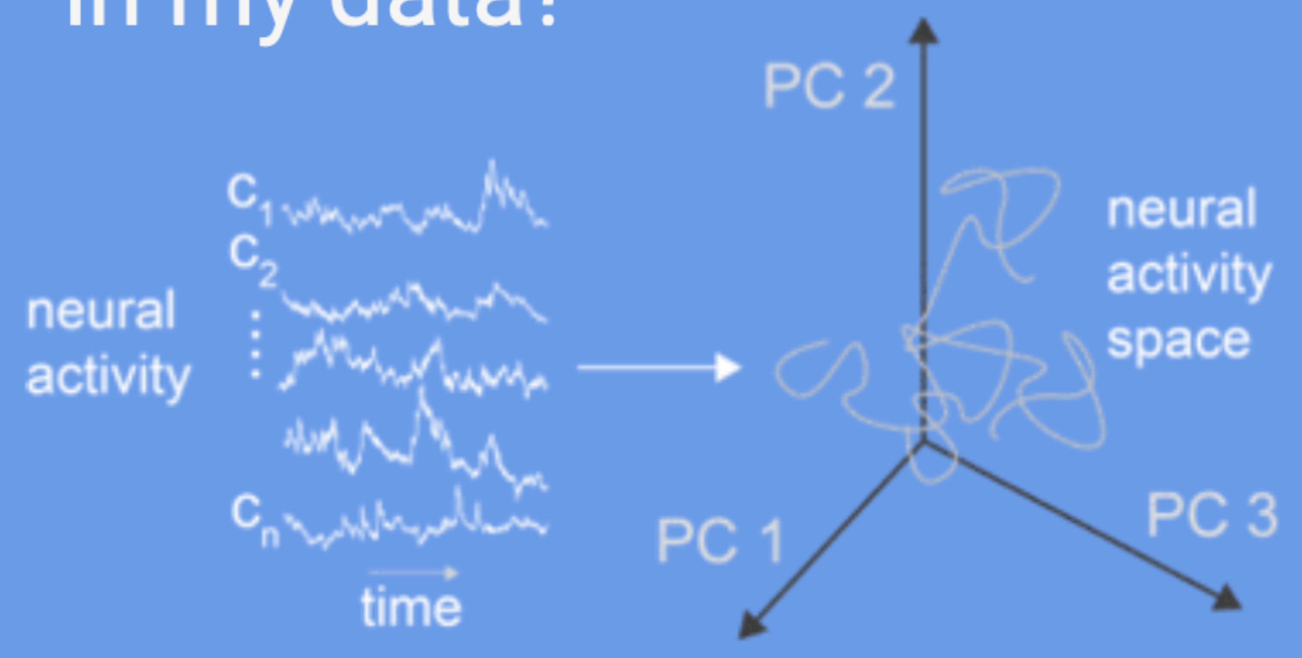
Machine Learning Basics

James Gornet

CHEN TIANQIAO & CHRISSY INSTITUTE

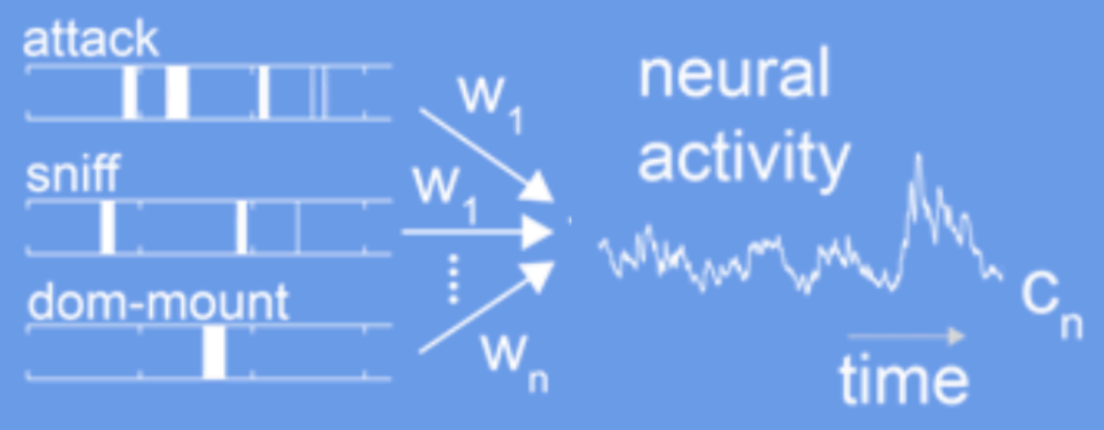
Machine learning provides useful interpretations of your experiments

what are the major signals in my data?



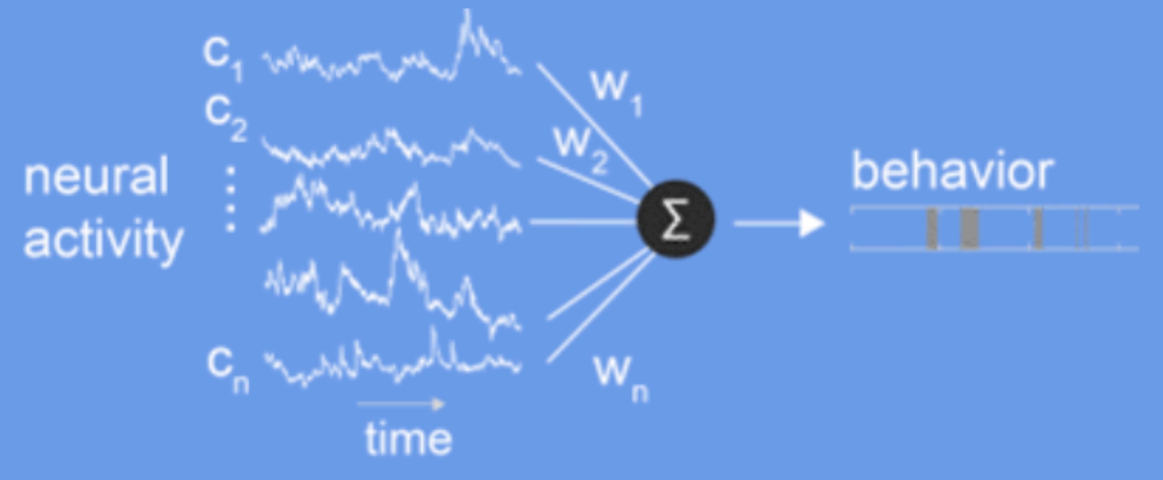
dimensionality reduction

what are single neurons tuned to?



linear encoding models

is information about behavior present in my data?



linear decoding models

Caltech
datasai_2023

Machine learning follows *three* simple principles

Statistics
how do I model my question?

Computational Mathematics
how do I tractably solve my question?

Programming
how do I implement my algorithm?

Statistics

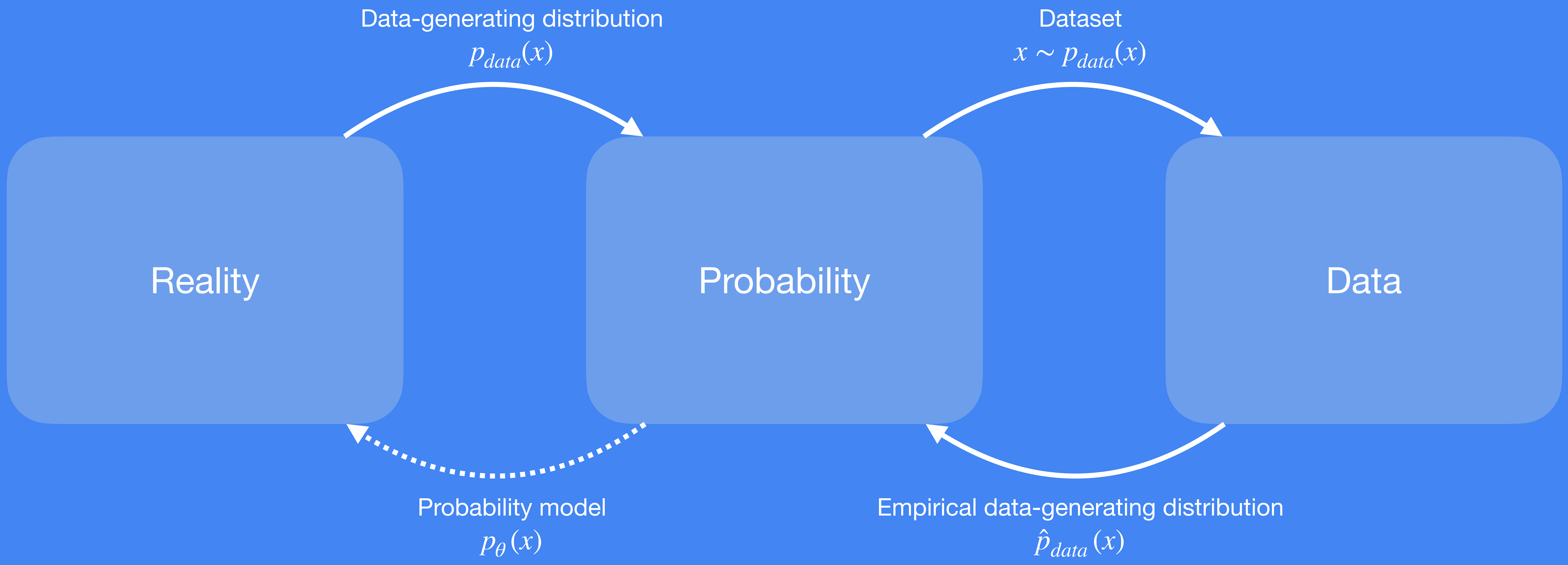
Computational Mathematics

Programming

Statistics is *inverse* probability

Reality $\xrightarrow[p_{data}(x)]{\text{Data-generating distribution}}$ Probability $\xrightarrow[x \sim p_{data}(x)]{\text{Dataset}}$ Data

Statistics is *inverse* probability



Probability identities

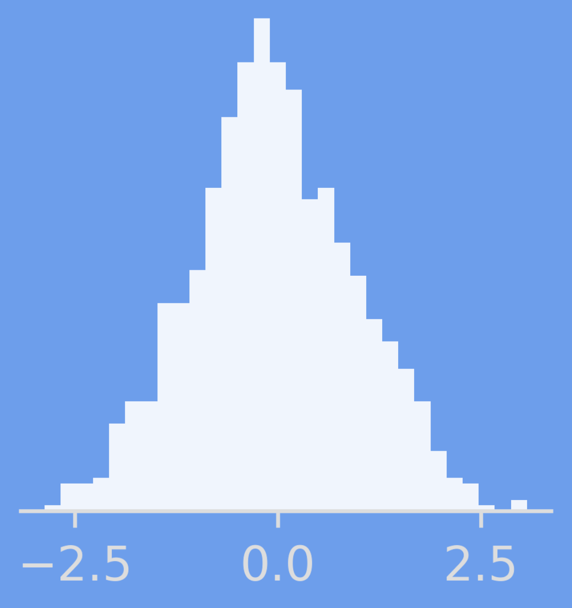
	Discrete	*Continuous*
Independence	$\mathbb{P}[x, y] = \mathbb{P}[x]\mathbb{P}[y]$	$p(x, y) = p(x)p(y)$
Conditioning	$\mathbb{P}[x, y] = \mathbb{P}[x \mid y]\mathbb{P}[x]$	$p(x, y) = p(x \mid y)p(x)$
Marginal	$\mathbb{P}[x] = \sum_y \mathbb{P}[x, y]$	$p(x) = \int_{y \in Y} p(x, y)\, dy$
Expectation	$\mathbb{E}[x] = \sum x\mathbb{P}[x]$	$\mathbb{E}[x] = \int_{x \in X} xp(x)\, dx$

Caltech
datasai_2023

Every probability distribution tells a story

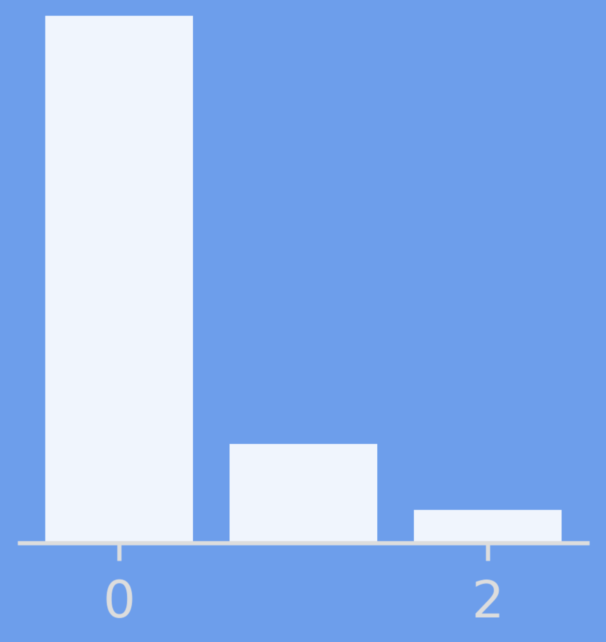
Normal distribution

if I have a quantity affected by sums of different noises, what do I expect to see?



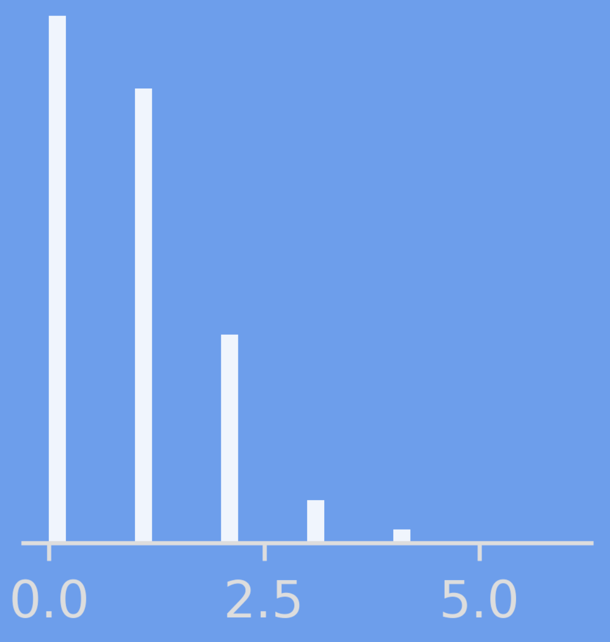
Categorical distribution

if I have a pick a set of objects *a,b,c* with probabilities *90%, 5%, 5%*, what do I expect to see?



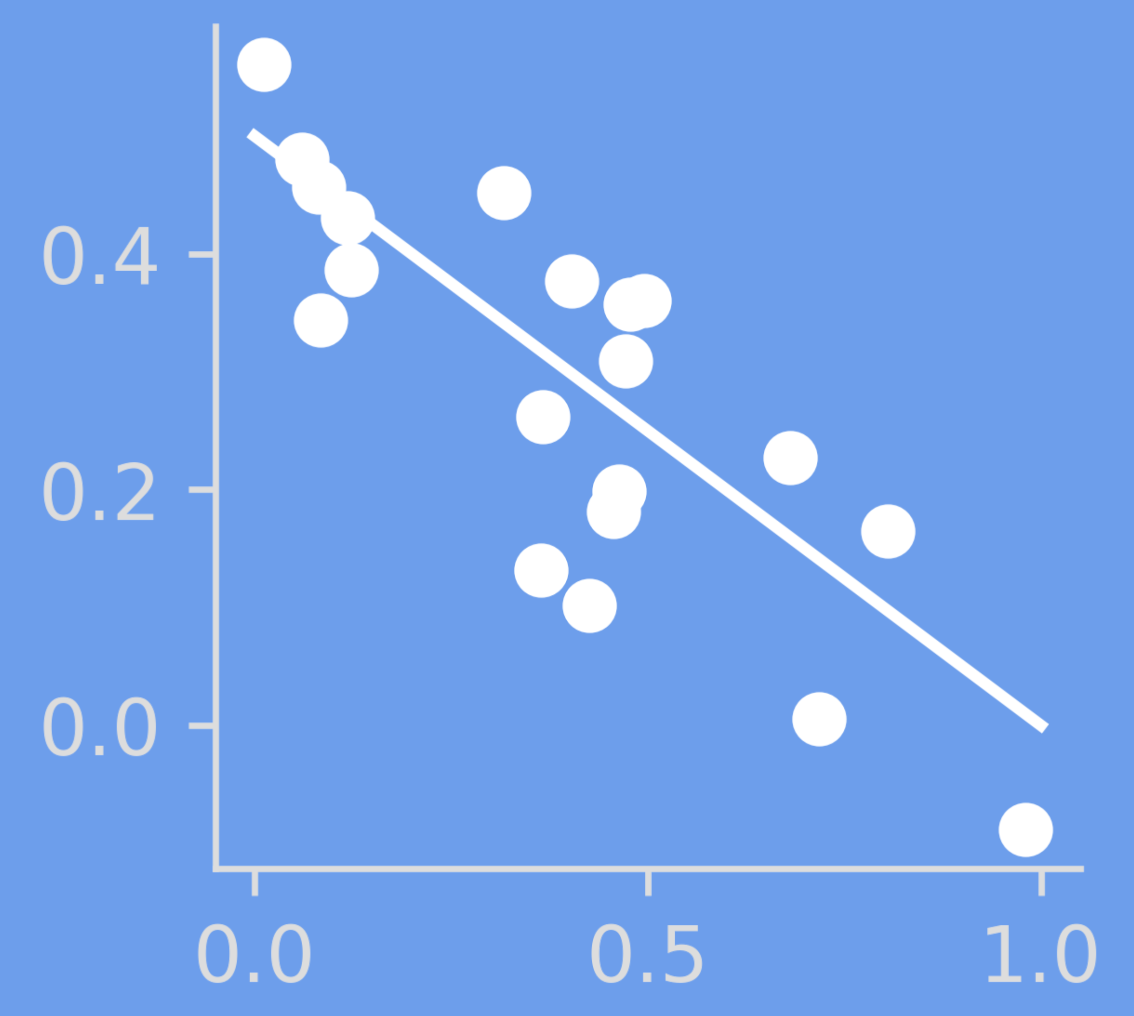
Poisson distribution

if an event occurs every 1.1 seconds, what do I expect to see?

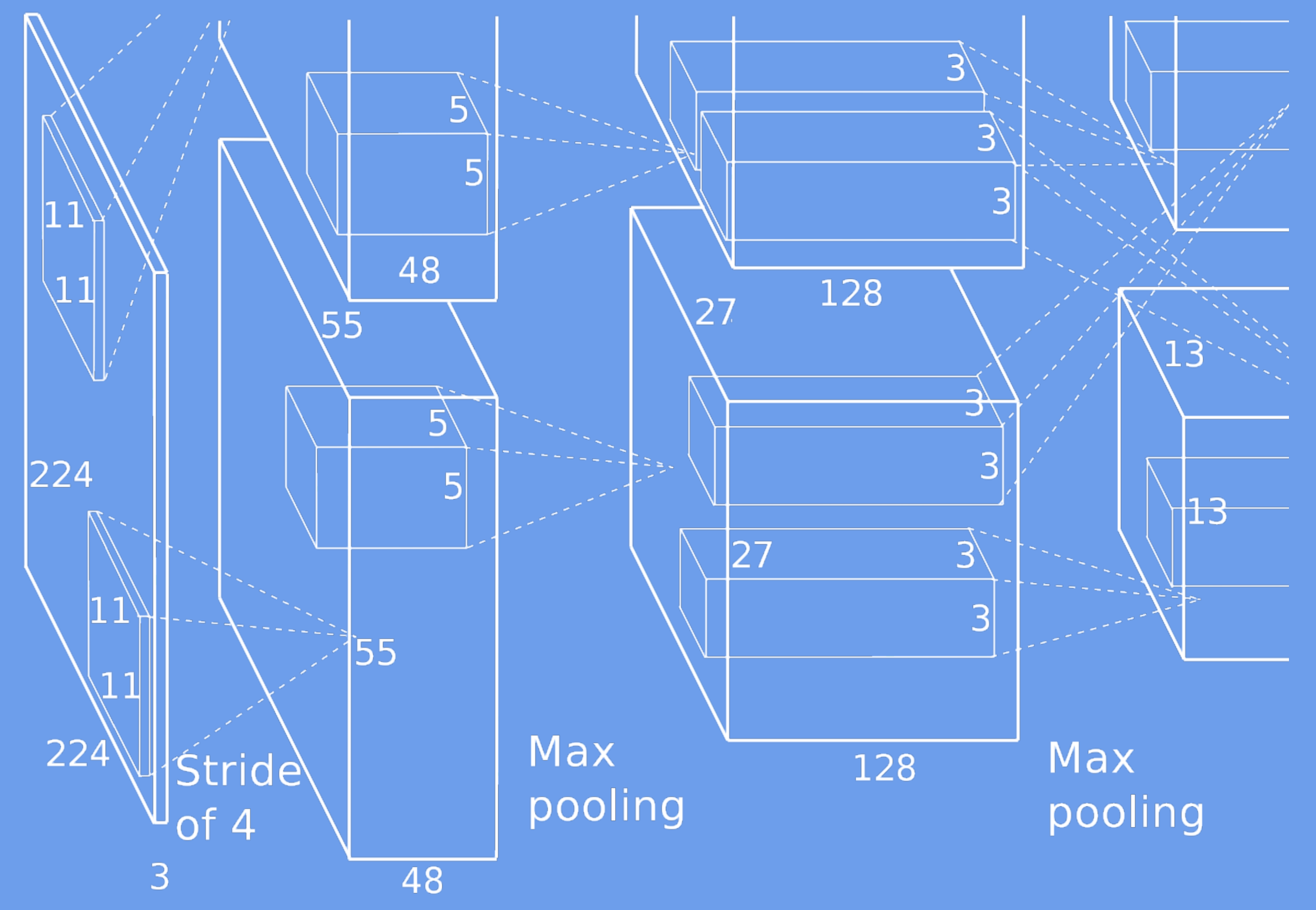


Every method in statistics has a model

Linear relationships



Pixel-wise neighbor relationships

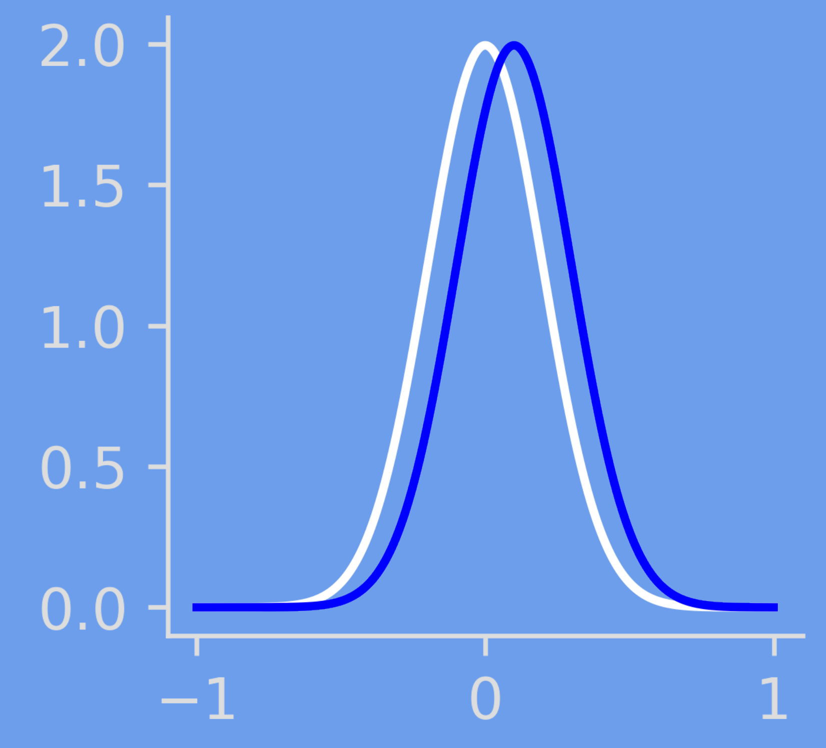


Caltech
datasai_2023

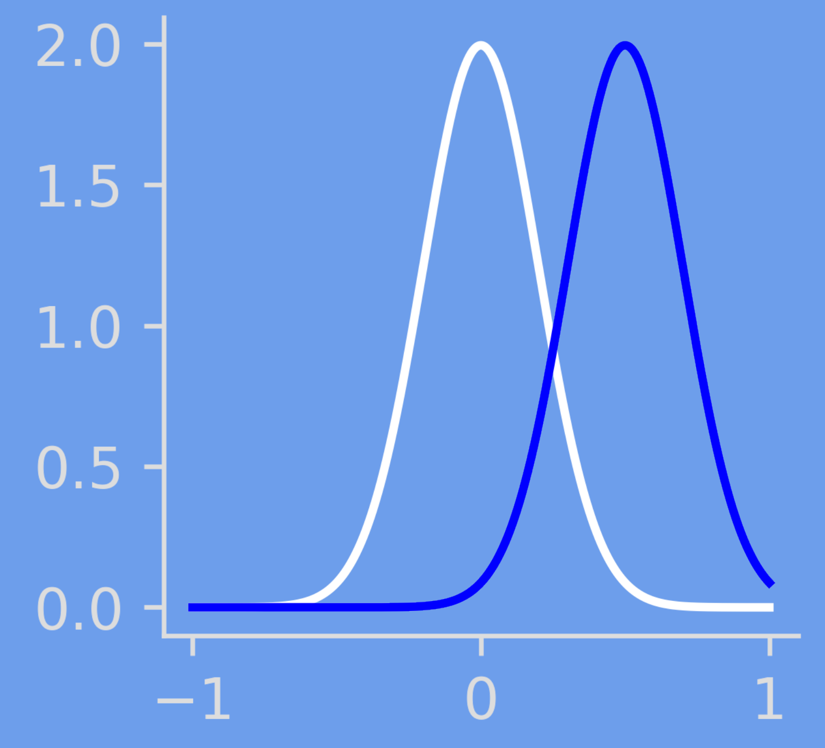
Maximum likelihood scores a probability model on data

Maximum likelihood

$$\arg\max_\theta \mathbb{E}_{p_{data}}[\log p_\theta(x)]$$



$>$



Statistics

Computational Mathematics

Programming

Gradient descent provides an algorithm for finding the best model

Gradient descent

$$\theta_{t+1} \leftarrow \theta_t - \nabla_\theta \mathbb{E}_{p_{data}}[-\log p_\theta(x)]$$

Example: Logistic regression

Model

$$p_\theta(x) = \frac{1}{1+e^{-z}}$$
$$z = Wx + b$$

Likelihood

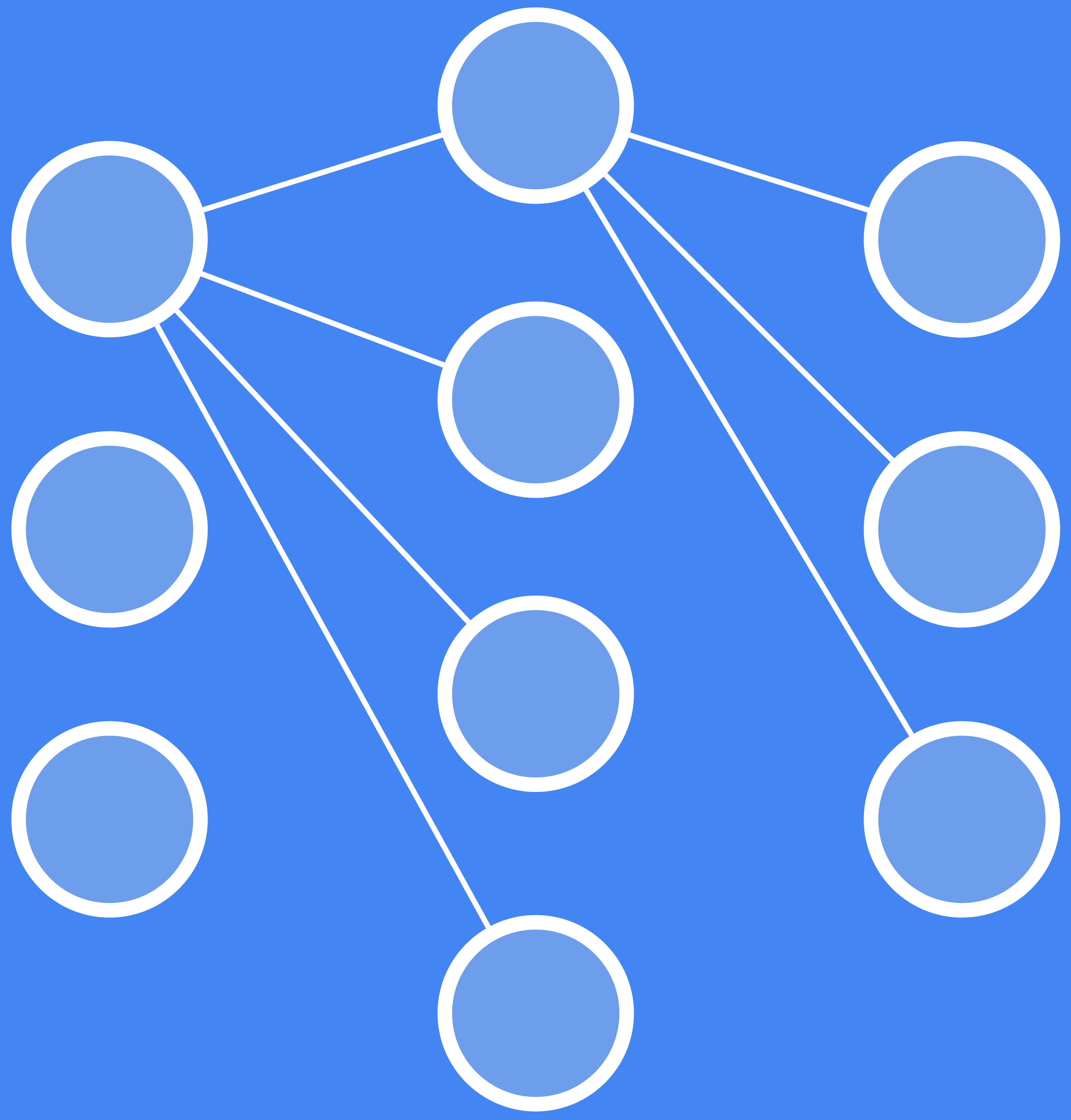
$$\mathbb{E}_{p_{data}}[\log p_\theta(x)]$$
$$= \sum_i y_i \log(z_i) + (1-y_i)\ \log(1-z_i)$$

Bernoulli distribution

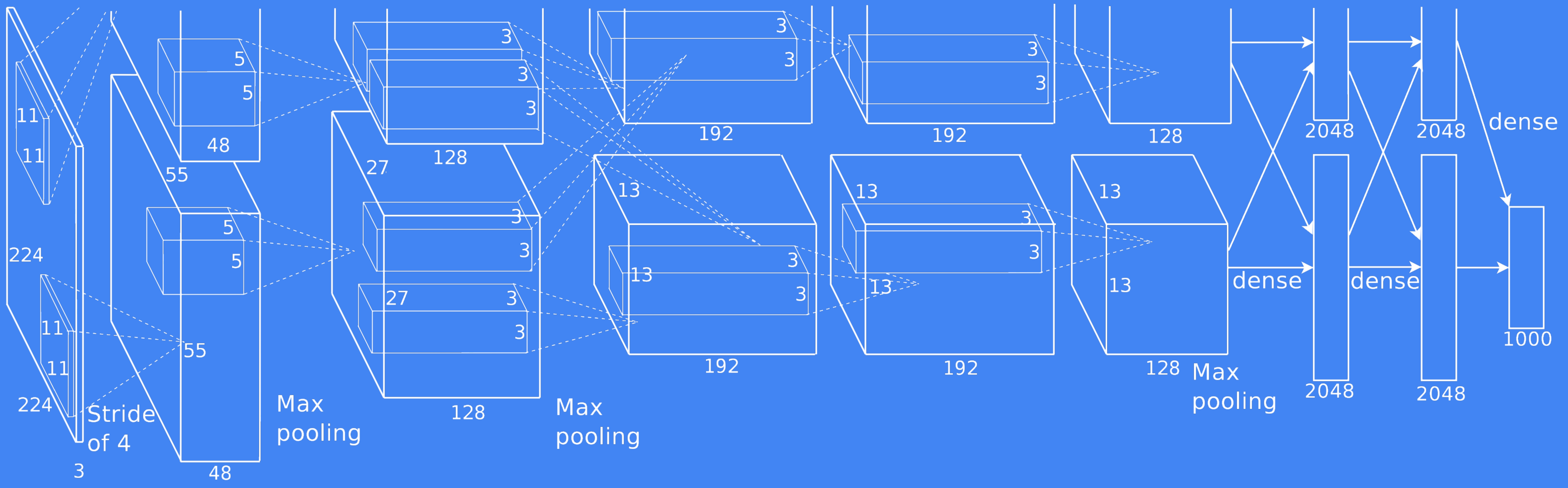
if an object has two possibilities with probabilities $y, (1-y)$, what do I expect to see?

Neural networks are linear transforms followed by nonlinear operations

$$f(x) = W_2 \max\{0, W_1 x + b_1\} + b_2$$



Deep neural networks are feature learners



Caltech
datasai_2023

MNIST Classification

Data

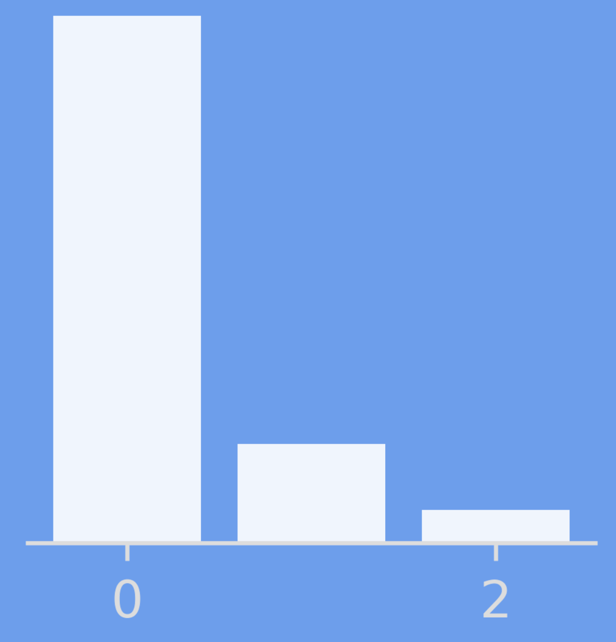
Model

$$p_\theta(z_i) = \frac{e^{z_i}}{\sum_j e^{zj}}$$

$$z = W_2 \max\{0, W_1 x + b_1\} + b_2$$

Categorical distribution

if an image can be the number *0,…,9* with probabilities $z_0, \ldots, z_9$, what do I expect to see?



Caltech
datasai_2023

Maximum likelihood provides a quantity to minimize

Model

$$p_\theta(z_i) = \frac{e^{z_i}}{\sum_j e^{zj}}$$

$$z = W_2 \max\{0, W_1 x + b_1\} + b_2$$

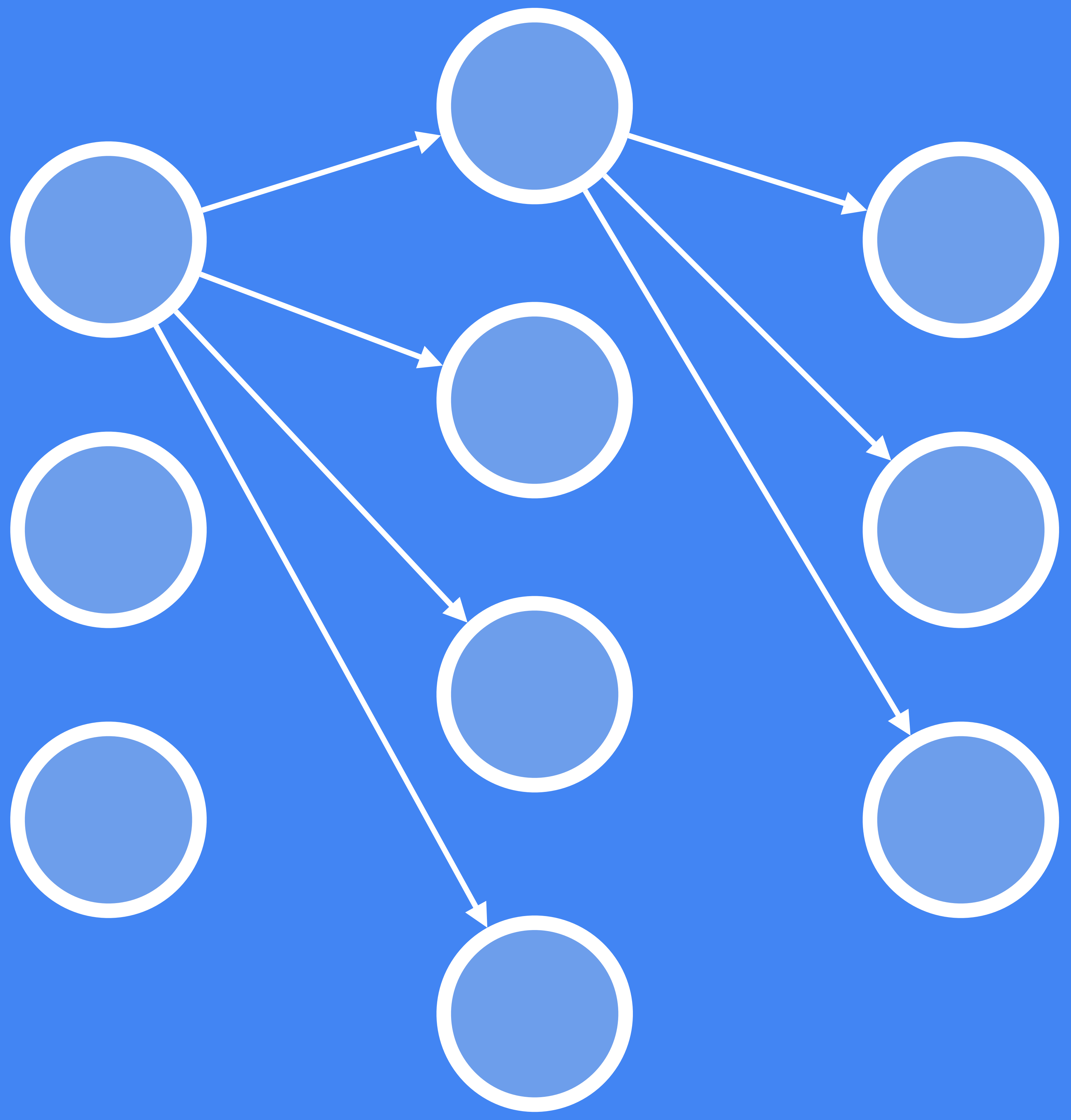
Likelihood

$$\mathbb{E}_{\hat{p}_{data}}[\log p_\theta(z)] = \sum_i y_i \log z_i$$

The chain rule of calculus allows us to perform maximum likelihood estimation

Chain rule

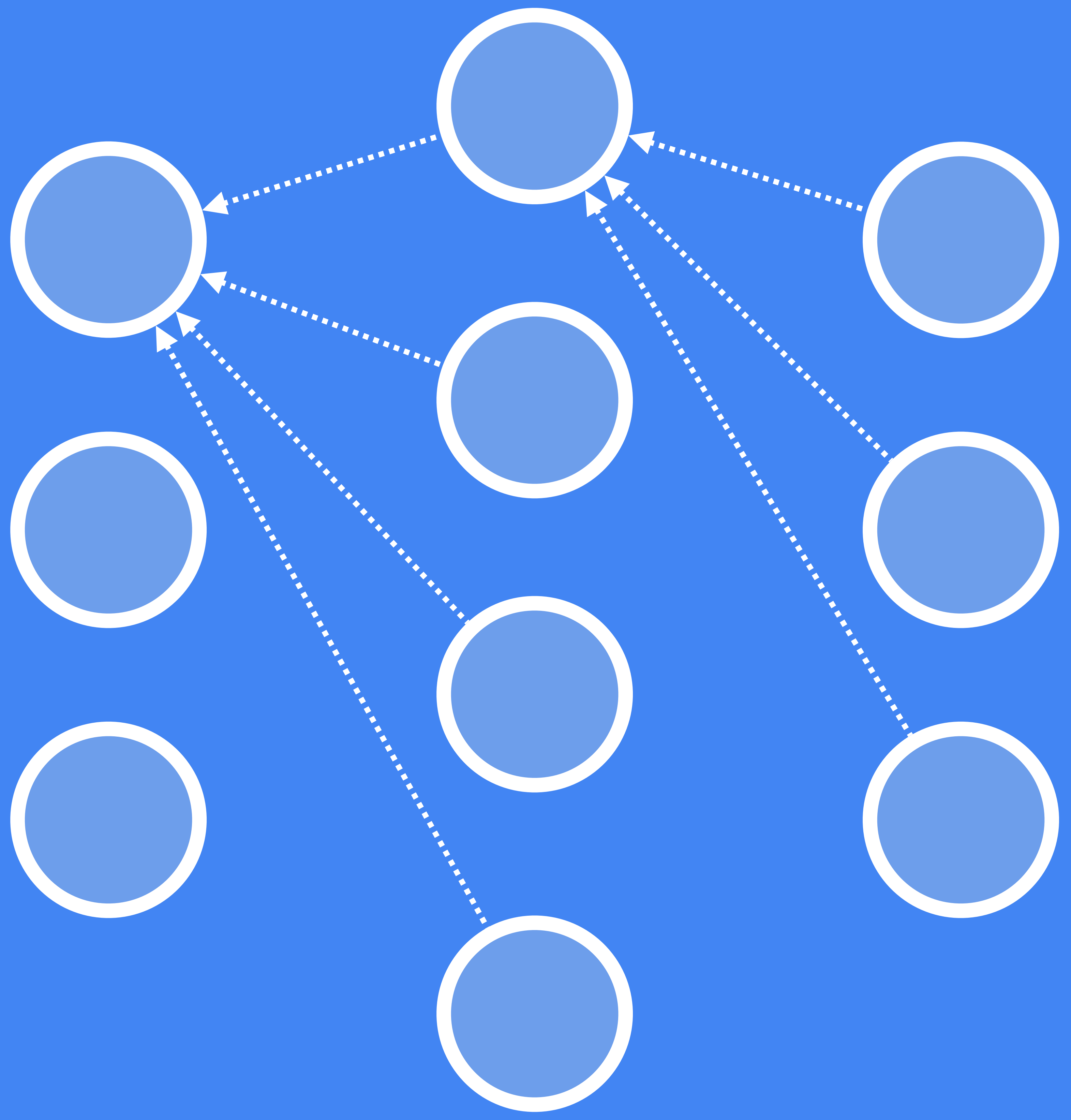
$$\nabla_x z(y(x)) = \left(\frac{\partial y}{\partial x}\right)^{\top} \nabla_x y(x)$$



The chain rule of calculus allows us to perform maximum likelihood estimation

Chain rule

$$\nabla_x z(y(x)) = \left(\frac{\partial y}{\partial x}\right)^\top \nabla_x y(x)$$



The chain rule of calculus allows us to perform maximum likelihood estimation

Model

$$p_\theta(z_i) = \frac{e^{z_i}}{\sum_j e^{zj}}$$

$$z = W_2 \max\{0, W_1 x + b_1\} + b_2$$

Maximum likelihood gradients

$$\nabla_z p_\theta(z) = z_i(\delta_{ij} - p_j)$$

for $y_i = 1$

$$\frac{\partial}{\partial x}\mathbb{E}_{p_{data}}[\log p_\theta(z)] = \sum_k y_k \frac{1}{(p_\theta(z))_k} \frac{\partial (p_\theta(z))_k}{\partial z}$$

$$= p_\theta(z)_i - y_i$$

Statistics

Computational Mathematics

Programming