# DSC291: Machine Learning with Few Labels

## Supervised / Unsupervised Learning

**Zhiting Hu**

Lecture 3, April 8, 2025

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

# Logistics

- Office Hour this week:
  - Thursday 2pm PT
- Office Hour in future weeks:
  - Tuesday 2pm PT

Office: HDSI 442

- TA's Office Hour TBA

- Will announce on Piazza later today

# Overview

# Machine learning solutions given few data (labels)

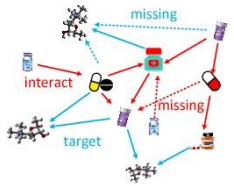
- (1) How can we make more efficient use of data?
  - Clean but small-size, Noisy, Out-of-domain

- (2) Can we incorporate other types of experience in learning?

*Data examples*

Type-2 diabetes is 90% more common than type-1

*Rules/Constraints*




*Knowledge graphs*


*Rewards*

*Auxiliary agents*

*Adversaries*

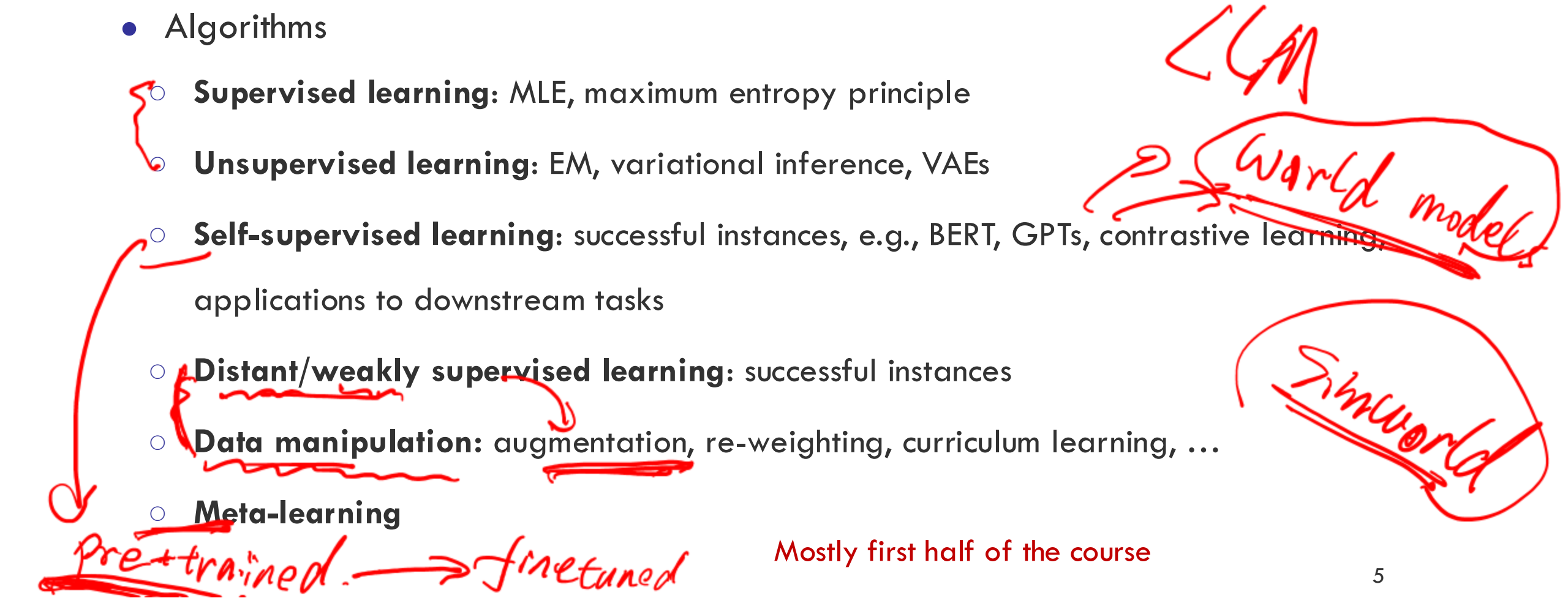

*Master classes*

...

*And all combinations thereof*

# Machine learning solutions given few data (labels)

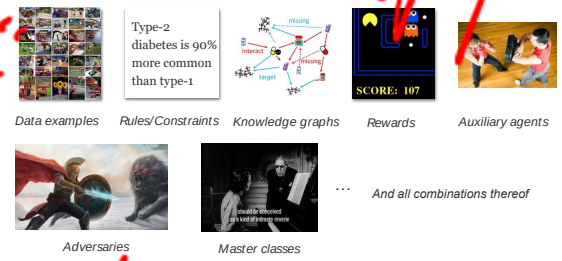
- (1) How can we make more efficient use of data?
  - Clean but small-size, Noisy, Out-of-domain, …
- Algorithms
  - **Supervised learning**: MLE, maximum entropy principle
  - **Unsupervised learning**: EM, variational inference, VAEs
  - **Self-supervised learning**: successful instances, e.g., BERT, GPTs, contrastive learning; applications to downstream tasks
  - **Distant/weakly supervised learning**: successful instances
  - **Data manipulation:** augmentation, re-weighting, curriculum learning, …
  - **Meta-learning**




Mostly first half of the course

# Machine learning solutions given few data (labels)

- (2) Can we incorporate other types of experience in learning?
  - Learning from auxiliary models, e.g., adversarial models:
    - Generative adversarial learning (GANs and variants), co-training, …
  - Learning from structured knowledge
    - Posterior regularization, constraint-driven learning, …
  - Learning from rewards
    - Reinforcement learning: model-free vs model-based, policy-based vs value-based, on-policy vs off-policy, extrinsic reward vs intrinsic reward, …
  - Learning in dynamic environment *(not covered)*
    - Online learning, lifelong/continual learning, …




Second half of the course

# Algorithm marketplace

Designs driven by: experience, task, loss function, training procedure …

maximum likelihood estimation

reinforcement learning as inference

data re-weighting

inverse RL

active learning

policy optimization

data augmentation

reward-augmented maximum likelihood

label smoothing

softmax policy gradient

imitation learning

actor-critic

adversarial domain adaptation

GANs

posterior regularization

knowledge distillation

intrinsic reward

constraint-driven learning

generalized expectation

prediction minimization

regularized Bayes

learning from measurements

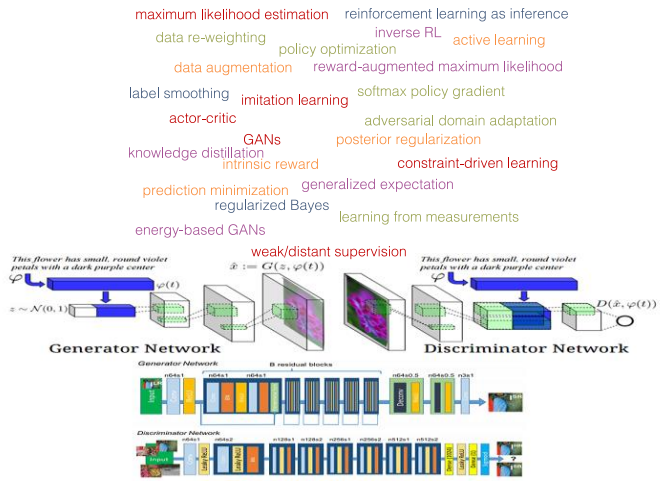
energy-based GANs

weak/distant supervision

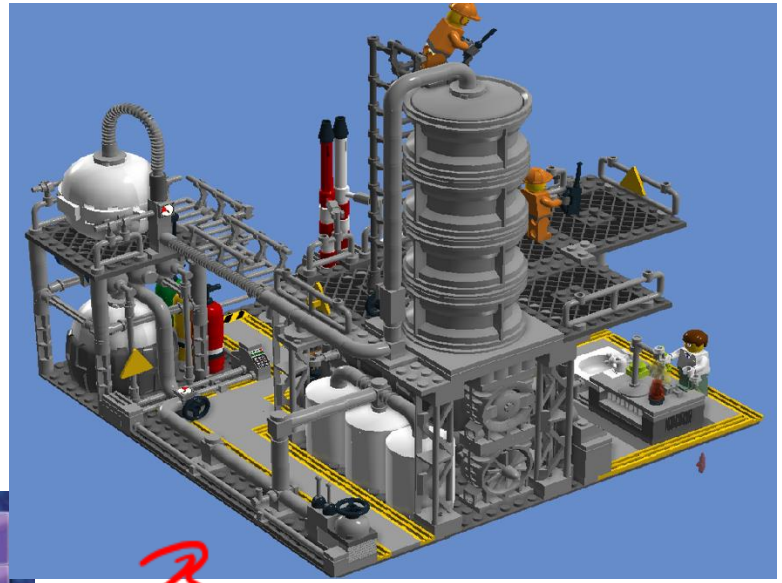
# Where we are now? Where we want to be?

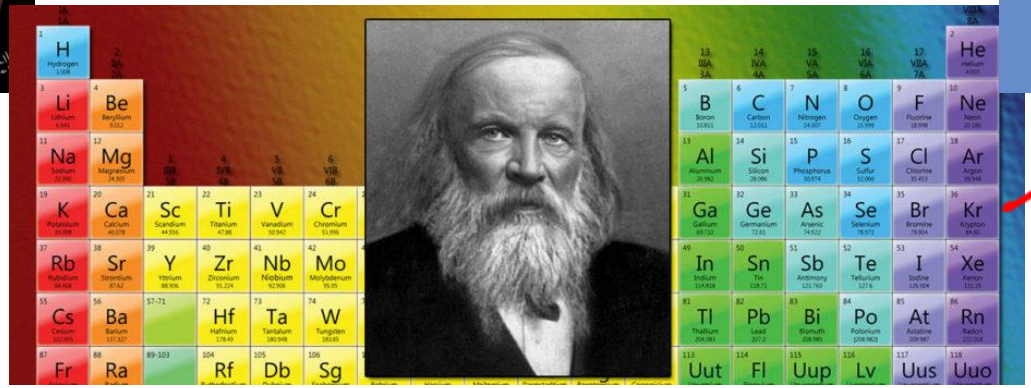
- Alchemy vs chemistry


≈




≠≠





# Quest for more standardized, unified ML principles

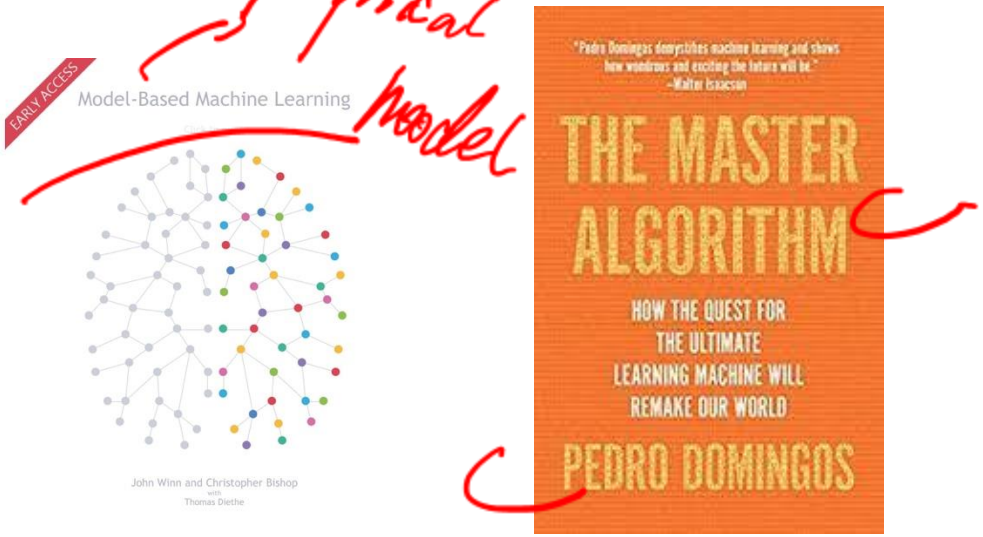
Machine Learning 3: 253–259, 1989


EDITORIAL

Toward a Unified Science of Machine Learning

[P. Langley, 1989]

Graphical model




*REVIEW* Communicated by Steven Nowlan

A Unifying Review of Linear Gaussian Models

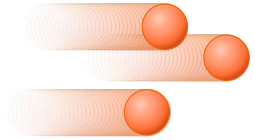
Sam Roweis*
*Computation and Neural Systems, California Institute of Technology, Pasadena, CA 91125, U.S.A.*

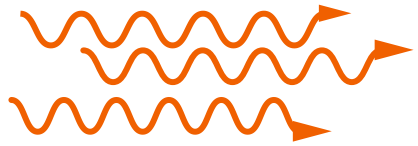
Zoubin Ghahramani*
*Department of Computer Science, University of Toronto, Toronto, Canada*

# Physics in the 1800's

- Electricity & magnetism:
  - Coulomb's law, Ampère, Faraday, ...

- Theory of light beams:
  - Particle theory: Isaac Newton, Laplace, Plank
  - Wave theory: Grimaldi, Chris Huygens, Thomas Young, Maxwell





- Law of gravity
  - Aristotle, Galileo, Newton, …

# “Standard equations” in Physics

EM + weak + strong

*Diverse electro-magnetic theories*

*Maxwell’s Eqns: original form*

| | | |
|---|---|---|
| $e+\frac{df}{dx}+\frac{dg}{dy}+\frac{dh}{dz}=0$ | (1) | Gauss’ Law |
| $\mu\alpha=\frac{dH}{dy}-\frac{dG}{dz}$ <br> $\mu\beta=\frac{dF}{dz}-\frac{dH}{dx}$ <br> $\mu\gamma=\frac{dG}{dx}-\frac{dF}{dy}$ | (2) | Equivalent to Gauss’ Law for magnetism |
| $P=\mu\left(\gamma\frac{dy}{dt}-\beta\frac{dz}{dt}\right)-\frac{dF}{dt}-\frac{d\Psi}{dx}$ <br> $Q=\mu\left(\alpha\frac{dz}{dt}-\gamma\frac{dx}{dt}\right)-\frac{dG}{dt}-\frac{d\Psi}{dy}$ <br> $R=\mu\left(\beta\frac{dx}{dt}-\alpha\frac{dy}{dt}\right)-\frac{dH}{dt}-\frac{d\Psi}{dz}$ | (3) | Faraday’s Law (with the Lorentz Force and Poisson’s Law) |
| $\frac{dy}{dy}-\frac{d\beta}{dz}=4\pi p'\quad p'=p+\frac{df}{dt}$ <br> $\frac{d\alpha}{dz}-\frac{dy}{dx}=4\pi q'\quad q'=q+\frac{dg}{dt}$ <br> $\frac{d\beta}{dx}-\frac{d\alpha}{dy}=4\pi r'\quad r'=r+\frac{dh}{dt}$ | (4) | Ampère-Maxwell Law |
| $P=-\xi p\quad Q=-\xi q\quad R=-\xi r$ | | Ohm’s Law |
| $P=kf\quad Q=kg\quad R=kh$ | | The electric elasticity equation (**E** = **D**/ε) |
| $\frac{de}{dt}+\frac{dp}{dx}+\frac{dq}{dy}+\frac{dr}{dz}=0$ | | Continuity of charge |

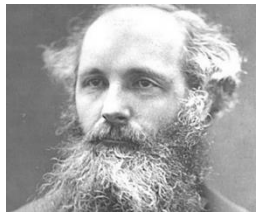
*Maxwell’s Eqns simplified w/ rotational symmetry*

$$\nabla\cdot\mathbf{D}=\rho_V$$

$$\nabla\cdot\mathbf{B}=0$$

$$\nabla\times\mathbf{E}=-\frac{\partial\mathbf{B}}{\partial t}$$

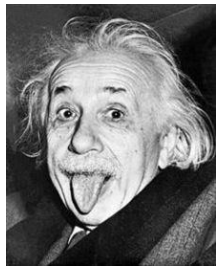
$$\nabla\times\mathbf{H}=\frac{\partial\mathbf{D}}{\partial t}+\mathbf{J}$$



*Maxwell’s Eqns further simplified w/ symmetry of special relativity*

$$\varepsilon^{uvk\lambda}\partial_v F_{k\lambda}=0$$

$$\partial_v F^{uV}=\frac{4\pi}{c}j^u$$



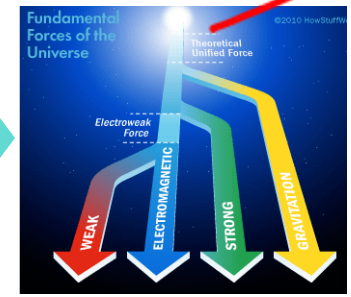
*Standard Model w/ Yang-Mills theory and US(3) symmetry*

$$\mathcal{L}_{\text{gf}}=-\frac{1}{2}\,\text{Tr}(F^2)$$

$$=-\frac{1}{4}F^{a\mu\nu}F^a_{\mu\nu}$$

*Unification of fundamental forces?*

Gravity




1861 1910s 1970s ?

# A "standard model" of ML

Data examples

Type-2 diabetes is 90% more common than type-1

Constraints


Rewards

Auxiliary agents

Adversaries


Imitation

$$\min_{q}\ -\mathbb{H} + \mathbb{D} - \mathbb{E}_{\theta}$$

Uncertainty | Divergence | Experience

- Panoramically learn from all types of experience
- Subsumes many existing algorithms as special cases

Will discuss in later in the class

# Lecture Schedule (tentative)

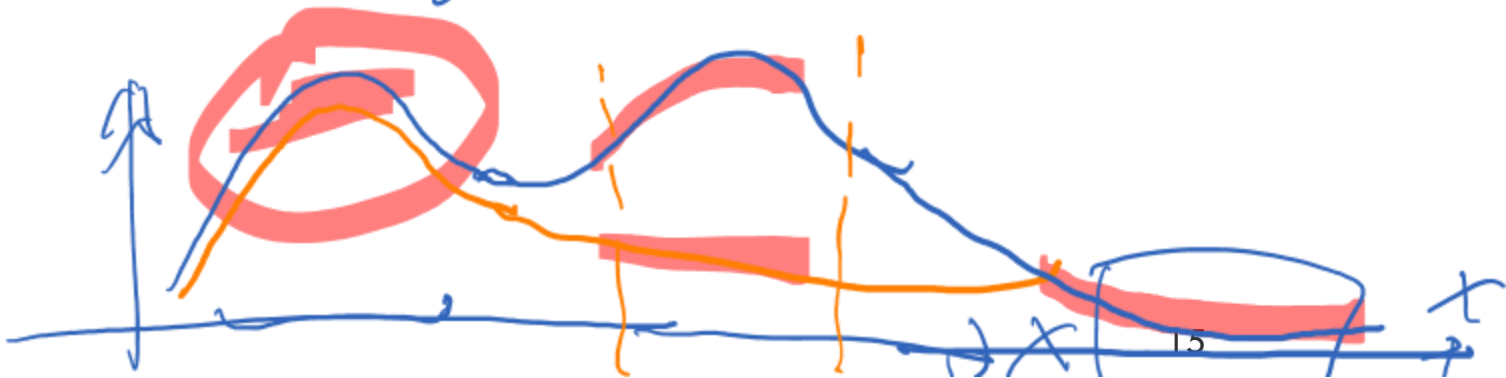
# Supervised Learning

# KL Divergence

- Kullback-Leibler (KL) divergence: measures the closeness of two distributions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$

$$\mathrm{KL}(q(\boldsymbol{x}) \,||\, p(\boldsymbol{x})) = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}$$

  - a.k.a. Relative entropy
  - KL >= 0 (Jensen's inequality) -> homework
  - **Questions:**
    - If $q$ is high and $p$ is high in a region, then KL divergence is _____ in this region.
    - If $q$ is high and $p$ is low in a region, then KL divergence is _____ in this region.
    - If $q$ is low in a region, then KL divergence is _____ in this region.



# KL Divergence

- Kullback-Leibler (KL) divergence: measures the closeness of two distributions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$

$$\mathrm{KL}(q(\boldsymbol{x}) \,||\, p(\boldsymbol{x})) = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}$$

  - a.k.a. Relative entropy
  - KL >= 0 (Jensen's inequality)
  - Intuitively:
    - If $q$ is high and $p$ is high, then we are happy (i.e. low KL divergence)
    - If $q$ is high and $p$ is low then we pay a price (i.e. high KL divergence).
    - If $q$ is low then we don't care (i.e. also low KL divergence, regardless of $p$)
  - not a true "distance":
    - not commutative (symmetric) $\mathrm{KL}(\mathrm{p}||\mathrm{q}) != \mathrm{KL}(\mathrm{q}||\mathrm{p})$
    - doesn't satisfy triangle inequality

# Supervised Learning

- Model to be learned $p_\theta(\boldsymbol{x})$
- Observe **full** data $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^N$
  - e.g., $\boldsymbol{x}_i$ includes both input (e.g., image) and output (e.g., image label)
  - $\mathcal{D}$ defines an empirical data distribution $\tilde{p}(\boldsymbol{x})$
    - $\boldsymbol{x} \sim \mathcal{D} \Leftrightarrow \boldsymbol{x} \sim \tilde{p}(\boldsymbol{x})$
- Maximum Likelihood Estimation (MLE)
  - The most classical learning algorithm

$$\min_\theta - \mathbb{E}_{x\sim\tilde{p}(x)}\left[\log p_\theta(\boldsymbol{x})\right]$$

# Supervised Learning

- Model to be learned $p_\theta(\boldsymbol{x})$
- Observe **full** data $\mathcal{D} = \{ \boldsymbol{x}_i \}_{i=1}^N$
  - e.g., $\boldsymbol{x}_i$ includes both input (e.g., image) and output (e.g., image label)
  - $\mathcal{D}$ defines an empirical data distribution $\tilde{p}(\boldsymbol{x})$
    - $\boldsymbol{x} \sim \mathcal{D} \;\Leftrightarrow\; \boldsymbol{x} \sim \tilde{p}(\boldsymbol{x})$
- Maximum Likelihood Estimation (MLE)
  - The most classical learning algorithm

$$\min_\theta - \mathbb{E}_{x \sim \tilde{p}(x)} \left[ \log p_\theta(\boldsymbol{x}) \right]$$

- **Question:** Show that MLE is minimizing the KL divergence between the empirical data distribution and the model distribution

# Supervised Learning

- Model to be learned $p_\theta(\boldsymbol{x})$
- Observe **full** data $\mathcal{D} = \{ \boldsymbol{x}_i \}_{i=1}^N$
  - e.g., $\boldsymbol{x}_i$ includes both input (e.g., image) and output (e.g., image label)
  - $\mathcal{D}$ defines an empirical data distribution $\tilde{p}(\boldsymbol{x})$
    - $\boldsymbol{x} \sim \mathcal{D} \Leftrightarrow \boldsymbol{x} \sim \tilde{p}(\boldsymbol{x})$

- Maximum Likelihood Estimation (MLE)
  - The most classical learning algorithm

$$\min_\theta - \mathbb{E}_{x \sim \tilde{p}(x)} \left[ \log p_\theta(\boldsymbol{x}) \right]$$

- **Question:** Show that MLE is minimizing the KL divergence between the empirical data distribution and the model distribution

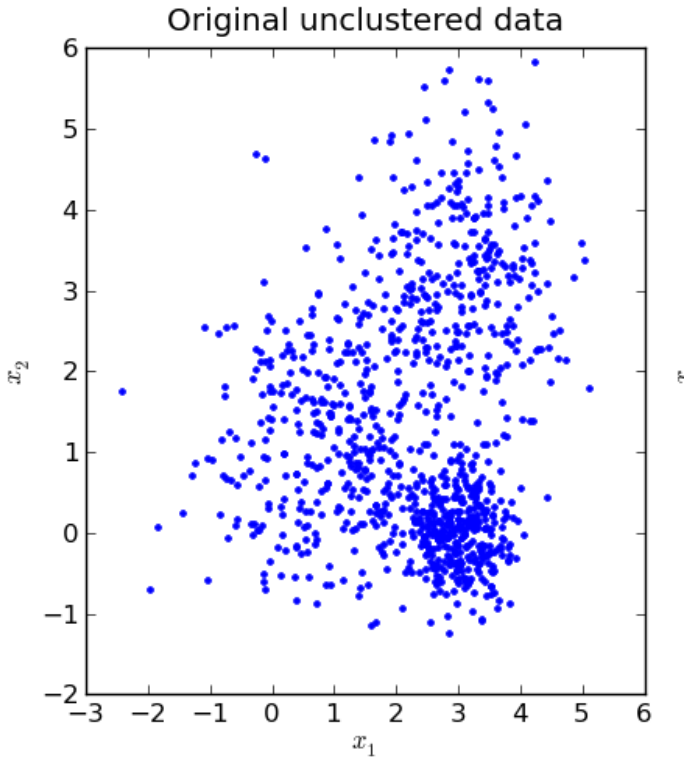
$$\mathrm{KL}(\tilde{p}(\boldsymbol{x}) \,||\, p_\theta(\boldsymbol{x})) = -\mathbb{E}_{\tilde{p}(x)}[\log p_\theta(\boldsymbol{x})] + H(\tilde{p}(\boldsymbol{x}))$$

Cross entropy

# Unsupervised Learning

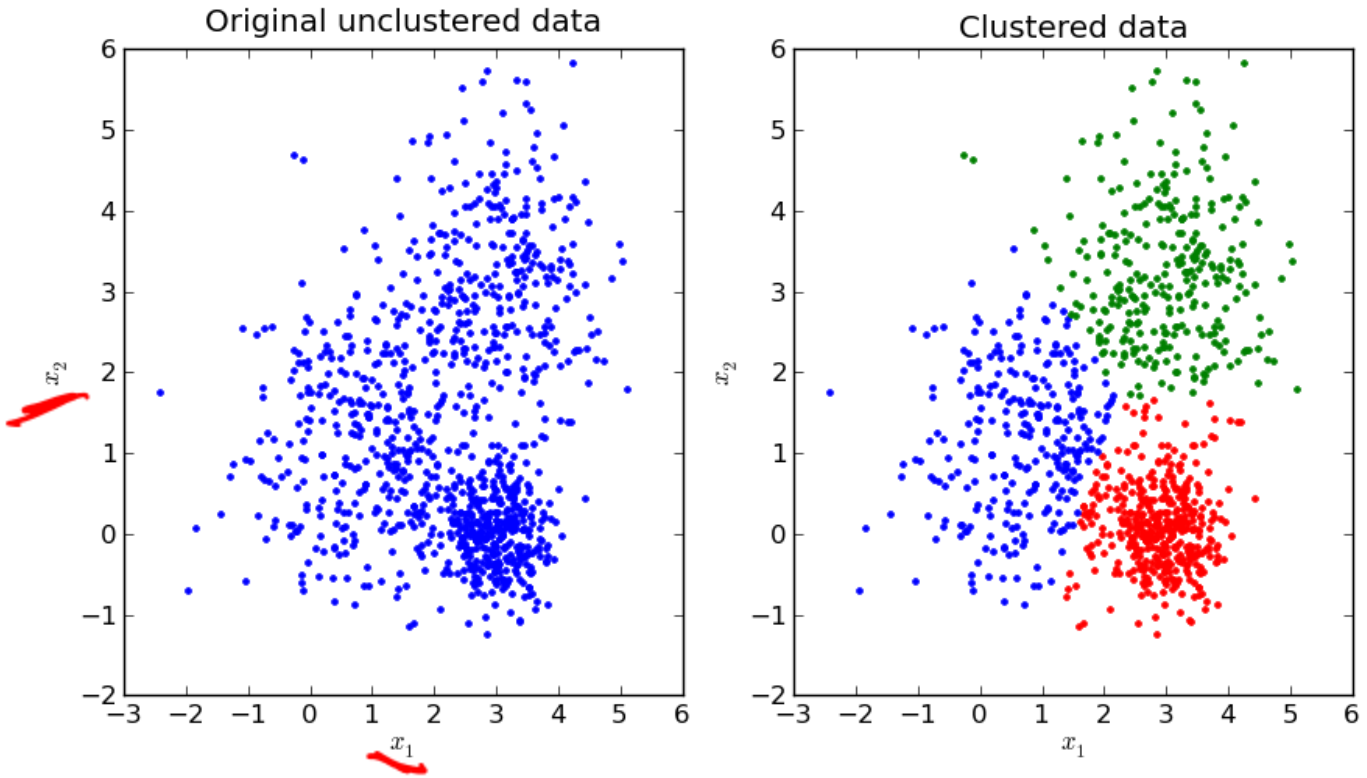
# Unsupervised Learning

- Each data instance is partitioned into two parts:
  - observed variables $\boldsymbol{x}$
  - latent (unobserved) variables $\boldsymbol{z}$
- Want to learn a model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$



# Unsupervised Learning

- Each data instance is partitioned into two parts:
  - observed variables $\boldsymbol{x}$
  - latent (unobserved) variables $\boldsymbol{z}$
- Want to learn a model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$

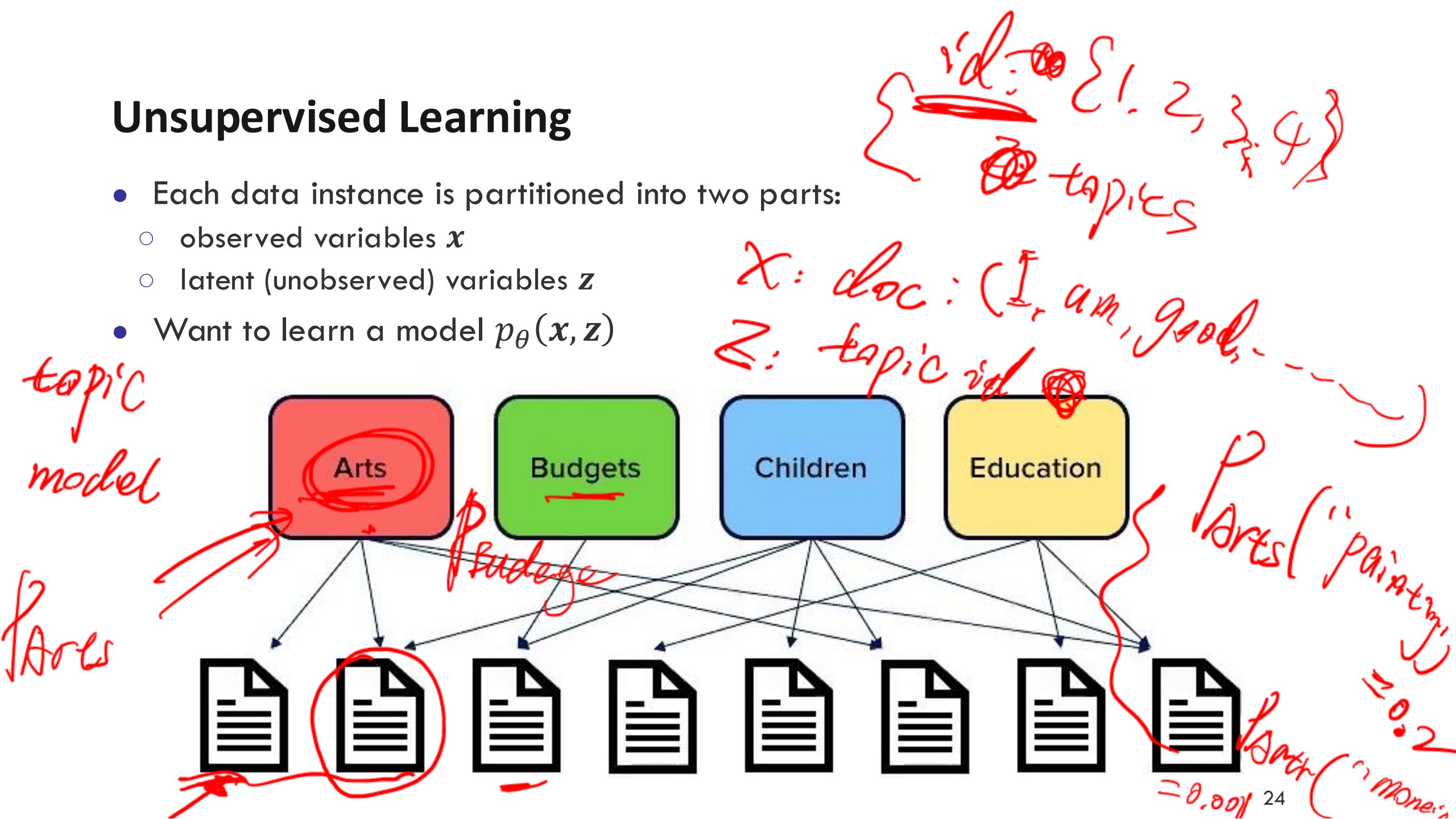


# Unsupervised Learning

- Each data instance is partitioned into two parts:
  - observed variables $\boldsymbol{x}$
  - latent (unobserved) variables $\boldsymbol{z}$
- Want to learn a model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$

# Unsupervised Learning

- Each data instance is partitioned into two parts:
  - observed variables $\boldsymbol{x}$
  - latent (unobserved) variables $\boldsymbol{z}$
- Want to learn a model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$

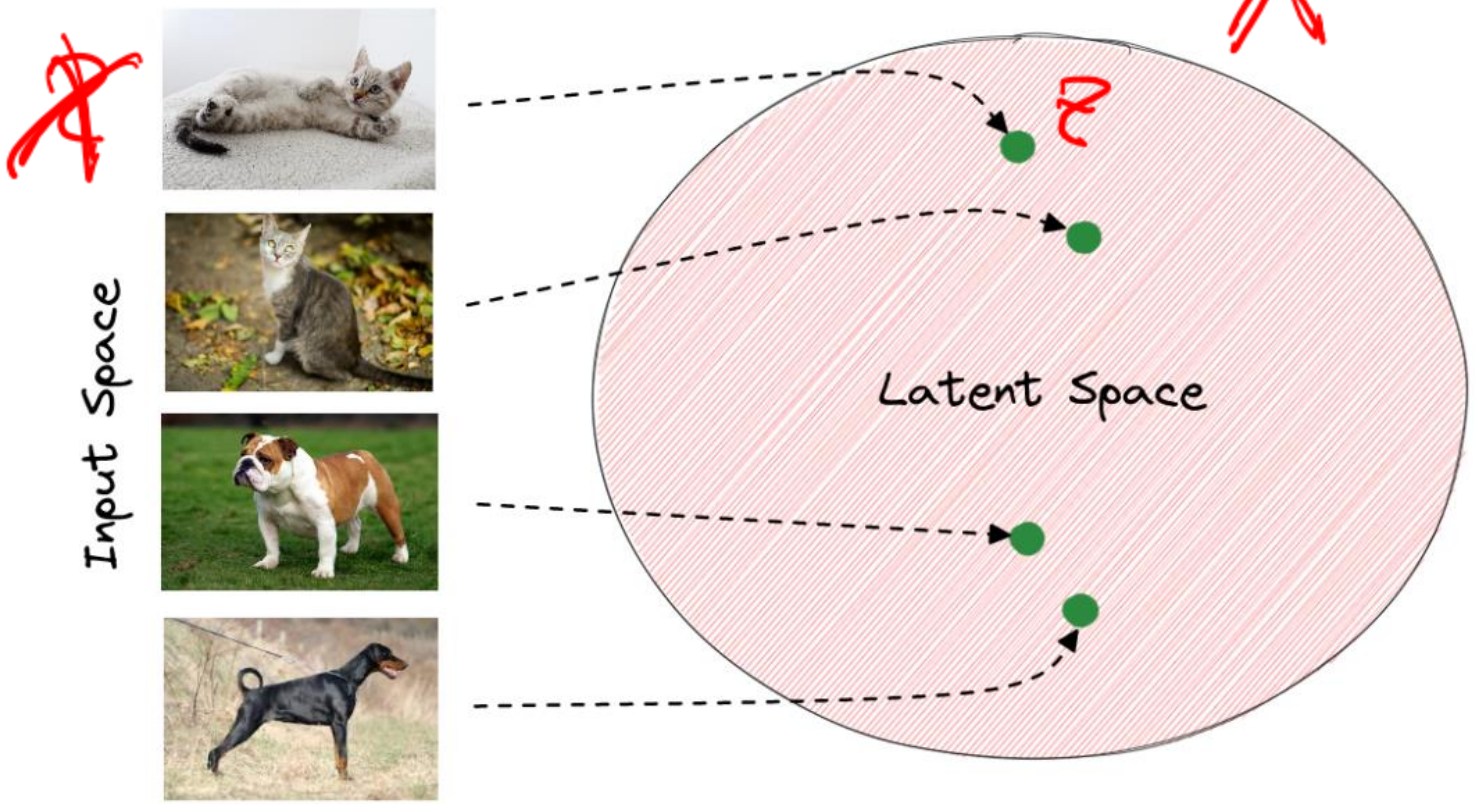


# Unsupervised Learning

- Each data instance is partitioned into two parts:
  - observed variables $\boldsymbol{x}$
  - latent (unobserved) variables $\boldsymbol{z}$
- Want to learn a model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$

# Unsupervised Learning

- Each data instance is partitioned into two parts:
  - observed variables $\boldsymbol{x}$
  - latent (unobserved) variables $\boldsymbol{z}$
- Want to learn a model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$



# Why is Unsupervised Learning Harder?

- **Complete log likelihood:** if both $\boldsymbol{x}$ and $\boldsymbol{z}$ can be observed, then

$$\ell_c(\theta; \boldsymbol{x}, \boldsymbol{z}) = \log p(\boldsymbol{x}, \boldsymbol{z}|\theta) = \log p(\boldsymbol{z}|\theta_z) + \log p(\boldsymbol{x}|\boldsymbol{z}, \theta_x)$$

  - Decomposes into a sum of factors, the parameter for each factor can be estimated separately

Now $\boldsymbol{z}$ is not observed:

- **Incomplete (or marginal) log likelihood:** with $\boldsymbol{z}$ unobserved, our objective becomes the log of a marginal probability:

$$\ell(\theta; \boldsymbol{x}) = \log p(\boldsymbol{x}|\theta)$$

# Why is Unsupervised Learning Harder?

- **Complete log likelihood:** if both $\boldsymbol{x}$ and $\boldsymbol{z}$ can be observed, then

$$\ell_c(\theta; \boldsymbol{x}, \boldsymbol{z}) = \log p(\boldsymbol{x}, \boldsymbol{z}|\theta) = \log p(\boldsymbol{z}|\theta_z) + \log p(\boldsymbol{x}|\boldsymbol{z}, \theta_x)$$

  - Decomposes into a sum of factors, the parameter for each factor can be estimated separately

Now $\boldsymbol{z}$ is not observed:

- **Incomplete (or marginal) log likelihood:** with $\boldsymbol{z}$ unobserved, our objective becomes the log of a marginal probability:

$$\ell(\theta; \boldsymbol{x}) = \log p(\boldsymbol{x}|\theta) = \log \sum_z p(\boldsymbol{x}, \boldsymbol{z}|\theta)$$

  - All parameters become coupled together
  - In other models when $\boldsymbol{z}$ is complex (continuous) variables (as we'll see later), marginalization over $\boldsymbol{z}$ is intractable.

# Expectation Maximization (EM): Intuition

# Expectation Maximization (EM): Intuition

- Supervised MLE is easy: $\max_\theta \ \ell_c(\theta; \boldsymbol{x}, \boldsymbol{z}) = \log p(\boldsymbol{x}, \boldsymbol{z}|\theta)$
  - Observe both $\boldsymbol{x}$ and $\boldsymbol{z}$
- Unsupervised MLE is hard: $\max_\theta \ \ell(\theta; \boldsymbol{x}) = \log p(\boldsymbol{x}|\theta) = \log \sum_z p(\boldsymbol{x}, \boldsymbol{z}|\theta)$
  - Observe only $\boldsymbol{x}$
- EM, intuitively:

**E-step:** $q(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x}, \theta)$

*We don't actually observe q, let's estimate it*

**M-step:** $\max_\theta \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}, \boldsymbol{z}|\theta)]$

*Let's "pretend" we also observe $\boldsymbol{z}$ (its distribution)*

# Expectation Maximization (EM): Intuition

- Supervised MLE is easy: $\max_\theta \ \ell_c(\theta; \boldsymbol{x}, \boldsymbol{z}) = \log p(\boldsymbol{x}, \boldsymbol{z}|\theta)$
  - Observe both $\boldsymbol{x}$ and $\boldsymbol{z}$
- Unsupervised MLE is hard: $\max_\theta \ \ell(\theta; \boldsymbol{x}) = \log p(\boldsymbol{x}|\theta) = \log \sum_z p(\boldsymbol{x}, \boldsymbol{z}|\theta)$
  - Observe only $\boldsymbol{x}$
- EM, intuitively:

**E-step:** $q^{t+1}(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x}, \theta^t)$

*We don't actually observe q, let's estimate it*

**M-step:** $\max_\theta \mathbb{E}_{q^{t+1}(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}, \boldsymbol{z}|\theta)]$

*Let's "pretend" we also observe $\boldsymbol{z}$ (its distribution)*

*This is an iterative process*

# Questions?