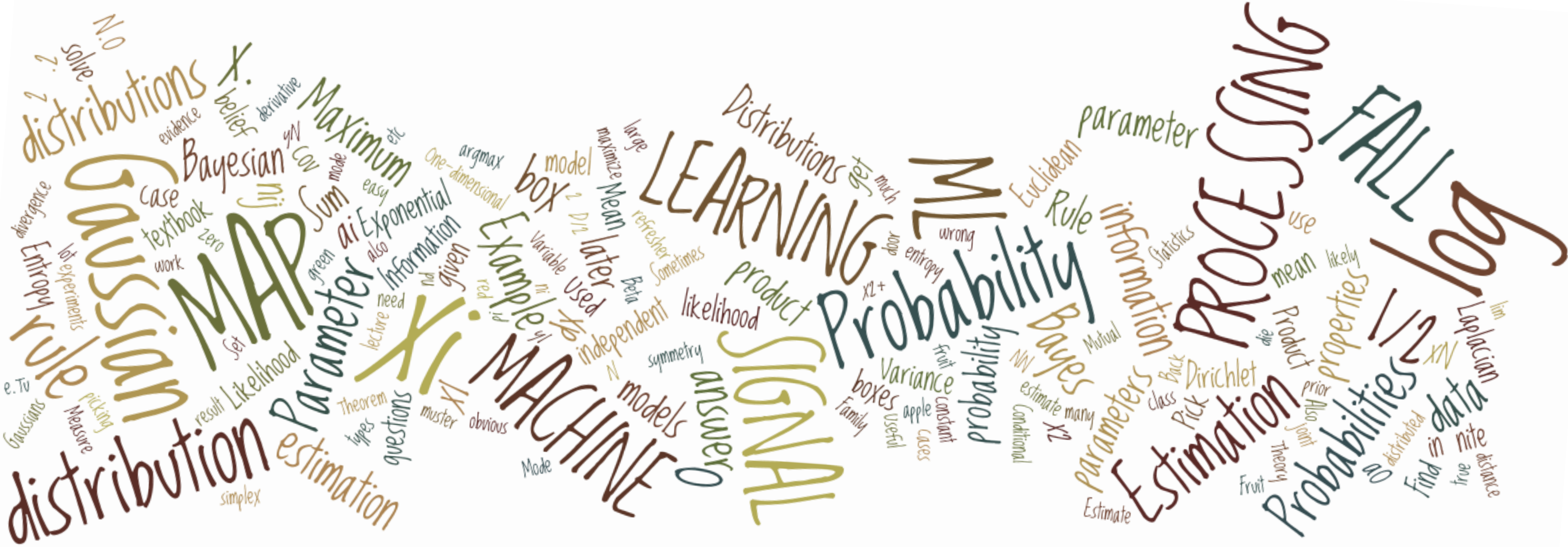
UNIVERSITY OF ILLINOIS @ URBANA-CHAMPAIGN



CS 598 Machine Learning for Signal Processing

Probability, Statistics & Parameter Estimation

27 August 2020

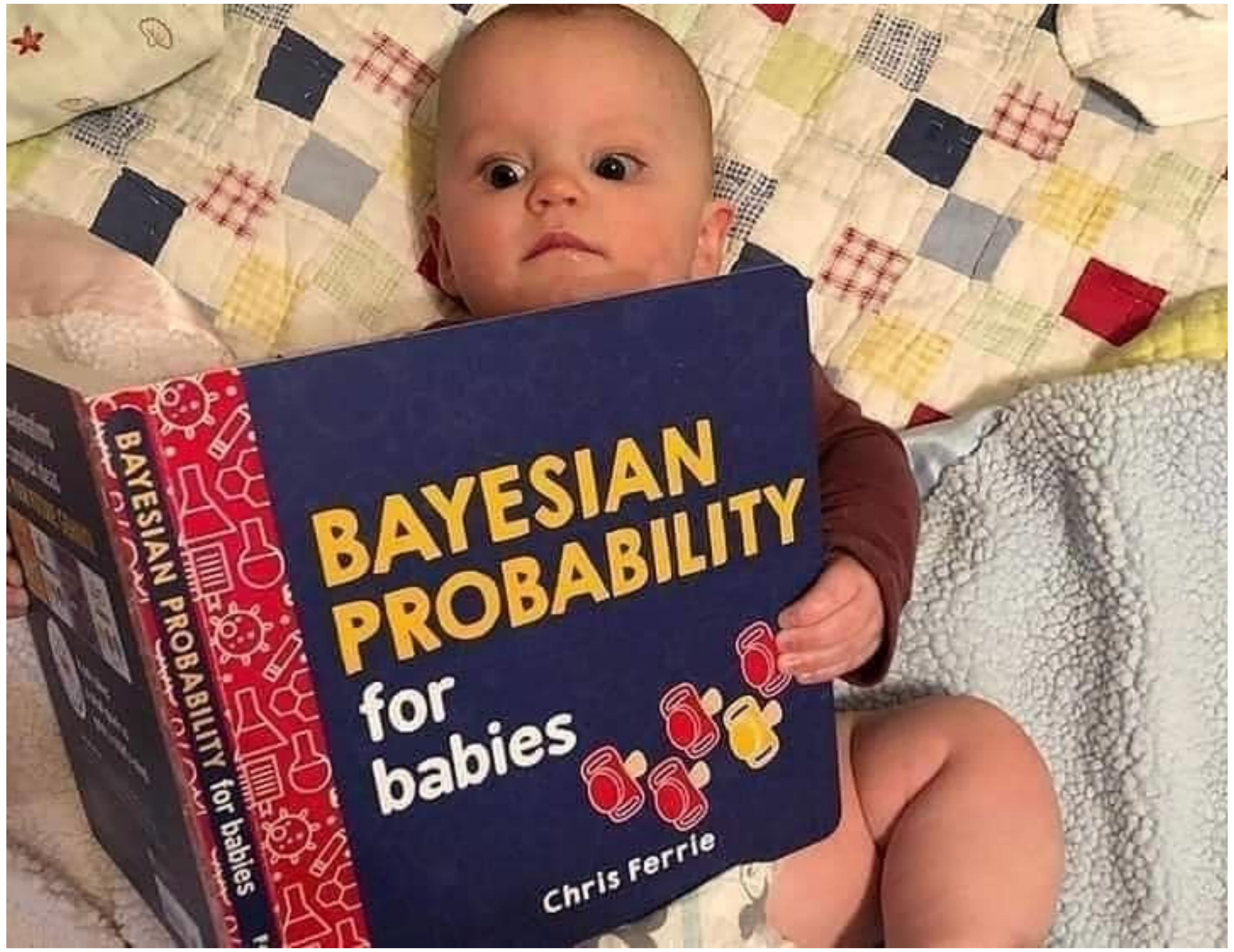
Logistics

- If you are new, add yourself to our piazza and slack
  - All class communications will use these interface
- Is there a waiting list to register for the class?
  - No. We don’t have a space crunch, so we can fit everyone in
  - CS-majors can/should register now, non-CS starts tomorrow
- We have TA office hours!
  - W/F 11:00AM CT, see zoom links in class calendar

Today's refresher

** Once again, this is all for future reference, don't expect to learn it all today*

- Probability
- Statistics
- Parameter Estimation


BAYESIAN PROBABILITY for babies
Chris Ferrie
BAYESIAN PROBABILITY for babies

Probability

- Probity
  - Measure of legal authority/nobility
    - Passed muster in the middle ages

- Probability
  - Measure of belief/likelihood
    - Passes muster today

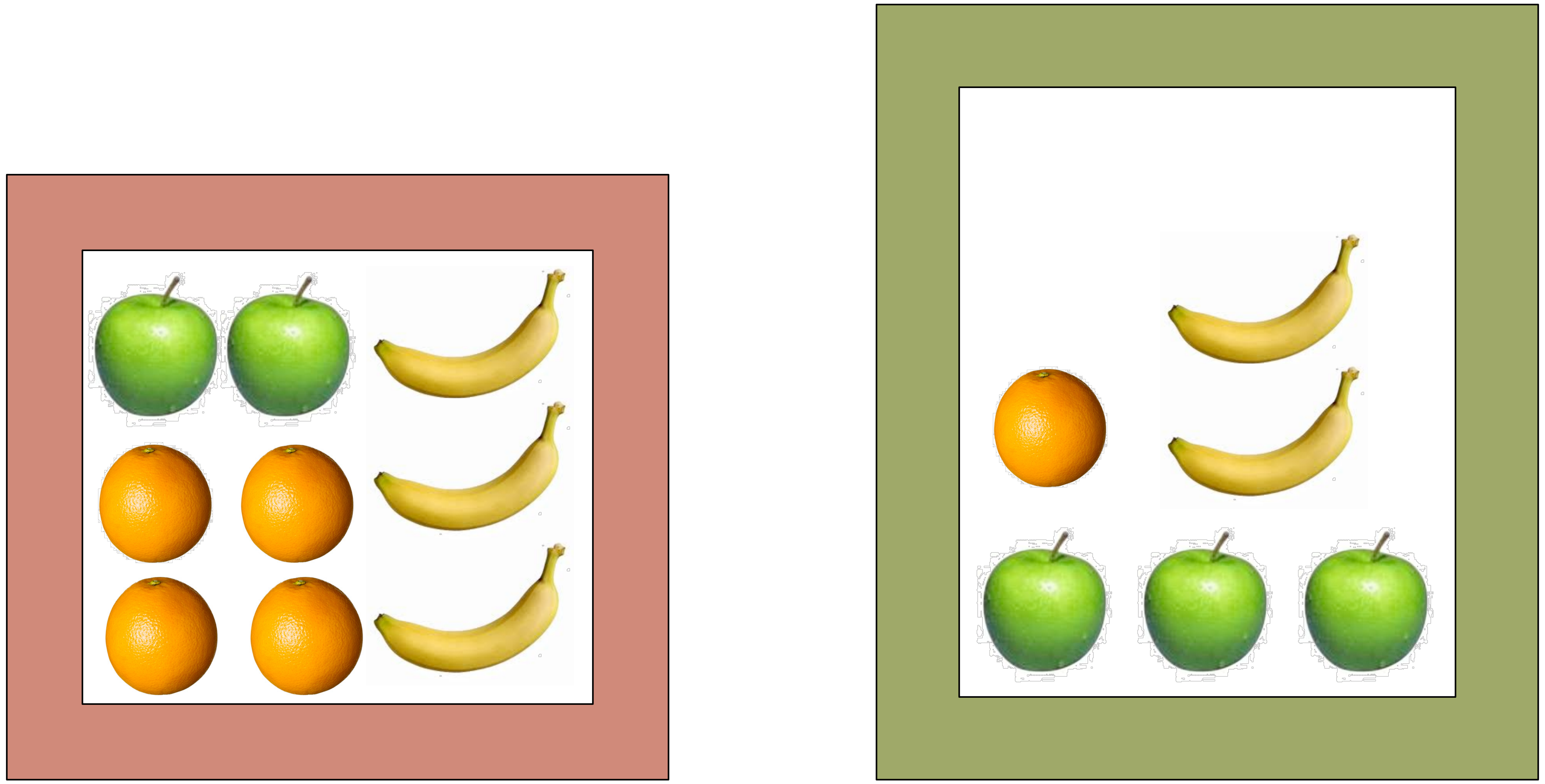
Goals of probability

- Characterize stochastic processes
  - How do dice roll?
  - What am I more likely to say next?

- Indicate belief given evidence
  - The suspect was nearby and there are feathers on his clothes. Was he the chicken thief?

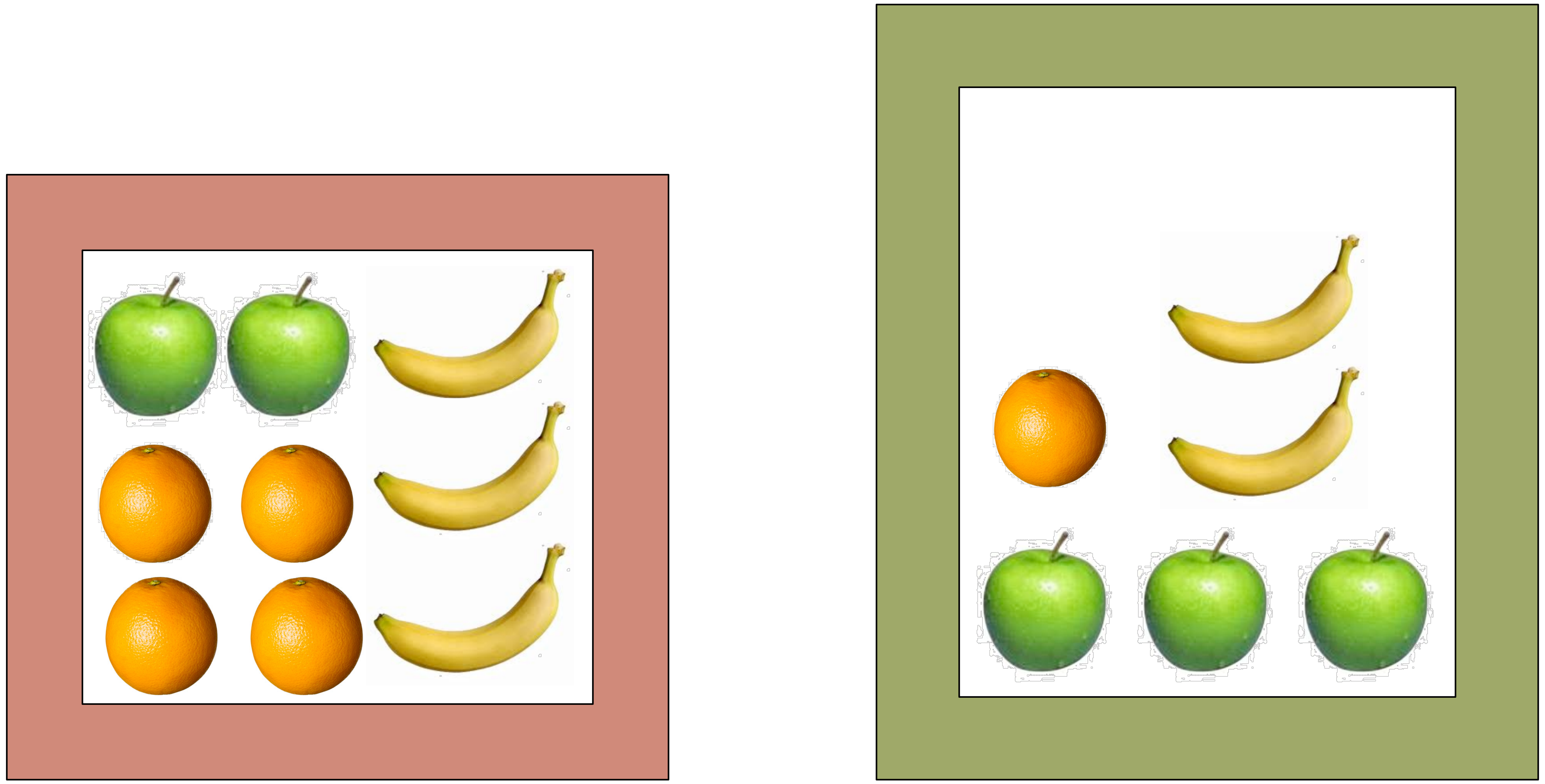
An example

- We start picking oranges, apples and bananas, from the two boxes below
  - 40% of the time pick from red box, 60% from green box



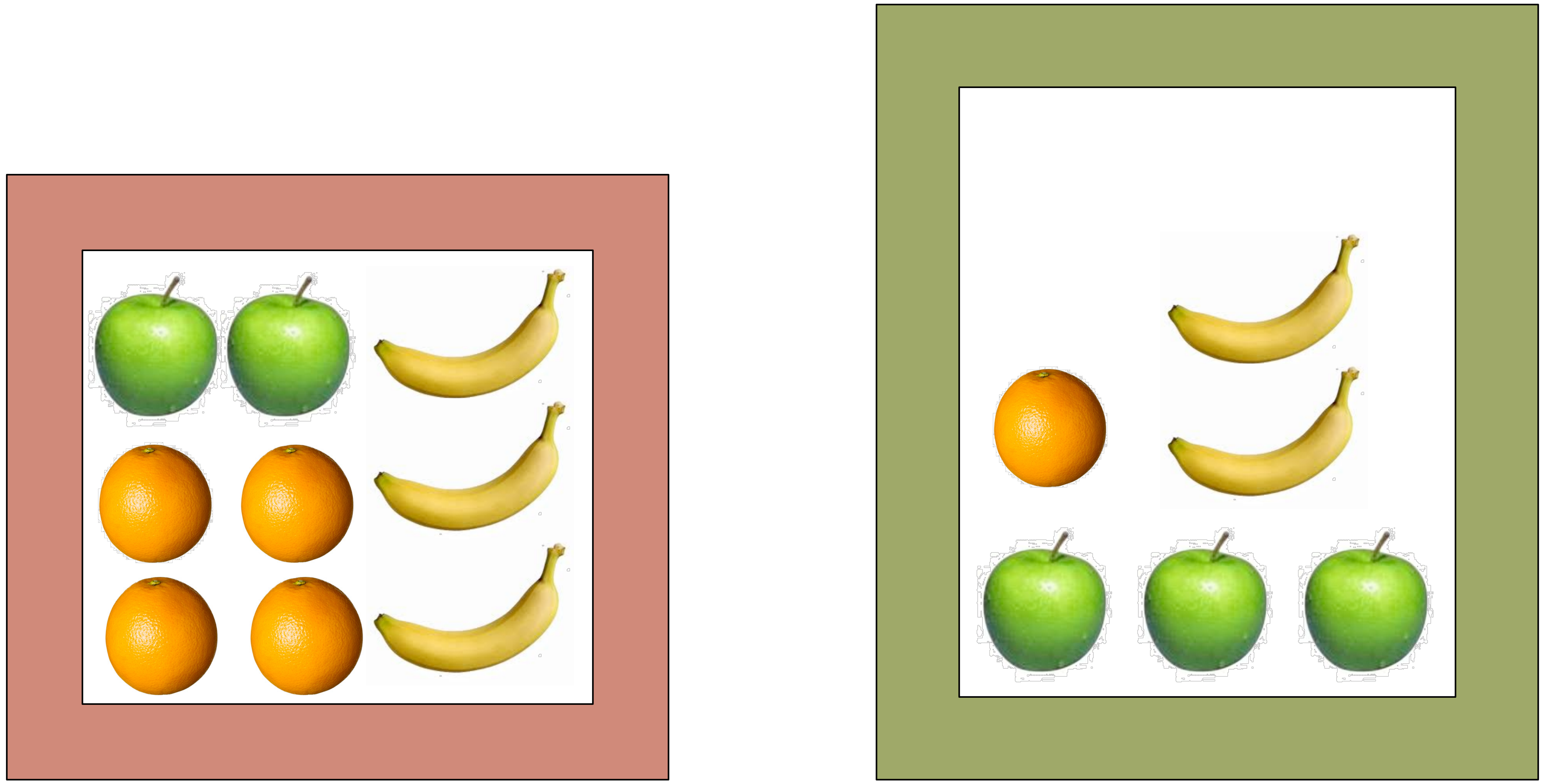
The random variables

- The box: $B = \{r, g\}$
- The fruit: $F = \{a, o, b\}$
  - What are the relevant probabilities?



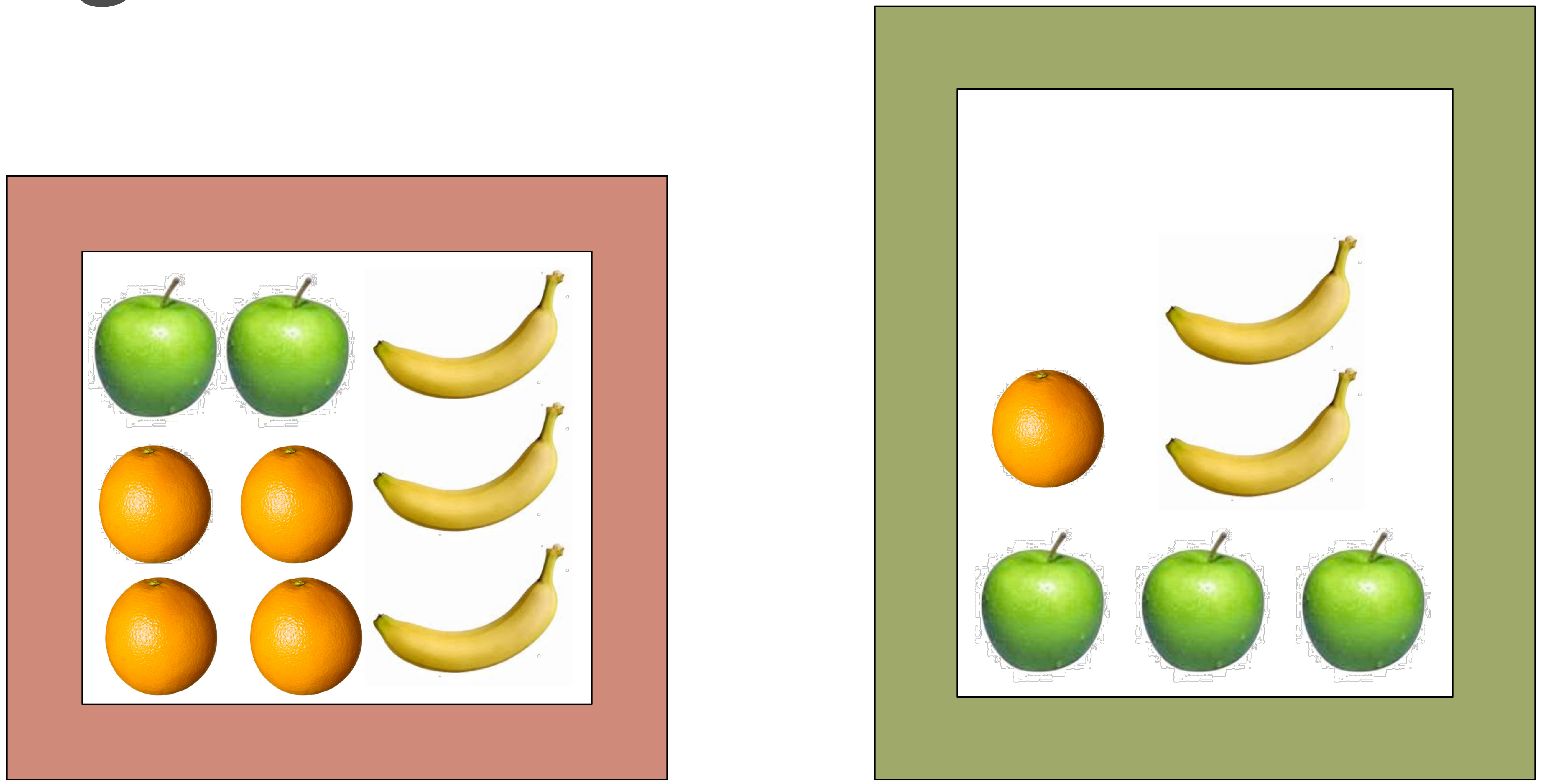
Box probabilities

- Obviously:
  - $P(B == g) = 60/100$
  - $P(B == r) = 40/100$
  - $P(\cdot) \in [0,1]$



Asking questions

- What is the probability of picking an apple?
- If we pick an orange, what is the probability that it came out of the green box?



Keeping track

- Do an experiment and accumulate counts in a table
  - The more experiments the better
  - e.g. pick a banana from red box
    - $n_{rb} = n_{rb} + 1$
    - $n_r = n_r + 1$
    - $n_b = n_b + 1$
    - $N = N + 1$

F

		Apple	Banana	Orange	Any fruit
	Green Box	n_{ga}	n_{gb}	n_{go}	n_g
B	Red Box	n_{ra}	n_{rb}	n_{ro}	n_r
	Any box	n_a	n_b	n_o	N

Single variable probabilities

$$P(B == i) = n_i \,/\, N$$

$$P(F == j) = n_j \,/\, N$$

		F			
		Apple	Banana	Orange	Any fruit
B	Green Box	n_{ga}	n_{gb}	n_{go}	n_g
	Red Box	n_{ra}	n_{rb}	n_{ro}	n_r
	Any box	n_a	n_b	n_o	N

Joint probabilities

$$P(B == i, F == j) = \frac{n_{ij}}{N}$$

$$P(B == i, F == j) = P(F == j, B == i)$$

F

		Apple	Banana	Orange	Any fruit
	Green Box	n_{ga}	n_{gb}	n_{go}	n_g
B	Red Box	n_{ra}	n_{rb}	n_{ro}	n_r
	Any box	n_a	n_b	n_o	N

The sum rule

$$n_i / N = \left(n_{ia} + n_{ib} + n_{io}\right) / N$$

$$P(B == i) = \sum_{\forall j} P(B == i, F == j)$$

			F		
		Apple	Banana	Orange	Any fruit
	Green Box	n_{ga}	n_{gb}	n_{go}	n_g
B	Red Box	n_{ra}	n_{rb}	n_{ro}	n_r
	Any box	n_a	n_b	n_o	N

Conditional probability

$$P(F == j \,|\, B == i) = \frac{n_{ij}}{n_i}$$

		F			
		Apple	Banana	Orange	Any fruit
	Green Box	n_{ga}	n_{gb}	n_{go}	n_g
B	Red Box	n_{ra}	n_{rb}	n_{ro}	n_r
	Any box	n_a	n_b	n_o	N

The product rule

$$P(B == i, F == j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{n_i}\frac{n_i}{N} = P(F == j \mid B == i)P(B == i)$$

		F			
		Apple	Banana	Orange	Any fruit
	Green Box	n_{ga}	n_{gb}	n_{go}	n_g
B	Red Box	n_{ra}	n_{rb}	n_{ro}	n_r
	Any box	n_a	n_b	n_o	N

The two basic rules

- Sum Rule:

$$P(X) = \sum_{Y} P(X,Y)$$

- Product Rule:

$$P(X,Y) = P(Y \mid X)P(X)$$

Bayes theorem

- From product rule & symmetry of joint

$$\overset{\textit{Posterior}}{P(Y \mid X)} = \frac{\overset{\textit{Likelihood}}{P(X \mid Y)}\overset{\textit{Prior}}{P(Y)}}{\underset{\textit{Normalizing constant}}{P(X)}}$$

- Will answer most of your questions!

Independence

- If:

$$P(B == i, F == j) = P(B == i)P(F == j)$$

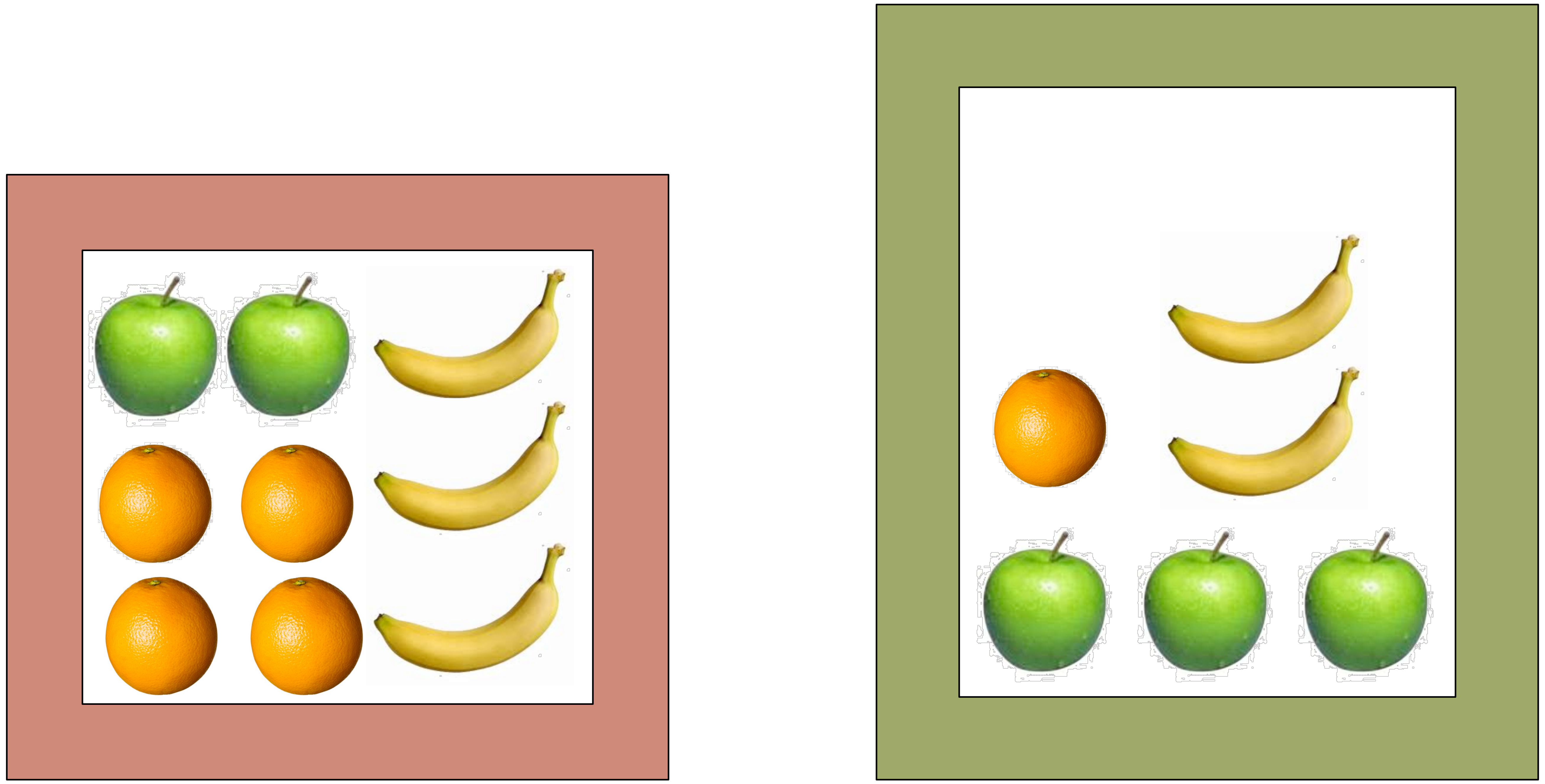
- Then B and F are *independent*
- Also means, via the product rule, that:

$$P(F \mid B) = P(F)$$

- If both boxes had the same fraction of fruits, then we would have independence

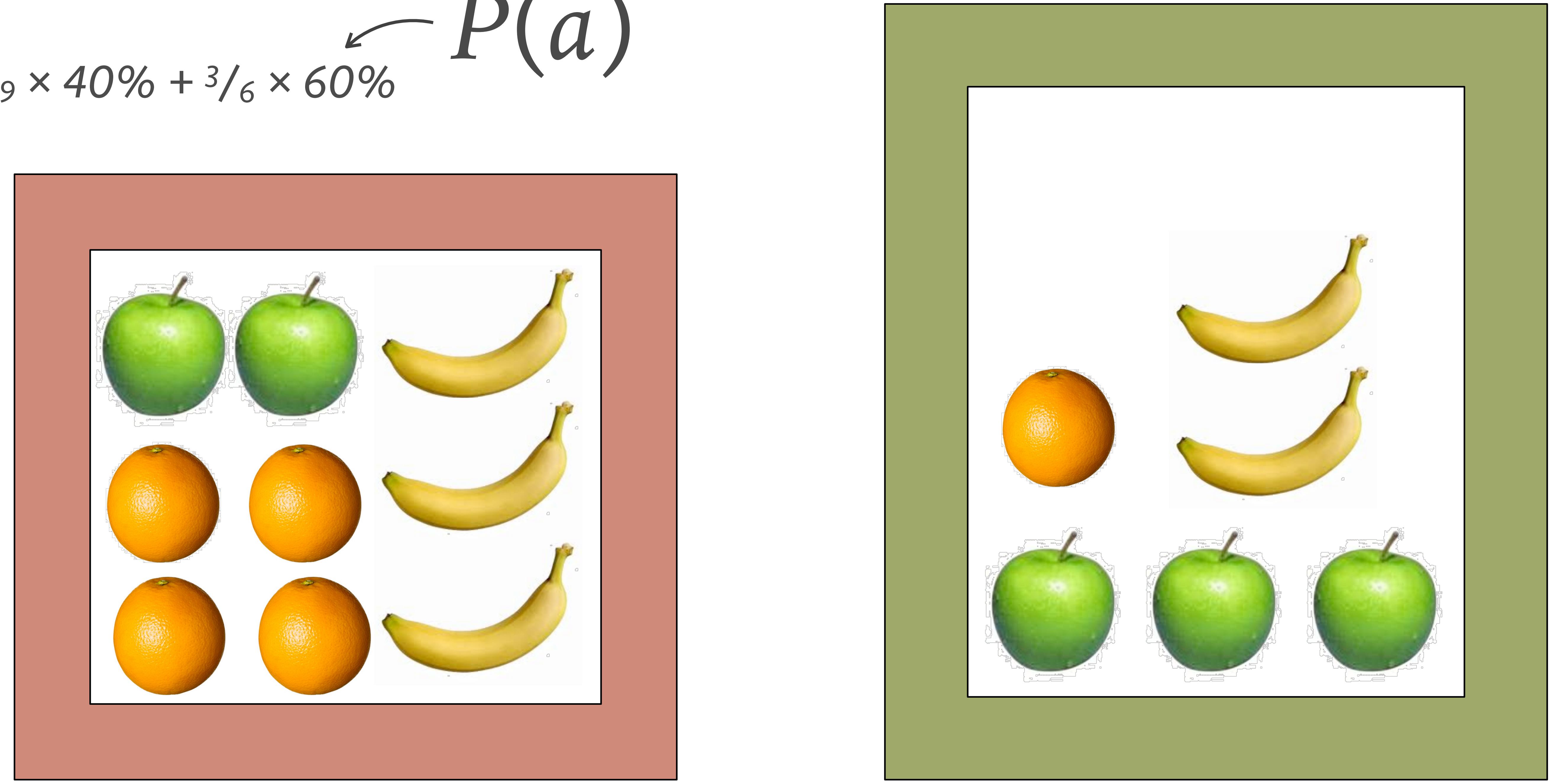
Back to the fruit

- What’s the probability of picking an apple?
  - Sum rule: $P(a) = \underset{\hookrightarrow\, {}^{2}/_{9} \times 40\%}{P(a,r)} + \underset{\hookrightarrow\, {}^{3}/_{6} \times 60\%}{P(a,g)}$



Back to the fruit

- What’s the probability that I picked a fruit from the red box given that I picked an apple?
  - Bayes rule: $P(r\,|\,a) = \dfrac{\overset{2/9}{P(a\,|\,r)}\,\overset{40\%}{P(r)}}{\underset{2/9 \times 40\% + 3/6 \times 60\%}{P(a)}}$



Schools of thought

- Frequentists
  - Probabilities are interpretations of frequencies of occurrence in experiments
    - There can only be one solution!
- Bayesians
  - Probabilities are a degree of belief, not a result of a counting experiment
    - What’s the distribution of the parameter? The priors?

Why belief?

- “Will a meteor hit earth?”
  - Frequentist: Let us wait until N is large …
- Using a Bayesian treatment we can find a likelihood given the evidence, not the data
  - But that requires models, priors, assumptions, … more later

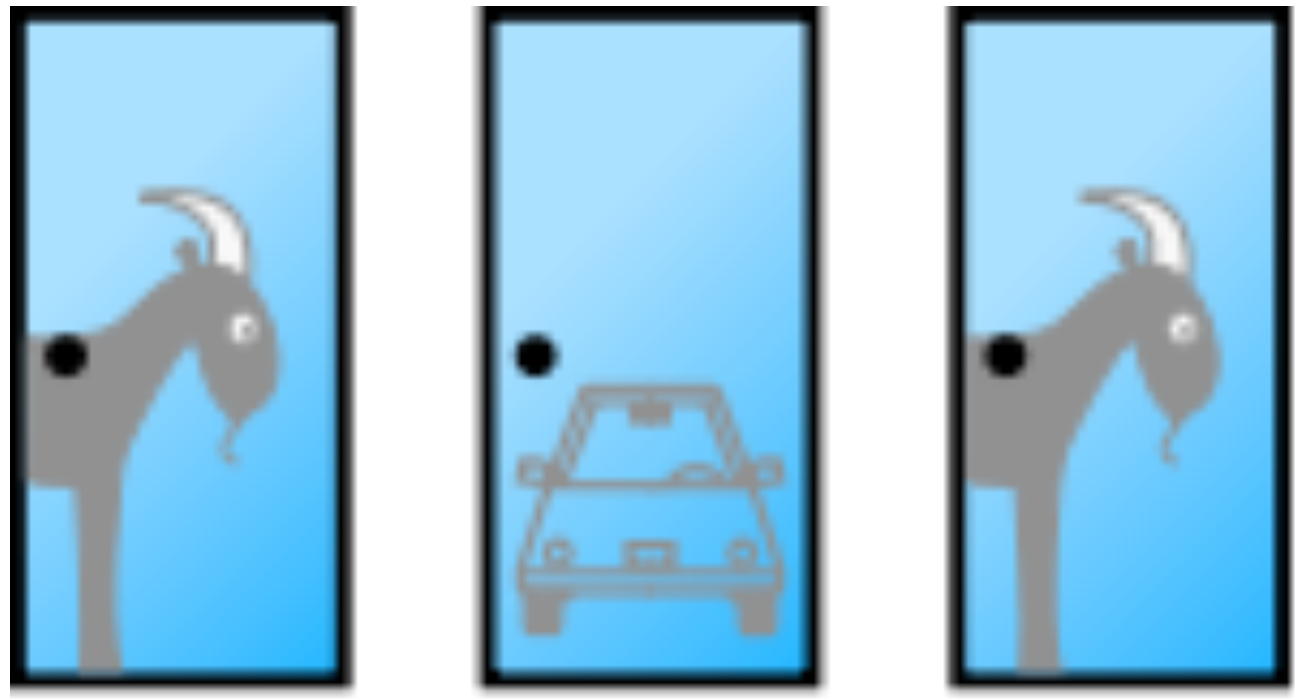
A practical application

Statistics in the Real World:
The Search for the USS Scorpion

http://www.youtube.com/watch?v=U9-G-noZrwc

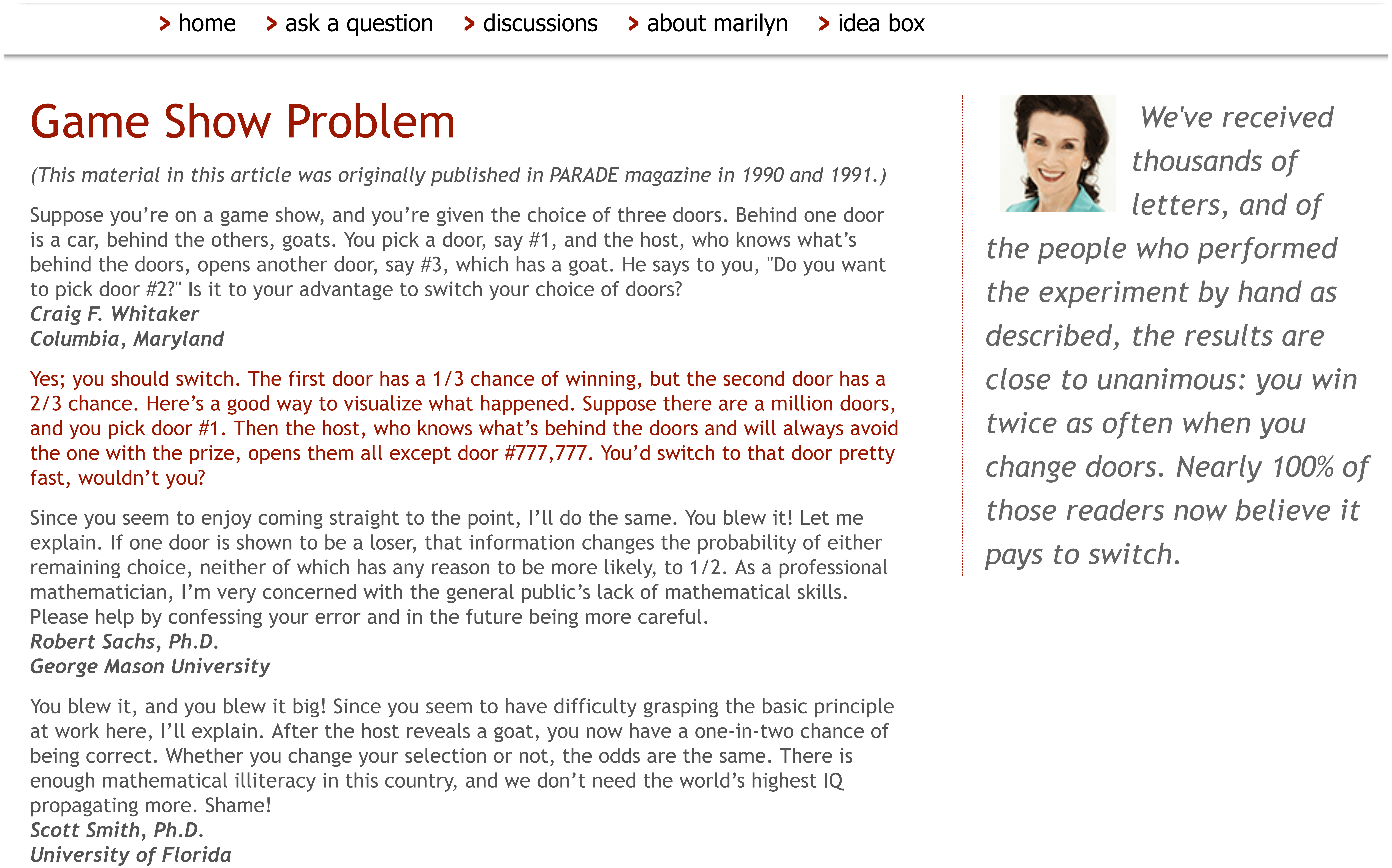
Getting lost? Don’t worry

- Probability is super tricky
  - Even seasoned professionals get it wrong!
    - "*In no other branch of mathematics is it so easy for experts to blunder as in probability theory*" - Martin Gardner
  - Case in point: *The Monty Hall problem*



PhDs being mean

http://marilynvossavant.com/game-show-problem/



Game Show Problem

(This material in this article was originally published in PARADE magazine in 1990 and 1991.)

Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say #1, and the host, who knows what's behind the doors, opens another door, say #3, which has a goat. He says to you, "Do you want to pick door #2?" Is it to your advantage to switch your choice of doors?
Craig F. Whitaker
Columbia, Maryland

Yes; you should switch. The first door has a 1/3 chance of winning, but the second door has a 2/3 chance. Here's a good way to visualize what happened. Suppose there are a million doors, and you pick door #1. Then the host, who knows what's behind the doors and will always avoid the one with the prize, opens them all except door #777,777. You'd switch to that door pretty fast, wouldn't you?

Since you seem to enjoy coming straight to the point, I'll do the same. You blew it! Let me explain. If one door is shown to be a loser, that information changes the probability of either remaining choice, neither of which has any reason to be more likely, to 1/2. As a professional mathematician, I'm very concerned with the general public's lack of mathematical skills. Please help by confessing your error and in the future being more careful.
Robert Sachs, Ph.D.
George Mason University

You blew it, and you blew it big! Since you seem to have difficulty grasping the basic principle at work here, I'll explain. After the host reveals a goat, you now have a one-in-two chance of being correct. Whether you change your selection or not, the odds are the same. There is enough mathematical illiteracy in this country, and we don't need the world's highest IQ propagating more. Shame!
Scott Smith, Ph.D.
University of Florida

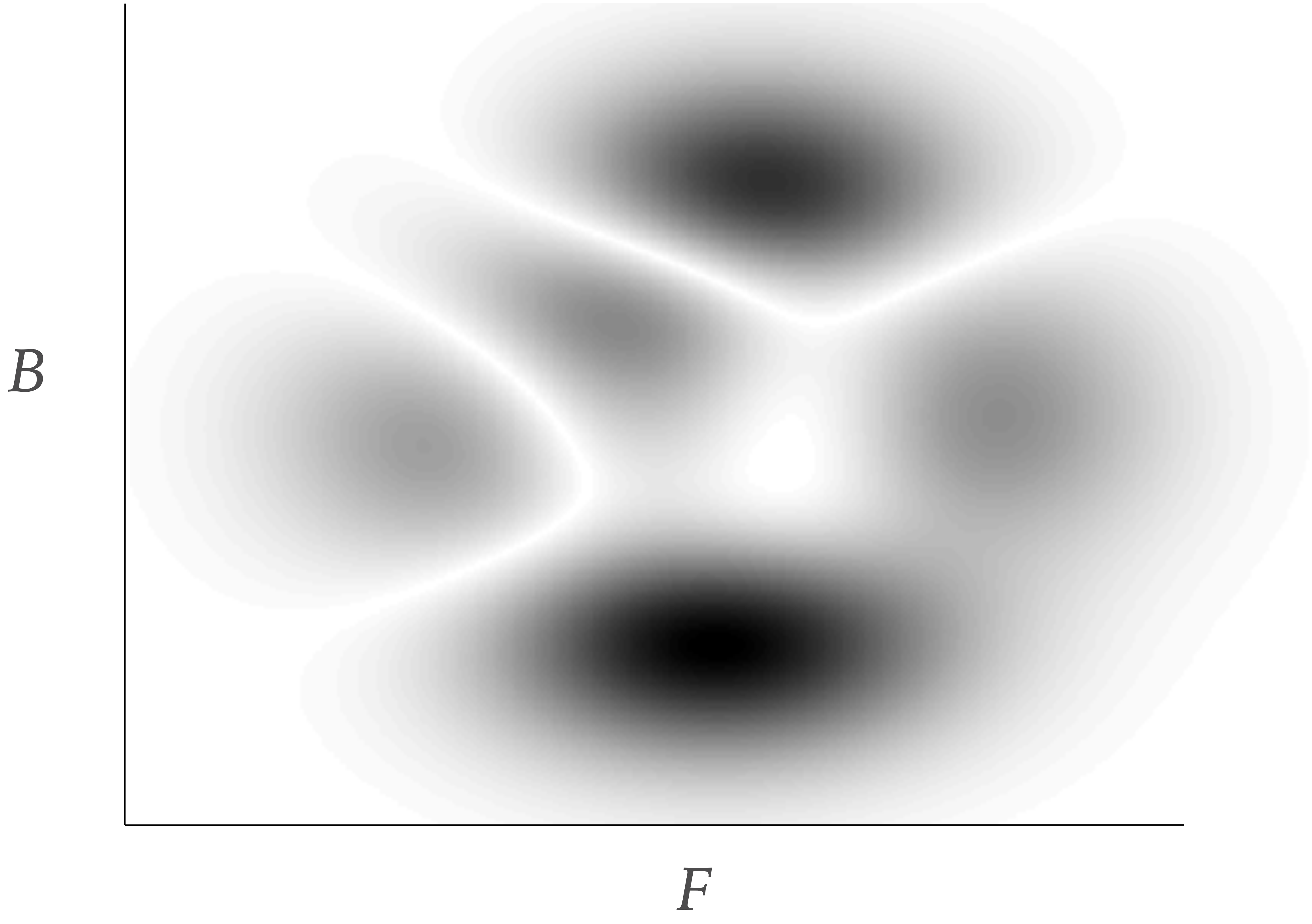
We've received thousands of letters, and of the people who performed the experiment by hand as described, the results are close to unanimous: you win twice as often when you change doors. Nearly 100% of those readers now believe it pays to switch.

Quick answer

		Door 1	Door 2	Door 3	Outcome
Pick door 1 and switch	1st case	Car	Goat	Goat	Switch & lose
	2nd case	Goat	Car	Goat	Switch & win
	3rd case	Goat	Goat	Car	Switch & win
Pick door 1 and stay	4th case	Car	Goat	Goat	Stay & win
	5th case	Goat	Car	Goat	Stay & lose
	6th case	Goat	Goat	Car	Stay & lose

Continuous distributions

- What if we have infinite colors of boxes, and infinite types of fruit?



Same(ish) rules (harder proofs)

- Sum rule: $P(x) = \int P(x,y)dy$

- Product rule: $P(x,y) = P(y\,|\,x)P(x)$

- Bayes rule: $P(x\,|\,y) = \dfrac{P(y\,|\,x)P(x)}{P(y)}$

Some properties

- Integration to unity

$$\int_{-\infty}^{\infty} P(x) = 1$$

  - You'll be amazed how many papers get this wrong!

- Probabilities are real and between 0 and 1

$$P(x) \in \mathbb{R} \qquad 0 \geq P(x) \leq 1$$

  - Well, they don't have to be. More on that later …

Some common operations

- Expectation: $\mathrm{E}\left(f(x)\right) = \int P(x) f(x) dx$

- Variance: $\mathrm{var}\left(f(x)\right) = \mathrm{E}\left[\left(f(x) - \mathrm{E}\left[f(x)\right]\right)^2\right] = \mathrm{E}\left[f(x)^2\right] - \mathrm{E}\left[f(x)\right]^2$

- Covariance: $\mathrm{cov}[x, y] = \mathrm{E}_{x,y}(xy) - \mathrm{E}(x)\mathrm{E}(y)$

Popular distributions

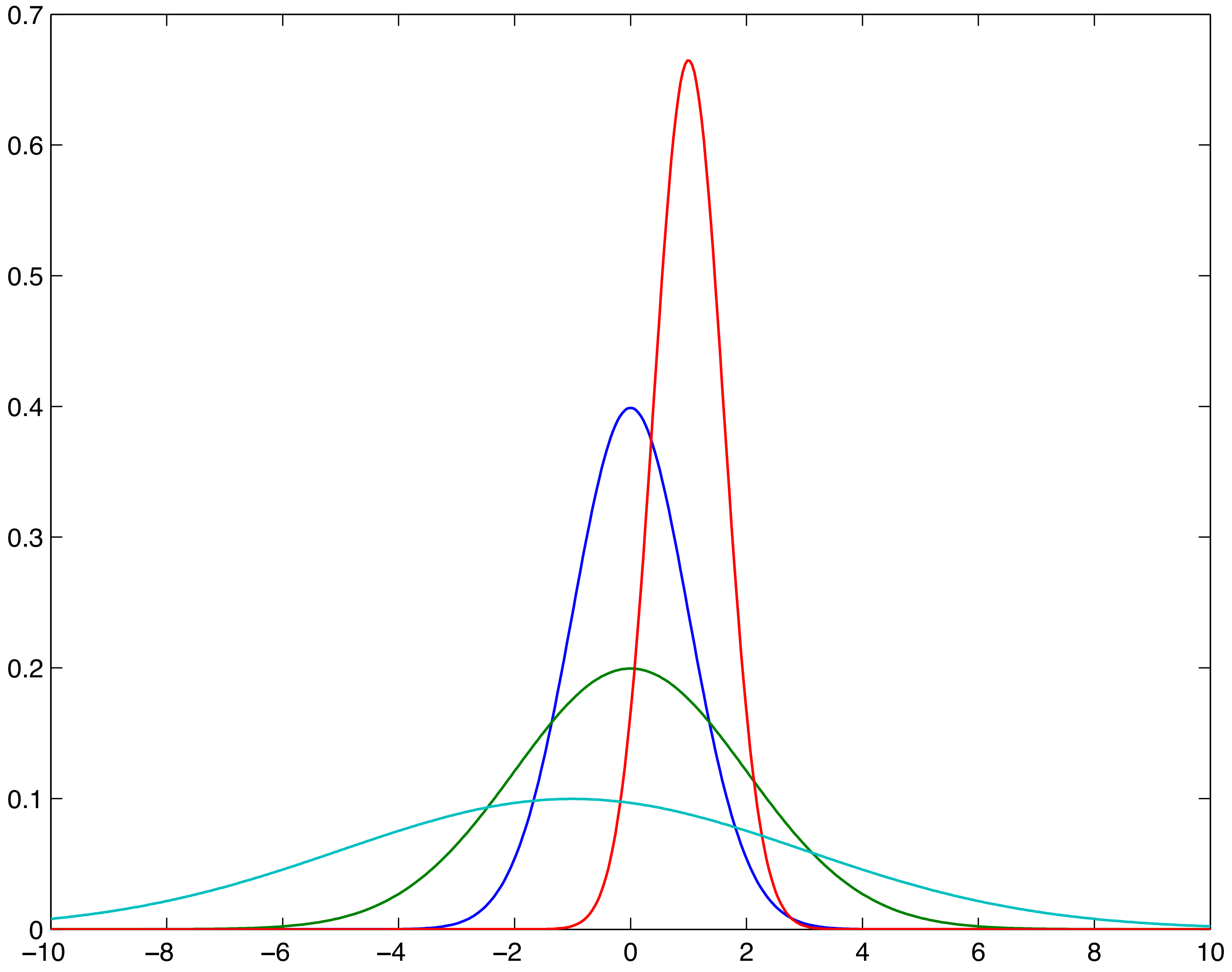
- We'll be seeing a lot of:
  - The Gaussian
    - Used pretty much everywhere
  - The Laplacian
    - Used for sparse models
  - The Dirichlet
    - Used for compositional models
  - The Exponential Family
    - Very useful properties!

The Gaussian

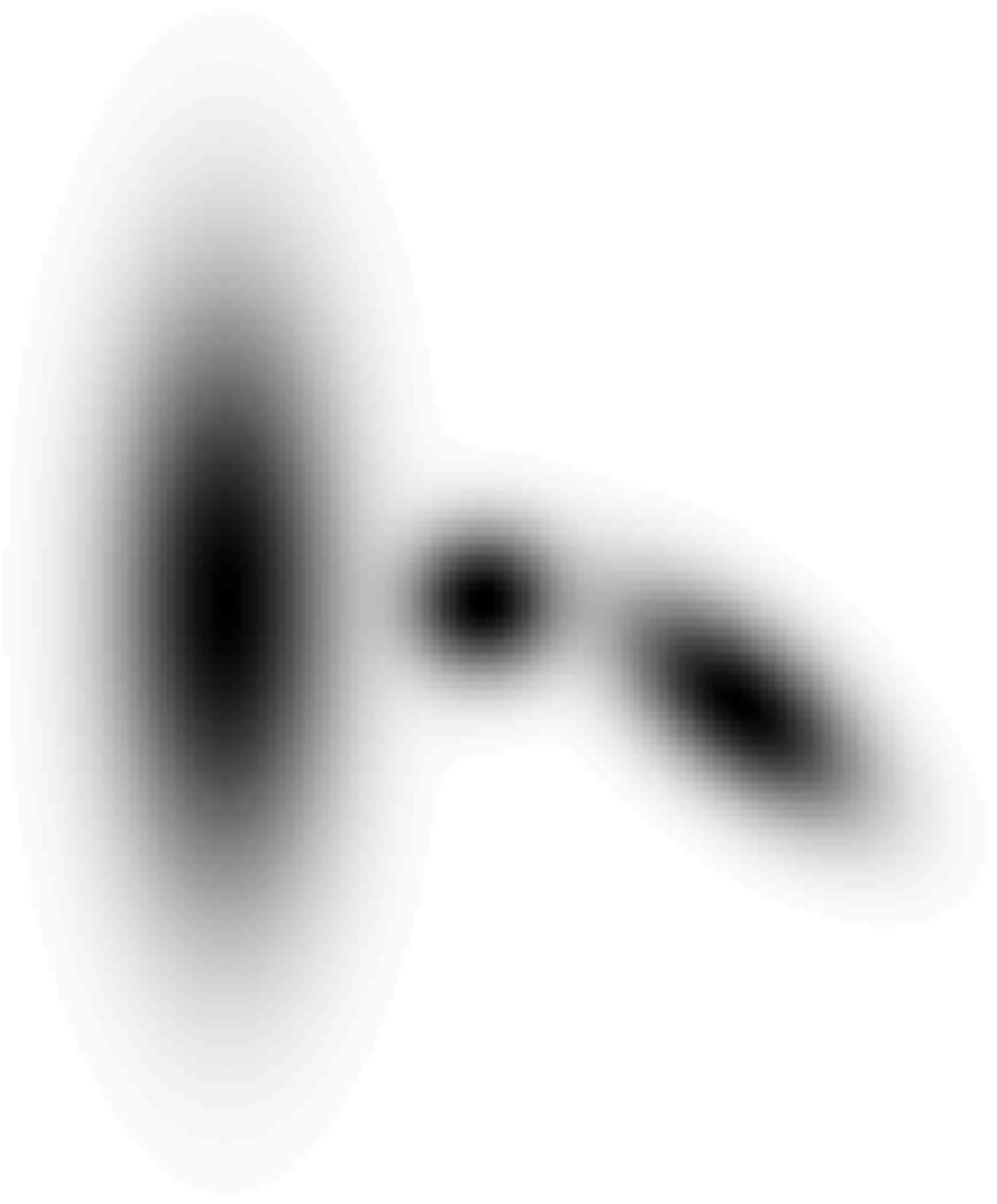
- Also known as the Normal distribution or the bell curve

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \qquad \mathbf{x} \in \mathbb{R}^D$$

One-dimensional Gaussians



Two-dimensional Gaussians



Why the Gaussian?

- Makes the Euclidean distance a distribution
  - e.g. in scalar case:

$$\mathcal{N}(x;\mu,\sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- If you assume squared Euclidean errors, then you are using a Gaussian

The Gaussian parameters

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \qquad \mathbf{x} \in \mathbb{R}^D$$

- The mean: $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$
- The covariance: $\mathrm{cov}(\mathbf{x}) = \boldsymbol{\Sigma}$
  - The mode: $\mathrm{mode}(\mathbf{x}) = \boldsymbol{\mu}$

Special case

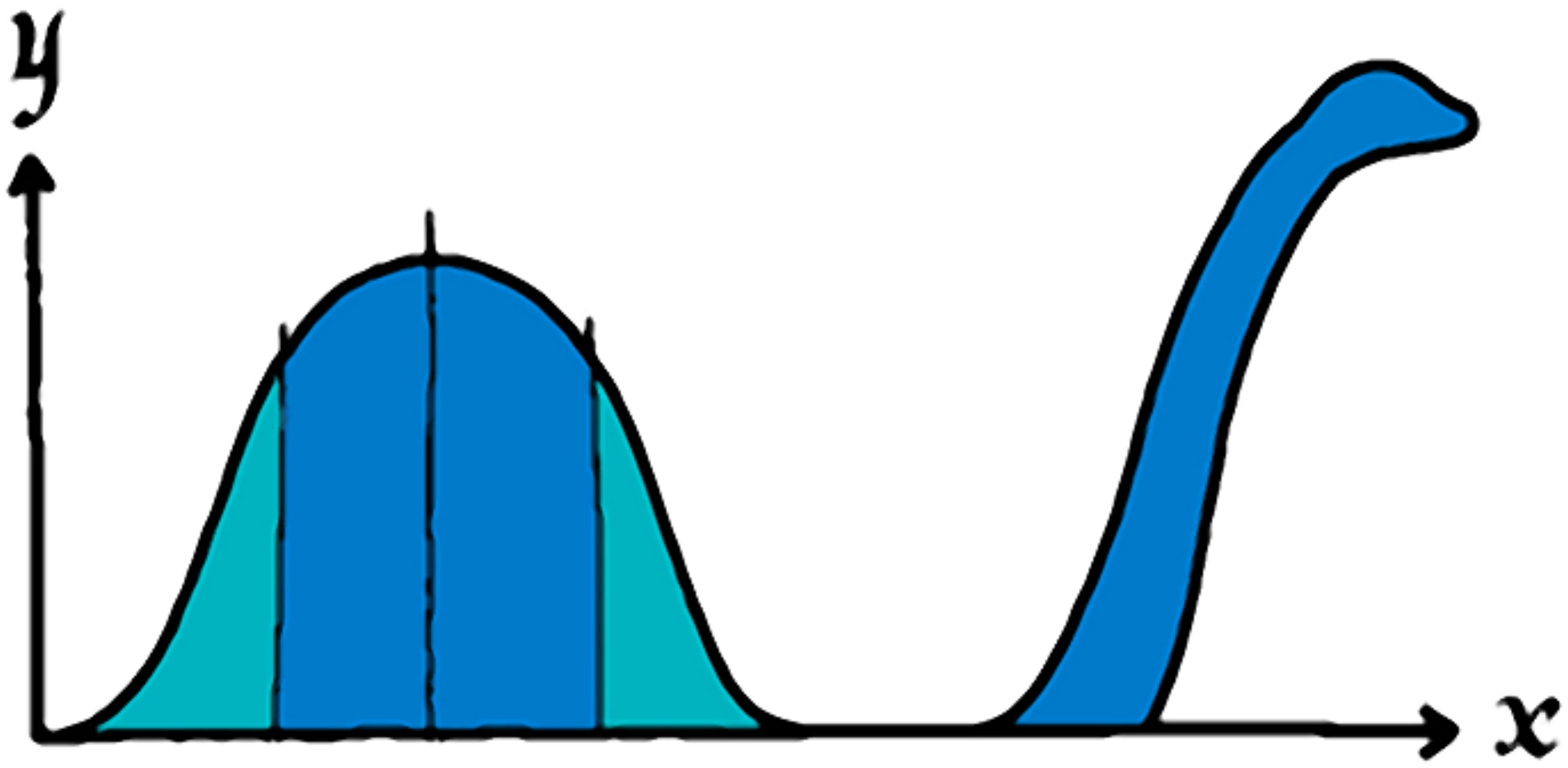


fig 1.0 The Extended Bell Curve.

– *by Tang Yau Hoong*

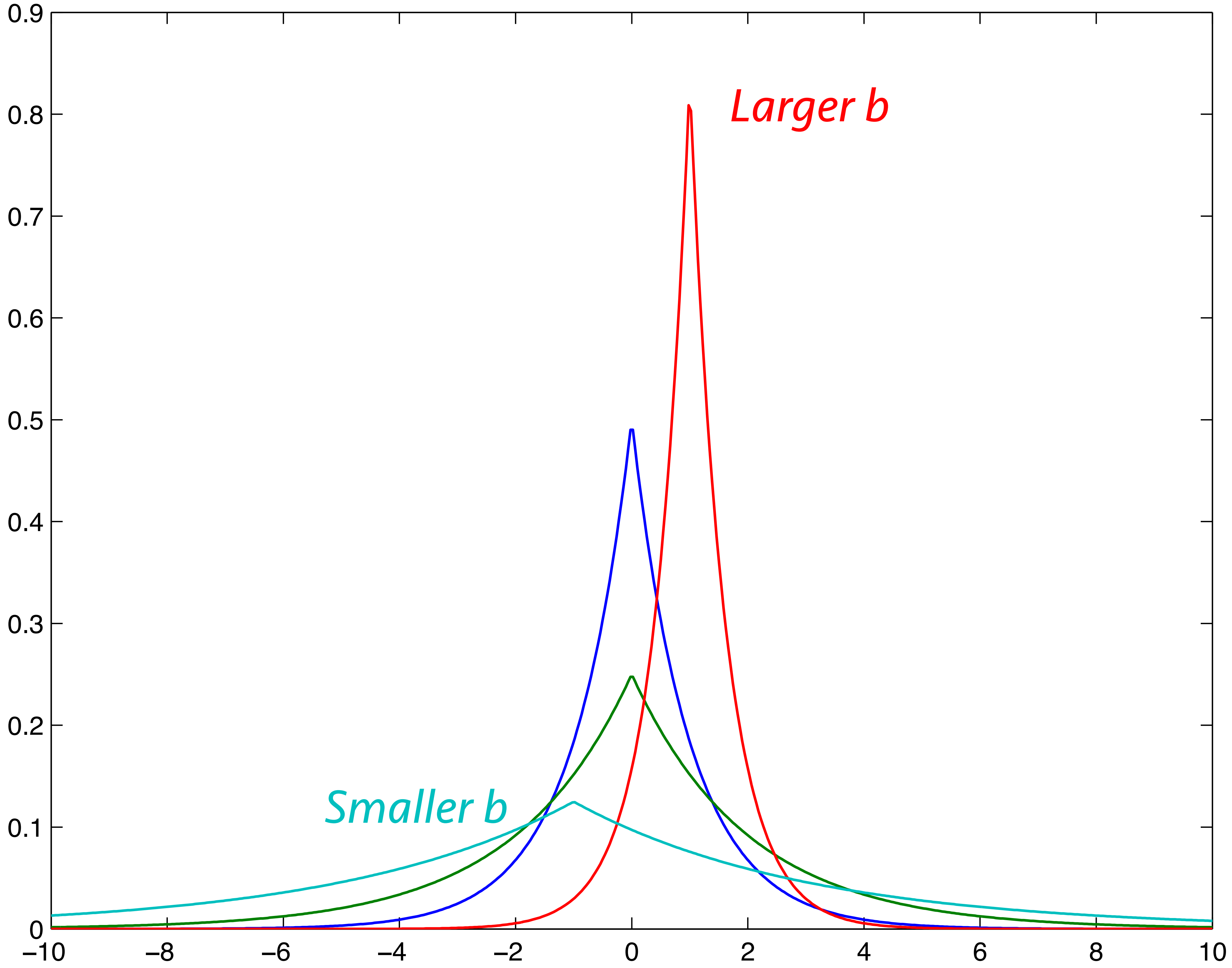
The Laplacian

- Sharper than the Gaussian
  - Uses absolute distance, instead of squared

$$P(x;\mu,b)=\frac{1}{2b}e^{-\frac{|x-\mu|}{b}}$$

- Mean: μ
- Variance: $2b^2$
  - Mode: μ

One-dimensional Laplacians



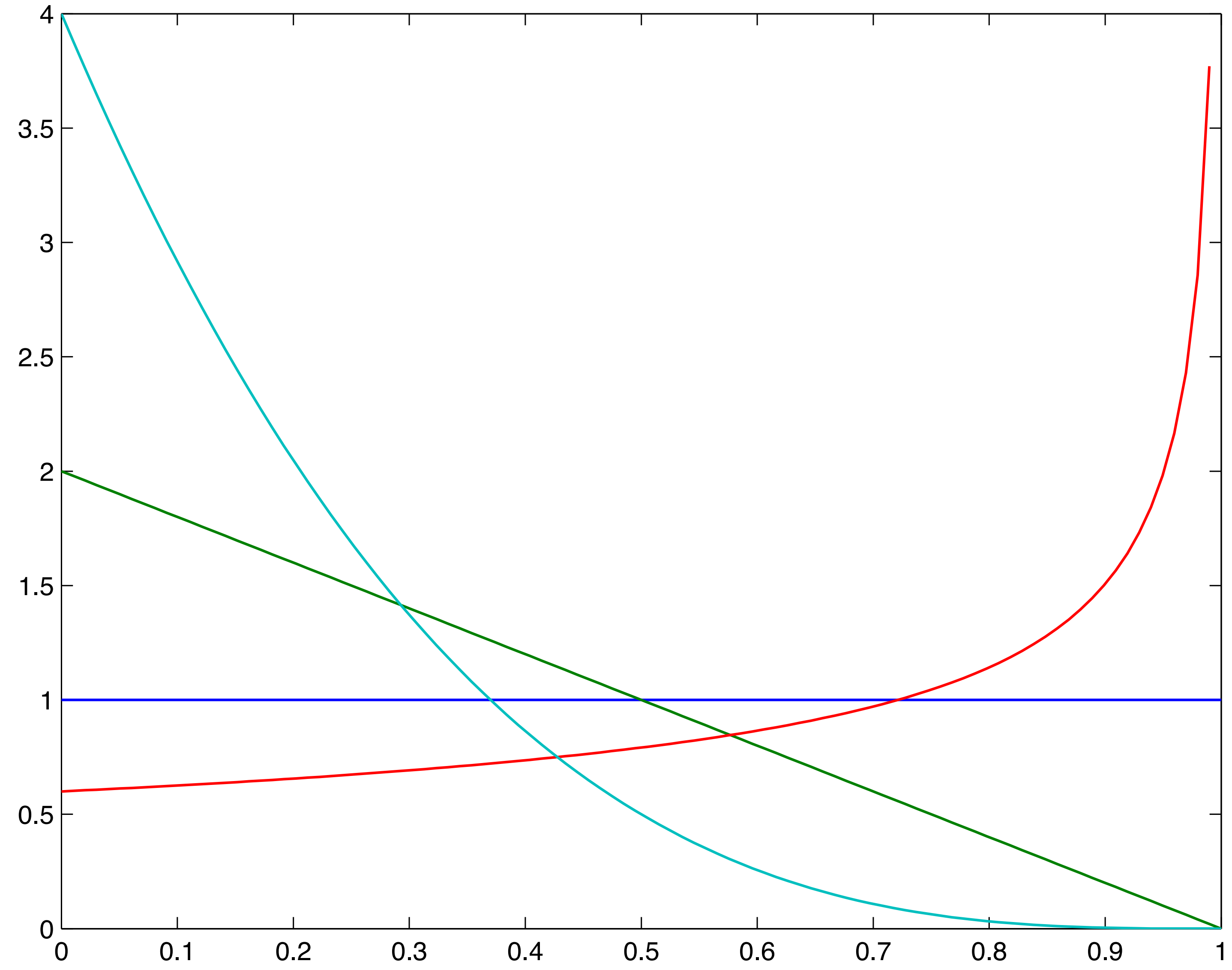
Beta/Dirichlet distributions

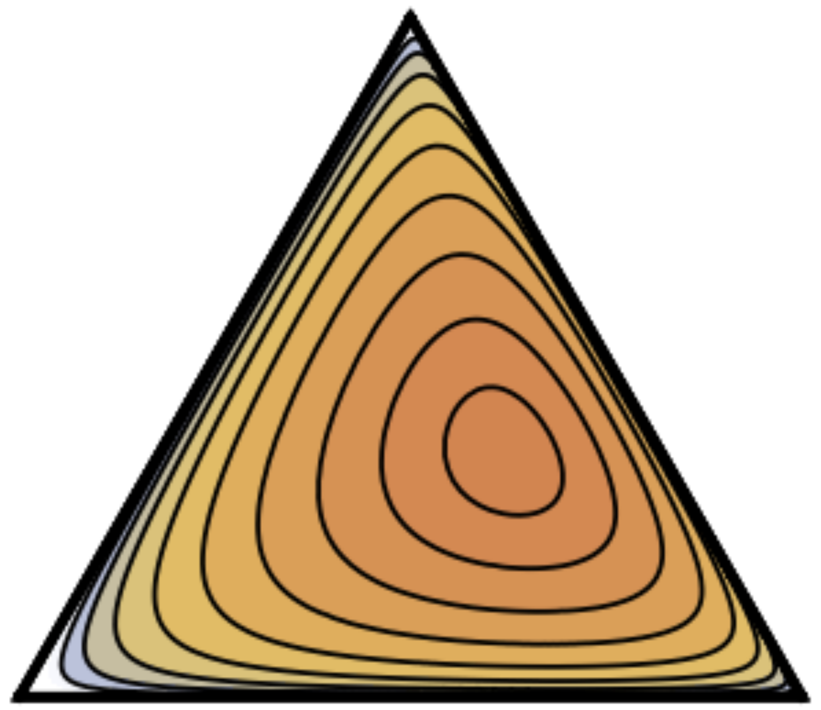
- Defined on a simplex
  - $x_1 + x_2 + x_3 + \ldots = 1$

$$P(\mathbf{x};\mathbf{a}) = \frac{\prod \Gamma(a_i)}{\Gamma\left(\sum a_i\right)} \prod x_i^{a_i - 1}$$

  - For 1D the Dirichlet is the Beta
- Mean: $\mathrm{E}[x_i] = a_i / a_0$
- Variance: $\operatorname{cov}[x_i, x_j] = \dfrac{-a_i a_j}{a_0{}^2 (a_0 + 1)}$
- Mode: $x_i = (a_i - 1) / (a_0 - K)$

Beta distributions on the {2-1}-simplex





Dirichlet distribution on a {3-1}-simplex

The exponential family

- Any distribution that can be written as:

$$P(\mathbf{x};\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})e^{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})}$$

- $\boldsymbol{\eta}$ contains the “natural” parameters
- $\mathbf{u}(\mathbf{x})$ is some function of $\mathbf{x}$
- $g(\boldsymbol{\eta})$ is just for normalization

Gaussian example

$$P(\mathbf{x};\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})e^{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})}$$

$$\mathbf{u}(\mathbf{x}) = \begin{bmatrix} x \\ x^2 \end{bmatrix}, \qquad h(\mathbf{x}) = (2\pi)^{-1/2}$$

$$\boldsymbol{\eta} = \begin{bmatrix} \mu / \sigma^2 \\ -1/2\sigma^2 \end{bmatrix}, \quad g(\boldsymbol{\eta}) = (-2\eta_2)^{1/2} e^{\eta_1{}^2/4\eta_2}$$

$$P(\mathbf{x};\mathrm{h}) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2} = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

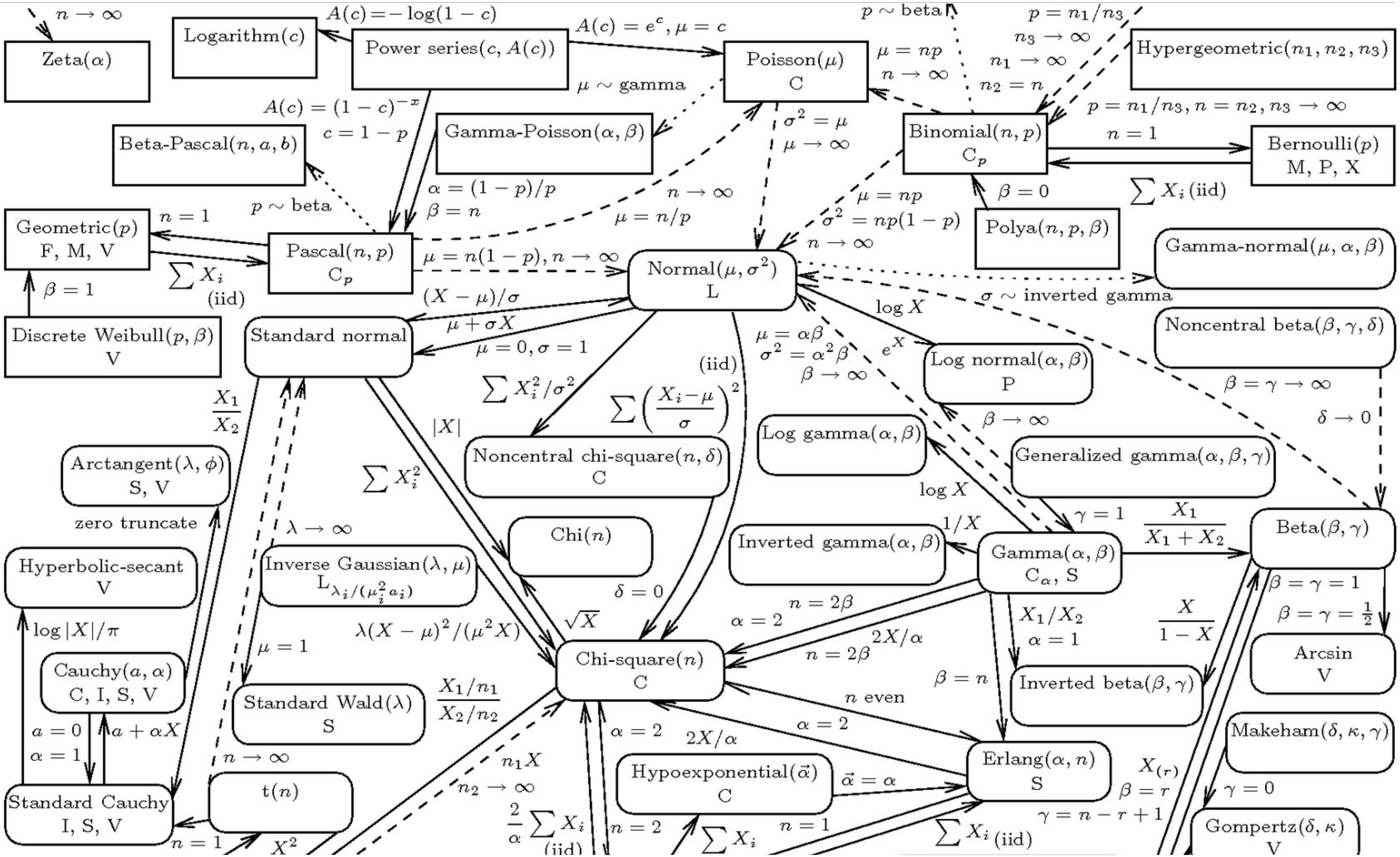
Why this mess???

- Allows us to see a broader picture

- Exponential distributions have convenient properties
  - Sufficiency
    - You won’t need more parameters for more data
  - Conjugate priors
    - Make life easy when we perform parameter estimation (more later)

And there's lots more …

http://www.math.wm.edu/~leemis/chart/UDR/UDR.html



Parameter estimation

- So what do we do with distributions?
  - We like to explain data with them
- To do so we need parameter estimation
  - Find the distribution parameters that result in explaining the observed data best
  - Various ways to go about it

Parameter estimation

- Given some independent samples:

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$$

- and a model:

$$P(\mathbf{X}; \theta)$$

- Find the parameters θ

Maximum likelihood

- The overall likelihood is:

$$P(\mathbf{X};\theta) = P(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N;\theta) = \prod_i P(\mathbf{x}_i;\theta)$$

- We want to find:

$$\theta_{ML} = \arg\max_{\theta} \prod_i P(\mathbf{x}_i;\theta)$$

- We can use straightforward solving

Maximum likelihood

- Set the derivative to zero:

$$\frac{\partial \prod_i P(\mathbf{x}_i;\theta)}{\partial \theta} = 0$$

- Go to the log domain to remove product:

$$\frac{\partial \log \prod_i P(\mathbf{x}_i;\theta)}{\partial \theta} = \sum_i \frac{\partial \log P(\mathbf{x}_i;\theta)}{\partial \theta} = \sum_i \frac{1}{P(\mathbf{x}_i;\theta)} \frac{\partial P(\mathbf{x}_i;\theta)}{\partial \theta} = 0$$

- Substitute your P and solve

Example

- Mean of Gaussian distributed data
  - Define the model:

$$P(\mathbf{x};\mu,\sigma^2) = \prod_{i=1}^{N} \mathcal{N}(x_i;\mu,\sigma^2)$$

  - Form log-likelihood:

$$\log P(\mathbf{x};\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\log\sigma^2 - \frac{N}{2}\log 2\pi$$

  - Set derivative to zero and solve:

$$\frac{\partial \log P(\mathbf{x};\mu,\sigma^2)}{\partial\mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\frac{\partial(x_n-\mu)^2}{\partial\mu} = 0 \Rightarrow \mu = \frac{1}{N}\sum_{i=1}^{N}x_i$$

Wait a minute!

- All that to prove the obvious?

- Yes, it is tedious
  - In many cases the answer will be obvious
    - But keep in mind that looks might be deceiving!

- In other cases the answer will not be easy
  - Requiring numerical/approximate optimization

Maximum a posteriori (MAP)

- Sometimes we have a prior belief
  - E.g. we believe the answer should be close to some value
  - Maximum likelihood doesn’t incorporate that
  - MAP estimation does

- Same setup as before but in addition to $P(x;\theta)$ we also have a $P(\theta)$

MAP estimation

- We use Bayes' theorem and maximize:

$$P(\theta \,|\, \mathbf{x}) = \frac{P(\theta)P(\mathbf{x} \,|\, \theta)}{P(\mathbf{x})}$$

- The denominator is constant so we only have to maximize the numerator:

$$\theta_{MAP} = \arg\max_{\theta} P(\theta)P(\mathbf{x} \,|\, \theta)$$

- Same story as before …

MAP estimation example

- Estimate the mean, but use a prior:

$$P(x;\mu,\sigma^2)=\prod_{i=1}^{N}\mathcal{N}(x_i;\mu,\sigma^2),\quad P(\mu;\mu_0,\sigma_\mu^2)=\mathcal{N}(\mu,\mu_0,\sigma_\mu^2)$$

- Take log, differentiate, solve:

$$\frac{\partial}{\partial\mu}\log\prod_{i=1}^{N}P(x_i\mid\mu)P(\mu)=0$$

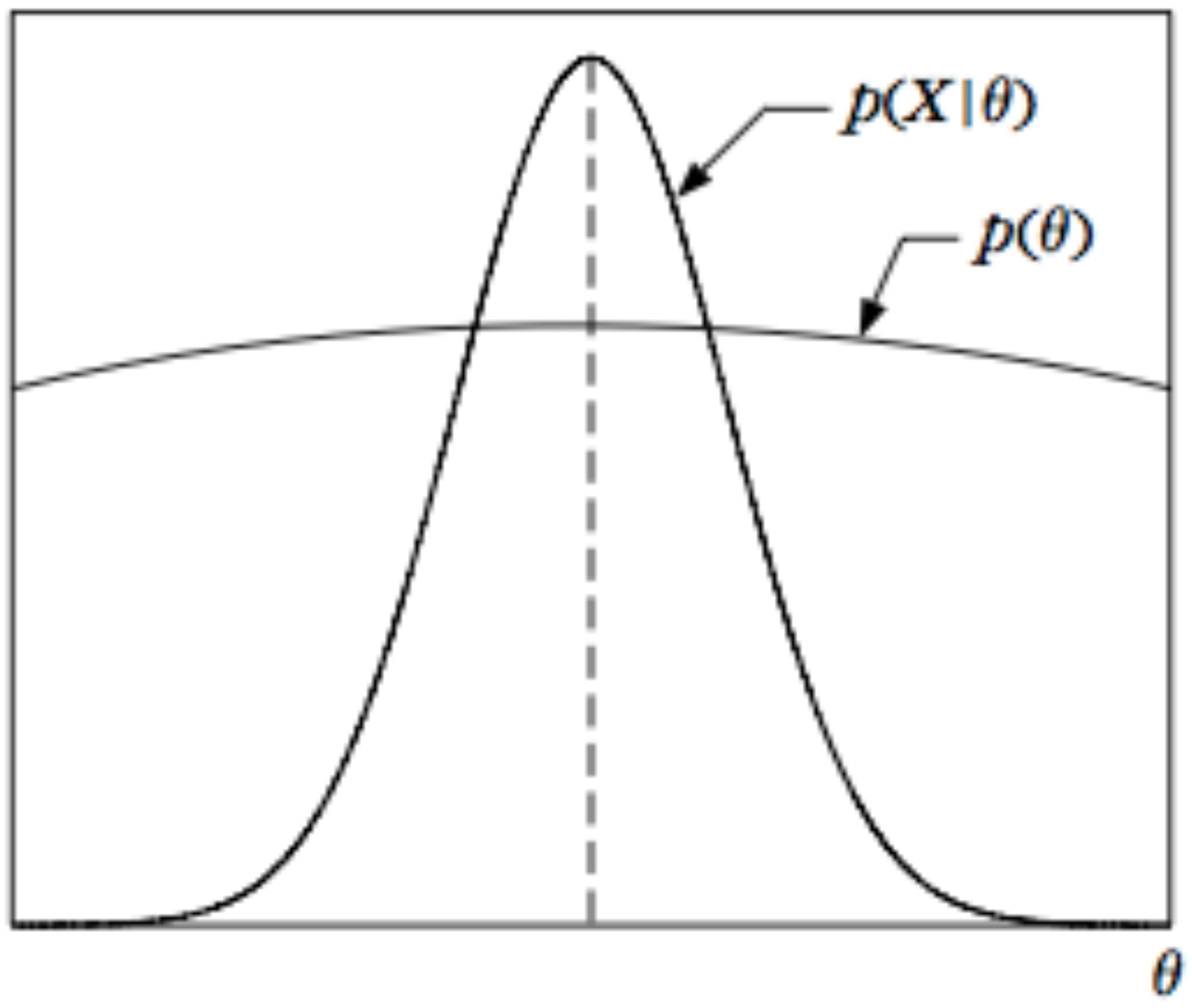
$$\sum_{i=1}^{N}\frac{1}{\sigma^2}(x_i-\mu)-\frac{1}{\sigma_\mu^2}(\mu-\mu_0)=0$$

$$\Rightarrow\mu_{MAP}=\frac{\mu_0+\frac{\sigma_\mu^2}{\sigma^2}\sum_{i=1}^{N}x_i}{1+\frac{\sigma_\mu^2}{\sigma^2}N}$$

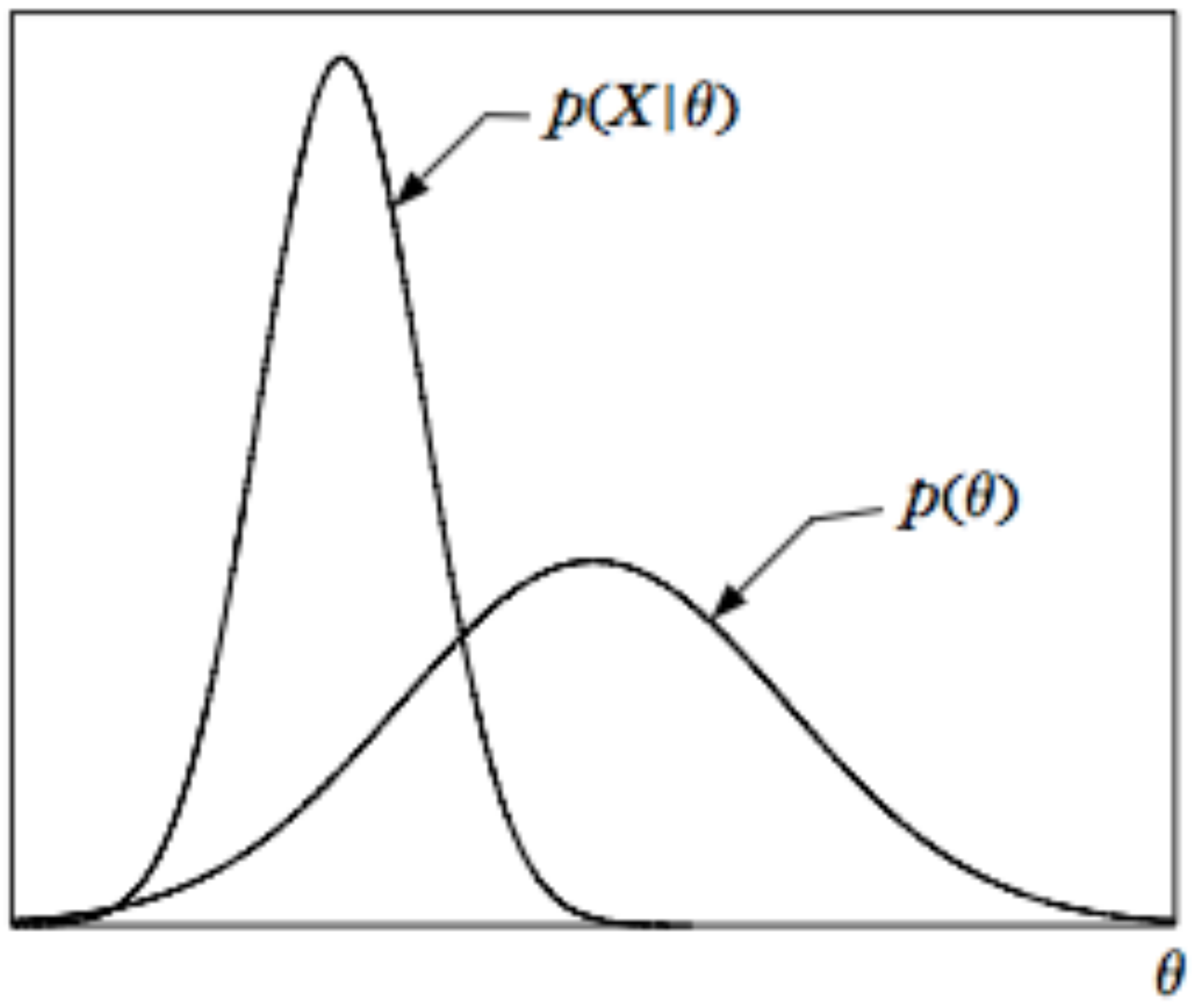
MAP vs. ML

- If $P(\theta)$ is uniform then MAP == ML
    - Otherwise they will most likely not coincide

ML and MAP will be the same

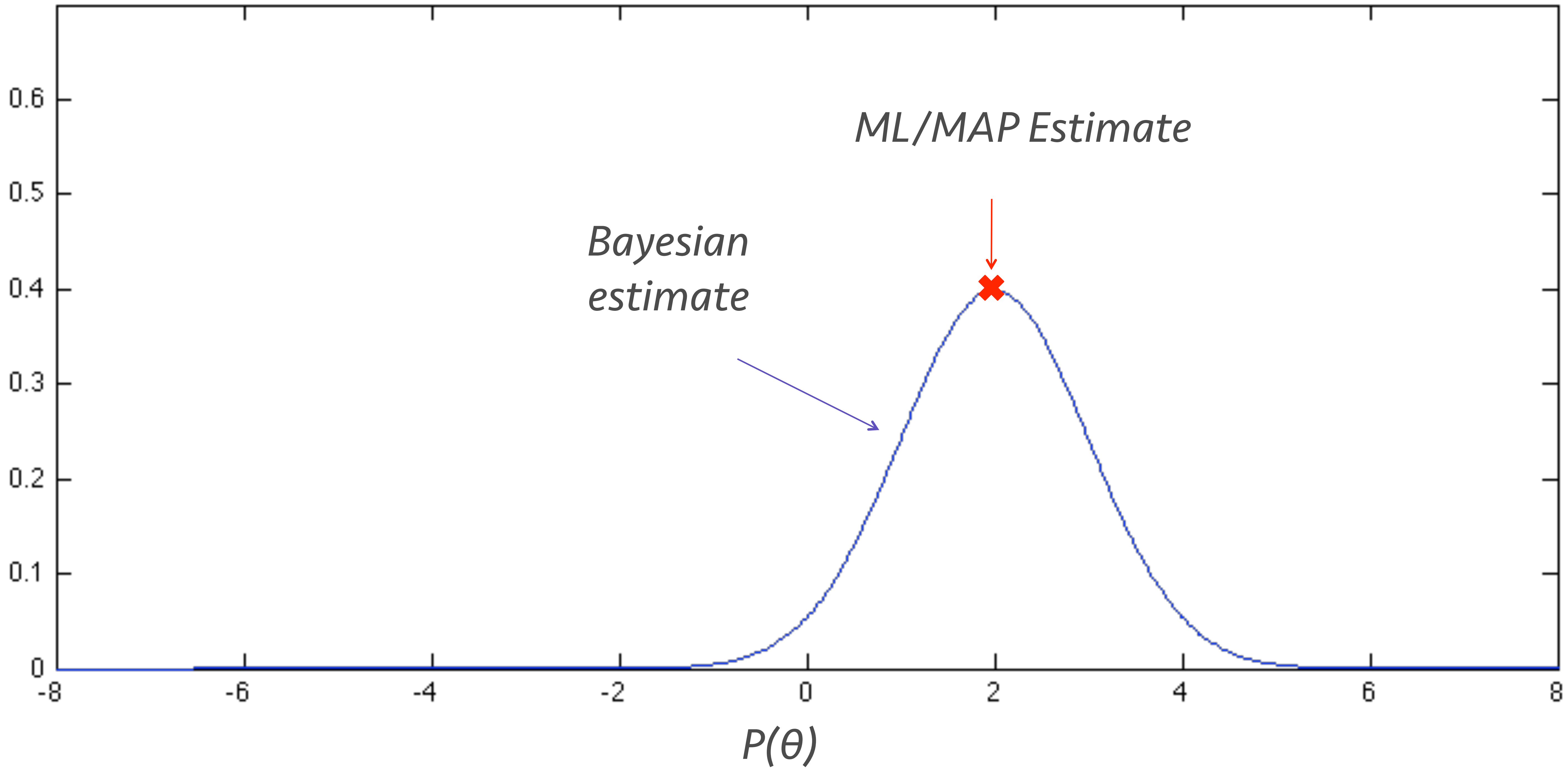


ML and MAP will be different



Bayesian inference

- Bayesian inference doesn't care about the most likely value, it cares about it's distribution



Example estimation

- Same setup as in the MAP case:

$$P(\mathbf{x};\mu,\sigma^2) = \prod_{i=1}^{N} \mathcal{N}(x_i;\mu,\sigma^2), \quad P(\mu;\mu_0,\sigma_\mu^2) = \mathcal{N}(\mu,\mu_0,\sigma_\mu^2)$$

- We now find the distribution of the mean:

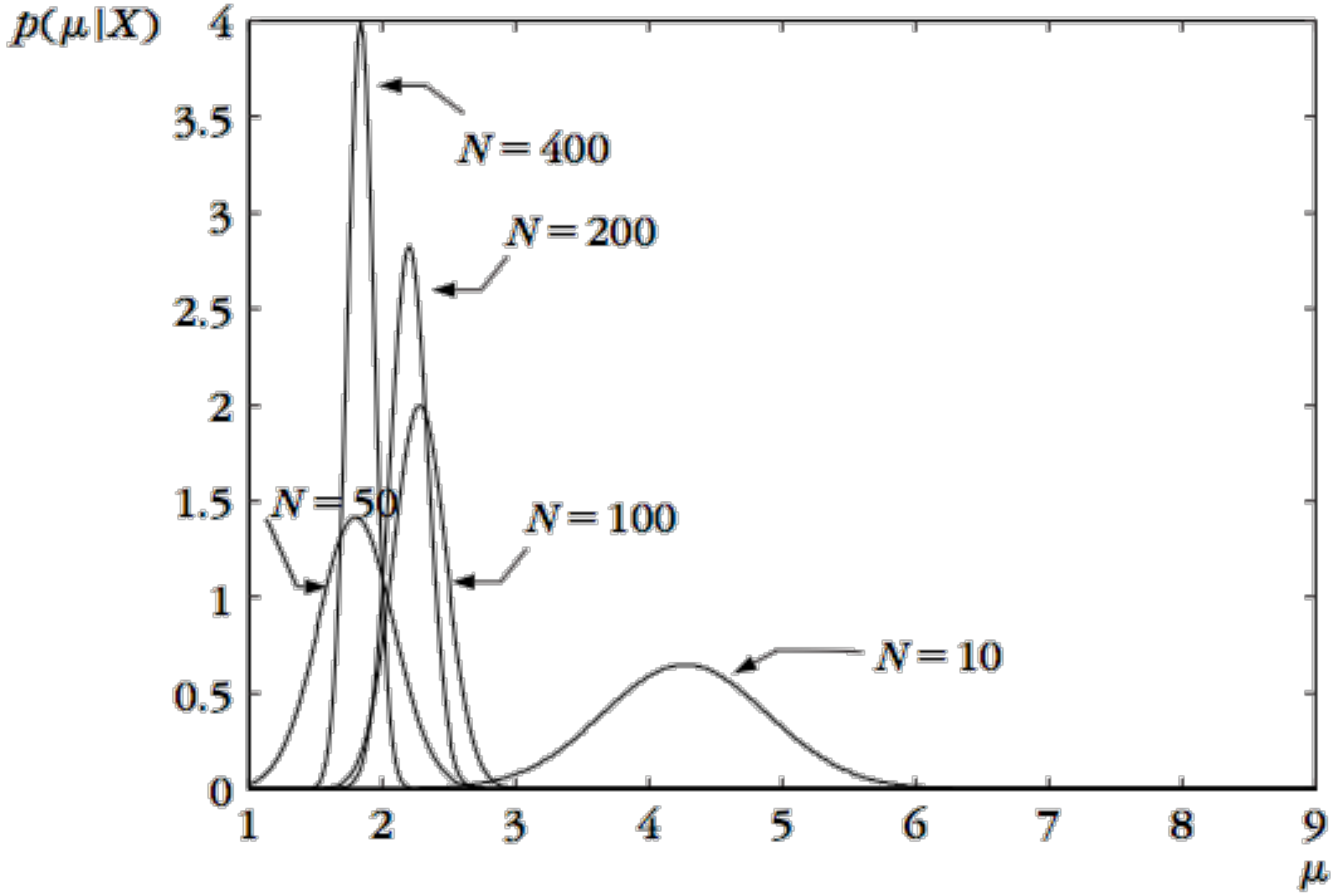
$$P(\mu\,|\,\mathbf{X}) = \frac{P(\mathbf{X}\,|\,\mu)P(\mu)}{P(\mathbf{X})} = \ldots = \mathcal{N}(\mu,\mu_N,\sigma_N^2)$$

$$\mu_N = \frac{N\sigma_0^2 \mathrm{E}[\mathbf{x}] + \sigma^2\mu_0}{N\sigma_0^2 + \sigma^2}, \qquad \sigma_N^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}$$

  - Which is also Gaussian!

Obtaining the estimate

- For different sample sizes N we obtain a different distribution of the parameter we estimate
  - The bigger the N the more sharp the distribution



And that was a clean case

- Often the distributions don’t work out
- We often resort to numerical solutions
  - Usually sampling (Monte Carlo, etc.)
- And there are many more estimation approaches!
  - We’ll see more later in the semester

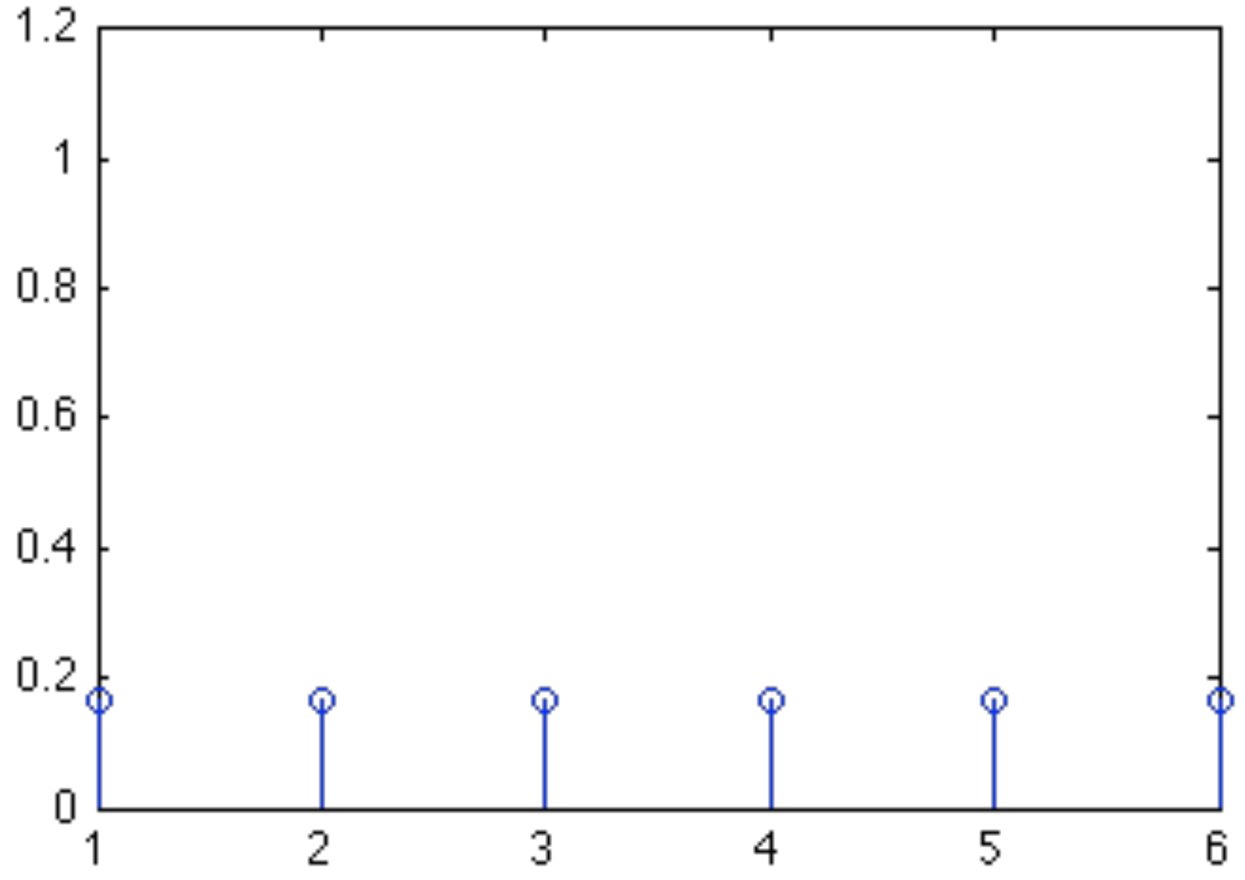
Examining the information in a signal

- Entropy: $H(x) = -\int P(x)\log P(x)dx$ or $-\sum_x P(x)\log P(x)$

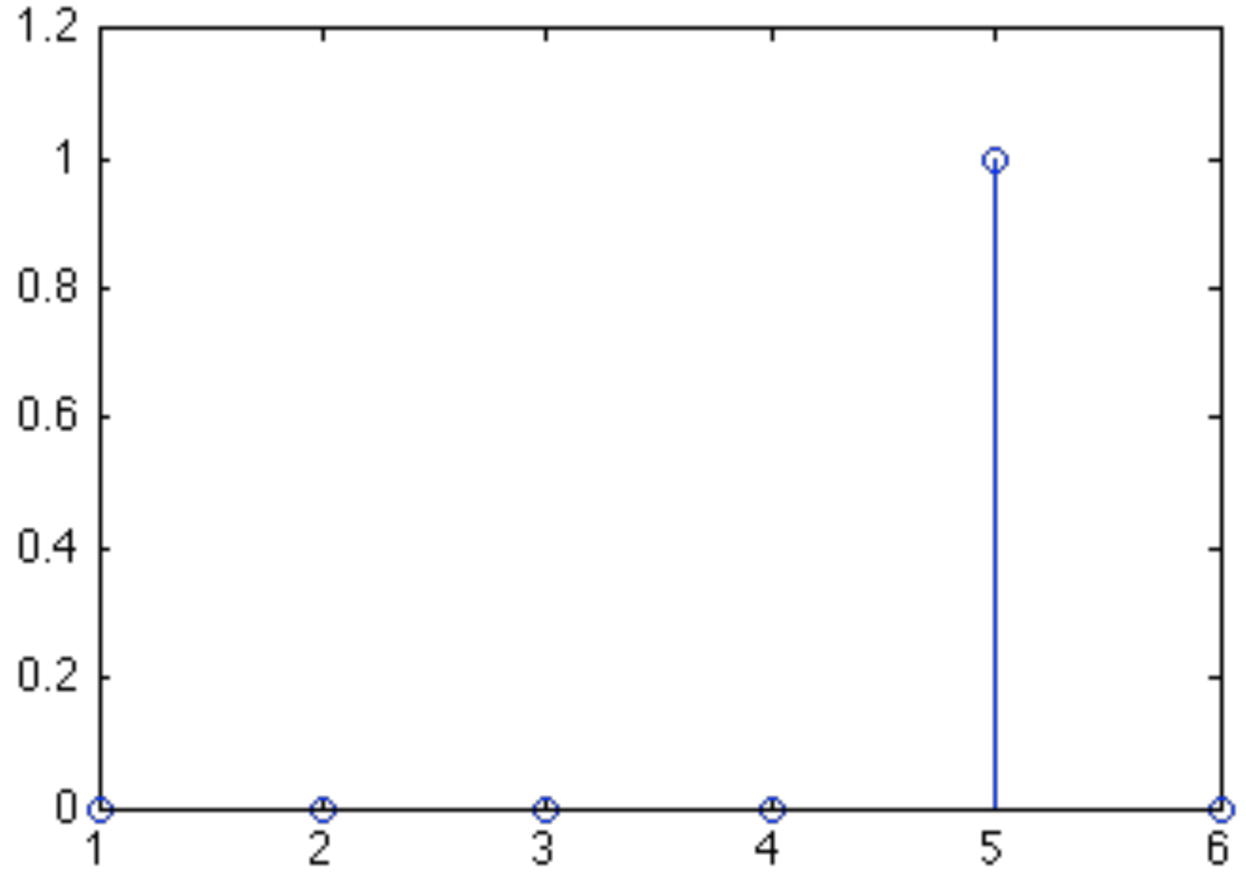
$$H(x,y) = -\int\int P(x,y)\log P(x,y)\,dx\,dy \quad \text{or} \quad -\sum_x\sum_y P(x,y)\log P(x,y)$$

  - A measure of "randomness" in a random variable

A fair die, H = 1.79
There is a lot of uncertainty
therefore more information



A heavily biased die, H = 0
no message to convey



Comparing information content

- Mutual information
  - Measures amount of shared information

$$I(x,y) = H(x) + H(y) - H(x,y)$$

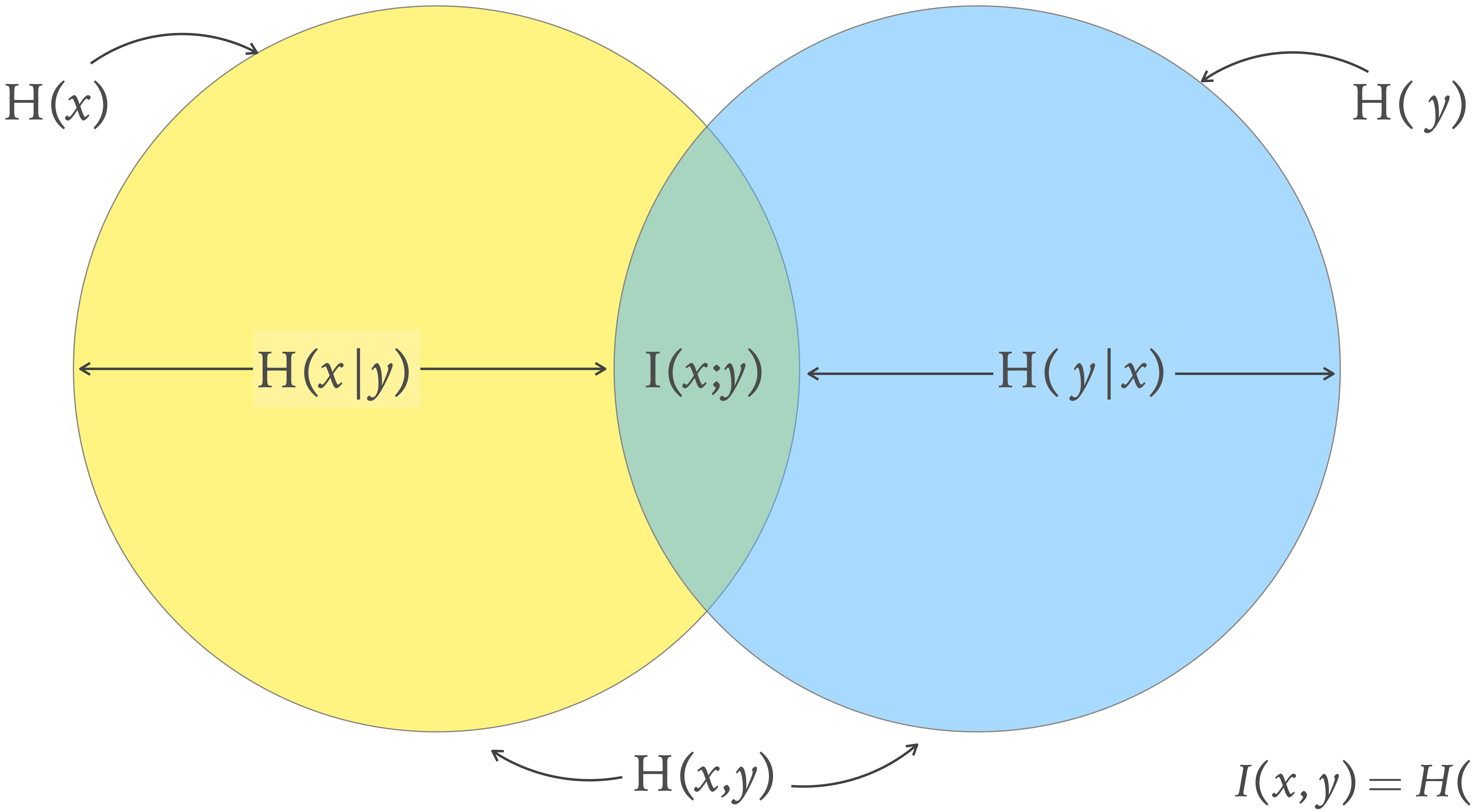
  - If it is zero then x,y are independent
- Kullback-Leibler divergence
  - a pseudo-distance for distributions

$$D(p\,||\,q) = \sum p_i \log \frac{p_i}{q_i} \quad \text{or} \quad \int p(x) \log \frac{p(x)}{q(x)} dx$$

$$D(P(x,y)\,||\,P(x)P(y)) = I(x,y)$$

  - If zero then p and q are the same

Entropy types

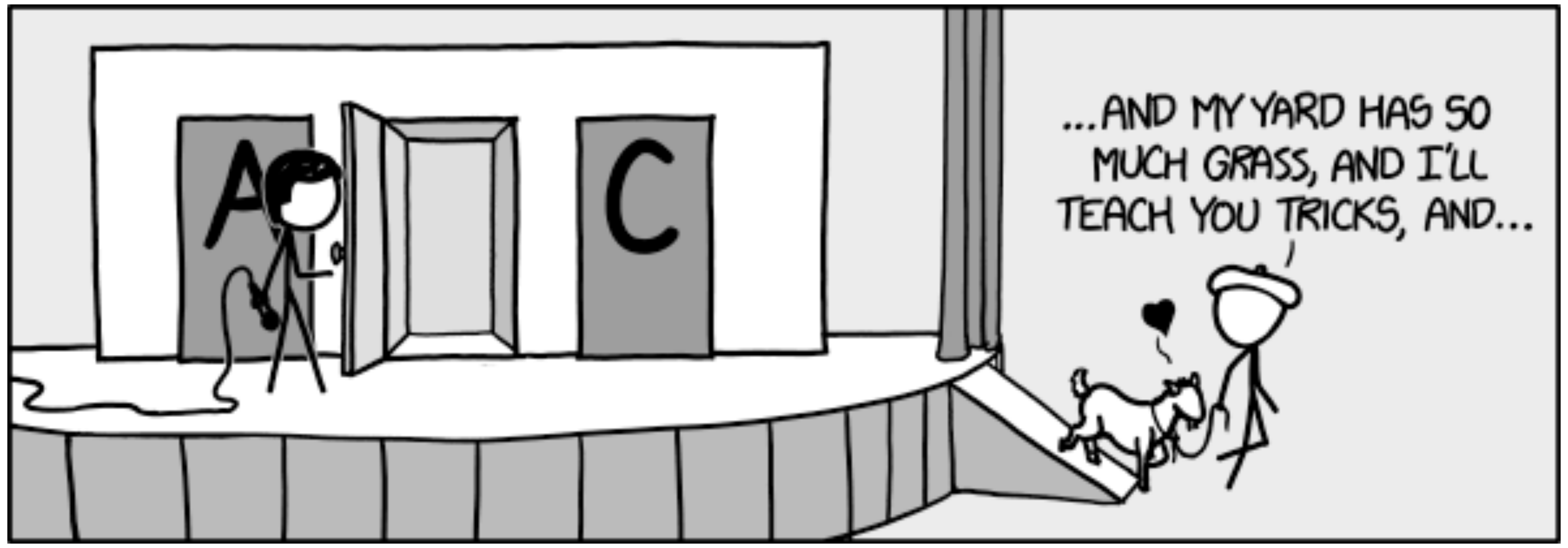


$$I(x,y) = H(x) + H(y) - H(x,y)$$

$$H(y \mid x) = H(x,y) - H(x)$$

Recap

- Probability
  - sum/product/Bayes rules
- Distributions
  - Gaussian, Laplacian, Dirichlet
- Parameter estimation
  - ML, MAP, Bayesian
- Information theory
  - Entropy, Mutual Info, KL divergence


A
C
...AND MY YARD HAS SO MUCH GRASS, AND I'LL TEACH YOU TRICKS, AND...

Too much information?

- You are not supposed to master all this
  - We will be seeing these ideas in context later
  - This lecture should serve as a reference

Some more reading

- Get textbook from class page
  - UIUC network access only (use VPN or UIUC library proxy)
- Probability basics
  - Appendix 1 of textbook
- Parameter estimation
  - Section 2.5 of textbook

Next week

- Signals refresher
    - “All of DSP in a lecture”