Welcome!

COGS 118B Winter 2024

Jason G. Fleischer, PhD
Department of Cognitive Science
University of California San Diego

https://jgfleischer.com
Book a slot in my office hours

How to succeed in this course

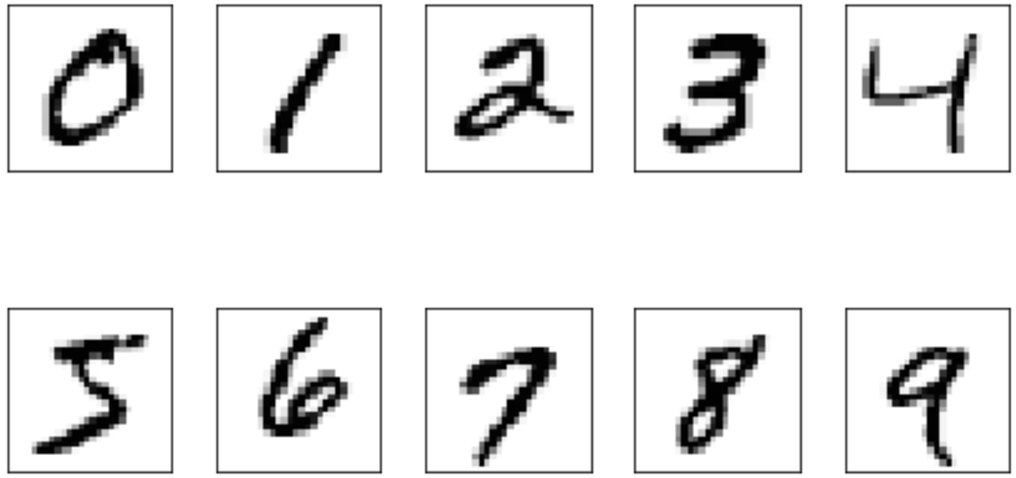
- Watch the videos before lecture
- Explore beyond the basic material
  - Readings and other extras in this course
  - Keep going!
- Play
  - with the Jupyter notebooks here
  - try things on your own with new data
- Participate in Piazza and discussion sections
- Try to figure things out for a while yourself, then ask for help

Machine learning

"A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

— Tom Mitchell, Professor at Carnegie Mellon University

Handwriting Recognition Example:



- Task T: ?
- Performance measure P : ?
- Training experience E: ?

Applications of Machine Learning

Your ideas…

Applications of Machine Learning

Your ideas…

Categories of ML

Supervised

- Labelled data
- Direct feedback
- Predict, classify, or fit a model

COGS 118A

Unsupervised

- No labels
- No feedback
- Find hidden structure using a model

COGS 118B

Reinforcement learning

- Feedback via reward (only label)
- Learns the series of actions that lead to reward

COGS 182

Neural Networks

- Mostly supervised and self-supervised. Some unsupervised
- Learn to predict/classify

COGS 181

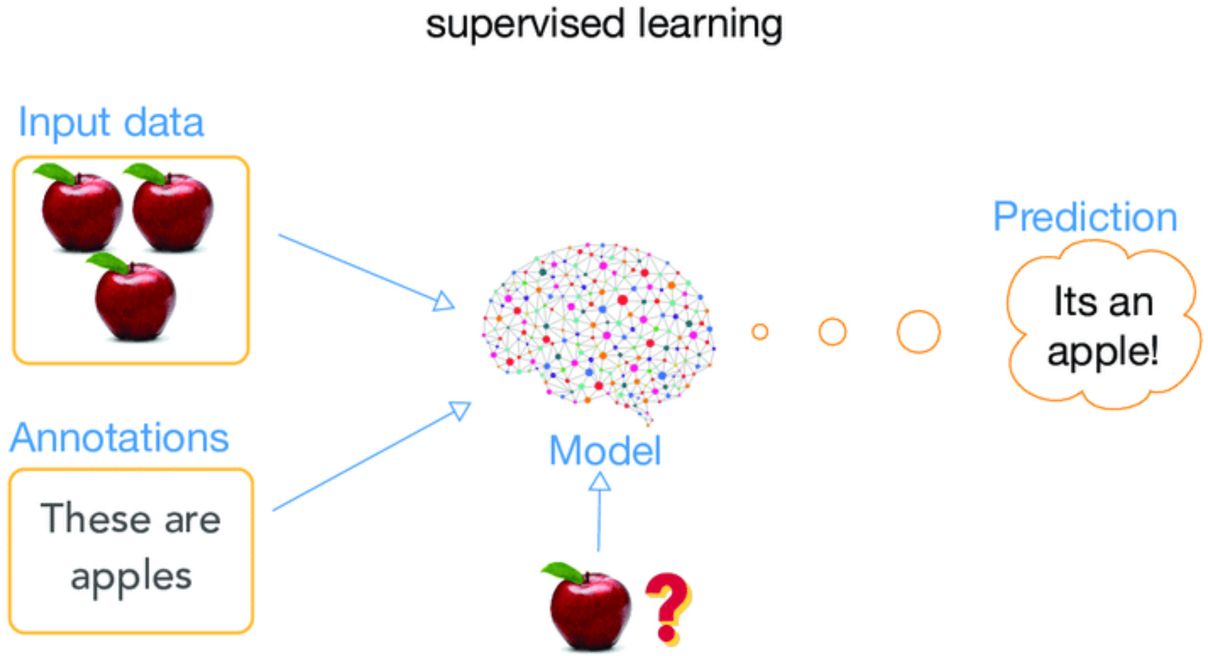
Genetic Algorithms

- Search for solutions through simulated natural selection
- Learn to predict/classify

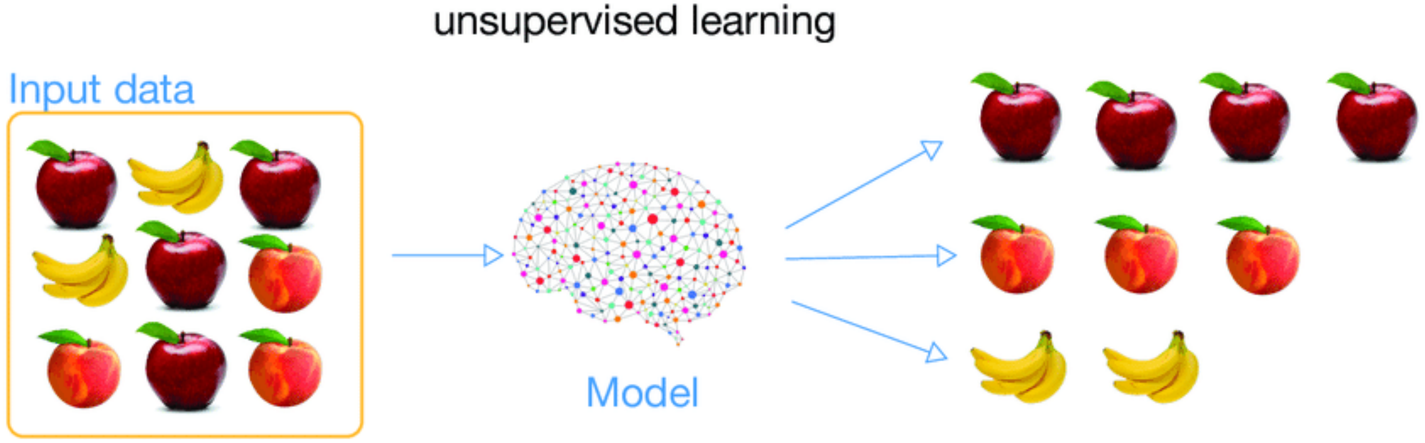
COGS 186

Supervised vs unsupervised

COGS 118A

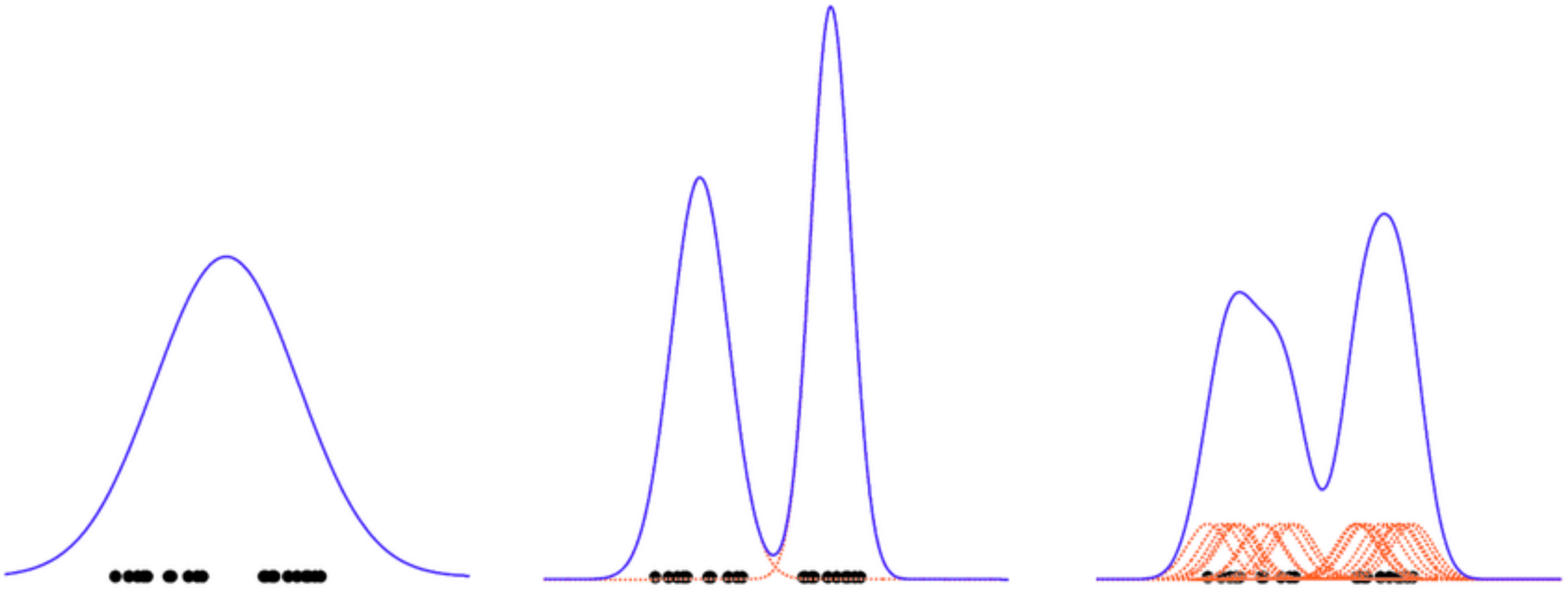


COGS 118B

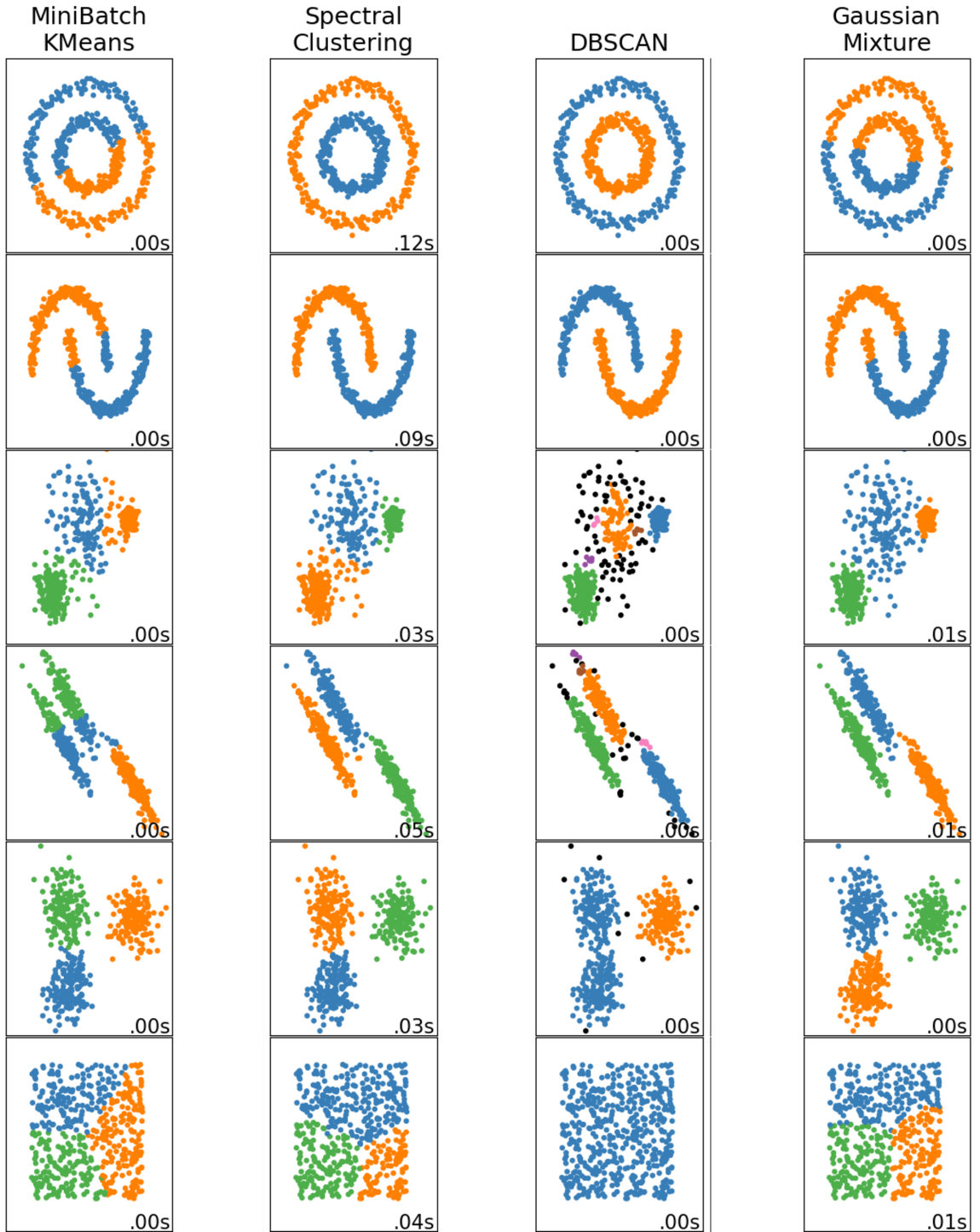


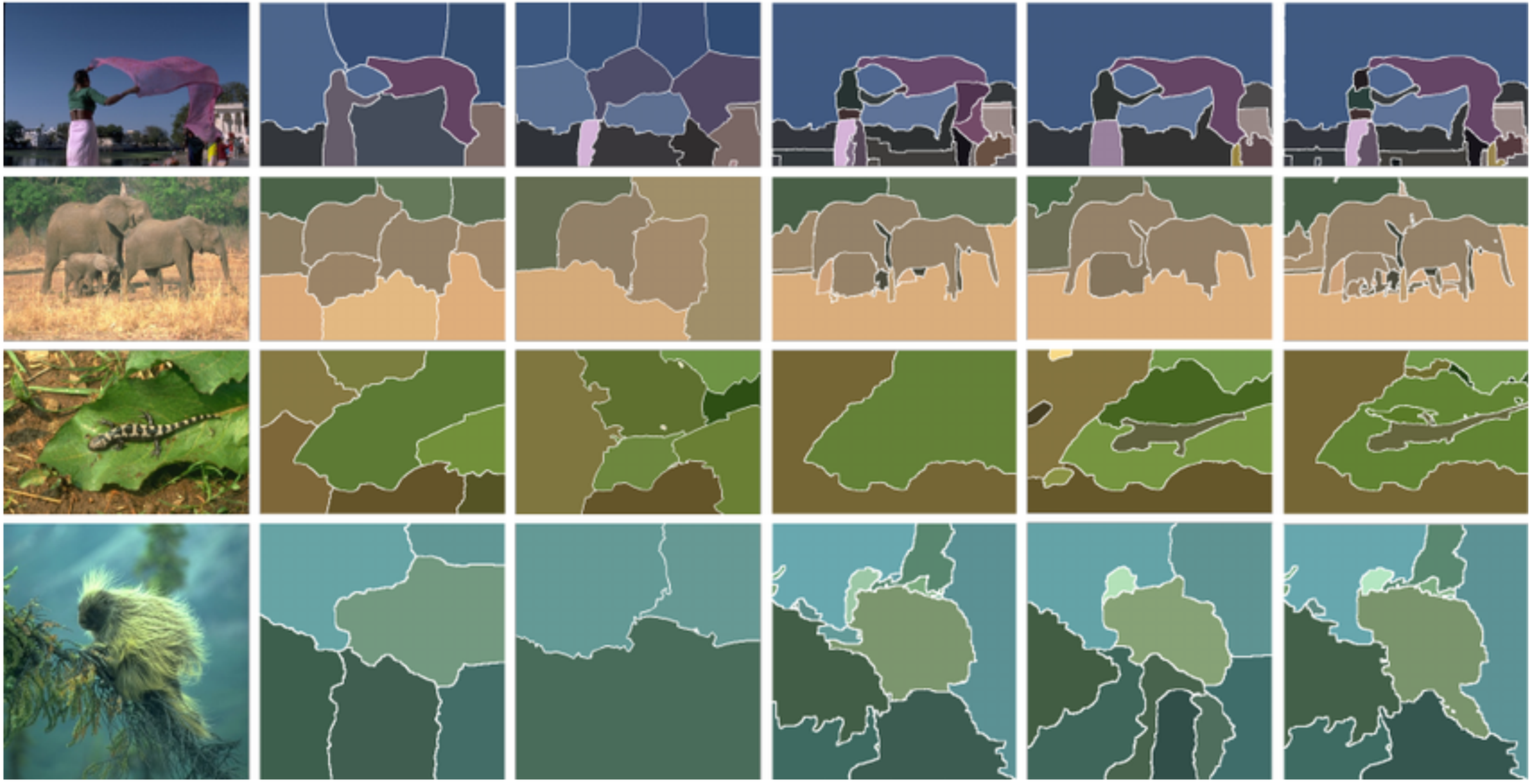
Some kinds of unsupervised learning

Density estimation

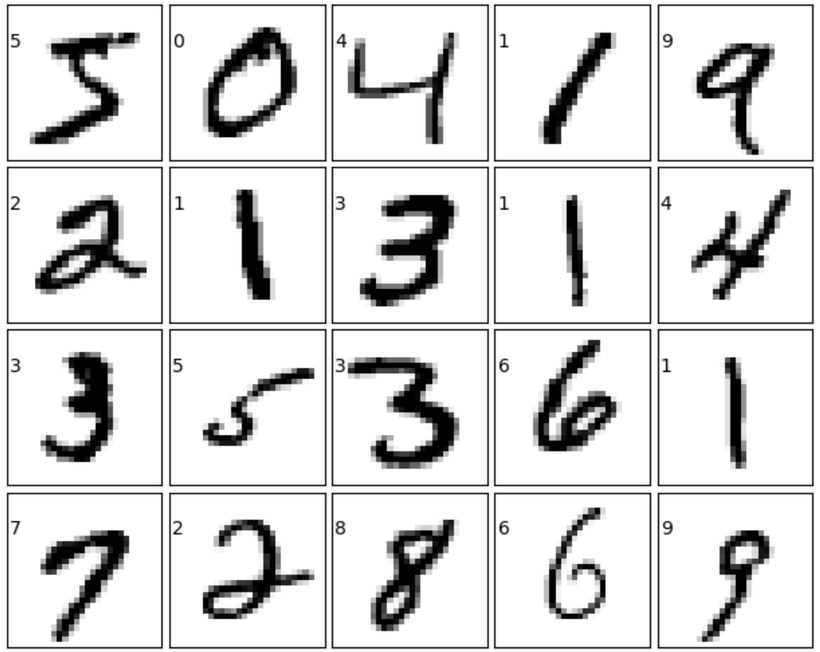


Clustering

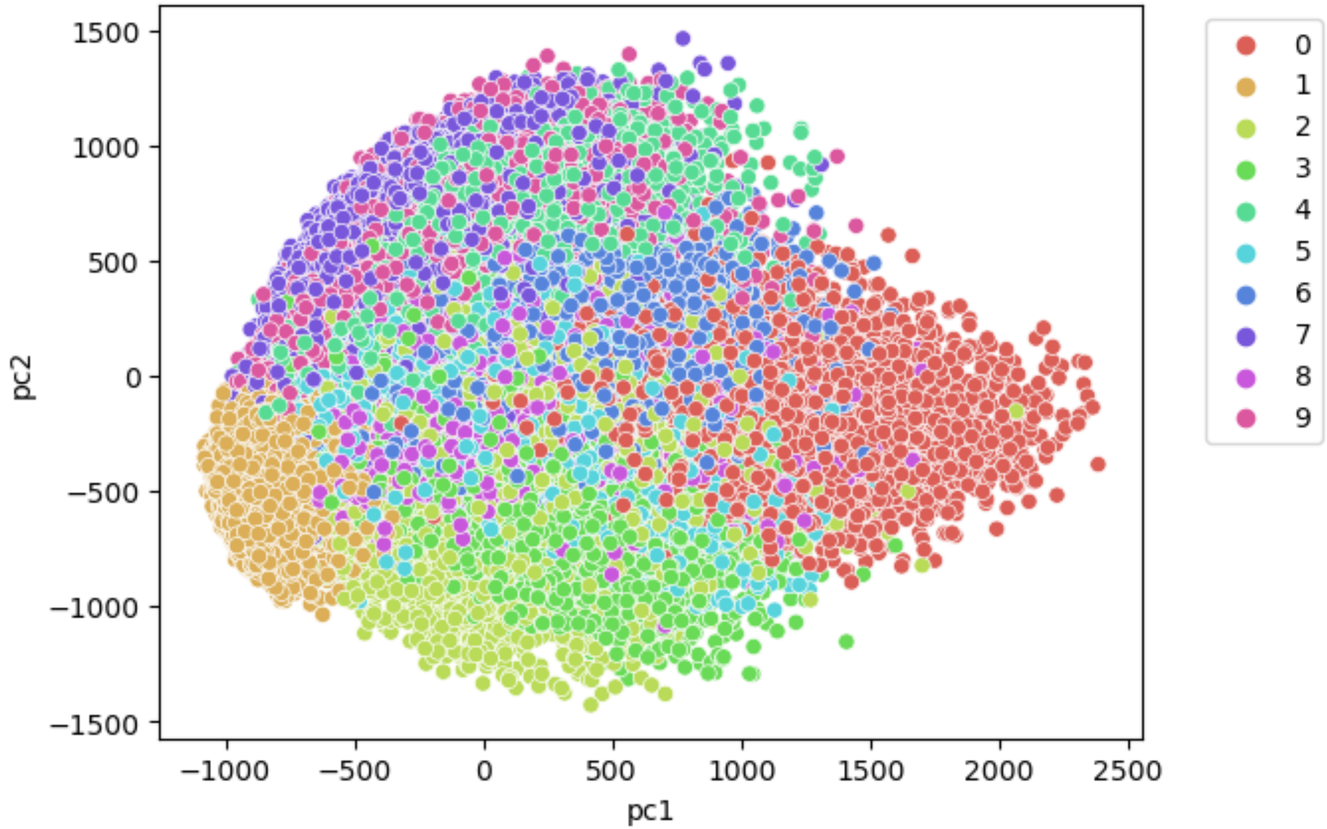

MiniBatch KMeans
Spectral Clustering
DBSCAN
Gaussian Mixture
.00s
.12s
.00s
.00s
.00s
.09s
.00s
.00s
.00s
.03s
.00s
.01s
.00s
.05s
.00s
.01s
.00s
.03s
.00s
.00s
.00s
.04s
.00s
.01s

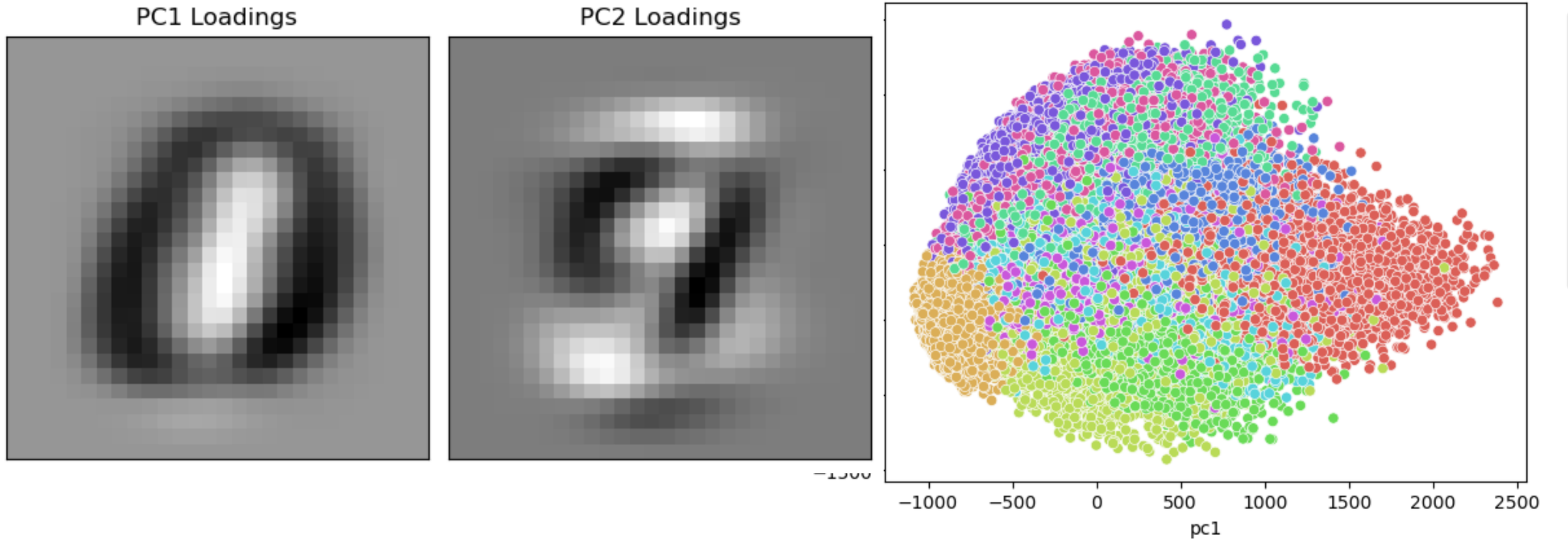


Dimensionality reduction



28x28 pixels = 784 dimensions




PC1 Loadings
PC2 Loadings
−1000
−500
0
500
1000
1500
2000
2500
pc1
0
1
2
3
4
5
6
7
8
9

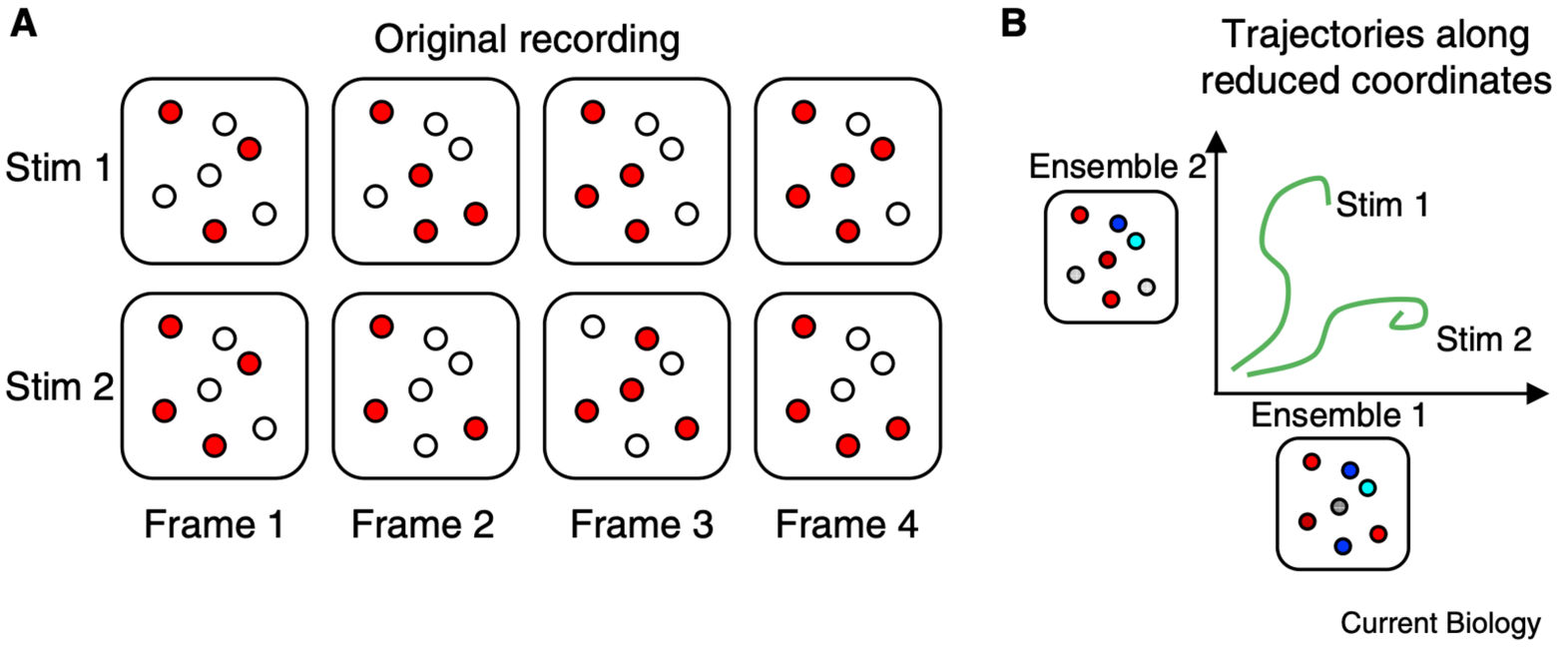
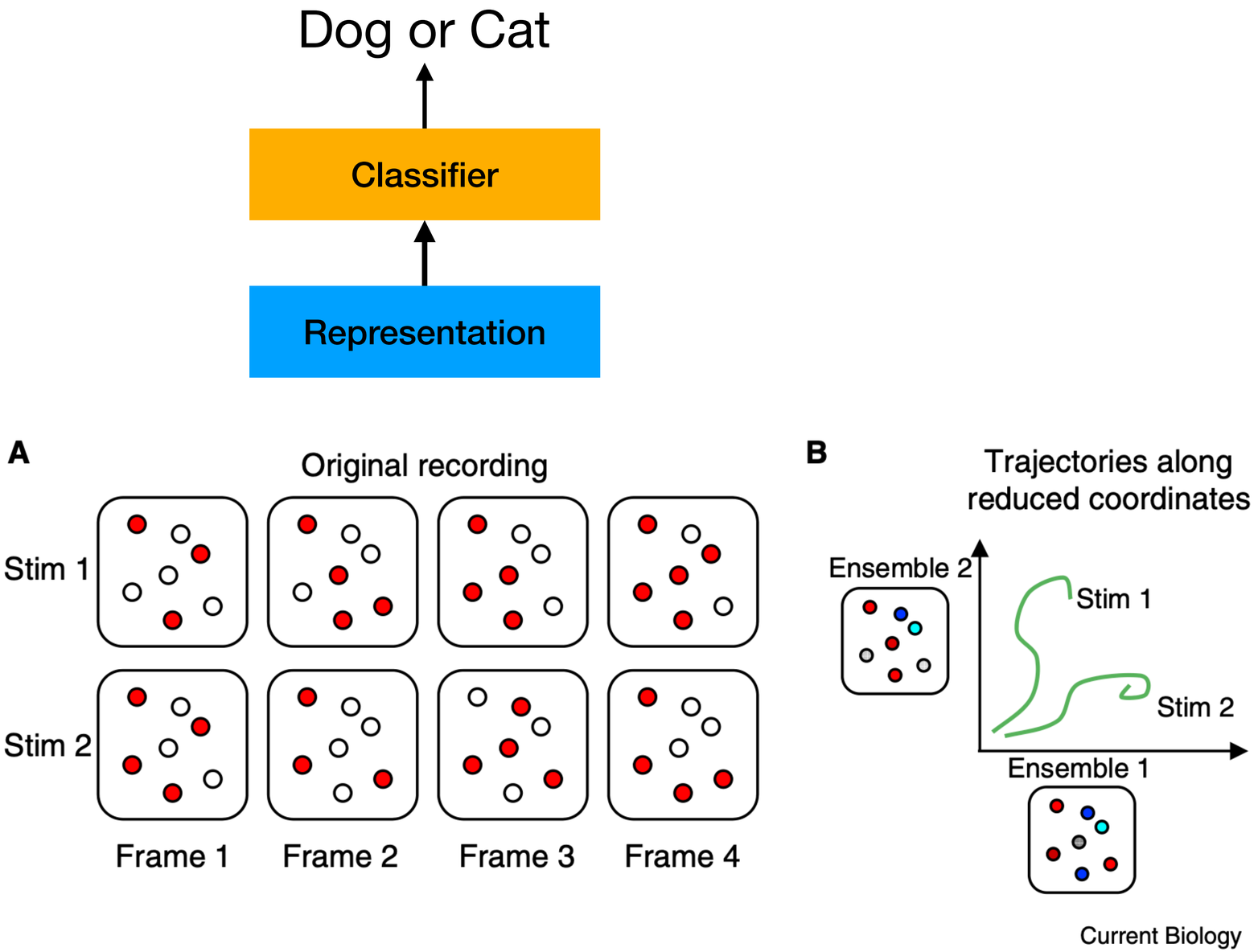


Figure 4. Reducing the dimensionality of a multi-neuron dataset to visualize stimulus-dependent responses.

Stimulus dependence can be hard to identify when looking at a high-dimensional dataset of many neurons recorded over many time points. (A) Diagram of a time-varying neural population recording one might obtain after presenting two different stimuli. The circles represent neurons, with red fill indicating their activity. (B) The time-varying activities of two ensembles (activation patterns) of neurons identified using PCA. The colors of the neurons in each ensemble represent the positive (red) or negative (blue) contribution of each neuron to the ensemble. In this two-dimensional view, the separability of the population responses to the two stimuli becomes more obvious.

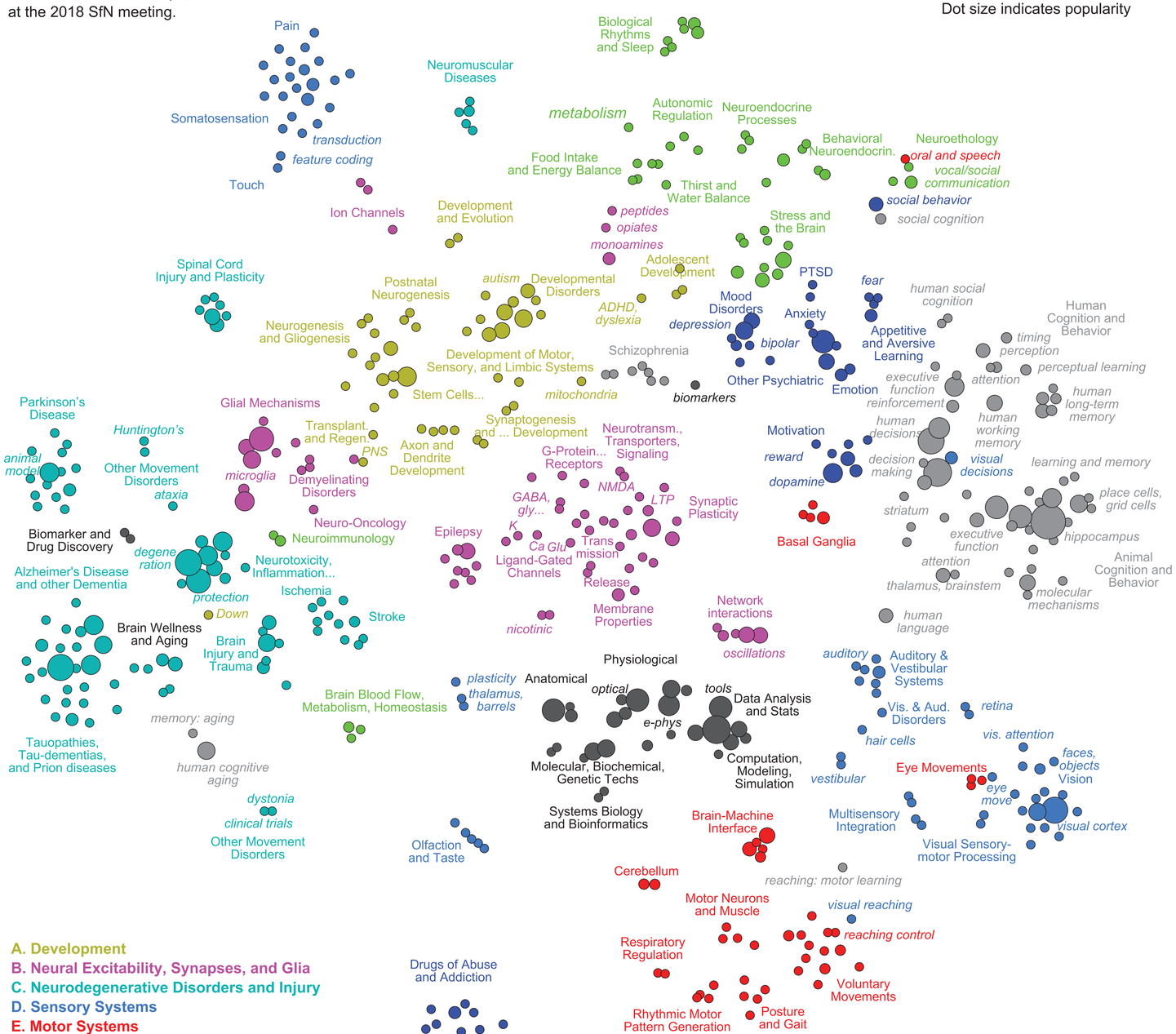
https://www.cell.com/current-biology/pdf/S0960-9822(16)30487-0.pdf


Dog or Cat
Classifier
Representation
A
Original recording
Stim 1
Stim 2
Frame 1
Frame 2
Frame 3
Frame 4
B
Trajectories along
reduced coordinates
Ensemble 2
Stim 1
Stim 2
Ensemble 1
Current Biology

The Structure of Neuroscience

From the itineraries of 8,329 attendees
at the 2018 SfN meeting.

Each dot is a Topic
Dot size indicates popularity

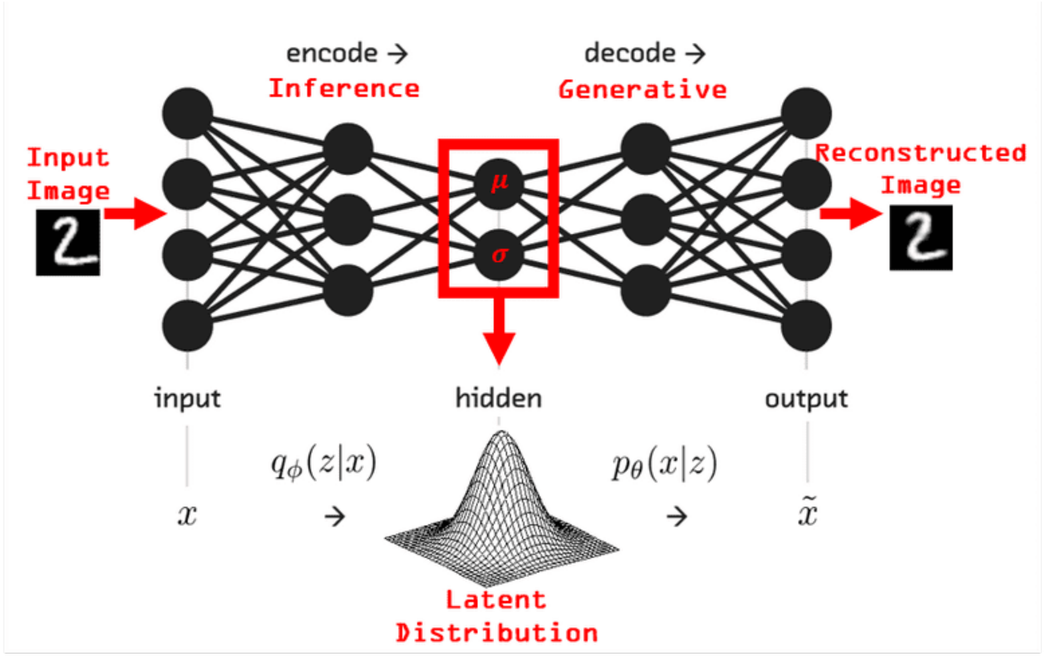


A. Development
B. Neural Excitability, Synapses, and Glia
C. Neurodegenerative Disorders and Injury
D. Sensory Systems
E. Motor Systems
F. Integrative Physiology and Behavior
G. Motivation and Emotion
H. Cognition
I. Techniques

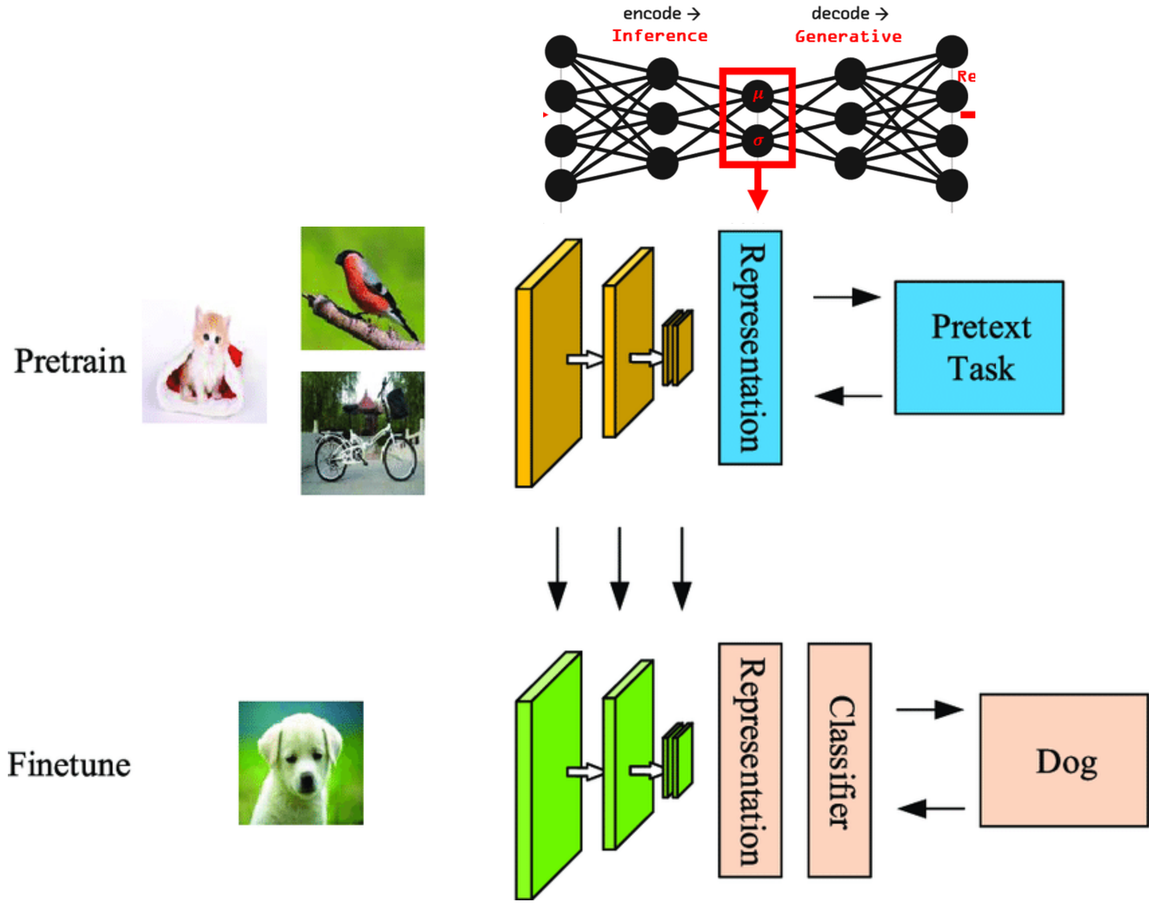
https://www.cell.com/neuron/pdf/S0896-6273(19)30437-4.pdf

Analysis: Matteo Carandini
Data: Society for Neuroscience

Supervised vs. unsupervised is a false dichotomy?

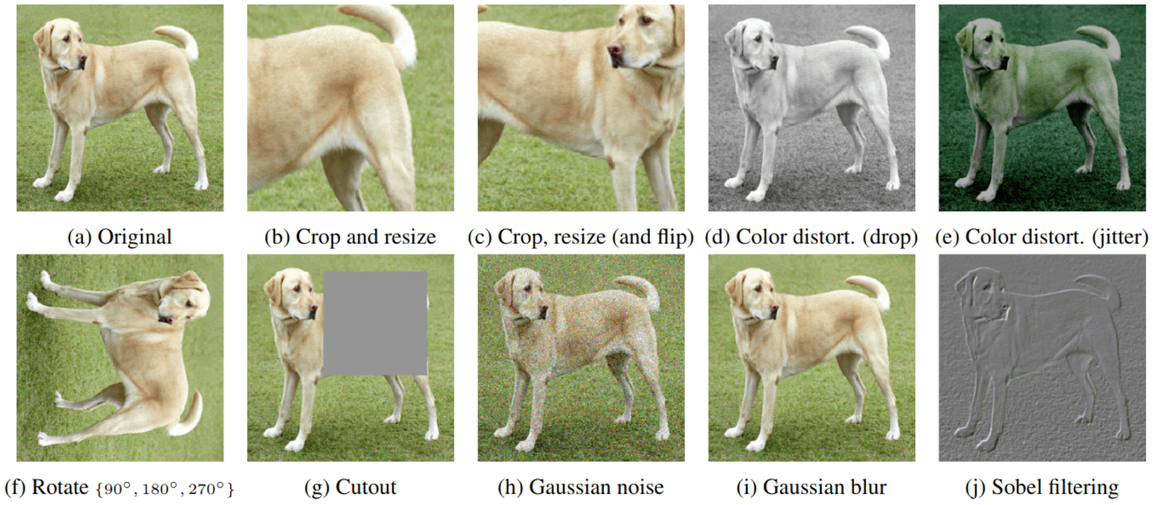


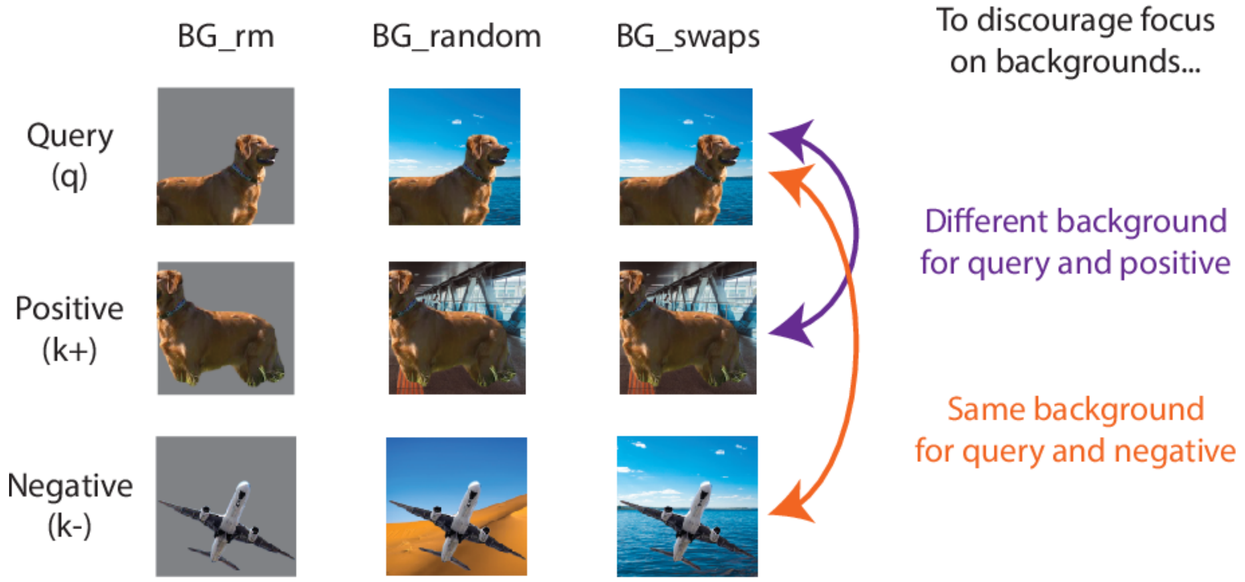
Autoencoders



Transfer learning

A false dichotomy continued…



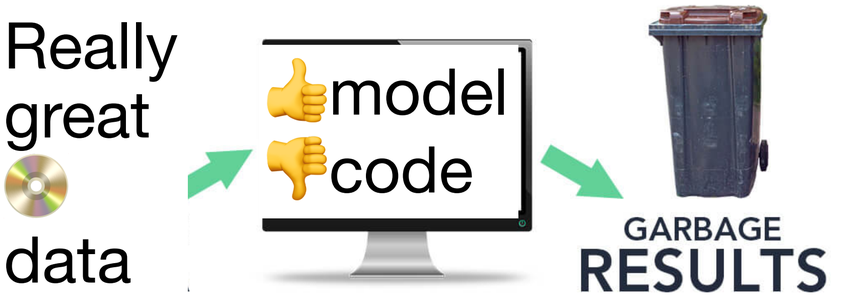


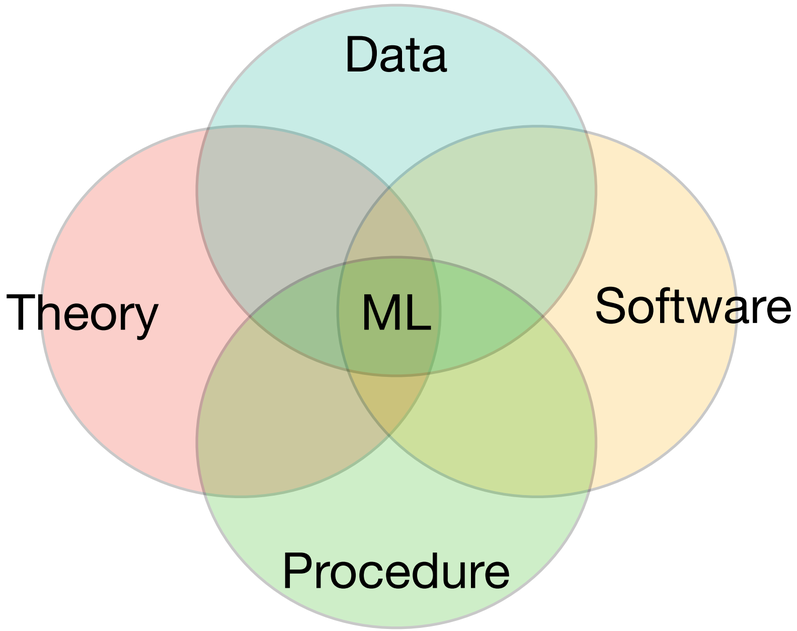
Contrastive learning via data augmentation

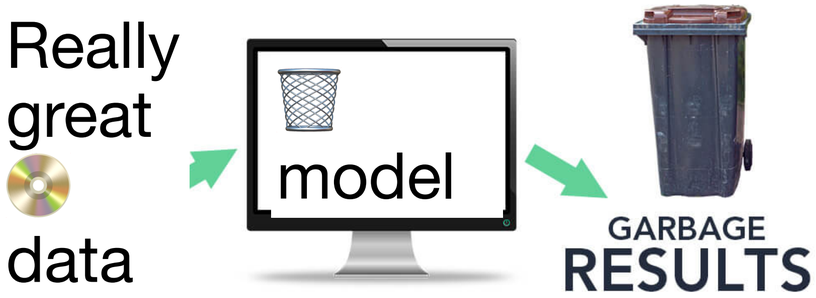
Challenges of ML

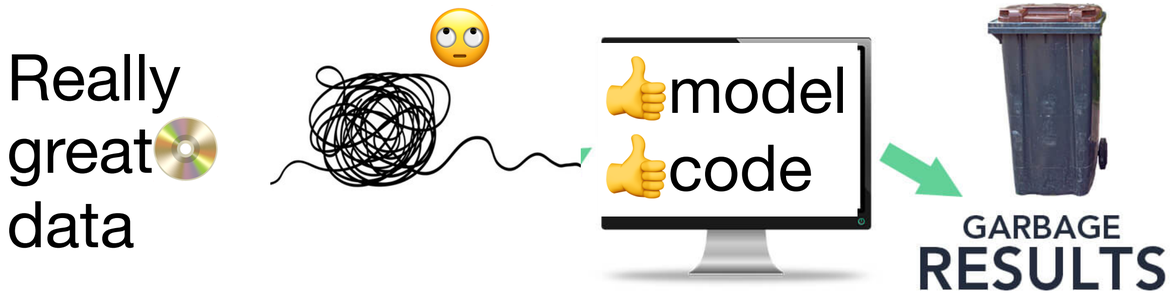
- Data
  - Insufficient
  - non-representative
  - poor quality
- Selecting a good enough/better/ best ML algorithm
  - Hyperparameter tuning
  - Model selection
- Testing how good your model is
  - How do you even measure performance? What's the metric?
  - ML is too powerful - overfit!
  - Generalization is hard, errors on new data are worse than expected!
- Dealing with SNAFU: distribution shifts, non-stationarity etc


GARBAGE
DATA
GREAT
MODEL
GARBAGE
RESULTS


Really
great
data
model
code
GARBAGE
RESULTS


Data
Theory
ML
Software
Procedure

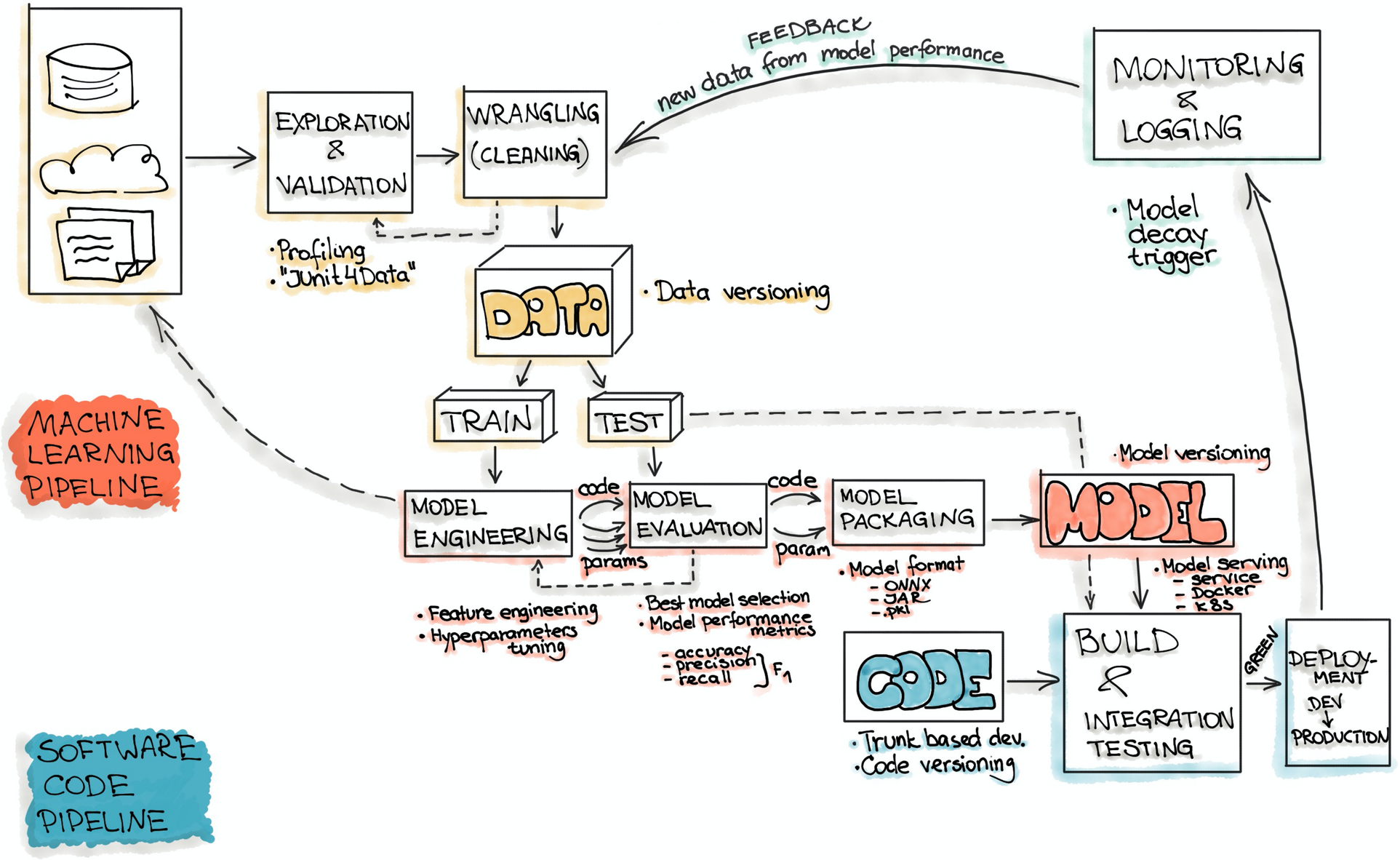

Really
great
data
model
GARBAGE
RESULTS


Really
great
data
model
code
GARBAGE
RESULTS

The process and art of machine learning
(At a small scale)

- Acquire data and knowledge about the subject
- Curate & clean the data
- Explore the data
- Make useful transformations of the data
- Split/resample your data (e.g., cross validation) to help maximize both how well the model is made and also how well you estimate the ability of the model
- “DO ML”
- Evaluate and interpret your results

MACHINE LEARNING ENGINEERING



ml-ops.org

Notation and math

Basic notation

Inputs/Data

A dataset contains n samples or observations. Each sample has m variables. We represent the i^{th} sample as a column vector $\mathbf{x}_i \in \mathcal{R}^m$ and the j^{th} measurement of that sample is represented as $x_{i,j}$. That is,

$$\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots x_{i,m})^{\mathrm{T}} = \begin{pmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,m} \end{pmatrix}$$

There are two different ways we might want to represent a whole dataset. The first method is using set notation: a dataset is a set of input vectors $S = \{\mathbf{x}_i, i \in [1, n]\}$. At other times (when we do linear algebra) we might want to represent the dataset as a so-called "design matrix". In that case, samples are rows and the columns are variables: $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n)^{\mathrm{T}}$. Or to write it another way....

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{pmatrix}$$

Basic notation

MODEL PARAMETERS

Model: $\mathbf{w} = (w_1, ..., w_m) \in \mathbb{R}^m$ (in the same dimension of input $\mathbf{x}$)

bias: $b \in \mathbb{R}$ (scalar)

Data sample $\mathbf{x} = (x_1, ..., x_m) \in \mathbb{R}^m$,

$$\mathbf{w} \cdot \mathbf{x} + b \qquad\qquad (w_1, w_2, .., w_m) \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ x_m \end{pmatrix} + b$$

"." refers to as the dot product between two vectors

Alternative notation 1: $< \mathbf{w}, \mathbf{x} > +b$

Alternative notation 2: $\mathbf{w}\mathbf{x}^T + b$ ($\mathbf{w}$ and $\mathbf{x}$ are row vectors).

$\mathbf{w}^T\mathbf{x} + b$ ($\mathbf{w}$ and $\mathbf{x}$ are column vectors).

Actually we often absorb the bias term into the weight vector:

$\mathbf{w} = (w_1, w_2, \ldots w_m, b) \in \mathscr{R}^{m+1}$

Which requires us to also modify the inputs:

$\mathbf{x} = (x_1, x_2, \ldots x_m, 1)^\mathsf{T} \in \mathscr{R}^{m+1}$

Basic notation

Probability

The probability of an event a happening can be written as

$Pr(a = \text{True})$ or equivalently $P(a)$

Likewise we can denote the probability of a random variable X having a specific value (such as being greater 3), as $P(X > 3)$. Obviously $X > 3$ is simply an event... which we might denote as a.

Joint probability

The probability that events a and b have happened simultaneously is $P(a, b)$. If and only if a and b are *independent* events then P(a,b) = P(a)P(b). Like if a is "A six sided die 🎲 rolls > 3" and b is "A second 🎲 rolls >3". Those are independent events because the rolls do not affect each other. It's rolls of 4,5,6 or 50% of the time that a die comes up >3. So the probability is $0.5 * 0.5 = 0.25$ for both to come up that way at the same time.

More generally (i.e., even when the events are **not** independent),

$$P(a, b) = P(a)P(b|a)$$

That last term would be read in English as "The probability of b happening *given* a has already happened". When a and b are independent, that last bit "given a has already happened" is irrelevant.

In those cases b happens with the same probability regardless of what happens with a. But sometimes a being true makes b more or less likely. For example, if $a =$ "I'm broke" is True, it makes $b =$ "I'm going out for dinner" much less likely.