is done in embedding space
 - e.g., I-JEPA learns y (masked portion) from x (unmasked portion) conditioned on z (position of mask)



Learning semantic representation better



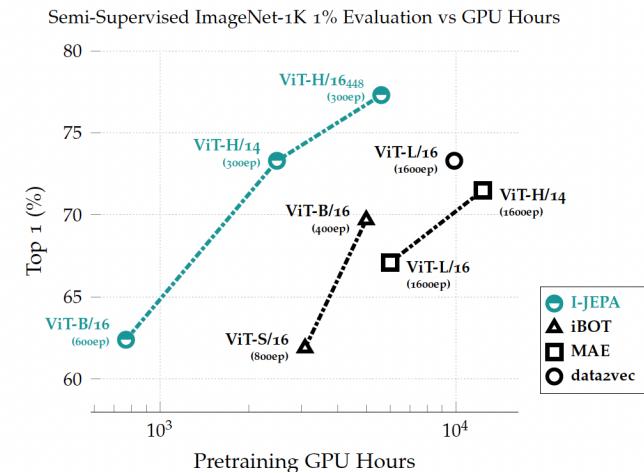
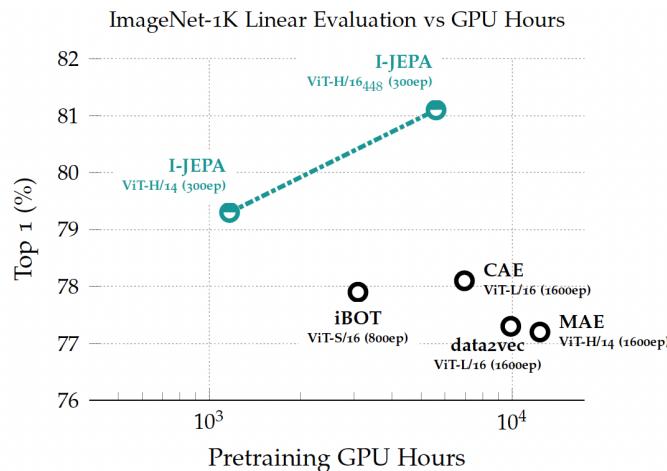
- I-JEPA

- predicts missing information in *abstract representation space*
- e.g., given single context block (unmasked part of the image), predict representations of various target blocks (masked regions of same image) where target representations computed by learned target-encoder
- *generates semantic representations* (not pixel-wise information) potentially eliminating unnecessary pixel-level details & allowing model to concentrate on learning more semantic features

I-JEPA outperforms other algorithms

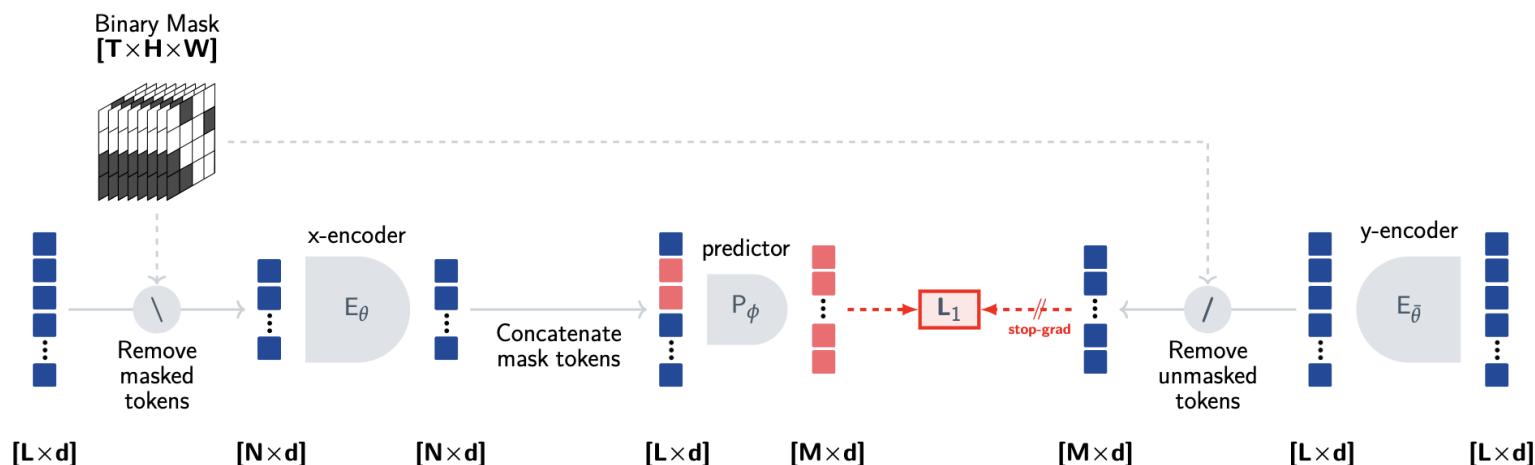
Method	Arch.	CIFAR100	Places205	iNat18
<i>Methods without view data augmentations</i>				
data2vec [8]	ViT-L/16	81.6	54.6	28.1
MAE [36]	ViT-H/14	77.3	55.0	32.9
I-JEPA	ViT-H/14	87.5	58.4	47.6
<i>Methods using extra view data augmentations</i>				
DINO [18]	ViT-B/8	84.9	57.9	55.9
iBOT [79]	ViT-L/16	88.3	60.4	57.3

Method	Arch.	Clevr/Count	Clevr/Dist
<i>Methods without view data augmentations</i>			
data2vec [8]	ViT-L/16	85.3	71.3
MAE [36]	ViT-H/14	90.5	72.4
I-JEPA	ViT-H/14	86.7	72.4
<i>Methods using extra data augmentations</i>			
DINO [18]	ViT-B/8	86.6	53.4
iBOT [79]	ViT-L/16	85.7	62.8



V-JEPA

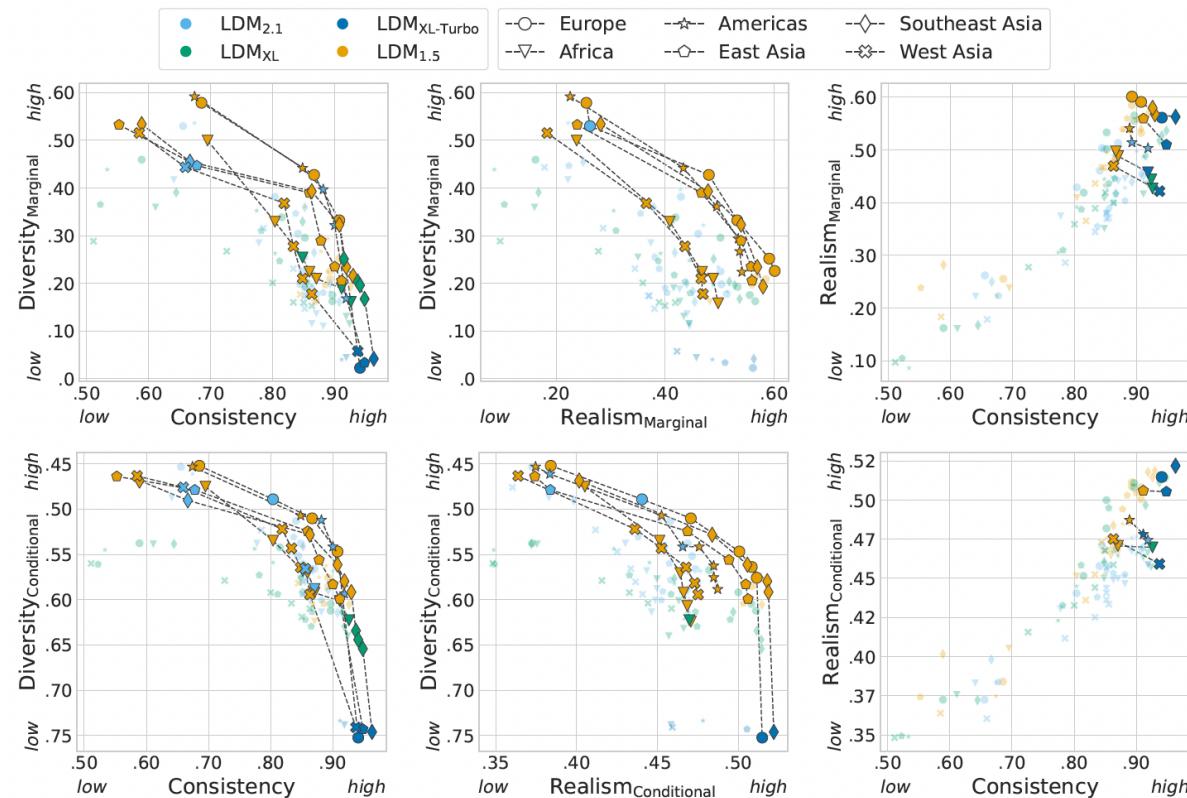
- Revisiting Feature Prediction for Learning Visual Representations from Video - Yann LeCun et al. - Feb-2024
 - essentially same ideas of JEPA - loss function is calculated in embedding space - for better semantic representation learning (rather than pixel-wise learning)



More realistic generative model becomes, less diverse it becomes

- Consistency-diversity-realism Pareto fronts of conditional image generative models - FAIR at Meta - Montreal, Paris & New York City labs, McGill University, Mila, Quebec AI institute, Canada CIFAR AI - Jun-2024
 - realism comes at the cost of coverage, *i.e.*, *the most realistic systems are mode-collapsed!*
 - intuition (or hunch)
 - world models should *not* be generative - should make predictions in representation space - in representation space, unpredictable or irrelevant information is absent
- main argument in favor of JEPA

Consistency-diversity-realism trade-off



Selected References & Sources

Selected references & sources

- Robert H. Kane “Quest for Meaning: Values, Ethics, and the Modern Experience” 2013
- Michael J. Sandel “Justice: What’s the Right Thing to Do?” 2009
- Daniel Kahneman “Thinking, Fast and Slow” 2011
- Yuval Noah Harari “Sapiens: A Brief History of Humankind” 2014
- M. Shanahan “Talking About Large Language Models” 2022
- A.Y. Halevy, P. Norvig, and F. Pereira “Unreasonable Effectiveness of Data” 2009
- A. Vaswani, et al. “Attention is all you need” @ NeurIPS 2017
- S. Yin, et. al. “A Survey on Multimodal LLMs” 2023
- Chris Miller “Chip War: The Fight for the World’s Most Critical Technology” 2022
- CEOs, CTOs, CFOs, COOs, CMOs & CCOs @ startup companies in Silicon Valley
- VCs on Sand Hill Road - Palo Alto, Menlo Park, Woodside in California, USA

References

References

- [ACH⁺24] Pietro Astolfi, Marlene Careil, Melissa Hall, Oscar Mañas, Matthew Muckley, Jakob Verbeek, Adriana Romero Soriano, and Michal Drozdzal. Consistency-diversity-realism pareto fronts of conditional image generative models, 2024.
- [ADM⁺23] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023.
- [BGP⁺24] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video, 2024.
- [Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.
- [BKP22] Abhaya Bhardwaj, Shristi Kishore, and Dhananjay K. Pandey. Artificial intelligence in biological sciences. *Life*, 12(1430), 2022.

- [BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.
- [DFJ22] Thomas A. Dixon, Paul S. Freemont, and Richard A. Johnson. A global forum on synthetic biology: The need for international engagement. *Nature Communications*, 13(3516), 2022.
- [GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016.
- [HM24] Guadalupe Hayes-Mota. Emerging trends in AI in biotech. *Forbes*, June 2024.
- [HNF09] Alon Halevy, Peter Norvig, and Nanediri Fernando. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24:8 – 12, 05 2009.
- [HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.
- [Kah11] Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, 2011.
- [KF09] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

- [Kui23] Todd Kuiken. Artificial intelligence in the biological sciences: Uses, safety, security, and oversight. *Congressional Research Service*, Nov 2023.
- [LG94] Alberto Leon-Garcia. *Probability and Random Processes for Electrical Engineering*. Addison-Wesley, 2nd edition, 1994.
- [LWV⁺24] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov-arnold networks, 2024.
- [Mur12] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.
- [RAB⁺23] Ziaur Rahman, Muhammad Aamir, Jameel Ahmed Bhutto, Zhihua Hu, and Yurong Guan. Innovative dual-stage blind noise reduction in real-world images using multi-scale convolutions and dual attention mechanisms. *Symmetry*, 15(11), 2023.
- [Say21] Kelley M. Sayler. Defense primer: Emerging technologies. *Congressional Research Service*, 2021.
- [SB18] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Bradford Books, 2nd edition, 2018.

- [Sha23] Murray Shanahan. Talking about large language models, 2023.
- [Toe23] Rob Toews. The next frontier for large language models is biology. *Forbes*, July 2023.
- [Wet23] Kris A. Wetterstrand. Dna sequencing costs: Data, 2023.