ORIE 4742 - Info Theory and Bayesian ML + decision theory

February 8, 2021

Semester: Spring 2021

essential course information

- *instructor*: Sid Banerjee, sbanerjee@cornell.edu
- *TA*: Spencer Peters, sp2473@cornell.edu
- *lectures*: MW 11:25am-12:40pm, Mann 107
- *Zoom link*
  https://cornell.zoom.us/j/93025504345 (pwd: Shannon)
- *website*
  https://piazza.com/cornell/spring2021/orie4742

the fine print

- *grading*
  50% homeworks, 20% prelim, 25% project,
  5% participation
- *homeworks*
  4-5 homeworks (on average 2 weeks for each)
  teams of 2
  submit single Jupyter notebook, with theory answers in Markdown
  submissions on `https://cmsx.cs.cornell.edu`
  4 late days across homeworks, lowest grade dropped
- *prelim*
  ~~in class~~, tentatively March 29 (most likely take-home)
  no final exam
- *project*
  use techniques learned in class on problem of your choosing
  teams of up to 4, report due before finals

what is this class about

- Q1. given data, how can we learn how it was generated? inference

what is this class about

- Q1. given data, how can we learn how it was generated?
- Q2. how can we translate data and models into future decisions?

what is this class about

- Q1. given data, how can we learn how it was generated?
- Q2. how can we translate data and models into future decisions?
- Q3. what are the fundamental limits and design principles of data-driven learning and decision-making

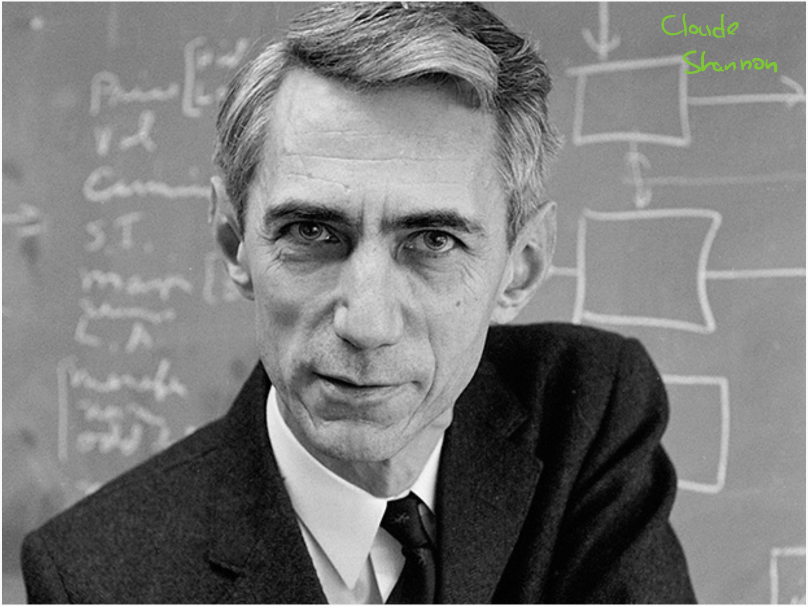
what is this class about

- Q1. given data, how can we learn how it was generated?
- Q2. how can we translate data and models into future decisions?
- Q3. what are the fundamental limits and design principles of data-driven learning and decision-making

our approach in this course: probabilistic modeling

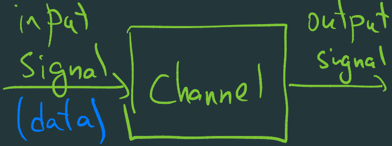
– bayesian inference: unified paradigm for learning and decision-making
– information theory: tool for designing and understanding data systems

problem: communicating over a noisy channel



communicating over channels

- mouth $\xrightarrow[\text{air}]{\text{mask}}$ ear
- ear $\xrightarrow{\text{nerves}}$ brain
- dna, dna $\xrightarrow{\text{reproduction}}$ dna
- cellphone $\xrightarrow{\text{air}}$ base tower
- data $\xrightarrow{\text{storage}}$ data in future
- generative model $\xrightarrow[\text{collection}]{\text{data}}$ data set

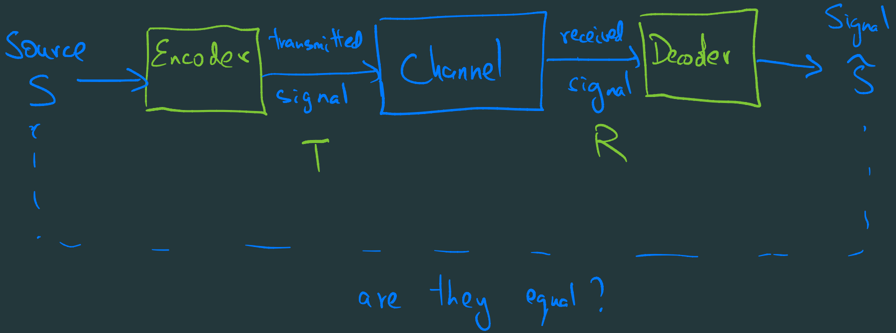


signal = data + noise

↳ learn signal

- change channel?
- 'systems approach'

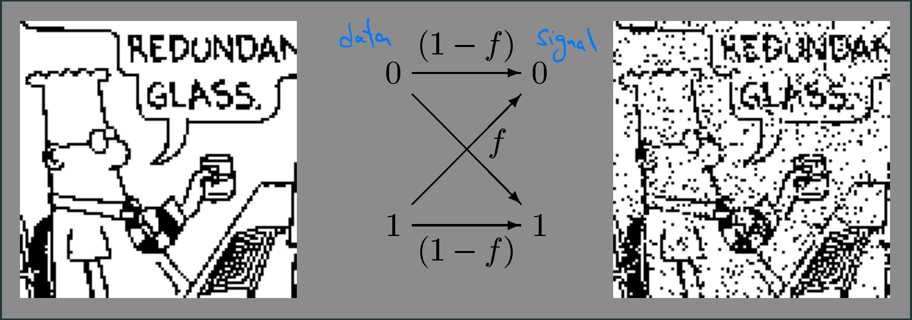
the system's solution



a toy model: the binary symmetric channel

data - binary $\{0, 1\}$

channel - flips data with prob f, else leaves it alone



credit: David Mackay

ideas for encoding

$s = 011$

$t = 000\ 111\ 111$

noise $= 001010001$

$r = t \oplus n$

↑ bitwise XOR

Eg $011 \longrightarrow$ [$n = 001$] $\longrightarrow 010$

$000\ 111\ 111 \longrightarrow 001\ \ 101\ \ 110$

 0 1 1

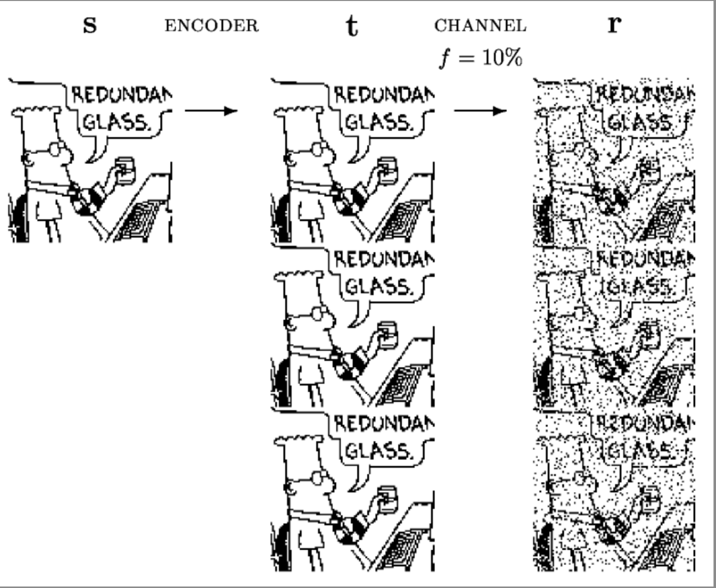
decision rule: majority

encoder - repitition, decoder - majority

encoder - parity, decoder - ?

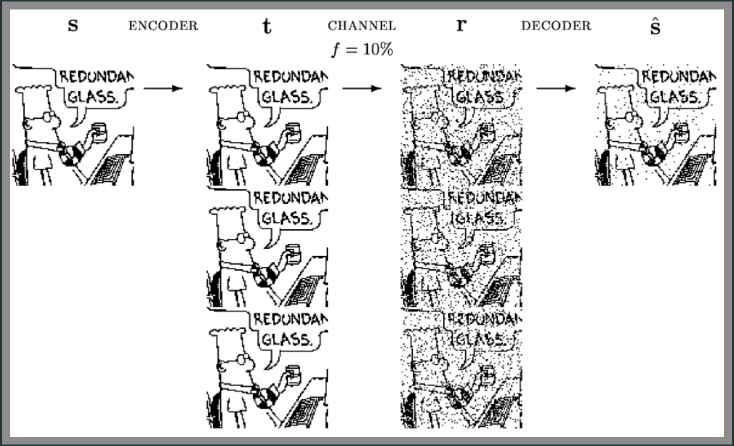
$011 \longrightarrow 0111 \longrightarrow 0101$

repetition codes: encoding



credit: David Mackay

repetition codes: decoding



credit: David Mackay

repetition codes: inference

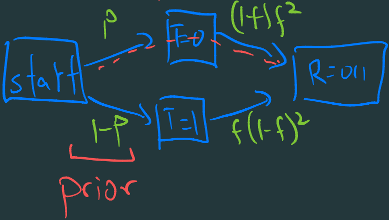
| $s = 0$ or 1, $t = sss$

to analyze – Bayes thm – inverse prob

- $\mathbb{P}[R=011 \mid T=0] = (1-f)\,f^{2}$ → # of flips
- $\mathbb{P}[R=011 \mid T=1] = f^{1}(1-f)^2$

$\Rightarrow$ majority rule $\equiv$ maximum likelihood detector

- Want – $\mathbb{P}[T=0 \mid R=011] = p(1-f)f^2 / Z$ ← MAP (compare)
- $\mathbb{P}[T=1 \mid R=011] = (1-p)f(1-f)^2 / Z$



$Z = \sout{p(1-f)f^2 + (1-p)f(1-f)^2}$ don't care

(normalizing constant / partition fn)

repetition codes: performance

If we use <u>optimal</u> inference rule, what is the prob of bit error (as a fn of # of reps)

- Eg. $R=5$ $\quad 0 \longrightarrow 00000 \xrightarrow{T} \longrightarrow R \longrightarrow$ majority

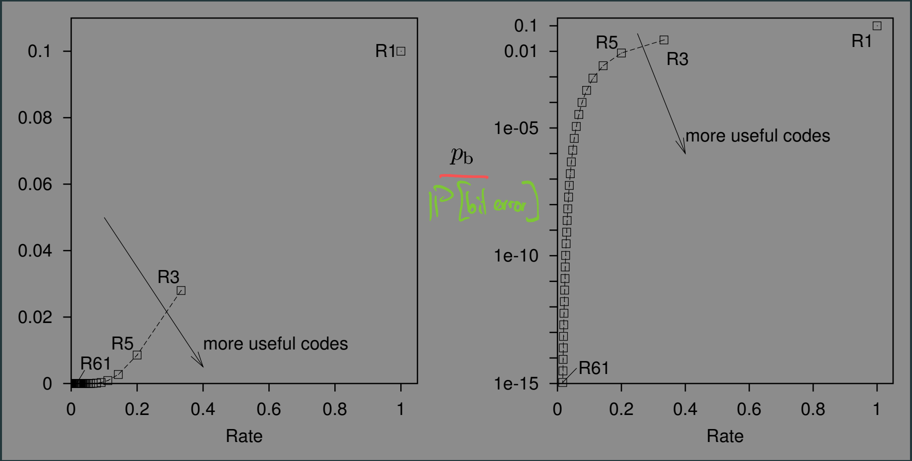
(assume $p = \mathbb{P}[s=0] = 1-p = \mathbb{P}[s=1]$)

$$\mathbb{P}[\hat{s} \neq s] = \binom{5}{3} f^3 (1-f)^2 + \binom{5}{4} f^4 (1-f) + f^5$$

$$\approx c f^3$$

In general - $\mathbb{P}[\hat{s} \neq s] \approx c f^{\lceil R/2 \rceil}$

repetition codes: the rate-error plot



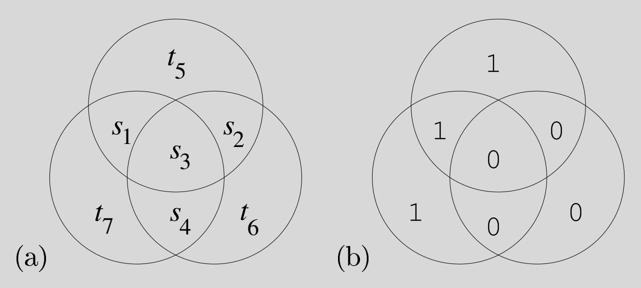
credit: David Mackay

Rate = # of bits of data / # of bits transmitted = 1/R for R reps

= S/S+1 if sending S + 1 parity bit

the (7,4) Hamming code

Rate $R = \frac{4}{7}$



credit: David Mackay

t = 4 signal bits + 3 parity bits

s = 0110 ⟹ t = 0110 001

parity bits

the (7,4) Hamming code: performance

s	t	s	t	s	t	s	t
0000	0000000	0100	0100110	1000	1000101	1100	1100011
0001	0001011	0101	0101101	1001	1001110	1101	1101000
0010	0010111	0110	0110001	1010	1010010	1110	1110100
0011	0011100	0111	0111010	1011	1011001	1111	1111111

credit: David Mackay

the (7,4) Hamming code: performance

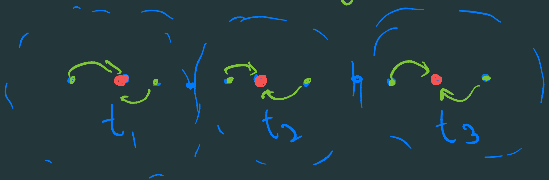
s	t	s	t	s	t	s	t
0000	0000000	0100	0100110	1000	1000101	1100	1100011
0001	0001011	0101	0101101	1001	1001110	1101	1101000
0010	0010111	0110	0110001	1010	1010010	1110	1110100
0011	0011100	0111	0111010	1011	1011001	1111	1111111

credit: David Mackay

distance between codewords

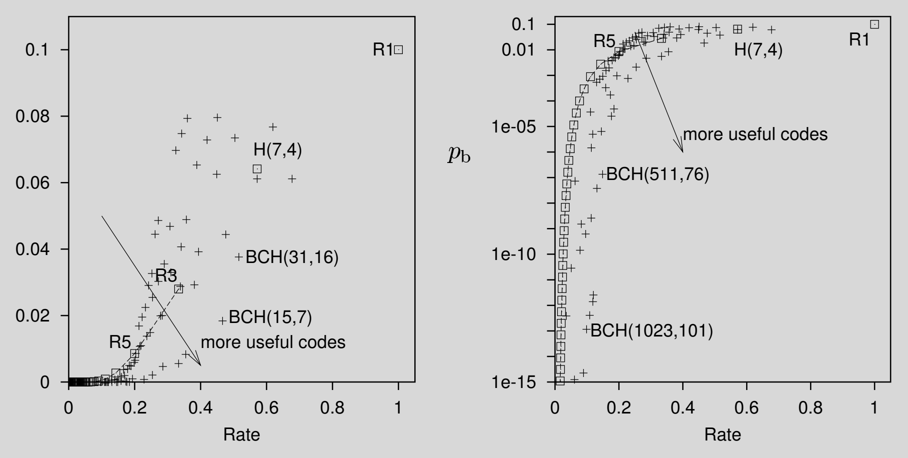
the minimal Hamming distance between any two correct codewords is 3

Hamming distance = # of different bits



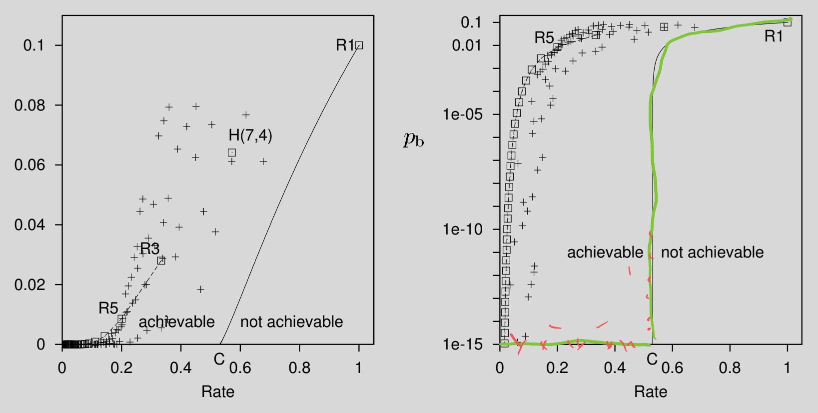
can correct a single bit error by moving to 'nearest' codeword

the rate-error plot



credit: David Mackay

Shannon's channel coding theorem (information theory)



Theorem (Claude Shannon, 1948)

for any channel, 0-error communication is possible at a rate up to $C > 0$

noisy channel communication ↔ machine learning

redundancy ⇒ inference

```
Emma Woodh*use, hands*me, clever* and rich,*with a
comfortab*e home an* happy di*position,*seemed to*unite som*
of the b*st bless*ngs of e*istence;*and had *ived nea*ly
twenty *ne year* in the*world w*th very*little *o distr*ss
or vex*her.  *he was*the yo*ngest *f the *wo dau*hters *f a
most *ffect*onate* indu*gent *ather* and *ad, i* cons*quenc*
of h*r si*ter'* mar*iage* bee* mis*ress*of h*s ho*se f*om a
ver* ea*ly *eri*d. *er *oth*r h*d d*ed *oo *ong*ago*for*her
to*ha*e *or* t*an*an*in*is*in*t *em*mb*an*e *f *er*ca*es*es*
a*d*h*r*p*a*e*h*d*b*e* *u*p*i*d*b* *n*e*c*l*e*t*w*m*n*a*
g**e**e**,**h**h** **l**n**i**l**s**r**o**a**o**e**i*
a***c***n***S***e***y***s***d***s***a***r***e***n***
W****o****s****i****l****a****g****n****t****a****e****v***
```

credit: David Mackay

redundancy ⇒ inference

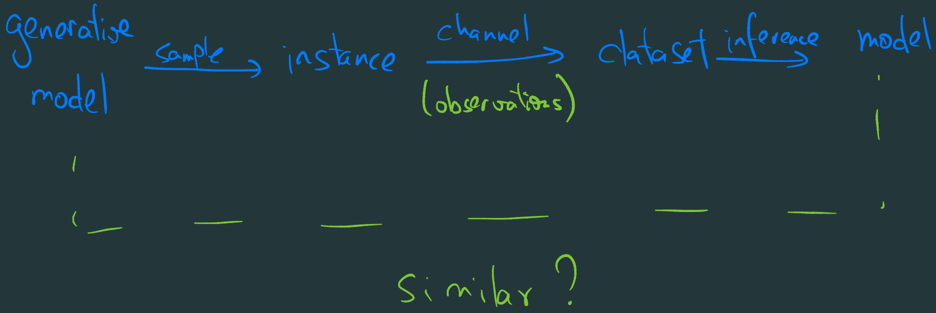
(deletion channel)

```
Emma Woodh*use, hands*me, clever* and rich,*with a
comfortab*e home an* happy di*position,*seemed to*unite som*
of the b*st bless*ngs of e*istence;*and had *ived nea*ly
twenty *ne year* in the*world w*th very*little *o distr*ss
or vex*her.  *he was*the yo*ngest *f the *wo dau*hters *f a
most *ffect*onate* indu*gent *ather* and *ad, i* cons*quenc*
of h*r si*ter'* mar*iage* bee* mis*ress*of h*s ho*se f*om a
ver* ea*ly *eri*d. *er *oth*r h*d d*ed *oo *ong*ago*for*her
to*ha*e *or* t*an*an*in*is*in*t *em*mb*an*e *f *er*ca*es*es*
a*d*h*r*p*a*e*h*d*b*e* *u*p*i*d*b* *n*e*c*l*e*t*w*m*n*a*
g**e**e**,**h**h** **l**n**i**l**s**r**o**a**o**e**i*
a***c***n***S***e***y***s***d***s***a***r***e***n***
W****o****s****i****l****a****g****n****t****a****e****v***
```

credit: David Mackay

```
Emma Woodhouse, handsome, clever, and rich, with a
comfortable home and happy disposition, seemed to unite some
of the best blessings of existence; and had lived nearly
twenty one years in the world with very little to distress
or vex her.  She was the youngest of the two daughters of a
most affectionate, indulgent father; and had, in consequence
of her sister's marriage, been mistress of his house from a
very early period. Her mother had died too long ago for her
to have more than an indistinct remembrance of her caresses;
and her place had been supplied by an excellent woman as
governess, who had fallen little short of a mother in
affection.  Sixteen years had Miss Taylor been in Mr
Woodhouse's family, less as a governess than a friend, very
```

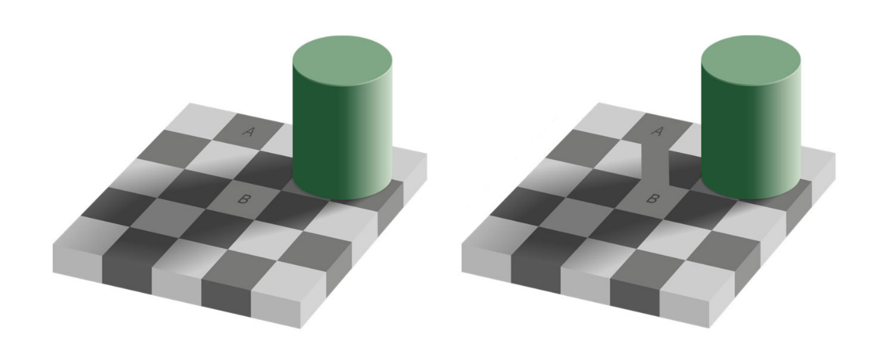
the noisy channel model in ML



noisy channels in ML

credit: MNIST dataset

we are inherently bayesian



credit: Quanta magazine, original image by Edward Adelson

"Tile A looks darker than tile B, though they are both the same shade (connecting the squares makes this clearer). The brain uses coloring of nearby tiles and location of the shadow to make inferences about the tile colors. . . lead to the perception that A and B are shaded differently."

what we hope to cover

– information theory: quantifying information and designing data systems
– bayesian inference: unified paradigm for learning from data
– decision theory: how to take actions based on what we have learned

- probability review, and introducing information measures
- data compression and the source coding theorem
- data transmission and the channel coding theorems

what we hope to cover

– information theory: quantifying information and designing data systems
– bayesian inference: unified paradigm for learning from data
– decision theory: how to take actions based on what we have learned

- probability review, and introducing information measures
- data compression and the source coding theorem
- data transmission and the channel coding theorems
- bayesian inference: priors, bayesian update, model selection
- bayesian classification/regression, gaussian processes, neural networks
- approximate inference: MCMC
- graphical models, markov random fields and causal inference

what we hope to cover

- information theory: quantifying information and designing data systems
- bayesian inference: unified paradigm for learning from data
- decision theory: how to take actions based on what we have learned

- probability review, and introducing information measures
- data compression and the source coding theorem
- data transmission and the channel coding theorems

info theory

- bayesian inference: priors, bayesian update, model selection
- bayesian classification/regression, gaussian processes, neural networks
- approximate inference: MCMC
- graphical models, markov random fields and causal inference

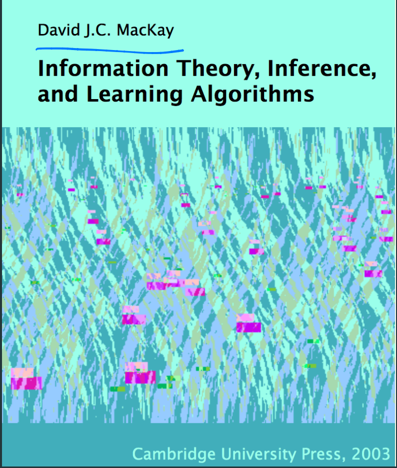
bayesian ML

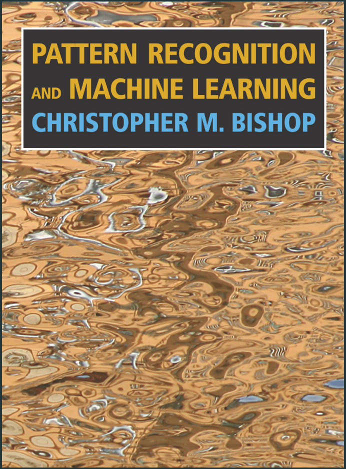
- models of decision-making
- bayesian optimization and bandit problems
- sequential decision-making and reinforcement learning

decision theory

aids in learning

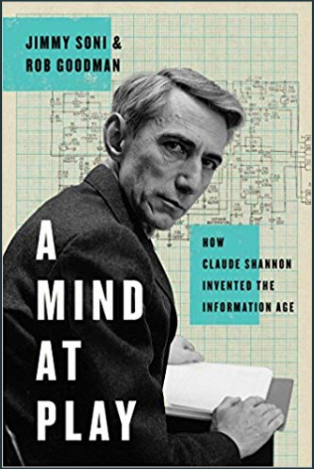
the following books are excellent references for most topics in the course

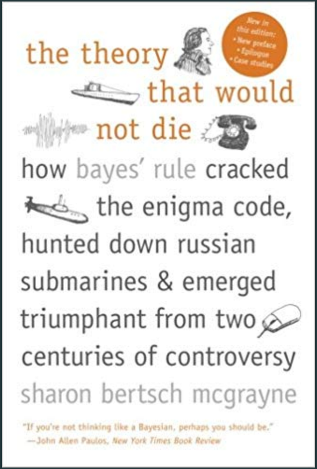




aids in getting excited about learning

the following help understand the larger context of what we will study





is this course right for you?

- prerequisites:
    - linear algebra, calculus
    - probability: ideally at the level of ORIE 3500
    - programming: python

is this course right for you?

Contact me!

- prerequisites:
  - linear algebra, calculus
  - probability: ideally at the level of ORIE 3500
  - programming: python
- caveat emptor:
  - may not be ideal as a first course in ML
  - we will focus on Bayesian methods, and ignore alternate 'frequentist' methods
  - will involve a fair bit of additional reading and programming, and some 'Bayesian philosophy'

something to puzzle on till next time

in a vaccine trial, scientists sequentially inject mice with a vaccine, and then the pathogen, and record if the mice show symptoms

- they report they tested 102 mice, of which 5 developed symptoms
  you use this to compute CIs for the vaccine's effectiveness

something to puzzle on till next time

in a vaccine trial, scientists sequentially inject mice with a vaccine, and then the pathogen, and record if the mice show symptoms

- they report they tested 102 mice, of which 5 developed symptoms
  you use this to compute CIs for the vaccine's effectiveness
- it later emerges that they kept doing trials till they got 5 negative cases (it just happened that it required 102 trials)
  do you change your estimates based on this?