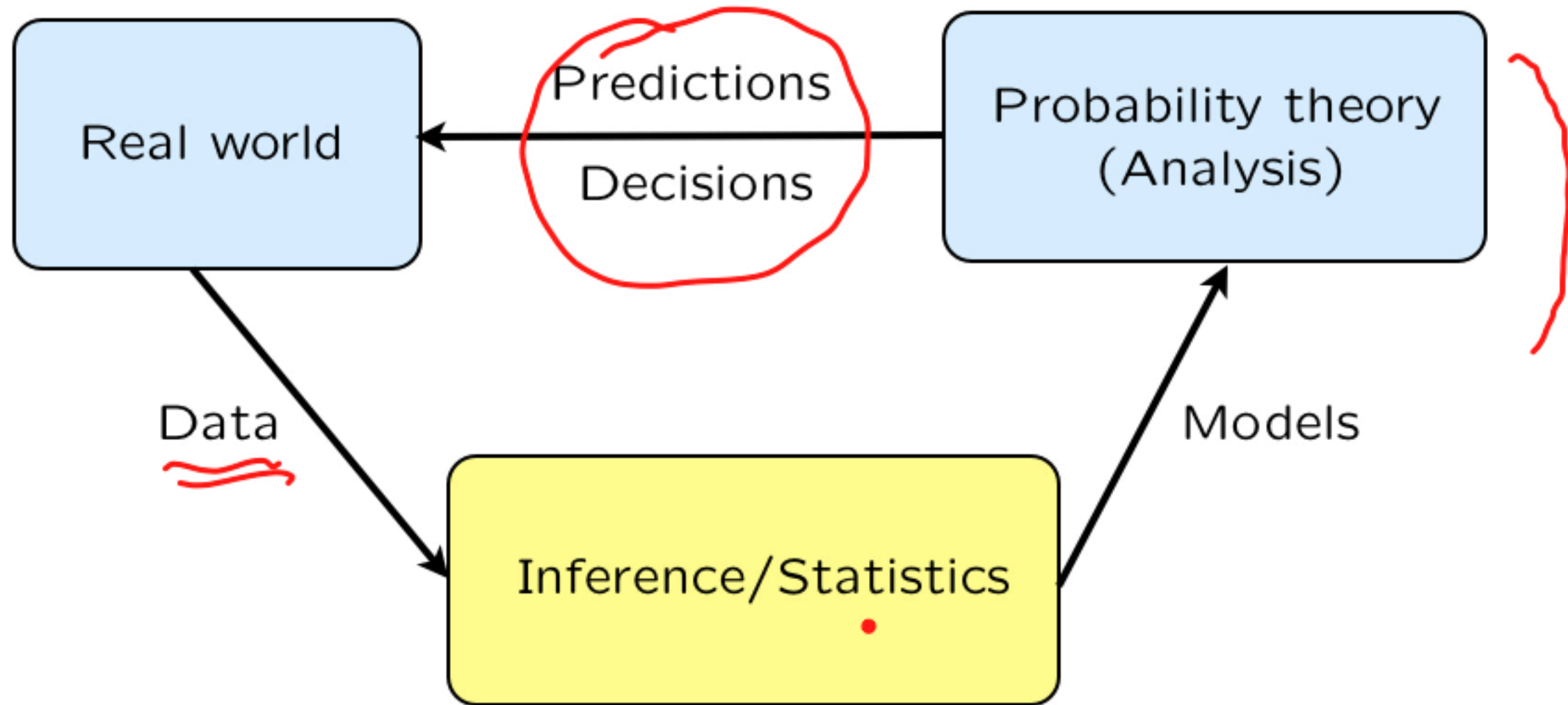


LECTURE 14: Introduction to Bayesian inference

- The big picture
 - motivation, applications
 - problem types (hypothesis testing, estimation, etc.)
- The general framework
 - Bayes' rule \rightarrow posterior
(4 versions)
 - point estimates (MAP, LMS)
 - performance measures)
(prob. of error; mean squared error)
 - examples

Inference: the big picture



Inference then and now

- Then:
10 patients were treated: 3 died
10 patients were not treated: 5 died
Therefore ...

Now:

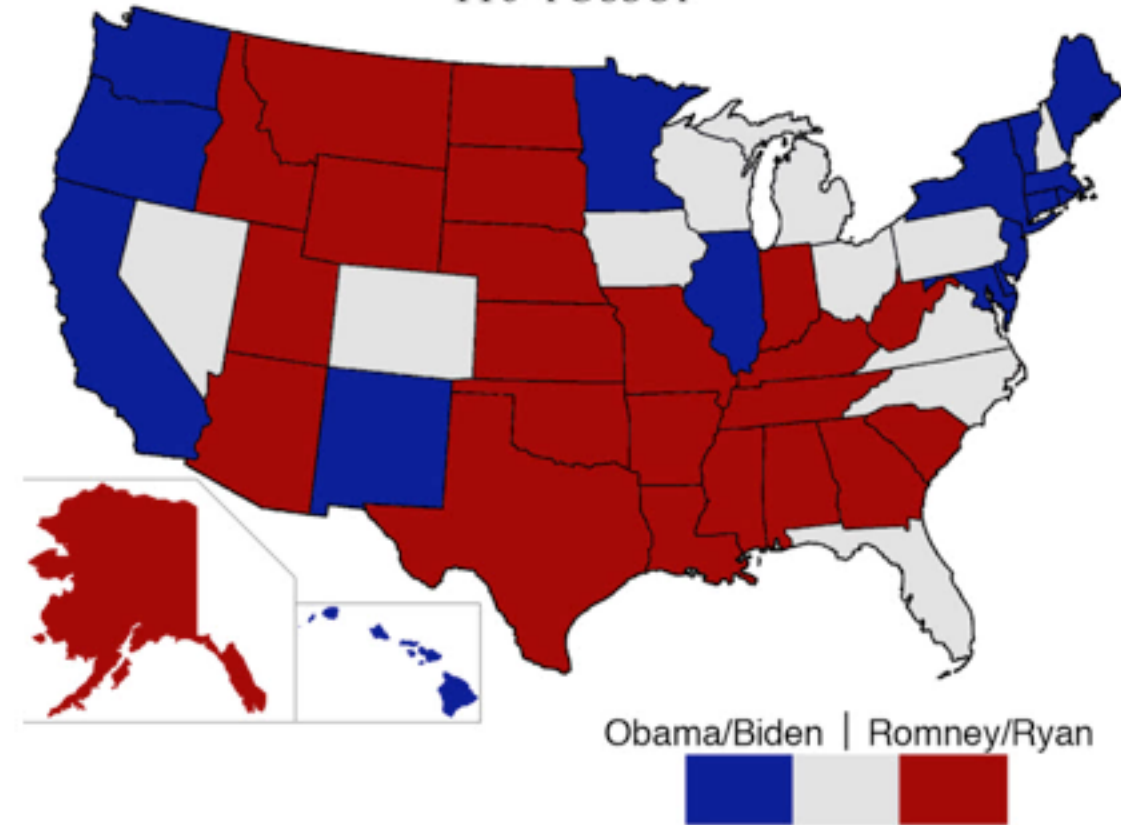
- Big data
- Big models
- Big computers

A sample of application domains

- Design and interpretation of experiments
 - polling •

STATE COUNTS (AND WASHINGTON, D.C.)
17 SOLIDLY DEMOCRATIC 23 SOLIDLY REPUBLICAN
11 TOSSUP

ELECTORAL VOTE COUNTS
237 LIKELY DEMOCRATIC 191 LIKELY REPUBLICAN
110 TOSSUP



A sample of application domains

- marketing, advertising
- recommendation systems
 - Netflix competition

persons

movie

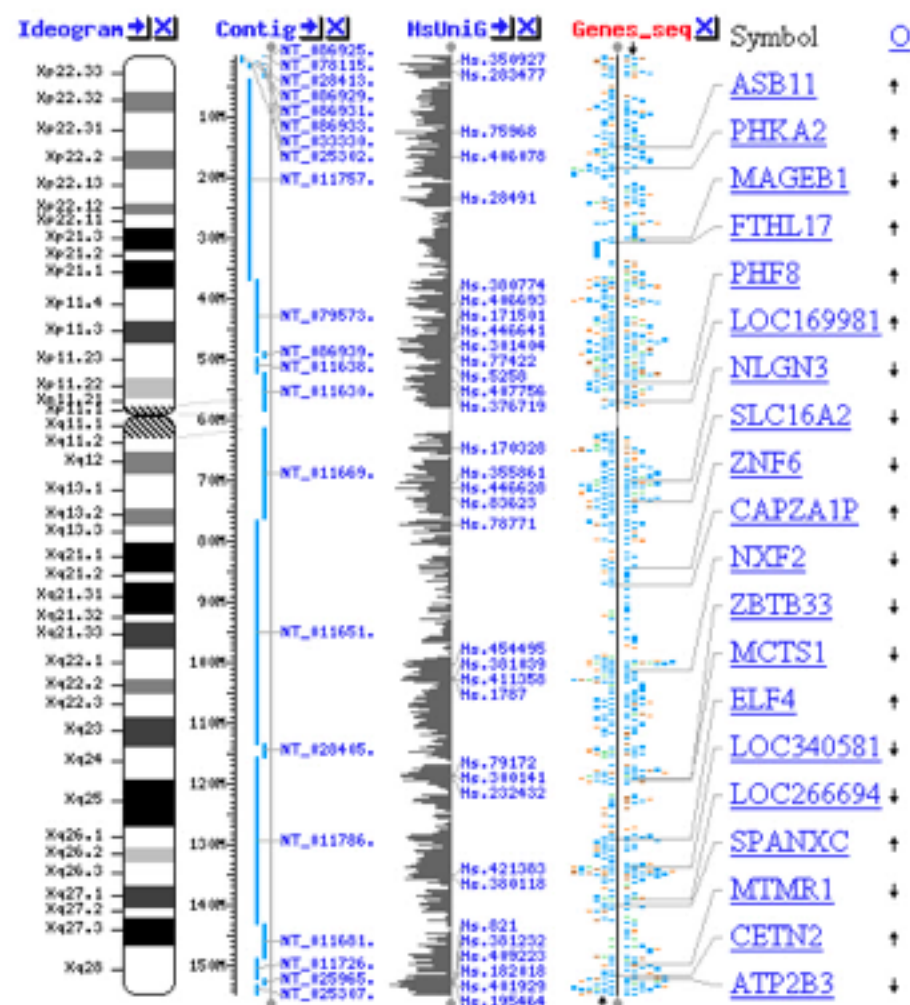
	2		1			4				5	
	5		4				?		1		3
		3		5			2				
4			?			5		3		?	
		4		1	3				5		
			2				1	?			4
	1					5		5		4	
		2		?	5		?		4		
	3		3		1		5		2		1
	3				1			2		3	
	4			5	1			3			
		3				3	?			5	
2	?		1		1						
		5			2	?		4		4	
	1		3		1	5		4		5	
1		2			4				5	?	

A sample of application domains

- Finance

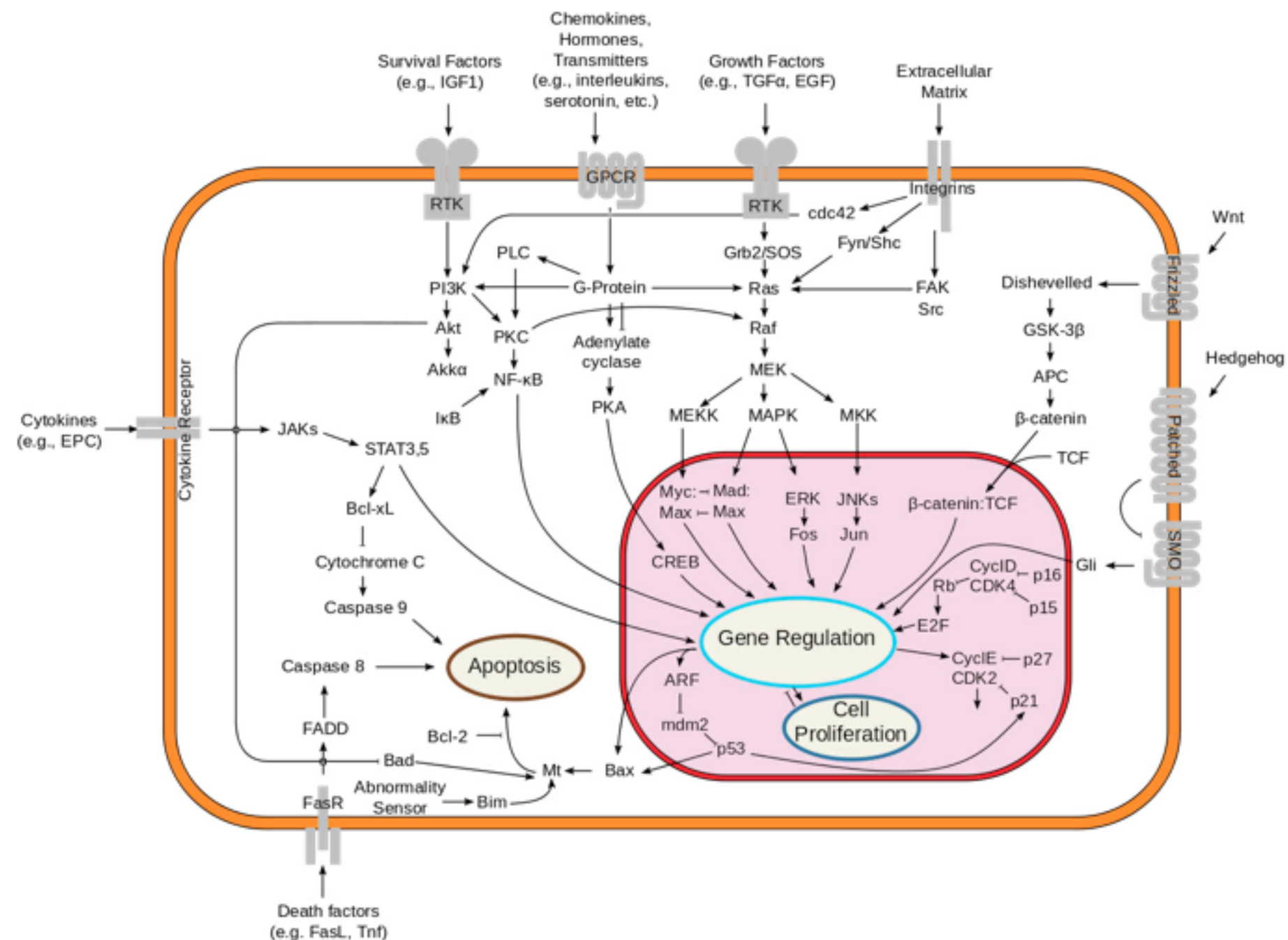


- genomics



- neuroscience, etc., etc.

- systems biology



A sample of application domains

- Modeling and monitoring the oceans
- Modeling and monitoring global climate
- Modeling and monitoring pollution
- Interpreting data from physics experiments •
- Interpreting astronomy data

A sample of application domains

- Signal processing
 - communication systems (noisy ...)
 - speech processing and understanding
 - image processing and understanding
 - tracking of objects
 - positioning systems (e.g., GPS)
 - detection of abnormal events

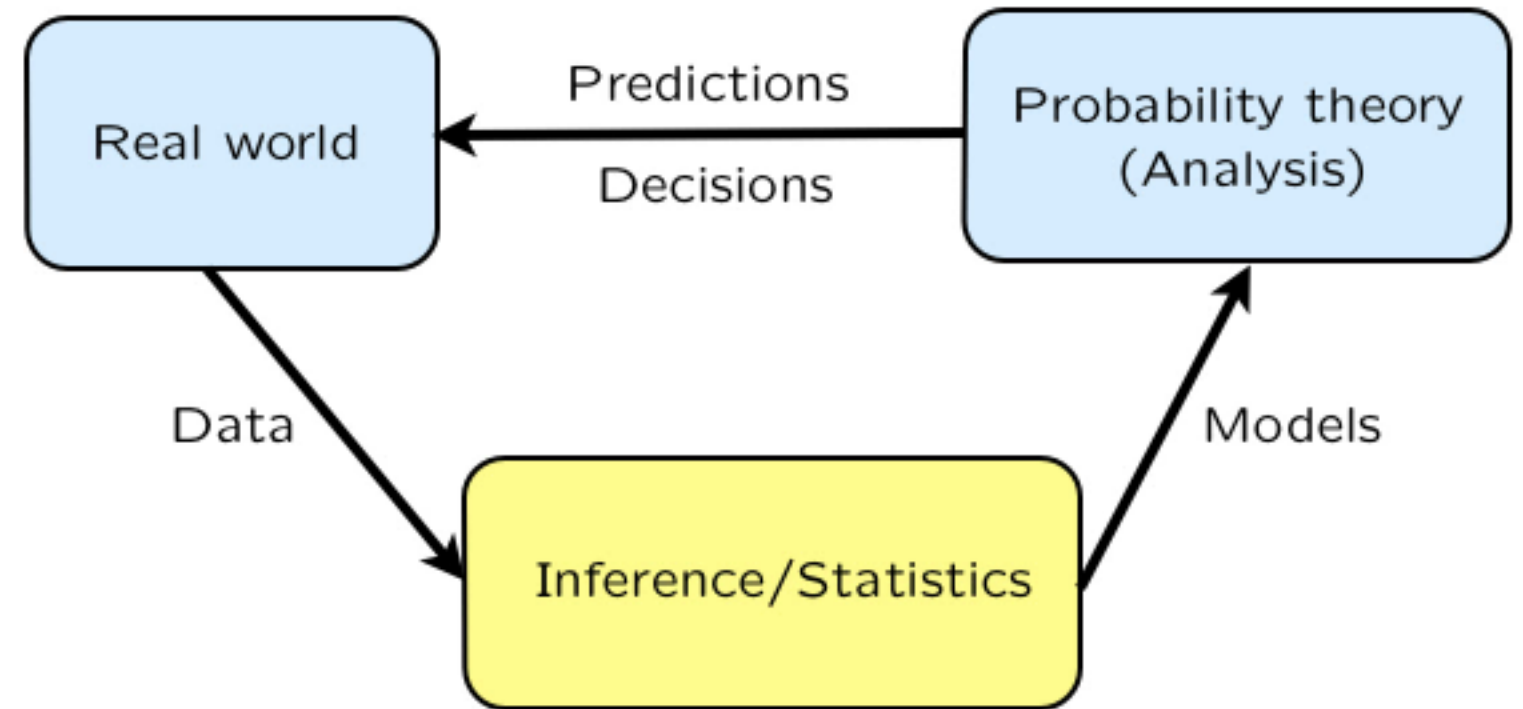


Model building versus inferring unobserved variables



$$X = aS + W$$

- Model building:
 - know “signal” S , observe X
 - infer a
- Variable estimation:
 - know a , observe X
 - infer S



Hypothesis testing versus estimation

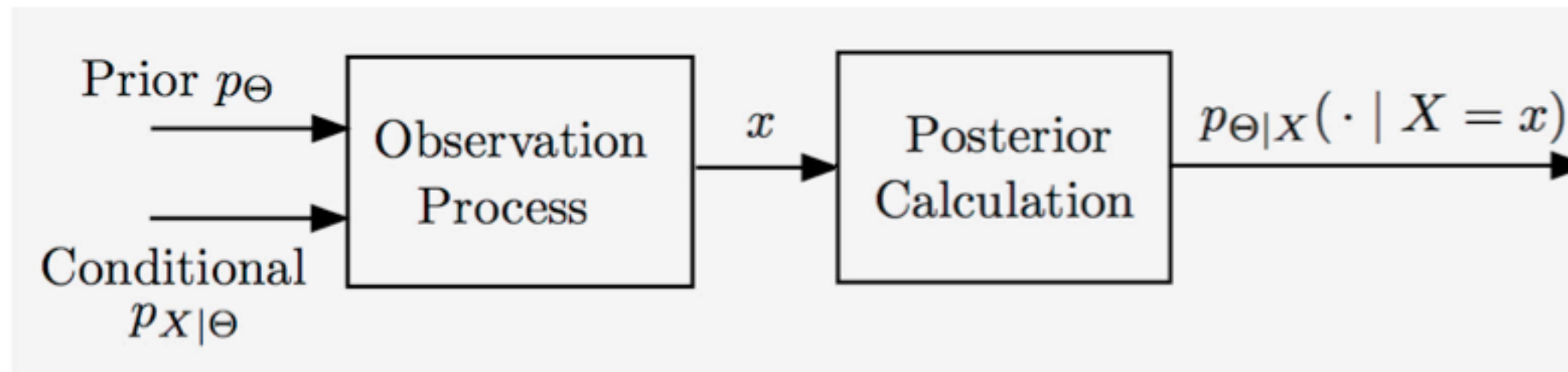
- Hypothesis testing:
 - unknown takes one of few possible values
 - aim at small probability of incorrect decision

Is it an airplane or a bird?

- Estimation:
 - numerical unknown(s)
 - aim at an estimate that is “close” to the true but unknown value

The Bayesian inference framework

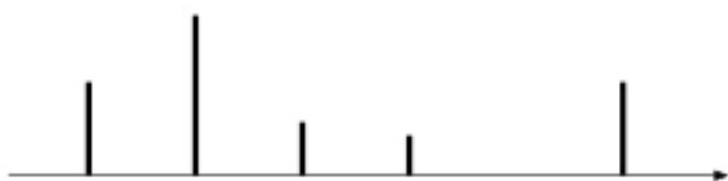
- Unknown Θ
 - treated as a random variable
 - prior distribution p_{Θ} or f_{Θ}
- Observation X
 - observation model $p_{X|\Theta}$ or $f_{X|\Theta}$
- Use appropriate version of the Bayes rule to find $p_{\Theta|X}(\cdot | X = x)$ or $f_{\Theta|X}(\cdot | X = x)$
- Where does the prior come from?
 - symmetry
 - known range
 - earlier studies
 - subjective or arbitrary



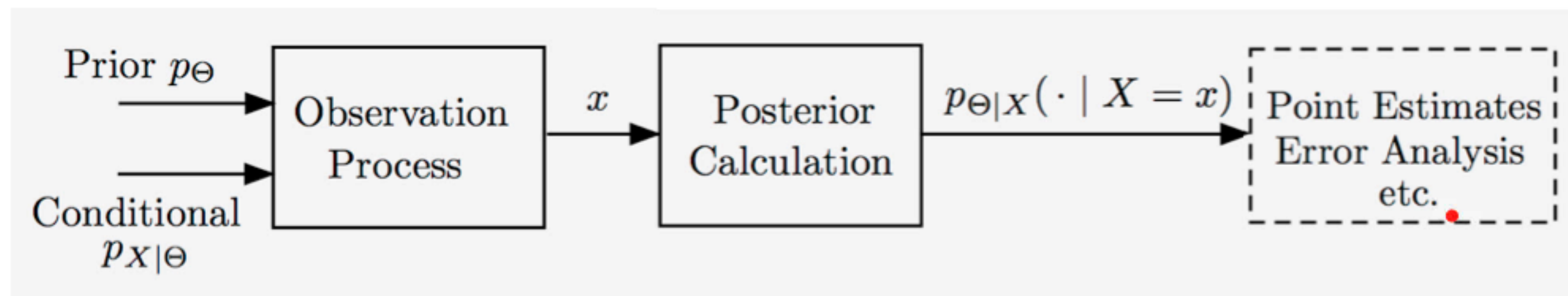
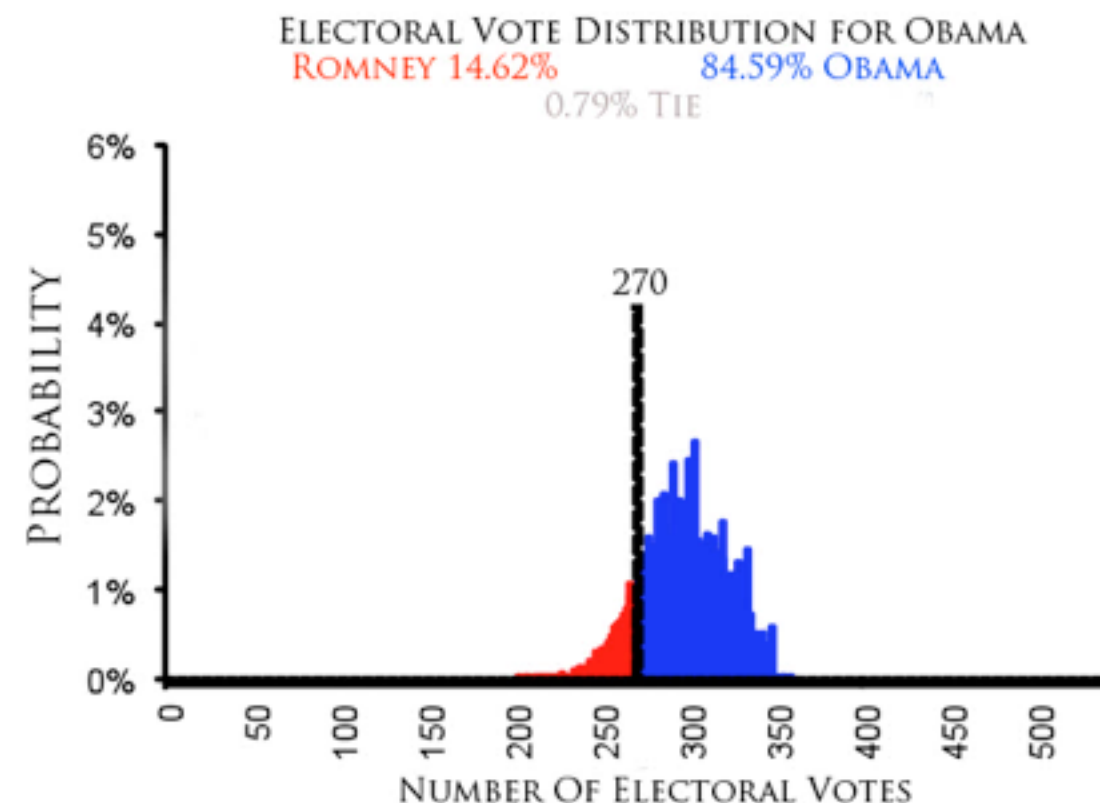
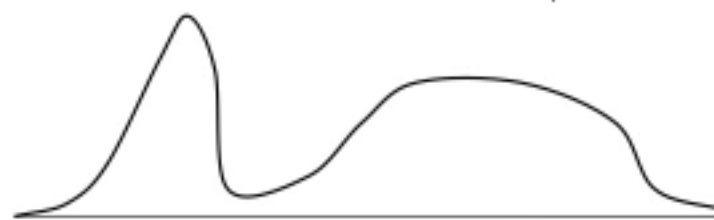
The output of Bayesian inference

The complete answer is a posterior distribution:
PMF $p_{\Theta|X}(\cdot | x)$ or PDF $f_{\Theta|X}(\cdot | x)$

$p_{\Theta|X}(\cdot | x)$

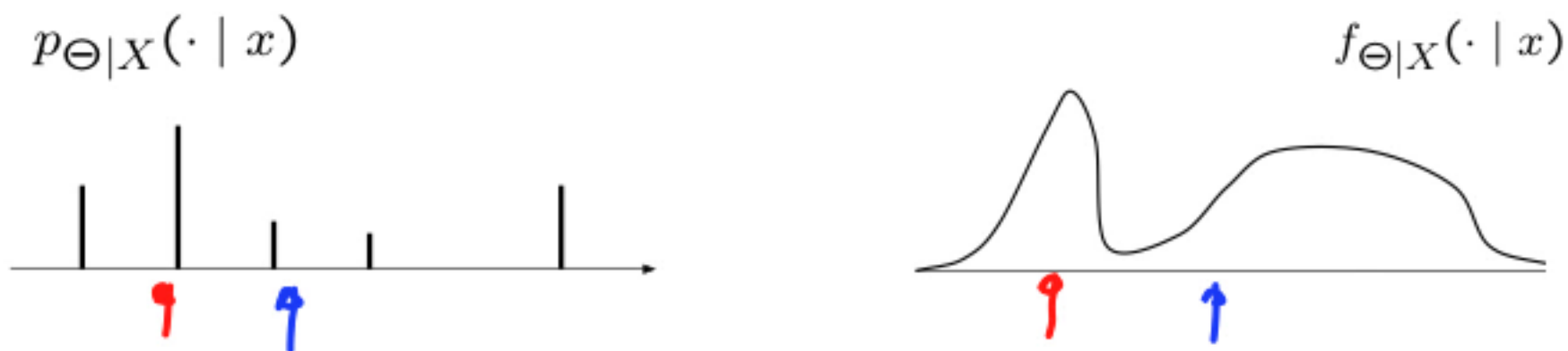


$f_{\Theta|X}(\cdot | x)$



Point estimates in Bayesian inference

The complete answer is a posterior distribution:
PMF $p_{\Theta|X}(\cdot | x)$ or PDF $f_{\Theta|X}(\cdot | x)$



estimate: $\hat{\theta} = g(x)$
(number)

estimator: $\hat{\Theta} = \underline{g(X)}$
(random variable)

- Maximum a posteriori probability (MAP):

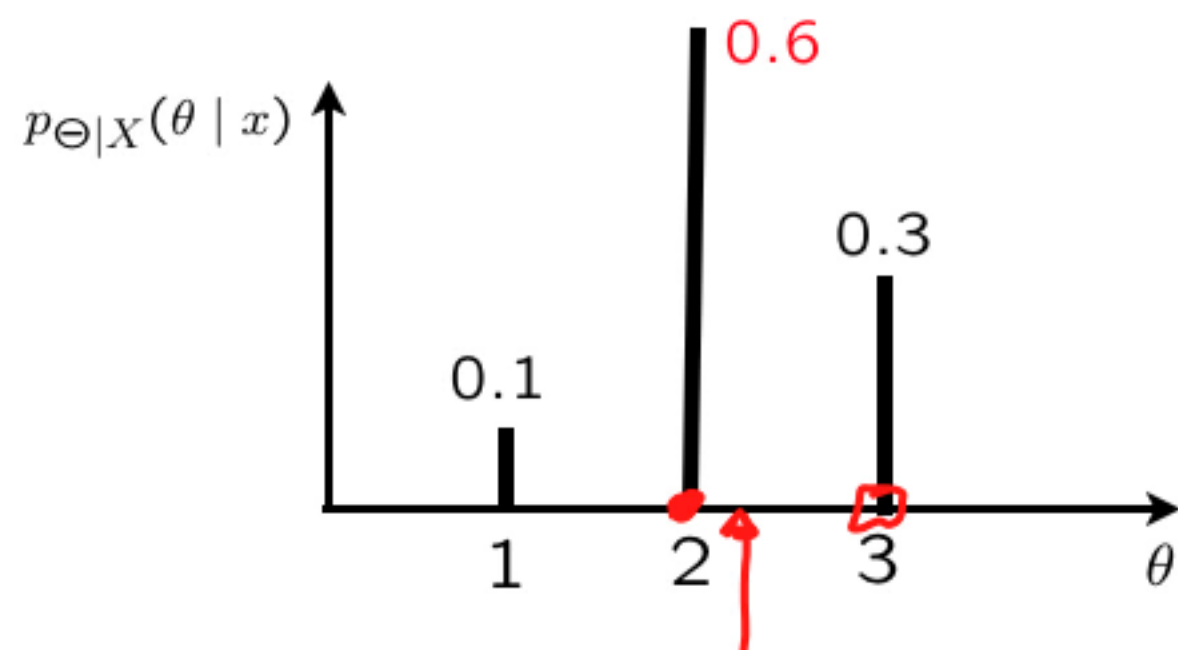
$$p_{\Theta|X}(\theta^* | x) = \max_{\theta} p_{\Theta|X}(\theta | x),$$

$$f_{\Theta|X}(\theta^* | x) = \max_{\theta} f_{\Theta|X}(\theta | x),$$

- Conditional expectation: $\mathbf{E}[\Theta | X = x]$ (LMS: Least Mean Squares)

Discrete Θ , discrete X

- values of Θ : alternative hypotheses



$$p_{\Theta|X}(\theta | x) = \frac{p_{\Theta}(\theta) p_{X|\Theta}(x | \theta)}{p_X(x)}$$

$$p_X(x) = \sum_{\theta'} p_{\Theta}(\theta') p_{X|\Theta}(x | \theta')$$

- conditional prob of error:

$$P(\hat{\theta} \neq \Theta | X = x) = 0.4$$

smallest under the MAP rule

- overall probability of error:

$$\begin{aligned} P(\hat{\Theta} \neq \Theta) &= \sum_x \underbrace{P(\hat{\Theta} \neq \Theta | X = x)}_{\text{smallest under the MAP rule}} p_X(x) \\ &= \sum_{\theta} P(\hat{\Theta} \neq \Theta | \Theta = \theta) p_{\Theta}(\theta) \end{aligned}$$

- MAP rule: $\hat{\theta} = 2$

$$LMS: \hat{\Theta} = E[\Theta | X=x] = 2.2$$

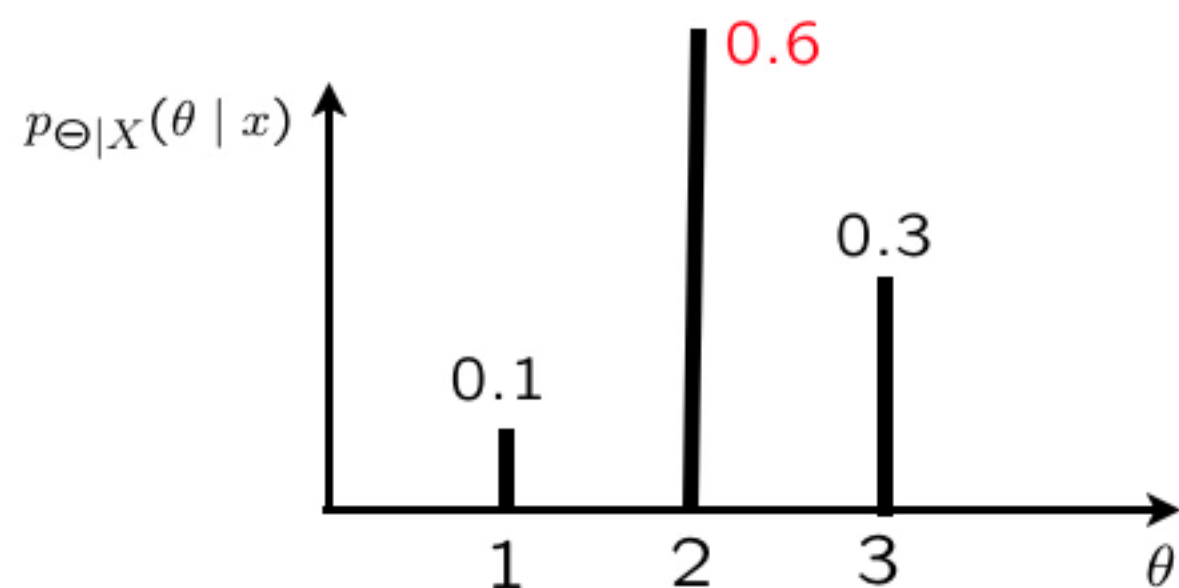
Discrete Θ , continuous X

- Standard example:
 - send signal $\Theta \in \{1, 2, 3\}$

$$X = \Theta + W$$

$W \sim N(0, \sigma^2)$, indep. of Θ

$$f_{X|\Theta}(x | \theta) = f_W(x - \theta)$$



- MAP rule: $\hat{\theta} = 2$

$$p_{\Theta|X}(\theta | x) = \frac{p_{\Theta}(\theta) f_{X|\Theta}(x | \theta)}{f_X(x)}$$

$$f_X(x) = \sum_{\theta'} p_{\Theta}(\theta') f_{X|\Theta}(x | \theta')$$

- conditional prob of error:

$$P(\hat{\theta} \neq \Theta | X = x)$$

→ **smallest under the MAP rule**

- overall probability of error:

$$\begin{aligned} P(\hat{\Theta} \neq \Theta) &= \int \underbrace{P(\hat{\Theta} \neq \Theta | X = x)}_{\text{MAP rule}} \underbrace{f_X(x) dx}_{\text{marginal density}} \\ &= \sum_{\theta} P(\hat{\Theta} \neq \theta | \Theta = \theta) p_{\Theta}(\theta) \end{aligned}$$

Continuous Θ , continuous X

- linear normal models
estimation of a noisy signal

$$X = \Theta + W$$

Θ and W : independent normals

multi-dimensional versions (many normal parameters, many observations)

- estimating the parameter of a uniform

$$X: \text{uniform}[0, \Theta]$$

$$\Theta: \text{uniform } [0, 1]$$

$$\underline{f_{\Theta|X}(\theta | x)} = \frac{f_{\Theta}(\theta) f_{X|\Theta}(x | \theta)}{f_X(x)}$$

$$f_X(x) = \int f_{\Theta}(\theta') f_{X|\Theta}(x | \theta') d\theta'$$

- $\hat{\Theta} = g(X)$ *MAP*
LMS

- interested in:

$$\left\{ \begin{array}{l} \mathbf{E}[(\hat{\Theta} - \Theta)^2 | X = x] \\ \mathbf{E}[(\hat{\Theta} - \Theta)^2] \end{array} \right.$$

Inferring the unknown bias of a coin and the Beta distribution

- Standard example:
 - coin with bias Θ ; prior $f_{\Theta}(\cdot)$
 - fix n ; K = number of heads
- Assume $f_{\Theta}(\cdot)$ is uniform in $[0, 1]$

$$f_{\Theta|K}(\theta | k) = \frac{f_{\Theta}(\theta) p_{K|\Theta}(k | \theta)}{p_K(k)}$$

$$p_K(k) = \int f_{\Theta}(\theta') p_{K|\Theta}(k | \theta') d\theta'$$

$$f_{\Theta|K}(\theta | k) = \frac{1 \cdot \binom{n}{k} \theta^k (1-\theta)^{n-k}}{p_K(k)}$$

$$\underline{\underline{\theta \in [0, 1]}}$$

$$= \frac{1}{d(n, k)} \theta^k (1-\theta)^{n-k} \quad \text{"Beta distribution, with parameters } (k+1, n-k+1)\text{"}$$

- If prior is Beta: $f_{\Theta}(\theta) = \frac{1}{c} \theta^{\alpha} (1-\theta)^{\beta}$ $\alpha, \beta \geq 0$

$$f_{\Theta|K}(\theta | k) = \frac{\frac{1}{c} \theta^{\alpha} (1-\theta)^{\beta} \binom{n}{k} \theta^k (1-\theta)^{n-k}}{p_K(k)} = d \theta^{\alpha+k} (1-\theta)^{\beta+n-k}$$

Inferring the unknown bias of a coin: point estimates

- Standard example:
 - coin with bias Θ ; prior $f_{\Theta}(\cdot)$
 - fix n ; K = number of heads

- Assume $f_{\Theta}(\cdot)$ is uniform in $[0, 1]$

$$f_{\Theta|K}(\theta | k) = \frac{1}{d(n, k)} \theta^k (1 - \theta)^{n-k}$$

- MAP estimate:

$$\hat{\theta}_{\text{MAP}} = k/n$$

$$\max_{\theta} [k \log \theta + (n-k) \log(1-\theta)]$$

$$k/\theta - (n-k)/(1-\theta) = 0$$

$$\hat{\Theta}_{\text{MAP}} = K/n$$

$$\int_0^1 \theta^{\alpha} (1 - \theta)^{\beta} d\theta = \frac{\alpha! \beta!}{(\alpha + \beta + 1)!}$$

$$\alpha \geq 0$$

$$\beta \geq 0$$

$$\mathbb{E}[\Theta | K = k] = \int_0^1 \theta f_{\Theta|K}(\theta | k) d\theta$$

$$= \frac{1}{d(n, k)} \int_0^1 \theta^{k+1} (1 - \theta)^{n-k} d\theta$$

$$= \frac{1}{\frac{k! (n-k)!}{(n+1)!}} \cdot \frac{(k+1)! (n-k)!}{(n+2)!}$$

$$= \frac{k+1}{n+2} \approx \frac{k}{n} \text{ (n large)}$$

Summary

- Problem data: $p_{\Theta}(\cdot)$, $p_{X|\Theta}(\cdot | \cdot)$
- Given the value x of X : **find**, e.g., $p_{\Theta|X}(\cdot | x)$
 - using appropriate version of the Bayes rule *(4 choices)*
- Estimator $\hat{\Theta} = g(X)$ Estimate $\hat{\theta} = g(x)$
 - **MAP**: $\hat{\theta}_{\text{MAP}} = g_{\text{MAP}}(x)$ maximizes $p_{\Theta|X}(\theta | x)$
 - **LMS**: $\hat{\theta}_{\text{LMS}} = g_{\text{LMS}}(x) = \underset{\bullet}{\mathbf{E}}[\Theta | X = x]$