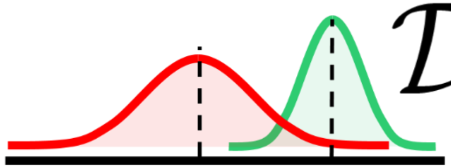
Probabilistic Methods-II

CS771: Introduction to Machine Learning

Purushottam Kar

Recap

The Generative Story



$\overset{?}{\mathcal{D}_{\boldsymbol{\theta}^*}} \Leftarrow \mathbf{Z}$

$\mathbb{P}\left[\mathbf{Z} \mid \boldsymbol{\theta}\right]$	Likelihood
$\mathbb{P}\left[\boldsymbol{\theta}\right]$	Prior
$\mathbb{P}\left[\boldsymbol{\theta} \mid \mathbf{Z}\right]$	Posterior

Defined for all parameters $\boldsymbol{\theta}$

Prediction with learnt models

 $= ?$

$\mathbf{Z} = \mathbf{y} =$ 

$\mathbb{P}\left[p\right]$

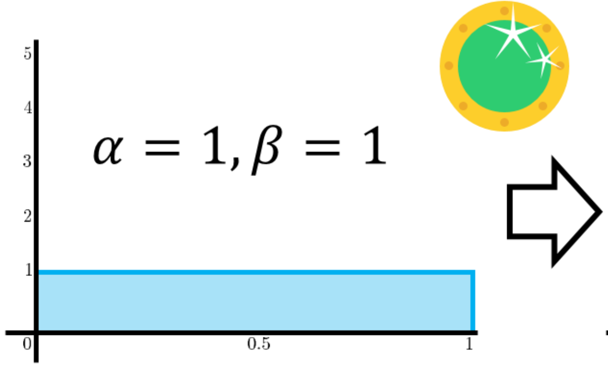
$$\mathbb{P}\left[y^i \mid p\right] = p^{y^i}(1-p)^{1-y^i}$$

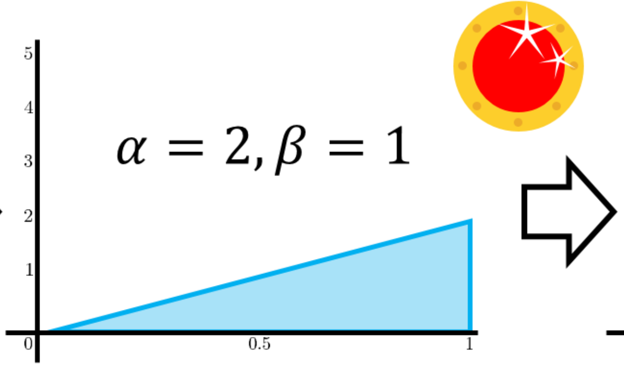
$\mathbb{P}[$  $= H \mid \mathbf{y}] = \int_p \mathbb{P}[$  $= H \mid p]\, \mathbb{P}[p \mid \mathbf{y}]\, dp$

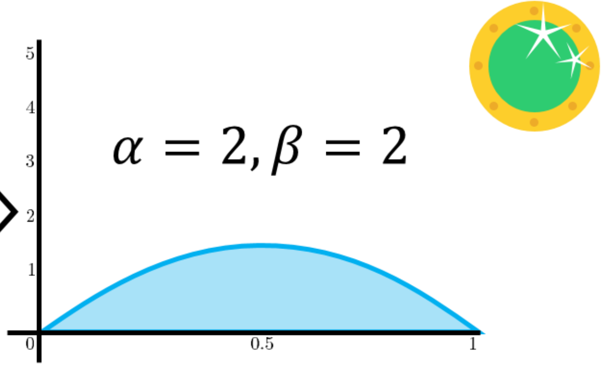
$\approx \mathbb{P}[$  $= H \mid \hat{p}_{\mathrm{MLE}}] = \frac{n_H}{n}$

$\approx \mathbb{P}[$  $= H \mid \hat{p}_{\mathrm{MAP}}] = \frac{n_H + \alpha - 1}{n + \alpha + \beta - 2}$

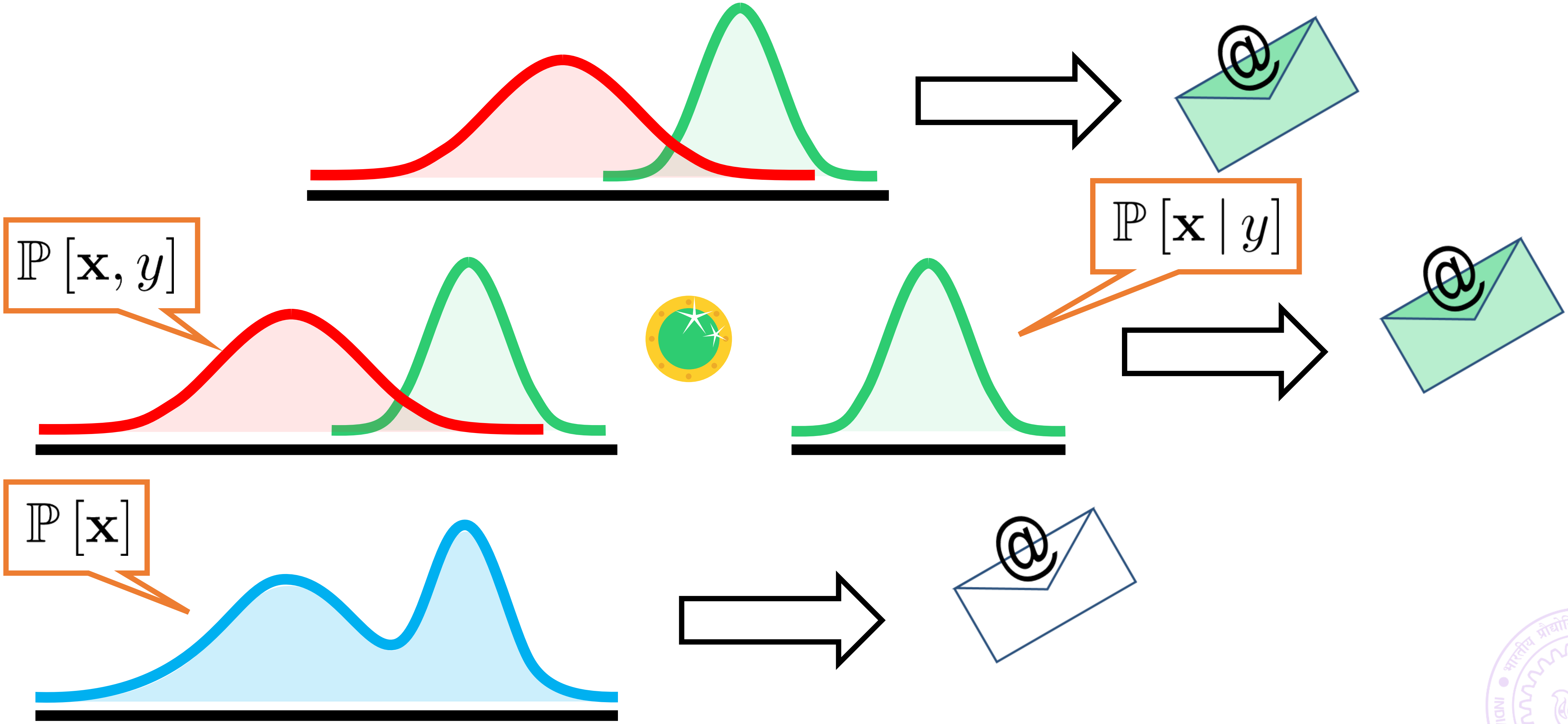
Online MAP!



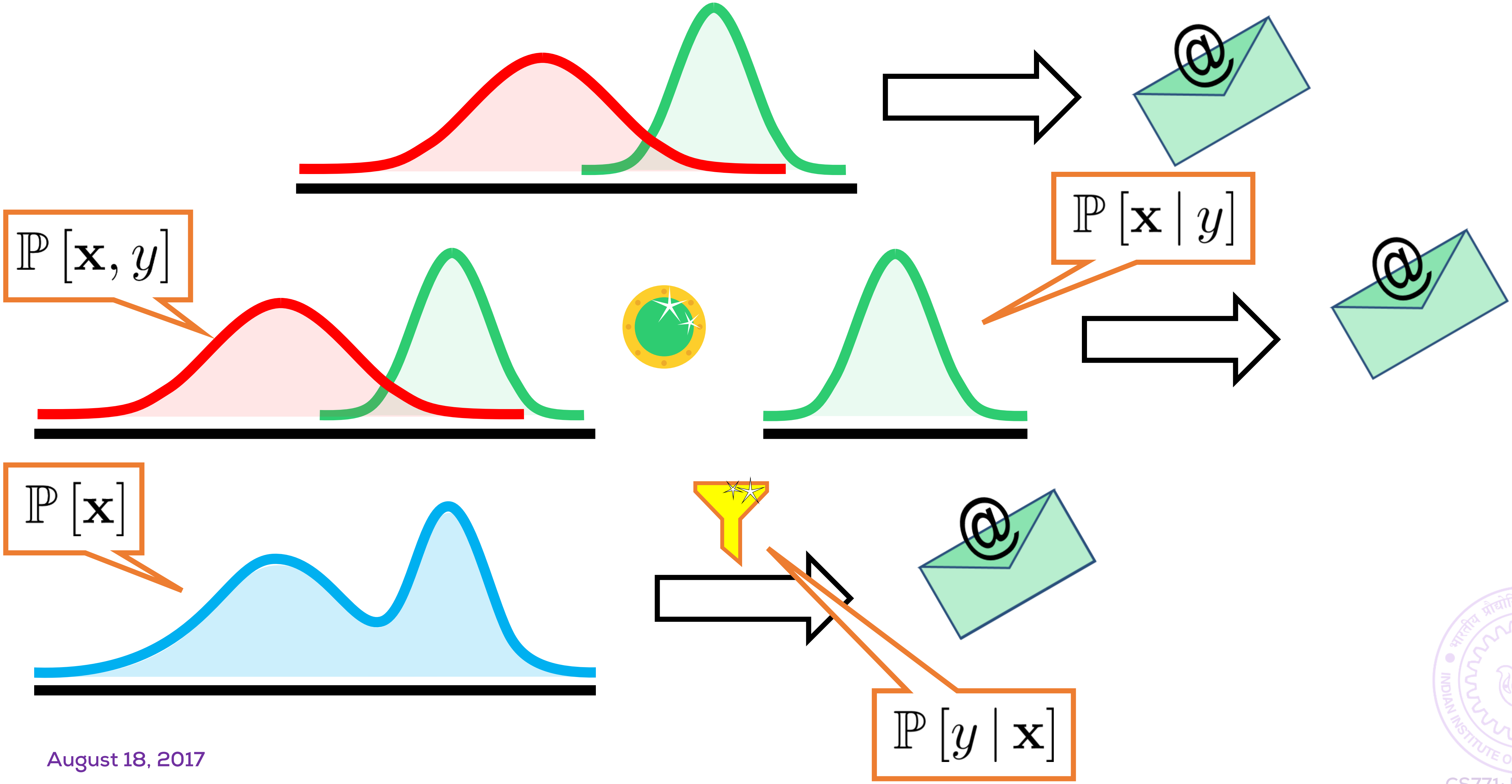




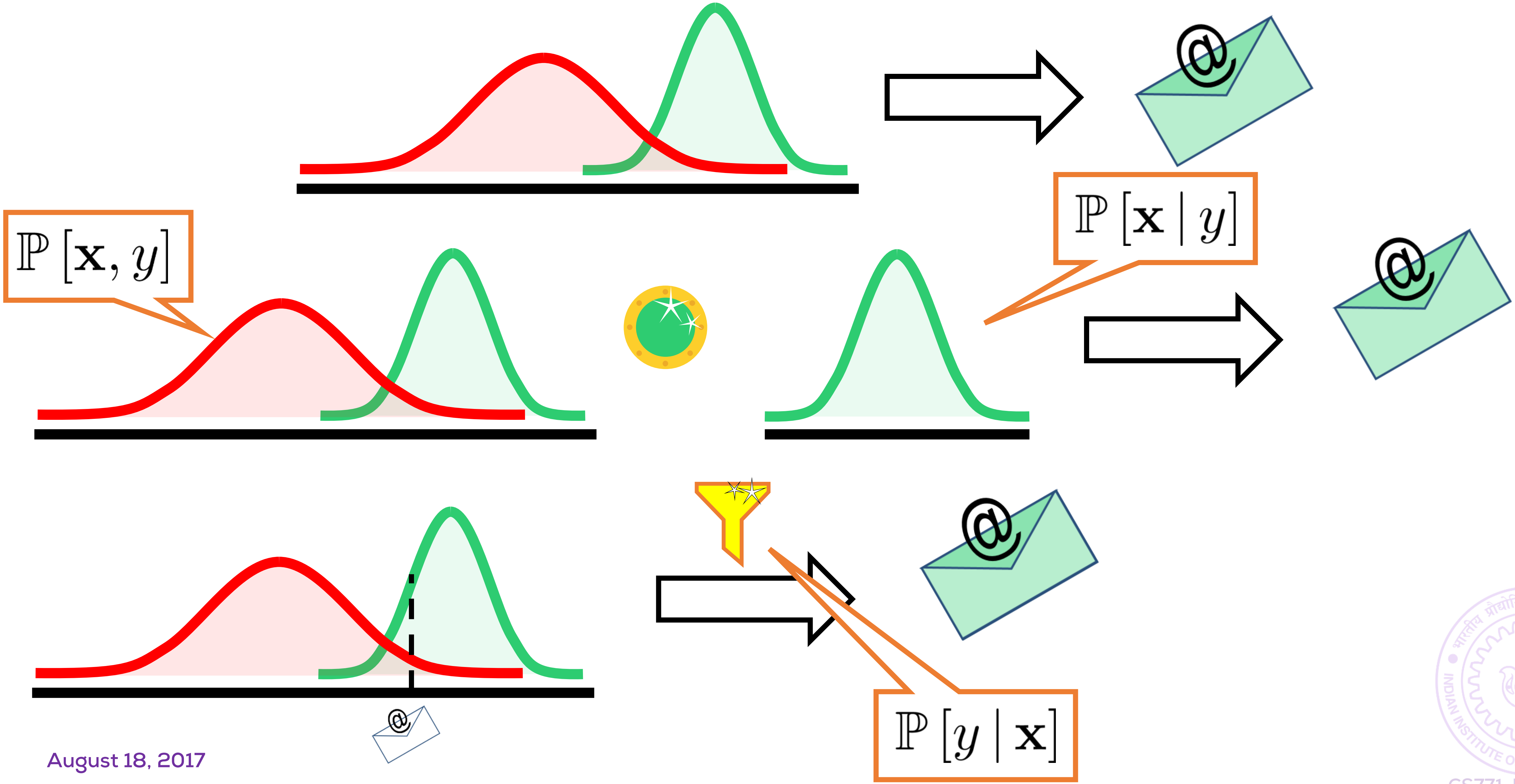
Generative Models for Labeled Data



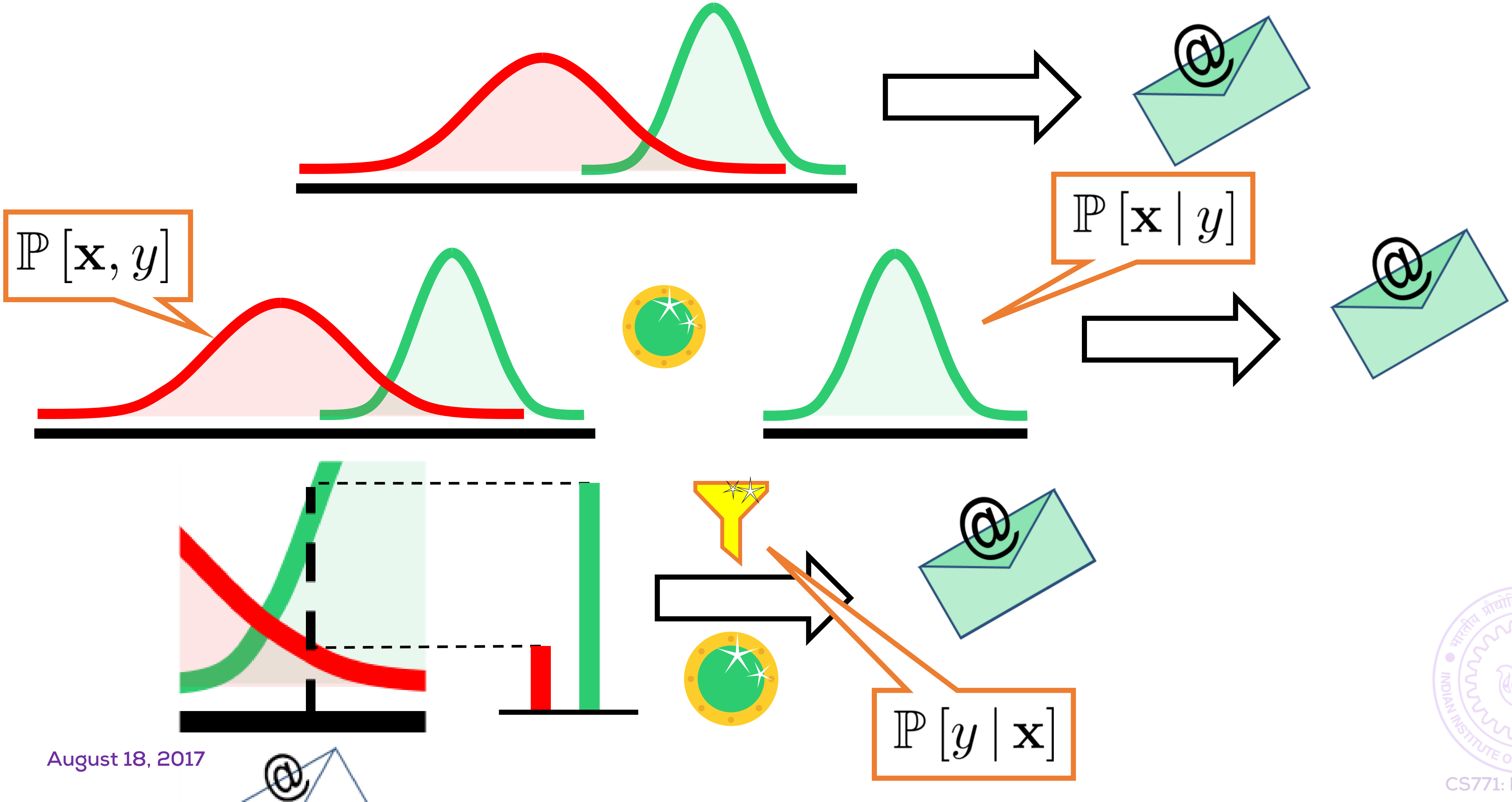
Generative Models for Labeled Data



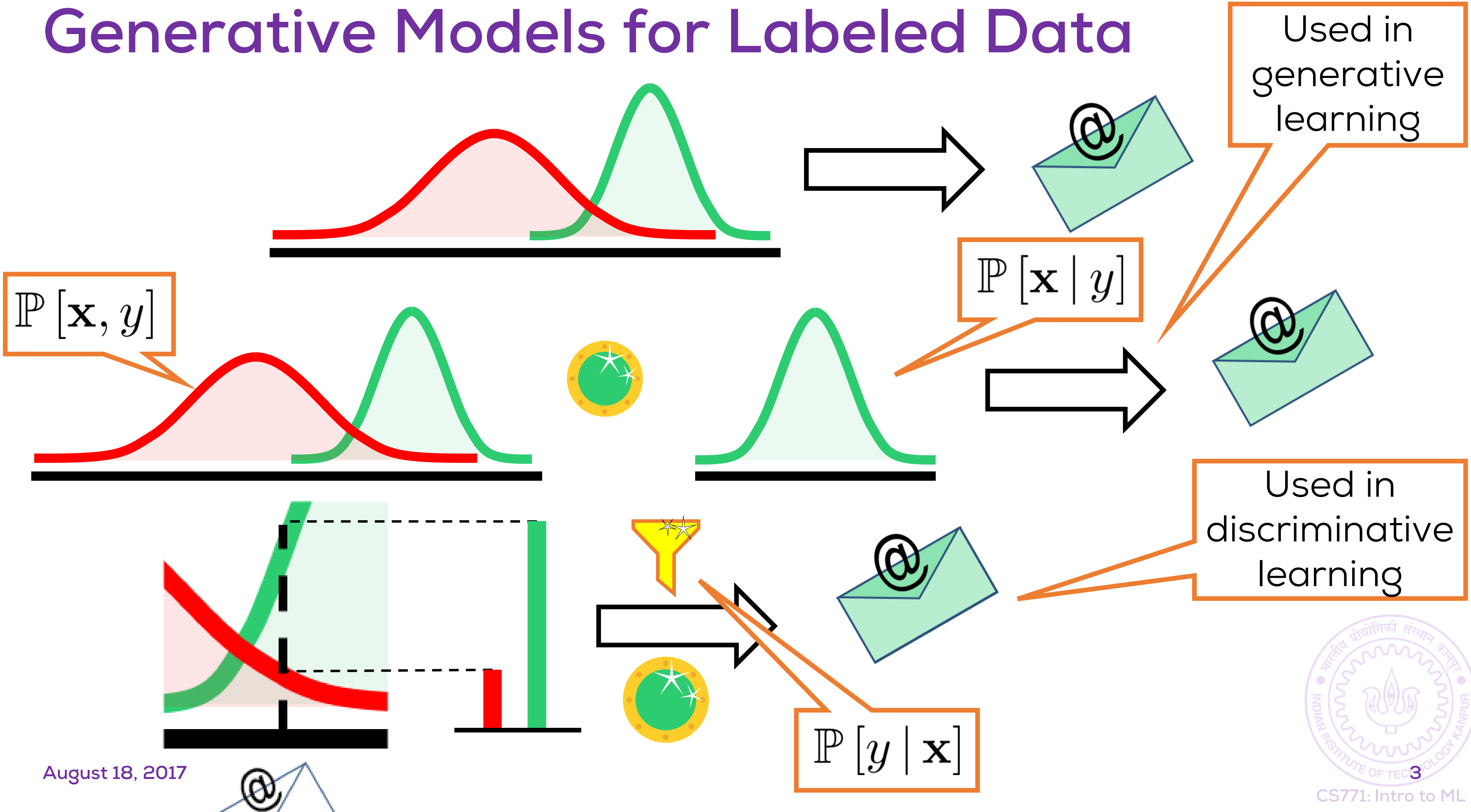
Generative Models for Labeled Data



Generative Models for Labeled Data

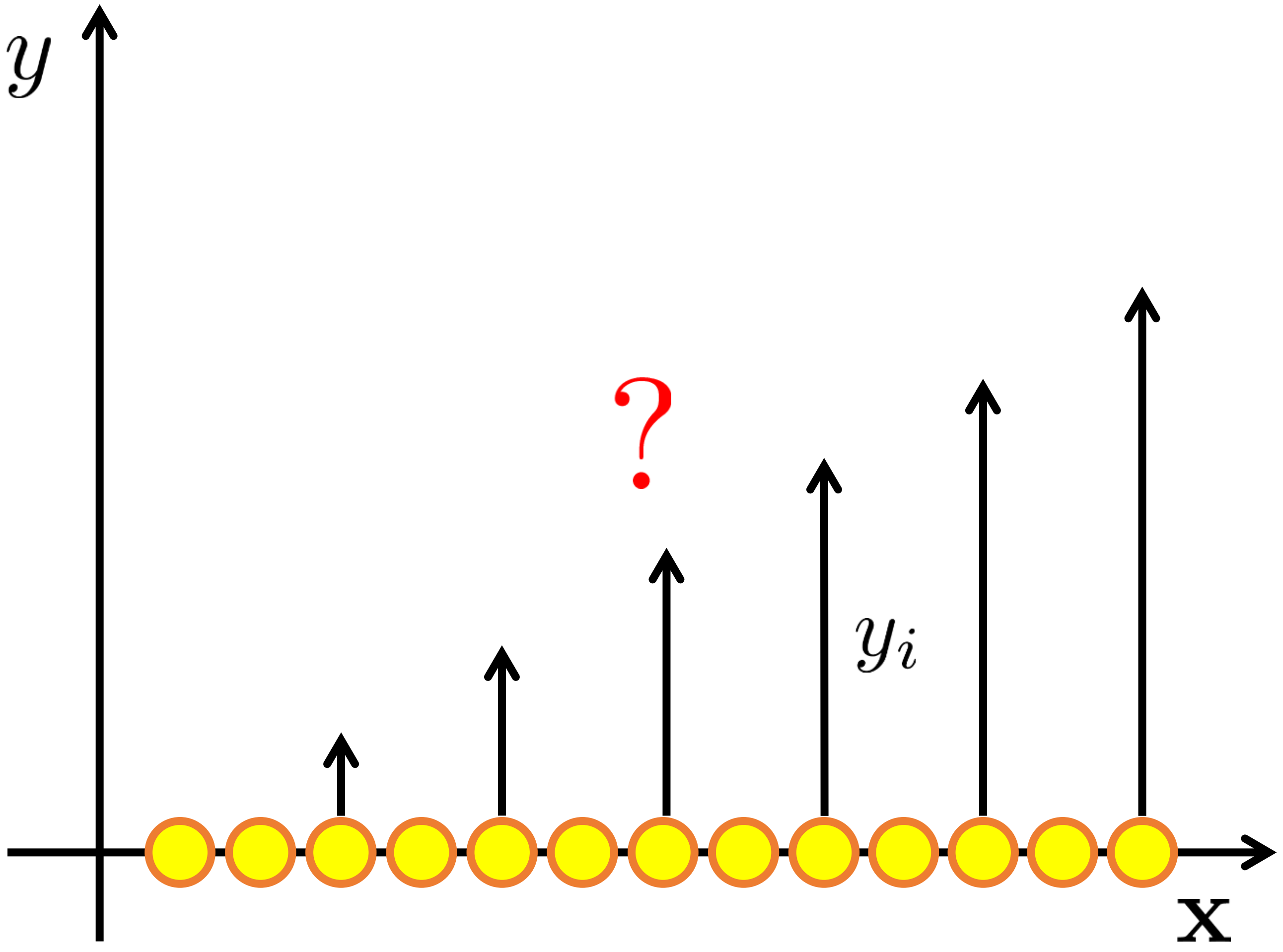


Generative Models for Labeled Data



Regression using PML

Linear Regression



Data: $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n \in \mathbb{R}^d$

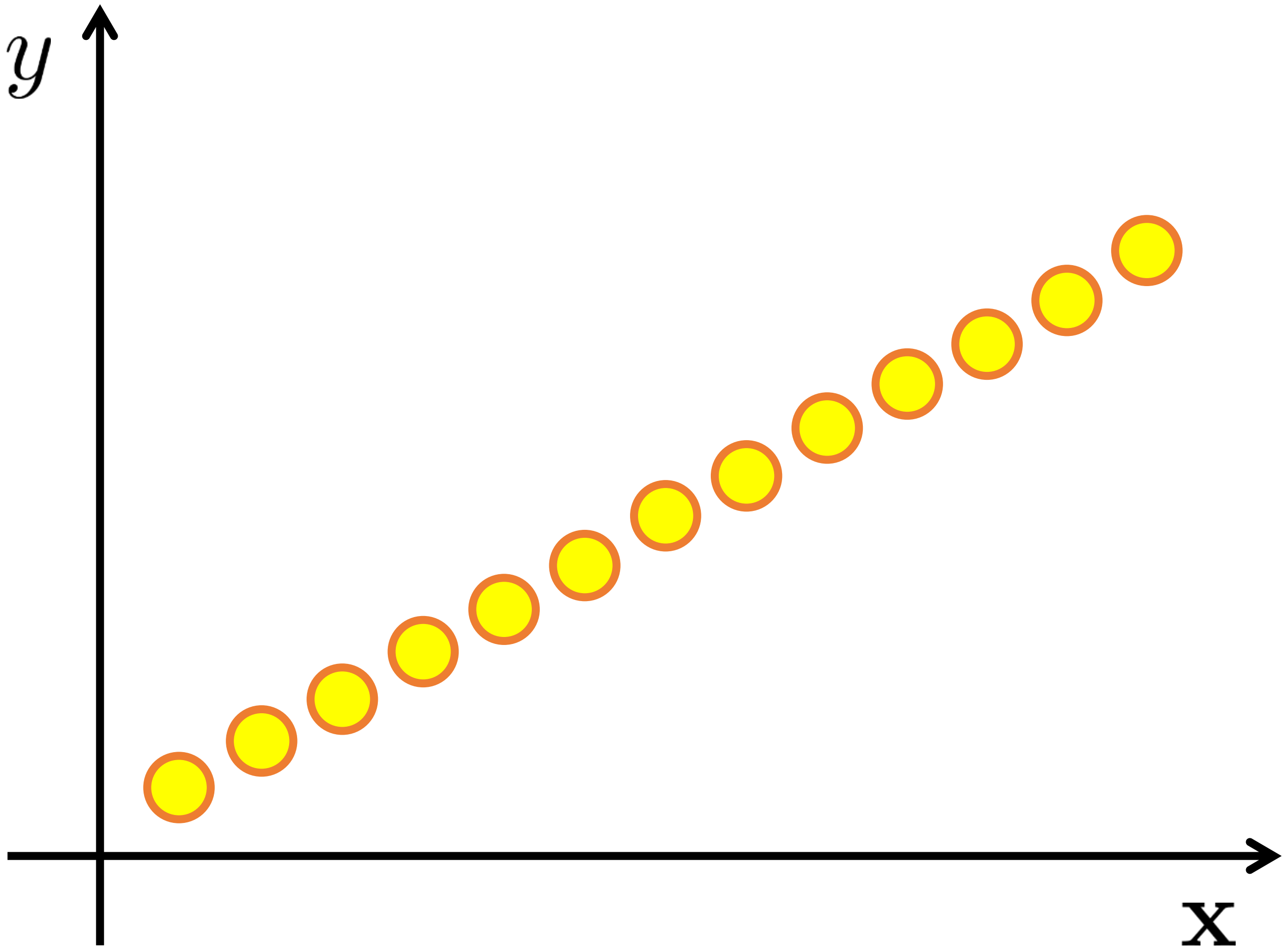
Model: $\mathbf{w}^*$ (hidden)

$$y^i = \langle \mathbf{w}^*, \mathbf{x}^i \rangle$$

Given: $(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^n, y^n)$

Recover $\mathbf{w}^*$?

Linear Regression



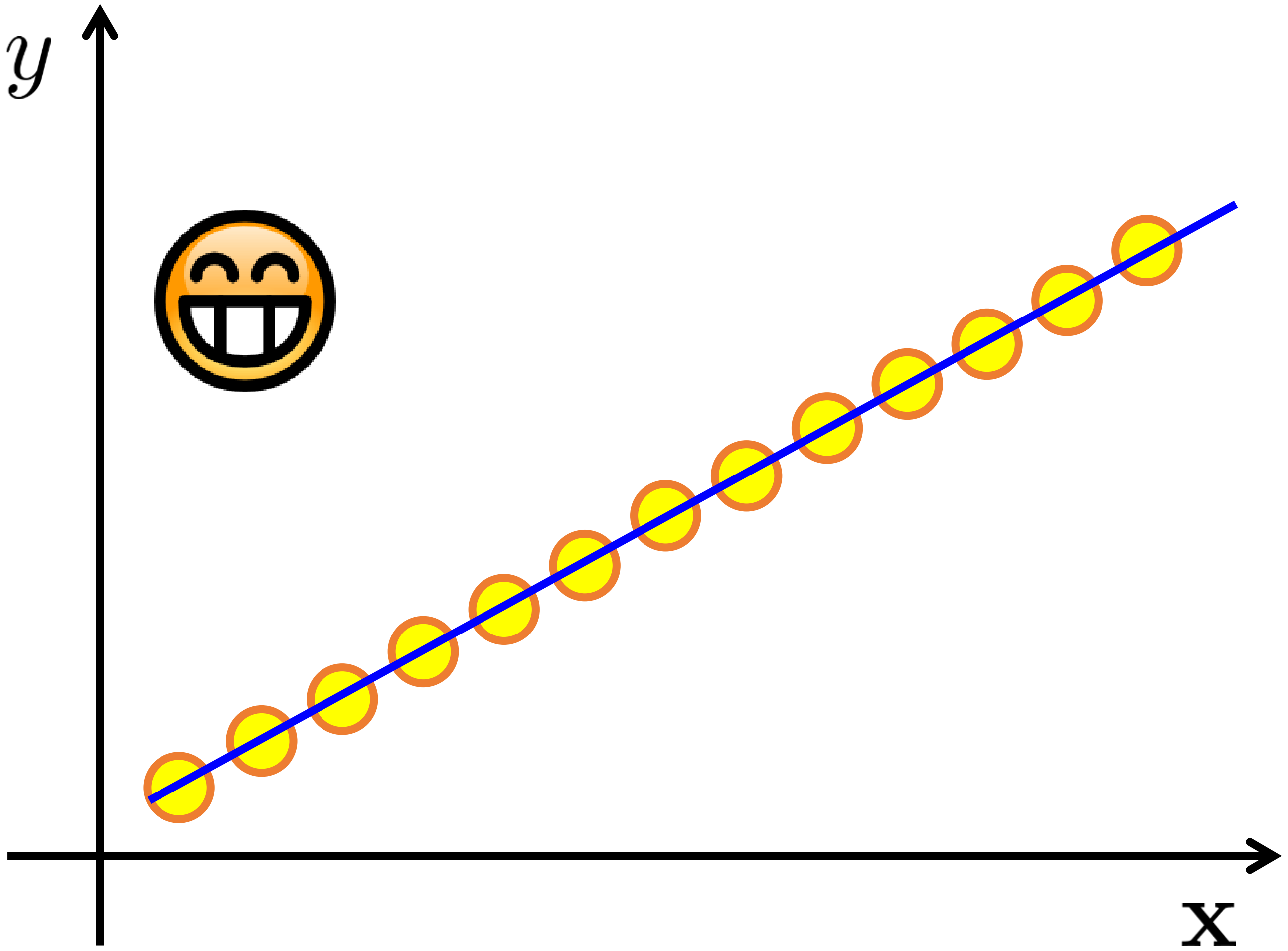
Given: $(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^n, y^n)$

$$y^i = \langle \mathbf{w}^*, \mathbf{x}^i \rangle$$

$$n \quad \underbrace{\begin{bmatrix} y^1 \\ y^2 \\ y^3 \\ \vdots \\ y^n \end{bmatrix}}_{\mathbf{y}} = \underbrace{\overbrace{\begin{bmatrix} \text{---} & \mathbf{x}^1 & \text{---} \\ \text{---} & \mathbf{x}^2 & \text{---} \\ \text{---} & \mathbf{x}^3 & \text{---} \\ & \vdots & \\ \text{---} & \mathbf{x}^n & \text{---} \end{bmatrix}}^{d}}_{\mathbf{X}^\top} \begin{bmatrix} | \\ \mathbf{w} \\ | \end{bmatrix} \quad d$$

?

Linear Regression



Given: $(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^n, y^n)$

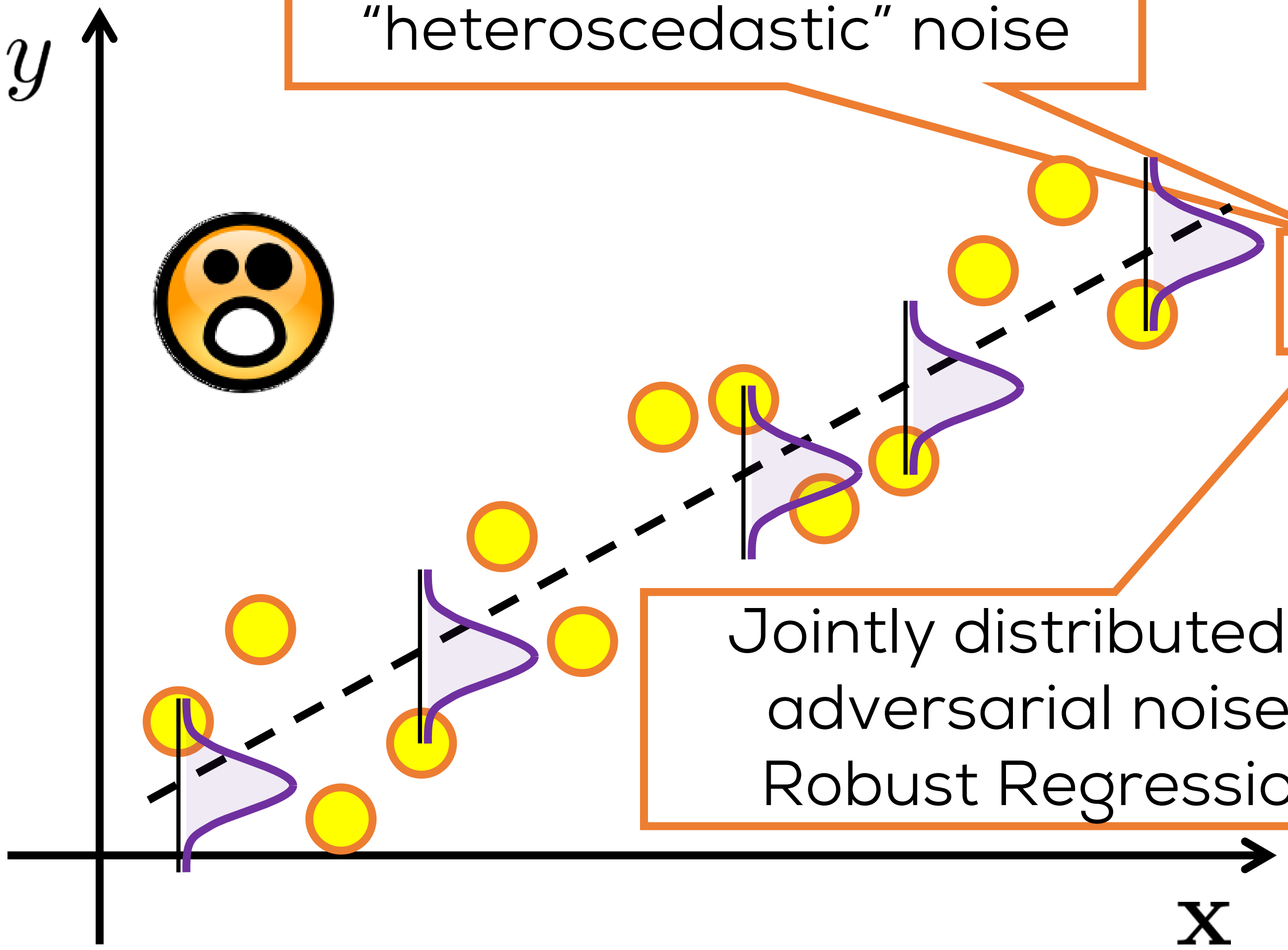
$$y^i = \langle \mathbf{w}^*, \mathbf{x}^i \rangle$$

$$\mathbf{y} = \mathbf{X}^\top \mathbf{w}$$

Linear system!!

$\mathbf{w}^*$ Recovered!!

Linear Regression with Noise



Non-identically distributed "heteroscedastic" noise

Given: $(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^n, y^n)$

$$y^i = \langle \mathbf{w}^*, \mathbf{x}^i \rangle + e^i$$

i.i.d noise

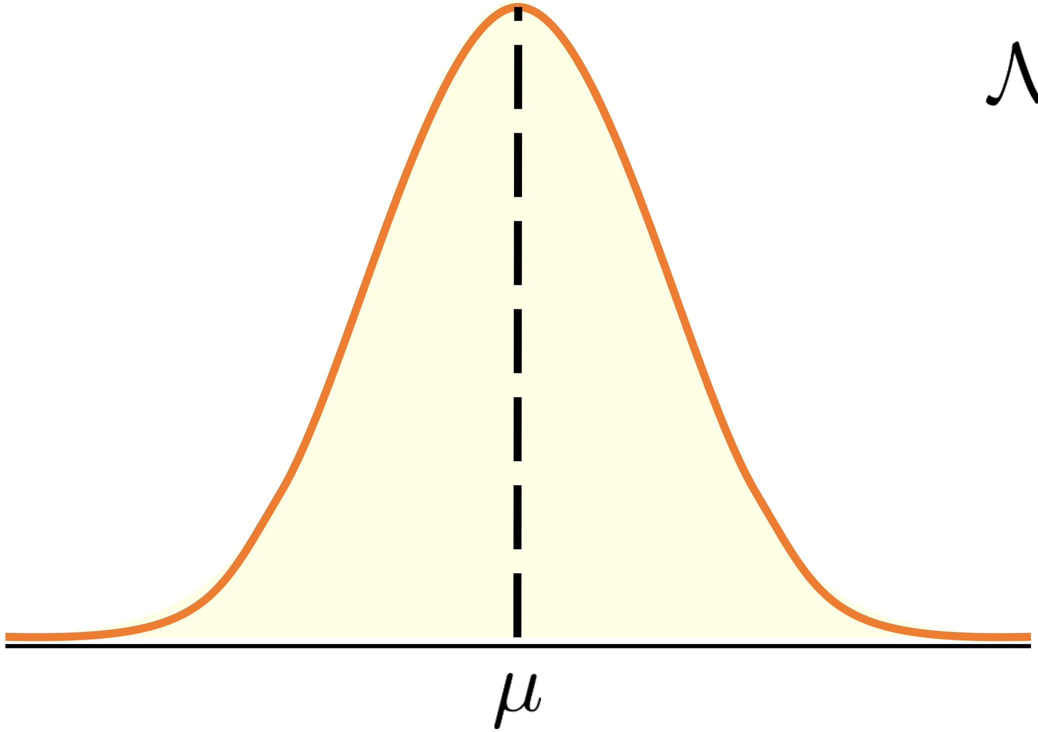
$$e^i \sim \mathcal{N}(0, \sigma^2)$$

$$y^i \sim \mathcal{N}(\langle \mathbf{w}^*, \mathbf{x}^i \rangle, \sigma^2)$$

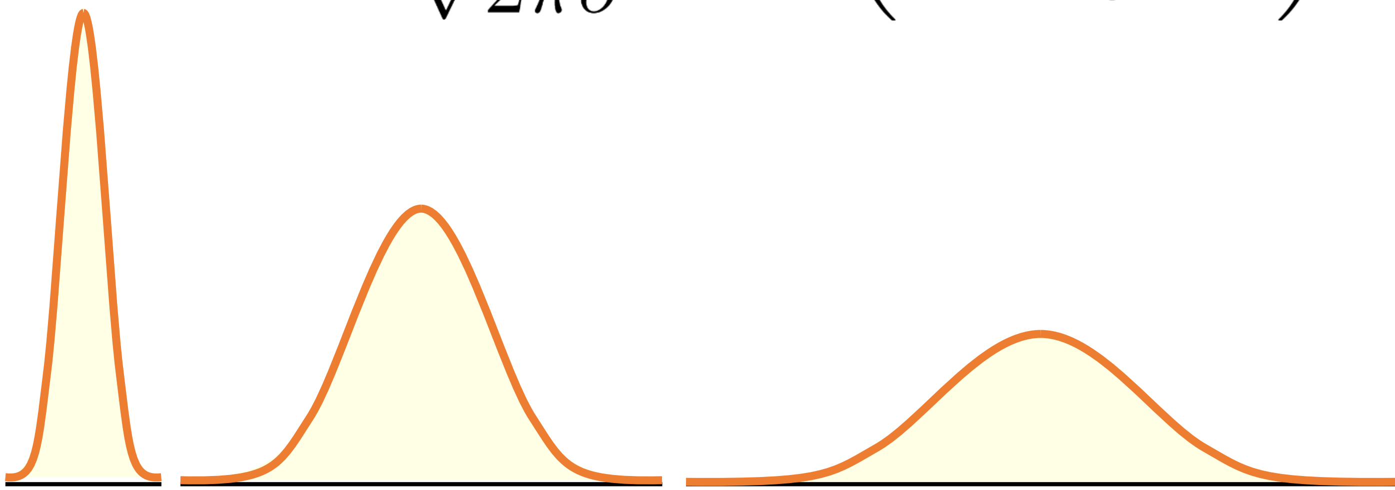
Jointly distributed or adversarial noise?
Robust Regression

Non-Gaussian noise?
Generalized models (GLM)!

The Gaussian Distribution



$$\mathcal{N}(t;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)$$



Multivariate Gaussian

$$\mathcal{N}(\mathbf{v};\boldsymbol{\mu},\Sigma)=\frac{1}{\sqrt{(2\pi)^d\,|\Sigma|}}\exp\left(-\frac{1}{2}(\mathbf{v}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{v}-\boldsymbol{\mu})\right)$$

Linear Regression using MLE

$$\mathbb{P}\left[y \,|\, \mathbf{x}^i, \mathbf{w}\right] = \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2}\right)$$

Linear function

Why isn't $\mathbf{x}$ getting modelled too?

Generative model

Likelihood

Log-likelihood

$$\log \mathbb{P}\left[\mathbf{y} \,|\, \mathbf{X}, \mathbf{w}\right] = C - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2$$

Least Squares!

$$\hat{\mathbf{w}}_{\mathrm{MLE}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 = (\mathbf{X}\mathbf{X}^\top)^\dagger \mathbf{X}\mathbf{y}$$

$\log \mathbb{P}\left[\mathbf{y}, \mathbf{X} \,|\, \boldsymbol{\theta}\right]$ Generative MLE??

Exercise

Linear Regression using MAP

$$\mathbb{P}\left[\mathbf{w}\right] = \mathcal{N}(\mathbf{0}, \rho^2 \cdot I_d) = \frac{1}{\sqrt{(2\pi)^d \rho^2}} \exp\left(-\frac{\|\mathbf{w}\|_2^2}{2\rho^2}\right)$$

Sparsity inducing priors

$$\log \mathbb{P}\left[\mathbf{w} \mid \mathbf{X}, \mathbf{y}\right] = C - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 - \frac{1}{2\rho^2}\|\mathbf{w}\|_2^2$$

$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\min \sum_{i=1}^{n}\left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \frac{\sigma^2}{\rho^2}\|\mathbf{w}\|_2^2 = (\mathbf{X}\mathbf{X}^\top + \lambda I)^{-1}\mathbf{X}\mathbf{y}$$

$\log \mathbb{P}\left[\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}\right]$
Generative MAP??

Ridge Regression

Exercise

Bayesian Linear Regression?

$$\mathbb{P}\left[\mathbf{w} \,|\, \mathbf{X}, \mathbf{y}\right] = \mathcal{N}(\boldsymbol{\nu}, \Lambda)$$

Cool! Posterior is Gaussian like the prior!

Wait! MAP?

$$\boldsymbol{\nu} = \left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\rho^2}\cdot I\right)^{-1}\mathbf{X}\mathbf{y}$$

$$\Lambda = \left(\frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^\top + \frac{1}{\rho^2}\cdot I\right)^{-1}$$

By definition, MAP is the mode of the posterior

$$\mathbb{P}\left[y \,|\, \mathbf{x}, \mathbf{X}, \mathbf{y}\right] = \int_{\mathbf{w}} \mathbb{P}\left[y \,|\, \mathbf{x}, \mathbf{w}\right]\mathbb{P}\left[\mathbf{w} \,|\, \mathbf{X}, \mathbf{y}\right] d\mathbf{w}$$

Predictive Posterior

$$= \mathcal{N}\left(\langle \boldsymbol{\nu}, \mathbf{x}\rangle, \sigma^2 + \mathbf{x}^\top\Lambda\mathbf{x}\right)$$

Extra Information

A few Thoughts

- Do I have to use these very forms for likelihood and prior?

$$\mathbb{P}\left[y \,|\, \mathbf{x}^i, \boldsymbol{\theta}\right] = \mathcal{D}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

  - No, however some choices make sense and make estimation easier
  - Conjugacy makes life very simple if possible to achieve
  - Look at your application and take these decisions
    - If your know your model is sparse, use a Laplacian prior
    - If you know your noise is not Gaussian, use a GLM
- Can I build probabilistic models for other ML tasks as well?
  - Yes, of course. We will look at classification problems next.
  - We can do probabilistic clustering, dim. redn., ranking
  - There are entire courses devoted to PML techniques: CS772, CS698S

A few Thoughts

- What about non-linear regression?

$$\mathbb{P}\left[y \,|\, \mathbf{x}^i, \boldsymbol{\theta}\right] = \mathcal{N}(f(\mathbf{x}^i, \boldsymbol{\theta}), \sigma^2)$$

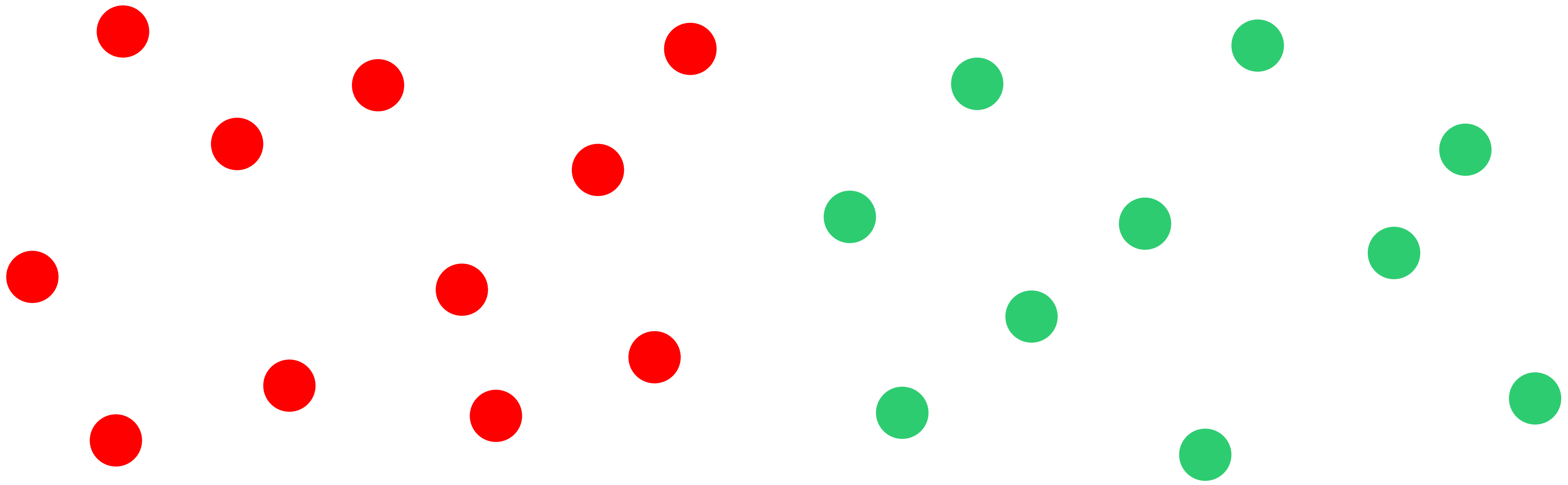
  - MLE, MAP estimation more challenging - kernel methods, deep learning
  - Gaussian processes more suited for non-linear PML – beyond the scope!
- Are Bayesian models better than non-Bayesian ones?
  - Bayesian models more informative $\mathcal{N}\left(\langle \boldsymbol{\nu}, \mathbf{x} \rangle, \sigma^2 + \mathbf{x}^\top \Lambda \mathbf{x}\right)$
  - Useful in settings like active learning, anomaly detection
  - Can be more expensive at training and prediction time
  - Ask your doctor if your application needs Bayesian reasoning or not!
- Bayesian ≠ Generative

A few Thoughts

- How do I solve all these optimization problems that MLE, MAP procedures keep throwing up?
  - We are in safe hands – optimization theory is very evolved
  - Wait for the next set of lectures on function approximation methods
- Are generatively trained models better or discriminative ones?
  - Discriminative models reason about $\mathbb{P}[y|\boldsymbol{x}]$
  - Generative models reason about $\mathbb{P}[y, \boldsymbol{x}]$
  - For prediction, all we need is $\mathbb{P}[y|\boldsymbol{x}]$
  - Generative models handle missing, erroneous data easily but bulky
  - Discriminative models are much lighter, easier to train
  - Models such GMMs, deep nets, can be trained in gen. and disc. manner

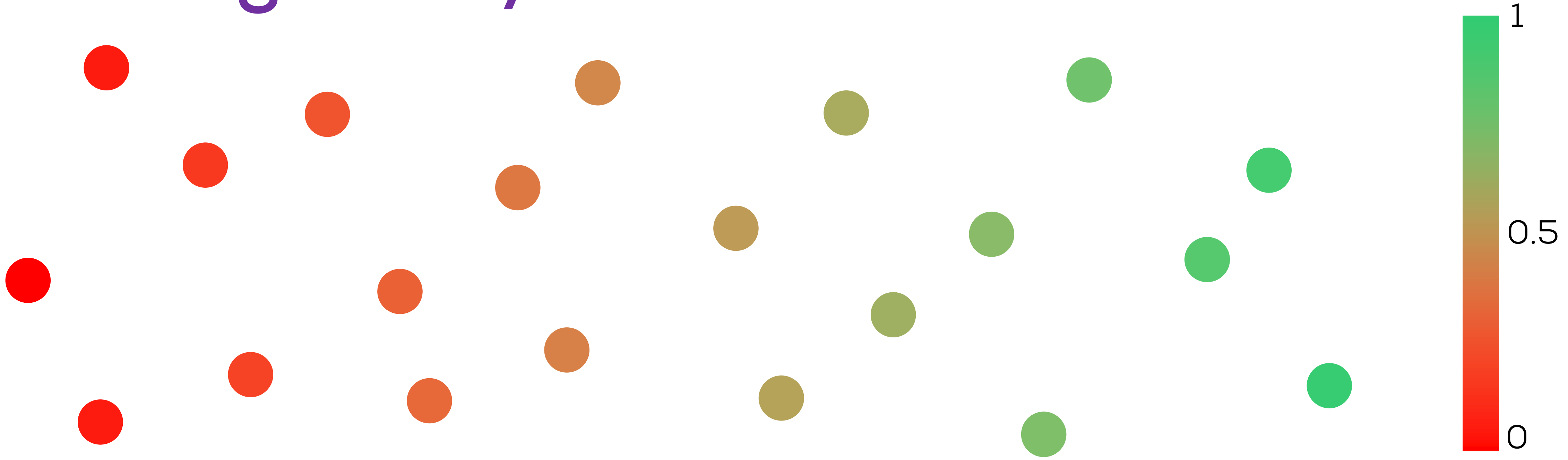
Classification using PML

Revisiting Binary Classification



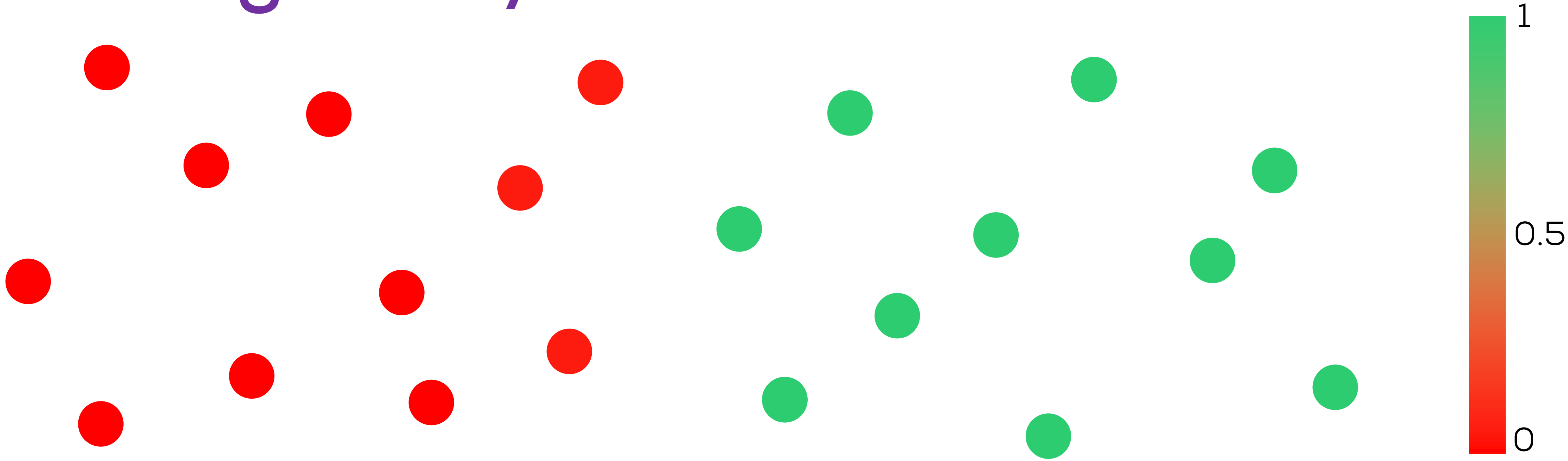
subscene.com, firstpost.com, wikipedia.com, wikia.com

Revisiting Binary Classification



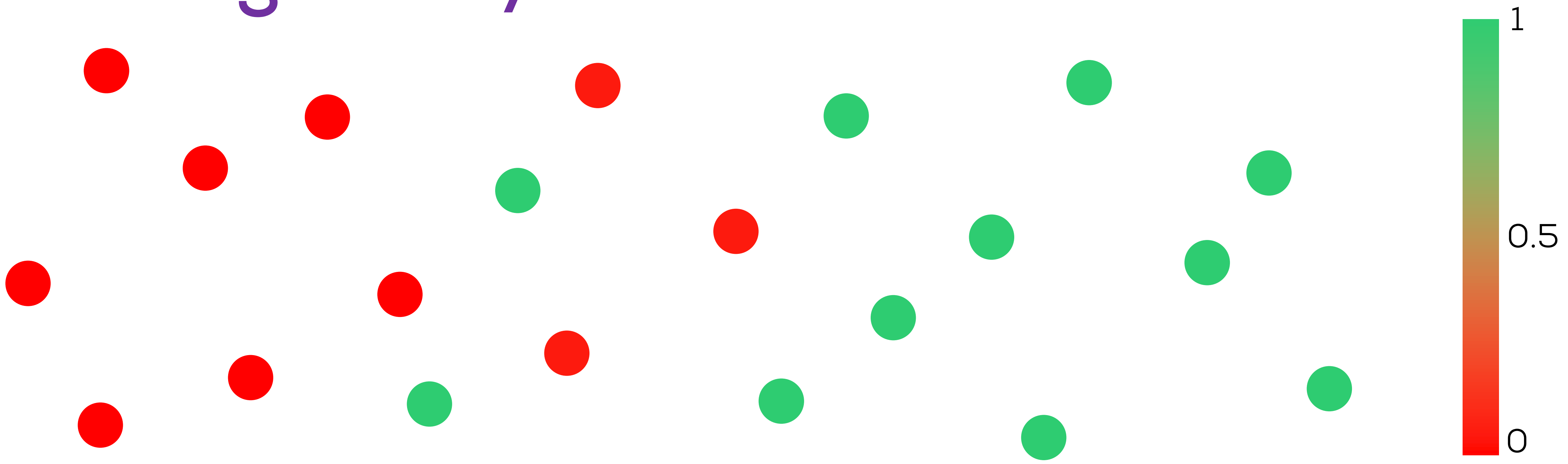
$$\eta(\mathbf{x}) = \mathbb{P}\left[y = 1 \,|\, \mathbf{x}\right]$$

Revisiting Binary Classification



$$\eta(\mathbf{x}) = \mathbb{P}\left[y = 1 \mid \mathbf{x}\right]$$

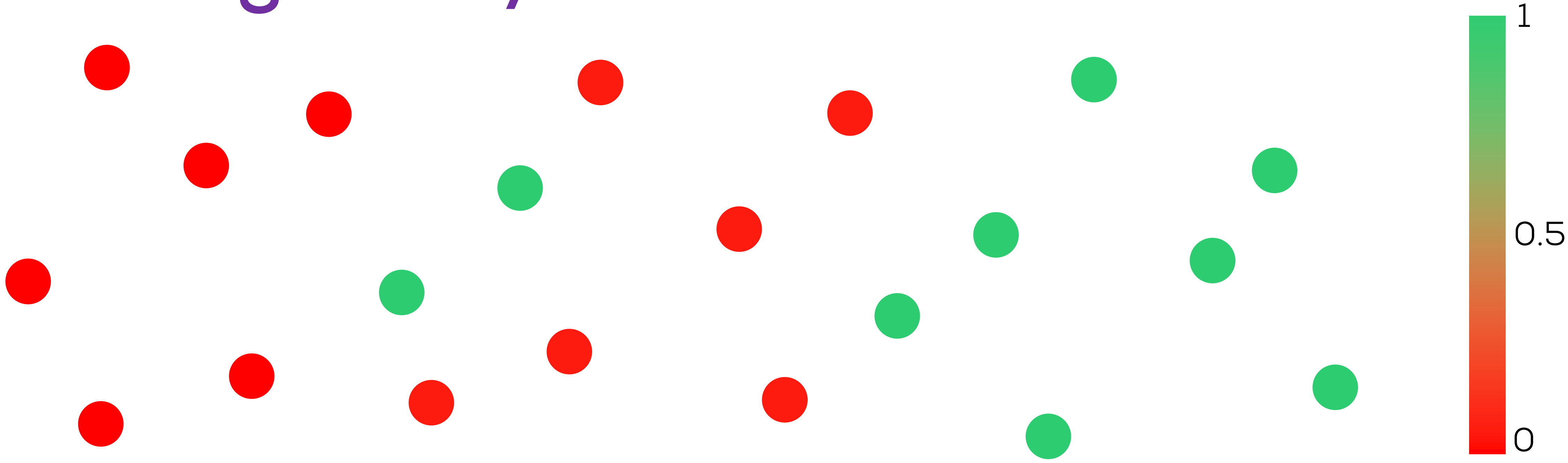
Revisiting Binary Classification



$$\eta(\mathbf{x}) = \mathbb{P}\left[y = 1 \,|\, \mathbf{x}\right]$$

subscene.com, firstpost.com, wikipedia.com, wikia.com

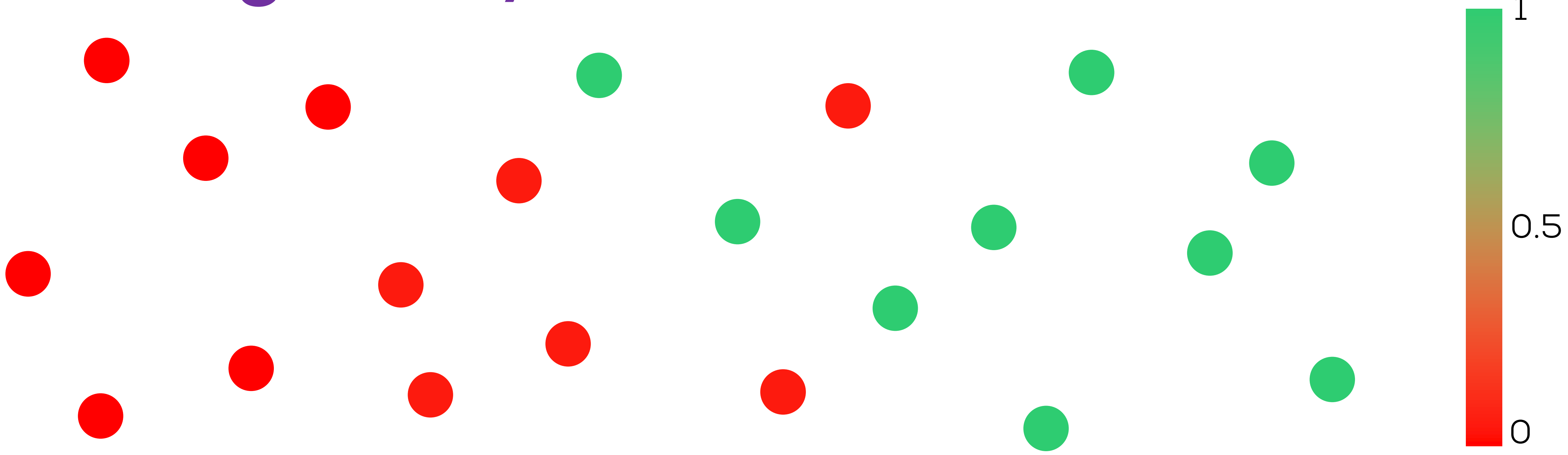
Revisiting Binary Classification



$$\eta(\mathbf{x}) = \mathbb{P}\left[y = 1 \mid \mathbf{x}\right]$$

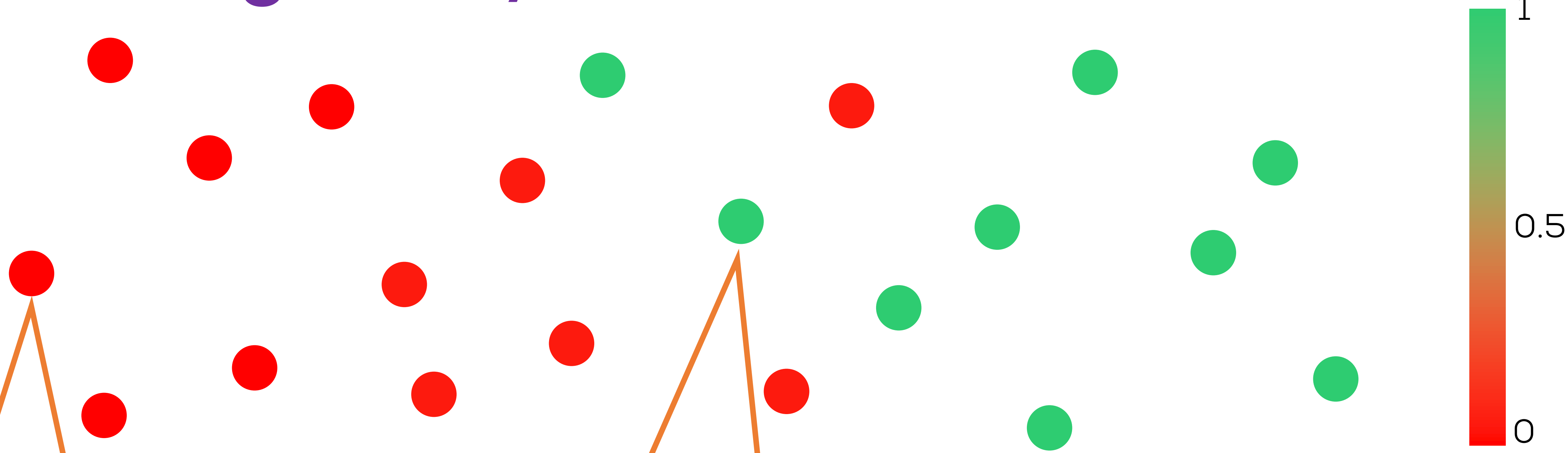
subscene.com, firstpost.com, wikipedia.com, wikia.com

Revisiting Binary Classification



$$\eta(\mathbf{x}) = \mathbb{P}\left[y = 1 \,|\, \mathbf{x}\right]$$

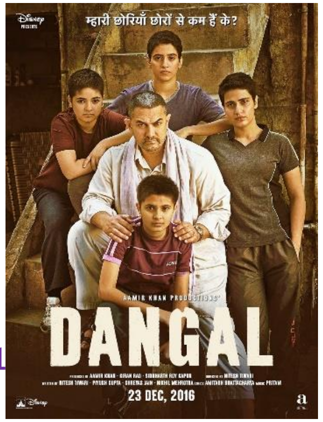
Revisiting Binary Classification



$$\eta(\mathbf{x}) = \mathbb{P}\left[y = 1 \mid \mathbf{x}\right]$$

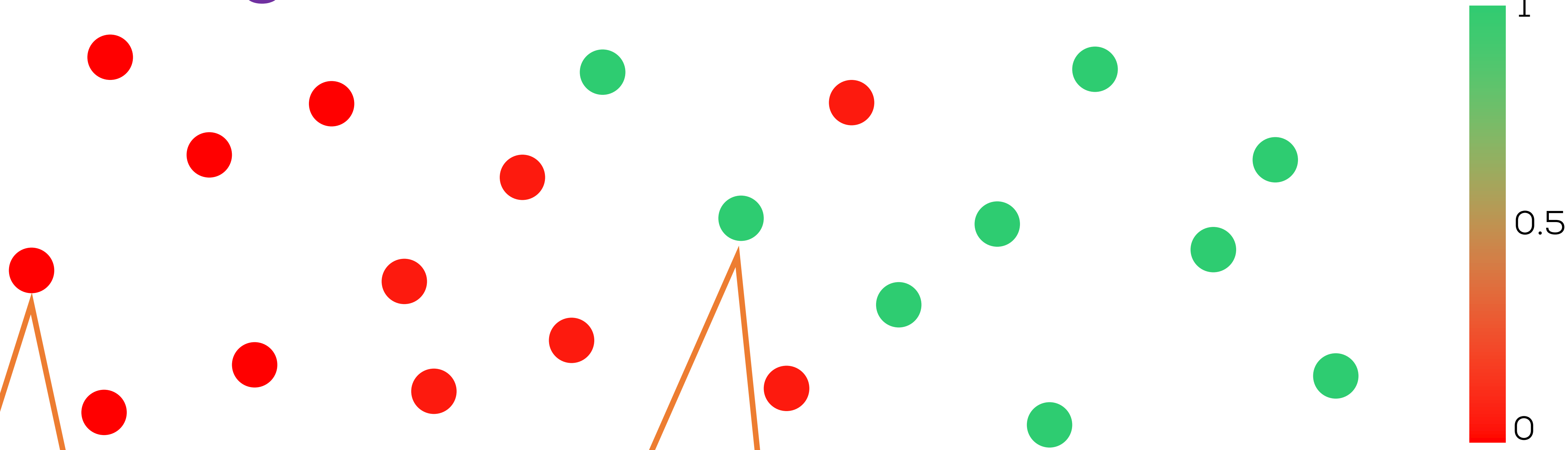
Others are not

Some decisions are hard



subscene.com, firstpost.com, wikipedia.com, wikia.com

Revisiting Binary Classification



Others are not

Some decisions are hard

Gives us a prediction rule!

$$\eta(\mathbf{x}) = \mathbb{P}\left[y = 1 \,|\, \mathbf{x}\right]$$

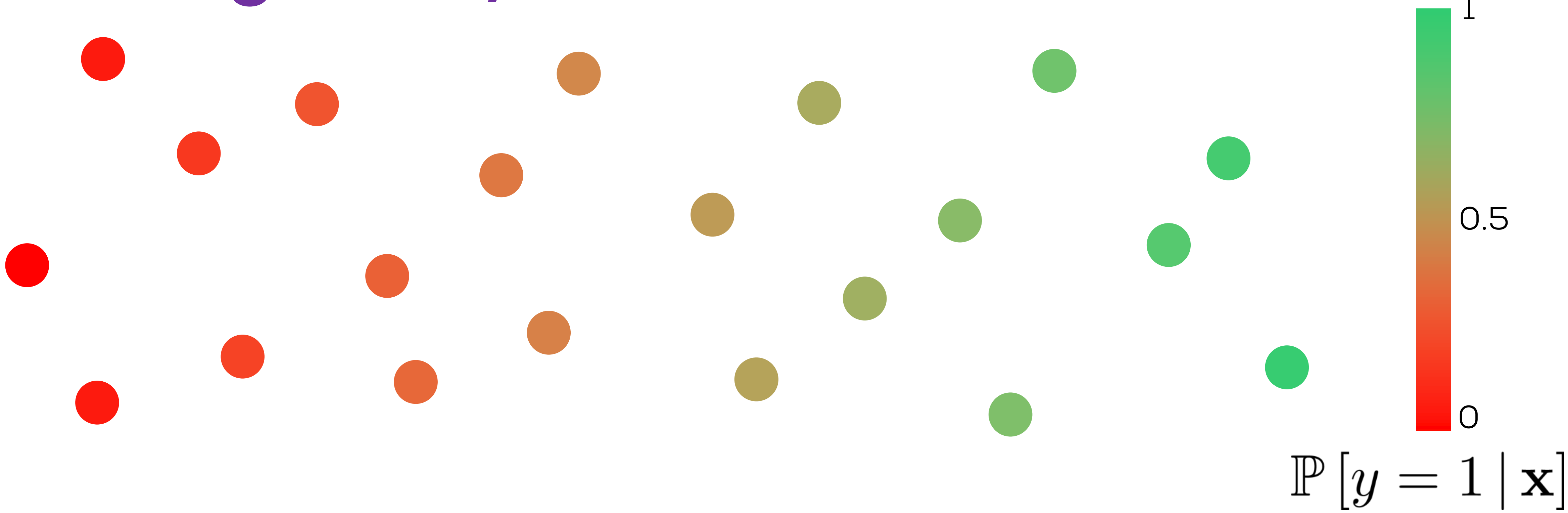
$$\mathbb{P}\left[y = 1 \,|\, \mathbf{x}\right] \geq 0.5 \quad \bullet$$

$$\mathbb{P}\left[\bullet \,|\, \mathbf{x}\right] > \mathbb{P}\left[\bullet \,|\, \mathbf{x}\right]$$

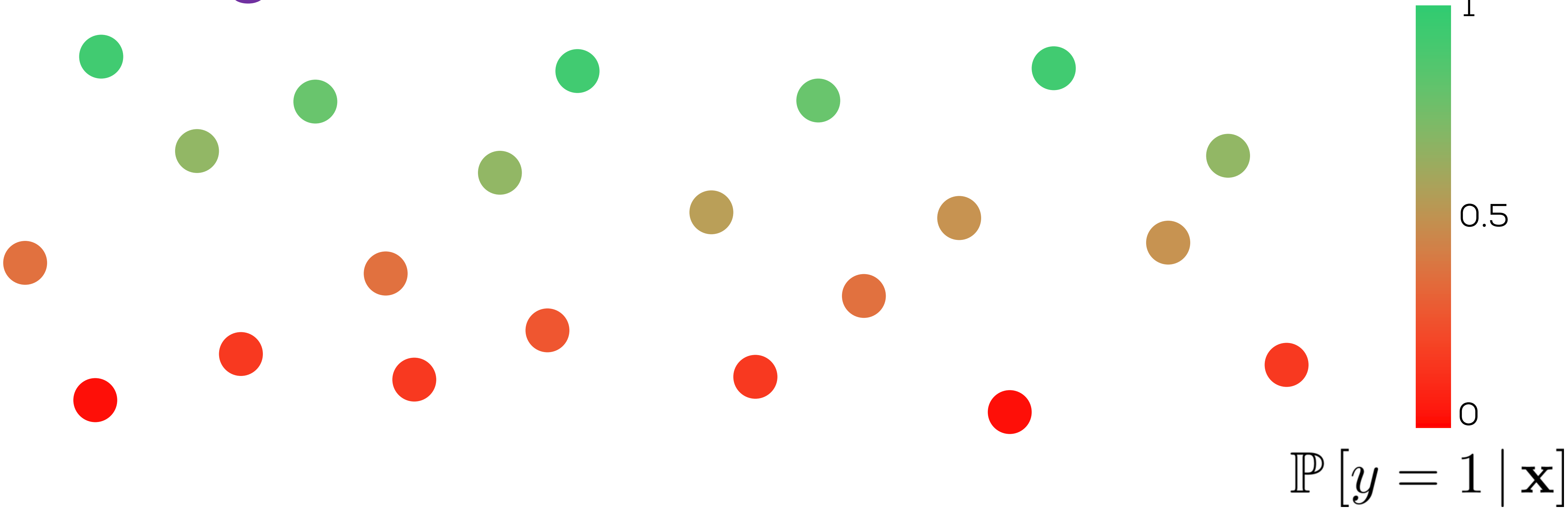
$$\mathbb{P}\left[y = -1 \,|\, \mathbf{x}\right] > 0.5 \quad \bullet$$

ost.com, wikipedia.com, wikia.com

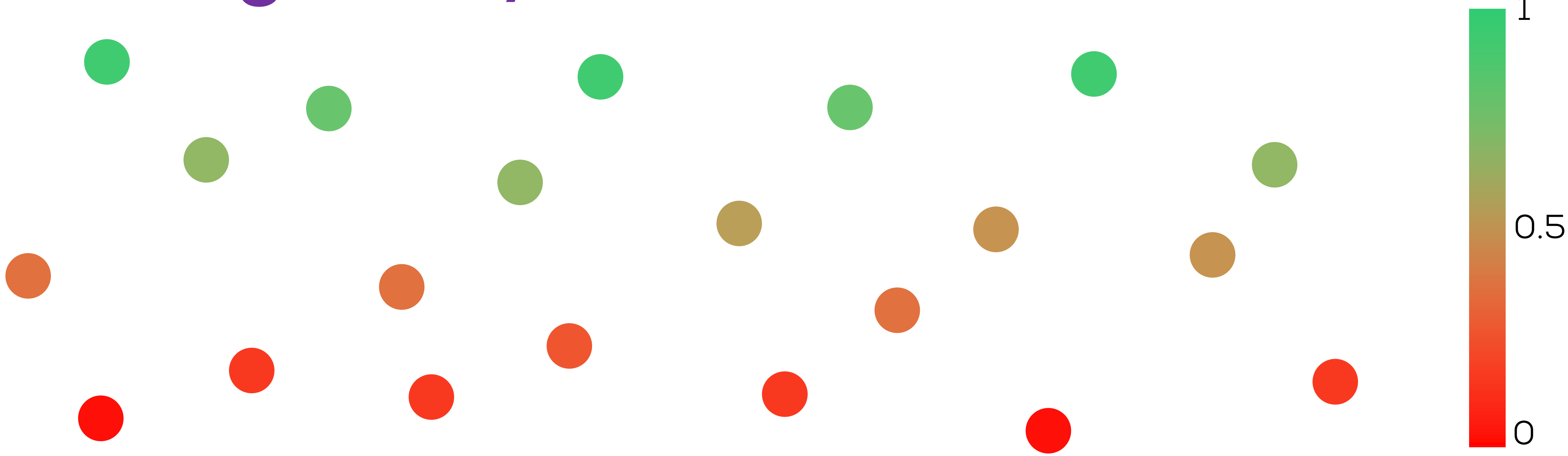
Revisiting Binary Classification



Revisiting Binary Classification



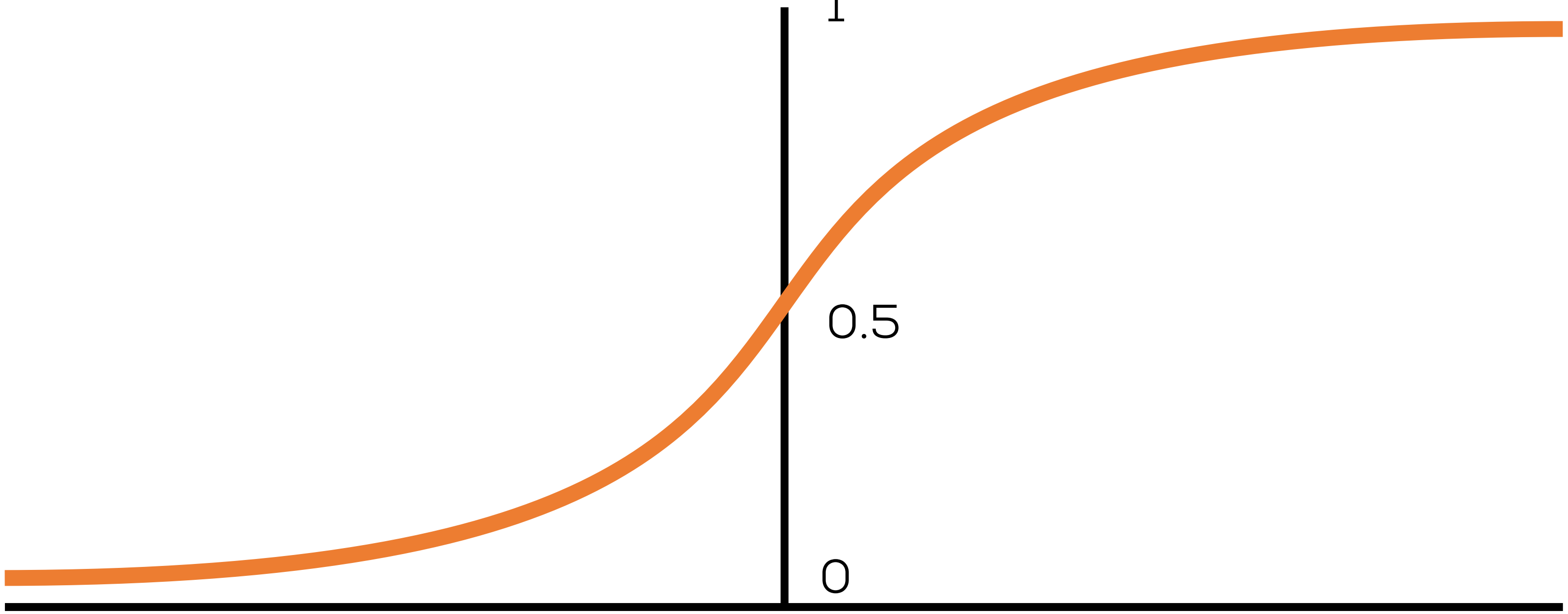
Revisiting Binary Classification



$\mathbb{P}\left[y = 1 \,|\, \mathbf{x}\right]$

Features Matter!

The Sigmoid Function



$$\sigma(t) = \frac{1}{1 + \exp(-t)} = \frac{\exp(t)}{\exp(t) + 1}$$

Classification using MLE

Shorthand
$\mathbb{P}[y^i = y] =: \mathbb{P}[y]$

$$\mathbb{P}\left[y \,|\, \mathbf{x}^i, \mathbf{w}\right] = \sigma\left(y\left\langle \mathbf{w}, \mathbf{x}^i \right\rangle\right) = \frac{1}{1 + \exp(-y\left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)}$$

Likelihood

$$\mathbb{P}\left[y^i = 1 \,|\, \mathbf{x}^i, \mathbf{w}\right] + \mathbb{P}\left[y^i = -1 \,|\, \mathbf{x}^i, \mathbf{w}\right] = 1$$

$$\log \frac{\mathbb{P}\left[y^i = 1 \,|\, \mathbf{x}^i, \mathbf{w}\right]}{\mathbb{P}\left[y^i = -1 \,|\, \mathbf{x}^i, \mathbf{w}\right]} = \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle$$

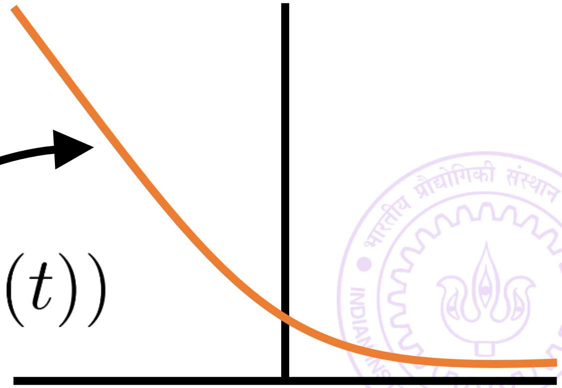
Log-odds

Log-likelihood

$$\log \mathbb{P}\left[\mathbf{y} \,|\, \mathbf{X}, \mathbf{w}\right] = \sum_{i=1}^{n} \log\left(\frac{1}{1 + \exp(-y^i\left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)}\right)$$

$$\hat{\mathbf{w}}_{\mathrm{MLE}} = ?$$

Wait for learning with FA



$$f(t) = -\log(\sigma(t))$$

Another possibility

- What happens if $y_i \in \{0,1\}$ instead of $y_i \in \{-1,1\}$?
- Binomial instead of Rademacher
- How to redefine likelihood?
- Need to ensure $\mathbb{P}\left[y^i = 1 \,|\, \mathbf{x}^i, \mathbf{w}\right] + \mathbb{P}\left[y^i = 0 \,|\, \mathbf{x}^i, \mathbf{w}\right] = 1$
- One way is the following

$$\hat{\eta}_i = \sigma(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

$$\mathbb{P}\left[y^i \,|\, \mathbf{x}^i, \mathbf{w}\right] = (\hat{\eta}_i)^{y^i} \cdot (1 - \hat{\eta}_i)^{1-y^i}$$

- Do you get the same MLE problem?
- What about MAP estimates? $\mathbb{P}\left[\mathbf{w}\right] = \mathcal{N}(\mathbf{0}, \rho^2 \cdot I_d)$

Exercise

Bayesian Logistic Regression?

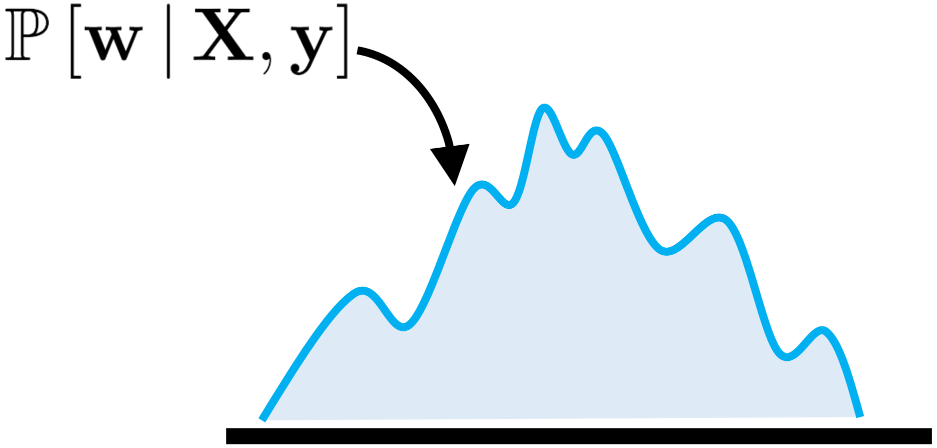
$$\mathbb{P}\left[y \,|\, \mathbf{x}, \mathbf{w}\right] = \sigma\left(y \left\langle \mathbf{w}, \mathbf{x} \right\rangle\right)$$

$$\mathbb{P}\left[\mathbf{w}\right] = \mathcal{N}(\mathbf{0}, \rho^2 \cdot I_d)$$

$$\mathbb{P}\left[\mathbf{w} \,|\, \mathbf{X}, \mathbf{y}\right] = \frac{\mathbb{P}\left[\mathbf{y} \,|\, \mathbf{X}, \mathbf{w}\right] \mathbb{P}\left[\mathbf{w}\right]}{\int_{\mathbf{w}'} \mathbb{P}\left[\mathbf{y} \,|\, \mathbf{X}, \mathbf{w}'\right] \mathbb{P}\left[\mathbf{w}'\right] d\mathbf{w}'} = \frac{\prod_{i=1}^{n} \sigma(y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)}{\int_{\mathbf{w}'} \prod_{i=1}^{n} \sigma(y^i \left\langle \mathbf{w}', \mathbf{x}^i \right\rangle) \exp\left(\frac{-\|\mathbf{w}'\|_2^2}{2\sigma^2}\right) d\mathbf{w}'}$$

Posterior Approximation

MCMC, Variational Inference,
Laplace approx



Bayesian Logistic Regression?

$$\mathbb{P}\left[y \,|\, \mathbf{x}, \mathbf{w}\right] = \sigma\left(y \left\langle \mathbf{w}, \mathbf{x} \right\rangle\right)$$

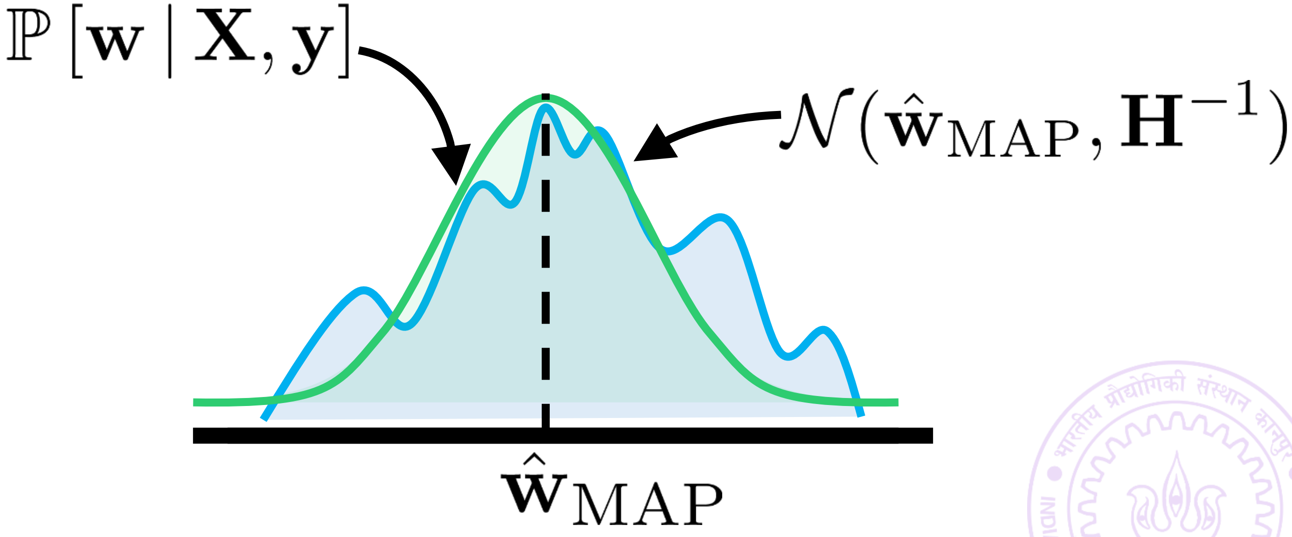
$$\mathbb{P}\left[\mathbf{w}\right] = \mathcal{N}(\mathbf{0}, \rho^2 \cdot I_d)$$

$$\mathbb{P}\left[\mathbf{w} \,|\, \mathbf{X}, \mathbf{y}\right] = \frac{\mathbb{P}\left[\mathbf{y} \,|\, \mathbf{X}, \mathbf{w}\right] \mathbb{P}\left[\mathbf{w}\right]}{\int_{\mathbf{w}'} \mathbb{P}\left[\mathbf{y} \,|\, \mathbf{X}, \mathbf{w}'\right] \mathbb{P}\left[\mathbf{w}'\right] d\mathbf{w}'} = \frac{\prod_{i=1}^{n} \sigma(y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)}{\int_{\mathbf{w}'} \prod_{i=1}^{n} \sigma(y^i \left\langle \mathbf{w}', \mathbf{x}^i \right\rangle) \exp\left(\frac{-\|\mathbf{w}'\|_2^2}{2\sigma^2}\right) d\mathbf{w}'}$$

Posterior Approximation

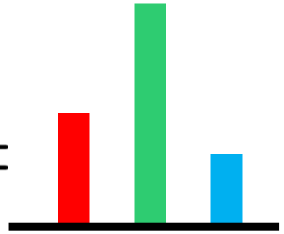
MCMC, Variational Inference, Laplace approx

However, even the Laplace predictive posterior is intractable ☹



Multi-class and Multi-label Classification using PML

Multi-classification using MLE

- $K > 2$ classes – need more detailed parameters
- For each point, its label profile is a vector $\boldsymbol{\eta}(\mathbf{x}) =$ 

$$\mathbb{P}\left[y^i = k \,|\, \mathbf{x}^i, \{\mathbf{w}^l\}_{1,\ldots,K}\right] \propto \exp(\langle \mathbf{w}^k, \mathbf{x}^i \rangle)$$

$$\mathbb{P}\left[y^i = k \,|\, \mathbf{x}^i, \{\mathbf{w}^l\}_{1,\ldots,K}\right] = \frac{\exp(\langle \mathbf{w}^k, \mathbf{x}^i \rangle)}{\sum_{l=1}^{K} \exp(\langle \mathbf{w}^l, \mathbf{x}^i \rangle)}$$

- Likelihood function is multinomial instead of binomial

$$\mathbb{P}\left[\mathbf{y} \,|\, \mathbf{X}, \mathbf{w}\right] = \prod_{i=1}^{n} \hat{\boldsymbol{\eta}}^i_{y^i}(\mathbf{x}) \qquad \hat{\boldsymbol{\eta}}^i_k(\mathbf{x}) = \frac{\exp(\langle \mathbf{w}^k, \mathbf{x}^i \rangle)}{\sum_{l=1}^{K} \exp(\langle \mathbf{w}^l, \mathbf{x}^i \rangle)}$$

Softmax Regression

Binary vs Multi-Classification using Linear LR

- Binary

$$\underset{y \in \{-1,1\}}{\arg\max}\ \mathbb{P}\left[y \mid \mathbf{x}\right]$$

- Multiclass

Binary vs Multi-Classification using Linear LR

- Binary

$$\mathbb{P}\left[y \mid \mathbf{x}\right] \geq 0.5$$

- Multiclass

Binary vs Multi-Classification using Linear LR

- Binary

$$\mathbb{P}\left[y \,|\, \mathbf{x}\right] \geq 0.5$$

$$\mathbb{P}\left[y \,|\, \mathbf{x}, \mathbf{w}\right] = \sigma\left(y \left\langle \mathbf{w}, \mathbf{x} \right\rangle\right)$$

- Multiclass

Binary vs Multi-Classification using Linear LR

- Binary

$$\sigma(y \langle \mathbf{w}, \mathbf{x}^i \rangle) \geq 0.5$$

- Multiclass

Binary vs Multi-Classification using Linear LR

- Binary

$$y \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \geq 0$$

- Multiclass

Binary vs Multi-Classification using Linear LR

- Binary

$$y = \operatorname{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

- Multiclass

Binary vs Multi-Classification using Linear LR

- Binary

$$y = \mathrm{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

- Multiclass

Binary vs Multi-Classification using Linear LR

- Binary

$$y = \mathrm{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

- Multiclass

$$\arg\max_{y \in [K]} \mathbb{P}\left[y \mid \mathbf{x}\right]$$

Binary vs Multi-Classification using Linear LR

- Binary

$$y = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

- Multiclass

$$\arg\max_{y \in [K]} \mathbb{P}\left[y \,|\, \mathbf{x}\right]$$

$$\mathbb{P}\left[y = k \,|\, \mathbf{x}, \mathbf{w}\right] = \sigma\left(\langle \mathbf{w}^k, \mathbf{x} \rangle\right)$$

Binary vs Multi-Classification using Linear LR

- Binary

$$y = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

- Multiclass

$$\arg\max_{y \in [K]} \sigma\left(\langle \mathbf{w}^k, \mathbf{x} \rangle\right)$$

Binary vs Multi-Classification using Linear LR

- Binary

$$y = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

- Multiclass

$$\arg\max_{y \in [K]} \langle \mathbf{w}^k, \mathbf{x} \rangle$$

Binary vs Multi-Classification using Linear LR

- Binary

$$y = \operatorname{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

- Multiclass

$$\underset{y \in [K]}{\arg\max} \; \langle \mathbf{w}^k, \mathbf{x} \rangle$$

Binary vs Multi-Classification using Linear LR

- Binary

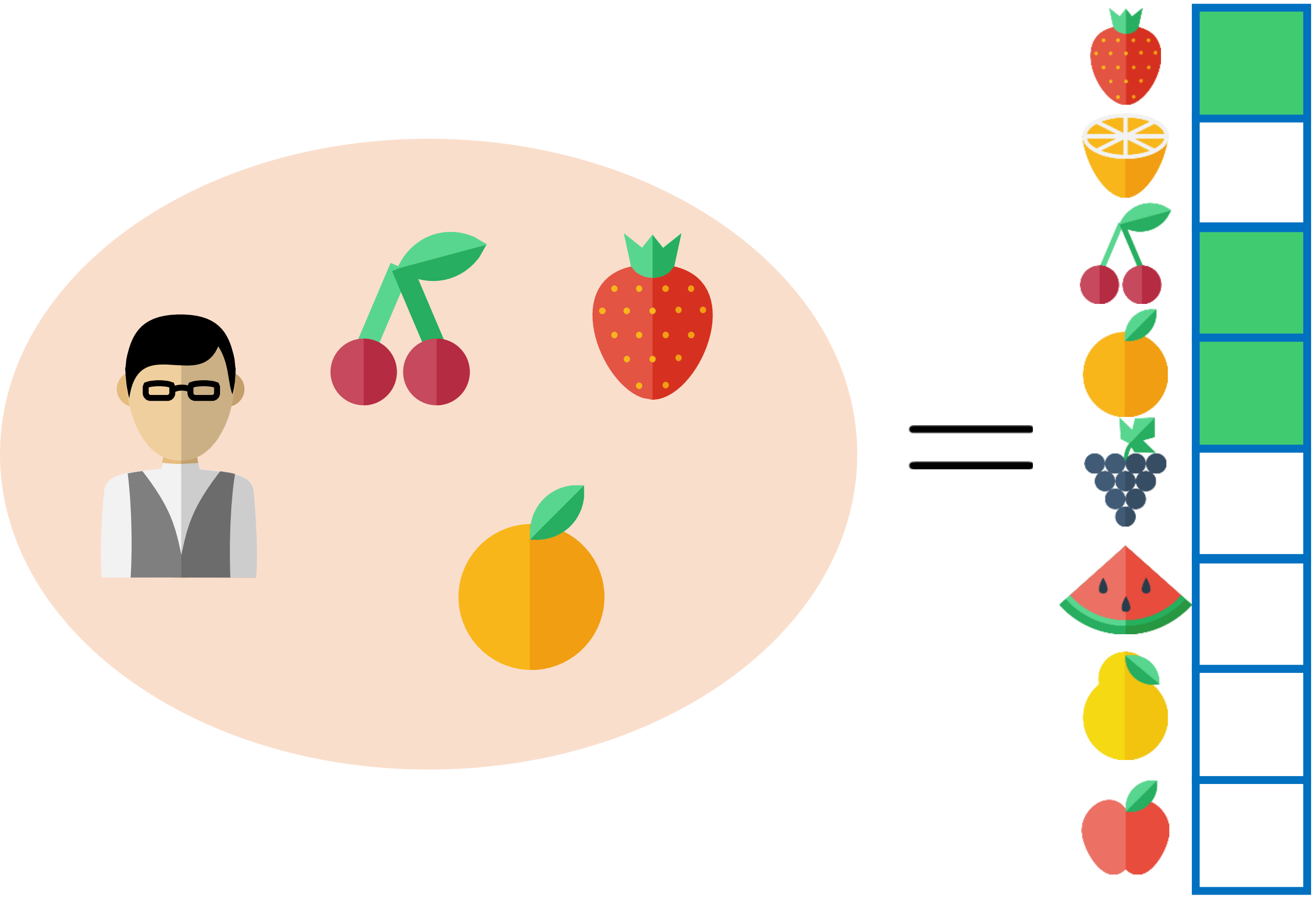
$$y = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

- Multiclass

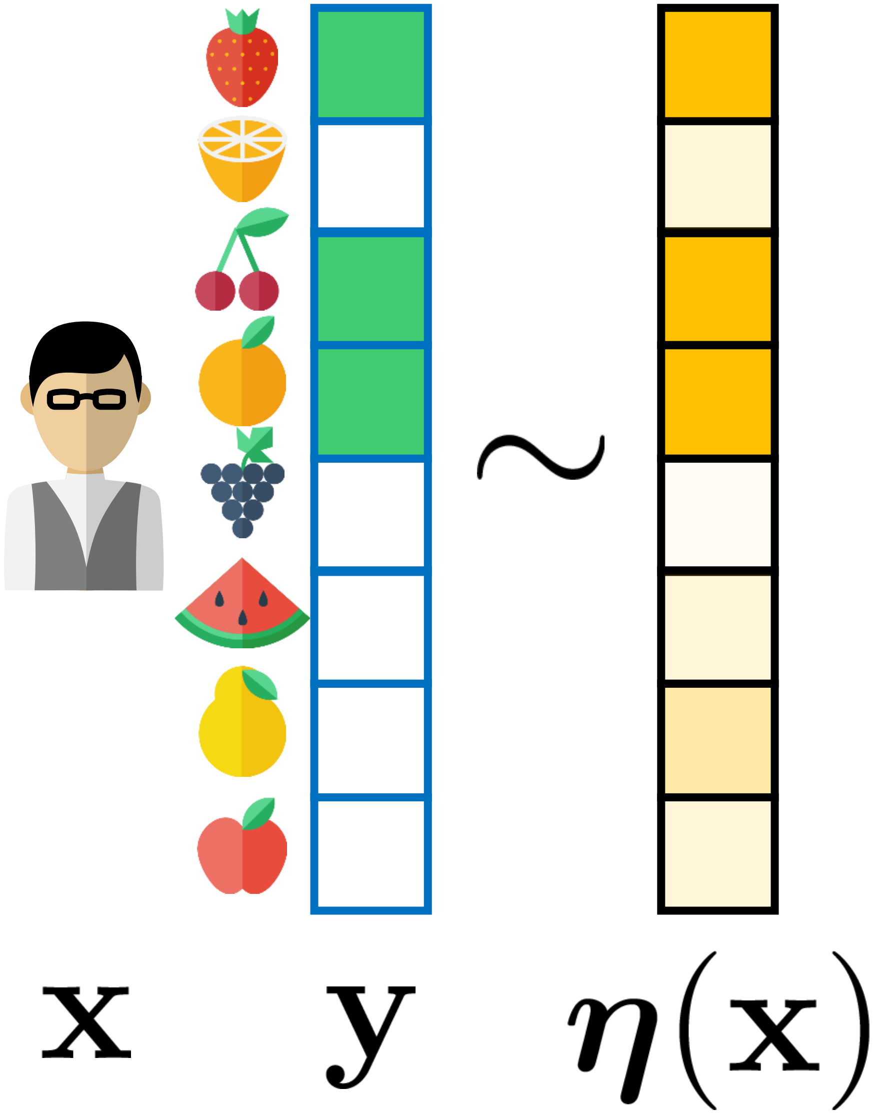
$$\arg\max_{y \in [K]} \langle \mathbf{w}^k, \mathbf{x} \rangle$$

Linear Classifier

Multi-label Learning using PML



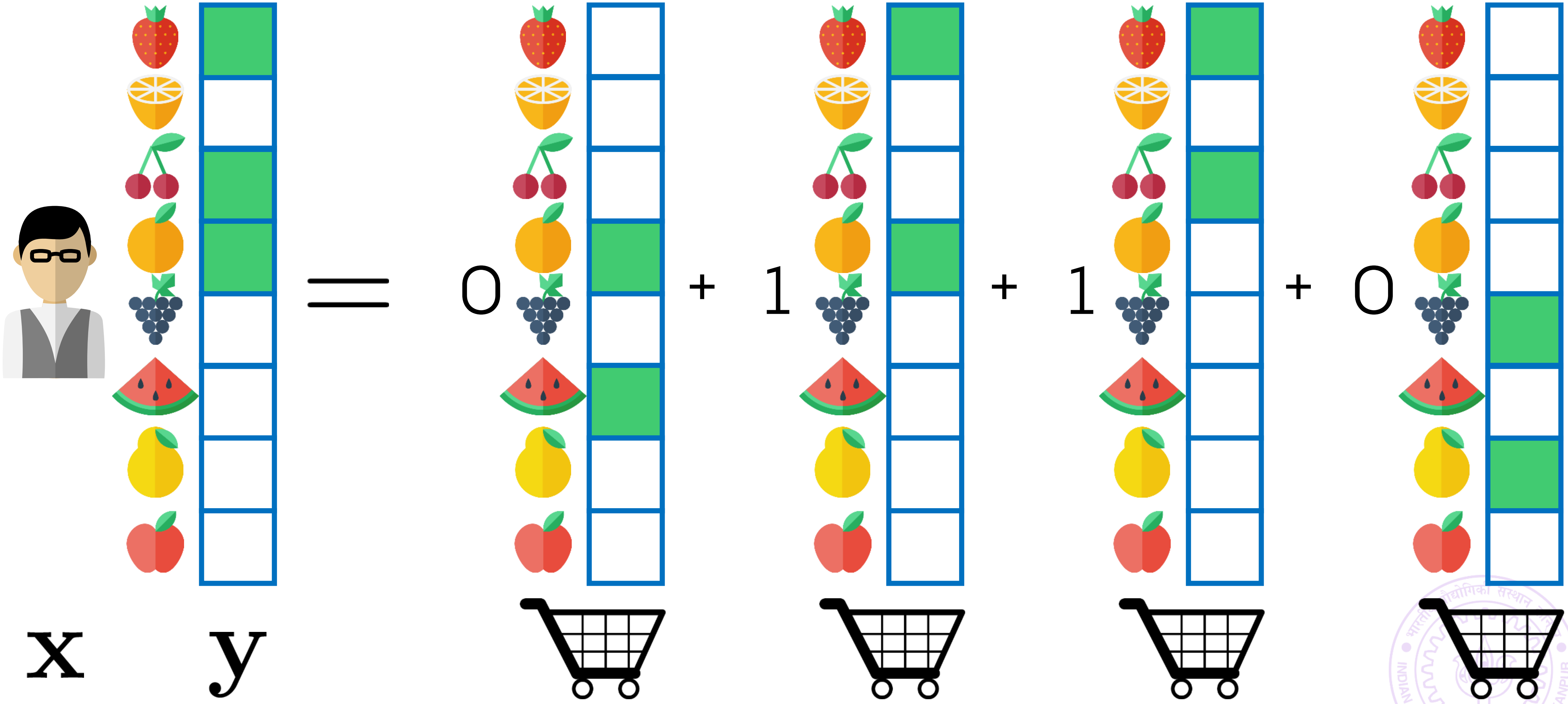
Multi-label Learning using PML



Solve as L independent Binary problems using Logistic regression!

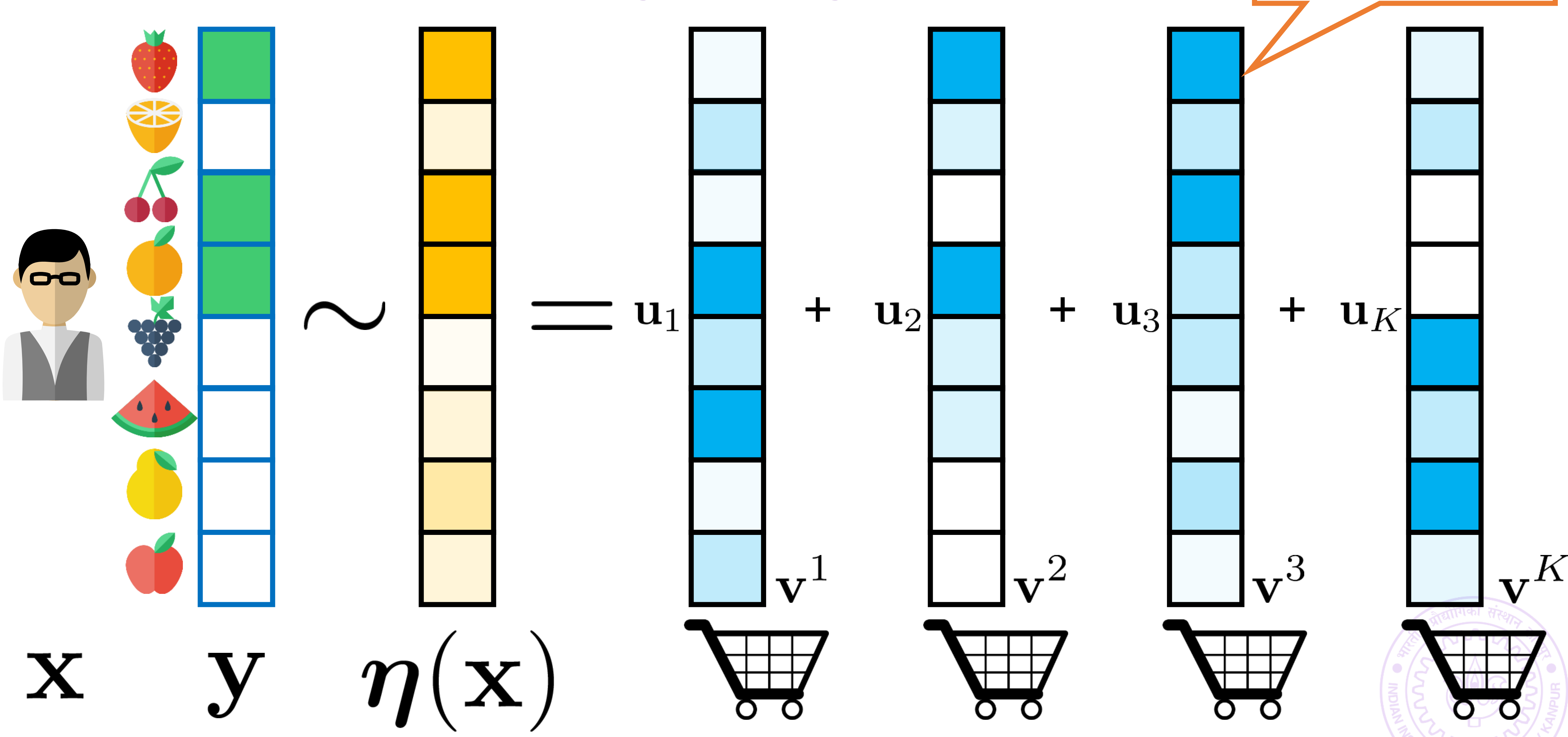
Does not scale!

Multi-label Learning using PML



Rai et al. Large-Scale Bayesian Multi-Label Learning via Topic-Based Label Embeddings, NIPS 2015

Multi-label Learning using PML



Rai et al. Large-Scale Bayesian Multi-Label Learning via Topic-Based Label Embeddings, NIPS 2015

Multi-label Learning using PML

$$\mathbf{y}^i \sim \text{Bernoulli}(\boldsymbol{\eta}^i)$$
$$\boldsymbol{\eta}^i = \mathbf{V}\mathbf{u}^i, \mathbf{V} = [\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^K]$$
$$\mathbf{u}^i_j \sim \mathcal{N}(\langle \mathbf{w}^j, \mathbf{x}^i \rangle, \sigma^2), j = 1, \ldots, K$$

Can fix $\boldsymbol{V}$ or else …

$$\mathbf{v}^j \sim \text{Dirichlet}(\boldsymbol{\alpha})$$
$$\mathbf{w}^j \sim \mathcal{N}(\mathbf{0}, \rho^2 \cdot I), j = 1, \ldots, K$$

Generalization of Beta distribution

$$\mathbb{P}[\mathbf{v}; \boldsymbol{\alpha}] = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{L} \mathbf{v}_i^{\boldsymbol{\alpha}_i - 1}$$
$$\mathbb{P}[p; \alpha, \beta] = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

Only K models, not L

Have fun posterioring!

Please give your Feedback

http://tinyurl.com/ml17-18afb