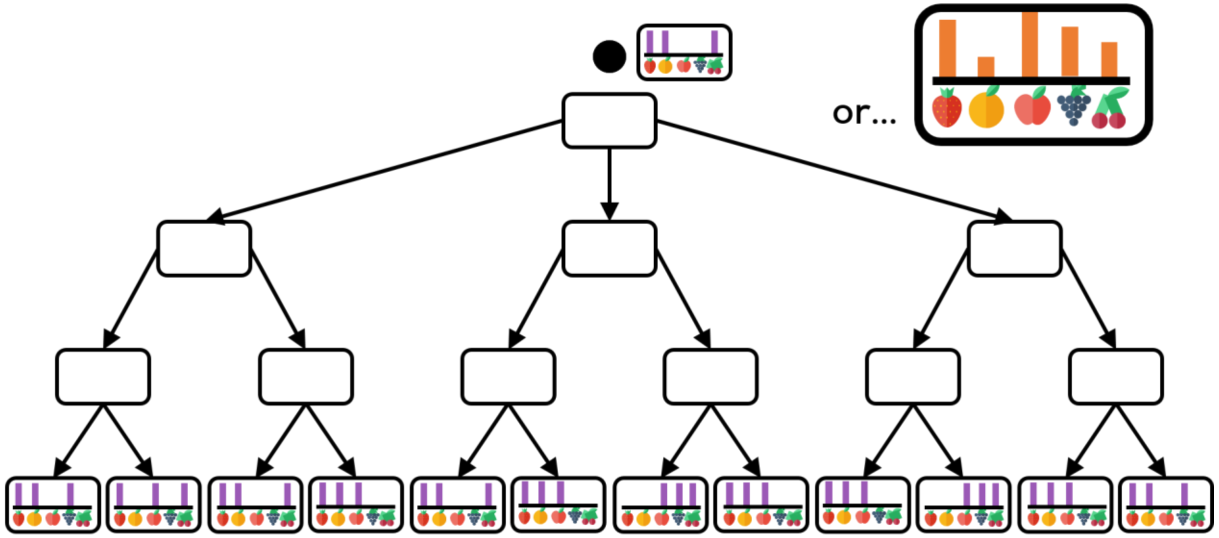
Probabilistic Methods-I

CS771: Introduction to Machine Learning

Purushottam Kar

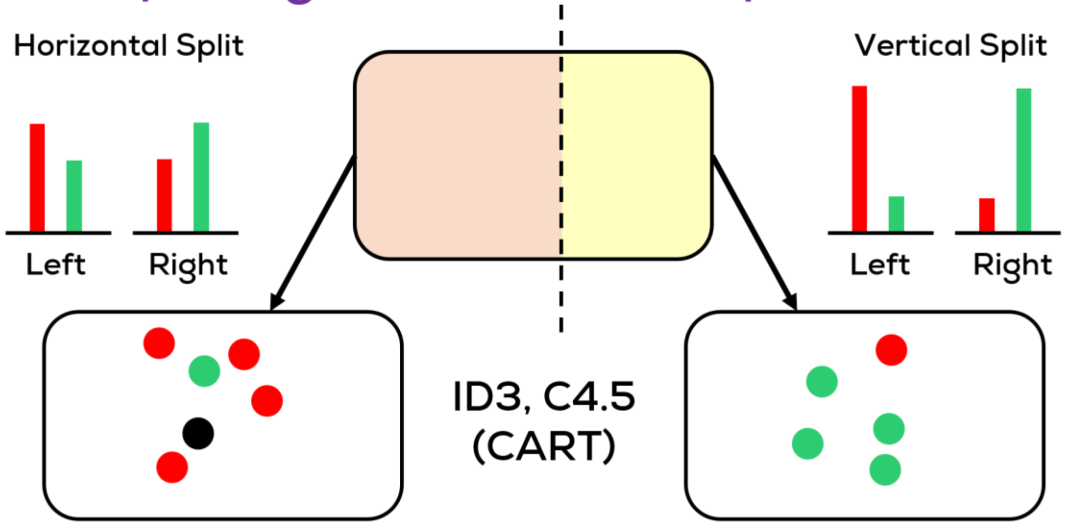
Recap

Multi-label Classification with Decision Trees

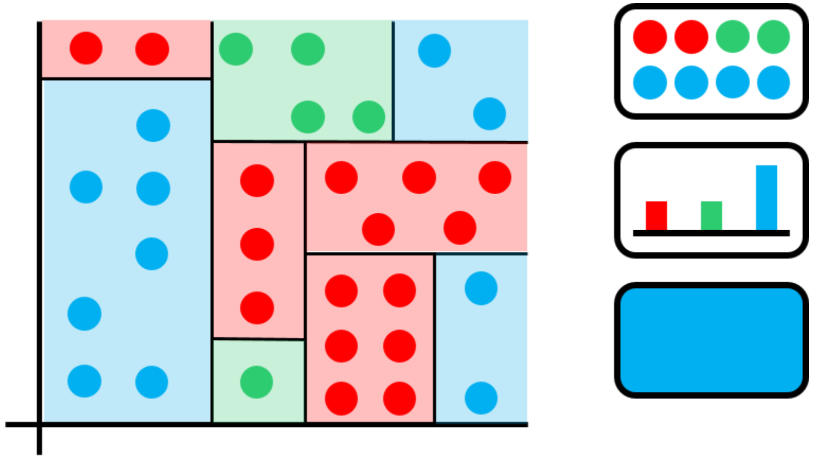


August 11, 2017 13

Node Splitting via feature stumps



August 11, 2017 **Exercise:** Think about how to deal with regression!! 22



Reconciling ID3

Insufficient Features

Guess the Movie!

Zero-shot learning problem

Original Language

DTs are excellent for NN search (kd-tree)

Year of Release
2010,2011,...2017

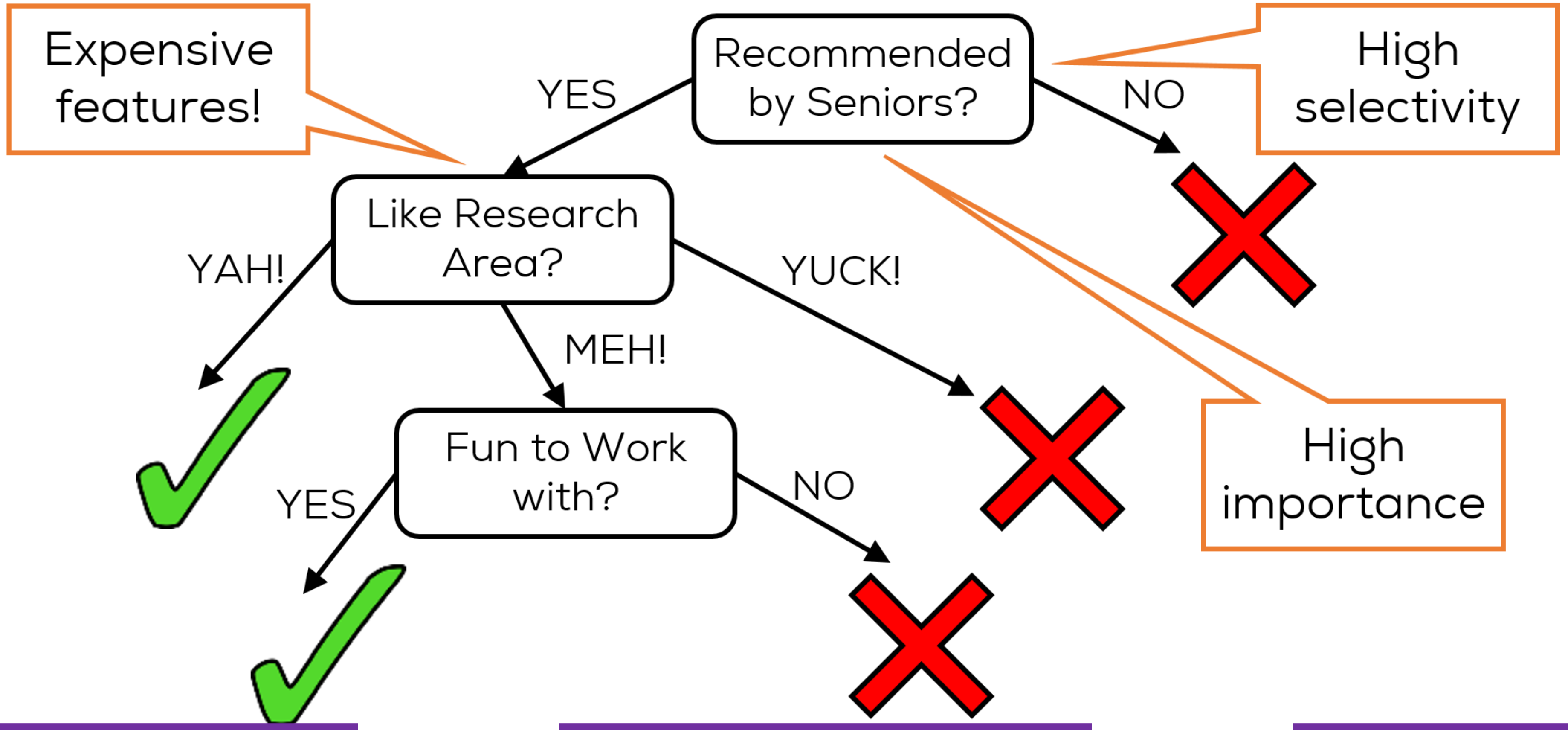
Box-Office Collection
Low (< INR 100 cr)
Medium (INR 100-1000 cr)
High (> 1000 cr)

shutterstock.com, inktalks.com, glamsham.com, youtube.com

Does narrow search space

Reconciling ID3

Choose an Adviser!



Cost-effectiveness key property

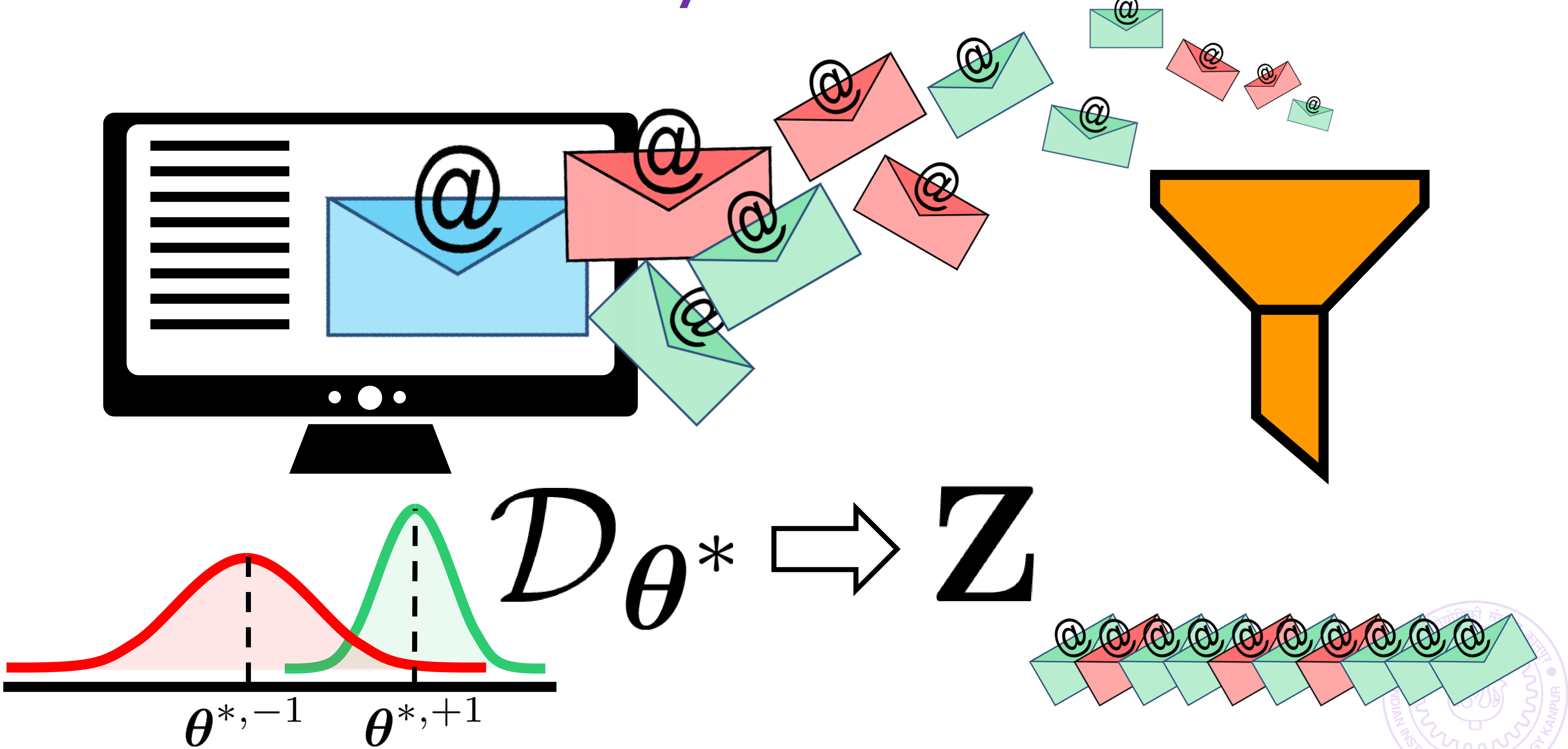
Helps narrow search space

Not a usual ML problem!

The Probabilistic Philosophy

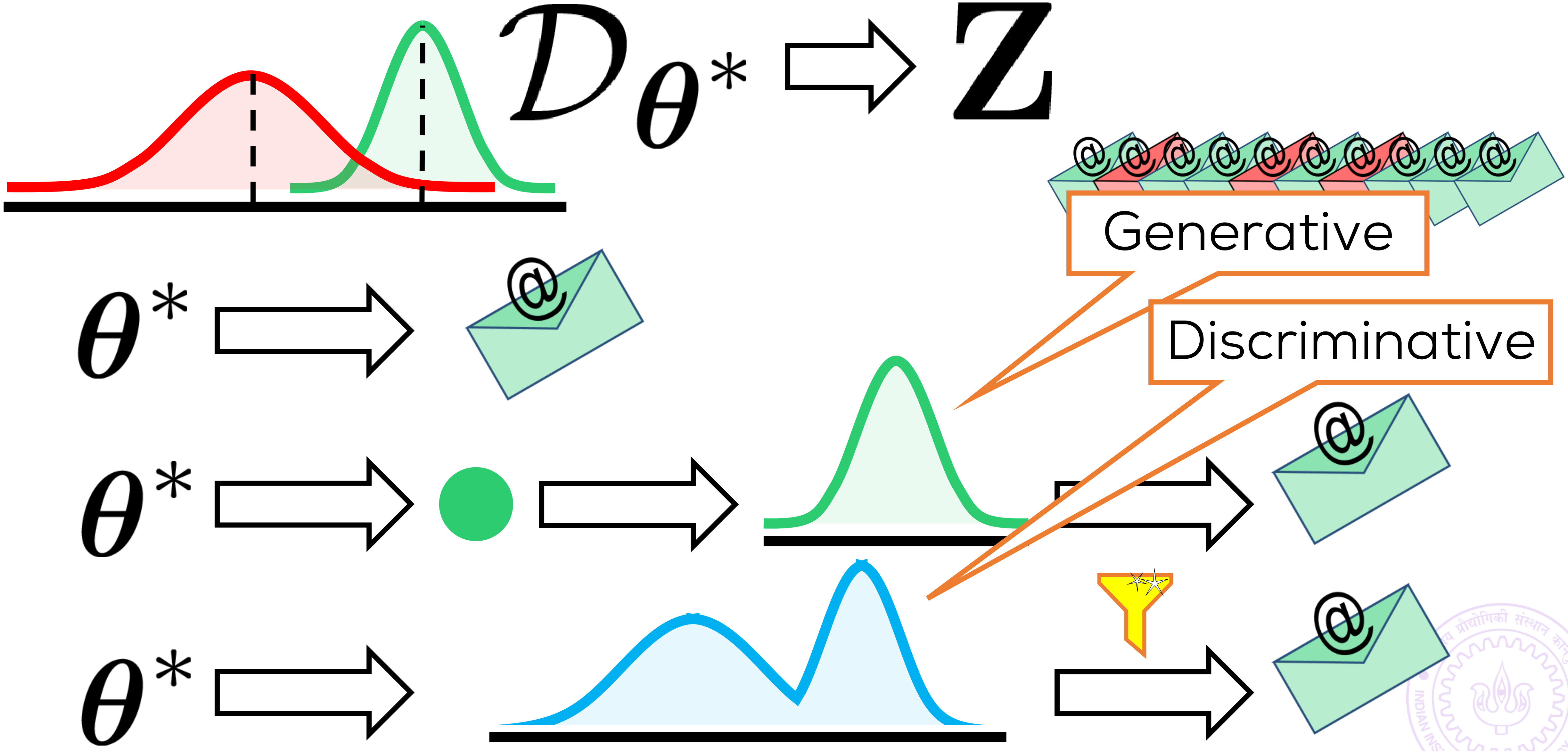
and the generative story …

The Generative Story



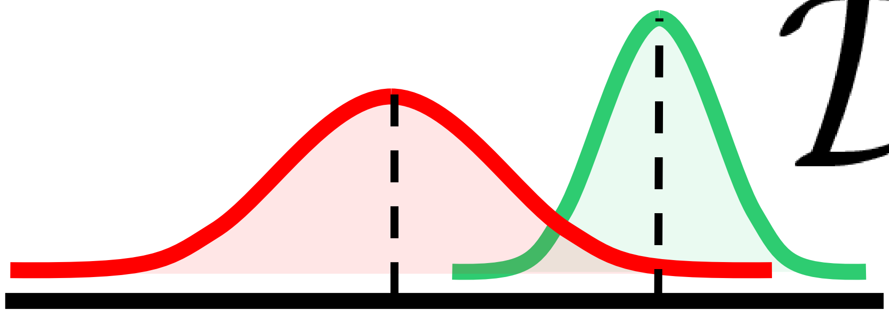
$$\mathcal{D}_{\theta^*} \Rightarrow \mathbf{Z}$$

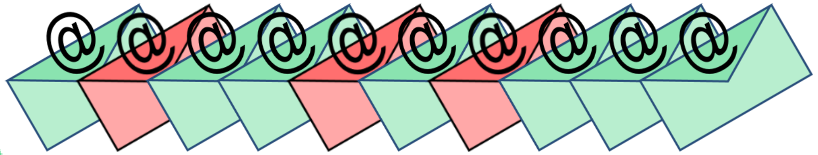
The Generative Story



The Generative Story

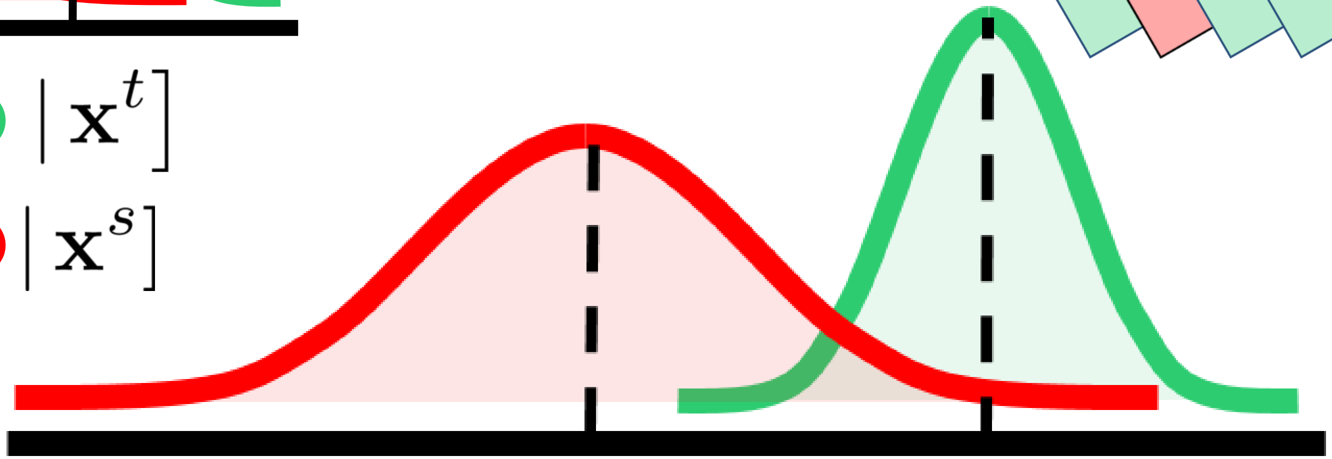
$$\mathcal{D}_{\theta^*} \overset{?}{\Longleftarrow} \mathbf{Z}$$





$$\mathbb{P}\left[\textcolor{red}{\bullet} \mid \mathbf{x}^t\right] > \mathbb{P}\left[\textcolor{green}{\bullet} \mid \mathbf{x}^t\right]$$

$$\mathbb{P}\left[\textcolor{green}{\bullet} \mid \mathbf{x}^s\right] > \mathbb{P}\left[\textcolor{red}{\bullet} \mid \mathbf{x}^s\right]$$



$\mathbf{x}^t$ 

$\mathbf{x}^s$ 

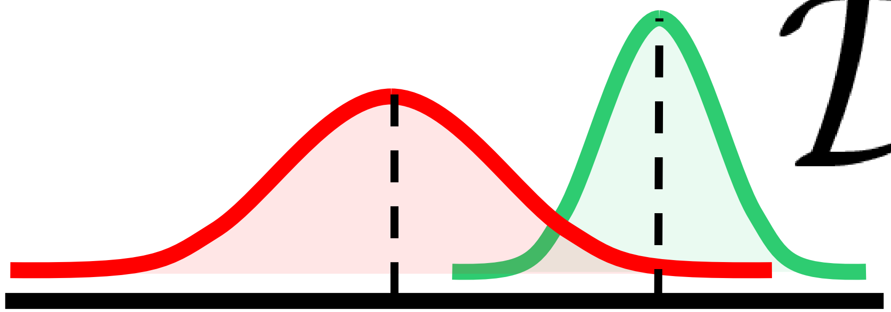
$$\mathbf{Z} = (\mathbf{X}, \mathbf{y})$$

$$\mathbf{X} = \left\{\mathbf{x}^1, \ldots, \mathbf{x}^n\right\}$$

$$\mathbf{y} = \left\{y^1, \ldots, y^n\right\}$$

The Generative Story

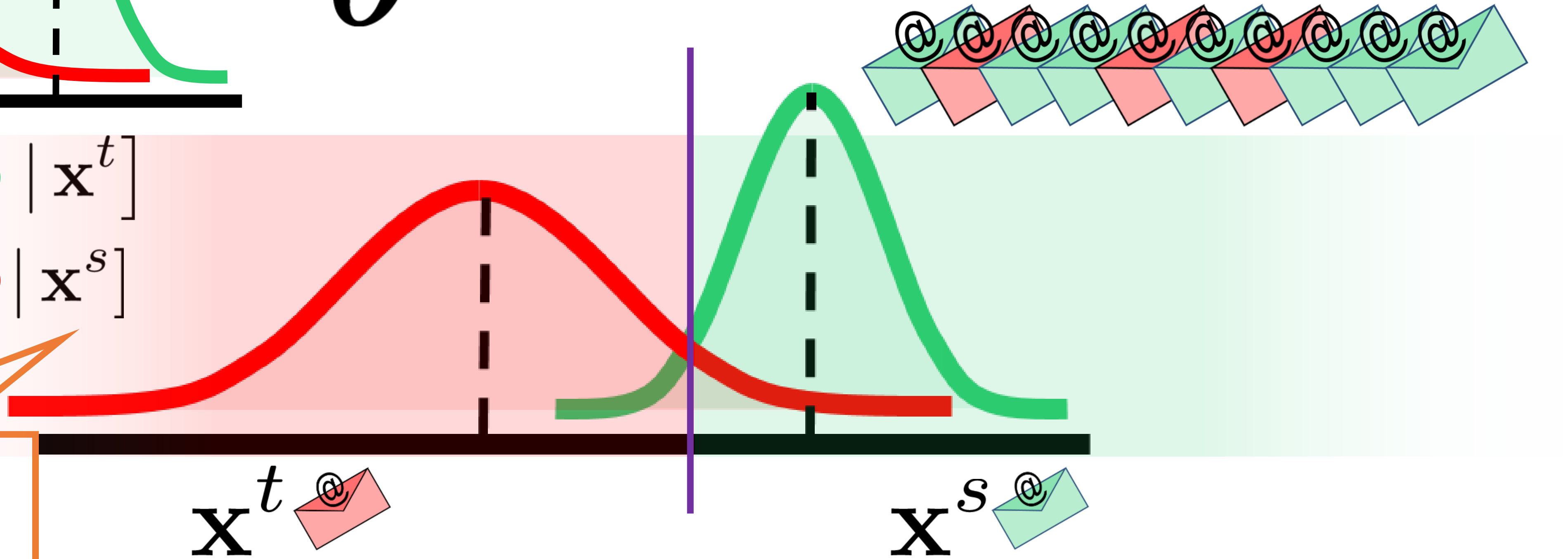
$$\mathcal{D}_{\boldsymbol{\theta}^*} \overset{?}{\Longleftarrow} \mathbf{Z}$$



$$\mathbb{P}\left[{\color{red}\bullet} \mid \mathbf{x}^t\right] > \mathbb{P}\left[{\color{green}\bullet} \mid \mathbf{x}^t\right]$$

$$\mathbb{P}\left[{\color{green}\bullet} \mid \mathbf{x}^s\right] > \mathbb{P}\left[{\color{red}\bullet} \mid \mathbf{x}^s\right]$$

Gives us a prediction rule!

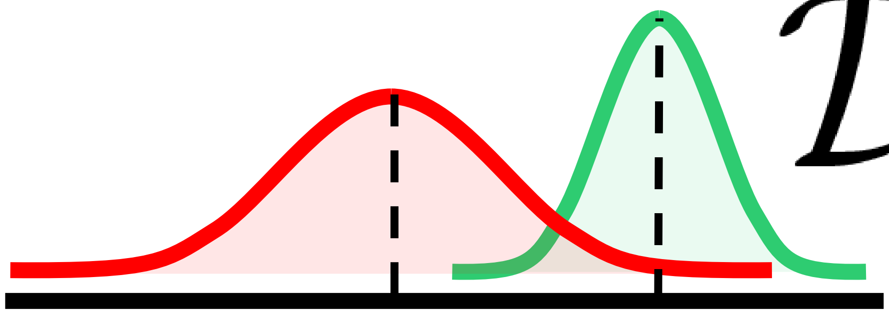


$$\mathbf{Z} = (\mathbf{X}, \mathbf{y})$$

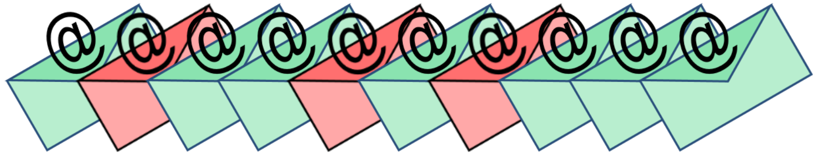
$$\mathbf{X} = \left\{\mathbf{x}^1, \ldots, \mathbf{x}^n\right\}$$

$$\mathbf{y} = \left\{y^1, \ldots, y^n\right\}$$

The Generative Story



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



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

$\mathbb{P}\left[\boldsymbol{\theta}\right]$ Prior

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

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

Learning a Coin

Learning the Bias of a Coin

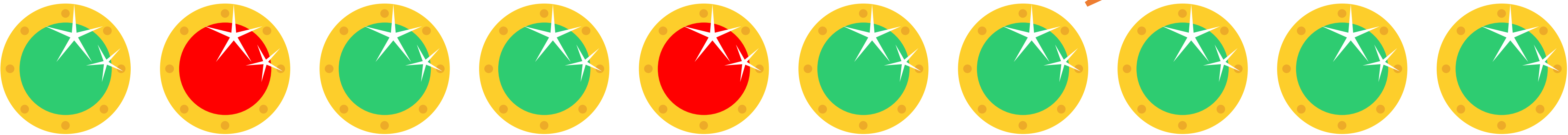
Bias p^*

The bias is the parameter $\boldsymbol{\theta}^*$

Bernoulli distribution

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

Independent tosses



1 0 1 1 0 1 1 1 1 1

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

How to estimate p^* using coin tosses $\boldsymbol{y}$?

Learning the Bias of a Coin

Bias p^*

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

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

Learning the Bias of a Coin

Bias p^*

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

Likelihood

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

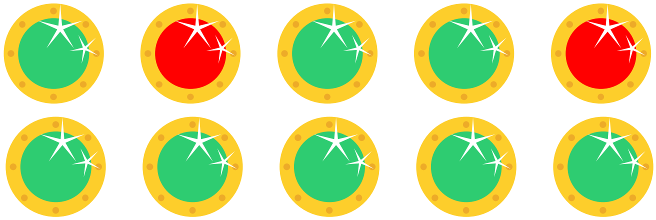
Log-likelihood

$$\log \mathbb{P}\left[\mathbf{y} \mid p\right] = n_H \log p + n_T \log(1-p)$$

Maximum Likelihood Estimate

often affectionately called the MLE

The ML Estimator

Bias p^* $\quad \mathbf{Z} = \mathbf{y} =$ 

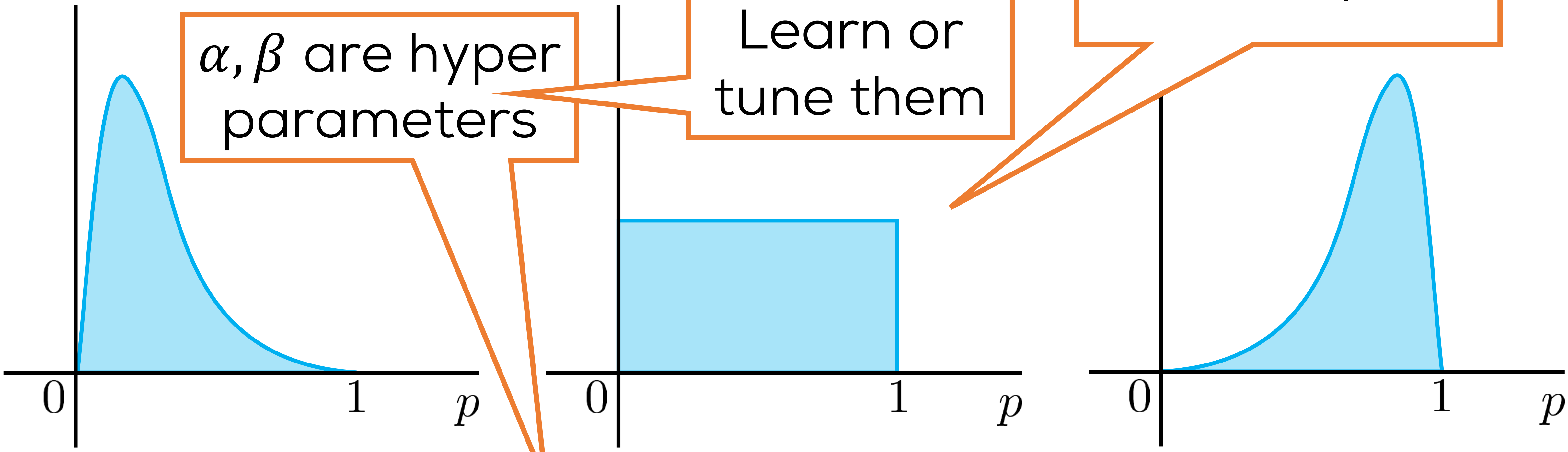
$$\mathbb{P}\left[y^i \,|\, p\right] = p^{y^i}(1-p)^{1-y^i}$$

$$\hat{p}_{\mathrm{MLE}} = \arg\max_p \mathbb{P}\left[\mathbf{y} \,|\, p\right]$$

$$= \arg\max_p \; \log \mathbb{P}\left[\mathbf{y} \,|\, p\right]$$

Exercise: Show that $\hat{p}_{\mathrm{MLE}} = \dfrac{n_H}{n}$

Here comes the Prior



α, β are hyper parameters

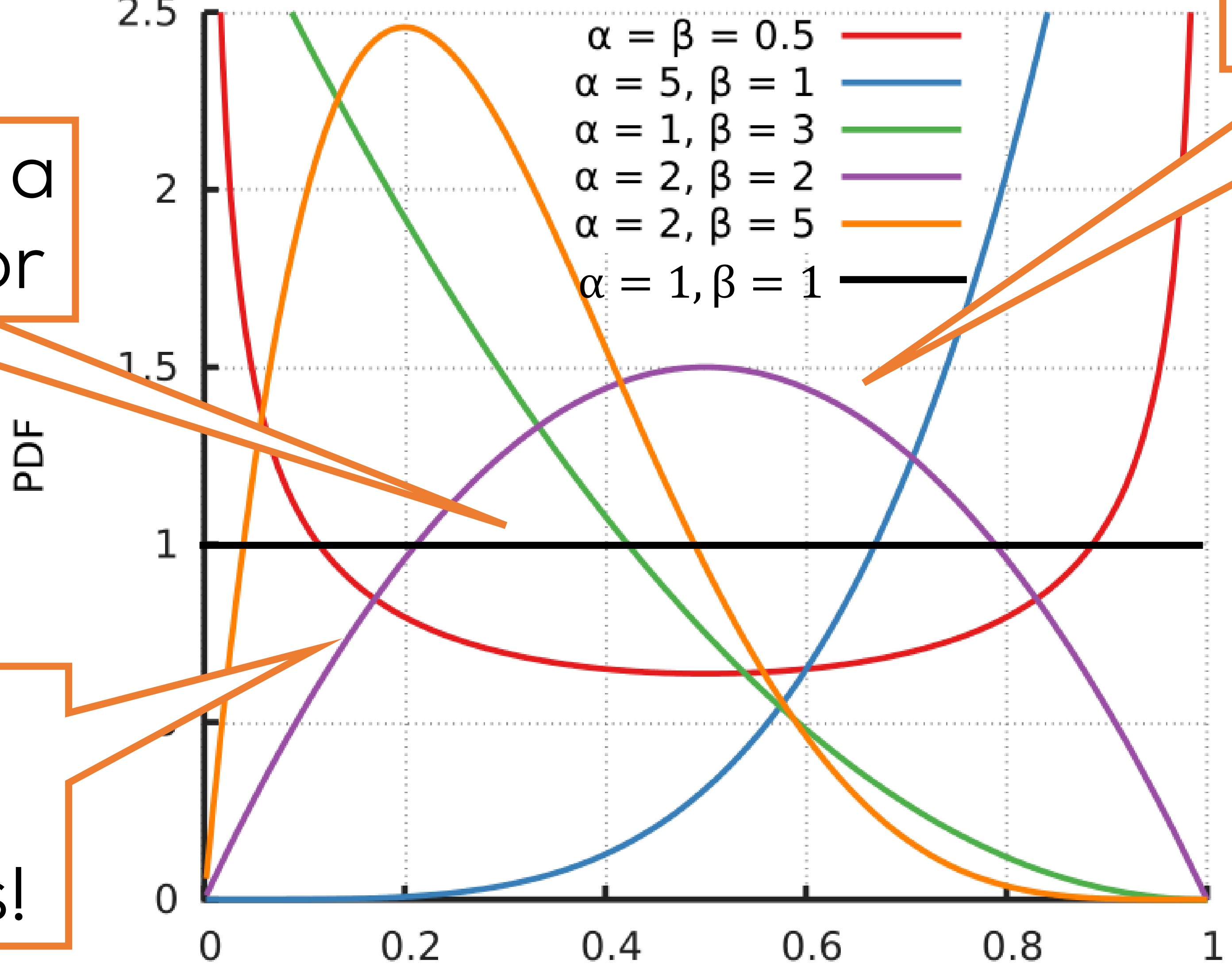
Learn or tune them

"Uniform prior"

$$\mathbb{P}\left[p\right] = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha - 1}(1-p)^{\beta - 1}$$

Beta prior
$\mathrm{Beta}(p; \alpha, \beta)$

Beta Distribution



wikipedia.com

Large α, β sharp peaks

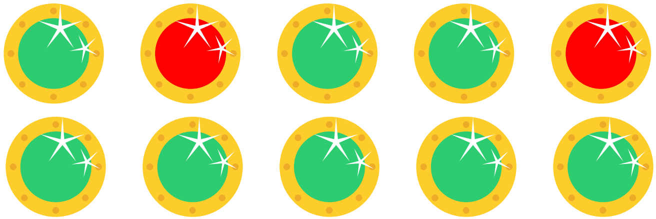
Can encode a uniform prior

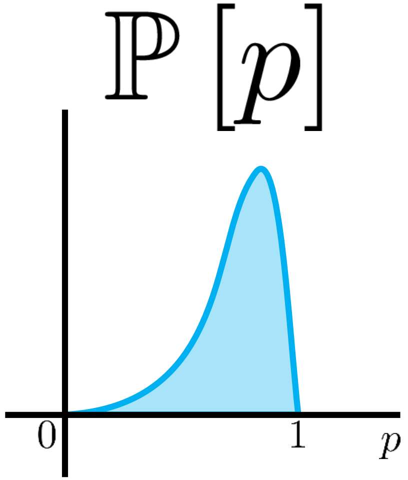
Encode previously seen tosses!

Maximum a-Posteriori

often affectionately called the MAP

The MAP Estimator

Bias p^* $\quad \mathbf{Z} = \mathbf{y} =$ 

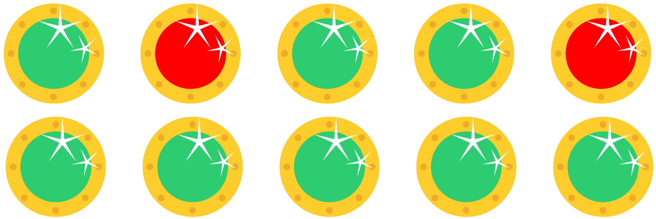


$$\mathbb{P}\left[y^i \,|\, p\right] = p^{y^i}(1-p)^{1-y^i}$$

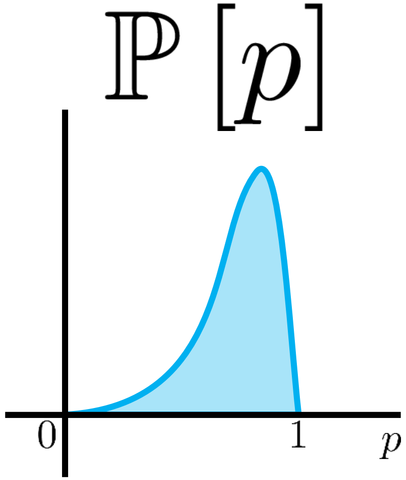
$$\hat{p}_{\text{MAP}} = \arg\max_p \mathbb{P}\left[p \,|\, \mathbf{y}\right]$$

$$= \arg\max_p \frac{\mathbb{P}\left[\mathbf{y} \,|\, p\right]\mathbb{P}\left[p\right]}{\mathbb{P}\left[\mathbf{y}\right]}$$

The MAP Estimator

Bias p^* $\quad \mathbf{Z} = \mathbf{y} =$ 

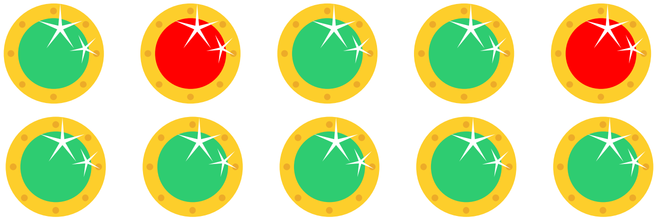
$$\mathbb{P}\left[y^i \,|\, p\right] = p^{y^i}(1-p)^{1-y^i}$$

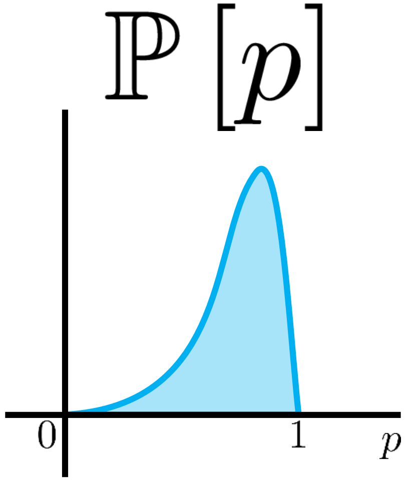


$$\hat{p}_{\text{MAP}} = \arg\max_p \mathbb{P}\left[p \,|\, \mathbf{y}\right]$$

$$= \arg\max_p \mathbb{P}\left[\mathbf{y} \,|\, p\right] \mathbb{P}\left[p\right]$$

The MAP Estimator

Bias p^* $\quad \mathbf{Z} = \mathbf{y} =$ 



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

$$\hat{p}_{\text{MAP}} = \arg\max_p \mathbb{P}\left[p \,|\, \mathbf{y}\right]$$

$$= \arg\max_p \; \log \mathbb{P}\left[\mathbf{y} \,|\, p\right] + \log \mathbb{P}\left[p\right]$$

Exercise: Show that $\hat{p}_{\text{MAP}} = \dfrac{n_H + \alpha - 1}{n + \alpha + \beta - 2}$

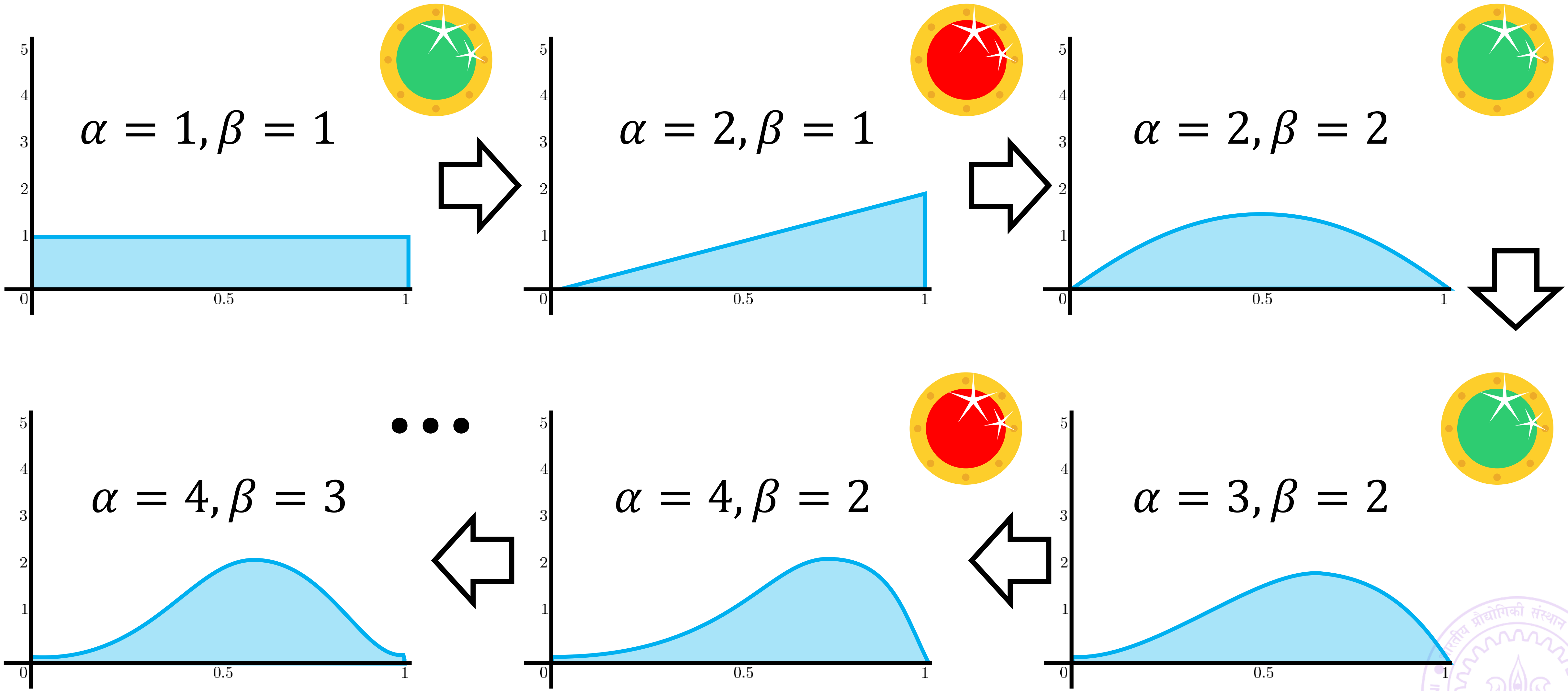
Online MAP!

- The posterior is the same "type" of distribution as the prior
- The posterior can be used as a prior by you
- The MAP update works even after witnessing a single toss
- Make incremental updates instead of one big update
- Online algorithm for MAP estimation!
- Efficient *online stochastic* updates

Conjugacy!

Exercise: Show that $\mathbb{P}\left[p \,|\, \mathbf{y}\right] = \mathrm{Beta}(p; \alpha + n_H, \beta + n_T)$

Online MAP!



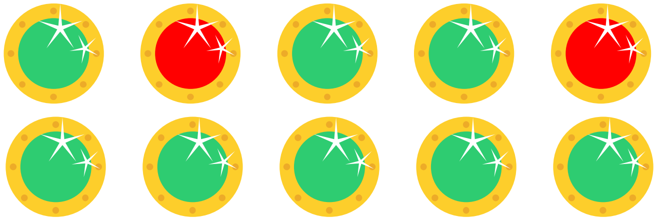
$\alpha = 1, \beta = 1$
$\alpha = 2, \beta = 1$
$\alpha = 2, \beta = 2$
$\alpha = 3, \beta = 2$
$\alpha = 4, \beta = 2$
$\alpha = 4, \beta = 3$

Posterior Averaging

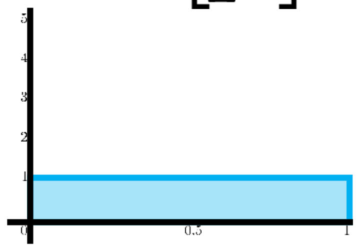
or more commonly known as Bayesian learning

Prediction with learnt models

 $= \; ?$

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

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



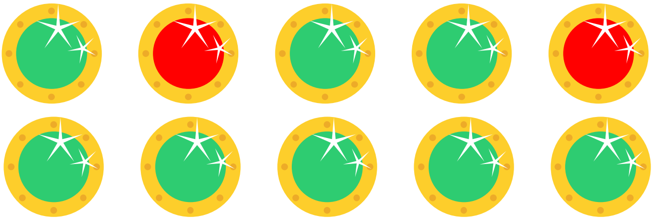
$$\mathbb{P}\left[y^i \mid p\right] = p^{y^i}(1-p)^{1-y^i}$$

$\mathbb{P}\left[\right.$  $= H \mid \mathbf{y}\left.\right] \approx \mathbb{P}\left[\right.$  $= H \mid \hat{p}_{\mathrm{MLE}}\left.\right] = \dfrac{n_H}{n}$

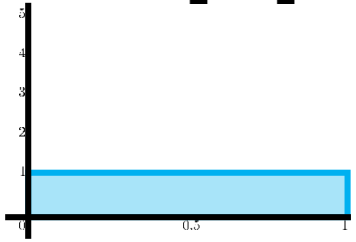
$\approx \mathbb{P}\left[\right.$  $= H \mid \hat{p}_{\mathrm{MAP}}\left.\right] = \dfrac{n_H + \alpha - 1}{n + \alpha + \beta - 2}$

Prediction with learnt models

 $= \ ?$

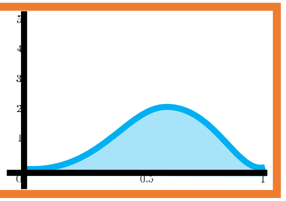
$\mathbf{Z} = \mathbf{y} =$ 

$\mathbb{P}[p]$



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

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



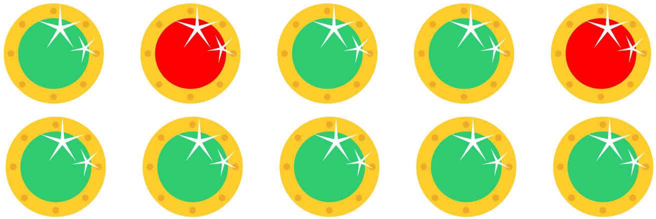
Apply Bayes rule properly

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

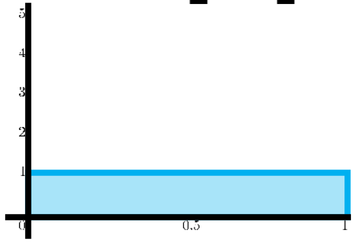
$\approx \mathbb{P}[$  $= H \mid \hat{p}_{\mathrm{MAP}}] = \dfrac{n_H + \alpha - 1}{n + \alpha + \beta - 2}$

Prediction with learnt models

 $= \; ?$

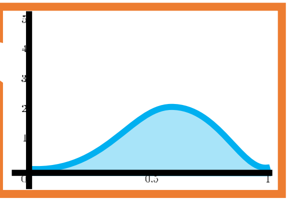
$\mathbf{Z} = \mathbf{y} =$ 

$\mathbb{P}[p]$



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

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

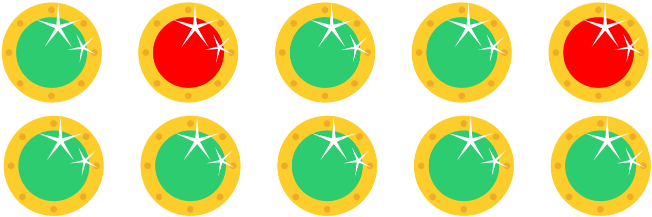


Apply Bayes rule properly

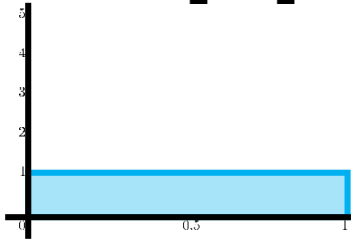
$$= \int_p p \cdot \mathbb{P}[p \mid \mathbf{y}]\, dp$$

Prediction with learnt models

 $= \, ?$

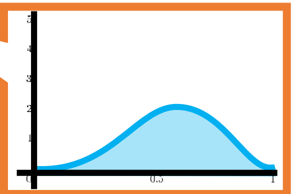
$\mathbf{Z} = \mathbf{y} =$ 

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



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

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

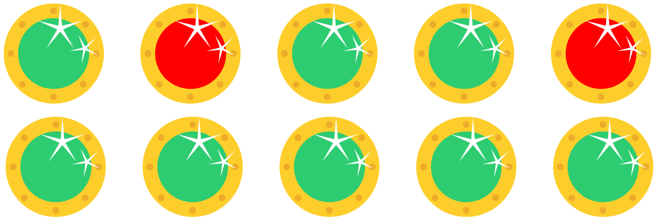


$$= \int_p p \cdot \mathrm{Beta}(\alpha + n_H, \beta + n_T)$$

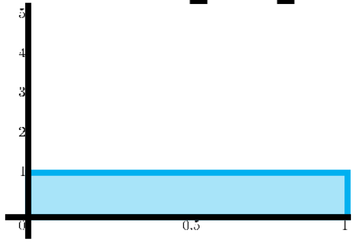
Apply Bayes rule properly

Prediction with learnt models

 $= \, ?$

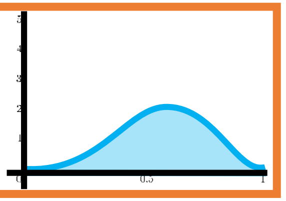
$\mathbf{Z} = \mathbf{y} =$ 

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



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

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



$$= \frac{\alpha + n_H}{\alpha + \beta + n}$$

Apply Bayes rule properly

Usually very challenging!

Exercise

Please give your Feedback

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