UVA CS 4501 - 001 / 6501 – 007 Introduction to Machine Learning and Data Mining

Lecture 12: Probability and Statistics Review

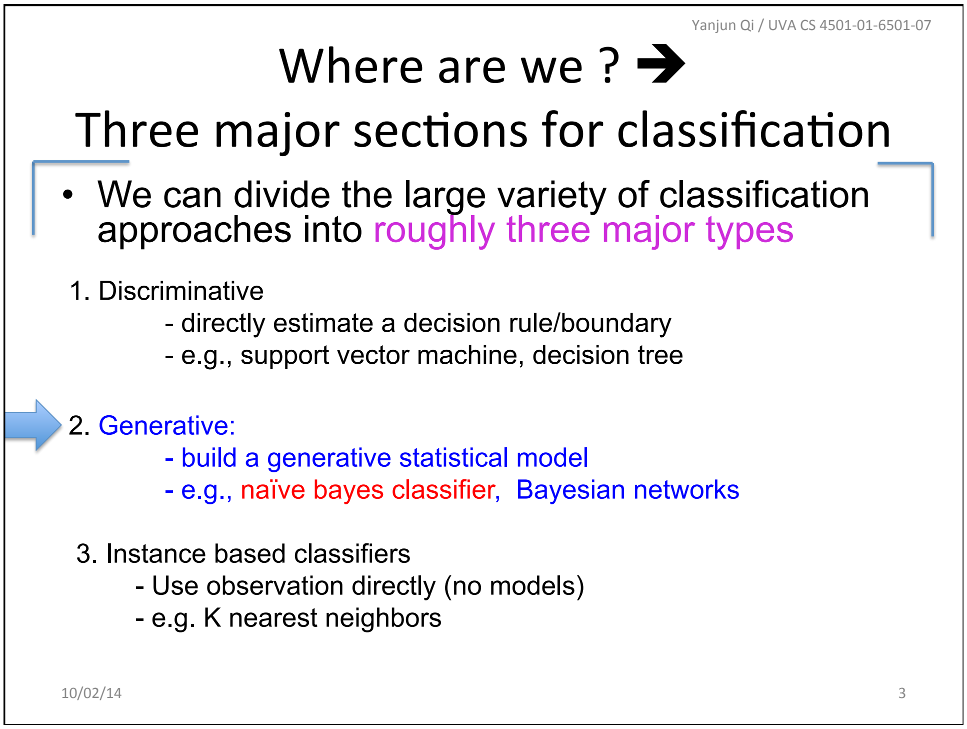
Yanjun Qi / Jane

University of Virginia

Department of Computer Science

Where are we ? ➔ Five major sections of this course

- ❑ Regression (supervised)
- ❑ Classification (supervised)
- ❑ Unsupervised models
- ❑ Learning theory
- ❑ Graphical models



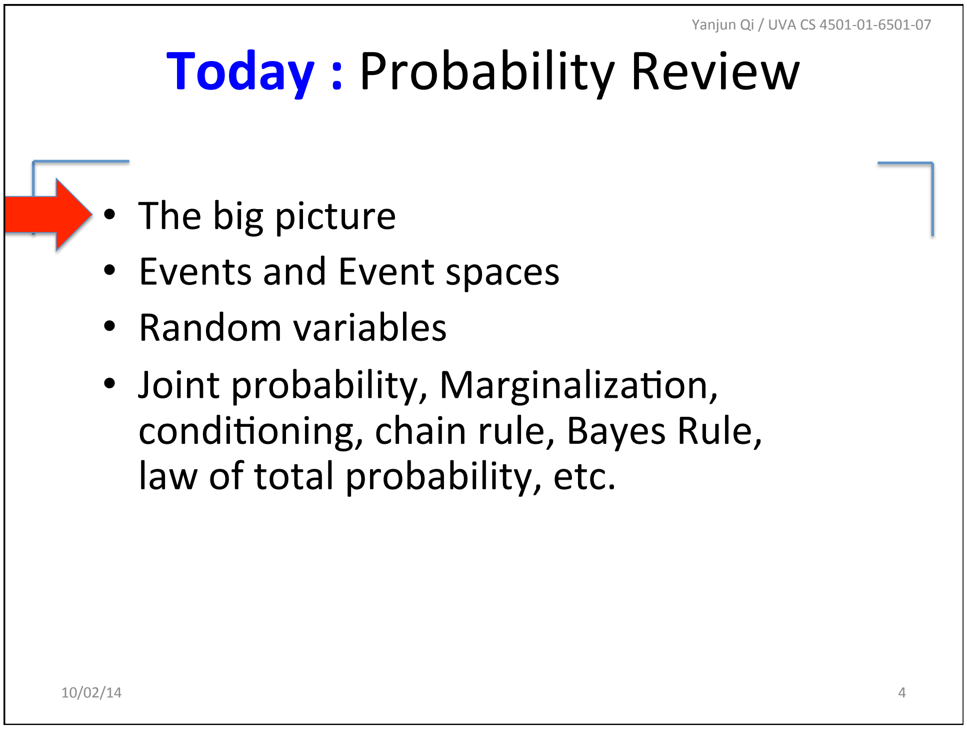
Where are we ? ➔ Three major sections for classification

- We can divide the large variety of classification approaches into roughly three major types

1. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., support vector machine, decision tree

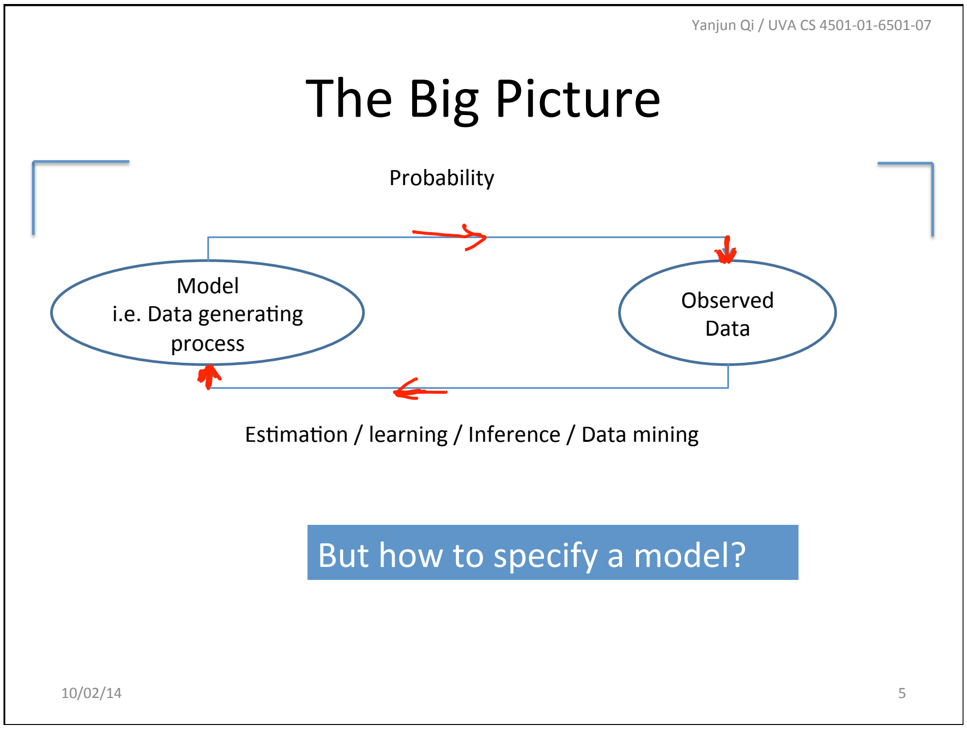
2. Generative:
   - build a generative statistical model
   - e.g., naïve bayes classifier, Bayesian networks

3. Instance based classifiers
   - Use observation directly (no models)
   - e.g. K nearest neighbors



Today : Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.



The Big Picture

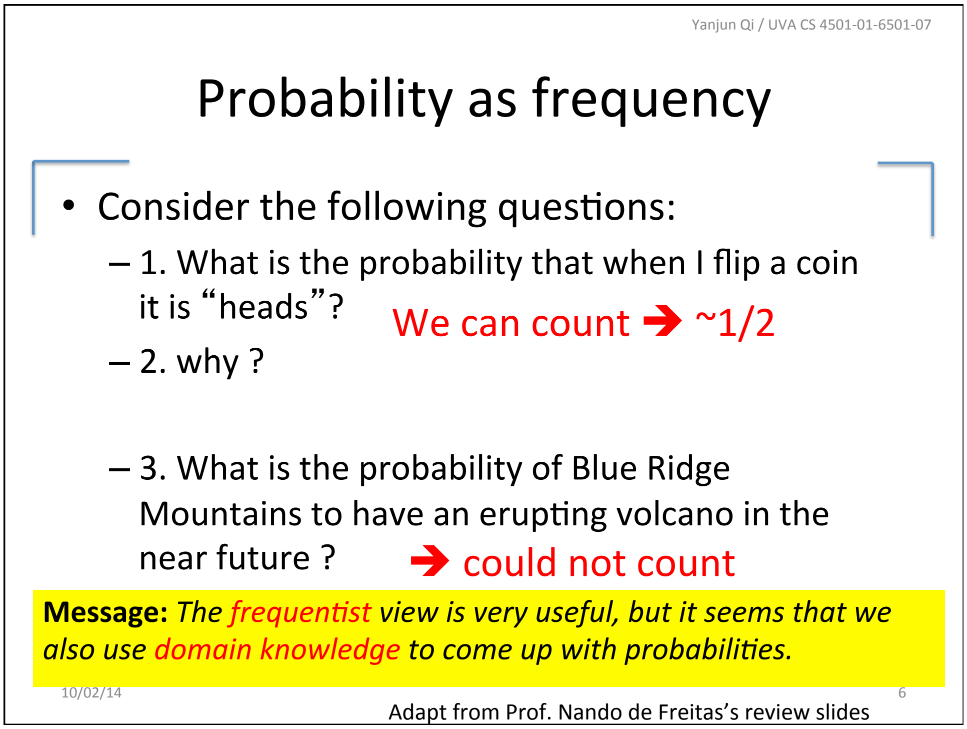
Probability

Model
i.e. Data generating process

Observed Data

Estimation / learning / Inference / Data mining

But how to specify a model?

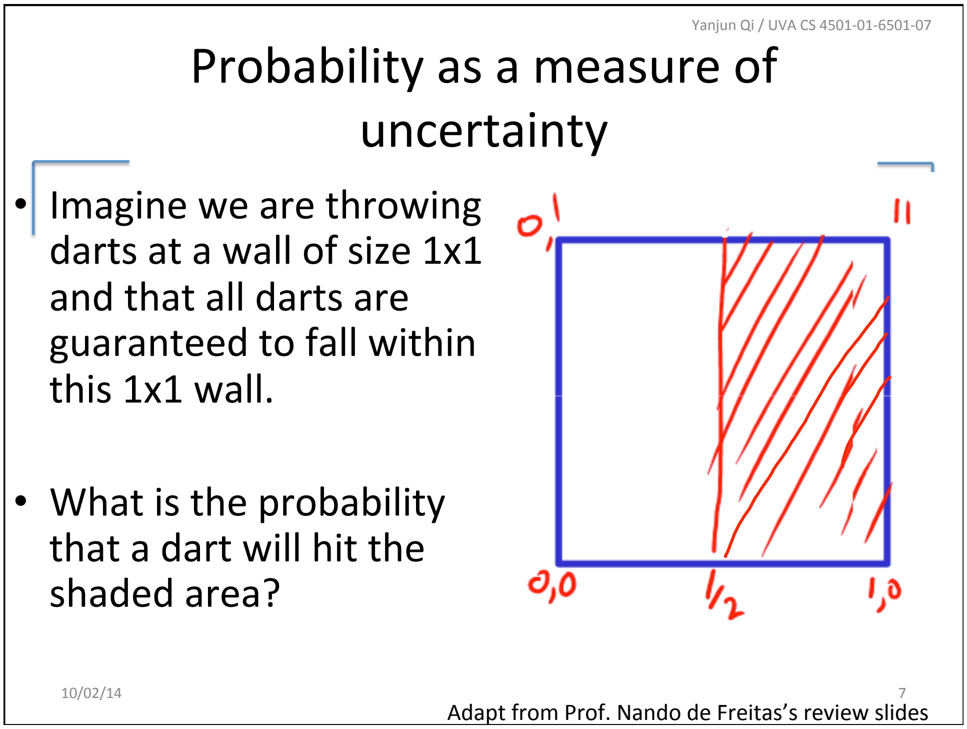


Probability as frequency

- Consider the following questions:
  - 1. What is the probability that when I flip a coin it is “heads”? We can count ➔ ~1/2
  - 2. why ?
  - 3. What is the probability of Blue Ridge Mountains to have an erupting volcano in the near future ? ➔ could not count

Message: *The frequentist view is very useful, but it seems that we also use domain knowledge to come up with probabilities.*

Adapt from Prof. Nando de Freitas’s review slides



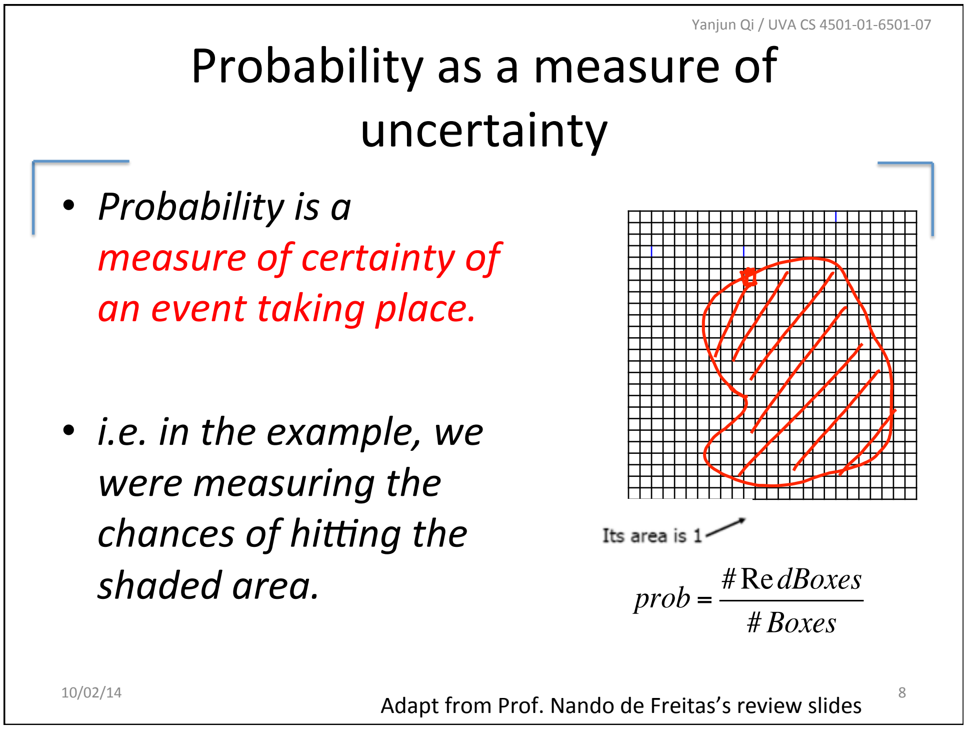
Yanjun Qi / UVA CS 4501-01-6501-07

Probability as a measure of uncertainty

- Imagine we are throwing darts at a wall of size 1x1 and that all darts are guaranteed to fall within this 1x1 wall.
- What is the probability that a dart will hit the shaded area?

10/02/14 7

Adapt from Prof. Nando de Freitas's review slides



Yanjun Qi / UVA CS 4501-01-6501-07

Probability as a measure of uncertainty

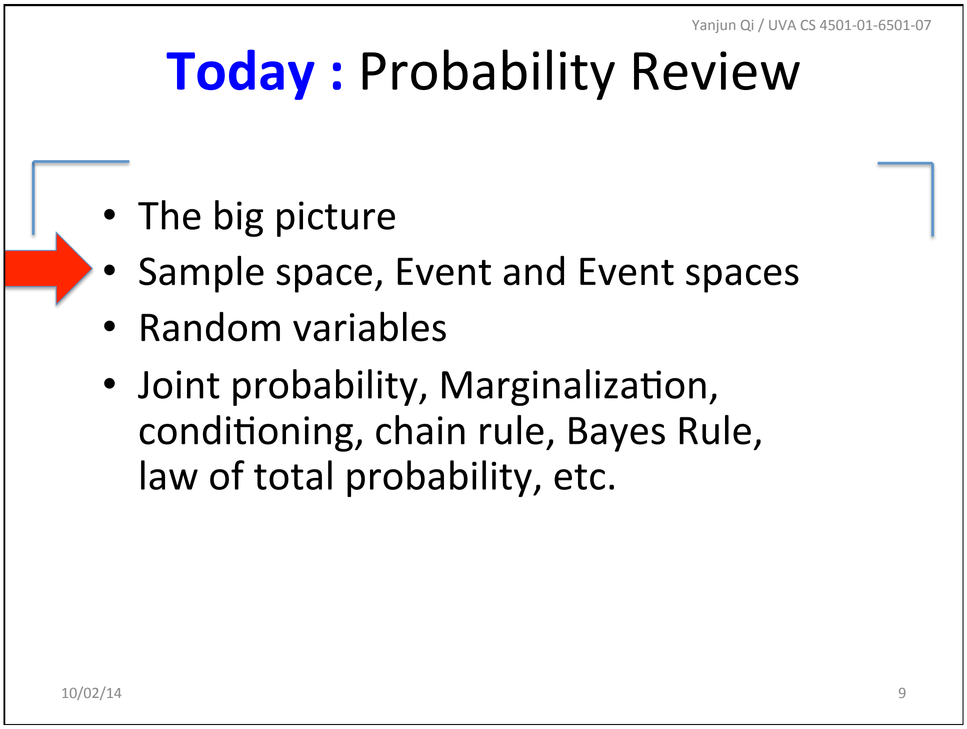
- *Probability is a measure of certainty of an event taking place.*
- *i.e. in the example, we were measuring the chances of hitting the shaded area.*

Its area is 1

$$prob = \frac{\# \mathrm{Re}\, dBoxes}{\# Boxes}$$

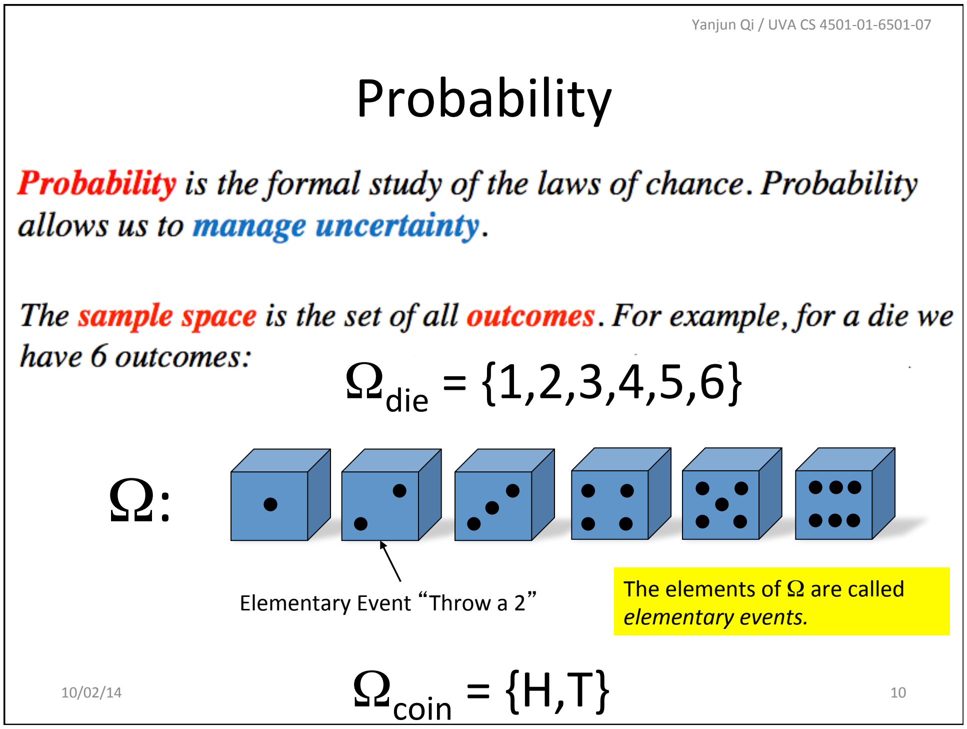
10/02/14 8

Adapt from Prof. Nando de Freitas's review slides



Today : Probability Review

- The big picture
- Sample space, Event and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.



Probability

***Probability** is the formal study of the laws of chance. Probability allows us to **manage uncertainty**.*

*The **sample space** is the set of all **outcomes**. For example, for a die we have 6 outcomes:*

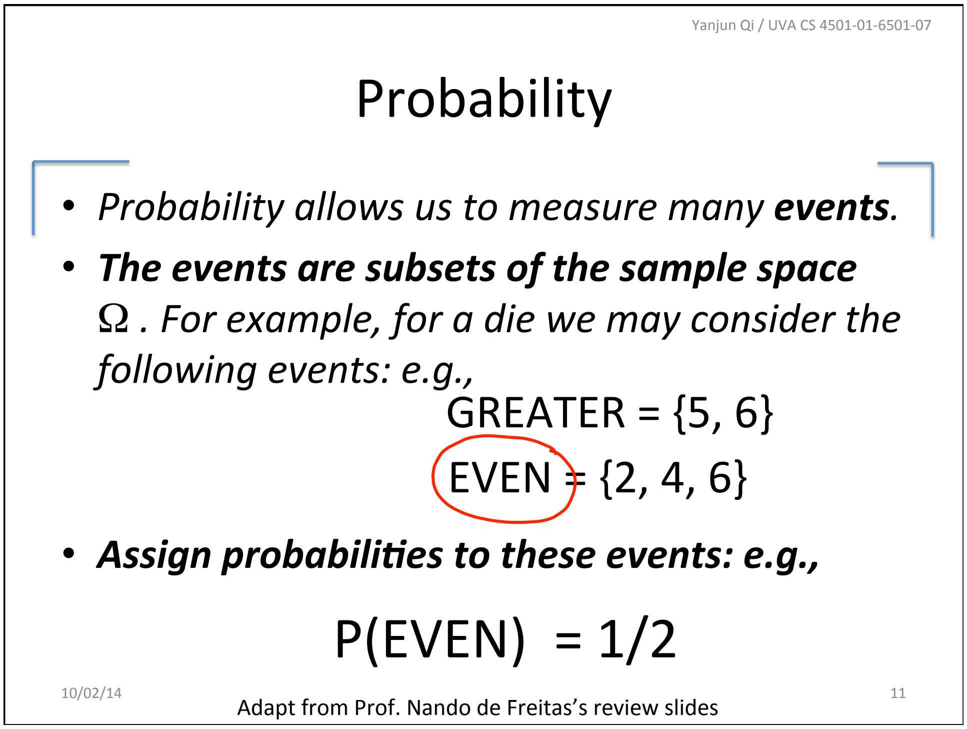
$$\Omega_{die} = \{1,2,3,4,5,6\}$$

Ω:

Elementary Event "Throw a 2"

The elements of Ω are called *elementary events.*

$$\Omega_{coin} = \{H,T\}$$



Probability

- *Probability allows us to measure many **events**.*
- ***The events are subsets of the sample space*** *Ω . For example, for a die we may consider the following events: e.g.,*

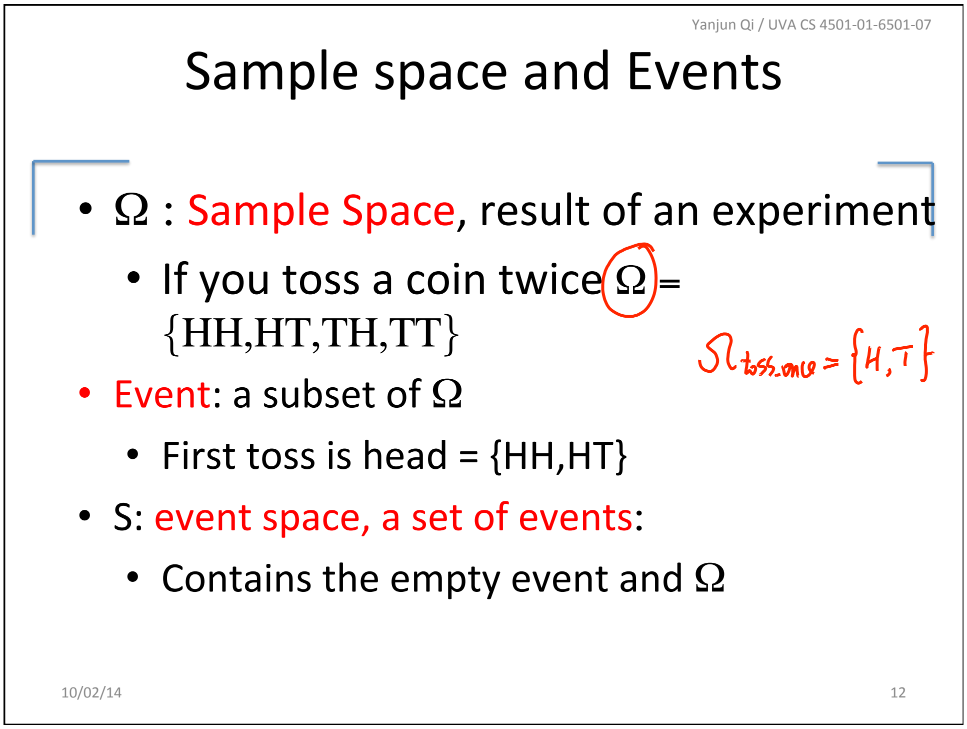
GREATER = {5, 6}

EVEN = {2, 4, 6}

- ***Assign probabilities to these events: e.g.,***

P(EVEN) = 1/2

Adapt from Prof. Nando de Freitas's review slides



Sample space and Events

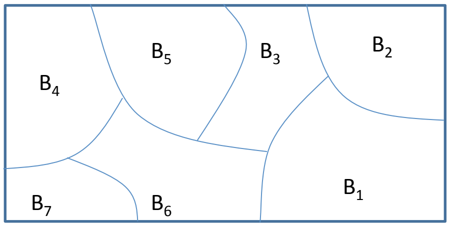
- Ω : Sample Space, result of an experiment
  - If you toss a coin twice Ω = {HH,HT,TH,TT}

$\Omega_{toss\_once} = \{H,T\}$

- Event: a subset of Ω
  - First toss is head = {HH,HT}
- S: event space, a set of events:
  - Contains the empty event and Ω

Axioms for Probability

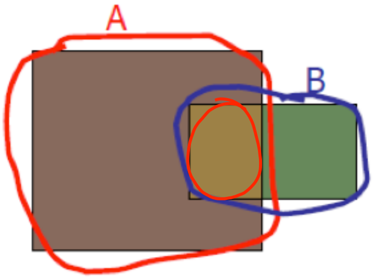
- Defined over (Ω, S) s.t.
  - $1 >= P(\alpha) >= 0$ for all α in S
  - $P(\Omega) = 1$
  - If A, B are disjoint, then
    - $P(A \cup B) = p(A) + p(B)$
- $P(\Omega) = \sum P(B_i)$

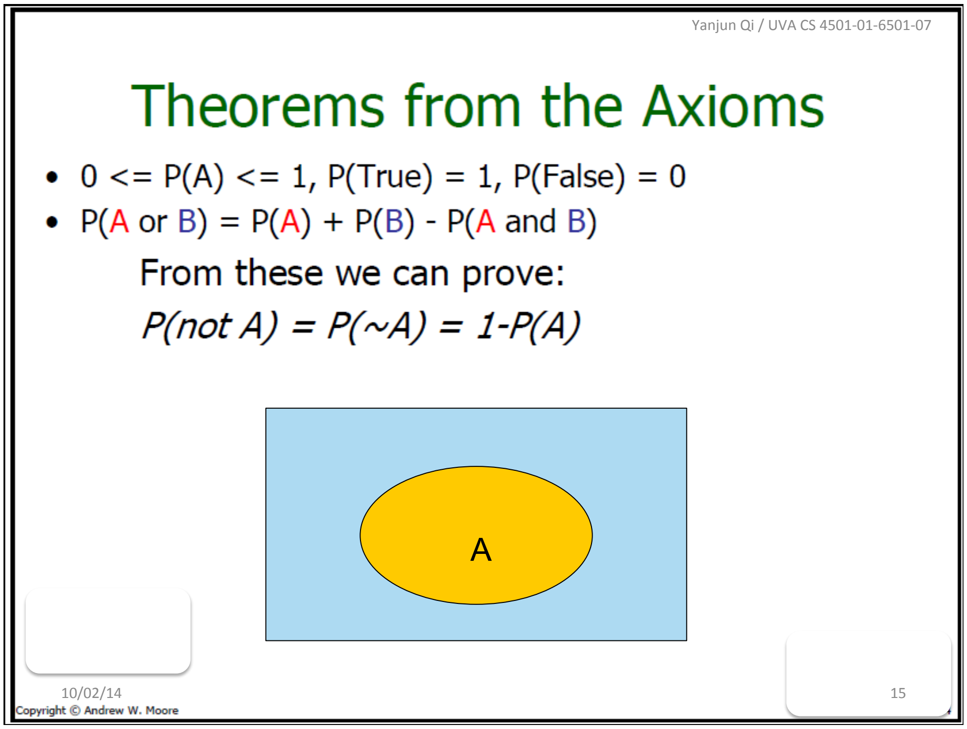


OR operation for Probability

- We can deduce other axioms from the above ones
  - Ex: $P(A \cup B)$ for non-disjoint events

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





Yanjun Qi / UVA CS 4501-01-6501-07

Theorems from the Axioms

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

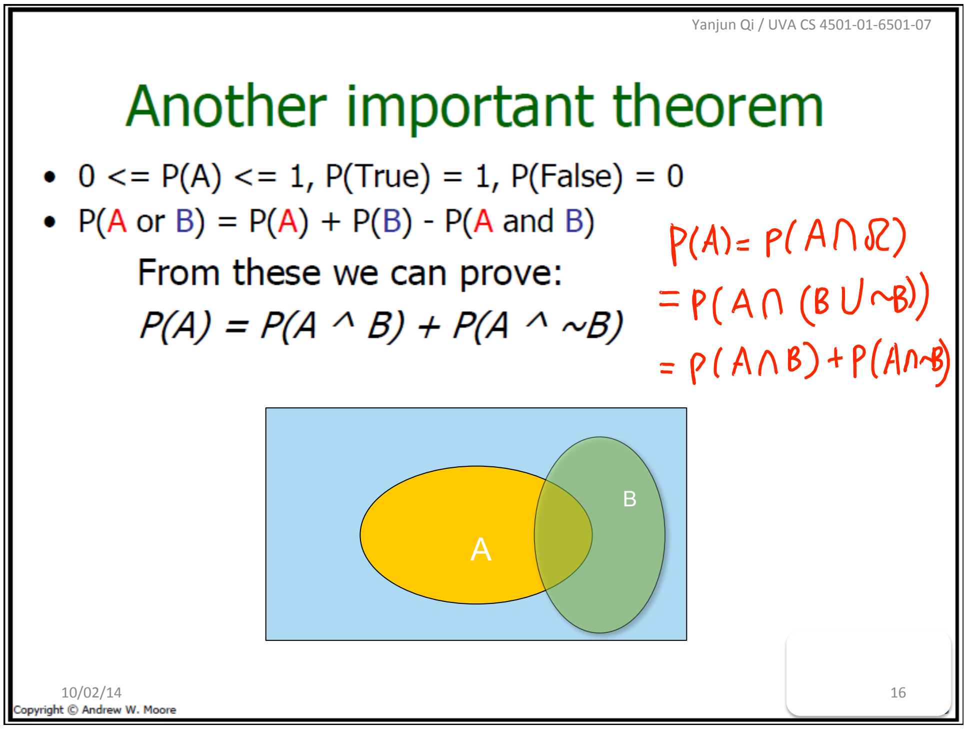
From these we can prove:

P(not A) = P(~A) = 1-P(A)

10/02/14

Copyright © Andrew W. Moore

15



Yanjun Qi / UVA CS 4501-01-6501-07

Another important theorem

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

From these we can prove:

P(A) = P(A ^ B) + P(A ^ ~B)

$P(A) = P(A \cap \Omega)$

$= P(A \cap (B \cup \sim B))$

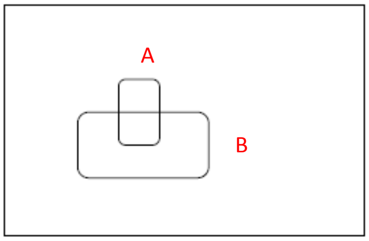
$= P(A \cap B) + P(A \cap \sim B)$

10/02/14

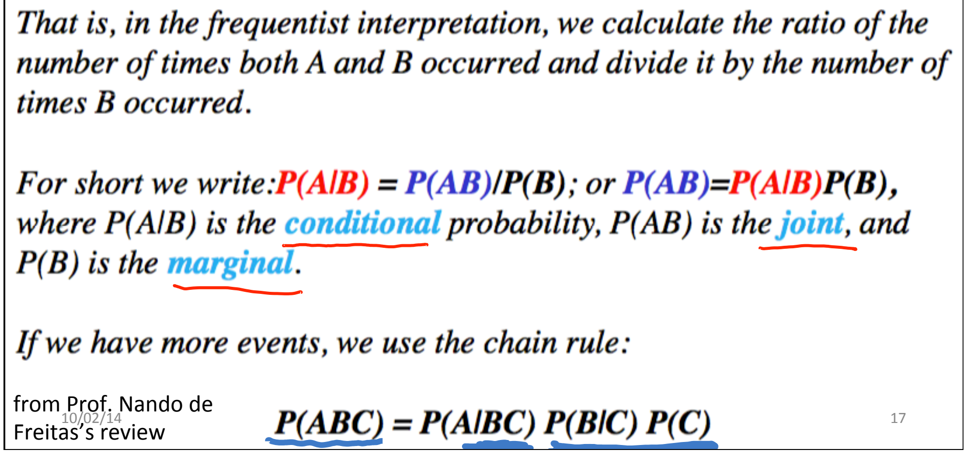
Copyright © Andrew W. Moore

16

Conditional Probability



$P(A \text{ given } B) = P(A \text{ and } B) \,/\, P(B)$



That is, in the frequentist interpretation, we calculate the ratio of the number of times both A and B occurred and divide it by the number of times B occurred.

For short we write: $P(A|B) = P(AB)/P(B)$; *or* $P(AB)=P(A|B)P(B)$, *where P(A|B) is the* ***conditional*** *probability, P(AB) is the* ***joint****, and P(B) is the* ***marginal****.*

If we have more events, we use the chain rule:

$$\underline{P(ABC)} = P(A|\underline{BC})\ \underline{P(B|C)\ P(C)}$$

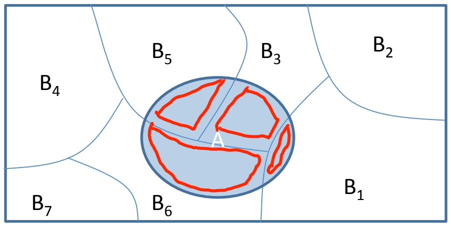
from Prof. Nando de Freitas's review

Conditional Probability / Chain Rule

- More ways to write out chain rule …

$$P(A,B) = \underline{p(B|A)}p(A)$$
$$P(A,B) = \underline{p(A|B)p(}B)$$

Rule of total probability => Marginalization



$$p(A) = \sum P(B_i) P(A \mid B_i)$$

WHY ???

$$P(A) = P(A \cap \Omega) = P(A \cap (B_1 \cup B_2 \cdots \cup B_K))$$
$$= P((A \wedge B_1) \cup (A \wedge B_2) \cup (A \wedge B_3) \ldots \cup (A \wedge B_K))$$
$$= P(A \wedge B_1) + P(A \wedge B_2) + \ldots + P(A \wedge B_K)$$
$$= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \ldots + P(B_K)P(A|B_K)$$

Today : Probability Review

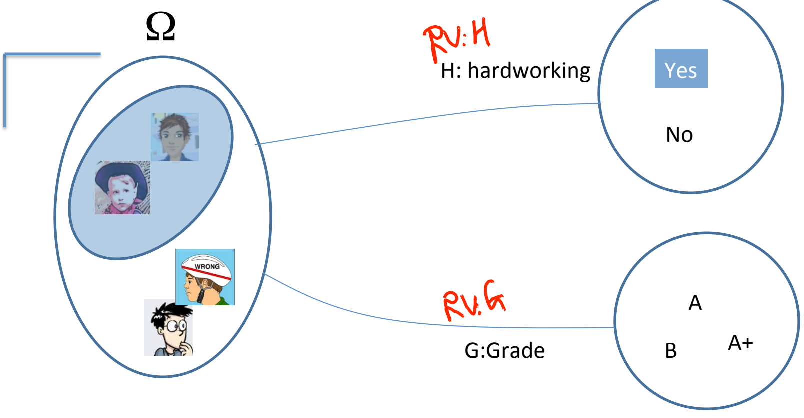
- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.

From Events to Random Variable

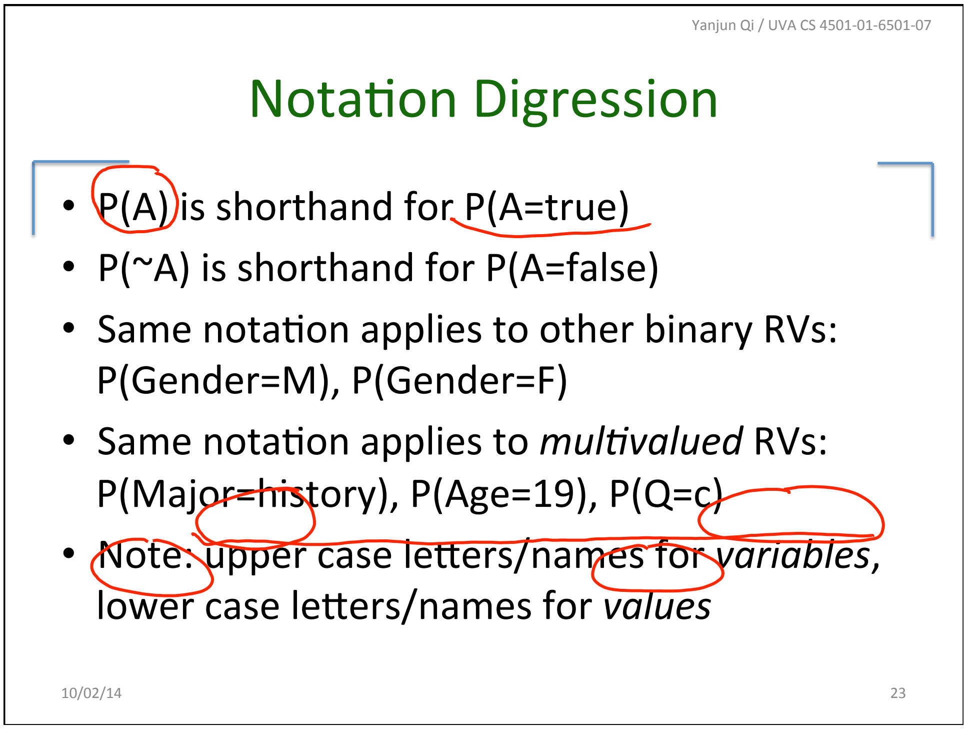
- Concise way of specifying attributes of outcomes
- Modeling students (Grade and Intelligence):
  - Ω = all possible students (sample space)
  - What are events (subset of sample space)
    - Grade_A = all students with grade A
    - Grade_B = all students with grade B
    - HardWorking_Yes = … who works hard
  - Very cumbersome
  - Need "functions" that maps from Ω to an attribute space T.
  - P(H = YES) = P({student ϵ Ω : H(student) = YES})

RV

Random Variables (RV)

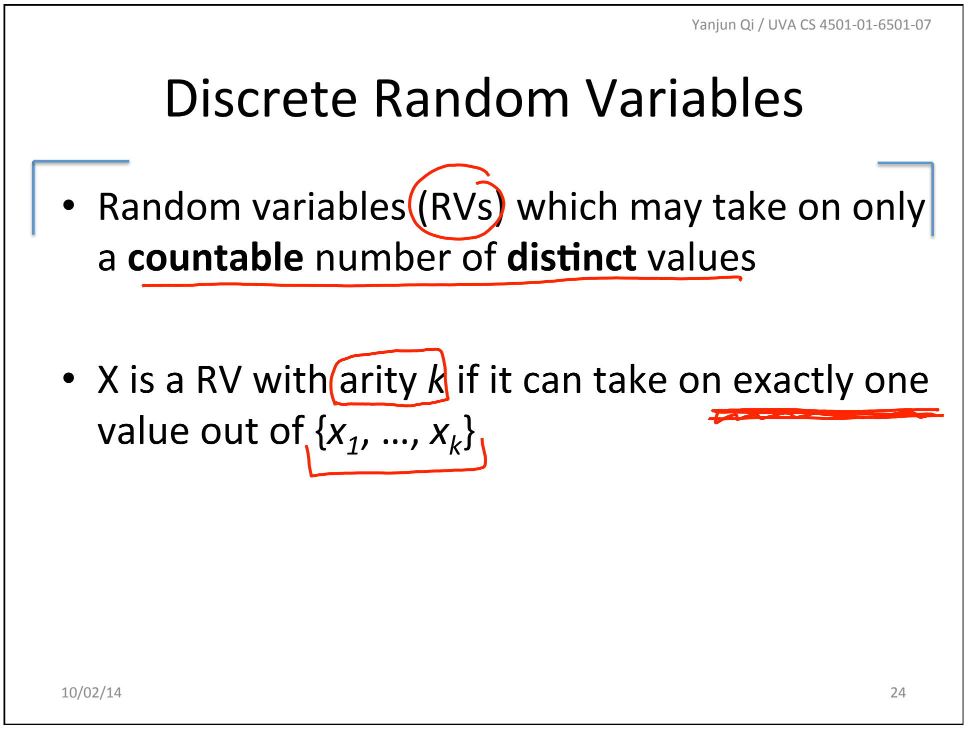


P(H = Yes) = P({all students who is working hard on the course})



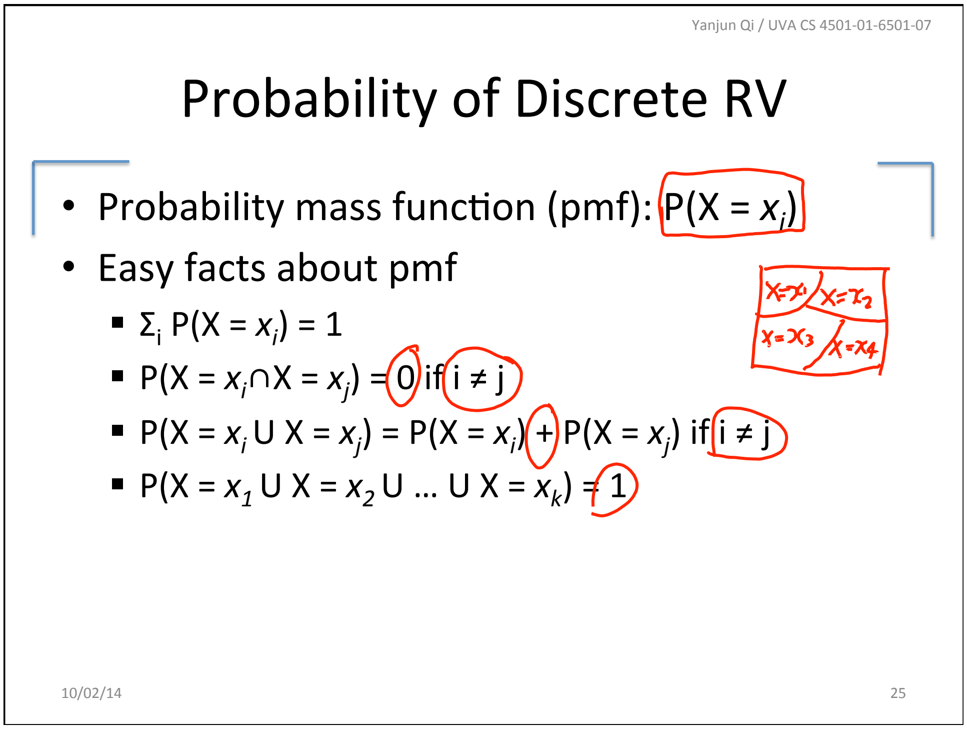
Notation Digression

- P(A) is shorthand for P(A=true)
- P(~A) is shorthand for P(A=false)
- Same notation applies to other binary RVs: P(Gender=M), P(Gender=F)
- Same notation applies to *multivalued* RVs: P(Major=history), P(Age=19), P(Q=c)
- Note: upper case letters/names for *variables*, lower case letters/names for *values*



Discrete Random Variables

- Random variables (RVs) which may take on only a **countable** number of **distinct** values
- X is a RV with arity *k* if it can take on exactly one value out of $\{x_1, \ldots, x_k\}$



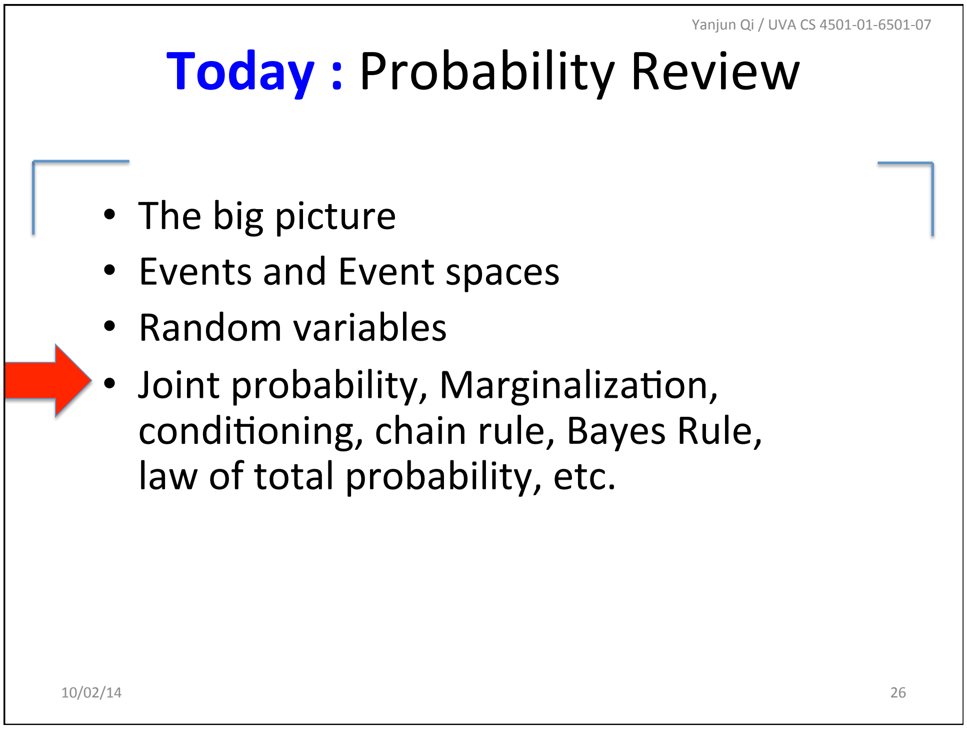
Yanjun Qi / UVA CS 4501-01-6501-07

Probability of Discrete RV

- Probability mass function (pmf): $P(X = x_i)$
- Easy facts about pmf
  - $\Sigma_i\ P(X = x_i) = 1$
  - $P(X = x_i \cap X = x_j) = 0$ if $i \neq j$
  - $P(X = x_i \cup X = x_j) = P(X = x_i) + P(X = x_j)$ if $i \neq j$
  - $P(X = x_1 \cup X = x_2 \cup \ldots \cup X = x_k) = 1$

$X=x_1$ $X=x_2$
$X=x_3$ $X=x_4$

10/02/14 25



Yanjun Qi / UVA CS 4501-01-6501-07

Today : Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.

10/02/14 26

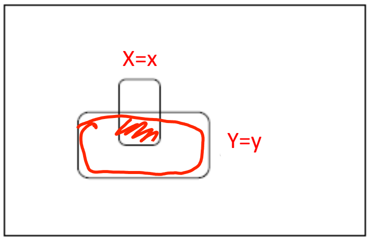
Conditional Probability

events

$$\mathrm{P}\left(\mathrm{X}=\underline{x} \mid \mathrm{Y}=\underline{y}\right)=\frac{\mathrm{P}\left(\mathrm{X}=x \cap \mathrm{Y}=y\right)}{\mathrm{P}\left(\mathrm{Y}=y\right)}$$

But we will always write it this way:

$$P(\underline{x} \mid \underline{y})=\frac{p(x,y)}{p(y)}$$



Marginalization

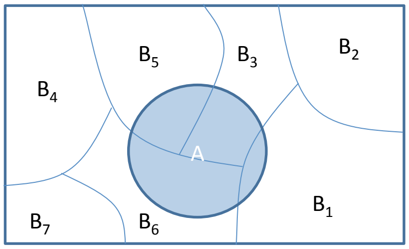
- We know p(X, Y), what is P(X=x)?
- We can use the law of total probability, why?

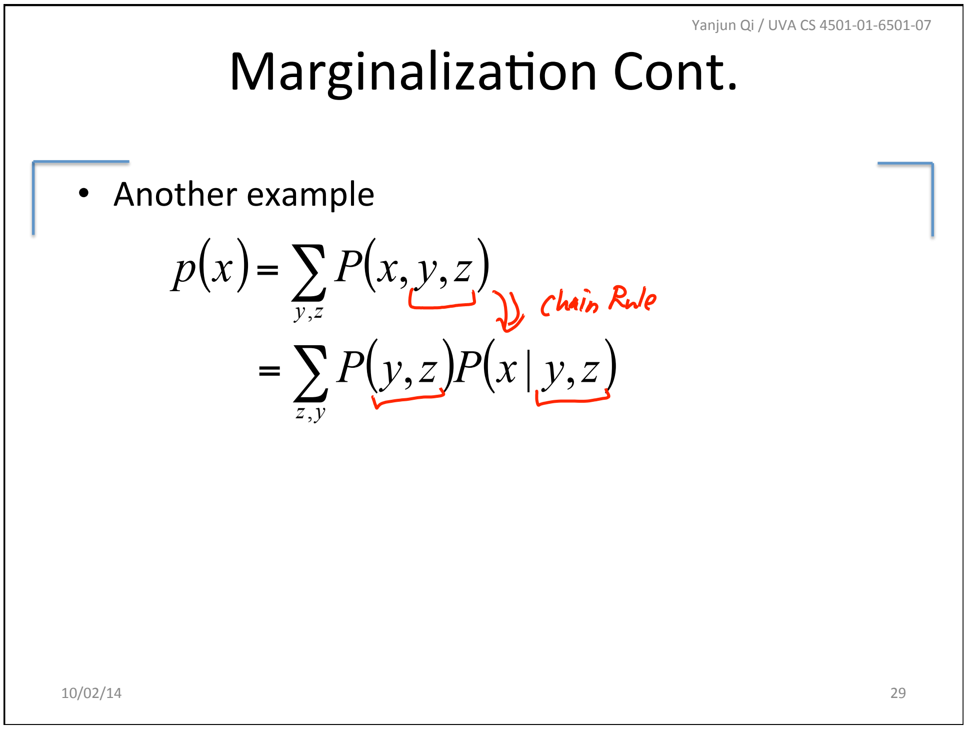
total prob. law

$$\underbrace{p(x)}_{\text{margin Prob.}}=\sum_y P(x,y)$$

⇓ chain rule

$$=\sum_y P(y)P(x \mid y)$$





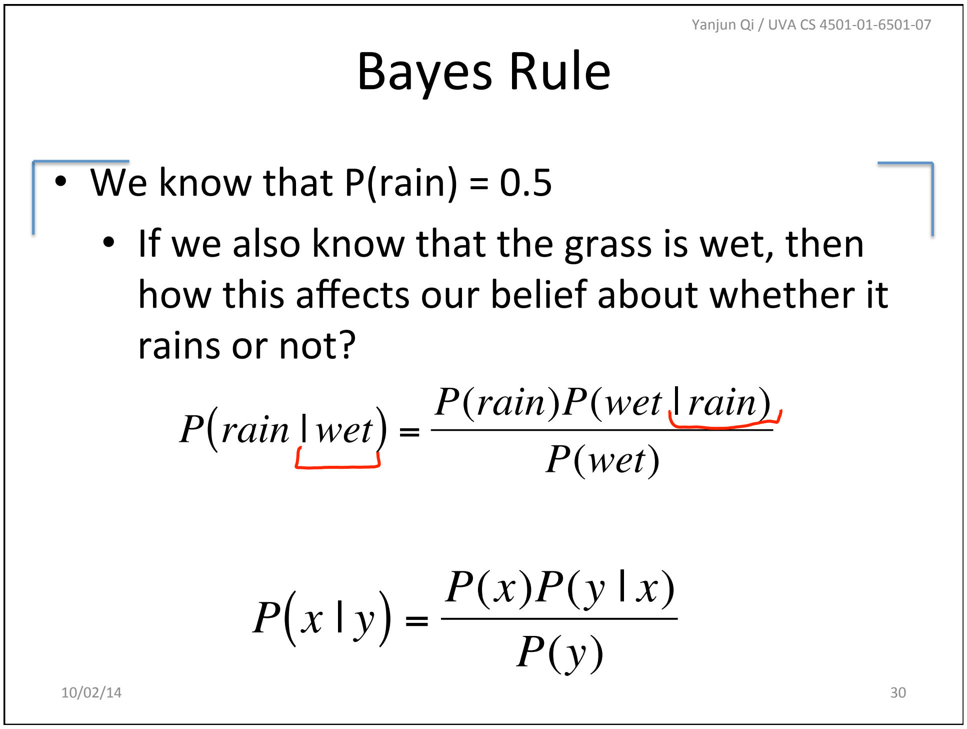
Marginalization Cont.

- Another example

$$p(x)=\sum_{y,z} P(x,\underbrace{y,z})$$

$\Downarrow$ Chain Rule

$$=\sum_{z,y} P(\underbrace{y,z})P(x\mid \underbrace{y,z})$$



Bayes Rule

- We know that P(rain) = 0.5
  - If we also know that the grass is wet, then how this affects our belief about whether it rains or not?

$$P(rain \mid \underbrace{wet}) = \frac{P(rain)P(wet \mid \underbrace{rain})}{P(wet)}$$

$$P(x\mid y) = \frac{P(x)P(y\mid x)}{P(y)}$$

What we just did…

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

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**

More General Forms of Bayes Rule

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

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

$$P(A = a_1 | B) = \frac{P(B | A = \underline{a_1})P(A = \underline{a_1})}{\sum_i P(B | A = \underline{a_i})P(A = \underline{a_i})}$$

Bayes Rule cont.

- You can condition on more variables

$$P(\underbrace{x \mid \underbrace{y}, z}) = \frac{P(x \mid z) P(y \mid x, z)}{P(y \mid z)}$$

Conditional Probability Example

Assume we have a dark box with 3 red balls and 1 blue ball. That is, we have the ***set*** $\underbrace{\{r,r,r,b\}}$. *What is the probability of drawing 2 red balls in the first 2 tries?*

Joint

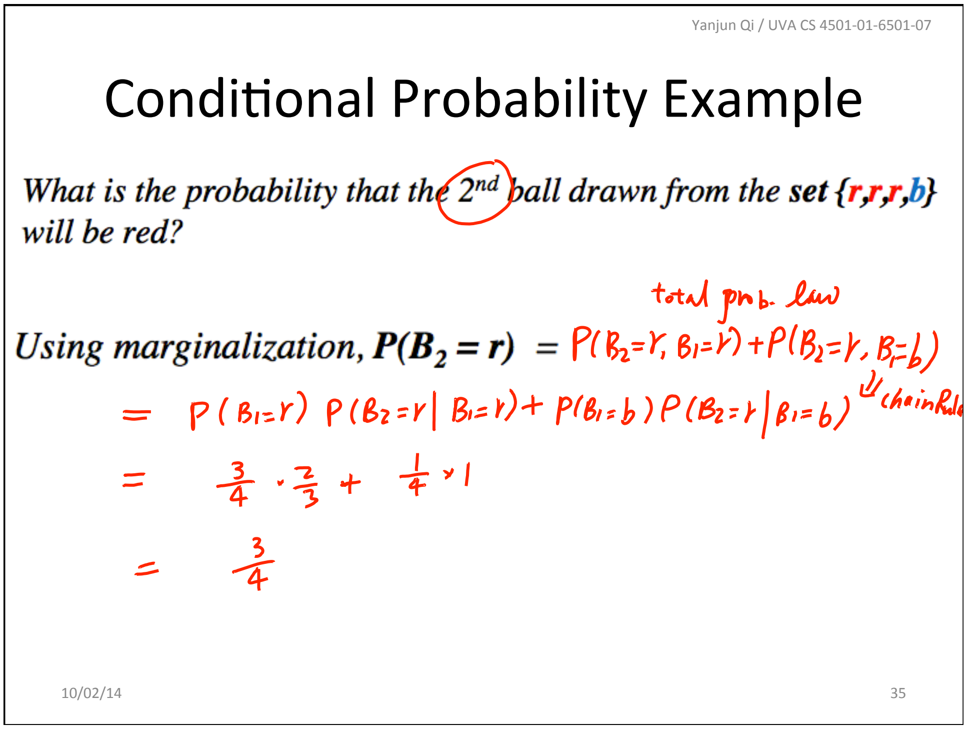
chain Rule

$$\boldsymbol{P(B_1 = r, B_2 = r)} = P(B_1 = r)\, P(B_2 = r \mid B_1 = r)$$

$$= \frac{3}{4} \times \frac{2}{3}$$

$$= \frac{1}{2}$$

Adapt from Prof. Nando de Freitas's review slides



Yanjun Qi / UVA CS 4501-01-6501-07

Conditional Probability Example

What is the probability that the 2nd ball drawn from the set **{r,r,r,b}** *will be red?*

total prob. law

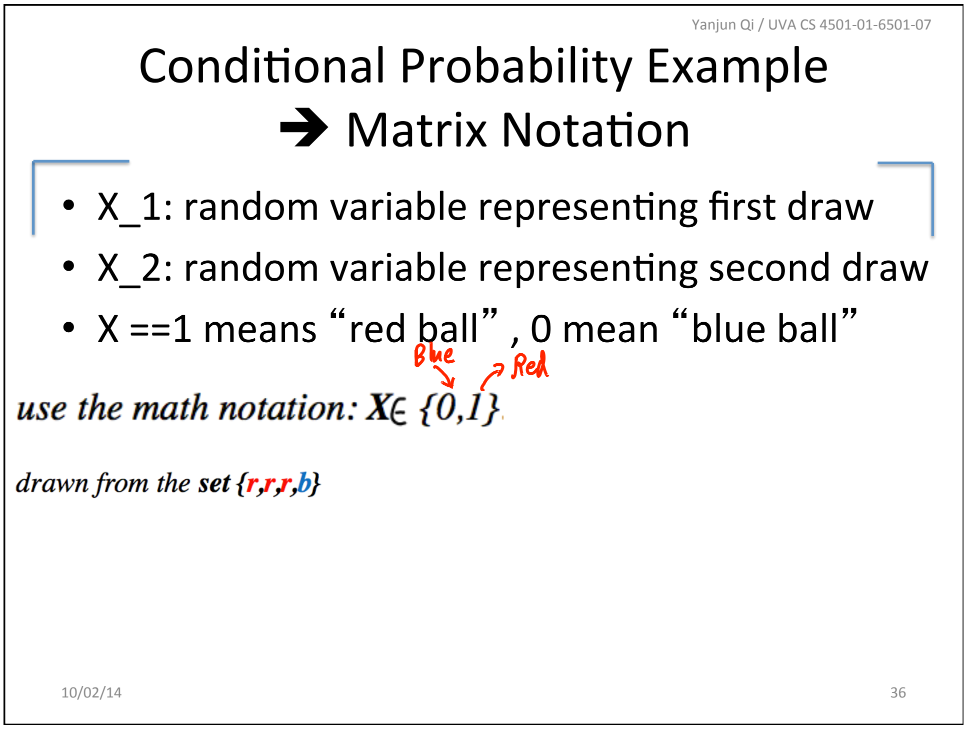
Using marginalization, $\mathbf{P(B_2 = r)} = P(B_2=r, B_1=r) + P(B_2=r, B_1=b)$

$= P(B_1=r)\, P(B_2=r \mid B_1=r) + P(B_1=b)\, P(B_2=r \mid B_1=b)$ chainRule

$= \frac{3}{4} \cdot \frac{2}{3} + \frac{1}{4} \times 1$

$= \frac{3}{4}$

10/02/14 35



Yanjun Qi / UVA CS 4501-01-6501-07

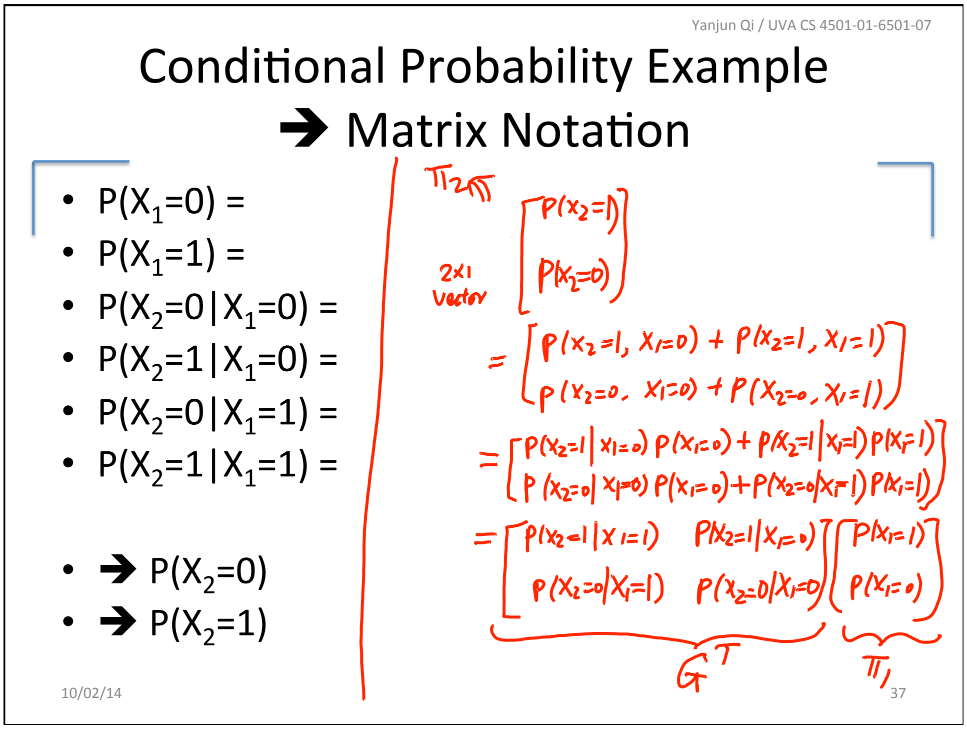
Conditional Probability Example ➔ Matrix Notation

- X_1: random variable representing first draw
- X_2: random variable representing second draw
- X ==1 means "red ball" , 0 mean "blue ball"

use the math notation: $\mathbf{X} \in \{\overset{\text{Blue}}{0},\overset{\text{Red}}{1}\}$

drawn from the set **{r,r,r,b}**

10/02/14 36



Yanjun Qi / UVA CS 4501-01-6501-07

Conditional Probability Example → Matrix Notation

- $P(X_1=0) =$
- $P(X_1=1) =$
- $P(X_2=0|X_1=0) =$
- $P(X_2=1|X_1=0) =$
- $P(X_2=0|X_1=1) =$
- $P(X_2=1|X_1=1) =$

- → $P(X_2=0)$
- → $P(X_2=1)$

π_2 π

2x1 vector

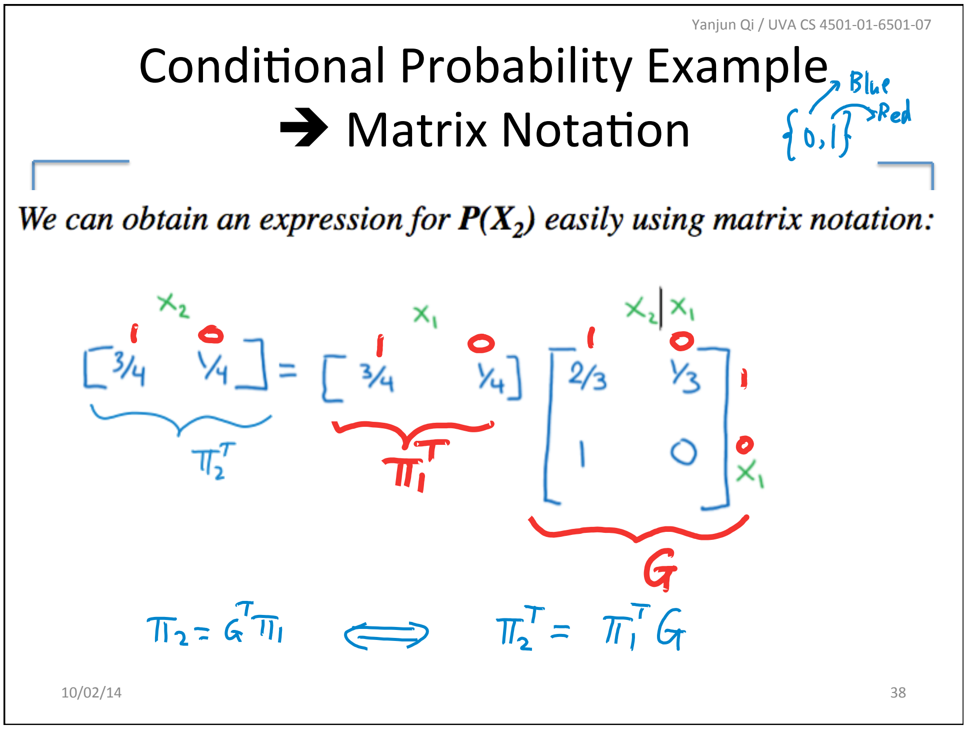
$$\begin{bmatrix} P(X_2=1) \\ P(X_2=0) \end{bmatrix}$$

$$= \begin{bmatrix} P(X_2=1, X_1=0) + P(X_2=1, X_1=1) \\ P(X_2=0, X_1=0) + P(X_2=0, X_1=1) \end{bmatrix}$$

$$= \begin{bmatrix} P(X_2=1|X_1=0)P(X_1=0) + P(X_2=1|X_1=1)P(X_1=1) \\ P(X_2=0|X_1=0)P(X_1=0) + P(X_2=0|X_1=1)P(X_1=1) \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} P(X_2=1|X_1=1) & P(X_2=1|X_1=0) \\ P(X_2=0|X_1=1) & P(X_2=0|X_1=0) \end{bmatrix}}_{G^T} \underbrace{\begin{bmatrix} P(X_1=1) \\ P(X_1=0) \end{bmatrix}}_{\pi_1}$$

10/02/14 37



Yanjun Qi / UVA CS 4501-01-6501-07

Conditional Probability Example → Matrix Notation

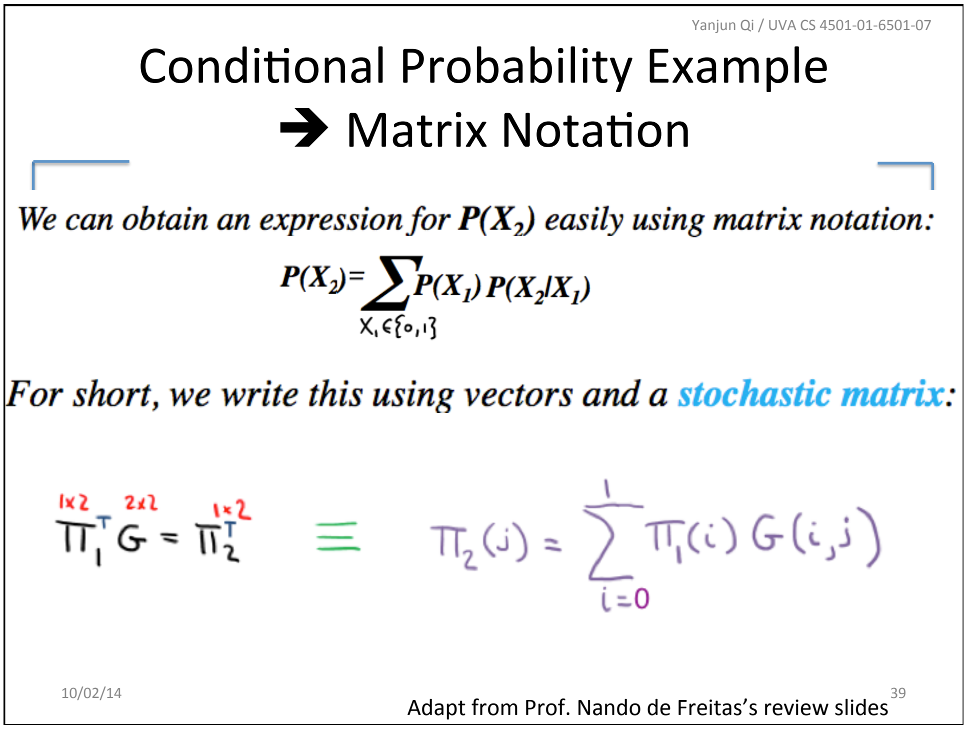
$\{0,1\}$ Blue Red

We can obtain an expression for ***P(X₂)*** *easily using matrix notation:*

$$\underbrace{\begin{array}{cc} \multicolumn{2}{c}{X_2} \\ 1 & 0 \\ \left[\frac{3}{4}\right. & \left.\frac{1}{4}\right] \end{array}}_{\pi_2^T} = \underbrace{\begin{array}{cc} \multicolumn{2}{c}{X_1} \\ 1 & 0 \\ \left[\frac{3}{4}\right. & \left.\frac{1}{4}\right] \end{array}}_{\pi_1^T} \underbrace{\begin{array}{ccc} \multicolumn{2}{c}{X_2|X_1} & \\ 1 & 0 & \\ \frac{2}{3} & \frac{1}{3} & 1 \\ 1 & 0 & 0 \\ & & X_1 \end{array}}_{G}$$

$$\pi_2 = G^T \pi_1 \iff \pi_2^T = \pi_1^T G$$

10/02/14 38



Yanjun Qi / UVA CS 4501-01-6501-07

Conditional Probability Example → Matrix Notation

We can obtain an expression for $P(X_2)$ *easily using matrix notation:*

$$P(X_2) = \sum_{X_1 \in \{0,1\}} P(X_1)\, P(X_2 | X_1)$$

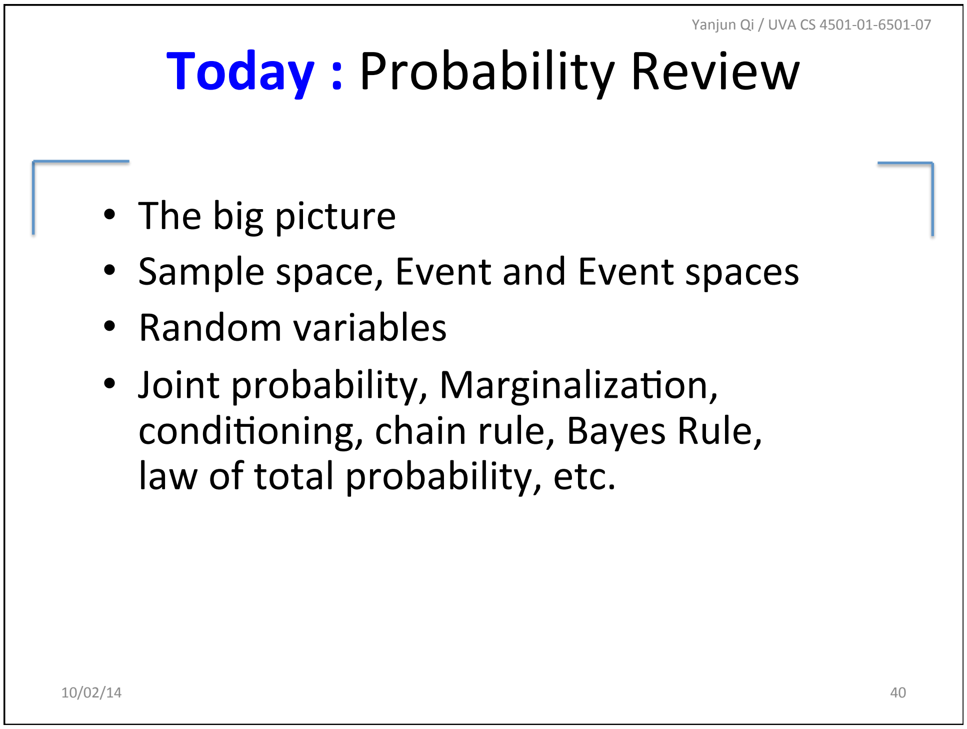
For short, we write this using vectors and a ***stochastic matrix****:*

$$\overset{1\times 2}{\pi_1^T}\,\overset{2\times 2}{G} = \overset{1\times 2}{\pi_2^T} \quad \equiv \quad \pi_2(j) = \sum_{i=0}^{1} \pi_1(i)\, G(i,j)$$

10/02/14

Adapt from Prof. Nando de Freitas's review slides

39



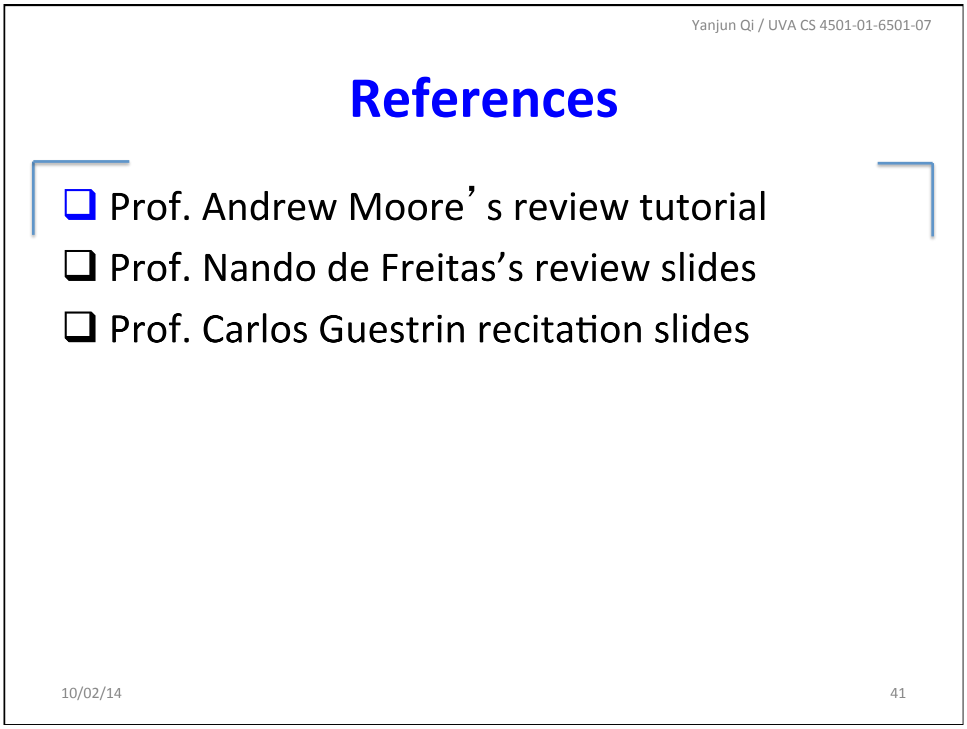
Yanjun Qi / UVA CS 4501-01-6501-07

Today : Probability Review

- The big picture
- Sample space, Event and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.

10/02/14

40



References

- ❑ Prof. Andrew Moore’ s review tutorial
- ❑ Prof. Nando de Freitas’s review slides
- ❑ Prof. Carlos Guestrin recitation slides