AM207

https://am207.github.io/2018spring/

When?

Tuesday 11.30am - 1pm, Lecture. Compulsory to attend. Maxwell Dworkin G115.

Thursday 11.30am - 1pm, Lecture. Compulsory to attend. Maxwell Dworkin G115.

Fridays 11am - 1pm Lab. Compulsory to attend. Pierce 301.

Who

Instructor:

Rahul Dave

TFs:

- Patrick Ohiomoba
- Will Claybaugh

Why take this course?

- learn how to think in principled ways of modeling..why..not just how..
- ..using bayesian statistics which is far more natural, and which has applications in almost every field
- understand deeply how and why machine leaning works
- learn generative models so that you can understand NNs, GANs better

- learn how to regularize models
- deal with data computationally large/small and statistically small/large
- learn how to optimize objective functions such as loss functions using Stochastic Gradient Descent and Simulated annealing
- Perform sampling and MCMC to solve a variety of problems
- Learn how and when to use parametric and non-parametric stochastic processes

What sorts of problems?

- machine learning hyperparameter optimization
- generalize A/B testing using Bandits (eg see https://support.google.com/analytics/answer/2844870?hl=en)
- generative modeling of images (see https://blog.openai.com/generative-models/)
- many problems in psychology, ecology, phylogenetics, public policy, etc

$$\mathbf{A}_t = \begin{array}{l} \\ \text{not yet entered} \\ \text{dead} \\ \text{offsite} \\ \text{onsite} \end{array} \begin{array}{c} \begin{array}{cccc} \text{not yet entered} & \text{dead} & \text{offsite} & \text{onsite} \end{array} \\ \begin{pmatrix} 1-\psi_t & 0 & 0 & 0 \\ 0 & 1 & 1-\phi_t & 1-\phi_t \\ \psi_t(1-\lambda_t) & 0 & \phi_t\gamma_t' & \phi_t\gamma_t'' \\ \psi_t\lambda_t & 0 & \phi_t(1-\gamma_t') & \phi_t(1-\gamma_t'') \end{pmatrix} \end{array}$$

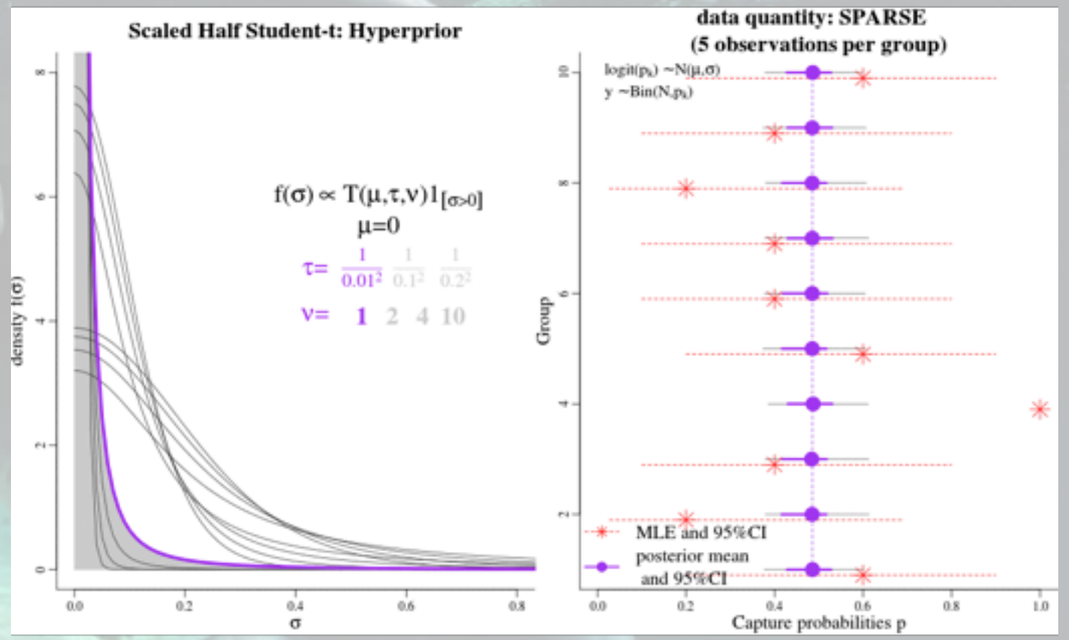
$$\mathbf{B}_{t,s} = \begin{array}{l} \\ \text{observed} \\ \text{unobserved} \end{array} \begin{array}{c} \begin{array}{cccc} \text{not yet entered} & \text{dead} & \text{offsite} & \text{onsite} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & p_{t,s} \\ 1 & 1 & 1 & 1-p_{t,s} \end{pmatrix} \end{array} \quad (1)$$

The most general model is represented as:

$$\begin{aligned} \text{initialize: } z_{0,i} &= 1 \text{ for } i = 1, \ldots, m \\ p(z_{t,i}|z_{t-1,i}, \mathbf{A}_t) &= \text{Cat}(\mathbf{A}_t[\cdot, z_{t-1,i}]) \text{ for } i = 1, \ldots, m; \\ & \quad t = 1, \ldots, T \\ p(y_{t,s,i}|z_{t,i}, \mathbf{B}_{t,s}) &= \text{Cat}(\mathbf{B}_{t,s}[\cdot, z_{t,i}]) \text{ for } i = 1, \ldots, m; \\ & \quad s_t = 1, \ldots, S_t;\ t = 1, \ldots, T \end{aligned}$$

$$\pi(\{\mathbf{A}\}_t^T, \{\mathbf{B}\}_s^{S_t}|\mathbf{Y}, \Lambda) \propto \left(\prod_i^m \left(\prod_{t=1}^T \left(\prod_{s_t=1}^{S_T} p(y_{t,s,i}|z_{t,i}, \mathbf{B}_{t,s})\right)\right.\right.$$
$$\left.\left. p(z_{t,i}|z_{t-1,i}, \mathbf{A}_t)\right)\right)\pi(\Lambda) \quad (2)$$

Bayesian Hierarchical Mark-Recapture Models (see https://www.frontiersin.org/articles/10.3389/fmars.2016.00025/full)



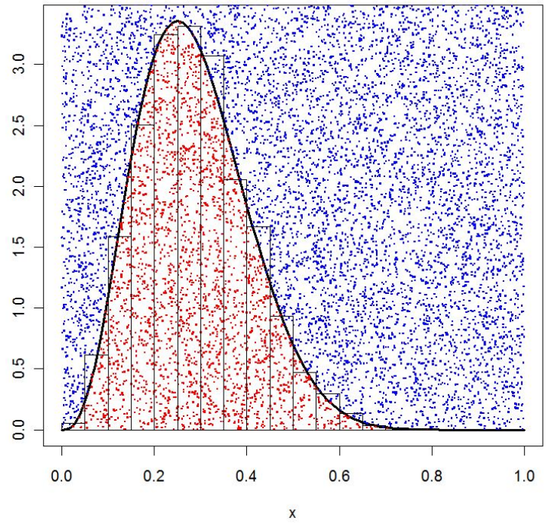
Why not?

- this is a hard course. you will have to work hard. especially on your own
- there is a lot of homework
- you do not have the requisite background
- you are a statistics expert

Modules

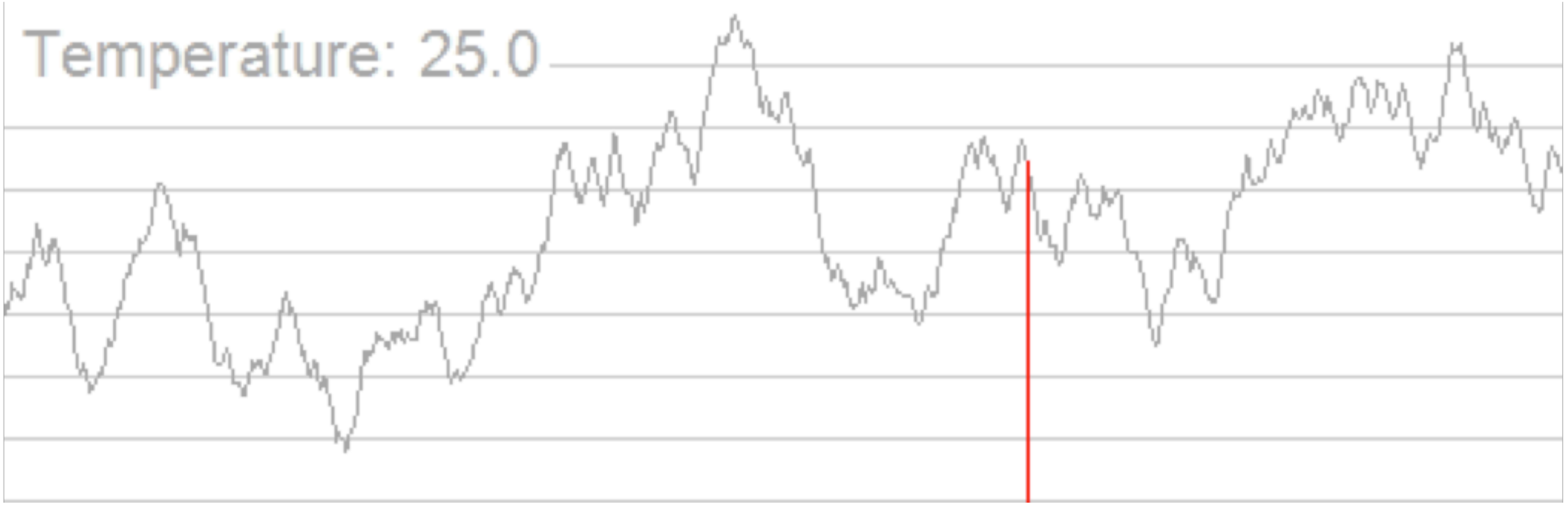
- stats review and sampling
- optimization and machine learning; stochastic optimization
- Bayesian concepts and density estimation
- MCMC and other algorithms to obtain posteriors
- Bayesian regression and glms
- Model checking, comparison, and selection
- Variational Bayes
- EXTRA: Time dependent, non-iid models, Non-parametric Bayes, or Autoencoders

sampling

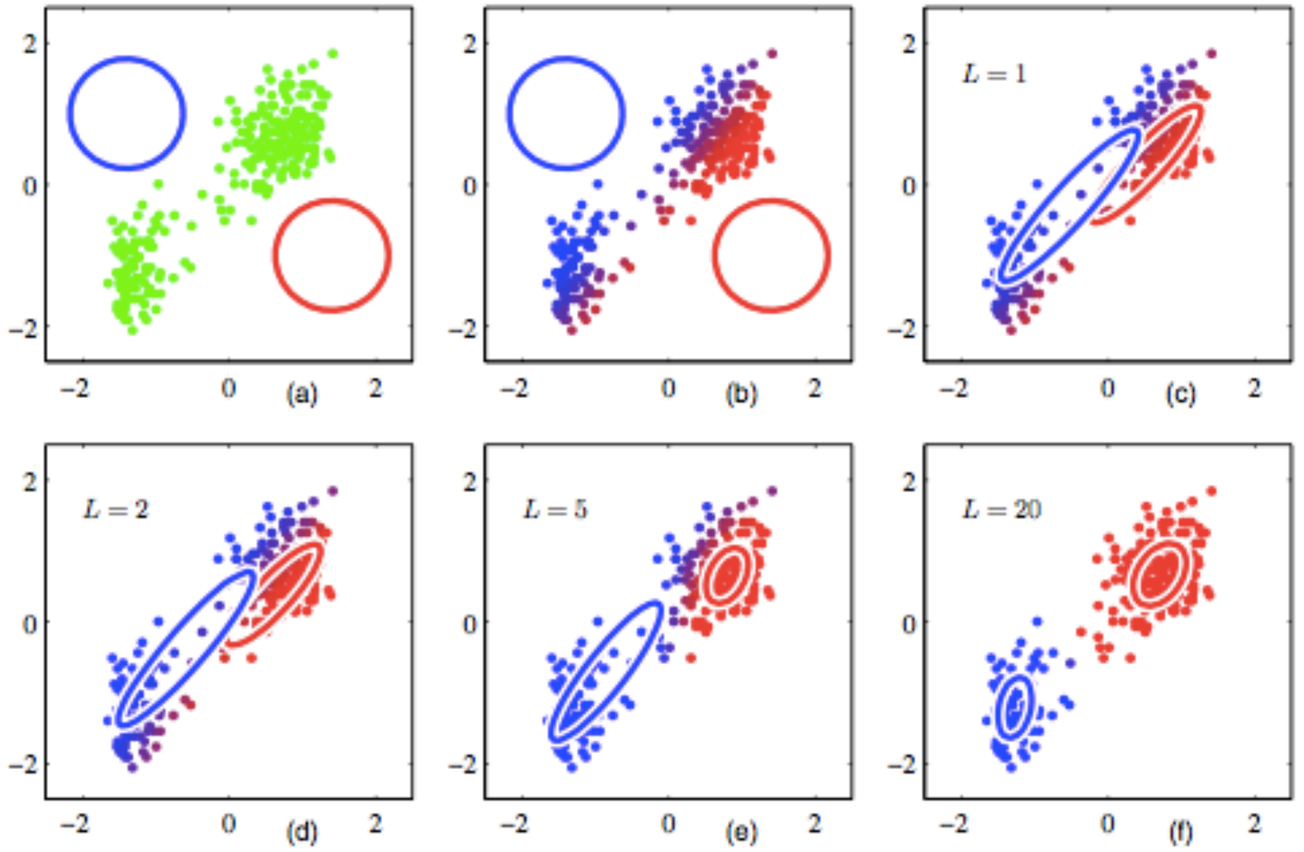


(see https://www.lancaster.ac.uk/pg/jamest/Group/index.html for nice brief introductions to some of our concepts)

optimization



(Simulated Annealing, Wikipedia)



Learns a generative model! Unsupervised learning! (EM, Bishop)

Differentiation vs Integration

- optimize a loss function: SGD, EM, etc

OR

- calculate an Expectation or a marginalization: numerical integration, monte carlo, MCMC
- two sides of the same coin

Bayesian statistics

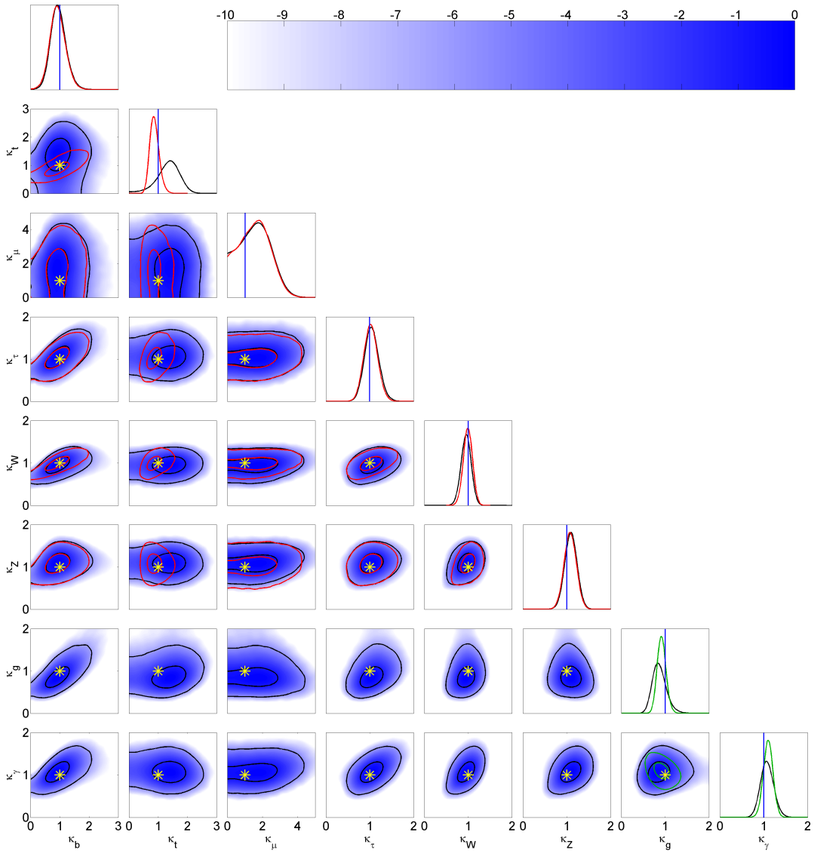
Small world:

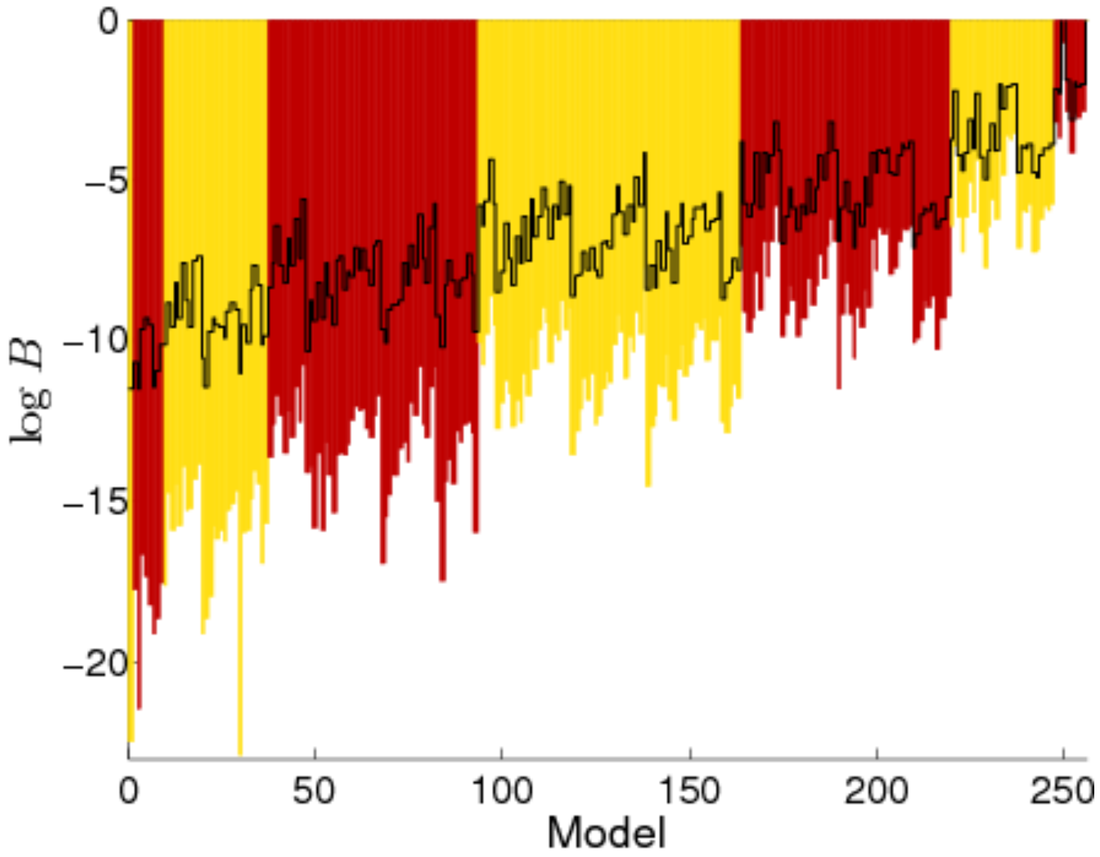
$$P(\theta \mid D) = \frac{P(D \mid \theta) \times P(\theta)}{P(D)}$$

Big World:

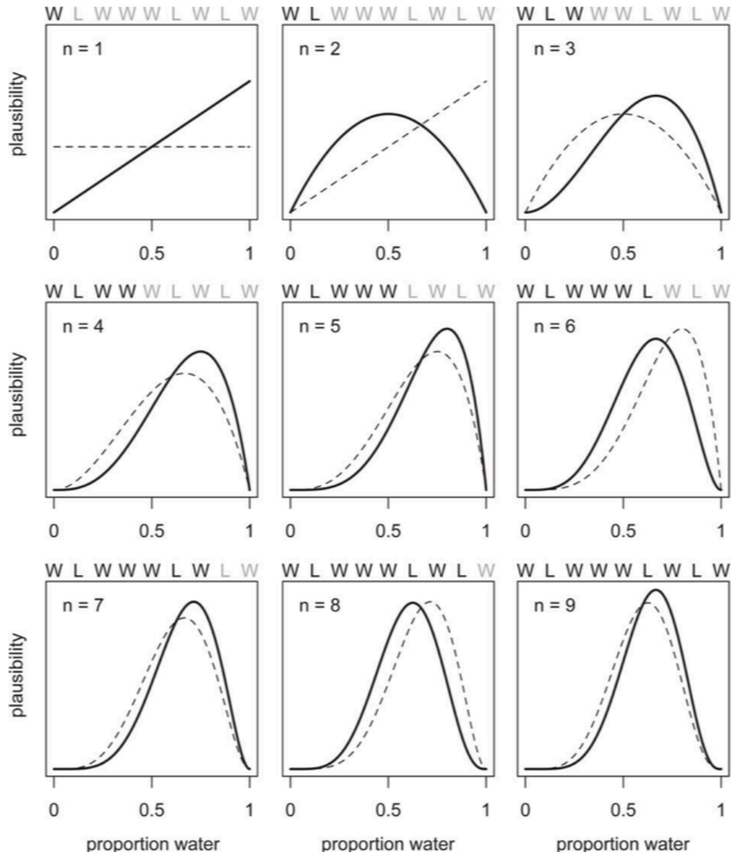
$$P(M \mid D) = \frac{P(D \mid M) \times P(M)}{P(D)}$$

Bayesian Analysis for Higgs

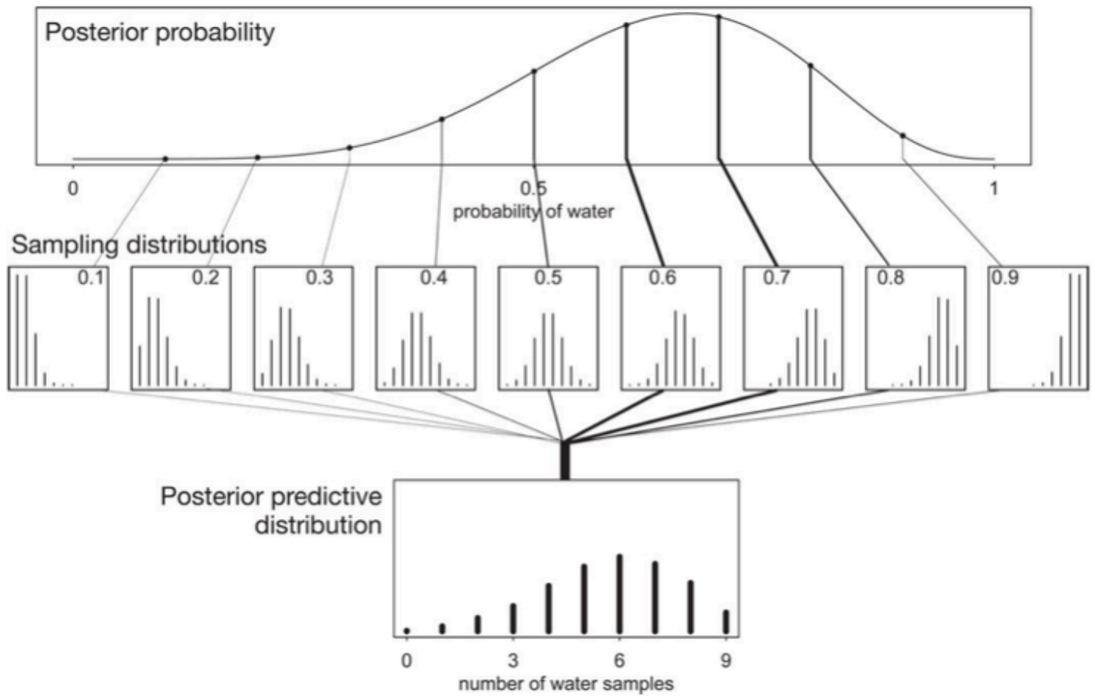


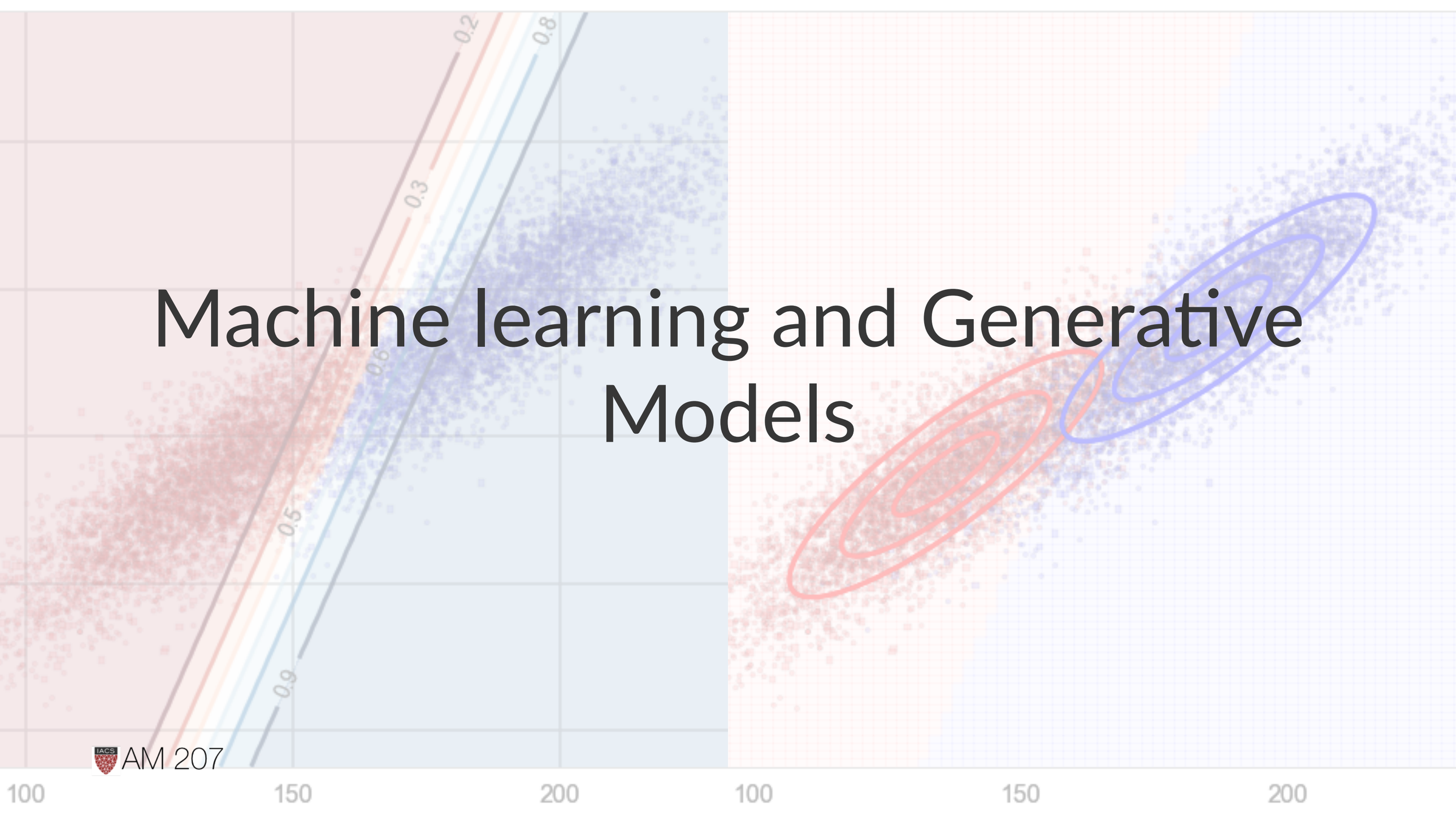


Posterior, updated



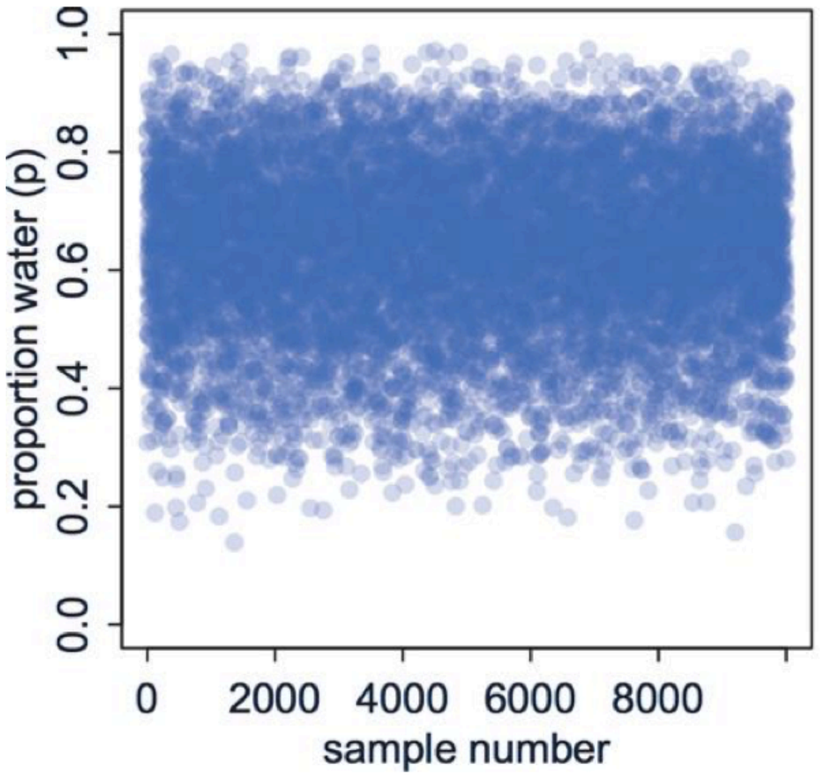
Posterior Predictive

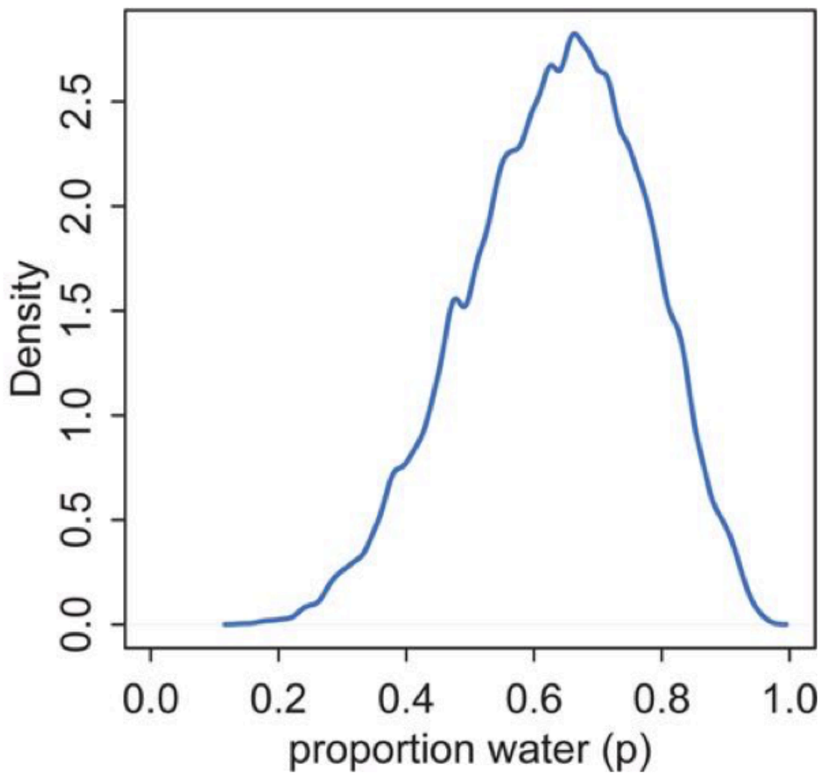




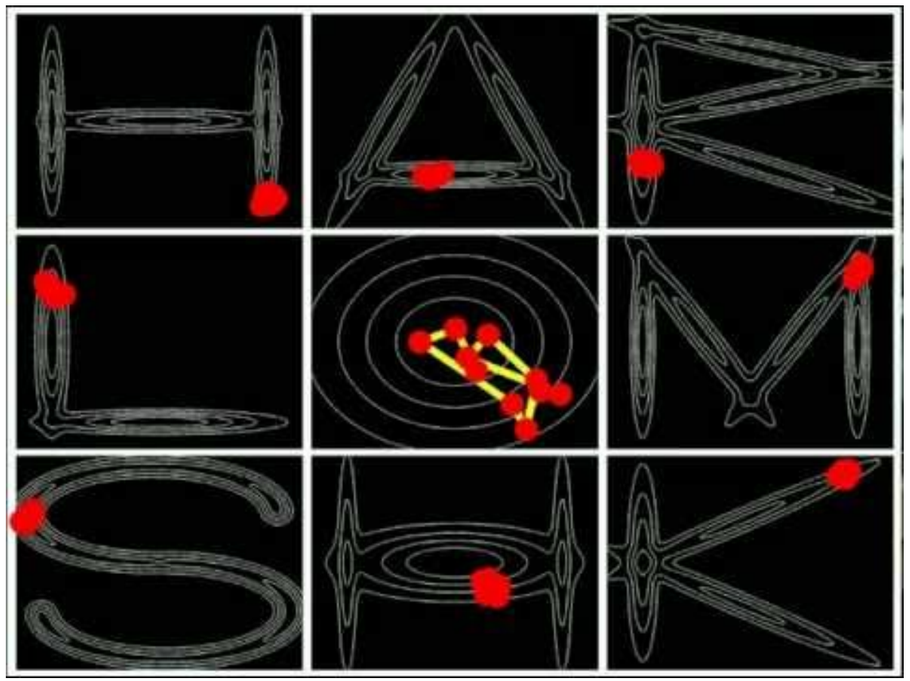
Machine learning and Generative Models

MCMC


proportion water (p)
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
sample number


Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
proportion water (p)

MCMC and HMC



(see at https://www.youtube.com/watch?v=Vv3f0QNWvWQ)

A GUIDE TO

INTEGRATION BY PARTS:

GIVEN A PROBLEM OF THE FORM:

$$\int f(x)\,g(x)\,dx = ?$$

CHOOSE VARIABLES u AND v SUCH THAT:

$$u = f(x)$$
$$dv = g(x)\,dx$$

NOW THE ORIGINAL EXPRESSION BECOMES:

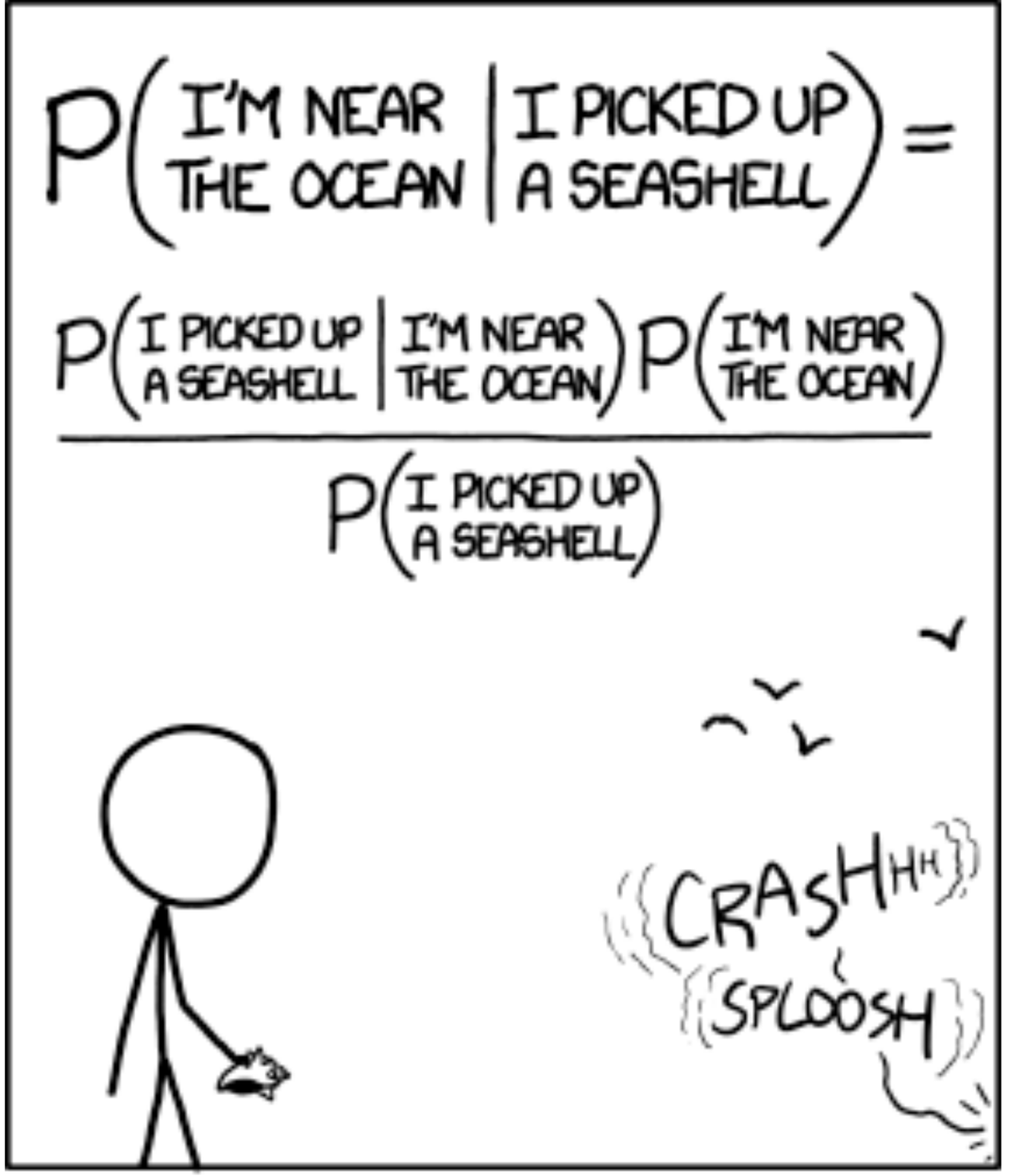
$$\int u\,dv = ?$$

WHICH *DEFINITELY* LOOKS EASIER.

ANYWAY, I GOTTA RUN.

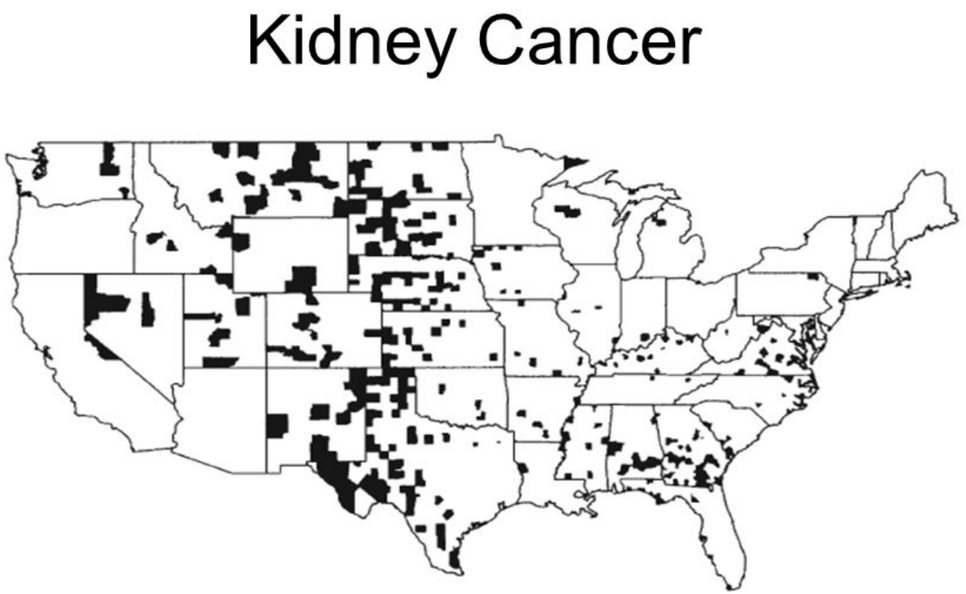
BUT GOOD LUCK!

$$P\left(\begin{array}{c}\text{I'M NEAR}\\ \text{THE OCEAN}\end{array}\middle|\begin{array}{c}\text{I PICKED UP}\\ \text{A SEASHELL}\end{array}\right) = \frac{P\left(\begin{array}{c}\text{I PICKED UP}\\ \text{A SEASHELL}\end{array}\middle|\begin{array}{c}\text{I'M NEAR}\\ \text{THE OCEAN}\end{array}\right) P\left(\begin{array}{c}\text{I'M NEAR}\\ \text{THE OCEAN}\end{array}\right)}{P\left(\begin{array}{c}\text{I PICKED UP}\\ \text{A SEASHELL}\end{array}\right)}$$



STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND *DON'T* HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

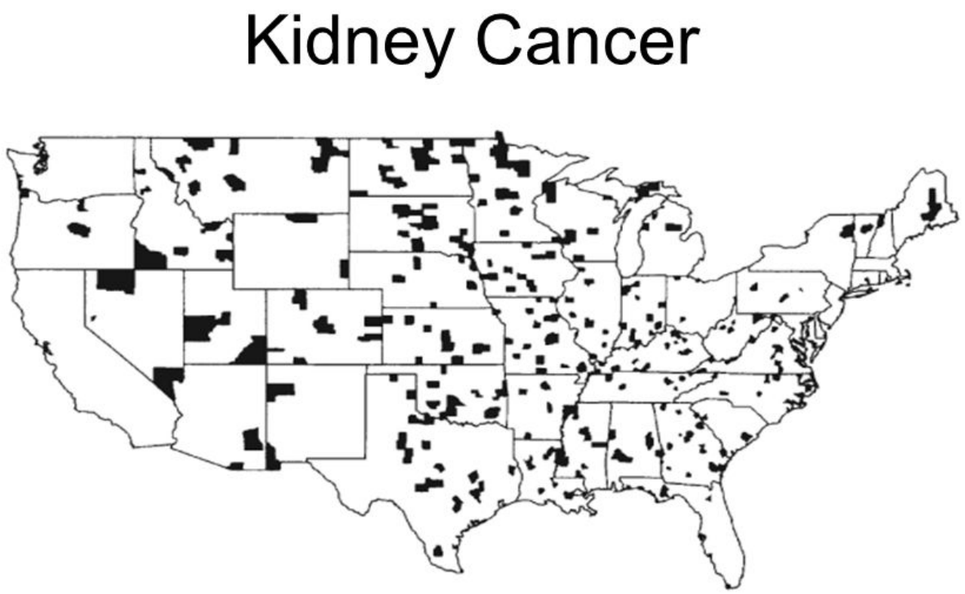
Whats up with these counties?



Counties with the lowest kidney cancer death rates

Source: Gelman et. al. Bayesian Data Anaylsis, CRC Press, 2004.

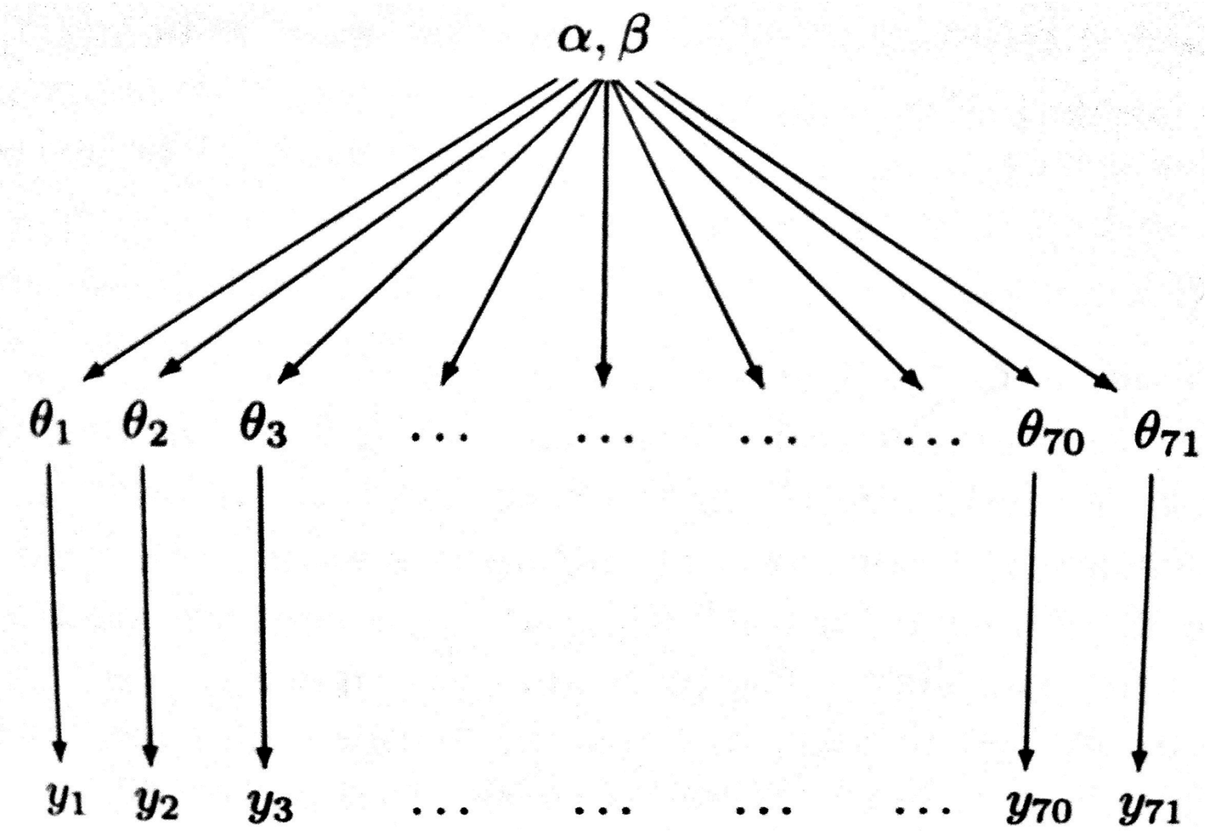
And with these?



Counties with the highest kidney cancer death rates

Source: Gelman et. al. Bayesian Data Anaylsis, CRC Press, 2004.

Hierarchical Model with regularization



glms

Monks in monastery i (indicator x_i) produce y_i manuscripts a day.

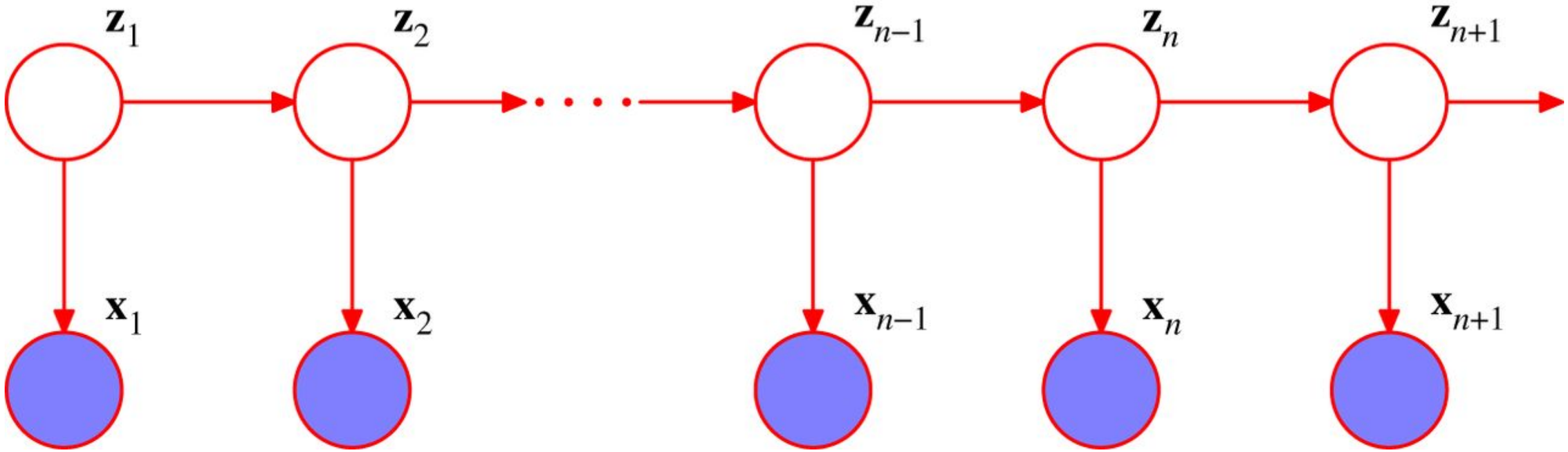
Poisson likelihood and logarithmic link

Model:

$$y_i \sim Poisson(\lambda_i)$$

$$log(\lambda_i) = \alpha + \beta x_i$$

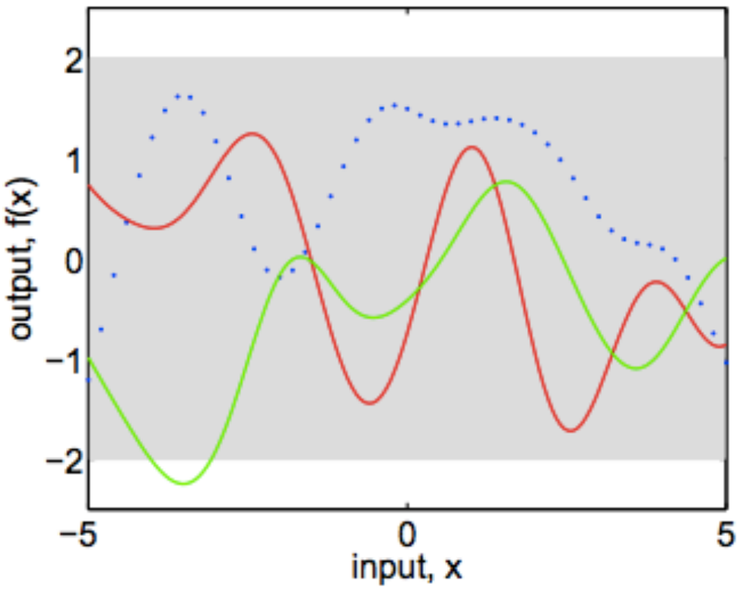
dynamical systems



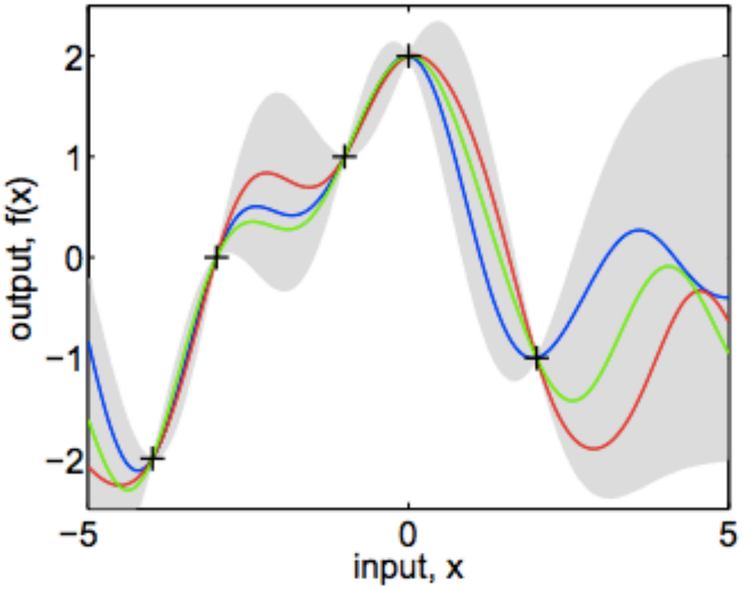
hidden markov models

gaussian processes

nonparametric, prior on functions...



(a), prior



(b), posterior

Concepts running through:

Hidden Variables

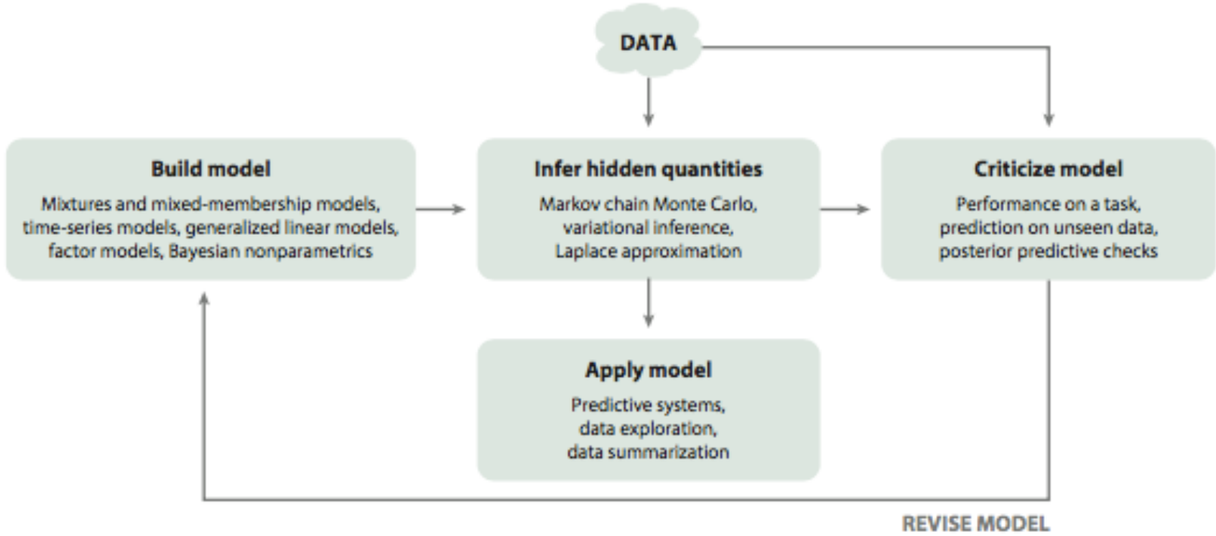
Marginalizing over nuisance parameters

Differentiation vs Integration

Frequentist vs Bayesian

Generative Models

Overall concept: Box's Loop



(image from David Blei's paper on hidden variables)

Requirements

- you will need to know how to program numerical python
- you will need to have a background in stats and simple distributions at least although we will review concepts whenever needed. Its better when you are reviewing concepts than learning it for the first time
- you should be comfortable with matrix manipulations and calculus. You should have a passing knowledge of multivariate calculus.

What kind of course?

- grad level course though nothing is really grad level hard
- if you have machine learning background you will make a lot of mental connections.
- i am your emcee; its my job to incorporate info and understanding from various places.
- probably harder than cs181 but simpler than cs281. Ideal in-between course.

Structure of the course

- lectures (2 per week), compulsory
- lab (you will play), compulsory
- homework (every week)
- paper
- final exam (a glorified project-ish homework)
- readings

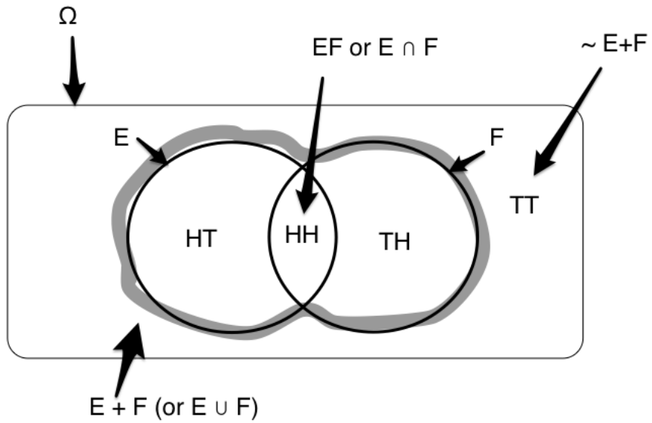
- there will be readings most weeks, some made available a lecture or two ahead
- preliminary notes will made available a lecture or two ahead..you should read these before class
- notes will be updated towards the time of the lecture
- lecture slides will be made available just before or after the lecture

- homework will be made available every week after lecture on thursday; is due every week thursday midnight. should take 6-7 hours
- expect another 6-7 hours of reading, including both before and after lecture

Probability

- from symmetry
- from a model, and combining beliefs and data: Bayesian Probability
- from long run frequency

- E is the event of getting a heads in a first coin toss, and F is the same for a second coin toss.
- Ω is the set of all possibilities that can happen when you toss two coins: {HH,HT,TH,TT}



Fundamental rules of probability:

1. $p(X) >= 0$; probability must be non-negative
2. $0 \leq p(X) \leq 1$
3. $p(X) + p(X^{-}) = 1$ either happen or not happen.
4. $p(X + Y) = p(X) + p(Y) - p(X, Y)$

Random Variables

Definition. A random variable is a mapping

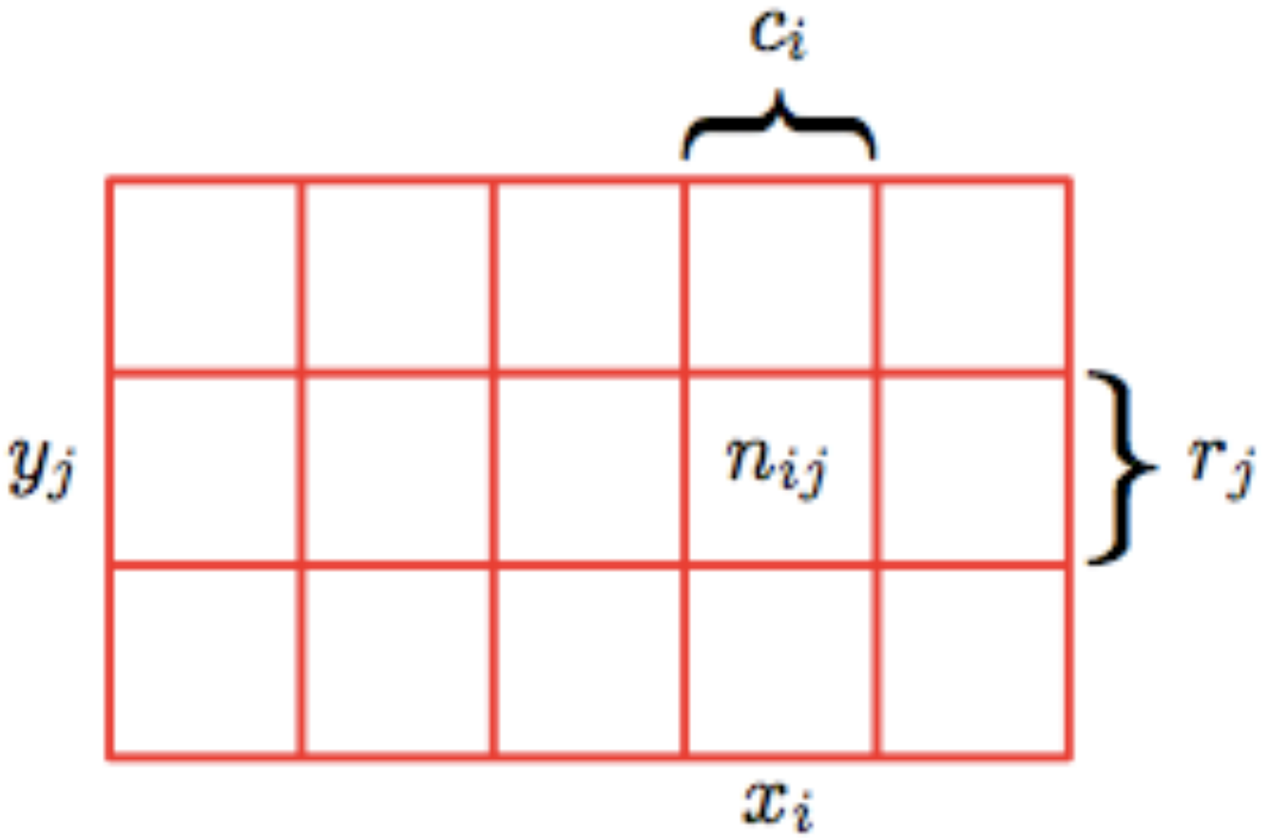
$$X : \Omega \to \mathbb{R}$$

that assigns a real number $X(\omega)$ to each outcome ω.

- Ω is the sample space. Points
- ω in Ω are called sample outcomes, realizations, or elements.
- Subsets of Ω are called Events.

- Say $\omega = HHTTTTHTT$ then $X(\omega) = 3$ if defined as number of heads in the sequence ω.
- We will assign a real number P(A) to every event A, called the probability of A.
- We also call P a probability distribution or a probability measure.

Marginals and Conditionals



$$p(X = x_i) = \sum_j p(X = x_i, Y = y_j)$$

$$p(Y = y_j \mid X = x_i) \times p(X = x_i) = p(X = x_i, Y = y_j).$$

More generally for hidden variables z:

$$p(x) = \sum_z p(x, z) = \sum_z p(x|z)p(z)$$

Bayes Theorem

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)} = \frac{p(x \mid y)\, p(y)}{\sum_{y'} p(x, y')} = \frac{p(x \mid y)\, p(y)}{\sum_{y'} p(x \mid y') p(y')}$$

Sally Clark, Convicted 1999, murder of her 2 babies.

The chance of one random infant dying from SIDS was about 1 in 1,300 during this period in Britain. The estimated odds of a second SIDS death in the same family was much larger, perhaps one in 100.

```
p(child 1 dying of sids) = 1/8500
P(child 2 dying of sids) = 1/100
p(S2 = both children dying of sids) =  0.000007
p(notS2 = not both dying of sids) =  0.999993
```

Data: both children died unexpectedly

Only about 30 children out of 650,000 annual births in England, Scotland, and Wales were known to have been murdered by their mothers. The number of double murders must be much lower, estimated as 10 times less likely.

$$p(\text{data} \mid S2) = 1$$

$$p(\text{data} \mid \text{not}S2) = 30/650000 \times 1/10 = 0.000005$$

Use Bayes Theorem

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)} = \frac{p(x \mid y)\, p(y)}{\sum_{y'} p(x, y')} = \frac{p(x \mid y)\, p(y)}{\sum_{y'} p(x \mid y') p(y')}$$

```
p(S2 | data) = P(data | S2) P(S2) /(P(data | S2) P(S2) + P(data|notS2)P(notS2))
= 1*0.000007/(1*0.000007 + 0.000005*0.999993)
=0.58
```

Sally Clark spent **3 years** in jail.

Died of acute alchohol intoxication in 2007.