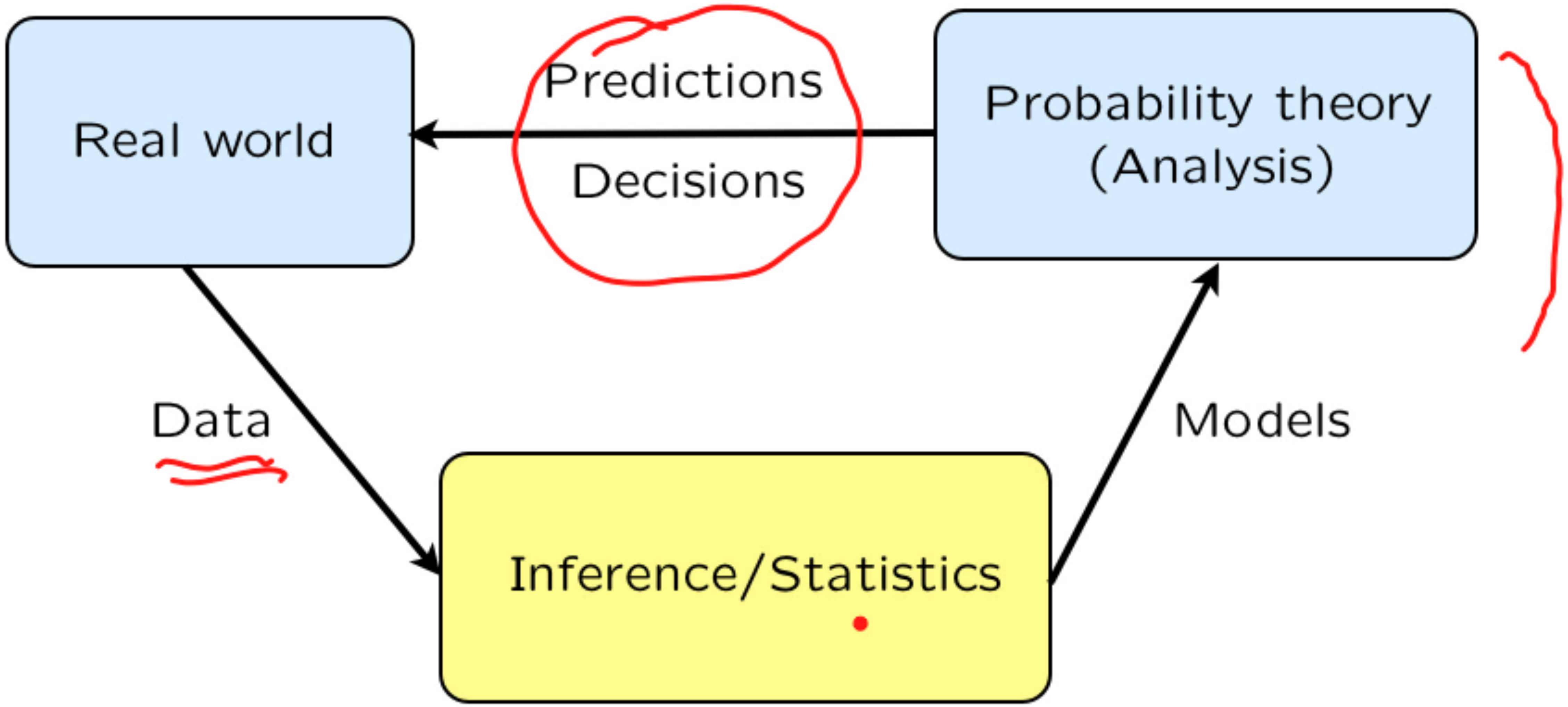
LECTURE 14: Introduction to Bayesian inference

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

Inference: the big picture



Inference then and now

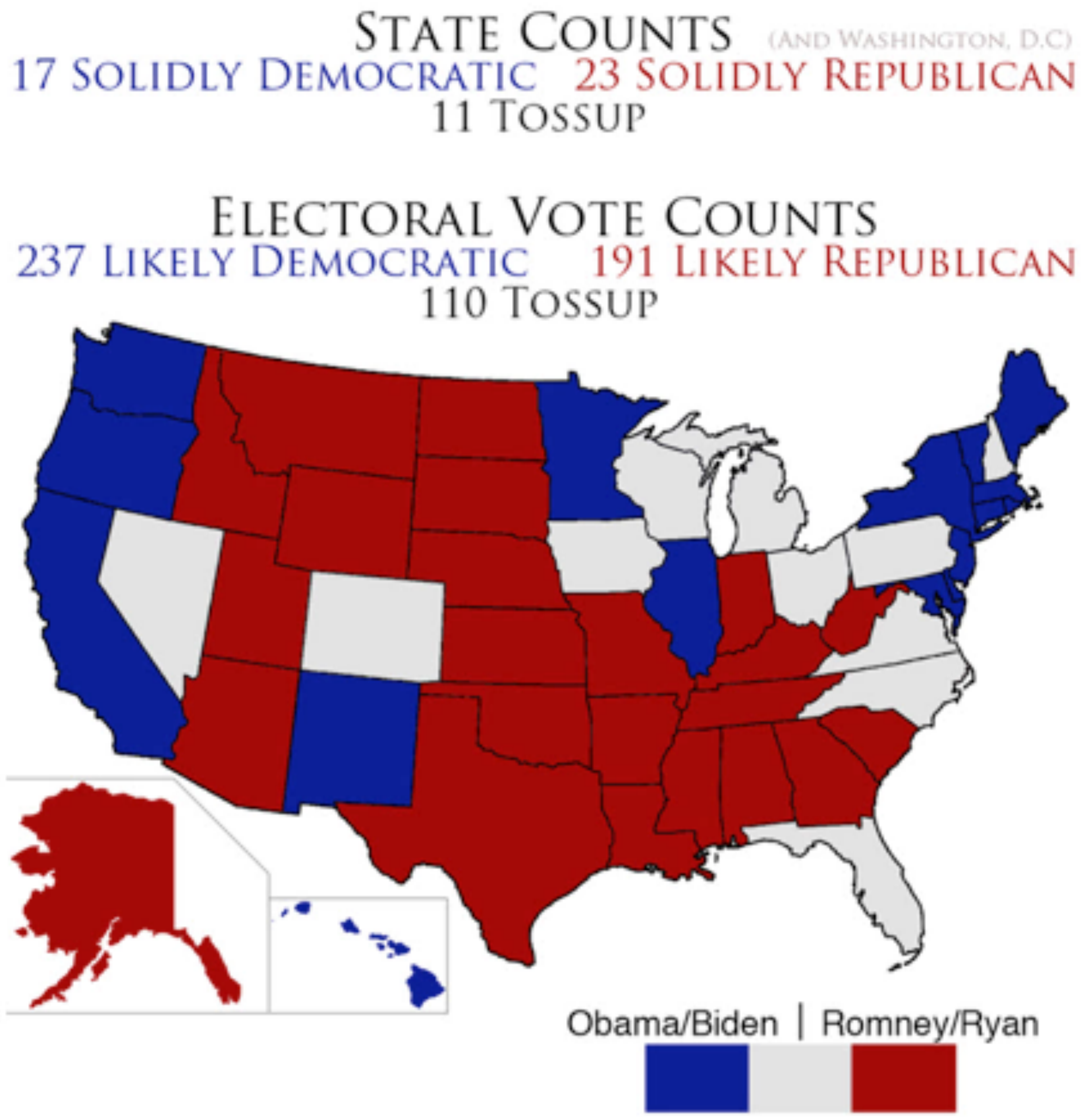
- Then:
  10 patients were treated: 3 died
  10 patients were not treated: 5 died
  Therefore ...

Now:

- Big data
- Big models
- Big computers

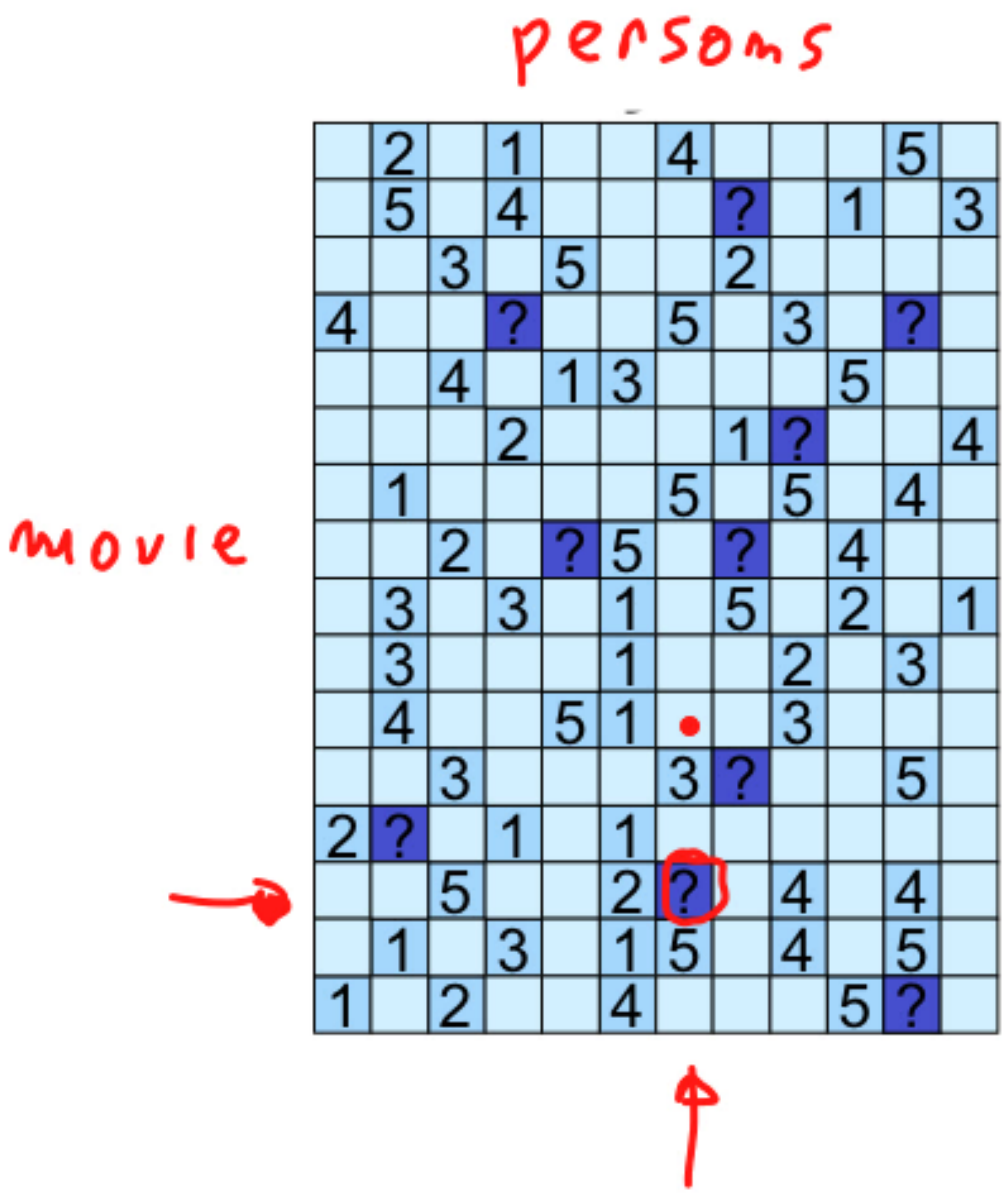
A sample of application domains

- Design and interpretation of experiments
  - polling



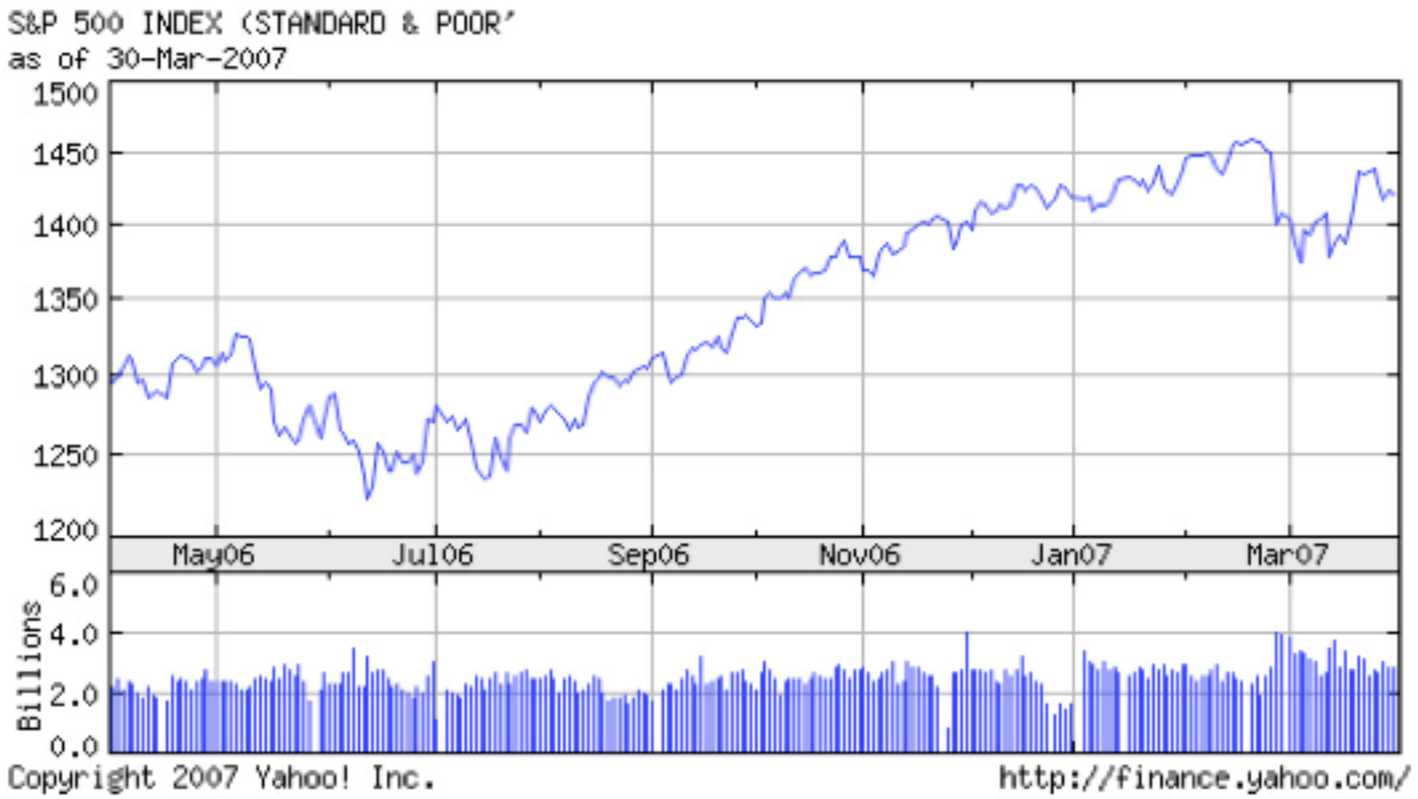
A sample of application domains

- marketing, advertising
- recommendation systems
  - Netflix competition



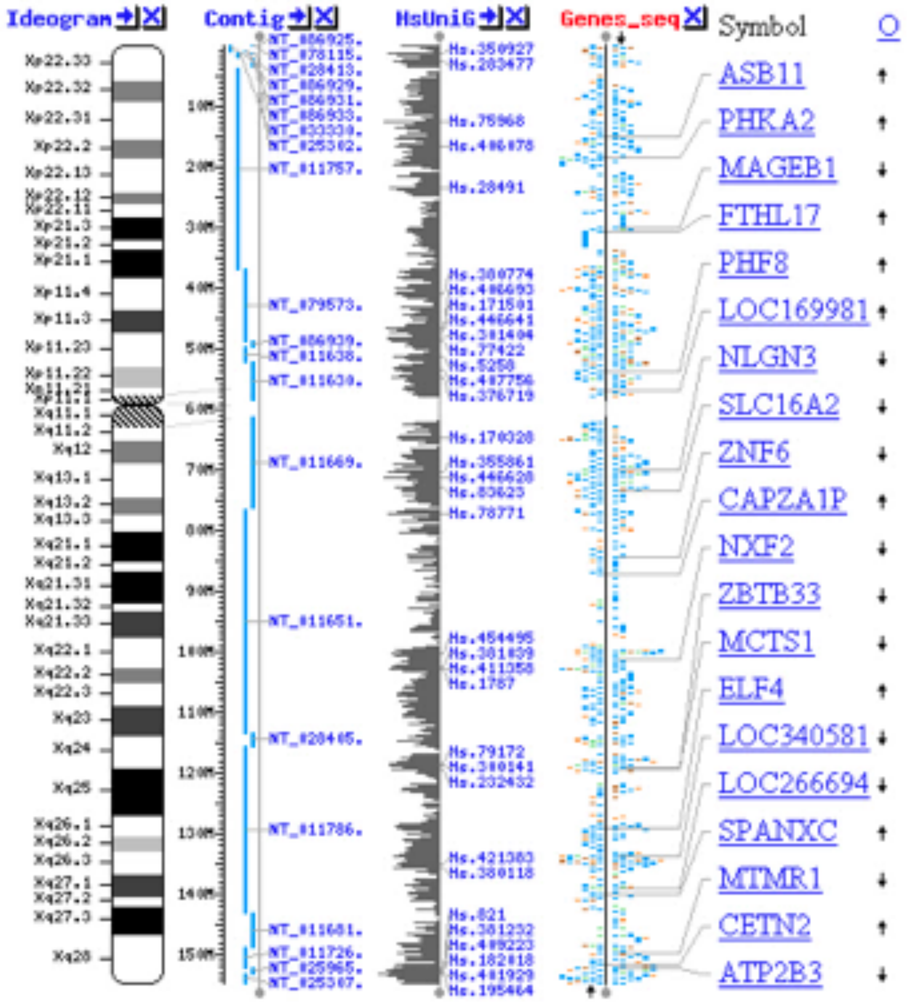
A sample of application domains

- Finance

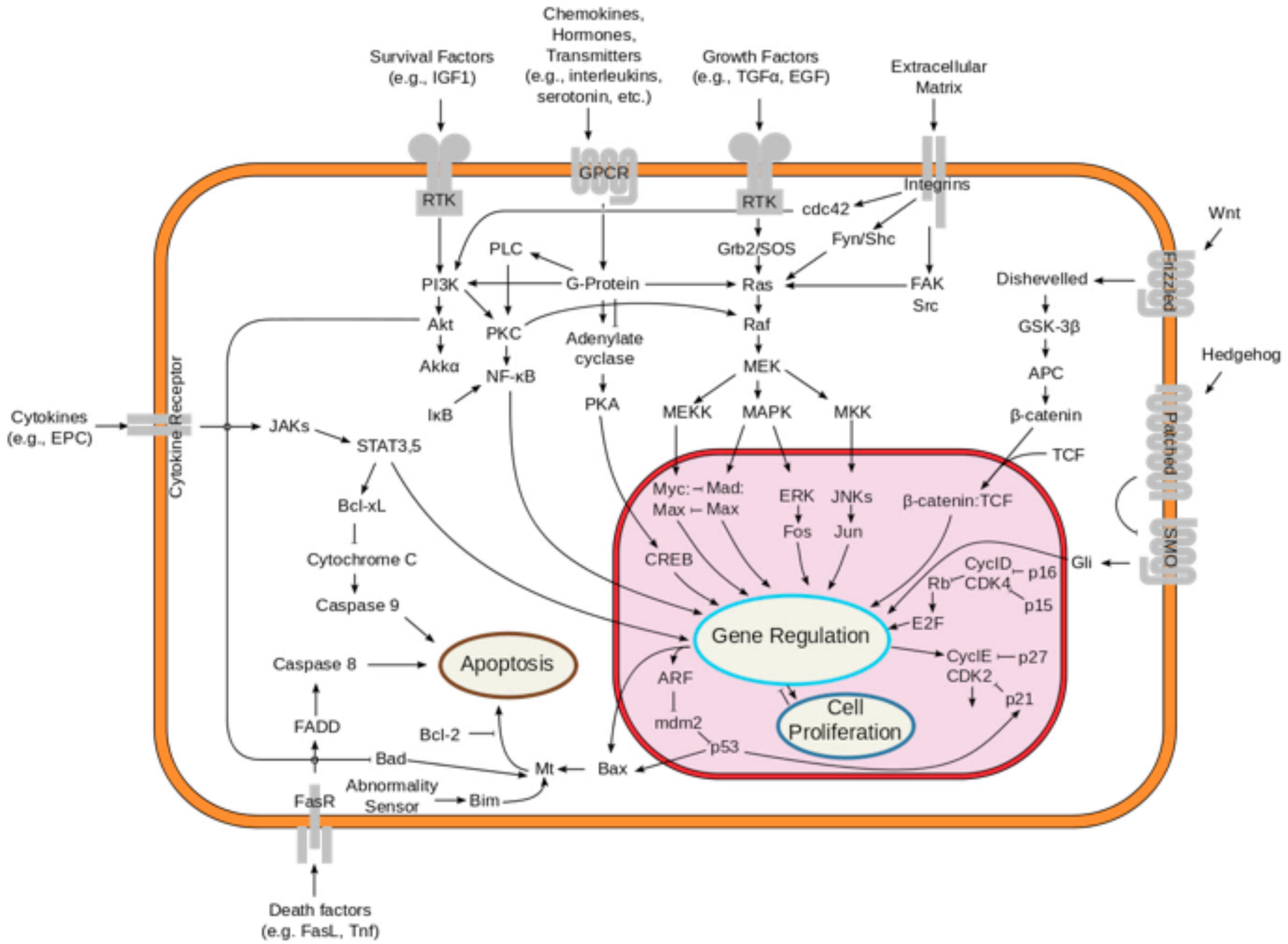


A sample of application domains

- Life sciences
  - genomics



  - systems biology



  - neuroscience, etc., etc.

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

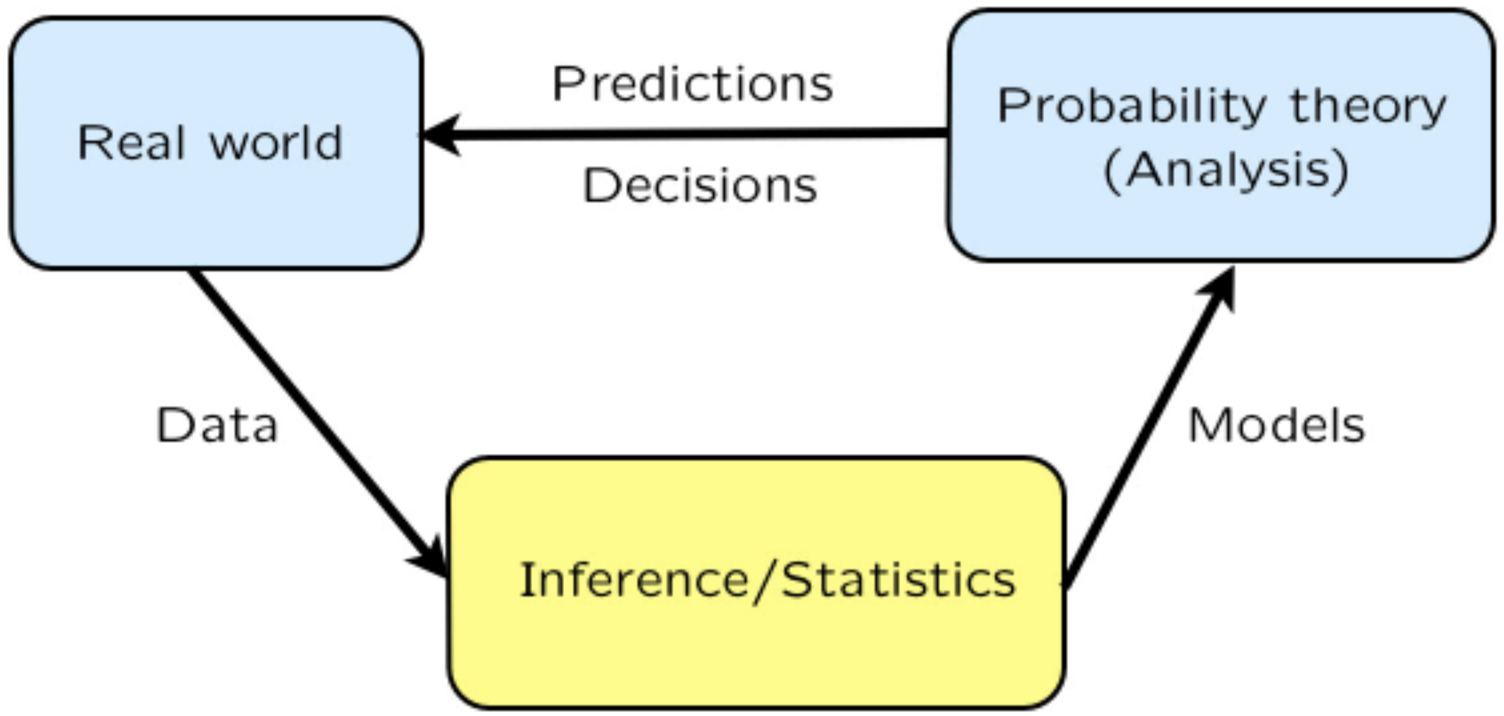
- Signal processing
  - communication systems (noisy . . .)
  - speech processing and understanding
  - image processing and understanding
  - tracking of objects
  - positioning systems (e.g., GPS)
  - detection of abnormal events

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

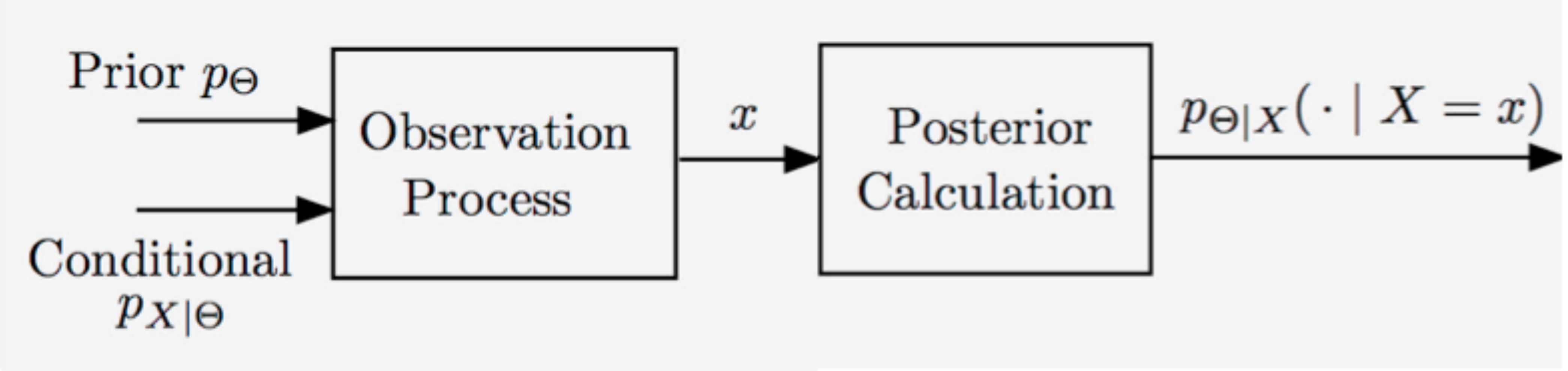
  Is it an airplane or a bird?

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

The Bayesian inference framework

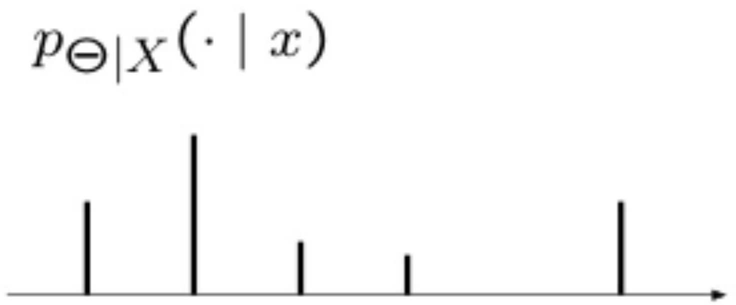
- Unknown Θ
  - treated as a random variable
  - prior distribution p_Θ or f_Θ
- Observation X
  - observation model $p_{X|\Theta}$ or $f_{X|\Theta}$
- Use appropriate version of the Bayes rule to find $p_{\Theta|X}(\cdot \mid X=x)$ or $f_{\Theta|X}(\cdot \mid X=x)$

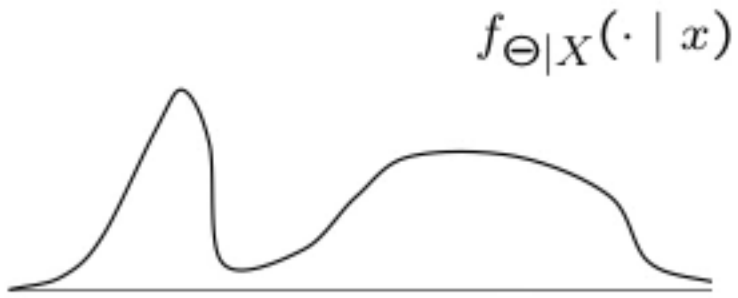
- Where does the prior come from?
  - symmetry
  - known range
  - earlier studies
  - subjective or arbitrary

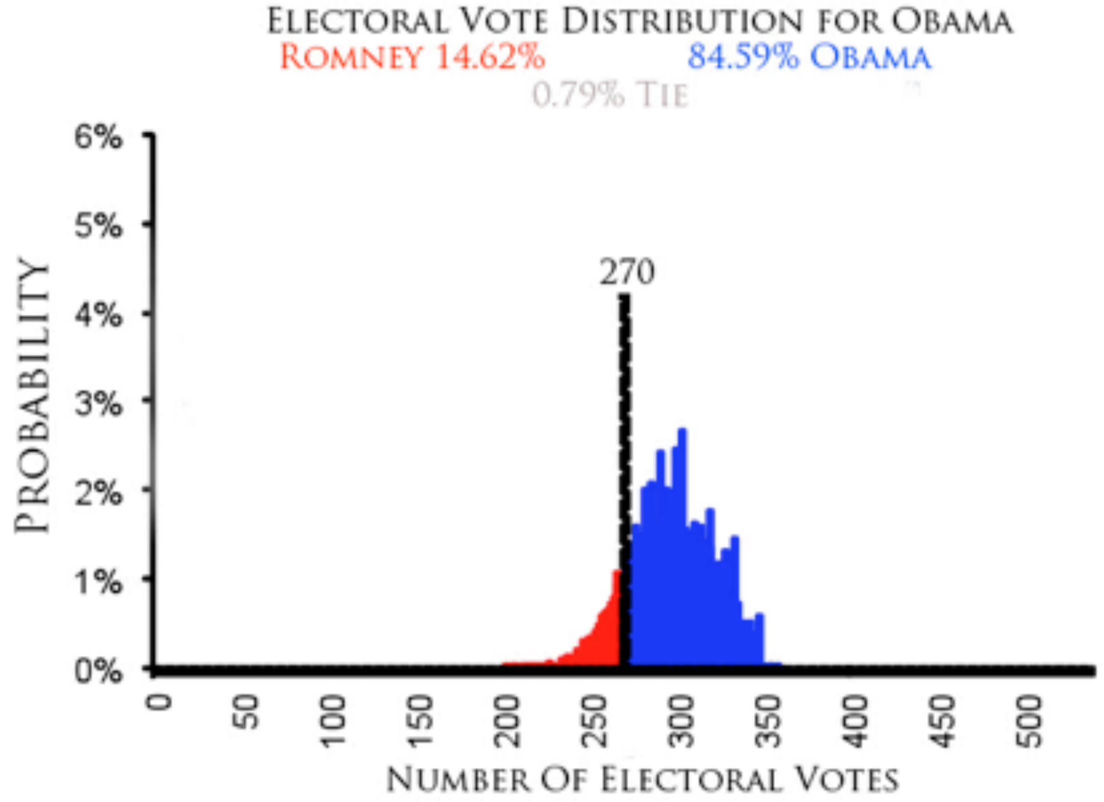


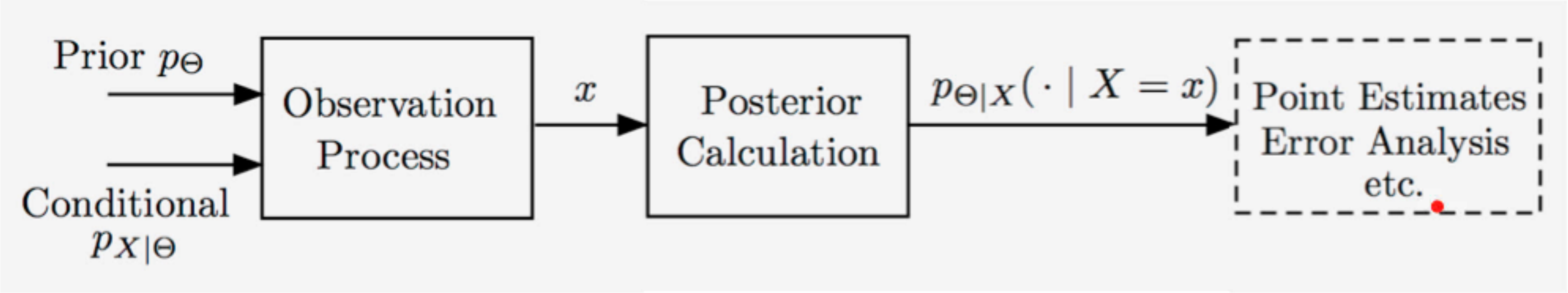
The output of Bayesian inference

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



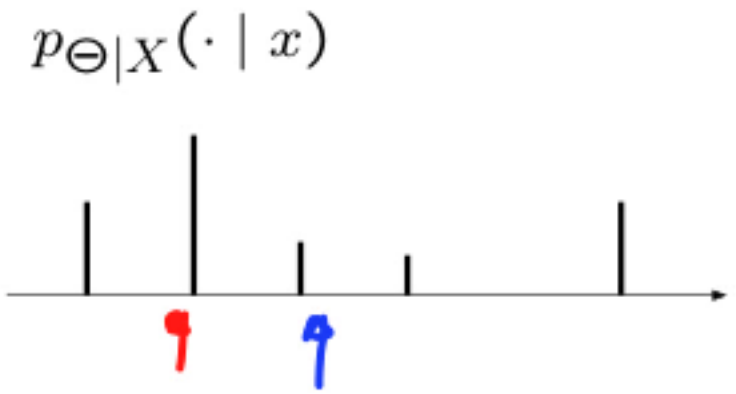


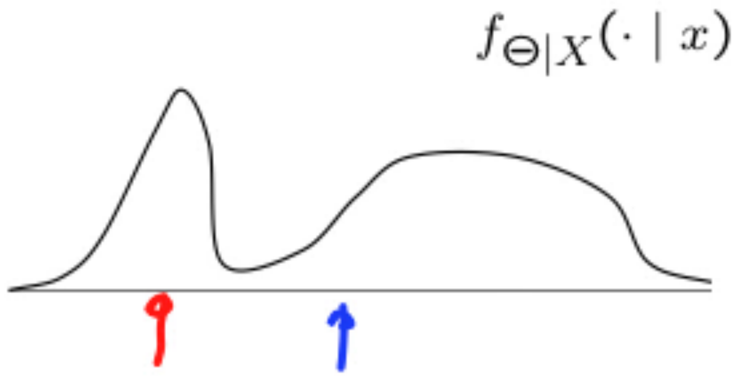




Point estimates in Bayesian inference

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





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

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

- Maximum a posteriori probability (MAP):

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

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

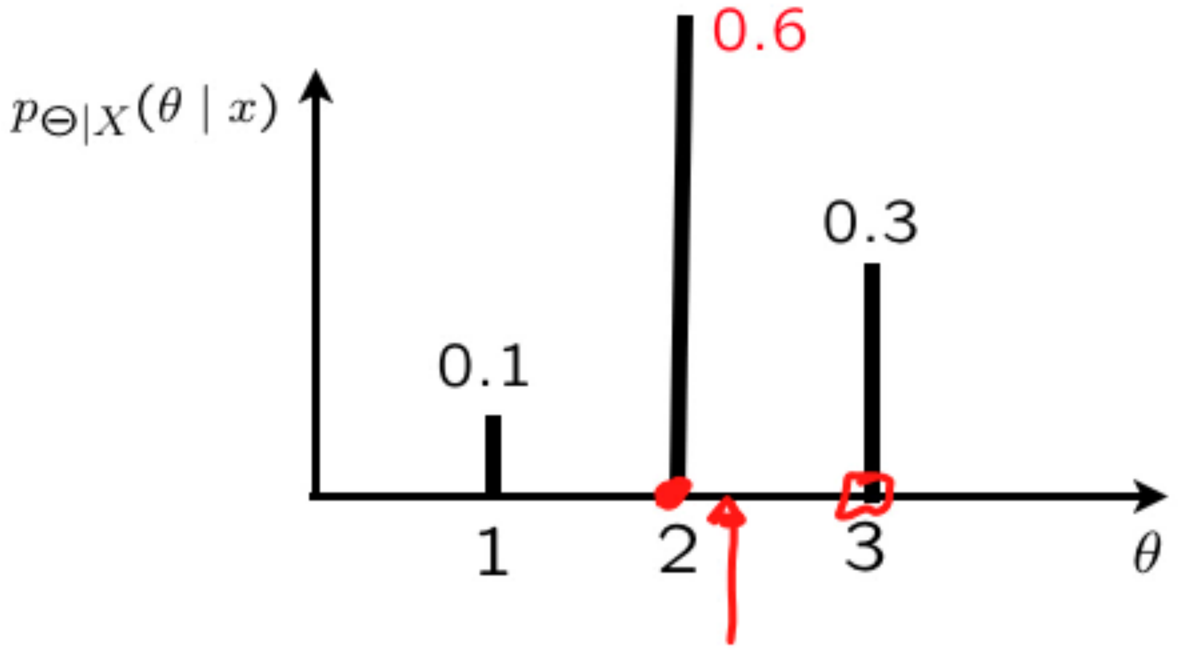
- Conditional expectation: $\mathbf{E}[\Theta \mid X = x]$ (LMS: Least Mean Squares)

Discrete Θ, discrete X

- values of Θ: alternative hypotheses

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

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



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

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

- conditional prob of error:

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

smallest under the MAP rule

- overall probability of error:

$$\mathbf{P}(\widehat{\Theta} \neq \Theta) = \sum_x \underbrace{\mathbf{P}(\widehat{\Theta} \neq \Theta \mid X = x)}\, p_X(x)$$

$$= \sum_\theta \mathbf{P}(\widehat{\Theta} \neq \Theta \mid \Theta = \theta)\, p_\Theta(\theta)$$

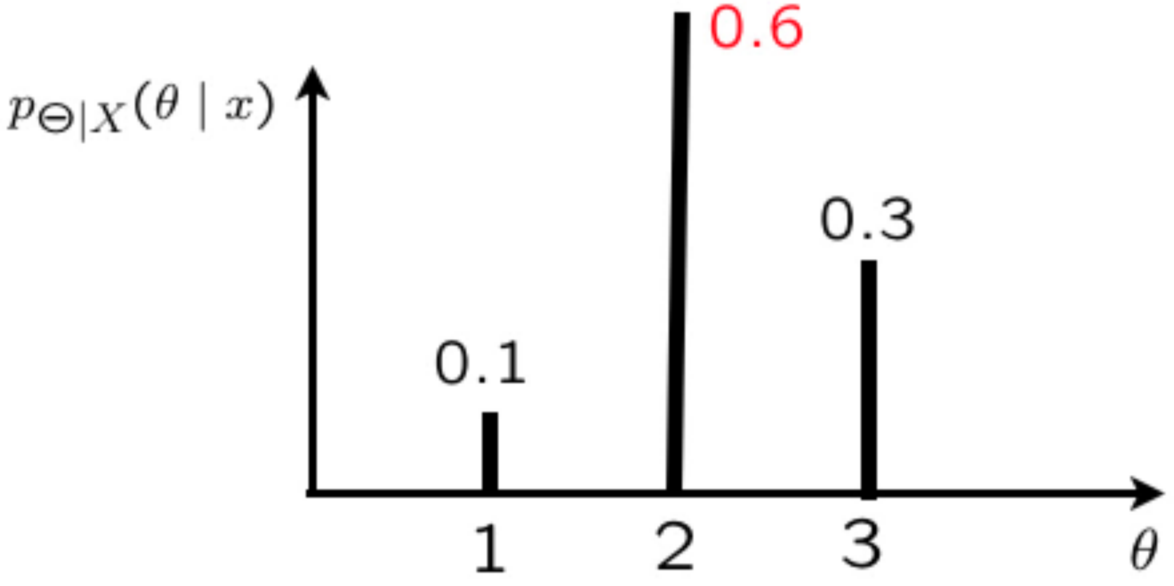
Discrete Θ, continuous X

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

    $X = \Theta + W$

    $W \sim N(0,\sigma^2)$, indep. of Θ

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



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

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

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

- conditional prob of error:

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

  smallest under the MAP rule

- overall probability of error:

$$\mathbf{P}(\widehat{\Theta} \neq \Theta) = \int \mathbf{P}(\widehat{\Theta} \neq \Theta \mid X = x) f_X(x)\, dx$$

$$= \sum_\theta \mathbf{P}(\widehat{\Theta} \neq \theta \mid \Theta = \theta)\, p_\Theta(\theta)$$

Continuous Θ, continuous X

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

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

- 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: uniform$[0, \Theta]$

  Θ: uniform $[0, 1]$

- $\widehat{\Theta} = g(X)$ MAP LMS
- interested in:

$$\mathbf{E}\Big[(\widehat{\Theta} - \Theta)^2 \mid X = x\Big]$$

$$\mathbf{E}\Big[(\widehat{\Theta} - \Theta)^2\Big]$$

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 $[0,1]$

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

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

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

$\theta \in [0,1]$

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

"Beta distribution, with parameters $(k+1, n-k+1)$"

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

$\alpha, \beta \geq 0$

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

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 $[0,1]$

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

- MAP estimate:

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

$$\max_\theta \left[k \log \theta + (n-k) \log(1-\theta)\right]$$

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

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

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

$\alpha \geq 0$

$\beta \geq 0$

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

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

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

$$\boxed{= \frac{k+1}{n+2}} \approx \frac{k}{n}$$

(n large)

Summary

- Problem data: $p_\Theta(\cdot)$, $p_{X|\Theta}(\cdot \mid \cdot)$
- Given the value x of X: find, e.g., $p_{\Theta|X}(\cdot \mid x)$
  - using appropriate version of the Bayes rule (4 choices)
- Estimator $\widehat{\Theta} = g(X)$ Estimate $\hat{\theta} = g(x)$
  - MAP: $\hat{\theta}_{\text{MAP}} = g_{\text{MAP}}(x)$ maximizes $p_{\Theta|X}(\theta \mid x)$
  - LMS: $\hat{\theta}_{\text{LMS}} = g_{\text{LMS}}(x) = \mathbf{E}[\Theta \mid X = x]$
- Performance evaluation of an estimator $\widehat{\Theta}$
  - $\mathbf{P}(\widehat{\Theta} \neq \Theta \mid X = x)$
  - $\mathbf{E}\big[(\widehat{\Theta} - \Theta)^2 \mid X = x\big]$
  - $\mathbf{P}(\widehat{\Theta} \neq \Theta)$
  - $\mathbf{E}\big[(\widehat{\Theta} - \Theta)^2\big]$ total $\left.\begin{matrix}\text{prob}\\ \text{exp}\end{matrix}\right\}$ thm.