POLITECNICO
MILANO 1863

Soft Computing – Probabilistic Reasoning

- Introduction to Probabilistic Reasoning-

Prof. Matteo Matteucci – *matteo.matteucci@polimi.it*

Classes Objectives

"These classes major goal is to provide students with the theoretical background and the practical skills to understand and use Probabilistic Graphical Models, and, at the same time, become familiar and with Probabilistic Reasoning"

⚠ This is the 1st edition of this set of lectures, there will be lectures you'll like and lectures you won't, there'll be topics clearly explained other not, there will be teaching styles you'll enjoy while others will just bore you. Keep with us until the end and help us in improving the course so next edition will be marvelous and unforgettable!

«Me, Myself, and I»

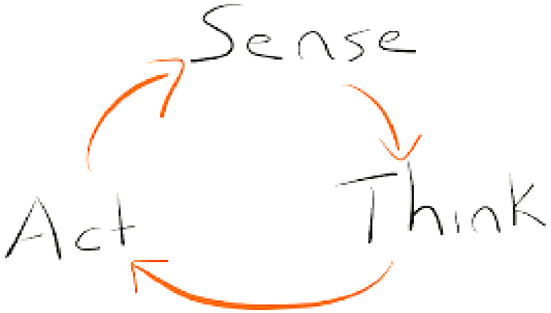
Matteo Matteucci, PhD
Associate Professor
Dept. of Electronics, Information &
Bioengineering
Politecnico di Milano
matteo.matteucci@polimi.it

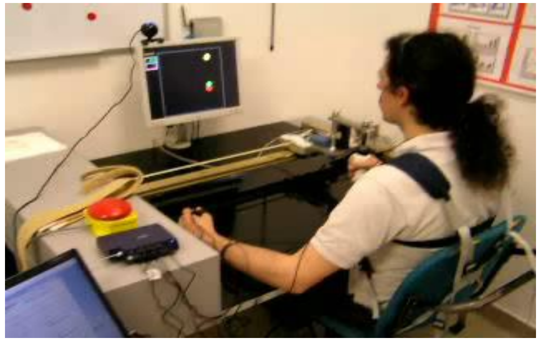
My research interests

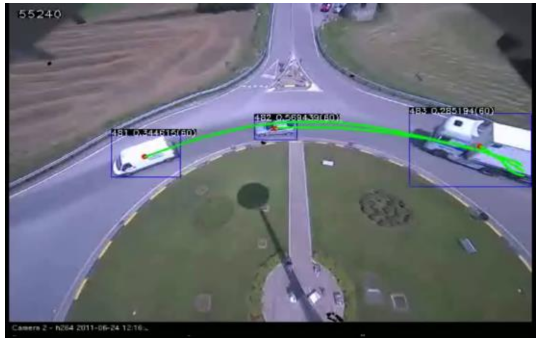
- Robotics & Autonomous Systems
- Machine Learning
- Pattern Recognition
- Computer Vision & Perception

Courses I teach

- Robotics (BS + MS)
- Cognitive Robotics (MS)
- Machine Learning (MS)
- Deep Learning (PhD)





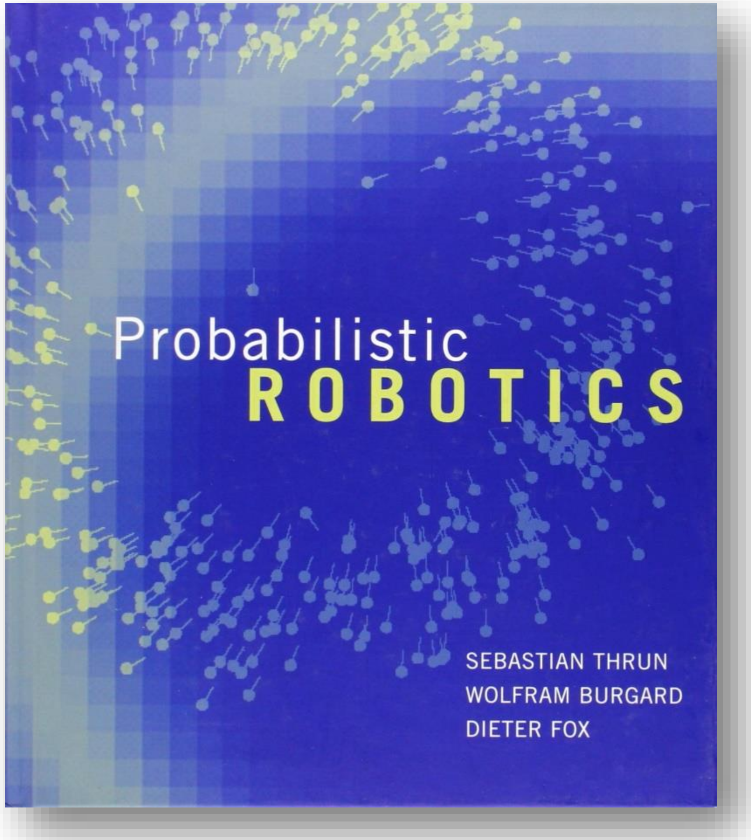


Enable physical and software autonomous systems to perceive, plan, and act without human intervention in the real world

Because the world is a very uncertain place …

Robots and Intelligent Systems have to deal with Uncertainty:

- *Environment*: the physical world is inherently unpredictable
- *Sensors*: inherently limited in what they can perceive due to physics and noise
- *Robots*: actuation involves motors that are, to some extent, unpredictable, due to effects like control noise and wear-and-tear.
- *Models*: inherently inaccurate, they are just abstractions of the real world
- *Computation*: Robots are real-time systems, which limits the amount of computation that can be carried out



Course Syllabus (Tentative)

Slides, in draft will be ready just before the lecture …

Probability basics (fast and furious)

- Frequentists vs Bayesians
- Joint and Naive Distributions

Probabilistic graphical models

- Directed graphical models (Bayesian Networks)
- Conditional independence and d-separation
- Inference in directed graphical models

Would like to discuss Causality, Markov Random Fields, Factor Graphs, etc., but …

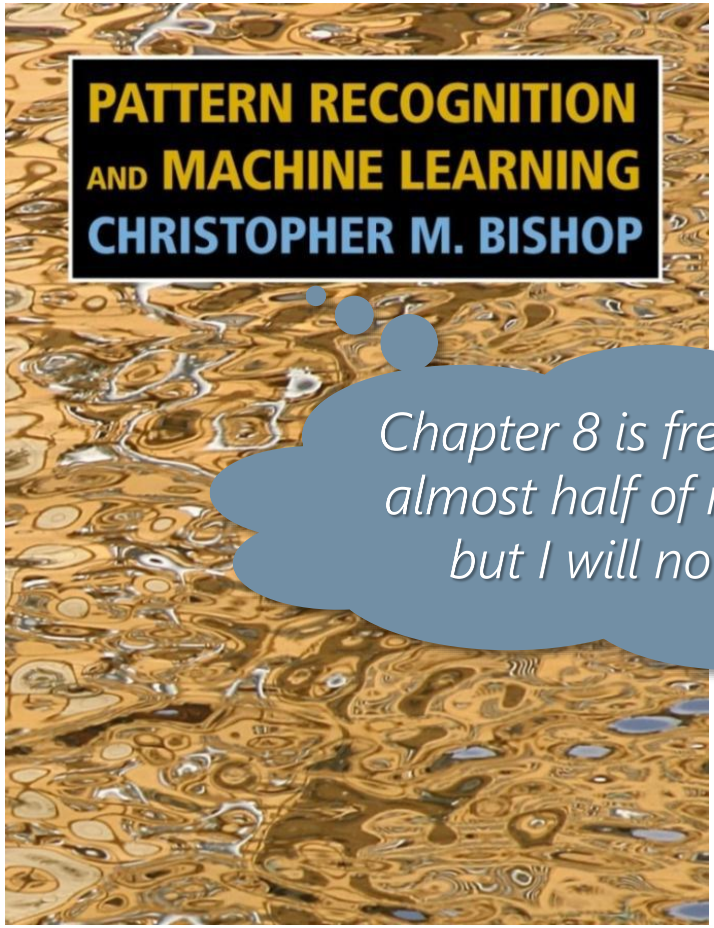
Dynamical graphical models

- Markov chains
- Hidden Markov models

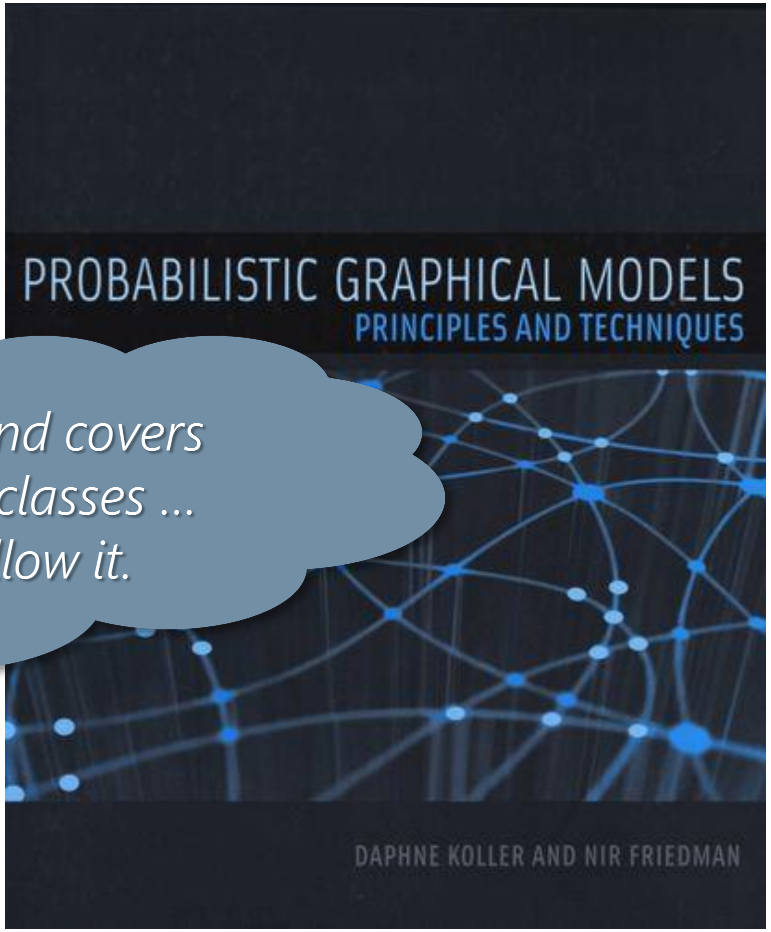
Learning directed graphical models …

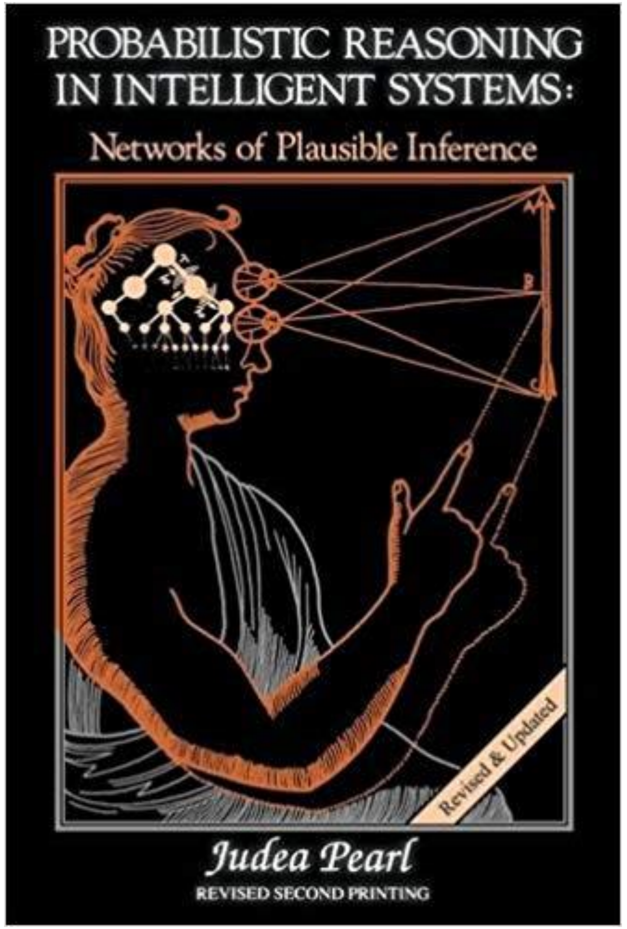
Course Material

Beside *last minute teacher slides* you can refer to many sources:



Chapter 8 is free and covers almost half of my classes … but I will not follow it.


PROBABILISTIC GRAPHICAL MODELS
PRINCIPLES AND TECHNIQUES
DAPHNE KOLLER AND NIR FRIEDMAN


PROBABILISTIC REASONING IN INTELLIGENT SYSTEMS:
Networks of Plausible Inference
Revised & Updated
Judea Pearl
REVISED SECOND PRINTING

POLITECNICO MILANO 1863

Soft Computing – Probabilistic Reasoning

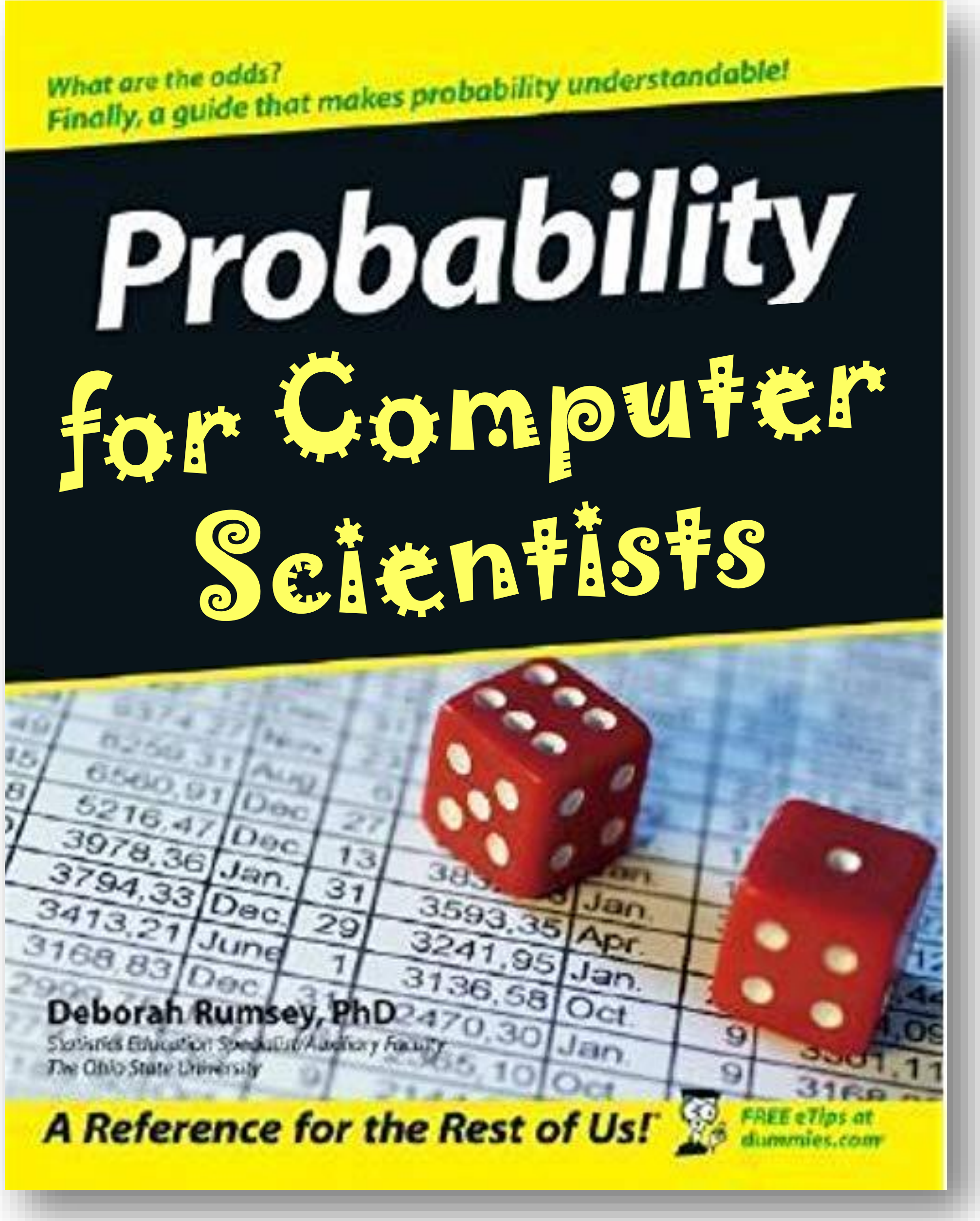
- Probability 101 -

Prof. Matteo Matteucci – *matteo.matteucci@polimi.it*

Just a Reminder …

You all had a class on probability, so the following should be just a simple refresh

- If you find the following brand new, you better revise your Probability & Statistics course notebook
- For sake of easyness I will use discrete variables, but things apply to contiunous distributions too
- If it seems boring and obvious, great, but still keep an eye on the notation …


What are the odds?
Finally, a guide that makes probability understandable!
Probability for Computer Scientists
Deborah Rumsey, PhD
Statistics Education Specialist/Auxiliary Faculty
The Ohio State University
A Reference for the Rest of Us!
FREE eTips at dummies.com

Probability and Boolean Random Variables

Boolean-valued random variable A is a Boolean-valued random variable if A denotes an event, and there is some degree of uncertainty as to whether A occurs.

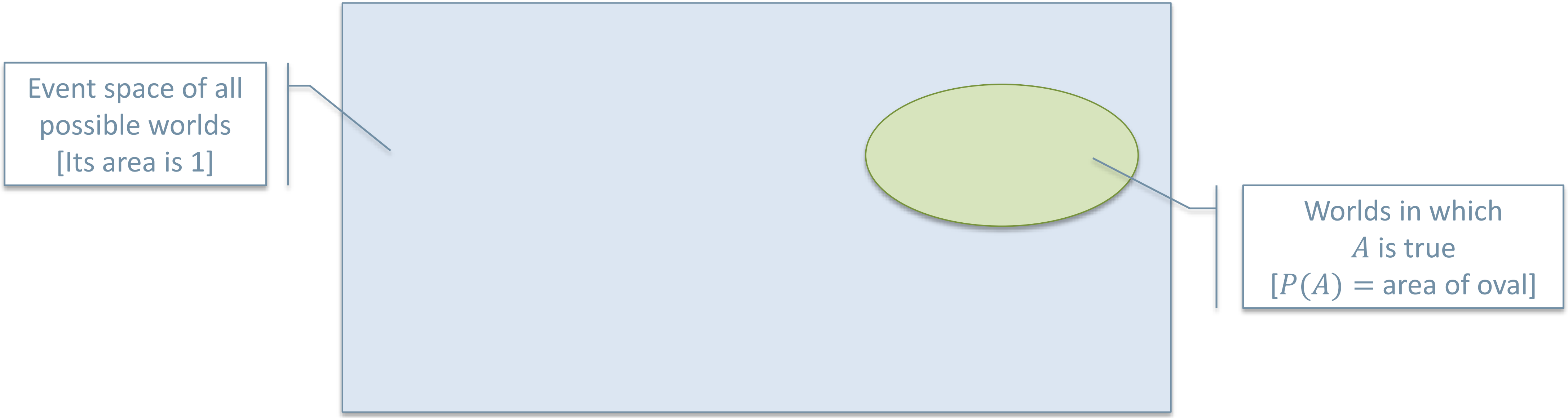
Examples:

- A = The US president in 2023 will be male
- A = You wake up tomorrow with a headache
- A = You like the "Gladiator" movie

Probability and Boolean Random Variables

Boolean-valued random variable $\boldsymbol{A}$ is a Boolean-valued random variable if $\boldsymbol{A}$ denotes an event, and there is some degree of uncertainty as to whether $\boldsymbol{A}$ occurs.

Probability of $\boldsymbol{A}$ "the fraction of possible worlds in which A is true"



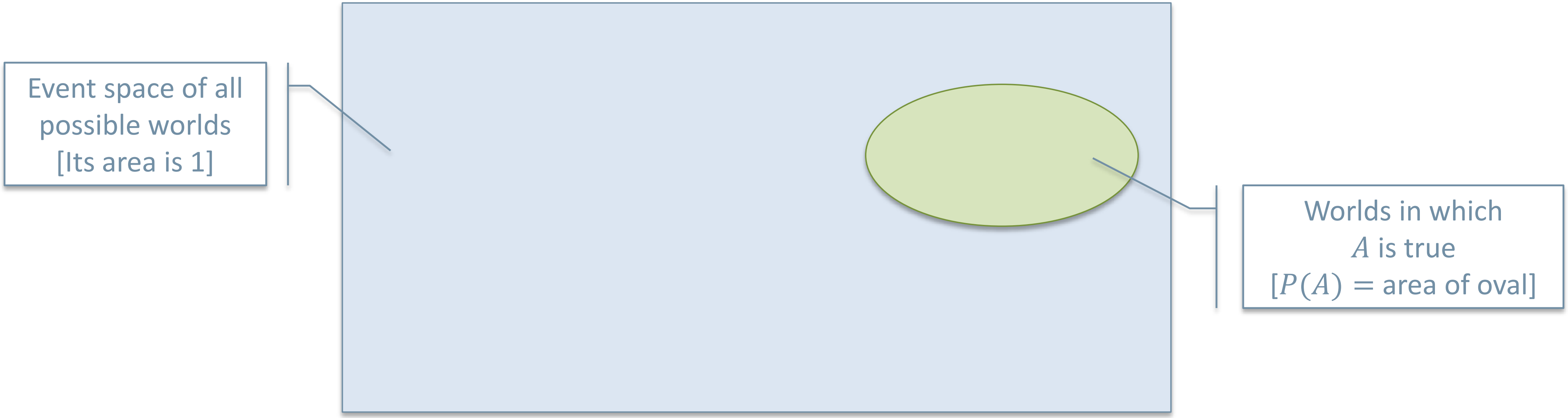
Event space of all possible worlds [Its area is 1]

Worlds in which A is true [$P(A) =$ area of oval]

Probability Axioms

Define the whole set of possible worlds with the label $\boldsymbol{TRUE}$ and the empty set with the label $\boldsymbol{FALSE}$:

- $0 \leq P(A) < 1$
- $P(A = TRUE) = 1$; $P(A = FALSE) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



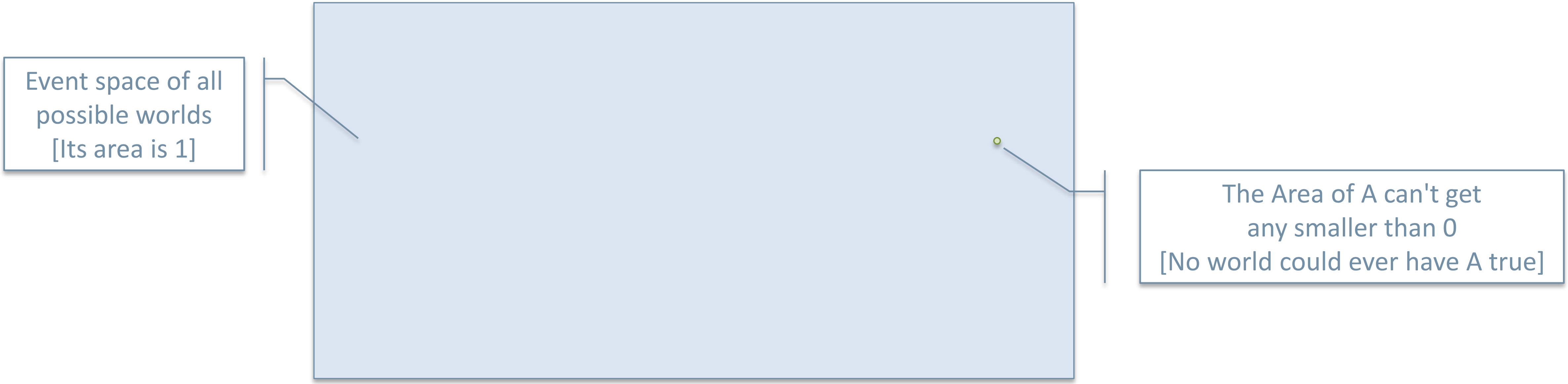
Event space of all possible worlds [Its area is 1]

Worlds in which A is true [$P(A) =$ area of oval]

Probability Axioms

Define the whole set of possible worlds with the label $TRUE$ and the empty set with label $FALSE$:

- $0 \leq P(A) < 1$
- $P(A = TRUE) = 1$; $P(A = FALSE) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Event space of all possible worlds [Its area is 1]

The Area of A can't get any smaller than 0 [No world could ever have A true]

Probability Axioms

Define the whole set of possible worlds with the label $TRUE$ and the empty set with label $FALSE$:

- $0 \leq P(A) < 1$
- $P(A = TRUE) = 1$; $P(A = FALSE) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

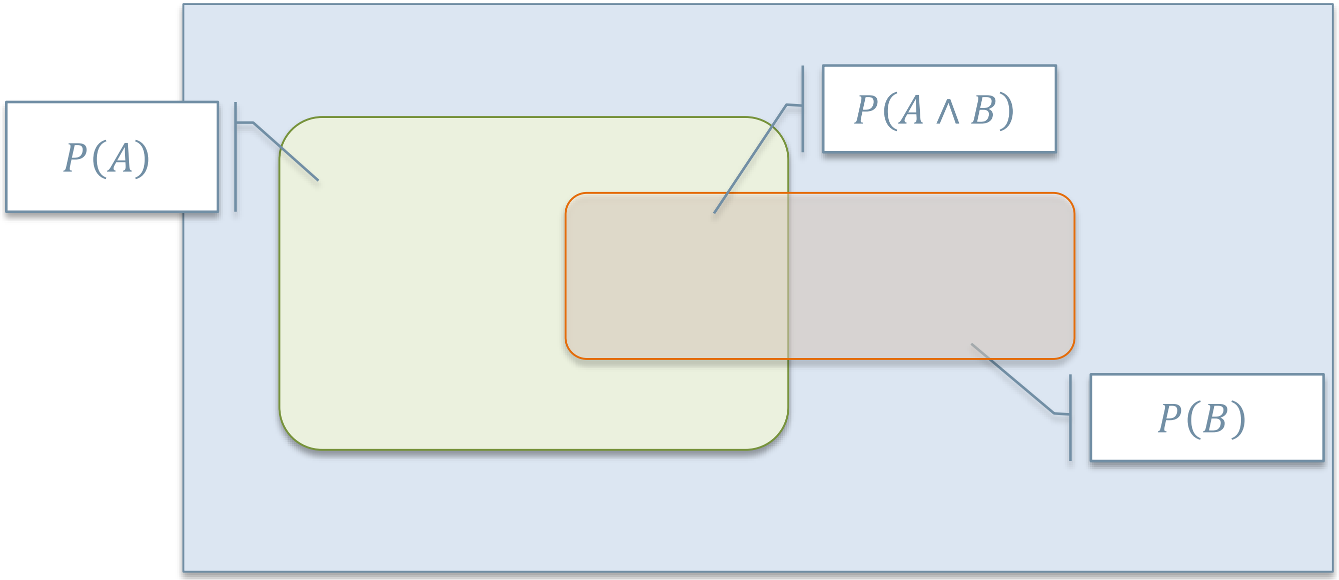
Event space of all possible worlds [Its area is 1]

The Area of A can't get any bigger than 1 [All worlds will have A true]

Probability Axioms

Define the whole set of possible worlds with the label $TRUE$ and the empty set with label $FALSE$:

- $0 \leq P(A) < 1$
- $P(A = TRUE) = 1$; $P(A = FALSE) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Theorems from the Axioms

Using Axioms

- $P(A = TRUE) = 1; P(A = FALSE) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

Prove: $P(\sim A) = P(\bar{A}) = 1 - P(A)$

$$\begin{aligned}
TRUE &= A \vee \bar{A} \\
P(TRUE) &= P(A \vee \bar{A}) \\
&= P(A) + P(\bar{A}) - P(A \wedge \bar{A}) \\
&= P(A) + P(\bar{A}) - P(FALSE) \\
1 &= P(A) + P(\bar{A}) - 0 \\
1 - P(A) &= P(\bar{A})
\end{aligned}$$

Theorems from the Axioms

Using Axioms

- $P(A = TRUE) = 1; P(A = FALSE) = 0$
- $P(A \wedge B) = P(A) + P(B) - P(A \vee B)$

Prove: $P(A) = P(A \wedge B) + P(A \wedge \bar{B})$

$$
\begin{aligned}
A &= A \wedge \mathrm{TRUE} \\
&= A \wedge (B \vee \bar{B}) \\
&= (A \wedge B) \vee (A \wedge \bar{B}) \\
P(A) &= P(A \wedge B) \vee P(A \wedge \bar{B}) \\
&= P(A \wedge B) + P(A \wedge \bar{B}) - P\big((A \wedge B) \wedge (A \wedge \bar{B})\big) \\
&= P(A \wedge B) + P(A \wedge \bar{B}) - P(FALSE) \\
&= P(A \wedge B) + P(A \wedge \bar{B})
\end{aligned}
$$

Multivalued Random Variables

A *multivalued random variable* $\boldsymbol{A}$ is a random variable of arity $\boldsymbol{k}$ if it can take on exactly one values out of $\{v_1, v_2, v_3, \dots, v_k\}$.

We still have the probability axioms plus

- $P(A = v_i \land A = v_j) = 0 \quad if \quad i \neq j$
- $P(A = v_1 \lor A = v_2 \lor A = v_3 \lor \cdots \lor A = v_k) = 1$

Proove: $P(A = v_1 \lor A = v_2 \lor \cdots \lor A = v_i) = \sum_{j=1}^{i} P(A = v_j)$

Proove: $\sum_{j=1}^{k} P(A = v_k) = 1$

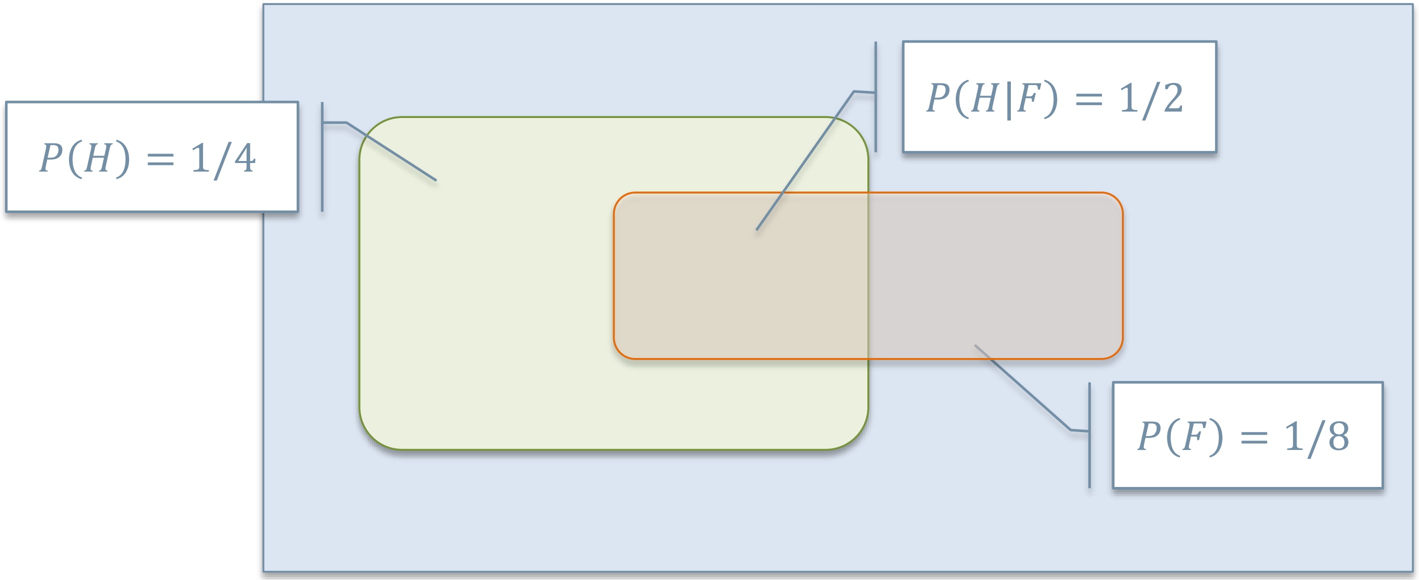
Proove: $P(B \land (A = v_1 \lor A = v_2 \lor \cdots \lor A = v_i)) = \sum_{j=1}^{i} P(B \land A = v_j)$

Proove: $P(B) = \sum_{j=1}^{k} P(B \land A = v_j)$

Conditional Probability

Probability of $\boldsymbol{A}$ *given* $\boldsymbol{B}$ is "the fraction of possible worlds in which $\boldsymbol{B}$ is true that also have $\boldsymbol{A}$ true"

"Sometimes I've the flu (F) and sometimes I've a headache (H), but half of the times I'm with the flu I've also a headache!"



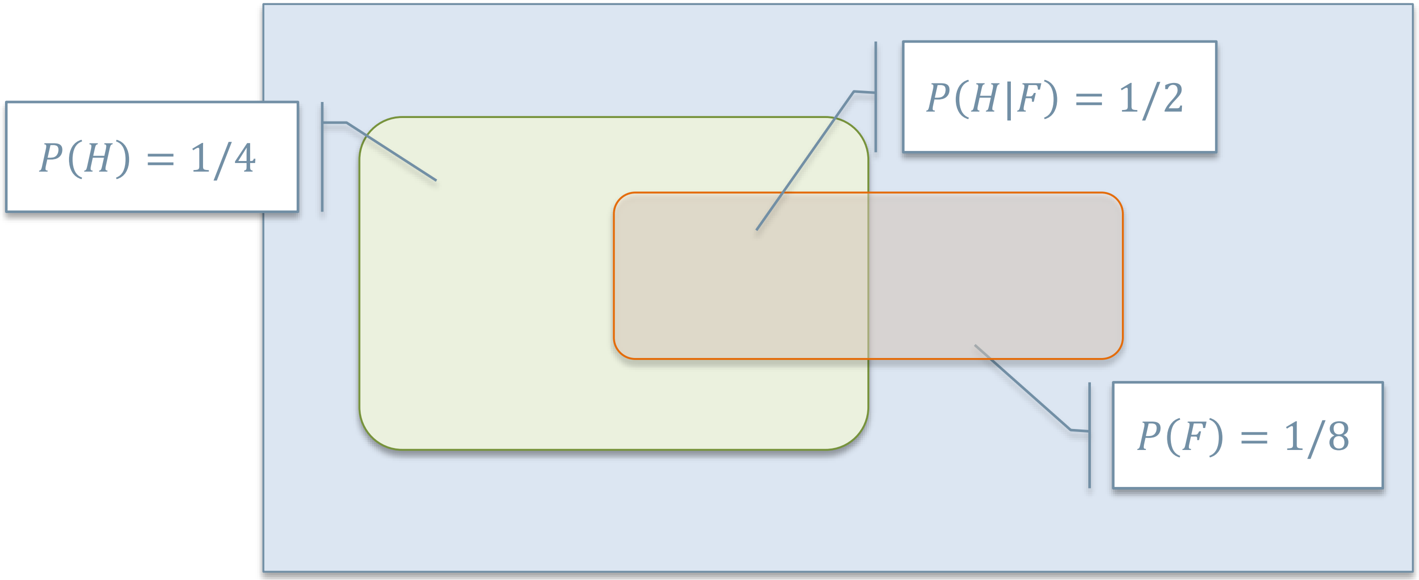
Probabilistic Inference

No!

Is this reasoning correct?

You wake up with a headache and you think: *"Half of the flus come with headaches so I must have 50% chance of getting the flu".*

$$P(F|H) = \frac{\#\ of\ worlds\ with\ F\ and\ H}{\#\ of\ worlds\ with\ H} = \frac{P(F \wedge H)}{P(H)} = \frac{P(H \wedge F)}{P(H)} = \frac{P(H|F) * P(F)}{P(H)} = \frac{\frac{1}{2} * \frac{1}{8}}{\frac{1}{4}} = \frac{1}{4}$$



$P(H) = 1/4$

$P(H|F) = 1/2$

$P(F) = 1/8$

Handy Theorems

In doing the previous inference we have used two famous theorems:

- Chain rule: $P(A \wedge B) = P(A|B) * P(B)$
- Bayes theorem: $P(A|B) = \frac{P(A \wedge B)}{P(B)} = \frac{P(B|A) * P(A)}{P(B)}$
- Sum rule: $\sum_b P(A \wedge B = b) = P(A)$ → $P(A) = \sum_b P(A \wedge B = b)$

We can have more general formulae:

- $P(A|B) = \frac{P(B|A) * P(A)}{P(B|A) * P(A) + P(B|\bar{A}) * P(\bar{A})}$
- $P(A|B \wedge X) = \frac{P(B|A \wedge X) * P(A \wedge X)}{P(B \wedge X)}$
- $P(A = v_i|B) = \frac{P(B|A = v_i) * P(A = v_i)}{P(B)}$

You need to know this stuff by heart!!!

Independent Variables

Independent variables: Assume A and B are boolean random variables; A and B are independent (denote it with $A \perp B$) if and only if:

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

Proove for indempendent variables: $P(A \wedge B) = P(A) * P(B)$

$$\begin{aligned} P(A \wedge B) &= P(A|B) * P(B) \\ &= P(A) * P(B) \end{aligned}$$

Proove for indempendent variables: $P(B|A) = P(B)$

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

POLITECNICO
MILANO 1863

Soft Computing – Probabilistic Reasoning

- A Case for Probabilistic Reasoning-

Prof. Matteo Matteucci – *matteo.matteucci@polimi.it*

What is Probabilistic Reasoning

Example credits to C. Bishop!

A fiendish murder has been committed. Whodunit?

There are two suspects

- The Butler
- The Cook

There are three possible murder weapons

- A butcher's Kife
- A Pistol
- A fireplace Poker

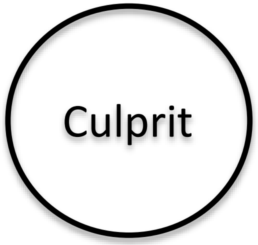
Prior Distribution

Butler has served family well for many years;
Cook, hired recently, rumors of dodgy history

$$P(Culprit = Butler) = 20\%$$

$$P(Culprit = Cook) = 80\%$$

Culprit is a random variable which probabilities add to 100%,



Culprit

$$P(Culprit)$$

$$Culprit\{Butler, Cook\}$$

Conditional Distribution

Butler is ex-army, keeps a gun in a locker drawer;

Cook has access to lot of knives;

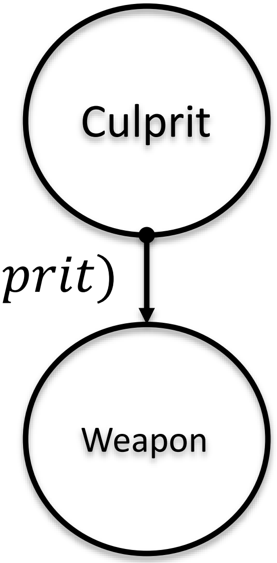
Butler is older, and getting frail.

Which weapon would each of them use?

$$Weapon = \{Pistol, Knife, Poker\}$$

$$P(Weapon|\,Culprit = Butler) = [80\%\ \ 10\%\ \ 10\%]$$

$$P(Weapon|\,Culprit = Cook) = [5\%\ \ 65\%\ \ 30\%]$$



$P(Culprit)$

$P(Weapon|Culprit)$

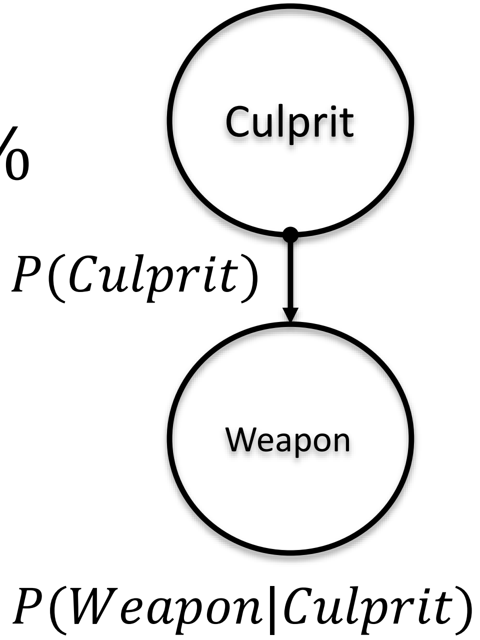
Joint Distribution

What is the probability that the Cook committed the murder using the Pistol?

$$P(Culprit = Cook) = 80\%$$

$$P(Weapon = Pistol|\, Culprit = Cook) = 5\%$$

$$P(Weapon = Pistol, Culprit = Butler) = 80\% \ * 5\% = 4\%$$



The same reasoning can be applied to all other 5 combinations of *Culprit* and *Weapon* …

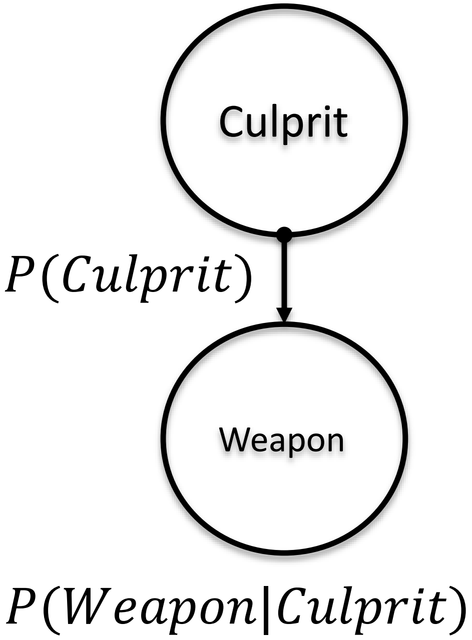
Joint Distribution

What is the probability that the **Cook** committed the murder using the **Pistol**?

	Pistol	Knife	Poker
Cook	4%	52%	24%
Butler	16%	2%	2%

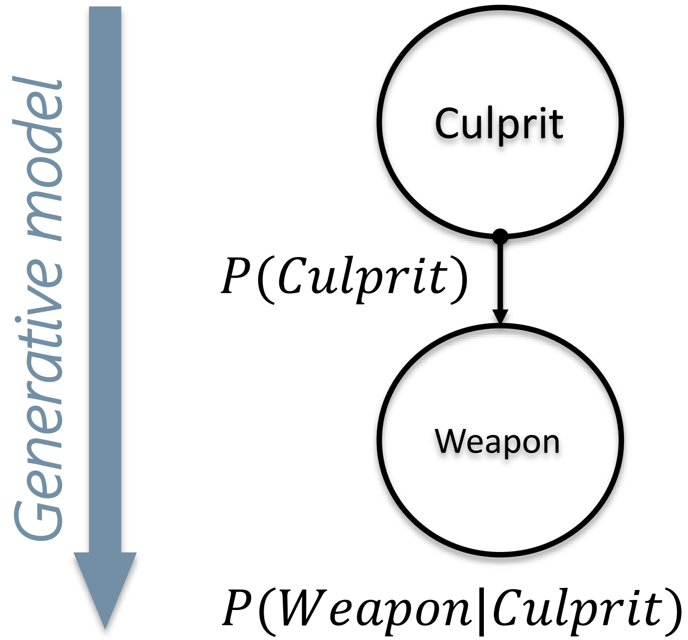
$$P(Weapon, Culprit) = P(Weapon|Culprit) * P(Culprit)$$

Chain Rule: $P(A, B) = P(A|B) * P(B)$



Graphical Models as Data Generators

Graphical models describe a factorization for the Join Distribution, and the process which possibly generates the data …



Murderer	Weapon
Cook	Knife
Butler	Knife
Cook	Pistol
Cook	Poker
Cook	Knife
Butler	Pistol
Cook	Poker
Cook	Knife
Butler	Pistol
Cook	Knife
Cook	Knife
Butler	Pistol
…	…

Marginal Distributions

Culprit marginal distribution

	Pistol	Knife	Poker	
Cook	4%	52%	24%	= 80%
Butler	16%	2%	2%	= 20%

Weapon marginal distribution

	Pistol	Knife	Poker
Cook	4%	52%	24%
Butler	16%	2%	2%
	= 20%	= 54%	= 26%

Sum Rule: $\sum_b P(A, B = b) = P(A)$

Posterior Distribution

We discover a Pistol at the scene of the crime

	Pistol	Knife	Poker	
Cook	4%	~~52%~~	~~24%~~	= ~~80%~~
Butler	16%	~~2%~~	~~2%~~	= ~~20%~~

Looks bad for the Butler

Murderer	Weapon
~~Cook~~	~~Knife~~
~~Butler~~	~~Knife~~
Cook	Pistol
~~Cook~~	~~Poker~~
~~Cook~~	~~Knife~~
Butler	Pistol
~~Cook~~	~~Poker~~
~~Cook~~	~~Knife~~
Butler	Pistol
~~Cook~~	~~Knife~~
~~Cook~~	~~Knife~~
Butler	Pistol
…	…

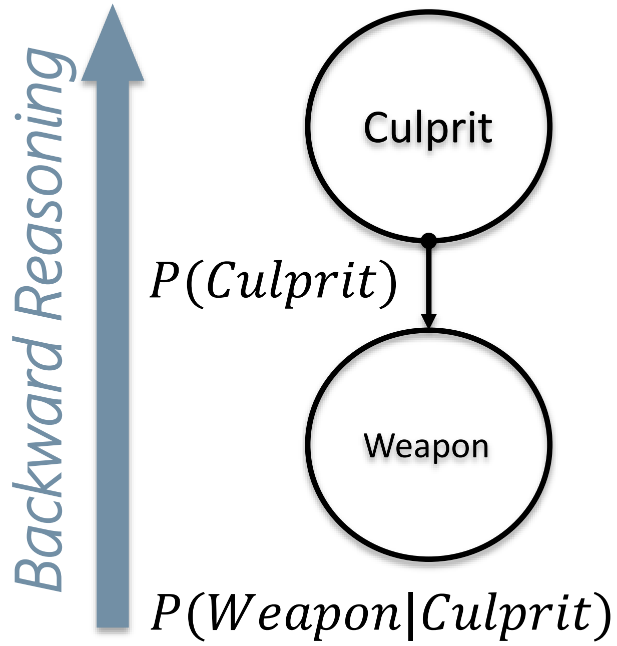
Graphical Models for Backward Reasoning

Graphical models describe a factorization for the Join Distribution, and can be used for back reasoning via Bayes Theorem …

Posterior Likelihood Prior

$$P(Culprit|Weapon) = \frac{P(Weapon|Culprit) * P(Culprit)}{P(Weapon)}$$

$$P(Weapon) = \sum_{c} P(Weapon|Culprit = c) * P(Culprit = c)$$

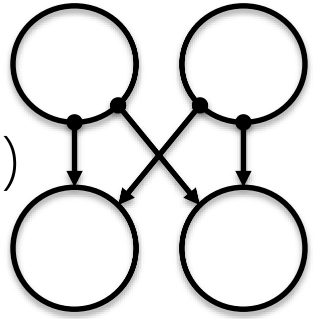


Bayes Theorem: $\mathrm{P(A|B)} = \mathrm{P}(B|A) * P(A)/P(B)$

Probabilistic Graphical Models

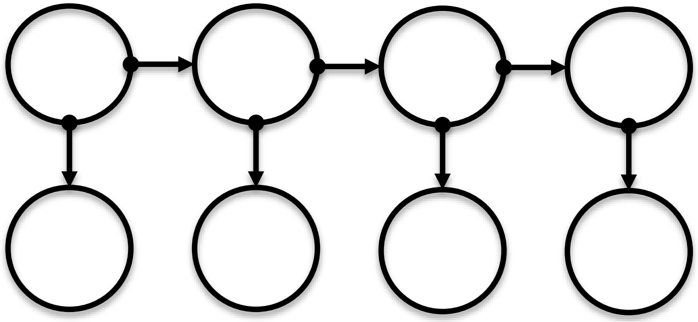
Combine probability theory with graphs

- Graph-based algorithms for calculation and computation (e.g., Feynman diagrams in physics)
- Efficient software implementation
- New insights into existing models
- Framework for designing new models



PCA, ICA,
Factor Analysis
Linear Regression
Logistic Regression
Mixture Models

Kalman Filters
Hidden Markov Models



Three types of graphs:

- Directed graphs (useful for designing models)
- Undirected graphs (good for some domains, e.g. computer vision)
- Factor graphs (useful for inference and learning)

POLITECNICO
MILANO 1863

Soft Computing – Probabilistic Reasoning

- Density Estimation -

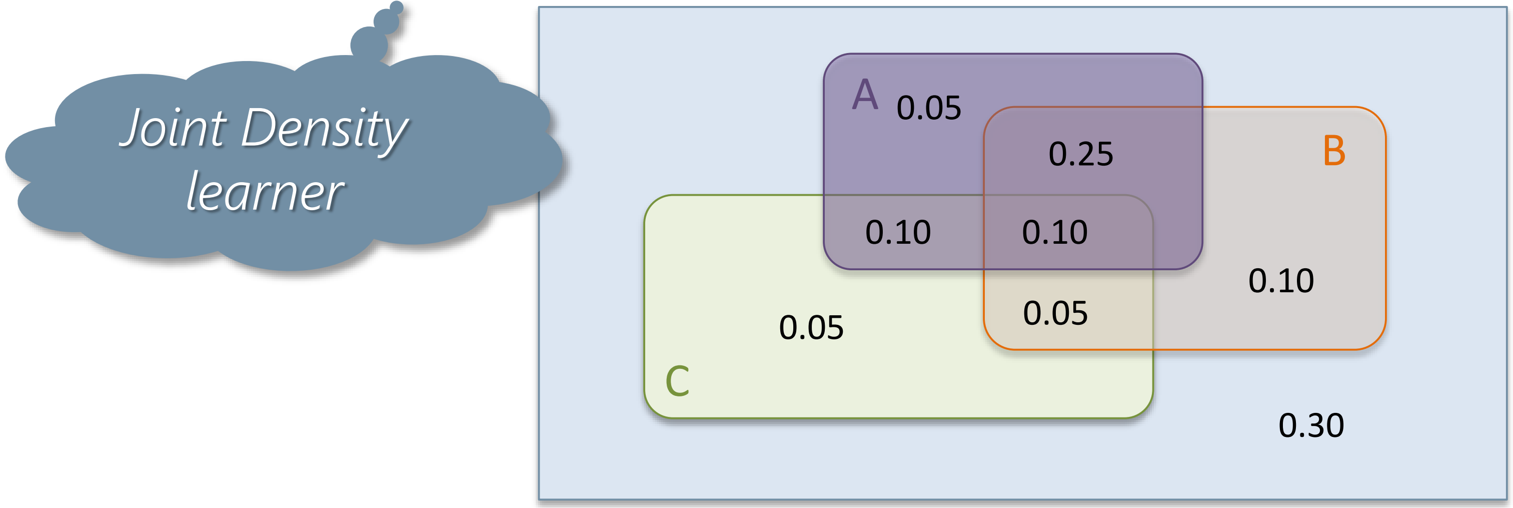
Prof. Matteo Matteucci – *matteo.matteucci@polimi.it*

The Joint Distribution

Given two random variables X and Y, the *joint distribution* of X and Y is the distribution of X and Y together: $P(X,Y)$

How to make a joint distribution of M variables:

- Make a truth table listing all combination of values
- For each combination compute how probable it is
- Check that all probabilities sum up to 1



A	B	C	P(A,B,C)
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

Using the Joint Distribution

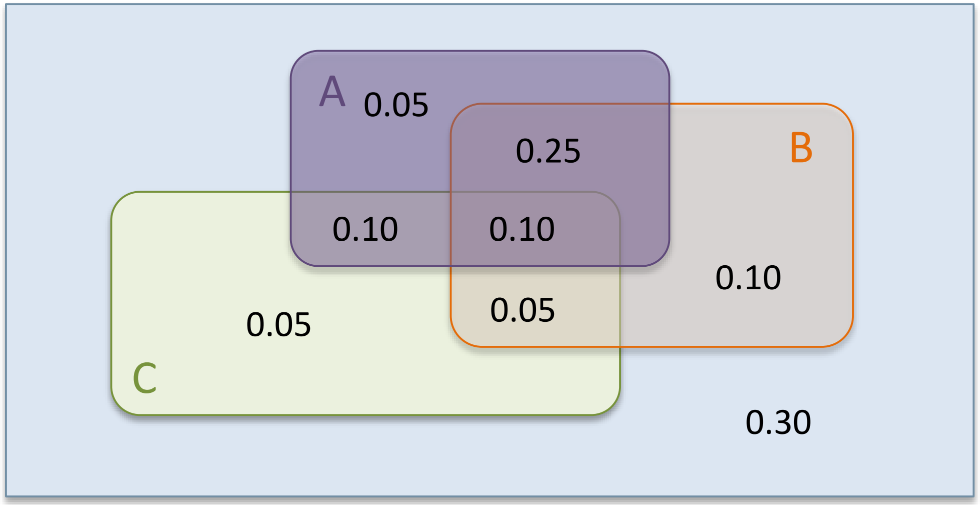
Compute probability for logic expression

$$P(E) = \sum_{row \sim E} P(row)$$

Examples:

- $P(A) = 0.05 + 0.10 + 0.25 + 0.10 = 0.5$
- $P(A \wedge B) = 0.25 + 0.10 = 0.35$
- $P(\bar{A} \vee B) = 0.30 + 0.05 + 0.10 + 0.05$
  $+0.25 + 0.10 = 0.85$

A	B	C	P(A,B,C)
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Using the Joint Distribution

Where does this come from?

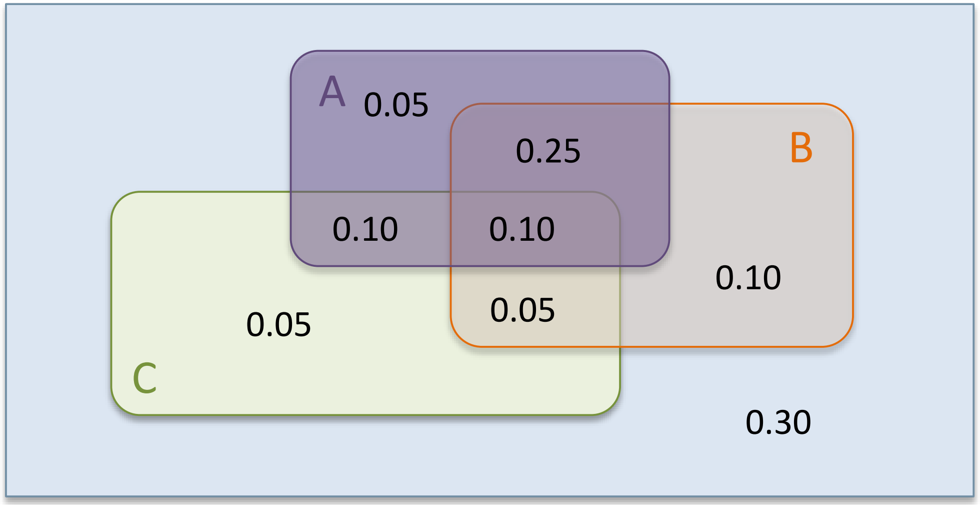
A	B	C	P(A,B,C)
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

Make Inference, a.k.a. probabilistic reasoning

$$P(E_1|E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{row\sim E_1 \wedge E_2} P(row)}{\sum_{row\sim E_2} P(row)}$$

Examples:

- $P(A|B) = \frac{(0.25+0.10)}{(0.10+0.05+0.25+0.10)} = \frac{0.35}{0.50} = 0.70$
- $P(C|A \wedge B) = \frac{(0.10)}{(0.25+0.10)} = \frac{0.10}{035} = 0.285$
- $P(\bar{A}|C) = \frac{(0.05+0.05)}{(0.05+0.05+0.10+0.10)} = \frac{0.10}{0.30} = 0.333$



The Joint Density Estimator

How can we evaluate it?

A *Density Estimator* learns a mapping from a set of attributes to a probability distribution over the attributes space

A	B	C	P(A,B,C)
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

Our Joint Distribution table is a first example:

- Build a Joint Distribution table for the attributes in which the probabilities are unspecified
- Then fill in each row with

$$\hat{P}(row) = \frac{\#\, records\ matching\ row}{total\ number\ of\ records}$$

	Pistol	Knife	Poker
Cook	4%	52%	24%
Butler	16%	2%	2%

We will come back to its formal definition at the end of the lecture …

Evaluating Density Estimators

We can use likelihood for evaluating density estimation:

- Given a record $\boldsymbol{x}$, a density estimator $\boldsymbol{M}$ tells you how likely it is

$$\hat{P}(x|M)$$

- Given a dataset with $\boldsymbol{N}$ records, a density estimator can tell how likely data is under the assumption that all records were independently generated from it

$$\hat{P}(dataset) = \hat{P}(x_1, x_2, \dots, x_N) = \prod_{n=1}^{N} \hat{P}(x|M)$$

- Since likelihood can get too small we usually use log-likelihood:

$$log\hat{P}(dataset) = log \prod_{n=1}^{N} \hat{P}(x_n|M) = \sum_{n=1}^{N} log\hat{P}(x_n|M)$$

Joint Estimator: the Good and the Bad

Density estimators can do many good things

- Can sort the records by probability, and thus spot weird records (e.g., anomaly/outliers detection)
- Can do inference: $P(E_1|E_2)$ (e.g., Automatic Doctor, Help Desk)
- Can be used for Bayes Classifiers (see later)

Joint Density estimators can badly overfit!

- Joint Estimator just mirrors the training data
- Suppose you see a ***new dataset***, its likelihood is going to be

$$\hat{P}(test\ dataset|M) = \sum_{n=1}^{N} log\hat{P}(x_n|M) = -\infty$$

Naïve Bayes Estimator

The naïve model assumes that each attribute is distributed independently of any of the other attributes.

- Let $\boldsymbol{x}[\boldsymbol{i}]$ denote the $\boldsymbol{i}^{\boldsymbol{th}}$ field of record $\boldsymbol{x}$
- The Naïve Density Estimator says that:

$$x[i] \perp \{x[1], x[2], .., x[i-1], x[i+1], \dots, x[M]\}$$

It is important to recall every time we use a Naïve Density that:

- Attributes are equally important
- Knowing one attribute says nothing about the value of another
- Independence assumption is almost never correct …
  but works quite well in practice!

Naïve Bayes Estimator and Joint Distribution

Suppose A, B, C, D independently distributed, what is $P(A, \bar{B}, C, \bar{D})$?

$$\begin{aligned} P(A, \bar{B}, C, \bar{D}) &= P(A|\bar{B}, C, \bar{D}) * P(\bar{B}, C, \bar{D}) \\ &= P(A) * P(\bar{B}, C, \bar{D}) \\ &= P(A) * P(\bar{B}|C, \bar{D}) * P(C, \bar{D}) \\ &= P(A) * P(\bar{B}) * P(C, \bar{D}) \\ &= P(A) * P(\bar{B}) * P(C|\bar{D}) * P(\bar{D}) \\ &= P(A) * P(\bar{B}) * P(C) * P(\bar{D}) \end{aligned}$$

Suppose to randomly shake a green dice and a red dice

- Dataset 1: A = red value, B = green value
- Dataset 2: A = red value, B = sum of values
- Dataset 3: A = sum of values, B = difference of values

Which violates the naïve assumption?

Learning a Naïve Bayes Estimator

Suppose $x[1], x[2], \dots, x[M]$ are independently distributed:

- We can construct any row of the implied Joint Distribution on demand

$$\hat{P}(x[1] = u_1, x[2] = u_2, \dots, x[M] = u_M) = \prod_{k=1}^{M} \hat{P}(x[k] = u_k)$$

- We can do any inference!

$$P(E_1 | E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{row \sim E_1 \wedge E_2} P(row)}{\sum_{row \sim E_2} P(row)}$$

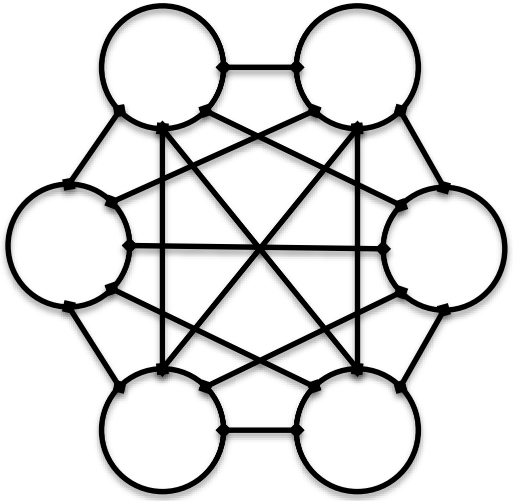
How do we learn a Naïve Density Estimator?

$$\hat{P}(x[i] = u) = \frac{\#\, records\ for\ which\ x[i] = u}{total\ number\ of\ records}$$

Joint Density vs. Naïve Density

Joint Distribution Estimator

- Can model anything
- Given 100 records and more than 6 Boolean attributes will perform poorly
- Can easily overfit the data



Is there anything in between?

Naïve Distribution Estimator

- Can model only very boring distributions
- Given 100 records and 10,000 multivalued attributes will be fine
- Quite robust to overfitting