Summaries and Exercises

Giulia Lasagni
University of Bologna

Marcello Di Bello
Arizona State University

AI for Judges

Agenda

PART I: Risk Models

PART II: LLMs

PART III: Bayesian Networks

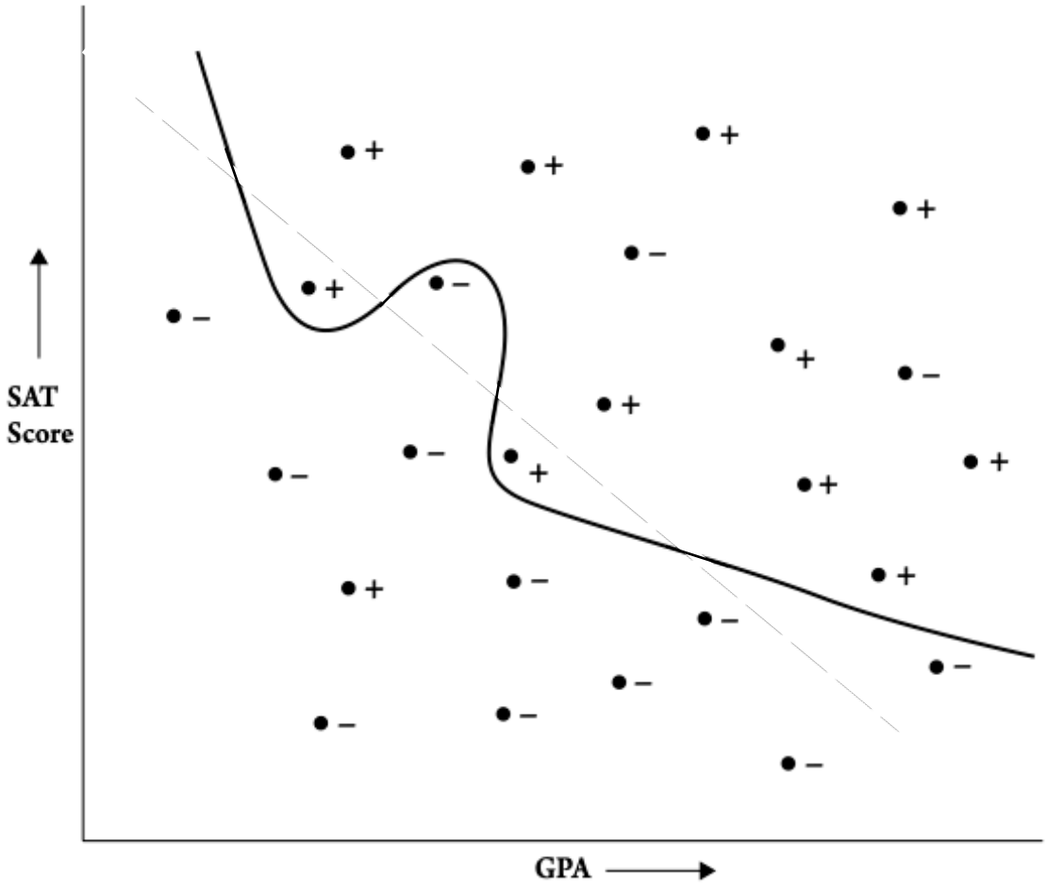
PART IV: The EU AI Act

Part I

Risk Models / Predictive Algorithms (recap and exercises)

Predictive Algorithms (or Predictive Models)

(*binary case*)

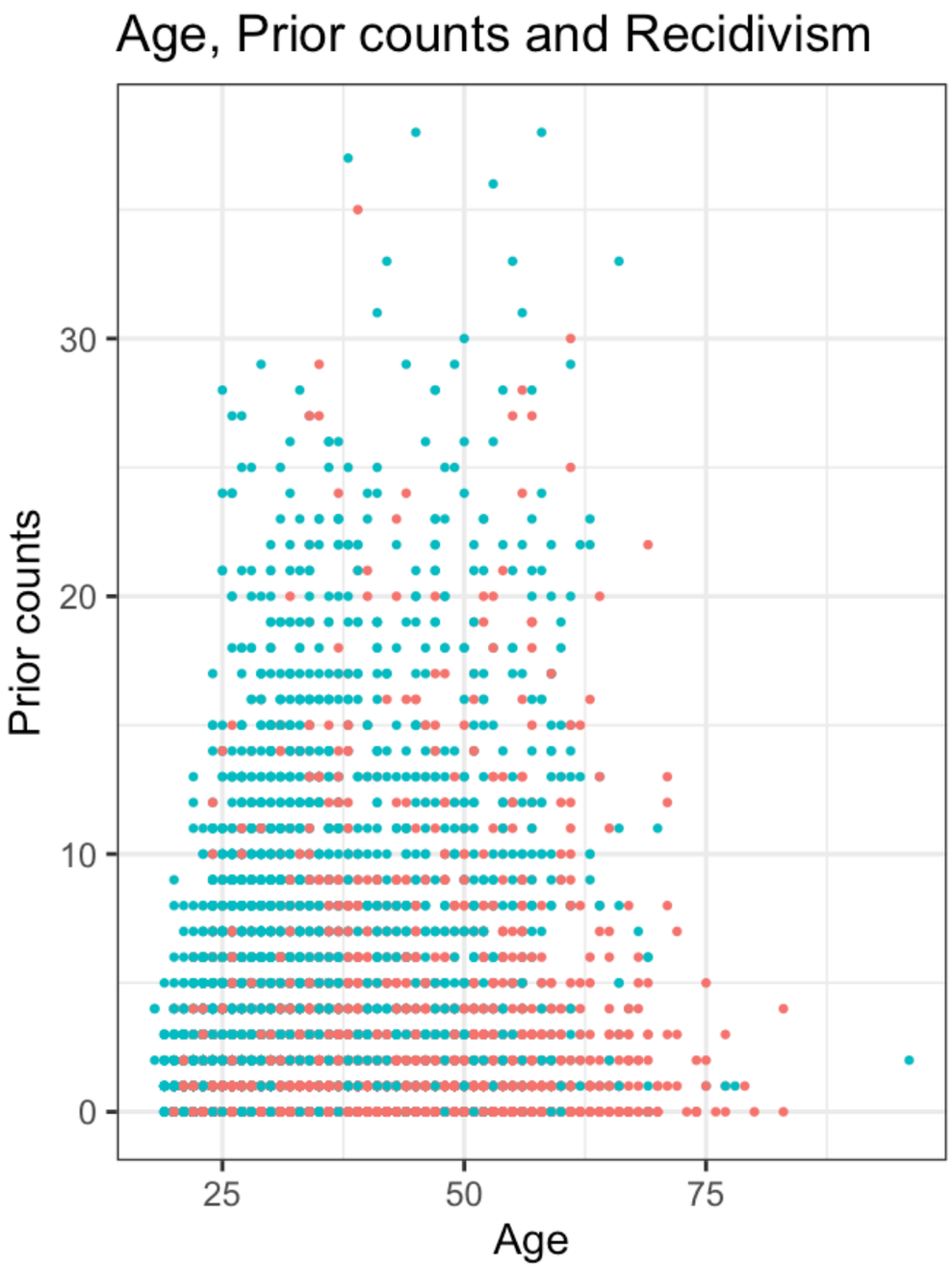


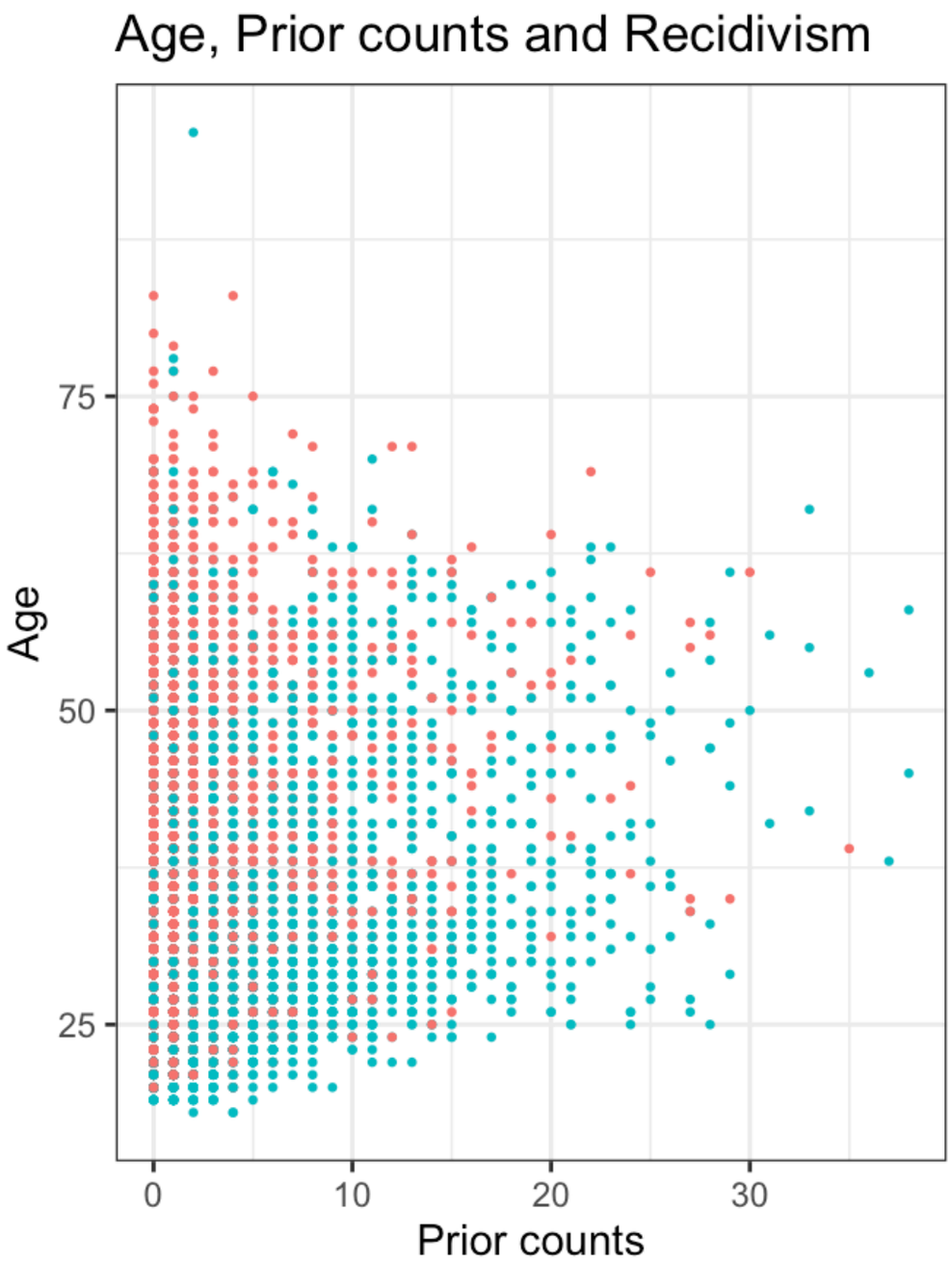
Suppose we aim to make predictions about a **binary outcome Y=1** or **Y=0** (e.g. college success, recidivism)

Machine learning algorithms (e.g. regression, SVM) mine the historical data and identify relationships between **predictive features** (e.g. GPA, income) and the outcome

Based on the features one possesses, the **predictive model classifies** individuals as **C=1** or **C=0**

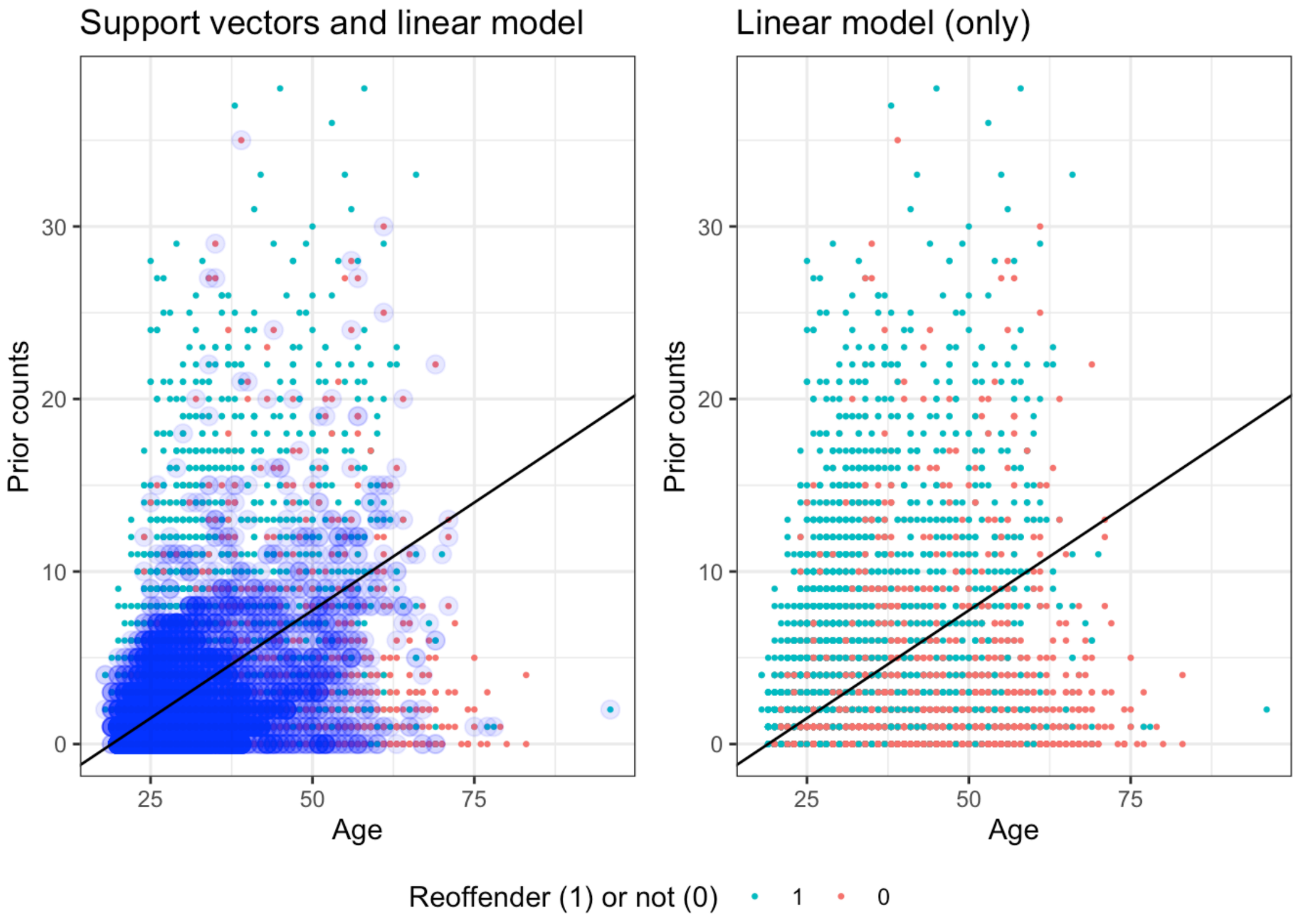
Historical Data: Age, Prior Counts, Reoffeding





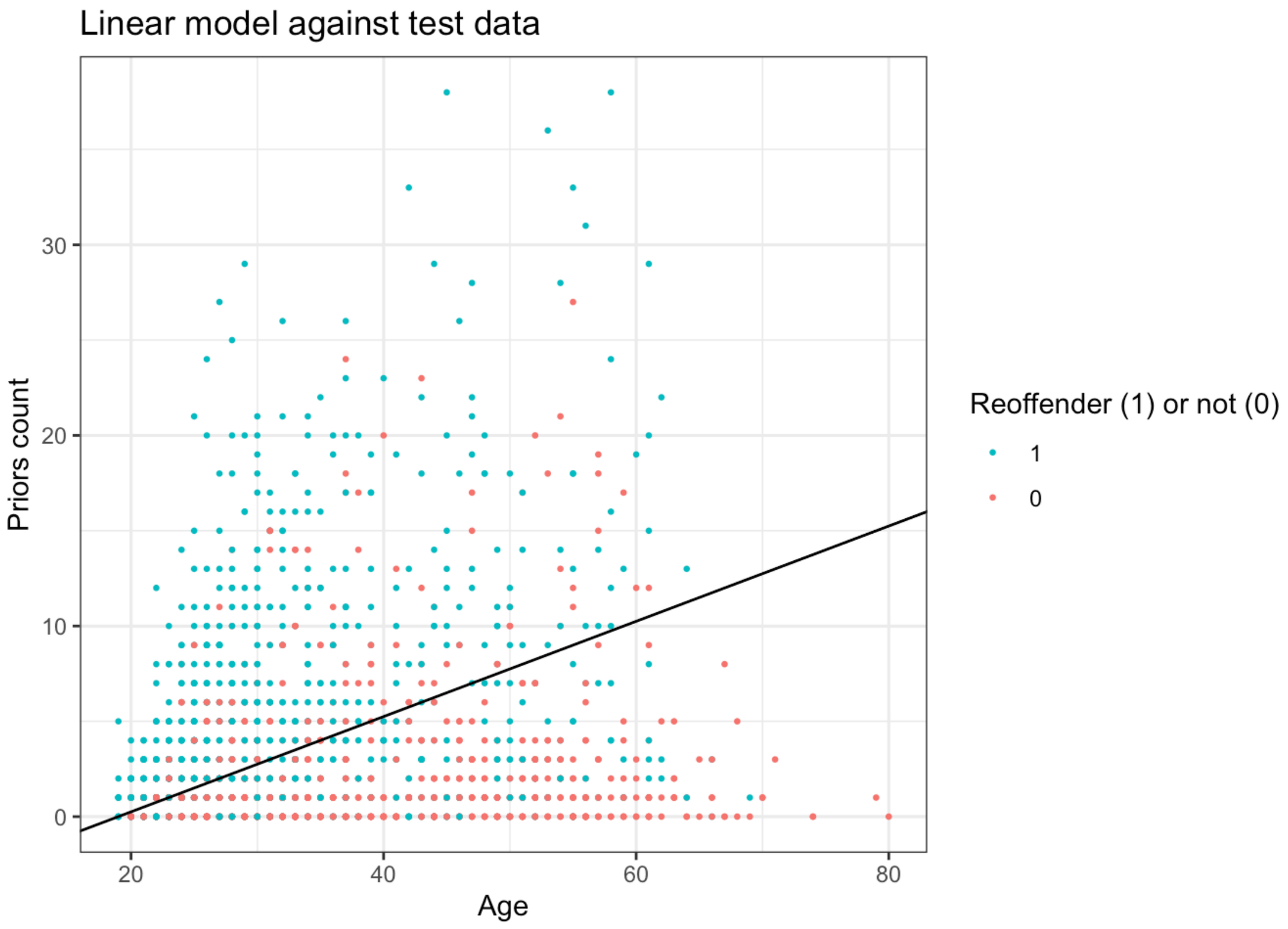
Reoffender (1) or not (0) 1 0

SVM Risk Model: Support Vectors and Line



Training Data v. Test Data

Validating Model Against Test Data



Example 1:
COMPAS

COMPAS (Northpoint Inc./Equivant): “static information (criminal history), with limited use of some dynamic variables (i.e. criminal associates, substance abuse)” + 137 interview questions + …?

The next few statements are about what you are like as a person, what your thoughts are, and how other people see you. There are no 'right or wrong' answers. Just indicate how much you agree or disagree with each statement.

112. "I am seen by others as cold and unfeeling."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

113. "I always practice what I preach."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

114. "The trouble with getting close to people is that they start making demands on you."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

115. "I have the ability to "sweet talk" people to get what I want."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

116. "I have played sick to get out of something."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

117. "I'm really good at talking my way out of problems."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

118. "I have gotten involved in things I later wished I could have gotten out of."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

119. "I feel bad if I break a promise I have made to someone."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

120. "To get ahead in life you must always put yourself first."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

WHICH APPLICATION?

- probation, alternative measures, etc.

- and what about sentencing?

The Loomis Case - State v. Loomis, 881 N.W.2d 749 (Wis. 2016)

Example 2:
Public Safety Assessment (PSA) (Printout)

WHICH REMEDY?

Art. 11 LED – Automated individual decision making

Decision based <u>solely</u> on automated processing
(including profiling)

which produces an adverse legal effect concerning the data subject or significantly affects him or her,
<u>shall be prohibited unless</u>

- authorised by Union or Member State law
- appropriate safeguards are provided, at least t**<u>he right to obtain human intervention on the part of the controller</u>**

Example

FRONTEX

- *European Travel Information Authorisation System* (ETIAS), fully operational by the end of 2022: automated assessment of third country citizens on the threat posed to national security or public health
- if positive assessment: need to have a second assessment by a human being

Do Human Overrides Improve Accuracy?

- “This study examines … the impact of overrides on the PCRA’s risk prediction effectiveness. Findings show that nearly all … tend to place **substantial numbers of persons under federal supervision** (especially those convicted of sex offenses) into the highest supervision categories, and that overrides result in a **deterioration of the PCRA’s risk prediction capacities**."

RISK ASSESSMENT OVERRIDES

Shuffling the Risk Deck Without Any Improvements in Prediction

THOMAS H. COHEN
CHRISTOPHER T. LOWENKAMP
Administrative Office of the U.S. Courts
KRISTIN BECHTEL
Arnold Ventures
ANTHONY W. FLORES
California State University, Bakersfield

In the federal supervision system, officers have discretion to depart from the risk designations provided by the Post Conviction Risk Assessment (PCRA) instrument. This component of the risk classification process is referred to as the supervision override. While the rationale for allowing overrides is that actuarial scores cannot always capture an individual’s unique characteristics, there is relatively limited literature on the actual effects of overrides on an actuarial tool’s predictive efficacies. This study examines overrides in the federal system by assessing the extent to which risk levels are adjusted through overrides as well as the impact of overrides on the PCRA’s risk prediction effectiveness. Findings show that nearly all overrides lead to an upward risk reclassification, that overrides tend to place substantial numbers of persons under federal supervision (especially those convicted of sex offenses) into the highest supervision categories, and that overrides result in a deterioration of the PCRA’s risk prediction capacities.

Keywords: supervision overrides; risk prediction; risk assessment tools; professional discretion

Exercise:

1. Familiarize yourself with a risk assessment tool. If you don't know which to pick, have a look at https://oxrisk.com/oxrec-9/
2. Input a few characteristics and see what prediction or risk assessment the tool returns
3. If you were a judge, would you follow the prediction? Explain.
4. What additional characteristics of the individual would make you override the prediction made by the risk assessment tool?

Part II

LLMs such as Chat GPT (recap and exercises)

GPT = generative pre-trained transformer

What Does Chat-GPT Do?

Chat-GTP is a **word completion** program on steroids.

It picks the next word based on reasonable probabilities, though it need not pick the most likely next word.

Complete the following:

“Plastic bags can…”

pollute	2%
save	3%
suffocate	3%
tables	0.0001%

One Word at a Time!

Chat-GPT carries out its completion task **one word at a time** until it hits a <stop> token that is assigned a reasonable probability.

Until it reaches <stop>, Chat-GPT continues its completion task using its previous output as part of the next input:

Plastic bags can …

Plastic bags can save …

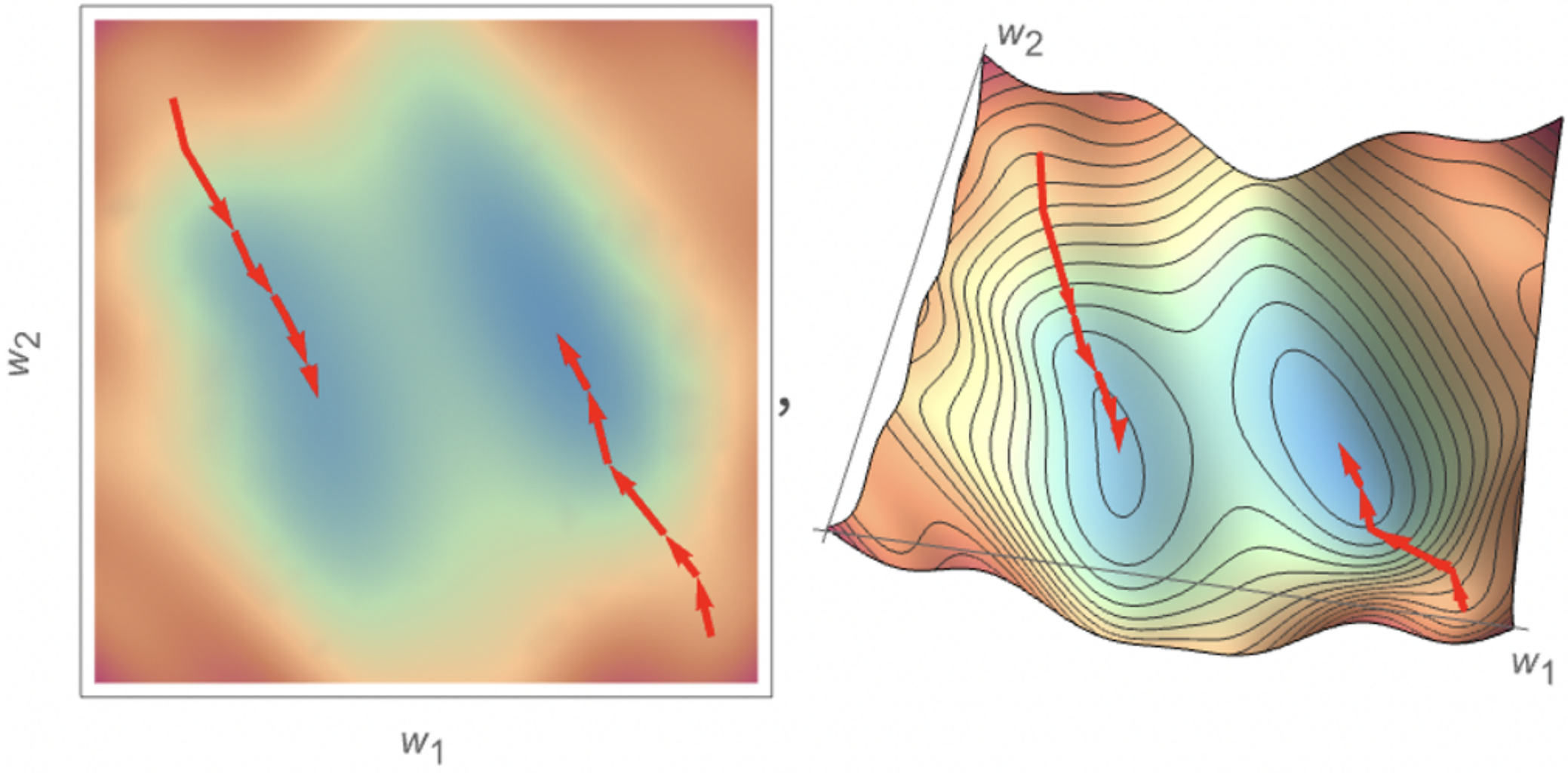
Plastic bags can save the …

How Does Chat-GPT Learn These “Next Word” Probabilities?

Pr(next word | past words)

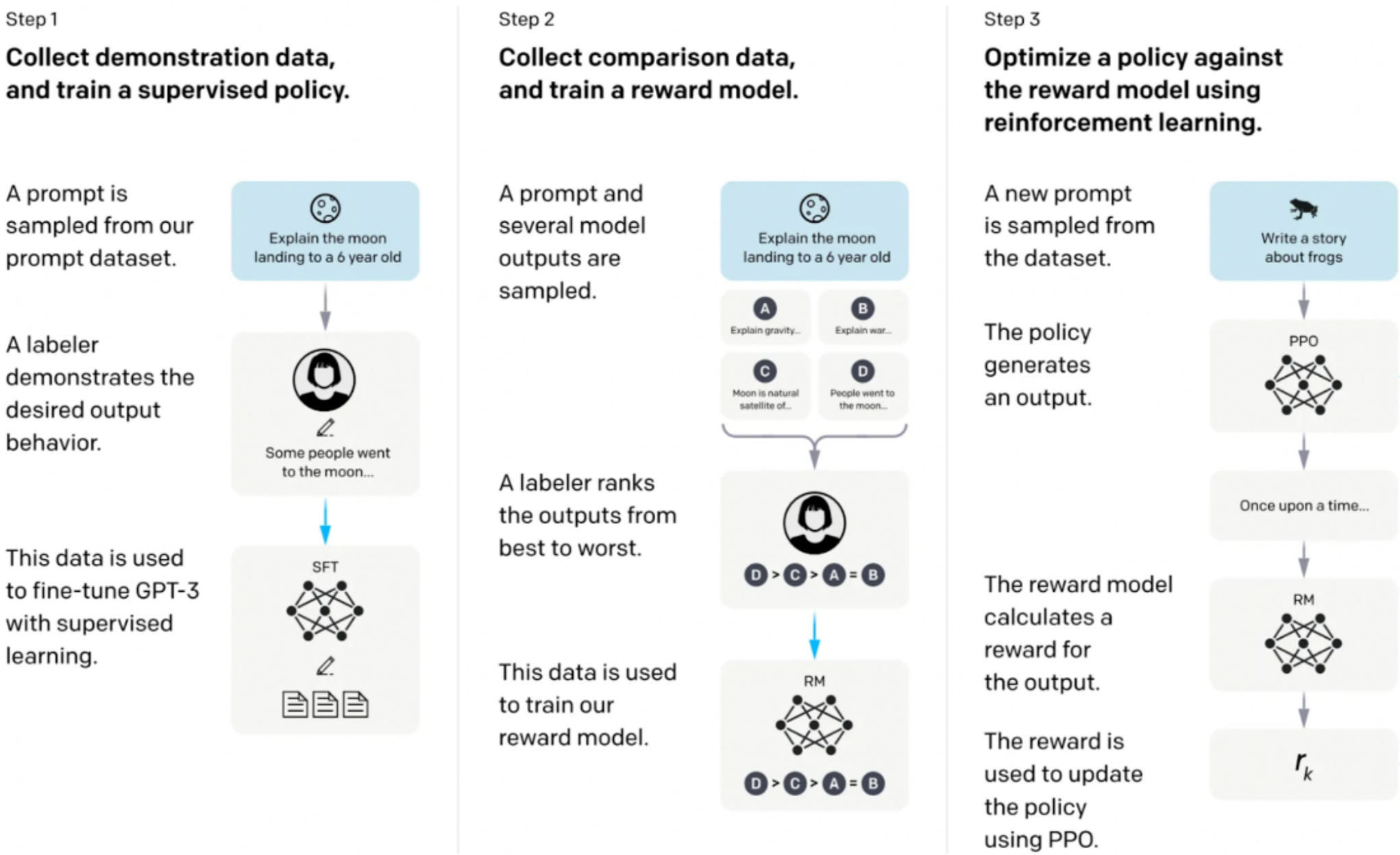
Minimizing Loss

$$
\begin{aligned}
&w_{511} f(w_{311} f(b_{11} + x w_{111} + y w_{112}) + w_{312} f(b_{12} + x w_{121} + y w_{122}) + \\
&\quad w_{313} f(b_{13} + x w_{131} + y w_{132}) + w_{314} f(b_{14} + x w_{141} + y w_{142}) + b_{31}) + \\
&w_{512} f(w_{321} f(b_{11} + x w_{111} + y w_{112}) + w_{322} f(b_{12} + x w_{121} + y w_{122}) + \\
&\quad w_{323} f(b_{13} + x w_{131} + y w_{132}) + w_{324} f(b_{14} + x w_{141} + y w_{142}) + b_{32}) + \\
&w_{513} f(w_{331} f(b_{11} + x w_{111} + y w_{112}) + w_{332} f(b_{12} + x w_{121} + y w_{122}) + \\
&\quad w_{333} f(b_{13} + x w_{131} + y w_{132}) + w_{334} f(b_{14} + x w_{141} + y w_{142}) + b_{33}) + b_{51}
\end{aligned}
$$



Source: [What Is ChatGPT Doing … and Why Does It Work?—Stephen Wolfram Writings]

Human Feedback



Source: https://openai.com/blog/chatgpt

Exercise: pick a court opinion that you know well, have Chat-GPT read it and then ask questions, such as:

- what is the holding? what are the key arguments?
- did the decision overrule any precedent?
- what precedent did it follow?
- etc.

Did Chat-GPT answer your questions correctly?

The Legal Bench Project

Project's Goals

Create a set of ***benchmark legal reasoning tasks***

Assess how LLMs like Chat GPT perform at executing legal reasoning tasks

LEGALBENCH: A COLLABORATIVELY BUILT BENCHMARK FOR MEASURING LEGAL REASONING IN LARGE LANGUAGE MODELS

Neel Guha[‡], Julian Nyarko[*1], Daniel E. Ho[*1], Christopher Ré[*1], Adam Chilton[2], Aditya Narayana[3], Alex Chohlas-Wood[1], Austin Peters[1], Brandon Waldon[1], Daniel N. Rockmore[4], Diego Zambrano[1], Dmitry Talisman[3], Enam Hoque[5], Faiz Surani[1], Frank Fagan[6], Galit Sarfaty[7], Gregory M. Dickinson[8], Haggai Porat[9], Jason Hegland[1], Jessica Wu[1], Joe Nudell[1], Joel Niklaus[1], John Nay[10], Jonathan H. Choi[11], Kevin Tobia[12], Margaret Hagan[13], Megan Ma[10], Michael Livermore[14], Nikon Rasumov-Rahe[3], Nils Holzenberger[15], Noam Kolt[7], Peter Henderson[1], Sean Rehaag[16], Sharad Goel[17], Shang Gao[20], Spencer Williams[18], Sunny Gandhi[19], Tom Zur[9], Varun Iyer , and Zehua Li[1]

[1]Stanford University, [2]University of Chicago, [3]Maxime Tools, [4]Dartmouth College, [5]LawBeta, [6]South Texas College of Law Houston, [7]University of Toronto, [8]St. Thomas University Benjamin L. Crump College of Law, [9]Harvard Law School, [10]Stanford Center for Legal Informatics - CodeX, [11]University of Southern California, [12]Georgetown University Law Center, [13]Stanford Law School, [14]University of Virginia, [15]Télécom Paris, Institut Polytechnique de Paris, [16]Osgoode Hall Law School, York University, [17]Harvard Kennedy School, [18]Golden Gate University School of Law, [19]Luddy School of Informatics - Indiana University Bloomington, [20]Casetext

August 23, 2023

ABSTRACT

The advent of large language models (LLMs) and their adoption by the legal community has given rise to the question: what types of legal reasoning can LLMs perform? To enable greater study

IRAC model of Legal Reasoning

Issue
spotting

Rule **Recall**

Rule
Application

Rule **Conclusion**

hearsay - train dataset for in-context prompting

index	answer	text	slice
0	No	On the issue of whether David is fast, the fact that David set a high school track record.	Non-assertive conduct
1	Yes	On the issue of whether Rebecca was ill, the fact that Rebecca told Ronald that she was unwell.	Standard hearsay
2	No	"To prove that Tim was a soccer fan, the fact that Tim told Jimmy that ""Real Madrid was the best soccer team in the world.""""	Not introduced to prove truth
3	No	"When asked by the attorney on cross-examination, Alice testified that she had ""never seen the plaintiff before, and had no idea who she was.""""	Statement made in-court
4	Yes	On the issue of whether Martin punched James, the fact that Martin smiled and nodded when asked if he did so by an officer on the scene.	Non-verbal hearsay

Example: hearsay - test dataset

huggingface.co/datasets/nguha/legalbench/blob/main/data/hearsay/test.tsv

main / legalbench / data / hearsay / test.tsv

nguha Data update cfb4055 about 1 year ago

raw | Copy download link | history | blame | contribute | delete | Safe | 16.1 kB

```
index   answer  text    slice
0   No  On the issue of whether James is an smart individual, the fact that James came first in his class in law school.    Non-assertive con
1   No  On the issue of whether Robert negligently drove, the fact that Robert fell asleep while driving.   Non-assertive conduct
2   No  On the issue of whether John knew about the conspiracy, the fact that John likes sweatpants.    Non-assertive conduct
3   No  On the issue of whether Michael was guilty of murder, the fact that Michael left the crime scene immediately.   Non-assertive conduct
4   No  On the issue of whether William was loved by his community, the fact that he was selected to speak at his graduation.   Non-assertive
5   No  On the issue of whether Mary robbed the bank, the fact that Mary went to the bank in disguise.  Non-assertive conduct
6   No  "On the issue of whether Patricia was a fan of Coldplay, the fact that she had a poster with the lyrics of ""Viva la Vida"" on her be
7   No  On the issue of whether Jennifer suffered reputational harm from Linda's article, the fact that Linda worked with several different e
8   No  On the issue of whether Elizabeth was misdiagnosed by Barbara, the fact that Barbara didn't consult with her usual charts while asses
9   No  On the issue of whether Richard had ever visited Chicago, the fact that he gave a speech there in 2005. Non-assertive conduct
10  No  On the issue of how long Joseph and Thomas had known each other, the fact that were neighbors during elementary school. Non-assertive
11  No  On the issue of whether Susan was familiar with Shakespeare, the fact that she had once played the role of Macbeth and recieved a sta
12  No  On the issue of whether Jessica was aware she was trespassing, the fact that Jessica had been diagnosed as near-sighted by her ophtha
13  No  On the issue of whether Sarah was acting as an agent for the corporation, the fact that Sarah had worked there previously for four ye
14  No  On the issue of whether Charles was responsible for the defamatory article published online, the fact that Charles had visited the we
15  No  On the issue of whether Karen negligently operated the forklift, the fact that Karen is a terrible driver who needed several tries to
16  No  On the issue of whether the patent was infringed, the fact that the corporation's chief scientist was known to cheat at card games. N
17  No  On the issue of whether Christopher acted with malice, the fact that Christopher was often moody and had a large temper.    Non-asser
18  No  On the issue of whether the parties had actually agreed to the contract, the fact that one of the parties had mistaken the identity o
19  No  On the issue of which car was responsible for a hit-and-run, the witness's statement in court that she believed it was the blue sedan
20  No  On the issue of the faultiness of the designed house, the drawing the witness made on the stand during testimony.   Statement made in
21  No  On the issue of which of the defendants was responsible for driving the get-away car, the fact that the witness on the stand turned a
22  No  On the issue of whether Ana lied to Jim, Jim's statement on cross-examination that he did not believe Ana to be an honest individual.
23  No  On the issue of whether Daniel drove negligently, the fact that Daniel testified during the trial that he told his wife he was tired
24  No  On the issue of whether Carl had knowledge of Amy's intentions, Carl told the questioning attorney on redirect examination that he kn
```

Exercise:
1. Think about a specific legal task that judges need to perform
2. Create a training set for in-context prompting and a test set
3. Assess Chat GPT with instances from test set (no prompting)
4. Train Chat CPT with in-context prompting
5. Assess Chat GPT with instances from test set (after prompting)

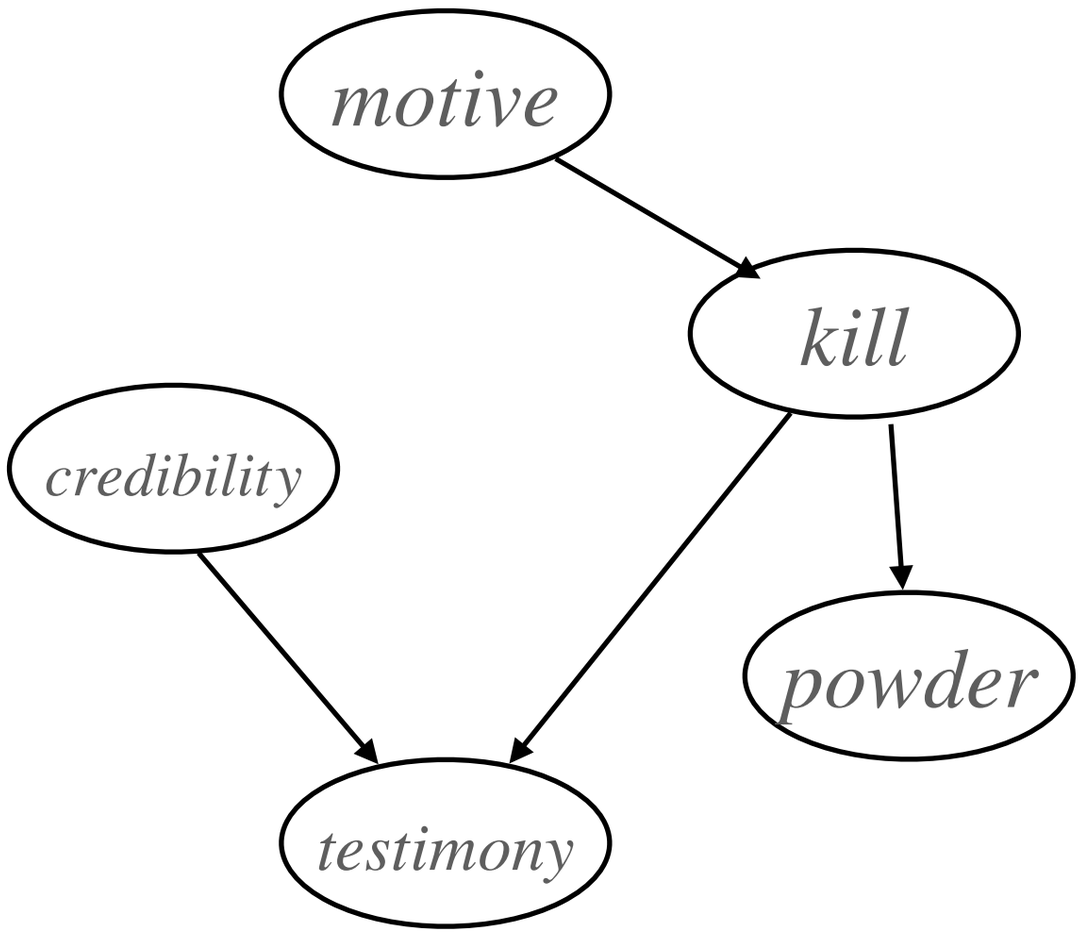
Find inspiration from Legal Bench project:
https://huggingface.co/datasets/nguha/legalbench

Part III

Bayesian Networks (recap and exercises)

Graphical Components of a Bayesian Network

Arrows



As a first approximation, think of **arrows** as *directions of causal influence* (though this interpretation is debated):

Whether or not the defendant had a motive to kill influences whether or not the defendant killed the victim

Whether or not the defendant killed the victim influences whether or not gunpowder was found on defendant

Whether or not the defendant killed the victim influences what the witness saw

Whether or not the witness is credible influences what the witness says

Consider this Stylized Legal Case

Chris is shot (clearly murder) on an island.

There are 100 possible perpetrators. One of them is Fred.

Gun shot residue is found on Fred's hands same day as the shooting took place.

> There are two possible explanations: Fred shot Chris or Fred was at the shooting range the same day. Both explanations can be true. Given the gun shot residue, it is impossible that both are false.

Fred goes to the shooting range 4 days a week.

Daniela, a woman who works at the shooting range, is asked if she saw Fred on the day in question, and she says that he was not at the range that day.

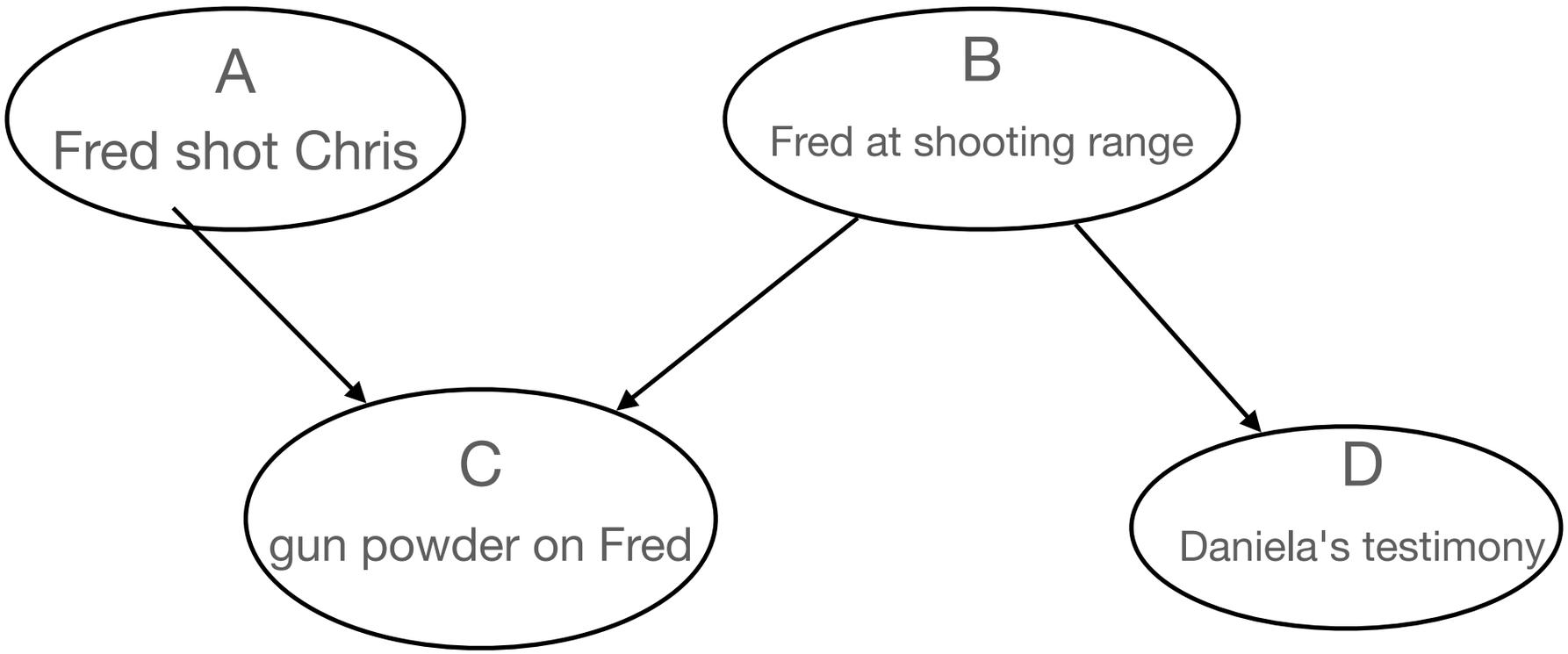
> Daniela's accuracy in correctly identifying and remembering Fred is 99%. In other words, if Fred was at the shooting range that day, there is a 1% chance that she will incorrectly report that he was not there, and if he was not, there is a 99% chance that she will correctly report that he was not there.

What is the probability that Fred shot Chris?

Graphs and Numbers

A=yes	1/100=1%
A=no	99%

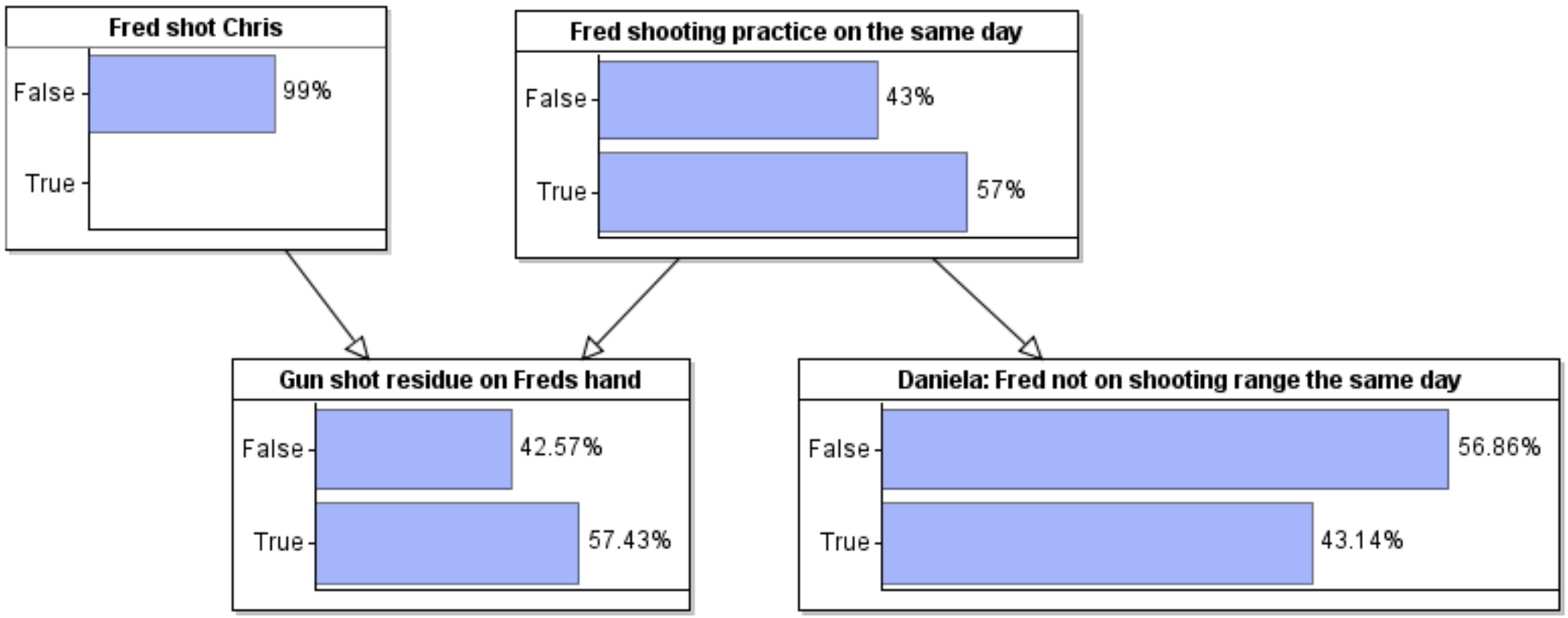
B=yes	4/7=57%
B=no	3/7=43%



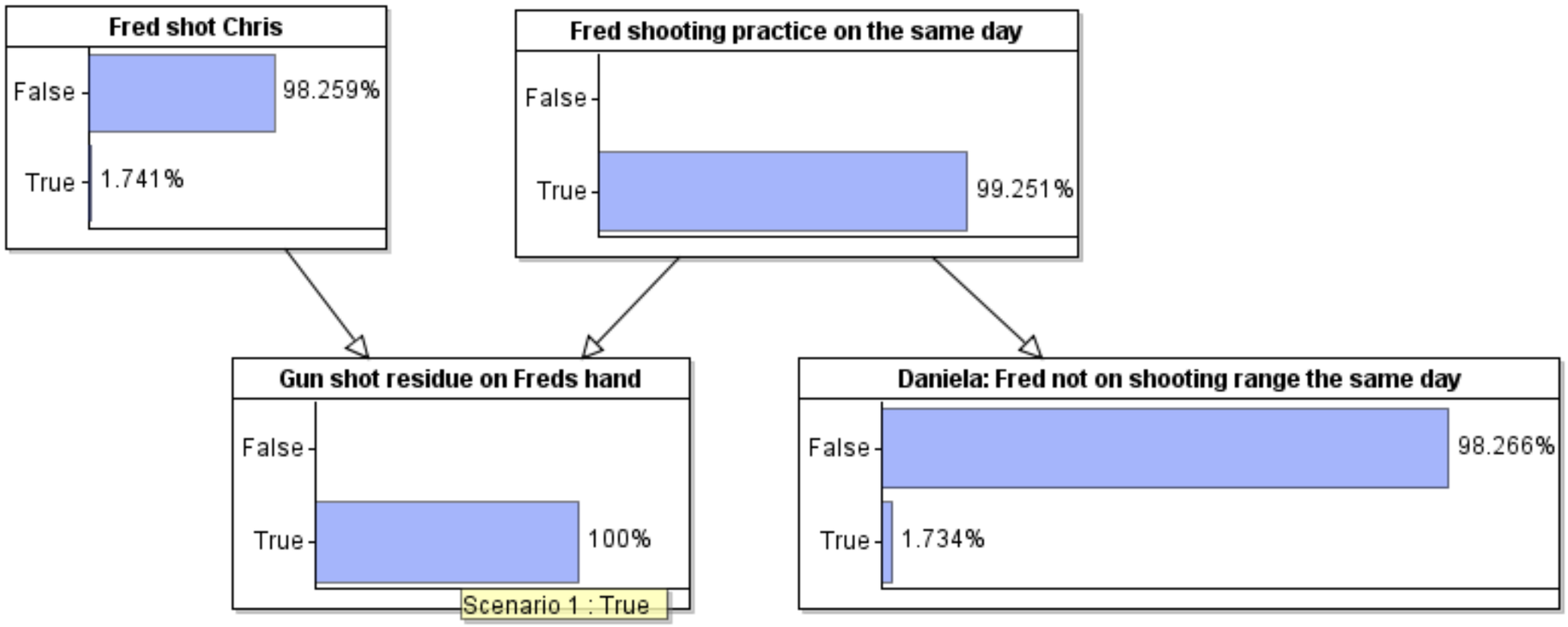
	A=yes & B=yes	**A=no & B=yes**	**A=yes & B=no**	**A=no & B=no**
C=yes	100%	100%	100%	0%
C=no	0%	0%	0%	100%

	B=yes	**B=no**
D=yes	99%	1%
D=no	1%	99%

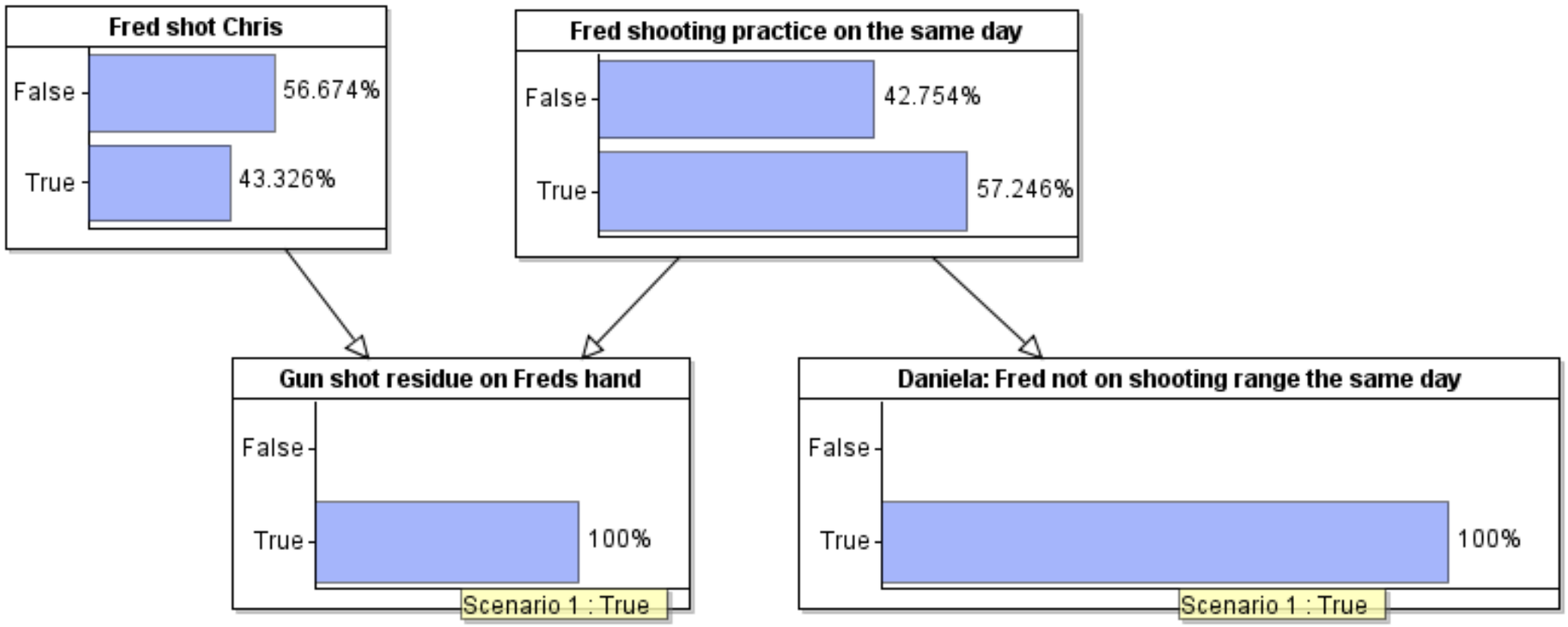
No Evidence:
Unlikely Fred Shot Chris



Gun Powder on Fred:
Still *Unlikely Fred Shot Chris*



Gun Powder on Fred *plus* Daniela's Testimony: Still *Unlikely Fred Shot Chris*



Exercise:
1. Think about a factual dispute in a legal case that you are familiar with
2. Draw a graphical model — arrows and nodes — of the evidence in the case
3. Add numbers if possible: fill in probabilities in the tables
4. Does the Bayesian network help in assessing the strength of the evidence and reaching a decision? Explain.

Part IV

The AI Act (separate slides)

Exercise:
1. Recall the AI systems we looked at, such as risk models, LLMs, multi-agent and Bayesian networks
2. Do they count as AI systems under the AI act?
3. Do they count as prohibited?
4. Do they count as high risk?
5. Do they count as low risk?