VBM683
Machine Learning

Pinar Duygulu

Slides are adapted from
Dhruv Batra

Plan for Today

- Review of Probability
  - Discrete vs Continuous Random Variables
  - PMFs vs PDF
  - Joint vs Marginal vs Conditional Distributions
  - Bayes Rule and Prior
  - Expectation, Entropy, KL-Divergence

Probability

- The world is a very uncertain place
- 30 years of Artificial Intelligence and Database research danced around this fact
- And then a few AI researchers decided to use some ideas from the eighteenth century

Slide Credit: Andrew Moore

Probability

- A is non-deterministic event
  - Can think of A as a boolean-valued variable

- Examples
  - A = your next patient has cancer
  - A = Donald Trump Wins the 2016 Presidential Election

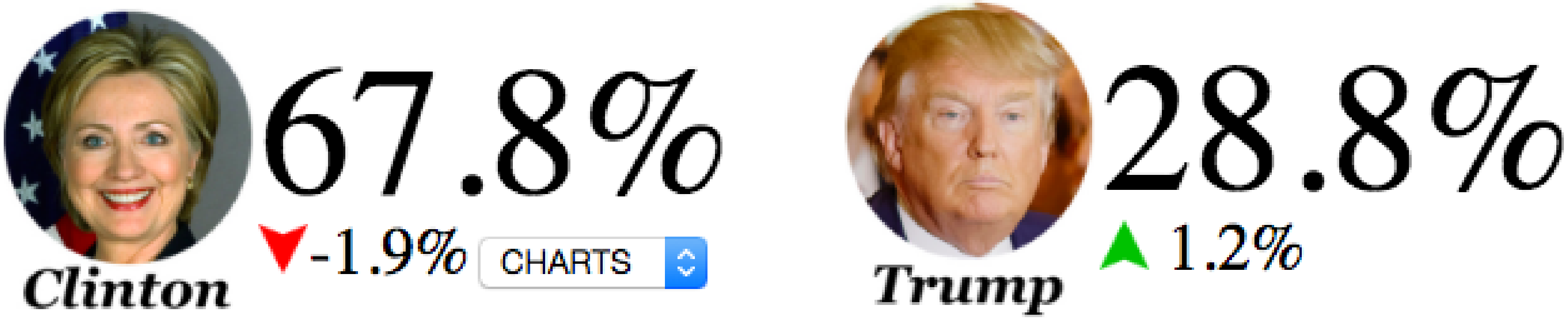
Interpreting Probabilities

- What does P(A) mean?

- Frequentist View
  - limit N→∞ #(A is true)/N
  - limiting frequency of a repeating non-deterministic event

- Bayesian View
  - P(A) is your “belief” about A

- Market Design View
  - P(A) tells you how much you would bet



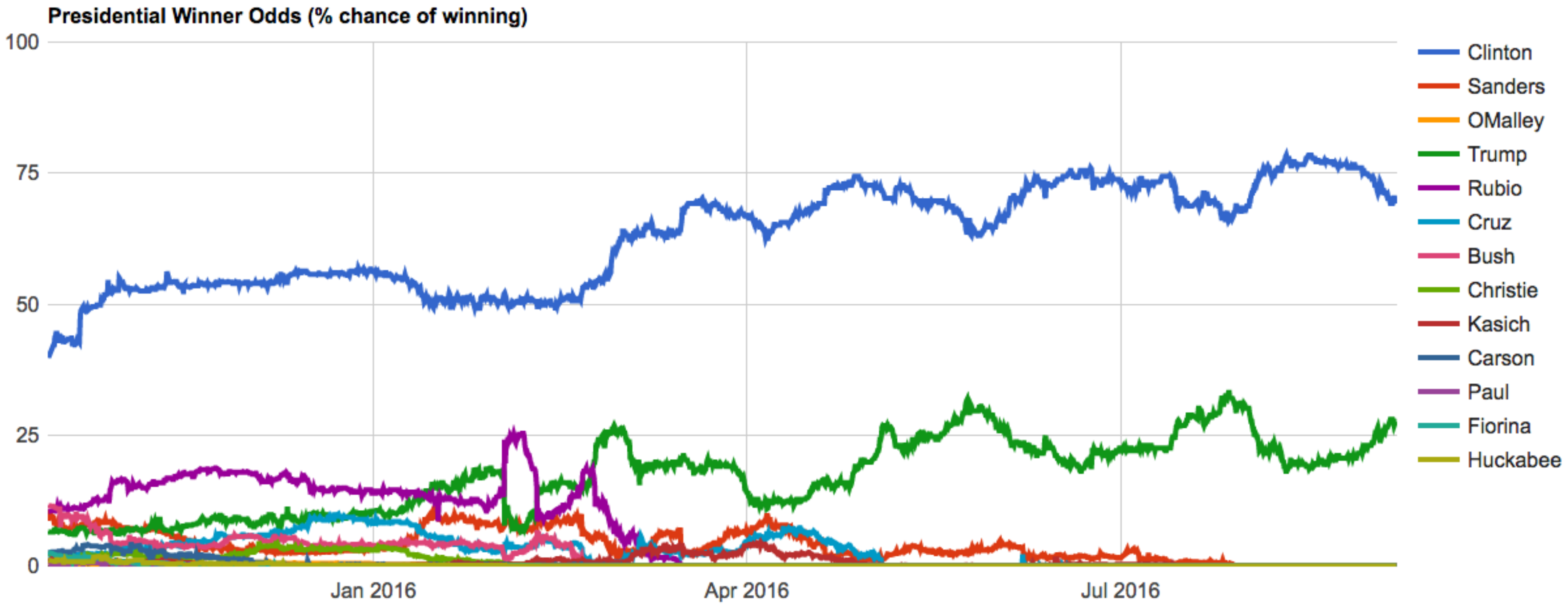


Image Credit: Intrade / NPR

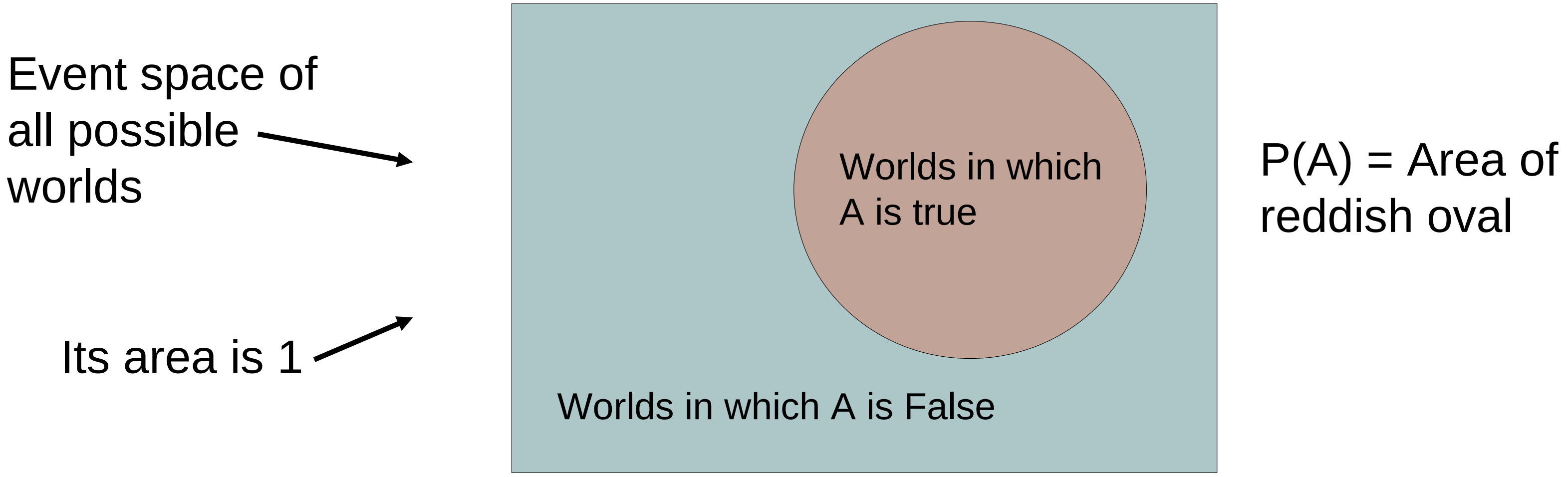
Slide Credit: Andrew Moore

Axioms of Probability

- $0 <= P(A) <= 1$
- P(empty-set) = 0
- P(everything) = 1
- P(A or B) = P(A) + P(B) – P(A and B)

Interpreting the Axioms

- 0<= P(A) <= 1
- P(empty-set) = 0
- P(everything) = 1
- P(A or B) = P(A) + P(B) – P(A and B)



Event space of all possible worlds

Its area is 1

Worlds in which A is true

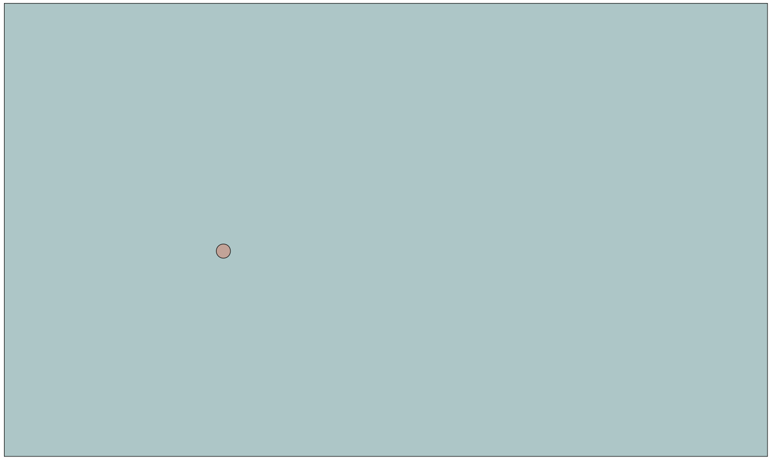
Worlds in which A is False

P(A) = Area of reddish oval

Image Credit: Andrew Moore

Interpreting the Axioms

- 0<= P(A) <= 1
- P(empty-set) = 0
- P(everything) = 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

Image Credit: Andrew Moore

Interpreting the Axioms

- 0<= P(A) <= 1
- P(empty-set) = 0
- P(everything) = 1
- P(A or B) = P(A) + P(B) – P(A and B)

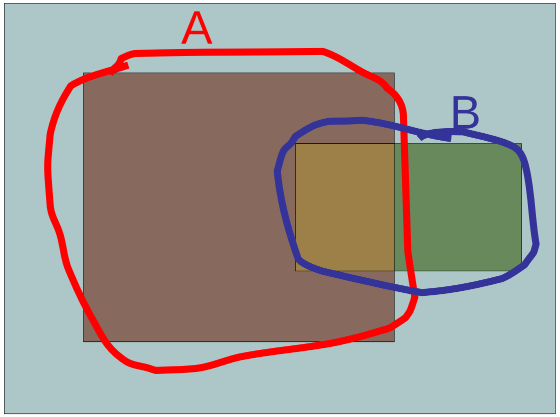
The area of A can't get any bigger than 1

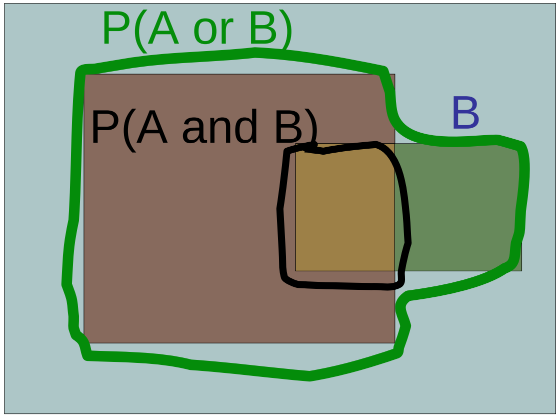
And an area of 1 would mean all worlds will have A true

Image Credit: Andrew Moore

Interpreting the Axioms

- 0<= P(A) <= 1
- P(empty-set) = 0
- P(everything) = 1
- P(A or B) = P(A) + P(B) – P(A and B)





Simple addition and subtraction

Image Credit: Andrew Moore

Concepts

- Sample Space
  - Space of events
- Random Variables
  - Mapping from events to numbers
  - Discrete vs Continuous
- Probability
  - Mass vs Density

Discrete Random Variables

X → discrete random variable

$\mathcal{X}$ or Val(X) → sample space of possible outcomes, which may be finite or countably infinite

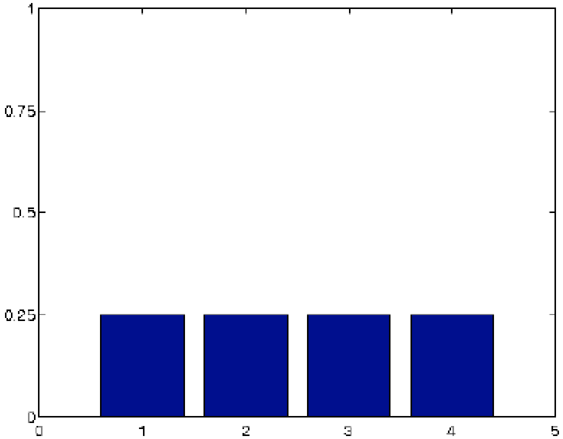
$x \in \mathcal{X}$ → outcome of sample of discrete random variable

$p(X = x)$ → probability distribution (probability mass function)

$p(x)$ → shorthand used when no ambiguity

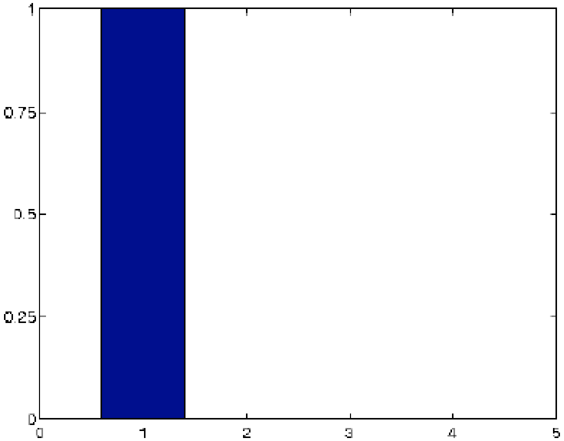
$$0 \leq p(x) \leq 1 \text{ for all } x \in \mathcal{X}$$

$$\sum_{x \in \mathcal{X}} p(x) = 1$$



uniform distribution

$$\mathcal{X} = \{1, 2, 3, 4\}$$



degenerate distribution

Slide Credit: Erik Suddherth

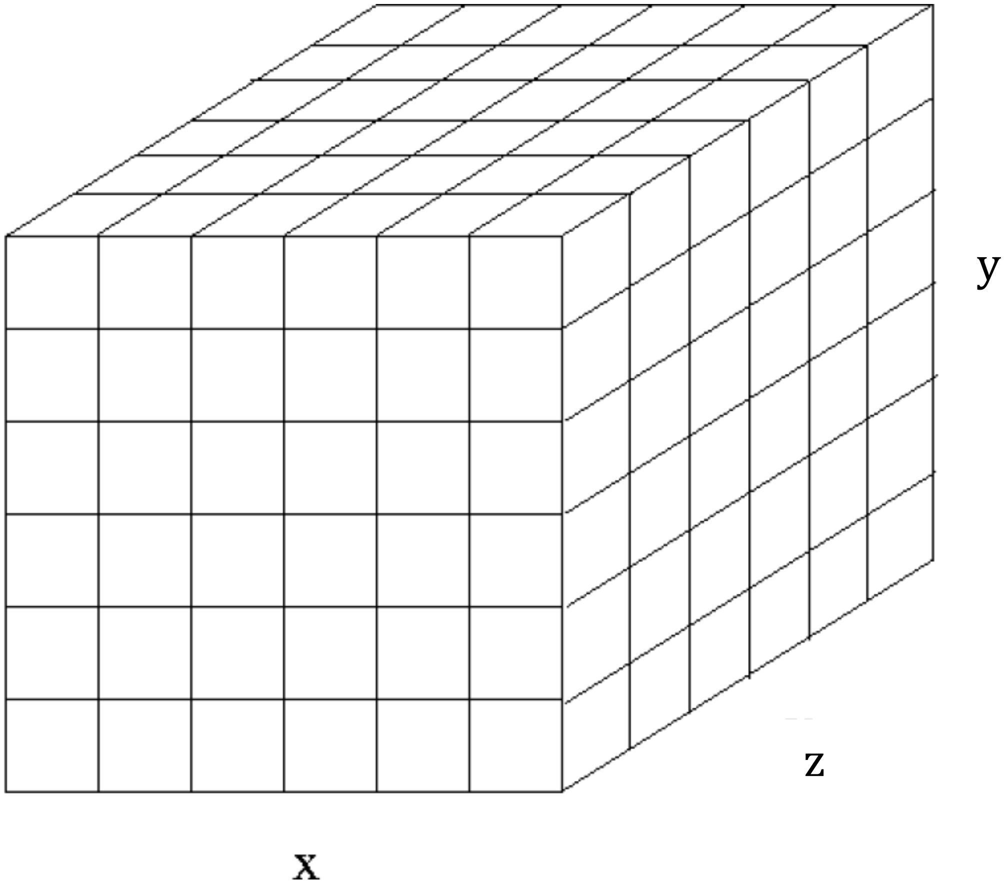
Concepts

- Expectation
- Variance

Most Important Concepts

- Marginal distributions / Marginalization
- Conditional distribution / Chain Rule
- Bayes Rule

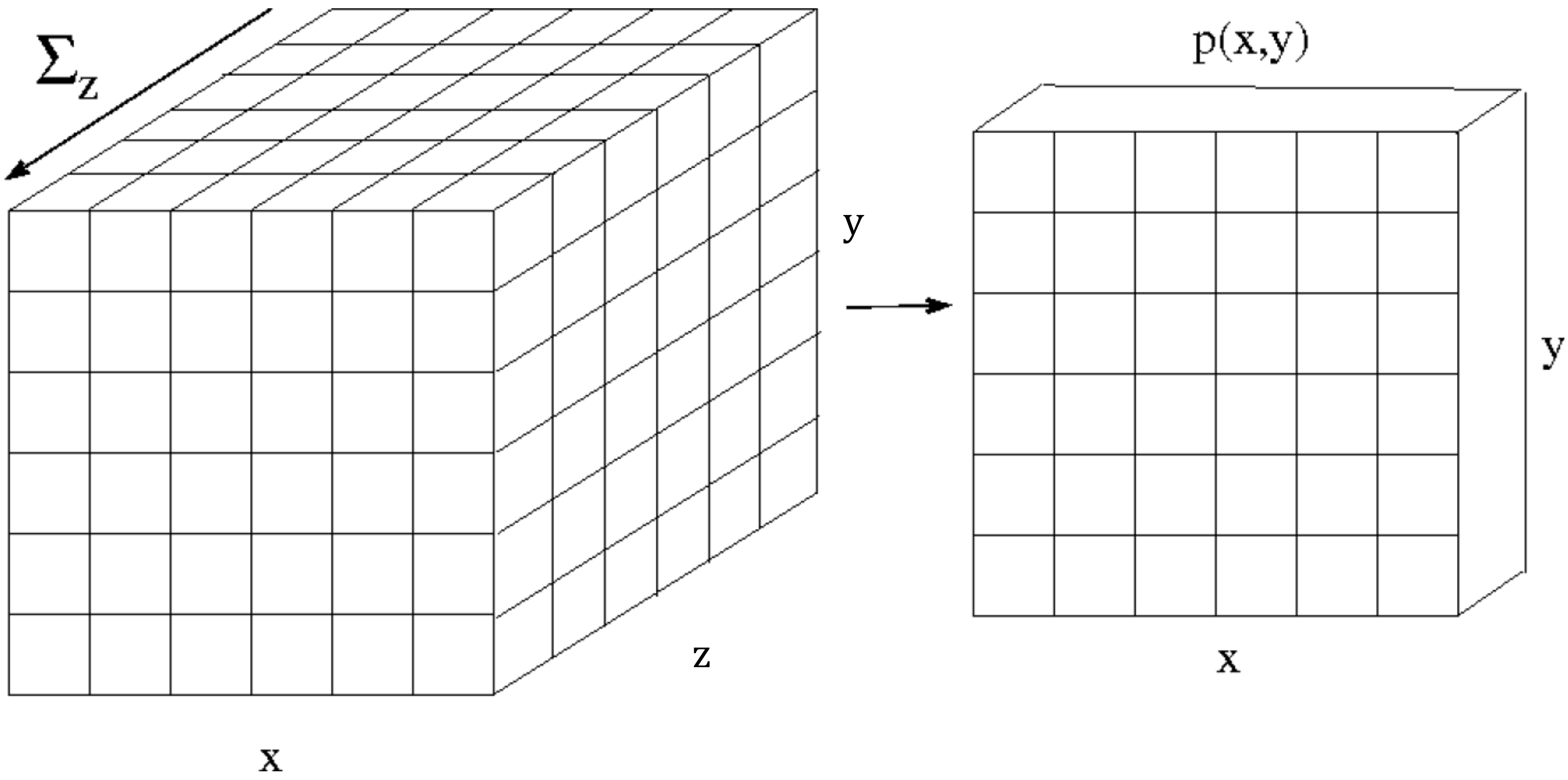
Joint Distribution



Marginalization

- Marginalization
  - Events: P(A) = P(A and B) + P(A and not B)
  - Random variables $P(X=x)=\sum_{y} P(X=x, Y=y)$

Marginal Distributions



$$p(x, y) = \sum_{z \in \mathcal{Z}} p(x, y, z) \qquad p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$$

Slide Credit: Erik Suddherth

Conditional Probabilities

- P(Y=y | X=x)

- What do you believe about Y=y, if I tell you X=x?

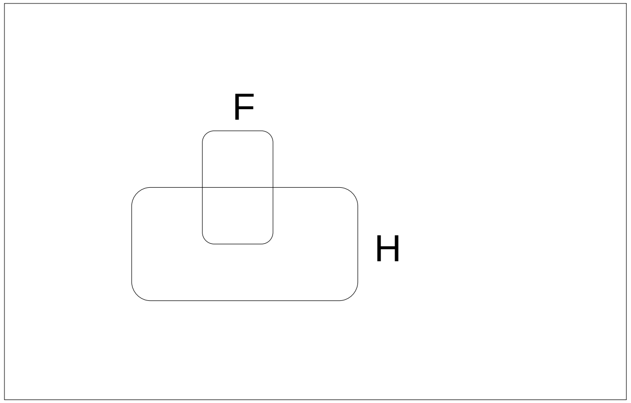
- P(Donald Trump Wins the 2016 Election)?

- What if I tell you:
  - He has the Republican nomination
  - His twitter history
  - The complete DVD set of The Apprentice

Conditional Probabilities

- P(A | B) = In worlds that where B is true,
  fraction where A is true

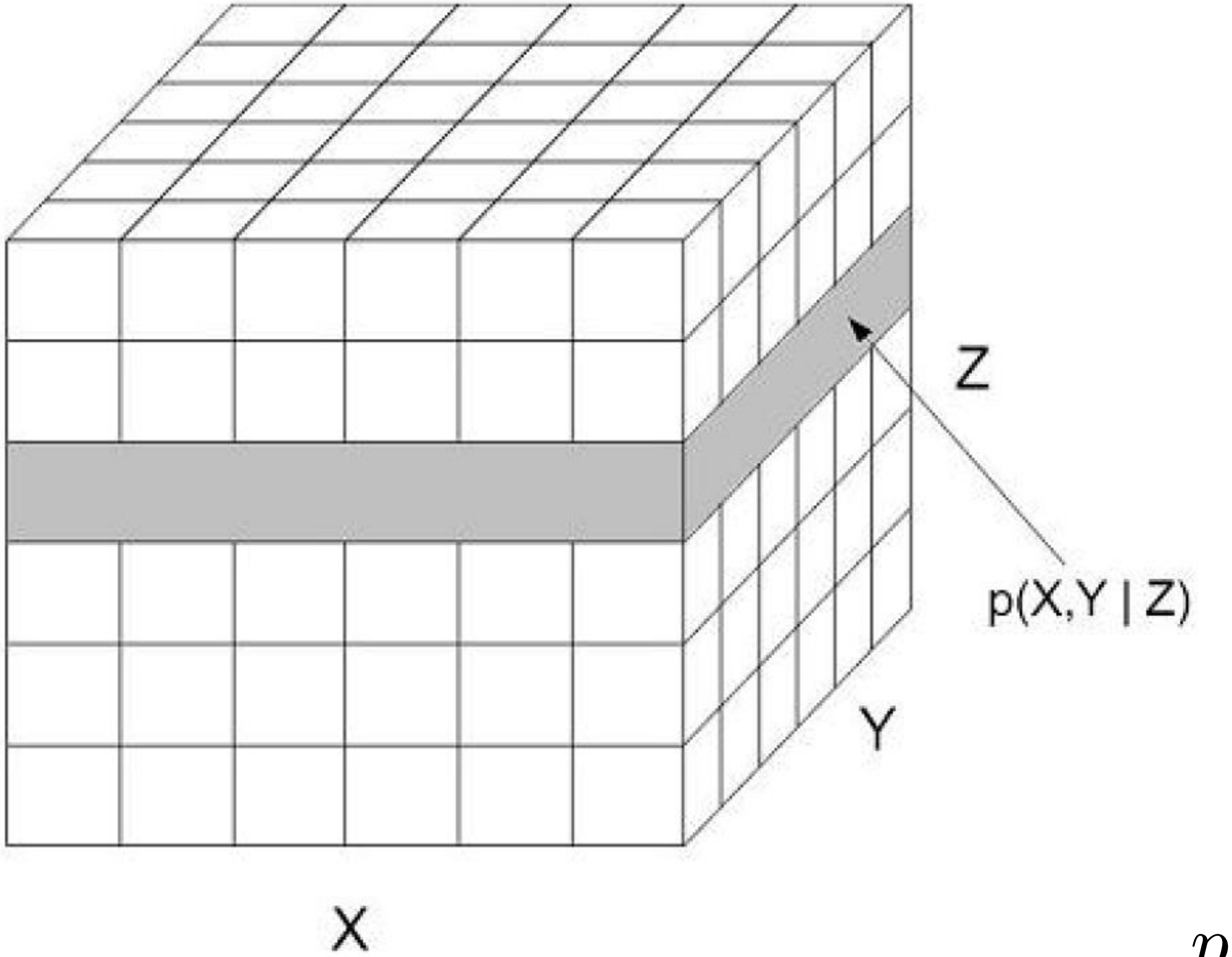
- Example
  - H: “Have a headache”
  - F: “Coming down with Flu”



P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

“Headaches are rare and flu is rarer, but if you’re coming down with ‘flu there’s a 50-50 chance you’ll have a headache.”

Conditional Distributions



$$p(x, y \mid Z = z) = \frac{p(x, y, z)}{p(z)}$$

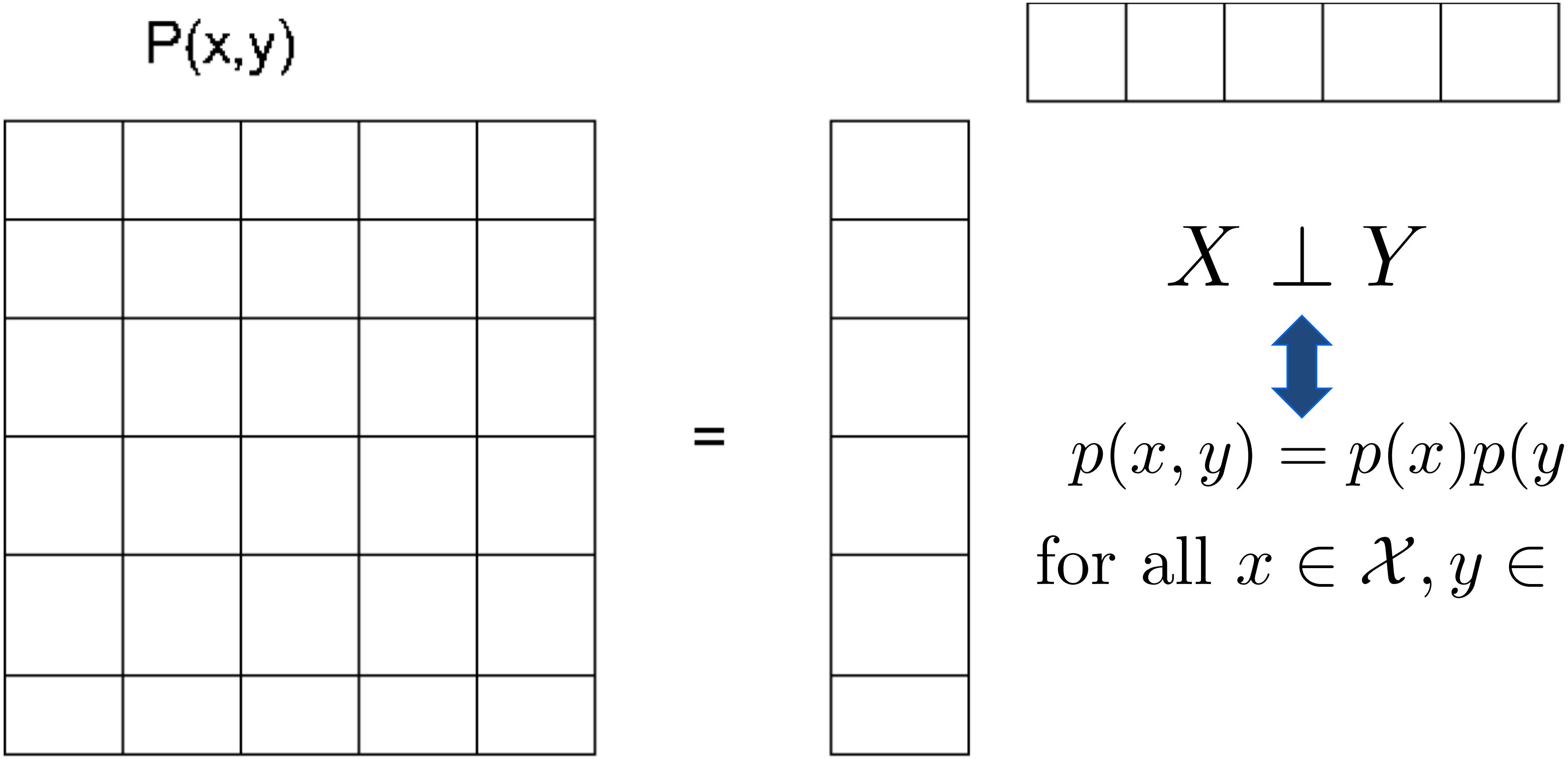
Slide Credit: Erik Sudderth

Conditional Probabilities

- Definition

- Corollary: Chain Rule

Independent Random Variables



$$X \perp Y$$

$$\Updownarrow$$

$$p(x, y) = p(x)p(y)$$

$$\text{for all } x \in \mathcal{X}, y \in \mathcal{Y}$$

Slide Credit: Erik Sudderth

Marginal Independence

- **Sets** of variables **X**, **Y**

- **X** is independent of **Y**
  - Shorthand: $P \vdash (\mathbf{X} \perp \mathbf{Y})$

- **Proposition:** *P* satisfies $(\mathbf{X} \perp \mathbf{Y})$ if and only if
  - $P(\mathbf{X=x},\mathbf{Y=y}) = P(\mathbf{X=x})\, P(\mathbf{Y=y})$, $\forall\, \boldsymbol{x} \in \boldsymbol{Val}(\boldsymbol{X}), \forall \boldsymbol{y} \in \boldsymbol{Val}(\boldsymbol{Y})$

Conditional independence

- **Sets** of variables **X**, **Y**, **Z**

- **X** is independent of **Y** given **Z** if
  - Shorthand: $P \vdash (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$
  - For $P \vdash (\mathbf{X} \perp \mathbf{Y} \mid \emptyset)$, write $P \vdash (\mathbf{X} \perp \mathbf{Y})$

- **Proposition:** *P* satisfies $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ if and only if
  - $P(\mathbf{X},\mathbf{Y}|\mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})\, P(\mathbf{Y}|\mathbf{Z})$, $\forall\ \forall \boldsymbol{x} \in \boldsymbol{Val}(\boldsymbol{X}), \forall \boldsymbol{y} \in \boldsymbol{Val}(\boldsymbol{Y}), \forall \boldsymbol{z} \in \boldsymbol{Val}(\boldsymbol{Z})$

Concept

- Bayes Rules
  - Simple yet fundamental

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

This is Bayes Rule

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**

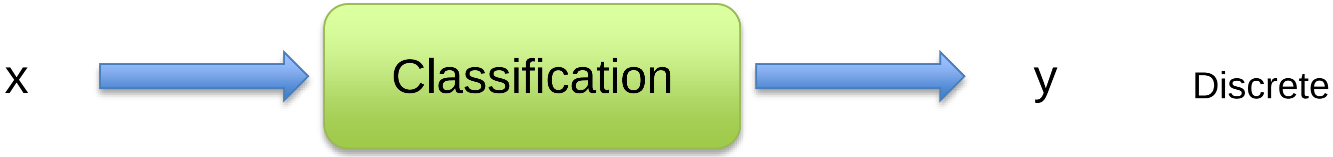
Image Credit: Andrew Moore

Bayes Rule

- Simple yet profound
  - Using Bayes Rules doesn’t make your analysis Bayesian!

- Concepts:
  - Likelihood
    - How much does a certain hypothesis explain the data?
  - Prior
    - What do you believe before seeing any data?
  - Posterior
    - What do we believe after seeing the data?

New Topic:
Naïve Bayes
(your first probabilistic classifier)



x

Classification

y

Discrete

Classification

- **Learn**: h:$\mathbf{X} \mapsto Y$
  - **X** – features
  - Y – target classes

- Suppose you know P(Y|**X**) exactly, how should you classify?
  - Bayes classifier:

Slide Credit: Carlos Guestrin

Generative vs. Discriminative

- Generative Approach
  - Assume some functional form for P(X|Y), P(Y)
  - Estimate p(X|Y) and p(Y)
  - Use Bayes Rule to calculate P(Y| X=x)
  - Indirect computation of P(Y|X) through Bayes rule
  - But, **can generate a sample**, $P(X) = \sum_y P(y) P(X|y)$
- Discriminative Approach
  - Estimate p(y|x) directly OR
  - Learn “discriminant” function h(x)
  - Direct but cannot obtain a sample of the data, because P(X) is not available

How hard is it to learn the optimal classifier?

- Categorical Data

- How do we represent these? How many parameters?
  - Class-Prior, P(Y):
    - Suppose Y is composed of *k* classes
  - Likelihood, P(**X**|Y):
    - Suppose **X** is composed of *d* binary features

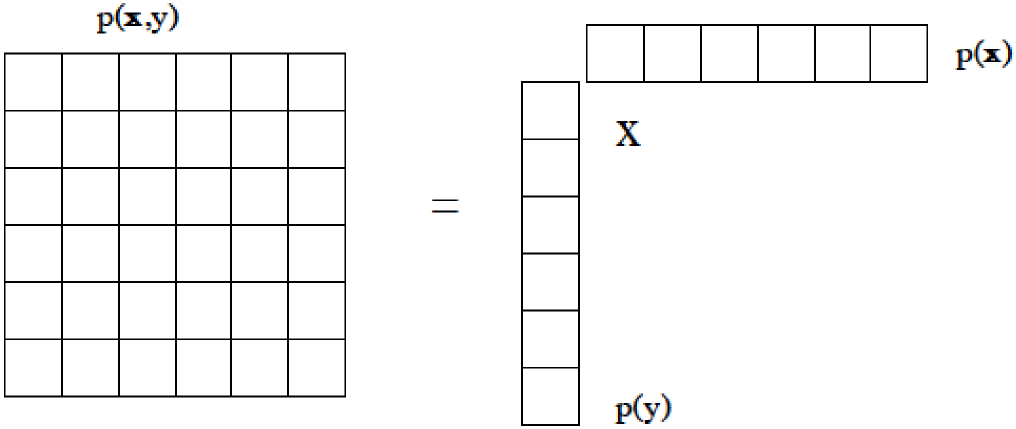
- Complex model → High variance with limited data!!!

Slide Credit: Carlos Guestrin

Independence to the rescue

- Two variables are independent iff their joint factors:

$$p(x, y) = p(x)p(y)$$



- Two variables are conditionally independent given a third one if for all values of the conditioning variable, the resulting slice factors:

$$p(x, y|z) = p(x|z)p(y|z) \qquad \forall z$$

Slide Credit: Sam Roweis

The Naïve Bayes assumption

- Naïve Bayes assumption:
  - Features are independent given class:

$$\begin{aligned} P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2|Y) \\ &= P(X_1|Y)P(X_2|Y) \end{aligned}$$

  - More generally:

$$\boxed{P(X_1 ... X_d|Y) = \prod_i P(X_i|Y)}$$

- How many parameters now?
  - Suppose **X** is composed of *d* binary features

Slide Credit: Carlos Guestrin

The Naïve Bayes Classifier

- Given:
  - Class-Prior P(Y)
  - *d* conditionally independent features **X** given the class Y
  - For each X_i, we have likelihood $P(X_i|Y)$

- Decision rule:

$$\begin{aligned} y^* = h_{NB}(\mathbf{x}) &= \arg\max_y P(y) P(x_1, \ldots, x_n \mid y) \\ &= \arg\max_y P(y) \prod_i P(x_i|y) \end{aligned}$$

- If assumption holds, NB is optimal classifier!

Slide Credit: Carlos Guestrin

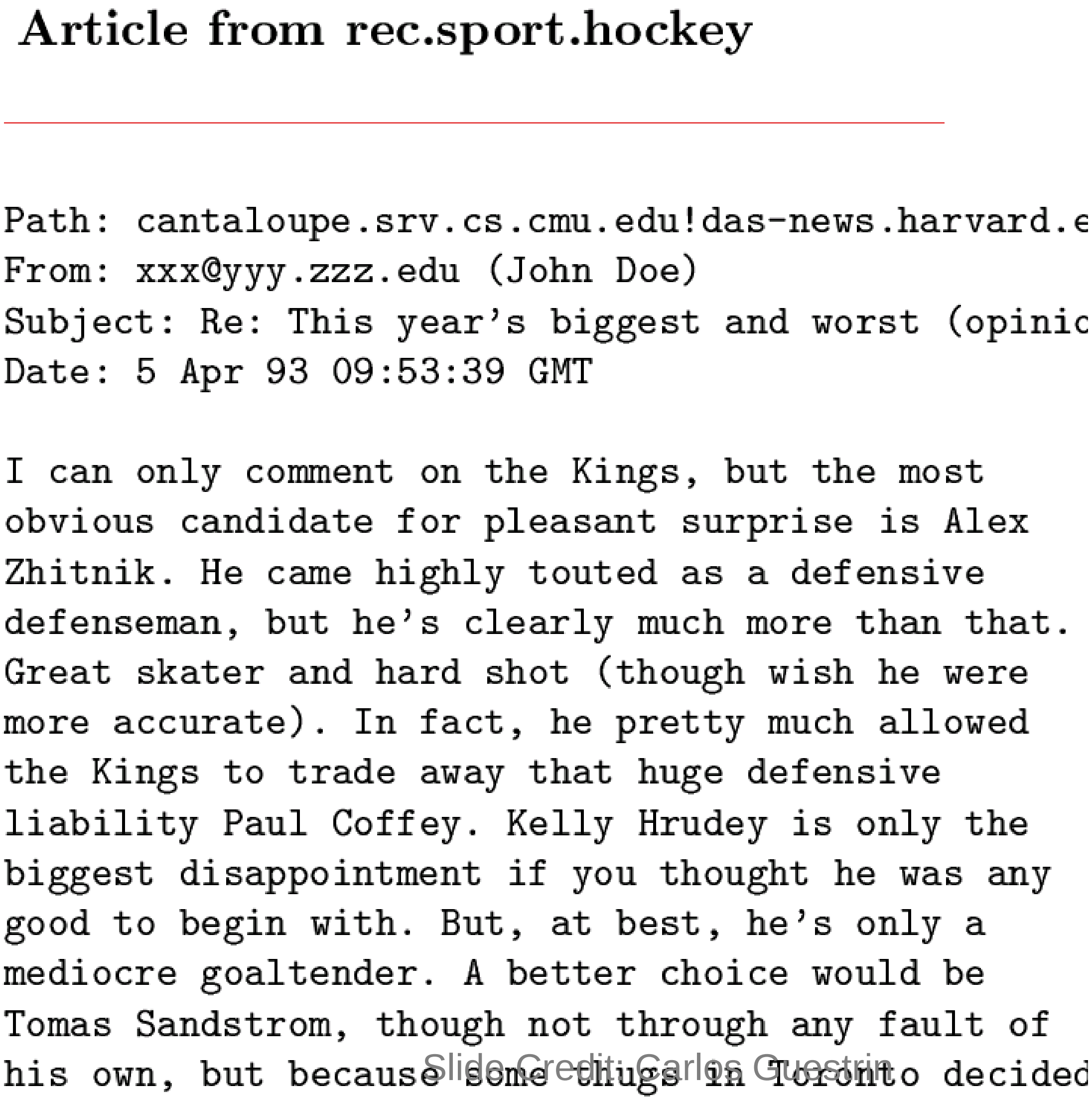
Text classification

- Classify e-mails
  - Y = {Spam,NotSpam}
- Classify news articles
  - Y = {what is the topic of the article?}
- Classify webpages
  - Y = {Student, professor, project, …}

- What about the features **X**?
  - The text!

Slide Credit: Carlos Guestrin

Features **X** are entire document – X_i for i^{th} word in article



Article from rec.sport.hockey

Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.e
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year's biggest and worst (opinic
Date: 5 Apr 93 09:53:39 GMT

I can only comment on the Kings, but the most obvious candidate for pleasant surprise is Alex Zhitnik. He came highly touted as a defensive defenseman, but he's clearly much more than that. Great skater and hard shot (though wish he were more accurate). In fact, he pretty much allowed the Kings to trade away that huge defensive liability Paul Coffey. Kelly Hrudey is only the biggest disappointment if you thought he was any good to begin with. But, at best, he's only a mediocre goaltender. A better choice would be Tomas Sandstrom, though not through any fault of his own, but because some thugs in Toronto decided

Slide Credit: Carlos Guestrin

NB for Text classification

- P(**X**|Y) is huge!!!
  - Article at least 1000 words, **X**={X_1,…,X_{1000}}
  - X_i represents i^th word in document, i.e., the domain of X_i is entire vocabulary, e.g., Webster Dictionary (or more), 10,000 words, etc.

- NB assumption helps a lot!!!
  - P(X_i=x_i|Y=y) is just the probability of observing word x_i in a document on topic y

$$h_{NB}(\mathbf{x}) = \arg\max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

Slide Credit: Carlos Guestrin

Bag of Words model

- Typical additional assumption:
  Position in document doesn't matter:
  $P(X_i=a|Y=y) = P(X_k=a|Y=y)$
  - "Bag of words" model – order of words on the page ignored
  - Sounds really silly, but often works very well!

$$P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

When the lecture is over, remember to wake up the person sitting next to you in the lecture room.

Slide Credit: Carlos Guestrin

Bag of Words model

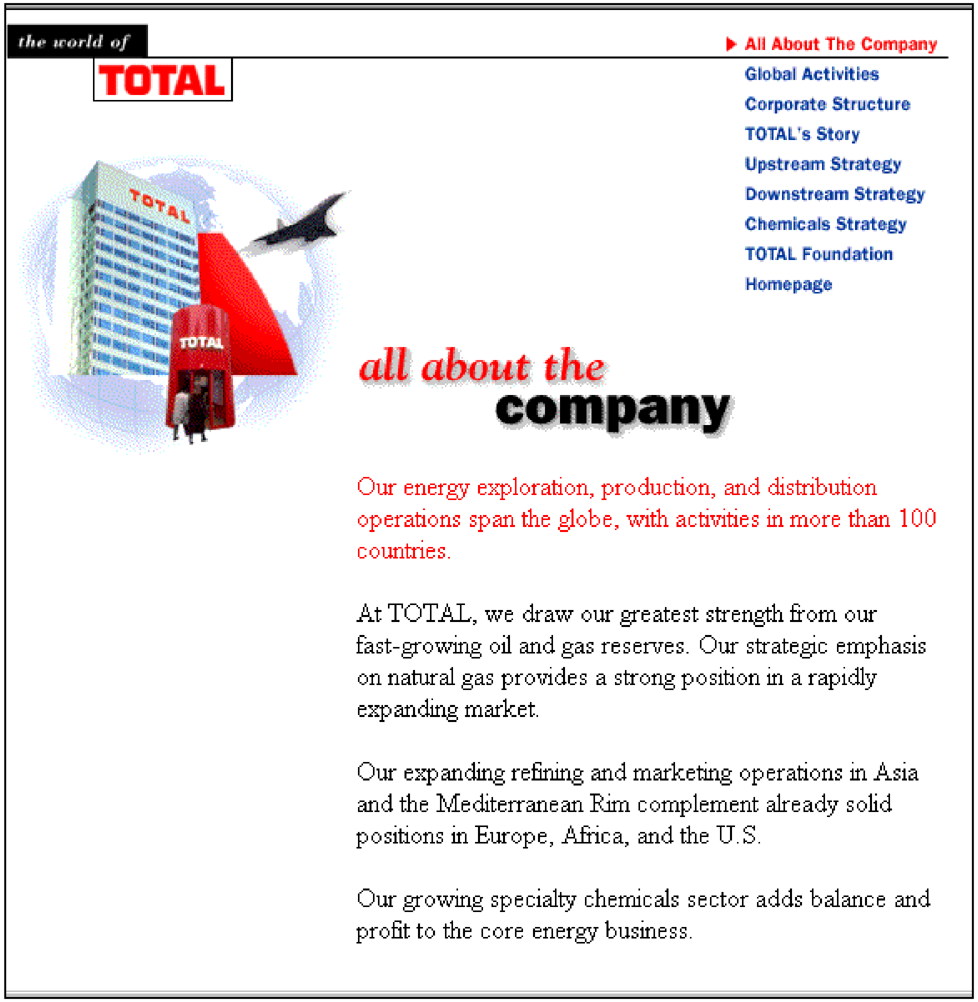
- Typical additional assumption:
  Position in document doesn't matter:
  $P(X_i=a|Y=y) = P(X_k=a|Y=y)$
  - "Bag of words" model – order of words on the page ignored
  - Sounds really silly, but often works very well!

$$P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

in is lecture lecture next over person remember room

sitting the the the to to up wake when you

Bag of Words model



all about the company

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

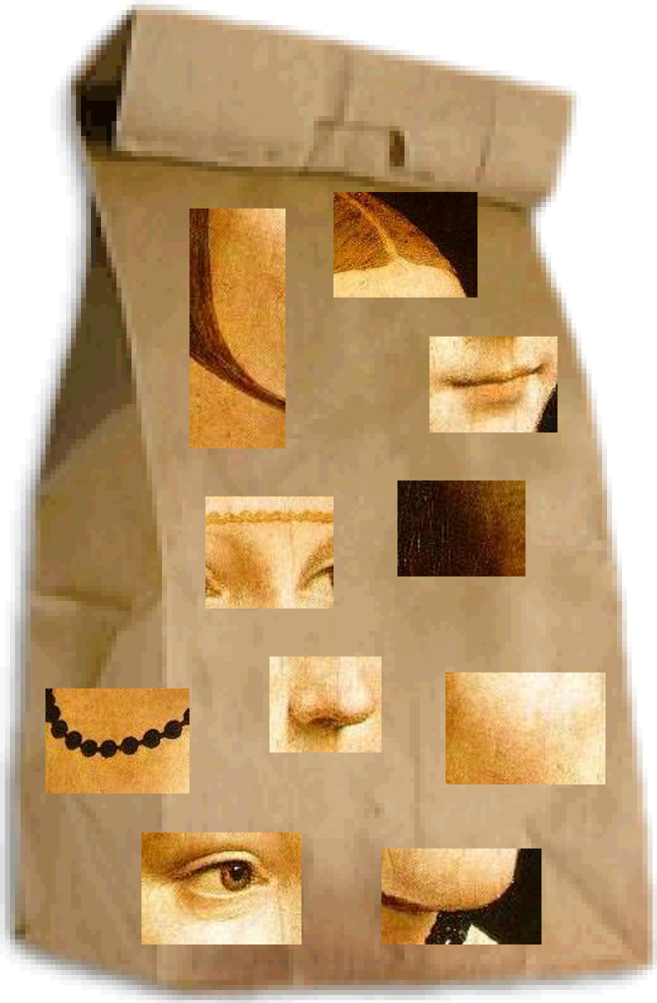
Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

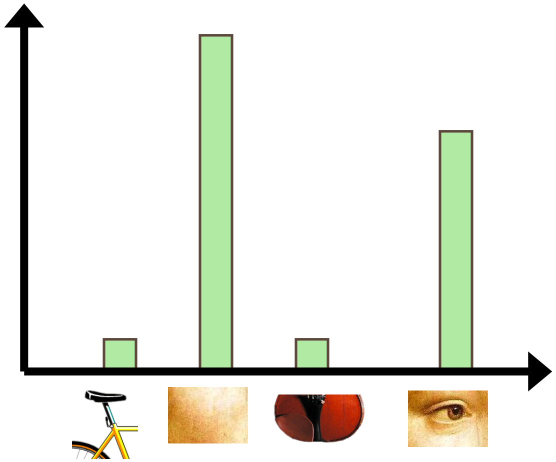
aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

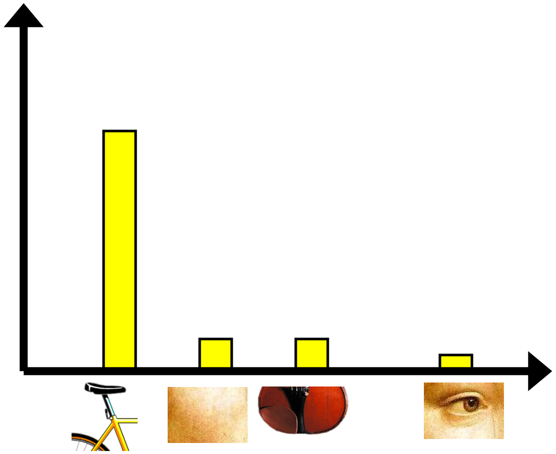
Slide Credit: Carlos Guestrin

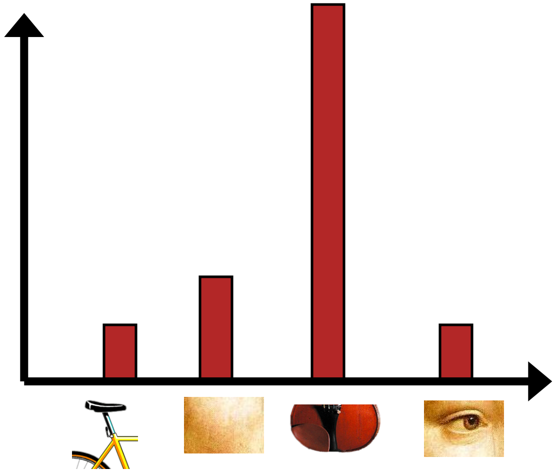
Object → **Bag of ‘words’**

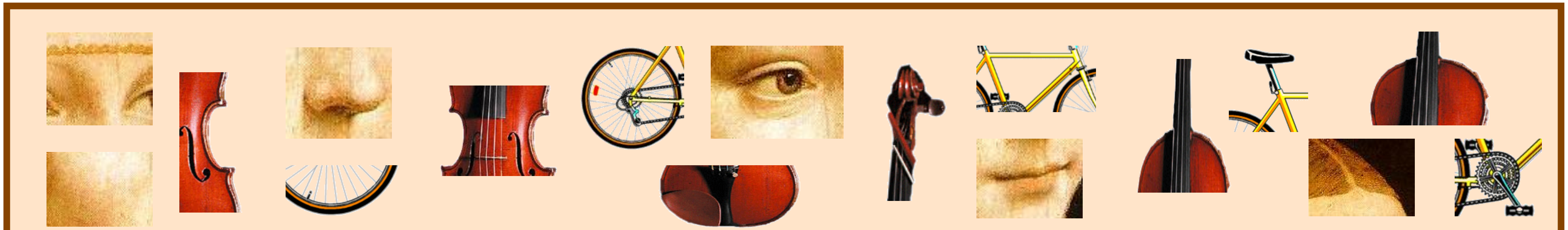


Slide Credit: Fei Fei Li



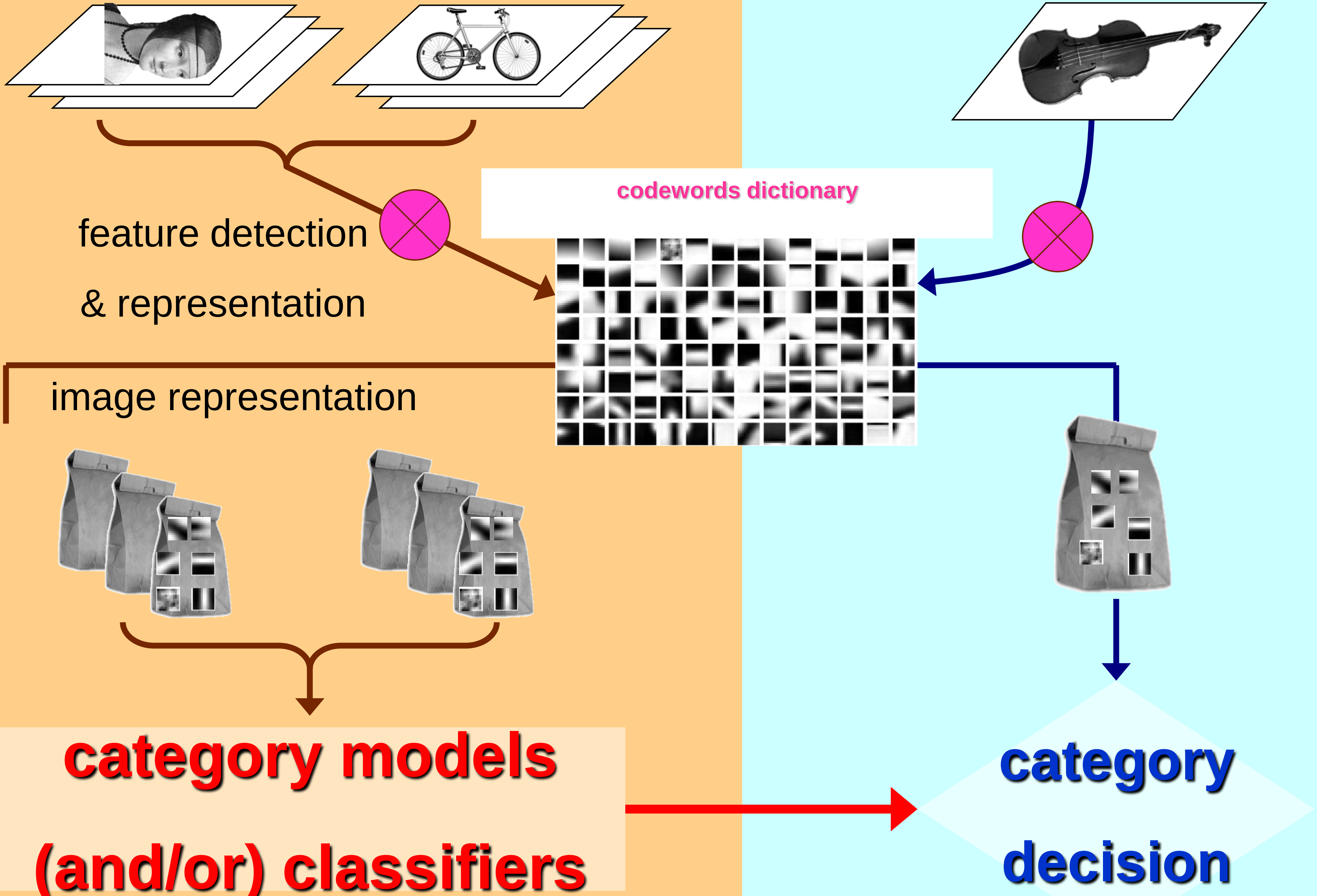






learning

recognition



codewords dictionary

feature detection

& representation

image representation

category models

(and/or) classifiers

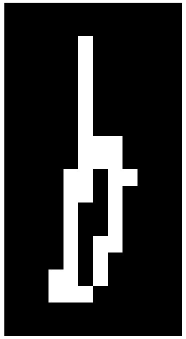
category

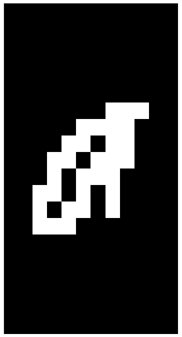
decision

Slide Credit: Fei Fei Li

What if we have continuous X_i ?

Eg., character recognition: X_i is ith pixel





Gaussian Naïve Bayes (GNB):

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \; e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

Sometimes assume variance

- is independent of Y (i.e., σ_i),
- or independent of X_i (i.e., σ_k)
- or both (i.e., σ)