Probability Overview

Machine Learning 10-601B

Many of these slides are derived from Tom Mitchell, William Cohen, Eric Xing. Thanks!

Course Overview

- Course website: http://www.cs.cmu.edu/~10601b
  - Lecture notes, recitation notes will be posted on this website
- All homework/project submissions should be uploaded to autolab folder
- Piazza for discussion

What do you need to know now?

- There are pre-requisites (see the course website), though not strictly enforced
- But… you should know how to do math and how to program:
  - Calculus (multivariate)
  - Probability/statistics
  - Linear algebra (matrices and vectors)
  - Programming:
    - You will implement some of the algorithms and apply them to datasets
    - Assignments will be mostly in Matlab and/or Octave (play with that now if you want)
    - All CMU students can download Matlab free of charge from CMU software website. Octave is open-source software.
- We may review these things but we will not teach them

What do you need to know now?

- In the first recitation, TA will review linear algebra and probability
- There is a “self-assessment” test on the class website
  - Won’t be graded
  - Everyone should take it to calibrate your prior knowledge

Grading

- Six homework assignments (60%)
  - Programming assignment (Matlab), written homework
  - Should be submitted
    - by 10:30am of the due date. (Two late days. 50% of the full grade for one-day late homework, 0 afterwards.)
    - To the autolab
      - Autolab website: https://autolab.cs.cmu.edu/courses/10601b-f15/assessments
- Project (20%)
  - Project proposal: Oct 22
  - Mid-report: Nov 24
  - Final report: Dec 17
  - Should be submitted by 10:30am of due date to autolab folder
  - No late days!! 50% of the full grade for one-day late submission, 0 afterwards
- Exam (20%): Nov 19 in class
- Pass/Fail: you should obtain at least B- to pass the course.
- Auditing: Should turn in at least 3 homework.

Collaboration Policy

- Discussion with fellow classmates are allowed, but only to *understand* better, not to save work.

- So:
  - *no notes* of the discussion are allowed to share
  - you should acknowledge who you got help from/did help in your homework (see the course website)

- This policy was also used previously in 10-601 taught by Roni Rosenfeld, William Cohen, and Eric Xing.

- We will take academic honesty seriously -- we will fail students.

Recitations and Office Hours

- Instructor's office hour: 10:30-11:30am Thursday
- TA office hours: location to be announced. Until then, the 8th floor common area
  - 5-6pm Monday
  - 11am-12pm Tuesday
  - 5-6pm Wednesday
- Recitations: 7:30-8:30pm Thursday, location to be announced

Main Topics for 10-601

- Supervised learning
  - Classifiers
    - Naïve Bayes, logistic regression, etc.
    - Extremely useful on many real tasks
  - Non-linear classifiers
    - Neural nets, decision trees, nearest-neighbor classifiers
  - Regression
- Unsupervised and semi-supervised learning
  - k-means, mixtures, SVD/PCA, …
- Graphical models
  - Bayes networks and Markov networks
  - Hidden Markov models
- Comparing and evaluating classifiers
  - Overfitting, cross validation, bias-variance trade off
  - Learning theory

Machine Learning:

Study of algorithms that

- improve their performance P
- at some task T
- with experience E

well-defined learning task: <P,T,E>

Learning to Predict Emergency C-Sections

[Sims et al., 2000]

Data:

Patient103 time=1	*Patient103* time=2	*Patient103* time=n
Age: 23	Age: 23	Age: 23
FirstPregnancy: no	FirstPregnancy: no	FirstPregnancy: no
Anemia: no	Anemia: no	Anemia: no
Diabetes: no	Diabetes: YES	Diabetes: no
PreviousPrematureBirth: no	PreviousPrematureBirth: no	PreviousPrematureBirth: no
Ultrasound: ?	Ultrasound: abnormal	Ultrasound: ?
Elective C–Section: ?	Elective C–Section: no	Elective C–Section: no
Emergency C–Section: ?	Emergency C–Section: ?	**Emergency C–Section: Yes**
...	...	...

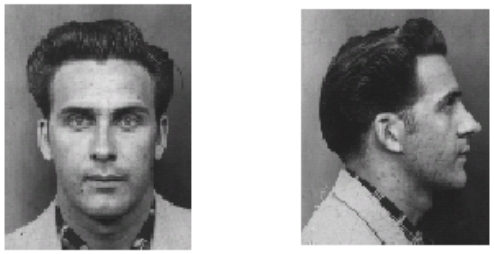
9714 patient records, each with 215 features

One of 18 learned rules:

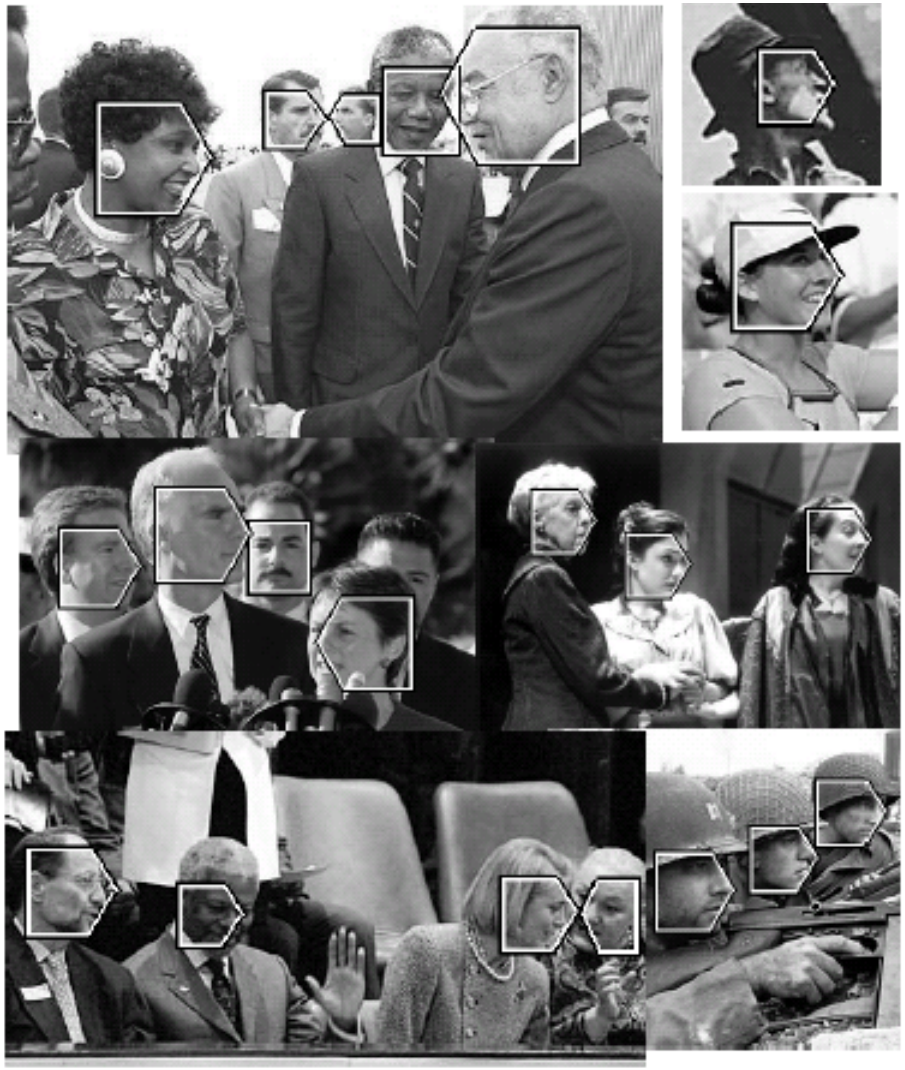
```
If   No previous vaginal delivery, and
     Abnormal 2nd Trimester Ultrasound, and
     Malpresentation at admission
Then Probability of Emergency C-Section is 0.6

 Over training data: 26/41 = .63,
 Over test data: 12/20 = .60
```

Learning to detect objects in images

(Prof. H. Schneiderman)



Example training images
for each orientation



Learning to classify text documents

→ Company home page

vs

Personal home page

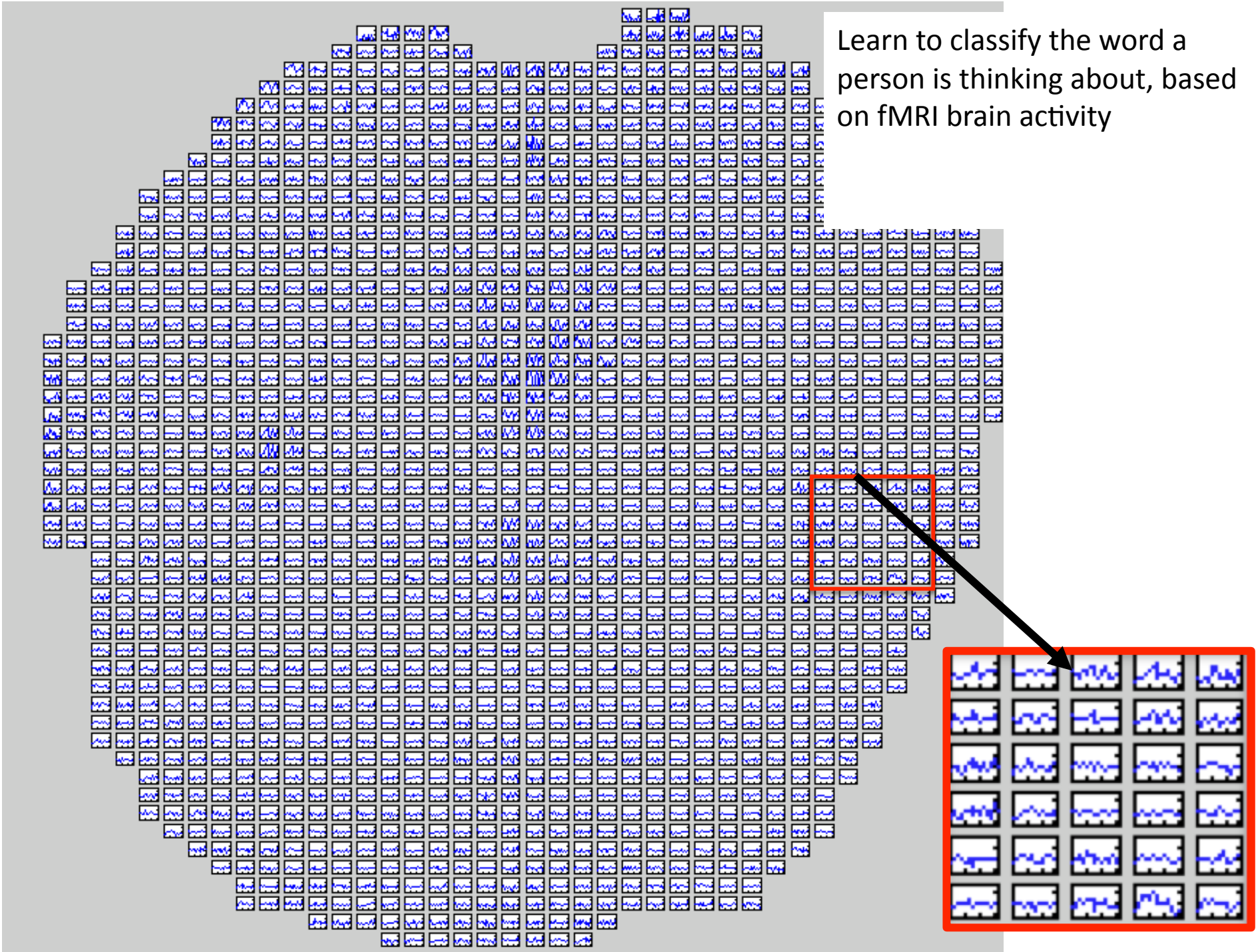
vs

University home page

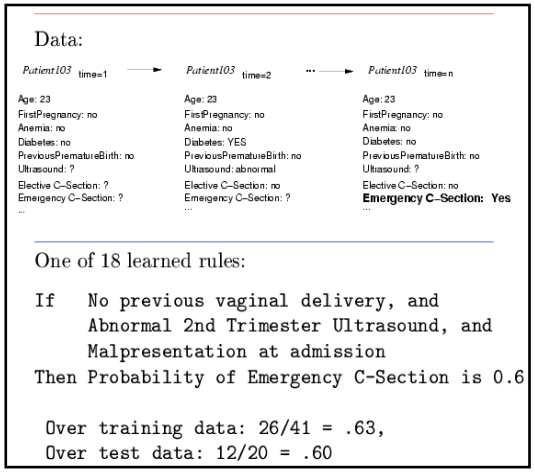
vs

…

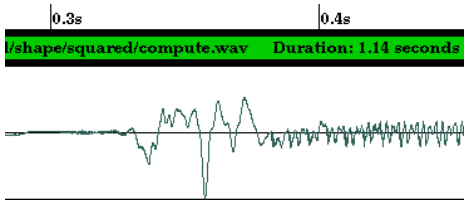
Learn to classify the word a person is thinking about, based on fMRI brain activity



Machine Learning - Practice

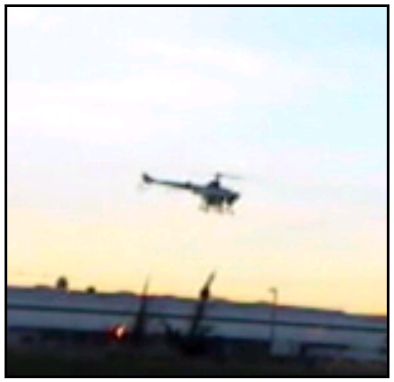


Mining Databases



Speech Recognition

Object recognition

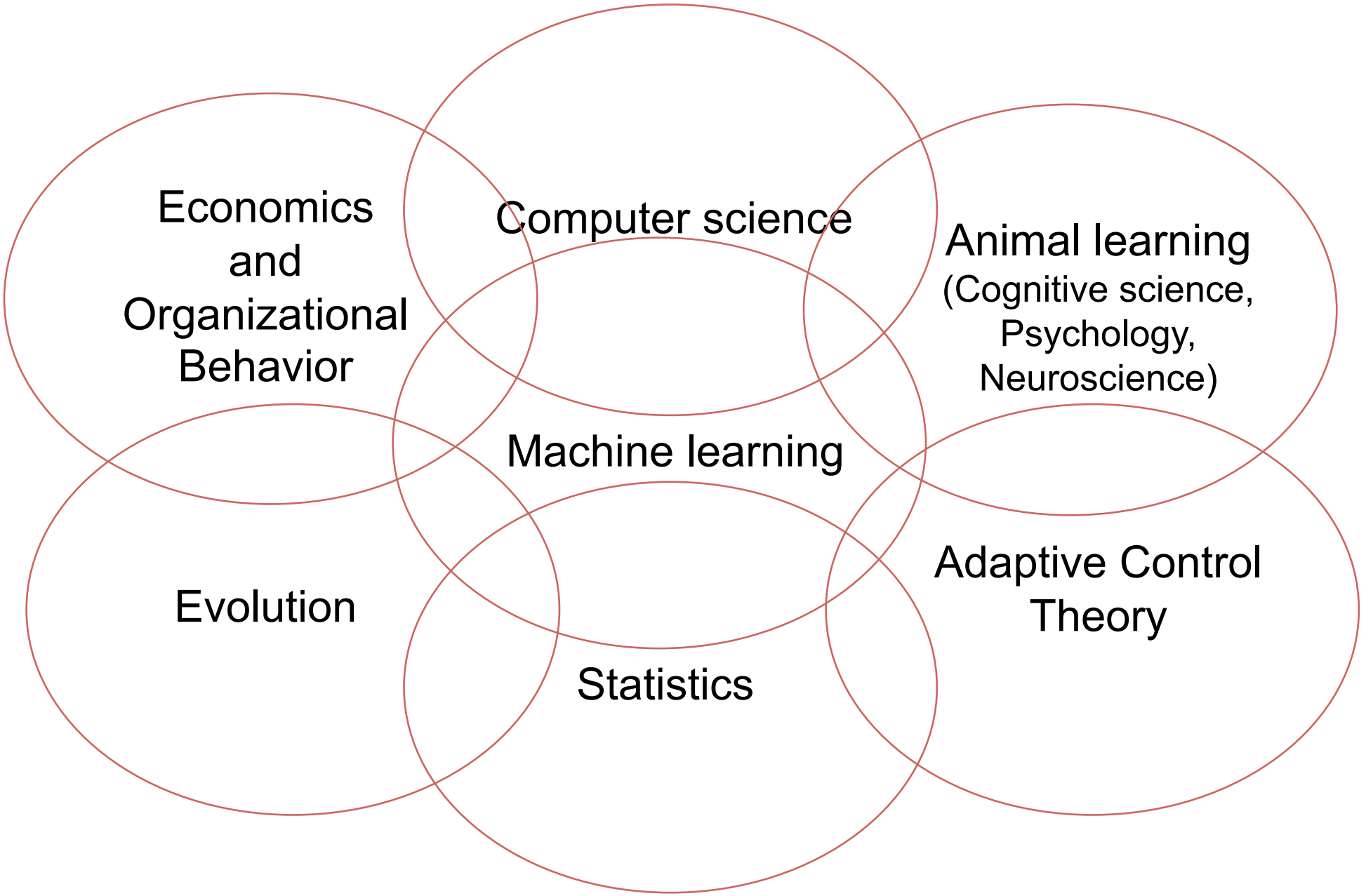


Control learning

Text analysis

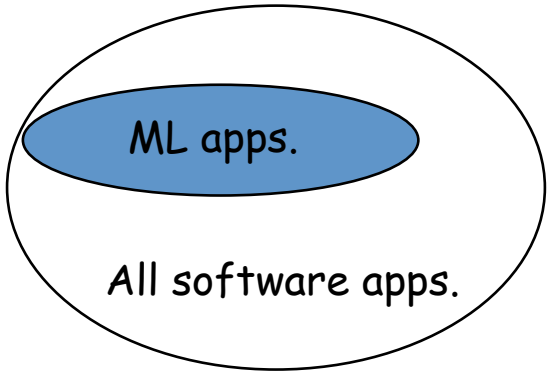
Peter H. van Oppen , **Chairman of the Board & Chief Executive Officer**
Mr. van Oppen has served as chairman of the board and chief executive officer of ADIC since its acquisition by Interpoint in 1994 and a director of ADIC since 1986. Until its acquisition by Crane Co. in October 1996, Mr. van Oppen served as chairman of the board of directors, president and chief executive officer of Interpoint . Prior to 1985, Mr. van Oppen worked as a consulting manager at Price Waterhouse LLP and at Bain & Company in Boston and London. He has additional experience in medical electronics and venture capital. Mr. van Oppen also serves as a director of Seattle FilmWorks Inc. and Spacelabs Medical, Inc.. He holds a B.A. from Whitman College and an M.B.A. from Harvard Business School, where he was a Baker Scholar.

- Supervised learning
- Bayesian networks
- Hidden Markov models
- Unsupervised clustering
- Reinforcement learning
- ....


Economics
and
Organizational
Behavior
Computer science
Animal learning
(Cognitive science,
Psychology,
Neuroscience)
Machine learning
Evolution
Adaptive Control
Theory
Statistics

Machine Learning in Computer Science

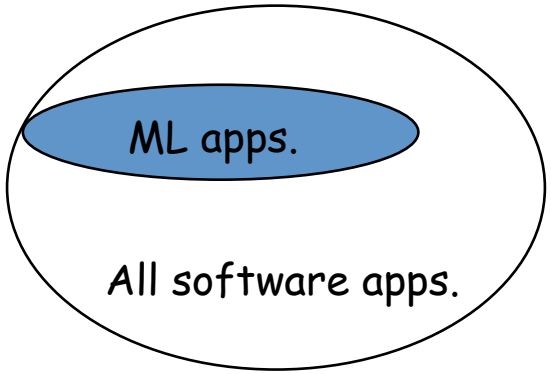
- Machine learning is already the preferred approach to
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - …



- This ML niche is growing (why?)

Machine Learning in Computer Science

- Machine learning is already the preferred approach to
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - …



- This ML niche is growing
  - Improved machine learning algorithms
  - Increased data capture, networking, new sensors
  - Demand for self-customization to user, environment

Probability Overview

- Events
  - discrete random variables, continuous random variables, compound events
- Axioms of probability
  - What defines a reasonable theory of uncertainty
- Independent events
- Conditional probabilities
- Independence, Conditional independence
- Bayes rule and beliefs

Random Variables

- Informally, A is a random variable if
  - A denotes something about which we are uncertain
  - perhaps the outcome of a randomized experiment

- Examples
  - A = True if a randomly drawn person from our class is female
  - A = The hometown of a randomly drawn person from our class
  - A = True if two randomly drawn persons from our class have same birthday

- Define P(A) as "the fraction of possible worlds in which A is true" or "the fraction of times A holds, in repeated runs of the random experiment"
  - the set of possible worlds is called the sample space, S

A little formalism

More formally, we have

- a sample space S (e.g., set of students in our class)
  - aka the set of possible worlds

- a random variable is a function defined over the sample space
  - Gender: S → { m, f }
  - Height: S → Reals
- an event is a subset of S
  - e.g., the subset of S for which Gender=f
  - e.g., the subset of S for which (Gender=m) AND (eyeColor=blue)
- We are often interested in probabilities of specific events and of specific events conditioned on other specific events

The Axioms of Probability

- Assume binary random variables A and B.
  - 0 <= P(A) <= 1
  - P(A=true) + P(A=false) = 1
  - P(A or B) = P(A) + P(B) – P(A and B)

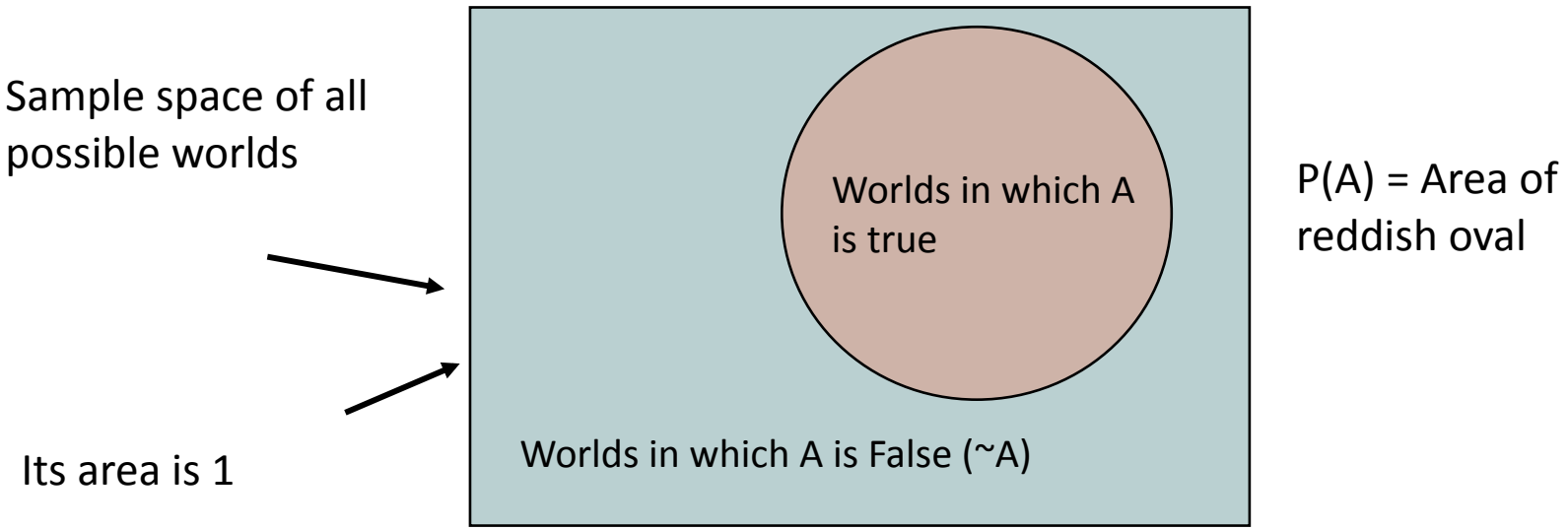
[di Finetti 1931]:

when gambling based on "uncertainty formalism A" you can be exploited by an opponent

iff

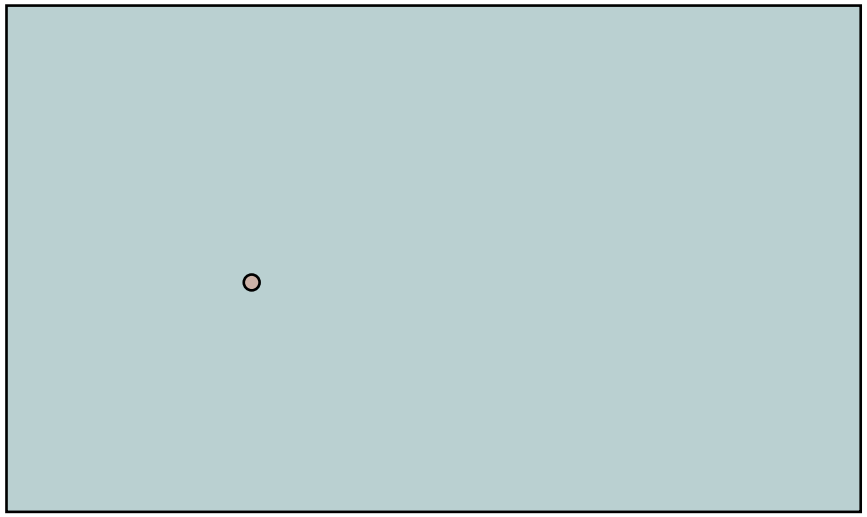
your uncertainty formalism A violates these axioms

Visualizing Probability Axioms



Interpreting the axioms

- 0 <= P(A) <= 1
- P(A or ~A) = P(True) = 1
- P(A or B) = P(A) + P(B) - P(A and B)

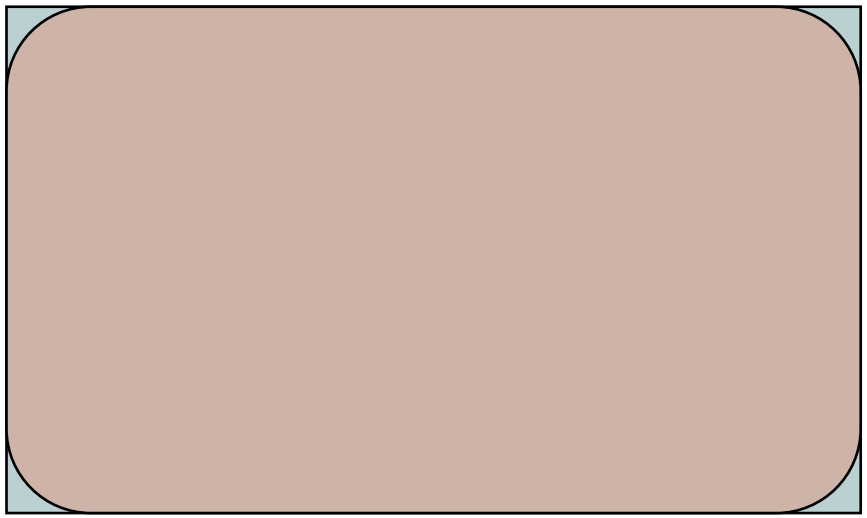


The area of A can't get any smaller than 0

And a zero area would mean no world could ever have A true

Interpreting the axioms

- 0 <= P(A) <= 1
- P(A or ~A) = P(True) = 1
- P(A or B) = P(A) + P(B) - P(A and B)

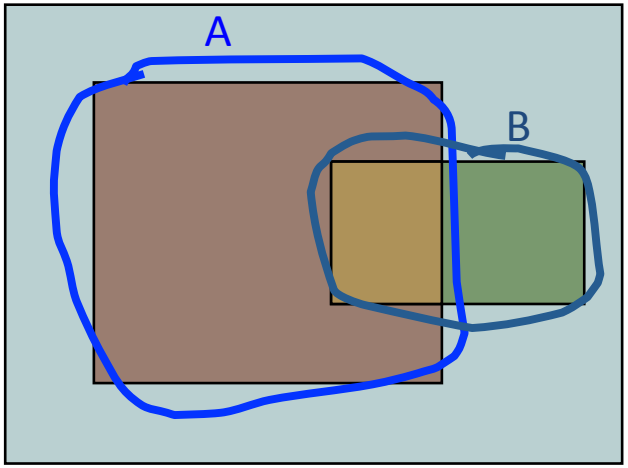


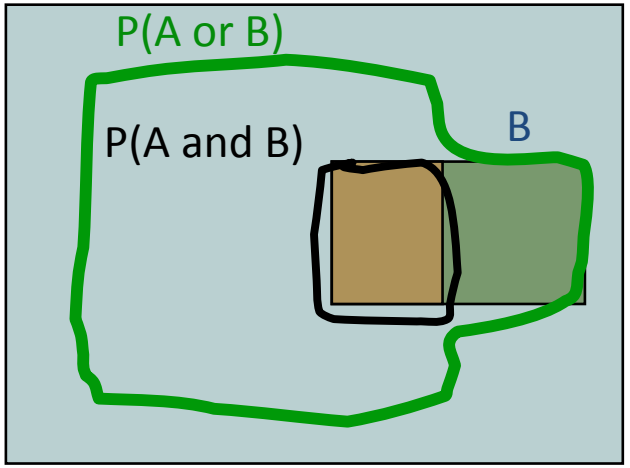
The area of A can't get any bigger than 1

And an area of 1 would mean all worlds will have A true

Interpreting the axioms

- 0 <= P(A) <= 1
- P(A or ~A) = P(True) = 1
- P(A or B) = P(A) + P(B) - P(A and B)





Simple addition and subtraction

Interpreting the axioms

- 0 <= P(A) <= 1
- P(A or ~A) = P(True) = 1
- P(A or B) = P(A) + P(B) - P(A and B)

Monotonicity: if A is a subset of B, then P(A) <= P(B)

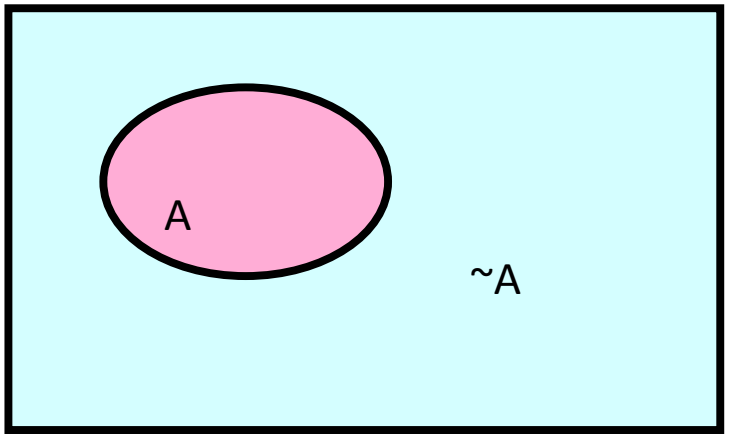
Proof:

- A subset of B ➔ B = A + C for C=B-A
- A and C are disjoint ➔ P(B) = P(A or C)=P(A) + P(C)
- P(C) >= 0

- So P(B) >= P(A)

Interpreting the axioms

- 0 <= P(A) <= 1
- P(A or ~A) = P(True) = 1
- P(A or B) = P(A) + P(B) - P(A and B)



Theorem: P(~A) = 1 - P(A)

Proof:

- P(A or ~A) = P(True) = 1
- A and ~A are disjoint ➔ P(A or ~A)=P(A) + P(~A)
  ➔ P(A) + P(~A) = 1

  ….then solve for P(~A)

Another useful theorem

- 0 <= P(A) <= 1, P(True) = 1, P(False) = 0,
- P(A or B) = P(A) + P(B) - P(A and B)

➔ P(A) = P(A and B) + P(A and ~B)

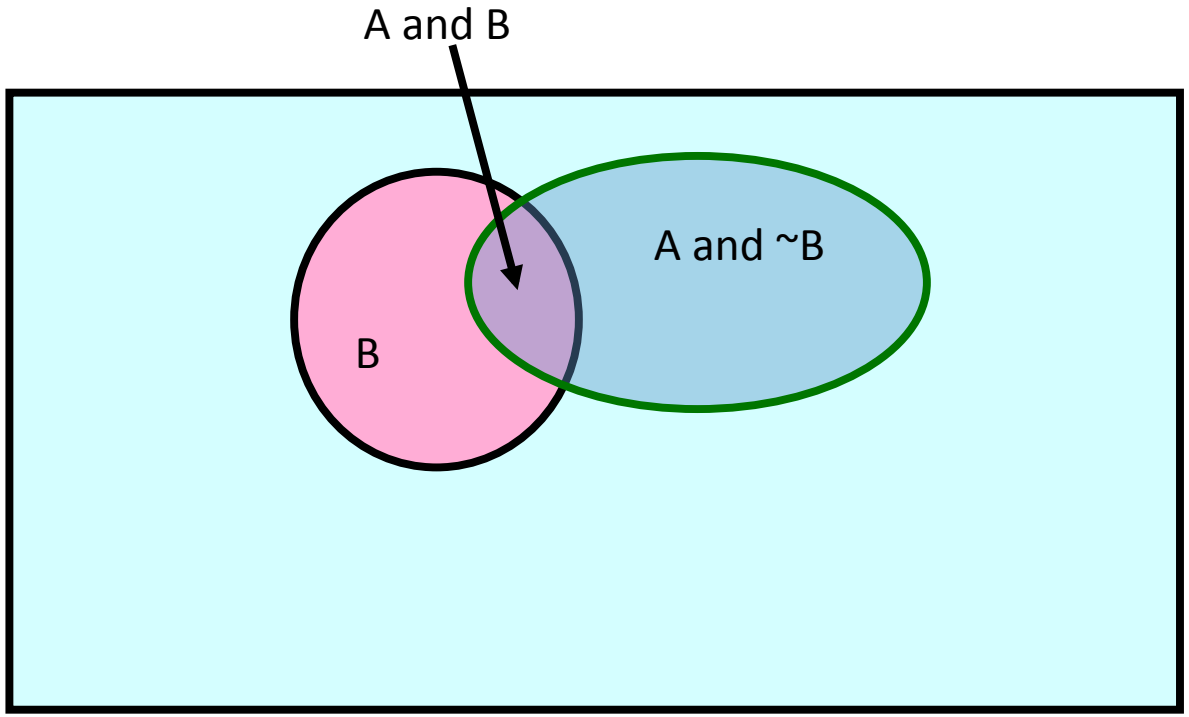
A = [A and (B or ~B)] = [(A and B) or (A and ~B)]

P(A) = P(A and B) + P(A and ~B) – P((A and B) and (A and ~B))

P(A) = P(A and B) + P(A and ~B) – P(A and B and A and ~B)

Elementary Probability in Pictures

- P(A) = P(A and B) + P(A and ~B)



Multivalued Discrete Random Variables

- Suppose A can take on more than 2 values
- A is a random variable with arity k if it can take on exactly one value out of $\{v_1, v_2, \ldots v_k\}$
  - Example: A={1,2,3….,20}: good for 20-sided dice games
  - Notation: let's write the event AHasValueOfv as "A=v"

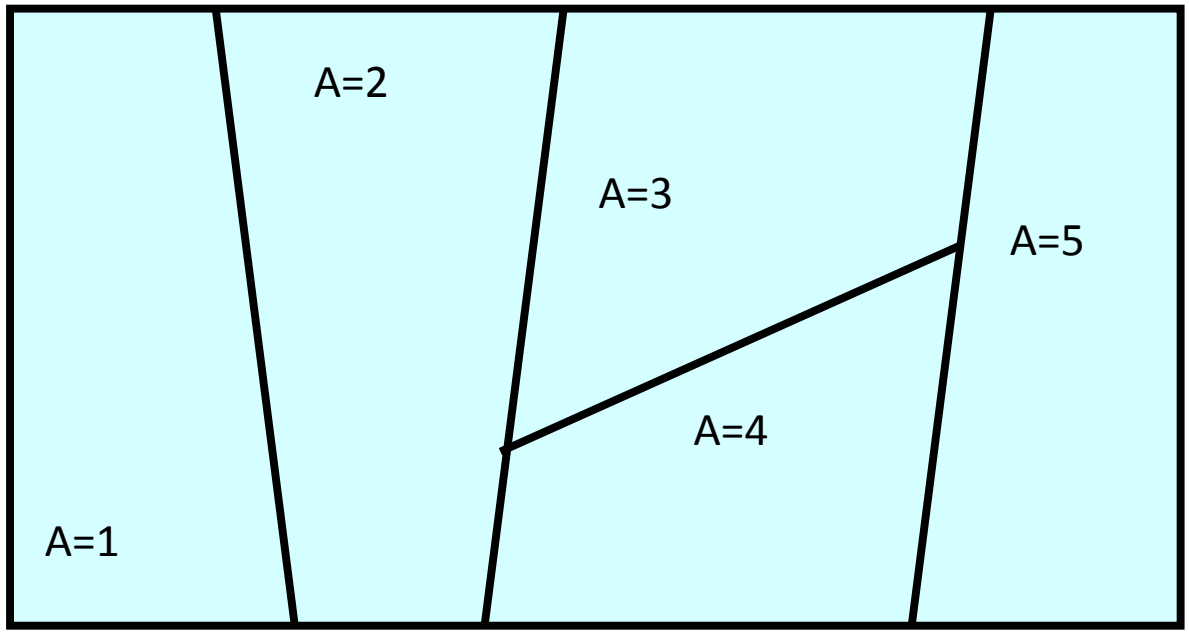
- Thus…

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$

$$P(A = v_1 \vee A = v_2 \vee \ldots \vee A = v_k) = 1$$

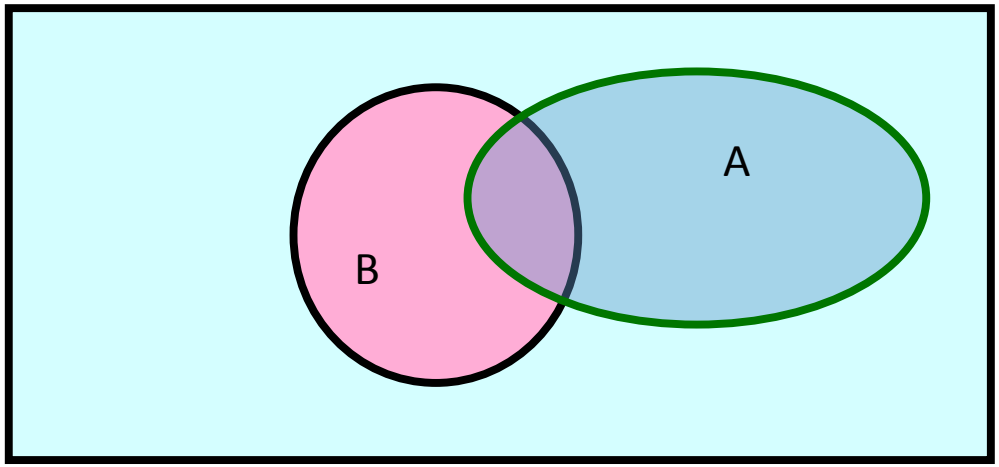
Elementary Probability in Pictures

$$\sum_{j=1}^{k} P(A = v_j) = 1$$



Definition of Conditional Probability

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$



Definition of Conditional Probability

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

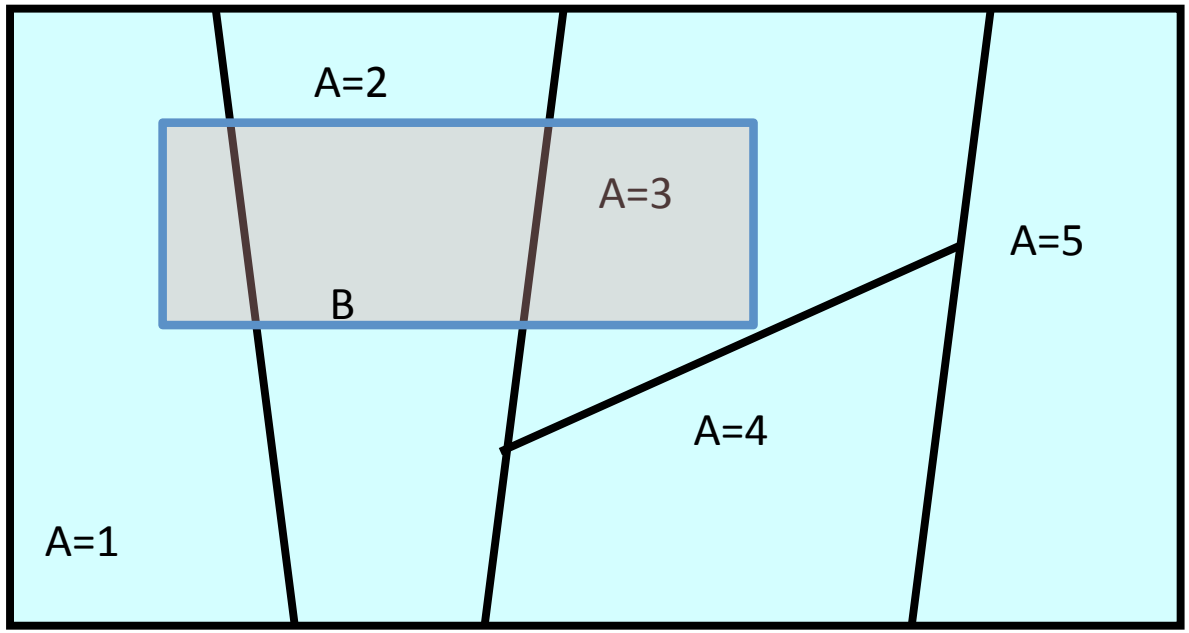
Corollary: The Chain Rule

$$P(A \wedge B) = P(A|B)\, P(B)$$

$$P(A \wedge B \wedge C) = P(A|B \wedge C)\, P(B|C)\, P(C)$$

Conditional Probability in Pictures

picture: P(B=true|A=2)



Independent Events

- Definition: two events A and B are *independent* if P(A and B)=P(A)*P(B)
- Intuition: knowing A tells us nothing about the value of B (and vice versa)
- From chain rule

$$P(A \wedge B) = P(A|B)\,P(B) = P(A)P(B)$$

$$-> P(A|B) = P(A)$$

- You frequently need to assume the independence of *something* to solve any learning problem.

Continuous Random Variables

- The discrete case: sum over all values of A is 1

$$\sum_{j=1}^{k} P(A = v_j) = 1$$

- The continuous case: infinitely many values for A and the *integral* is 1

$$\int_{-\infty}^{\infty} f_P(x)\,dx = 1$$

f(x) is a probability density function (pdf)

1. 0<=P(A) <= 1
2. P(True) = 1
3. P(A or B) = P(A) + P(B) - P(A and B)

also….

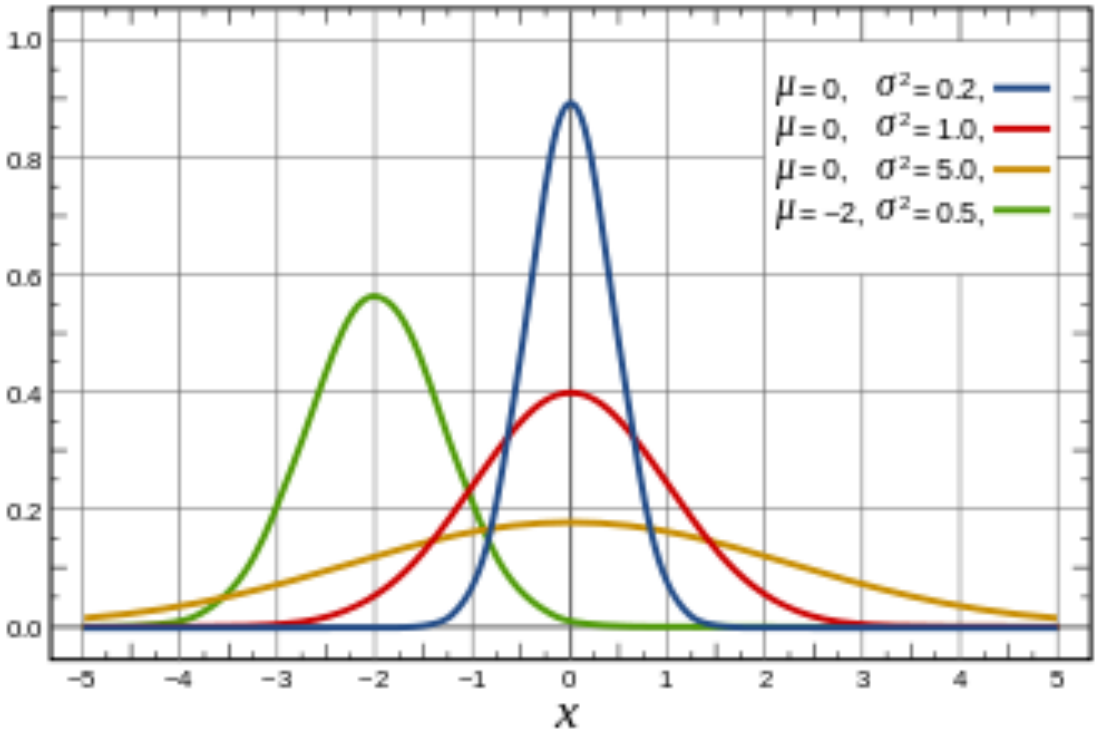
$$\forall x, f_P(x) \geq 0$$

Continuous Random Variables

Gaussian probability density with parameters

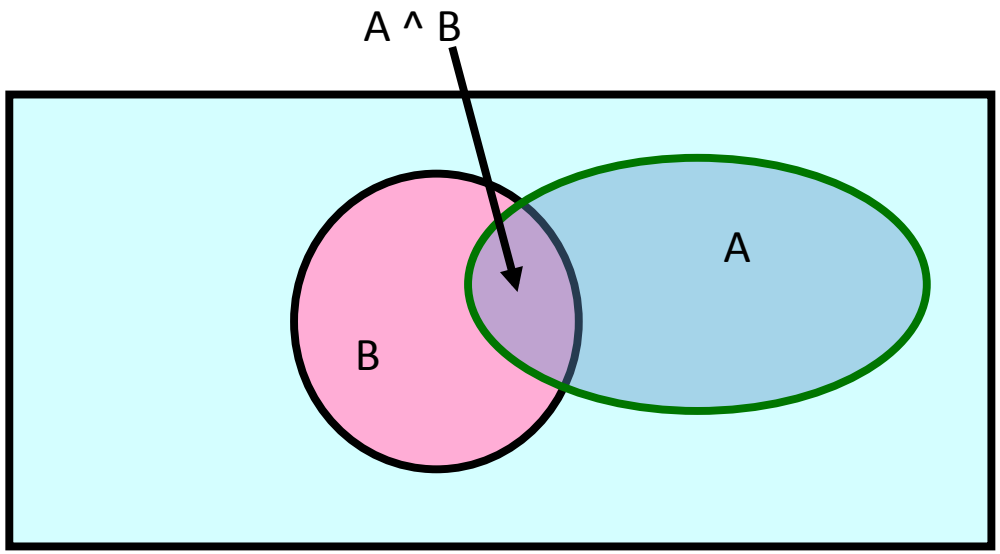
- mean μ
- standard deviation σ

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$



Bayes Rule

- let's write two expressions for P(A ^ B)



$$P(A \wedge B) = P(A|B)\,P(B)$$

$$P(A \wedge B) = P(B|A)P(A)$$

$$P(A|B)\,P(B) = P(B|A)P(A)$$

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$ Bayes' rule

we call P(A) the "prior"

and P(A|B) the "posterior"

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning…*

Other Forms of Bayes Rule

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \sim A)P(\sim A)}$$

$$P(A \mid B \wedge X) = \frac{P(B \mid A \wedge X)P(A \wedge X)}{P(B \wedge X)}$$

Applying Bayes Rule

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \sim A)P(\sim A)}$$

A = you have the flu, B = you just coughed

Assume:
P(A) = 0.05

Also assume the following information is known to you
P(B|A) = 0.80
P(B| ~A) = 0.2

what is P(flu | cough) = P(A|B)?

Bayes Rule in Machine Learning

- D: data (evidence)
- θ: unknown quantities
  - e.g., model parameters, predictions

likelihood: How likely is the observed data under the particular unknown quantity θ

posterior belief on the unknown quantity **after** you see data D

Prior belief on the unknown quantity **Before** you see data D

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

You should know

- Events
  - discrete random variables, continuous random variables, compound events
- Axioms of probability
  - What defines a reasonable theory of uncertainty
- Independent events
- Conditional probabilities
- Bayes rule and beliefs