10-601

Machine Learning

http://www.cs.cmu.edu/afs/cs/academic/class/10601-f10/index.html

Course data

- All up-to-date info is on the course web page:

http://www.cs.cmu.edu/afs/cs/academic/class/10601-f10/index.html

- Instructor:
  - Ziv Bar-Joseph
- TAs:
  - Anthony Gitter
  - Hai-Son Le
  - Yang Xu
- See web page for contact info, office hours, etc.
- Mailing list - send email to: yx1@cs.cmu.edu
- Follow us on Twitter: ML10601

- 8/23/10 – Intro to ML and probablity
- 8/25/10 – Density estimation, classification theory
- 8/30/10 – Classification
- 09/01/10 – Bayes and Naïve Bayes classifiers / PS1

9/8 (Wednesday): No class (Matlab recitation)

- 09/15/10 – SVM 1 / PS1 in PS2 out
- 09/20/10 – SVM 2
- 09/22/10 – Boosting
- 09/27/10 – Learning theory 1
- 09/29/10 – Learning theory 2 / PS2 in PS3 out

Intro and classification (A.K.A. 'supervised learning')

- 10/04/10 – Decision trees
- 10/06/10 – NN
- 10/11/10 - Hierarchical clustering
- 10/13/10 – Kmeans and Gaussian mixtures / PS3 in PS4 out
- 10/20/10 – Model selection, feature selection
- 10/18/10 – semi supervised learning / Project proposal

Clustering ('Unsupervised learning')

- 10/25/10 – ML in industry 1
- 10/27/10 – ML in industry 2 / PS4 in

11/3 (Wednesday): Midterm (1:30-3:30)

- 11/10/10 – HMM2 structure learning / Project progress report
- 11/15/10 – MDPs
- 11/17/10 – PCA, SVD / PS5 in

Probabilistic representation and modeling ('reasoning under uncertainty')

- 11/22/10 – Graph clustering? RL?
- 11/24/10 – Thanksgiving

12/1 (Wednesday): Poster session in the afternoon (class as usual)

Applications of ML

Grading

- **5 Problem sets - 40%**
- **Project - 30%**
- **Midterm - 25%**
- **Class participation - 5%**

Class assignments

- 5 Problem sets

  - Each containing both theoretical and programming assignments

- Projects

  - Groups of 1 or 2

  - Implement and apply an algorithm discussed in class to a new domain

  - Extend algorithms discussed in class in various directions

  - New theoretical results (for example, for a new setting of a problem)

  - More information on website

- Recitations

  - Monday, 5-6:20pm, NSH 1305

  - Expand on material learned in class, go over problems from previous classes etc.

What is Machine Learning?

Easy part: Machine

Hard part: Learning

- Short answer: Methods that can help generalize information in observed data so that it can be used to make better decisions in the future

What is Machine Learning?

Longer answer: The term Machine Learning is used to characterize a number of different approaches for generalizing from observed data:

- Supervised learning
  - Given a set of features and labels learn a model that will predict a label to a new feature set
- Unsupervised learning
  - Discover patterns in data
- Reasoning under uncertainty
  - Determine a model of the world either from samples or as you go along
- Active learning
  - Select not only model but also which examples to use

Paradigms of ML

- Supervised learning

  - Given $D = \{X_i, Y_i\}$ learn a model (or function) $F: X_k \rightarrow Y_k$

- Unsupervised learning

  Given $D = \{X_i\}$ group the data into Y classes using a model (or function) $F: X_i \rightarrow Y_j$

- Reinforcement learning (reasoning under uncertainty)

  Given D = {environment, actions, rewards} learn a policy and utility functions:

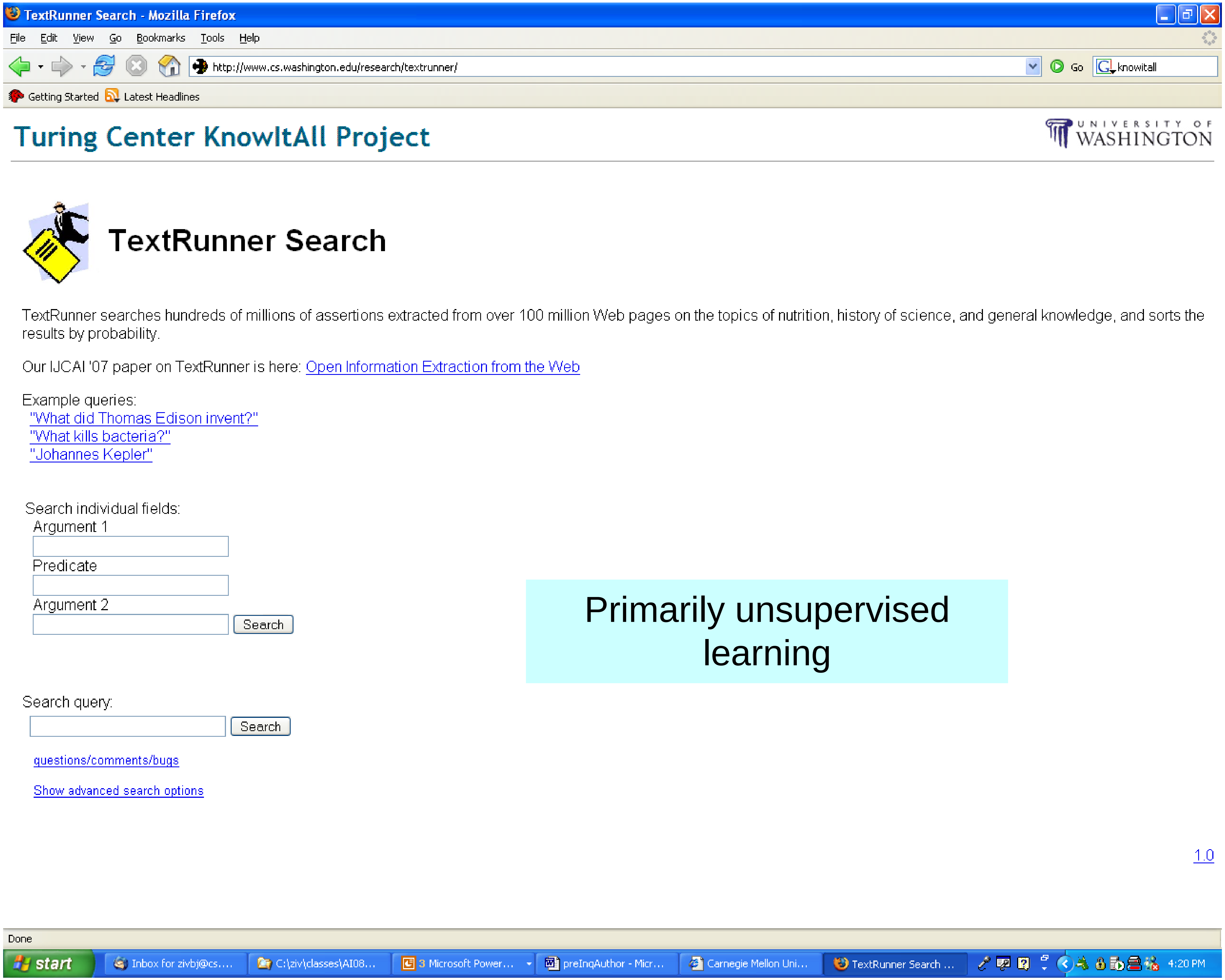
  policy: $F1: \{e,r\} \rightarrow a$

  utility: $F2: \{a,e\} \rightarrow R$

- Active learning

  - Given $D = \{X_i, Y_i\}$, $\{X_j\}$ learn a function $F1 : \{X_j\} \rightarrow x_k$ *to* maximize the success of the supervised learning function $F2: \{X_i , x_k\} \rightarrow Y$

Web search

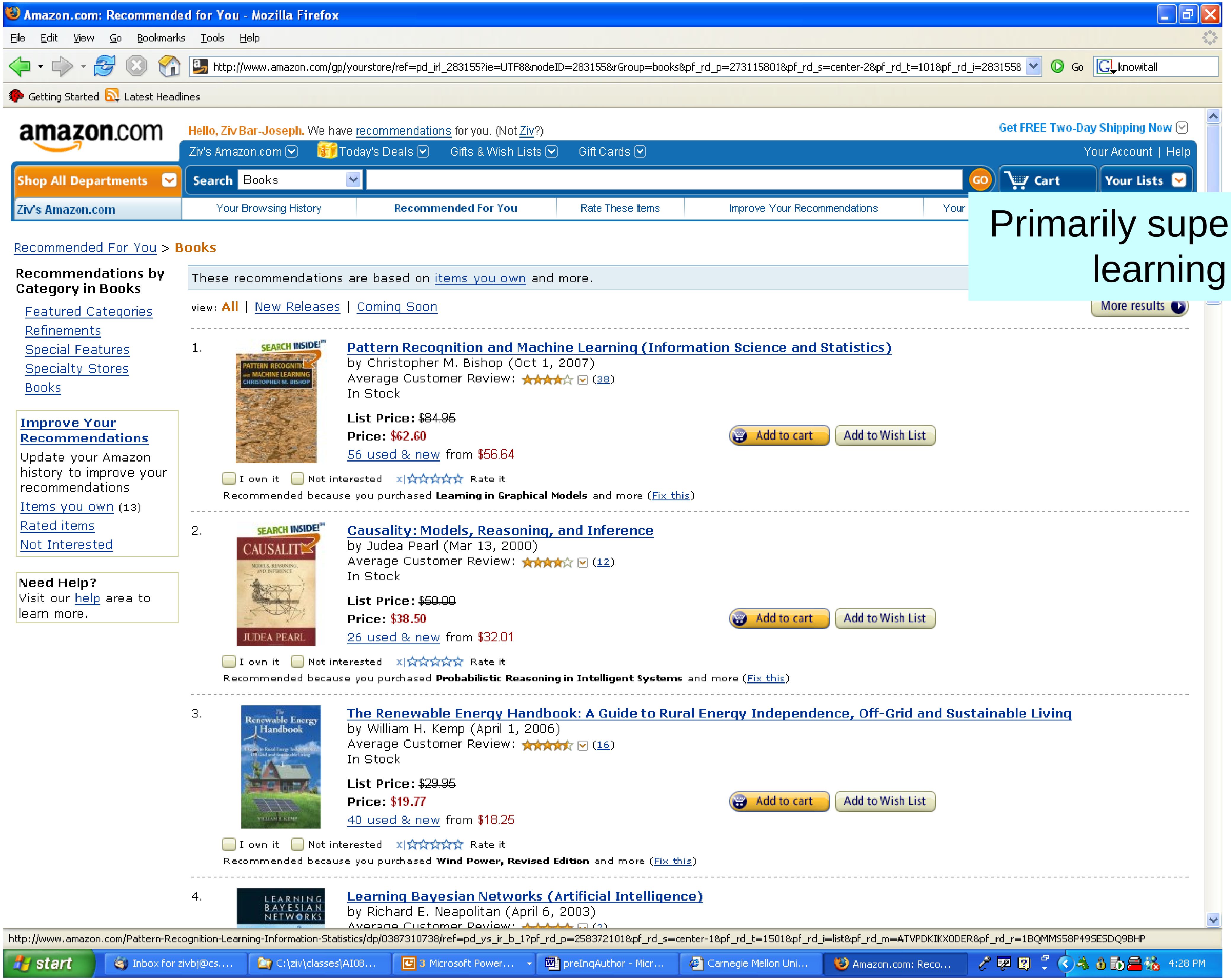


Primarily unsupervised learning

Web search, cont'd

Primarily supervised learning

Recommender systems



Primarily supervised learning

Grand and Urban Challenges road race

Supervised and
reinforcement learning

Helicopter control

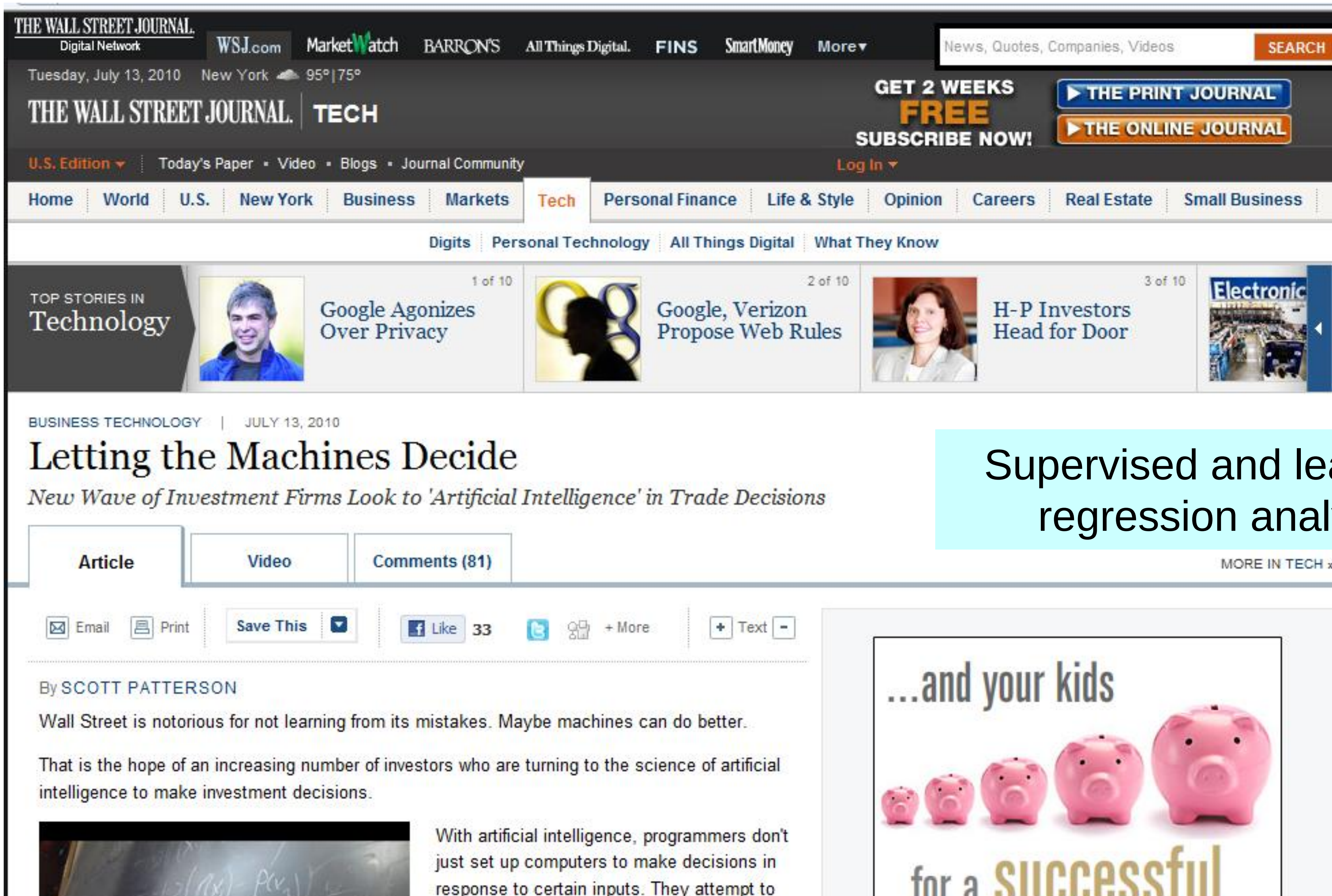
Reinforcement learning

Biology

ACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAATTCGATAGCAATTC
GATAACGCTGAGCAATCGGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAACG
CTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATCGGATATCGATAGCAATTCGATAAATC
GGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAATTCGATAGC
AATTCGATAACGCTGAGCAATCGGATATCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCA
ATTCGATAGCAATTCGATAACGCTGAGCAATCGGATAACGCTGAGCAATTCGATAGCATTCGAT
AACGCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATCGGATAACGCTG
AGCAATTCGATAGCAATTCGATAACGCTGAGCTGAGCAATTCGATAGCAATTCGATAACGCTGA
GCAATCGGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAATTC
GATAGCAATTCGATAGCAATTCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAATTCGAT
AGCAATTCGATAACGCTGAGCAATCGGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCT
GAGCAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATCGGATATCGATAGCAATT
CGATAACGCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATCGGATAAC
GC … CGCTGAGCTGAGCAATTCGATAGCAATTCGATAACG
CT … CGATAGCAATTCGATAACGCTGAGCAACGCTGAGCA
AT … AATCGGATATCGATAGCAATTCGATAACGCTGAGCA
ACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATCGGATAACGCTGAGCAATTCGAT
AGCATTCGATAACGCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATCGGATAACGCTGAGC
AATTCGATAGCAATTCGATAACGCTGAGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCA
ATCGGATAACGCTGAGCAATTCGATAGCA … TGAGCAATTCGAT
AGCAATTCGATAACGCTGAGCAATCGGAT … GAGCAACGCTGA
GCAATTCGATAGCAATTCGATAACGCTGA … TTCGATAGCATTC
GATAACGCTGAGCAACGCTGAGCAATTCG … CAATCGGATAACG
CTGAGCAATTCGATAGCAATTCGATAACG … ATTCGATAACGC
TGAGCAATCGGATAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAA
TTCGATAGCAATTCGATAGCAATTCGATAGCAATTCGATAACGCTGAGCAACGCTGAGCAATTC
GATAGCAATTCGATAACGCTGAGCAATCGGATAACGCTGAGCAATTCGATAGCAATTCGATAAC
GCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATCGGATATCGATAGCA
ATTCGATAACGCTGAGCAACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCAATCGGAT
AACGCTGAGCAATTCGATAGCAATTCGATAACGCTGAGCTGAGCAATTCGATAGCAATTCGATA
ACGCTGAGCAATCGGA

Which part is the gene?

Supervised and unsupervised learning (can also use active learning)



Supervised and learning, regression analysis

Most investors trying the approach are using "machine learning," a branch of artificial intelligence in which a computer program analyzes huge chunks of data and makes predictions about the future.

Common Themes

- Mathematical framework

  - Well defined concepts based on explicit assumptions
- Representation

  - How do we encode text? Images?
- Model selection

  - Which model should we use? How complex should it be?
- Use of prior knowledge

  - How do we encode our beliefs? How much can we assume?

(brief) intro to probability

Basic notations

- Random variable

  - referring to an element / event whose status is unknown:

  A = “it will rain tomorrow”

- Domain (usually denoted by Ω)

  - The set of values a random variable can take:

  - “A = The stock market will go up this year”: Binary

  - “A = Number of Steelers wins in 2007”: Discrete

  - “A = % change in Google stock in 2007”: Continuous

Axioms of probability (Kolmogorov's axioms)

A variety of useful facts can be derived from just three axioms:

1. $0 \leq P(A) \leq 1$
2. $P(\text{true}) = 1$, $P(\text{false}) = 0$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

There have been several other attempts to provide a foundation for probability theory. Kolmogorov's axioms are the most widely used.

Using the axioms

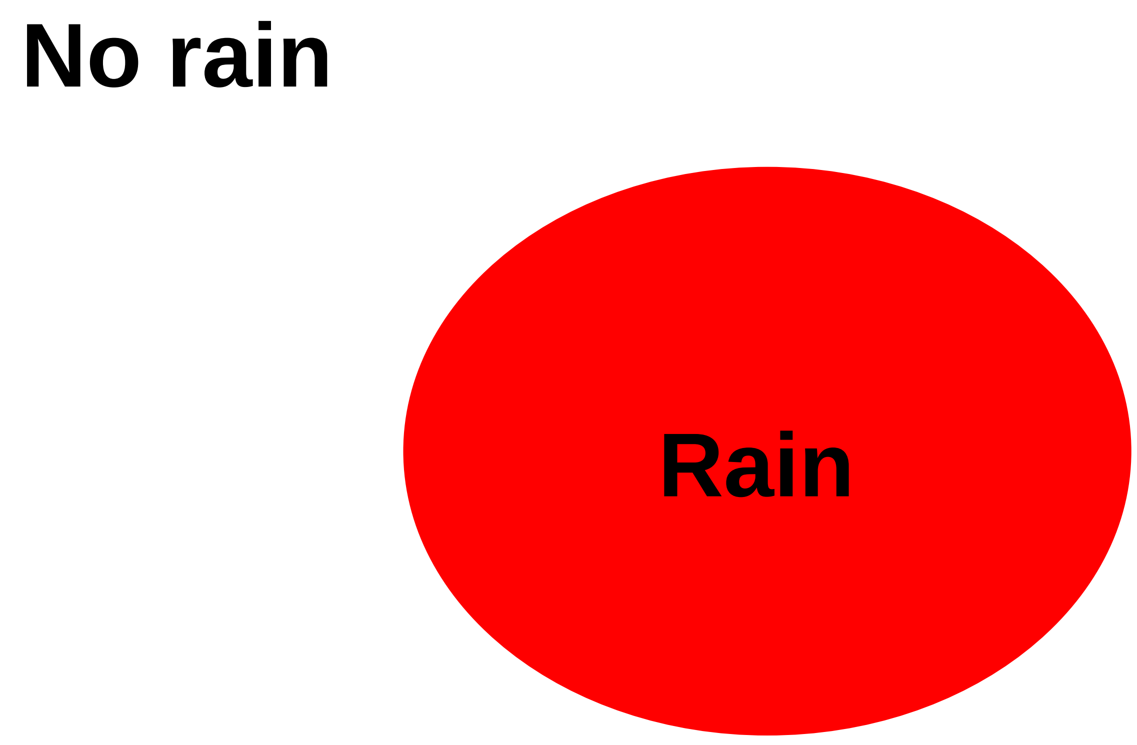
- How can we use the axioms to prove that:

  $P(\neg A) = 1 - P(A)$

  ?

Priors

Degree of belief in an event in the absence of any other information



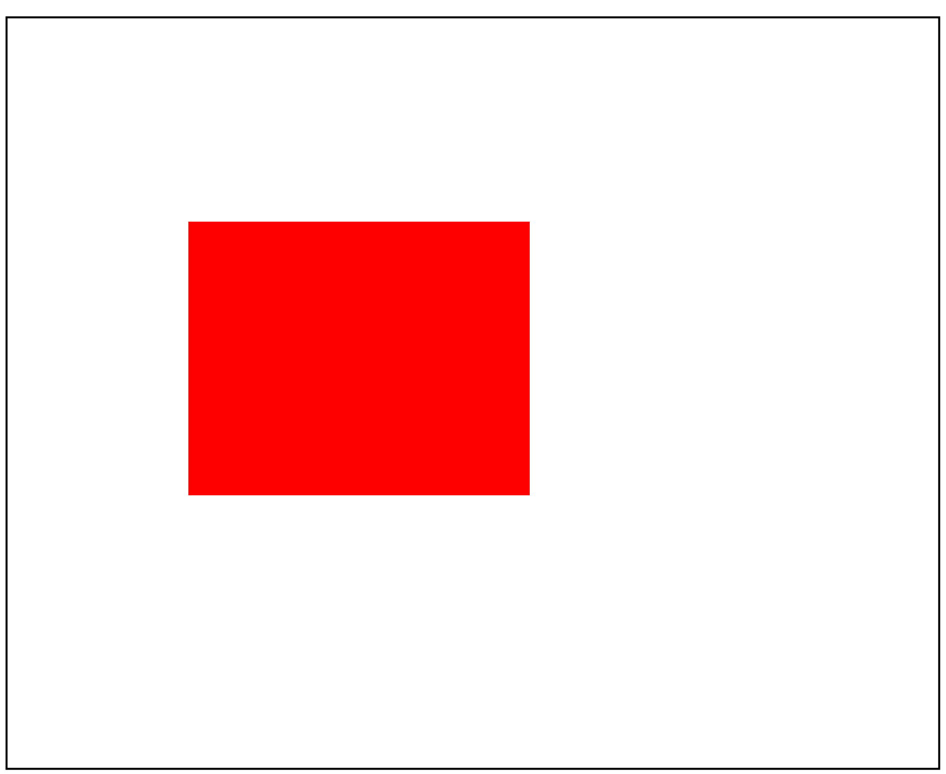
P(rain tomorrow) = 0.2

P(no rain tomorrow) = 0.8

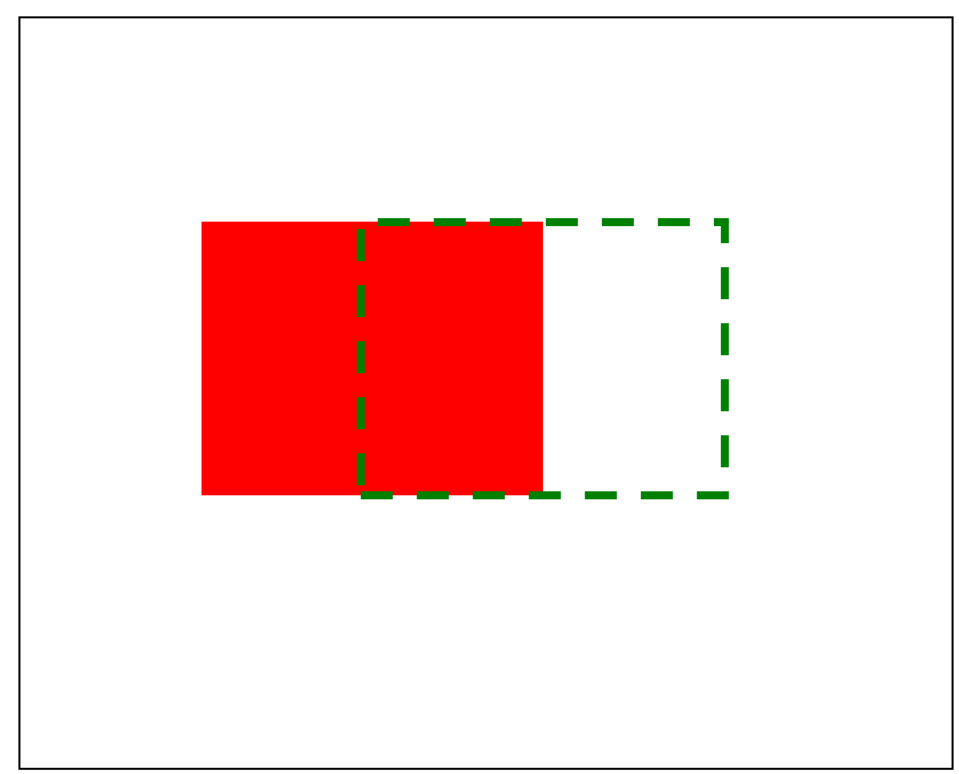
Conditional probability

- P(A = 1 | B = 1): The fraction of cases where A is true if B is true

P(A = 0.2)



P(A|B = 0.5)



Conditional probability

- In some cases, given knowledge of one or more random variables we can improve upon our prior belief of another random variable
- For example:

p(slept in movie) = 0.5

p(slept in movie | liked movie) = 1/4

p(didn't sleep in movie | liked movie) = 3/4

Slept	**Liked**
1	0
0	1
1	1
1	0
0	0
1	0
0	1
0	1

Joint distributions

- The probability that a set of random variables will take a specific value is their joint distribution.
- Notation: $P(A \wedge B)$ or $P(A,B)$
- Example: P(liked movie, slept)

If we assume independence then

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

However, in many cases such an assumption maybe too strong (more later in the class)

Joint distribution (cont)

P(class size > 20) = 0.6

P(summer) = 0.4

P(class size > 20, summer) = ?

Evaluation of classes

Size	Time	Eval
30	R	2
70	R	1
12	S	2
8	S	3
56	R	1
24	S	2
10	S	3
23	R	3
9	R	2
45	R	1

Joint distribution (cont)

P(class size > 20) = 0.6

P(summer) = 0.4

P(class size > 20, summer) = 0.1

Evaluation of classes

Size	Time	Eval
30	R	2
70	R	1
12	S	2
8	S	3
56	R	1
24	S	2
10	S	3
23	R	3
9	R	2
45	R	1

Joint distribution (cont)

P(class size > 20) = 0.6

P(eval = 1) = 0.3

P(class size > 20, eval = 1) = 0.3

Size	Time	Eval
30	R	2
70	R	1
12	S	2
8	S	3
56	R	1
24	S	2
10	S	3
23	R	3
9	R	2
45	R	1

Joint distribution (cont)

P(class size > 20) = 0.6

P(eval = 1) = 0.3

P(class size > 20, eval = 1) = 0.3

Evaluation of classes

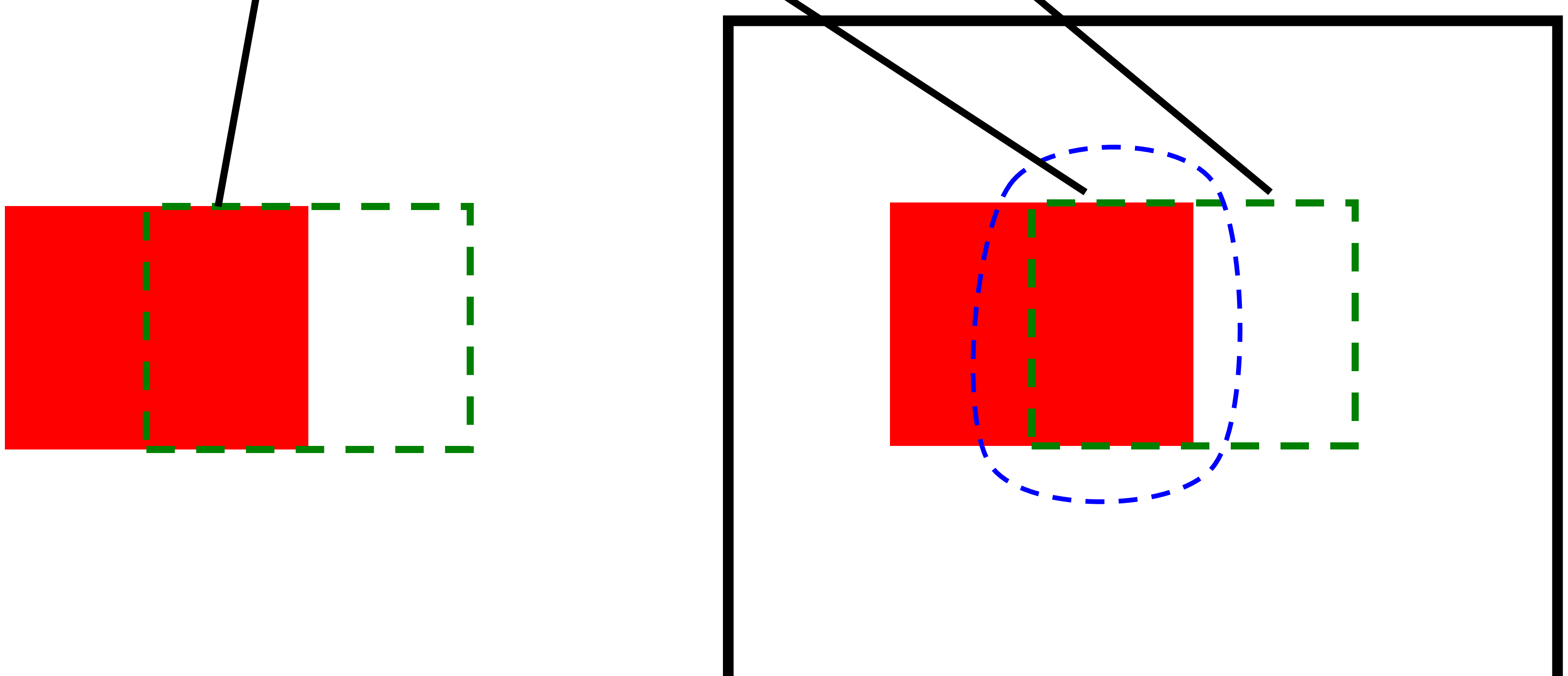
Size	Time	Eval
30	R	2
70	R	1
12	S	2
8	S	3
56	R	1
24	S	2
10	S	3
23	R	3
9	R	2
45	R	1

Chain rule

- The joint distribution can be specified in terms of conditional probability:

  P(A,B) = P(A|B)*P(B)

- Together with Bayes rule (which is actually derived from it) this is one of the most powerful rules in probabilistic reasoning



Bayes rule

- One of the most important rules for AI usage.
- Derived from the chain rule:

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

- Thus,

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

Thomas Bayes was an English clergyman who set out his theory of probability in 1764.

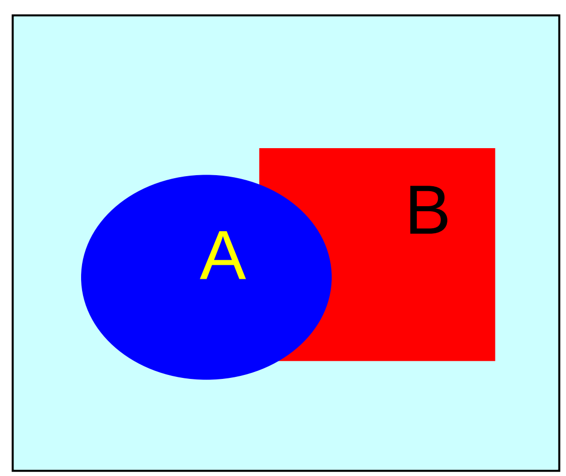
Bayes rule (cont)

Often it would be useful to derive the rule a bit further:

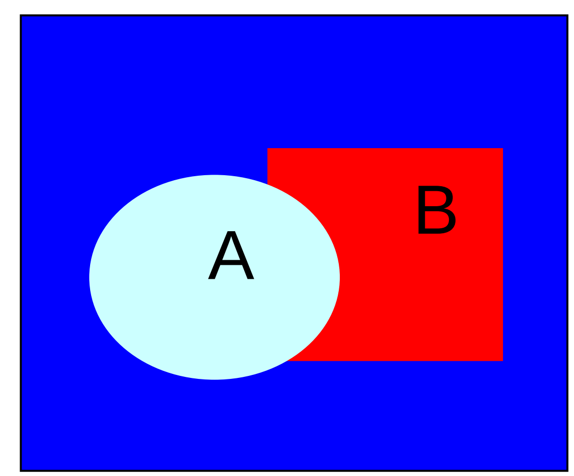
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_A P(B|A)P(A)}$$

This results from:
$P(B) = \sum_A P(B,A)$

P(B,A=1)

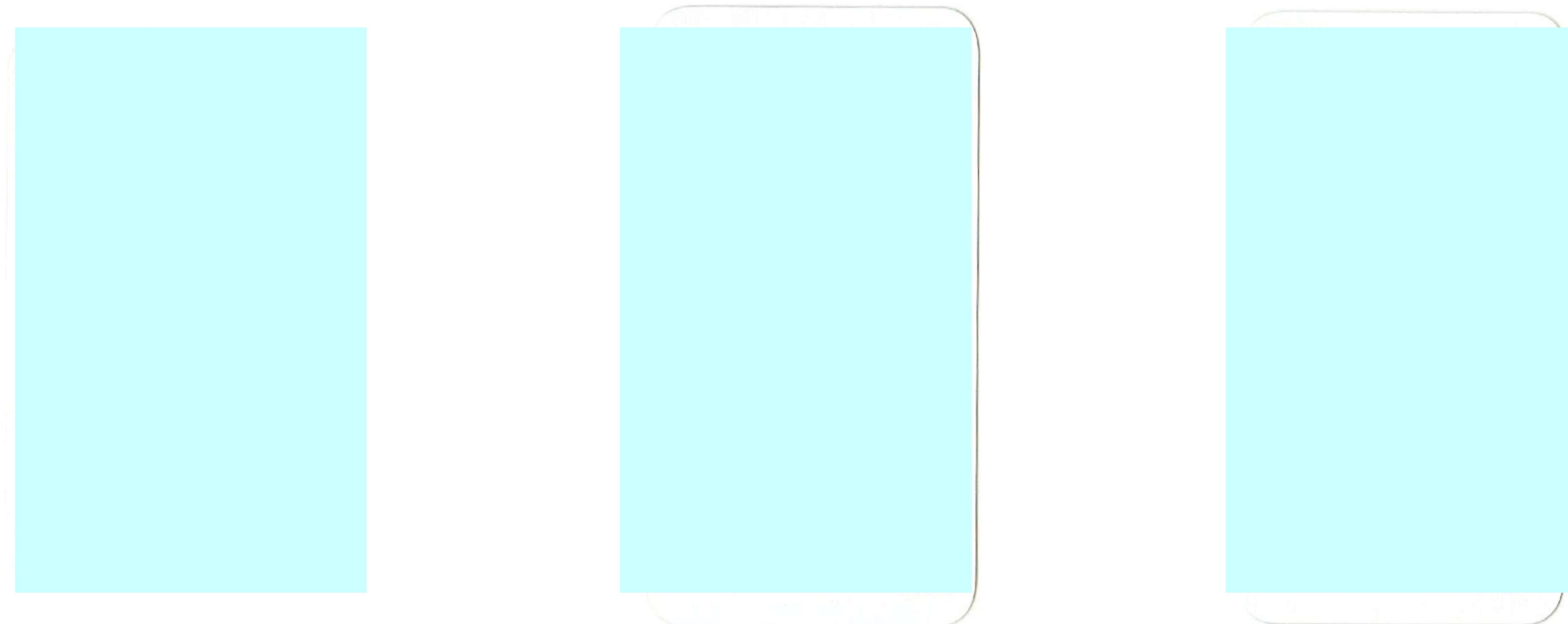


P(B,A=0)



Using Bayes rule

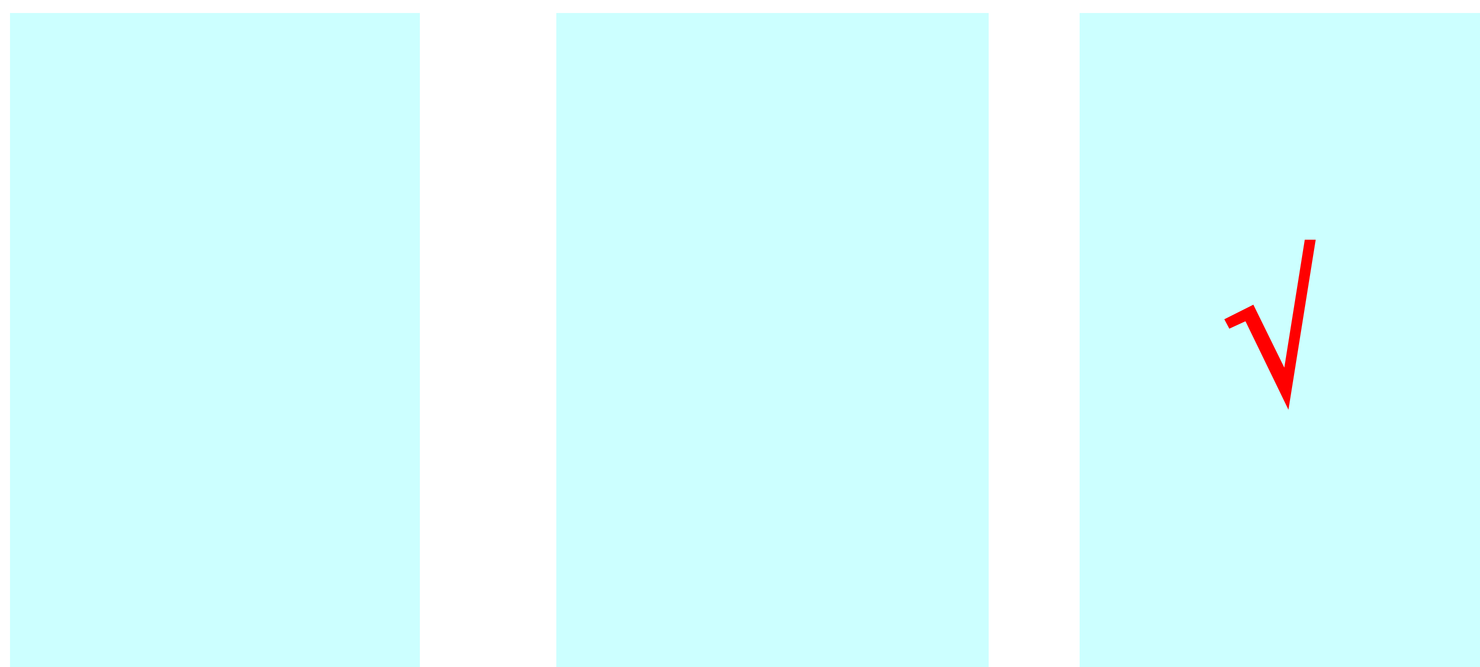
- Cards game:



Place your bet on the location of the King!

Using Bayes rule

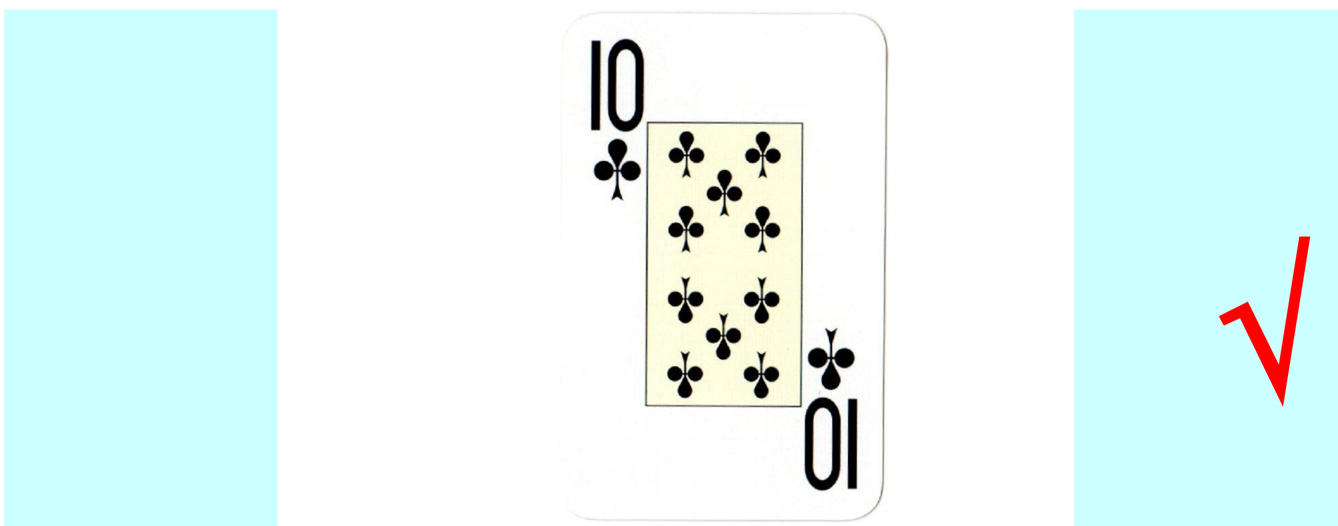
- Cards game:



Do you want to change your bet?

Using Bayes rule

A B C



Bayes rule

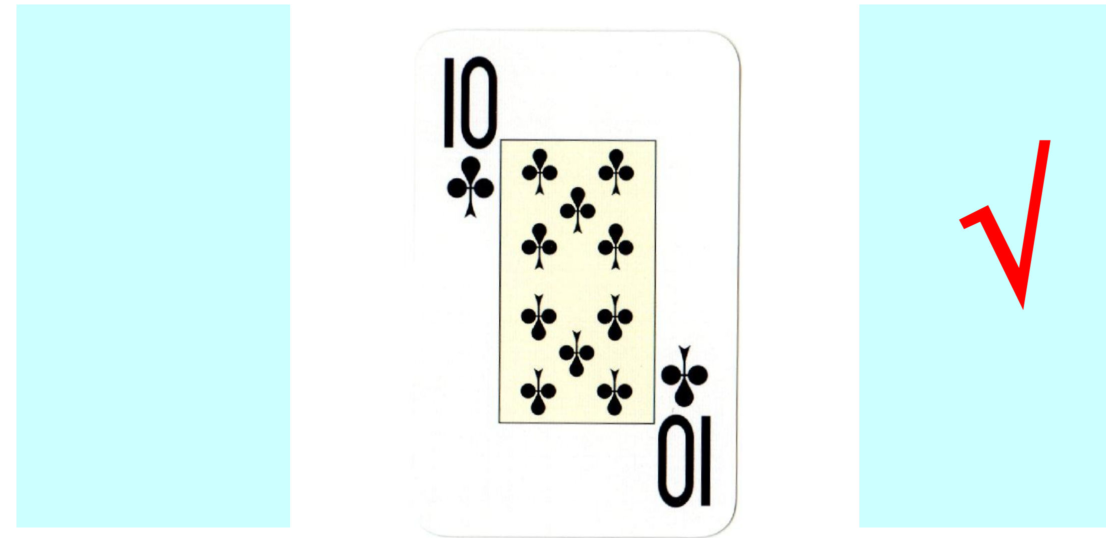
Computing the (posterior) probability: P(C = k | selB)

$$P(C=k \mid selB) = \frac{P(selB \mid C=k)P(C=k)}{P(selB)}$$

$$= \frac{P(selB \mid C=k)P(C=k)}{P(selB \mid C=k)P(C=k) + P(selB \mid C=10)P(C=10)}$$

Using Bayes rule

A B C

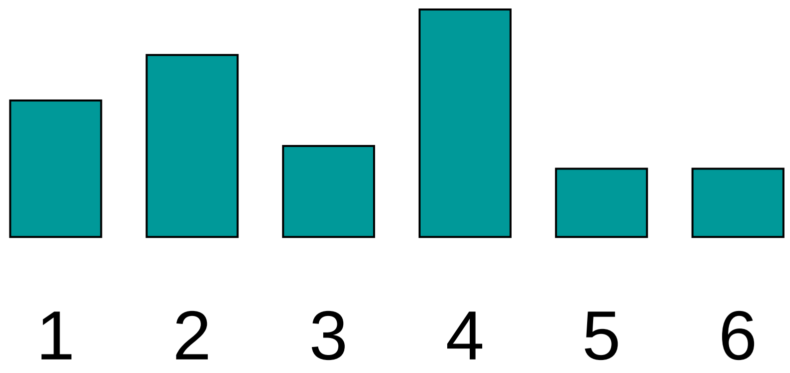


$P(C{=}k \mid selB) =$

$$= \frac{\overbrace{P(selB \mid C=k)}^{1/2}\,\overbrace{P(C=k)}^{1/3}}{\underbrace{P(selB \mid C=k)}_{1/2}\,\underbrace{P(C=k)}_{1/3} + \underbrace{P(selB \mid C=10)}_{1/2}\,\underbrace{P(C=10)}_{2/3}} = \mathbf{1/3}$$

Probability Density Function

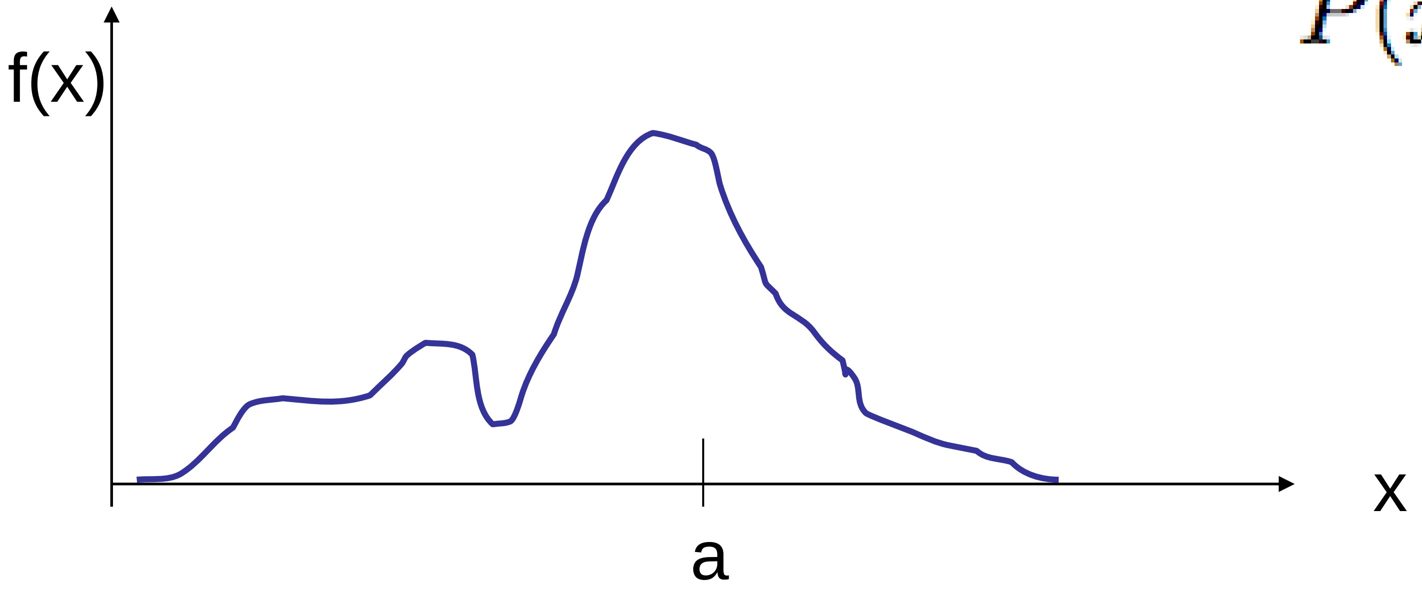
- Discrete distributions



$$\sum_i P(X = x_i) = 1$$

- Continuous: Cumulative Density Function (CDF): *F(a)*

$$P(x \leq a) = \int_{-\infty}^{a} f(\tau) d\tau$$



Cumulative Density Functions

- Total probability
- Probability Density Function (PDF)

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

- Properties:

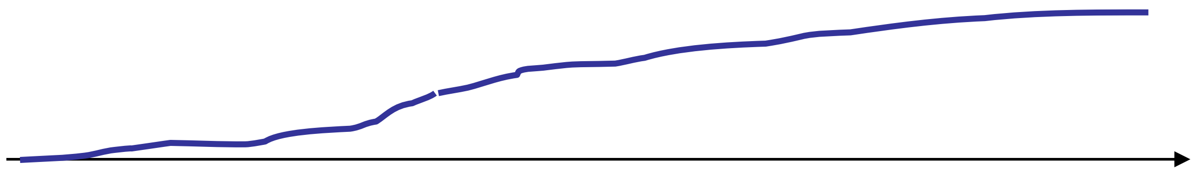
$$\frac{d}{dx}F(x) = f(x)$$

$$P(a \leq x \leq b) = \int_b^a f(x)dx = F(b) - F(a)$$

$$\lim_{x \to -\infty} F(x) = 0$$

$$\lim_{x \to \infty} F(x) = 1$$

$$F(a) \geq F(b) \; \forall a \geq b$$



Expectations

- Mean/Expected Value:
- Variance:
  - Note:
- In general:

$$E[x] = \bar{x} = \int x f(x) dx$$

$$Var(x) = E[(x - \bar{x})^2] = E[x^2] - (\bar{x})^2$$

$$E[x^2] = \int x^2 f(x) dx$$

$$E[g(x)] = \int g(x) f(x) dx$$

Multivariate

- Joint for (x,y)
- Marginal:
- Conditionals:
- Chain rule:

$$P\left((x,y) \in A\right) = \int\int_A f(x,y)dxdy$$

$$f(x) = \int f(x,y)dy$$

$$f(x|y) = \frac{f(x,y)}{f(y)}$$

$$f(x,y) = f(x|y)f(y) = f(y|x)f(x)$$

Bayes Rule

- Standard form:

- Replacing the bottom:

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}$$

$$f(x|y) = \frac{f(y|x)f(x)}{\int f(y|x)f(x)dx}$$

Binomial

- Distribution:

$$x \sim Binomial(p, n)$$

$$P(x = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

- Mean/Var:

$$E[x] = np$$

$$Var(x) = np(1 - p)$$

Uniform

- Anything is equally likely in the region [a,b]

- Distribution:

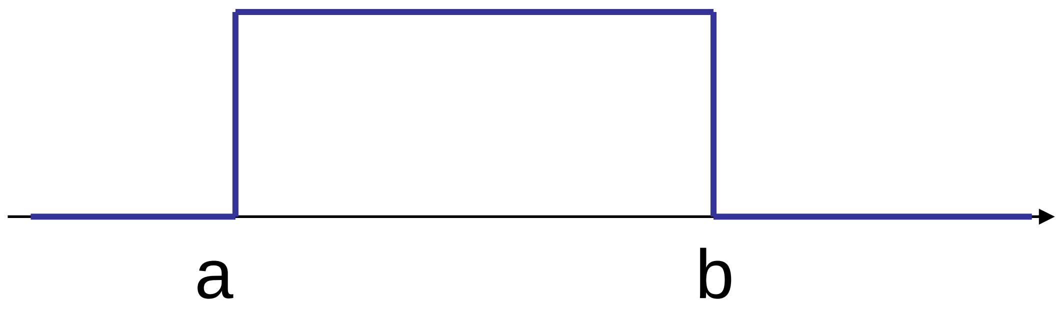
$$x \sim U(a,b)$$

- Mean/Var

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & otherwise \end{cases}$$

$$E[x] = \frac{a+b}{2}$$

$$Var(x) = \frac{a^2+ab+b^2}{3}$$



Gaussian (Normal)

- If I look at the height of women in country xx, it will look approximately Gaussian
- Small random noise errors, look Gaussian/Normal

- Distribution:

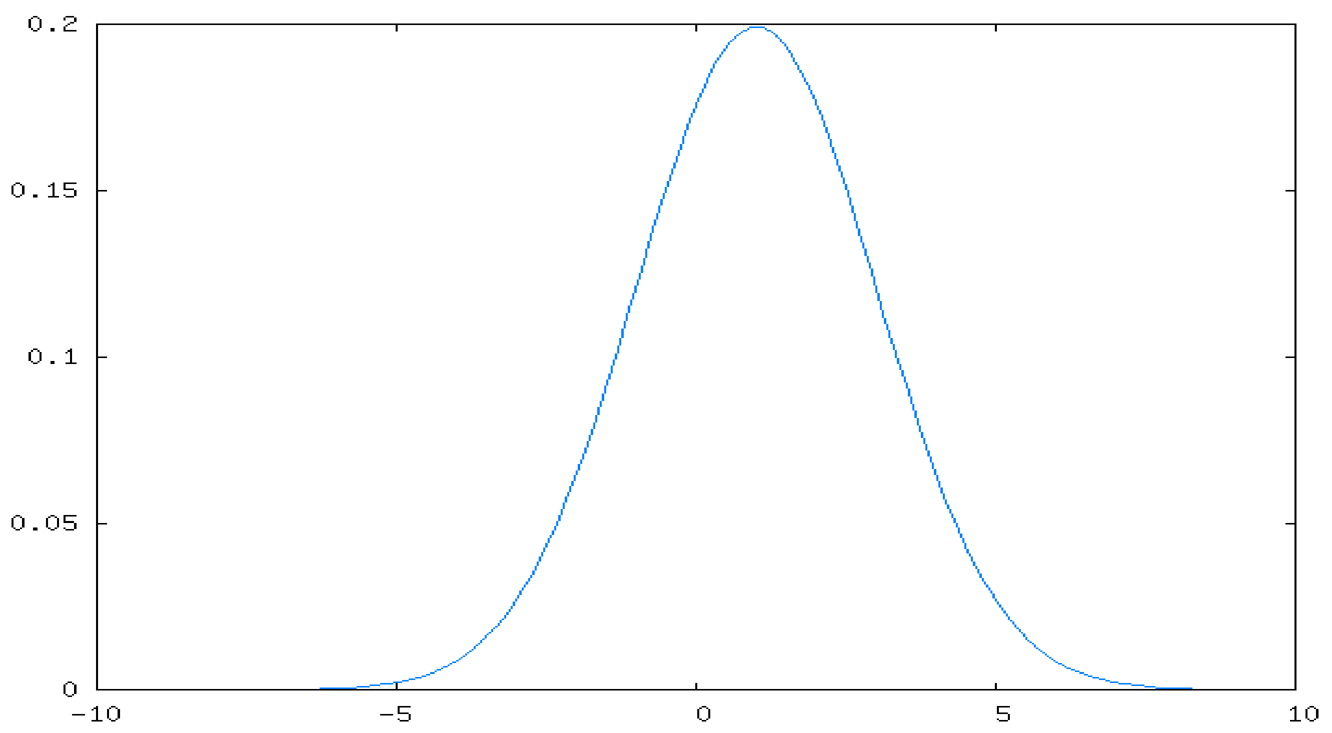
$$x \sim N(\mu, \sigma^2)$$

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

- Mean/var

$$E[x] = \mu$$

$$Var(x) = \sigma^2$$



Why Do People Use Gaussians

- Central Limit Theorem: (loosely)
  - Sum of a large number of IID random variables is approximately Gaussian

Multivariate Gaussians

- Distribution for vector x

$$x = (x_1, \ldots, x_N)^T, \quad x \sim N(\mu, \Sigma)$$

- PDF:

$$f(x) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$$E[x] = \mu = (E[x_1], \ldots, E[x_N])^T$$

$$Var(x) \rightarrow \Sigma = \begin{pmatrix} Var(x_1) & Cov(x_1, x_2) & \ldots & Cov(x_1, x_N) \\ Cov(x_2, x_1) & Var(x_2) & \ldots & Cov(x_2, x_N) \\ \vdots & & \ddots & \vdots \\ Cov(x_N, x_1) & Cov(x_N, x_2) & \ldots & Var(x_N) \end{pmatrix}$$

Multivariate Gaussians

$$f(x) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

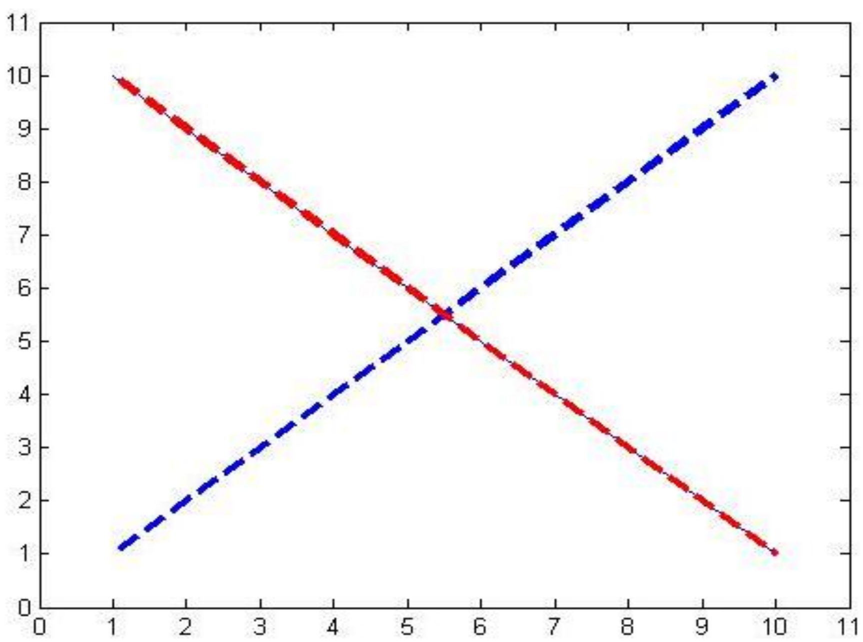
$$E[x] = \mu = (E[x_1], \ldots, E[x_N])^T$$

$$Var(x) \rightarrow \Sigma = \begin{pmatrix} Var(x_1) & Cov(x_1, x_2) & \ldots & Cov(x_1, x_N) \\ Cov(x_2, x_1) & Var(x_2) & \ldots & Cov(x_2, x_N) \\ \vdots & & \ddots & \vdots \\ Cov(x_N, x_1) & Cov(x_N, x_2) & \ldots & Var(x_N) \end{pmatrix}$$

$$\mathrm{cov}(\chi_1, \chi_2) = \frac{1}{n} \sum_{i=1}^{n} (x_{1,i} - \mu_1)(x_{2,i} - \mu_2)$$

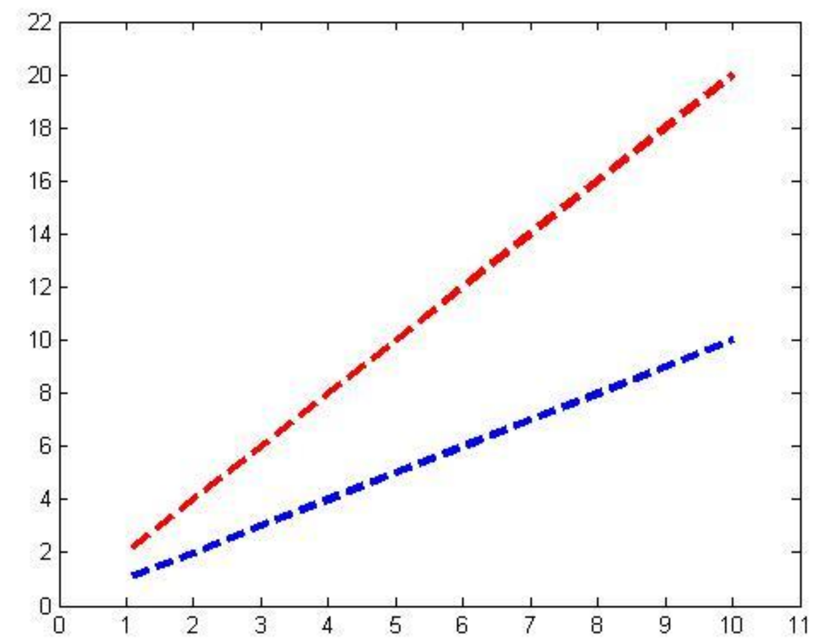
Covariance examples

Anti-correlated



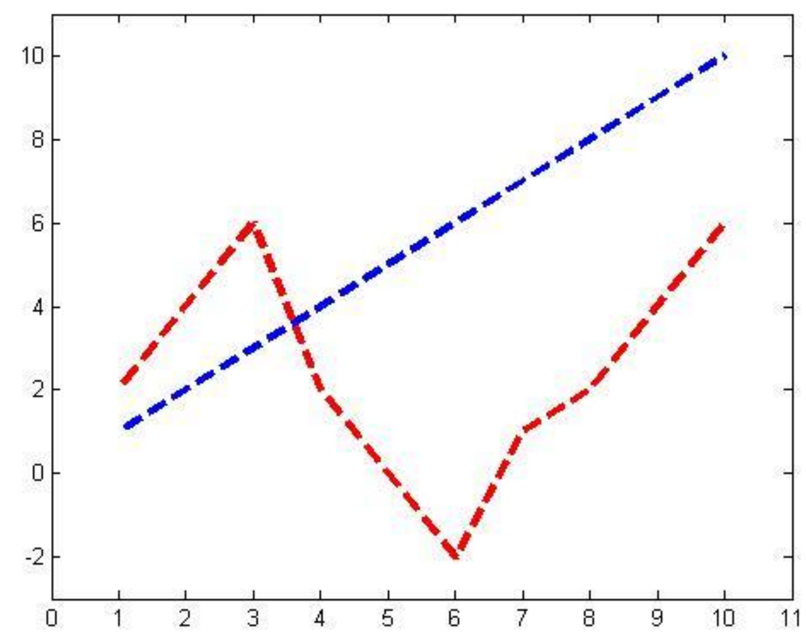
Covariance: -9.2

Correlated



Covariance: 18.33

Independent (almost)



Covariance: 0.6

Sum of Gaussians

- The sum of two Gaussians is a Gaussian:

$$x \sim N(\mu, \sigma^2) \quad y \sim N(\mu_y, \sigma_y^2)$$

$$ax + b \sim N(a\mu + b, (a\sigma)^2)$$

$$x + y \sim N(\mu + \mu_y, \sigma^2 + \sigma_y^2)$$

Important points

- Random variables
- Chain rule
- Bayes rule
- Joint distribution, independence, conditional independence