ORIE 4742 - Info Theory and Bayesian ML

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*

  5-6 homeworks (on average 2 weeks for each)

  teams of up to 3

  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

  no final exam
- *project*

  use techniques learned in class on problem of your choosing

  teams of up to 3, report due before finals

what is this class about

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

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

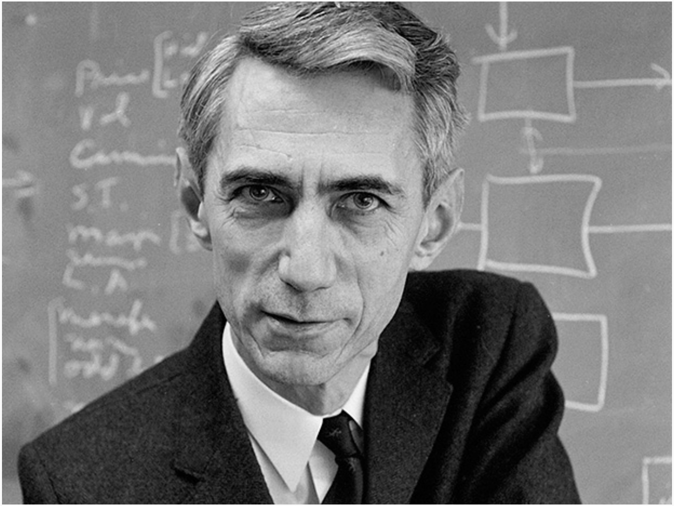
what is this class about

- 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

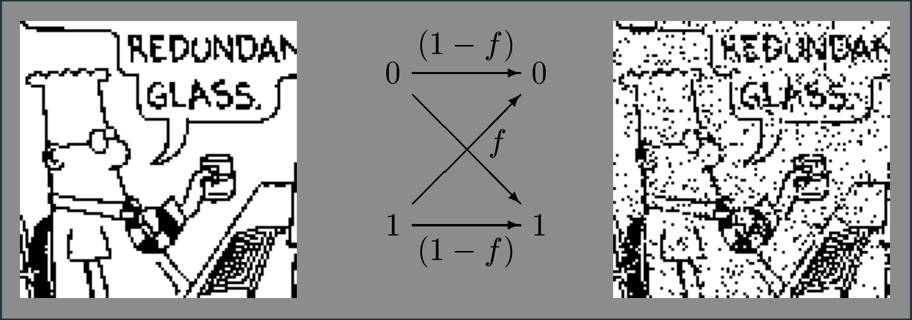


reading assignment: chapter 1 of Mackay

communicating over channels

the system's solution

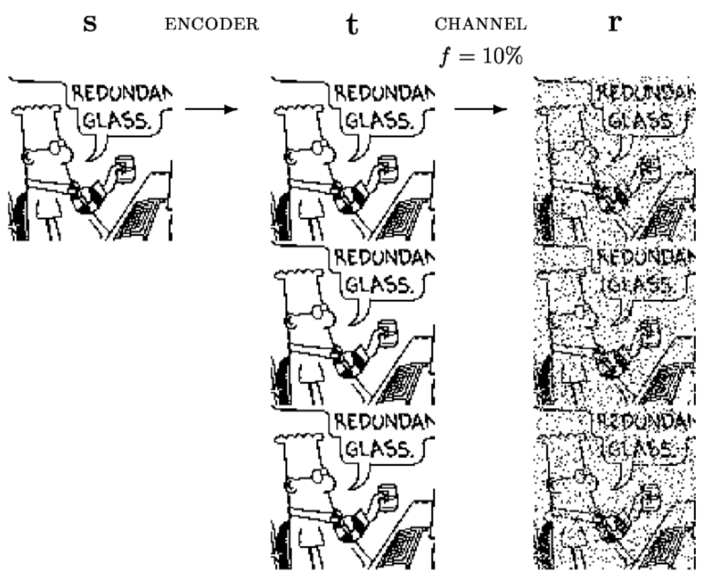
a toy model: the binary symmetric channel



credit: David Mackay

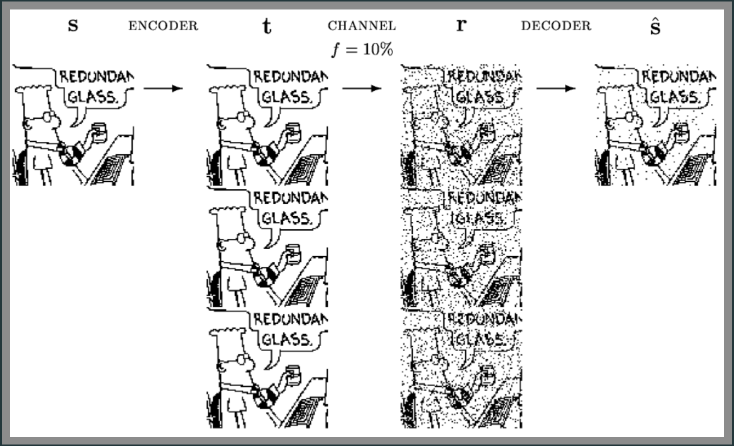
ideas for encoding

repetition codes: encoding



credit: David Mackay

repetition codes: decoding

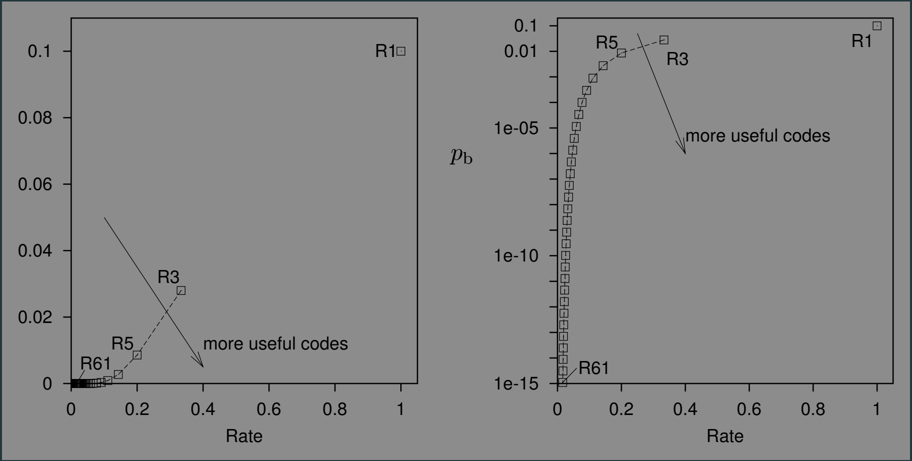


credit: David Mackay

repetition codes: inference

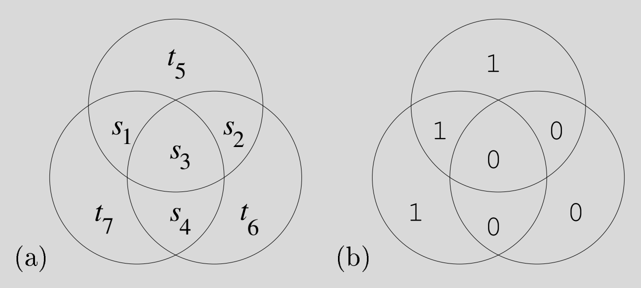
repetition codes: performance

repetition codes: the rate-error plot



credit: David Mackay

the (7,4) Hamming code



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

the (7,4) Hamming code: performance

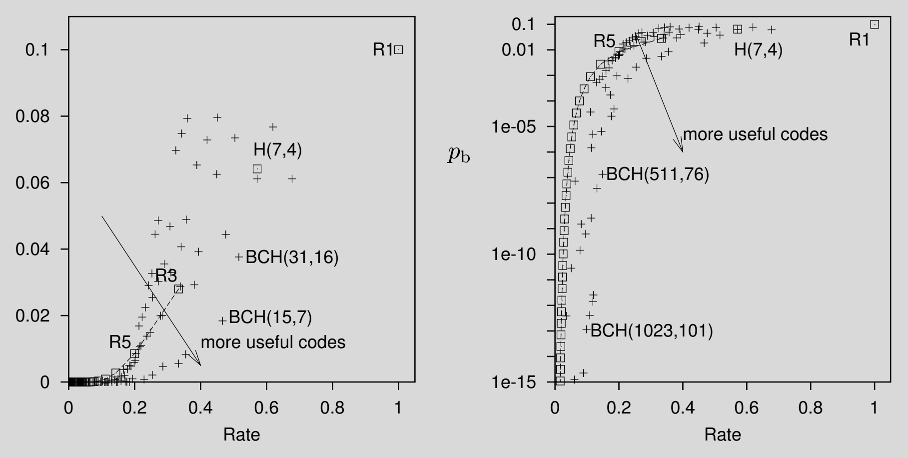
s	t	s	t	s	t	s	t
0000	0000000	0100	0100110	1000	1000101	1100	1100011
0001	0001011	0101	0101101	1001	1001110	1101	1101000
0010	0010111	0110	0110001	1010	1010010	1110	1110100
0011	0011100	0111	0111010	1011	1011001	1111	1111111

credit: David Mackay

distance between codewords

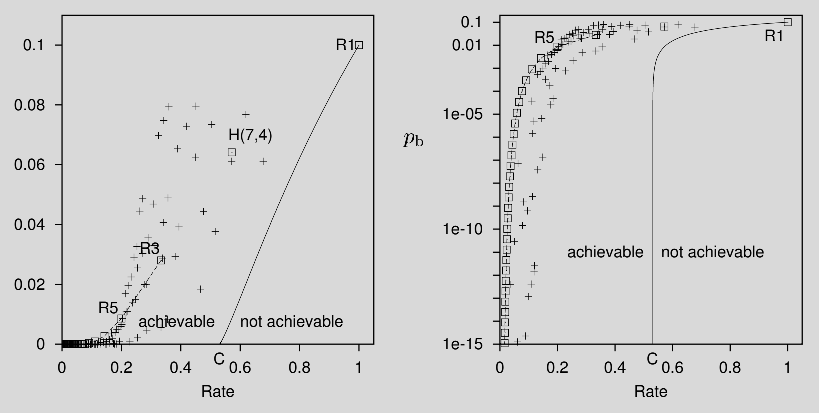
the minimal Hamming distance between any two correct codewords is 3

the rate-error plot



credit: David Mackay

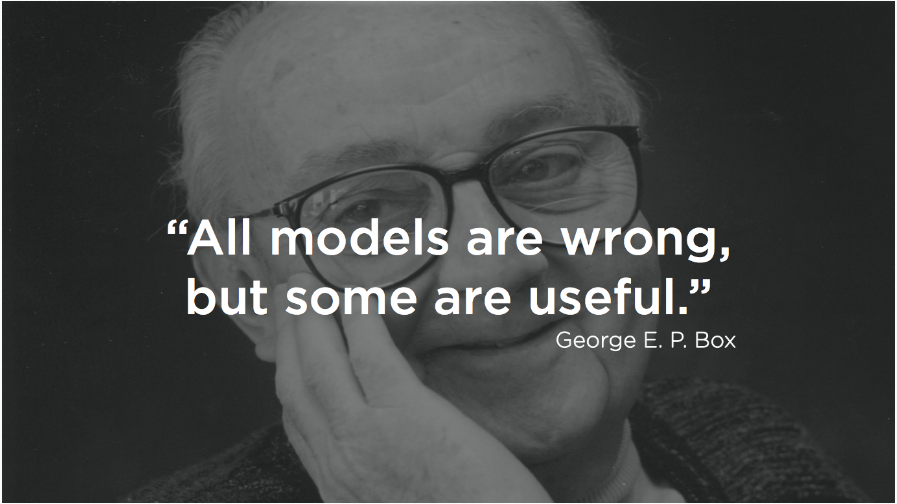
Shannon's channel coding theorem



Theorem (Claude Shannon, 1948)

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

noisy channel communication $\leftrightarrow$ machine learning



redundancy $\Rightarrow$ 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

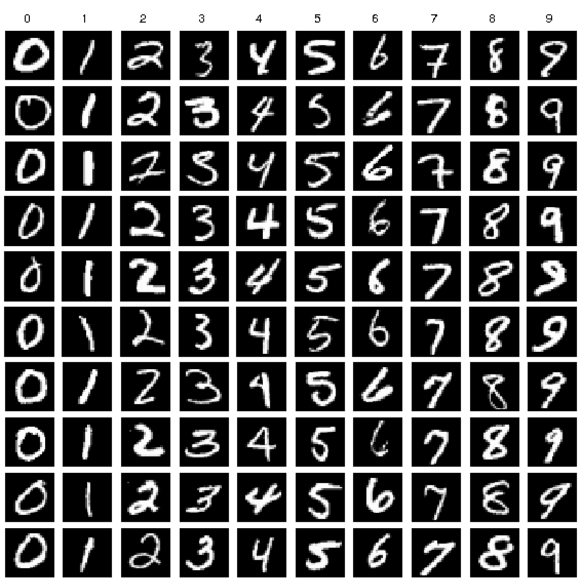
```
Emma Woodh*use, hands*me, clever* and rich,*with a
comfortab*e home an* happy di*position,*seemed to*unite som*
of the b*st bless*ngs of e*istence;*and had *ived nea*ly
twenty *ne year* in the*world w*th very*little *o distr*ss
or vex*her.  *he was*the yo*ngest *f the *wo dau*hters *f a
most *ffect*onate* 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
```

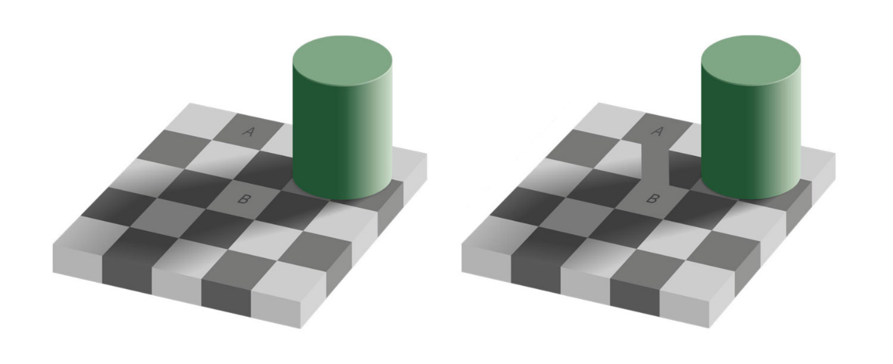
the noisy channel model in ML

noisy channels in ML



credit: MNIST dataset

we are inherently bayesian



credit: quantamagazine.org, 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

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

- basic probability review, and introducing information measures
- the source and channel coding theorems

what we hope to cover

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

- basic probability review, and introducing information measures
- the source and channel coding theorems
- bayesian inference: priors, bayesian update, model selection
- bayesian classification and regression

what we hope to cover

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

- basic probability review, and introducing information measures
- the source and channel coding theorems
- bayesian inference: priors, bayesian update, model selection
- bayesian classification and regression
- complex models: gaussian processes, artificial neural networks
- graphical models and markov random fields

what we hope to cover

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

- basic probability review, and introducing information measures
- the source and channel coding theorems
- bayesian inference: priors, bayesian update, model selection
- bayesian classification and regression
- complex models: gaussian processes, artificial neural networks
- graphical models and markov random fields
- approximate inference: MCMC

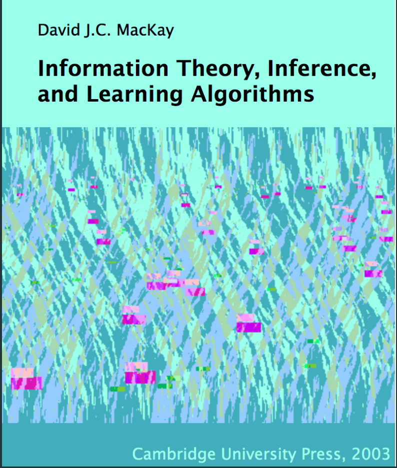
what we hope to cover

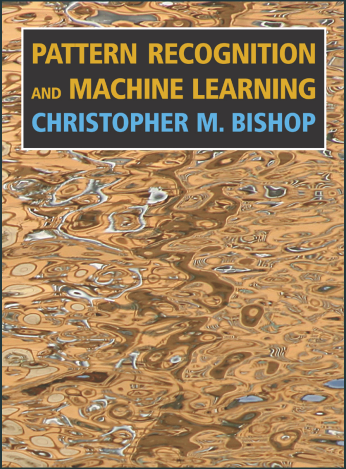
– bayesian inference: unified paradigm for learning and decision-making
– information theory: tool for designing and understanding data systems

- basic probability review, and introducing information measures
- the source and channel coding theorems
- bayesian inference: priors, bayesian update, model selection
- bayesian classification and regression
- complex models: gaussian processes, artificial neural networks
- graphical models and markov random fields
- approximate inference: MCMC
- model-based decision-making: bayesian optimization, causal inference, sequential decision-making and reinforcement learning

aids in learning

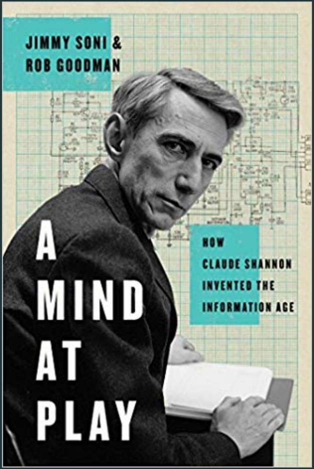
the following books are excellent references for most topics in the course

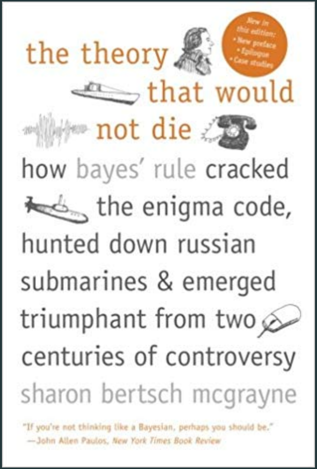




aids in getting excited about learning

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?

- 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'