

WELCOME

CSC380 - Principles of Data Science

TODAY'S PLAN

- Syllabus overview: Check complete syllabus at D2L. Send me an email if any part is not clear.
- Introduction to Probability

COURSE INSTRUCTOR

- Cesim Erten, PhD from University of Arizona in CS (2004)
- Research:
Bioinformatics, Computational network biology, Graph algorithms
- Teaching
Data Science, Artificial Intelligence, Algorithms, Automata Theory
Formal Languages, Data Structures, Bioinformatics ...
- Office Hours: TBA

COURSE TAs

- Yao Zhao, Ruoshan Lan, Zachary Hansen, Desmond Thomas Goodman-Ahearn, Advait Khopade

COURSE COMMUNICATIONS

- Outside-lecture communications will be **through Piazza**.
Signup link: In D2L
- Assignment/Project submissions **through gradescope**.
Signup code: In D2L
- The instructor and TAs can also be reached **through email**.
- Come to instructor/TA **office hours**.

COURSE MATERIALS

- Textbooks:
 - Watkins, J., "An Introduction to the Science of Statistics: From Theory to Implementation"
 - Wasserman, L. "All of Statistics: A Concise Course in Statistical Inference."
 - James, G. et al. "An Introduction to Statistical Learning with Applications in Python"
 - Steven S. Skiena, "The Data Science Design Manual"
 - Lau, S. et al. "Learning Data Science"
- Lecture slides will be shared through D2L.

GRADING ITEMS

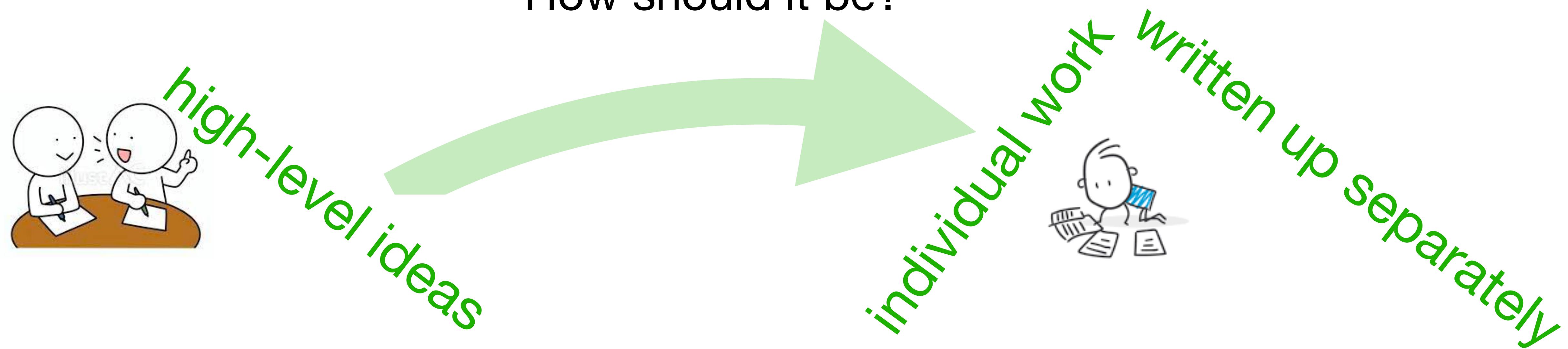
- **Assignments (5% each):**
 - Problem solving + Coding. Must be done individually.
 - 7 assignments (lowest grade dropped)
- **Project (16% total)**
 - Can be done in pairs
- **Quizzes (1% each)**
 - 12 quizzes (lowest two dropped). One question in each.
 - Announced one day before. Previous 2 lectures. 5 mins.
- **Exams (44% total)**
 - 1 midterm (20%) + final (24%)

BONUS OPPORTUNITIES

- **5% bonus:** All conditions below must be satisfied.
 - ≥ 4 “participation instances” in lectures
 - ≥ 4 “participation instances” in piazza
 - ≥ 4 appearances in instructor’s office hours
 - ≥ 4 SI session participations
- “participation instance” in lecture:
Answer a group (with neighbor) discussion question correctly **PAIR UP & DISCUSS**
- “participation instance” in piazza:
A correct answer to a technical question (at most one/month counted)

MORE ON ASSIGNMENTS

- Use of external sources: For instance, possible collaboration with friends.
How should it be?



- If you discuss high-level ideas with a friend mention his/her name in your solution.
- **Verbatim copying of solutions** from any external source (web search, ChatGPT, etc.) is **not acceptable**. If you make use of any external source mention the source in your solution and **make sure that your solution is your own work**.
- Submissions after the due date and time are **not accepted**.

FURTHER IMPORTANT POINTS

Classroom Behavior Policy

To foster a positive learning environment, students and instructors have a shared responsibility. We want a safe, welcoming, and inclusive environment where all of us feel comfortable with each other and where we can challenge ourselves to succeed. To that end, our focus is on the tasks at hand and not on extraneous activities (e.g., texting, chatting, reading a newspaper, making phone calls, web surfing, etc.).

Students are asked to refrain from disruptive conversations with people sitting around them during lecture. Students observed engaging in disruptive activity will be asked to cease this behavior. Those who continue to disrupt the class will be asked to leave lecture or discussion and may be reported to the Dean of Students.

Safety on Campus and in the Classroom

For a list of emergency procedures for all types of incidents, please visit the website of the Critical Incident Response Team (CIRT): <https://cirt.arizona.edu/case-emergency/overview>

Also watch the video available at https://arizona.sabacloud.com/Saba/Web_spf/NA7P1PRD161/common/learningeventdetail/crtfy000000000003560

UNIVERSITY/DEPARTMENT POLICIES

University-wide Policies link

Links to the following UA policies are provided here,

[http://catalog.arizona.edu/syllabus-policies:](http://catalog.arizona.edu/syllabus-policies)

- Absence and Class Participation Policies
- Threatening Behavior Policy
- Accessibility and Accommodations Policy
- Code of Academic Integrity
- Nondiscrimination and Anti-Harassment Policy

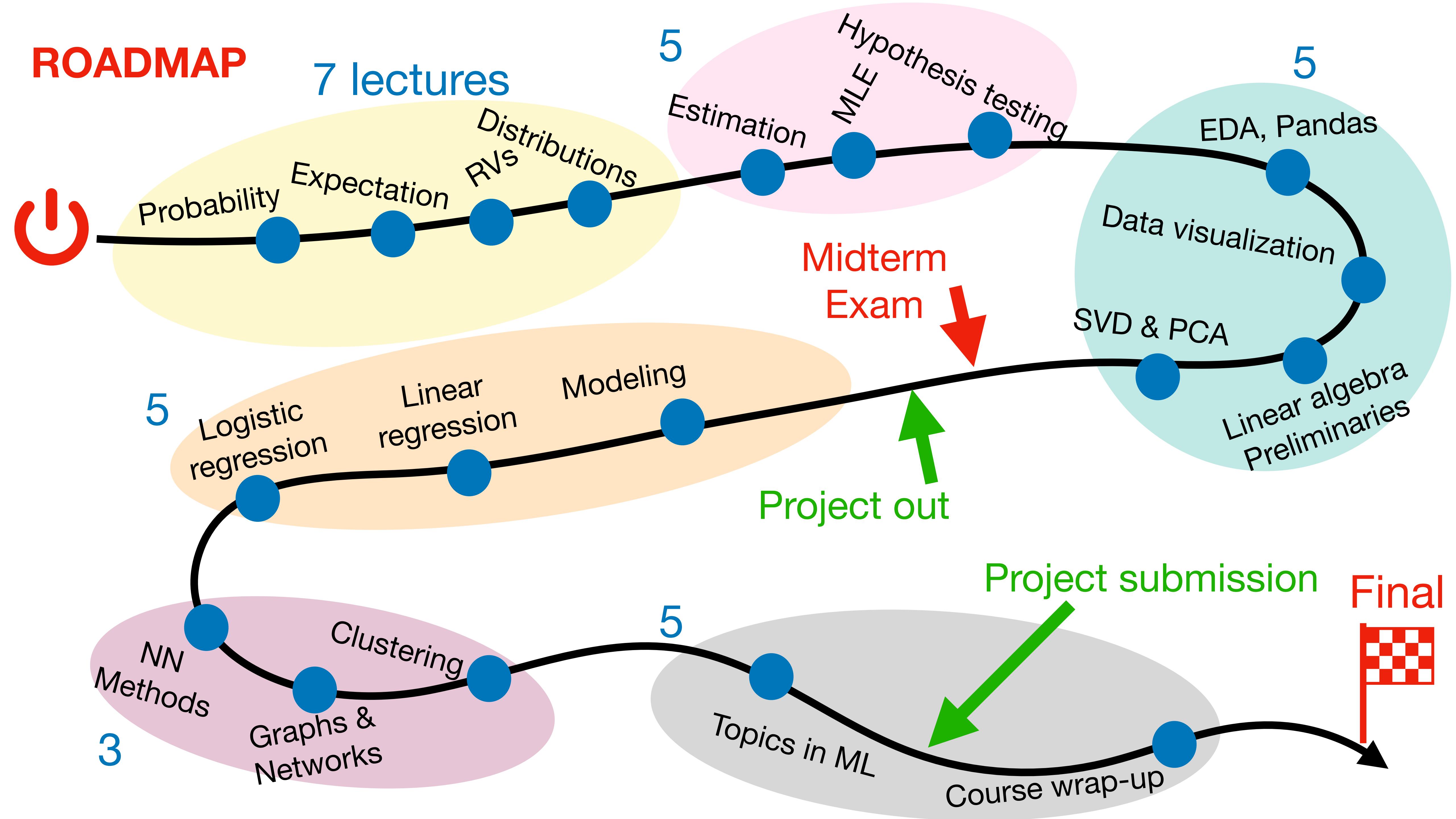
Department-wide Syllabus Policies and Resources link

Links to the following departmental syllabus policies and resources are provided here,

[https://www.cs.arizona.edu/cs-course-syllabus-policies :](https://www.cs.arizona.edu/cs-course-syllabus-policies)

- Department Code of Conduct
- Class Recordings
- Illnesses and Emergencies
- Obtaining Help
- Preferred Names and Pronouns
- Confidentiality of Student Records
- Additional Resources
- Land Acknowledgement Statement

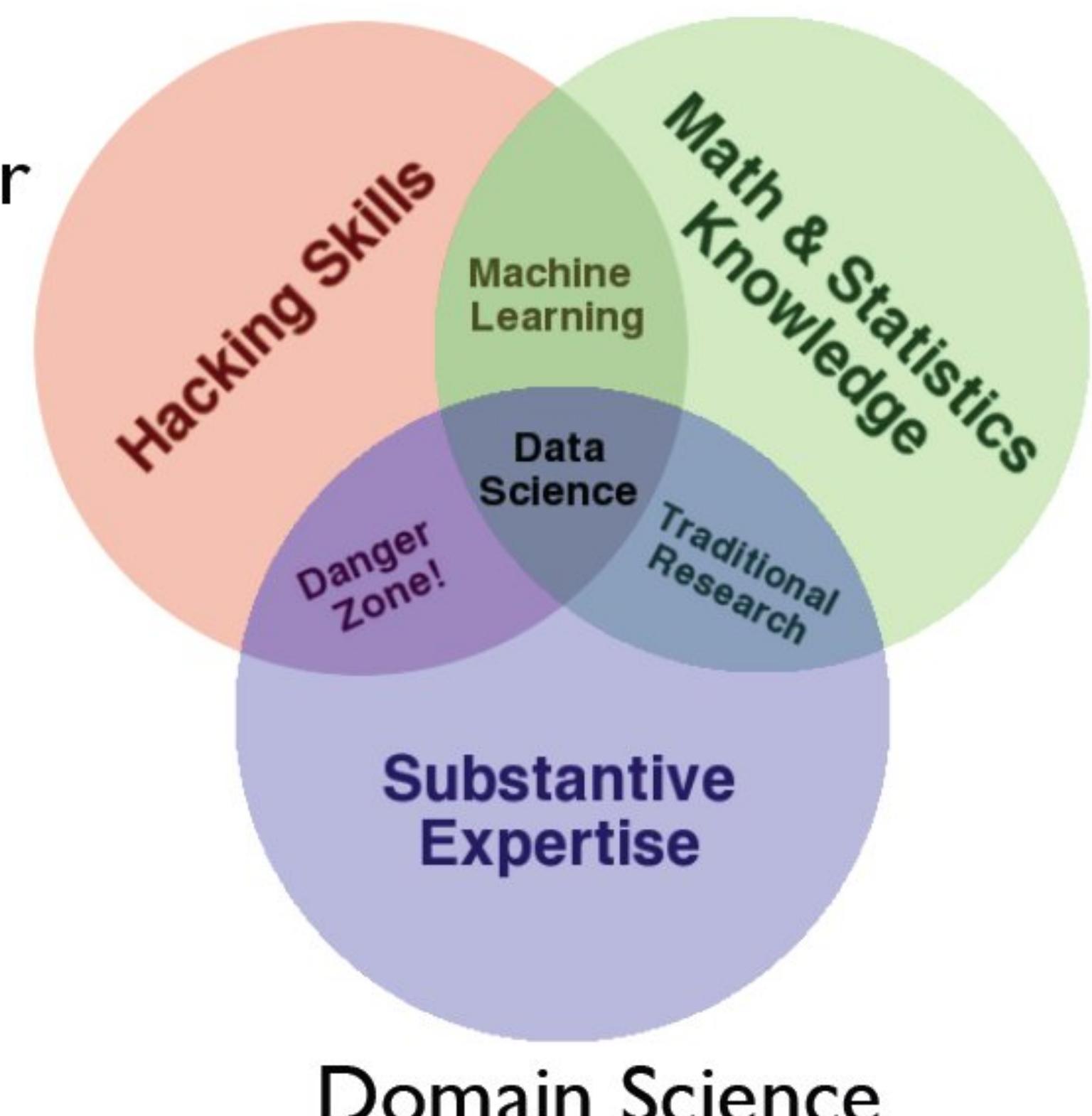
ROADMAP



INTRODUCTION TO DATA SCIENCE

- **Broad & Interdisciplinary :**
not a single discipline, at the intersection of multiple disciplines:
statistics, computer science,
operations research, statistical
and machine learning, data
warehousing, visualization,
mathematics, information
science, ...

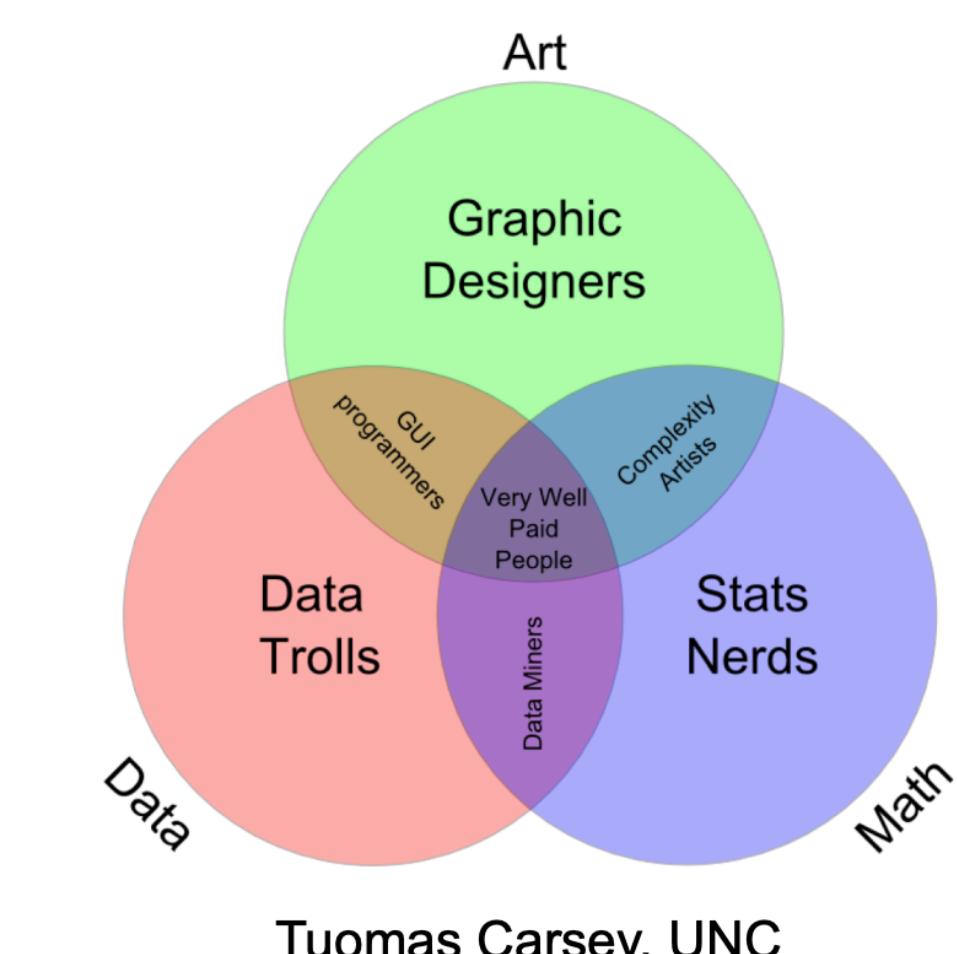
Computer
Science



Statistics

Domain Science

Drew Conway

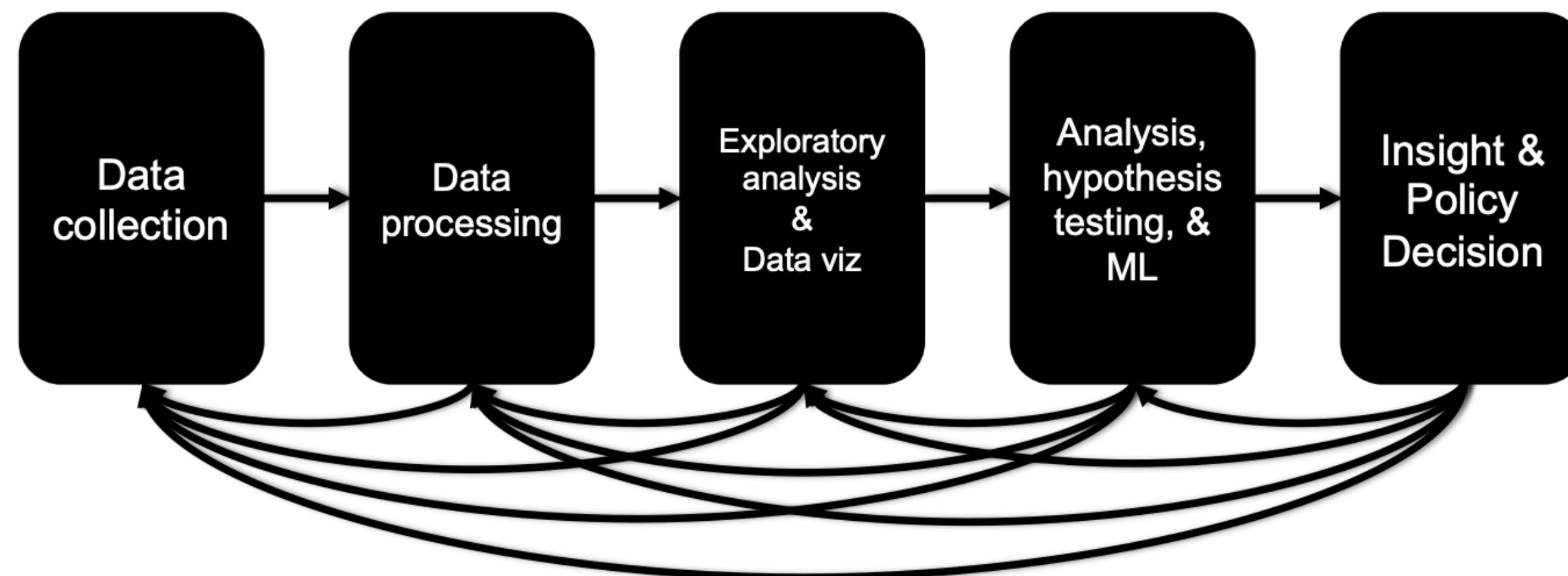


Data
Art
Math

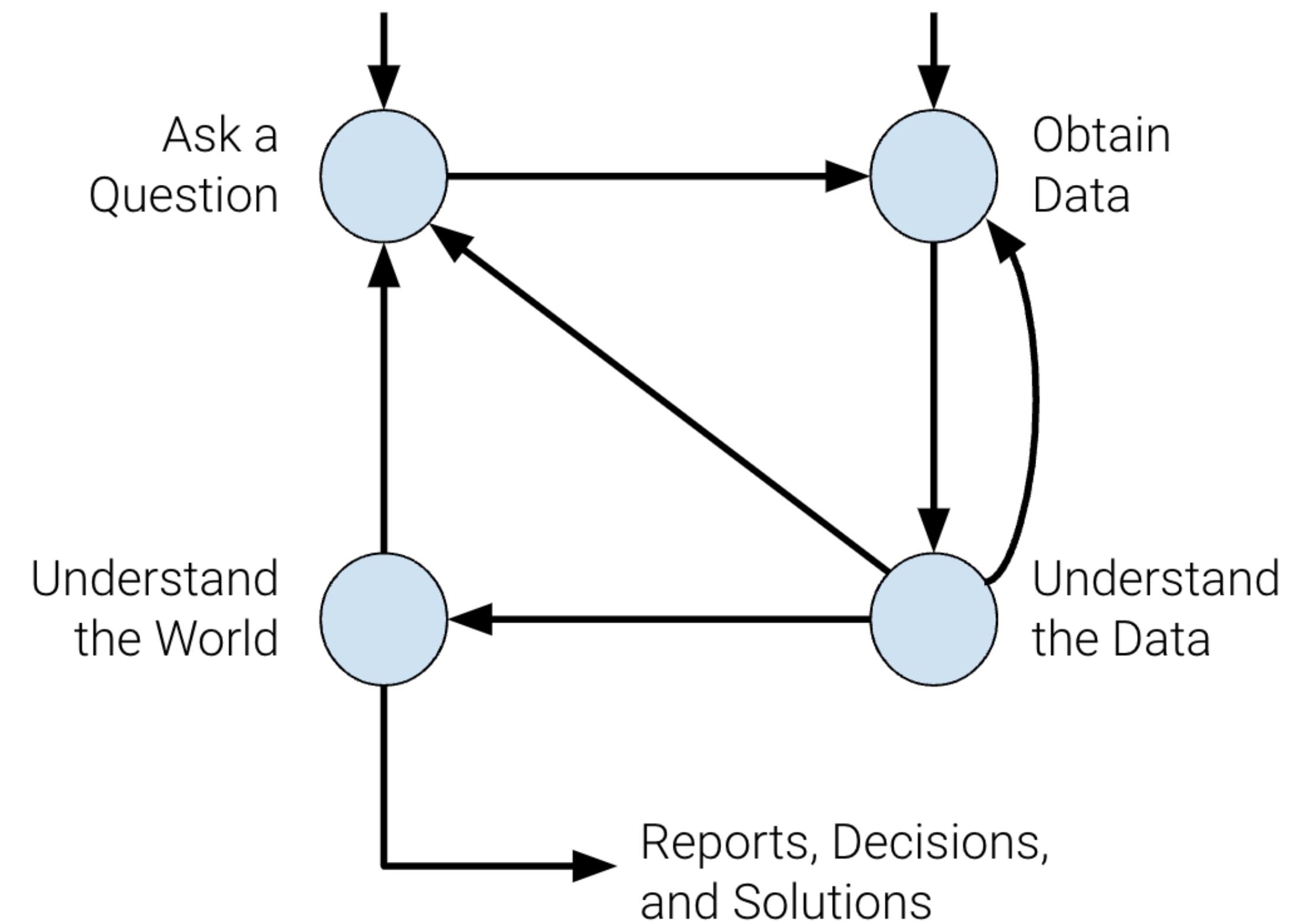
Tuomas Carsey, UNC

THE DATA SCIENCE LIFE CYCLE

Many different data science life cycles: core ideas similar



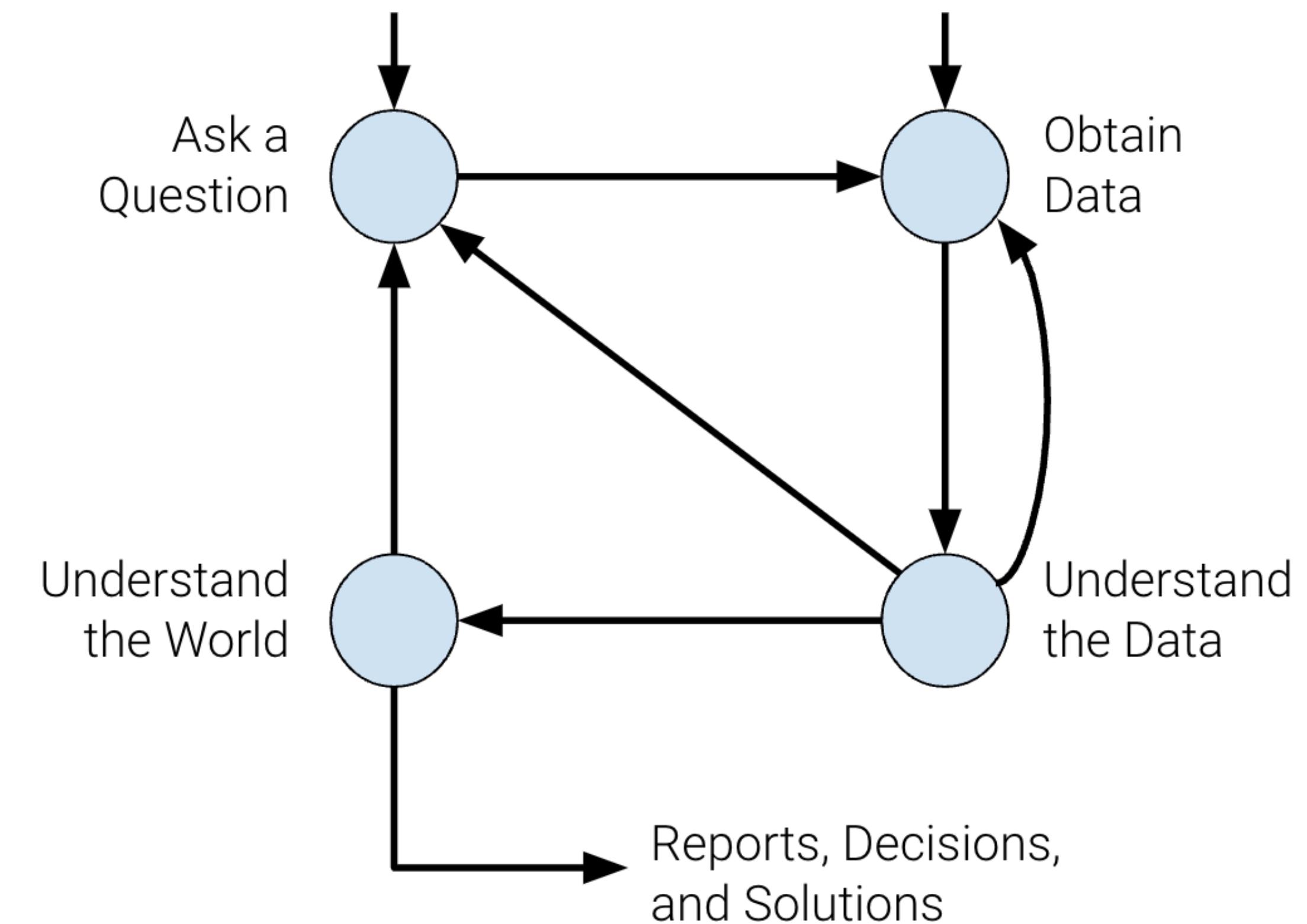
J. Dickerson, UMD



Data 100, UC Berkeley

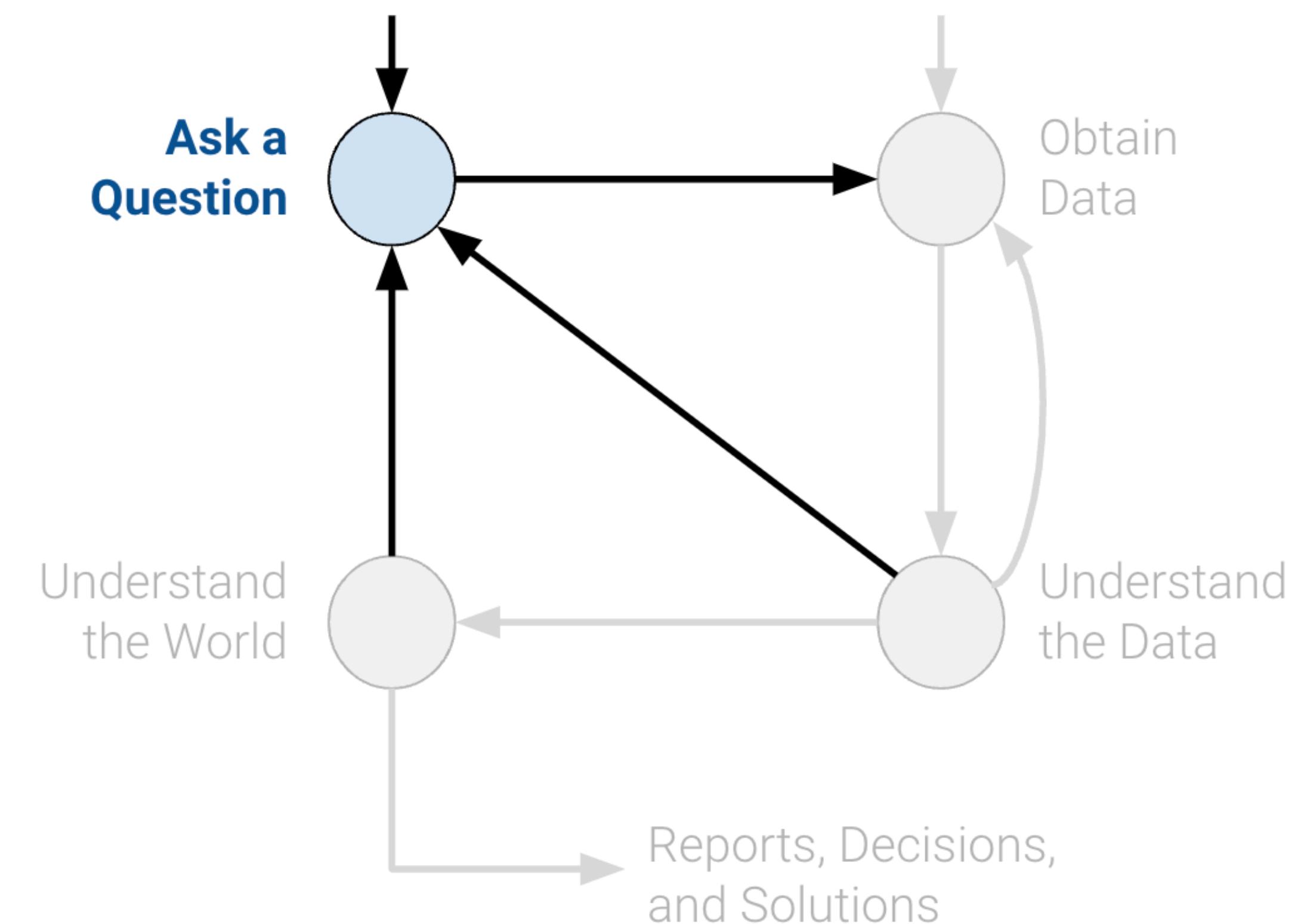
THE DATA SCIENCE LIFE CYCLE

- Many different data science life cycles: core ideas similar



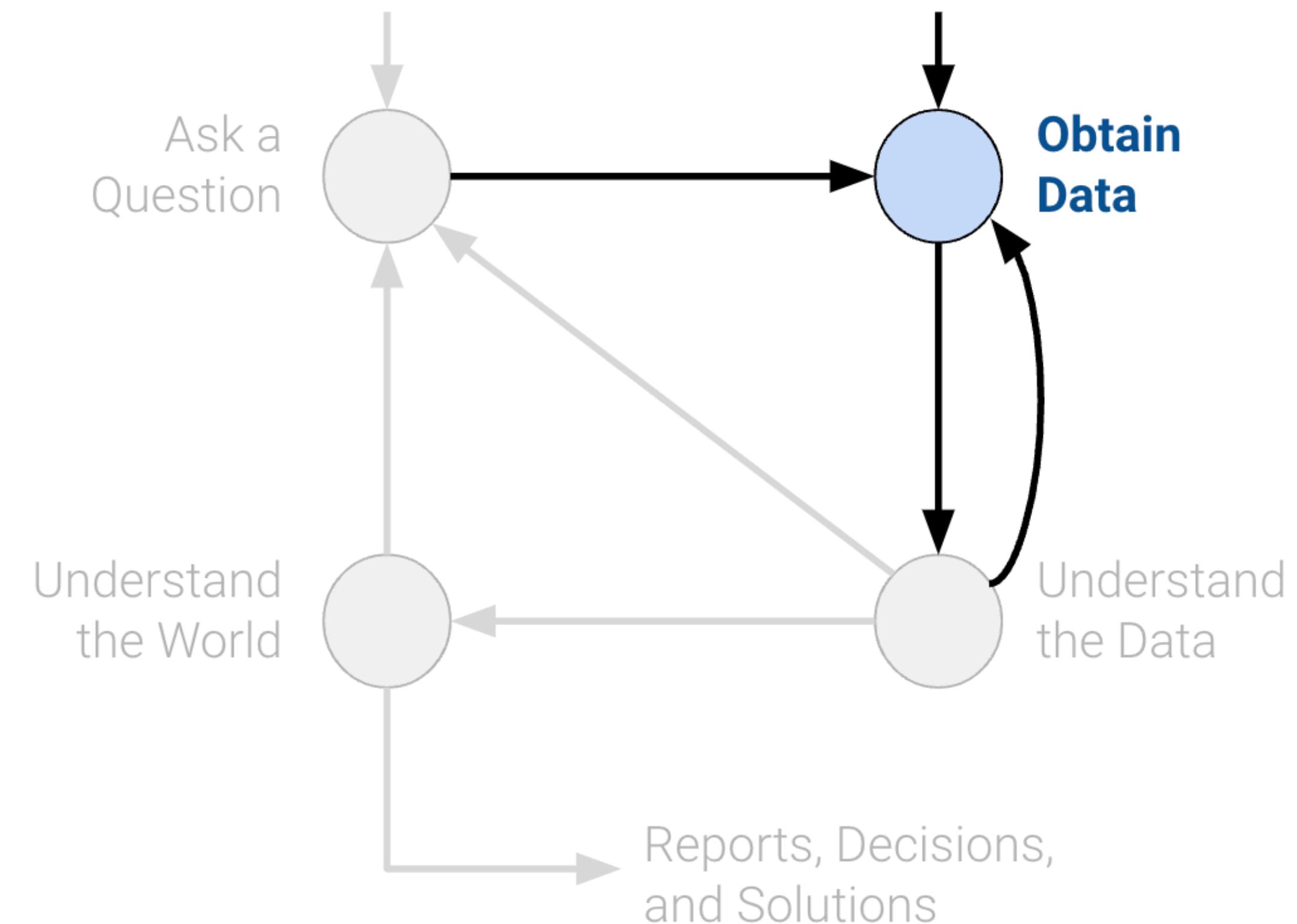
THE DATA SCIENCE LIFE CYCLE

- What do we want to know?
- What problems are we solving?
- What hypotheses are we testing?
- What are our success metrics?



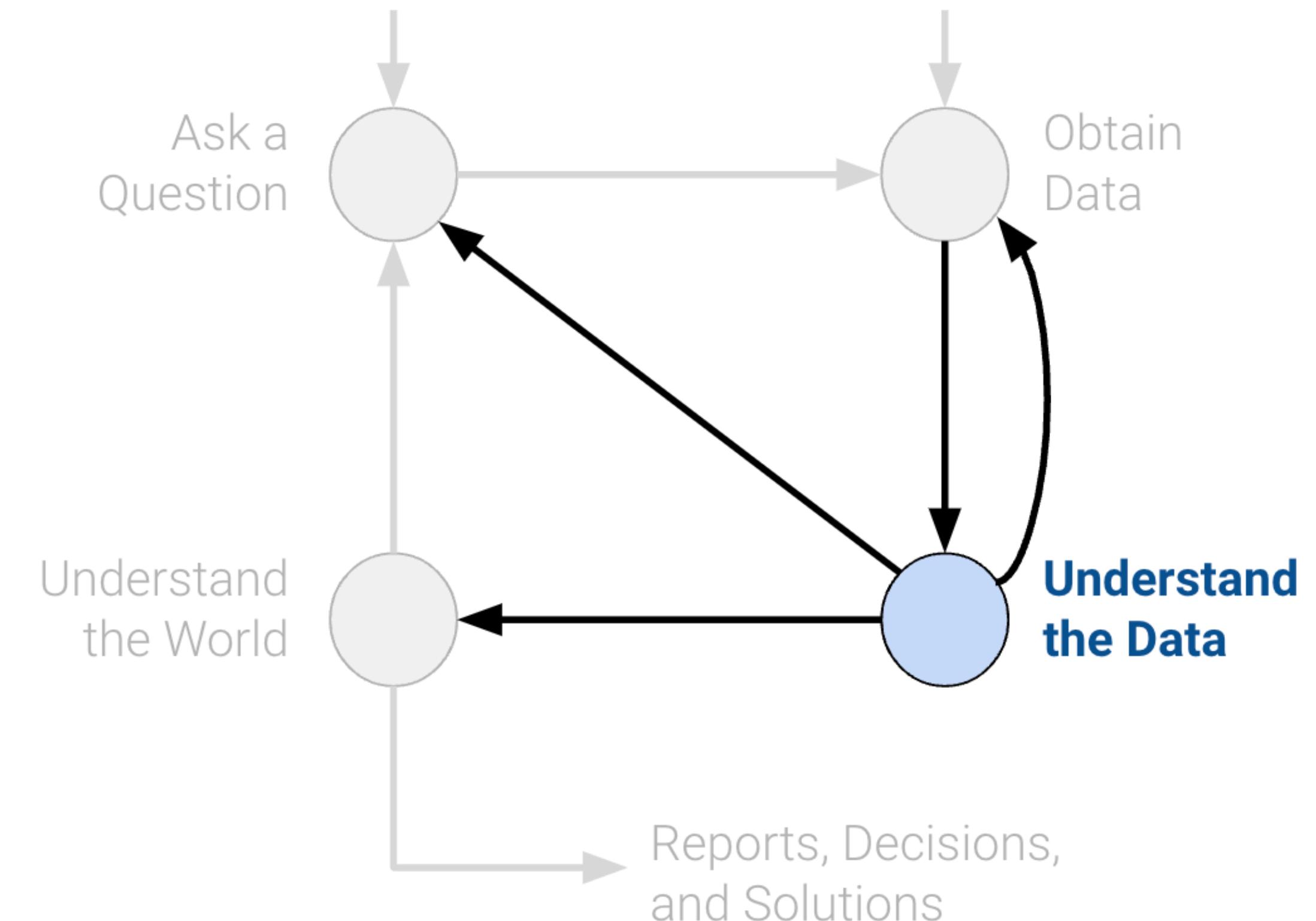
THE DATA SCIENCE LIFE CYCLE

- What data do we have and what data do we need?
- How will we sample more data?
- Is our data representative of the population we want to study?



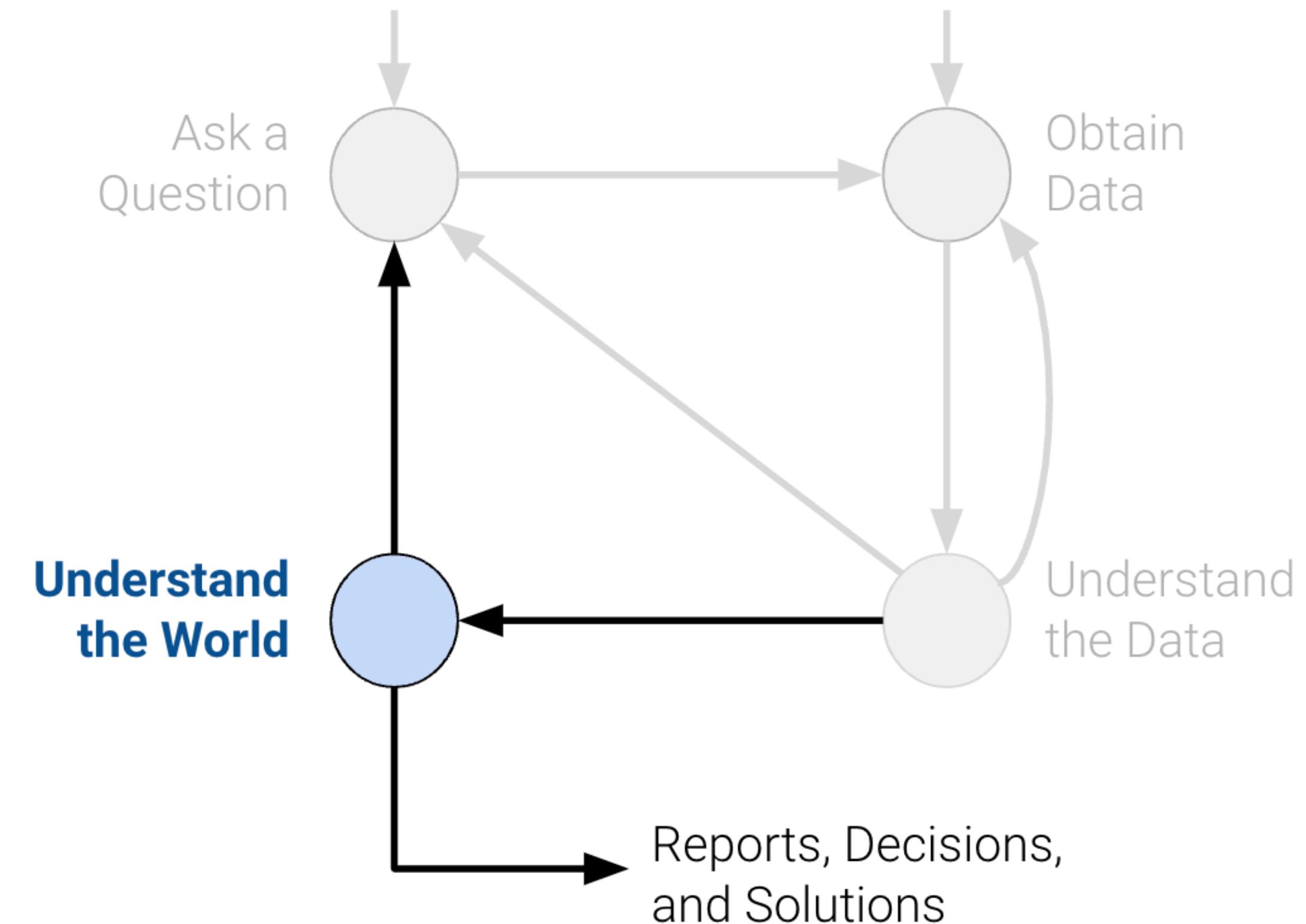
THE DATA SCIENCE LIFE CYCLE

- How is our data organized and what does it contain?
- Do we already have relevant data?
- What are the biases, anomalies, or other issues with the data?
- How do we transform the data to enable effective analysis?



THE DATA SCIENCE LIFE CYCLE

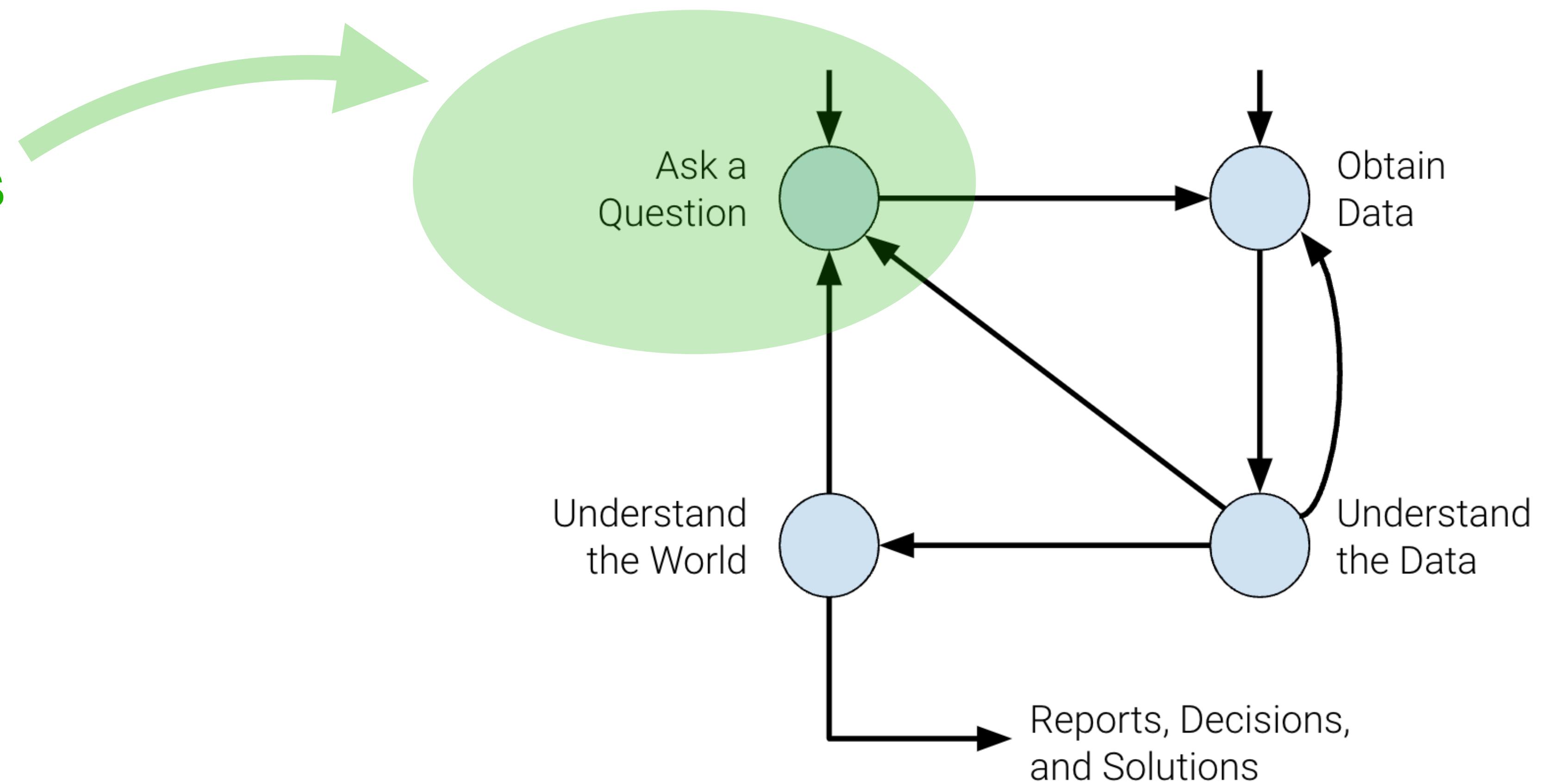
- What does the data say about the world?
- Does it answer our questions or accurately solve the problem?
- How robust are our conclusions and can we trust the predictions?



THE DATA SCIENCE LIFE CYCLE

- Many different data science life cycles: core ideas similar

Let's focus on this



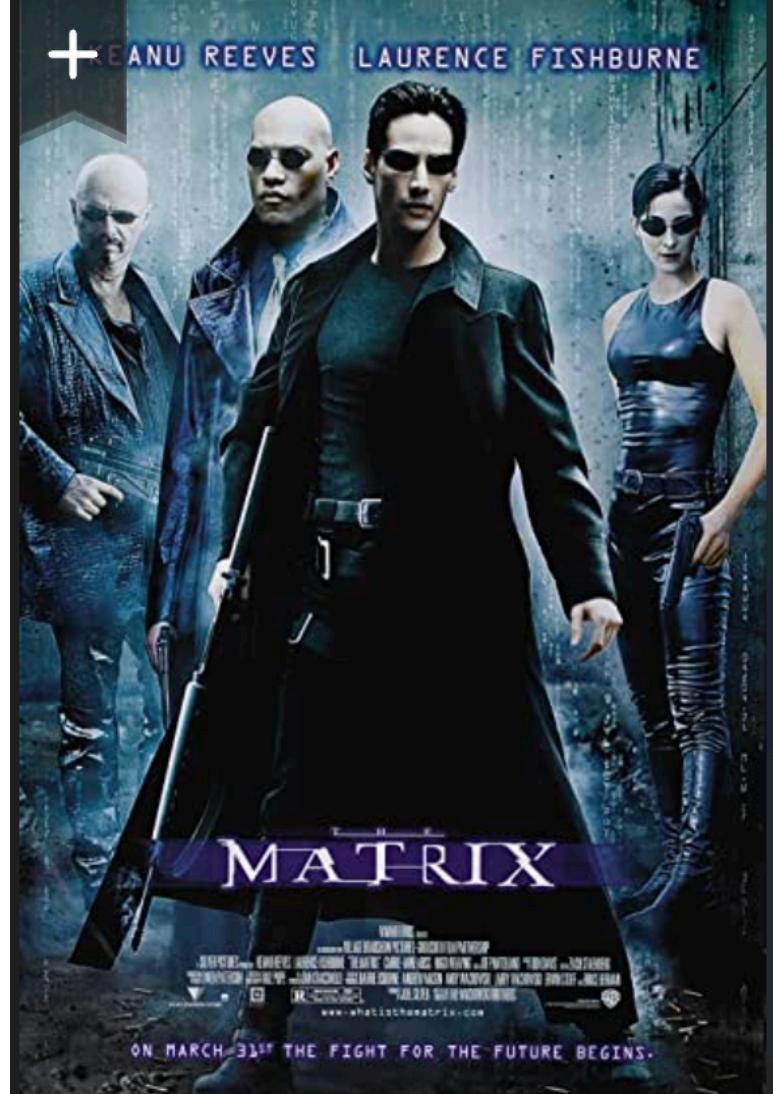
ASKING QUESTIONS

- The IMDb movie database: Movies

imdb.com/title/tt0133093/?ref_=chttp_i_16

The Matrix

1999 · R · 2h 16m



Action

Sci-Fi

When a beautiful stranger leads computer hacker Neo to a forbidding underworld, he discovers the shocking truth--the life he knows is the elaborate deception of an evil cyber-intelligence.

Directors [Lana Wachowski](#) · [Lilly Wachowski](#)

Writers [Lilly Wachowski](#) · [Lana Wachowski](#)

IMDb RATING
8.7 / 10
2M

STREAMING
max
PRIME VIDEO CHANNEL

+ Add to
Added

Screenshot

imdb.com/title/tt0133093/?ref_=chttp_i_16

Details

Release date [March 31, 1999 \(United States\)](#)

Countries of origin [United States](#) · [Australia](#)

Official sites [HBO Max \(United States\)](#) · [Official Facebook](#)

Language English

Also known as Ma Trận

Filming locations [Nashville, Tennessee, USA](#) (exterior scenes: skyline in opening Trinity rooftop chase)

Production companies [Warner Bros.](#) · [Village Roadshow Pictures](#) · [Groucho Film Partnership](#)

See more company credits at [IMDbPro](#)

Edit

Box office

Edit

Budget

\$63,000,000 (estimated)

Opening weekend US & Canada

\$27,788,331 · Apr 4, 1999

Gross US & Canada

\$172,076,928

Gross worldwide

\$467,222,728

Screenshot

ASKING QUESTIONS

- The IMDb movie database: Movies



Keanu Reeves

Biography

Jump to ▾

Overview

Born September 2, 1964 · Beirut, Lebanon

Birth name Keanu Charles Reeves

Nicknames The Wall · The One

Height 6' 1" (1.86 m)

Family

Children

No Children

Parents

Samuel Nowlin Reeves

Patric Reeves

Relatives

Kim Reeves (Sibling)

Karina Miller (Half Sibling)

Emma Reeves (Half Sibling)

Trademarks

Intense contemplative gaze

Deep husky voice

Known for playing stoic reserved characters

Friendly, down-to-earth personality



The Matrix (1999)

Full Cast & Crew

IMDbPro See agents for this cast & crew on IMDbPro

Directed by

Lana Wachowski

... (as The Wachowski Brothers)

Lilly Wachowski

... (as The Wachowski Brothers)

Writing Credits (WGA)

Lilly Wachowski

... (written by) (as The Wachowski Brothers) &

Lana Wachowski

... (written by) (as The Wachowski Brothers)

Cast (in credits order) verified as complete



Keanu Reeves

...

Neo



Laurence Fishburne

...

Morpheus



Carrie-Anne Moss

...

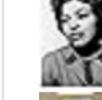
Trinity



Hugo Weaving

...

Agent Smith



Gloria Foster

...

Oracle



Joe Pantoliano

...

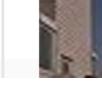
Cypher



Marcus Chong

...

Tank



Julian Arahanga

...

Apoc

ASKING QUESTIONS

- Stars live longer/shorter as compared to average person?
- Can we predict the rating of a new movie?
- Can we predict the filming locations?
Which features are important for it? Genre, budget, ...?
- What is the age distribution of actors and actresses in film?
When the characters they played are couples?
Any genres showing different distribution than the overall?
- What does the social network of actresses/actors look like?

FORMAT OF DATA



The screenshot shows the Wikipedia article for "Data science". The page title is "Data science". The main content discusses Data science as an interdisciplinary field that uses statistics, scientific computing, scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from potentially noisy, structured, or unstructured data. It also integrates domain knowledge from the underlying application domain (e.g., natural sciences, information technology, and medicine). The page includes a sidebar with sections like "Foundations", "Relationship to statistics", "Etymology", "Early usage", and "Modern usage". A callout box on the right side of the page provides a visual example of data analysis, showing a starfield with red dots representing the discovery of Comet NEOWISE.



	“interdisciplinary”	“Statistics”	...
“Data science”	5	10	
...			

Unstructured data

Structured data

- We will mostly assume the use of structured data.

THINGS TO WATCH OUT

- Quantitative vs Categorical Data

	“interdisciplinary”	“Statistics”	Document Category
“Data science”	5	10	1
...			

Document category:

1- Science, 2- Arts, 3- Politics, ...

Cannot treat these values as numbers, for anything other than simple identity testing

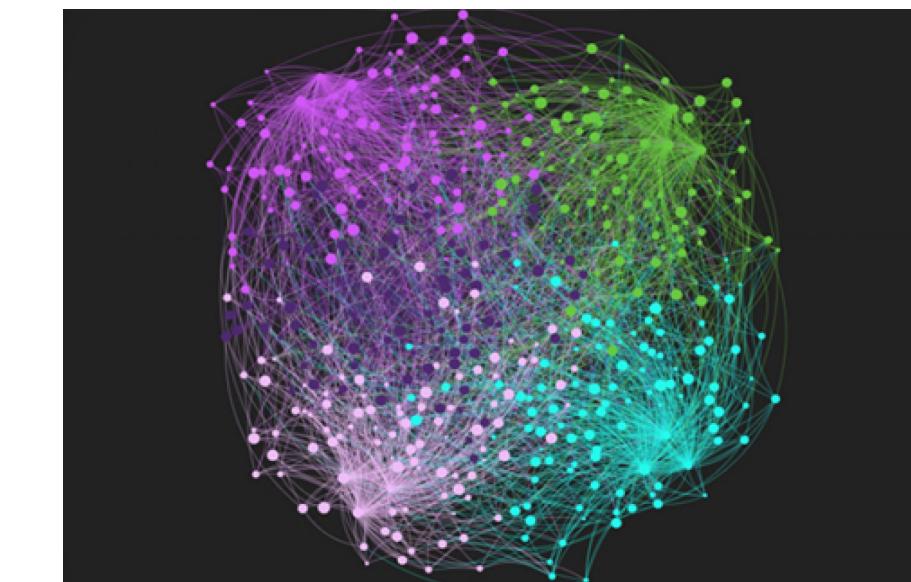
- Big data vs little data: Challenges,

The analysis cycle time slows as data size grows

Large data sets are complex to visualize

Simple models do not require massive data to fit or evaluate

Priority: Look for the right data rather than arbitrary big data



Max de Marzi

- Classification vs Regression

SOME TECHNOLOGIES/LIBRARIES WE WILL USE



MATH PRELIMINARIES: PROBABILITY

- Readings for the next few lectures:

Ch 5, 6 (Watkins)

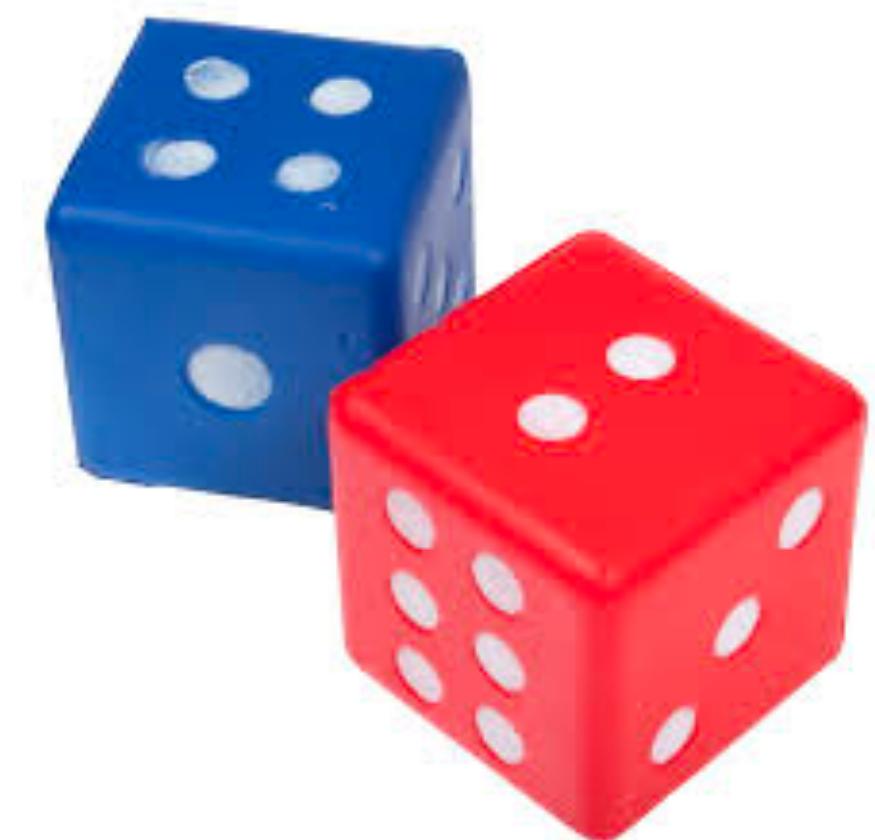
Ch 1 (Wasserman, for more formal treatment)

Ch 1, 2.1 (Skiena, for light motivational reading)

MATH PRELIMINARIES: PROBABILITY

- The *experiment* of tossing two dice: one red, one blue.

Any process in which possible outcomes can be identified ahead of time

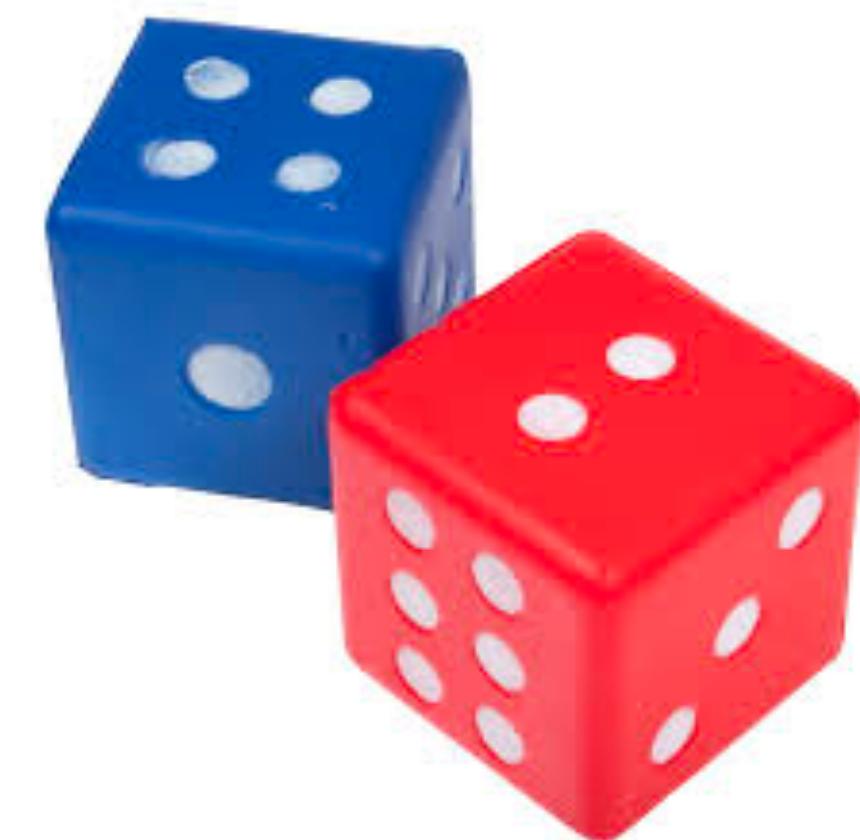


- Sample space Ω : set of all possible outcomes of an experiment
In our experiment, each outcome in the sample:
 $(\text{Value of red die}, \text{value of blue die})$
- Our example: What is $|\Omega|$?

MATH PRELIMINARIES: PROBABILITY

- Our sample space Ω :

$\{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}.$



- Event A : specified subset of the outcomes of an experiment

- Ex: Event A (Sum of dice = 7 or Sum of dice = 11)

$$A = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1), (5,6), (6,5)\}$$

AXIOMS OF PROBABILITY

- For any event $A \subseteq \Omega$:

$P(A)$: the probability that the event A occurs.

- Axiom-1: For every event A , $P(A) \geq 0$
- Axiom-2: $P(\Omega) = 1$
- Axiom-3: For any sequence of disjoint events A_1, A_2, \dots

$$P\left(\bigcup_{i \geq 1} A_i\right) = \sum_{i \geq 1} P(A_i)$$

PROBABILITY AND SET THEORY

- Formal mathematical model for events: Set theory

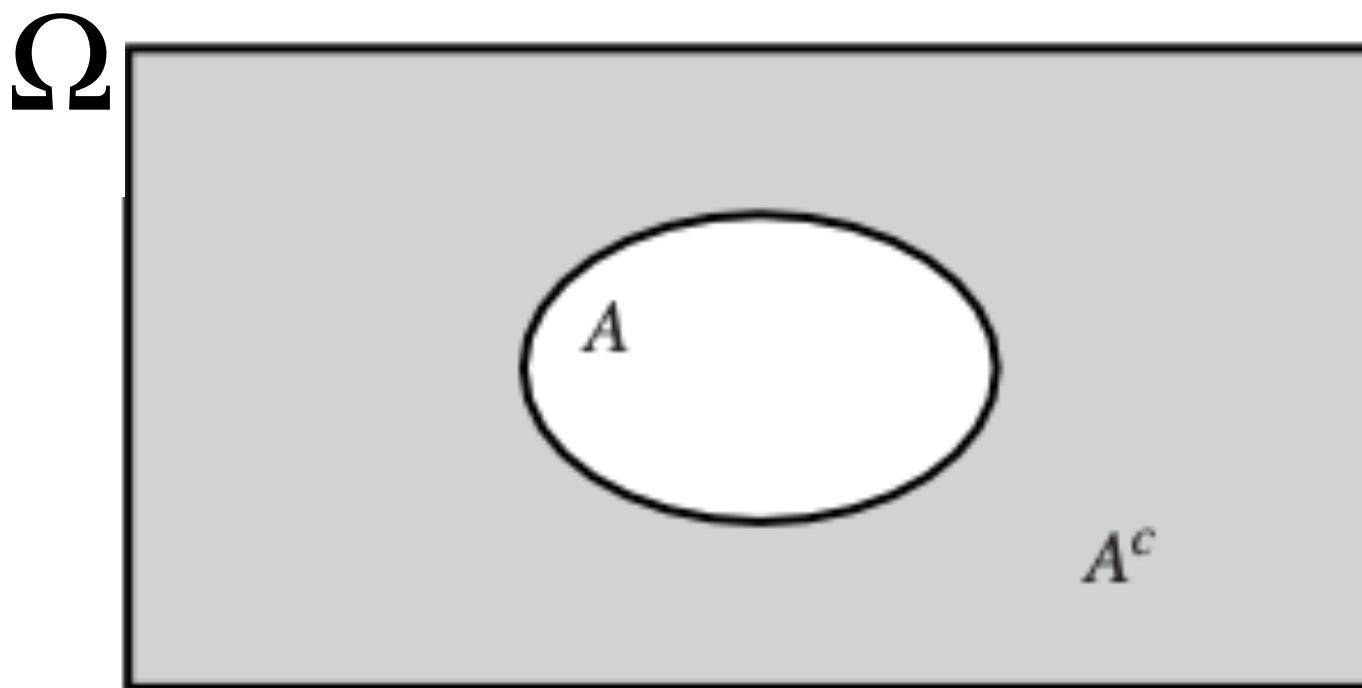
Sample space Ω : a set S

Outcome: an element of S

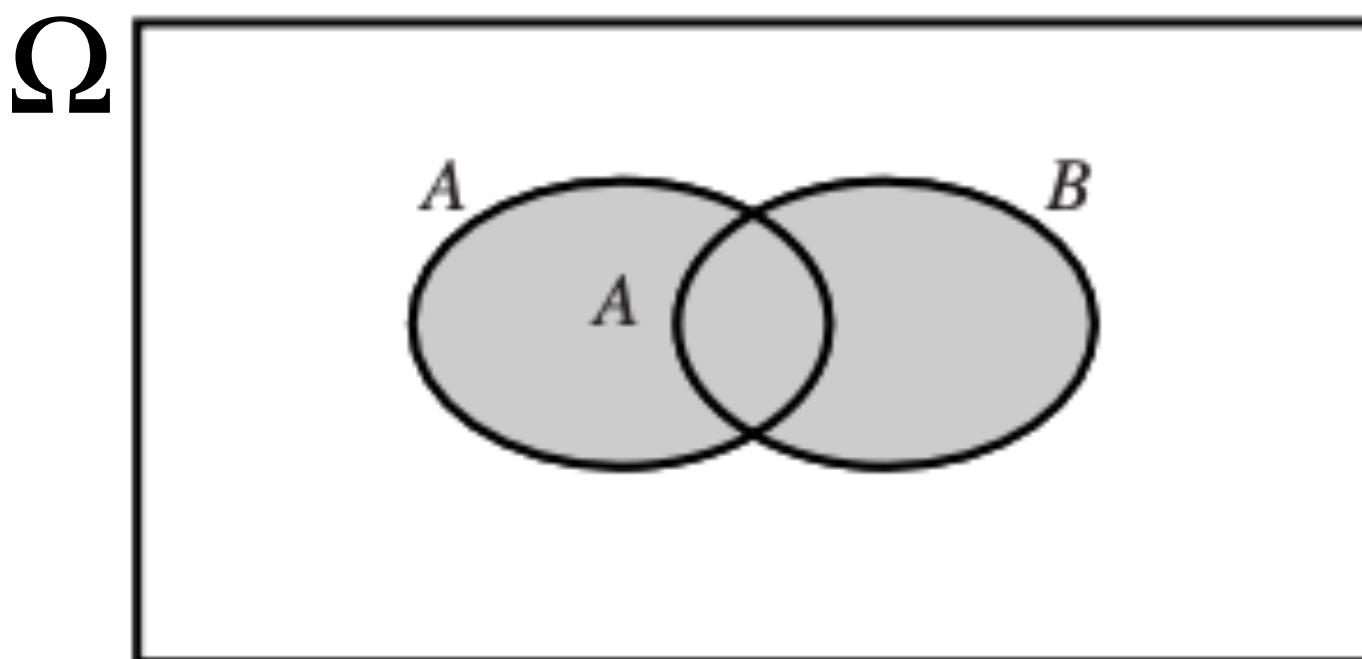
Event: a subset of S

- Further properties of probabilities:
deduce from axioms and set theory

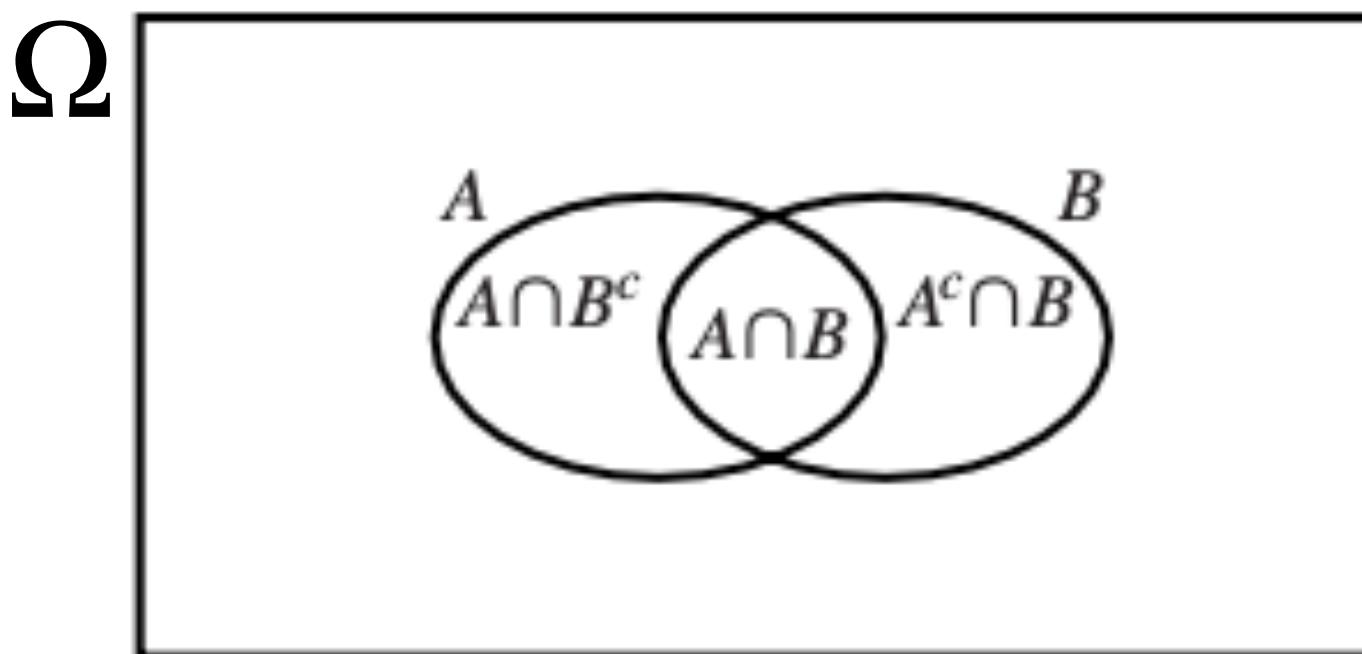
SET THEORY: VERY BRIEF REVIEW



The event A^c



The event $A \cup B$



The partition of $A \cup B$

FURTHER PROPERTIES OF PROBABILITY

- $P(\emptyset) = 0$ (\emptyset refers to an impossible event, e.g. 0 on a die)
- For an event A , $P(A) = 1 - P(A^c)$ (complement)
- $A \subset B \Rightarrow P(A) \leq P(B)$ (monotonicity)
- For any events A, B :
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
 (inclusion-exclusion)
- ...

SIMPLE SAMPLE SPACE

- Finite # of outcomes: s_1, s_2, \dots, s_n and each **equally likely**:

Event A has m outcomes $\Rightarrow P(A) = \frac{m}{n}$

Ex: Event A : (Sum of the dice = 7 or Sum of the dice = 11)

$$A = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1), (5,6), (6,5)\}$$

$$P(A) = \frac{8}{36}$$

FURTHER EXAMPLES

Ex: Disease diagnosis

Patient's symptoms: sore throat, fever

Doctor's diagnosis: viral or bacterial or both

Probability of viral = 0.8, probability of bacterial=0.4



Probability that the patient has both?

$$P(B \cup V) = P(B) + P(V) - P(B \cap V)$$

$$1 = 0.4 + 0.8 - P(B \cap V) \Rightarrow P(B \cap V) = 0.2$$

FURTHER EXAMPLES

Ex: Sampling without replacement

- A president (P) and a treasurer (T) chosen from 20 people including Alice, Bob.
- Probability that Alice is president and Bob is treasurer?

$$\text{\# of ways to select (P, T): } 20 \times 19 = 380 \Rightarrow \frac{1}{380}$$

EXAMPLE FROM PREVIOUS LECTURE

Ex: Sampling without replacement

- A president (P) and a treasurer (T) chosen from 20 people including Alice, Bob.
- Probability that Alice is president and Bob is treasurer?

$$\text{\# of ways to select (P, T): } 20 \times 19 = 380 \Rightarrow \frac{1}{380}$$

IN GENERAL

If ordered selection of k items out of n is done one at a time **without replacement**, there are $n \times (n - 1) \times \dots \times (n - k + 1) = \frac{n!}{(n - k)!}$ outcomes (k items in order of selection)

FURTHER EXAMPLES

Ex: Birthday problem

- Probability that 2 in a group of 20 have same birthday?

Size of $\Omega = 365^{20}$ (Sampling with replacement)

of outcomes s.t. all birthdays are different = $\frac{365!}{(365 - 20)!}$
(Sampling without replacement)

$$\Rightarrow \text{Probability} = 1 - \frac{365!}{(365 - 20)! \times 365^{20}} = 0.411$$

PROBABILITY IN LIGHT OF NEW INFORMATION

Judgment under Uncertainty: Heuristics and Biases

Biases in judgments reveal some heuristics of thinking under uncertainty.

Amos Tversky and Daniel Kahneman

Individual described by:

former neighbor as follows: “Steve is very shy and withdrawn, invariably helpful, but with little interest in people, or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.”

Is Steve a librarian or a farmer?

PROBABILITY IN LIGHT OF NEW INFORMATION

Judgment under Uncertainty: Heuristics and Biases

Biases in judgments reveal some heuristics of thinking under uncertainty.

Amos Tversky and Daniel Kahneman

Tempting to say Steve is a librarian, but ignoring prior probabilities is a mistake.

Individual described by:

former neighbor as follows: “Steve is very shy and withdrawn, invariably helpful, but with little interest in people, or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.”

Is Steve a librarian or a farmer?

PROBABILITY IN LIGHT OF NEW INFORMATION

Judgment under Uncertainty: Heuristics and Biases

Biases in judgments reveal some heuristics of thinking under uncertainty.

Amos Tversky and Daniel Kahneman

Tempting to say Steve is a librarian, but ignoring prior probabilities is a mistake.

- Assume:

- Ω only farmers (F), librarians (L). $|F| = 100$, $|L| = 5$
- 40% of L and 10% of F fit the description.

Individual described by:

former neighbor as follows: “Steve is very shy and withdrawn, invariably helpful, but with little interest in people, or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.”

Is Steve a librarian or a farmer?

PROBABILITY IN LIGHT OF NEW INFORMATION

Judgment under Uncertainty: Heuristics and Biases

Biases in judgments reveal some heuristics of thinking under uncertainty.

Amos Tversky and Daniel Kahneman

Tempting to say Steve is a librarian, but ignoring prior probabilities is a mistake.

- Assume:

- Ω only farmers (F), librarians (L). $|F| = 100$, $|L| = 5$
- 40% of L and 10% of F fit the description.

People fitting the description: 2 L and 10 F .

Individual described by:

former neighbor as follows: “Steve is very shy and withdrawn, invariably helpful, but with little interest in people, or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.”

Is Steve a librarian or a farmer?

PROBABILITY IN LIGHT OF NEW INFORMATION

- From the opposite direction:
Also a mistake to ignore new information and
not update prior probabilities.

PROBABILITY IN LIGHT OF NEW INFORMATION

- From the opposite direction:
Also a mistake to ignore new information and
not update prior probabilities.
- Probability a randomly picked person *Farmer* is $20/21$.

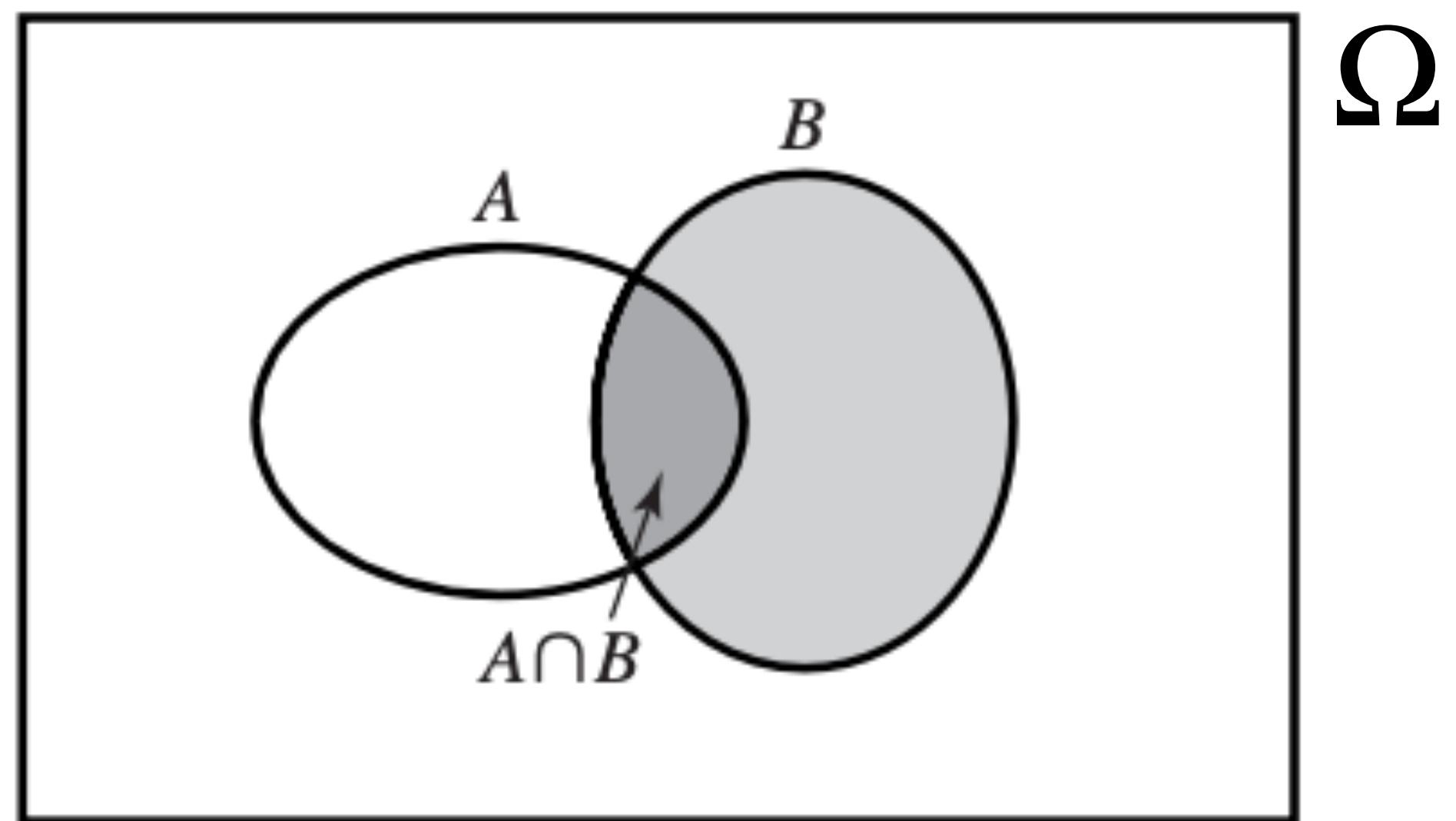
PROBABILITY IN LIGHT OF NEW INFORMATION

- From the opposite direction:
Also a mistake to ignore new information and not update prior probabilities.
- Probability a randomly picked person *Farmer* is 20/21.
Say new information. We know person picked fits description:

former neighbor as follows: “Steve is very shy and withdrawn, invariably helpful, but with little interest in people, or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.”
- Probability that picked person is *F* still 20/21? No! Decreases.

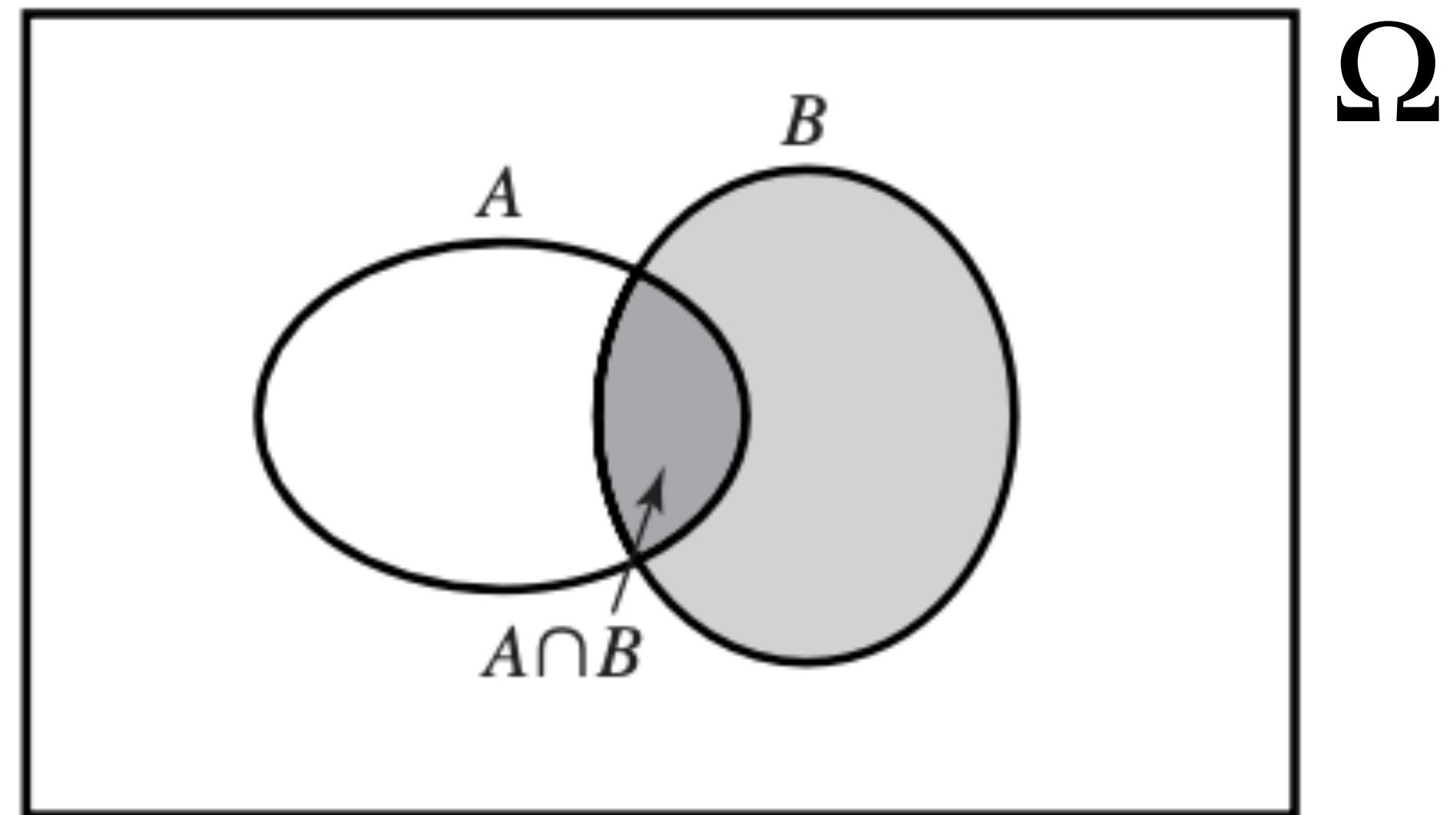
CONDITIONAL PROBABILITIES

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



CONDITIONAL PROBABILITIES

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



Interpret with the previous example:

- Ω only farmers (F), librarians (L). $|F| = 100$, $|L| = 5$
- 40% of L and 10% of F fit the description.

People fitting the description: 2 L and 10 F .

$$P(F | D) = \frac{P(F \cap D)}{P(D)} = (10/105)/(12/105) = 10/12$$

CONDITIONAL PROBABILITIES

Ex: Two fair dice rolled.

A : At least one of two dice is even,

B : Sum of two dice is 7 or 11. $P(A | B), P(B | A)$?



CONDITIONAL PROBABILITIES

Ex: Two fair dice rolled.

A : At least one of two dice is even,

B : Sum of two dice is 7 or 11. $P(A | B), P(B | A)$?



$$B \subset A \Rightarrow P(A \cap B) = P(B) \Rightarrow P(A | B) = P(A \cap B)/P(B) = 1$$

$$P(A) = 1 - P(A^c) = 1 - 9/36 = 27/36, \quad P(B) = 8/36$$

$$P(B | A) = P(A \cap B)/P(A) = (8/36)/(27/36) = 8/27$$

CONDITIONAL PROBABILITIES

In effect we are restricting the sample space.

Ex: 2 dice rolled repeatedly, sum T observed for each roll.

Probability that $T = 5$ observed before $T = 6$?

CONDITIONAL PROBABILITIES

In effect we are restricting the sample space.

Ex: 2 dice rolled repeatedly, sum T observed for each roll.

Probability that $T = 5$ observed before $T = 6$?

Restrict attention to new space: outcome either $T = 5$ or $T = 6$.

Let A : event that $T = 5$, B : event that either $T = 5$ or $T = 6$.

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{4/36}{4/36 + 5/36} = \frac{4}{9}$$

MULTIPLICATION PRINCIPLE

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

MULTIPLICATION PRINCIPLE

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

Note that we also have

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

MULTIPLICATION PRINCIPLE

Ex: 2 balls drawn, without replacement, from box with r red and b blue balls. Probability that 1st ball red and 2nd ball blue?

MULTIPLICATION PRINCIPLE

Ex: 2 balls drawn, without replacement, from box with r red and b blue balls. Probability that 1st ball red and 2nd ball blue?

Let A : event that 2nd ball blue, B : event that 1st ball red.

$$P(B) = \frac{r}{r+b}, P(A | B) = \frac{b}{r+b-1} \Rightarrow P(A \cap B) = \frac{rb}{(r+b)(r+b-1)}$$

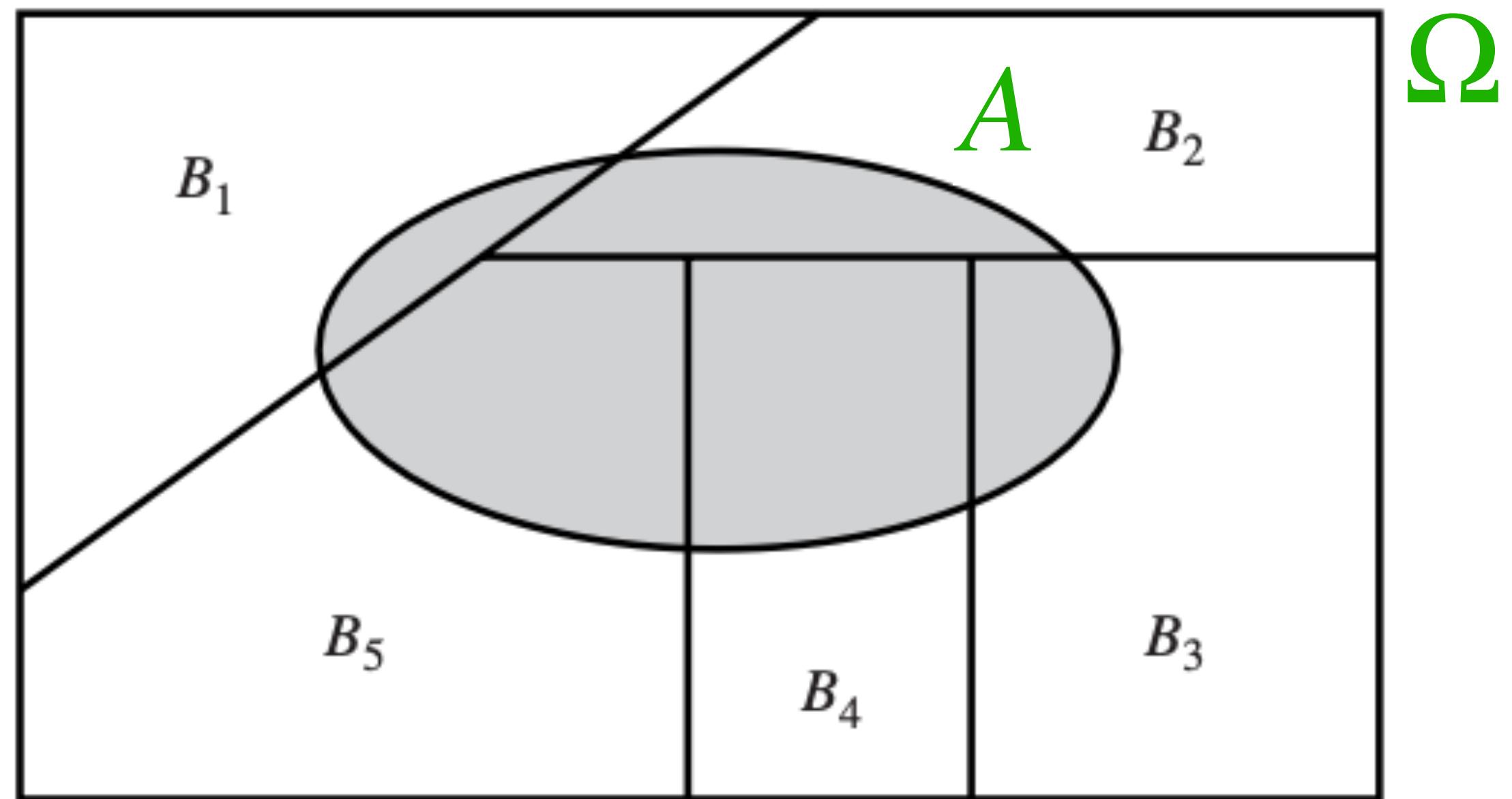
MULTIPLICATION PRINCIPLE

We can also generalize to three or more events. For three:

$$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B)$$

LAW OF TOTAL PROBABILITY

Partition of sample space makes probability calculations easy.



Ω

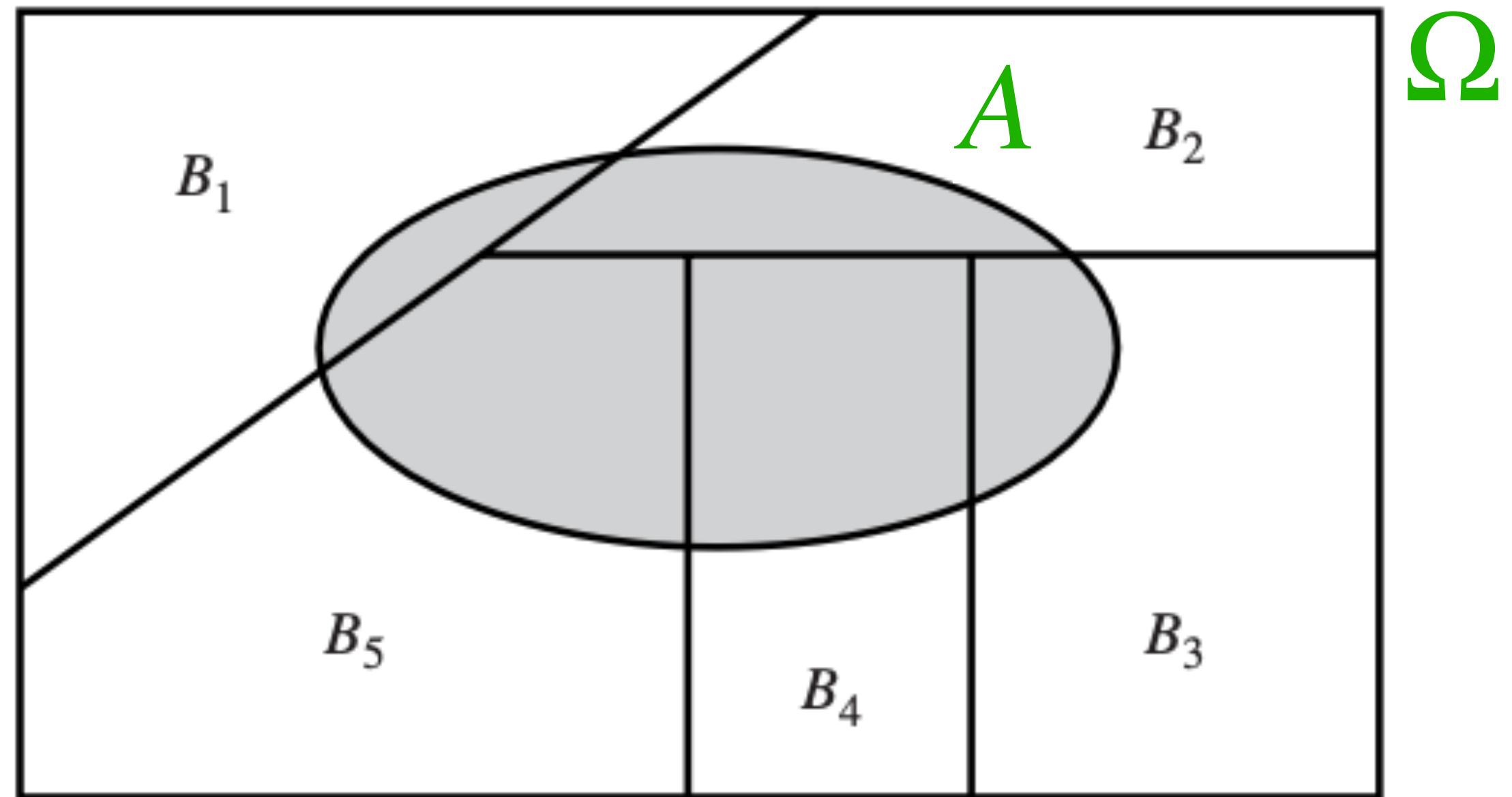
$$i \neq j \Rightarrow B_i \cap B_j = \emptyset$$

AND

$$\bigcup_{\forall i} B_i = \Omega$$

LAW OF TOTAL PROBABILITY

Partition of sample space makes probability calculations easy.



Ω

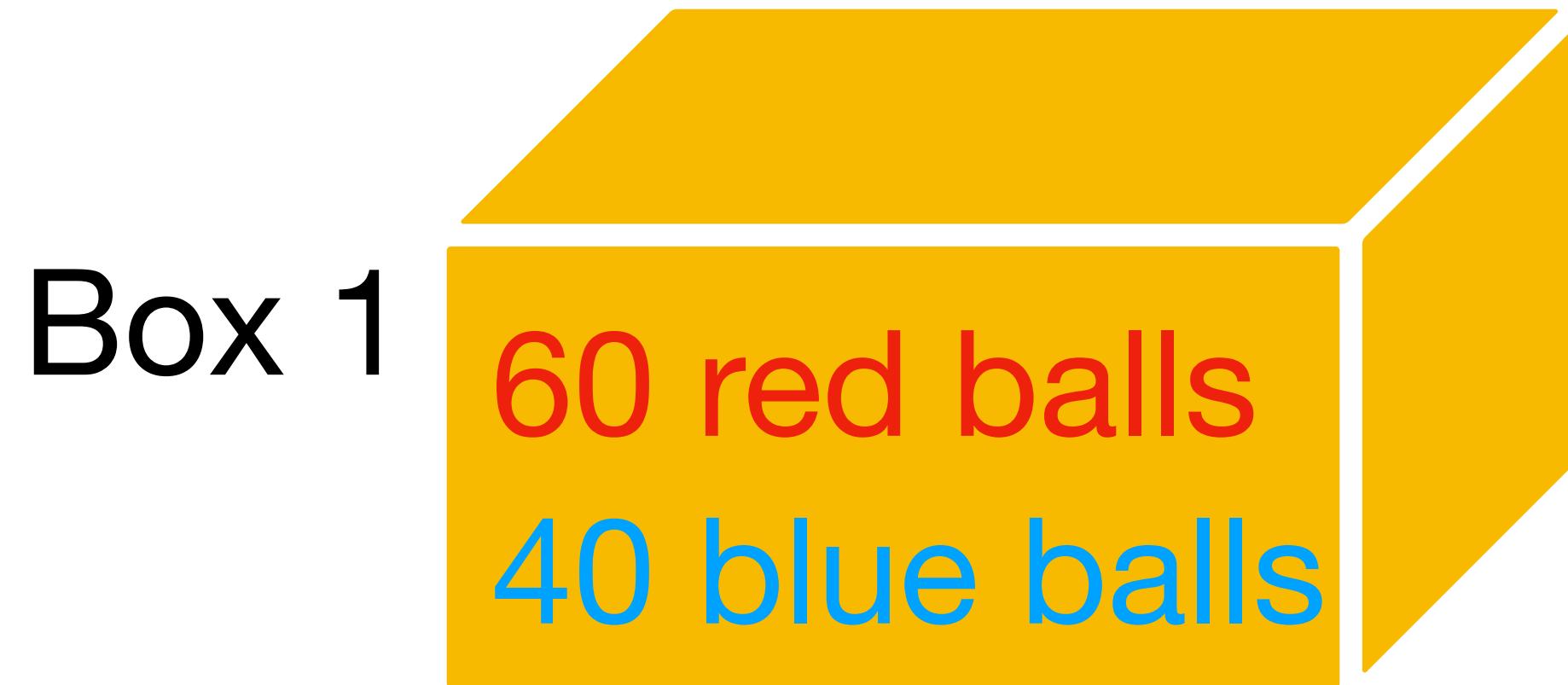
$$i \neq j \Rightarrow B_i \cap B_j = \emptyset$$

AND

$$\bigcup_{\forall i} B_i = \Omega$$

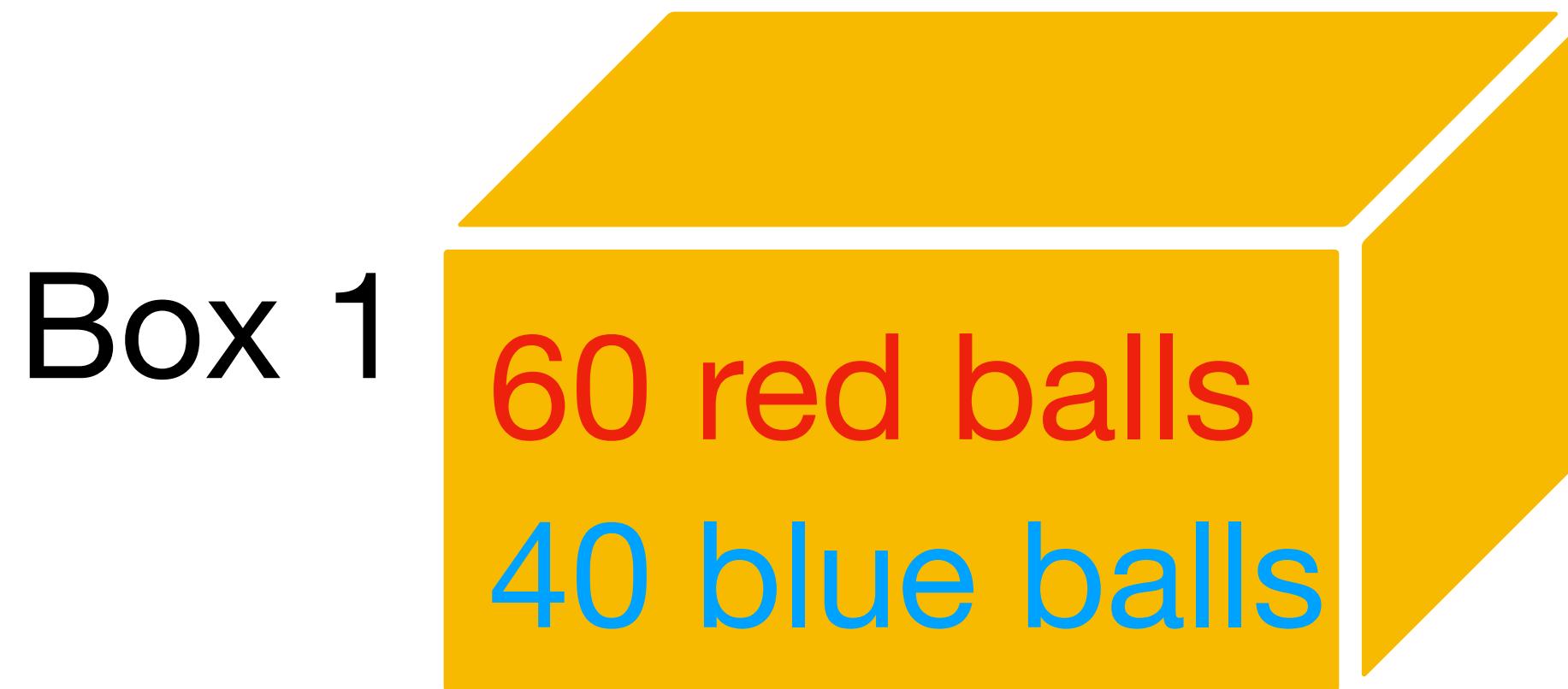
For any event A , $P(A) = \sum_{\forall i} P(B_i)P(A | B_i)$

LAW OF TOTAL PROBABILITY



Ex: Select a box randomly and select a ball from it randomly.
Probability that the selected ball is red?

LAW OF TOTAL PROBABILITY



Ex: Select a box randomly and select a ball from it randomly.
Probability that the selected ball is red?

B_1 : Box 1 selected, B_2 : Box 2 selected, A : a red ball is selected.

$$P(A) = \sum_{i=1,2} P(B_i)P(A | B_i) = \frac{1}{2} \times \frac{60}{100} + \frac{1}{2} \times \frac{10}{30} = \frac{7}{15}$$

BAYES' THEOREM

Box 1



Box 2



Ex: Someone selects a box, selects a ball from it (both random).
We are told it is red. Probability that Box 1 was selected?

BAYES' THEOREM

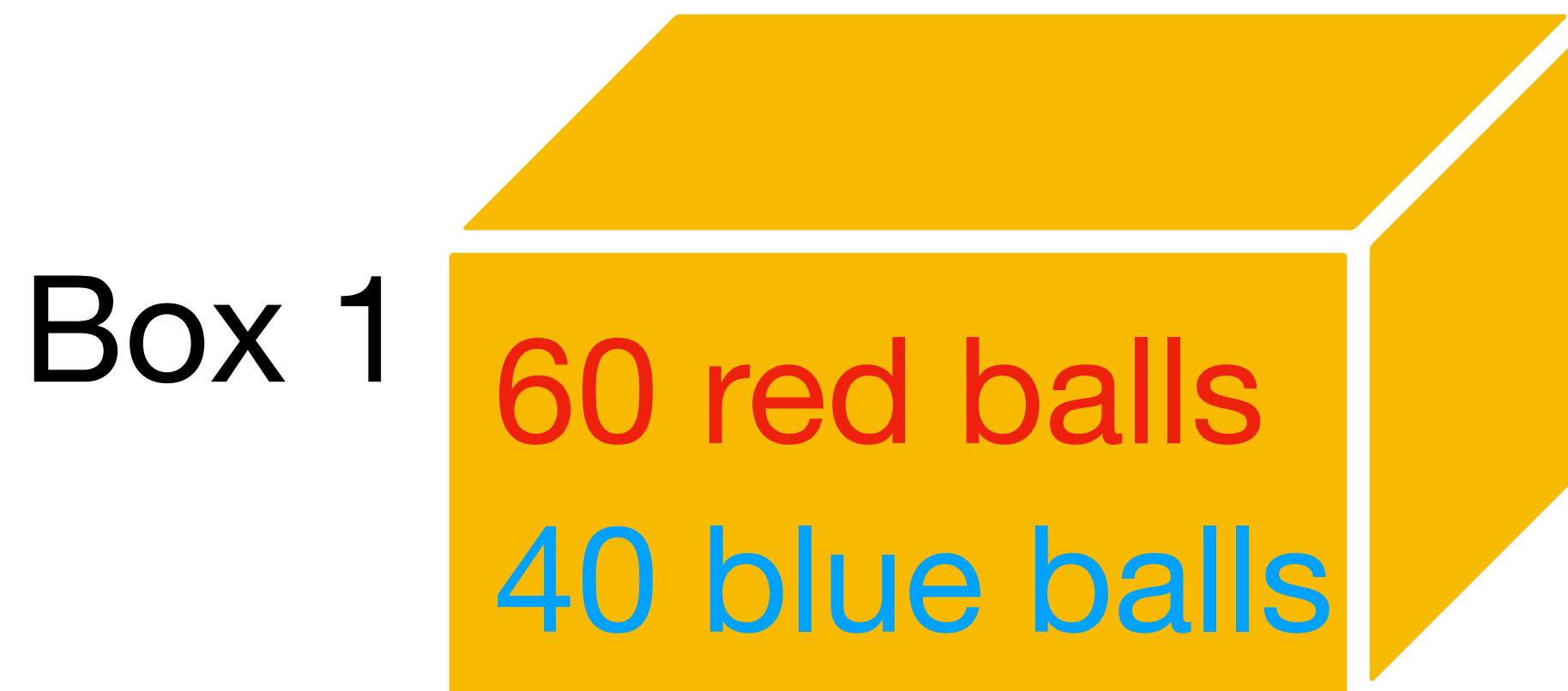


Ex: Someone selects a box, selects a ball from it (both random). We are told it is red. Probability that Box 1 was selected?

B_1 : Box 1 selected, B_2 : Box 2 selected, A : a red ball is selected.

$$P(B_1 | A) = \frac{P(A \cap B_1)}{P(A)} = \frac{P(B_1)P(A | B_1)}{\sum_{i=1,2} P(B_i)P(A | B_i)} = \frac{\frac{1}{2} \times \frac{3}{5}}{\frac{7}{15}} = \frac{9}{14}$$

BAYES' THEOREM



Ex: Someone selects a box, selects a ball from it (both random). We are told it is red. Probability that Box 1 was selected?

B_1 : Box 1 selected, B_2 : Box 2 selected, A : a red ball is selected.

$$P(B_1 | A) = \frac{P(A \cap B_1)}{P(A)} = \frac{P(B_1)P(A | B_1)}{\sum_{i=1,2} P(B_i)P(A | B_i)} = \frac{\frac{1}{2} \times \frac{3}{5}}{\frac{7}{15}} = \frac{9}{14}$$

BAYES' THEOREM

$$P(B_i | A) = \frac{P(B_i)P(A | B_i)}{\sum_{j=1}^k P(B_j)P(A | B_j)}$$

where $B_1, B_2 \dots, B_k$ partition Ω .

BAYES' THEOREM

Posterior probability

Prior probability

$$P(B_i | A) = \frac{P(B_i)P(A | B_i)}{\sum_{j=1}^k P(B_j)P(A | B_j)}$$

where $B_1, B_2 \dots, B_k$ partition Ω .

Updates prior probability of B_i

based on information gathered from A

REVIEW

$$\text{Conditional Probability: } P(A | B) = \frac{P(A \cap B)}{P(B)}$$

$$\text{Multiplication Rule: } P(A \cap B) = P(A)P(B | A) = P(B)P(A | B)$$

$$\text{Law of Total Probability: } P(A) = \sum_{\forall i} P(B_i)P(A | B_i)$$

$$\text{Bayes' Theorem: } P(B_i | A) = \frac{P(B_i)P(A | B_i)}{\sum_{j=1}^k P(B_j)P(A | B_j)}$$

Note: Conditional probabilities are probabilities.

$$P(A \cap B | C) = P(A | C)P(B | A \cap C)$$

INDEPENDENCE

Learning that B has occurred doesn't change probability of A .

Formally $P(A | B) = P(A)$. (or vice versa)

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

INDEPENDENCE

Learning that B has occurred doesn't change probability of A .

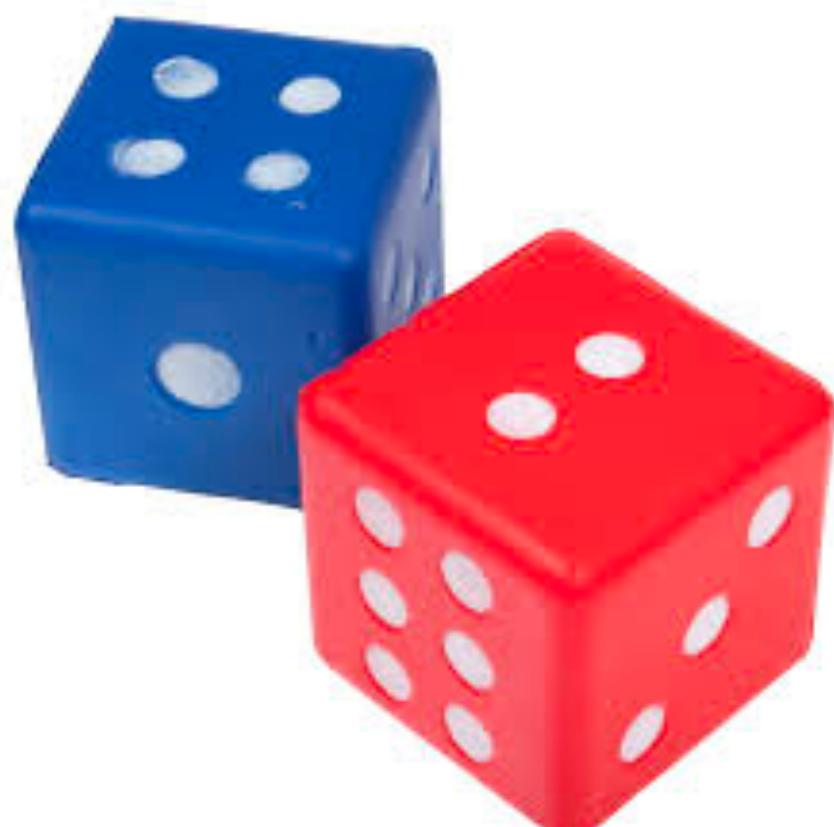
Formally $P(A | B) = P(A)$. (or vice versa)

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

Ex: Rolling two fair dice.

A : Red die is a 1, B : Blue die is a 2.

$$P(A \cap B) = \frac{1}{36} \text{ and } P(A)P(B) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$



INDEPENDENCE

Ex: Flipping two coins. First coin c_1 flipped, then coin c_2 .

c_1 : Fair coin c_2 : If c_1 is H , flip a fair coin

o.w. flip unfair coin with $P(H) = 0.2$

A : Coin c_1 is H , B : Coin c_2 is H . Are A, B independent?

INDEPENDENCE

Ex: Flipping two coins. First coin c_1 flipped, then coin c_2 .

c_1 : Fair coin c_2 : If c_1 is H , flip a fair coin

o.w. flip unfair coin with $P(H) = 0.2$

A : Coin c_1 is H , B : Coin c_2 is H . Are A, B independent?

$$P(B|A) = \frac{1}{2}, P(B) = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{2}{10} = \frac{7}{20}$$

$\Rightarrow P(B|A) \neq P(B) \Rightarrow$ not independent.

INDEPENDENCE

Ex: Flipping two coins. First coin c_1 flipped, then coin c_2 .

A : Coin c_1 is H , B : Coin c_2 is H .

For $r \neq 1$, what q (in terms of p, r) makes A, B independent?

$q = p$ makes them independent.

INDEPENDENCE

Ex: Flipping two coins. First coin c_1 flipped, then coin c_2 .

c_1 : Fair coin c_2 : If c_1 is H , flip a fair coin

o.w. flip unfair coin with $P(H) = 0.2$

A : Coin c_1 is H , B : Coin c_2 is H . Are A, B independent?

$$P(B|A) = \frac{1}{2}, P(B) = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{2}{10} = \frac{7}{20}$$

$\Rightarrow P(B|A) \neq P(B) \Rightarrow$ not independent. What is $P(A \cap B)$?

INDEPENDENCE OF SEVERAL EVENTS

A_1, A_2, \dots, A_n are independent if \forall subsets $A_{i1}, A_{i2}, \dots, A_{ij}$,

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_j}) = \prod_{1 \leq k \leq j} P(A_{i_k})$$

INDEPENDENCE OF SEVERAL EVENTS

Ex: The house or the player?

- 4 dice are rolled:
House wins if at least one die is a 6,
otherwise player wins.
- What is the probability that the house wins?



INDEPENDENCE OF SEVERAL EVENTS

Ex: The house or the player?

- 4 dice are rolled:
House wins if at least one die is a 6,
otherwise player wins.
- What is the probability that the house wins?
- Multiple ways:

Events on the dice: A : Die-1 is a 6, B : Die-2 is a 6, ...

$$P(A \cup B \cup C \cup D) = P(A) + P(B) + \dots - P(A \cap B) - P(A \cap C) \dots + P(A \cap B \cap C) + \dots - P(A \cap B \cap C \cap D)$$



INDEPENDENCE OF SEVERAL EVENTS

Ex: The house or the player?

- 4 dice are rolled:
House wins if at least one die is a 6,
otherwise player wins.
- What is the probability that the house wins?
- Multiple ways:

Events on the dice: A : Die-1 is a 6, B : Die-2 is a 6, ...

$$P(A \cup B \cup C \cup D) = P(A) + P(B) + \dots - P(A \cap B) - P(A \cap C) \dots + P(A \cap B \cap C) + \dots - P(A \cap B \cap C \cap D)$$

Events on # of 6s: A : exactly one die 6, B : exactly two dice 6s, ...

$$\binom{4}{1}5^3/6^4 + \binom{4}{2}5^2/6^4 + \binom{4}{3}5^1/6^4 + \binom{4}{4}5^0/6^4$$



INDEPENDENCE OF SEVERAL EVENTS

Ex: The house or the player?

- 4 dice are rolled:
House wins if at least one die is a 6,
otherwise player wins.
- What is the probability that the house wins?
- Much easier way: Think of the complement

H : At least one die is a 6 $\Rightarrow H^c$: No die is a 6.

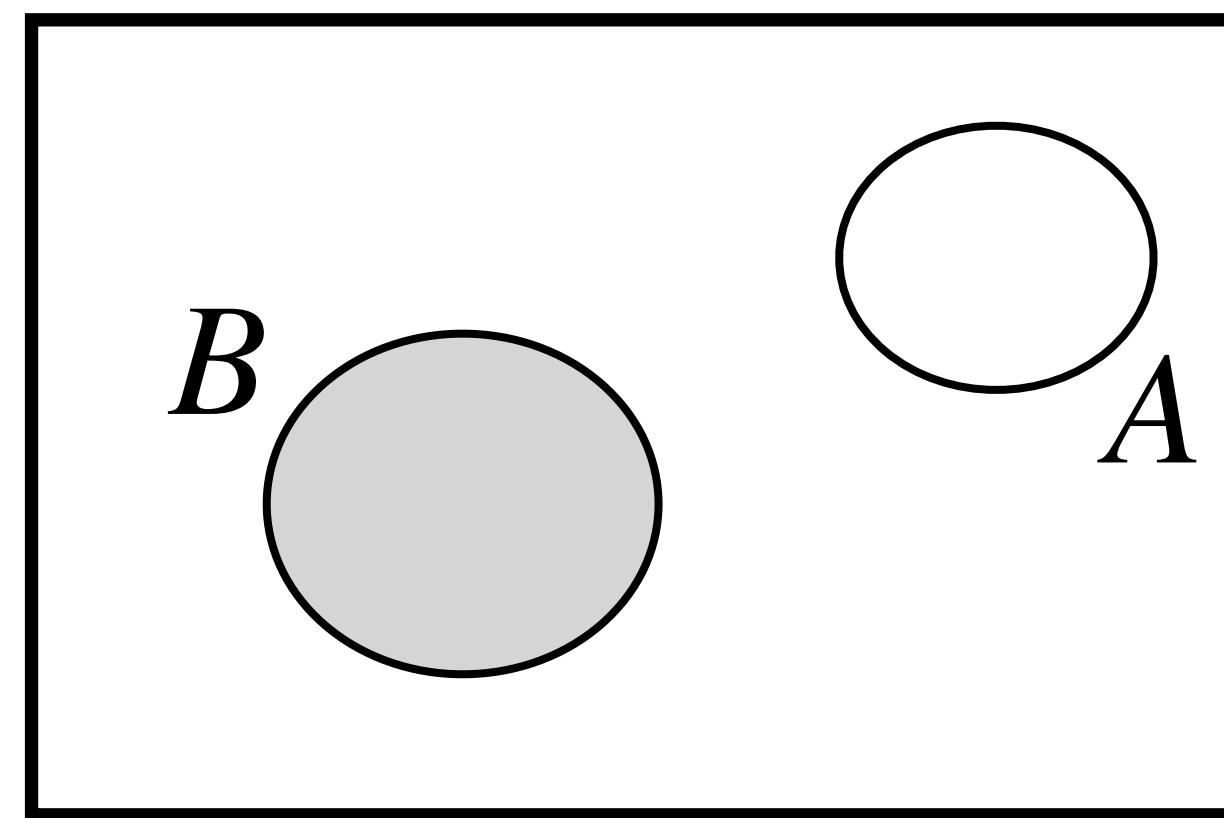
$$P(H) = 1 - P(H^c) = 1 - (5/6)^4 = 0.518$$



Take-home message: The house always wins!

DISJOINT EVENTS VS INDEPENDENT EVENTS

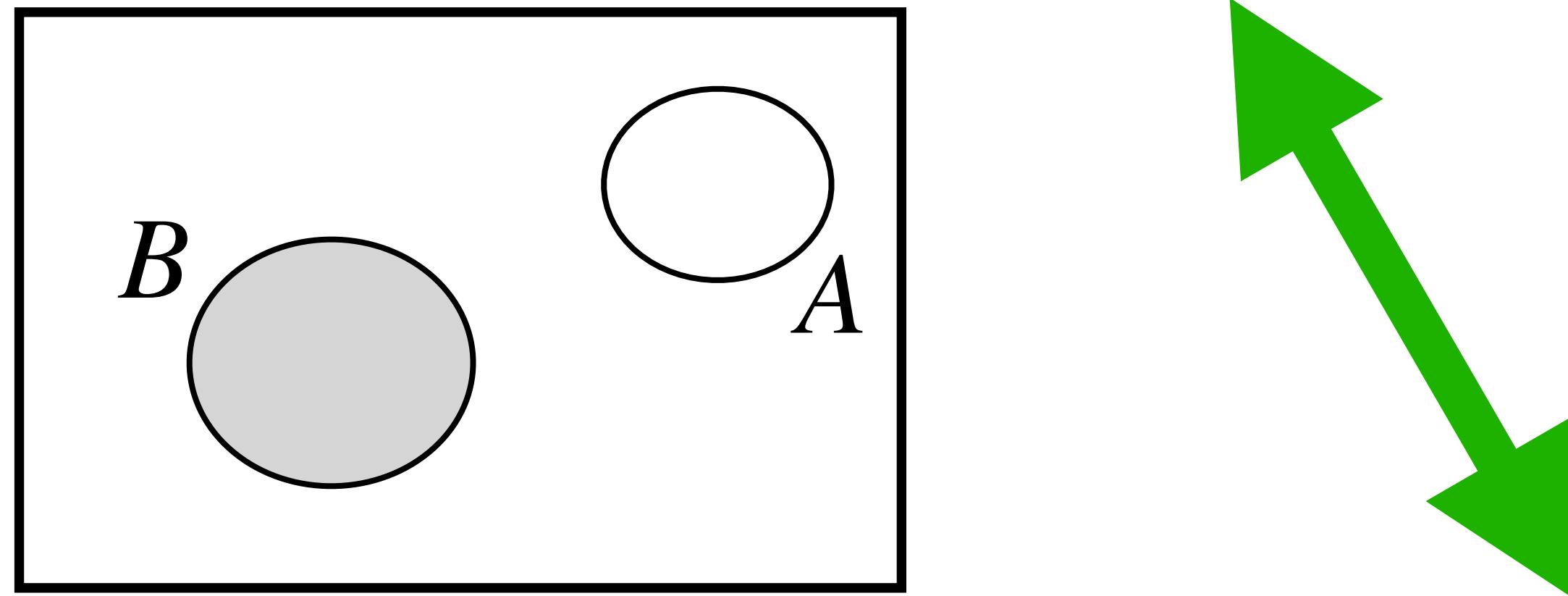
A, B disjoint $\Rightarrow A, B$ dependent (almost always)



Knowing B occurred changes $P(A)$ to 0 (unless it already was 0)

DISJOINT EVENTS VS INDEPENDENT EVENTS

A, B disjoint $\Rightarrow A, B$ dependent (almost always)



Knowing B occurred changes $P(A)$ to 0 (unless it already was 0)

How about the other direction?

Does $A \cap B \neq \emptyset$ imply A, B independent? **NO!**

RANDOM VARIABLES AND DISTRIBUTIONS

- Readings for the next few lectures:

Ch 7 (Watkins)

Ch 2 (Wasserman, for more formal treatment)

RANDOM VARIABLES

- A real-valued function defined on Ω is a **random variable**.

Examples:

X : # of heads when a coin is tossed 10 times
gives an integer in range [0,10].

$Y = 10 - X$ also a random variable (# of tails)

RANDOM VARIABLES

- A real-valued function defined on Ω is a **random variable**.

Examples:

X : # of heads when a coin is tossed 10 times
gives an integer in range [0,10].

$Y = 10 - X$ also a random variable (# of tails)

X : the sum of the values of two fair dice
gives an integer in range [2,12].

$P(X = 1) = 0, P(X = 2) = 1/36, P(X = 3) = 2/36, \dots$

RANDOM VARIABLES

- A real-valued function defined on Ω is a **random variable**.

Examples:

X : # of heads when a coin is tossed 10 times
gives an integer in range [0,10].

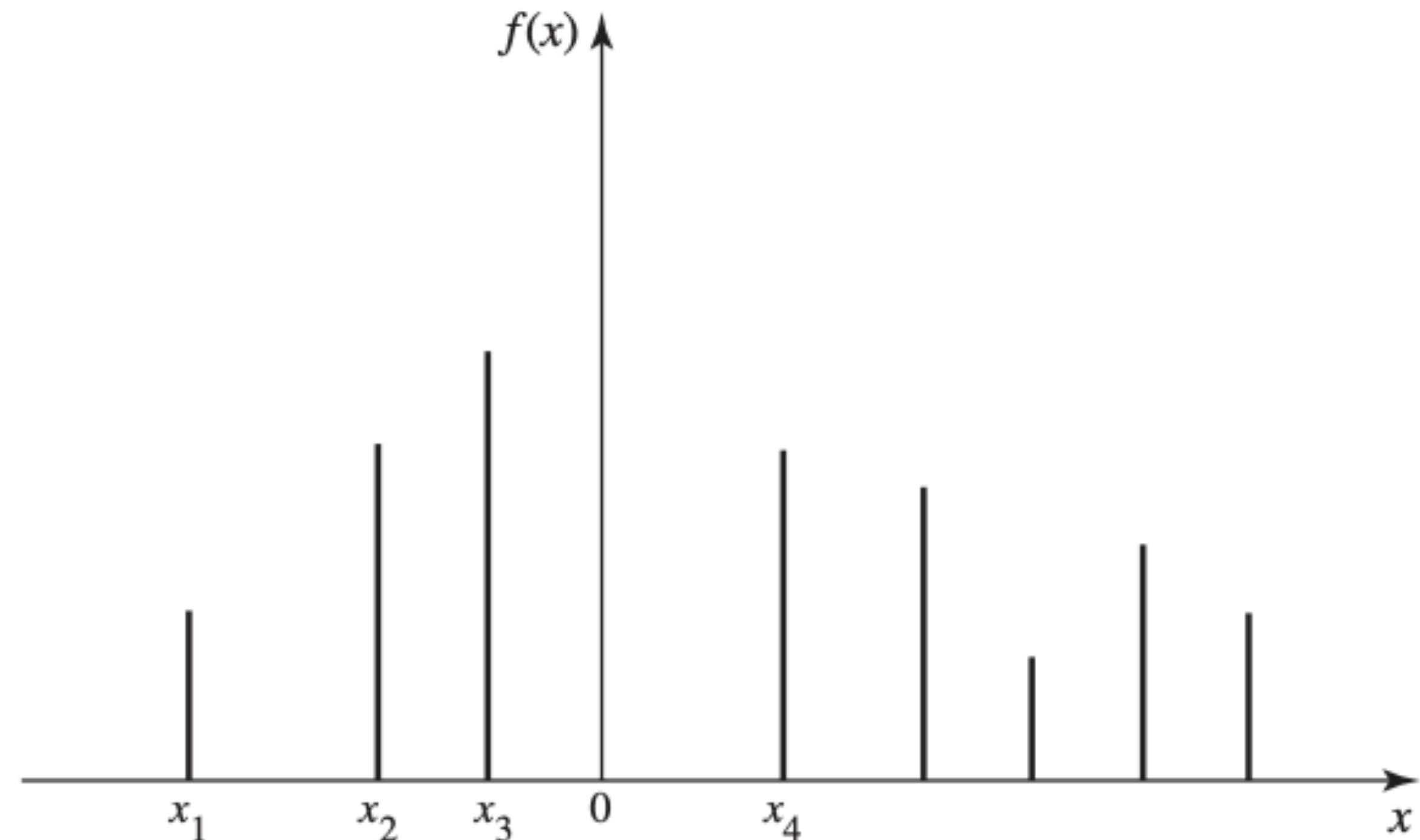
$Y = 10 - X$ also a random variable (# of tails)

X : the sum of the values of two fair dice
gives an integer in range [2,12].

$P(X = 1) = 0, P(X = 2) = 1/36, P(X = 3) = 2/36, \dots$

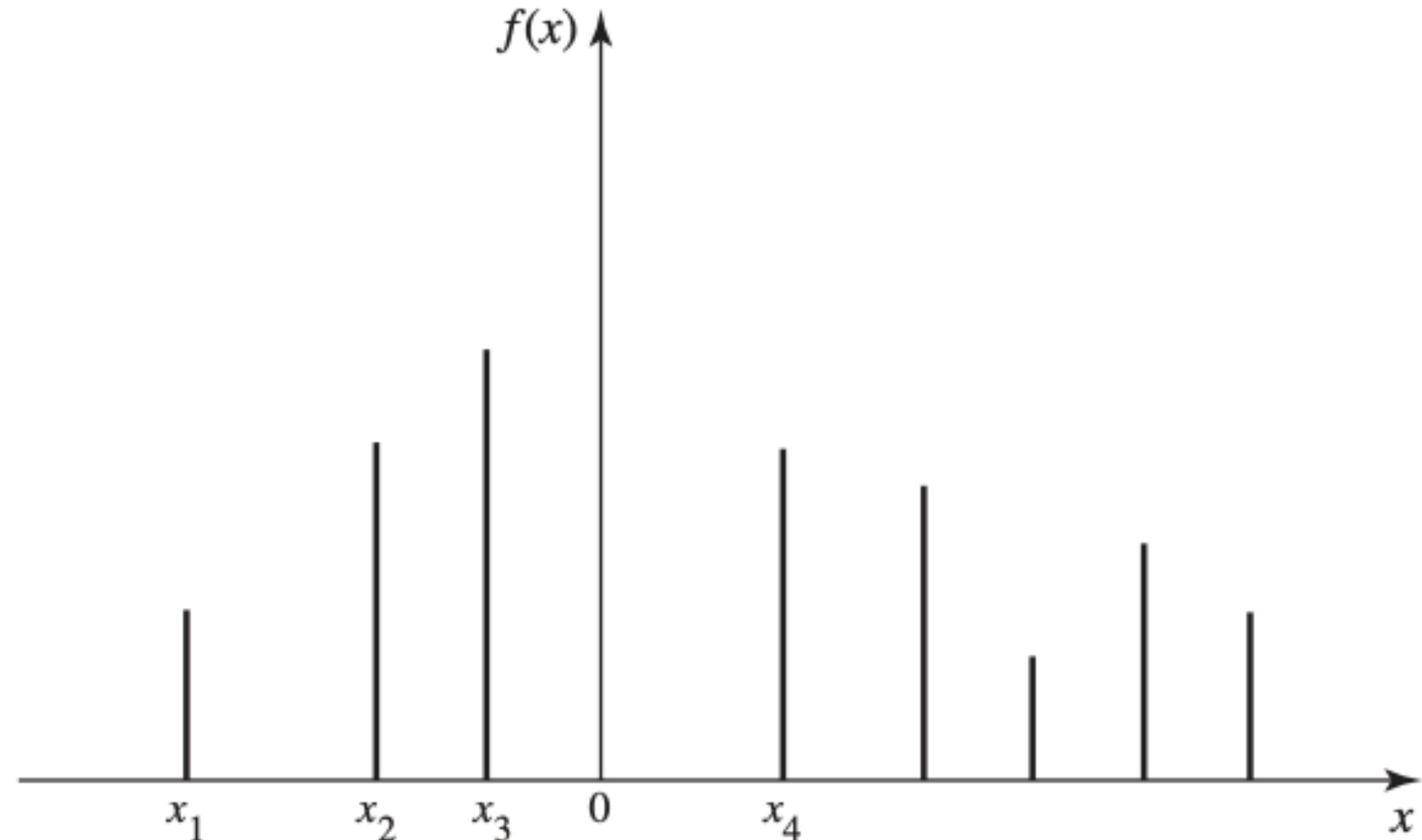
X : height of a person randomly selected from a population

DISCRETE DISTRIBUTIONS AND RANDOM VARIABLES



- If \exists finite (or countably infinite) values X can take on,
 $\Rightarrow X$ discrete random variable.

DISCRETE DISTRIBUTIONS AND RANDOM VARIABLES

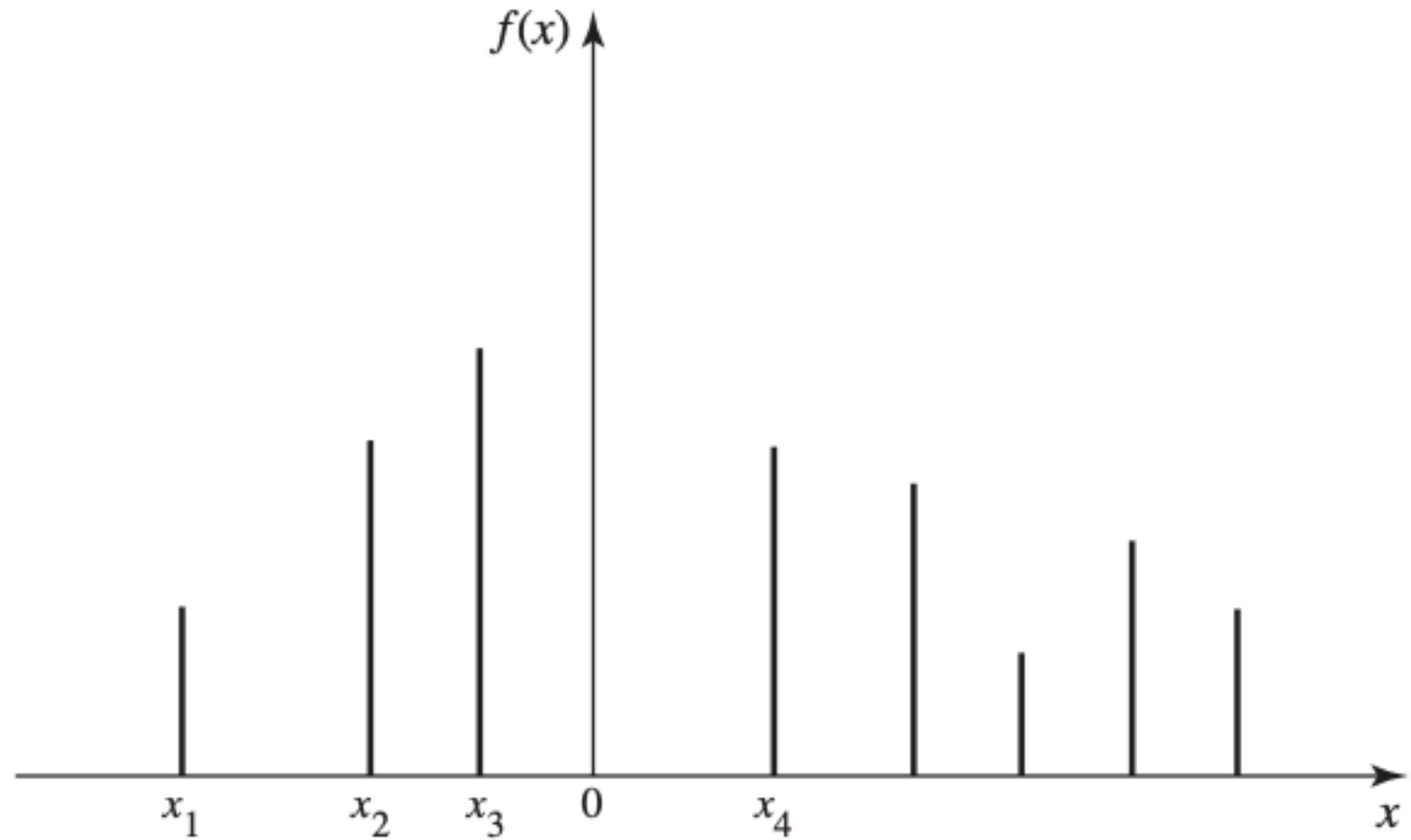


- If \exists finite (or countably infinite) values X can take on,
 $\Rightarrow X$ discrete random variable.

$f(x) = P(X = x)$ $\forall x \in \mathbb{R}$ is the probability mass function of X

$\{x : f(x) > 0\}$ is the support of X .

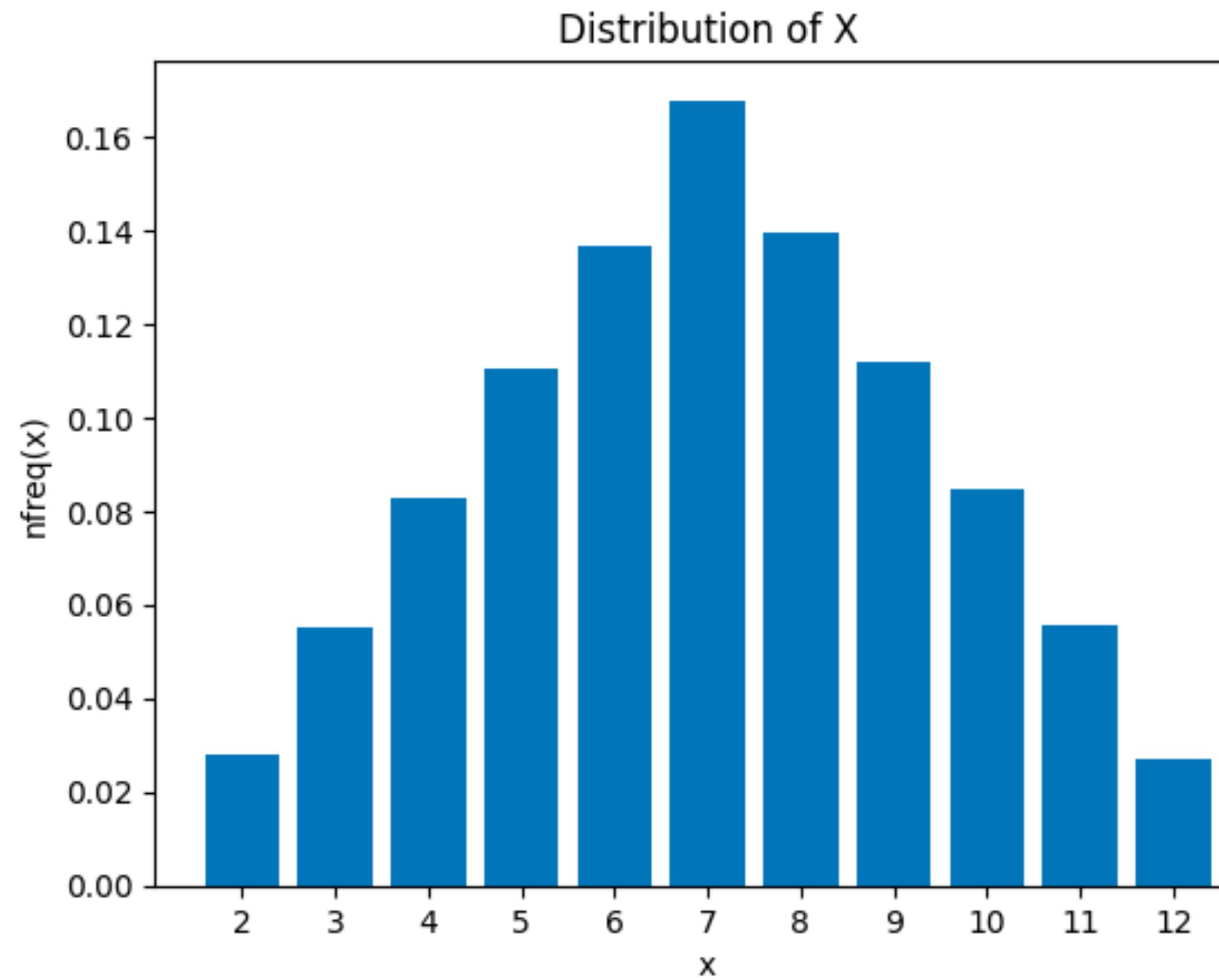
DISCRETE DISTRIBUTIONS AND RANDOM VARIABLES



- For any pmf:
 - $f(x) \geq 0$
 - $\sum_x f(x) = 1$

Note that for an interval: $P(x_2 \leq X \leq x_4)$ $\sum_{x \in \{x_2, x_3, x_4\}} f(x)$

PMF VS FREQUENCIES VIA SIMULATIONS



X : the sum of two fair dice

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
np.random.seed(5)
N=100_000
dice1 = np.random.randint(1,7,size=N)
dice2 = np.random.randint(1,7,size=N)
dicepairs= [(dice1[i], dice2[i]) for i in range(N)]
```

Find # of dice pairs with sum $X = x$.

Normalize & plot.

COMPLETE NOTEBOOK:

<https://colab.research.google.com/drive/1CJps2wxNUDqswSosaGaPb2Km4URcdqho?usp=sharing>

CODING PROBLEMS

Note: Python code in notebooks. Two alternatives:

- Colab: hosted notebooks within Google Drive.
 - No installation
 - <https://colab.research.google.com/>
- Run notebook on your own machine:
 - Install Anaconda
 - Use JupyterLab or Jupyter Notebook

Reading: Python for Data Analysis, Ch 4: NumPy Basics

DISCRETE DISTRIBUTIONS AND RANDOM VARIABLES

Ex: X : # of heads when a fair coin is tossed 10 times

Probability mass function of X :

DISCRETE DISTRIBUTIONS AND RANDOM VARIABLES

Ex: X : # of heads when a fair coin is tossed 10 times

Probability mass function of X :

$$f(x) = \begin{cases} \binom{10}{x} \frac{1}{2^{10}} & \text{if } x=0, 1, \dots, 10 \\ 0 & \text{otherwise} \end{cases}$$

and $\{0, 1, \dots, 10\}$ is the support of X .

(Binomial distribution)

DISCRETE DISTRIBUTIONS AND RANDOM VARIABLES

Ex: Class with B boys and G girls.

n selected randomly without replacement.

X : # of boys in the selection. Probability mass function of X :

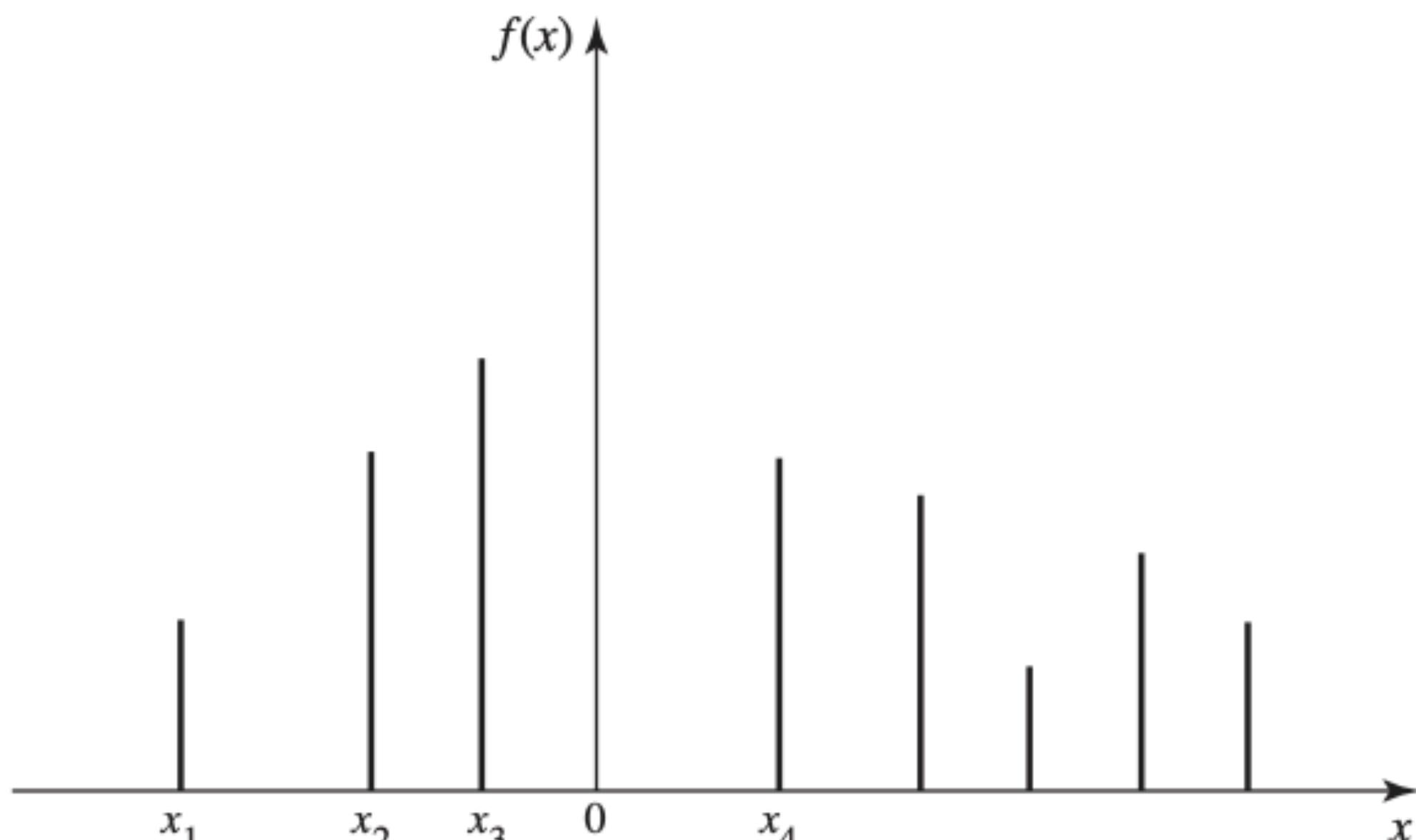
$$f(x) = \begin{cases} \frac{\binom{B}{x} \binom{G}{n-x}}{\binom{B+G}{n}} & \text{for } \max(0, n-G) \leq x \leq \min(n, B) \\ 0 & \text{otherwise} \end{cases}$$

(Hypergeometric distribution)

QUIZ 1

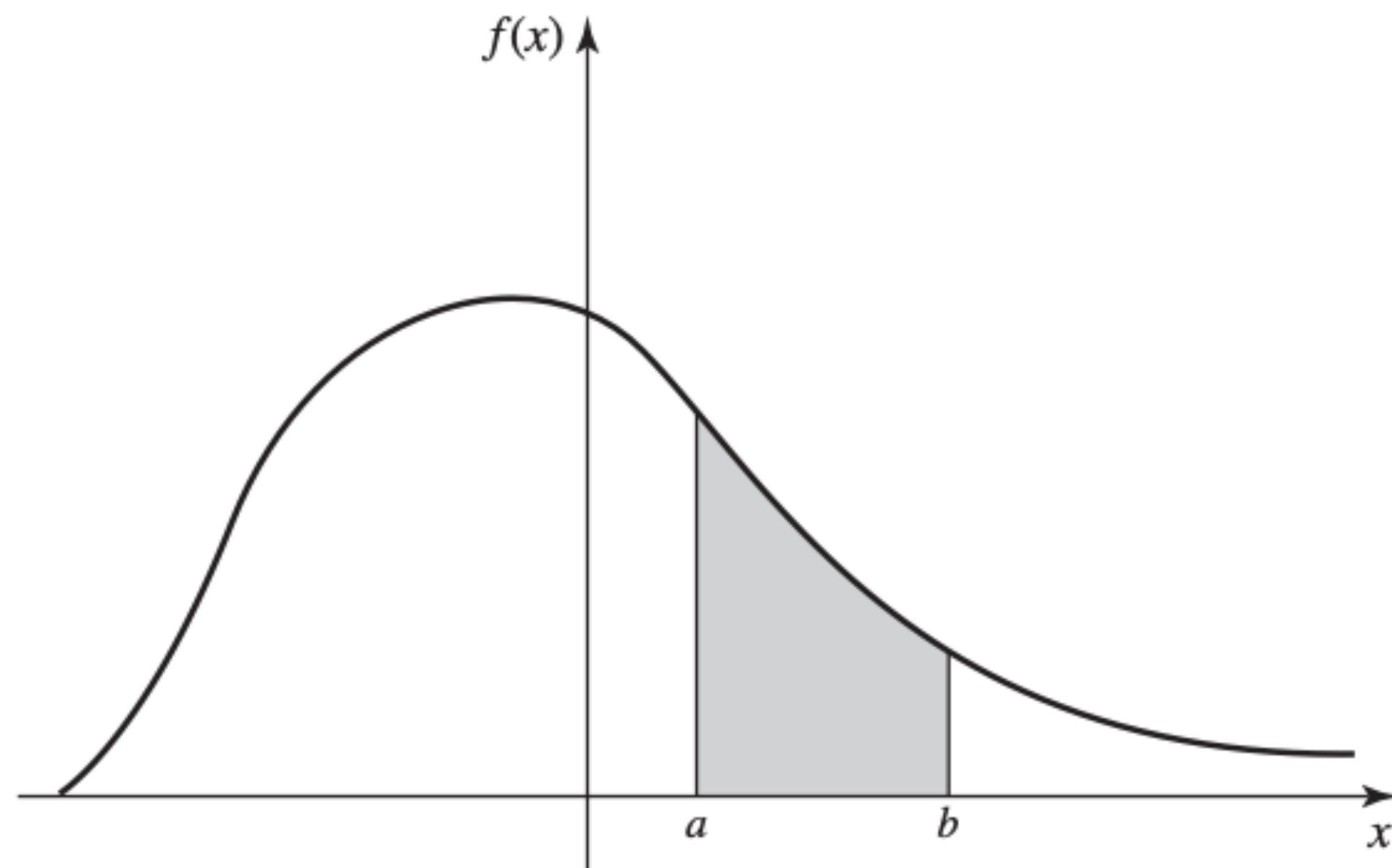
What is the probability of x heads in n tosses of a fair coin?
(x can be any one of $0, 1, \dots, n$)

REVIEW OF (LAST PART OF) LECTURE 3

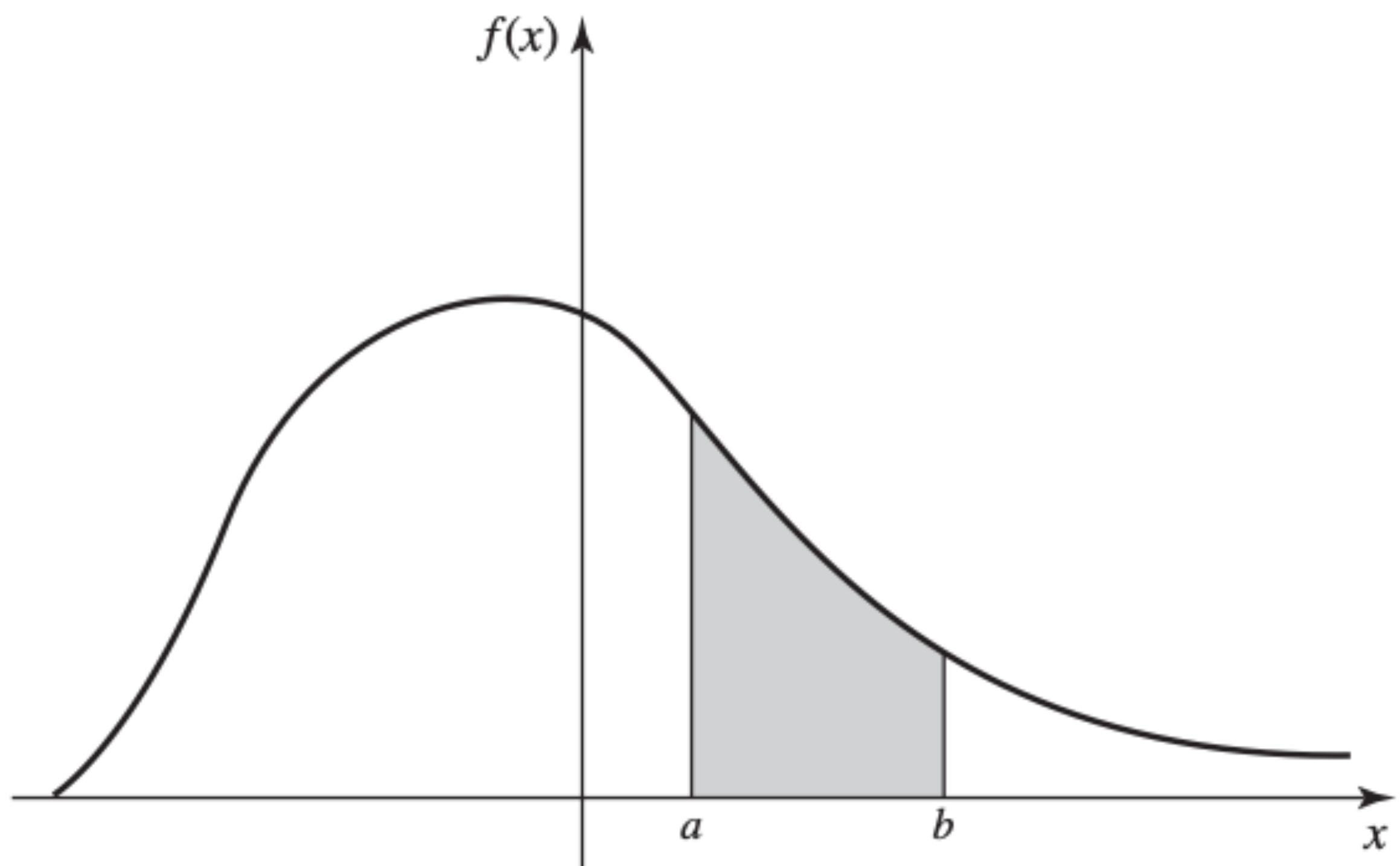


- pmf of a discrete random variable X :
$$f(x) = P(X = x)$$
- For any pmf:
 - $f(x) \geq 0$
 - $\sum_x f(x) = 1$
- For an interval: $P(x_2 \leq X \leq x_4) = \sum_{x \in \{x_2, x_3, x_4\}} f(x)$

CONTINUOUS DISTRIBUTIONS AND RANDOM VARIABLES



CONTINUOUS DISTRIBUTIONS AND RANDOM VARIABLES

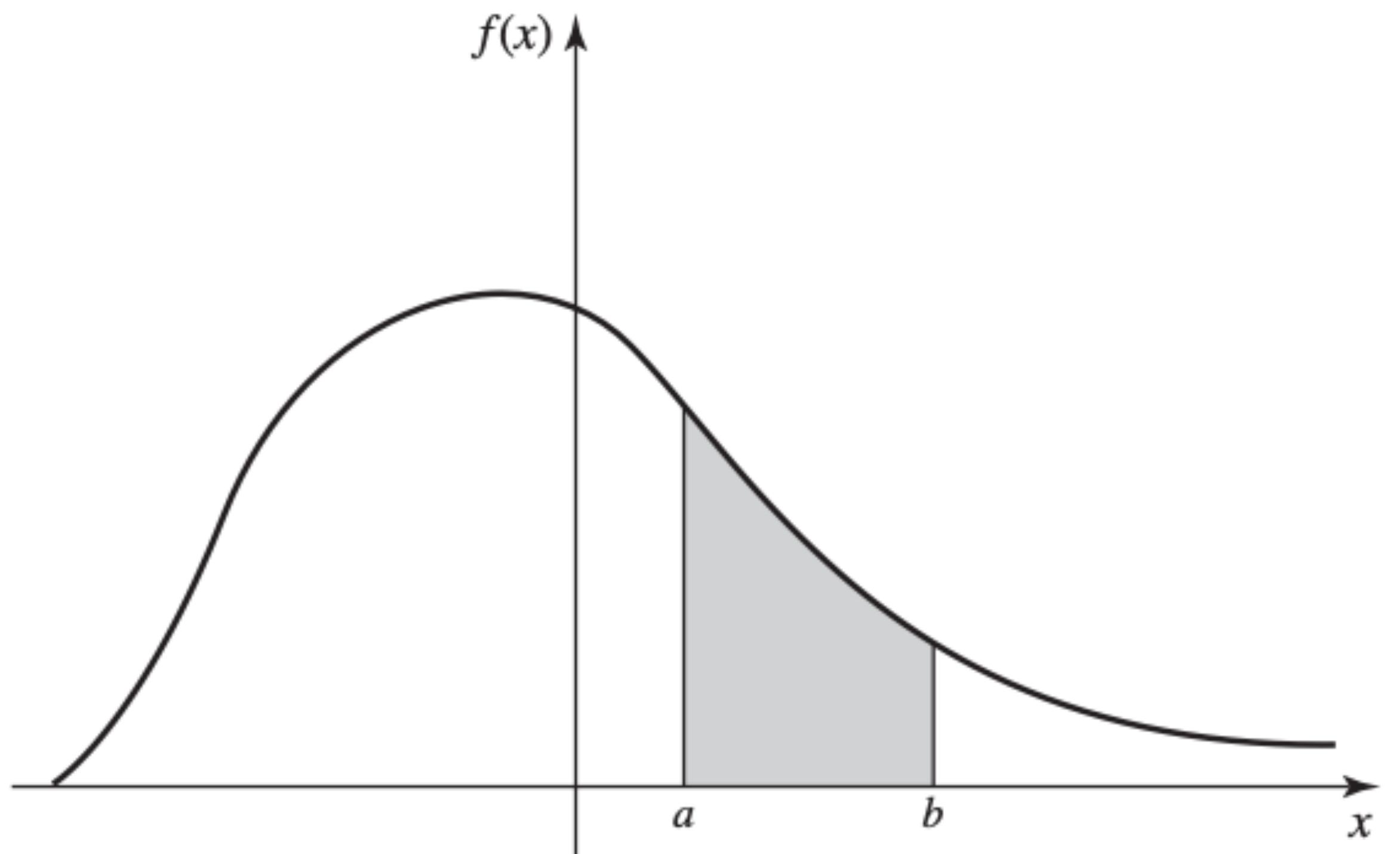


- If $\exists f$ s.t. $\forall [a, b]$ s.t. $a, b \in \mathbb{R}$,

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

$\Rightarrow X$ continuous random variable

CONTINUOUS DISTRIBUTIONS AND RANDOM VARIABLES



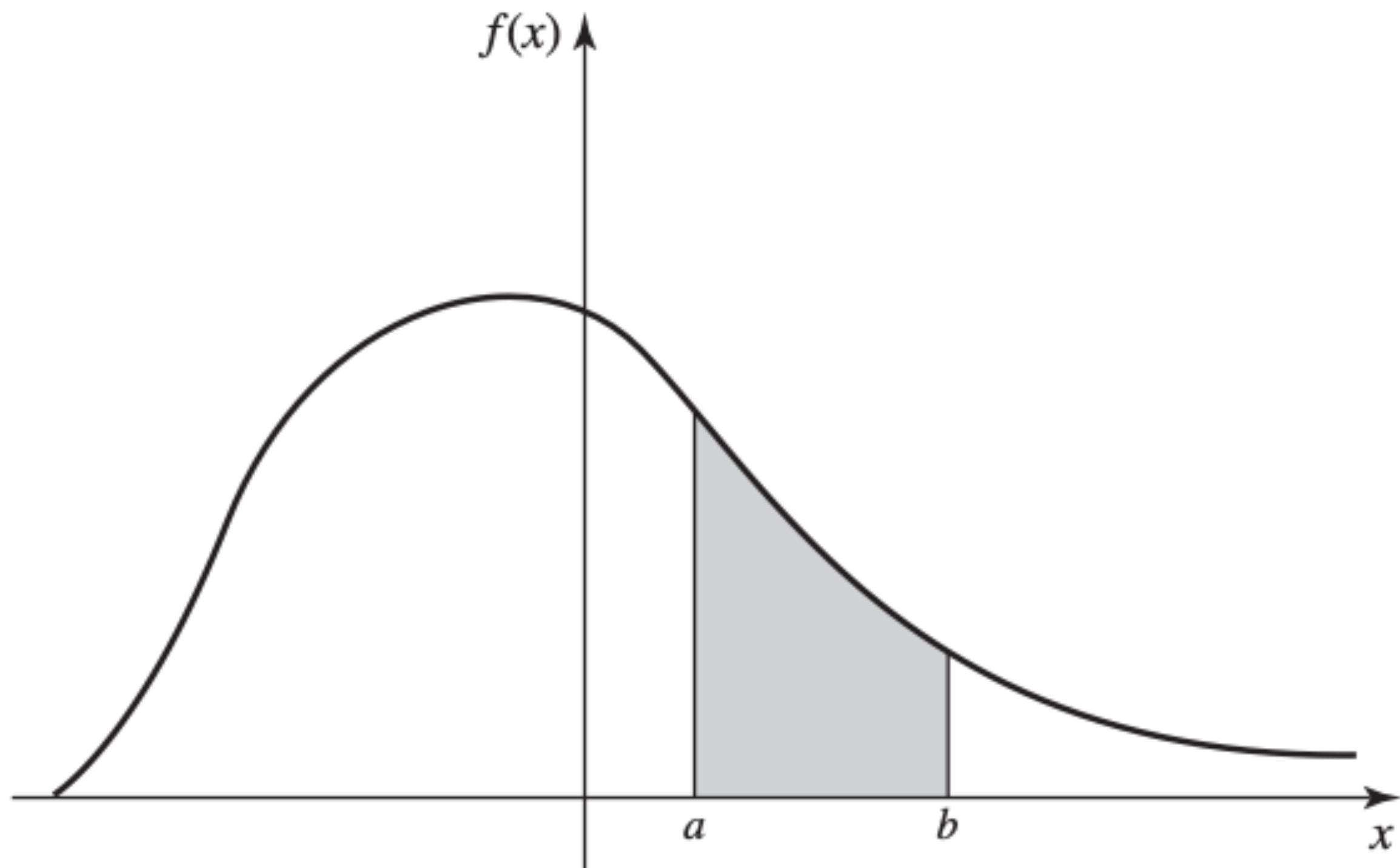
- If $\exists f$ s.t. $\forall [a, b]$ s.t. $a, b \in \mathbb{R}$,

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

$\Rightarrow X$ continuous random variable

Reminder: $P(a \leq X \leq b) = \sum_{x=a}^b f(x)$ for discrete RV X .

CONTINUOUS DISTRIBUTIONS AND RANDOM VARIABLES



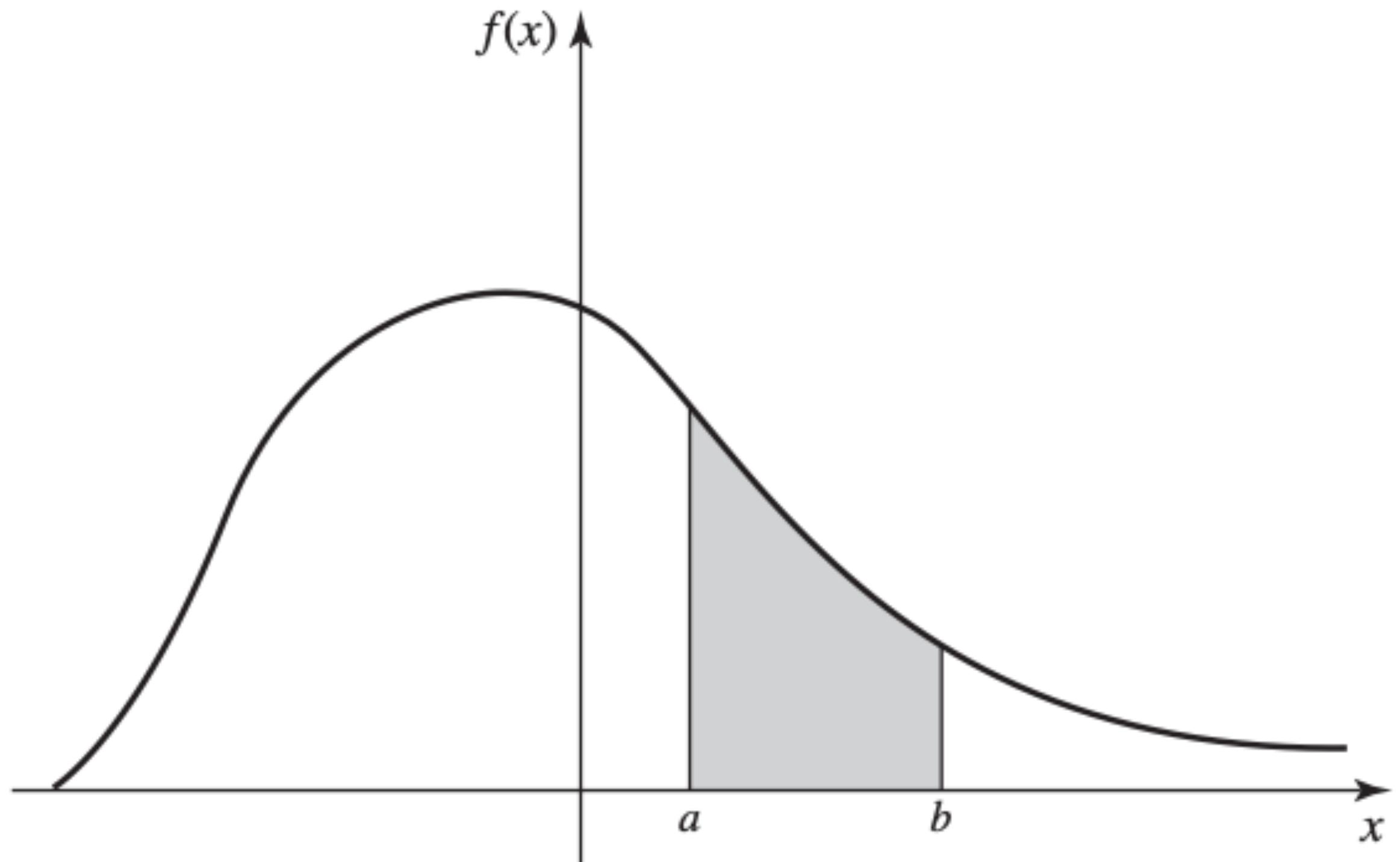
- If $\exists f$ s.t. $\forall [a, b]$ s.t. $a, b \in \mathbb{R}$,

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

$\Rightarrow X$ continuous random variable

- $f(x)$ is the probability density function (pdf) of X .
- $\{x : f(x) > 0\}$ is the support of X .

CONTINUOUS DISTRIBUTIONS AND RANDOM VARIABLES

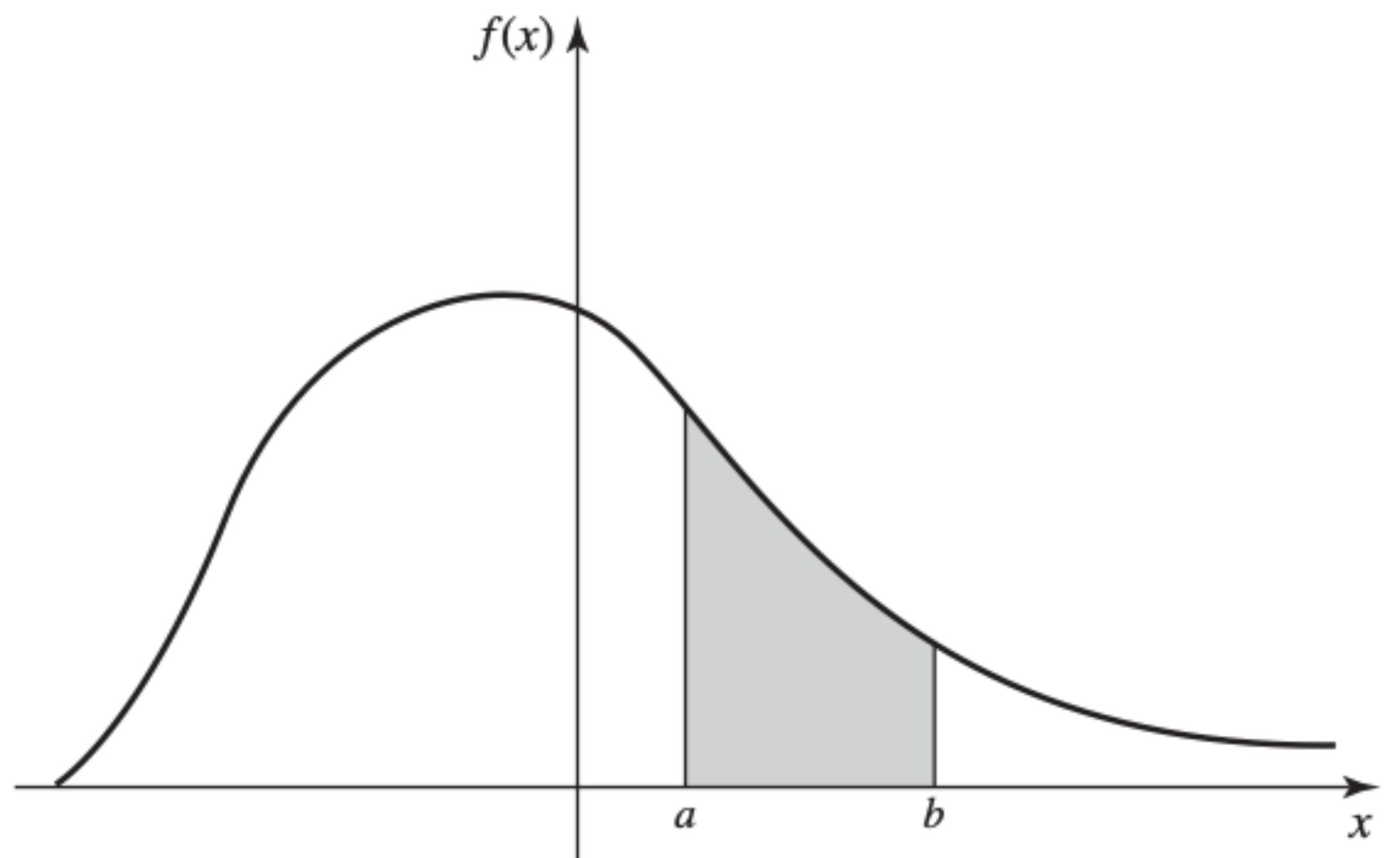


- If $\exists f$ s.t. $\forall [a, b]$ s.t. $a, b \in \mathbb{R}$,

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

Note: $P(x = a) = 0, \forall a.$

CONTINUOUS DISTRIBUTIONS AND RANDOM VARIABLES



- For any pdf:
 - $f(x) \geq 0, \quad \forall x$
 - $\int_{-\infty}^{\infty} f(x) \, dx = 1$

CONTINUOUS DISTRIBUTIONS AND RANDOM VARIABLES

Ex: $a \leq X \leq b$ and \forall subinterval, probability X in it is proportional to the length of subinterval

$\Rightarrow X$ has uniform distribution on $[a, b]$. What is the pdf of X ?

CONTINUOUS DISTRIBUTIONS AND RANDOM VARIABLES

Ex: $a \leq X \leq b$ and \forall subinterval, probability X in it is proportional to the length of subinterval

$\Rightarrow X$ has uniform distribution on $[a, b]$. What is the pdf of X ?

Probability that X is in any interval with same length is the same

$\Rightarrow f(x)$ is some constant c throughout $[a, b]$. Furthermore,

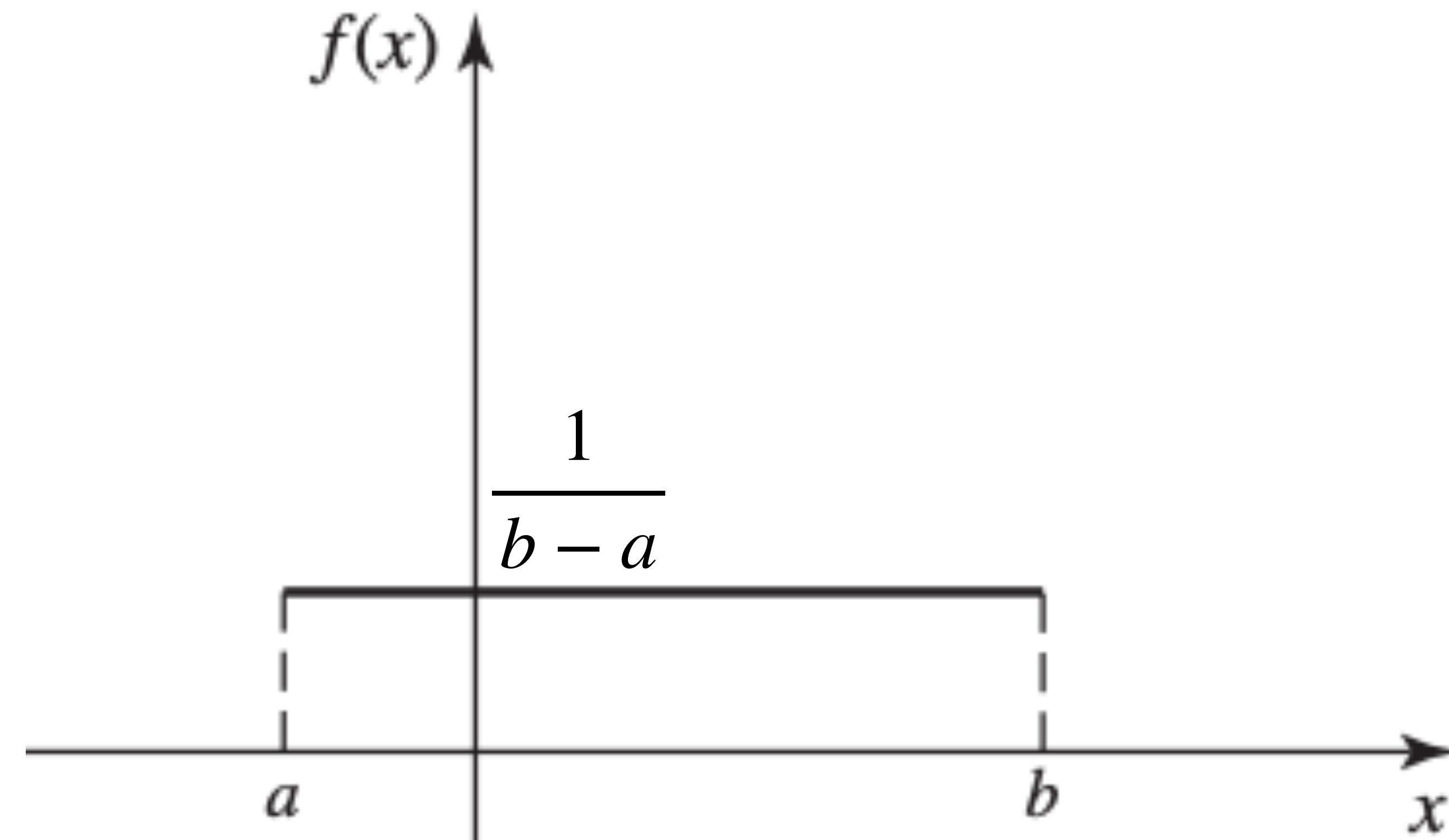
$$\int_{-\infty}^{\infty} f(x) dx = \int_a^b c dx = 1 \Rightarrow f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

CONTINUOUS DISTRIBUTIONS AND RANDOM VARIABLES

Ex: $a \leq X \leq b$ and \forall subinterval, probability X in it is proportional to the length of subinterval

$\Rightarrow X$ has uniform distribution on $[a, b]$. What is the pdf of X ?

That is,



CONTINUOUS DISTRIBUTIONS AND RANDOM VARIABLES

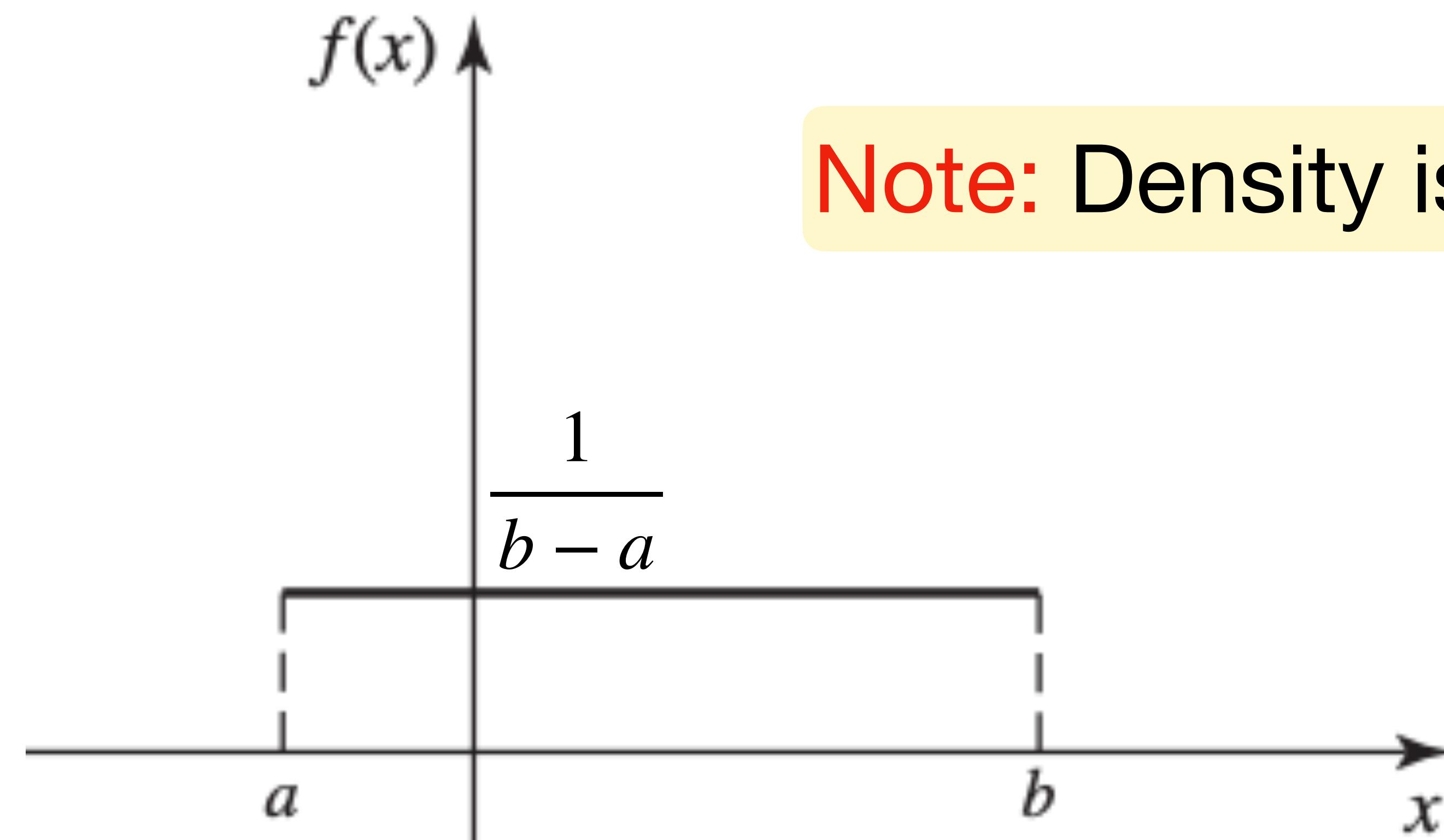
Ex: $a \leq X \leq b$ and \forall subinterval, probability X in it is proportional to the length of subinterval

$\Rightarrow X$ has uniform distribution on $[a, b]$. What is the pdf of X ?

$$f(x)$$

Note: Density is not probability.

That is,



CUMULATIVE DISTRIBUTION FUNCTION

- $F(x) = P(X \leq x)$ for $-\infty < x < \infty$

is the cumulative distribution function (cdf) of X .

CUMULATIVE DISTRIBUTION FUNCTION

- $F(x) = P(X \leq x)$ for $-\infty < x < \infty$

is the cumulative distribution function (cdf) of X .

- For continuous distributions this becomes

$$F(x) = \int_{-\infty}^x f(t) dt \quad \text{and} \quad \frac{dF(x)}{dx} = f(x)$$

CUMULATIVE DISTRIBUTION FUNCTION

- $F(x) = P(X \leq x)$ for $-\infty < x < \infty$

is the cumulative distribution function (cdf) of X .

- For continuous distributions this becomes

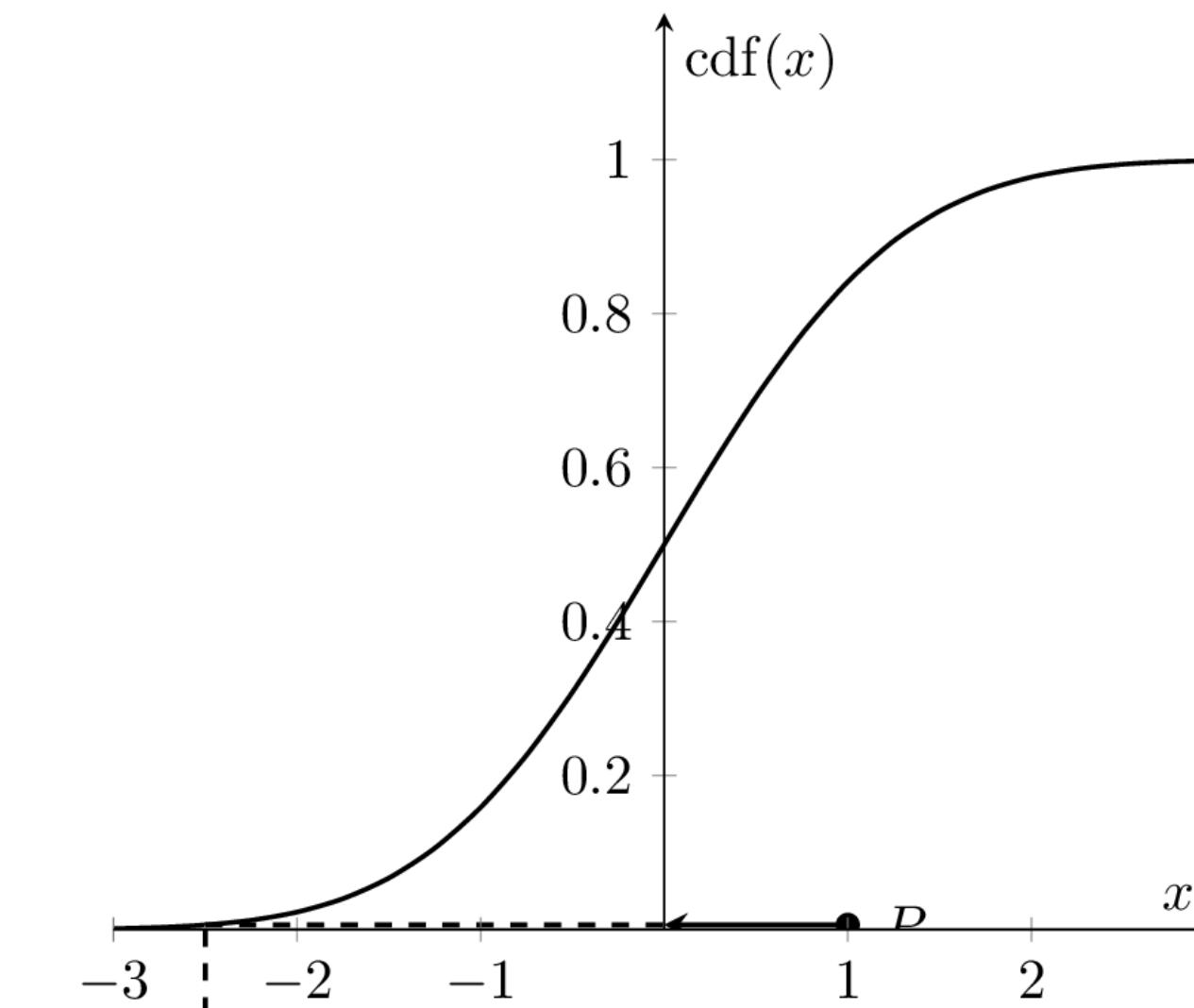
$$F(x) = \int_{-\infty}^x f(t) dt \quad \text{and} \quad \frac{dF(x)}{dx} = f(x)$$

- $P(X > x) = 1 - F(x)$ and $P(a < X \leq b) = F(b) - F(a)$

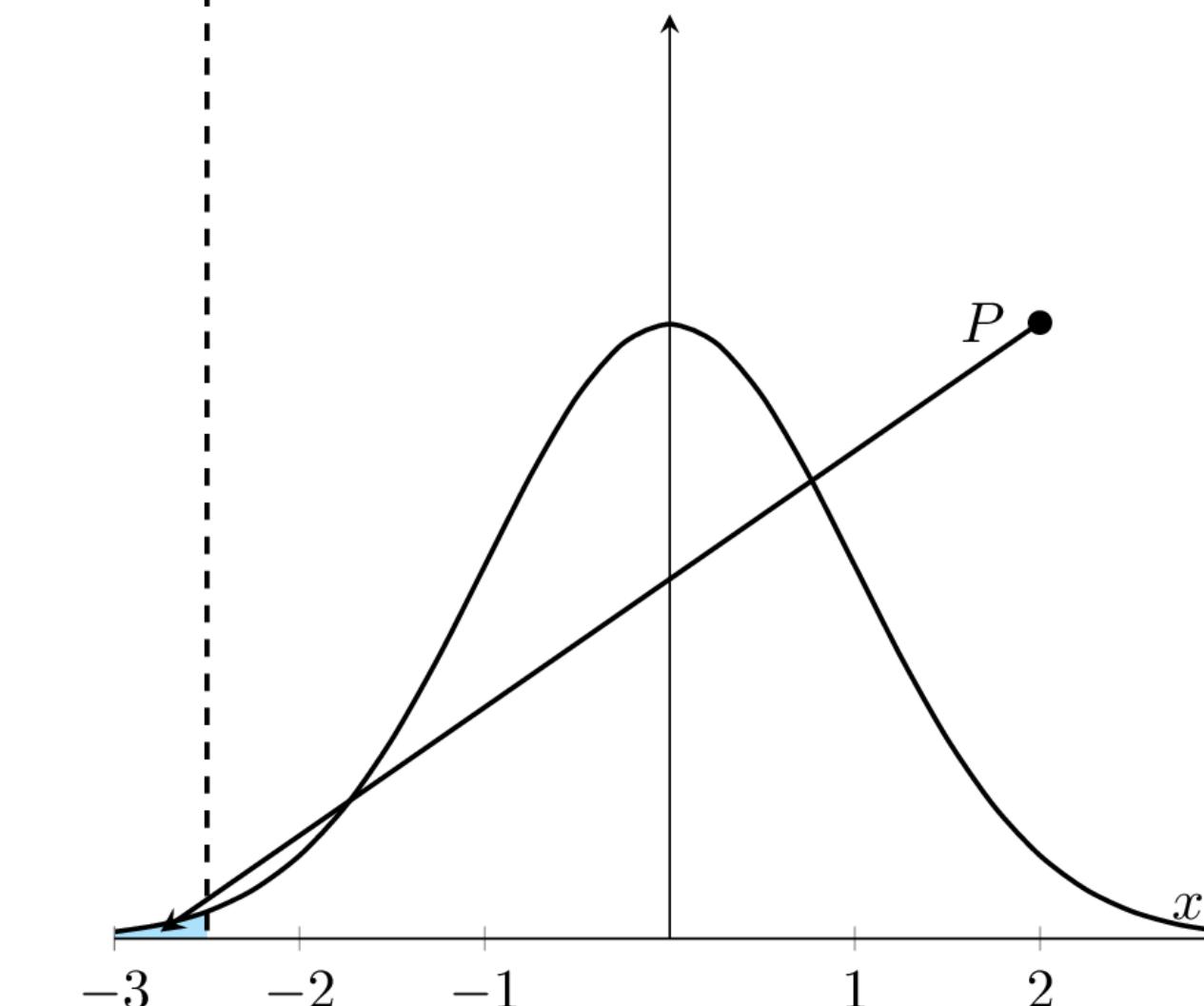
CUMULATIVE DISTRIBUTION FUNCTION

- Relationship between cdf and pdf

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$$



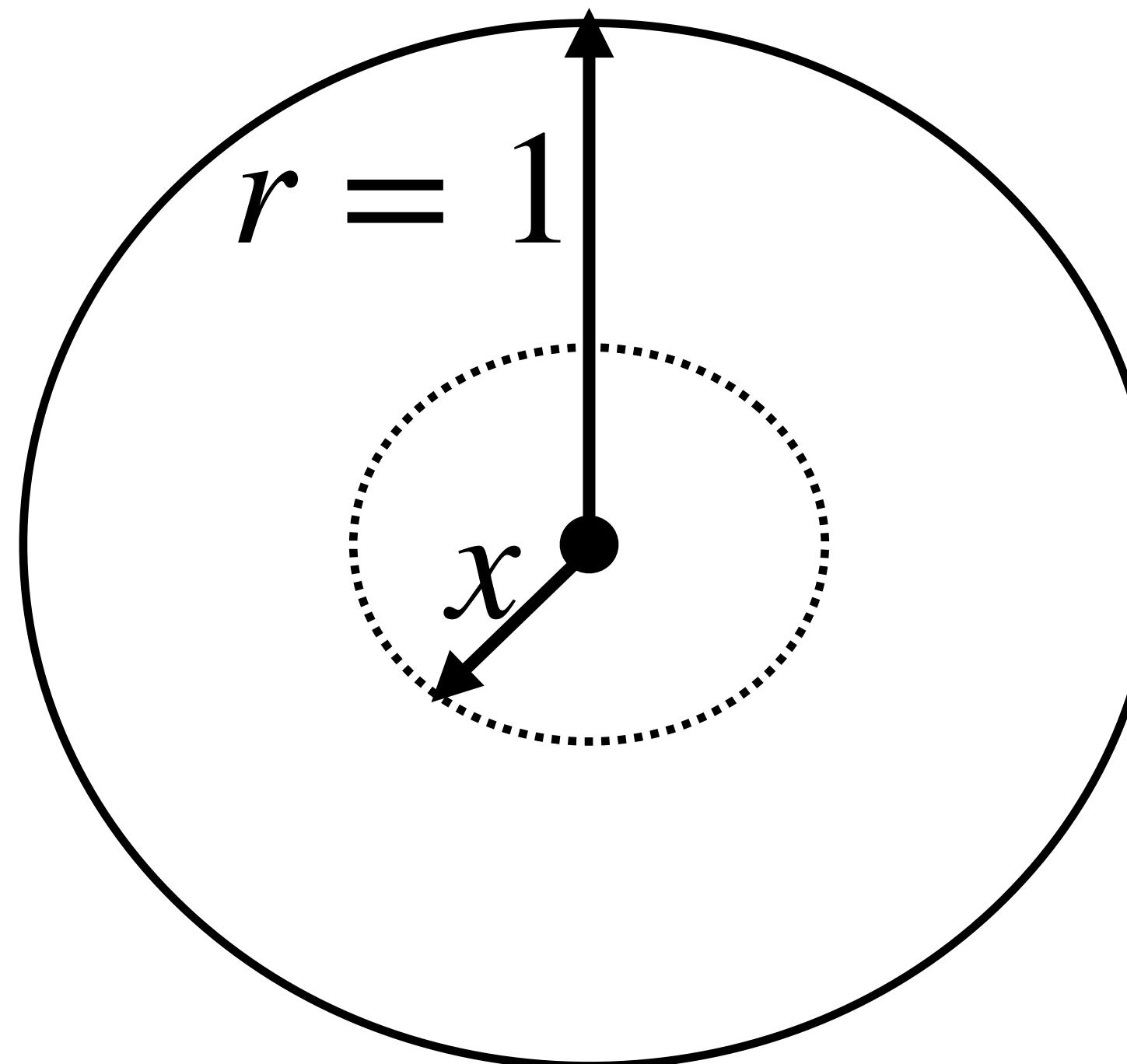
cdf $F(x)$



pdf $f(x)$

CUMULATIVE DISTRIBUTION FUNCTION

Ex: Dart with radius 1. Lands randomly uniformly on the board.

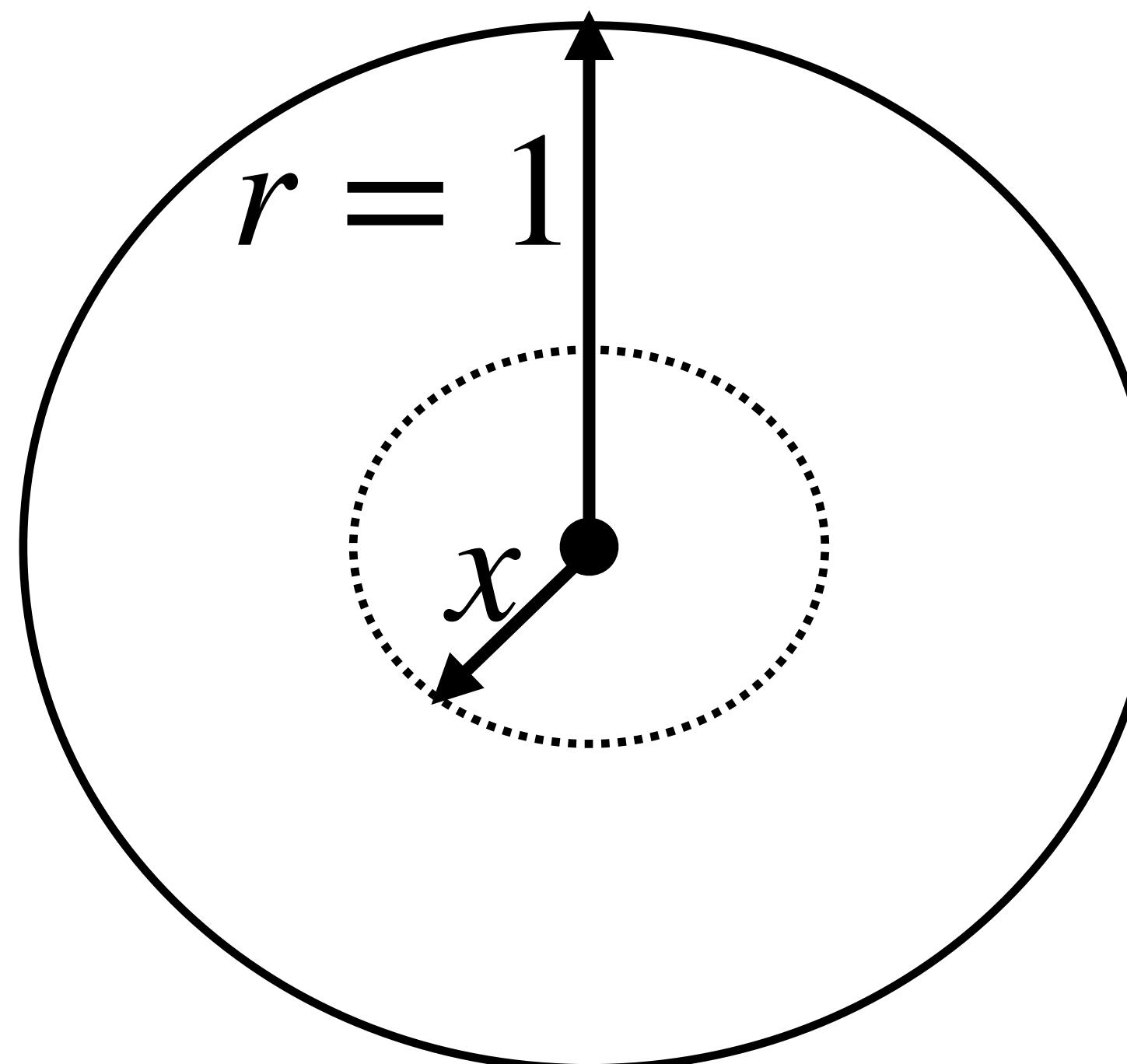


X : distance from the center

What is the cdf of X , that is $F(x)$?

CUMULATIVE DISTRIBUTION FUNCTION

Ex: Dart with radius 1. Lands randomly uniformly on the board.



X : distance from the center

What is the cdf of X , that is $F(x)$?

$$P(X \leq x) = \frac{\pi x^2}{\pi 1^2} = x^2 \quad \Rightarrow$$

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x^2 & \text{if } 0 < x \leq 1 \\ 1 & \text{if } 1 < x \end{cases}$$

CUMULATIVE DISTRIBUTION FUNCTION

Ex: If X has uniform distribution on $[a, b]$, we found the pdf:

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

What is the cdf of X ?

CUMULATIVE DISTRIBUTION FUNCTION

Ex: If X has uniform distribution on $[a, b]$, we found the pdf:

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

What is the cdf of X ?

$$F(x) = \begin{cases} 0 & \text{if } x \leq a \\ \int_a^x \frac{1}{b-a} dt & \text{if } a < x \leq b \\ 1 & \text{if } x > b \end{cases} \Rightarrow F(x) = \frac{x-a}{b-a}, \quad \forall a < x \leq b$$

JOINT DISTRIBUTIONS: DISCRETE

- The joint pmf of discrete random variables X, Y :

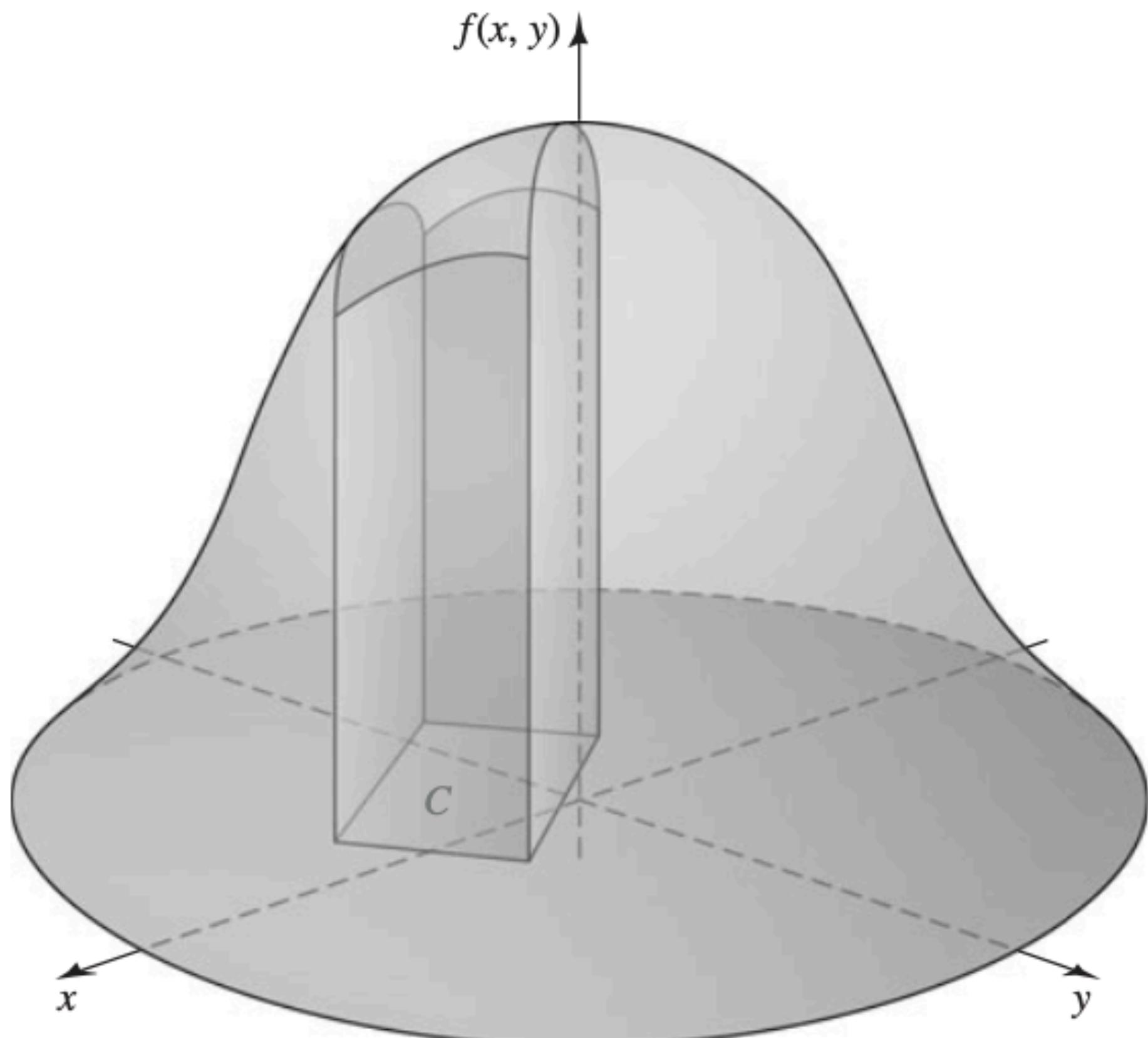
$$f(x, y) = P(X = x \text{ and } Y = y)$$

Note that $\sum_{\forall(x,y)} f(x, y) = 1.$

JOINT DISTRIBUTIONS: CONTINUOUS

- A joint pdf satisfies:

$$f(x, y) \geq 0 \text{ AND } \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$$



$$P((X, Y) \in C) = \iint_C f(x, y) dx dy$$

JOINT DISTRIBUTIONS: DISCRETE

Ex: X : # of cars owned by a randomly selected household

Y : # of computers owned by the same household

Joint pmf shown with a table. For our example:

x	y			
	1	2	3	4
1	0.1	0	0.1	0
2	0.3	0	0.1	0.2
3	0	0.2	0	0

JOINT DISTRIBUTIONS: DISCRETE

Ex: X : # of cars owned by a randomly selected household

Y : # of computers owned by the same household

Joint pmf shown with a table. For our example:

x	y			
	1	2	3	4
1	0.1	0	0.1	0
2	0.3	0	0.1	0.2
3	0	0.2	0	0

Probability that a randomly selected household has ≥ 2 cars and computers?

JOINT DISTRIBUTIONS: DISCRETE

Ex: X : # of cars owned by a randomly selected household

Y : # of computers owned by the same household

Joint pmf shown with a table. For our example:

x	y			
	1	2	3	4
1	0.1	0	0.1	0
2	0.3	0	0.1	0.2
3	0	0.2	0	0

Probability that a randomly selected household has ≥ 2 cars and computers?

Sum of values gives $P(X \geq 2 \text{ and } Y \geq 2)$
= 0.5

MARGINAL DISTRIBUTIONS: DISCRETE AND CONTINUOUS

- Given joint distribution of X, Y need distribution of one, say X . Such a distribution is **marginal distribution of X** .
- f : joint pmf of X, Y . Marginal pmf of X :

$$f_1(x) = \sum_{\forall y} f(x, y)$$

MARGINAL DISTRIBUTIONS: DISCRETE AND CONTINUOUS

- Given joint distribution of X, Y need distribution of one, say X . Such a distribution is **marginal distribution of X** .

f : joint pmf of X, Y . Marginal pmf of X :

$$f_1(x) = \sum_{\forall y} f(x, y)$$

- For the continuous case:

f : joint pdf of X, Y . Marginal pdf of X :

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

MARGINAL DISTRIBUTIONS: DISCRETE

Ex: X : # of cars owned by a randomly selected household

Y : # of computers owned by the same household

Joint pmf shown with the table.

x	y			
	1	2	3	4
1	0.1	0	0.1	0
2	0.3	0	0.1	0.2
3	0	0.2	0	0

MARGINAL DISTRIBUTIONS: DISCRETE

Ex: X : # of cars owned by a randomly selected household

Y : # of computers owned by the same household

Joint pmf shown with the table.

	y				Total
x	1	2	3	4	
1	0.1	0	0.1	0	0.2
2	0.3	0	0.1	0.2	0.6
3	0	0.2	0	0	0.2
Total	0.4	0.2	0.2	0.2	1.0

Marginal pmf of X : $f_1(X)$

Marginal pmf of Y : $f_2(Y)$

REVIEW OF (LAST PART OF) LECTURE 4

- Joint pmf of X, Y : $f(x, y) = P(X = x \text{ and } Y = y)$
- Joint pdf of X, Y : $f(x, y) \geq 0$ AND $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$
$$P((X, Y) \in C) = \int_C \int f(x, y) dx dy$$
- Marginal pmf of discrete X : $f_1(x) = \sum_{\forall y} f(x, y)$
- Marginal pdf of continuous X : $f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy$

INDEPENDENT RANDOM VARIABLES

- X, Y are independent (denoted with $X \perp\!\!\!\perp Y$) if and only if

$$f(x, y) = f_1(x)f_2(y) \quad \forall x, y$$

Ex: Which of these corresponds to independent X, Y ?

	$Y = 0$	$Y = 1$	
$X=0$	1/4	1/4	1/2
$X=1$	1/4	1/4	1/2
	1/2	1/2	1

	$Y = 0$	$Y = 1$	
$X=0$	1/2	0	1/2
$X=1$	0	1/2	1/2
	1/2	1/2	1

INDEPENDENT RANDOM VARIABLES

- X, Y are independent (denoted with $X \perp\!\!\!\perp Y$) if and only if

$$f(x, y) = f_1(x)f_2(y) \quad \forall x, y$$

Ex: Which of these corresponds to independent X, Y ?

	$Y = 0$	$Y = 1$	
$X=0$	1/4	1/4	1/2
$X=1$	1/4	1/4	1/2
	1/2	1/2	1

X, Y independent.

	$Y = 0$	$Y = 1$	
$X=0$	1/2	0	1/2
$X=1$	0	1/2	1/2
	1/2	1/2	1

X, Y not independent.

$$f_1(0)f_2(1) = \frac{1}{4} \text{ whereas } f(0,1) = 0$$

CONDITIONAL DISTRIBUTIONS

- **Discrete:** X, Y have joint pmf f . Y has marginal pmf f_2 .

Conditional pmf of X given Y :
$$g_1(x | y) = \frac{f(x, y)}{f_2(y)}$$

Note: $P(X = x | Y = y) = g_1(x | y)$

CONDITIONAL DISTRIBUTIONS

- **Discrete:** X, Y have joint pmf f . Y has marginal pmf f_2 .

Conditional pmf of X given Y :
$$g_1(x | y) = \frac{f(x, y)}{f_2(y)}$$

Note: $P(X = x | Y = y) = g_1(x | y)$

- **Continuous:** X, Y have joint pdf f . Y has marginal pdf f_2 .

Conditional pdf of X given Y :
$$g_1(x | y) = \frac{f(x, y)}{f_2(y)}$$

Note: $P(X \in C | Y = y) = \int_C g_1(x | y) dx$

DISCRETE CONDITIONAL DISTRIBUTIONS

Ex: $X = 0$: Car not stolen, $X = 1$: Car stolen

The table corresponding to joint pmf of X, Y :

Stolen X	Brand Y					Total
	1	2	3	4	5	
0	0.129	0.298	0.161	0.280	0.108	0.976
1	0.010	0.010	0.001	0.002	0.001	0.024
Total	0.139	0.308	0.162	0.282	0.109	1.000

Give the table corresponding to conditional pmf of X given Y .

DISCRETE CONDITIONAL DISTRIBUTIONS

Ex: $X = 0$: Car not stolen, $X = 1$: Car stolen

The table corresponding to joint pmf of X, Y :

Stolen X	Brand Y					Total
	1	2	3	4	5	
0	0.129	0.298	0.161	0.280	0.108	0.976
1	0.010	0.010	0.001	0.002	0.001	0.024
Total	0.139	0.308	0.162	0.282	0.109	1.000

Give the table corresponding to conditional pmf of X given Y .

Stolen X	Brand Y				
	1	2	3	4	5
0	0.928	0.968	0.994	0.993	0.991
1	0.072	0.032	0.006	0.007	0.009

CONTINUOUS CONDITIONAL DISTRIBUTIONS

Ex: Y : time required to do necessary reading

X : time required to do the homework (includes Y)

$$f(x, y) = \begin{cases} e^{-x} & \text{for } 0 \leq y \leq x < \infty \\ 0 & \text{otherwise} \end{cases}$$

Useful integral:

$$\int_a^{\infty} e^{-x} dx = e^{-a}$$

Find $P(X \geq 10 | Y = 4)$?

CONTINUOUS CONDITIONAL DISTRIBUTIONS

Ex: Y : time required to do necessary reading

X : time required to do the homework (includes Y)

Useful integral:
$$\int_a^{\infty} e^{-x} dx = e^{-a}$$

$$f(x, y) = \begin{cases} e^{-x} & \text{for } 0 \leq y \leq x < \infty \\ 0 & \text{otherwise} \end{cases}$$

Find $P(X \geq 10 | Y = 4)$?

marginal pmf of Y : $f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_y^{\infty} e^{-x} dx = e^{-y}$

$$g_1(x | y) = \frac{f(x, y)}{f_2(y)} = \frac{e^{-x}}{e^{-y}} = e^{y-x} \Rightarrow P(X \geq 10 | Y = 4) = \int_{10}^{\infty} e^{4-x} dx = e^{-6}$$

GENERALIZING REST OF THE RULES ON EVENTS

- Multiplication Rule: $f(x, y) = g_1(x | y)f_2(y)$
- Law of Total Probability:

$$f_1(x) = \sum g_1(x | y)f_2(y) \quad \text{for discrete } Y$$

$$f_1(x) = \int_{-\infty}^y g_1(x | y)f_2(y) dy \quad \text{for continuous } Y$$

- Bayes' Theorem :

$$g_1(x | y) = \frac{g_2(y | x)f_1(x)}{f_2(y)}$$

EXPECTATION

- Readings for the next few lectures:

Ch 8 (Watkins)

Ch 3 (Wasserman, for more formal treatment)

EXPECTATION

- Mean \equiv expected value \equiv expectation
- Weighted average of values of a random variable where weights are probabilities.

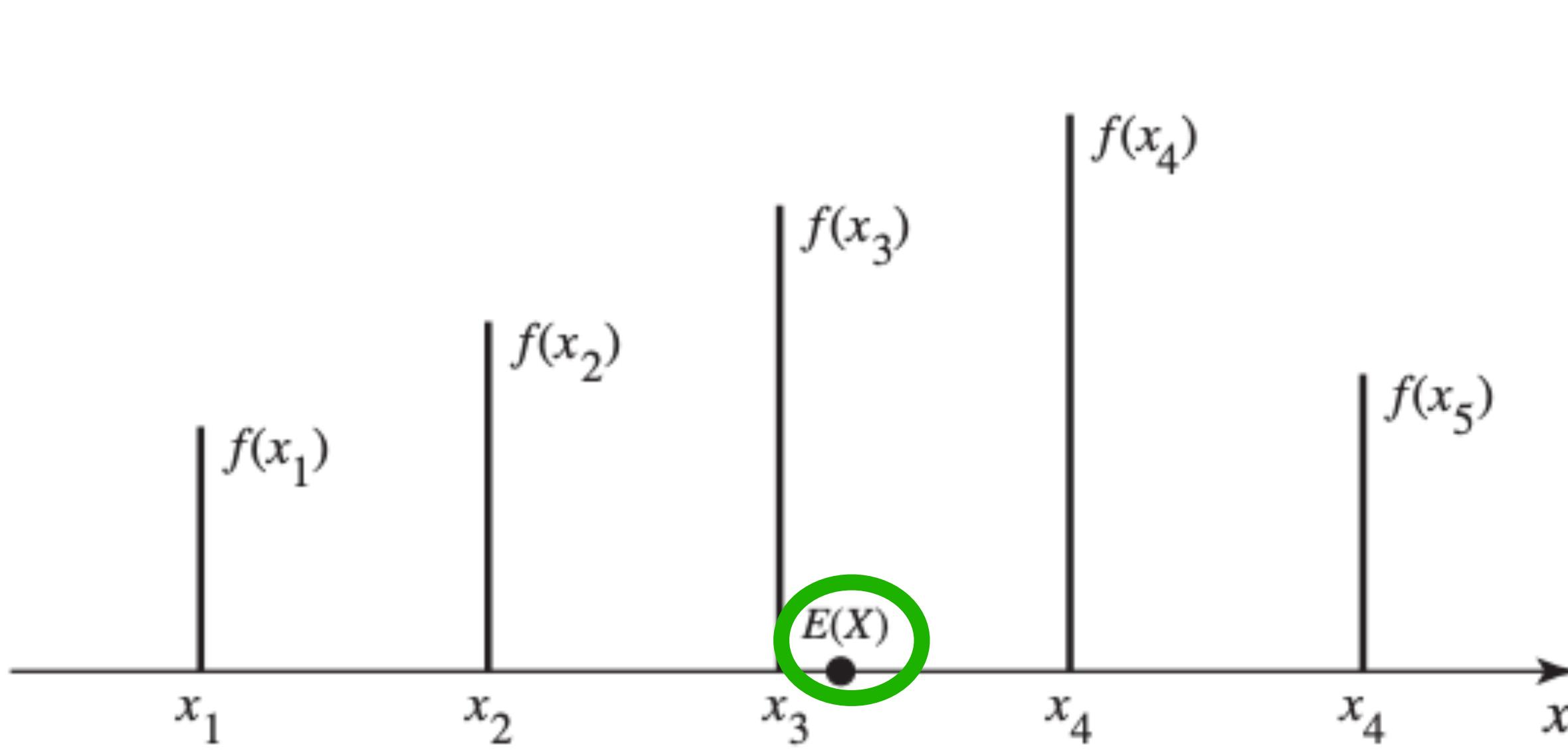
$$E(X) = \sum_{\forall x} xf(x) \quad \text{for discrete } X$$

$$E(X) = \int xf(x) dx \quad \text{for continuous } X$$

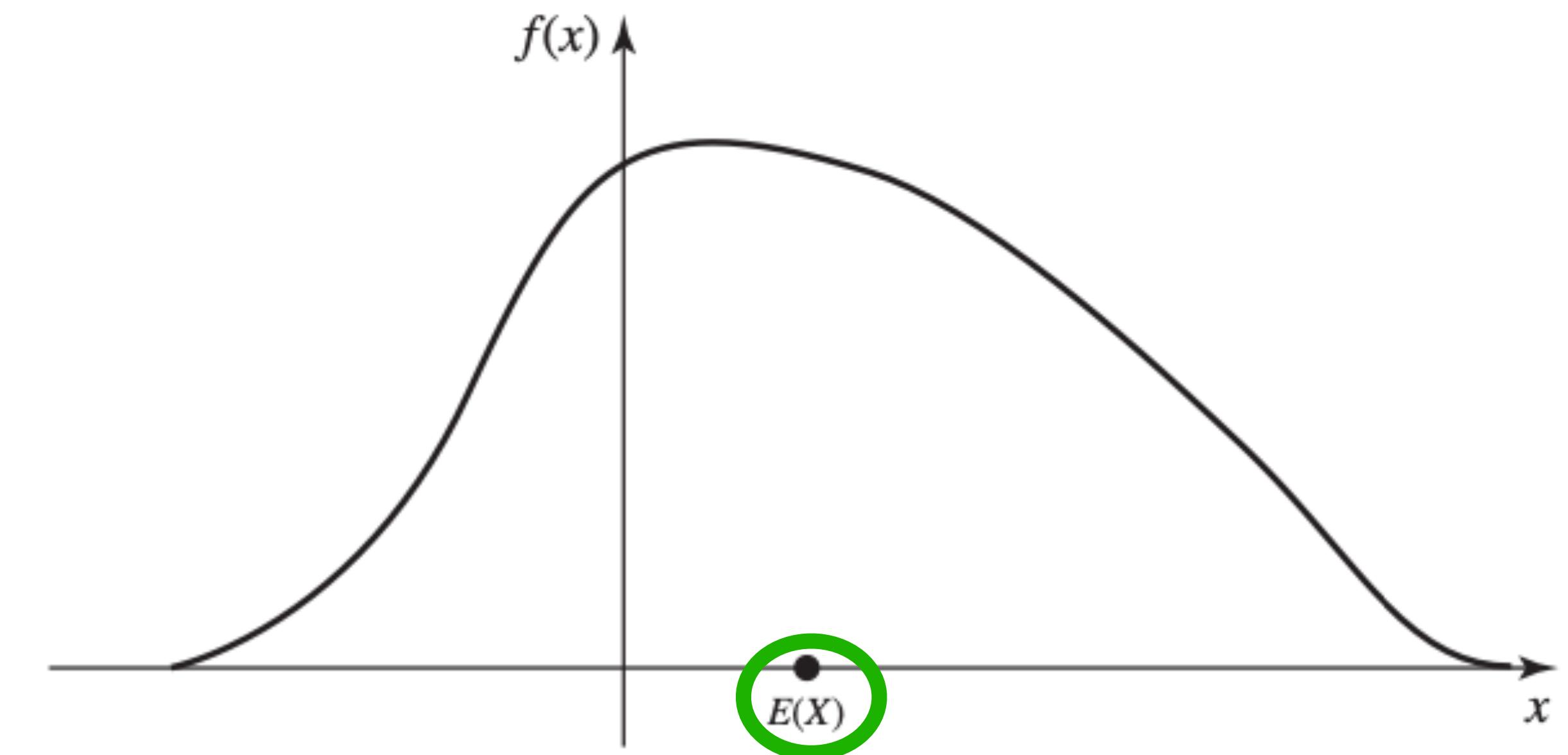
- Also denoted with μ .

EXPECTATION

- Expectation as the center of gravity



Discrete

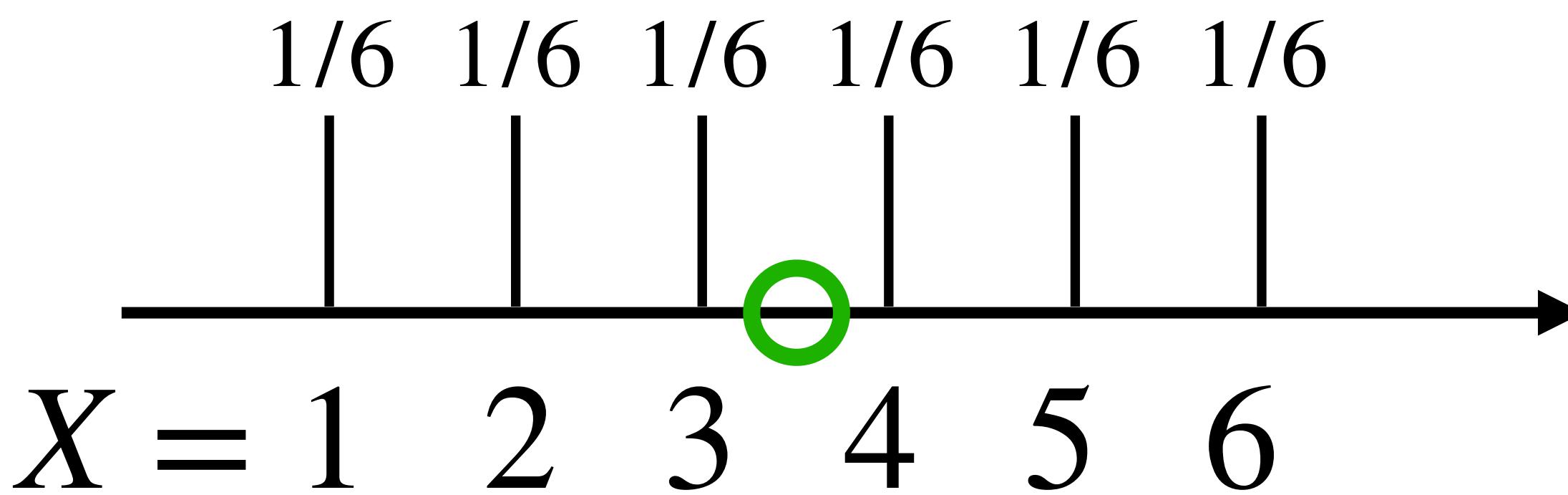


Continuous

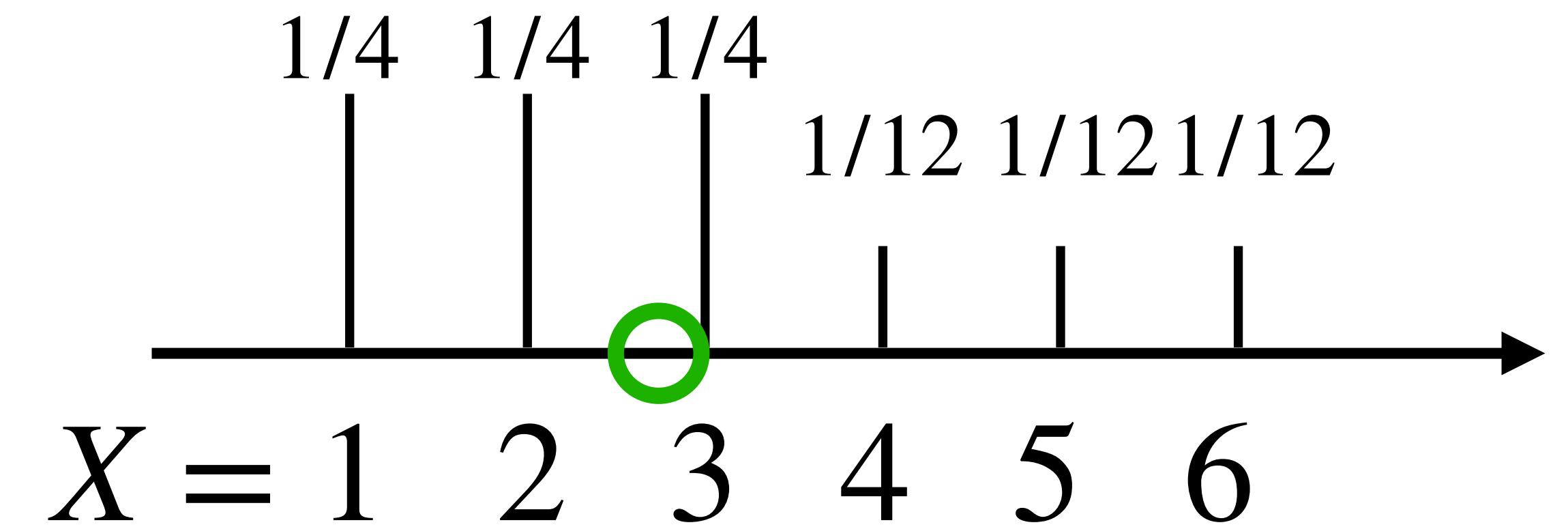
EXPECTATION

Ex: Expectation of fair die vs unfair die. X : value on die.

Fair die



Unfair die



$$E(X) = 1 \times 1/6 + 2 \times 1/6 + \dots + 6 \times 1/6 = 7/2$$

$$E(X) = 1 \times 1/4 + 2 \times 1/4 + \dots + 6 \times 1/12 = 11/4$$

EXPECTATION

Ex: Expectation of Bernoulli random variable X

X has Bernoulli distribution with parameter $p \Rightarrow$

$$P(X = 1) = p, P(X = 0) = 1 - p$$

EXPECTATION

Ex: Expectation of Bernoulli random variable X

X has Bernoulli distribution with parameter $p \Rightarrow$

$$P(X = 1) = p, P(X = 0) = 1 - p$$

- Throwing a coin with probability of Head p , Tail $1 - p$

EXPECTATION

Ex: Expectation of Bernoulli random variable X

X has Bernoulli distribution with parameter $p \Rightarrow$

$$P(X = 1) = p, P(X = 0) = 1 - p$$

- Throwing a coin with probability of Head p , Tail $1 - p$
- Sampling with replacement: **R**, **B** balls in a box.
Proportion of **R** is p . $X_i = 1$, if i^{th} ball is **R**, 0 otherwise.

EXPECTATION

Ex: Expectation of Bernoulli random variable X

X has Bernoulli distribution with parameter $p \Rightarrow$

$$P(X = 1) = p, P(X = 0) = 1 - p$$

- Throwing a coin with probability of Head p , Tail $1 - p$
- Sampling with replacement: **R**, **B** balls in a box.
Proportion of **R** is p . $X_i = 1$, if i^{th} ball is **R**, 0 otherwise.

What is $E(X)$?

EXPECTATION

Ex: Expectation of Bernoulli random variable X

X has Bernoulli distribution with parameter $p \Rightarrow$

$$P(X = 1) = p, P(X = 0) = 1 - p$$

- Throwing a coin with probability of Head p , Tail $1 - p$
- Sampling with replacement: **R**, **B** balls in a box.
Proportion of **R** is p . $X_i = 1$, if i^{th} ball is **R**, 0 otherwise.

What is $E(X)$?

$$E(X) = \sum_{x=\{0,1\}} xf(x) = 0 \times (1 - p) + 1 \times p = p$$

EXPECTATION

Ex: X : Time until a lightbulb fails. Its pdf $f(x) = 2x$, $0 < x < 1$.

What is $E(X)$?

EXPECTATION

Ex: X : Time until a lightbulb fails. Its pdf $f(x) = 2x$, $0 < x < 1$.

What is $E(X)$?

$$E(X) = \int xf(x) \, dx = \int_0^1 x(2x) \, dx = \int_0^1 2x^2 \, dx = \frac{2}{3}$$

EXPECTATION OF A FUNCTION OF A RANDOM VARIABLE

- For continuous random variable X , let $r(X)$: a function of X

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

(Rule of the lazy statistician: could also find it by first finding pdf of $r(X)$ which would require many further calculations. Lazy prefers easy.)

- Same for discrete RV, except replace integral with sum.

EXPECTATION OF A FUNCTION OF A RANDOM VARIABLE

Ex: Assume pdf of previous example, i.e. $f(x) = 2x$, $0 < x < 1$

Find $E(\sqrt{X})$.

$$E(\sqrt{X}) = \int \sqrt{x}f(x) \, dx = \int_0^1 \sqrt{x}2x \, dx = \int_0^1 2x^{3/2} \, dx = \frac{4}{5}$$

FURTHER PROPERTIES

- Sums: Expectation of sum is sum of expectations

For random variables X_1, X_2

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

Note: Generalizes to n variables.

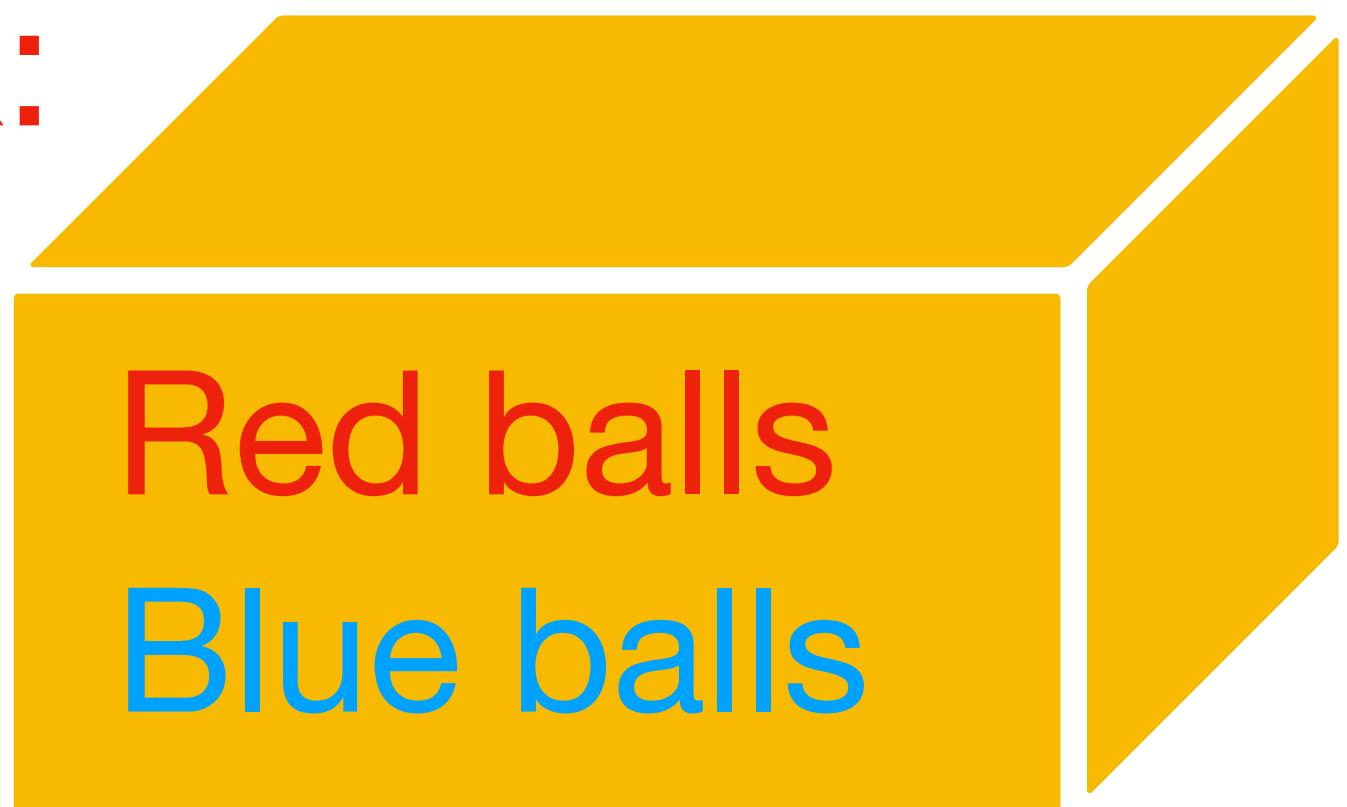
- Also implies linearity of expectation

Let $Y = aX + b$ for constants a, b

$$E(Y) = aE(X) + b$$

FURTHER PROPERTIES

Ex:

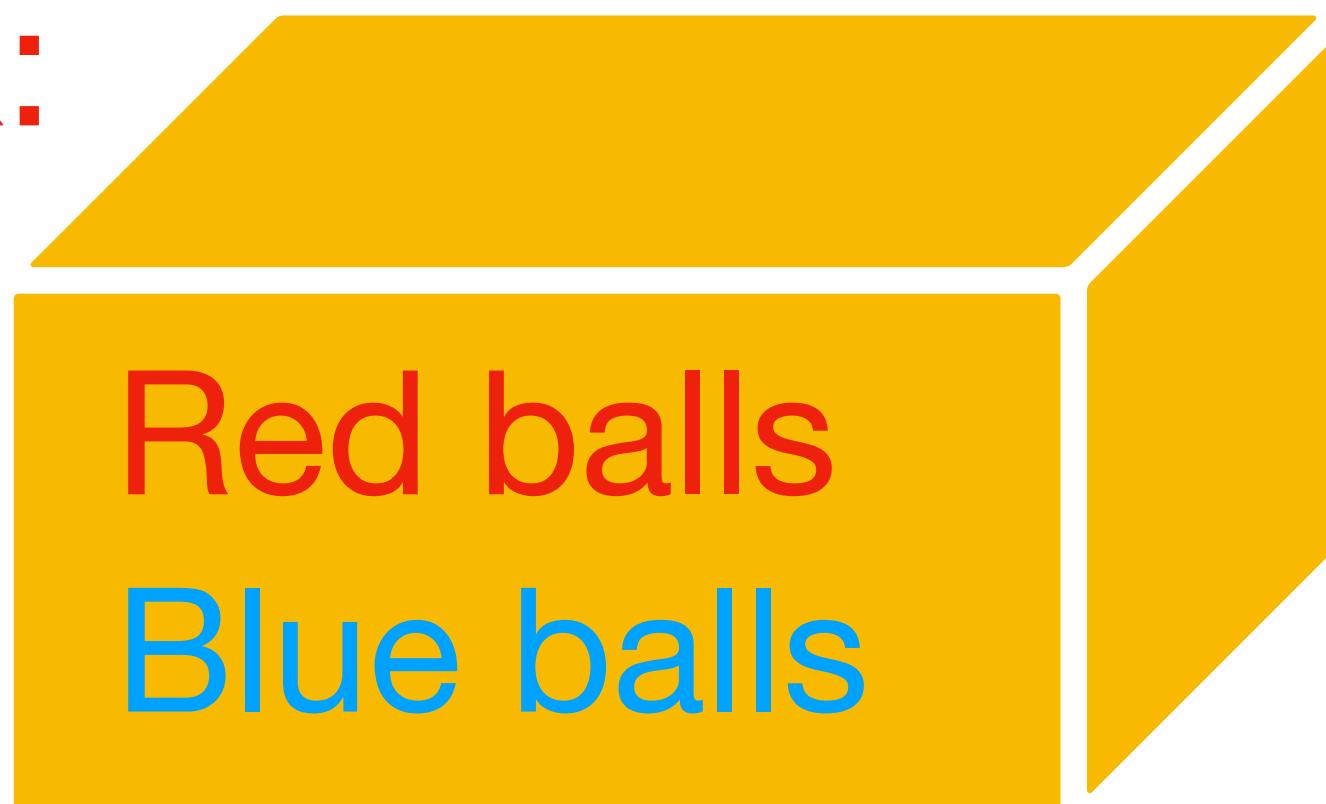


Proportion of R is p . Random sample of n balls selected **with replacement**.

X : # of R balls in the sample. $E(X) = ?$

FURTHER PROPERTIES

Ex:



Proportion of R is p . Random sample of n balls selected **with replacement**.

X : # of R balls in the sample. $E(X) = ?$

Let $X_i = 1$ if i^{th} ball is red, 0 otherwise (for $i = 1, \dots, n$).

$$\Rightarrow X = X_1 + X_2 + \dots + X_n$$

X_i has **Bernoulli distribution** with parameter p .

$$\Rightarrow E(X) = E(X_1 + \dots + X_n) = E(X_1) + \dots + E(X_n) = np$$

CONDITIONAL EXPECTATION

- Conditional expectation of X given $Y = y$:

$$E(X | Y = y) = \sum_x x g_1(x | y) \quad (\text{if } X \text{ has discr. cond. distr. given } Y = y)$$

$$E(X | Y = y) = \int_{-\infty}^{\infty} x g_1(x | y) \quad (\text{if } X \text{ has cont. cond. distr. given } Y = y)$$

CONDITIONAL EXPECTATION

Ex: Roll 2 fair dice. Expected value of die 1 given their sum is 5?

X : outcome of die 1, Y : Sum of 2 dice

$$E(X | Y = y) = \sum_x x g_1(x | y) = \sum_x x \frac{f(x, y)}{f_2(y)}$$

$$E(X | Y = 5) = \sum_{x=1}^{x=4} x \frac{1/36}{4/36} = \sum_{x=1}^{x=4} x \frac{1}{4} = 2.5$$

EXPECTATION

- Readings for the next few lectures:

Ch 8 (Watkins)

Ch 3 (Wasserman, for more formal treatment)

REVIEW OF (LAST PART OF) LECTURE 5

- Mean \equiv expected value \equiv expectation $E(X)$ or μ
$$\sum_{\forall x} xf(x), \text{ discrete } X \quad \int xf(x) dx, \text{ continuous } X$$

- Expectation of a function of a random variable:

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

- Properties:

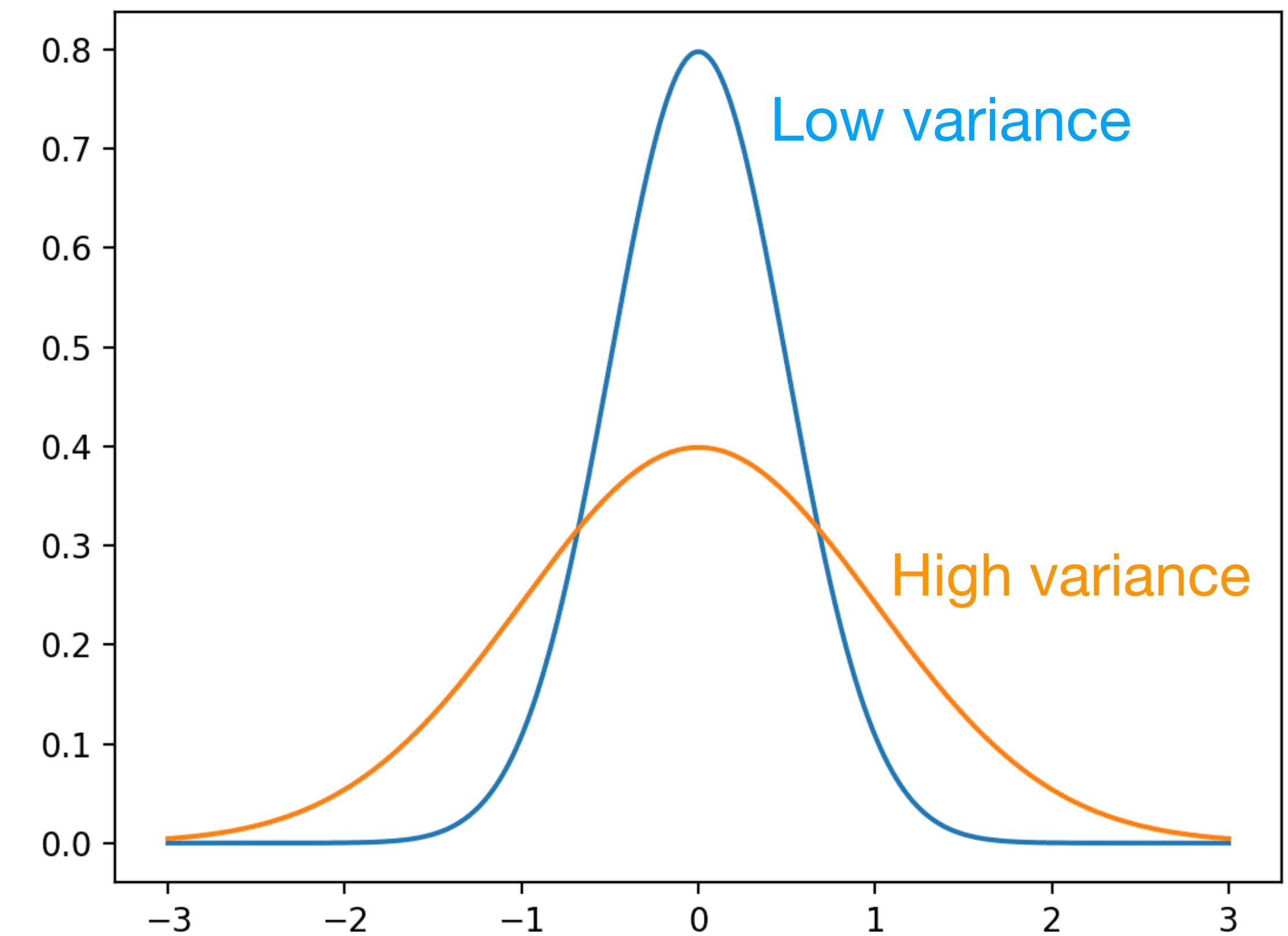
$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

$$E(Y) = aE(X) + b \quad \text{for } Y = aX + b$$

VARIANCE

- Variance of X measures how spread out distribution of X is.

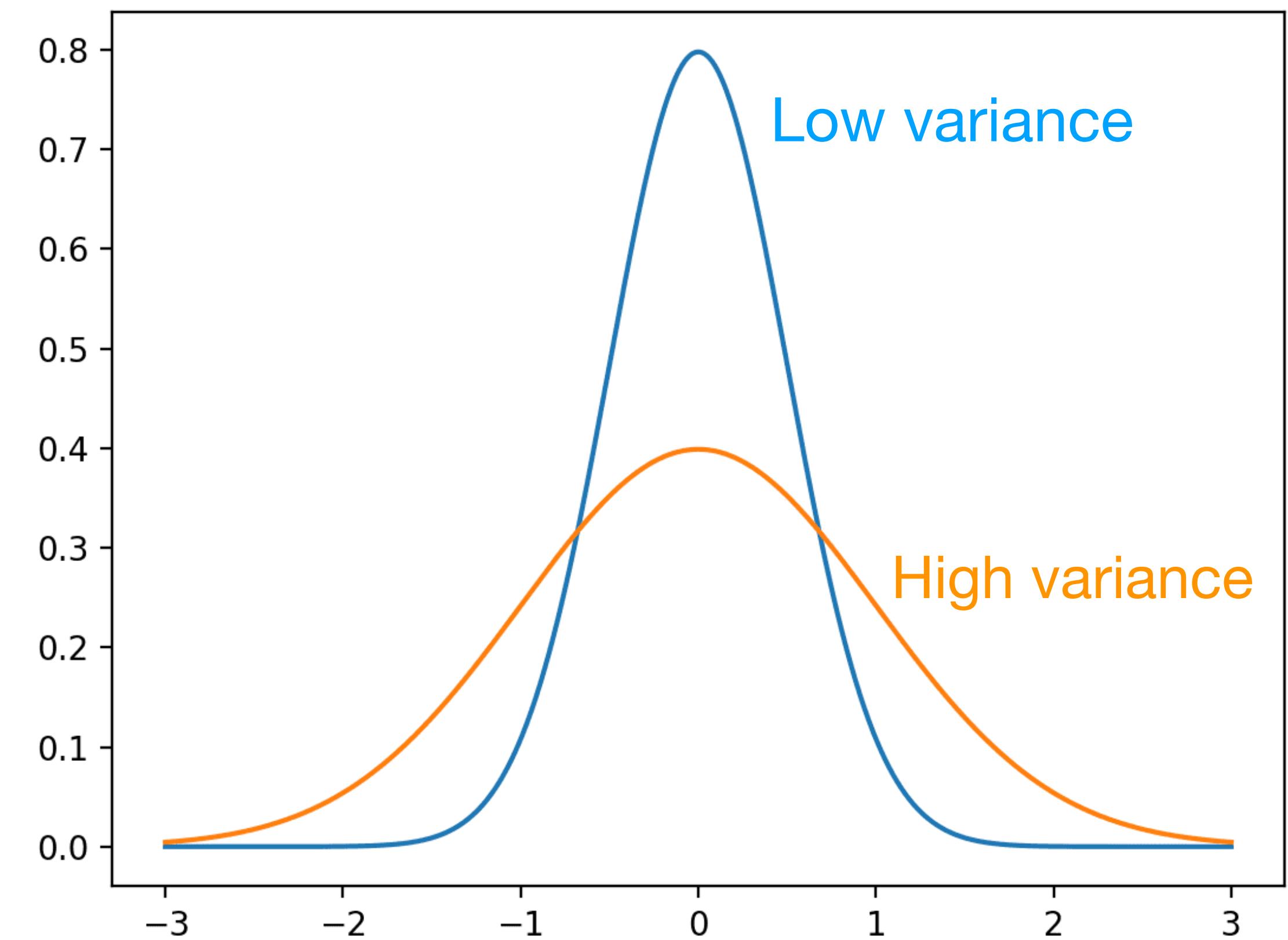
$$Var(X) = \sigma^2 = E[(X - \mu)^2]$$



VARIANCE

- Variance of X measures how spread out distribution of X is.

$$Var(X) = \sigma^2 = E[(X - \mu)^2]$$



- Related measure: Standard deviation of $X = \sigma_x = \sqrt{Var(X)}$

VARIANCE

Ex: Variance of X : outcome of fair six-sided die?

$$E(X) = \frac{7}{2}$$

$$Var(X) = E[(X - E(x))^2]$$

$$= \sum_{x=1}^{x=6} (x - \frac{7}{2})^2 \frac{1}{6}$$

$$= \frac{1}{6}((-5/2)^2 + (-3/2)^2 + (-1/2)^2 + (1/2)^2 + (3/2)^2 + (5/2)^2)$$

$$\approx 2.92$$

Note that $E[r(X)] = \sum_x r(x)f(x)$

VARIANCE

- An equivalent formula for variance:

$$Var(X) = E(X^2) - [E(X)]^2$$

Proof:

$$\begin{aligned} Var(X) &= E[(X - \mu)^2] = E[X^2 - 2X\mu + \mu^2] \\ &= E(X^2) - E(2X\mu) + E(\mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2 \end{aligned}$$

- More useful in calculations.

VARIANCE

Ex: Variance of X : outcome of fair n-sided die.

Helpful formula: $\sum_{x=1}^n x = n(n + 1)/2$, $\sum_{x=1}^n x^2 = n(n + 1)(2n + 1)/6$

$$\begin{aligned}Var(X) &= E(X^2) - [E(X)]^2 = \sum_{x=1}^n x^2 \frac{1}{n} - \left(\sum_{x=1}^n x \frac{1}{n} \right)^2 \\&= \frac{1}{n} \times \frac{n(n + 1)(2n + 1)}{6} - \left(\frac{1}{n} \times \frac{n(n + 1)}{2} \right)^2 = \frac{n^2 - 1}{12}\end{aligned}$$

PROPERTIES OF VARIANCE

- Let $Y = aX + b$ for constants a, b .

$$Var(Y) = a^2 Var(X)$$

Proof:

$$Var(Y) = E[(aX + b - E(aX + b))^2]$$

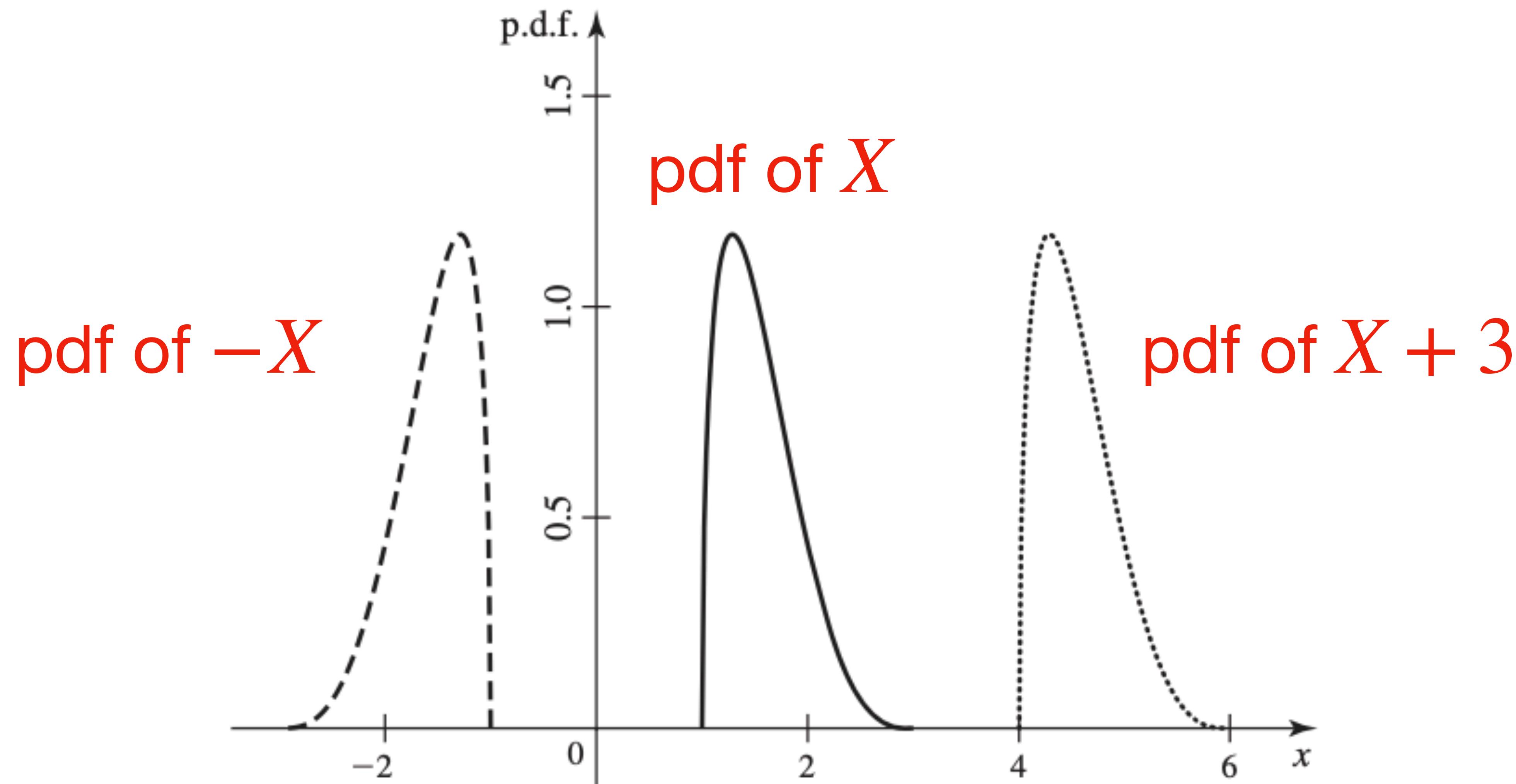
$$= E[(aX + b - aE(X) - b)^2]$$

$$= E[(aX - aE(X))^2]$$

$$= E[a^2(X - E(X))^2] = a^2 E[(X - E(X))^2] = a^2 Var(X)$$

PROPERTIES OF VARIANCE

- Previous result implies $\text{Var}(X + b) = \text{Var}(X)$, for constant b .



SUMMARIES OF ASSOCIATION BETWEEN TWO VARIABLES

- Covariance of X, Y : numerical measure of the degree to which X and Y vary together. Let $E(X) = \mu_x, E(Y) = \mu_y$.

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

SUMMARIES OF ASSOCIATION BETWEEN TWO VARIABLES

- Covariance of X, Y : numerical measure of the degree to which X and Y vary together. Let $E(X) = \mu_x, E(Y) = \mu_y$.

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

- Equivalent formula: $Cov(X, Y) = E(XY) - \mu_x\mu_y$

Proof: $Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$

$$= E(XY) - \mu_y E(X) - \mu_x E(Y) + \mu_x \mu_y$$

$$= E(XY) - 2\mu_x \mu_y + \mu_x \mu_y = E(XY) - \mu_x \mu_y$$

SUMMARIES OF ASSOCIATION BETWEEN TWO VARIABLES

- Covariance of X, Y : numerical measure of the degree to which X and Y vary together

BUT

is also influenced by the overall magnitudes of X, Y .

$$Cov(2X, Y) = 2Cov(X, Y) \text{ for instance.}$$

SUMMARIES OF ASSOCIATION BETWEEN TWO VARIABLES

- Covariance of X, Y : numerical measure of the degree to which X and Y vary together

BUT

is also influenced by the overall magnitudes of X, Y .

$$\text{Cov}(2X, Y) = 2\text{Cov}(X, Y) \text{ for instance.}$$

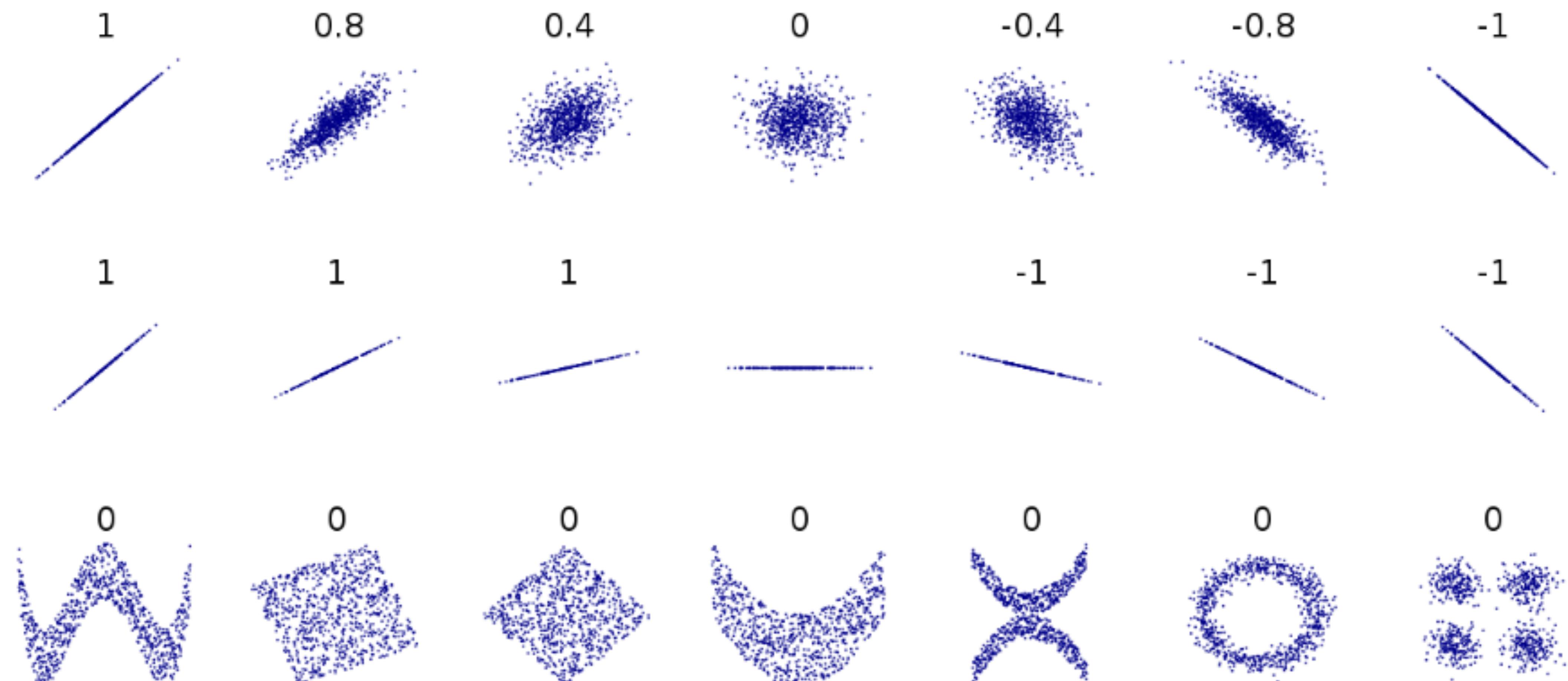
- Better measure, independent of arbitrary changes in scales:

$$\text{Correlation of } X, Y = \rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Measures linear association of X, Y . Always between -1 and 1 .

SUMMARIES OF ASSOCIATION BETWEEN TWO VARIABLES

- Example instances of $\rho(X, Y)$:



SUMMARIES OF ASSOCIATION BETWEEN TWO VARIABLES

- Note: Correlation does not imply causation!

Assume strong ρ between X, Y . Possible scenarios:

- X causes Y .
- Y causes X .
- Z (another factor/set of factors) causes changes in X, Y .

Am I happy because I ate more pizza?

Am I eating more pizza because I am happy?

Am I happy and eating more pizza because prices went down?



VARIANCE OF SUM OF TWO VARIABLES

- For expectation we had: $E(X + Y) = E(X) + E(Y)$.
Does it hold for variance? NO!
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

VARIANCE OF SUM OF TWO VARIABLES

- For expectation we had: $E(X + Y) = E(X) + E(Y)$.
Does it hold for variance? NO!
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

Proof: We know $E(X + Y) = \mu_X + \mu_Y$

$$\begin{aligned}Var(X + Y) &= E[(X + Y - E(X + Y))^2] = E[(X + Y - \mu_X - \mu_Y)^2] \\&= E[(X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)] \\&= \text{E}[(X - \mu_X)^2] + \text{E}[(Y - \mu_Y)^2] + 2\text{E}[(X - \mu_X)(Y - \mu_Y)]\end{aligned}$$

INDEPENDENT RANDOM VARIABLES

- For expectation we had: $E(X + Y) = E(X) + E(Y)$.
(Holds whether X, Y independent or not)

INDEPENDENT RANDOM VARIABLES

- For expectation we had: $E(X + Y) = E(X) + E(Y)$.
(Holds whether X, Y independent or not)
- How about their product? Is $E(XY) = E(X)E(Y)$ always true?

If X, Y **independent** random variables $E(XY) = E(X)E(Y)$

Note: Generalizes to n independent variables.

INDEPENDENT RANDOM VARIABLES

- For expectation we had: $E(X + Y) = E(X) + E(Y)$.
(Holds whether X, Y independent or not)
- How about their product? Is $E(XY) = E(X)E(Y)$ always true?

If X, Y **independent** random variables $E(XY) = E(X)E(Y)$

Proof: (Only for discrete case. Continuous case is similar.)

$$\begin{aligned} E(XY) &= \sum_{\forall x,y} xyf(x,y) = \sum_{\forall x} \sum_{\forall y} xyf(x)f(y) = \sum_{\forall x} xf(x) \sum_{\forall y} yf(y) \\ &= \sum_{\forall x} xf(x)\mu_y = \mu_y \sum_{\forall x} xf(x) = \mu_y \mu_x \end{aligned}$$

$\xrightarrow{X \perp Y \Rightarrow f(x,y) = f(x)f(y)}$

Note: Generalizes to n independent variables.

INDEPENDENT RANDOM VARIABLES

Ex: Roll 2 fair dice. X : outcome of die 1, Y : outcome of die 2.

Mean of their product?

$$X, Y \text{ independent} \Rightarrow E(XY) = E(X)E(Y) = \frac{7}{2} \times \frac{7}{2} = \frac{49}{4}$$

QUIZ

Write down the formula for calculating the mean of continuous random variable X .

INDEPENDENT RANDOM VARIABLES

- We saw that:

If X, Y independent random variables $E(XY) = E(X)E(Y)$

- Now variance under independence:

If X, Y independent then $Var(X + Y) = Var(X) + Var(Y)$

Proof:

INDEPENDENT RANDOM VARIABLES

- We saw that:

If X, Y independent random variables $E(XY) = E(X)E(Y)$

- Now variance under independence:

If X, Y independent then $Var(X + Y) = Var(X) + Var(Y)$

Proof:

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

$$Cov(X, Y) = E(XY) - \mu_X\mu_Y = \mu_X\mu_Y - \mu_X\mu_Y = 0$$

INDEPENDENT RANDOM VARIABLES

- Result of the previous proof:

If X, Y independent then $\text{Cov}(X, Y) = \rho(X, Y) = 0$

- Is the converse also always true? NO!

INDEPENDENT RANDOM VARIABLES

- Result of the previous proof:

If X, Y independent then $\text{Cov}(X, Y) = \rho(X, Y) = 0$

- Is the converse also always true? NO!

Counter Example: $X \in \{-1, 0, 1\}$ each equally likely.

Let $Y = X^2$. $\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = E(X^3) - 0$

X^3 same variable as $X \Rightarrow E(X^3) = 0 \Rightarrow \text{Cov}(X, Y) = 0$

X, Y not correlated but dependent (Y determined by X).

WIDELY USED DISCRETE DISTRIBUTIONS: BERNOULLI

X has Bernoulli distribution with parameter p , if

- Only outcomes are $X = 0$ or $X = 1$ and
- $P(X = 1) = p \quad P(X = 0) = 1 - p$

Denoted with $X \sim \text{Bernoulli}(p)$

WIDELY USED DISCRETE DISTRIBUTIONS: BERNOULLI

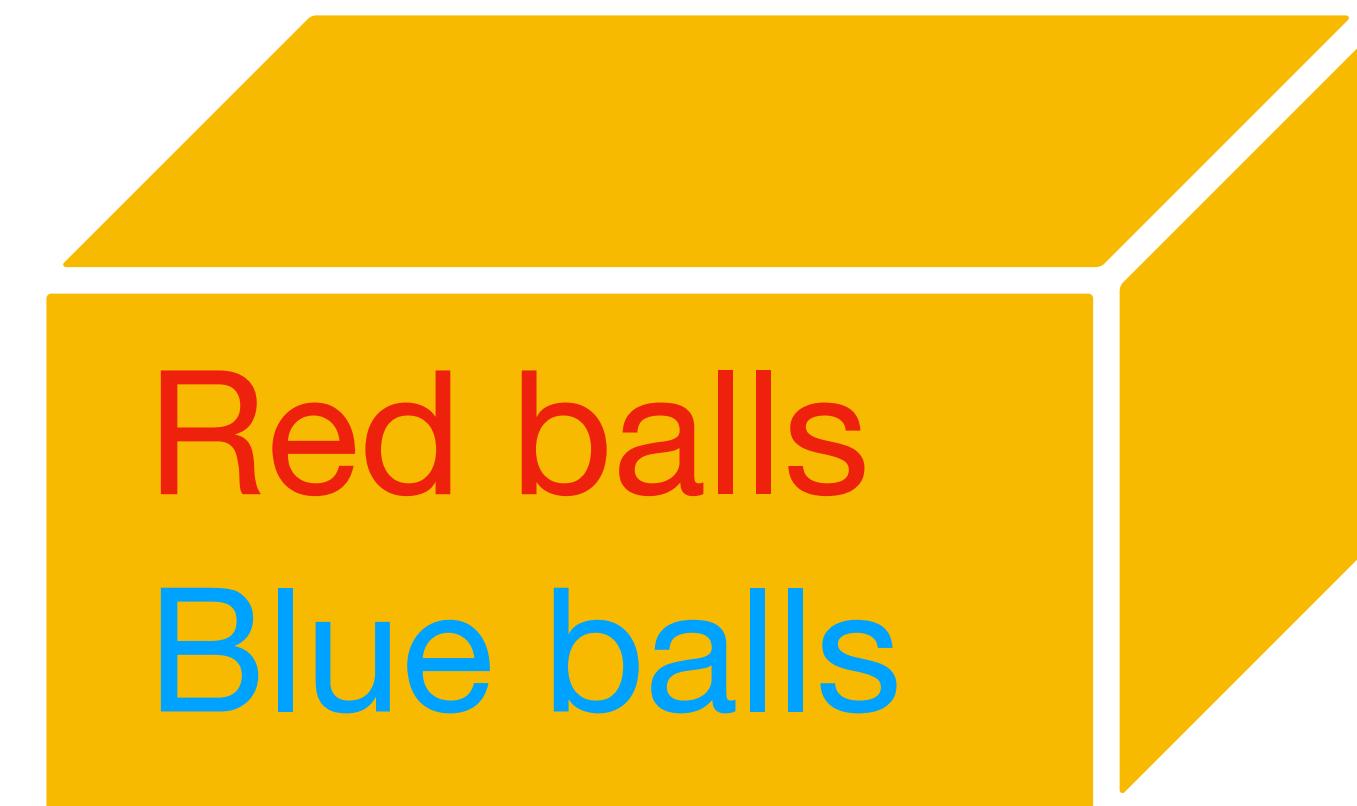
- Examples:



Tossing a coin ($X = 1$ if Head, $X = 0$ if Tail and $p = 1/2$ if fair)

WIDELY USED DISCRETE DISTRIBUTIONS: BERNOULLI

- Examples:



Drawing one R ball where proportion of R balls in the box is p

($X = 1$ if R , $X = 0$ if B and parameter p)

WIDELY USED DISCRETE DISTRIBUTIONS: BERNOULLI

- Examples:



Clinical trial: Treatment success

($X = 1$ if Successful, $X = 0$ if Fails and parameter p)

WIDELY USED DISCRETE DISTRIBUTIONS: BERNOULLI

- **Expectation and Variance of Bernoulli distribution:**

$$E(X) = 1 \times p + 0 \times (1 - p) = p$$

Since X and X^2 are the same random variable: $E(X^2) = E(X)$

$$\text{Var}(X) = E(X^2) - (E(X))^2 = p - p^2 = p(1 - p)$$

WIDELY USED DISCRETE DISTRIBUTIONS: BERNOULLI

- Bernoulli Trial:

If X_1, X_2, \dots are iid (independently and identically distributed) and each X_i has Bernoulli distribution with parameter p
 $\Rightarrow X_1, X_2, \dots$ are Bernoulli trials with parameter p .

WIDELY USED DISCRETE DISTRIBUTIONS: BINOMIAL

X has Binomial distribution with parameters n, p , if

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, 2, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

Denoted with $X \sim \text{Binomial}(n, p)$

or $X \sim B_{n,p}$

WIDELY USED DISCRETE DISTRIBUTIONS: BINOMIAL

- Example:



A coin with probability of Head p . Toss it n times. Tosses are iid.

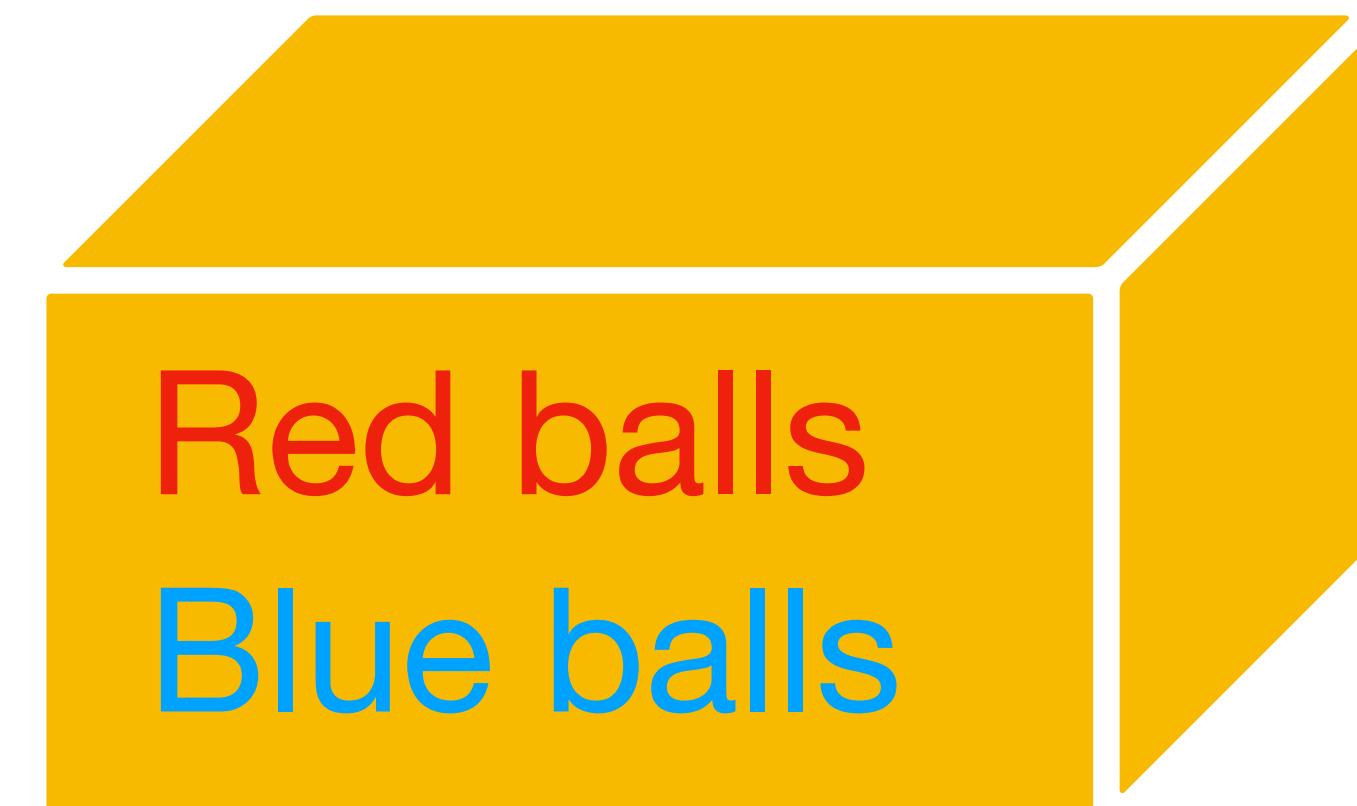
X : # of Heads in n tosses. $X \sim \text{Binomial}(n, p)$

For a particular sequence of x Heads and $(n - x)$ Tails:

probability is $p^x(1 - p)^{n-x} \cdot \binom{n}{x}$ ways of choosing a sequence.

WIDELY USED DISCRETE DISTRIBUTIONS: BINOMIAL

- Example:



Proportion of R balls in a box containing R, B balls is p .

X : # of R balls in n draws with replacement.

$X_i = 1$, if i^{th} draw a R ball, $X_i = 0$ otherwise.

$$X = X_1 + X_2 + \cdots + X_n, \quad X \sim \text{Binomial}(n, p)$$

WIDELY USED DISCRETE DISTRIBUTIONS: BINOMIAL

- Can generalize the previous example:

Theorem

Let X_1, X_2, \dots, X_n form n Bernoulli trials and $X = X_1 + \dots + X_n$

$$X \sim \text{Binomial}(n, p)$$

WIDELY USED DISCRETE DISTRIBUTIONS: BINOMIAL

- Expectation and Variance of Binomial distribution:

$$E(X) = \sum_{i=1}^n E(X_i) = \sum_{i=1}^n p = np$$

$$Var(X) = \sum_{i=1}^n Var(X_i) = \sum_{i=1}^n p(1 - p) = np(1 - p)$$

WIDELY USED DISCRETE DISTRIBUTIONS: BINOMIAL

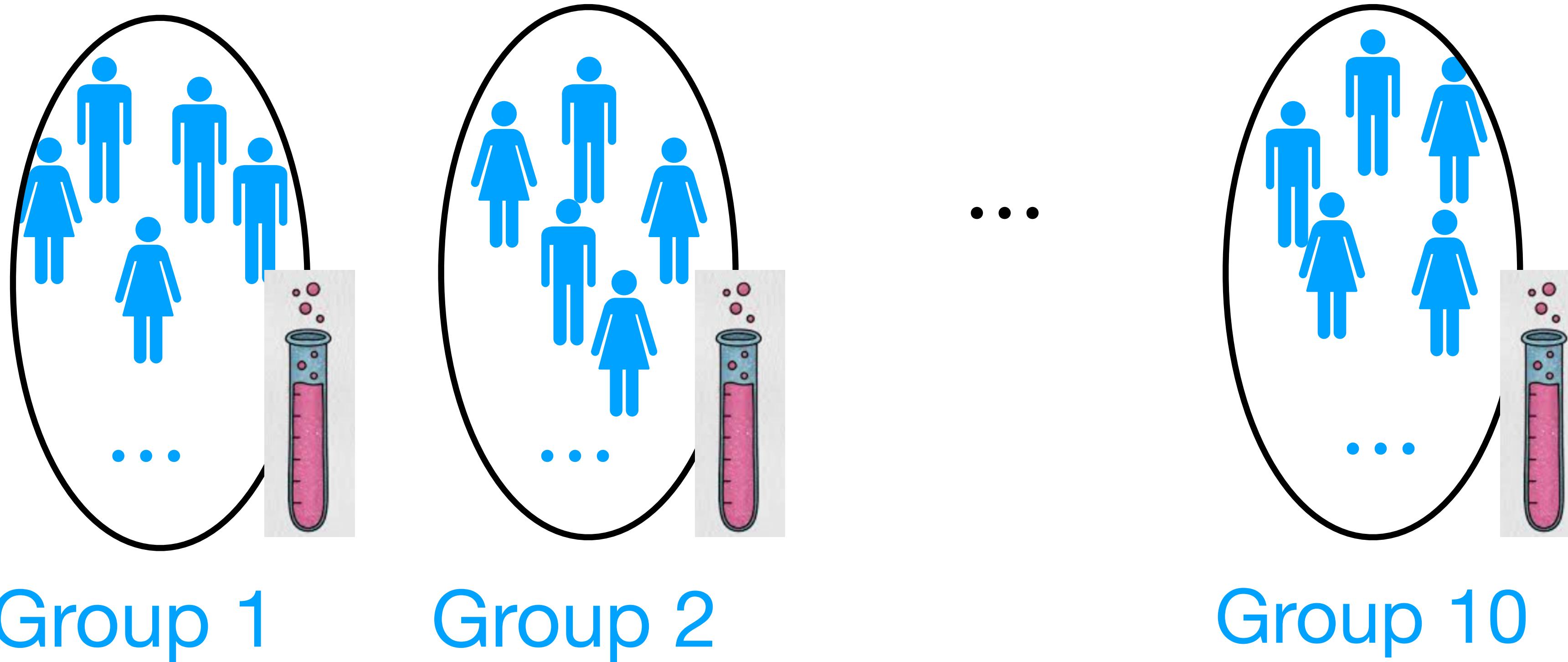
Ex: Test large group (1000) for a rare disease (affects 2 in 1000).

Naive testing requires 1000 tests.

WIDELY USED DISCRETE DISTRIBUTIONS: BINOMIAL

Ex: Test large group (1000) for a rare disease (affects 2 in 1000).

Naive testing requires 1000 tests. A better way:

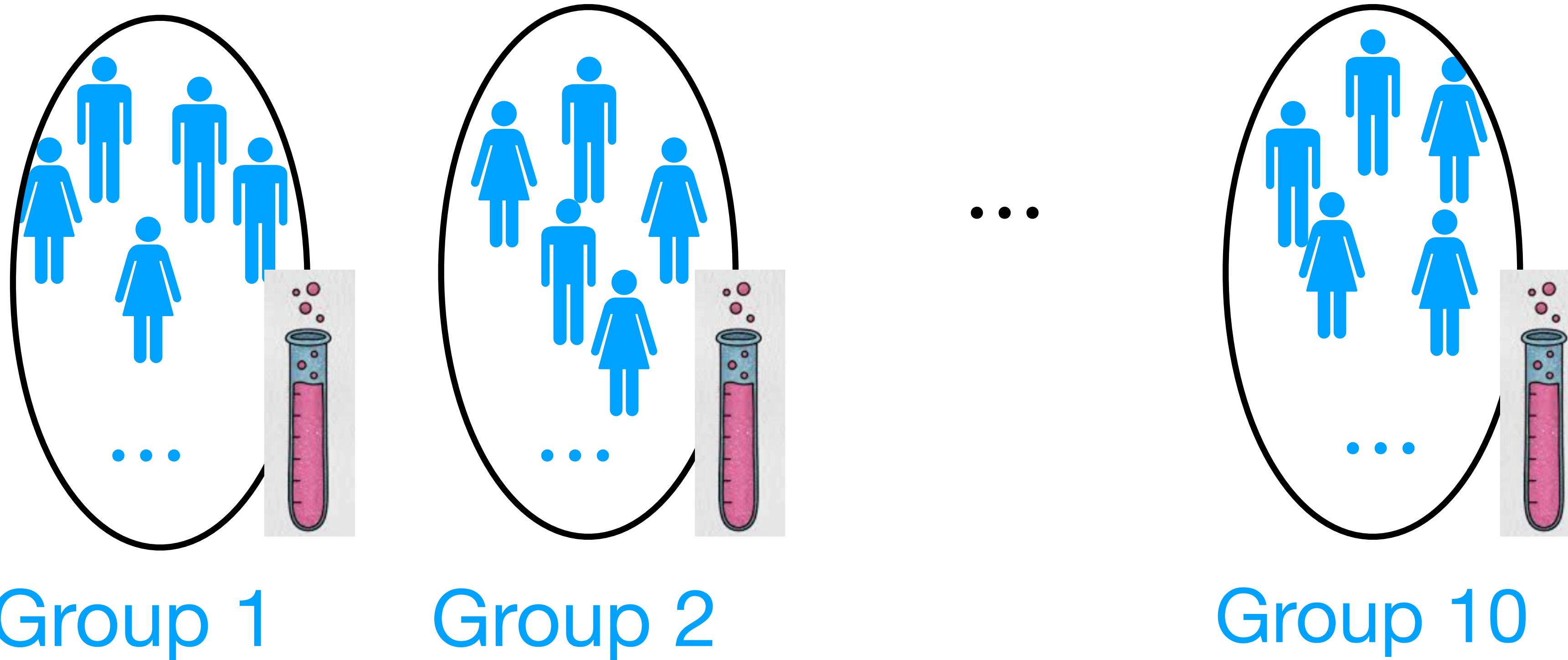


100 people in each group. Single mixed tube for each group.

WIDELY USED DISCRETE DISTRIBUTIONS: BINOMIAL

Ex: Test large group (1000) for a rare disease (affects 2 in 1000).

Naive testing requires 1000 tests. A better way:

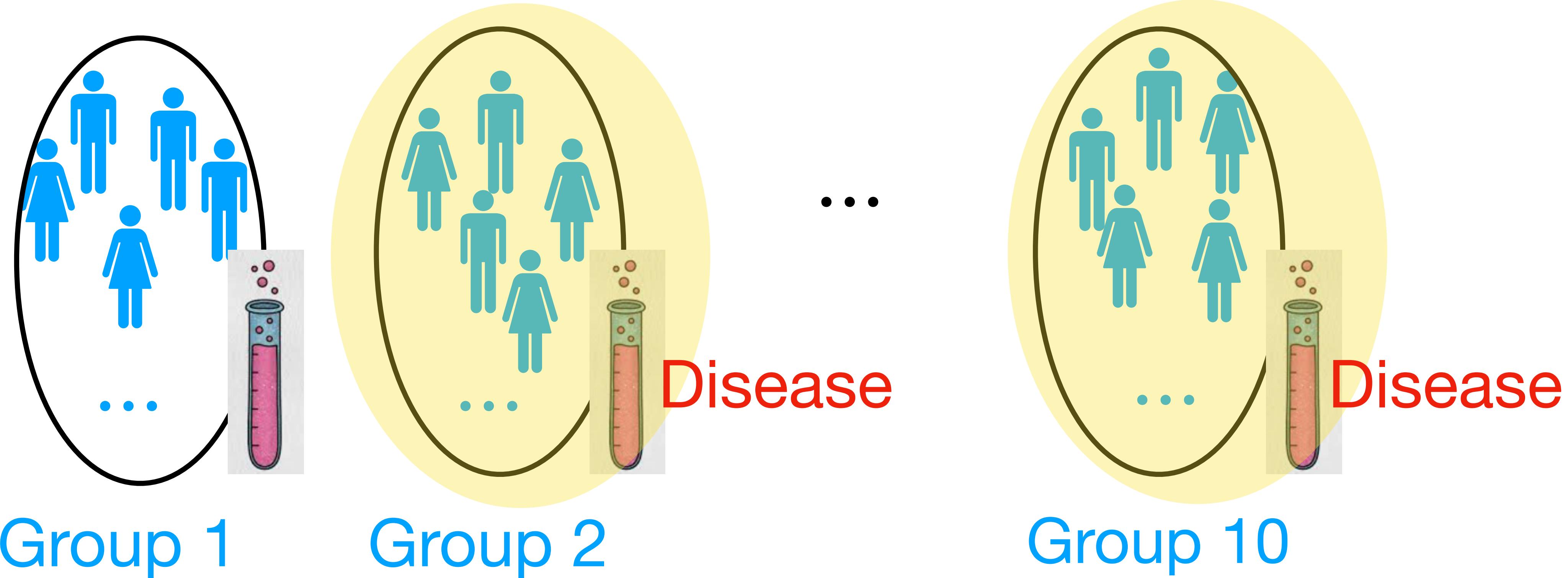


10 tests so far. If no group test shows disease we are done!

WIDELY USED DISCRETE DISTRIBUTIONS: BINOMIAL

Ex: Test large group (1000) for a rare disease (affects 2 in 1000).

Naive testing requires 1000 tests. A better way:

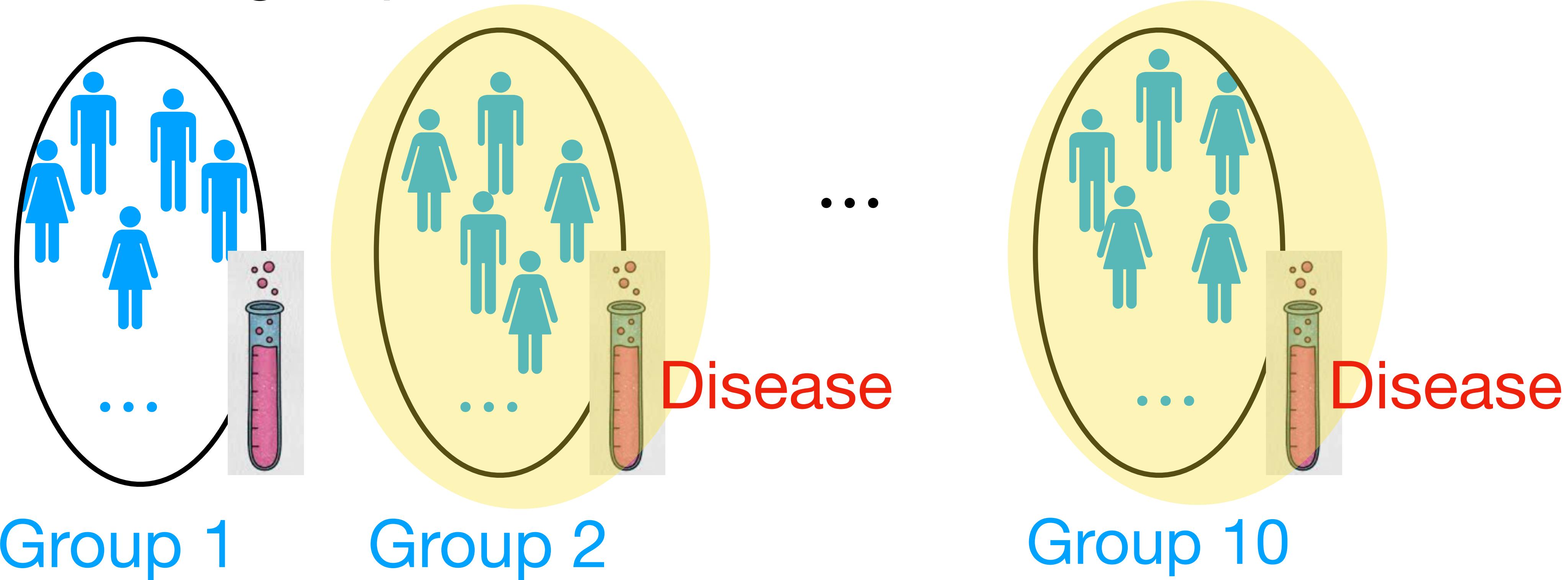


If a group's test shows disease,
test all individuals in group.

WIDELY USED DISCRETE DISTRIBUTIONS: BINOMIAL

Ex: Test large group (1000) for a rare disease (affects 2 in 1000).

Naive testing requires 1000 tests. A better way:



Expected number of tests?

WIDELY USED DISCRETE DISTRIBUTIONS: BINOMIAL

Ex: Test large group (1000) for a rare disease (affects 2 in 1000).

Naive testing requires 1000 tests. A better way:

$X_j = 1$ if j has disease, o.w. $X_j = 0 \Rightarrow$

$X_j \sim Bernoulli(0.002)$, iid for $j = 1, \dots, 1000$



WIDELY USED DISCRETE DISTRIBUTIONS: BINOMIAL

Ex: Test large group (1000) for a rare disease (affects 2 in 1000).

Naive testing requires 1000 tests. A better way:



Group i

$X_j = 1$ if j has disease, o.w. $X_j = 0 \Rightarrow$

$X_j \sim Bernoulli(0.002)$, iid for $j = 1, \dots, 1000$

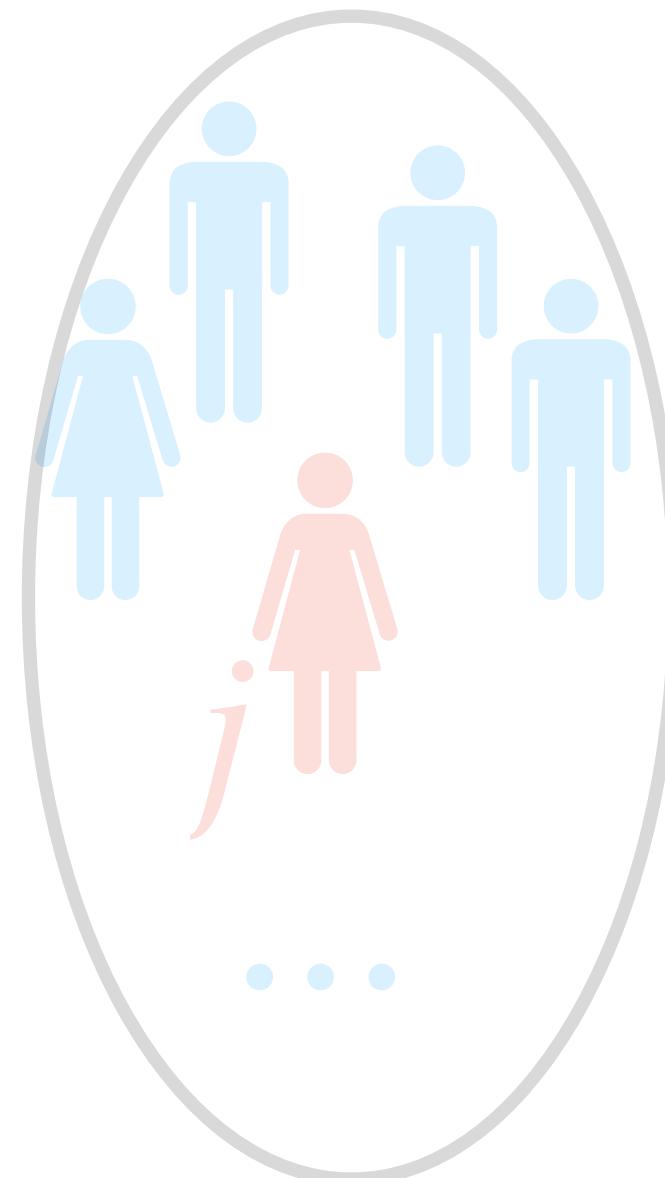
Y_i : # of diseased people in Group $i \Rightarrow$

$Y_i \sim Binomial(100, 0.002)$ for $i = 1, \dots, 10$

WIDELY USED DISCRETE DISTRIBUTIONS: BINOMIAL

Ex: Test large group (1000) for a rare disease (affects 2 in 1000).

Naive testing requires 1000 tests. A better way:



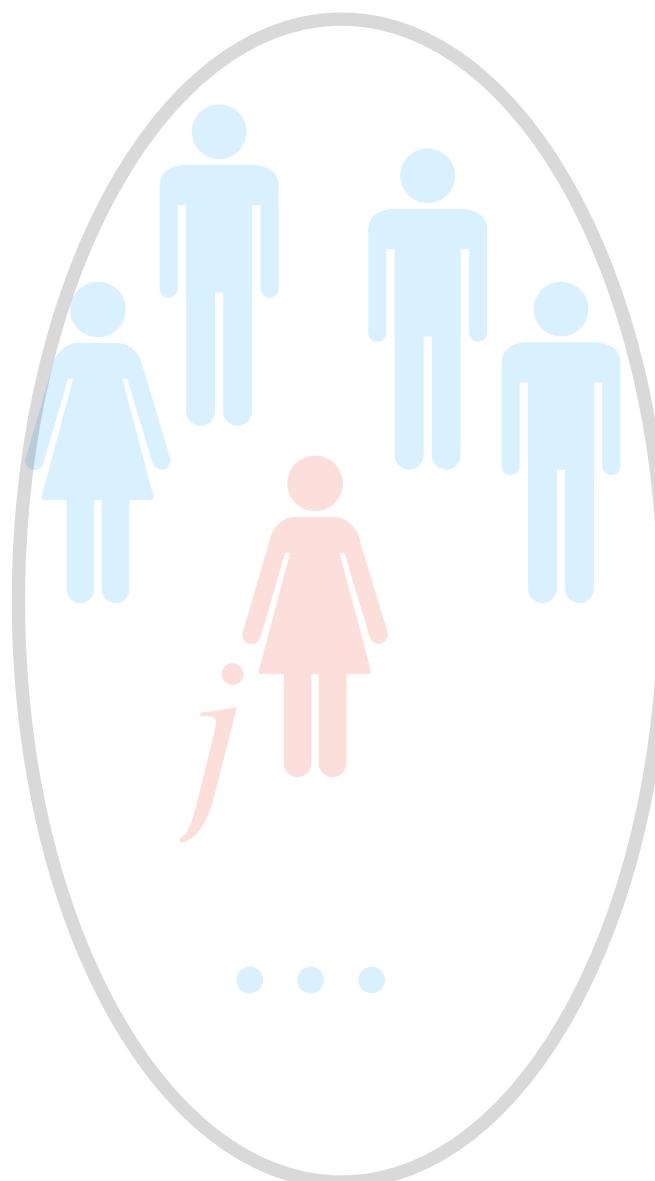
Y_i : # of diseased people in Group i \Rightarrow

$Y_i \sim \text{Binomial}(100, 0.002)$ for $i = 1, \dots, 10$

WIDELY USED DISCRETE DISTRIBUTIONS: BINOMIAL

Ex: Test large group (1000) for a rare disease (affects 2 in 1000).

Naive testing requires 1000 tests. A better way:



Y_i : # of diseased people in Group i \Rightarrow

$Y_i \sim \text{Binomial}(100, 0.002)$ for $i = 1, \dots, 10$

$Z_i = 1$ if $Y_i > 0$, o.w. $Z_i = 0 \Rightarrow Z_i \sim \text{Bernoulli}(p)$,

where $p = 1 - P(Y_i = 0)$

```
from scipy.stats import binom  
print(1-binom.pmf(0, 100, 0.002))
```

0.18143319531157243

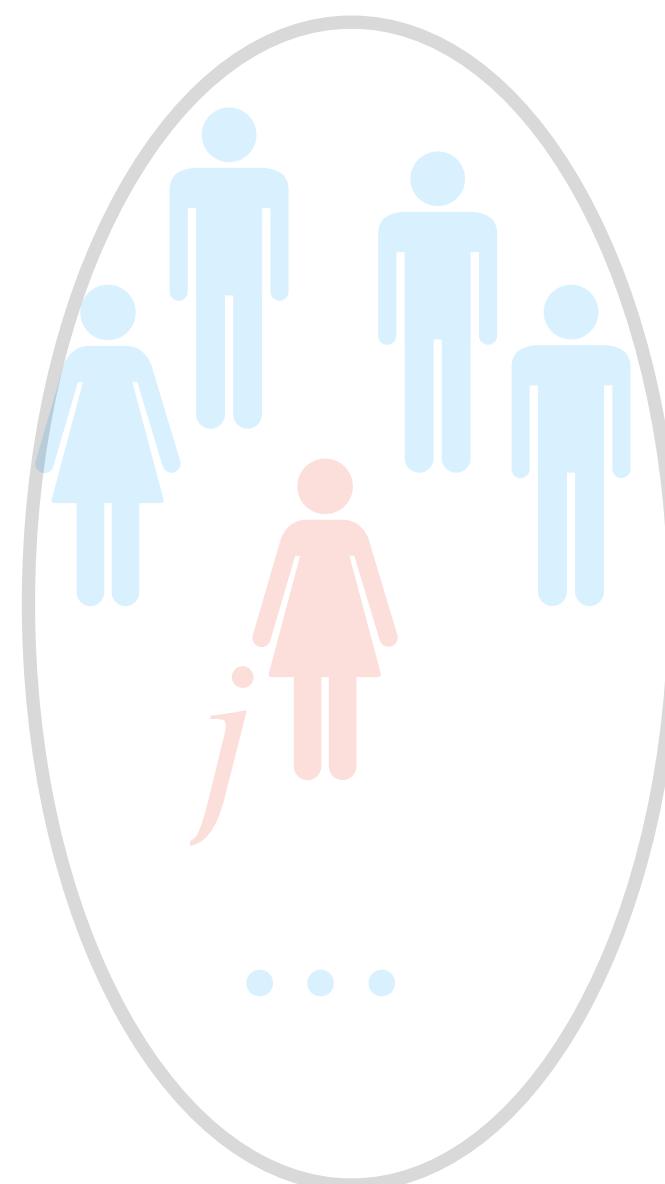
WIDELY USED DISCRETE DISTRIBUTIONS: BINOMIAL

Ex: Test large group (1000) for a rare disease (affects 2 in 1000).

Naive testing requires 1000 tests. A better way:

$$Z_i = 1 \text{ if } Y_i > 0, \text{ o.w. } Z_i = 0 \Rightarrow Z_i \sim \text{Bernoulli}(p),$$

where $p = 1 - P(Y_i = 0) = 0.181$

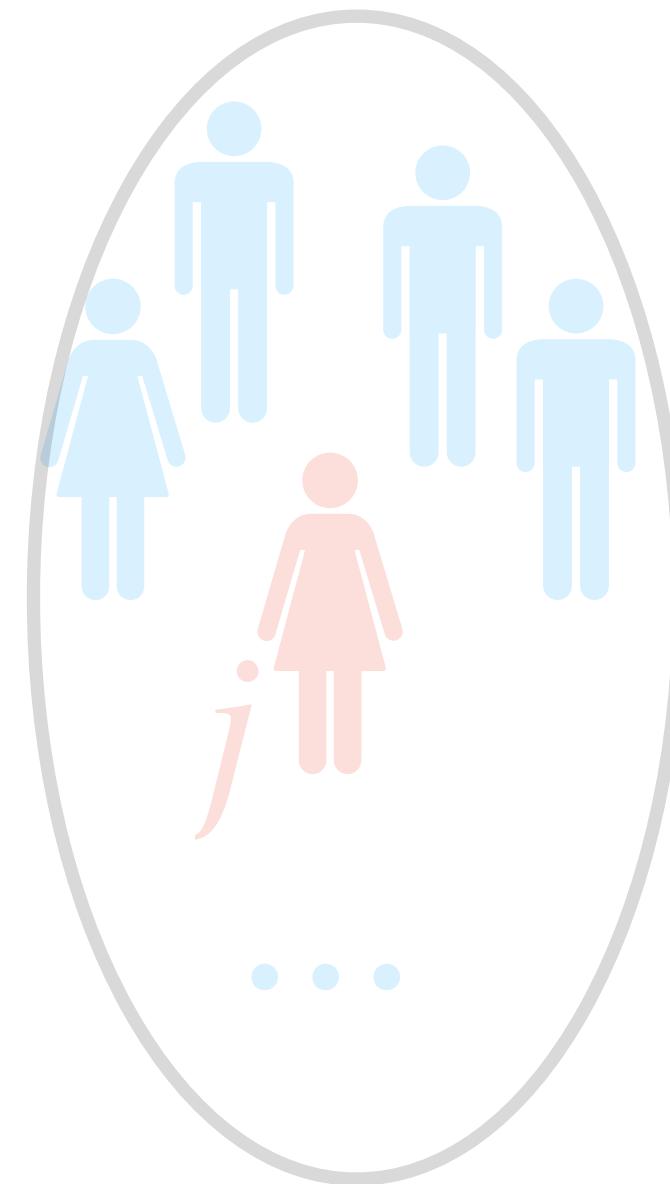


Group i

WIDELY USED DISCRETE DISTRIBUTIONS: BINOMIAL

Ex: Test large group (1000) for a rare disease (affects 2 in 1000).

Naive testing requires 1000 tests. A better way:



$$Z_i = 1 \text{ if } Y_i > 0, \text{ o.w. } Z_i = 0 \Rightarrow$$

$$\text{where } p = 1 - P(Y_i = 0) = 0.181$$

$$W = Z_1 + \dots + Z_{10} \Rightarrow W \sim \text{Binomial}(10, 0.181)$$

$$E(W) = 10 \times 0.181 = 1.81$$

$$E(100W) = 181, \text{ Expected \# of tests} = 191$$

REVIEW OF PREVIOUS LECTURE

- If X, Y independent random variables:

$$E(XY) = E(X)E(Y), \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y), \text{Cov}(X, Y) = 0$$

- Bernoulli random variable: $X \sim \text{Bernoulli}(p)$

$$P(X = 1) = p \quad P(X = 0) = 1 - p$$

$$E(X) = p, \text{Var}(X) = p(1 - p)$$

- Binomial random variable: $X \sim \text{Binomial}(n, p)$

$$\binom{n}{x} p^x (1 - p)^{n-x} \quad E(X) = np, \text{Var}(X) = np(1 - p)$$

WIDELY USED DISCRETE DISTRIBUTIONS: HYPERGEOMETRIC

X has Hypergeometric distribution with parameters n, A, B , if

$$f(x) = \begin{cases} \frac{\binom{A}{x} \binom{B}{n-x}}{\binom{A+B}{n}} & \text{for } \max(0, n-B) \leq x \leq \min(n, A) \\ 0 & \text{otherwise} \end{cases}$$

- **Example:** A red, B blue balls.
 n balls selected randomly without replacement.
 X : # of red balls selected.

WIDELY USED DISCRETE DISTRIBUTIONS: HYPERGEOMETRIC

- Expectation of Hypergeometric distribution:

Let $p = \frac{A}{A + B}$. Let $X_i = 1$ if i^{th} ball red, $X_i = 0$ otherwise.

WIDELY USED DISCRETE DISTRIBUTIONS: HYPERGEOMETRIC

- Expectation of Hypergeometric distribution:

Let $p = \frac{A}{A + B}$. Let $X_i = 1$ if i^{th} ball red, $X_i = 0$ otherwise.

Although X_i are dependent, we have

$$P(X_i = 1) = p, P(X_i = 0) = 1 - p \quad E(X_i) = p$$

WIDELY USED DISCRETE DISTRIBUTIONS: HYPERGEOMETRIC

- Expectation of Hypergeometric distribution:

Let $p = \frac{A}{A + B}$. Let $X_i = 1$ if i^{th} ball red, $X_i = 0$ otherwise.

Although X_i are dependent, we have

$$P(X_i = 1) = p, P(X_i = 0) = 1 - p \quad E(X_i) = p$$

$$E(X) = \sum_{i=1}^n E(X_i) = \sum_{i=1}^n p = np$$

Same as Binomial Distribution (Sampling with replacement).

WIDELY USED DISCRETE DISTRIBUTIONS: HYPERGEOMETRIC

- **Variance** of Hypergeometric distribution:

We will not derive it. Since X_i dependent, as expected, variance will differ from that of Binomial:

$$Var(X) = np(1 - p)\left(\frac{T - n}{T - 1}\right)$$

$T = A + B$ is size of population, $p = \frac{A}{A + B}$, n is sample size

Variance smaller than that of Binomial (Sampling with replacement).

WIDELY USED DISCRETE DISTRIBUTIONS: HYPERGEOMETRIC

- Comparison of Binomial and Hypergeometric:

Although their pmf and variances are different:

If Population size $A + B \gg$ Sample size $n \Rightarrow$

Sampling with replacement \equiv Sampling without replacement

(Binomial distr. almost same as Hypergeometric distr.)

WIDELY USED DISCRETE DISTRIBUTIONS: HYPERGEOMETRIC

- Comparison of Binomial and Hypergeometric:

Although their pmf and variances are different:

If Population size $A + B \gg$ Sample size $n \Rightarrow$

Sampling with replacement \equiv Sampling without replacement
(Binomial distr. almost same as Hypergeometric distr.)

- How big? n is less than 5 % of population size.

```
from scipy.stats import binom  
binom.pmf(0, 100, 0.002)
```

0.8185668046884276

```
from scipy.stats import hypergeom  
hypergeom.pmf(0, 2000, 4, 100)
```

0.8143774997343618

WIDELY USED DISCRETE DISTRIBUTIONS: POISSON

X has Poisson distribution with parameter λ , if

$$f(x) = \begin{cases} e^{-\lambda} \frac{\lambda^x}{x!} & \text{for } x = 0, 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}$$

WIDELY USED DISCRETE DISTRIBUTIONS: POISSON

X has Poisson distribution with parameter λ , if

$$f(x) = \begin{cases} e^{-\lambda} \frac{\lambda^x}{x!} & \text{for } x = 0, 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}$$

Can model occurrences of random arrivals:

Customer arrivals, traffic accidents, flood occurrences, etc.

Assume # of arrivals in any interval is independent.

Used to approximate Binomial if large n , small p (rare event).

Note: Both expectation and variance = λ .

WIDELY USED DISCRETE DISTRIBUTIONS: POISSON

Ex: Rare disease with probability 0.01.

Probability that in a random group of 200, 3 people have disease

Let X : # of people with disease.

With Binomial (exact): $P(X = 3) = \binom{200}{3} (0.01^3)(0.99)^{197} \approx 0.181$

With Poisson (approximation):

$$\lambda = 200 \times 0.01 = 2 \quad P(X = 3) = e^{-2} \frac{2^3}{3!} \approx 0.18$$

WIDELY USED DISCRETE DISTRIBUTIONS: POISSON

Ex: Customers arrive at a store at a rate of 3/hour on average.

Assume customer arrivals in different time periods independent.

Probability that 10 or more customers arrive between 2pm-4pm

of customers in any 2 hour period, $\lambda = 6$.

```
from scipy.stats import poisson  
print(1-poisson.cdf(9, 6))
```

0.08392401699487584

WIDELY USED CONTINUOUS DISTRIBUTIONS: NORMAL

X has Normal (Gaussian) distribution with parameter μ, σ^2 , if

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

Denoted with $X \sim N(\mu, \sigma^2)$

WIDELY USED CONTINUOUS DISTRIBUTIONS: NORMAL

X has Normal (Gaussian) distribution with parameter μ, σ^2 , if

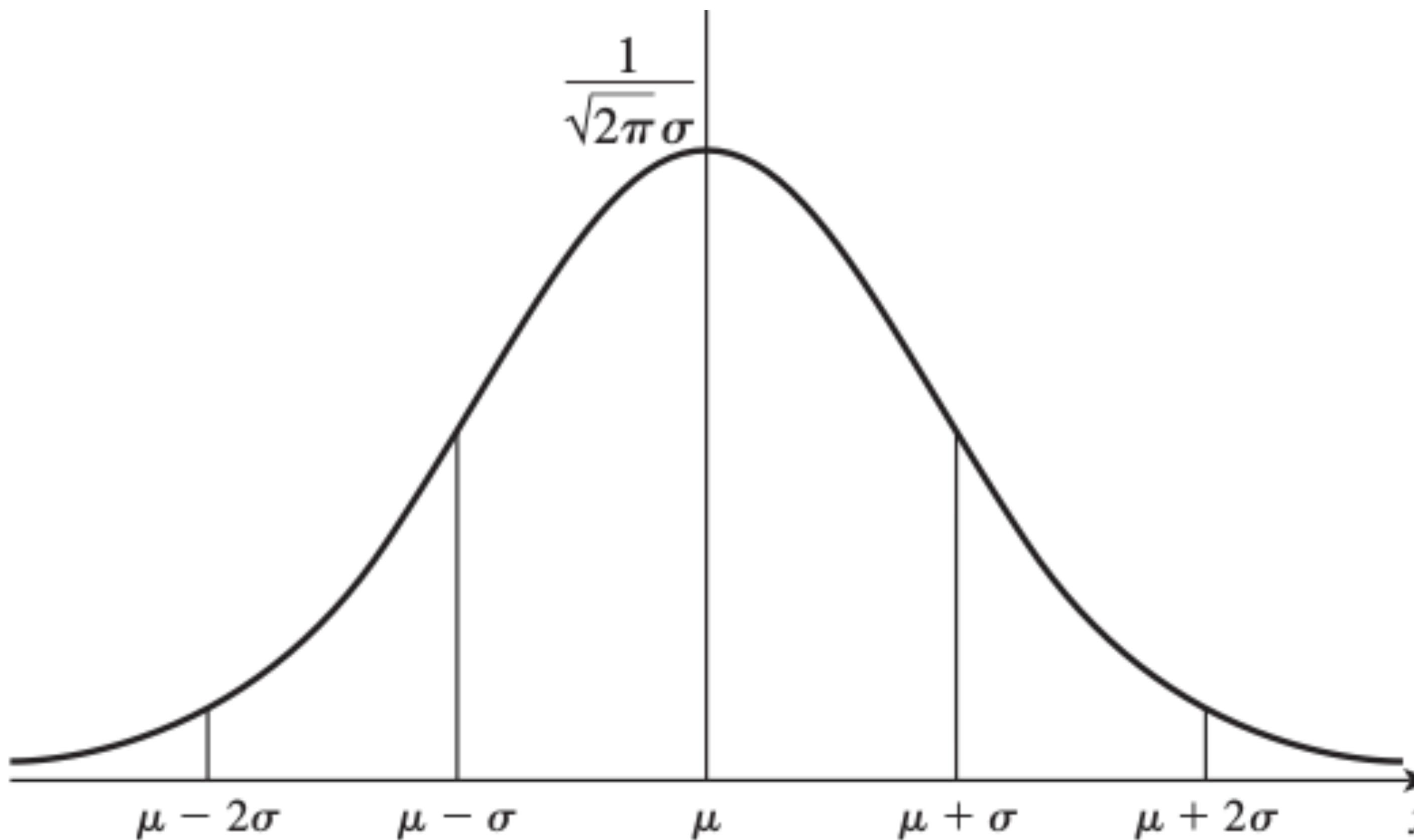
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Denoted with $X \sim N(\mu, \sigma^2)$

Most common distribution:

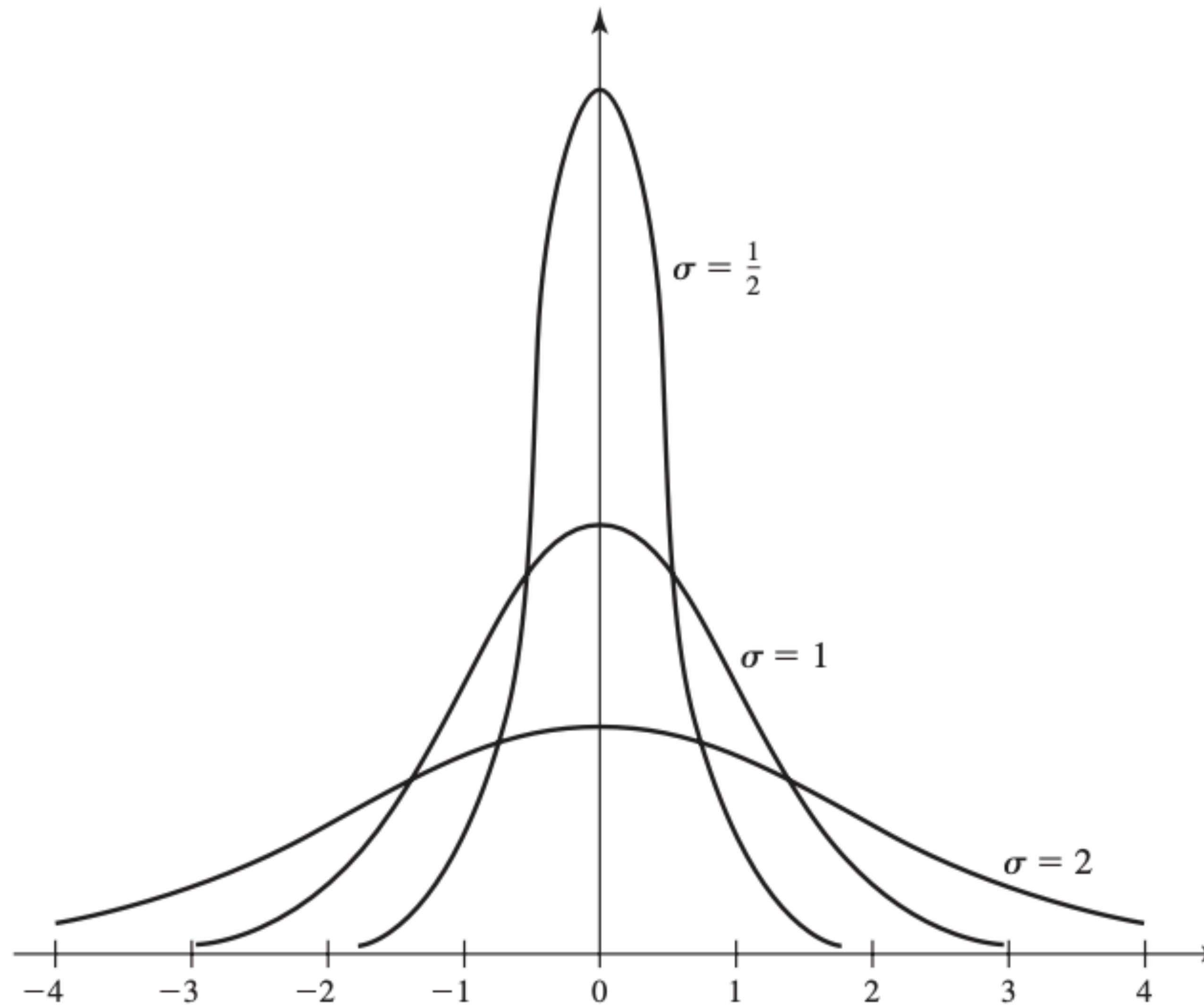
- Mathematical convenience
- Many physical phenomena approximately normal
- Central Limit Theorem: Mean of Sample from **any distribution** is approximately normal (More in a little while)

WIDELY USED CONTINUOUS DISTRIBUTIONS: NORMAL



Symmetric around mean

WIDELY USED CONTINUOUS DISTRIBUTIONS: NORMAL



As σ grows,
pdf gets more flat

WIDELY USED CONTINUOUS DISTRIBUTIONS: NORMAL

Linear Transformations of normal independent variables:

$$X_1 \sim N(\mu_1, \sigma_1^2) \text{ and}$$

$$X_2 \sim N(\mu_2, \sigma_2^2)$$

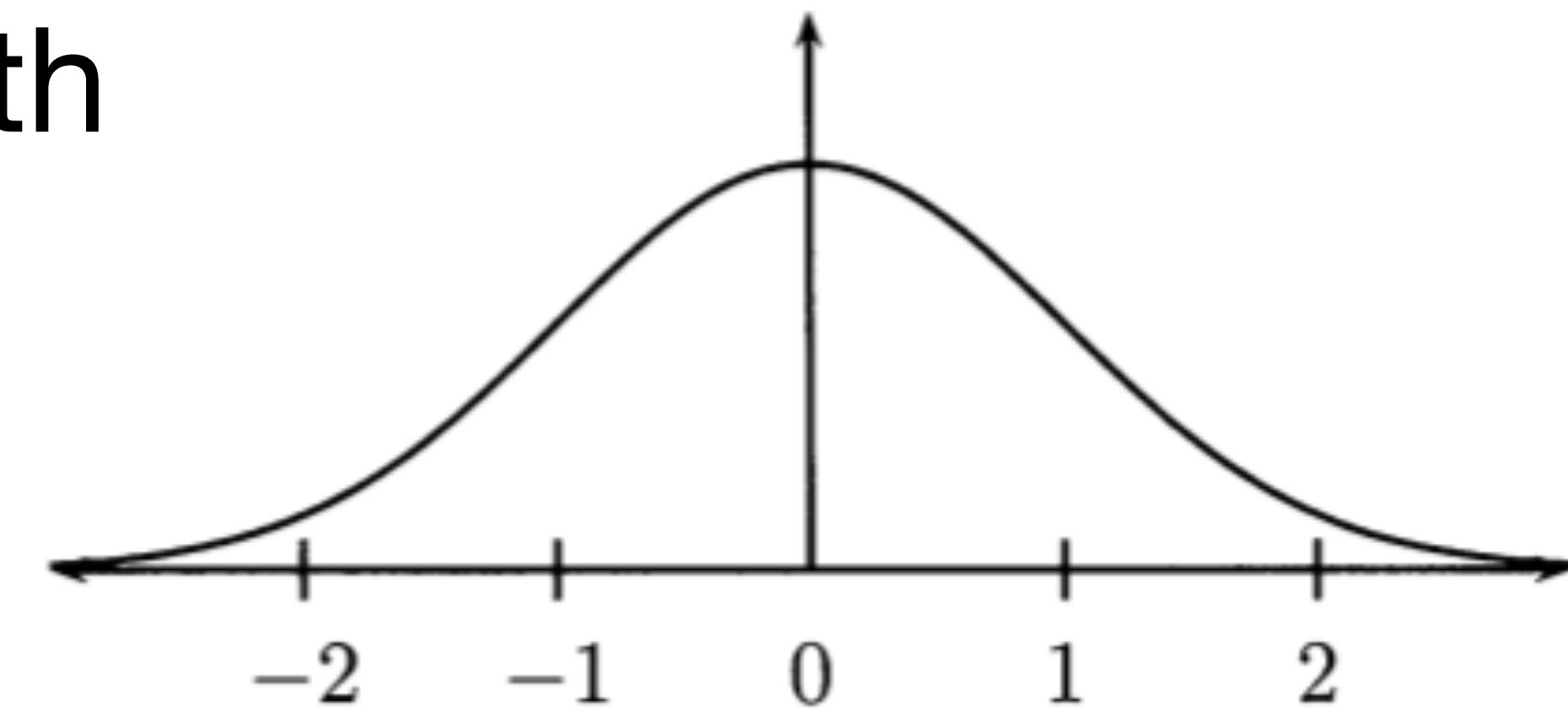
$$\Rightarrow Y = a_1 X_1 + a_2 X_2 + b \sim N(a_1 \mu_1 + a_2 \mu_2 + b, a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2)$$

(Can be generalized to $k > 2$ random variables)

WIDELY USED CONTINUOUS DISTRIBUTIONS: NORMAL

Standard Normal: Normal distribution with

$$\mu = 0, \sigma^2 = 1$$



- Denoted with $Z \sim N(0,1)$.
- Pdf denoted with $\phi(z)$. Cdf denoted with $\Phi(z)$.

WIDELY USED CONTINUOUS DISTRIBUTIONS: NORMAL

From normal to standard normal:

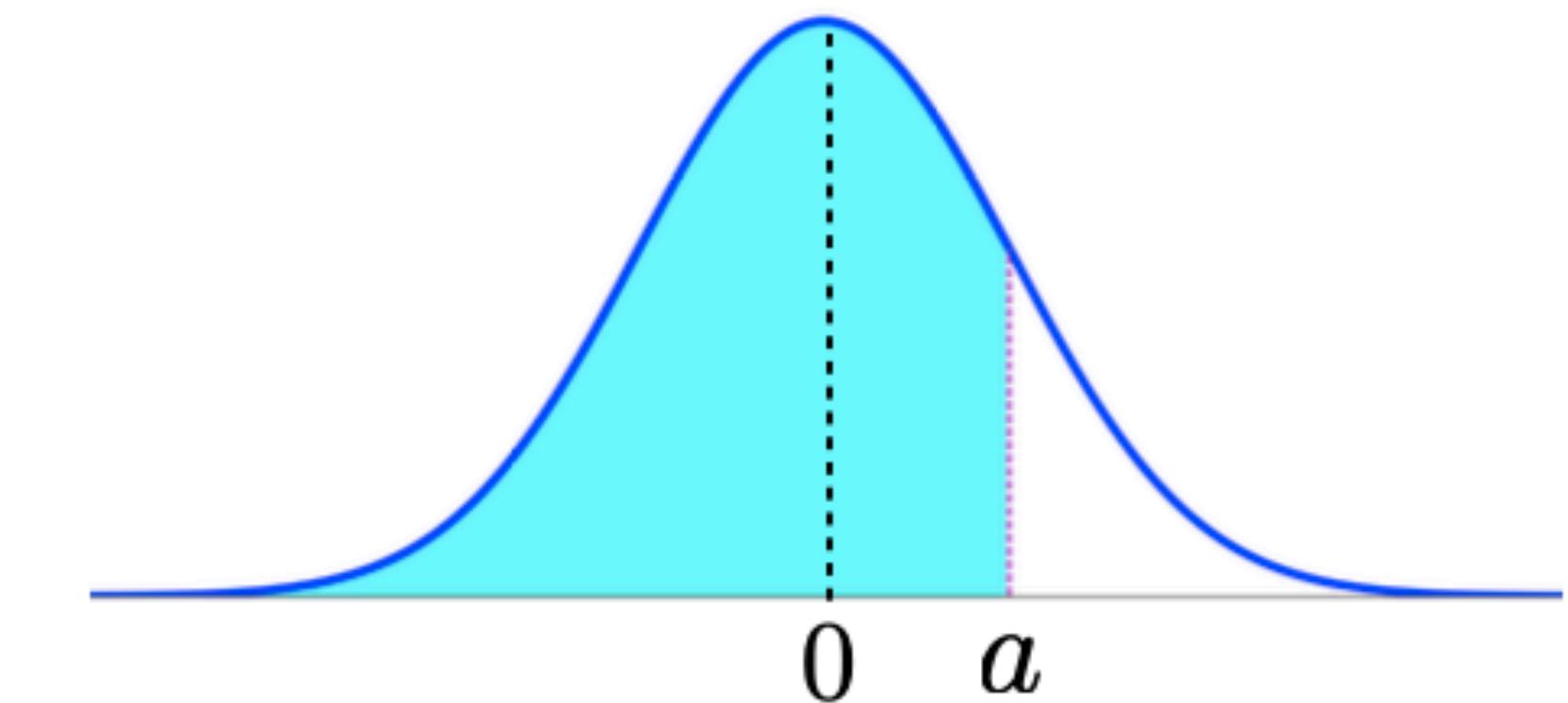
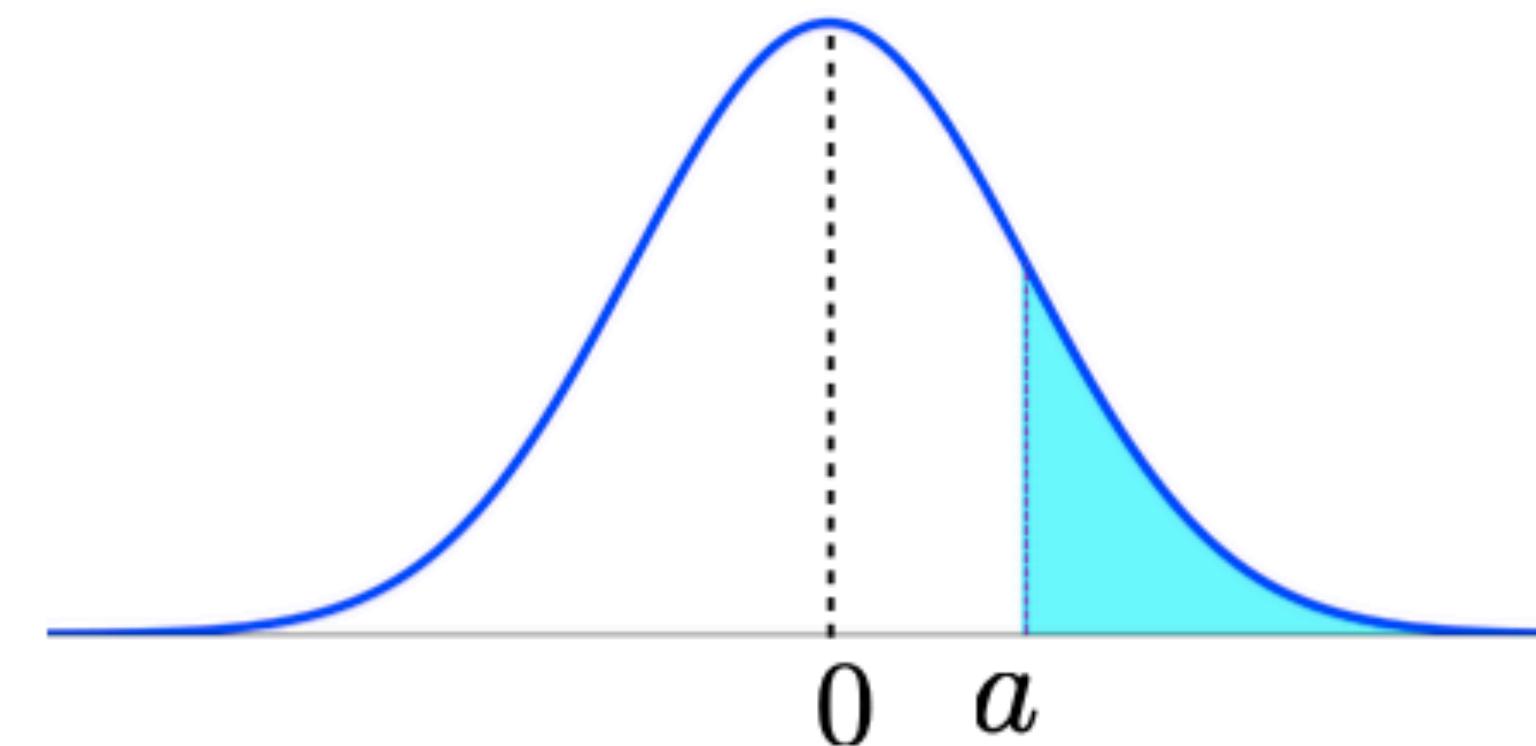
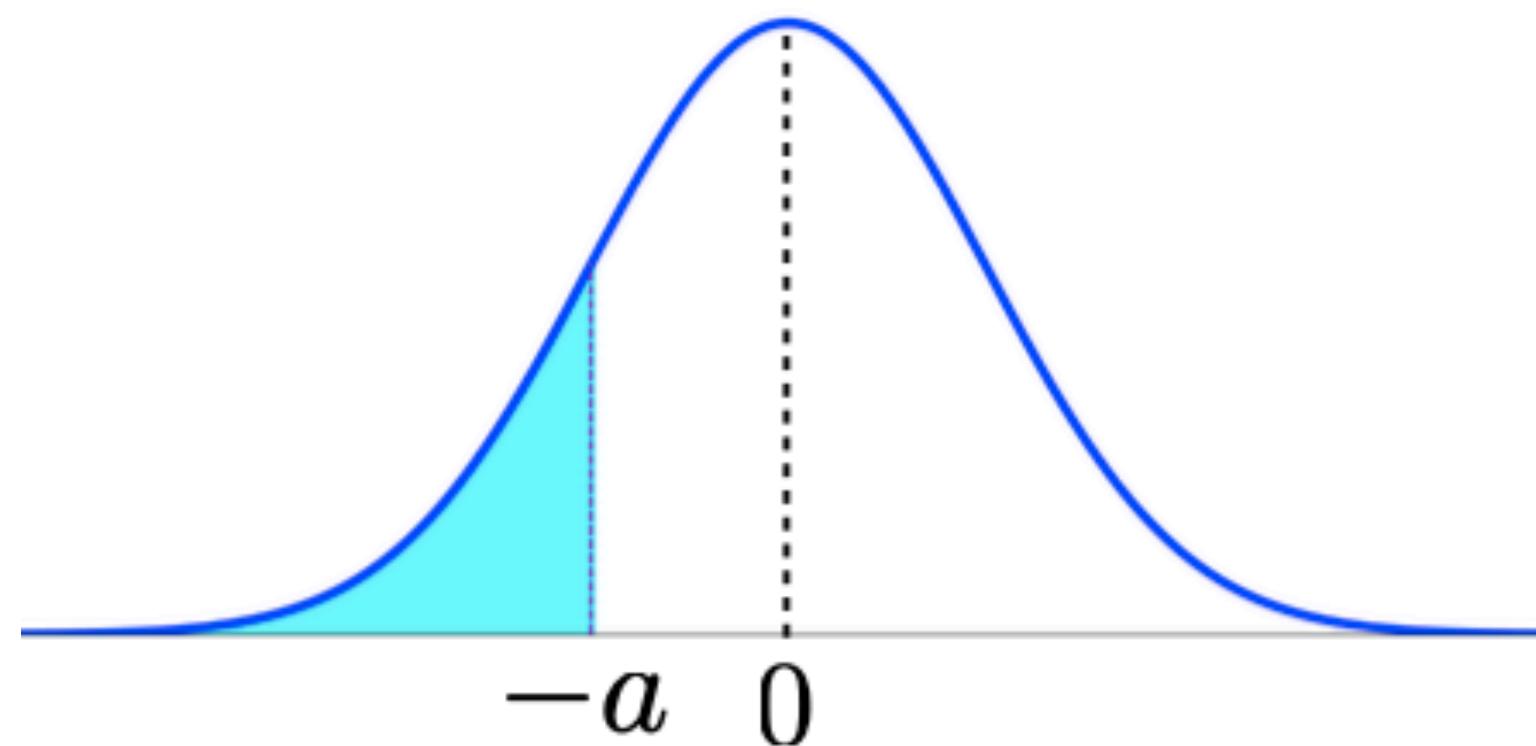
$$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

(Follows from linear transformations of normal independent RVs)

$$\begin{aligned}\Rightarrow P(a < X < b) &= P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)\end{aligned}$$

WIDELY USED CONTINUOUS DISTRIBUTIONS: NORMAL

Standard Normal: Consequence of symmetry



$$\begin{aligned}\Phi(-a) &= P(Z \leq -a) \\ &= P(Z \geq a) \\ &= 1 - P(Z \leq a) = 1 - \Phi(a)\end{aligned}$$

WIDELY USED CONTINUOUS DISTRIBUTIONS: NORMAL

Ex: $X \sim N(5,4)$ $\Rightarrow P(1 < X < 8) = ?$

$$\begin{aligned} P(1 < X < 8) &= P\left(\frac{1 - 5}{2} < \frac{X - 5}{2} = Z < \frac{8 - 5}{2}\right) \\ &= \Phi(1.5) - \Phi(-2) = \Phi(1.5) - (1 - \Phi(2)) \end{aligned}$$

WIDELY USED CONTINUOUS DISTRIBUTIONS: NORMAL

Ex: $X \sim N(5,4)$ $\Rightarrow P(1 < X < 8) = ?$

$$\begin{aligned} P(1 < X < 8) &= P\left(\frac{1-5}{2} < \frac{X-5}{2} = Z < \frac{8-5}{2}\right) \\ &= \Phi(1.5) - \Phi(-2) = \Phi(1.5) - (1 - \Phi(2)) \end{aligned}$$

```
from scipy.stats import norm
print(norm.cdf(1.5)-(1-norm.cdf(2)))
```

0.9104426667829627

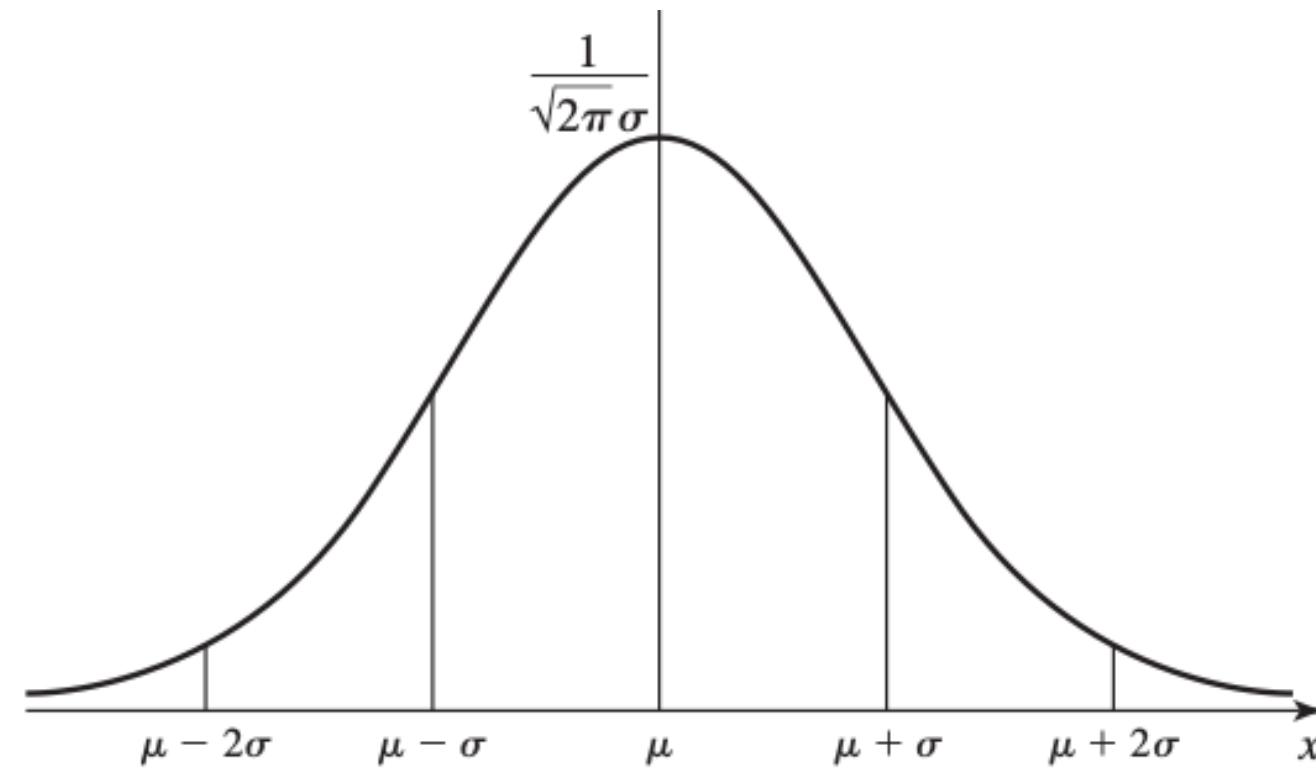
QUIZ

True or False:

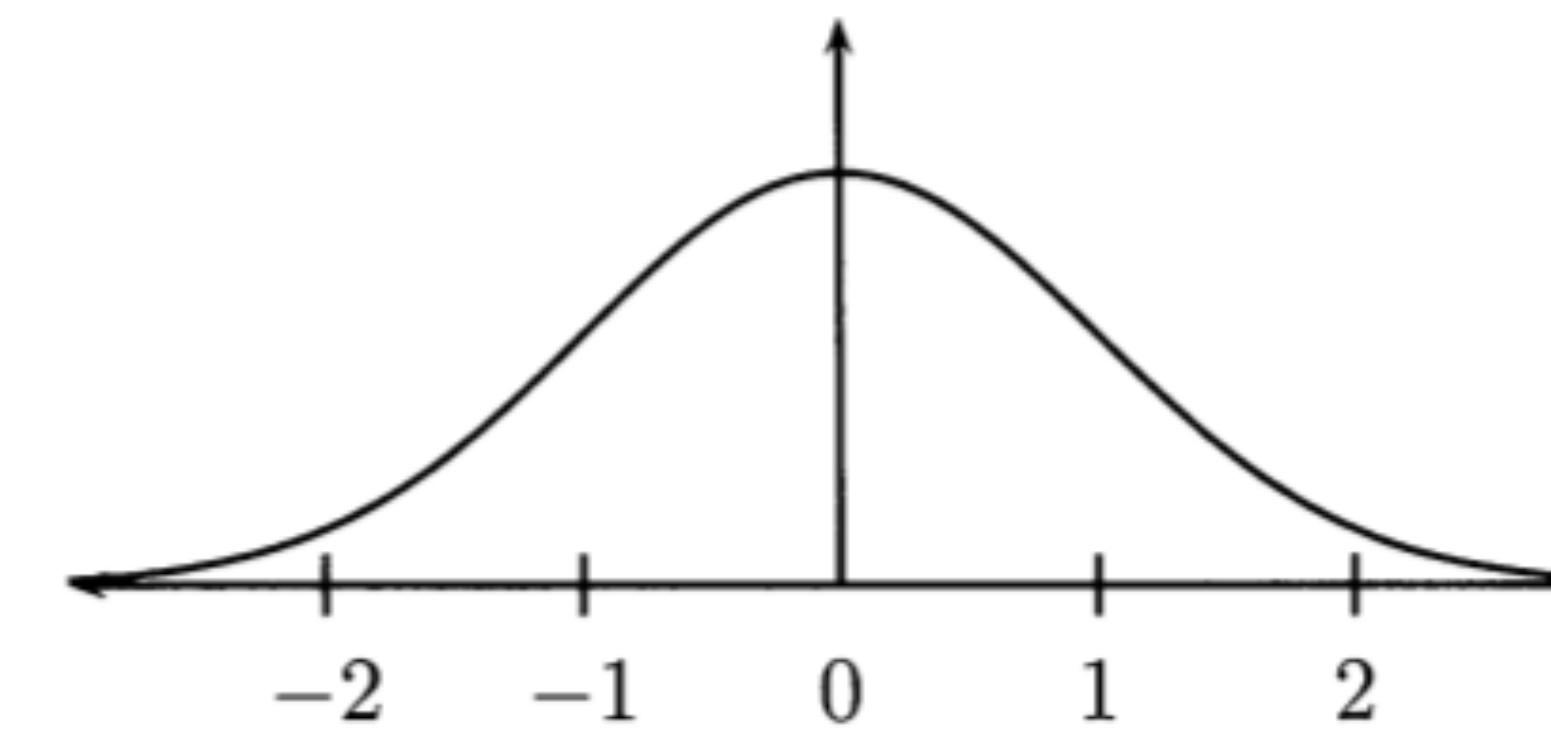
If two random variables are independent, then they must be uncorrelated.

Explain with one sentence.

REVIEW OF (LAST PART OF) THE PREVIOUS LECTURE



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

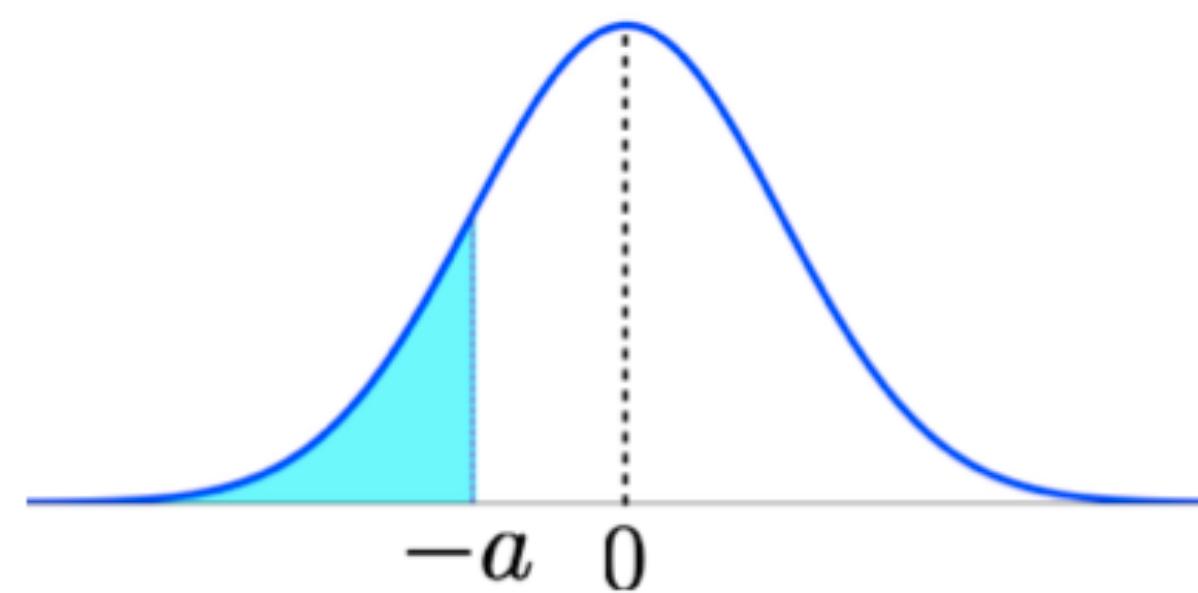


$$Z \sim N(0,1)$$

Normal to standard normal:

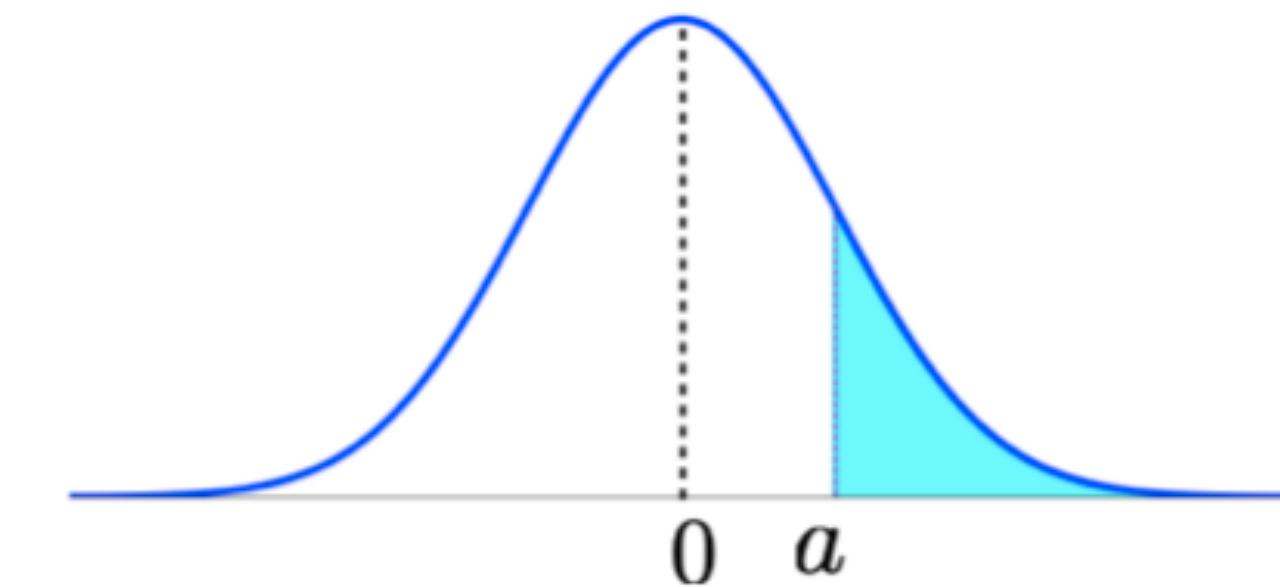
$$X \sim N(\mu, \sigma^2) \Rightarrow$$

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$



$$\Phi(-a)$$

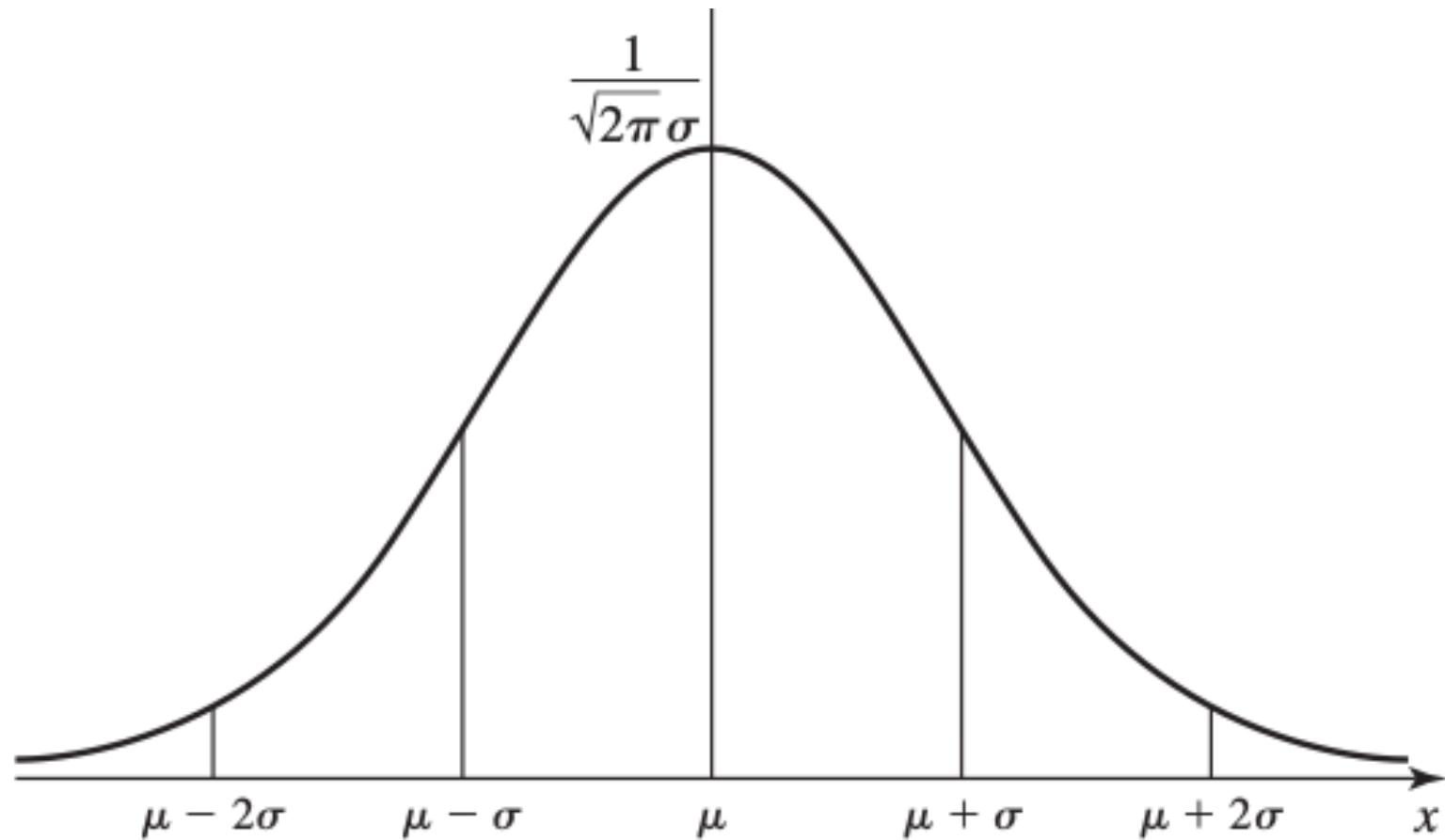
=



$$1 - \Phi(a)$$

WIDELY USED CONTINUOUS DISTRIBUTIONS: NORMAL

Easy to find: Probability that normal X within k std of mean.



k	p_k
1	0.6826
2	0.9544
3	0.9974
4	0.99994

p_k : Probability that normal RV
within $k \times \sigma$ of μ

How? $P(\mu - k\sigma \leq X \leq \mu + k\sigma) = P(-k \leq Z \leq k)$

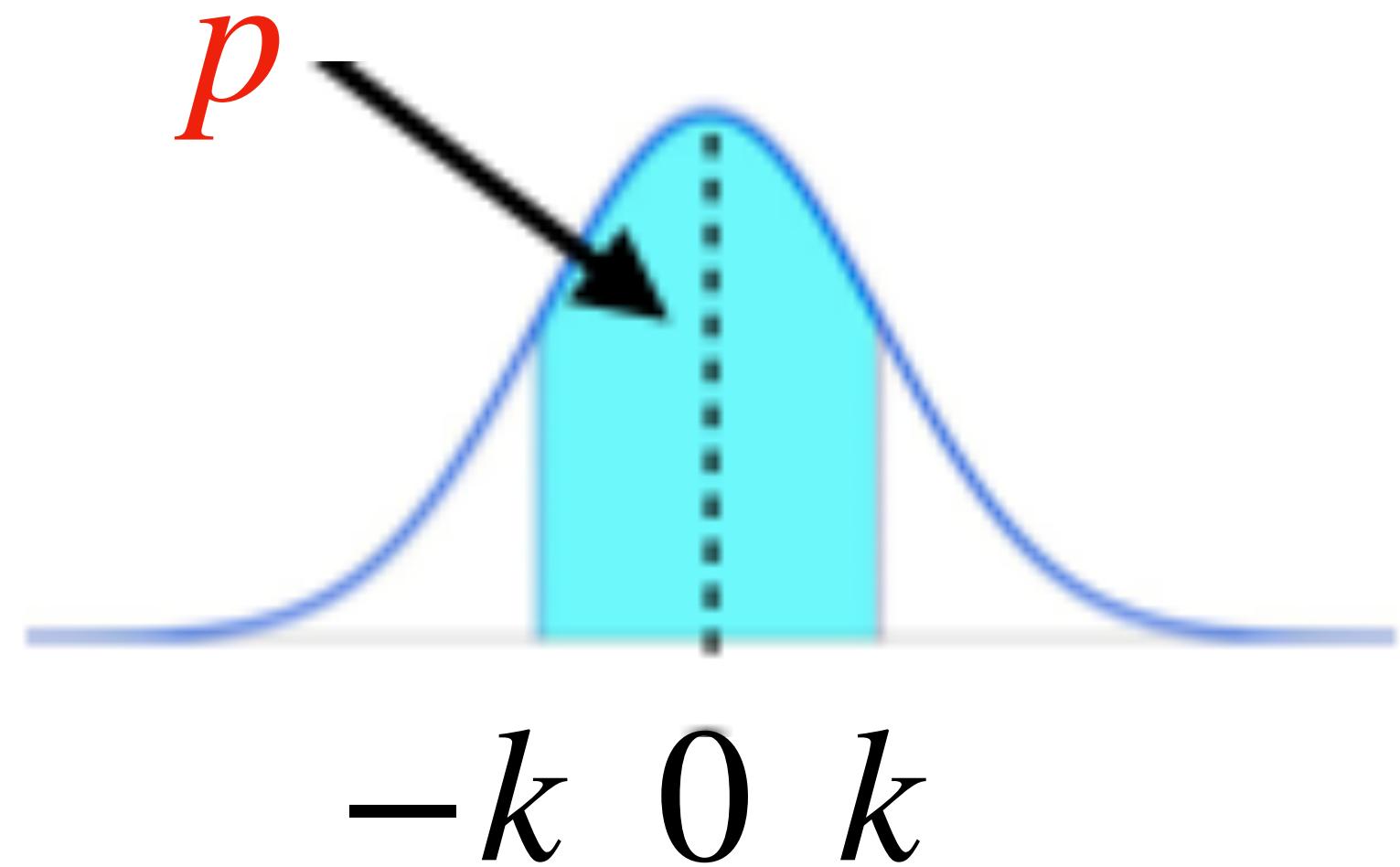
$$= 2\Phi(k) - 1$$

THE QUANTILE FUNCTION

Common problem:

Given desired probability p , find k s.t. $P(-k \leq Z \leq k) = p$

$$p = P(-k \leq Z \leq k) = 2\Phi(k) - 1$$



$$\Rightarrow \Phi(k) = \frac{p+1}{2} \text{ and } \Phi^{-1}\left(\frac{p+1}{2}\right) = k$$

THE QUANTILE FUNCTION

```
from scipy.stats import norm  
norm.ppf(0.9)
```

1.2815515655446004

$$\Phi^{-1}(p) : \text{norm.ppf}(p)$$

Percent point function

$$\Phi^{-1}(0.9) \approx 1.28 \Rightarrow \Phi(1.28) = P(Z \leq 1.28) \approx 0.9$$

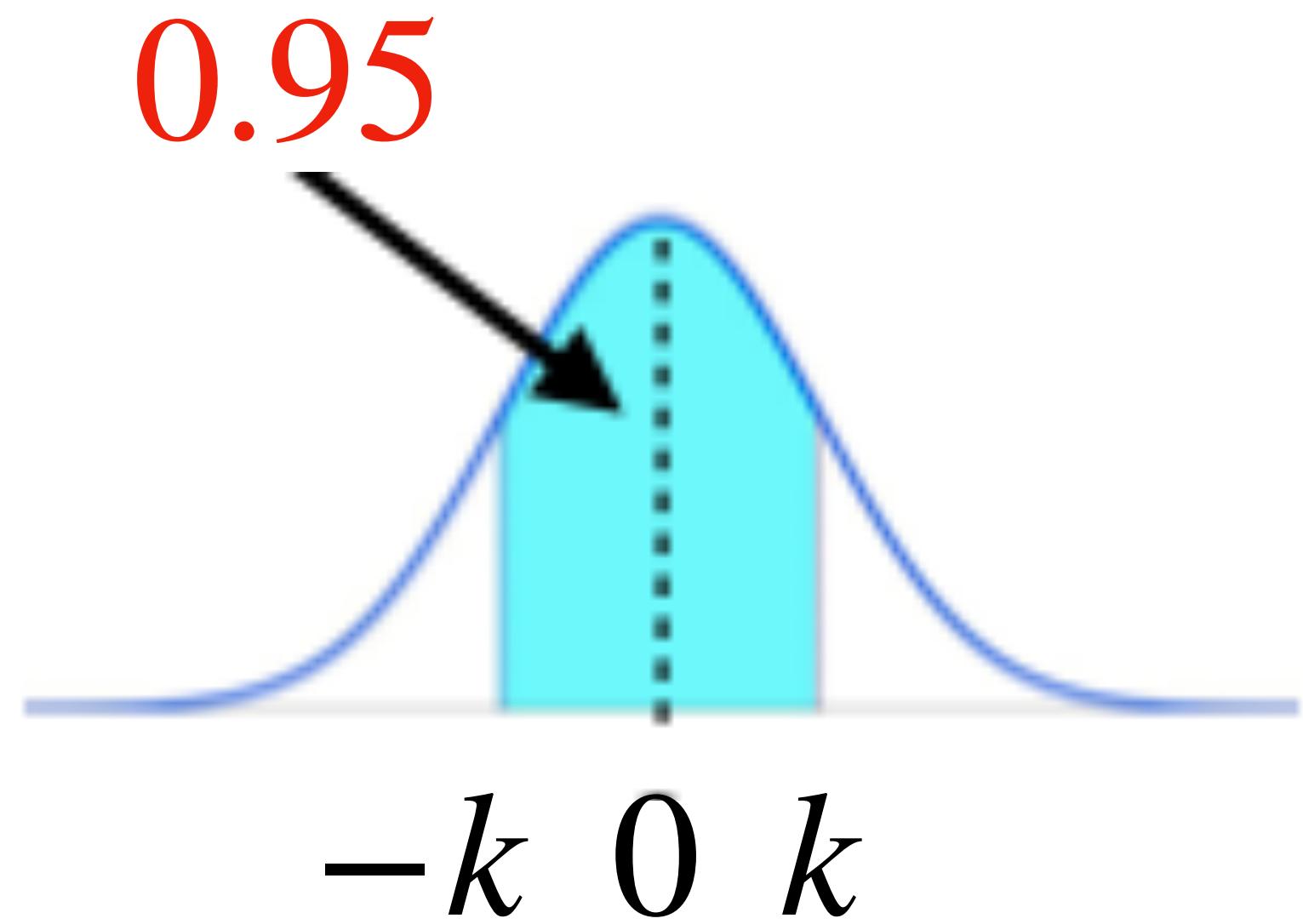
Can use the *ppf* of other distributions to compute their $F^{-1}(p)$.

THE QUANTILE FUNCTION

Example: Find k s.t. $P(-k \leq Z \leq k) = 0.95$

$$p = 0.95$$

$$k = \Phi^{-1}\left(\frac{p+1}{2}\right) = \Phi^{-1}(0.975)$$



```
from scipy.stats import norm  
norm.ppf(0.975)
```

1.959963984540054

THE QUANTILE FUNCTION

How to generate synthetic data from a distribution with cdf F ?

Inverse transform:

Repeatedly pick random p in $[0..1]$ and return $F^{-1}(p)$.

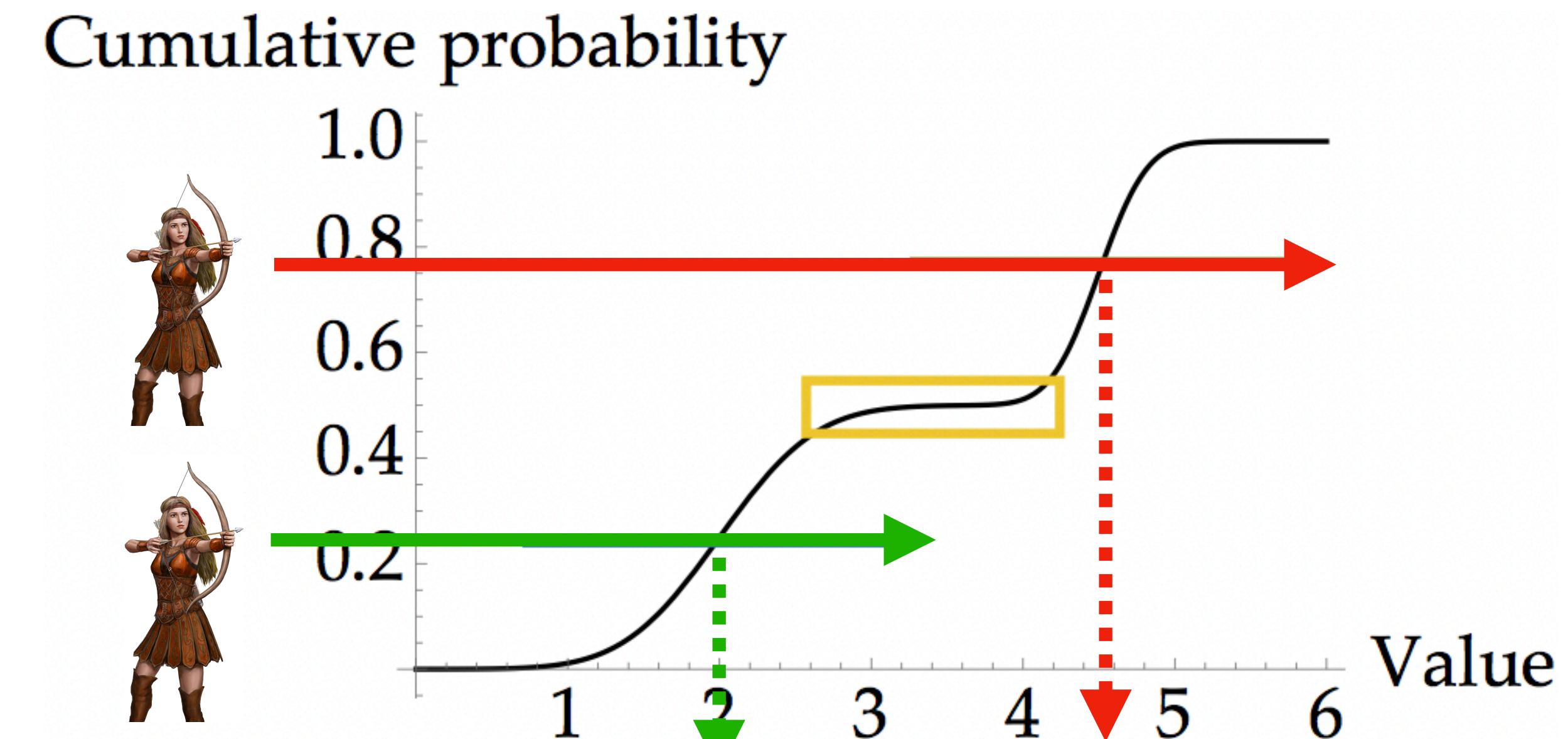
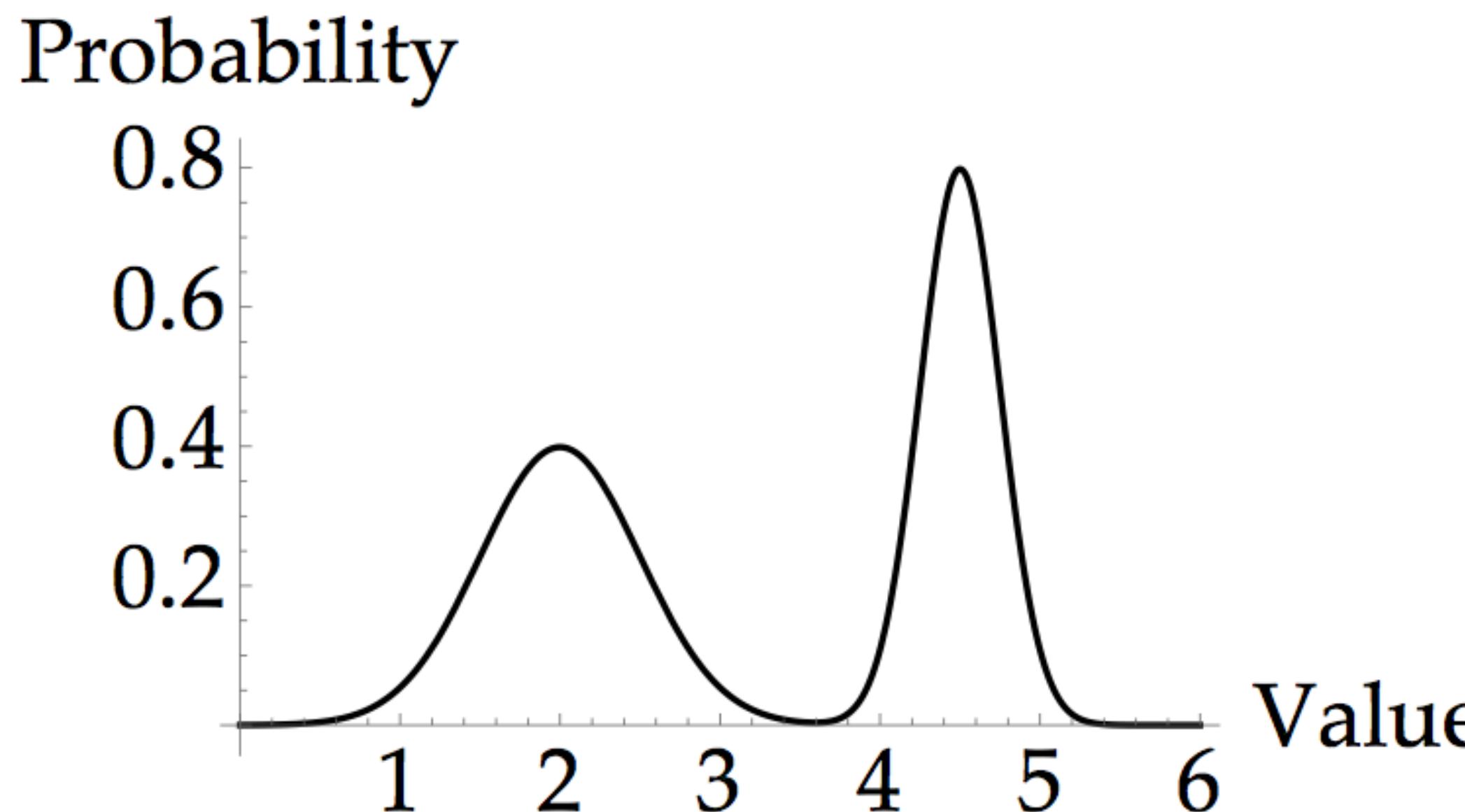
THE QUANTILE FUNCTION

How to generate synthetic data from a distribution with cdf F ?

Inverse transform:

Repeatedly pick random p in $[0..1]$ and return $F^{-1}(p)$.

Why does it work?



PROBABILITY INEQUALITIES

Markov's Inequality:

$$X \text{ a non-negative random variable} \Rightarrow \forall t > 0, P(X \geq t) \leq \frac{\mu}{t}$$

Proof:

X discrete random variable. Similar proof works for continuous.

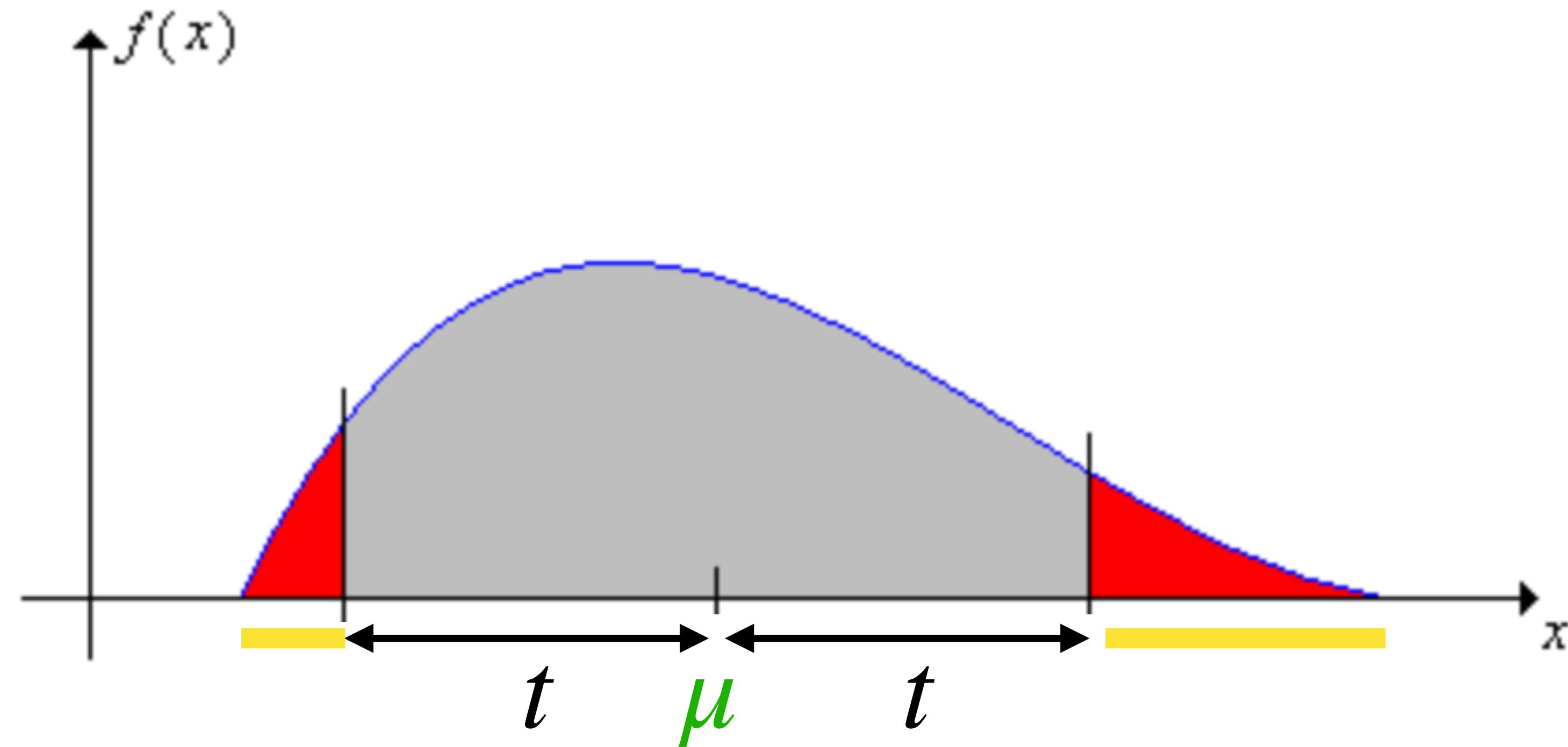
$$\mu = \sum_{\forall x} xf(x) \geq \sum_{x \geq t} xf(x) \geq \sum_{x \geq t} tf(x) = t \sum_{x \geq t} f(x) = tP(X \geq t)$$

Interesting only for $t > \mu$.

PROBABILITY INEQUALITIES

Chebyshev's Inequality:

$$X \text{ a random variable} \Rightarrow \forall t > 0, P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

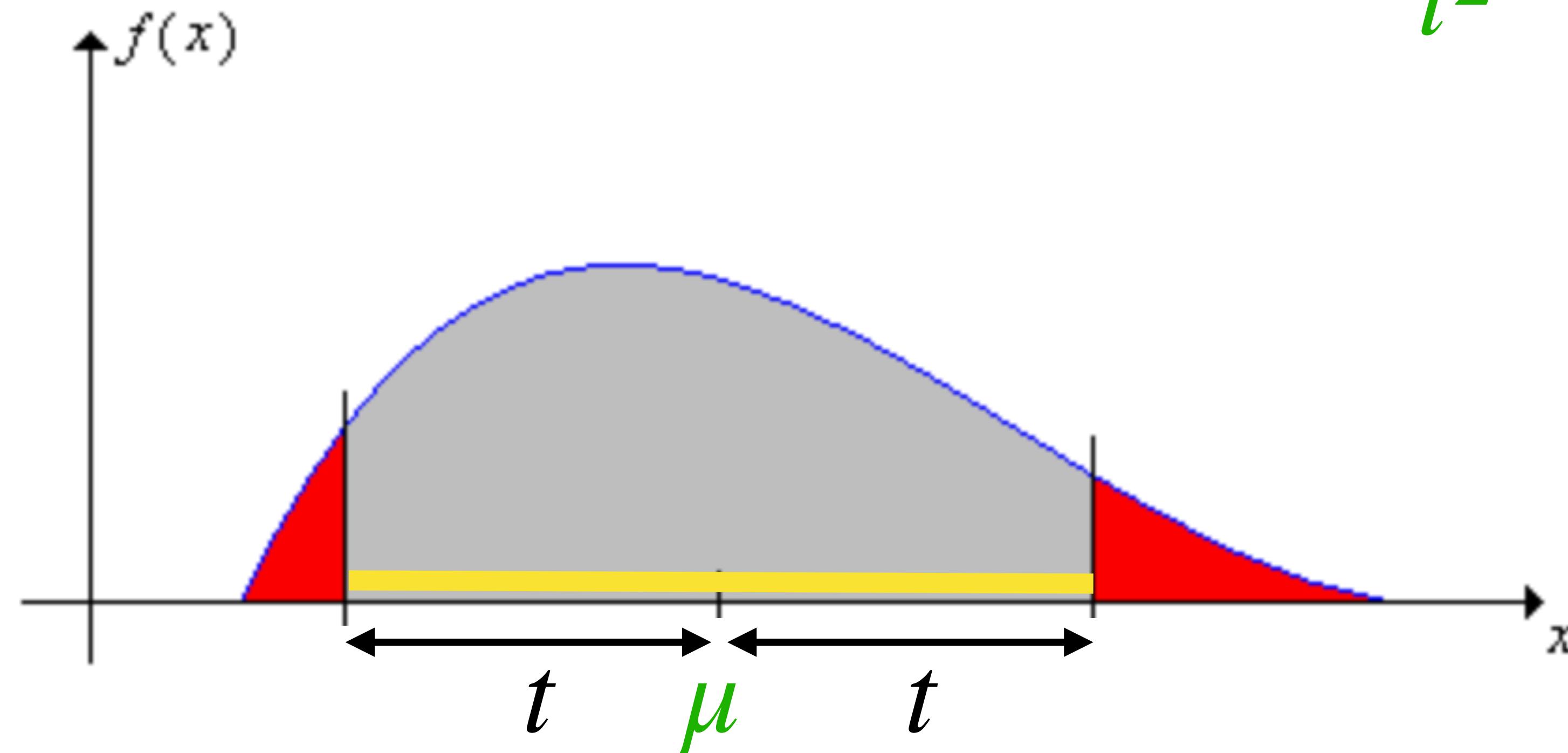


$$P(X \text{ lies in yellow interval}) \leq \frac{\sigma^2}{t^2} \quad (\text{area in red})$$

PROBABILITY INEQUALITIES

Chebyshev's Inequality: Another interpretation

$$P(|X - \mu| < t) \geq 1 - \frac{\sigma^2}{t^2}$$

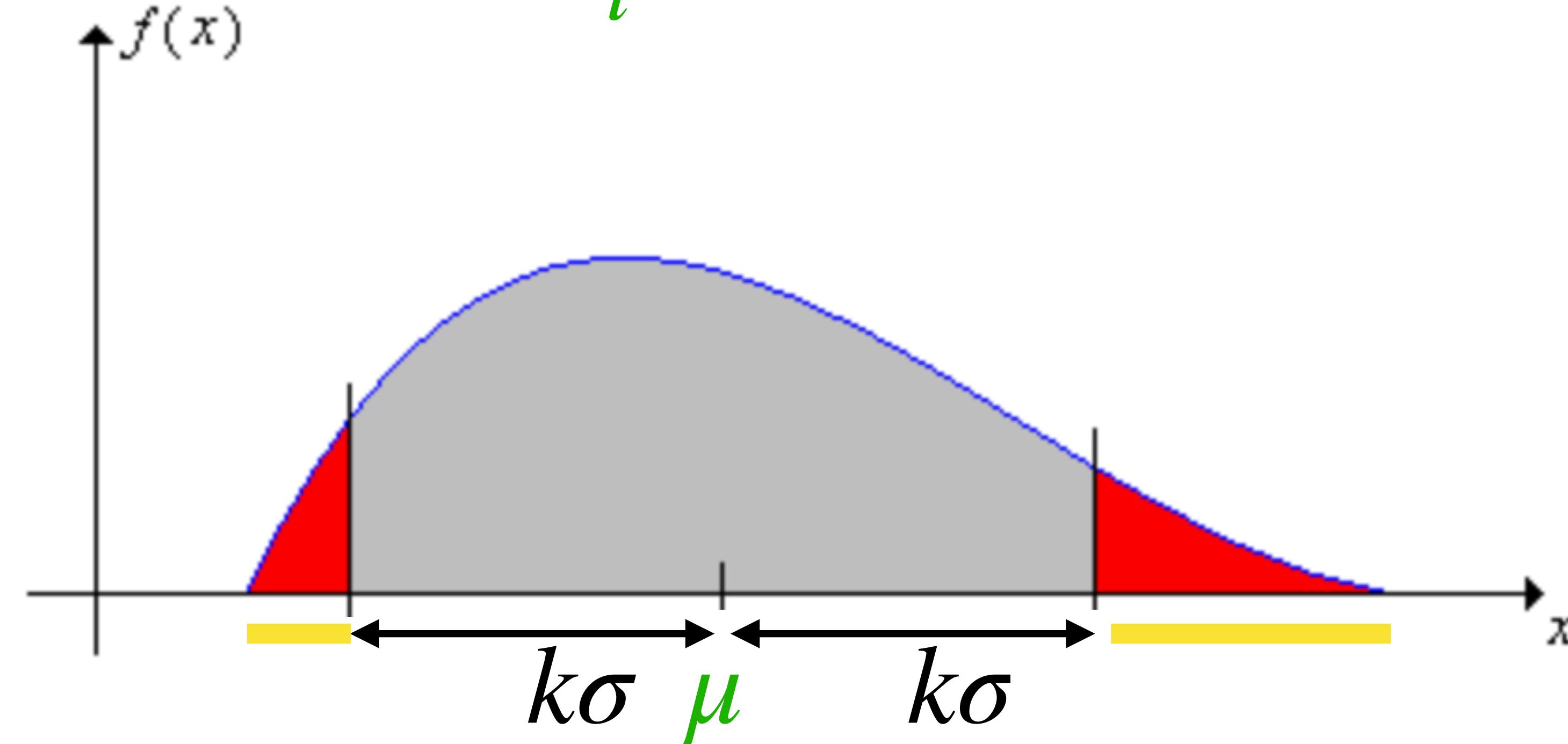


$$P(X \text{ lies in yellow interval}) \geq 1 - \frac{\sigma^2}{t^2} \quad (\text{area in gray})$$

PROBABILITY INEQUALITIES

Chebyshev's Inequality: Being at least k std away from mean

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \quad \Rightarrow \quad P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

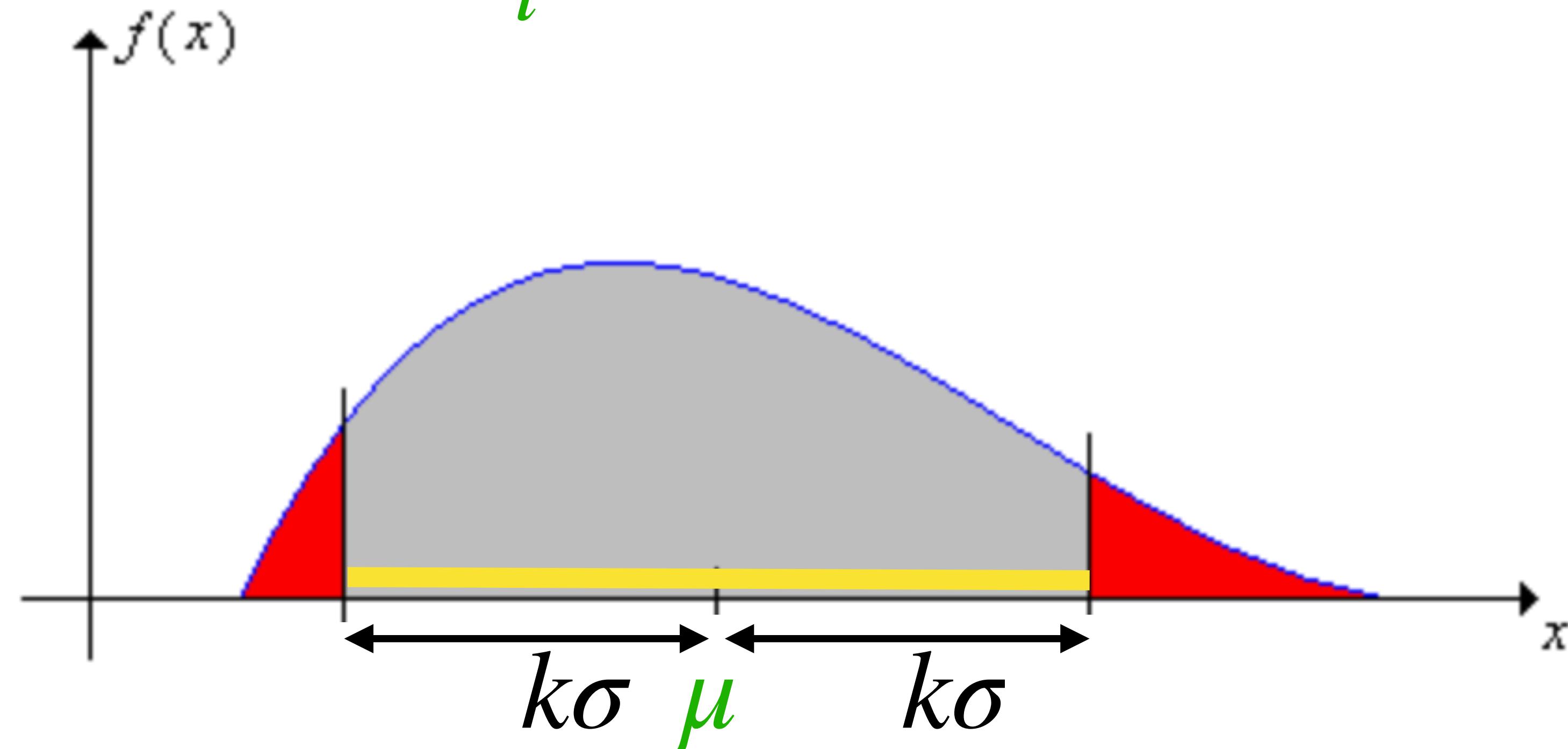


$$P(X \text{ lies in yellow interval}) \leq \frac{1}{k^2} \quad (\text{area in red})$$

PROBABILITY INEQUALITIES

Chebyshev's Inequality: Being within k std of the mean

$$P(|X - \mu| < t) \geq 1 - \frac{\sigma^2}{t^2} \quad \Rightarrow \quad P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$



$$P(X \text{ lies in yellow interval}) \geq 1 - \frac{1}{k^2} \quad (\text{area in gray})$$

PROBABILITY INEQUALITIES

Chebyshev's Inequality:

$$X \text{ a random variable} \Rightarrow \forall t > 0, P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

Proof:

$$P(|X - \mu| \geq t) = P((X - \mu)^2 \geq t^2) \leq \frac{E[(X - \mu)^2]}{t^2} = \frac{\sigma^2}{t^2}$$

PROBABILITY INEQUALITIES

Example: Let X denote the number of citations of a paper.

Let $\mu = 8$, $\sigma = 5$. What is $P(X \geq 58)$?

With Markov's Inequality:

$$P(X \geq 58) \leq \frac{8}{58} \approx 0.14$$

With Chebyshev's Inequality:

$$P(X \geq 58) = P(X - \mu \geq 50) \leq P(|X - \mu| \geq 50) \leq \left(\frac{\sigma}{50}\right)^2 = 0.01$$

SAMPLE MEAN

X_1, \dots, X_n iid with mean μ and variance σ^2 . Sample mean \bar{X}_n

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Note that \bar{X}_n is a random variable.

$$E(\bar{X}_n) = E\left(\frac{1}{n} \sum X_i\right) = \frac{1}{n} \sum E(X_i) = \frac{\mu n}{n} = \mu$$

$$Var(\bar{X}_n) = Var\left(\frac{1}{n} \sum X_i\right) = \frac{1}{n^2} \sum Var(X_i) = \frac{\sigma^2 n}{n^2} = \frac{\sigma^2}{n}$$

SAMPLE MEAN

Example: Poll 100,000 people. Assume each votes for candidate C independently with probability p .

Bound the probability that the poll is off by more than 1 % .

$$X_i = 1 \text{ if } i \text{ votes for } C, X_i = 0 \text{ o.w.} \quad X_i \sim \text{Bernoulli}(p)$$

First find a bound for σ^2 . What is max possible value for it?

$$\mu = p, \sigma^2 \leq 1/4 \quad \text{since } p(1 - p) \text{ is maximum when } p = 1/2$$

[find p where derivative of $p(1 - p)$ is 0]

REVIEW OF (LAST PART OF) THE PREVIOUS LECTURE

Markov's Inequality:

$$X \text{ a non-negative random variable} \Rightarrow \forall t > 0, P(X \geq t) \leq \frac{\mu}{t}$$

REVIEW OF (LAST PART OF) THE PREVIOUS LECTURE

Markov's Inequality:

$$X \text{ a non-negative random variable} \Rightarrow \forall t > 0, P(X \geq t) \leq \frac{\mu}{t}$$

Chebyshev's Inequality:

$$X \text{ a random variable} \Rightarrow \forall t > 0, P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

REVIEW OF (LAST PART OF) THE PREVIOUS LECTURE

Markov's Inequality:

$$X \text{ a non-negative random variable} \Rightarrow \forall t > 0, P(X \geq t) \leq \frac{\mu}{t}$$

Chebyshev's Inequality:

$$X \text{ a random variable} \Rightarrow \forall t > 0, P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

Sample Mean:

$$X_1, \dots, X_n \text{ iid with mean } \mu \text{ and variance } \sigma^2. \quad \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$
$$E(\bar{X}_n) = \mu \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

REVIEW OF (LAST PART OF) THE PREVIOUS LECTURE

Markov's Inequality:

X a non-negative random variable $\Rightarrow \forall t > 0, P(X \geq t) \leq \frac{\mu}{t}$

Chebychev's Inequality:

X a random variable $\Rightarrow \forall t > 0, P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$

Sample Mean:

X_1, \dots, X_n iid with mean μ and variance σ^2 . $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$

$$E(\bar{X}_n) = \mu \quad Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

Chebychev's Applied on \bar{X}_n : $P(|\bar{X}_n - \mu| \geq t) \leq \frac{\sigma^2}{nt^2}$

READINGS ON THE LAST LECTURE AND THE NEXT FEW LECTURES

10.1, 11.2, 11.2, 11.5, Ch 12, 14.1, 14.2, 14.3, 15.1-15.5

[Watkins]

4.1 (Markov's, Chebyshev's Inequalities),

5.1, 5.3, 5.4, Ch 6, 9.1, 9.3-9.5

[Wasserman]

SAMPLE MEAN

Example: Poll 100,000 people. Assume each votes for candidate C independently with probability p .

Bound the probability that the poll is off by more than 1 % .

SAMPLE MEAN

Example: Poll 100,000 people. Assume each votes for candidate C independently with probability p .

Bound the probability that the poll is off by more than 1 % .

$X_i = 1$ if i votes for C , $X_i = 0$ o.w. $X_i \sim \text{Bernoulli}(p)$

$$P(|\bar{X}_n - \mu| \geq 0.01) = ?$$

SAMPLE MEAN

Example: Poll 100,000 people. Assume each votes for candidate C independently with probability p .

Bound the probability that the poll is off by more than 1 % .

$X_i = 1$ if i votes for C , $X_i = 0$ o.w. $X_i \sim \text{Bernoulli}(p)$

$$P(|\bar{X}_n - \mu| \geq 0.01) = ?$$

$\mu = p$, $\sigma^2 \leq 1/4$ since $p(1 - p)$ is maximum when $p = 1/2$

SAMPLE MEAN

Example: Poll 100,000 people. Assume each votes for candidate C independently with probability p .

Bound the probability that the poll is off by more than 1 % .

$X_i = 1$ if i votes for C , $X_i = 0$ o.w. $X_i \sim \text{Bernoulli}(p)$

$$P(|\bar{X}_n - \mu| \geq 0.01) = ?$$

$\mu = p$, $\sigma^2 \leq 1/4$ since $p(1 - p)$ is maximum when $p = 1/2$

\bar{X}_n has mean p , variance $\leq \frac{1/4}{100,000}$. Chebyshev's Inequality on \bar{X}_n :

$$P(|\bar{X}_n - p| \geq 0.01) \leq \frac{1/4}{0.01^2 \times 100,000} = 2.5 \%$$

SAMPLE MEAN

Example: Want to take a random sample from a distribution with unknown μ and variance σ^2 is 8.

Size of sample s.t. \bar{X}_n within 2 units of μ with ≥ 0.99 probability?

SAMPLE MEAN

Example: Want to take a random sample from a distribution with unknown μ and variance σ^2 is 8.

Size of sample s.t. \bar{X}_n within 2 units of μ with ≥ 0.99 probability?

\bar{X}_n has mean μ and variance $\frac{8}{n}$. Chebychev's Inequality on \bar{X}_n :

$$P(|\bar{X}_n - \mu| < 2) \geq 1 - \frac{8}{4n} = 1 - \frac{2}{n}$$

SAMPLE MEAN

Example: Want to take a random sample from a distribution with unknown μ and variance σ^2 is 8.

Size of sample s.t. \bar{X}_n within 2 units of μ with ≥ 0.99 probability?

\bar{X}_n has mean μ and variance $\frac{8}{n}$. Chebychev's Inequality on \bar{X}_n :

$$P(|\bar{X}_n - \mu| < 2) \geq 1 - \frac{8}{4n} = 1 - \frac{2}{n}$$

$$1 - \frac{2}{n} = 0.99 \Rightarrow 0.01n = 2 \Rightarrow n \text{ at least } 200.$$

WEAK LAW OF LARGE NUMBERS

Informally:

As number of samples increases,

Sample mean converges to distribution mean (in probability).

WEAK LAW OF LARGE NUMBERS

Informally:

As number of samples increases,

Sample mean converges to distribution mean (in probability).

Formally: Let X_1, \dots, X_n iid with mean μ , standard deviation σ

$$P(|\bar{X}_n - \mu| < t) \geq 1 - \frac{\sigma^2}{nt^2}, \text{ for any } t > 0$$

$$\Rightarrow \lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < t) = 1$$

CENTRAL LIMIT THEOREM

Informally:

For a large random sample, sample mean has an approximately normal distribution with mean μ , variance σ^2/n .

Note: Nothing is assumed regarding distribution of X_i .

CENTRAL LIMIT THEOREM

Informally:

For a large random sample, sample mean has an approximately normal distribution with mean μ , variance σ^2/n .

Note: Nothing is assumed regarding distribution of X_i .

Formally: Let X_1, \dots, X_n iid with mean μ , standard deviation σ ,

As $n \rightarrow \infty$

$$Z_n = \frac{\bar{X}_n - \mu}{\sqrt{Var(\bar{X}_n)}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0, 1)$$

CENTRAL LIMIT THEOREM

Equivalent statements:

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

CENTRAL LIMIT THEOREM

Equivalent statements:

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\bar{X}_n - \mu \sim N\left(0, \frac{\sigma^2}{n}\right)$$

CENTRAL LIMIT THEOREM

Equivalent statements:

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\bar{X}_n - \mu \sim N\left(0, \frac{\sigma^2}{n}\right)$$

$$\sqrt{n}(\bar{X}_n - \mu) \sim N(0, \sigma^2)$$

CENTRAL LIMIT THEOREM

Equivalent statements:

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\bar{X}_n - \mu \sim N\left(0, \frac{\sigma^2}{n}\right)$$

$$\sqrt{n}(\bar{X}_n - \mu) \sim N(0, \sigma^2)$$

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0, 1)$$

CENTRAL LIMIT THEOREM

Example: # of errors per program has Poisson distribution with mean 4. Let X_1, \dots, X_{100} be the numbers of errors in 100 programs. How to approximate $P(\bar{X}_n < 4.2)$?

CENTRAL LIMIT THEOREM

Example: # of errors per program has Poisson distribution with mean 4. Let X_1, \dots, X_{100} be the numbers of errors in 100 programs. How to approximate $P(\bar{X}_n < 4.2)$?

$$\lambda = 4 \text{ is both mean and variance of } X_i. \quad \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0, 1)$$

CENTRAL LIMIT THEOREM

Example: # of errors per program has Poisson distribution with mean 4. Let X_1, \dots, X_{100} be the numbers of errors in 100 programs. How to approximate $P(\bar{X}_n < 4.2)$?

$\lambda = 4$ is both mean and variance of X_i . $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0,1)$

$$\Rightarrow P(\bar{X}_n < 4.2) = P\left(\frac{\sqrt{100}(\bar{X}_n - 4)}{\sqrt{4}} < \frac{\sqrt{100}(4.2 - 4)}{\sqrt{4}}\right)$$

$$\approx P(Z < 1) = 0.84$$

CENTRAL LIMIT THEOREM

Example: X_i : customer spending with $\mu = 80, \sigma = 40$

Approximate the probability that the average spending of 100 customers is 10 % or more below average.

CENTRAL LIMIT THEOREM

Example: X_i : customer spending with $\mu = 80, \sigma = 40$

Approximate the probability that the average spending of 100 customers is 10 % or more below average.

$$P(\bar{X}_n \leq 72) = ?$$

CENTRAL LIMIT THEOREM

Example: X_i : customer spending with $\mu = 80, \sigma = 40$

Approximate the probability that the average spending of 100 customers is 10 % or more below average.

$$P(\bar{X}_n \leq 72) = ?$$

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0,1)$$

CENTRAL LIMIT THEOREM

Example: X_i : customer spending with $\mu = 80, \sigma = 40$

Approximate the probability that the average spending of 100 customers is 10 % or more below average.

$$P(\bar{X}_n \leq 72) = ?$$

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0,1)$$

$$\begin{aligned} \Rightarrow P(\bar{X}_n \leq 72) &= P\left(\frac{\sqrt{100}(\bar{X}_n - 80)}{40} \leq \frac{\sqrt{100}(72 - 80)}{40}\right) \\ &\approx P(Z \leq -2) \approx 0.023 \end{aligned}$$

PROBABILITY AND STATISTICS

Probability:

Distribution → Samples

Given distribution find probabilities of data/events.

Ex: $X \sim \text{Hypergeometric}(r = 70, b = 30, n = 10)$

The probability of a random sample of size n having 5 red balls?

PROBABILITY AND STATISTICS

Probability:

Distribution → Samples

Given distribution find probabilities of data/events.

Ex: $X \sim \text{Hypergeometric}(r = 70, b = 30, n = 10)$

The probability of a random sample of size n having 5 red balls?

Statistics:

Sample → Distribution

Given data find parameters/properties of distribution.

Ex: We observed $X_1 = 0, X_2 = 1, \dots, X_{10} = 0$

What is the distribution parameter p , i.e. probability of heads?

POINT ESTIMATION

Find single “good estimate” of a quantity of interest of a distribution/population using statistics.

Statistics: Any function of the sample. avg, max, max-min, ⋯

POINT ESTIMATION

Find single “good estimate” of a quantity of interest of a distribution/population using statistics.

Statistics: Any function of the sample. avg, max, max-min, ...

Formally, X_1, \dots, X_n iid data points from some distribution.

An **estimate** of parameter θ of the distribution: $\hat{\theta}_n = r(X_1, \dots, X_n)$
for some appropriate function r .

Note: θ : fixed, unknown quantity. $\hat{\theta}_n$: random variable.

POINT ESTIMATION

Ex: Estimate $\theta = \mu = \sum x f(x)$ of an unknown distribution.

Say true (unknown) value $\theta = 3.5$.

Sample 4 data points X_1, X_2, X_3, X_4 , say 3,6,5, - 2.

POINT ESTIMATION

Ex: Estimate $\theta = \mu = \sum x f(x)$ of an unknown distribution.

Say true (unknown) value $\theta = 3.5$.

Sample 4 data points X_1, X_2, X_3, X_4 , say 3,6,5, - 2.

Can try to estimate θ with **any** function of X_1, \dots, X_4 :

$$\hat{\theta}_n: \frac{X_1 + \dots + X_n}{n} \quad \frac{\min(X_1, \dots, X_n) + \max(X_1, \dots, X_n)}{2} \quad X_1 \cdot X_n$$

PLUG-IN ESTIMATOR

Property of distribution: θ

Min: $\min_{\forall x} \{x : f(x) > 0\}$ 

Plug-in sample: $\hat{\theta}_n$

Sample min: $\min(X_1, \dots, X_n)$

PLUG-IN ESTIMATOR

Property of distribution: θ

Min: $\min_{\forall x} \{x : f(x) > 0\}$ 

Plug-in sample: $\hat{\theta}_n$

Mean: $\mu = \sum_x x f(x)$ 

Sample mean: $\frac{1}{n} \sum_{i=1}^n X_i$

Sample min: $\min(X_1, \dots, X_n)$

PLUG-IN ESTIMATOR

Property of distribution: θ

Min: $\min_{\forall x} \{x : f(x) > 0\}$ \rightarrow

Plug-in sample: $\hat{\theta}_n$

Mean: $\mu = \sum_x x f(x)$ \rightarrow

Sample mean: $\frac{1}{n} \sum_{i=1}^n X_i$

Correlation:

Sample Correlation:

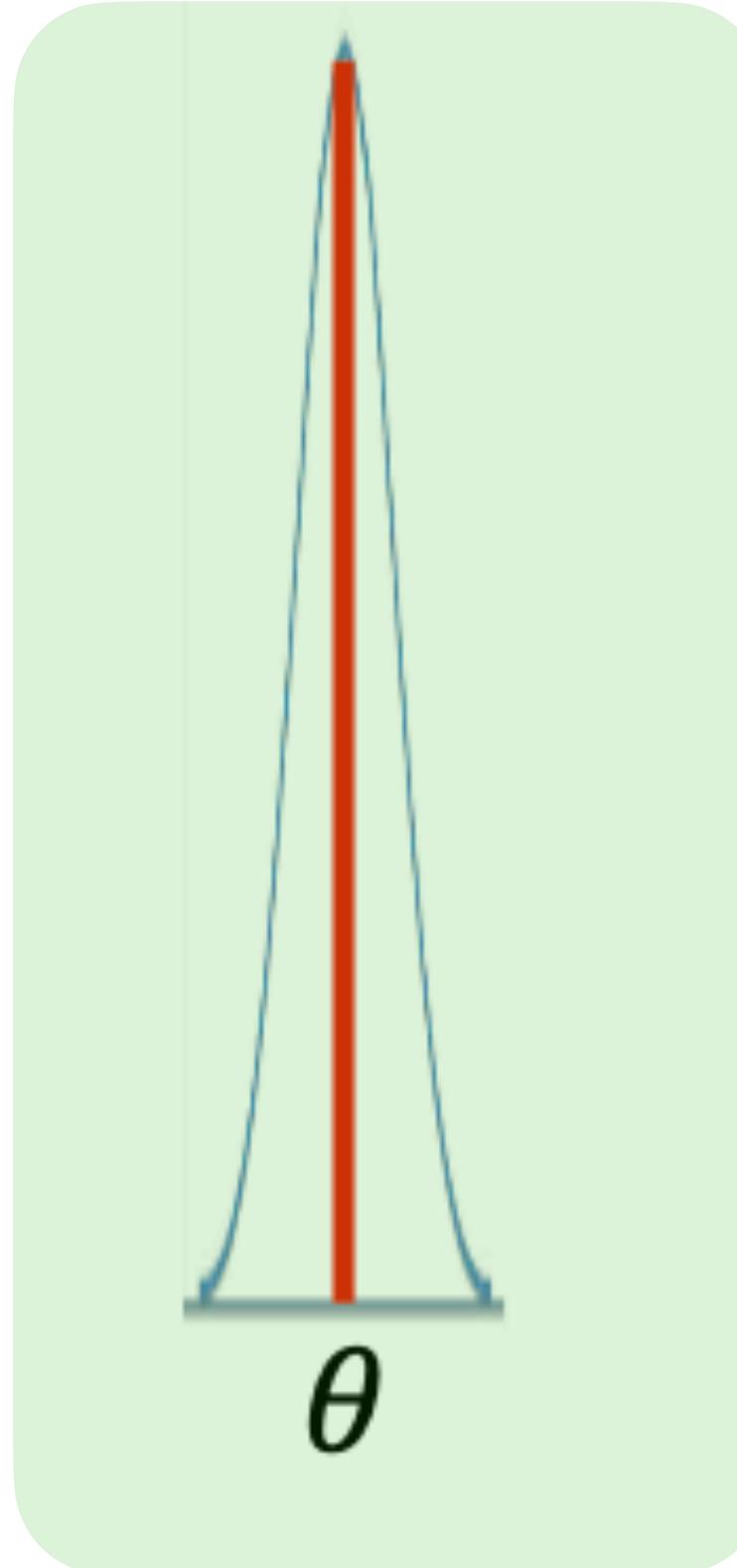
$$\rho(X, Y) = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{E[(X - \mu_X)^2]}\sqrt{E[(Y - \mu_Y)^2]}}$$

$$\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

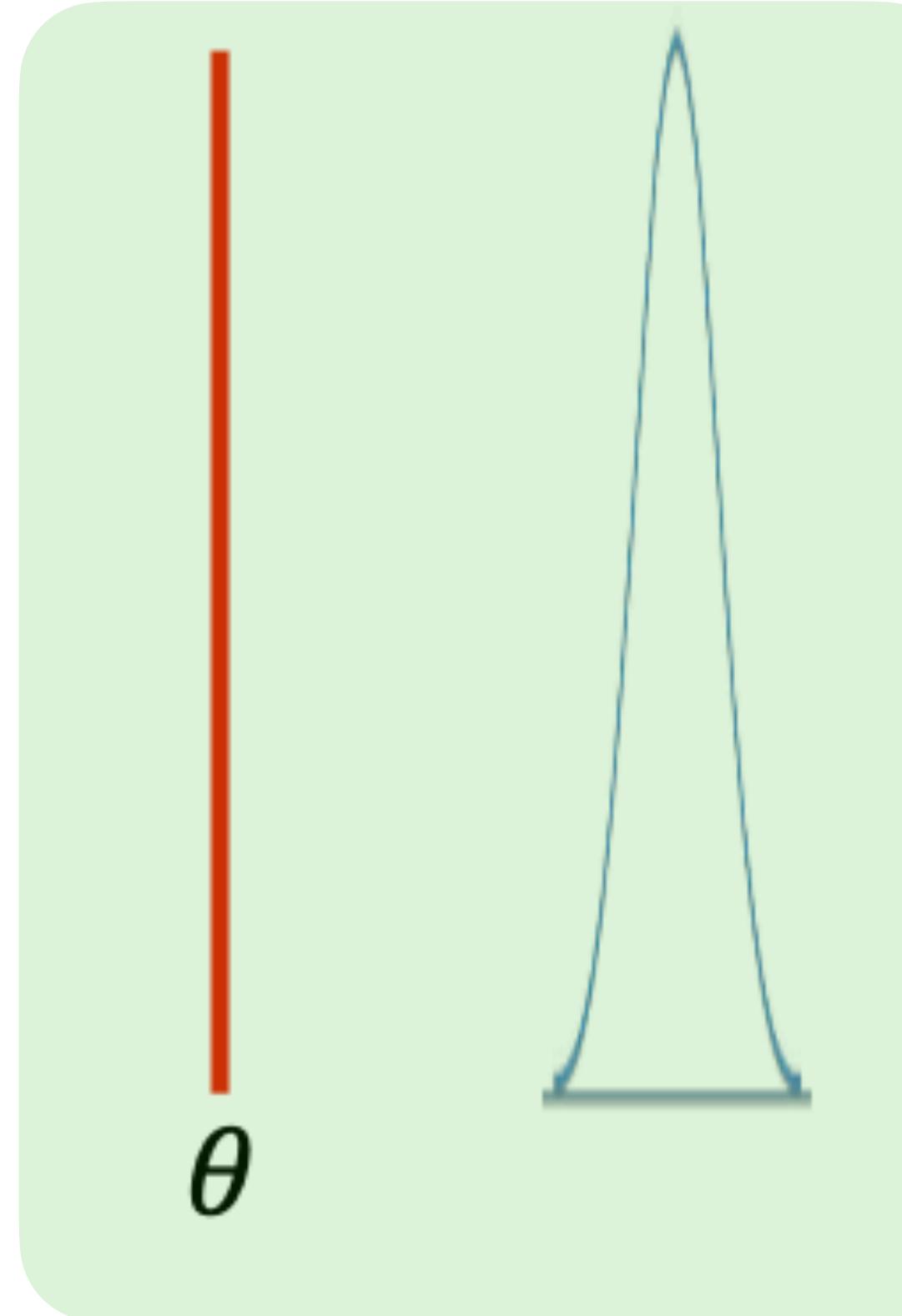
HOW GOOD IS AN ESTIMATOR?

Several possible estimators. Quality of an estimator?

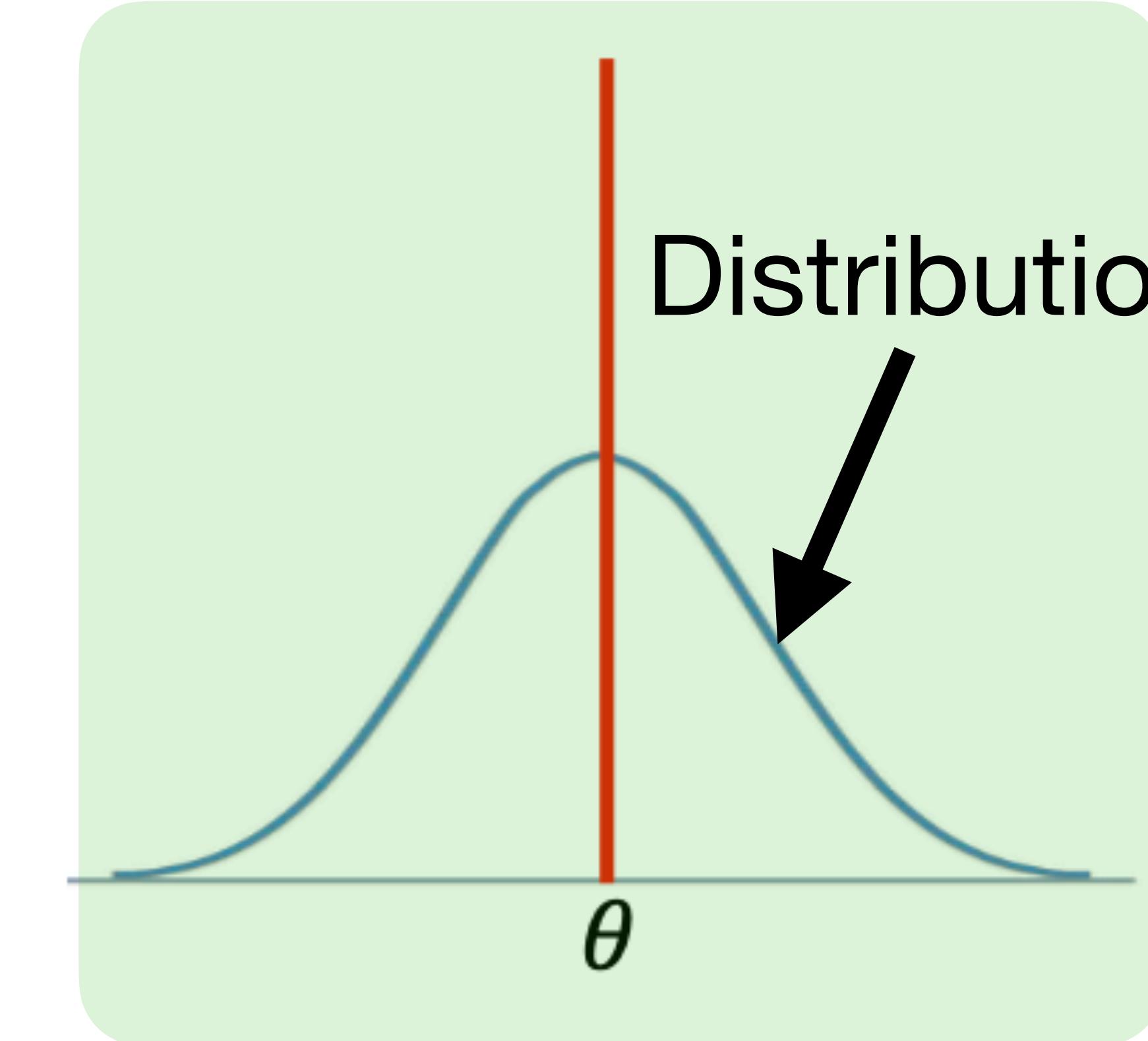
Good



Low bias
Low variance



Bias



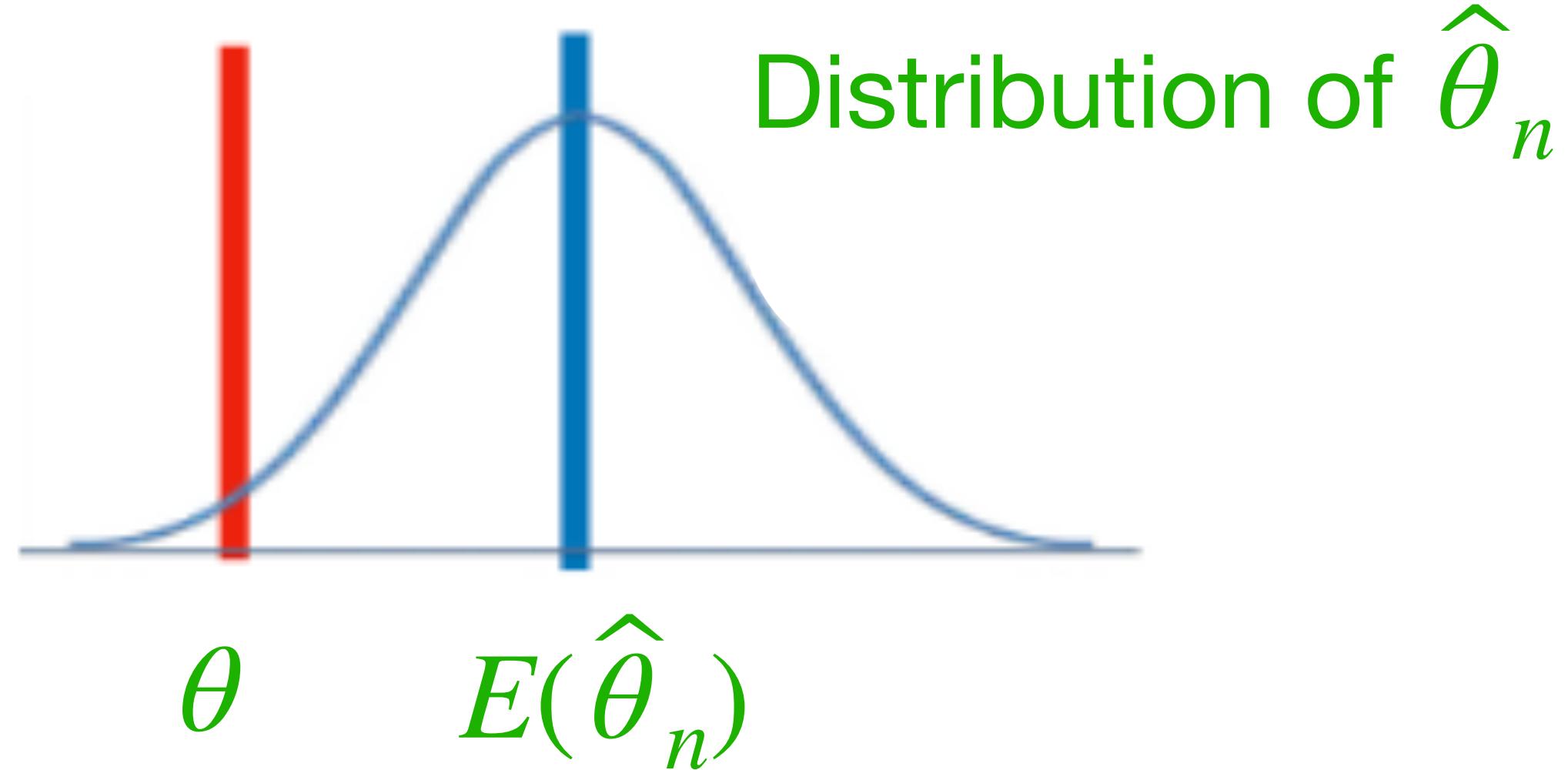
Variance

Distribution of $\hat{\theta}_n$

BIAS, VARIANCE

Bias: Expected overestimate of θ :

$$Bias(\hat{\theta}_n) = E(\hat{\theta}_n - \theta) = E(\hat{\theta}_n) - \theta$$

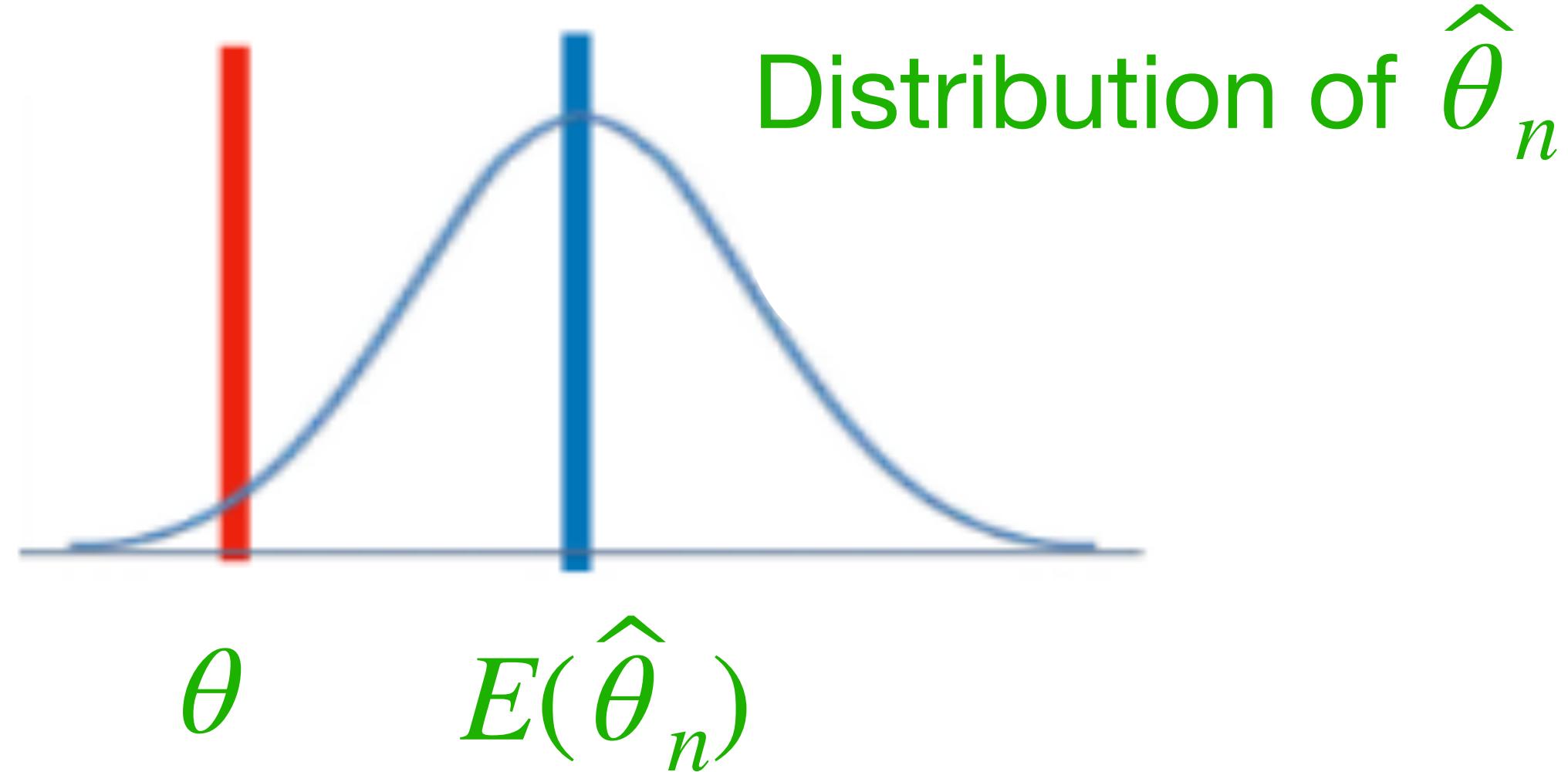


An estimator $\hat{\theta}_n$ is **unbiased** if $Bias(\hat{\theta}_n) = 0$

BIAS, VARIANCE

Bias: Expected overestimate of θ :

$$Bias(\hat{\theta}_n) = E(\hat{\theta}_n - \theta) = E(\hat{\theta}_n) - \theta$$



An estimator $\hat{\theta}_n$ is **unbiased** if $Bias(\hat{\theta}_n) = 0$

Variance: $Var(\hat{\theta}_n) = E[(\hat{\theta}_n - E(\hat{\theta}_n))^2]$

Bias-Variance tradeoff is typical.

BIAS, VARIANCE

Ex: Observe n coin flips $X_1, \dots, X_n \sim Bernoulli(p)$.

True value of p unknown. Want to estimate it.

Possible estimator: Sample mean $\bar{X}_n = \frac{1}{n} \sum X_i$

BIAS, VARIANCE

Ex: Observe n coin flips $X_1, \dots, X_n \sim Bernoulli(p)$.

True value of p unknown. Want to estimate it.

Possible estimator: Sample mean $\bar{X}_n = \frac{1}{n} \sum X_i$

Recall: $E(\bar{X}_n) = \frac{1}{n} \sum E(X_i) = p$,

Sample mean is an unbiased estimator.

$$Var(\bar{X}_n) = \frac{1}{n^2} \sum Var(X_i) = \frac{p(1-p)}{n}$$

HOW GOOD IS AN ESTIMATOR?

Quality of an estimator $\hat{\theta}_n$:

$$\text{mean squared error (mse)} = E[(\hat{\theta}_n - \theta)^2]$$

Theorem: $mse = bias^2(\hat{\theta}_n) + Var(\hat{\theta}_n)$

HOW GOOD IS AN ESTIMATOR?

Quality of an estimator $\hat{\theta}_n$:

$$\text{mean squared error (mse)} = E[(\hat{\theta}_n - \theta)^2]$$

Theorem: $mse = bias^2(\hat{\theta}_n) + Var(\hat{\theta}_n)$

Proof:

$$\begin{aligned} E[(\hat{\theta}_n - \theta)^2] &= E[(\hat{\theta}_n - \mu_{\hat{\theta}_n} + \mu_{\hat{\theta}_n} - \theta)^2] \\ &= E[(\hat{\theta}_n - \mu_{\hat{\theta}_n})^2] + E[(\mu_{\hat{\theta}_n} - \theta)^2] + 2E[\hat{\theta}_n - \mu_{\hat{\theta}_n}]E[\mu_{\hat{\theta}_n} - \theta] \end{aligned}$$

HOW GOOD IS AN ESTIMATOR?

Quality of an estimator $\hat{\theta}_n$:

$$\text{mean squared error (mse)} = E[(\hat{\theta}_n - \theta)^2]$$

Theorem: $mse = bias^2(\hat{\theta}_n) + Var(\hat{\theta}_n)$

Proof:

$$\begin{aligned} E[(\hat{\theta}_n - \theta)^2] &= E[(\hat{\theta}_n - \mu_{\hat{\theta}_n} + \mu_{\hat{\theta}_n} - \theta)^2] \\ &= E[(\hat{\theta}_n - \mu_{\hat{\theta}_n})^2] + E[(\mu_{\hat{\theta}_n} - \theta)^2] + 2E[\hat{\theta}_n - \mu_{\hat{\theta}_n}]E[\mu_{\hat{\theta}_n} - \theta] \\ &\quad \text{Var}(\hat{\theta}_n) \qquad \qquad \qquad \downarrow \\ &\quad (\mu_{\hat{\theta}_n} - \theta)^2 = Bias^2(\hat{\theta}_n) \end{aligned}$$

BIAS-VARIANCE TRADEOFF

Ex: Observe n coin flips $X_1, \dots, X_n \sim Bernoulli(p)$.

True value of p unknown. Want to estimate it.

Found: $E(\bar{X}_n) = p$, $Var(\bar{X}_n) = \frac{p(1-p)}{n} \Rightarrow mse = \frac{p(1-p)}{n}$

\bar{X}_n is unbiased estimator, call it $\hat{\theta}_U$.

BIAS-VARIANCE TRADEOFF

Ex: Observe n coin flips $X_1, \dots, X_n \sim Bernoulli(p)$.

True value of p unknown. Want to estimate it.

Alternative estimator, call it $\hat{\theta}_B$:
$$\frac{1 + \sum_{i=1}^n X_i}{n + 2}$$

e.g., 7 successes out of 10 trials, estimate of p with:

$$\hat{\theta}_B : 8/12 \approx 0.67, \quad \hat{\theta}_U : \frac{7}{10} = 0.7$$

BIAS-VARIANCE TRADEOFF

Ex: Observe n coin flips $X_1, \dots, X_n \sim Bernoulli(p)$.

True value of p unknown. Want to estimate it.

Alternative estimator, call it $\hat{\theta}_B$:
$$\frac{1 + \sum_{i=1}^n X_i}{n + 2}$$

Bias of $\hat{\theta}_B$:
$$\frac{1 + np}{n + 2} - p$$
, **biased.**

Variance of $\hat{\theta}_B$:
$$\frac{1}{(n + 2)^2} Var\left(\sum_{i=1}^n X_i\right) = \frac{np(1 - p)}{(n + 2)^2}$$

BIAS-VARIANCE TRADEOFF

Ex: Observe n coin flips $X_1, \dots, X_n \sim Bernoulli(p)$.

True value of p unknown. Want to estimate it.

Which one is a better estimator for $n = 10$?

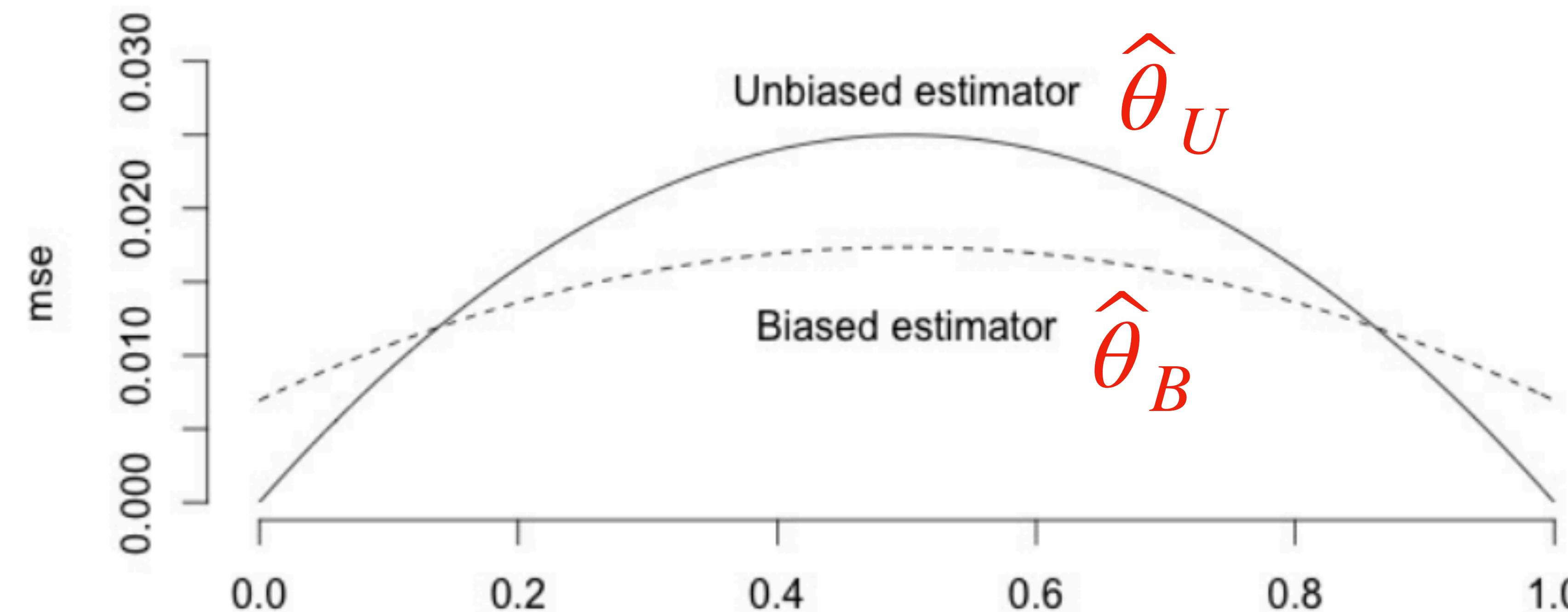
$$\text{mse of } \hat{\theta}_U: \frac{p(1-p)}{10}, \quad \text{mse of } \hat{\theta}_B: \frac{-6p^2 + 6p + 1}{144}$$

BIAS-VARIANCE TRADEOFF

Ex: Observe n coin flips $X_1, \dots, X_n \sim \text{Bernoulli}(p)$.

True value of p unknown. Want to estimate it.

Which one is a better estimator for $n = 10$?



REVIEW OF PREVIOUS LECTURE

Estimator $\hat{\theta}_n$:

$$Bias(\hat{\theta}_n) = E(\hat{\theta}_n - \theta) = E(\hat{\theta}_n) - \theta$$

$$Var(\hat{\theta}_n) = E[(\hat{\theta}_n - E(\hat{\theta}_n))^2]$$

Quality of $\hat{\theta}_n$:

$$\text{mean squared error (mse)} = E[(\hat{\theta}_n - \theta)^2]$$

$$mse = bias^2(\hat{\theta}_n) + Var(\hat{\theta}_n)$$

X_1, \dots, X_n iid \Rightarrow sample mean is an unbiased estimator of μ .

UNBIASED ESTIMATION OF VARIANCE

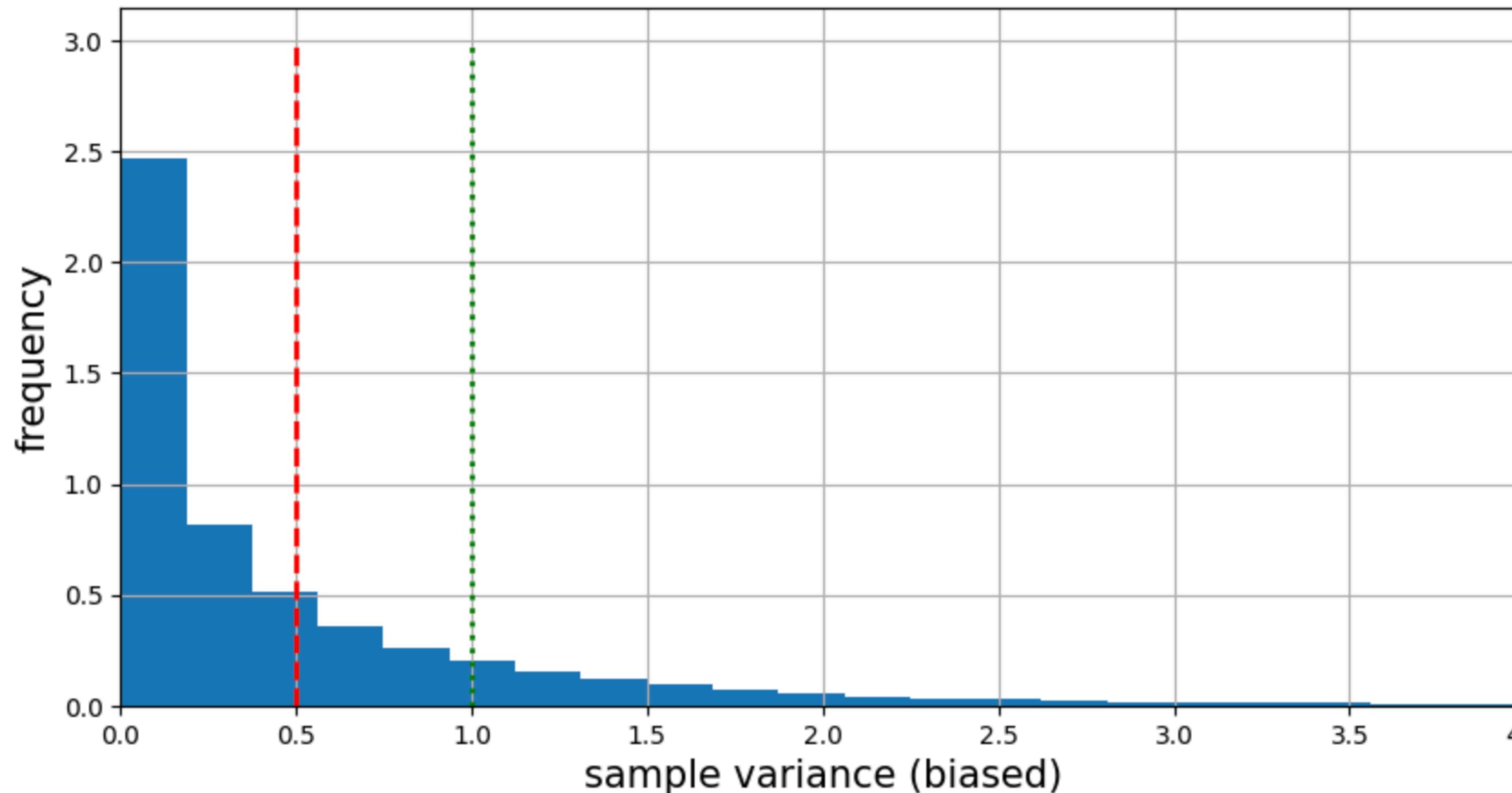
Reminder: $\text{Var}(X) = \sigma^2 = E(X - \mu_x)^2$

Estimate σ^2 :

1st try, plug-in estimate: $\widehat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$

UNBIASED ESTIMATION OF VARIANCE

```
n=2  
s = 100000  
X = np.random.normal(0,1,[n,s])  
# ddof is 0(1) for dividing by n (n-1)  
svar_b = np.var(X, axis=0, ddof=0)  
mean_svar_b = np.mean(svar_b)
```



$$\widehat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

UNBIASED ESTIMATION OF VARIANCE

$\widehat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ is biased.

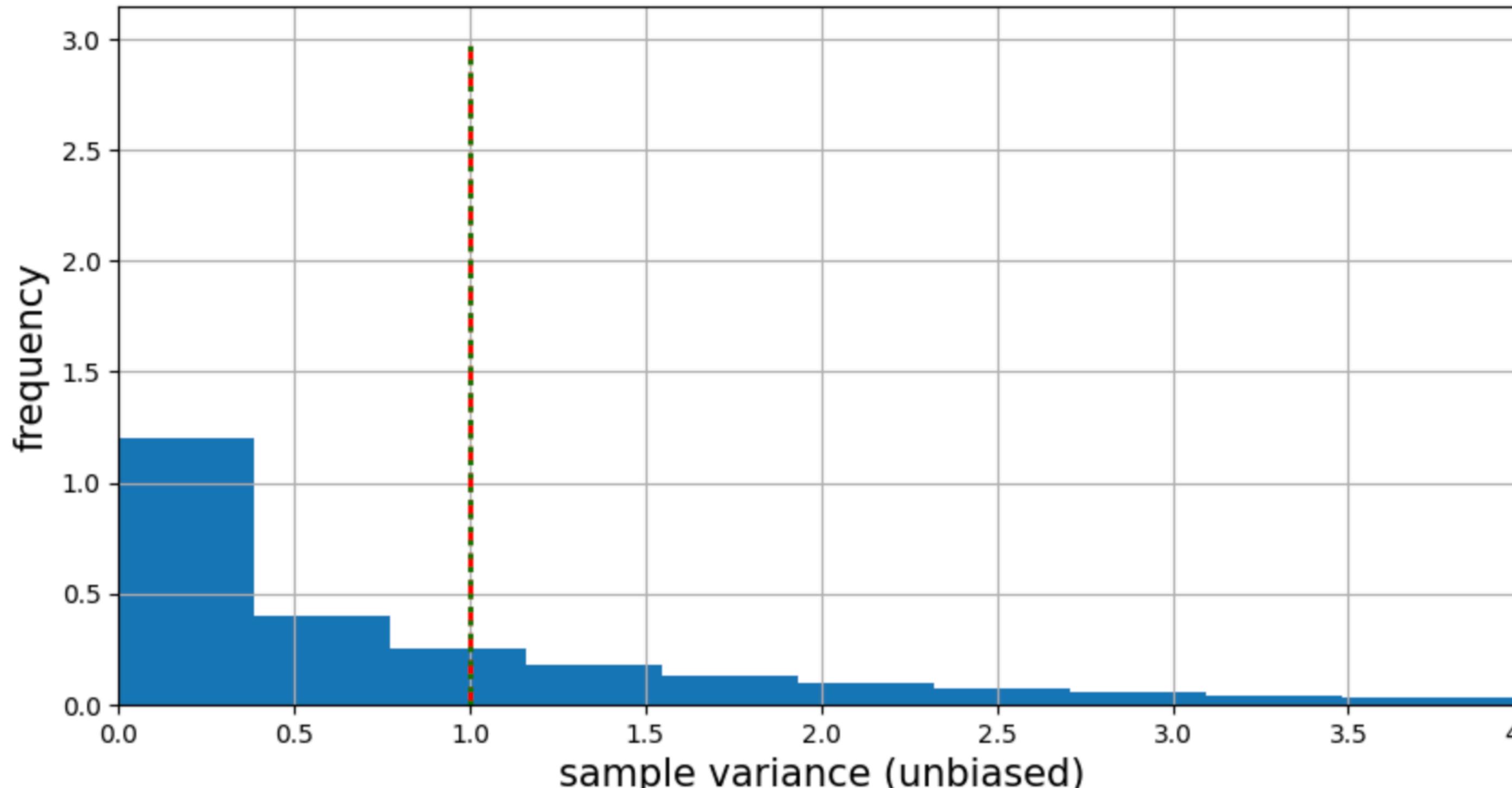
Theorem: $E(\widehat{\sigma}_0^2) = \frac{n-1}{n} \cdot \sigma^2$ (Check textbook for details)

Consider $\widehat{\sigma}_1^2 = \frac{n}{n-1} \cdot \widehat{\sigma}_0^2$:

$\widehat{\sigma}_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ is unbiased.

UNBIASED ESTIMATION OF VARIANCE

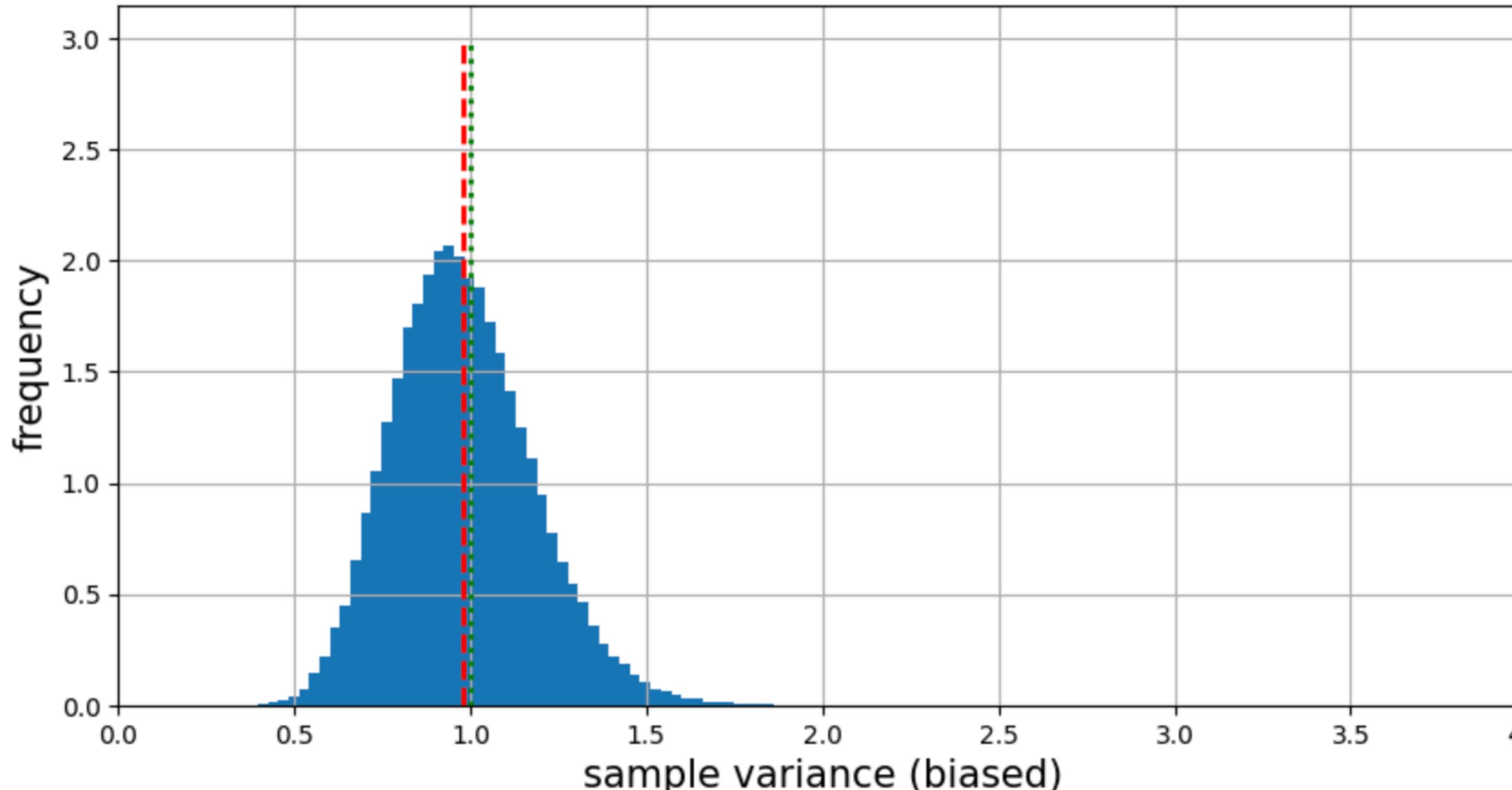
```
n=2  
s = 100000  
X = np.random.normal(0,1,[n,s])  
# ddof is 0(1) for dividing by n (n-1)  
svar_b = np.var(X,axis=0,ddof=1)  
mean_svar_b = np.mean(svar_b)
```



$$\widehat{\sigma}_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

UNBIASED ESTIMATION OF VARIANCE

```
n=50  
s = 100000  
X = np.random.normal(0,1,[n,s])  
# ddof is 0(1) for dividing by n (n-1)  
svar_b = np.var(X,axis=0,ddof=0)  
mean_svar_b = np.mean(svar_b)
```



$$\widehat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

For large n :

$$\widehat{\sigma}_0^2 \approx \widehat{\sigma}_1^2$$

MAXIMUM LIKELIHOOD ESTIMATORS

Ex: 4 flips of a coin $\Rightarrow [H, T, H, H]$. What would be our best estimate \hat{p} of probability of heads p given this sample?

MAXIMUM LIKELIHOOD ESTIMATORS

Ex: 4 flips of a coin $\Rightarrow [H, T, H, H]$. What would be our best estimate \hat{p} of probability of heads p given this sample?

Under $\hat{p} = 0.75$, the likelihood of observing the sample:

$$0.75^3 \cdot 0.25^1$$

MAXIMUM LIKELIHOOD ESTIMATORS

Ex: 4 flips of a coin $\Rightarrow [H, T, H, H]$. What would be our best estimate \hat{p} of probability of heads p given this sample?

Under $\hat{p} = 0.75$, the likelihood of observing the sample:

$$0.75^3 \cdot 0.25^1$$

Under $\hat{p} = 0.25$, the likelihood of observing the sample:

$$0.25^3 \cdot 0.75^1$$

\Rightarrow Our best estimate would be $\hat{p} = 0.75$.

MAXIMUM LIKELIHOOD ESTIMATORS

Likelihood function: Joint density of the data as a function of θ .

More specifically, let X_1, \dots, X_n iid with pdf $f(x; \theta)$:

$$L(\theta) = \prod_{i=1}^n f(X_i; \theta)$$

MAXIMUM LIKELIHOOD ESTIMATORS

Likelihood function: Joint density of the data as a function of θ .

More specifically, let X_1, \dots, X_n iid with pdf $f(x; \theta)$:

$$L(\theta) = \prod_{i=1}^n f(X_i; \theta)$$

Goal: Find parameter θ that maximizes the likelihood.

Note: θ maximizing log-likelihood maximizes likelihood too.

MAXIMUM LIKELIHOOD ESTIMATORS

Likelihood function: Joint density of the data as a function of θ .

More specifically, let X_1, \dots, X_n iid with pdf $f(x; \theta)$:

$$L(\theta) = \prod_{i=1}^n f(X_i; \theta)$$

Goal: Find parameter θ that maximizes the likelihood.

Note: θ maximizing log-likelihood maximizes likelihood too.

Maximizing log-likelihood is easier; summation rather than product

Simpler Goal: Find parameter θ that maximizes the log-likelihood.

$$\text{Solve for } \theta \text{ in } \frac{d \log L(\theta)}{d\theta} = 0$$

MAXIMUM LIKELIHOOD ESTIMATORS

Example: $X_1, \dots, X_n \sim Bernoulli(\theta)$. How to estimate unknown θ ?

$$L(\theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i}$$

e.g. with 1,0,1,1 the probability $\theta \cdot (1 - \theta) \cdot \theta \cdot \theta$

MAXIMUM LIKELIHOOD ESTIMATORS

Example: $X_1, \dots, X_n \sim Bernoulli(\theta)$. How to estimate unknown θ ?

$$L(\theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} \quad \text{e.g. with } 1,0,1,1 \text{ the probability } \theta \cdot (1 - \theta) \cdot \theta \cdot \theta$$

$$\log L(\theta) = \log \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i}$$

MAXIMUM LIKELIHOOD ESTIMATORS

Example: $X_1, \dots, X_n \sim Bernoulli(\theta)$. How to estimate unknown θ ?

$$L(\theta) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} \quad \text{e.g. with } 1,0,1,1 \text{ the probability } \theta \cdot (1-\theta) \cdot \theta \cdot \theta$$

$$\log L(\theta) = \log \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i}$$

$$= \sum_{i=1}^n \log \theta^{x_i} + \sum_{i=1}^n \log (1-\theta)^{1-x_i} = \log \theta \sum_{i=1}^n x_i + \log (1-\theta) \sum_{i=1}^n 1 - x_i$$

MAXIMUM LIKELIHOOD ESTIMATORS

Example: $X_1, \dots, X_n \sim Bernoulli(\theta)$. How to estimate unknown θ ?

$$L(\theta) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} \quad \text{e.g. with } 1,0,1,1 \text{ the probability } \theta \cdot (1-\theta) \cdot \theta \cdot \theta$$

$$\log L(\theta) = \log \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i}$$

$$= \sum_{i=1}^n \log \theta^{x_i} + \sum_{i=1}^n \log(1-\theta)^{1-x_i} = \log \theta \sum_{i=1}^n x_i + \log(1-\theta) \sum_{i=1}^n 1 - x_i$$

$$= \log \theta \sum_{i=1}^n x_i + \log(1-\theta)(n - \sum_{i=1}^n x_i)$$

MAXIMUM LIKELIHOOD ESTIMATORS

Example: $X_1, \dots, X_n \sim Bernoulli(\theta)$. How to estimate unknown θ ?

$$L(\theta) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} \quad \text{e.g. with } 1,0,1,1 \text{ the probability } \theta \cdot (1-\theta) \cdot \theta \cdot \theta$$

$$\log L(\theta) = \log \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i}$$

$$= \sum_{i=1}^n \log \theta^{x_i} + \sum_{i=1}^n \log(1-\theta)^{1-x_i} = \log \theta \sum_{i=1}^n x_i + \log(1-\theta) \sum_{i=1}^n 1 - x_i$$

$$= \log \theta \sum_{i=1}^n x_i + \log(1-\theta)(n - \sum_{i=1}^n x_i) \quad \begin{aligned} &\text{Derivative wr.t. } \theta \text{ equal to 0 gives} \\ &\theta = \bar{X}_n \end{aligned}$$

MAXIMUM LIKELIHOOD ESTIMATORS

Example: $X_1, \dots, X_n \sim Bernoulli(\theta)$. How to estimate unknown θ ?

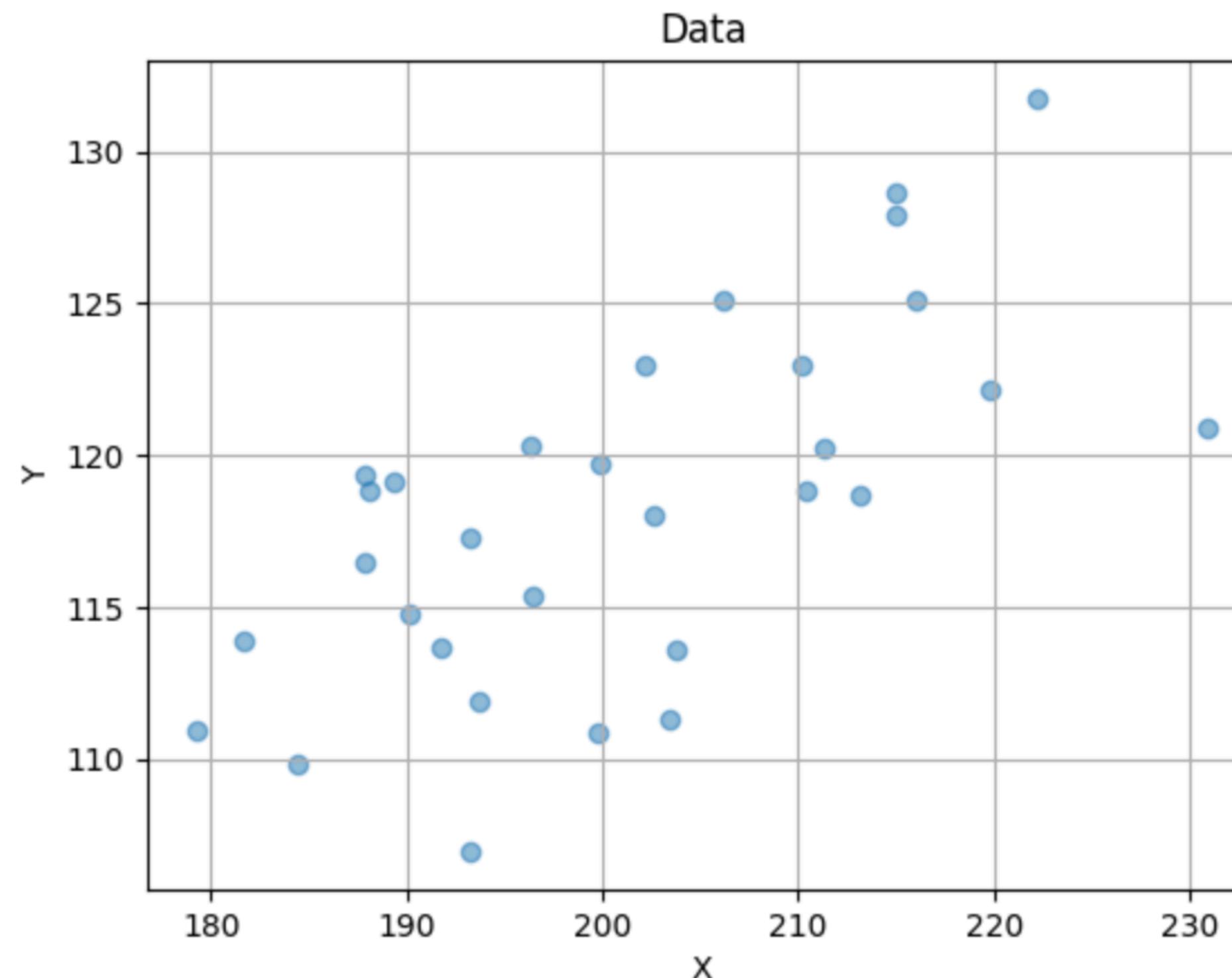
Same estimate as the plug-in estimator: sample mean.

Bias is 0.

Derivative wr.t. θ equal to 0 gives

$$\theta = \bar{X}_n$$

SIMULATION APPROXIMATION OF BOOTSTRAP ESTIMATES



X, Y have bivariate joint distribution.

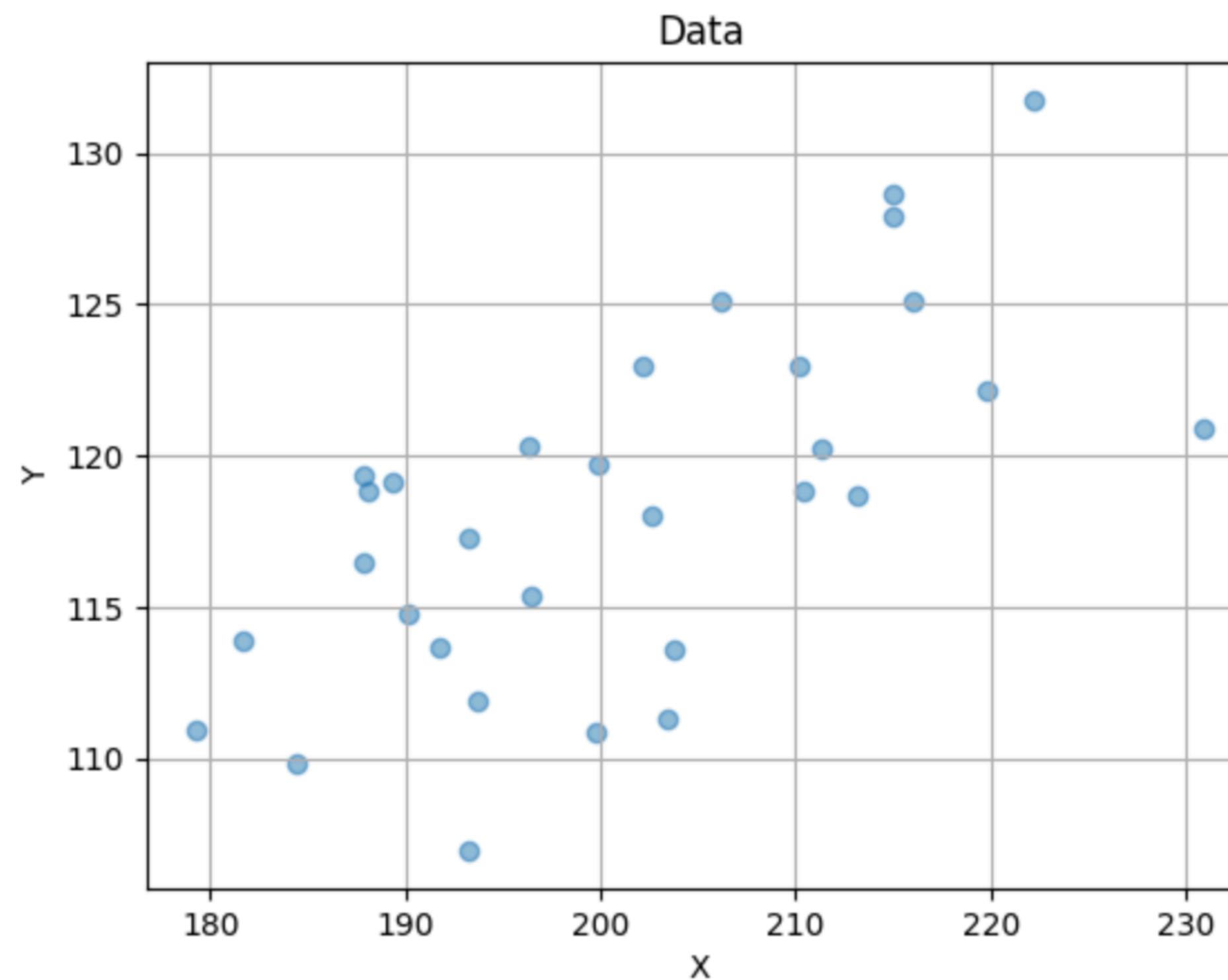
We observe a random sample (Data)

Sample Correlation:

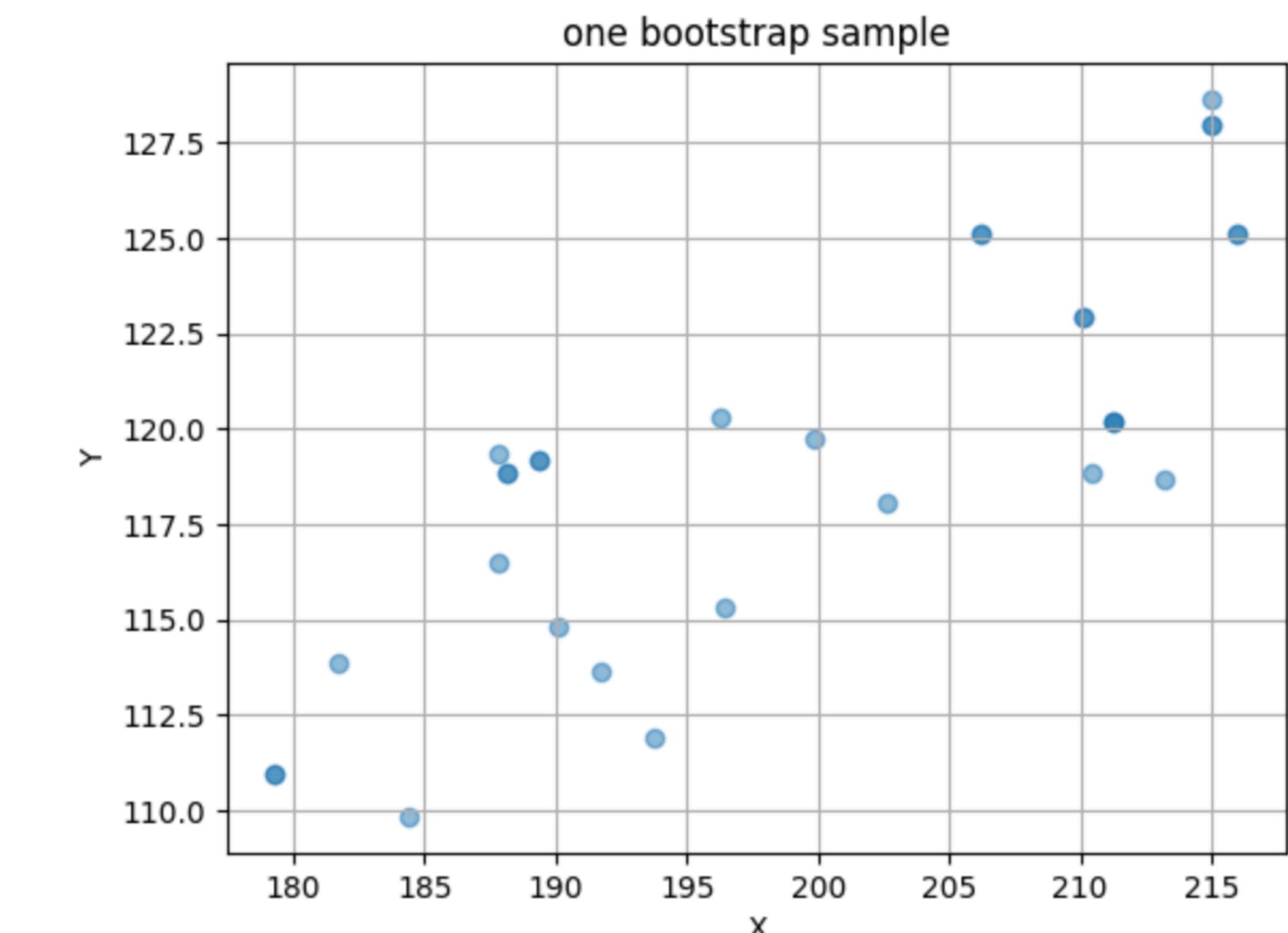
$$R = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Say we're interested in bias of R in estimating ρ .

SIMULATION APPROXIMATION OF BOOTSTRAP ESTIMATES



PearsonRResult(statistic=0.6649024611769977, pvalue=4.498954341386631e-05)



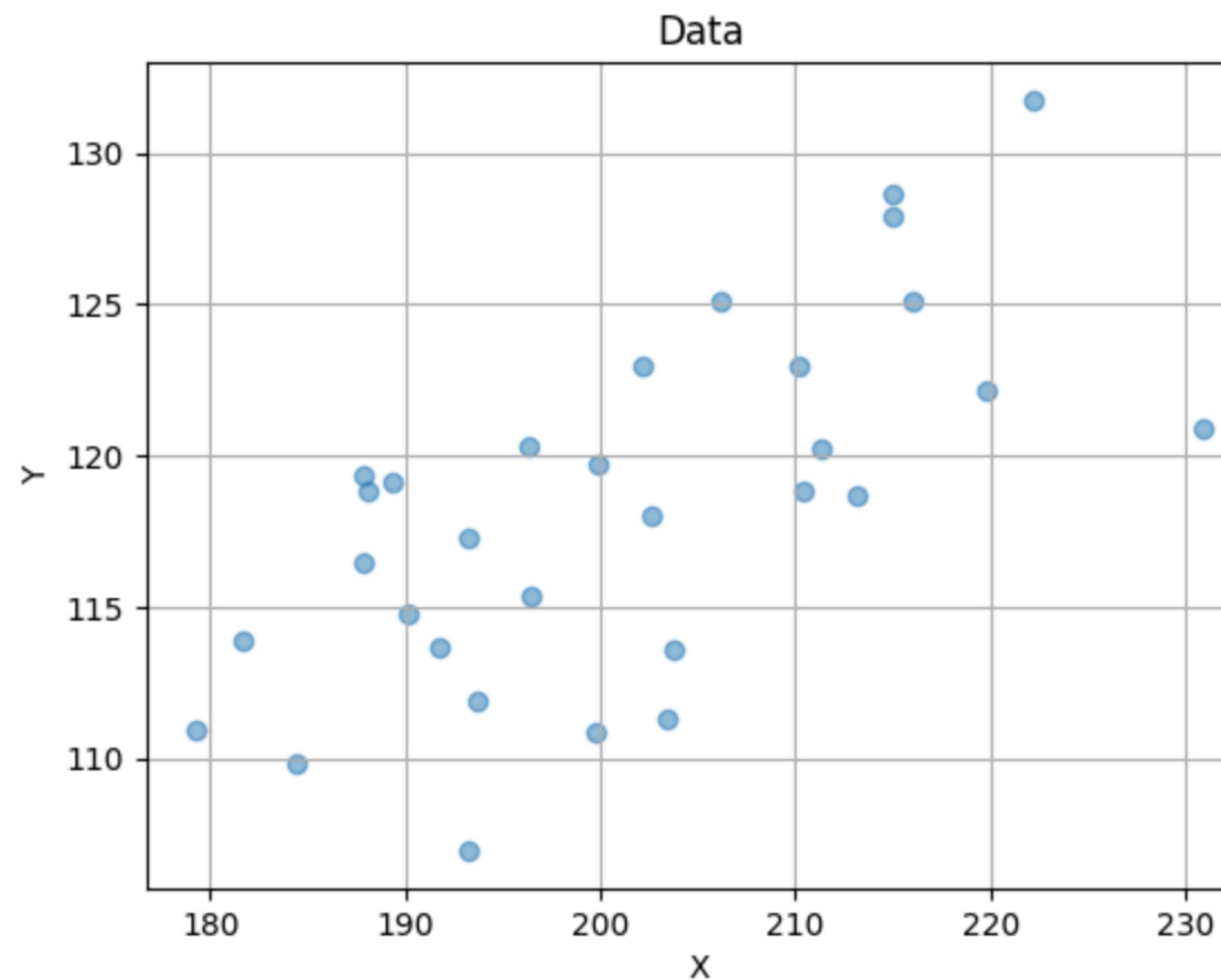
PearsonRResult(statistic=0.8015767674698848, pvalue=5.992620251964453e-08)

A bootstrap sample

Sample data with replacement

```
n_r = data.shape[0]
rand_i = np.random.choice(n_r, size=n_r, replace=True)
new_b = data[rand_i]
```

SIMULATION APPROXIMATION OF BOOTSTRAP ESTIMATES



```
n_samples=1000
n_r = data.shape[0]
diff_sum = 0
for _ in range(n_samples):
    rand_i = np.random.choice(n_r, size=n_r, replace=True)
    new_b = data[rand_i]
    corr = scipy.stats.pearsonr(new_b[:, 0], new_b[:, 1])
    diff_sum += corr[0]-corr_data
diff_sum/n_samples
```

0.0038447409039712647

R^i : Correlation of i^{th} bootstrap sample

Bias of sample correlation: Avg of $R^i - R$ to estimate mean of $R - \rho$

CONFIDENCE INTERVALS

Can apply to any parameter but most common is mean μ .

Have seen: \bar{X}_n is a sensible point estimator for mean μ .

Precise estimate, like $\bar{X}_n = 4$

but how much confidence do we have that $\mu = 4$?

Confidence interval:

Make statements like

with 95 % confidence $\mu \in (\bar{X}_n - m, \bar{X}_n + m)$

CONFIDENCE INTERVALS

For μ with known σ :

Reminder 1:
$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0,1)$$

Reminder 2:

$$p = P(-k \leq Z \leq k) \Rightarrow k = \Phi^{-1}\left(\frac{1+p}{2}\right)$$

CONFIDENCE INTERVALS

Given p for desired confidence find interval for μ using \bar{X}_n .

$$p = P(-k \leq Z \leq k) = P\left(-k \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq k\right)$$

CONFIDENCE INTERVALS

Given p for desired confidence find interval for μ using \bar{X}_n .

$$p = P(-k \leq Z \leq k) = P\left(-k \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq k\right)$$

Note:

$$P\left(-k \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq k\right) = P\left(\bar{X}_n - \frac{k\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{k\sigma}{\sqrt{n}}\right)$$

CONFIDENCE INTERVALS

Given p for desired confidence find interval for μ using \bar{X}_n .

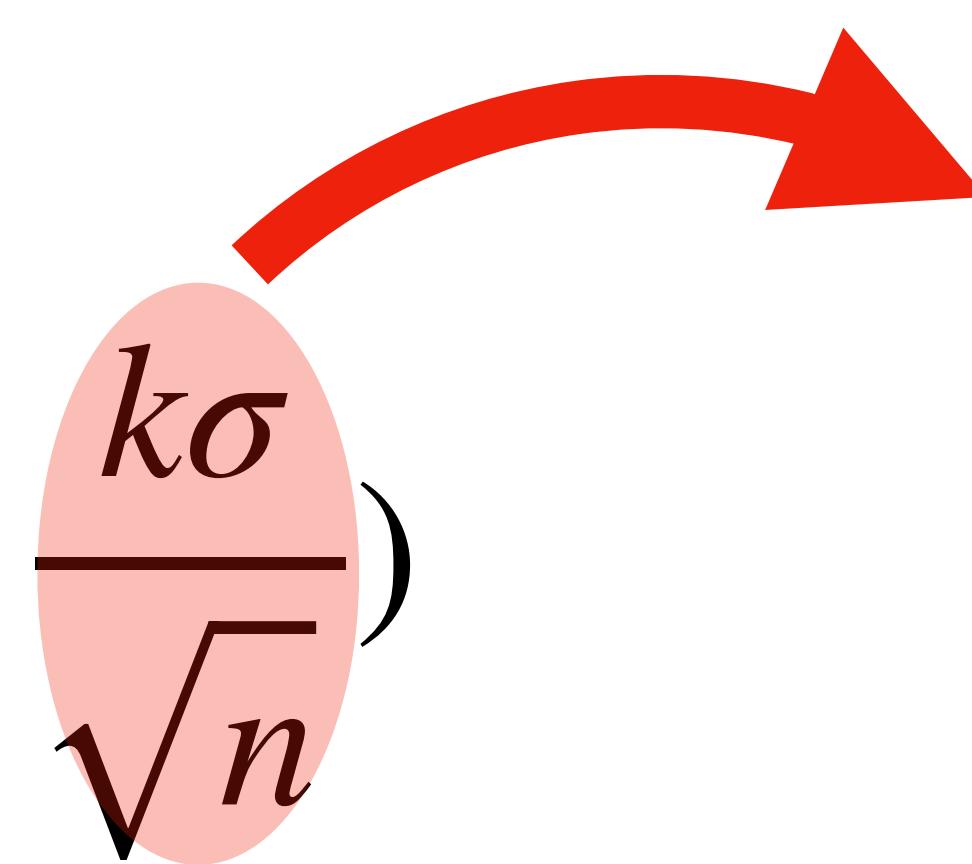
$$p = P(-k \leq Z \leq k) = P\left(-k \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq k\right)$$

Note:

$$P\left(-k \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq k\right) = P\left(\bar{X}_n - \frac{k\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{k\sigma}{\sqrt{n}}\right)$$

Step 1: Find $k = \Phi^{-1}\left(\frac{1+p}{2}\right)$

Step 2: Find interval $(\bar{X}_n - \frac{k\sigma}{\sqrt{n}}, \bar{X}_n + \frac{k\sigma}{\sqrt{n}})$



Margin of error

QUIZ

Write down the formula for mean squared error (mse) of estimator $\hat{\theta}_n$ that relates mse with $bias(\hat{\theta}_n)$ and $var(\hat{\theta}_n)$.

REVIEW OF (LAST PART OF) PREVIOUS LECTURE

Make statements like

with 95 % confidence $\mu \in (\bar{X}_n - m, \bar{X}_n + m)$

$$p = P(-k \leq Z \leq k) \Rightarrow k = \Phi^{-1}\left(\frac{1+p}{2}\right)$$

$$P\left(-k \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq k\right) = P\left(\bar{X}_n - \frac{k\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{k\sigma}{\sqrt{n}}\right)$$

Margin of error

CONFIDENCE INTERVALS

Ex: # of citations of a random paper is a random variable with $\sigma = 5$.

In a sample of 100 papers sample mean \bar{X}_n is 7.5.

What is the 95 % confidence interval for mean μ ?

$$\text{Step 1: } k = \Phi^{-1}\left(\frac{1 + p}{2}\right) = \Phi^{-1}(0.975) = 1.96$$

CONFIDENCE INTERVALS

Ex: # of citations of a random paper is a random variable with $\sigma = 5$.

In a sample of 100 papers sample mean \bar{X}_n is 7.5.

What is the 95 % confidence interval for mean μ ?

$$\text{Step 1: } k = \Phi^{-1}\left(\frac{1+p}{2}\right) = \Phi^{-1}(0.975) = 1.96$$

Step 2: 95 % confidence interval for mean μ :

$$\left(\bar{X}_n - \frac{k\sigma}{\sqrt{n}}, \bar{X}_n + \frac{k\sigma}{\sqrt{n}}\right) = \left(7.5 - \frac{1.96 \times 5}{10}, 7.5 + \frac{1.96 \times 5}{10}\right) = (6.52, 8.48)$$

CONFIDENCE INTERVALS

For μ with unknown σ :

We have seen $\widehat{\sigma}_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ is unbiased.

CONFIDENCE INTERVALS

For μ with unknown σ :

We have seen $\widehat{\sigma}_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ is unbiased.

Replace σ in Step 2 with $\sqrt{\widehat{\sigma}_1^2}$.

Theorem: $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\widehat{\sigma}_1}$ has *t distribution* with $n - 1$ degrees of freedom.

CONFIDENCE INTERVALS

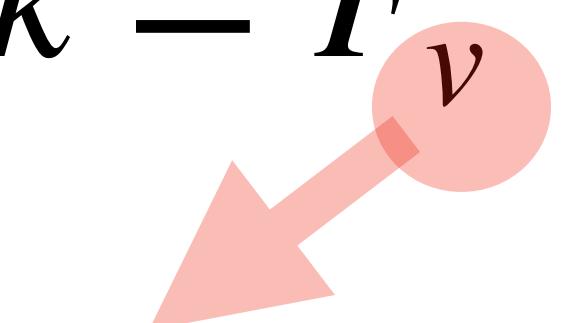
For μ with unknown σ :

We have seen $\widehat{\sigma}_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ is unbiased.

Replace σ in Step 2 with $\sqrt{\widehat{\sigma}_1^2}$.

Theorem: $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\widehat{\sigma}_1}$ has *t distribution* with $n - 1$ degrees of freedom.

Replace $k = \Phi^{-1}\left(\frac{1+p}{2}\right)$ with $k = F^{-1}\left(\frac{1+p}{2}\right)$ in Step 1.



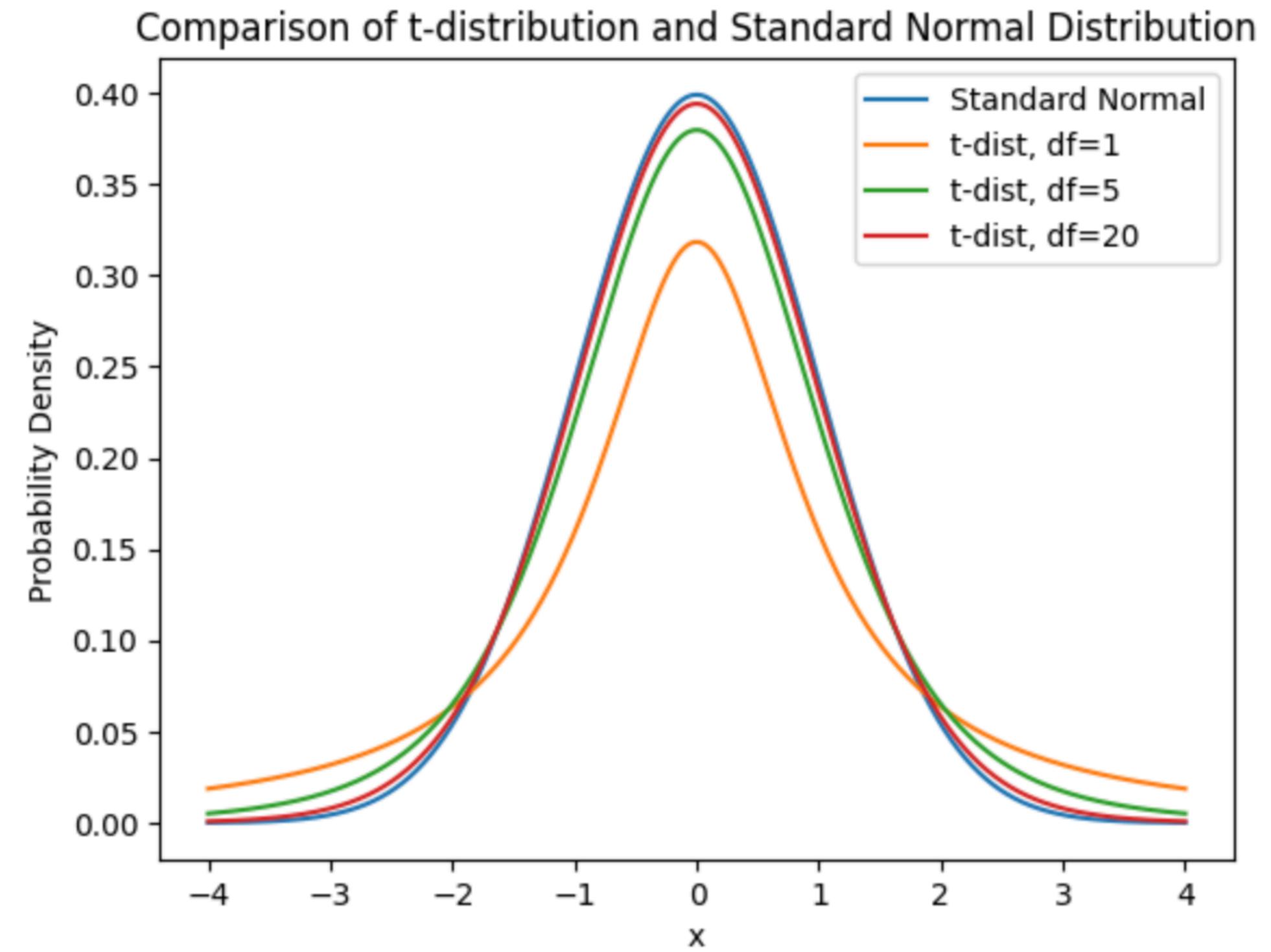
$v = n - 1$ is the degrees of freedom

CONFIDENCE INTERVALS

(Student's)
t distribution

How to find F_t^{-1} ?

$F_t^{-1}(1)$ with 20
degrees of freedom.



William S Gosset
(aka *Student*)



```
from scipy.stats import t  
t.cdf(1,20)
```

0.83537114141455

While working at
Guinness

CONFIDENCE INTERVALS

Ex: Heights of men. Sample of size 9: 168, 176, 195, 182, 188, 150, 165, 170, 158

Find 95 % confidence interval for distribution mean.

CONFIDENCE INTERVALS

Ex: Heights of men. Sample of size 9: 168, 176, 195, 182, 188, 150, 165, 170, 158

Find 95 % confidence interval for distribution mean.

Find $k = F_{\nu}^{-1}\left(\frac{1+p}{2}\right)$ for $\nu = 8, p = 0.95$

$$k = F_8^{-1}\left(\frac{1+0.95}{2}\right) = F_8^{-1}(0.975) \approx 2.306$$

```
from scipy.stats import t  
t.ppf(0.975,8)
```

2.306004135204166

CONFIDENCE INTERVALS

Ex: Heights of men. Sample of size 9: 168, 176, 195, 182, 188, 150, 165, 170, 158

Find 95 % confidence interval for distribution mean.

Find $k = F_v^{-1}\left(\frac{1+p}{2}\right)$ for $v = 8, p = 0.95$

```
from scipy.stats import t  
t.ppf(0.975, 8)
```

$k = F_8^{-1}\left(\frac{1+0.95}{2}\right) = F_8^{-1}(0.975) \approx 2.306$

```
2.306004135204166
```

$\bar{X}_n = 172.4, \widehat{\sigma}_1^2 \approx 206.03, \widehat{\sigma}_1 \approx 14.35,$ Margin of error: $\frac{k\widehat{\sigma}_1}{\sqrt{n}} \approx 11.03$

Confidence interval $(\bar{X}_n - \frac{k\widehat{\sigma}_1}{\sqrt{n}}, \bar{X}_n + \frac{k\widehat{\sigma}_1}{\sqrt{n}}) = (161.37, 183.43)$

CONFIDENCE INTERVALS

How to interpret confidence intervals?

Not right to say:

μ is between 161.37 and 183.43 with 95% probability.

CONFIDENCE INTERVALS

How to interpret confidence intervals?

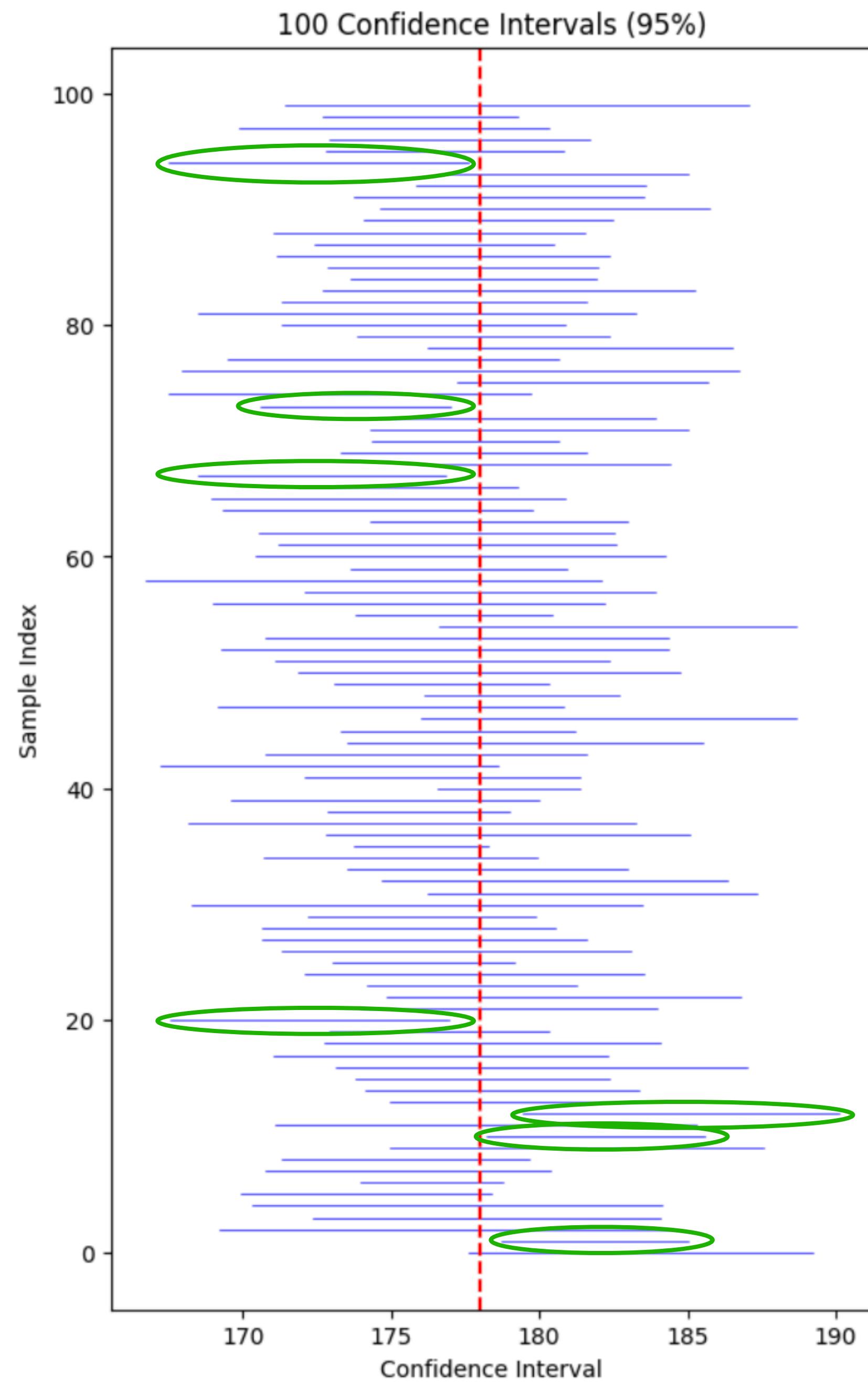
Not right to say:

μ is between 161.37 and 183.43 with 95% probability.

100 observed 95 % confidence intervals based on samples of size 9, $\mu = 178$, $\sigma = 7$.

```
samples = [np.random.normal(178, 7, 9) for _ in range(100)]  
  
# Calculate confidence intervals  
intervals = []  
for sample in samples:  
    k = stats.t.ppf((1 + conf_level) / 2, df=sample_size - 1)  
    s_mean = np.mean(sample)  
    s_std = np.std(sample, ddof=1)  
    mar_of_err = k * s_std / np.sqrt(sample_size)  
    intervals.append((s_mean - mar_of_err, s_mean + mar_of_err))
```

For this sample 93 % of intervals contain $\mu = 178$.



Number of intervals outside 95% CI: 7

HYPOTHESIS TESTING

Statements about parameter/property θ of a distribution/population

Average GPA is < 2.8

Probability of heads of a coin is > 0.6

People eat more on weekends than in weekdays.

HYPOTHESIS TESTING

Statements about parameter/property θ of a distribution/population

Average GPA is < 2.8

Probability of heads of a coin is > 0.6

People eat more on weekends than in weekdays.

Simple vs composite hypotheses

$\theta = 3.2$ (simple), $\theta \in \{3, 4.2\}$ (composite), $\theta \in [0.5, 1.4]$ (composite)

HYPOTHESIS TESTING

Statements about parameter/property θ of a distribution/population

Average GPA is < 2.8

Probability of heads of a coin is > 0.6

People eat more on weekends than in weekdays.

Simple vs composite hypotheses

$\theta = 3.2$ (simple), $\theta \in \{3, 4.2\}$ (composite), $\theta \in [0.5, 1.4]$ (composite)

One-sided vs Two-sided

$\theta > 3.2$ (one-sided), $\theta < -1.5$ or $\theta > 2$ (two-sided), $\theta \neq 2$ (two-sided)

HYPOTHESIS TESTING

Two hypothesis

Null Hypothesis: H_0

Status quo, assumption believed to be true

Coin in my pocket, probability of heads $p = 0.5$

HYPOTHESIS TESTING

Two hypothesis

Null Hypothesis: H_0

Status quo, assumption believed to be true

Coin in my pocket, probability of heads $p = 0.5$

Alternative Hypothesis: H_A

Complement of H_0

Novel finding after research

Coin has probability of heads $p \neq 0.5$

HYPOTHESIS TESTING

How to test?

Design experiment, collect data. Check:

If data shows strong evidence against H_0

Reject H_0

(In favor of H_A)

Else

Do not reject H_0

(Note: doesn't mean accept H_0)

HYPOTHESIS TESTING

How to test?

Design experiment, collect data. Check:

If data shows strong evidence against H_0

Reject H_0

(In favor of H_A)

Else

Do not reject H_0

(Note: doesn't mean accept H_0)

Analogy with the legal principle:

Presumed **innocent** (H_0) until proven **guilty** (H_A) with strong evidence against innocence.

HYPOTHESIS TESTING

More specifically,

Design experiment

Define **test statistic** T (related to hypothesis)

Distribution of T under H_0

Compute value t of the statistic T applied on the data

Reject H_0 if $t \in R$, for ‘reasonable’ **rejection region** R

HYPOTHESIS TESTING

Ex: Mean of a normal distribution with known variance

$X_1, \dots, X_n \sim N(\mu, \sigma^2)$ with unknown μ . Test the hypotheses:

$$H_0 : \mu = \mu_0 \quad H_A : \mu \neq \mu_0$$

HYPOTHESIS TESTING

Ex: Mean of a normal distribution with known variance

$X_1, \dots, X_n \sim N(\mu, \sigma^2)$ with unknown μ . Test the hypotheses:

$$H_0 : \mu = \mu_0 \quad H_A : \mu \neq \mu_0$$

Informally, reject H_0 if \bar{X}_n far from μ_0 (strong evidence against H_0)

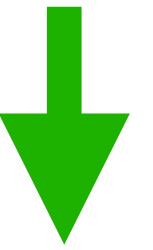
Formally, define test statistic, $T = |\bar{X}_n - \mu_0|$. Reject H_0 if $T \geq c$.

HYPOTHESIS TESTING

How to choose c ?

Significance level α :

$$P_{H_0}(T \geq c) \leq \alpha$$



Type I error:
We reject H_0
since \bar{X}_n too
distant from μ_0
but H_0 is true.

Usually $\alpha = 0.05$

HYPOTHESIS TESTING

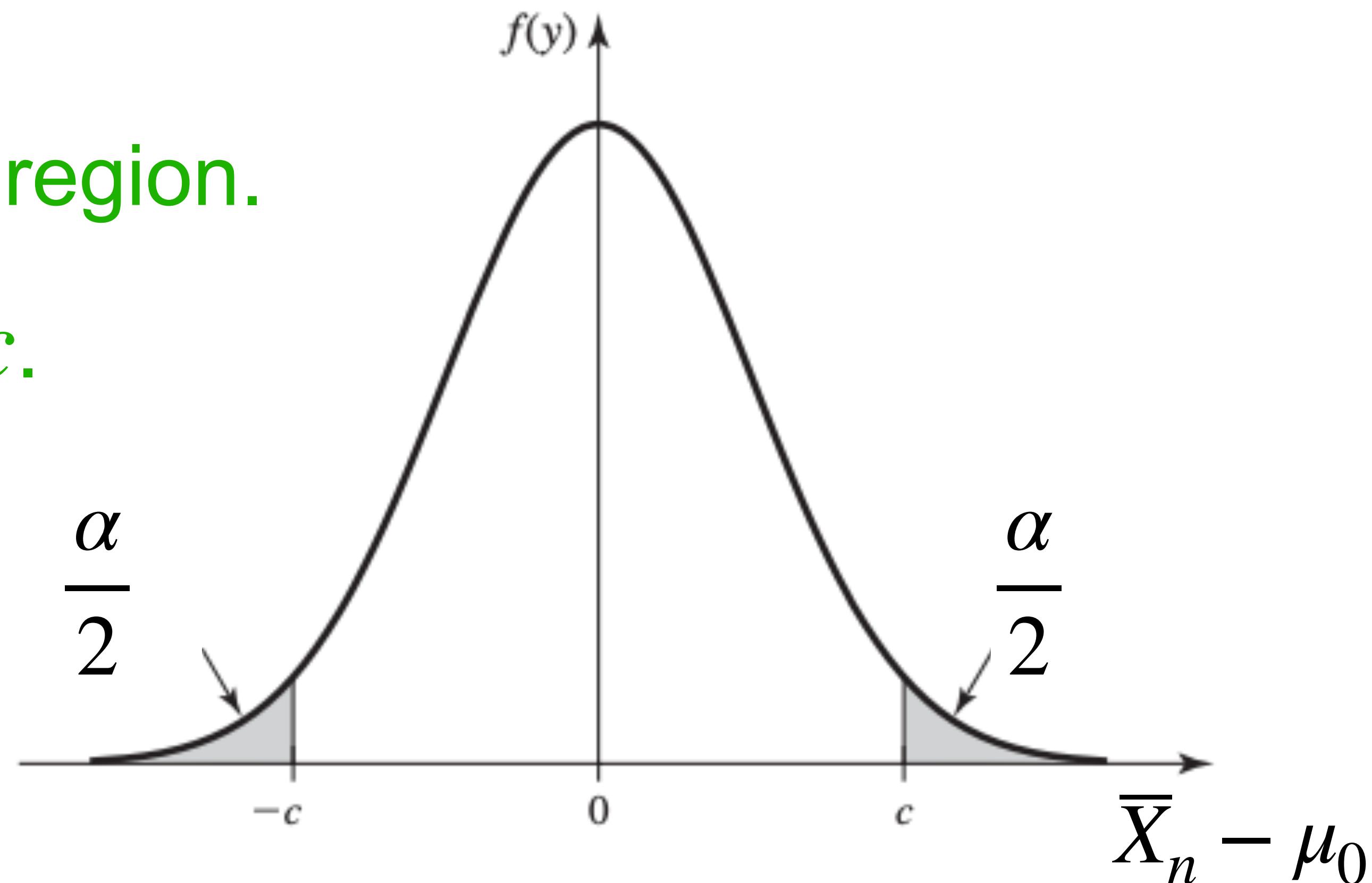
How to choose c ?

Significance level α :

$$P_{H_0}(T \geq c) \leq \alpha$$

Reject H_0 if \bar{X}_n lies anywhere in shaded region.

Use quantile to find c .



QUIZ

Why do we maximize **log likelihood** rather than **likelihood** directly while computing a maximum likelihood estimator?

(INFORMAL) COURSE EVALUATION

How can we improve the course for the remainder of the semester?

Don't write your name on the paper.

1- Lectures in terms of:

- a- Density b- Slides c- Anything else

2- Homeworks in terms of:

- a- Difficulty b- Length c- Fair evaluation d- Anything else

3- SI sessions in terms of:

- a- Usefulness b- Content c- Anything else

4- Office hours / Piazza in terms of:

- a- Office hour times b- Piazza usefulness c- Anything else

5- Any other suggestions for improving the course.

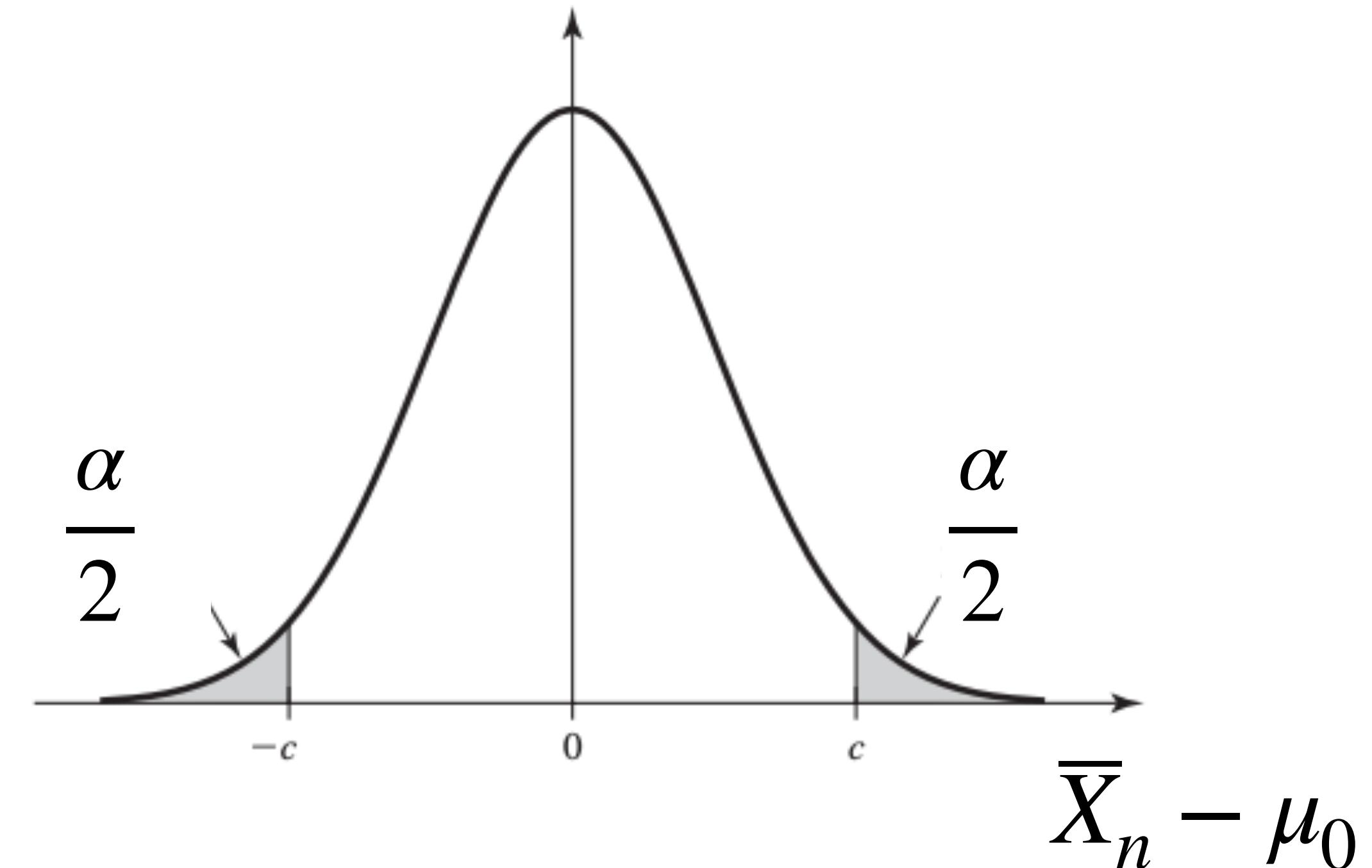
HYPOTHESIS TESTING

Hypothesis regarding mean of a normal distribution with known σ :

$$\text{Test } H_0 : \mu = \mu_0 \quad H_A : \mu \neq \mu_0$$

$$T = |\bar{X}_n - \mu_0|, \quad \text{Significance level } \alpha: P_{H_0}(T \geq c) \leq \alpha$$

Reject H_0 if $\bar{X}_n - \mu_0$ lies in shaded region.



HYPOTHESIS TESTING

Hypothesis regarding mean of a normal distribution with known σ :

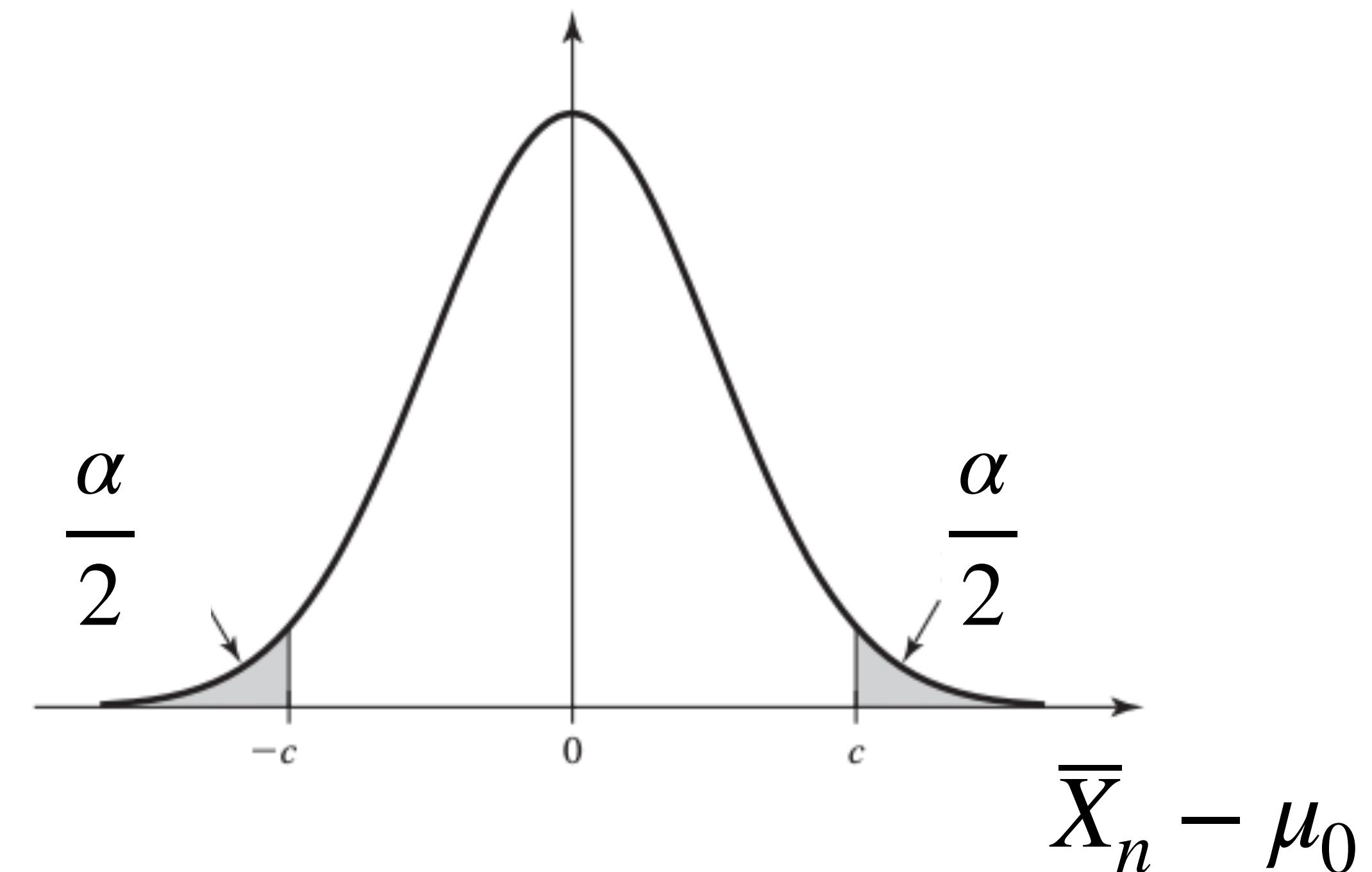
$$\text{Test } H_0 : \mu = \mu_0 \quad H_A : \mu \neq \mu_0$$

$$T = |\bar{X}_n - \mu_0|, \quad \text{Significance level } \alpha: P_{H_0}(T \geq c) \leq \alpha$$

Reject H_0 if $\bar{X}_n - \mu_0$ lies in shaded region.

Use quantile to find c .

$$c = \Phi^{-1}(1 - \alpha/2) \frac{\sigma}{\sqrt{n}}$$



HYPOTHESIS TESTING

Hypothesis regarding mean of a normal distribution with known σ :

$$\text{Test } H_0 : \mu = \mu_0 \quad H_A : \mu \neq \mu_0$$

$$T = |\bar{X}_n - \mu_0|, \quad \text{Significance level } \alpha: P_{H_0}(T \geq c) \leq \alpha$$

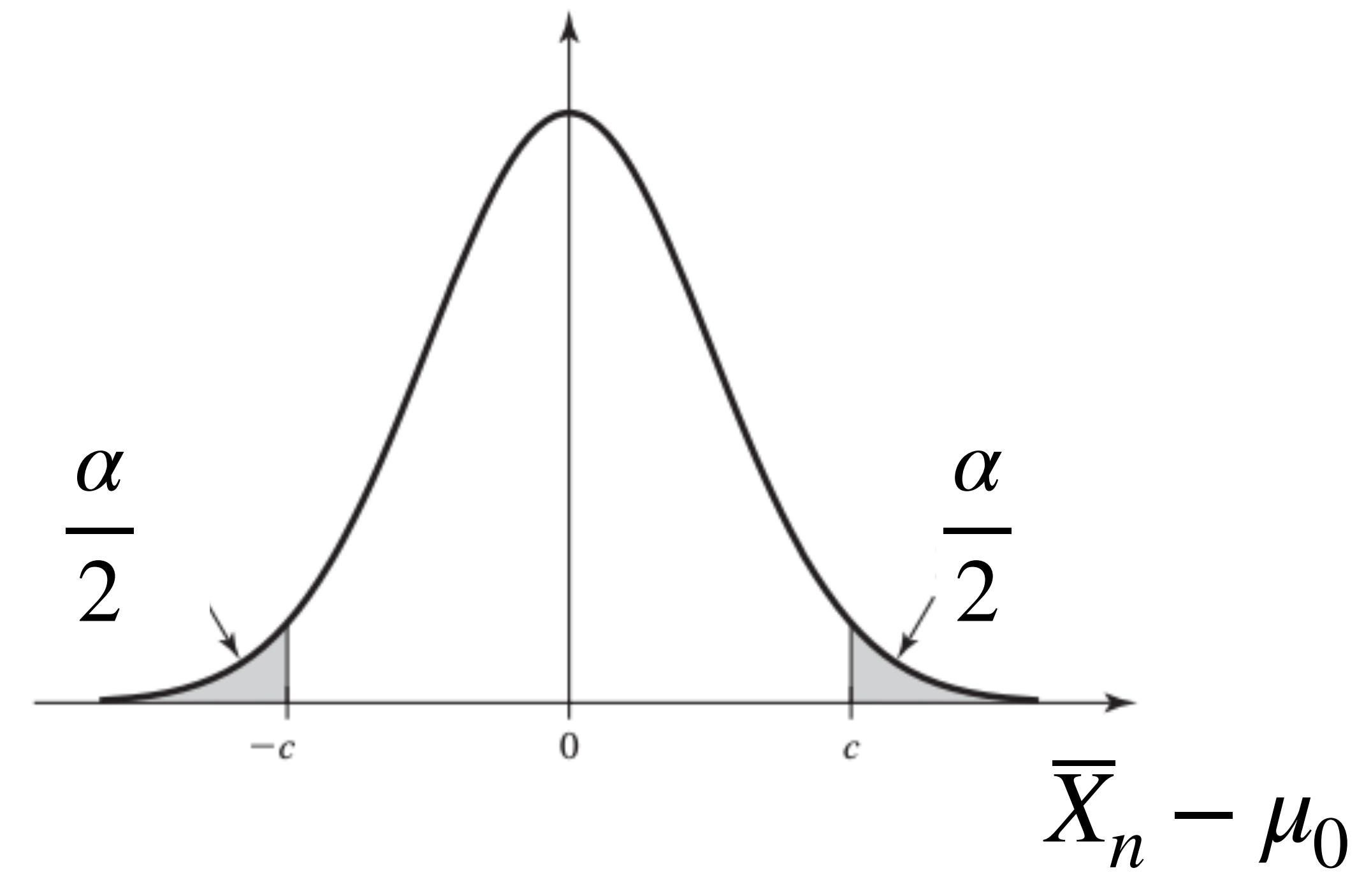
Reject H_0 if $\bar{X}_n - \mu_0$ lies in shaded region.

Use quantile to find c .

$$c = \Phi^{-1}(1 - \alpha/2) \frac{\sigma}{\sqrt{n}}$$

Calculate T for sample \bar{X}_n

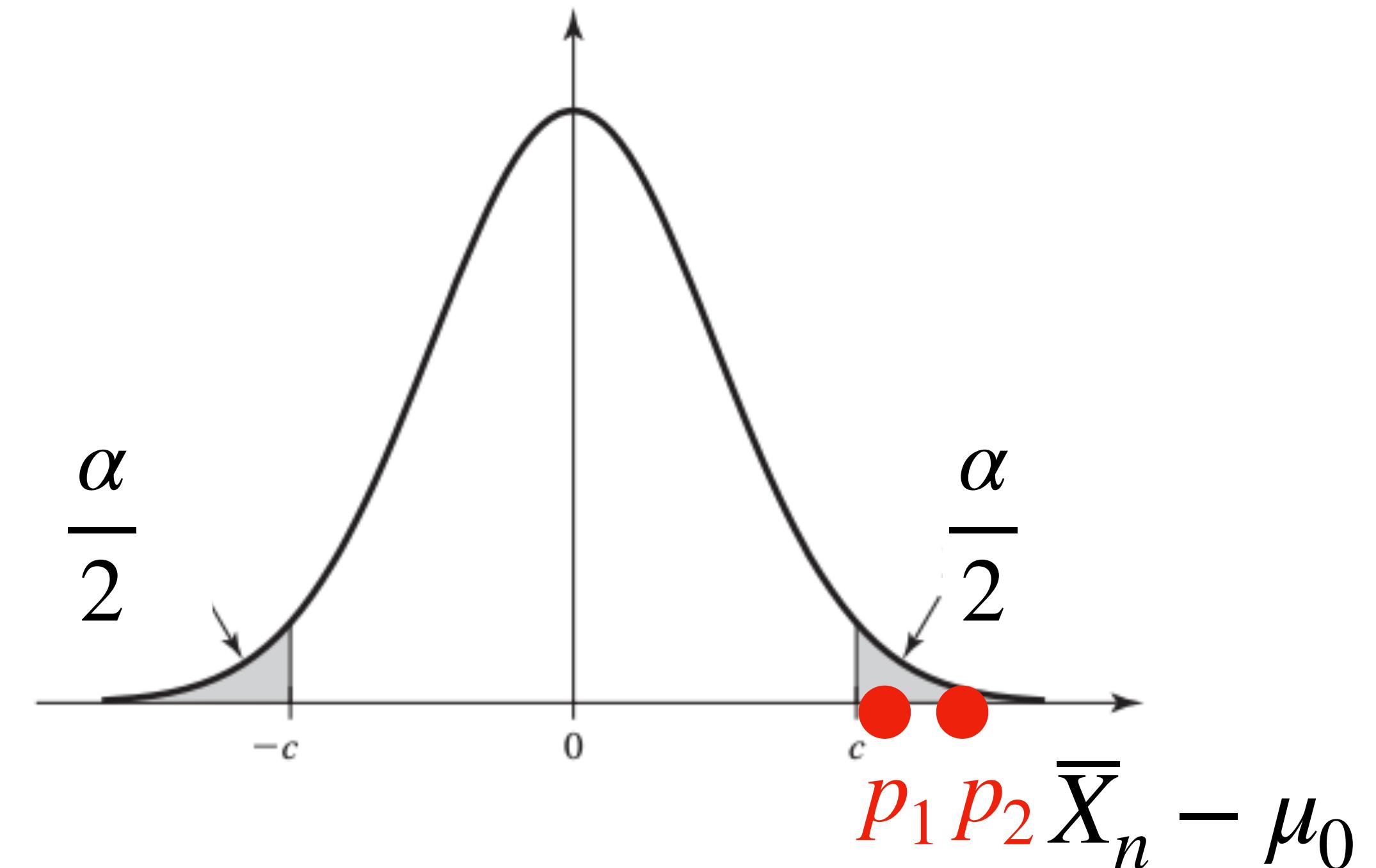
If $T \geq c$ reject H_0 at significance α
else do not reject H_0 .



HYPOTHESIS TESTING

Note: No difference between p_1, p_2 :

' H_0 rejected at significance level α ' in both cases. Not so good!



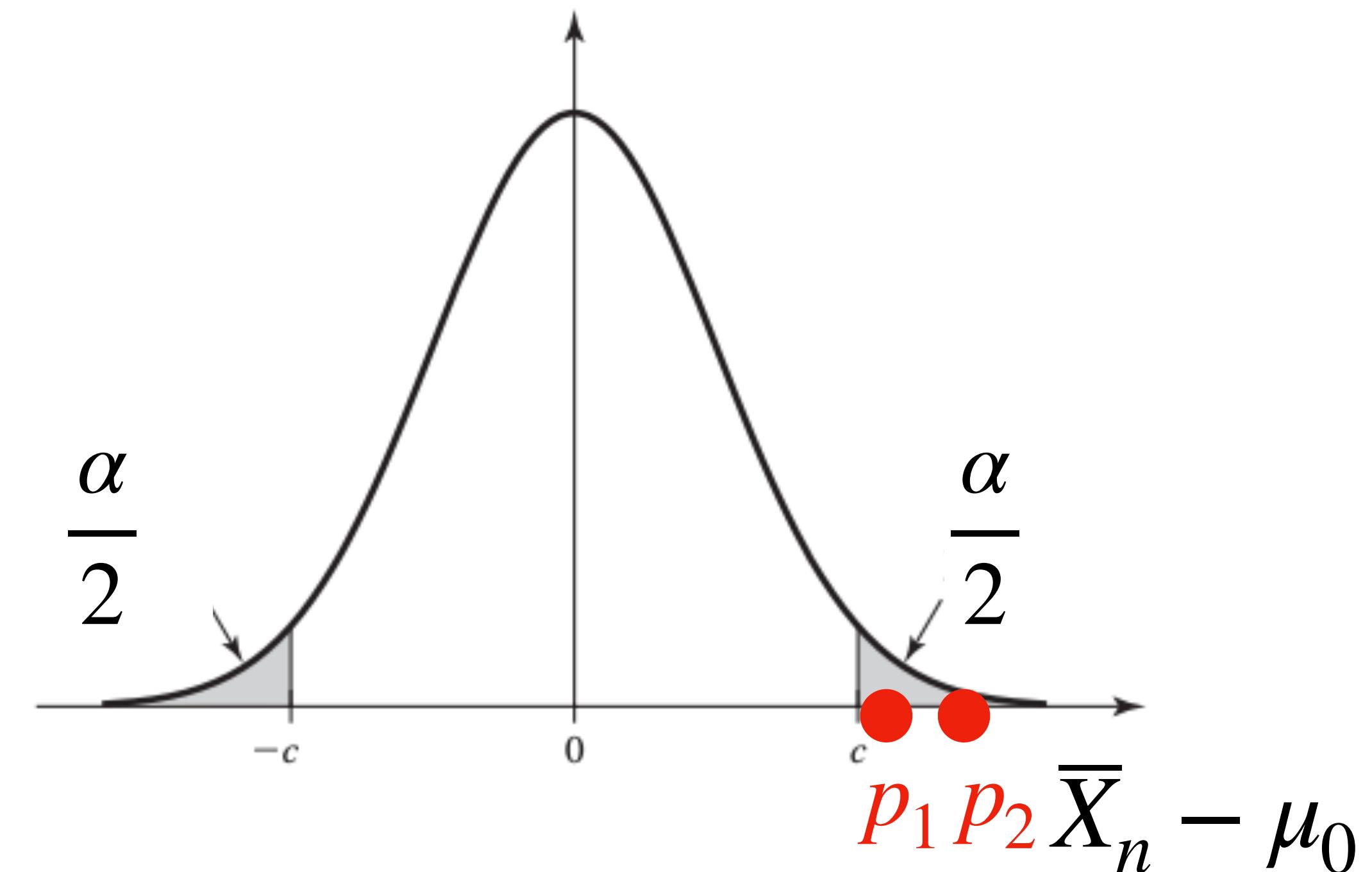
HYPOTHESIS TESTING

Note: No difference between p_1, p_2 :

' H_0 rejected at significance level α ' in both cases. Not so good!

Alternative Find probability specifically for the data: **p-value**

Probability of observing data at least as extreme as \bar{X}_n .



HYPOTHESIS TESTING

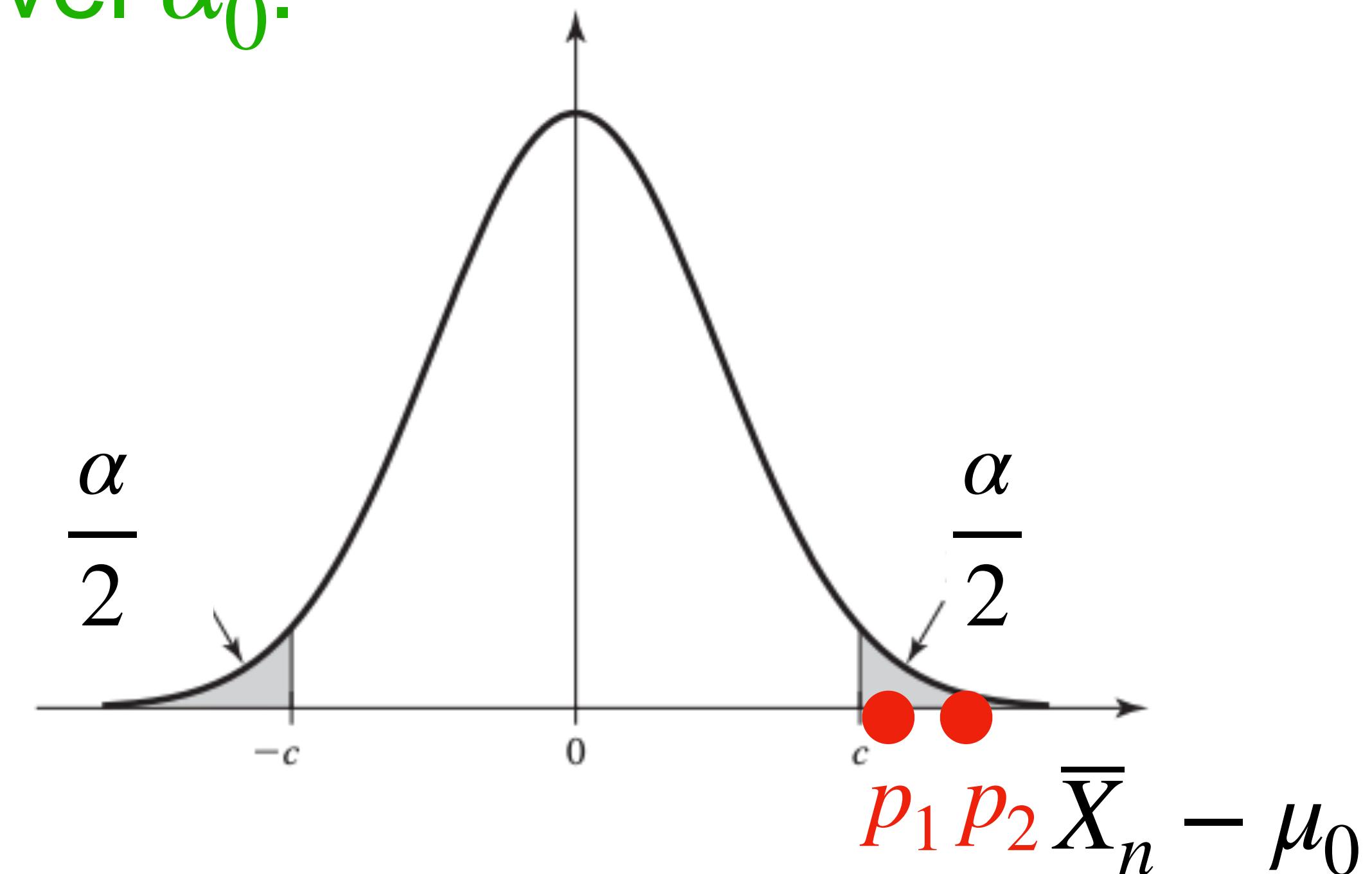
Note: No difference between p_1, p_2 :

' H_0 rejected at significance level α ' in both cases. Not so good!

Alternative Find probability specifically for the data: **p-value**

Probability of observing data at least as extreme as \bar{X}_n .

Smallest α_0 s.t. we would reject H_0 at level α_0 .



HYPOTHESIS TESTING

Note: No difference between p_1, p_2 :

' H_0 rejected at significance level α ' in both cases. Not so good!

Alternative Find probability specifically for the data: **p-value**

Probability of observing data at least as extreme as \bar{X}_n .

Smallest α_0 s.t. we would reject H_0 at level α_0 .

How? For this example:

$$Z = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sigma}, \text{ say } Z = 2.78. \quad \text{Smallest } \alpha_0 \text{ s.t. } 2.78 \geq \Phi^{-1}(1 - \alpha_0/2)$$

HYPOTHESIS TESTING

Note: No difference between p_1, p_2 :

' H_0 rejected at significance level α ' in both cases. Not so good!

Alternative Find probability specifically for the data: **p-value**

Probability of observing data at least as extreme as \bar{X}_n .

Smallest α_0 s.t. we would reject H_0 at level α_0 .

How? For this example:

$$Z = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sigma}, \text{ say } Z = 2.78. \quad \text{Smallest } \alpha_0 \text{ s.t. } 2.78 \geq \Phi^{-1}(1 - \alpha_0/2)$$

```
from scipy.stats import norm  
2*(1-norm.cdf(2.78))
```

0.005435889845402553

p value = 0.0054. Since p value ≤ 0.05
we reject H_0 at significance 0.05.

HYPOTHESIS TESTING

Note: Many different tests. We only cover a subset. Each may be one-sided/two-sided. In examples one version (the other similar).

HYPOTHESIS TESTING

Note: Many different tests. We only cover a subset. Each may be one-sided/two-sided. In examples one version (the other similar).

For this example, for instance:

Test $H_0 : \mu = \mu_0$ $H_A : \mu \neq \mu_0$

Reject H_0 if:

$$\left| \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sigma} \right| \geq \Phi^{-1}(1 - \alpha/2)$$

Test $H_0 : \mu \geq \mu_0$ $H_A : \mu < \mu_0$

$$\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sigma} \leq \Phi^{-1}(\alpha)$$

Test $H_0 : \mu \leq \mu_0$ $H_A : \mu > \mu_0$

$$\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sigma} \geq \Phi^{-1}(1 - \alpha)$$

HYPOTHESIS TESTING

Note: Many different tests. We only cover a subset. Each may be one-sided/two-sided. In examples one version (the other similar).

For this example, for instance:

$$\text{Test } H_0 : \mu = \mu_0 \quad H_A : \mu \neq \mu_0 \quad \text{p value} = 2(1 - \Phi\left(\frac{\sqrt{n}|\bar{X}_n - \mu_0|}{\sigma}\right))$$

$$\text{Test } H_0 : \mu \geq \mu_0 \quad H_A : \mu < \mu_0 \quad \text{p value} = \Phi\left(\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sigma}\right)$$

$$\text{Test } H_0 : \mu \leq \mu_0 \quad H_A : \mu > \mu_0 \quad \text{p value} = 1 - \Phi\left(\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sigma}\right)$$

HYPOTHESIS TESTING

Testing hypothesis on the mean of normal with unknown σ^2 .

HYPOTHESIS TESTING

Testing hypothesis on the mean of normal with unknown σ^2 .

Ex: # of medical in-patient days in nursing homes.

Sample of $n = 18$ normal random variables, unknown μ and σ^2 .

$\bar{X}_n = 182.17$, $\hat{\sigma}_1 = 72.22$. Test $H_0: \mu \geq 200$, $H_A: \mu < 200$

HYPOTHESIS TESTING

Testing hypothesis on the mean of normal with unknown σ^2 .

Ex: # of medical in-patient days in nursing homes.

Sample of $n = 18$ normal random variables, unknown μ and σ^2 .

$\bar{X}_n = 182.17$, $\widehat{\sigma}_1 = 72.22$. Test $H_0: \mu \geq 200$, $H_A: \mu < 200$

Already seen $\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\widehat{\sigma}_1}$ has *t-distribution* with $n - 1$ dof.

Observed value: $\frac{\sqrt{18}(182.7 - 200)}{72.22} = -1.018$

HYPOTHESIS TESTING

Testing hypothesis on the mean of normal with unknown σ^2 .

Ex: # of medical in-patient days in nursing homes.

Sample of $n = 18$ normal random variables, unknown μ and σ^2 .

$\bar{X}_n = 182.17$, $\hat{\sigma}_1 = 72.22$. Test $H_0: \mu \geq 200$, $H_A: \mu < 200$

Alternative-1 to present results:

```
from scipy.stats import t  
t.ppf(0.05, 17)
```

-1.7396067260750676

Since observed value = $-1.018 > -1.73$ we don't reject H_0 .

PROBLEM FROM PREVIOUS LECTURE

Testing hypothesis on the mean of normal with unknown σ^2 .

Ex: # of medical in-patient days in nursing homes.

Sample of $n = 18$ normal random variables, unknown μ and σ^2 .

$\bar{X}_n = 182.17$, $\widehat{\sigma}_1 = 72.22$. Test $H_0: \mu \geq 200$, $H_A: \mu < 200$

Already seen $\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\widehat{\sigma}_1}$ has *t-distribution* with $n - 1$ dof.

Observed value: $\frac{\sqrt{18}(182.7 - 200)}{72.22} = -1.018$

PROBLEM FROM PREVIOUS LECTURE

Testing hypothesis on the mean of normal with unknown σ^2 .

Ex: # of medical in-patient days in nursing homes.

Sample of $n = 18$ normal random variables, unknown μ and σ^2 .

$\bar{X}_n = 182.17$, $\hat{\sigma}_1 = 72.22$. Test $H_0: \mu \geq 200$, $H_A: \mu < 200$

Alternative-1 to present results:

```
from scipy.stats import t  
t.ppf(0.05, 17)
```

-1.7396067260750676

Since observed value = $-1.018 > -1.73$ we don't reject H_0 .

HYPOTHESIS TESTING

Testing hypothesis on the mean of normal with unknown σ^2 .

Ex: # of medical in-patient days in nursing homes.

Sample of $n = 18$ normal random variables, unknown μ and σ^2 .

$\bar{X}_n = 182.17$, $\hat{\sigma}_1 = 72.22$. Test $H_0: \mu \geq 200$, $H_A: \mu < 200$

Alternative-2 to present results:

HYPOTHESIS TESTING

Testing hypothesis on the mean of normal with unknown σ^2 .

Ex: # of medical in-patient days in nursing homes.

Sample of $n = 18$ normal random variables, unknown μ and σ^2 .

$\bar{X}_n = 182.17$, $\hat{\sigma}_1 = 72.22$. Test $H_0: \mu \geq 200$, $H_A: \mu < 200$

Alternative-2 to present results:

```
from scipy.stats import t  
t.cdf(-1.018, 17)
```

0.16147397056756546

Since p-value = 0.16 > 0.05 we don't reject H_0 .

HYPOTHESIS TESTING

Comparing sample means

Measure the arm lengths of two groups. Does it show one group has longer arms, on average? Is the difference significant?

HYPOTHESIS TESTING

Comparing sample means

Measure the arm lengths of two groups. Does it show one group has longer arms, on average? **Is the difference significant?**

Two sample t-test: Are means of two samples different significantly?

2-sided example:

H_0 : Same means

H_1 : different means

HYPOTHESIS TESTING

Comparing sample means Unknown but common variance σ^2

HYPOTHESIS TESTING

Comparing sample means Unknown but common variance σ^2

The two sample t test:

$X = (X_1, \dots, X_m)$ random sample from normal, μ_1, σ^2 unknown

$Y = (Y_1, \dots, Y_n)$ random sample from normal, μ_2, σ^2 unknown

HYPOTHESIS TESTING

Comparing sample means Unknown but common variance σ^2

Test hypotheses: $H_0: \mu_1 \leq \mu_2, H_1: \mu_1 > \mu_2$

HYPOTHESIS TESTING

Comparing sample means Unknown but common variance σ^2

Test hypotheses: $H_0: \mu_1 \leq \mu_2, H_1: \mu_1 > \mu_2$

Theorem: $T = \frac{\sqrt{m+n-2}(\bar{X}_m - \bar{Y}_n)}{\sqrt{\frac{1}{m} + \frac{1}{n}}\sqrt{\sum_{i=1}^m (X_i - \bar{X}_m)^2 + \sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}$ has *t distribution* with $m+n-2$ dof.

At significance α , reject H_0 if $T \geq F_v^{-1}(1 - \alpha)$, for $v = m + n - 2$ dof.

Alternatively present p value too: $1 - F_v(T)$, for $v = m + n - 2$ dof.

HYPOTHESIS TESTING

Comparing sample means Unknown variances σ_1^2, σ_2^2 different.

HYPOTHESIS TESTING

Comparing sample means Unknown variances σ_1^2, σ_2^2 different.
Use Welch's t statistic.

HYPOTHESIS TESTING

Comparing sample means Unknown variances σ_1^2, σ_2^2 different.

Use Welch's t statistic.

Ex: Effect of a calcium supplement on blood pressure.

10 people received calcium supplement. 11 received placebo.

Calcium	7	-4	18	17	-3	-5	1	10	11	-2
Placebo	-1	12	-1	-3	3	5	2	-11	-1	-3

Each entry:

Blood pressure after 12 weeks –
Blood pressure before experiment

HYPOTHESIS TESTING

Comparing sample means Unknown variances σ_1^2, σ_2^2 different.

Use Welch's t statistic.

Ex: Effect of a calcium supplement on blood pressure.

10 people received calcium supplement. 11 received placebo.

Calcium	7	-4	18	17	-3	-5	1	10	11	-2
Placebo	-1	12	-1	-3	3	5	2	-11	-1	-3

Each entry:

Blood pressure after 12 weeks –
Blood pressure before experiment

Does the calcium supplement increase blood pressure?

μ_1 : mean change in blood pressure of calcium supplement group

μ_2 : mean change in blood pressure of placebo group

$H_0: \mu_1 \leq \mu_2, H_A: \mu_1 > \mu_2$

Test significance at $\alpha = 0.05$.

HYPOTHESIS TESTING

Calcium supplement group vs placebo

```
import scipy.stats as stats
import numpy as np

calcium_sample = np.array([7, -4, 18, 17, -3, -5, 1,
                           10, 11, -2])
np.mean(calcium_sample)
```

5.0

```
np.var(calcium_sample)
```

68.8

```
placebo_sample = np.array([-1, 12, -1, -3, 3, -5, 5,
                           2, -11, -1, -3])
np.mean(placebo_sample)
```

-0.2727272727272727

```
np.var(placebo_sample)
```

31.65289256198347

```
stats.ttest_ind(calcium_sample, placebo_sample, equal_var = False, alternative = 'greater')
```

TtestResult(statistic=1.6037172876755148, pvalue=0.06441968481096698, df=15.590512968733774)

HYPOTHESIS TESTING

Calcium supplement group vs placebo

```
import scipy.stats as stats
import numpy as np

calcium_sample = np.array([7, -4, 18, 17, -3, -5, 1,
                          10, 11, -2])
np.mean(calcium_sample)
```

5.0

```
np.var(calcium_sample)
```

68.8

```
placebo_sample = np.array([-1, 12, -1, -3, 3, -5, 5,
                           2, -11, -1, -3])
```

```
np.mean(placebo_sample)
```

-0.2727272727272727

```
np.var(placebo_sample)
```

31.65289256198347

Note 1: For Welch's t test, set to False

```
stats.ttest_ind(calcium_sample, placebo_sample, equal_var = False, alternative = 'greater')
```

TtestResult(statistic=1.6037172876755148, pvalue=0.06441968481096698, df=15.590512968733774)

HYPOTHESIS TESTING

Calcium supplement group vs placebo

```
import scipy.stats as stats
import numpy as np

calcium_sample = np.array([7, -4, 18, 17, -3, -5, 1,
                          10, 11, -2])
np.mean(calcium_sample)
```

5.0

```
np.var(calcium_sample)
```

68.8

```
placebo_sample = np.array([-1, 12, -1, -3, 3, -5, 5,
                           2, -11, -1, -3])
```

```
np.mean(placebo_sample)
```

-0.2727272727272727

```
np.var(placebo_sample)
```

31.65289256198347

Note 2: When H_A is $\mu_1 > \mu_2$.

```
stats.ttest_ind(calcium_sample, placebo_sample, equal_var = False, alternative = 'greater')
```

TtestResult(statistic=1.6037172876755148, pvalue=0.06441968481096698, df=15.590512968733774)

HYPOTHESIS TESTING

Calcium supplement group vs placebo

```
import scipy.stats as stats
import numpy as np

calcium_sample = np.array([7, -4, 18, 17, -3, -5, 1,
                           10, 11, -2])
np.mean(calcium_sample)
```

5.0

```
np.var(calcium_sample)
```

68.8

```
placebo_sample = np.array([-1, 12, -1, -3, 3, -5, 5,
                           2, -11, -1, -3])
np.mean(placebo_sample)
```

-0.2727272727272727

```
np.var(placebo_sample)
```

31.65289256198347

Can not reject H_0 at 0.05 significance level

```
stats.ttest_ind(calcium_sample, placebo_sample, equal_var = False, alternative = 'greater')
```

TtestResult(statistic=1.6037172876755148, pvalue=0.06441968481096698, df=15.590512968733774)

THE χ^2 TEST

Given explanatory variable and response variable (both categorical)
test for association between the variables.

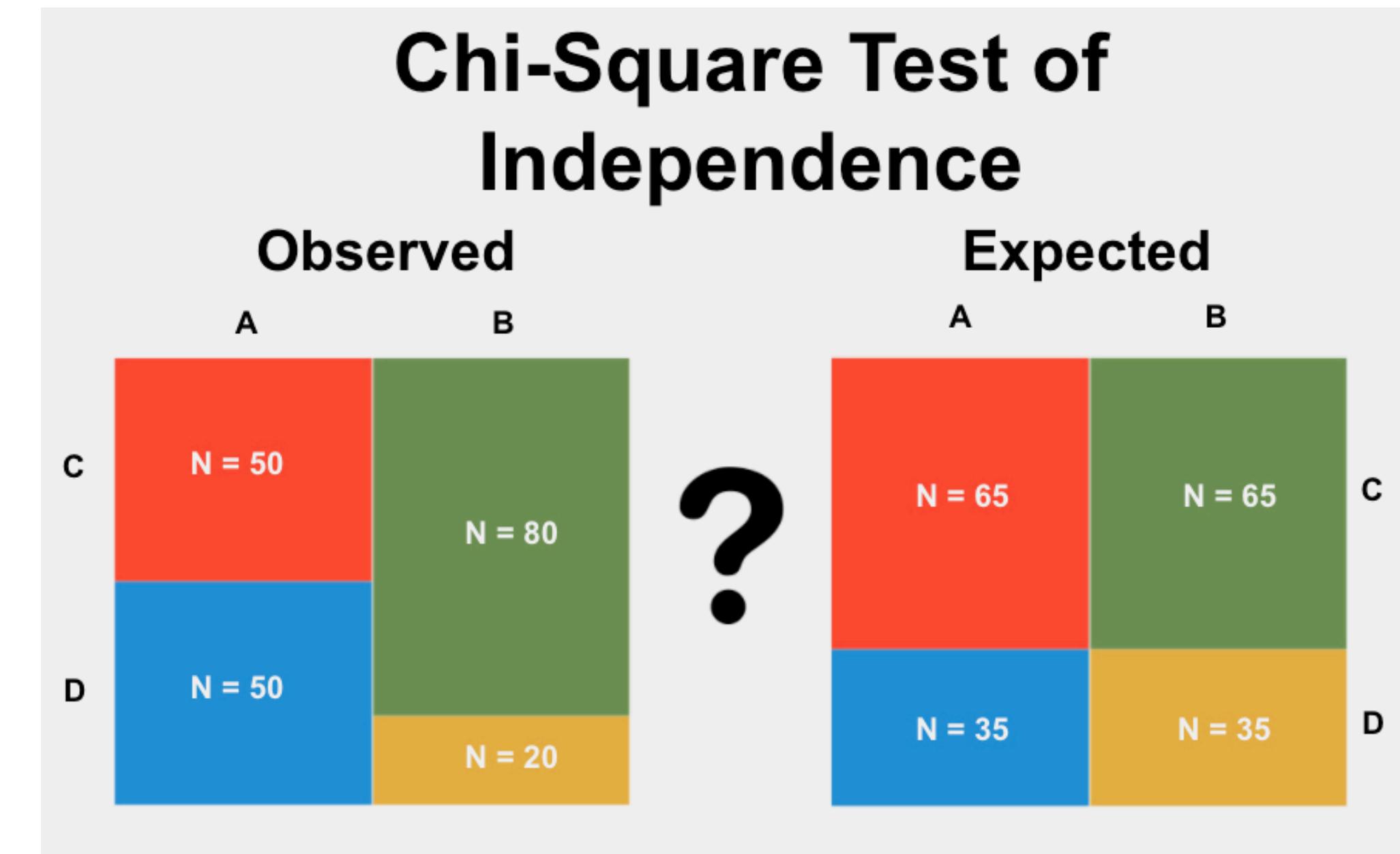
H_0 : No association between variables

THE χ^2 TEST

Given explanatory variable and response variable (both categorical) test for association between the variables.

H_0 : No association between variables

- Build contingency table
- Find out expected values
- Compute p value of χ^2 statistic
- If p value significant
reject H_0 and claim association
else
not enough evidence for association



THE χ^2 TEST

Ex: Is cancer independent of marital status?

Marital Status	Cancer	No Cancer	Total
Single	29 (38.1)	47 (37.9)	76
Married	116 (112.3)	108 (111.7)	224
Widowed	67 (61.6)	56 (61.4)	123
Div/Sep	5 (5.0)	5 (5.0)	10
Total	217	216	433

How to find expected values?
(Green shaded values)

Assuming independence:
(row sum x column sum) / total

THE χ^2 TEST

Ex: Is cancer independent of marital status?

Marital Status	Cancer	No Cancer	Total
Single	29 (38.1)	47 (37.9)	76
Married	116 (112.3)	108 (111.7)	224
Widowed	67 (61.6)	56 (61.4)	123
Div/Sep	5 (5.0)	5 (5.0)	10
Total	217	216	433

How to find expected values?
(Green shaded values)

Assuming independence:
(row sum x column sum) / total

How to compute χ^2 statistics?

$$X^2 = \sum \frac{(observed - expected)^2}{expected} = \frac{(29 - 38.1)^2}{38.1} + \frac{(47 - 37.9)^2}{37.9} + \dots + \frac{(5 - 5.0)^2}{5.0} = 5.53$$

THE χ^2 TEST

Ex: Is cancer independent of marital status?

Marital Status	Cancer	No Cancer	Total
Single	29 (38.1)	47 (37.9)	76
Married	116 (112.3)	108 (111.7)	224
Widowed	67 (61.6)	56 (61.4)	123
Div/Sep	5 (5.0)	5 (5.0)	10
Total	217	216	433

How to find expected values?
(Green shaded values)

Assuming independence:
(row sum x column sum) / total

How to compute χ^2 statistics?

$$X^2 = \sum \frac{(observed - expected)^2}{expected} = \frac{(29 - 38.1)^2}{38.1} + \frac{(47 - 37.9)^2}{37.9} + \dots + \frac{(5 - 5.0)^2}{5.0} = 5.53$$

```
from scipy.stats import chi2  
1-chi2.cdf(5.53, 3)
```

0.13685525539157362

THE χ^2 TEST

Ex: Is cancer independent of marital status?

Marital Status	Cancer	No Cancer	Total
Single	29 (38.1)	47 (37.9)	76
Married	116 (112.3)	108 (111.7)	224
Widowed	67 (61.6)	56 (61.4)	123
Div/Sep	5 (5.0)	5 (5.0)	10
Total	217	216	433

How to find expected values?
(Green shaded values)

Assuming independence:
(row sum x column sum) / total

How to compute χ^2 statistics?

$$X^2 = \sum \frac{(observed - expected)^2}{expected} = \frac{(29 - 38.1)^2}{38.1} + \frac{(47 - 37.9)^2}{37.9} + \dots + \frac{(5 - 5.0)^2}{5.0} = 5.53$$

```
from scipy.stats import chi2  
1-chi2.cdf(5.53, 3)
```

dof = (row_count - 1)(column_count - 1)

0.13685525539157362

THE χ^2 TEST

Ex: Is cancer independent of marital status?

Marital Status	Cancer	No Cancer	Total
Single	29 (38.1)	47 (37.9)	76
Married	116 (112.3)	108 (111.7)	224
Widowed	67 (61.6)	56 (61.4)	123
Div/Sep	5 (5.0)	5 (5.0)	10
Total	217	216	433

How to find expected values?
(Green shaded values)

Assuming independence:
(row sum x column sum) / total

How to compute χ^2 statistics?

$$X^2 = \sum \frac{(observed - expected)^2}{expected} = \frac{(29 - 38.1)^2}{38.1} + \frac{(47 - 37.9)^2}{37.9} + \dots + \frac{(5 - 5.0)^2}{5.0} = 5.53$$

```
from scipy.stats import chi2  
1-chi2.cdf(5.53, 3)
```

0.13685525539157362

p value of test statistic: 0.137

At significance level 0.05 won't reject H_0

THE χ^2 TEST

Ex: Is cancer independent of marital status?

Marital Status	Cancer	No Cancer	Total
Single	29 (38.1)	47 (37.9)	76
Married	116 (112.3)	108 (111.7)	224
Widowed	67 (61.6)	56 (61.4)	123
Div/Sep	5 (5.0)	5 (5.0)	10
Total	217	216	433

How to find expected values?
(Green shaded values)

Assuming independence:
(row sum x column sum) / total

How to compute χ^2 statistics?

$$X^2 = \sum \frac{(observed - expected)^2}{expected} = \frac{(29 - 38.1)^2}{38.1} + \frac{(47 - 37.9)^2}{37.9} + \dots + \frac{(5 - 5.0)^2}{5.0} = 5.53$$

```
from scipy.stats import chi2  
1-chi2.cdf(5.53, 3)
```

0.13685525539157362

p value of test statistic: 0.137

At significance level 0.05 won't reject H_0

⇒ Can't conclude marital status associated with cancer.

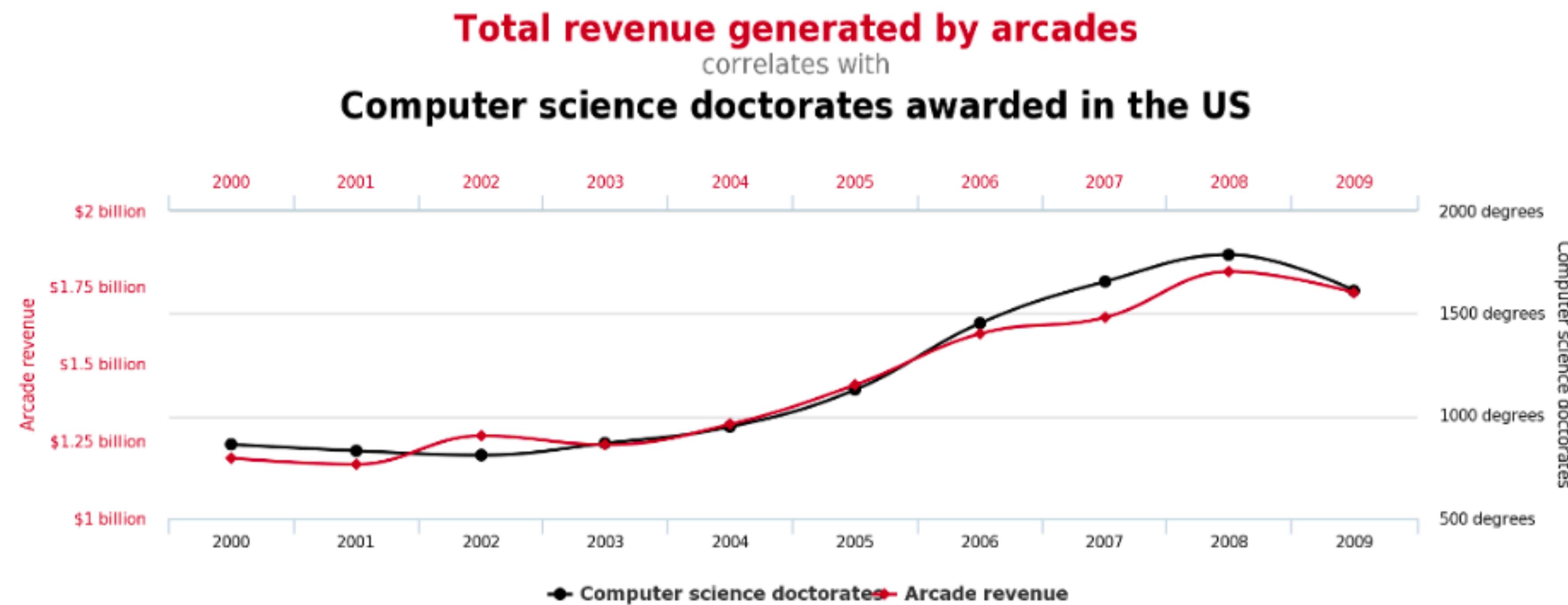
READINGS ON CONFIDENCE INTERVALS, HYPOTHESIS TESTING

Wasserman: 6.3, 10.2, 10.3 , 10.7

Watkins: 16.1, 16.2, 17.1, 18.1, 18.2, 18.3, 18.5, 20.1, 20.2, 20.3,
20.5, 20.6, 21, 2

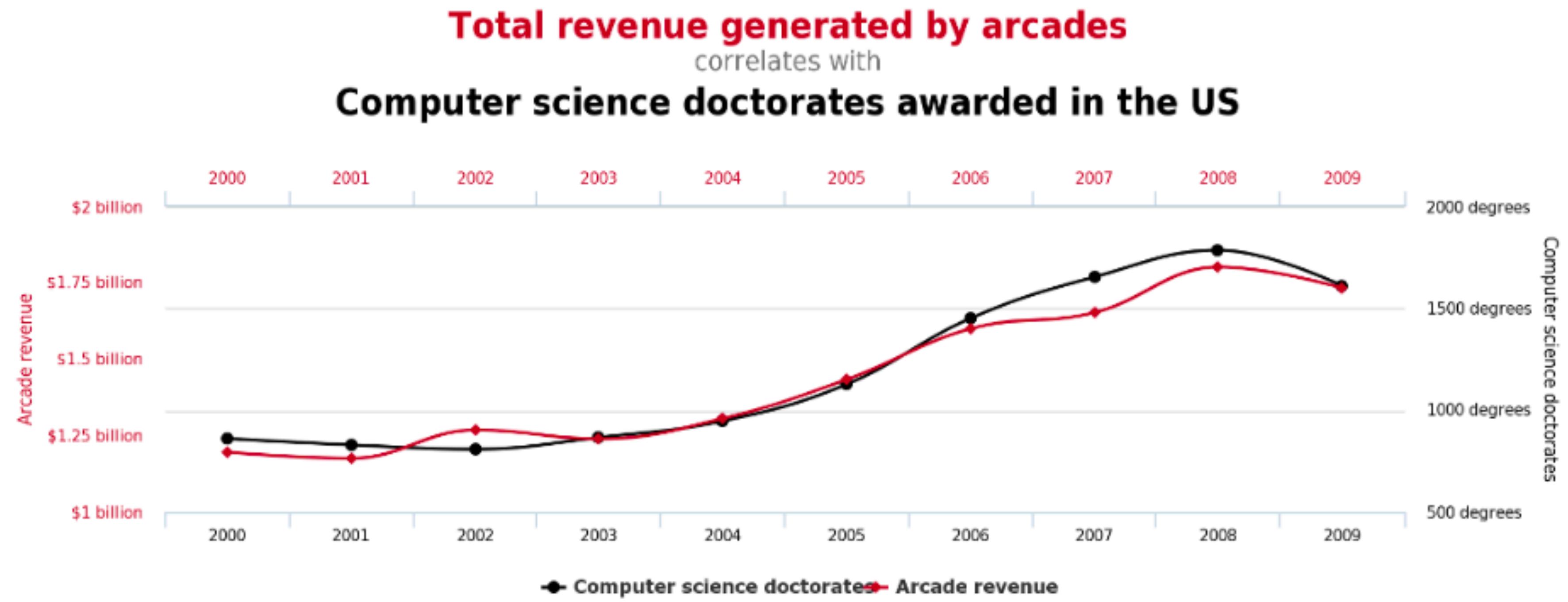
One last topic before we end our discussion on hypothesis testing!

MULTIPLE TESTING CORRECTION



tylervigen.com

MULTIPLE TESTING CORRECTION



Thousands of hypotheses tested (correlations of pairs of variables).

Significance level of 0.05 not a high standard in this case:
Buying a single lottery ticket and winning vs
Buying millions of tickets and winning.

MULTIPLE TESTING CORRECTION

Bonferroni Correction:

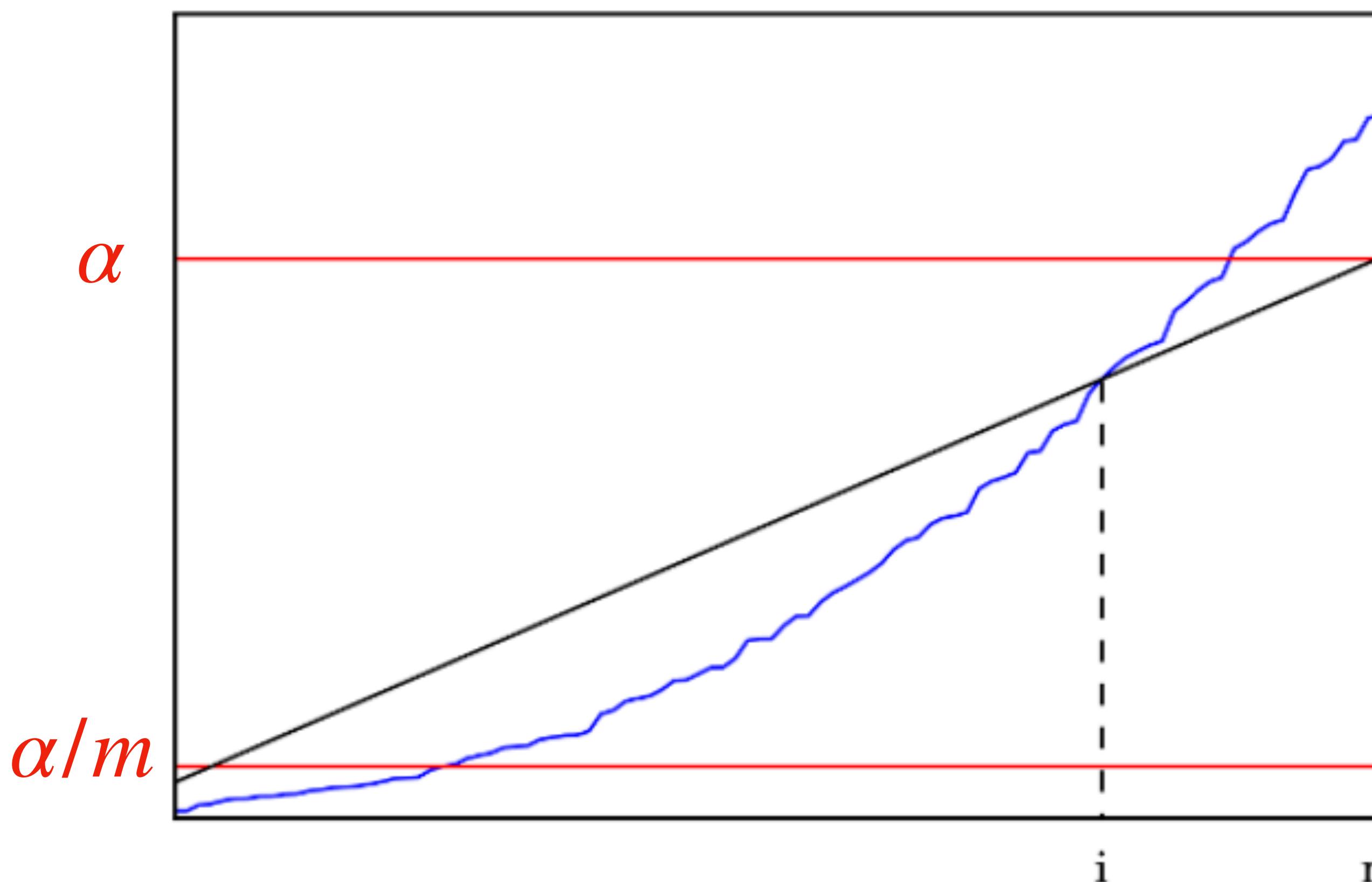
If testing m hypothesis \Rightarrow to be significant at α , p-value against α/m .

MULTIPLE TESTING CORRECTION

Bonferroni Correction:

If testing m hypothesis \Rightarrow to be significant at α , p-value against α/m .

Benjamini Hochberg Procedure:



p_i : p-value of the i^{th} hypothesis

Sort all p_i in ascending order.

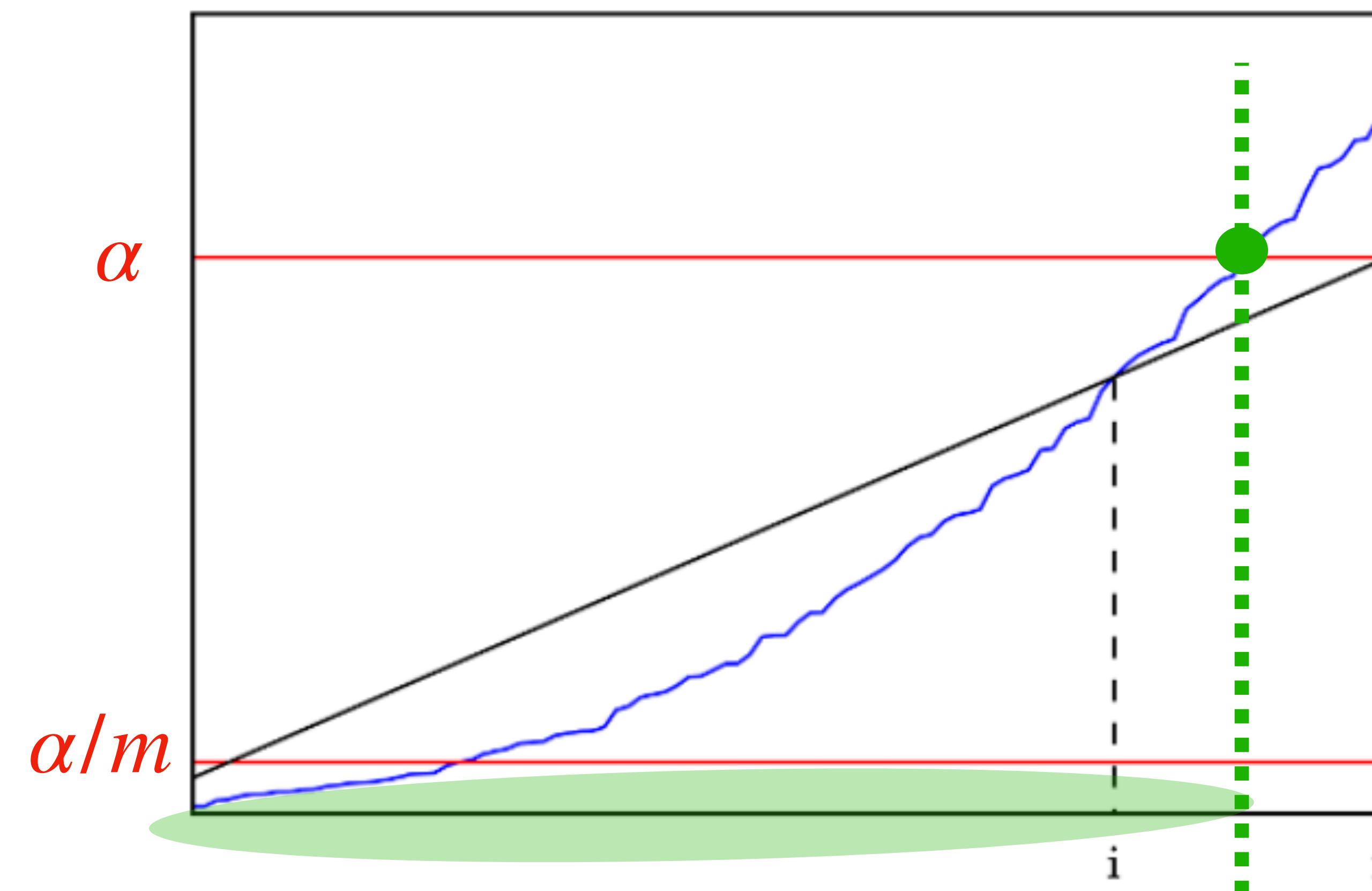
Blue curve: p-values

MULTIPLE TESTING CORRECTION

Bonferroni Correction:

If testing m hypothesis \Rightarrow to be significant at α , p-value against α/m .

Benjamini Hochberg Procedure:



p_i : p-value of the i^{th} hypothesis

Sort all p_i in ascending order.

Blue curve: p-values

If check against α , each j in shaded region is significant.

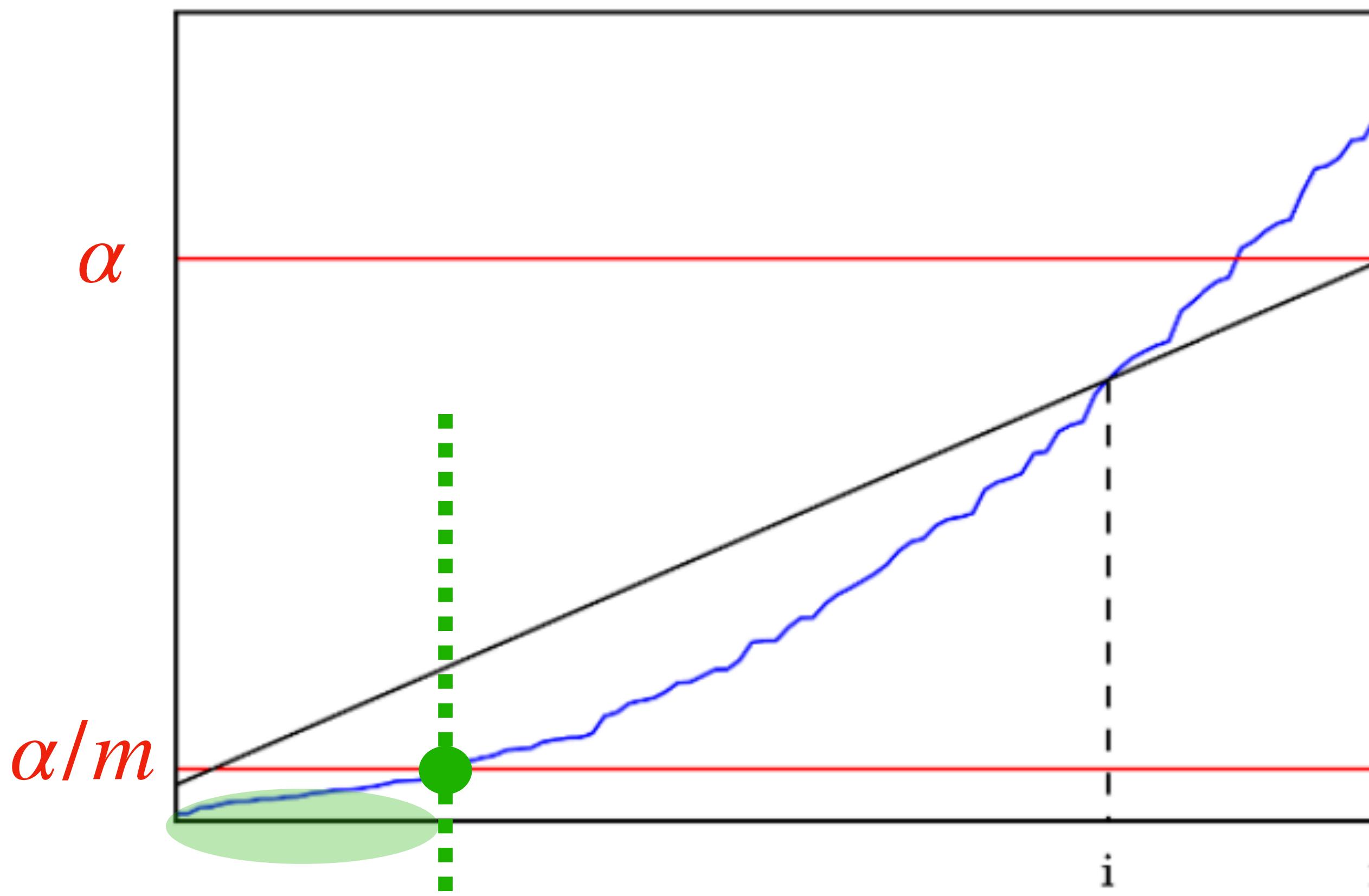
No correction! Too relaxed.

MULTIPLE TESTING CORRECTION

Bonferroni Correction:

If testing m hypothesis \Rightarrow to be significant at α , p-value against α/m .

Benjamini Hochberg Procedure:



p_i : p-value of the i^{th} hypothesis

Sort all p_i in ascending order.

Blue curve: p-values

If check against α/m , each j in shaded region is significant.

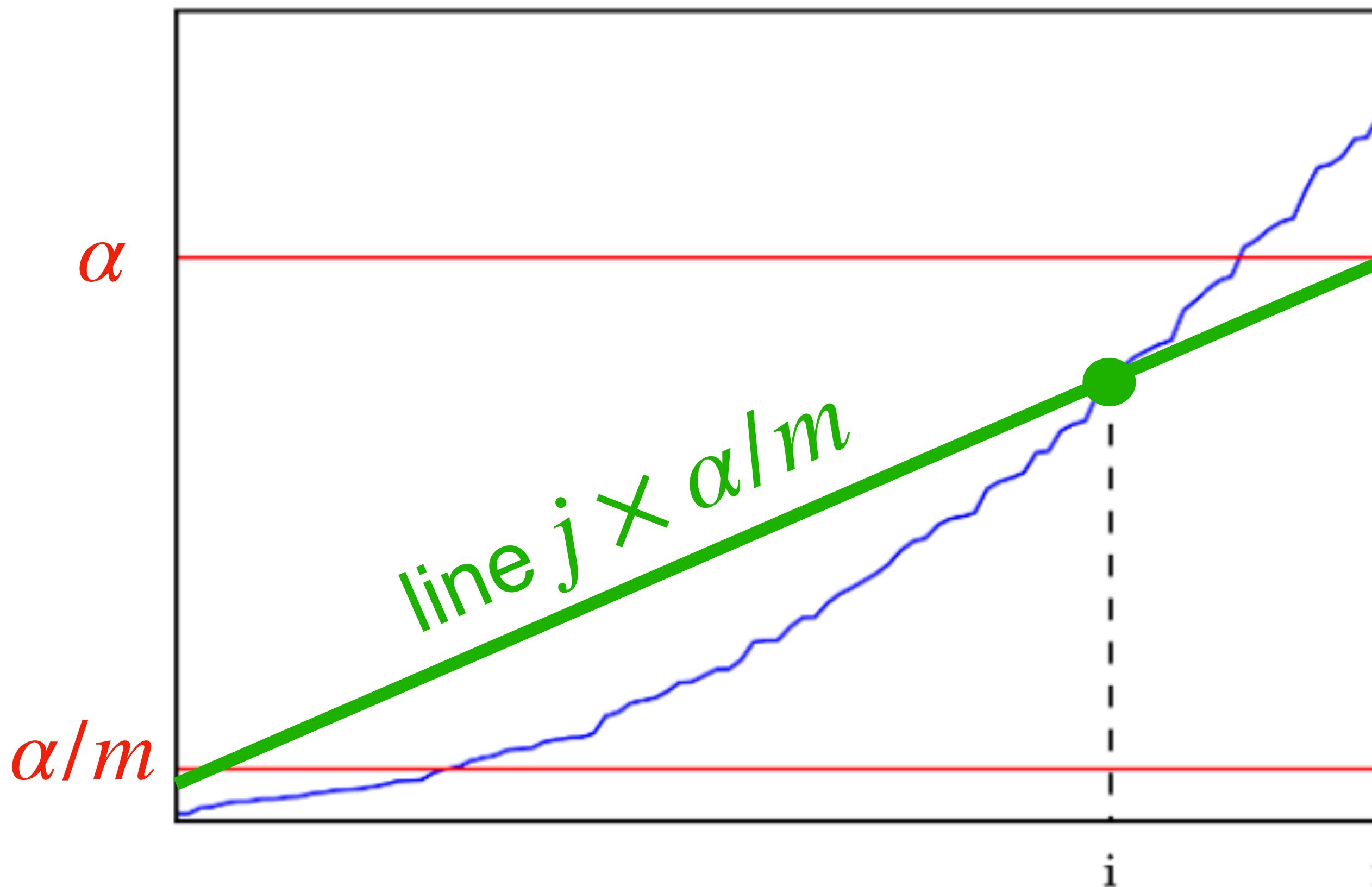
Bonferroni correction! Too strict.

MULTIPLE TESTING CORRECTION

Bonferroni Correction:

If testing m hypothesis \Rightarrow to be significant at α , p-value against α/m .

Benjamini Hochberg Procedure:



p_i : p-value of the i^{th} hypothesis

Sort all p_i in ascending order.

Blue curve: p-values

Let j represent x-coord.

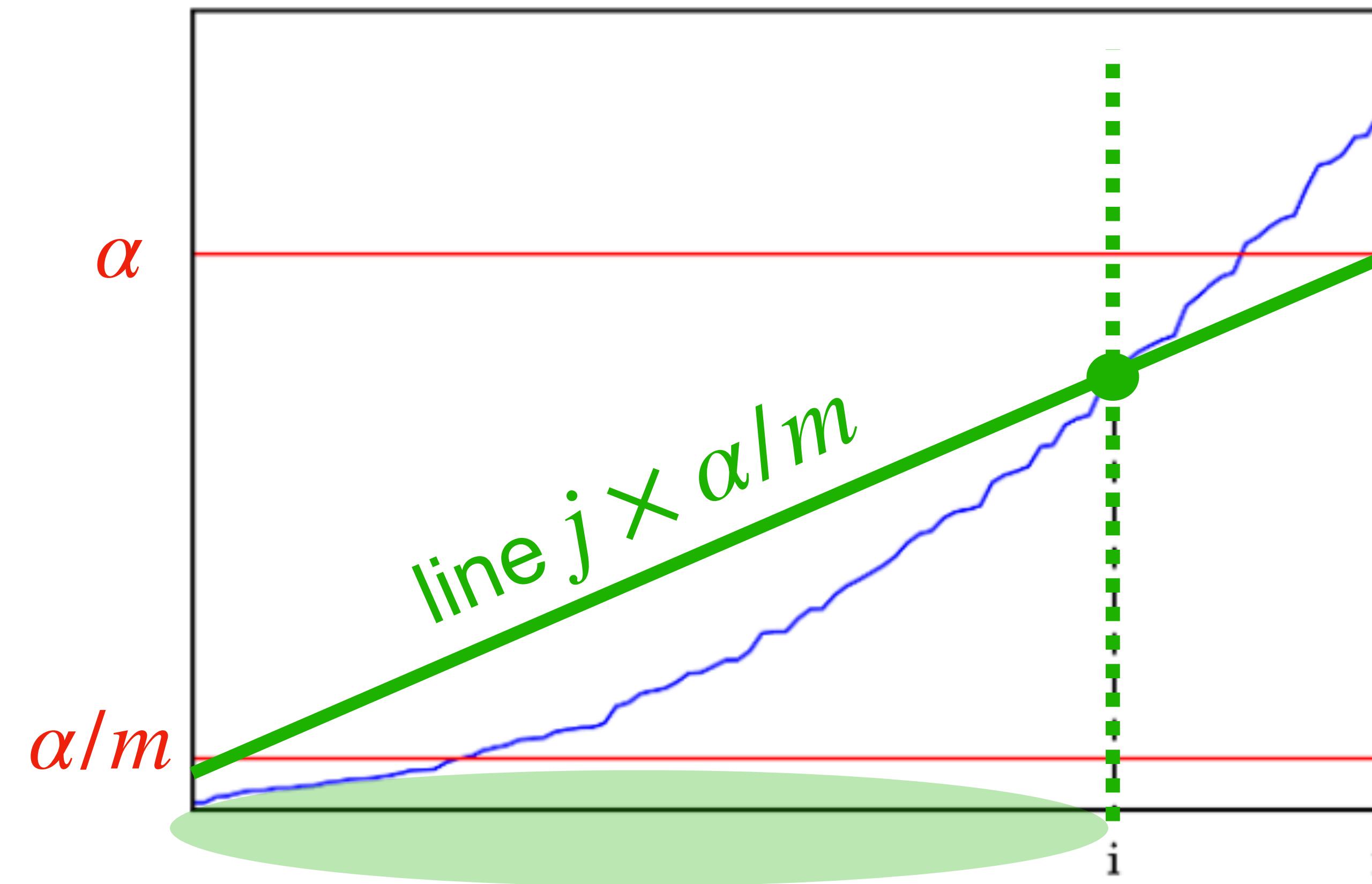
Consider line $j \times \alpha/m$.

MULTIPLE TESTING CORRECTION

Bonferroni Correction:

If testing m hypothesis \Rightarrow to be significant at α , p-value against α/m .

Benjamini Hochberg Procedure:



p_i : p-value of the i^{th} hypothesis

Sort all p_i in ascending order.

Blue curve: p-values

Let j represent x-coord.

Consider line $j \times \alpha/m$.

Variables up to i , where

$\forall_{j=1}^i p_j \leq j \times \alpha/m$ are significant.

PRELIMINARIES FOR EDA IN PYTHON

A notebook for a quick review of NumPy and Pandas at:

https://colab.research.google.com/drive/17Zx64-WzB6Mokvy3CN_fsQmyGCP7dj_C?usp=sharing

For further practice on NumPy:

Chapter 4 from Python for Data Analysis by Wes McKinney

<https://wesmckinney.com/book/numpy-basics>

For further practice on Pandas:

Chapter 5 from Python for Data Analysis by Wes McKinney

<https://wesmckinney.com/book/pandas-basics>

Chapter 6 from Learning Data Science by Sam Lau et al.

https://learningds.org/ch/06/pandas_intro.html

MORE ON DATAFRAMES

A notebook for an exploratory analysis of TED talks data at:

<https://colab.research.google.com/drive/1Ohqj99tdIPNkZ6fsrvKs6jjVMJ335S8X?usp=sharing>

MORE ON DATAFRAMES

A notebook for an exploratory analysis of TED talks data at:

<https://colab.research.google.com/drive/1Ohqj99tdIPNkZ6fsrvKs6jjVMJ335S8X?usp=sharing>

DATA WRANGLING

Reading: Ch3 from Skiena

- Acquiring data and preparing it for analysis
- Data Pipelines:

Notebooks make it easier to maintain data pipelines, the sequence of processing steps from start to finish.

Expect to have to redo your analysis from scratch, so build your code to make it possible.

LANGUAGES

- Python: many libraries and features (e.g regular expressions)
- R: programming language of statisticians.
- Matlab: fast and efficient matrix operations.
- Java/C: for Big Data systems.
- Excel: good for EDA
- ...

COMMON DATA FORMATS

- CSV (comma separated value):

A small part of ted_main.csv file we have seen earlier

```
comments,description,duration,event,film_date,languages,main_speaker,name,num_speaker,published_date,ratings,related_talks,speaker_occupation,tags,title,url,views
4553,Sir Ken Robinson makes an entertaining and profoundly moving case for creating an education system that nurtures (rather than undermines) creativity.,1164,TED2006,1140825600,60,Ken Robinson,Ken Robinson: Do schools kill creativity?,1,1151367060,"[{"id": 7, "name": "Funny", "count": 19645}, {"id": 1, "name": "Beautiful", "count": 4573}, {"id": 9, "name": "Ingenious", "count": 6073}, {"id": 3, "name": "Courageous", "count": 3253}, {"id": 11, "name": "Longwinded", "count": 387}, {"id": 2, "name": "Confusing", "count": 242}, {"id": 8, "name": "Informative", "count": 7346}, {"id": 22, "name": "Fascinating", "count": 10581}, {"id": 21, "name": "Unconvincing", "count": 300}, {"id": 24, "name": "Persuasive", "count": 10704}, {"id": 23, "name": "Jaw-dropping", "count": 4439}, {"id": 25, "name": "OK", "count": 1174}, {"id": 26, "name": "Obnoxious", "count": 209}, {"id": 10, "name": "Inspiring", "count": 24924}]", "[{"id": 865, "hero": "https://pe.tedcdn.com/images/ted/172559_800x600.jpg", "speaker": "Ken Robinson", "title": "Bring on the learning revolution!", "duration": 1008, "slug": "sir_ken_robinson_bring_on_the_revolution", "viewed_count": 7266103}, {"id": 1738, "hero": "https://pe.tedcdn.com/images/ted/de98b161ad1434910ff4b56c89de71af04b8b873_1600x1200.jpg", "speaker": "Ken Robinson", "title": "\"How to escape education's death valley\"", "duration": 1151, "slug": "ken_robinson_how_to_escape_education_s_death_valley", "viewed_count": 6657572}, {"id": 2276, "hero": "https://pe.tedcdn.com/images/ted/3821f3728e0b755c7b9aea2e69cc093eca41abe1_2880x1620.jpg", "speaker": "Linda Cliatt-Wayman", "title": "How to fix a broken school? Lead fearlessly, love hard", "duration": 1027, "slug": "linda_cliatt_wayman_how_to_fix_a_broken_school_lead_fearlessly_love_hard", "viewed_count": 1617101}, {"id": 892, "hero": "https://pe.tedcdn.com/images/ted/e79958940573cc610ccb583619a54866c41ef303_2880x1620.jpg", "speaker": "Charles Leadbeater", "title": "Education innovation in the slums", "duration": 1138, "slug": "charles_leadbeater_on_education", "viewed_count": 772296}, {"id": 1232, "hero": "https://pe.tedcdn.com/images/ted/
```

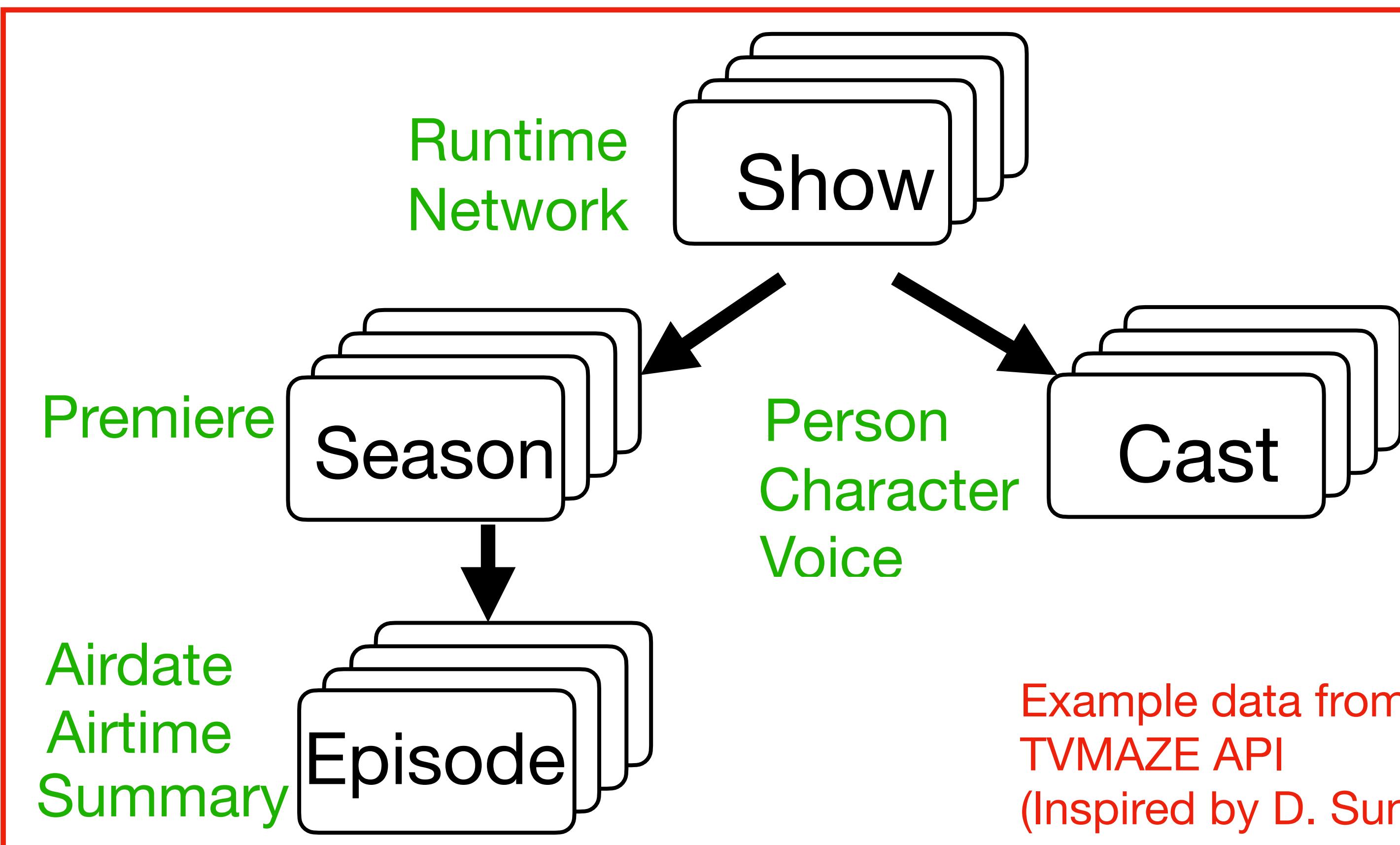
Can easily convert to a pandas dataframe for further processing:

```
ted_main = pd.read_csv('content/drive/My  
Drive/datasets/ted_main.csv')
```

COMMON DATA FORMATS

- JSON (JavaScript Object Notation):

```
[{'name': 'Black Mirror', 'runtime': 60, 'network': {'name': '', ...},  
 'cast': [{'person': {'name': ...}, 'character': {'name': ...}, 'voice': False}, ...],  
 'seasons': [{'premiereDate': '2012-04-15', 'episodes': [...]}, ...], ...]
```



Hierarchical data

COMMON DATA FORMATS

- XML (eXtensible Markup Language):
- Like JSON, can represent hierarchical data:
 - Fields shown with tags
 - Each tag with open <...> and close </...>
 - Children shown with nested tags
 - Repeated fields shown with repeated tags
- Can use BeautifulSoup library to read XML data into a BeautifulSoup object, which represents the data as a tree.

```
<?xml version="1.0" encoding="UTF-8"?>
<root>
  <show>
    <name>Black Mirror</name>
    <runtime>60</runtime>
    <cast>
      <person>...</person>
      <character>....</character>
    </cast>
    <cast>
      ...
    </cast>
    <season>
      <episode>...</episode>
      <episode>...</episode>
      ...
    </season>
    <season>
      ...
    </season>
    <show>
  </root>
```

Example data
Inspired by D. Sun

DATA SOURCES

- Academic data sets
Some journals require datasets to be published, some digging will be necessary. Economic, medical, demographic, historical, and scientific data.
- Sensor data
- Other important data sources in more detail in the next slides...

DATA SOURCES

- Government data sets

Data from City, State, and Federal governments. May differ from city/state...

ST_NUM ↗	CITY ↗	ZIP_CODE ↗	OWN_OCC ↗	OWNER ↗	MAIL_STREET_ADDRESS ↗	LAND_SF ↗	GROSS_AREA ↗	LIVING_AREA ↗	LAND_VALUE ↗	BLDG_VALUE ↗	TOTAL_VALUE ↗	GROSS_TAX ↗	YR_BUILT ↗	YR_REMODEL ↗
104	EAST BOSTON	02128	Y	PASCUCCI CARLO	195 LEXINGTON ST	1,150	3353	2202	197,600	594,400	792,000	\$8,632.80	1900	
197	EAST BOSTON	02128	N	SEMBRANO RODERICK	197 LEXINGTON ST	1,150	3047	2307	198,500	619,700	818,200	\$8,918.38	1920	2000
199	EAST BOSTON	02128	Y	GUERRA CHEVARRIA ANA S	199 LEXINGTON ST	1,150	3392	2268	199,100	605,300	804,400	\$8,767.96	1905	1985
201	EAST BOSTON	02128	N	JB REALTY TRUST	PO BOX 557 #	1,150	3108	2028	199,700	535,600	735,300	\$8,014.77	1900	1991
203	EAST BOSTON	02128	Y	MARKS TRAVIS JOSEPH	203 Lexington ST	2,010	3700	2546	230,200	501,400	731,600	\$7,974.44	1900	1978
205	EAST BOSTON	02128	N	205 LEXINGTON LLC	28 LAUDHOLM RD	2,500	6278	4362	263,800	1,037,400	1,301,200	\$14,183.08	1900	2018
209	EAST BOSTON	02128	N	YOON SUNG PIL	-211 209 LEXINGTON ST	2,500	6432	4296	264,700	1,003,200	1,267,900	\$13,820.11	1900	2009
213	EAST BOSTON	02128	Y	CASTALDINI ANTONIO	213 LEXINGTON ST	2,500	6048	4080	265,300	885,400	1,150,700	\$12,542.63	1900	

PROPERTY ASSESSMENT data from data.boston.gov. For instance, no such data available for Tucson?. Neighborhood income doesn't have this information at this detail level.

- data.gov has almost 300K datasets available!

DATA SOURCES

- Proprietary data sources

Usually via rate-limited application program interfaces (APIs)

- For an example check:

<https://colab.research.google.com/drive/19pC9y6DpN-Y5677kr6AuJ18Jff2QOgkI?usp=sharing>

```
import pandas as pd
df_mirror = pd.json_normalize(pmirror)
df_mirror.head()
```

```
import requests
jmirror = requests.get("http://api.tvmaze.com/search/shows?q=mirror")
jmirror.text[:500]
' [{"score":0.7032747,"show":{"id":305,"url":"https://www.tvmaze.com/sh
ish","genres":["Drama","Science-Fiction","Thriller"],"status":"Running
ficialSite":"https://www.netflix.com/title/70264888","schedule":{"time
el":{"id":1,"name":"Netflix","country":null,"officialSite'
pmirror = jmirror.json()
pmirror
[{'score': 0.7032747,
 'show': {'id': 305,
 'url': 'https://www.tvmaze.com/shows/305/black-mirror',
 'name': 'Black Mirror',
 'type': 'Scripted',
 'language': 'English',
 'genres': ['Drama', 'Science-Fiction', 'Thriller'],
 'status': 'Running',
 'runtime': None,
 'averageRuntime': 63,
```

DATA SOURCES

- Scraping websites: Which US town has seen the largest growth 2010-2020?
- For an example check:

https://colab.research.google.com/drive/1pcsmyp0fcq8kL14Y4hy47StocV4jvF_f?usp=sharing

```
import requests
response = requests.get(
    "https://en.wikipedia.org/wiki/Tucson,_Arizona")
```

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(response.text, "html.parser")
tucson_hist_pop = soup.find('table', {'class': 'us-census-pop us-census-pop-right'})
```

```
tucson_hist_pop






```

```
data = []
for row in tucson_hist_pop.find_all('tr'):
    row_data = []
    for cell in row.find_all('td'):
        row_data.append(cell.text)
    data.append(row_data)
```

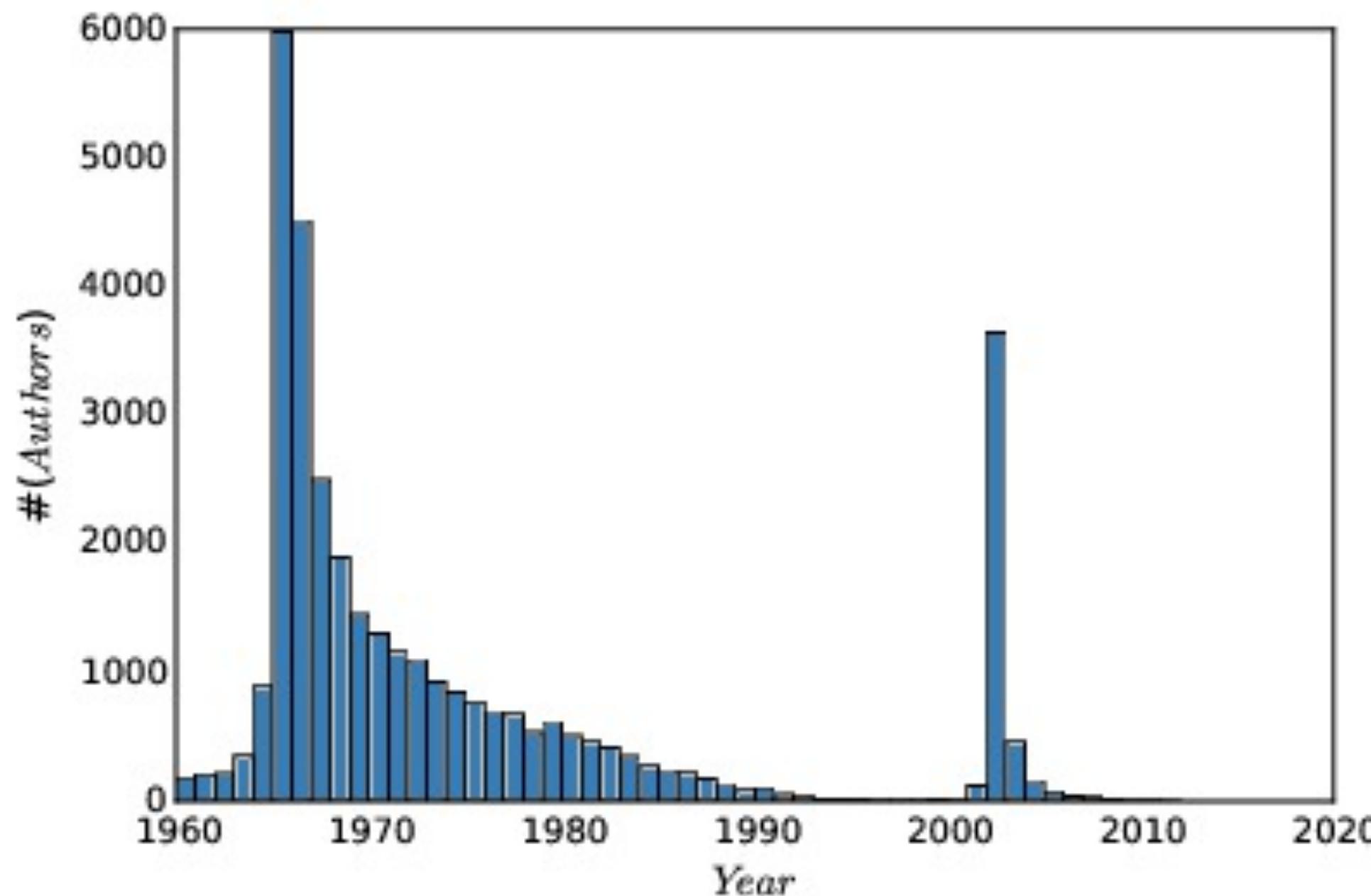
DATA CLEANING

- Many potential issues:
 - Distinguishing errors from artifacts.
 - Data compatibility / unification.
 - Imputation of missing values.
 - Estimating unobserved (zero) counts.
 - Outlier detection.



DATA CLEANING

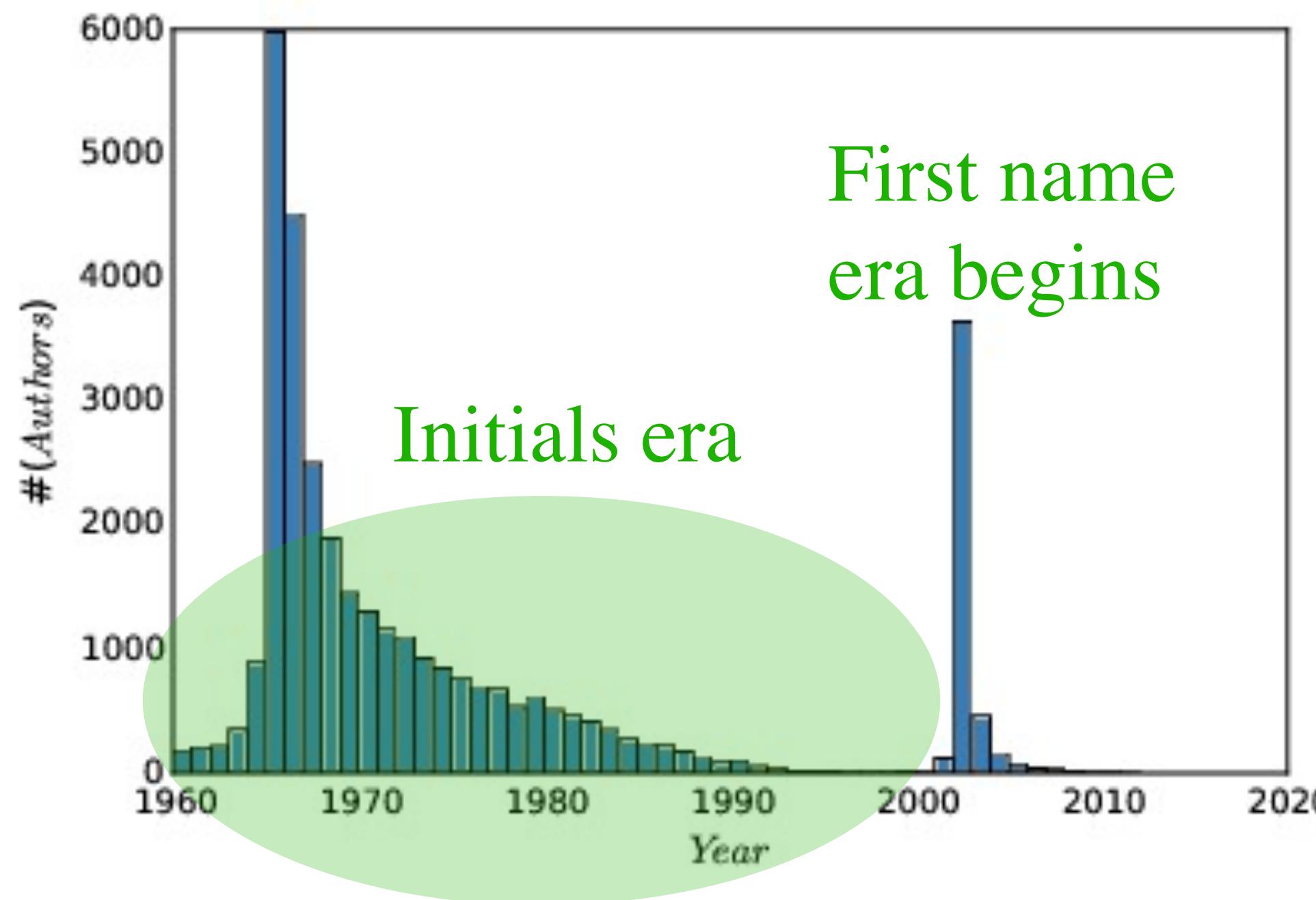
- Errors: Information lost in acquisition (measurements etc.)
- Artifacts: Problems arising from processing done to raw data.



100,000 most prolific authors, binned according to
the year of their first paper appearing in Pubmed

DATA CLEANING

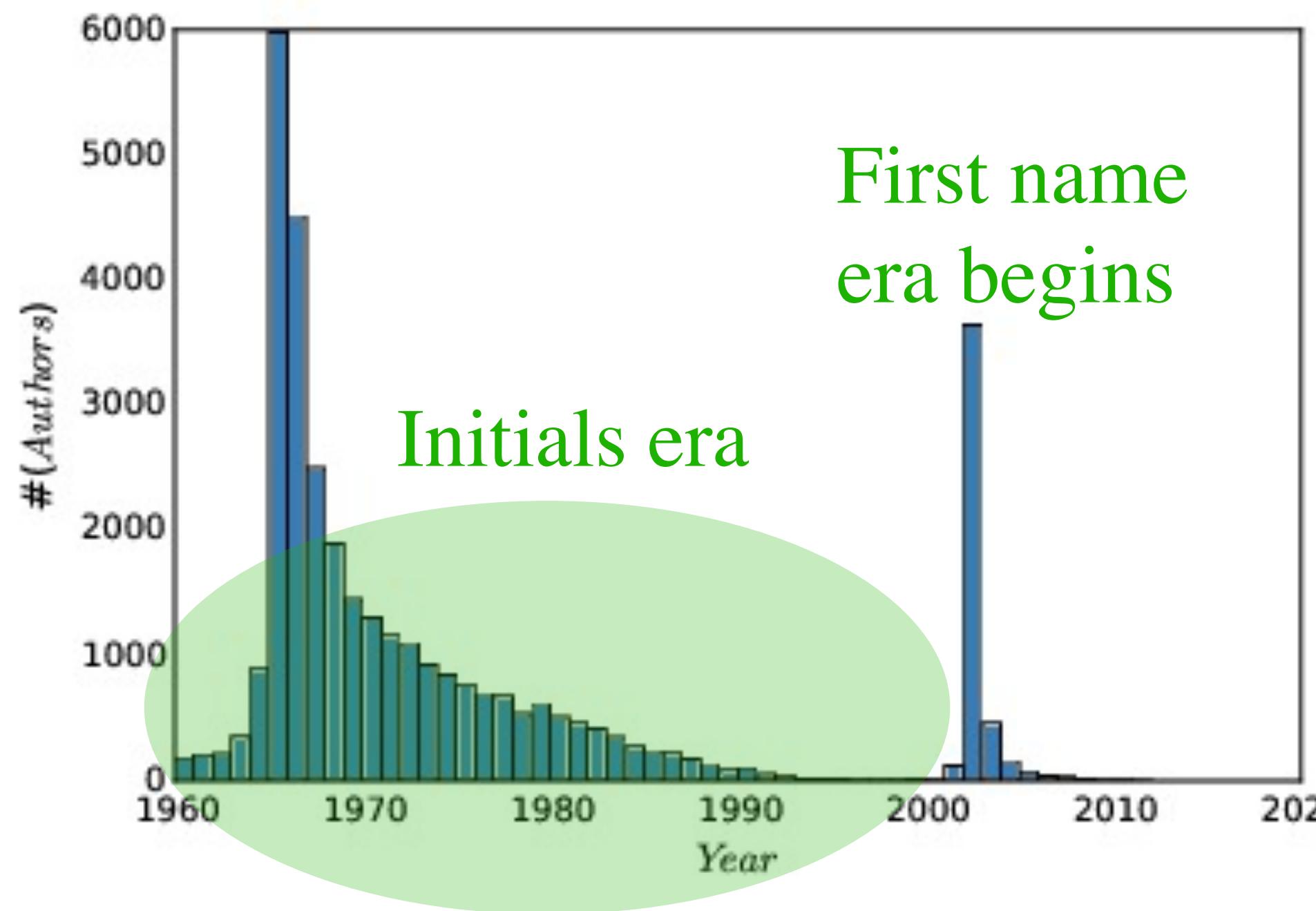
- Errors: Information lost in acquisition (measurements etc.)
- Artifacts: Problems arising from processing done to raw data.



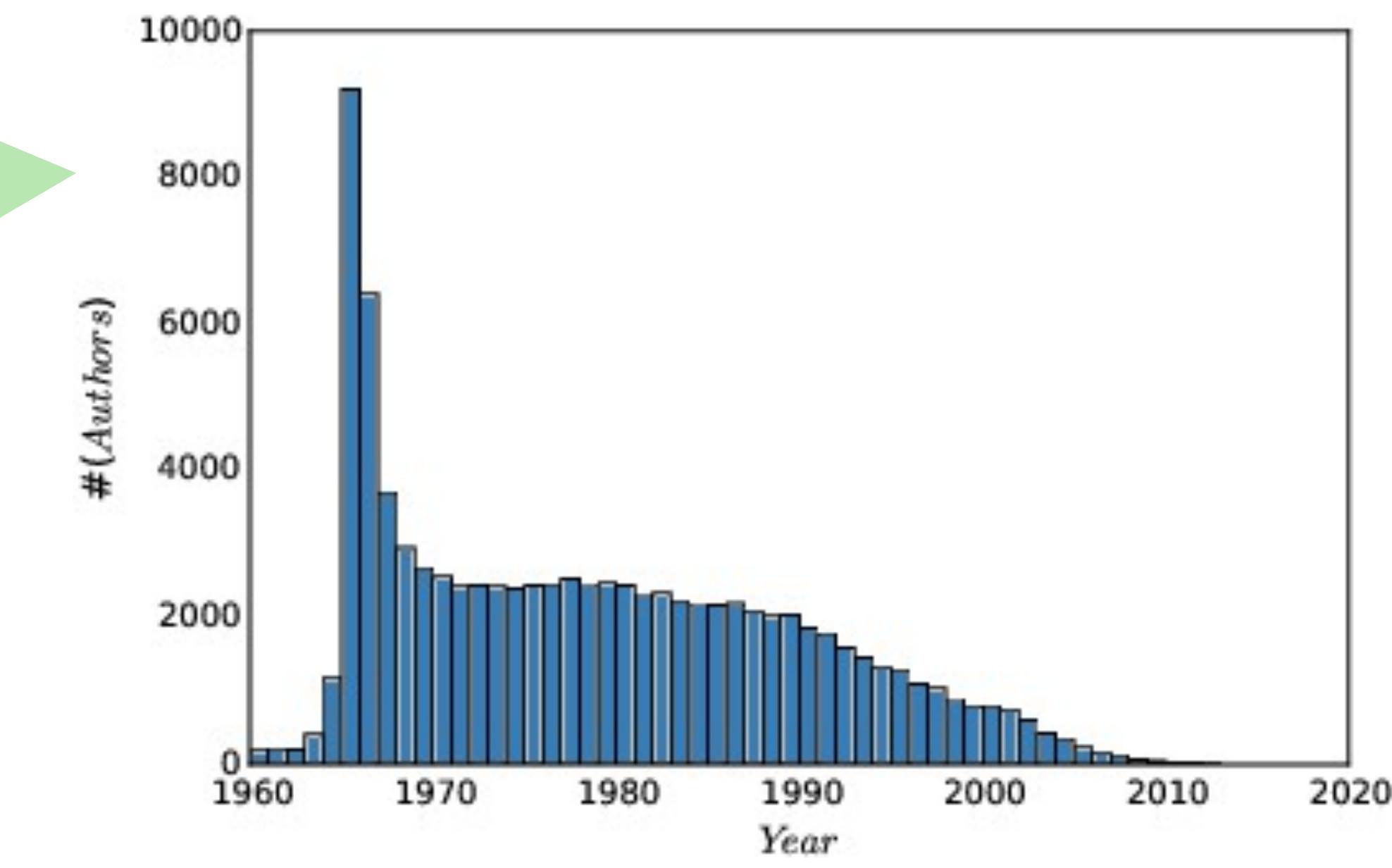
100,000 most prolific authors, binned according to the year of their first paper appearing in Pubmed

DATA CLEANING

- Errors: Information lost in acquisition (measurements etc.)
- Artifacts: Problems arising from processing done to raw data.



100,000 most prolific authors, binned according to the year of their first paper appearing in Pubmed



After cleaning the data (merging different strings representing same names)

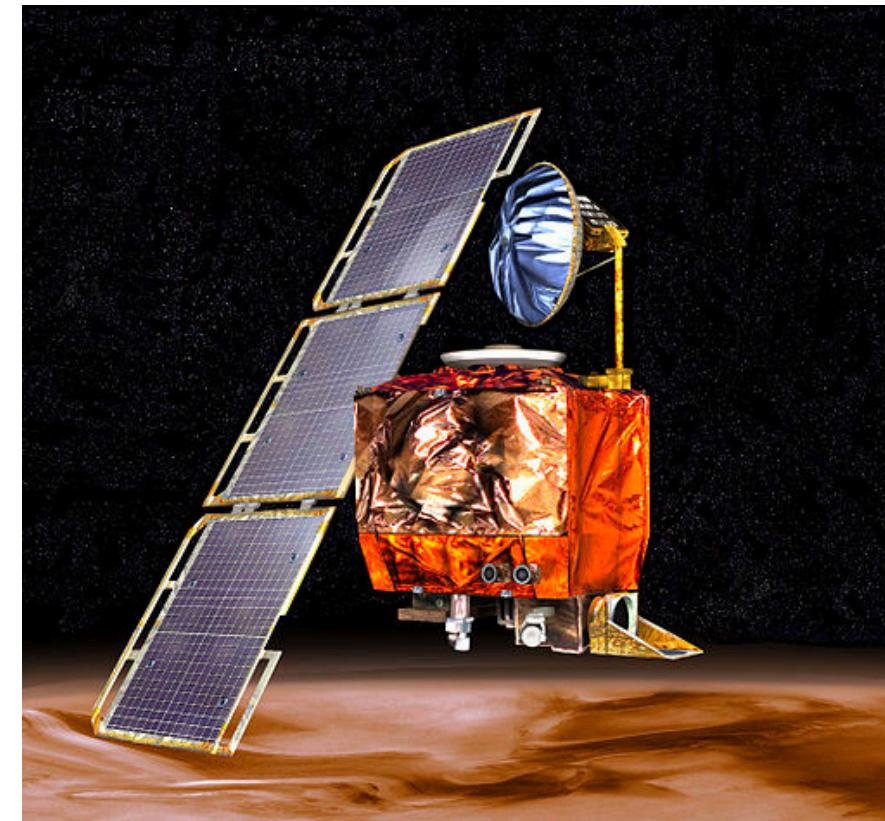
DATA CLEANING

- Data compatibility issues to be aware of:
 - Unit conversions
 - Number / character code representations
 - Name unification
 - Time/date unification
 - Financial unification

DATA CLEANING

- Unit conversions

- Can have disastrous effects: Mars Climate Orbiter

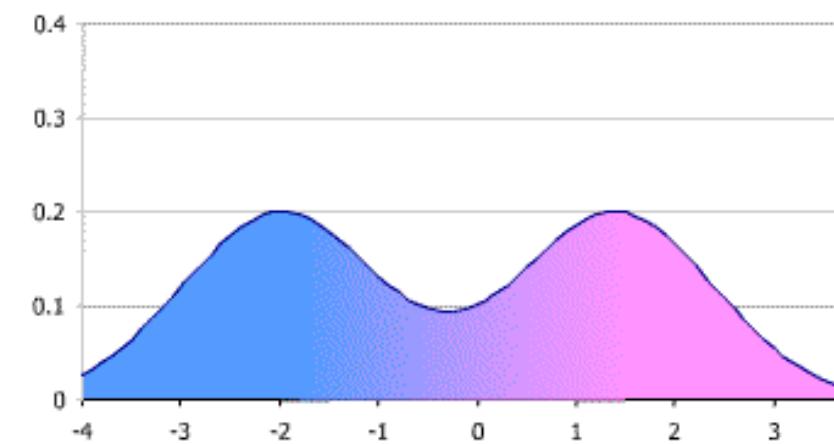


entered an orbit around the Sun.^[2] An investigation attributed the failure to a measurement mismatch between two measurement systems: [SI units](#) (metric) by NASA and [US customary](#) units by spacecraft builder [Lockheed Martin](#).^[3]

from Wikipedia

Even if decide on metric: cm, m, km?

- For same variable: watch out if there is a bimodal distribution

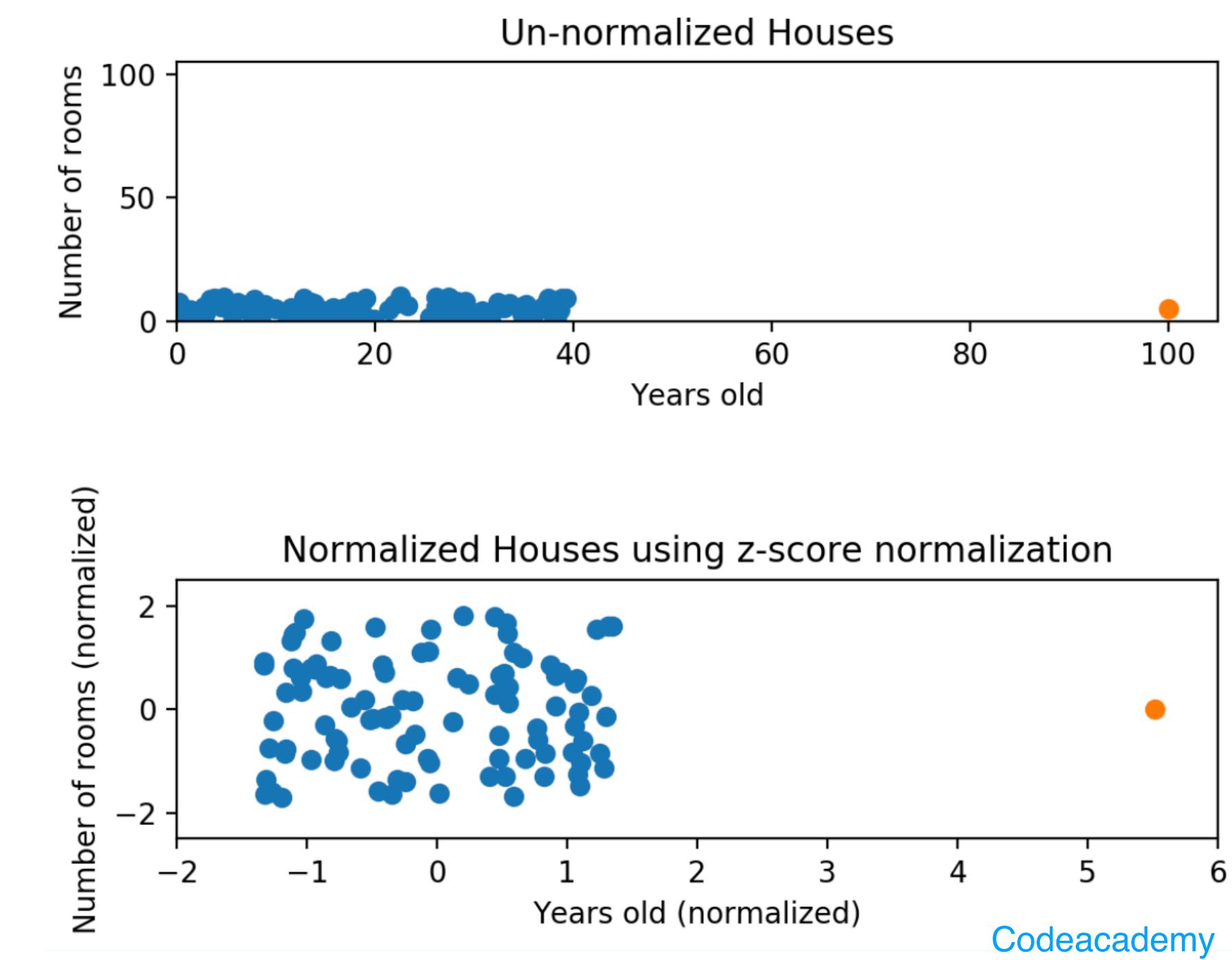


- Multiple variables: Normalization is important.

DATA CLEANING

- Normalization with Z-scores
 - Z-scores are dimensionless quantities
 - Normalize different variables:
range/distribution comparable.

$$Z_i = (X_i - \bar{X}) / \hat{\sigma}$$



- Z-scores of height in inches same as height measured in miles.
- Z-scores have mean 0 and sigma=1:

$$\begin{aligned}\mu(B) &= 21.9 & \sigma(B) &= 1.92 \\ \mu(Z) &= 0 & \sigma(Z) &= 1\end{aligned}$$

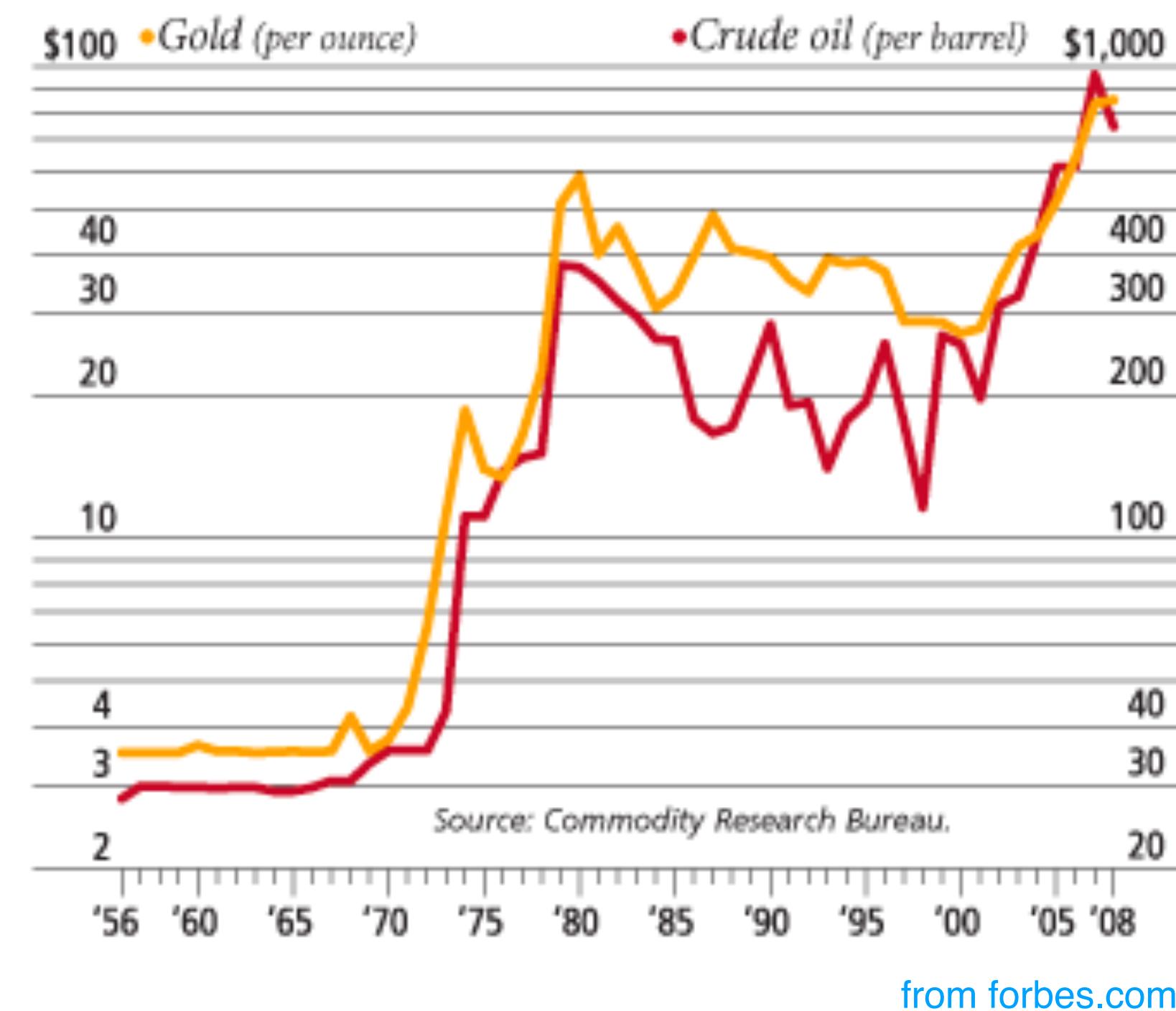
B	19	22	24	20	23	19	21	24	24	23
Z	-1.51	0.05	1.09	-0.98	0.57	-1.51	-0.46	1.09	1.09	0.57

DATA CLEANING

- Time/Date
 - Aligning temporal events from different datasets problematic:
 - Coordinated Universal Time (UTC), GMT, Unix time.
 - Algorithms for converting Gregorian calendar dates to others.

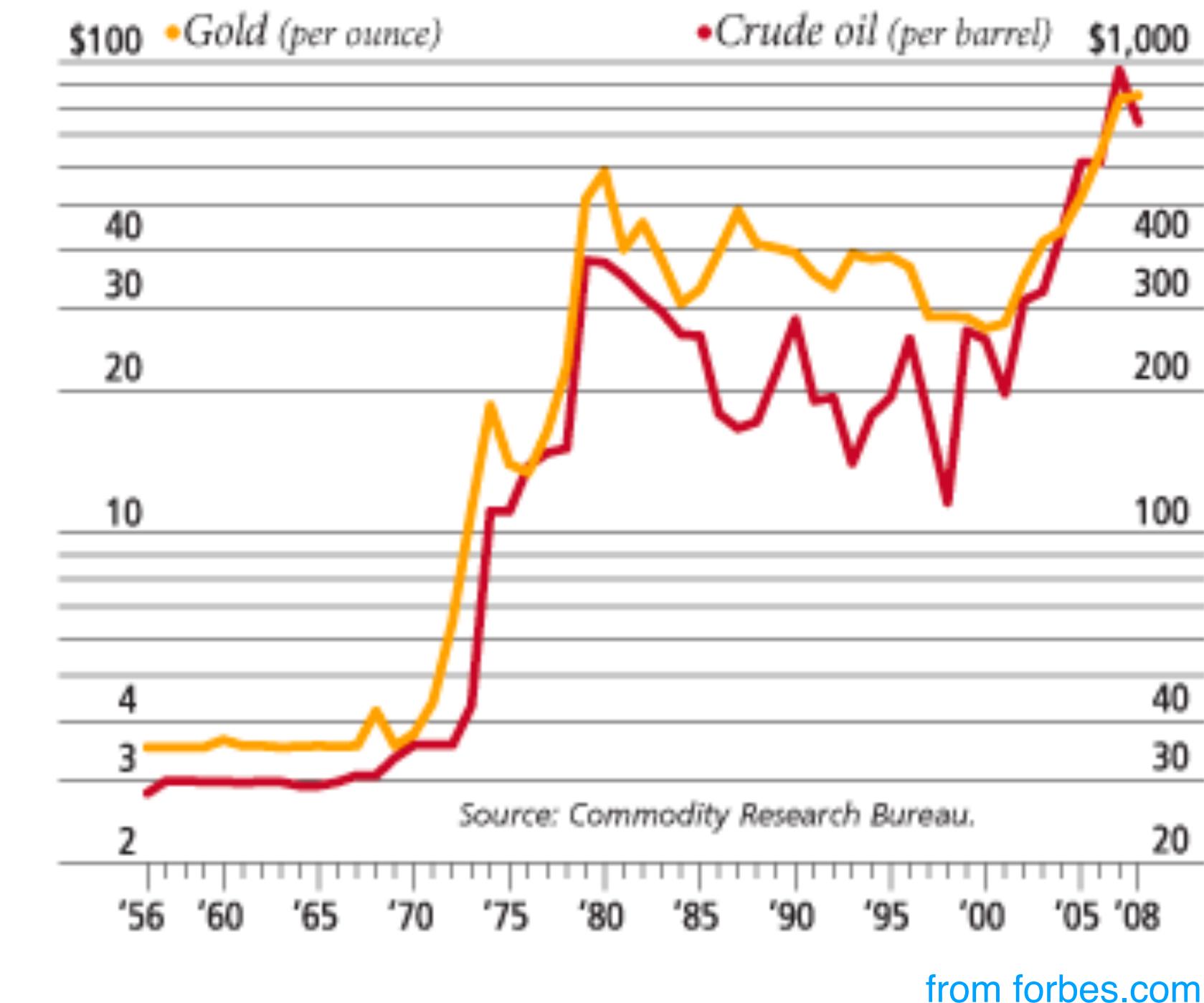
DATA CLEANING

- Financial data
 - Must be careful when units of measurement can change with time
 - Gold/oil prices really correlated?



DATA CLEANING

- Financial data
 - Must be careful when units of measurement can change with time
 - Gold/oil prices really correlated?



- Both measured in dollars: Value of dollar changes with time!
Time series of prices of any pair of items correlated!
- Solution: Use returns $r_i = \frac{p_{i+1} - p_i}{p_i}$

DATA CLEANING

- Dealing with missing values
 - Year of death of a living person?
 - Survey question left blank?

DATA CLEANING

- Dealing with missing values
 - Year of death of a living person?
 - Survey question left blank?

Filling in with values such as 0 or –1 is problematic.

⇒ will be misinterpreted as actual data.

DATA CLEANING

- Dealing with missing values

- Year of death of a living person?
- Survey question left blank?

Filling in with values such as 0 or –1 is problematic.

⇒ will be misinterpreted as actual data.

Dropping records with missing values is another option

⇒ other data fields of those records might be valuable

Imputation is a good option.

DATA CLEANING

- Imputation methods

Heuristic-based imputation: death year(living person)=birth year+80

DATA CLEANING

- Imputation methods

Heuristic-based imputation: death year(living person)=birth year+80

Mean value (of the column) imputation: leaves mean the same.

DATA CLEANING

- Imputation methods

Heuristic-based imputation: death year(living person)=birth year+80

Mean value (of the column) imputation: leaves mean the same.

Imputation by nearest neighbor: better than mean when there are systematic reasons to explain variance.

DATA CLEANING

- Imputation methods

Heuristic-based imputation: death year(living person)=birth year+80

Mean value (of the column) imputation: leaves mean the same.

Imputation by nearest neighbor: better than mean when there are systematic reasons to explain variance.

Random value (from the same column) imputation: repeatedly selecting random values permits statistical evaluation of the impact of imputation.

DATA CLEANING

- Imputation methods

Heuristic-based imputation: death year(living person)=birth year+80

Mean value (of the column) imputation: leaves mean the same.

Imputation by nearest neighbor: better than mean when there are systematic reasons to explain variance.

Random value (from the same column) imputation: repeatedly selecting random values permits statistical evaluation of the impact of imputation.

Imputation by interpolation: using linear regression to predict missing values works well if few fields are missing per record.

DATA CLEANING

- Outlier Detection

Dinosaur vertebra 1500mm \Rightarrow 188 feet long beast

Next largest is 122 feet.

If tallest person recorded was 8 feet 9 inches,

Wouldn't you be surprised of a claim of 12 feet person?

DATA CLEANING

- Outlier Detection

Dinosaur vertebra 1500mm \Rightarrow 188 feet long beast

Next largest is 122 feet.

If tallest person recorded was 8 feet 9 inches,

Wouldn't you be surprised of a claim of 12 feet person?

\Rightarrow Detecting outliers in normally distributed data should be easy
(No large outliers)

\Rightarrow the dinosaur example was probably measurement error

\Rightarrow Detecting outliers harder with power-law distributions

DATA CLEANING

- Outlier Detection

How to detect them?

Clustering: a data point too far from its cluster center

DATA CLEANING

- Outlier Detection

How to detect them?

Clustering: a data point too far from its cluster center

What to do about them?

Fix why you have outliers.

May hint at systematic problems elsewhere.

Deleting outliers prior to fitting?

May give better models

(if outliers are due to measurement error)

May give worse models

(if removed due to not being explained by your simple model)

DATA VISUALIZATION

Importance of Data Visualization

Anscombe's Quartet: 4 data sets with identical statistical properties.

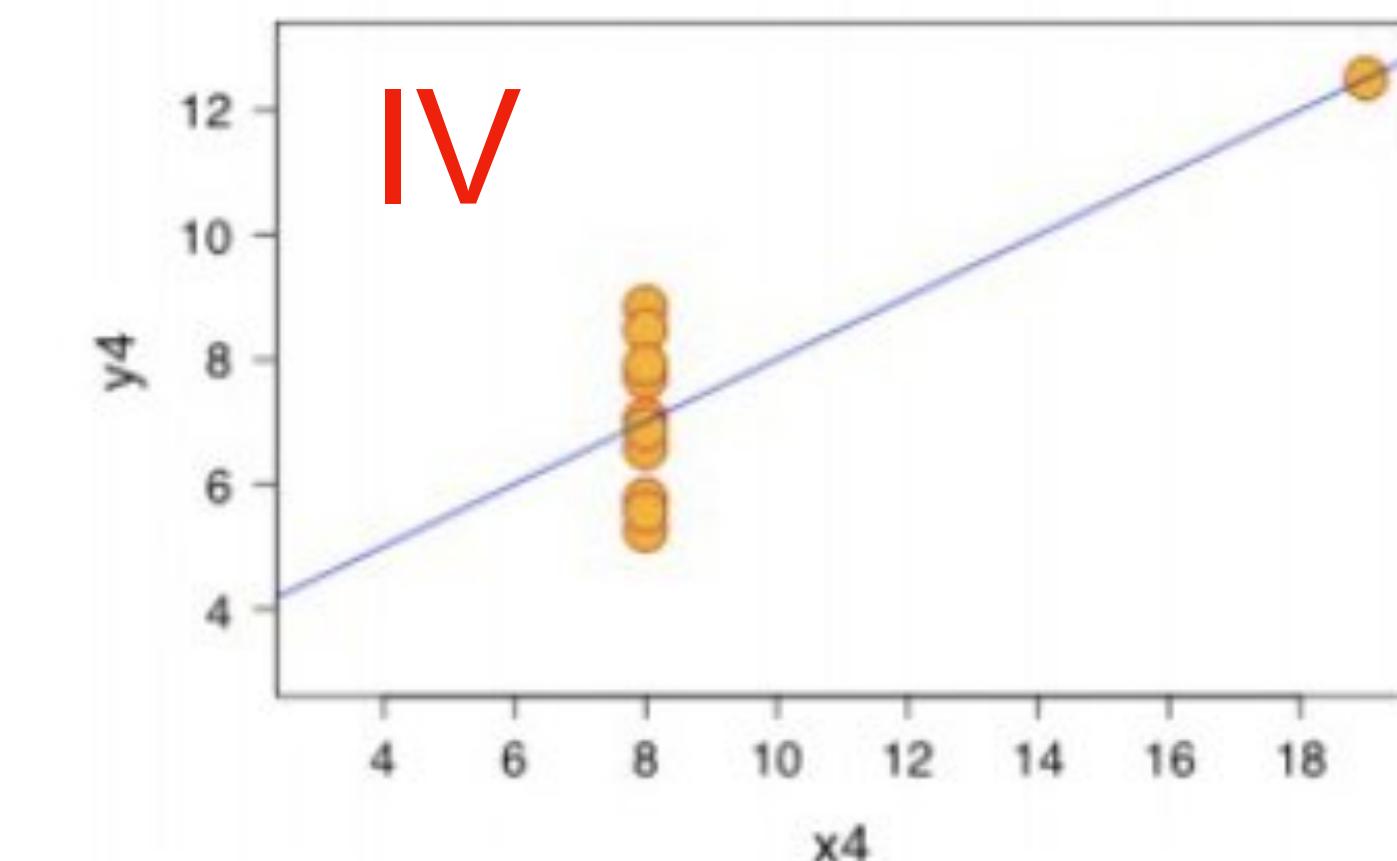
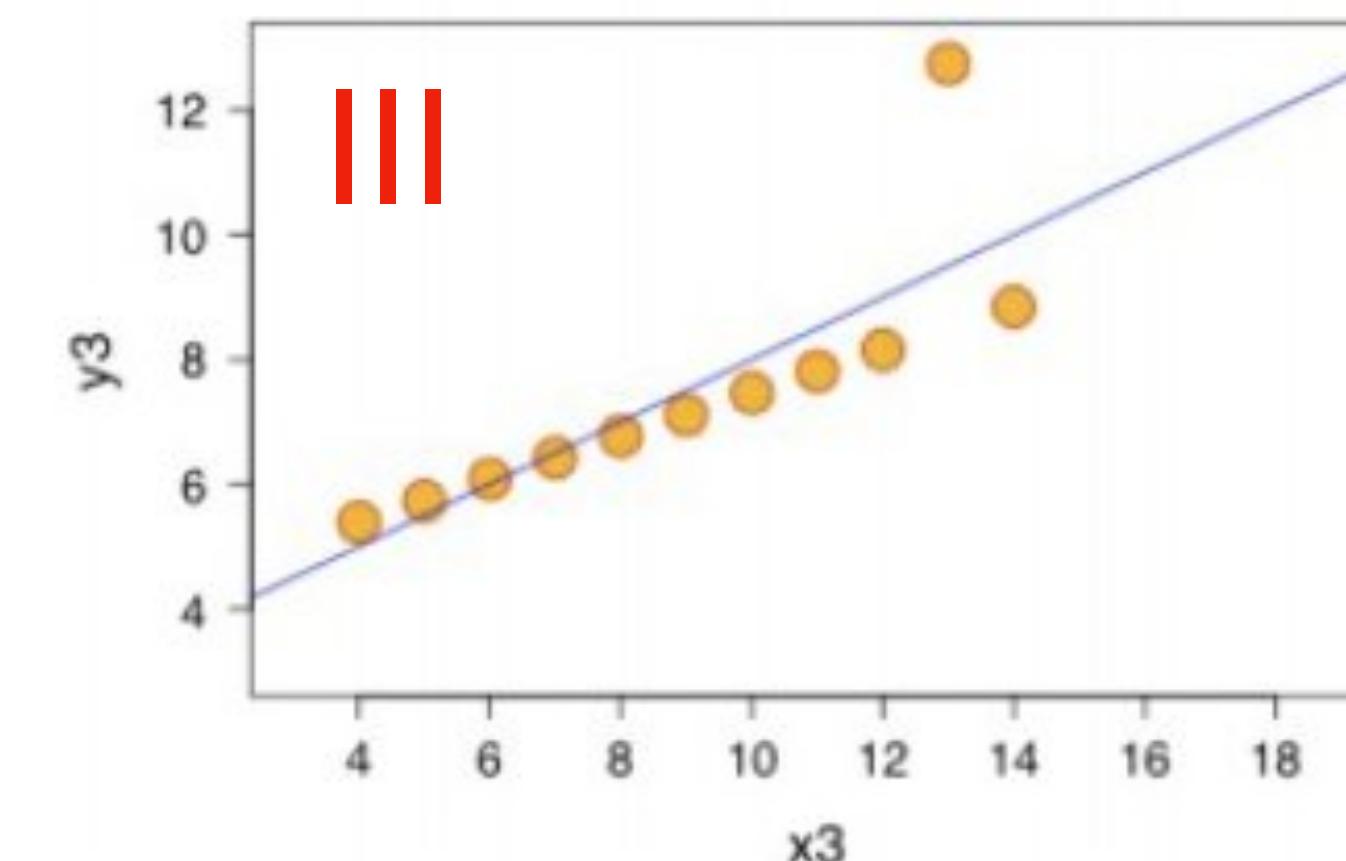
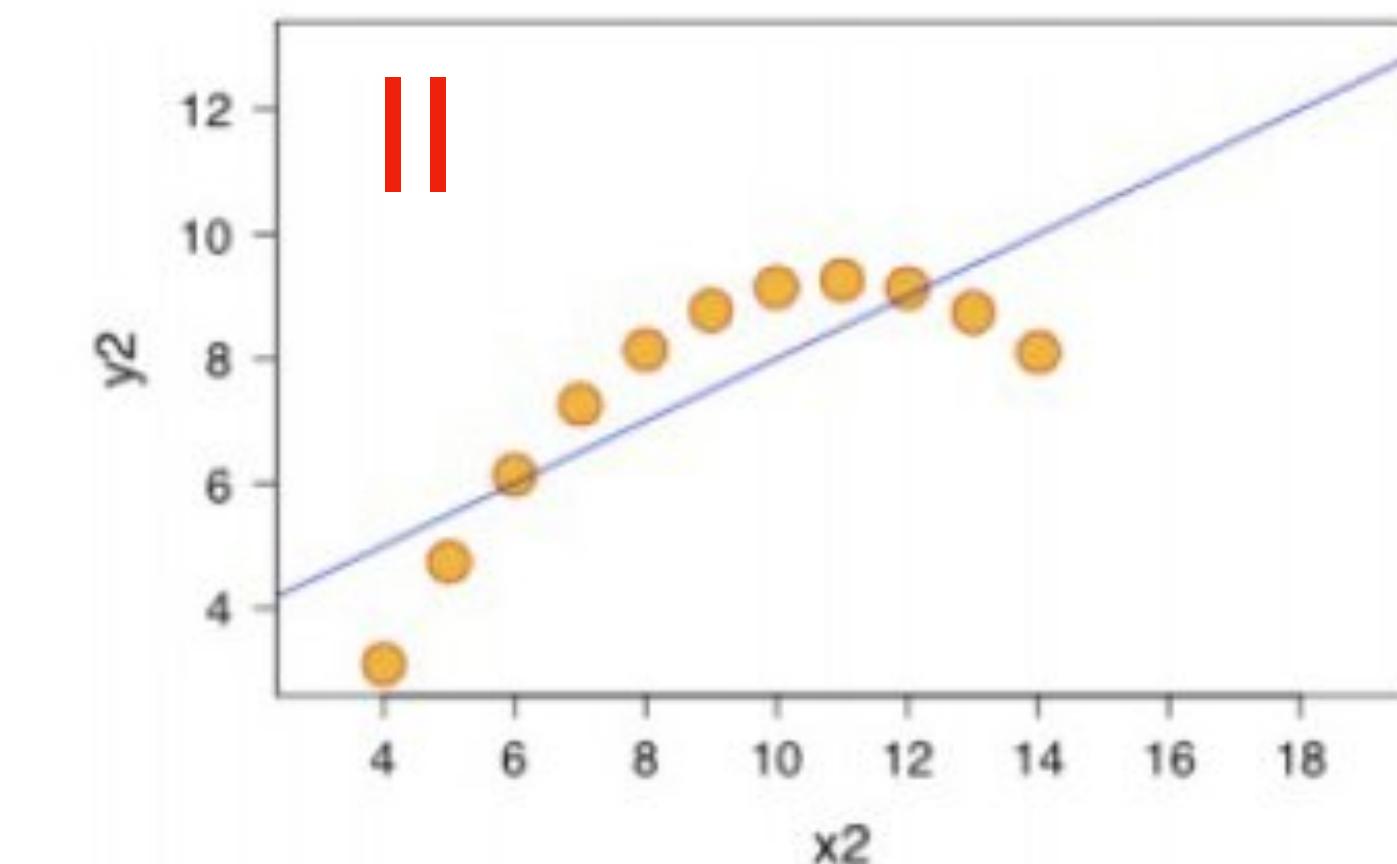
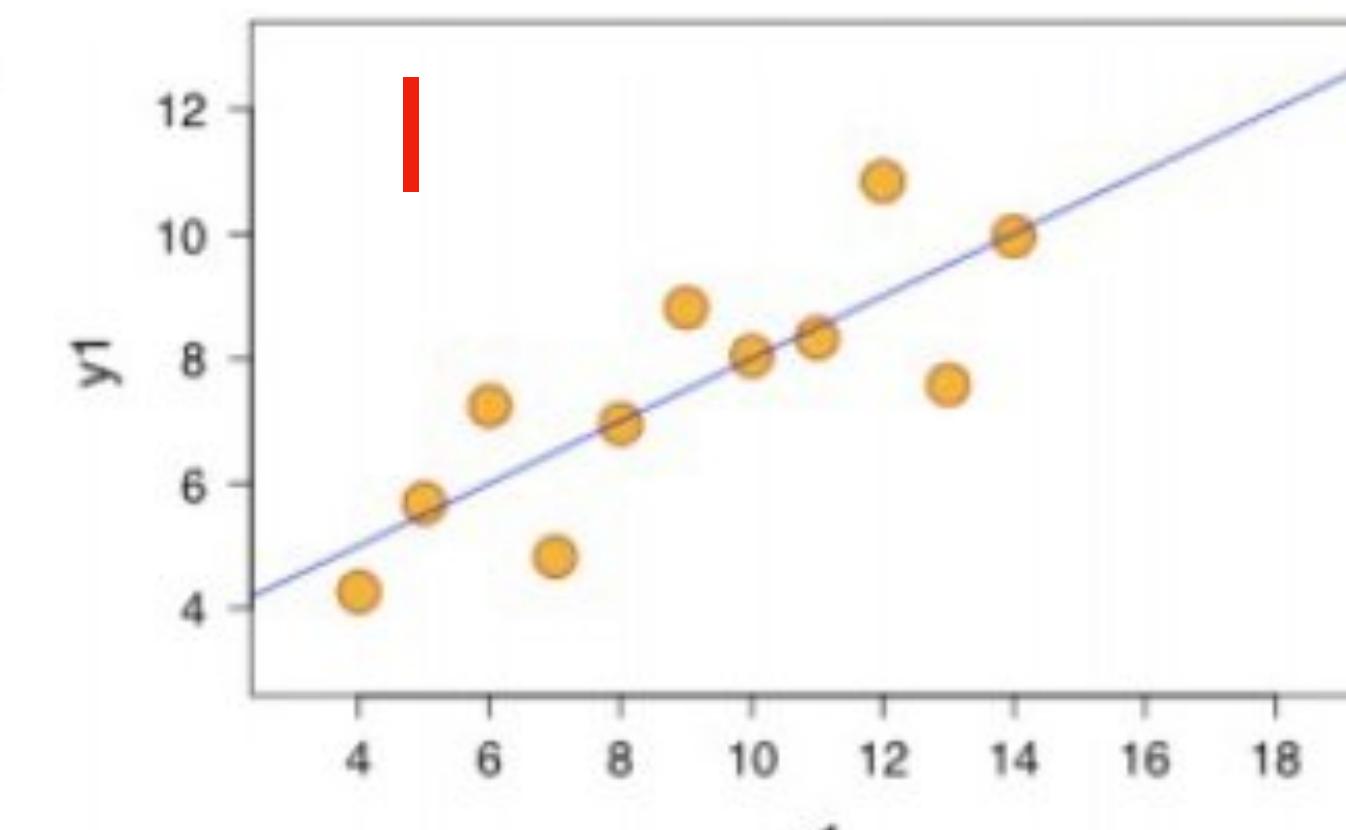
I		II		III		IV		
x	y	x	y	x	y	x	y	
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58	
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76	
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71	
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84	
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47	
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04	
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25	
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50	
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56	
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91	
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89	
mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
corr.		0.816		0.816		0.816		0.816

DATA VISUALIZATION

Importance of Data Visualization

Anscombe's Quartet: 4 data sets with identical statistical properties.

I		II		III		IV		
x	y	x	y	x	y	x	y	
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58	
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76	
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71	
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84	
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47	
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04	
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25	
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50	
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56	
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91	
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89	
mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
corr.	0.816		0.816		0.816		0.816	



DATA VISUALIZATION

Importance of Data Visualization

- EDA:
 - Identify mistakes in data collection/preprocessing
 - Identify violations of statistical assumptions
 - Observe patterns in the data
 - Construct hypothesis

DATA VISUALIZATION

Importance of Data Visualization

- EDA:
 - Identify mistakes in data collection/preprocessing
 - Identify violations of statistical assumptions
 - Observe patterns in the data
 - Construct hypothesis
- Error Detection:
 - Good idea not to feed unvisualized data into a ML algorithm.
 - Outliers, insufficient cleaning erroneous assumptions etc.

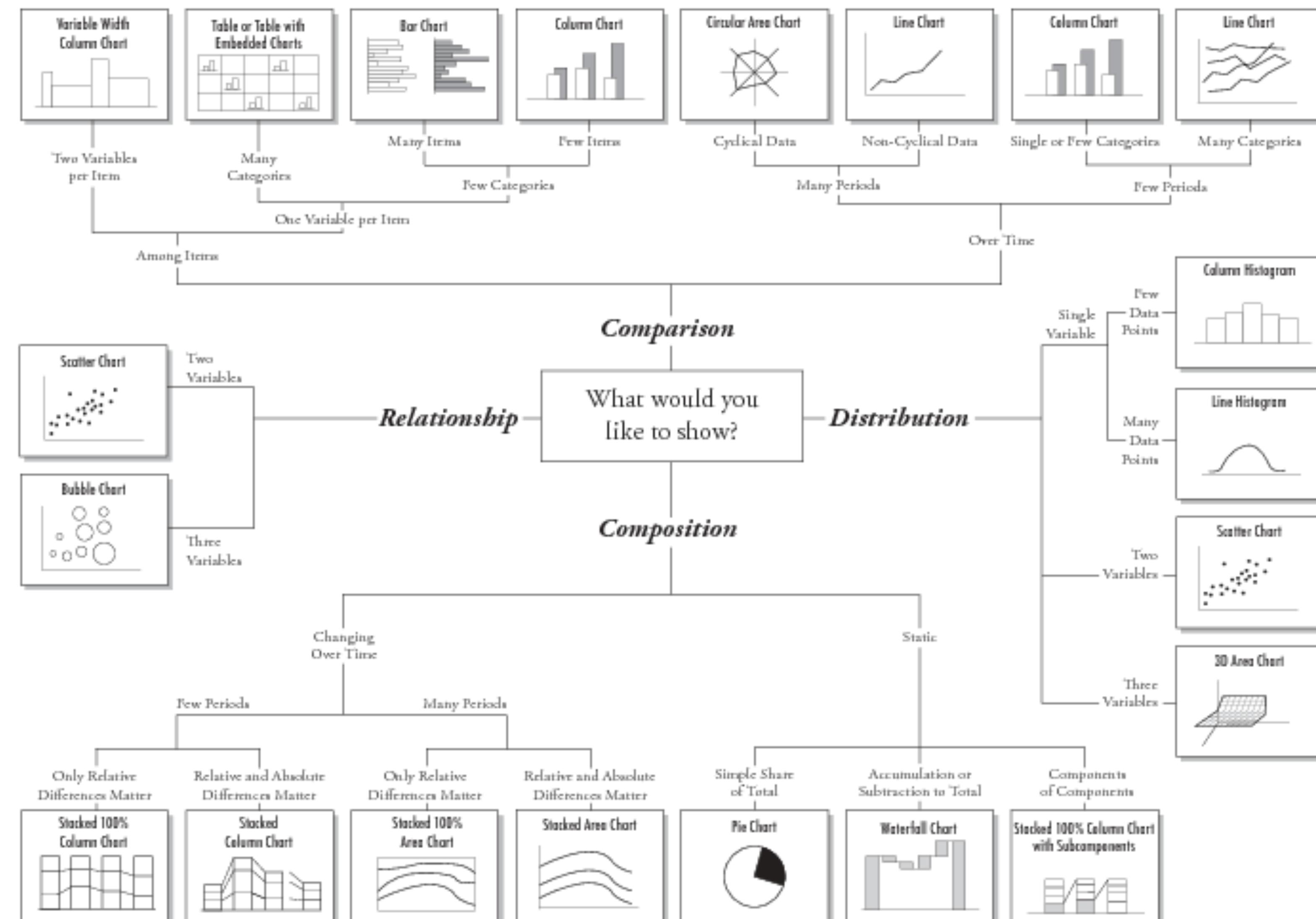
DATA VISUALIZATION

Importance of Data Visualization

- EDA:
 - Identify mistakes in data collection/preprocessing
 - Identify violations of statistical assumptions
 - Observe patterns in the data
 - Construct hypothesis
- Error Detection:
 - Good idea not to feed unvisualized data into a ML algorithm.
 - Outliers, insufficient cleaning, erroneous assumptions etc.
- Communication of findings:
 - A picture is worth 1000 words.

PRACTICE OF DATA VISUALIZATION

Chart Suggestions—A Thought-Starter



PRACTICE OF DATA VISUALIZATION

Visualizing Distributions

Distribution describes:

- The set of values that a variable can possibly take.
- The frequency with which each value occurs.

PRACTICE OF DATA VISUALIZATION

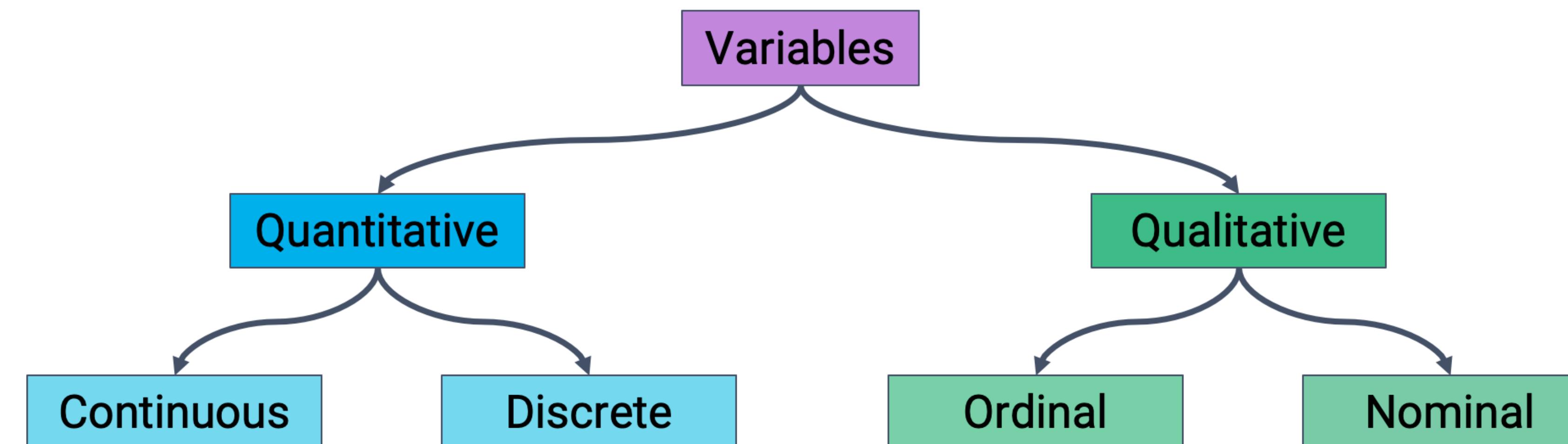
Visualizing Distributions

Distribution describes:

- The set of values that a variable can possibly take.
- The frequency with which each value occurs.

How to visualize a distribution?

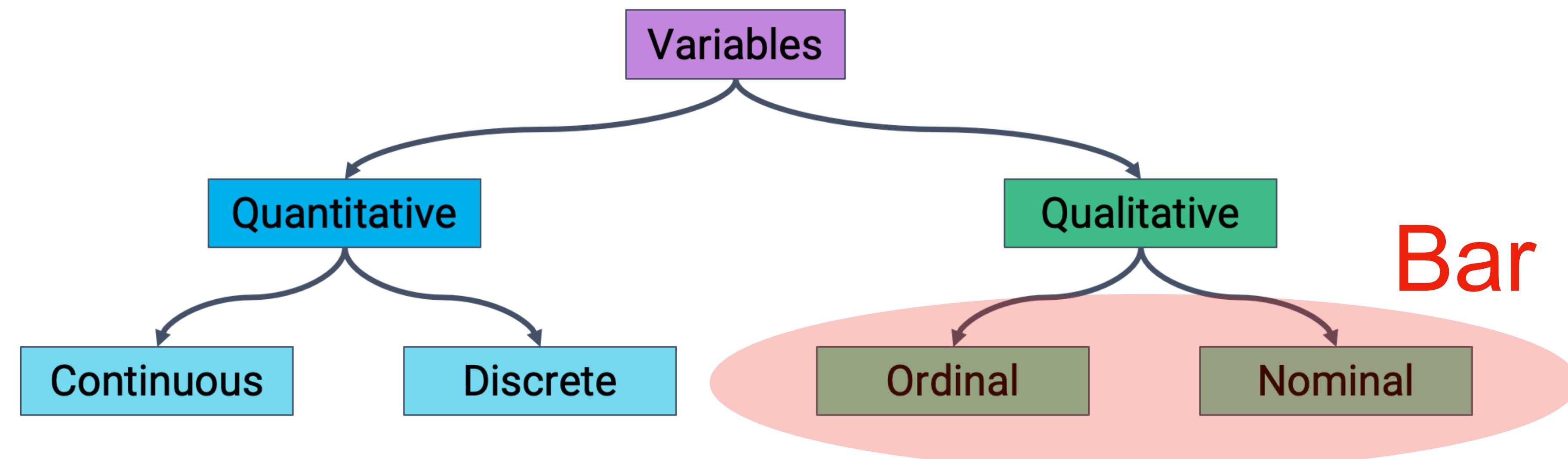
- Identify the variables being visualized.
- Select a plot type accordingly.



PRACTICE OF DATA VISUALIZATION

The Python notebook with the codes in the slides can be accessed at:

https://colab.research.google.com/drive/1g3c5Bgr3ynet4_-Tz4CD9jOP4Cg56YE?usp=sharing



Bar Plots

PRACTICE OF DATA VISUALIZATION

Bar plots

```
import matplotlib.pyplot as plt
events = ted_main["event"].value_counts().head(5).sort_index()
```

TED2009	83
TED2013	77
TED2014	84
TED2015	75
TED2016	77
Name:	event, dtype: int64

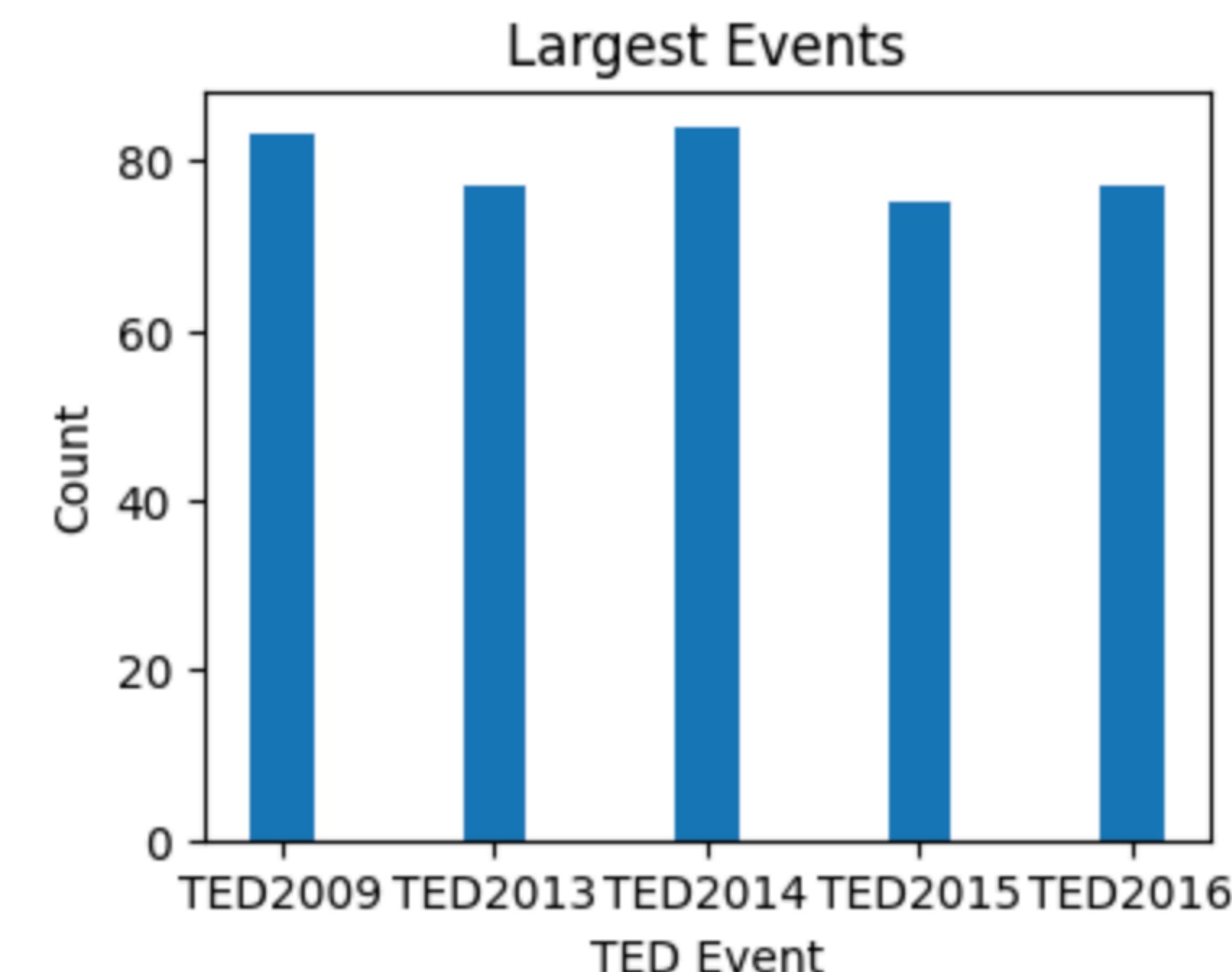
PRACTICE OF DATA VISUALIZATION

Bar plots

```
import matplotlib.pyplot as plt
events = ted_main["event"].value_counts().head(5).sort_index()
```

TED2009	83
TED2013	77
TED2014	84
TED2015	75
TED2016	77
Name: event, dtype: int64	

```
plt.figure(figsize=(4, 3))
plt.bar(events.index, events.values, width=0.3)
plt.xlabel("TED Event")
plt.ylabel("Count")
plt.title("Largest Events")
```



PRACTICE OF DATA VISUALIZATION

Stacked bar plots:

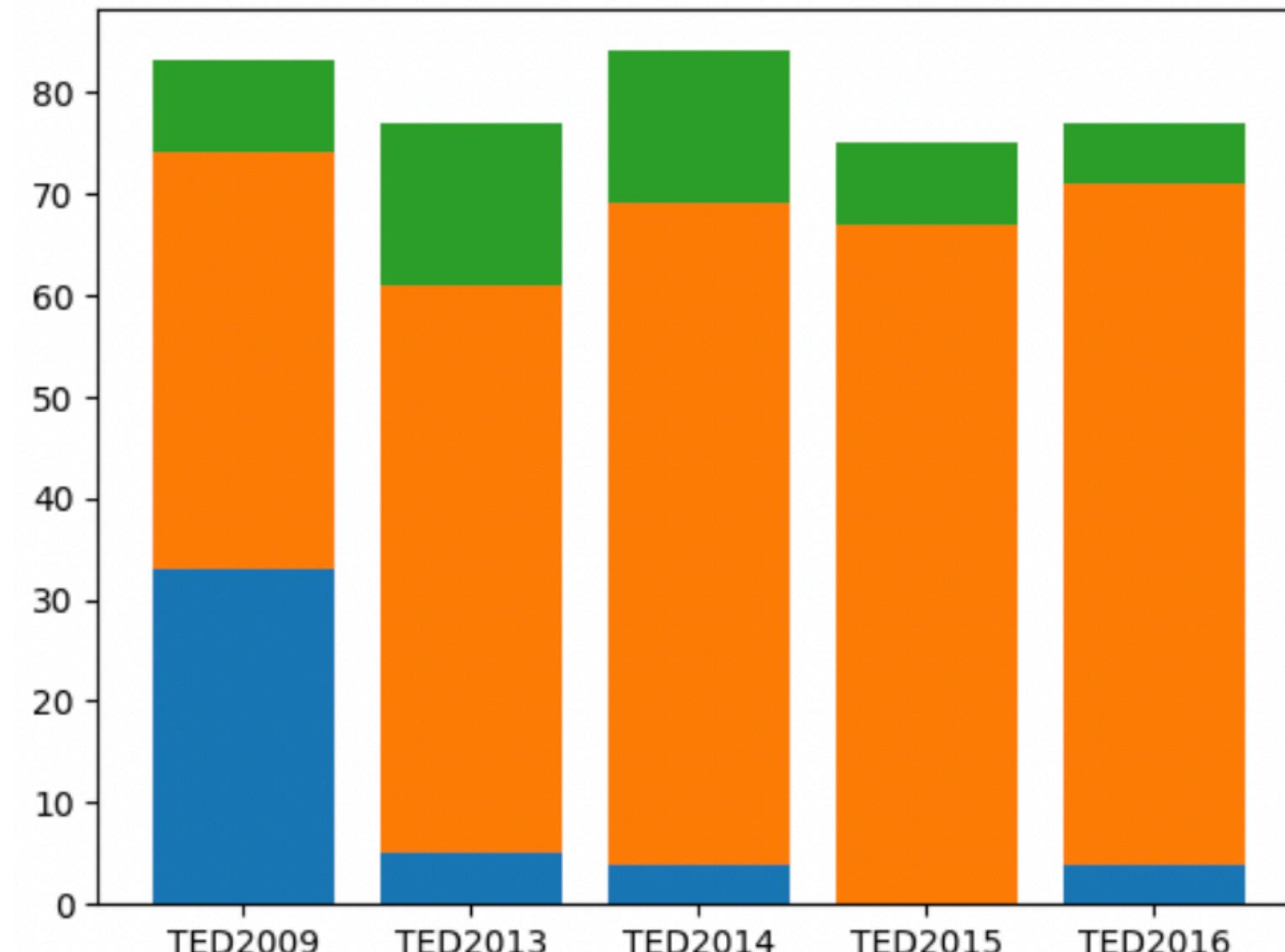
```
p1 = [(x, len(ted_main[(ted_main["event"]==x) &
                     (ted_main["views"]<=800000)]))
       for x in events.index]
p2 = [(x, len(ted_main[(ted_main["event"]==x) &
                     (ted_main["views"]>800000) &
                     (ted_main["views"]<=3000000)]))
       for x in events.index]
p3 = [(x, len(ted_main[(ted_main["event"]==x) &
                     (ted_main["views"]>3000000)]))
       for x in events.index]
```

PRACTICE OF DATA VISUALIZATION

Stacked bar plots:

```
p1 = [(x, len(ted_main[(ted_main["event"]==x) &
                     (ted_main["views"]<=800000)]))
       for x in events.index]
p2 = [(x, len(ted_main[(ted_main["event"]==x) &
                     (ted_main["views"]>800000) &
                     (ted_main["views"]<=3000000)]))
       for x in events.index]
p3 = [(x, len(ted_main[(ted_main["event"]==x) &
                     (ted_main["views"]>3000000)]))
       for x in events.index]
```

```
plt.bar([d[0] for d in p1], [d[1] for d in p1])
plt.bar([d[0] for d in p2], [d[1] for d in p2],
        bottom=[d[1] for d in p1])
plt.bar([d[0] for d in p3], [d[1] for d in p3],
        bottom=np.array([d[1] for d in p1])+np.array([d[1] for d in p2]))
```



PRACTICE OF DATA VISUALIZATION

Ex: Bar plots not suitable for quantitative variables.

```
cd_ratio = ted_main["comments"] / ted_main["duration"]
cd_ratio = cd_ratio.value_counts().head(100)
cd_ratio
```

```
plt.figure(figsize=(8, 3))
plt.bar(cd_ratio.index, cd_ratio.values, width=0.005)
plt.xlabel("C/D Ratios")
plt.ylabel("Count")
plt.title("Comments/Duration Ratio Counts")
```

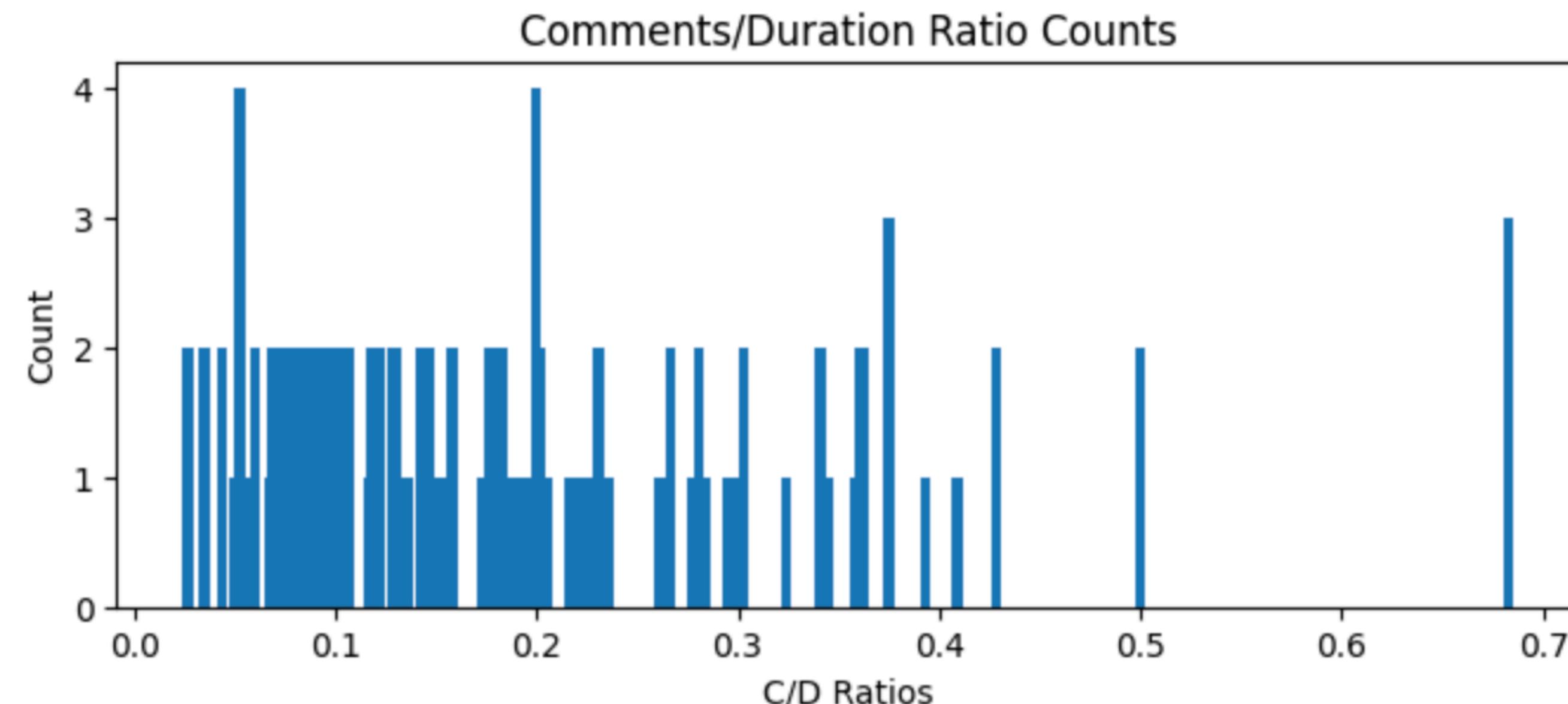
PRACTICE OF DATA VISUALIZATION

Ex: Bar plots not suitable for quantitative variables.

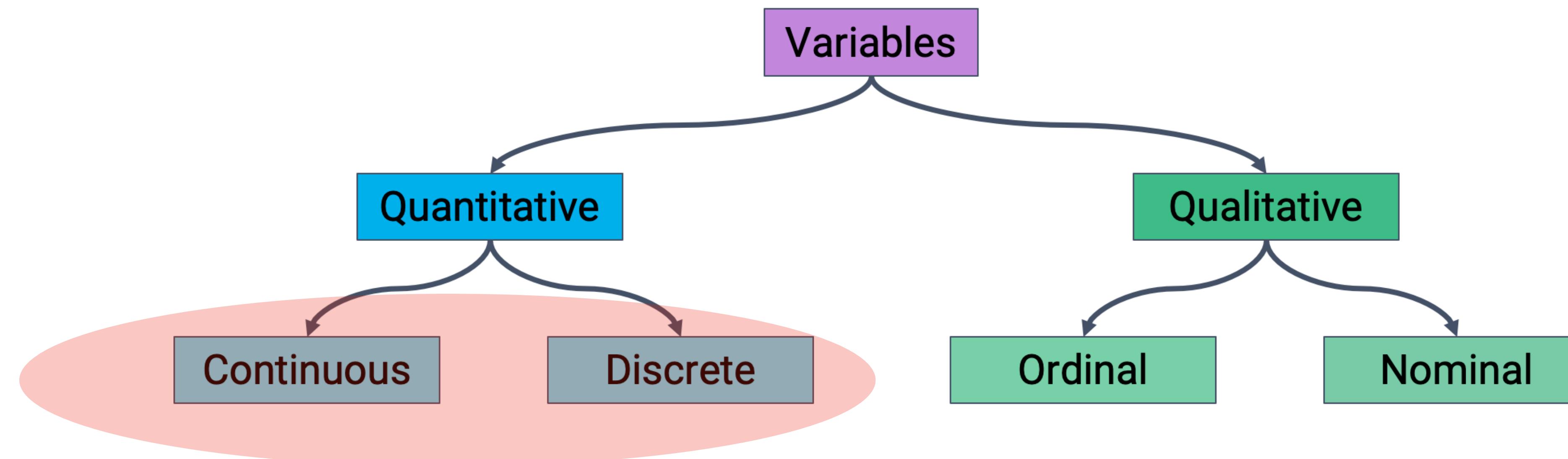
```
cd_ratio = ted_main["comments"] / ted_main["duration"]
cd_ratio = cd_ratio.value_counts().head(100)
cd_ratio
```

```
plt.figure(figsize=(8, 3))
plt.bar(cd_ratio.index, cd_ratio.values, width=0.005)
plt.xlabel("C/D Ratios")
plt.ylabel("Count")
plt.title("Comments/Duration Ratio Counts")
```

Separate bar for each unique value. Too many bars for continuous data or discrete data with many unique values.



PRACTICE OF DATA VISUALIZATION



How to
visualize?

PRACTICE OF DATA VISUALIZATION

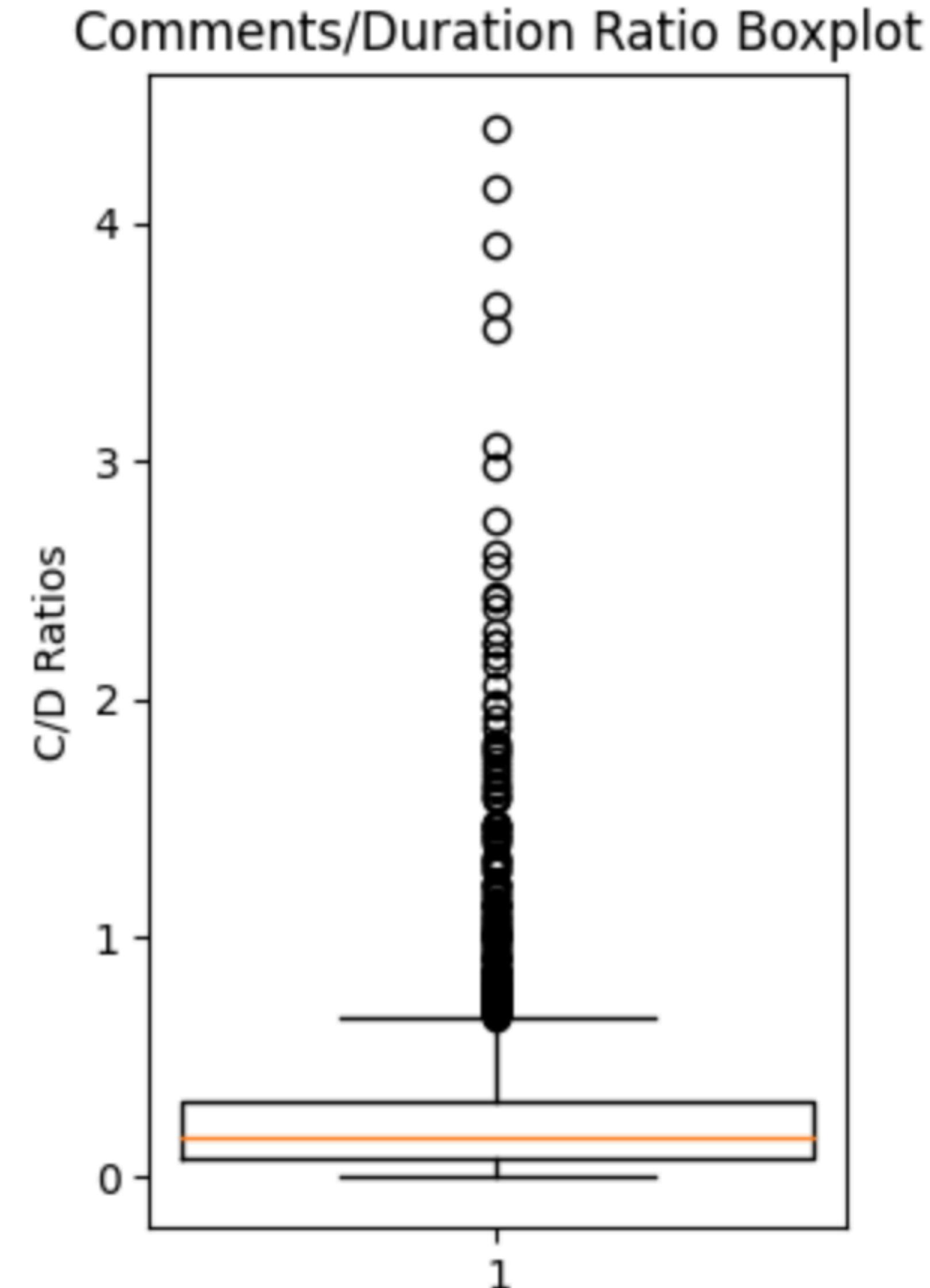
Box Plots: for visualizing a continuous quantitative variable.

```
plt.figure(figsize=(3, 5))
plt.boxplot(ted_main["comments"] / ted_main["duration"], widths=(1.5))
plt.ylabel("C/D Ratios")
plt.title("Comments/Duration Ratio Boxplot")
```

PRACTICE OF DATA VISUALIZATION

Box Plots: for visualizing a continuous quantitative variable.

```
plt.figure(figsize=(3, 5))
plt.boxplot(ted_main["comments"] / ted_main["duration"], widths=(1.5))
plt.ylabel("C/D Ratios")
plt.title("Comments/Duration Ratio Boxplot")
```



PRACTICE OF DATA VISUALIZATION

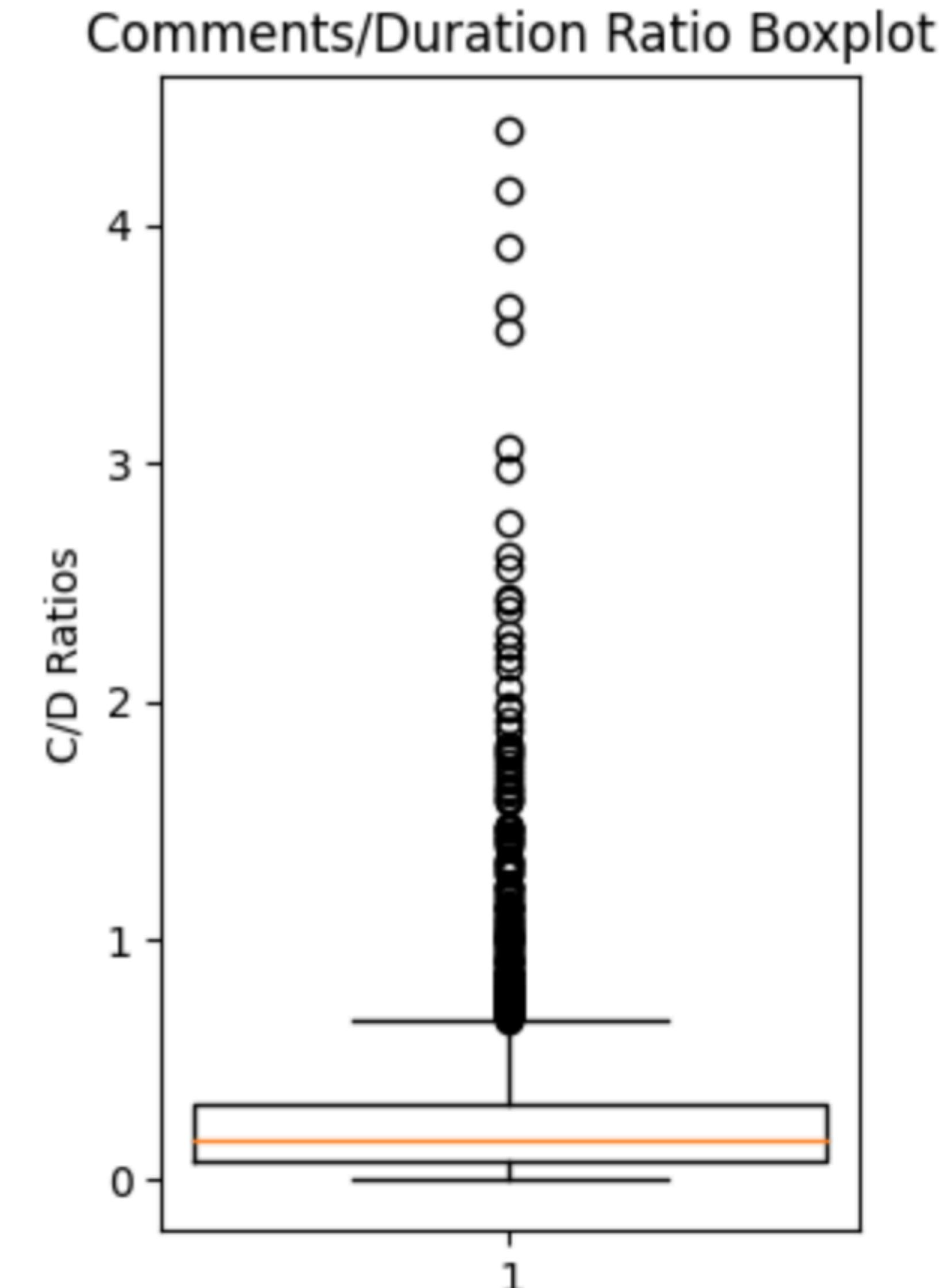
Box Plots: for visualizing a continuous quantitative variable.

Display distribution using quartiles:

- First or lower quartile: 25th percentile.
- Second quartile: 50th percentile (median).
- Third or upper quartile: 75th percentile.

[1^{st} quartile, 3^{rd} quartile]: “middle 50 % ” of data

IQR: 3^{rd} quartile – 1^{st} quartile



Width of the box encodes no information.

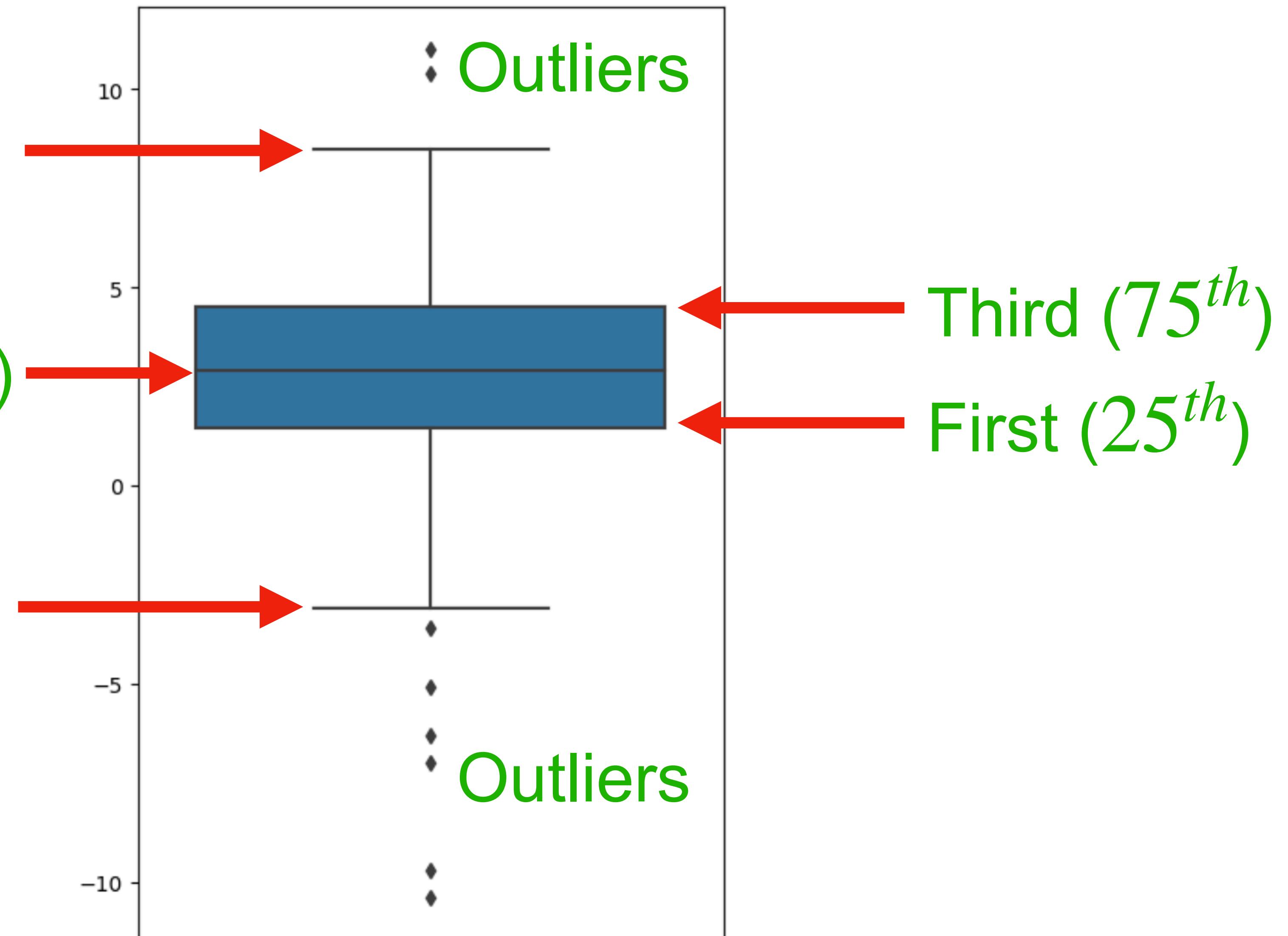
PRACTICE OF DATA VISUALIZATION

Box Plots: for visualizing a continuous quantitative variable.

Whisker: $3^{rd} + 1.5IQR$

Second quartile (median)

Whisker: $1^{st} - 1.5IQR$



PRACTICE OF DATA VISUALIZATION

Violin Plots: for visualizing quantitative variable.

This time let's use the seaborn library.

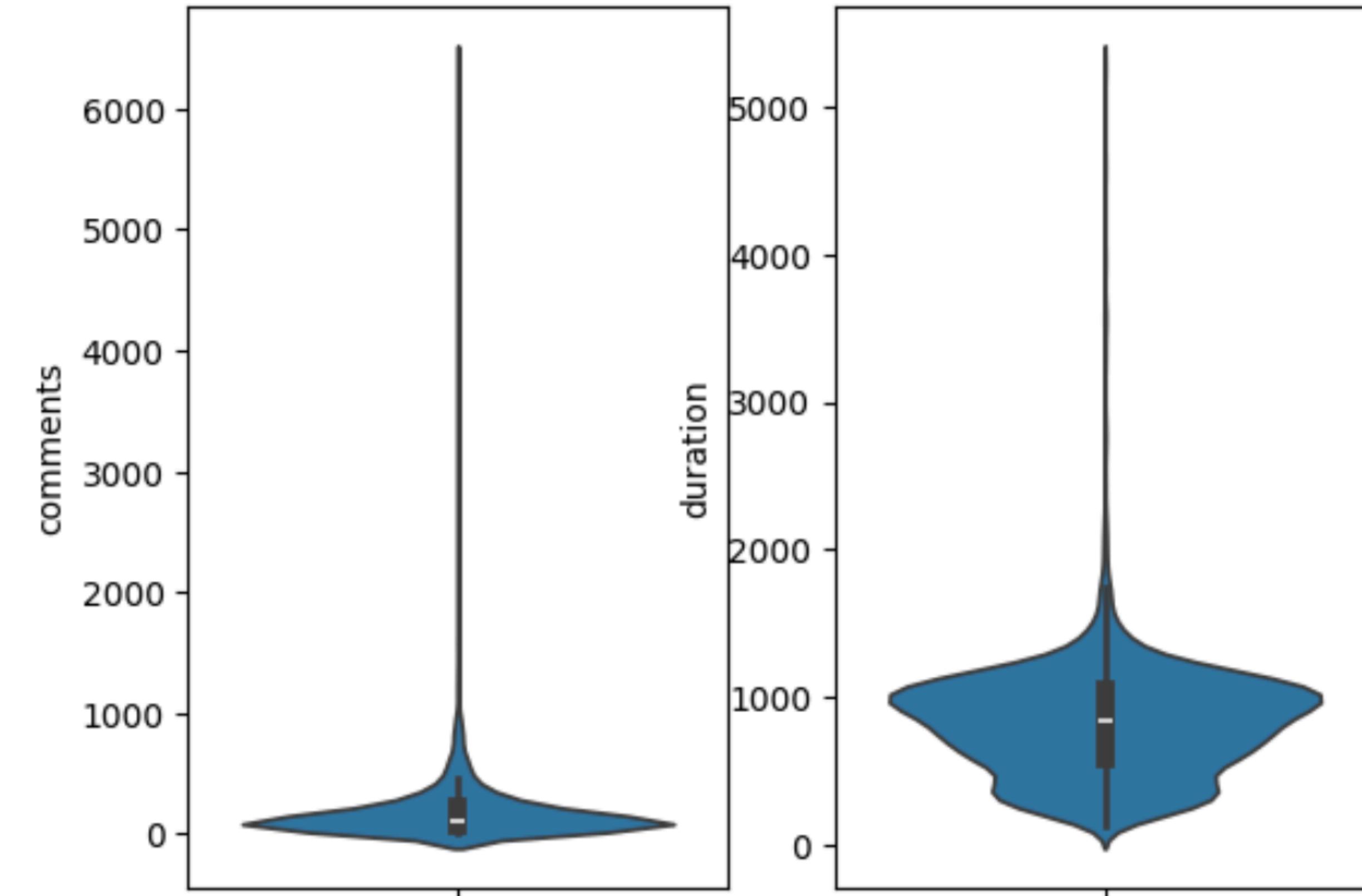
```
import seaborn as sns
fig, (ax1, ax2) = plt.subplots(1, 2)
sns.violinplot(ted_main["comments"], ax=ax1)
sns.violinplot(ted_main["duration"], ax=ax2)
```

PRACTICE OF DATA VISUALIZATION

Violin Plots: for visualizing quantitative variable.

This time let's use the seaborn library.

```
import seaborn as sns
fig, (ax1, ax2) = plt.subplots(1, 2)
sns.violinplot(ted_main["comments"], ax=ax1)
sns.violinplot(ted_main["duration"], ax=ax2)
```



Also shows smoothed density curves: Width has a meaning
Quartiles/whiskers still there.

PRACTICE OF DATA VISUALIZATION

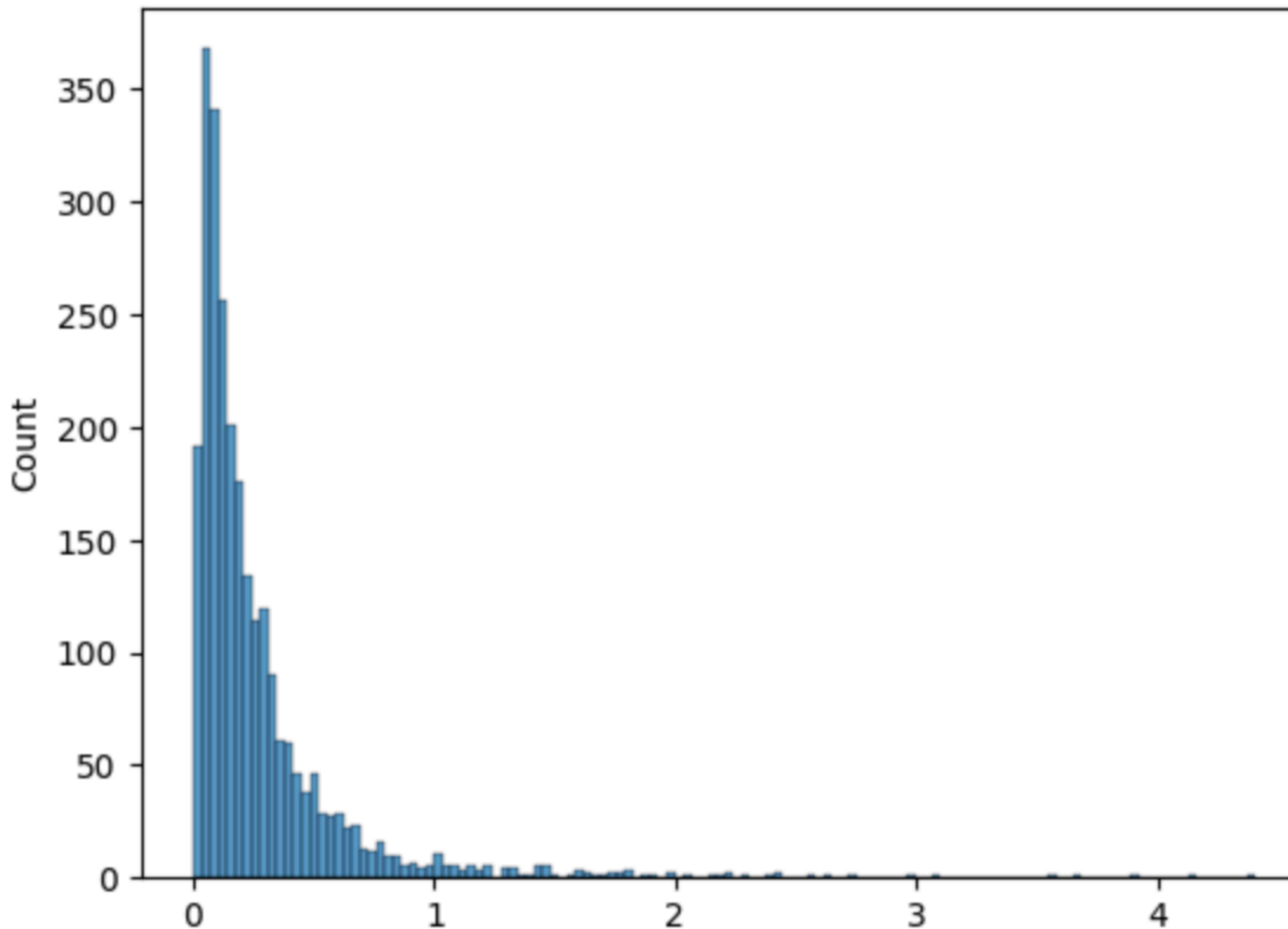
Histogram: collects data points with similar values into bins

```
sns.histplot(ted_main["comments"]/ted_main["duration"])
```

PRACTICE OF DATA VISUALIZATION

Histogram: collects data points with similar values into bins

```
sns.histplot(ted_main["comments"]/ted_main["duration"])
```

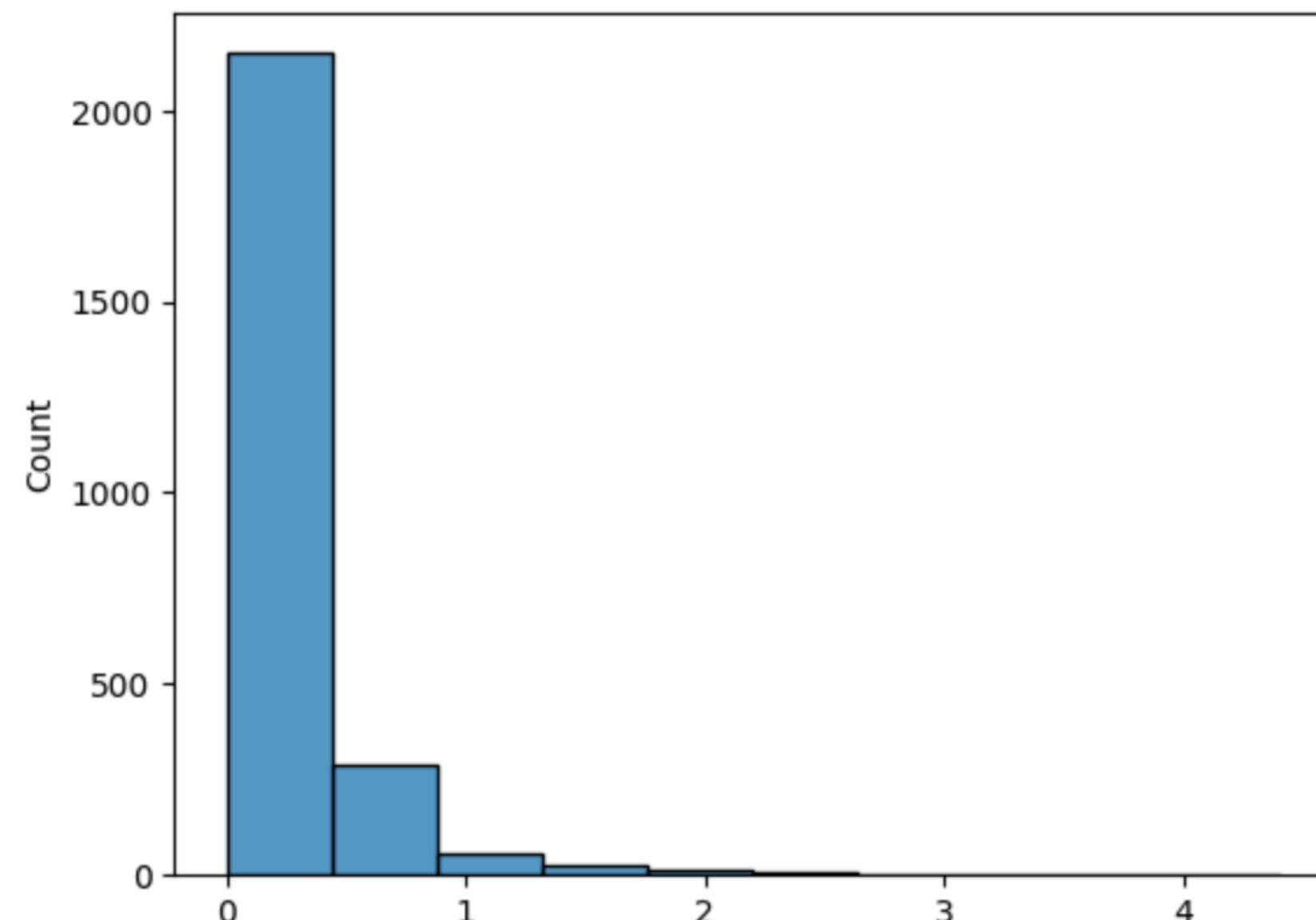


Bin area = % of datapoints in it.

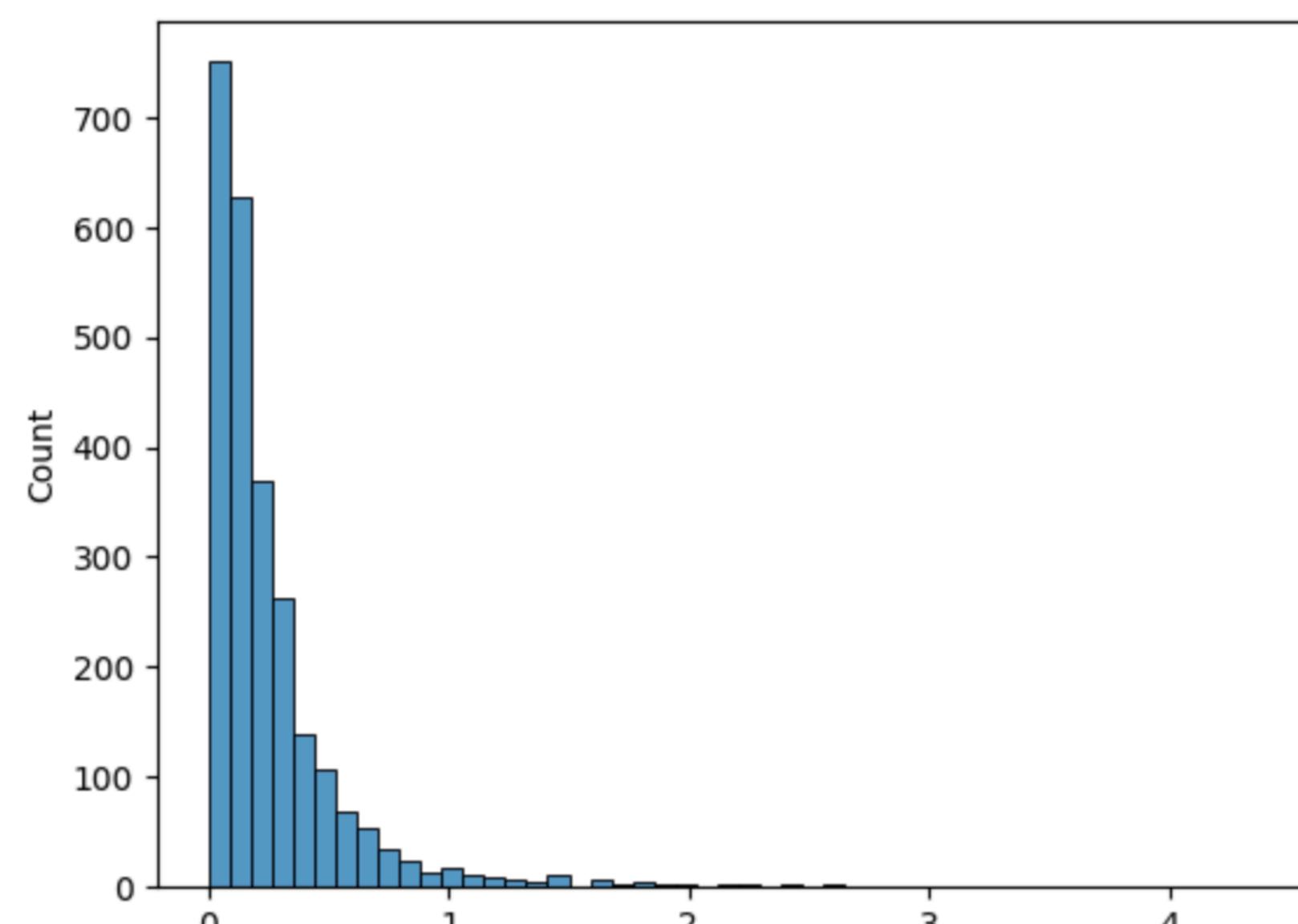
Ex: For the 1st bin:
width = 0.033, height=190
 $\Rightarrow 0.033 \times 190 = 6.3\%$
of the data points in that bin

PRACTICE OF DATA VISUALIZATION

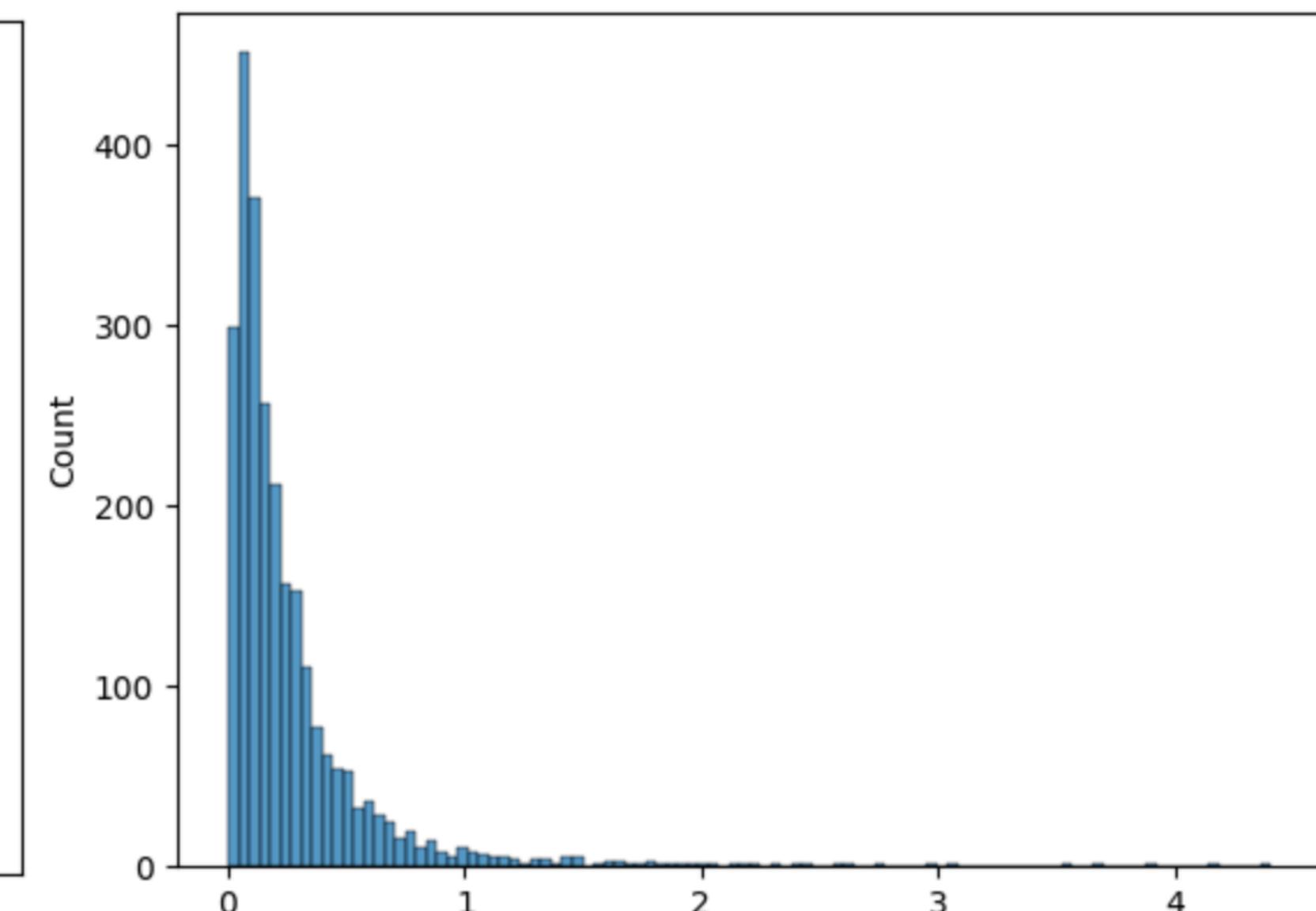
Impact of Bin size in a Histogram



10 bins



50 bins

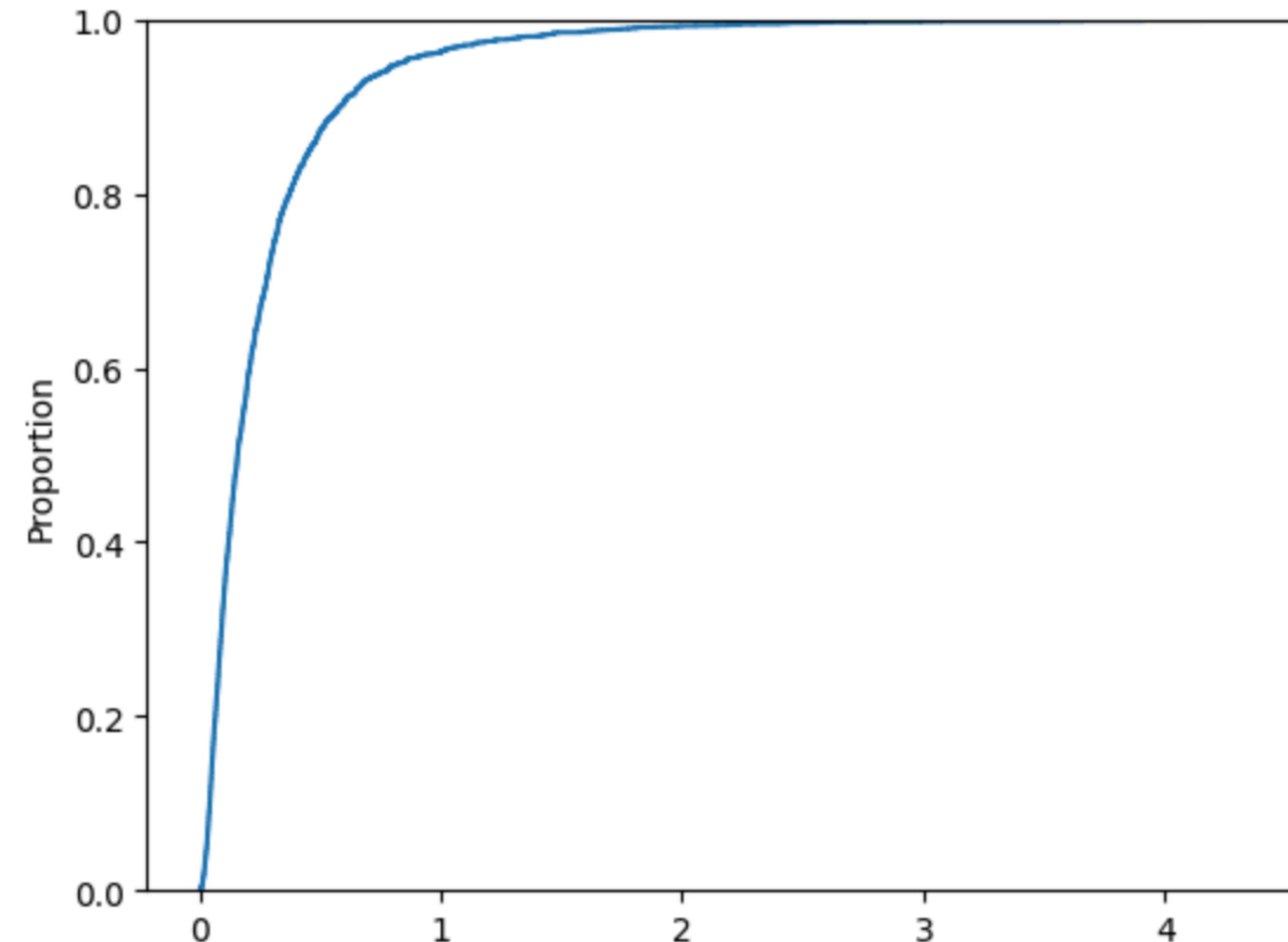


100 bins

PRACTICE OF DATA VISUALIZATION

ECDF: empirical cumulative distribution function

```
sns.ecdfplot(ted_main["comments"]/ted_main["duration"])
```

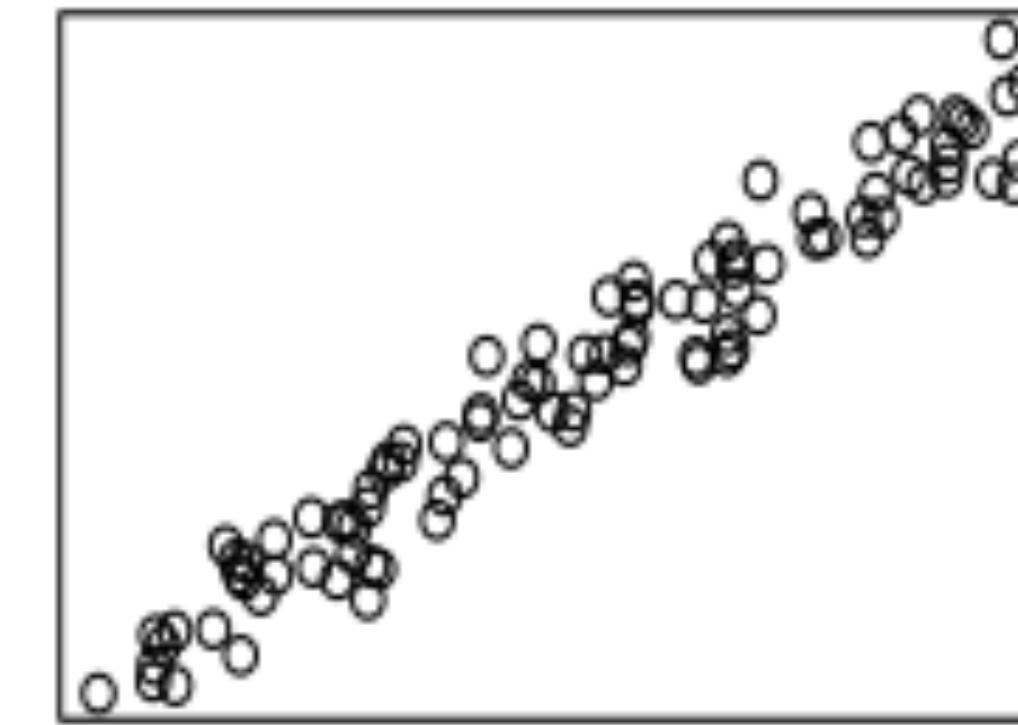


PRACTICE OF DATA VISUALIZATION

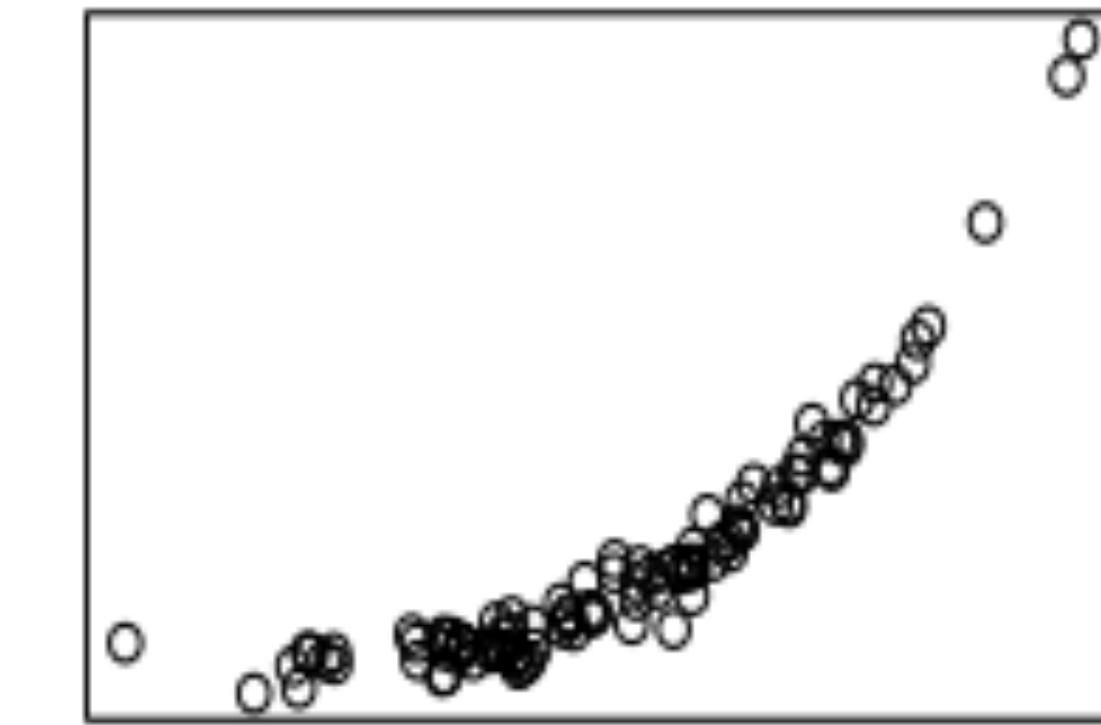
Visualizing Relationship Between Two Variables

Scatter plots

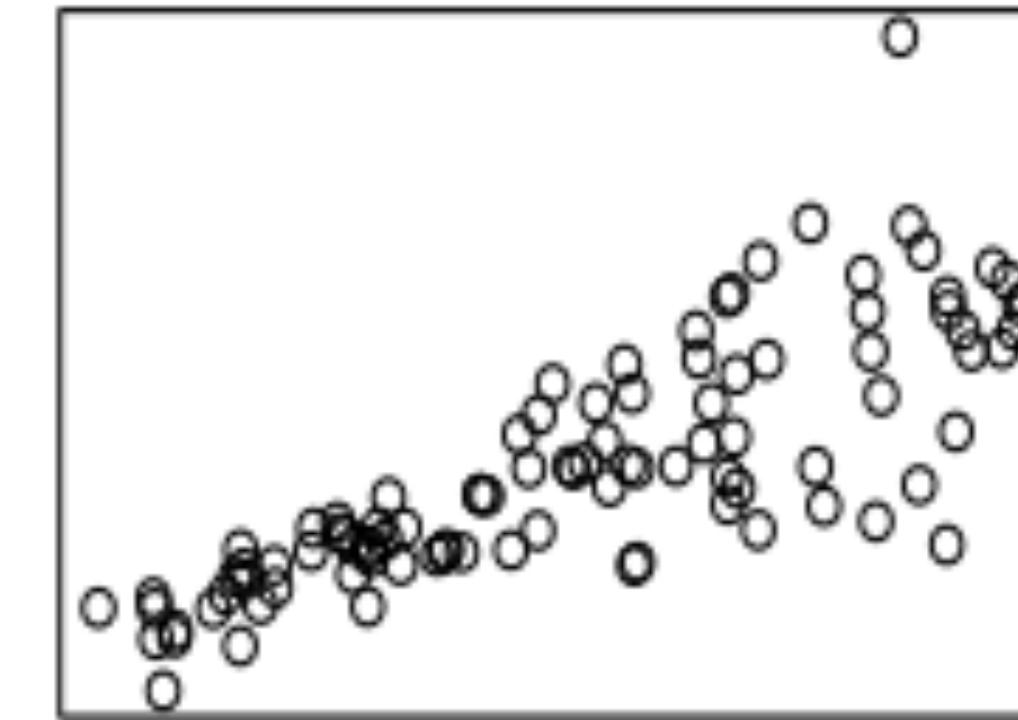
simple linear



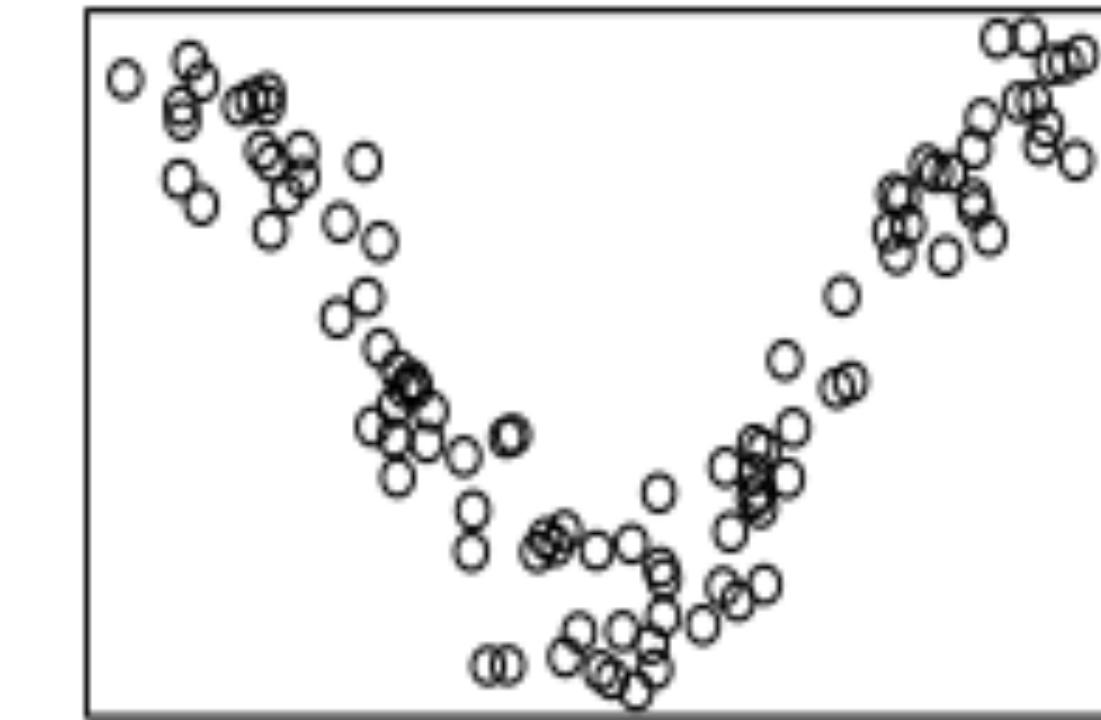
simple nonlinear



unequal spread



complex nonlinear



PRACTICE OF DATA VISUALIZATION

Visualizing Relationship Between Two Variables

Scatter plots

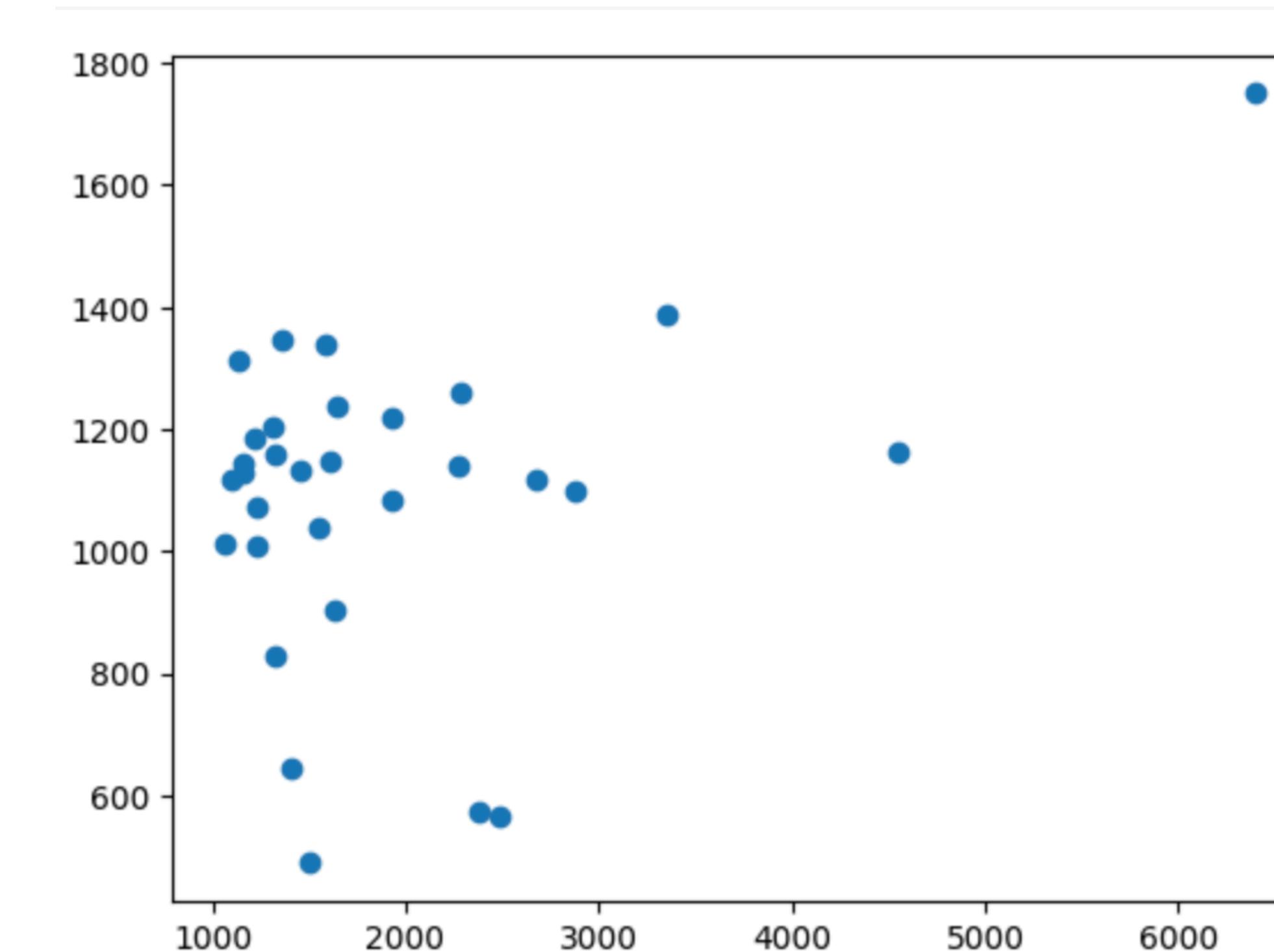
```
plt.scatter(ted_main['comments'][ted_main['comments']>1000],  
            ted_main['duration'][ted_main['comments']>1000])  
plt.show()
```

PRACTICE OF DATA VISUALIZATION

Visualizing Relationship Between Two Variables

Scatter plots

```
plt.scatter(ted_main['comments'][ted_main['comments']>1000],  
            ted_main['duration'][ted_main['comments']>1000])  
plt.show()
```



PRACTICE OF DATA VISUALIZATION

Visualizing Relationship Between Two Variables

Hex plots

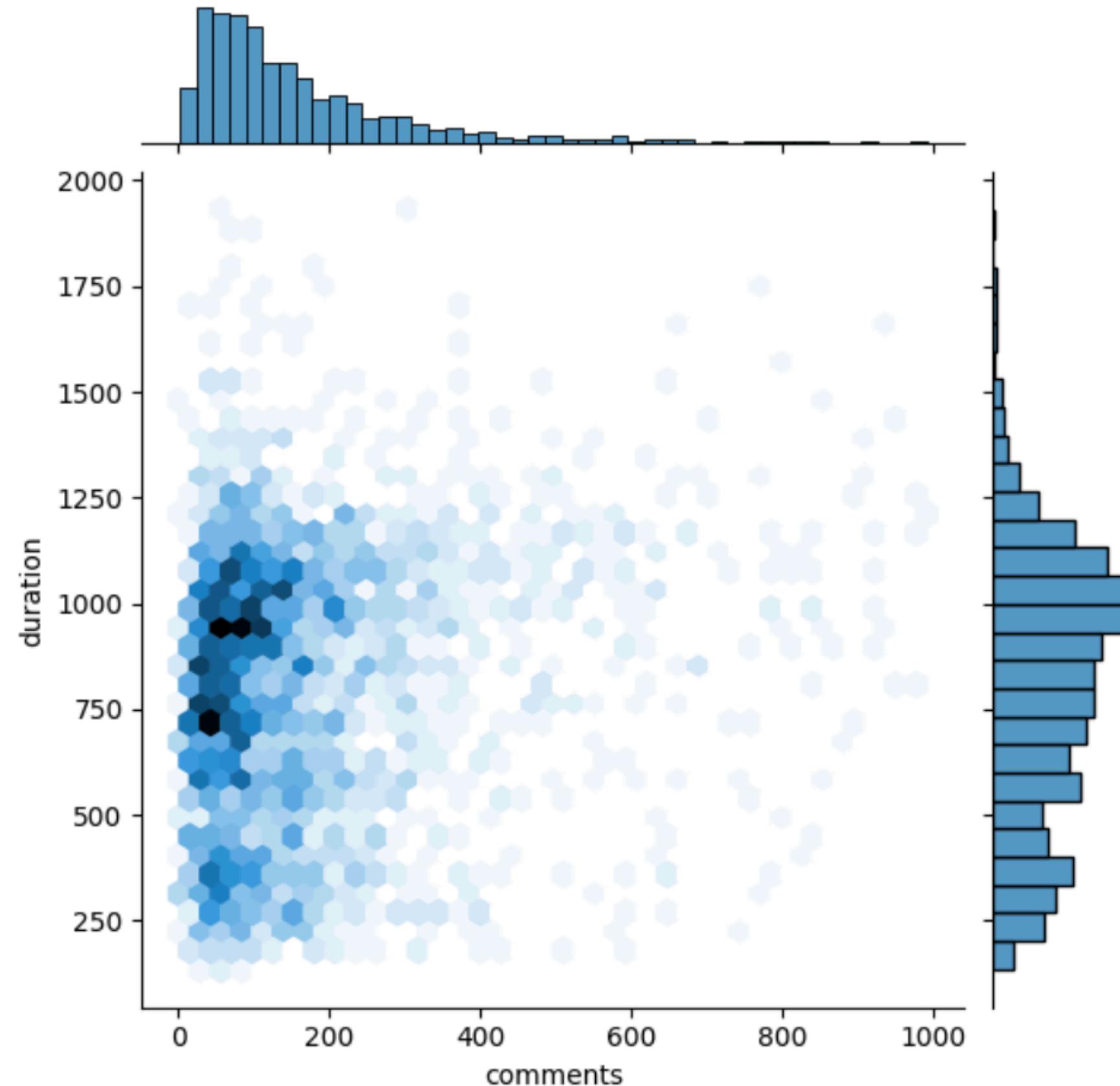
```
ted_filtered = ted_main[(ted_main['comments']<1000) &  
                         (ted_main['duration']<2000)]  
sns.jointplot(data=ted_filtered, x="comments", y="duration", kind="hex")
```

PRACTICE OF DATA VISUALIZATION

Visualizing Relationship Between Two Variables

Hex plots

```
ted_filtered = ted_main[(ted_main['comments']<1000) &  
                         (ted_main['duration']<2000)]  
sns.jointplot(data=ted_filtered, x="comments", y="duration", kind="hex")
```



Rather than individual datapoints, plot the density of their joint distribution.

- a two dimensional histogram.
- xy plane binned into hexagons.

Darker hexagons \Rightarrow more datapoints

PRACTICE OF DATA VISUALIZATION

Visualizing Relationship Between Two Variables

Contour plots

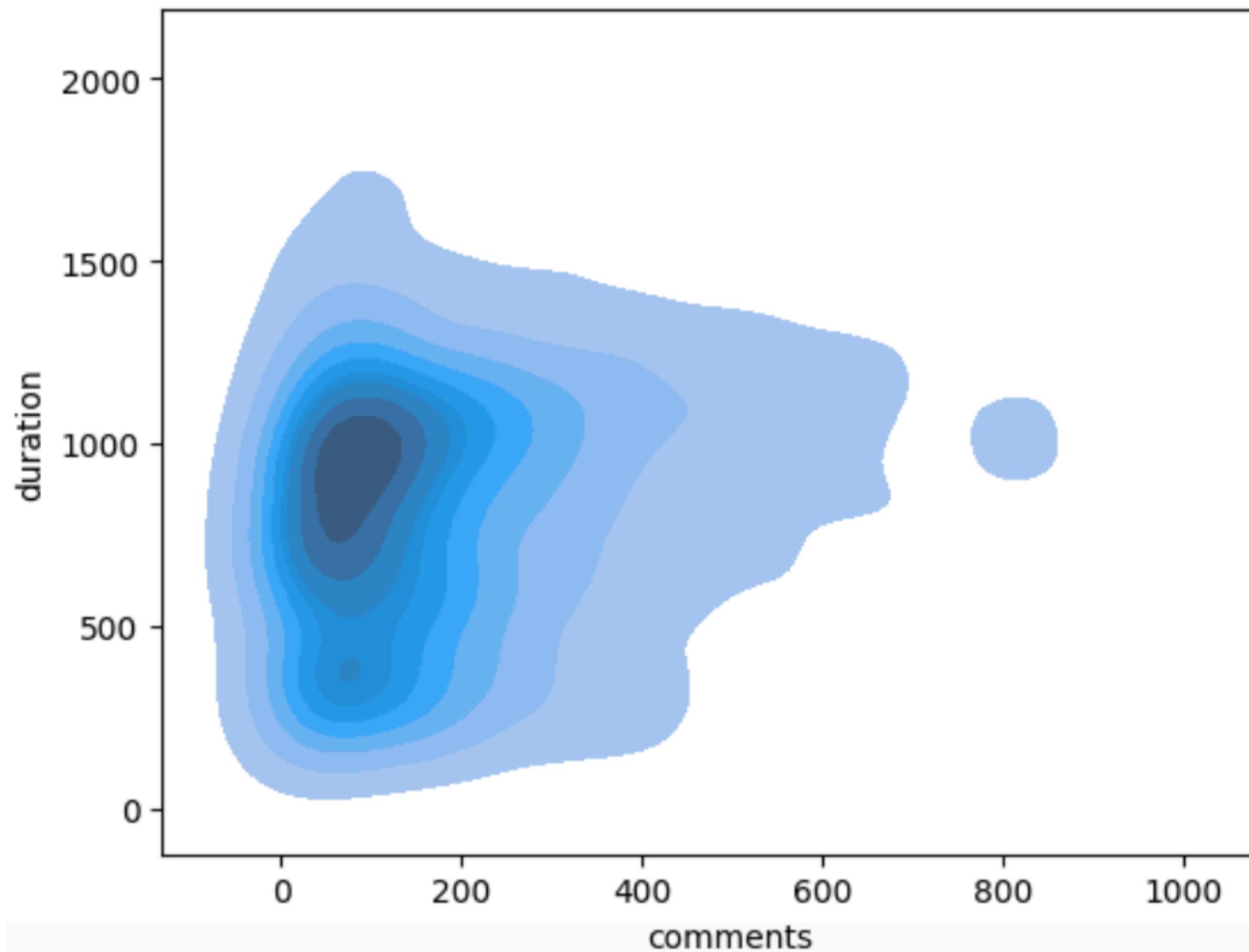
```
sns.kdeplot(data=ted_filtered, x="comments", y="duration", fill=True)
```

PRACTICE OF DATA VISUALIZATION

Visualizing Relationship Between Two Variables

Contour plots

```
sns.kdeplot(data=ted_filtered, x="comments", y="duration", fill=True)
```



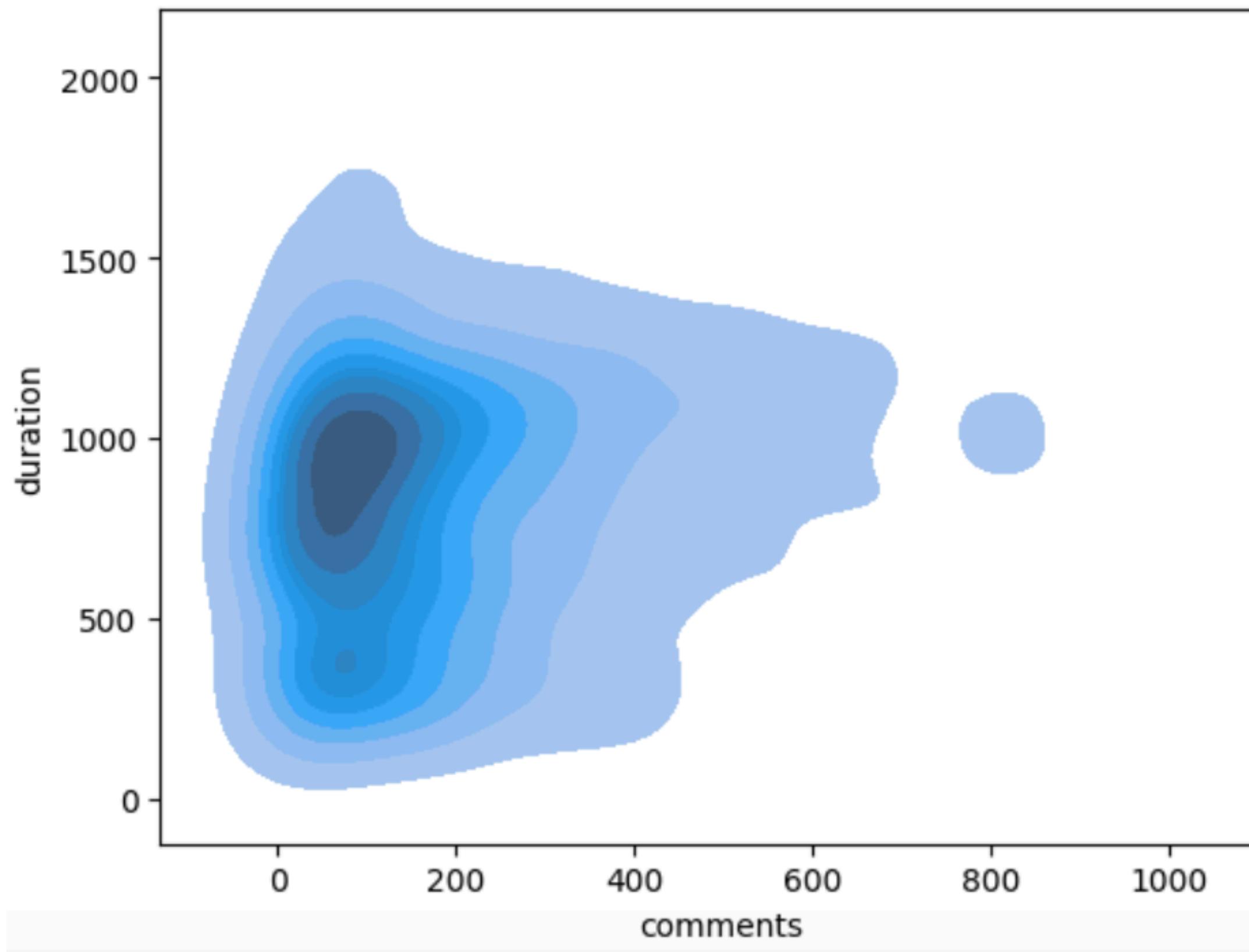
Contour lines represent an area with the same density of datapoints in it.

PRACTICE OF DATA VISUALIZATION

Visualizing Relationship Between Two Variables

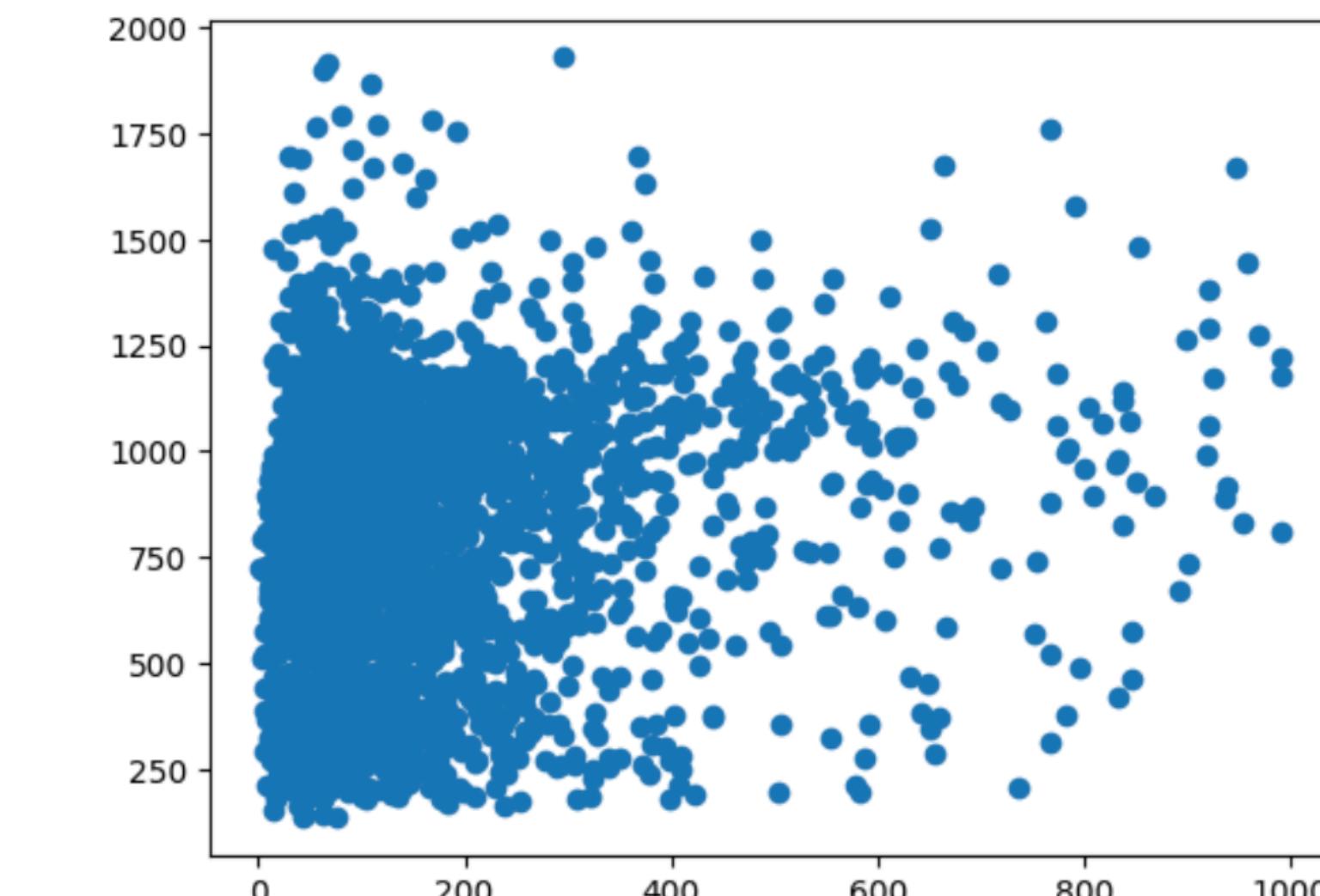
Contour plots

```
sns.kdeplot(data=ted_filtered, x="comments", y="duration", fill=True)
```



Contour lines represent an area with the same density of datapoints in it.

Compare against the scatter plot:



DATA VISUALIZATION (CONT.)

Reading:
Chapter 6 from Skiena.

PRINCIPLES OF DATA VISUALIZATION

'Graphical excellence begins with telling the truth about the data' E. Tufte

Slides will make heavy use of 'The Visual Display of Quantitative Information' by Edward Tufte.

PRINCIPLES OF DATA VISUALIZATION

'Graphical excellence begins with telling the truth about the data' E. Tufte

Slides will make heavy use of 'The Visual Display of Quantitative Information' by Edward Tufte.

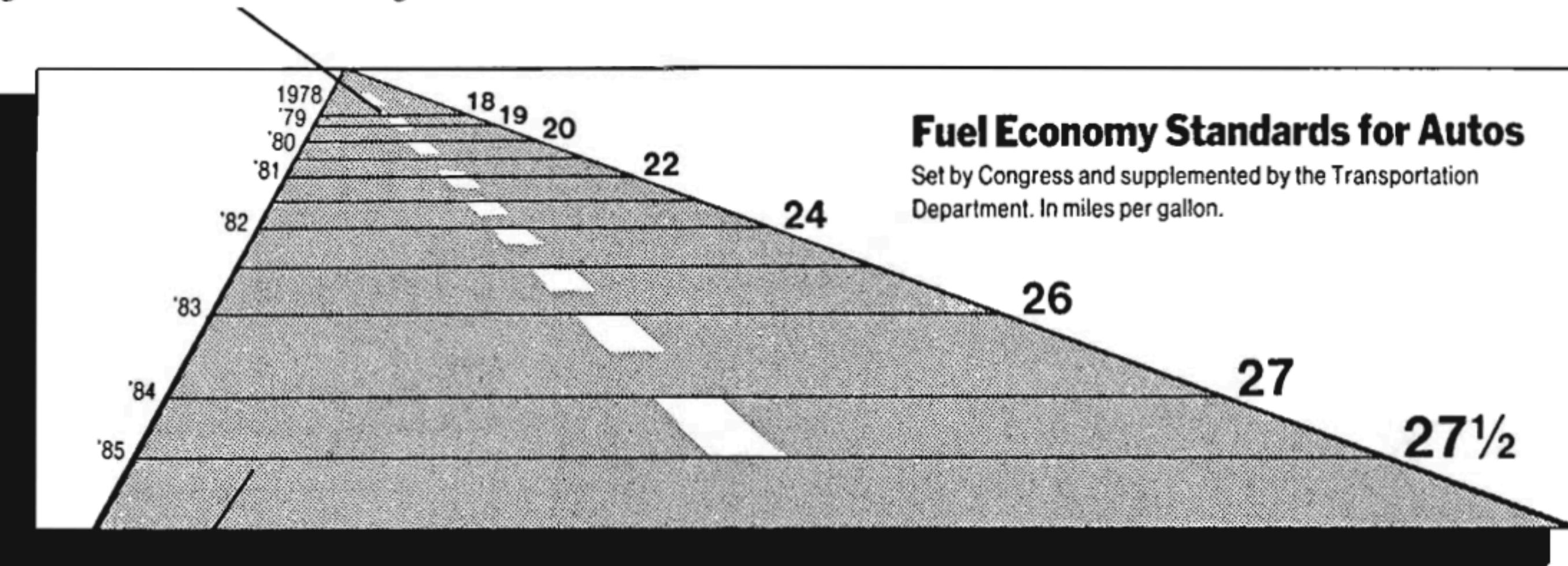
Lie Factor = size of effect shown in graphic / size of effect in data

PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Design and Data Variation

Is this a good visualization?

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

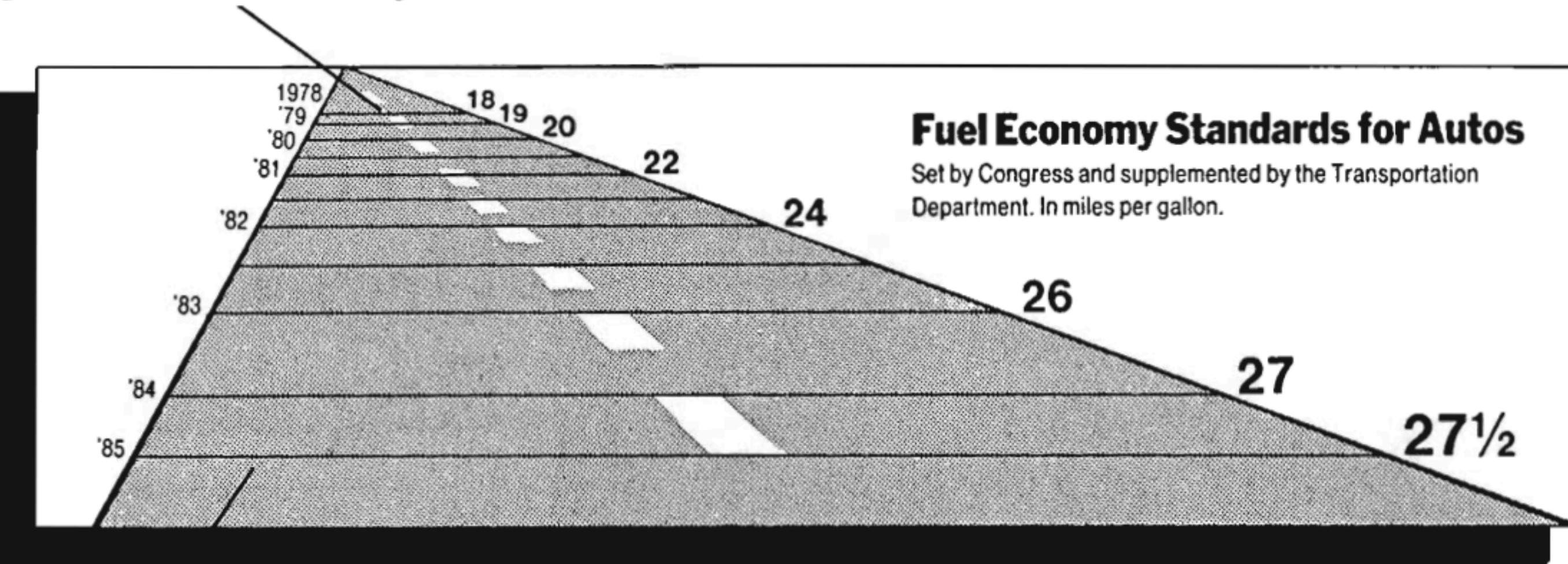
E. Tufte

PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Design and Data Variation

Is this a good visualization?

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

E. Tufte

- Which direction is the future?
- Perspective on left numbers
- Road width: perspective or change in values

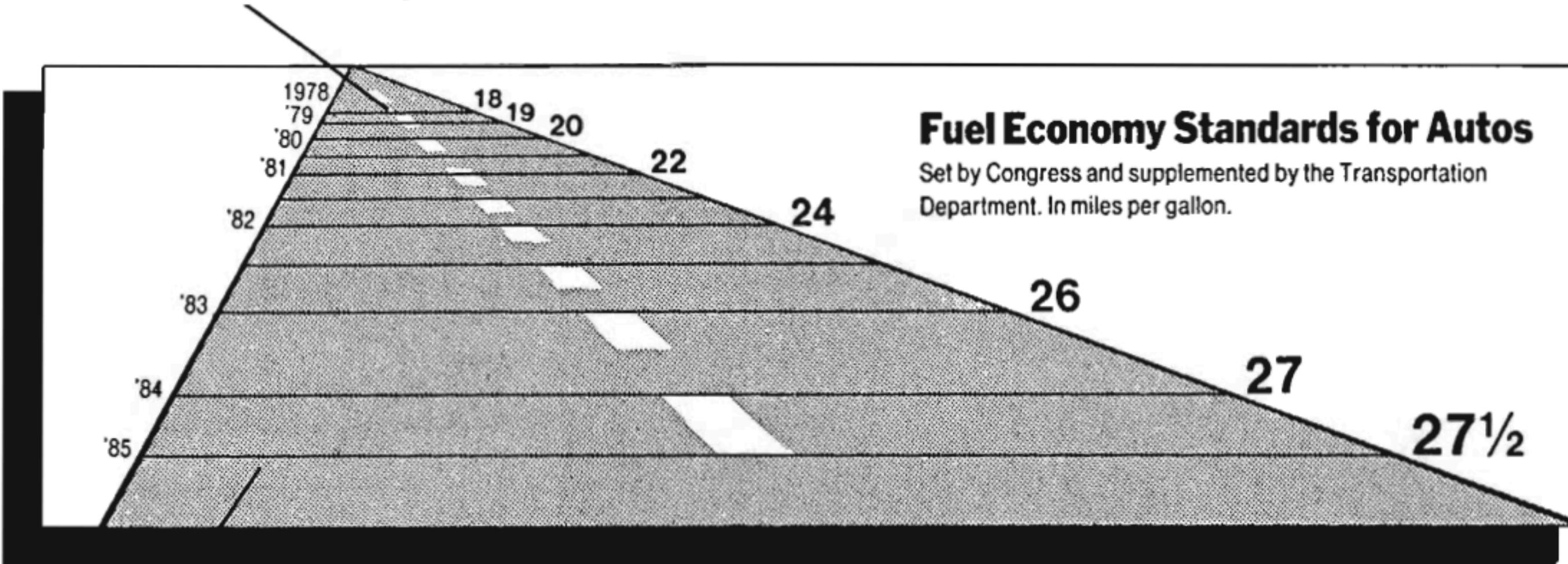
Biggest mistake: Lie factor

PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Design and Data Variation

Is this a good visualization?

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

E. Tufte

Actual data:

$$\frac{27.5 - 18.0}{18.0} = 53\%$$

In graphic:

$$\frac{5.3 - 0.6}{0.6} = 783\%$$

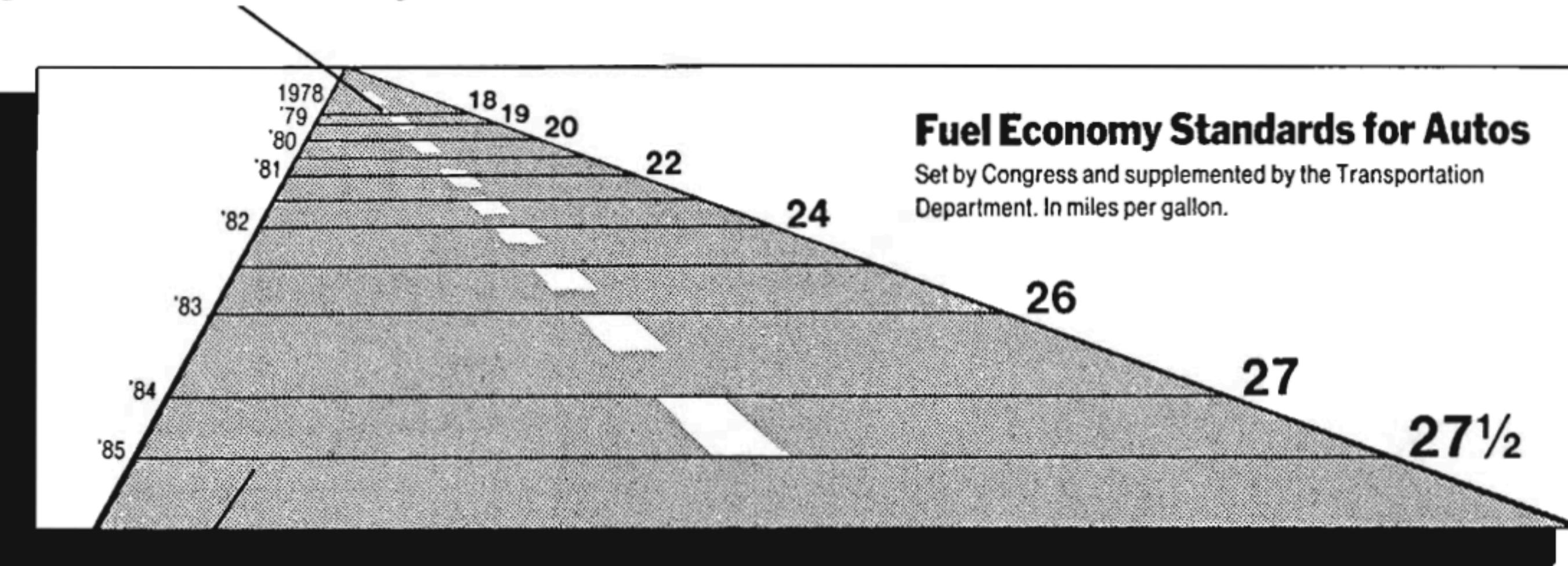
Lie factor = $783/53 = 14.8$

PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Design and Data Variation

Is this a good visualization?

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



E. Tufte

Problem with the lie factor:

Trying to capture both design variation and data variation.

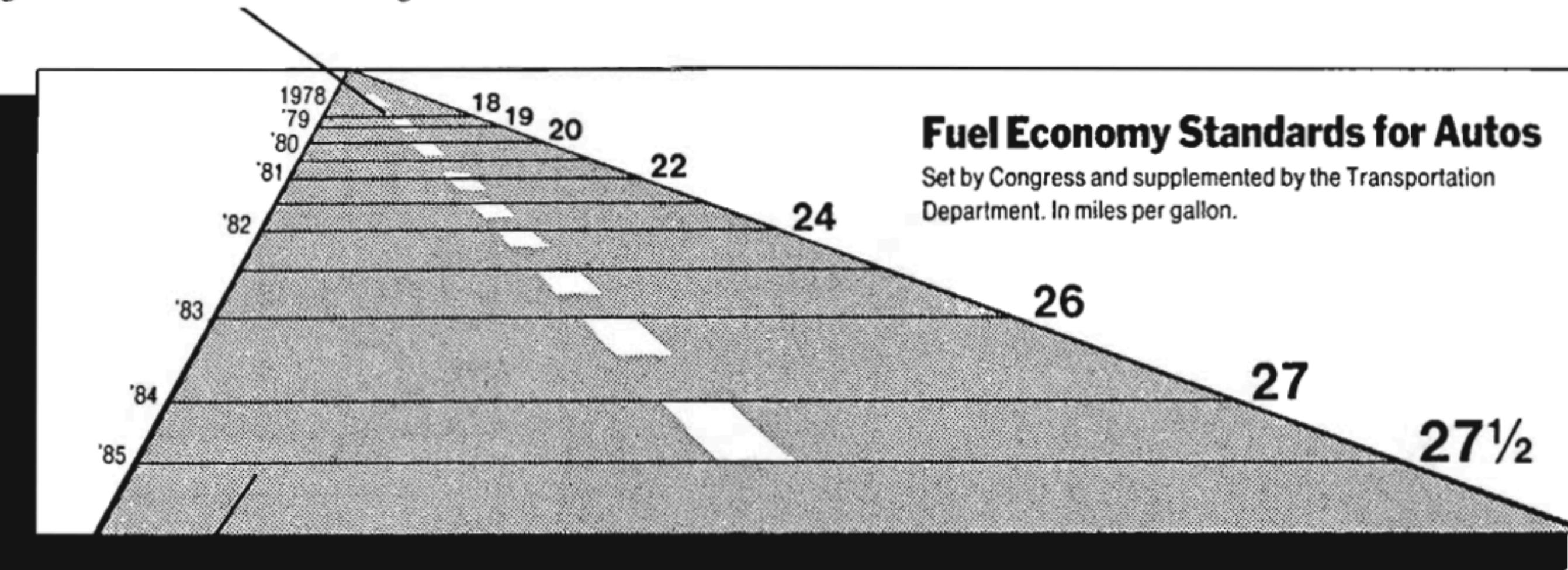
Need to show data variation not design variation!

PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Design and Data Variation

Is this a good visualization?

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.

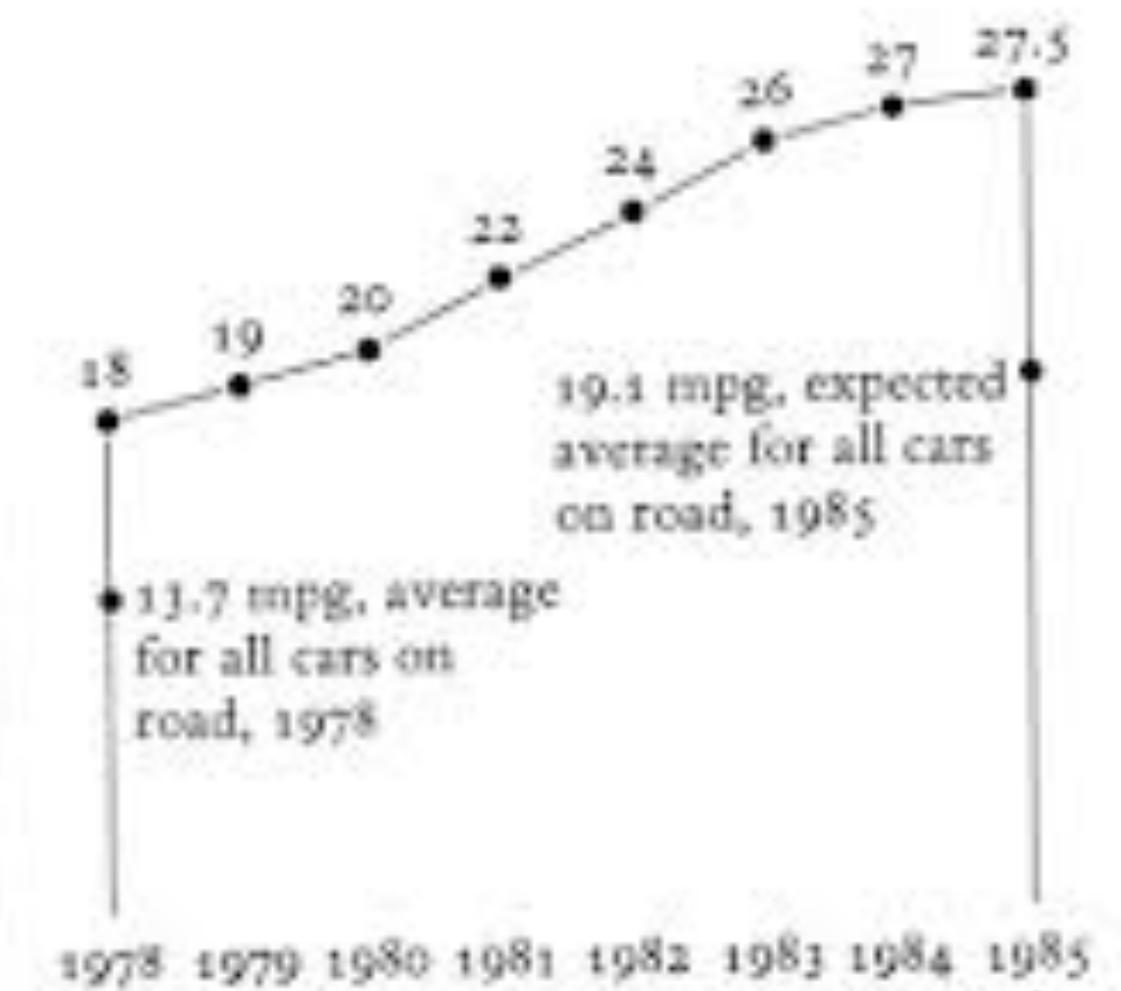


This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

E. Tufte

Visualization matching data

REQUIRED FUEL ECONOMY STANDARDS:
NEW CARS BUILT FROM 1978 TO 1985



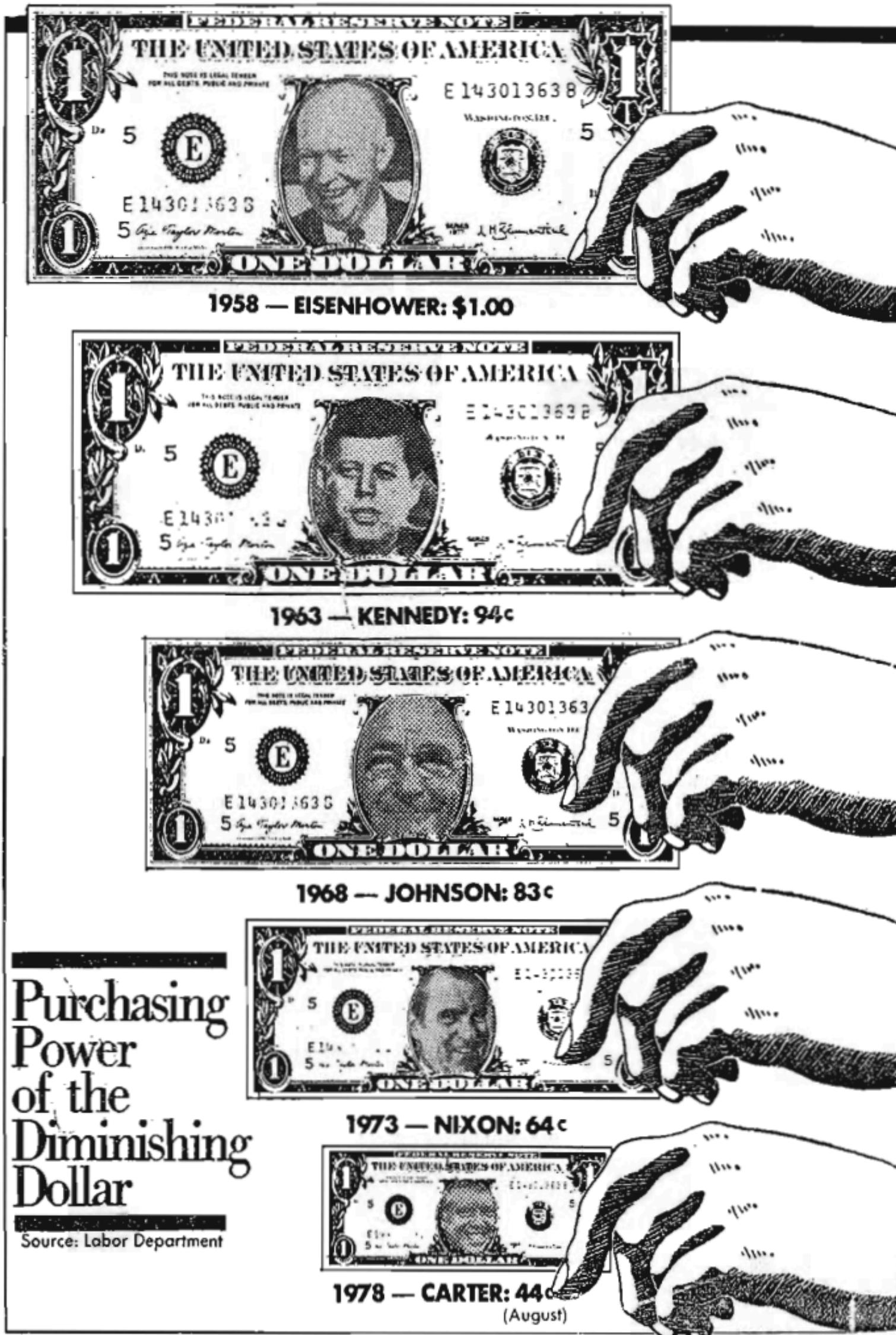
PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Visual Area and Numerical Measure



PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Visual Area and Numerical Measure



Not a good idea to use area/volume to show one-dimensional data.

Bottom dollar's area vs top dollar's area?

PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Visual Area and Numerical Measure



Not a good idea to use area/volume to show one-dimensional data.

Bottom dollar's area vs top dollar's area?

Bottom dollar's value 44, top dollar's value 100

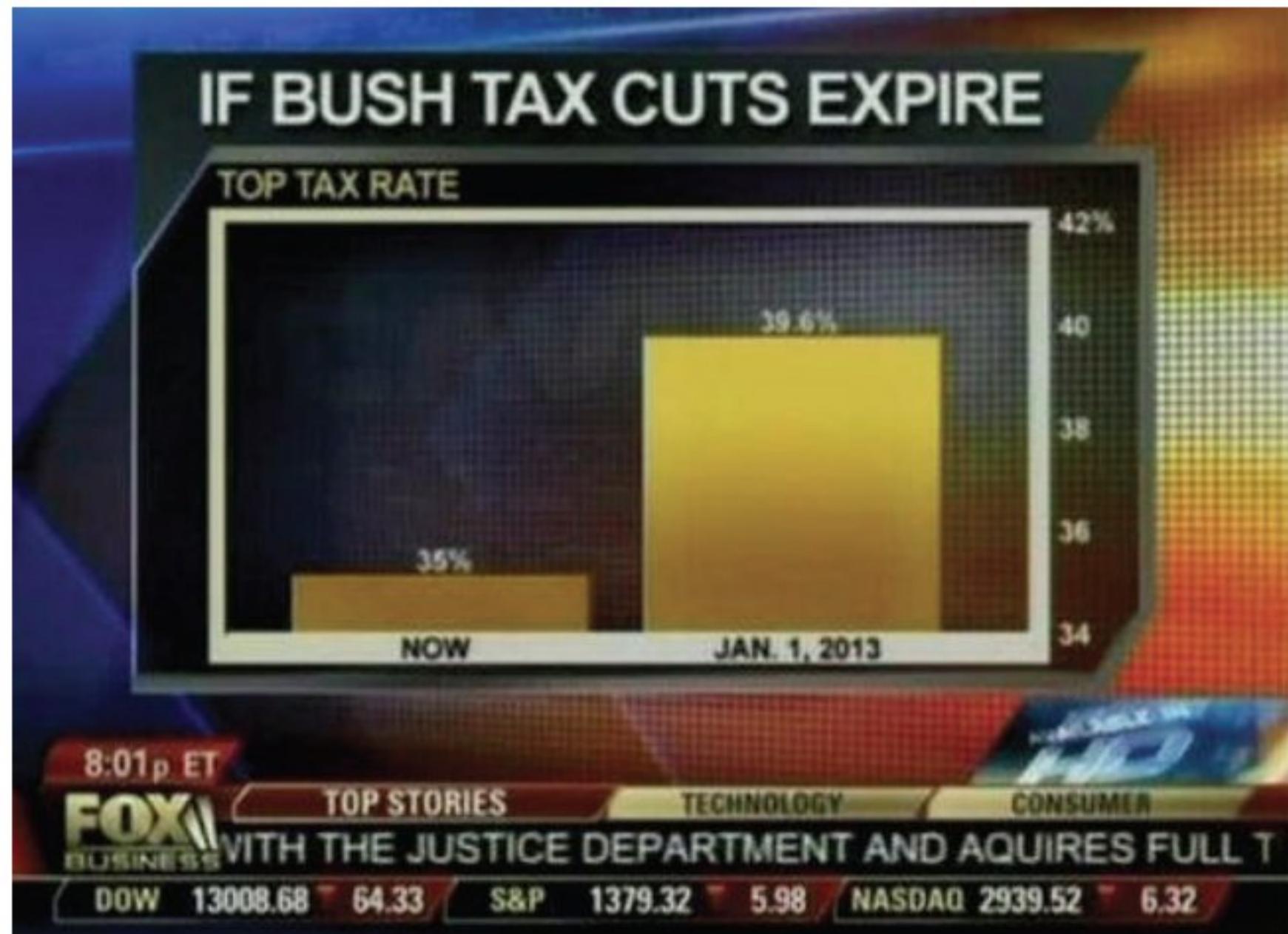
Bottom dollar had to be twice in size.

PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Offset Distortion

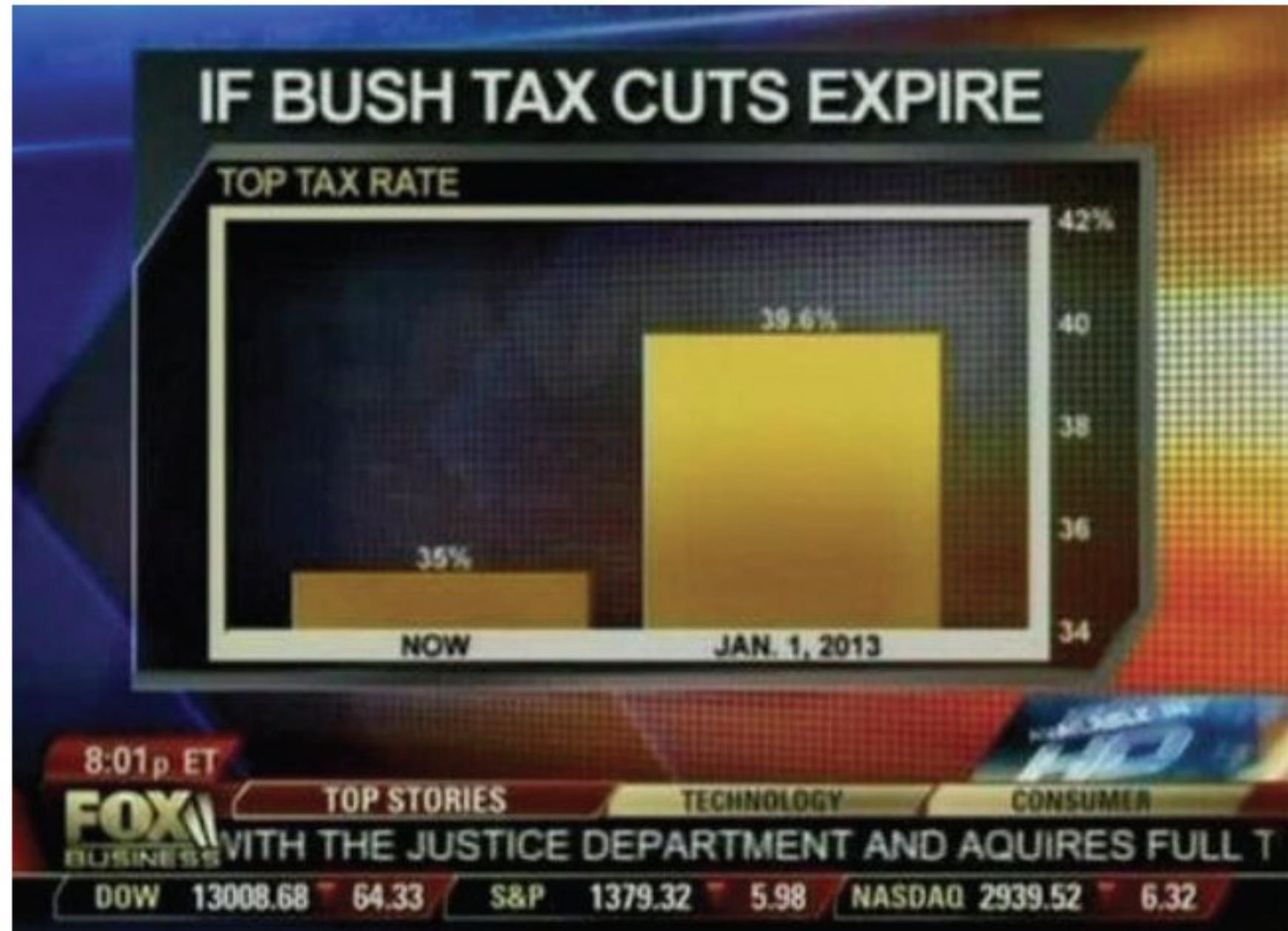
PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Offset Distortion



PRINCIPLES OF DATA VISUALIZATION

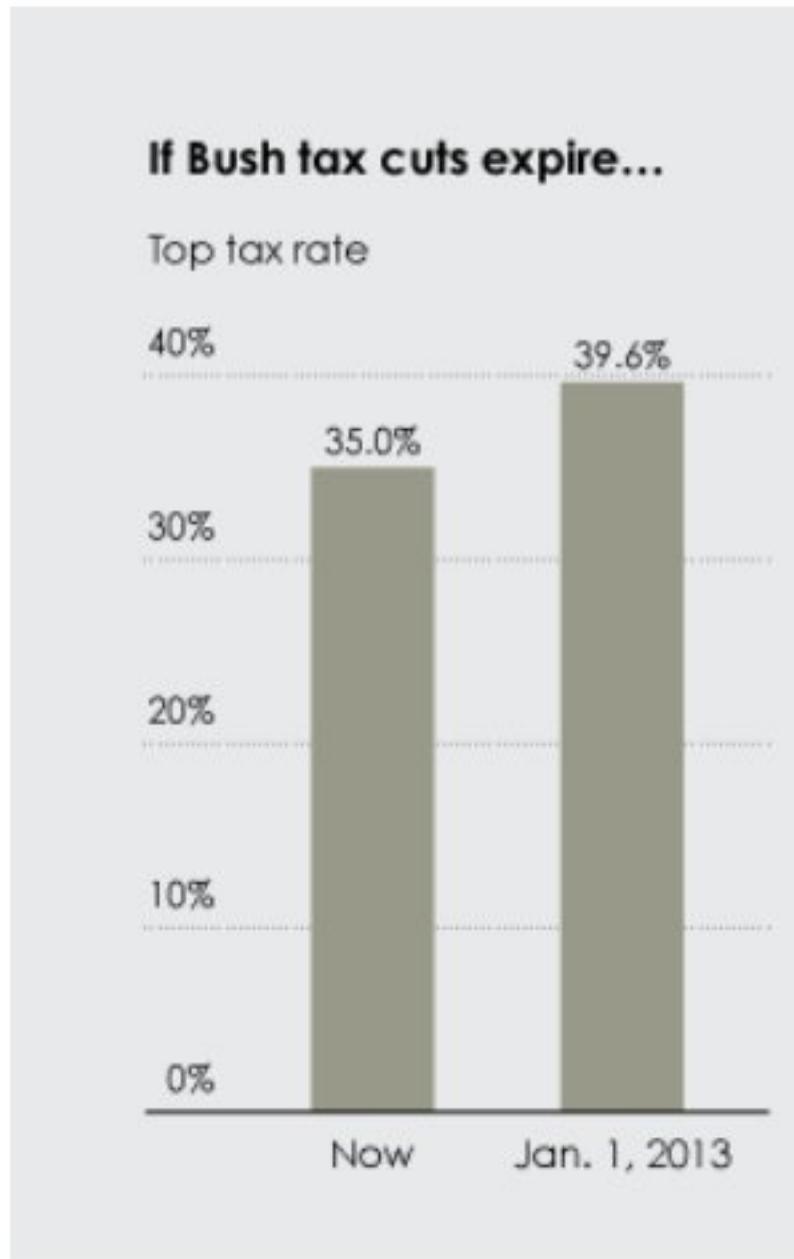
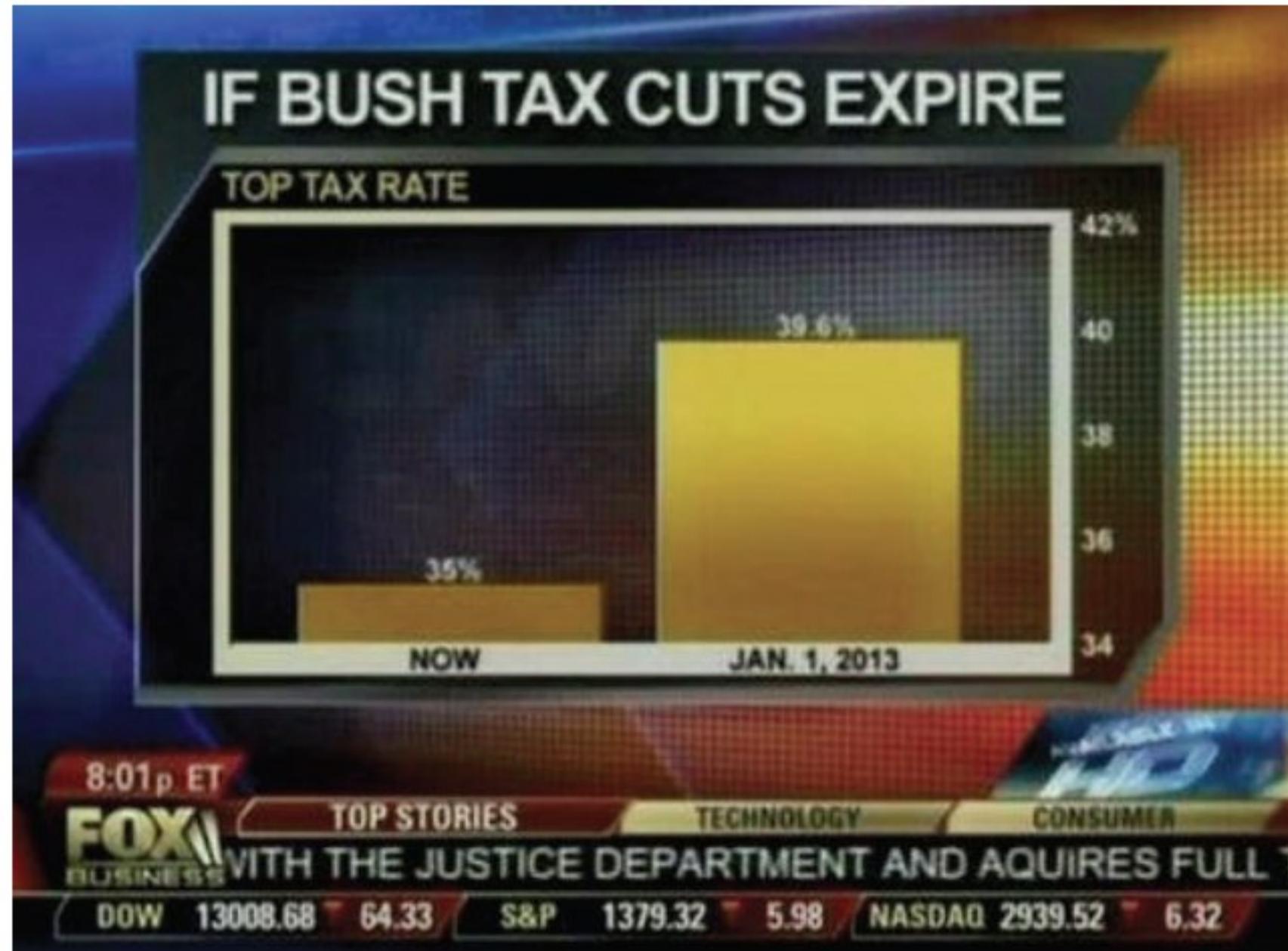
Graphical Integrity: Offset Distortion



What a difference!

PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Offset Distortion



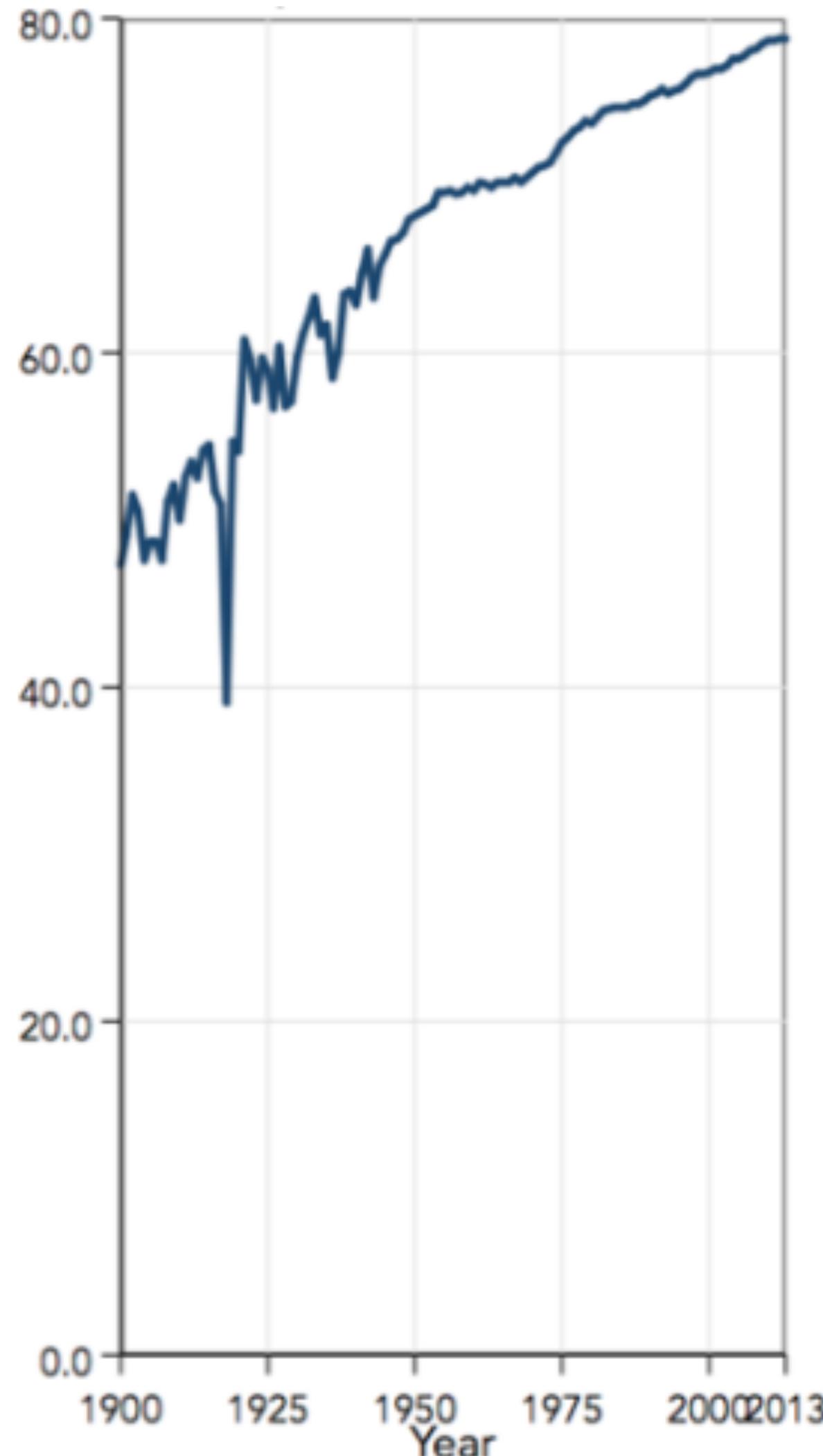
What a difference!

Actual difference not so spectacular!

Start bar graphs at 0.

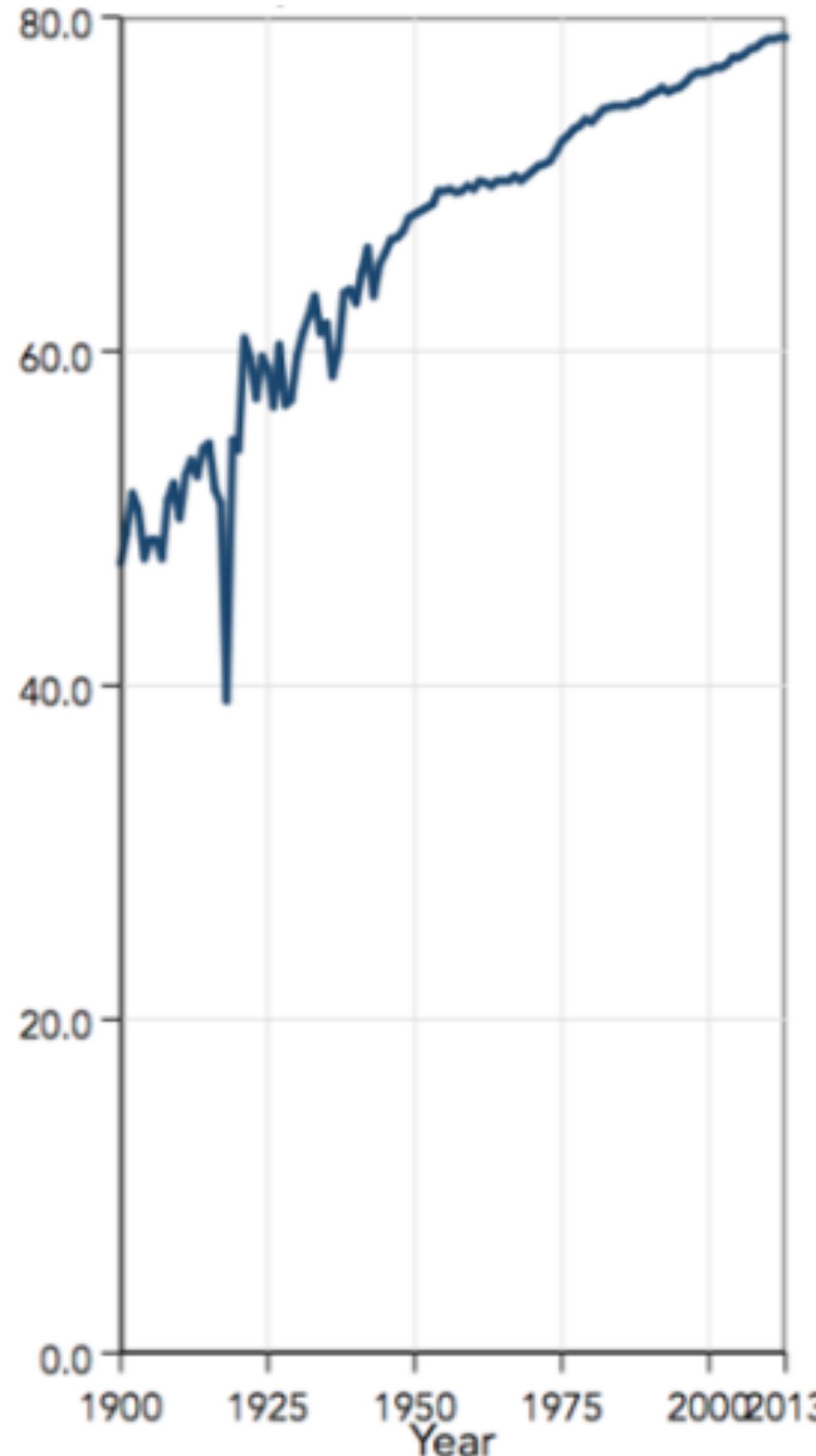
PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Scale Distortion



PRINCIPLES OF DATA VISUALIZATION

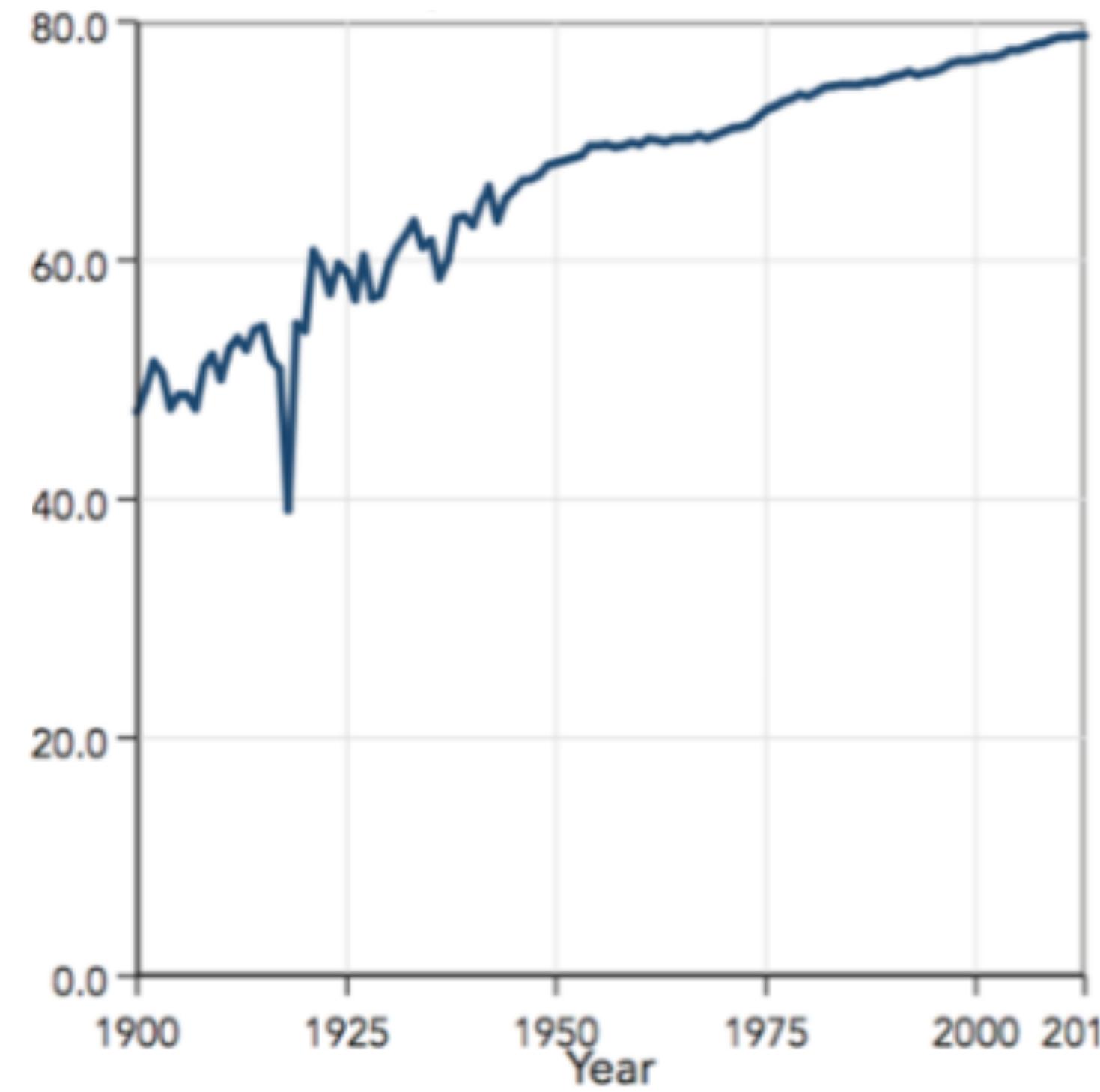
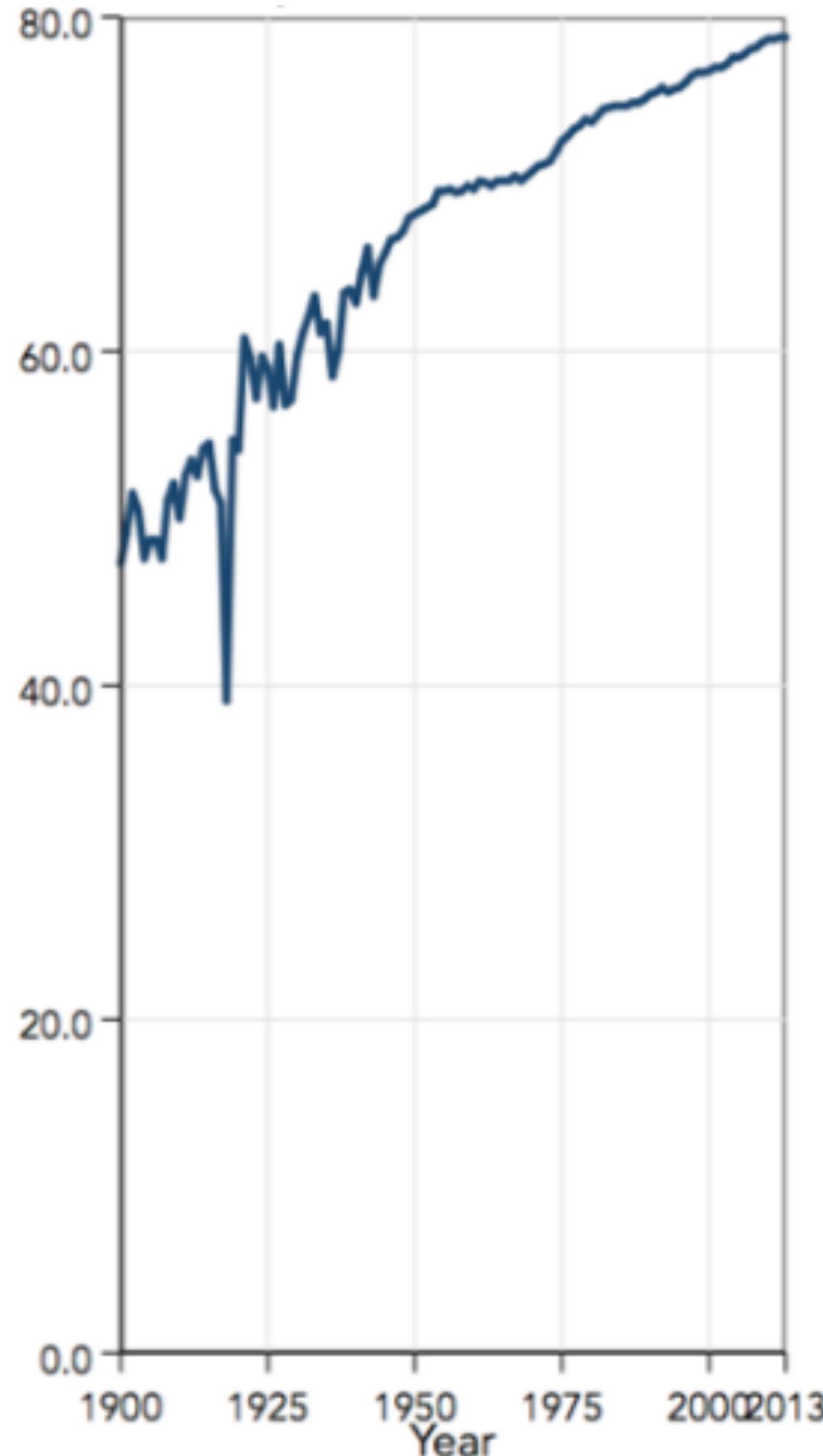
Graphical Integrity: Scale Distortion



Something is steeply increasing.
How serious is the problem?

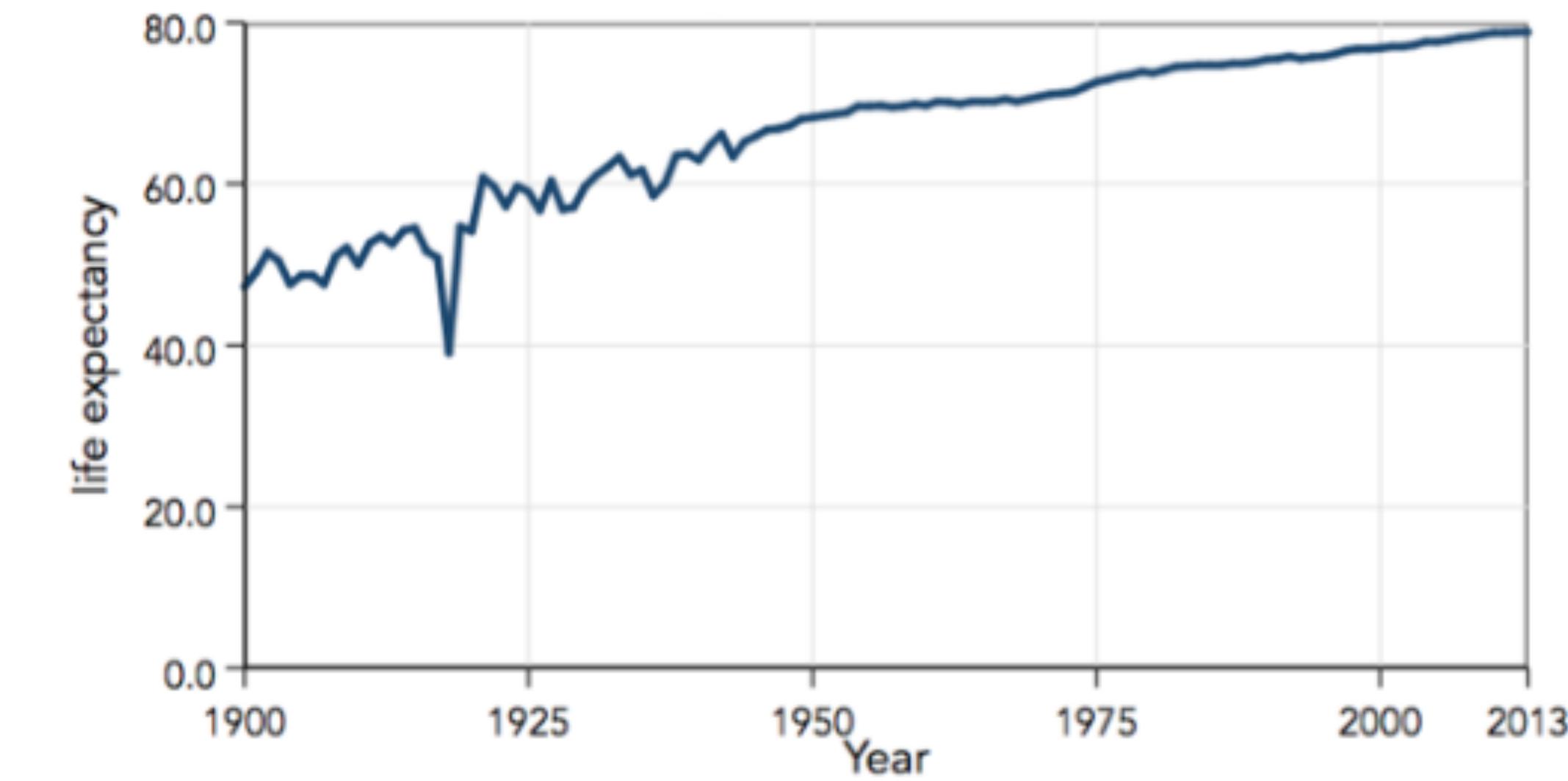
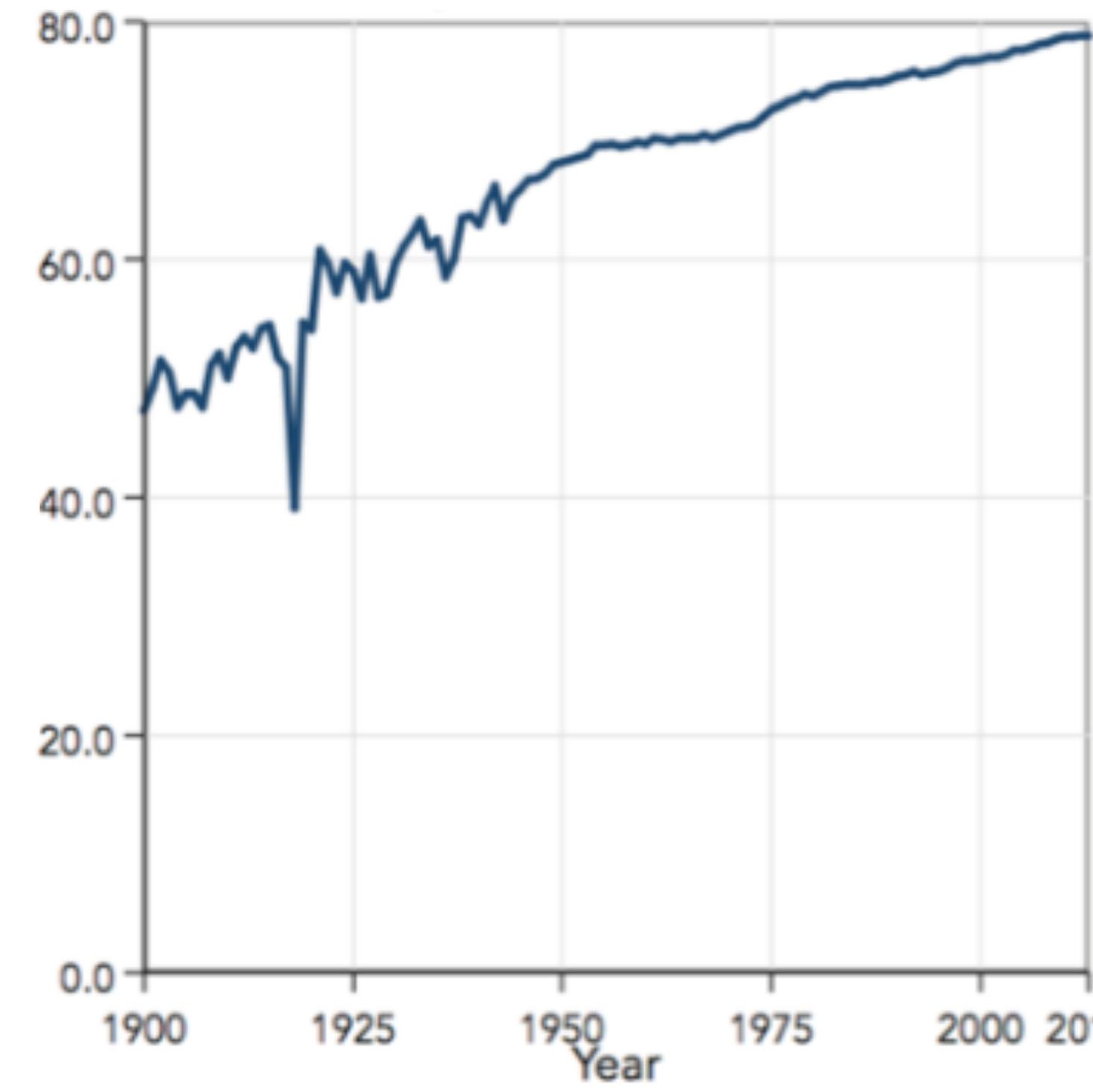
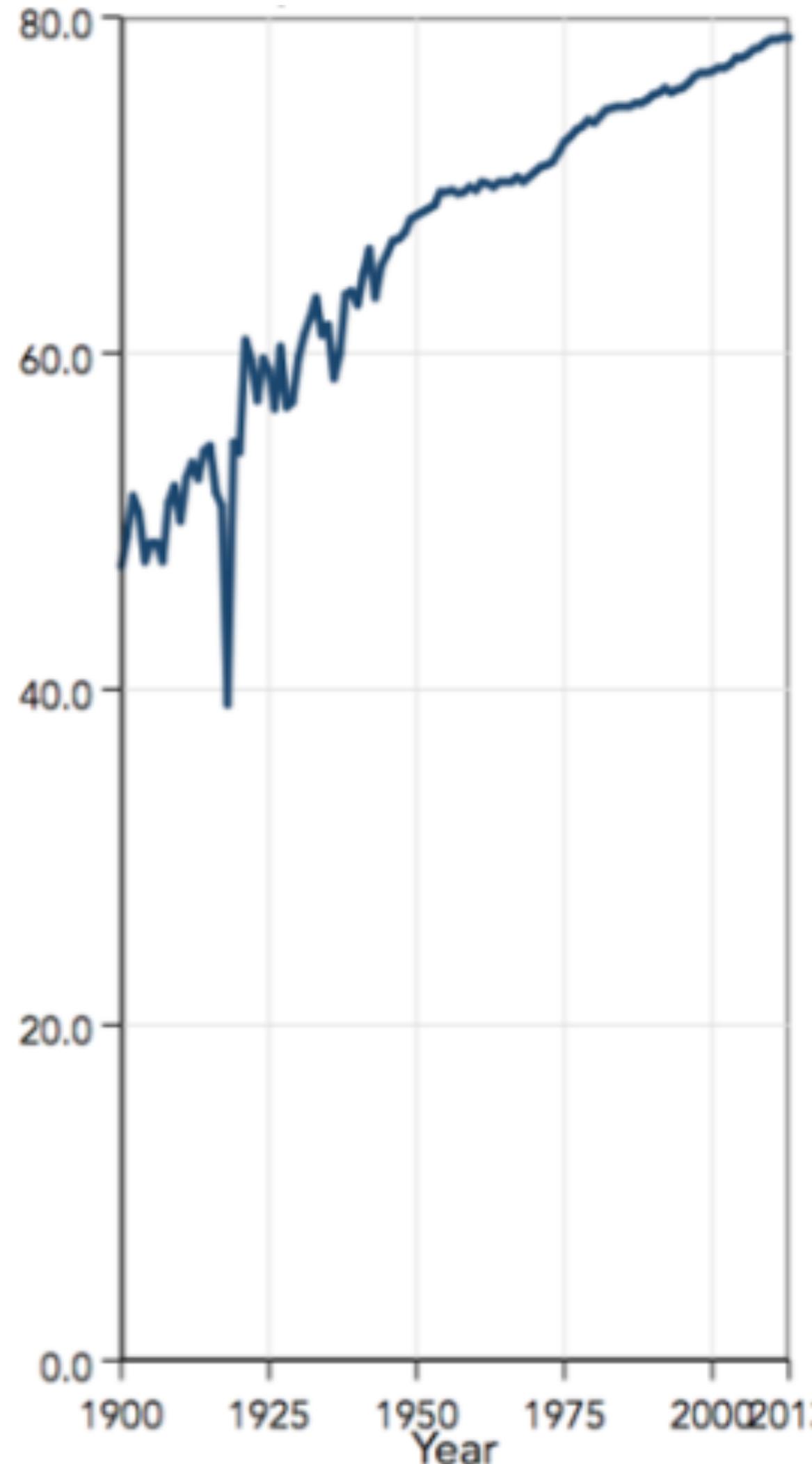
PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Scale Distortion



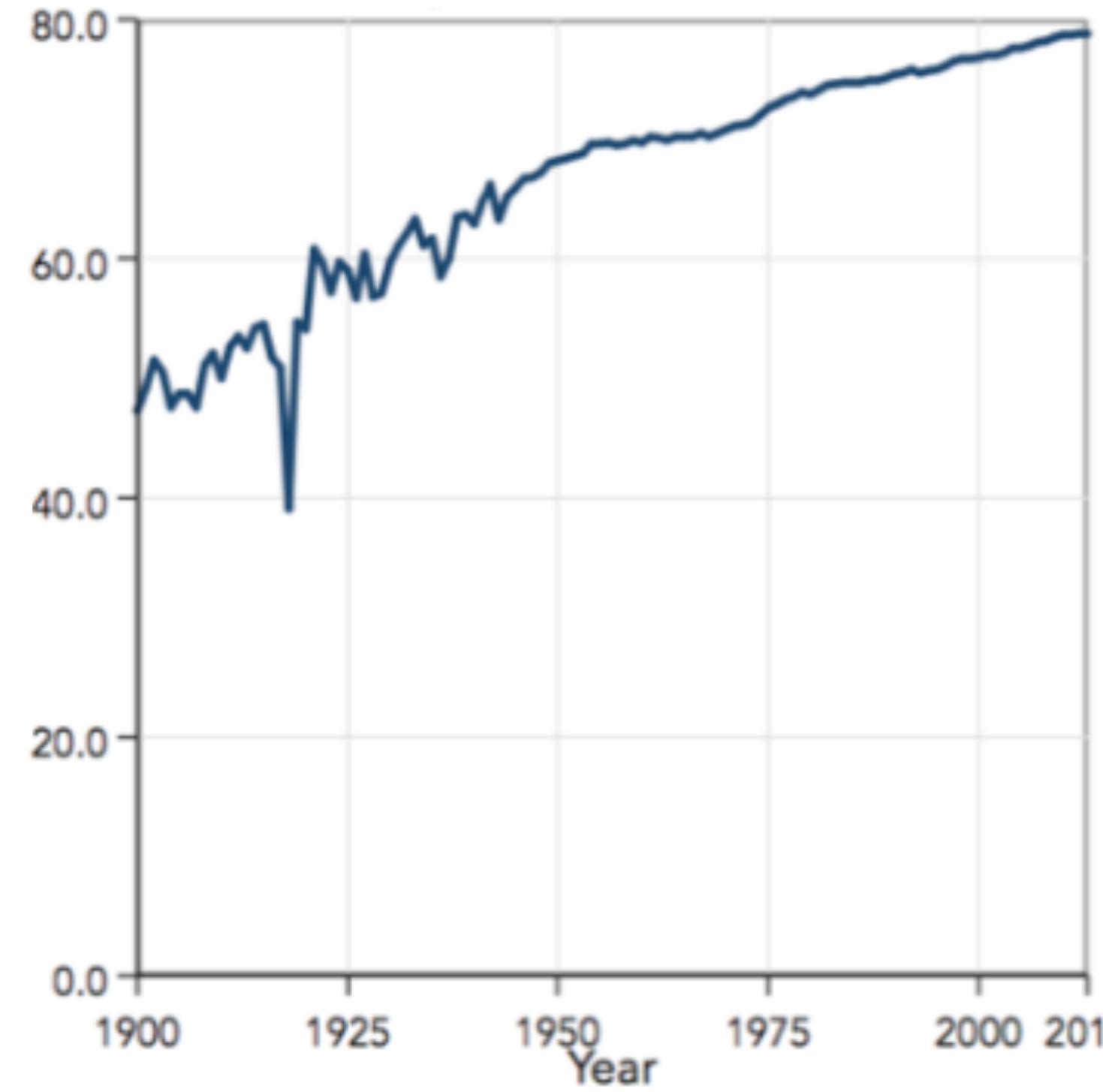
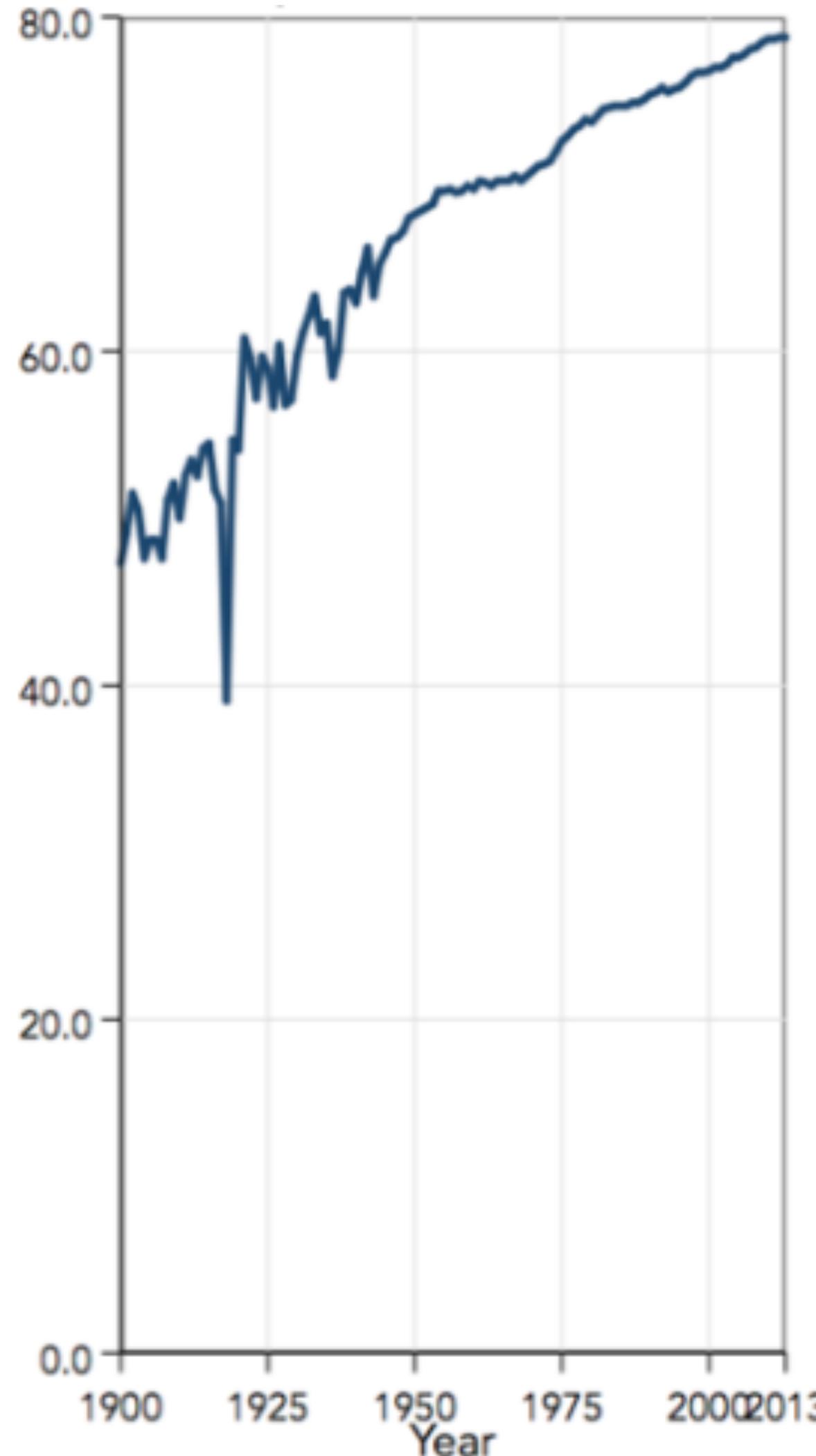
PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Scale Distortion



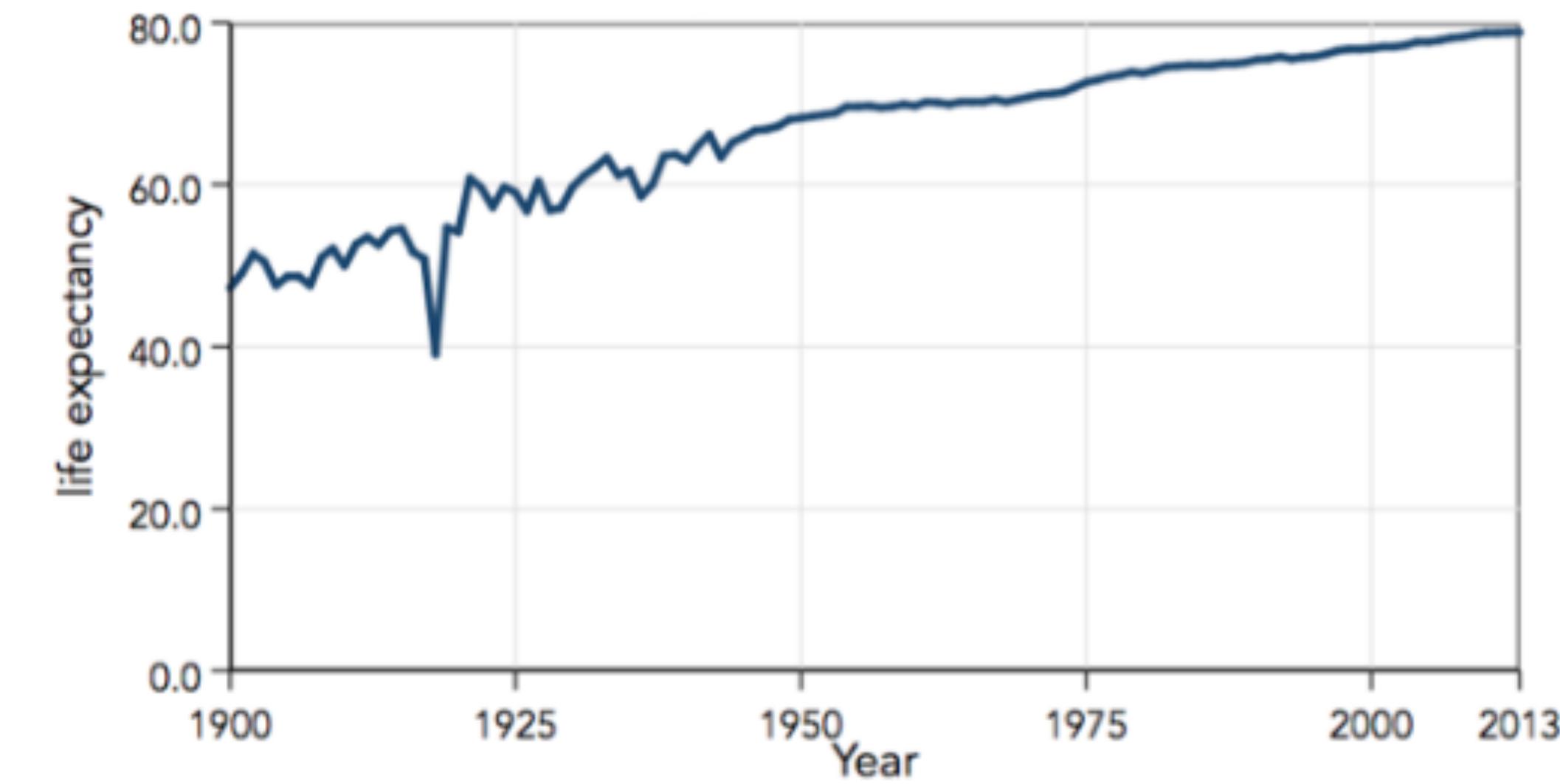
PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Scale Distortion



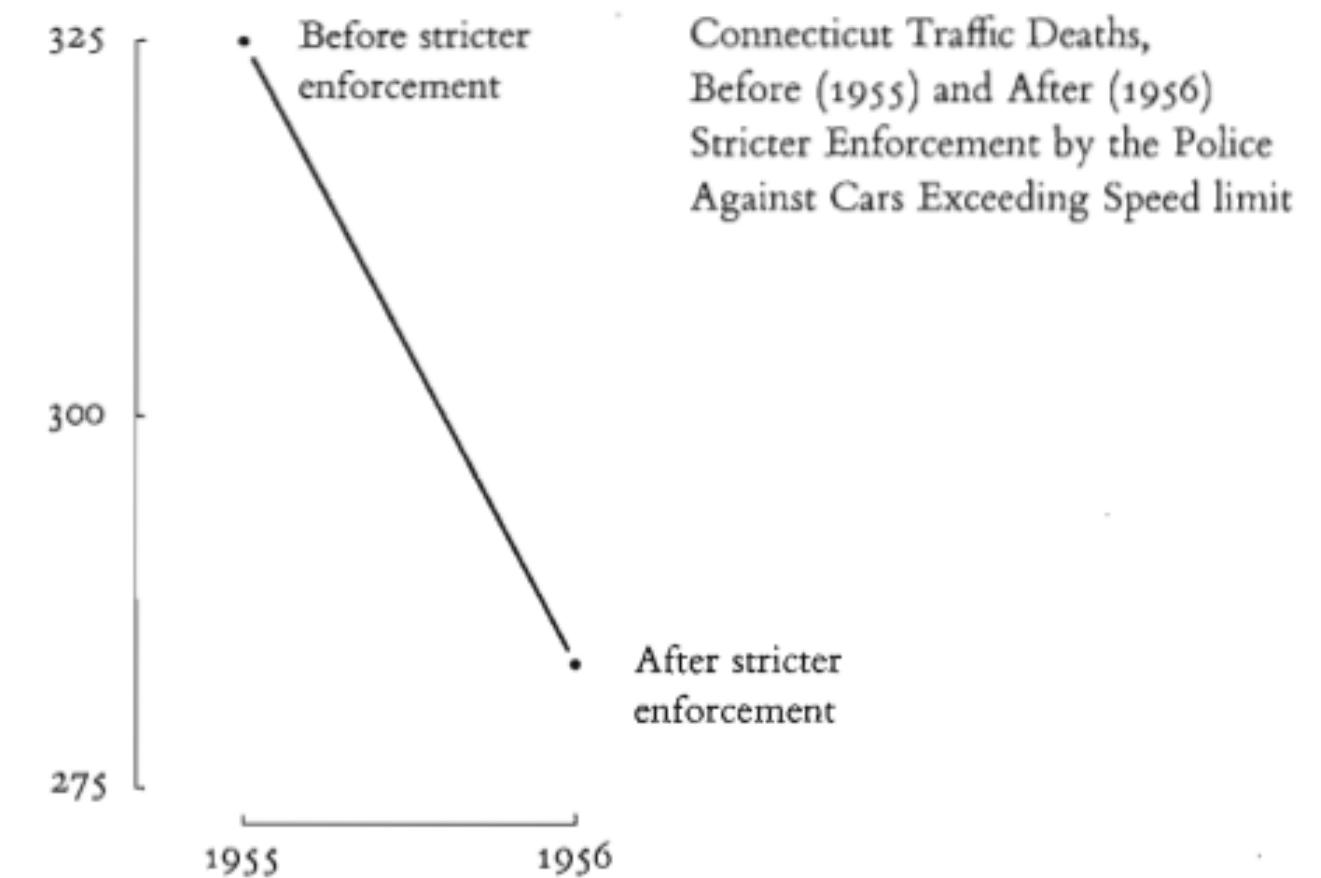
Aspect ratio when visualizing variables with different scales?

Depends. One possible rule:
Golden ratio ($W=1.6 \times H$)



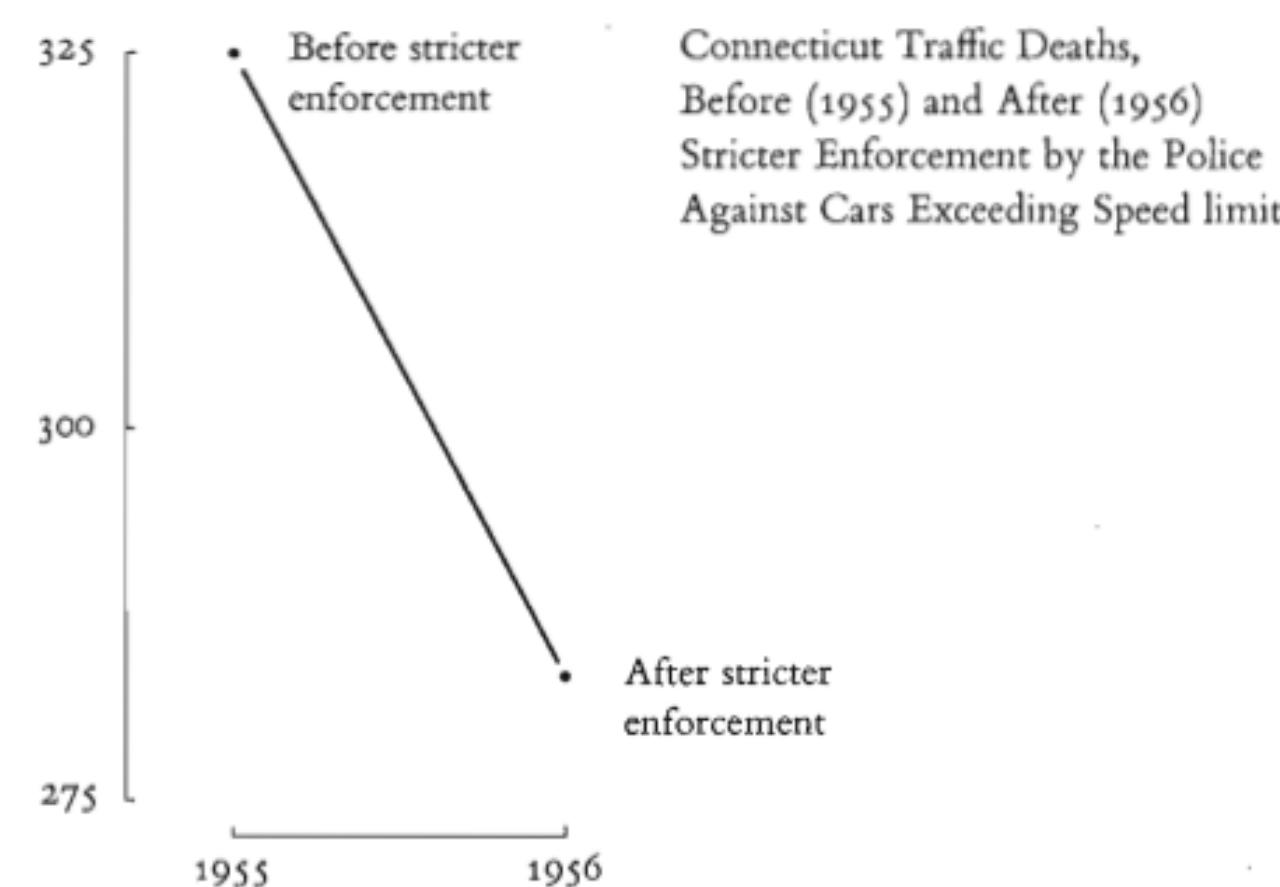
PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Omitting Context



PRINCIPLES OF DATA VISUALIZATION

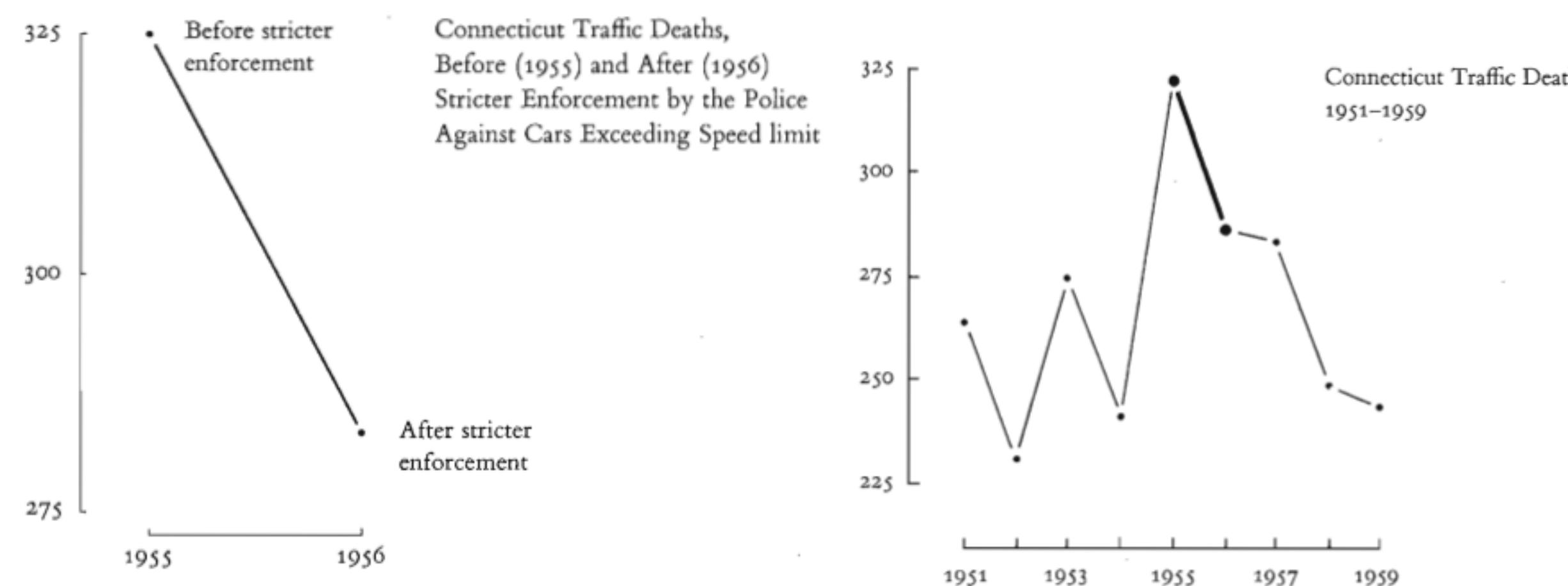
Graphical Integrity: Omitting Context



A large decline in traffic deaths in Connecticut: Due to stricter enforcement?

PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Omitting Context

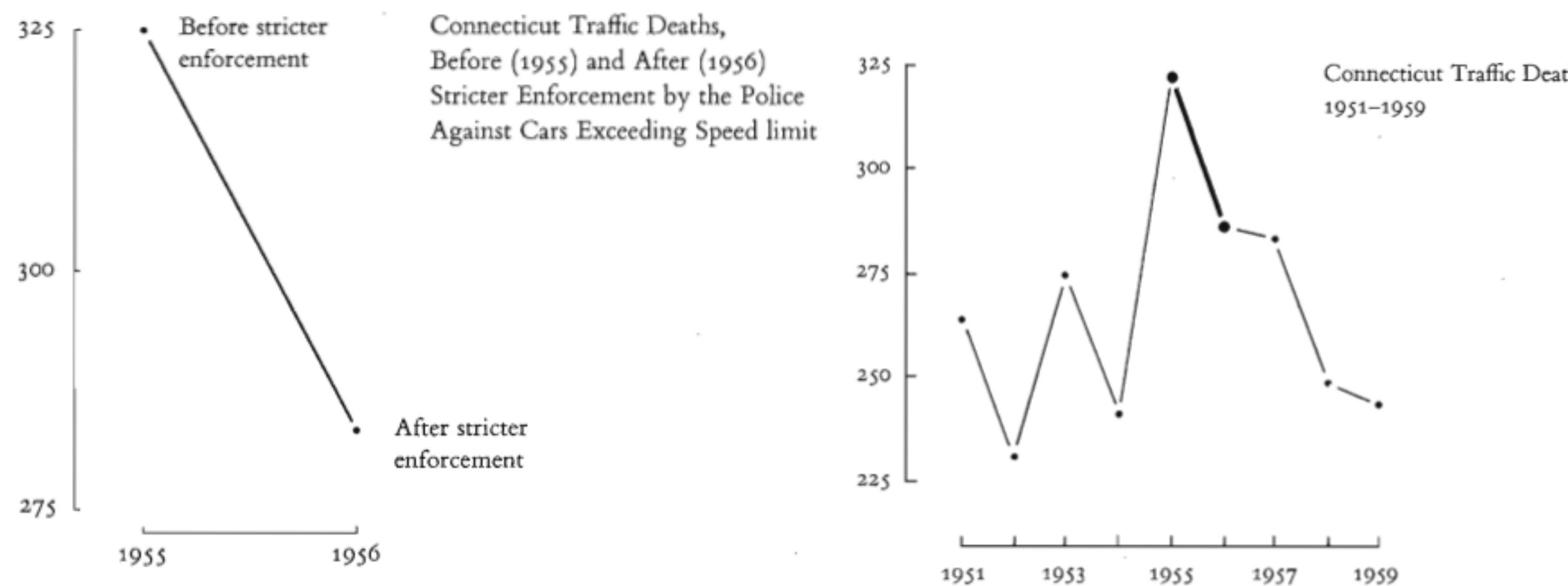


A large decline in traffic deaths in Connecticut: Due to stricter enforcement?

Adding context changes the story

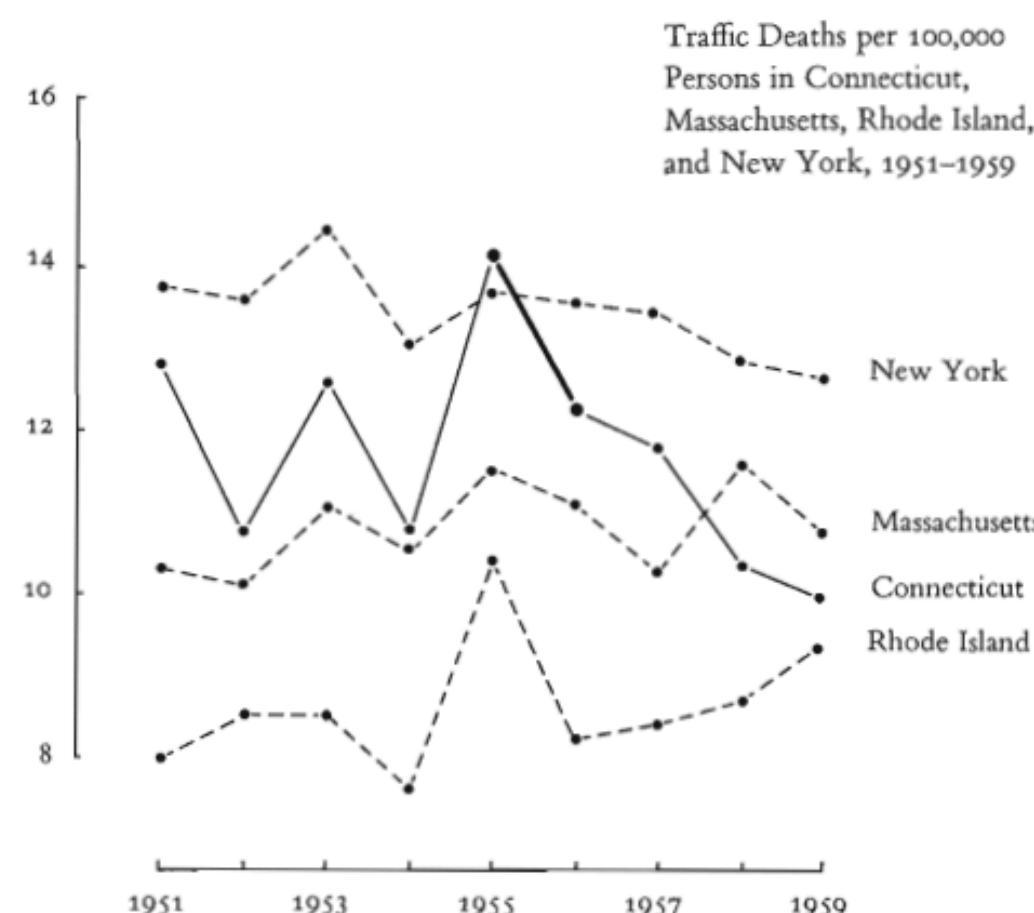
PRINCIPLES OF DATA VISUALIZATION

Graphical Integrity: Omitting Context



A large decline in traffic deaths in Connecticut: Due to stricter enforcement?

Adding context changes the story
Adding further context: not only Connecticut had a decline.



Graphics must not quote data out of context: There should always be a 'Compared to what?'

PRINCIPLES OF DATA VISUALIZATION

Occam's Razor: If visualization with less ink possible, do so.

Data-ink ratio = data-ink / total ink used for the graphic

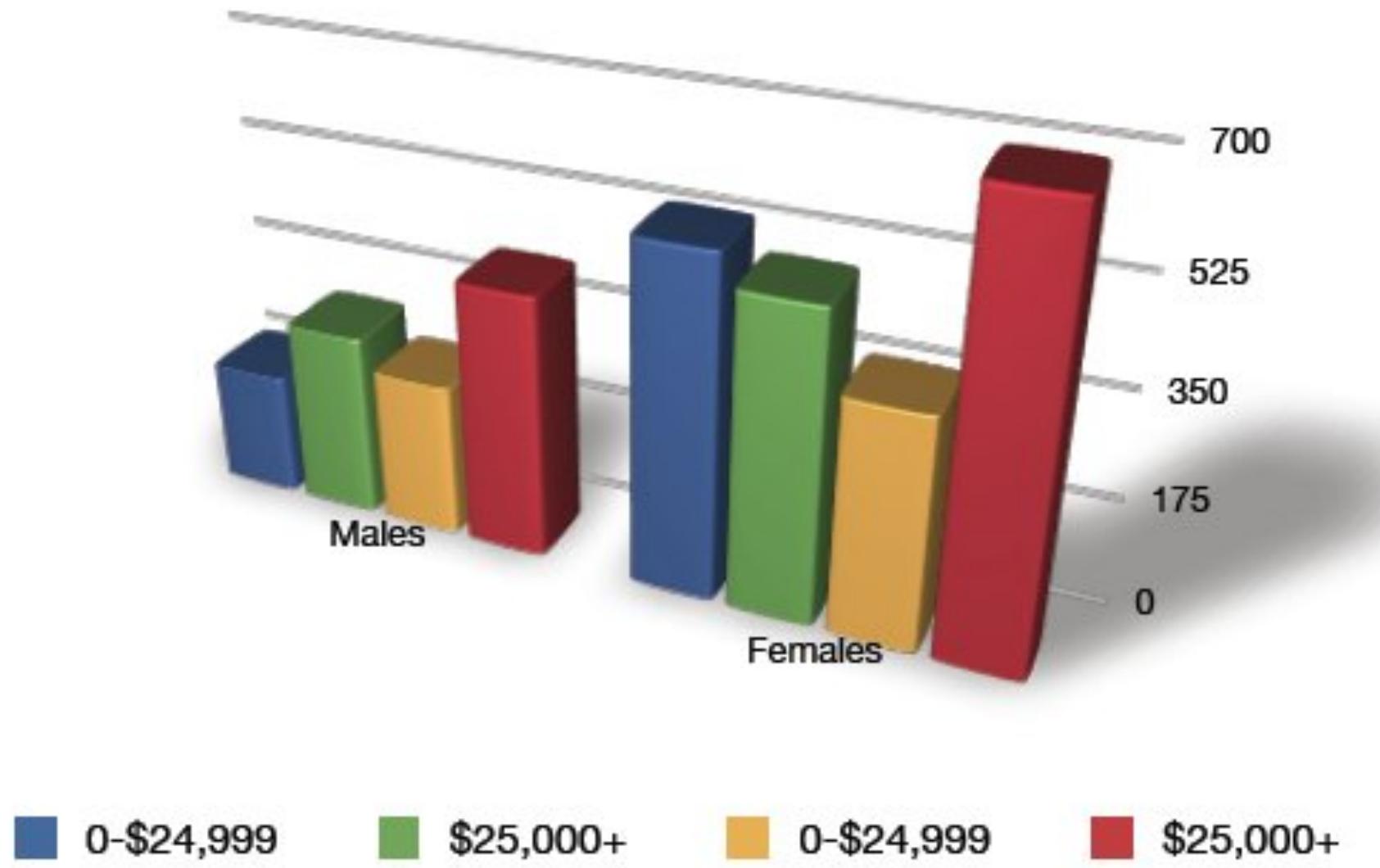
Goal: Maximize the ratio.

PRINCIPLES OF DATA VISUALIZATION

Occam's Razor: If visualization with less ink possible, do so.

Data-ink ratio = data-ink / total ink used for the graphic

Goal: Maximize the ratio.



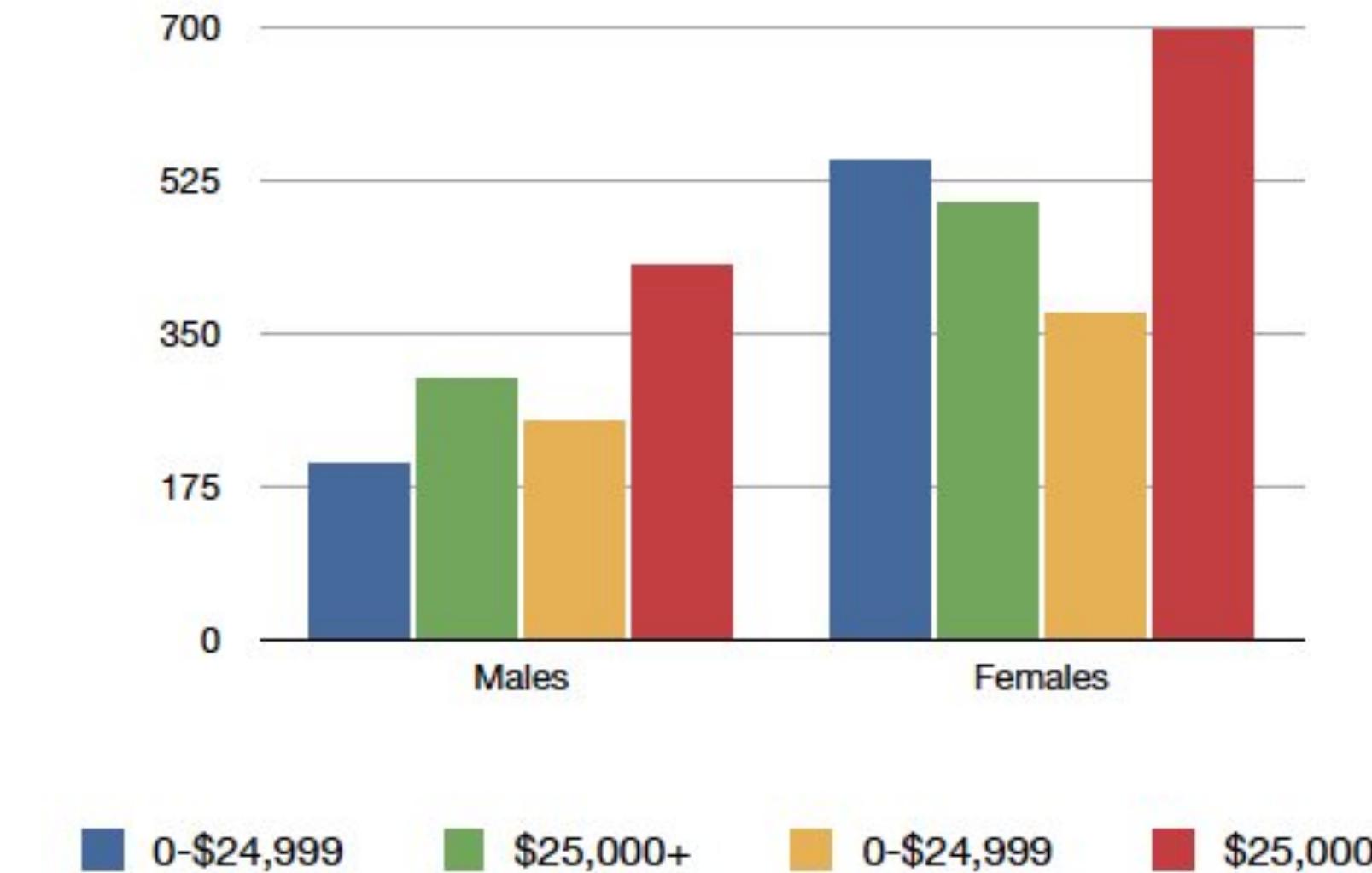
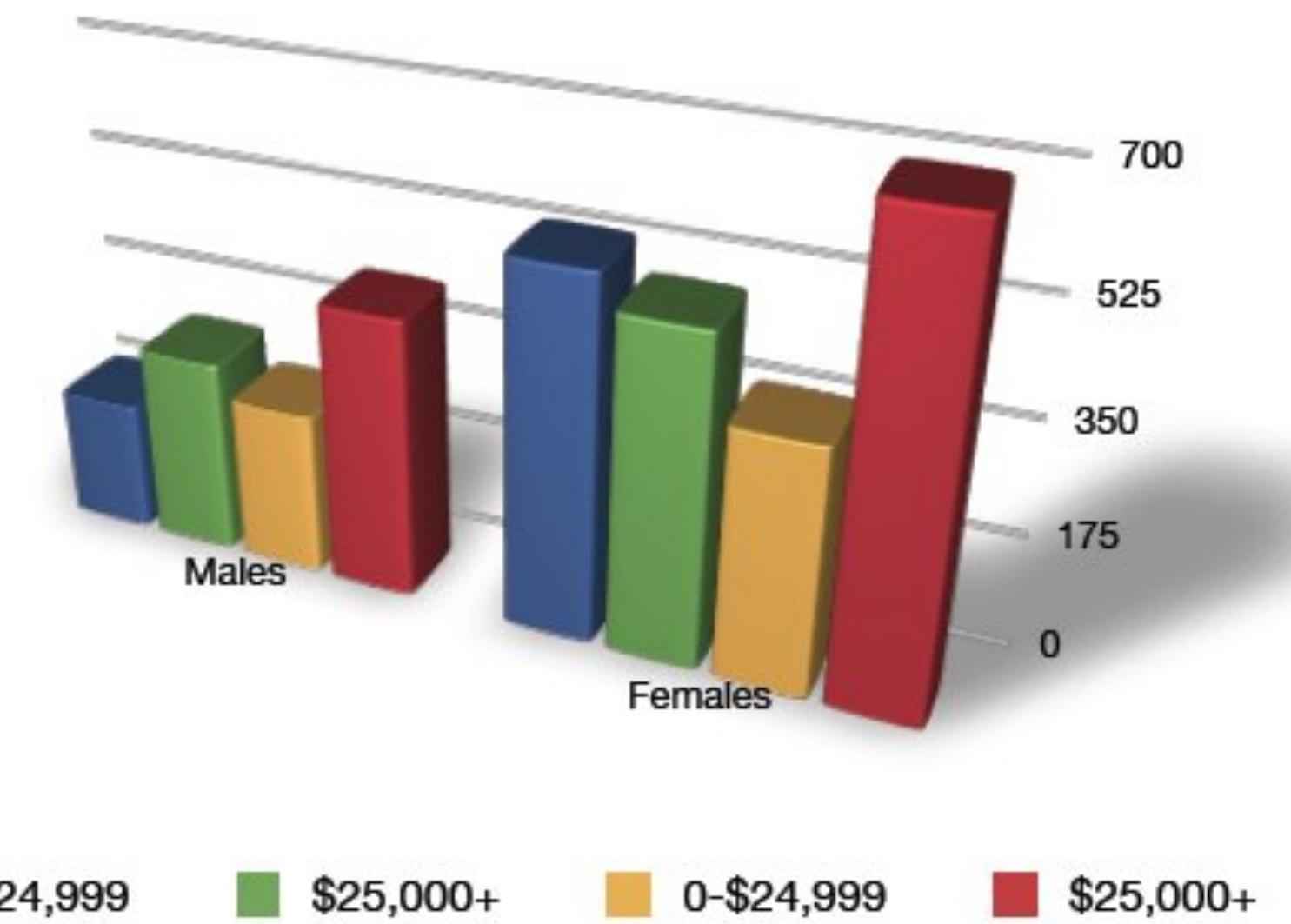
Looks cool in 3D,
but seeing differences easy?

PRINCIPLES OF DATA VISUALIZATION

Occam's Razor: If visualization with less ink possible, do so.

Data-ink ratio = data-ink / total ink used for the graphic

Goal: Maximize the ratio.



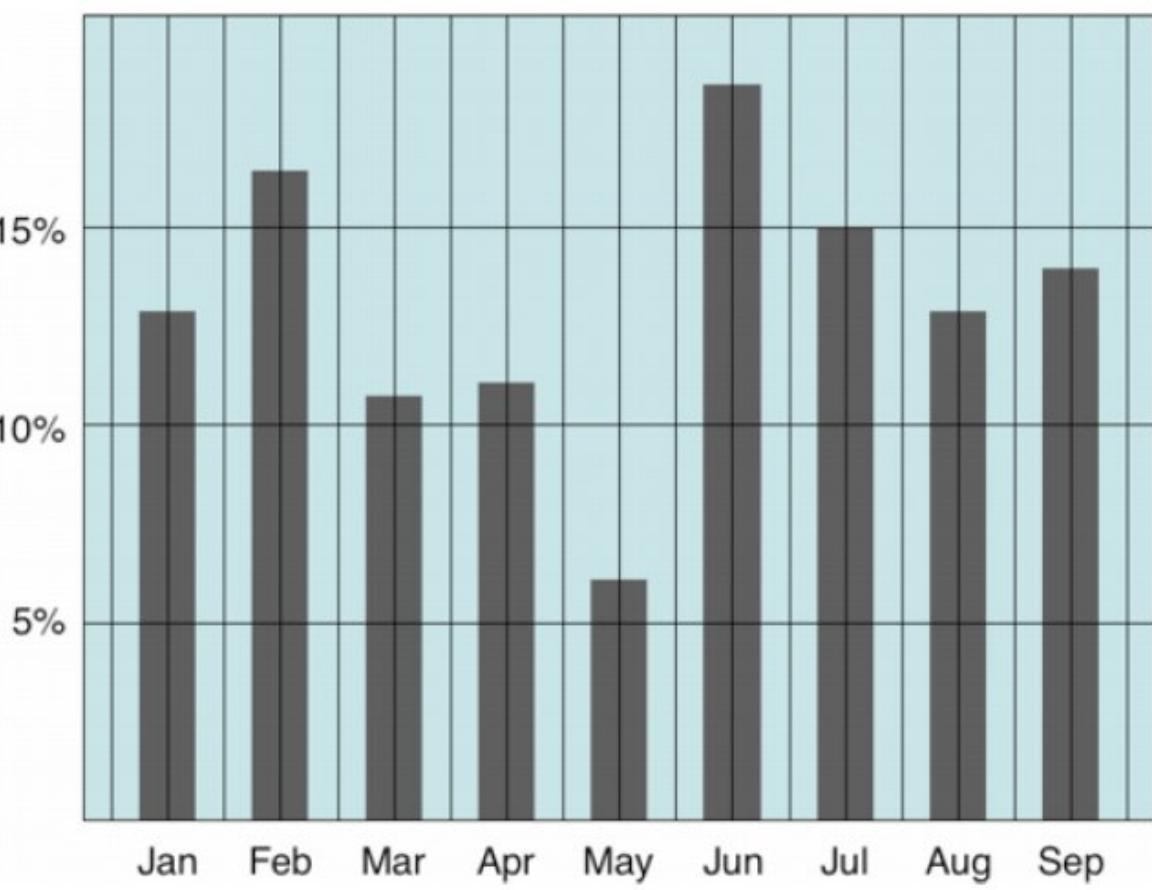
Looks cool in 3D,
but seeing differences easy?

Much better when most ink is
used to present the data.

PRINCIPLES OF DATA VISUALIZATION

Occam's Razor: If visualization with less ink possible, do so.

Further examples: Erase non-data ink.

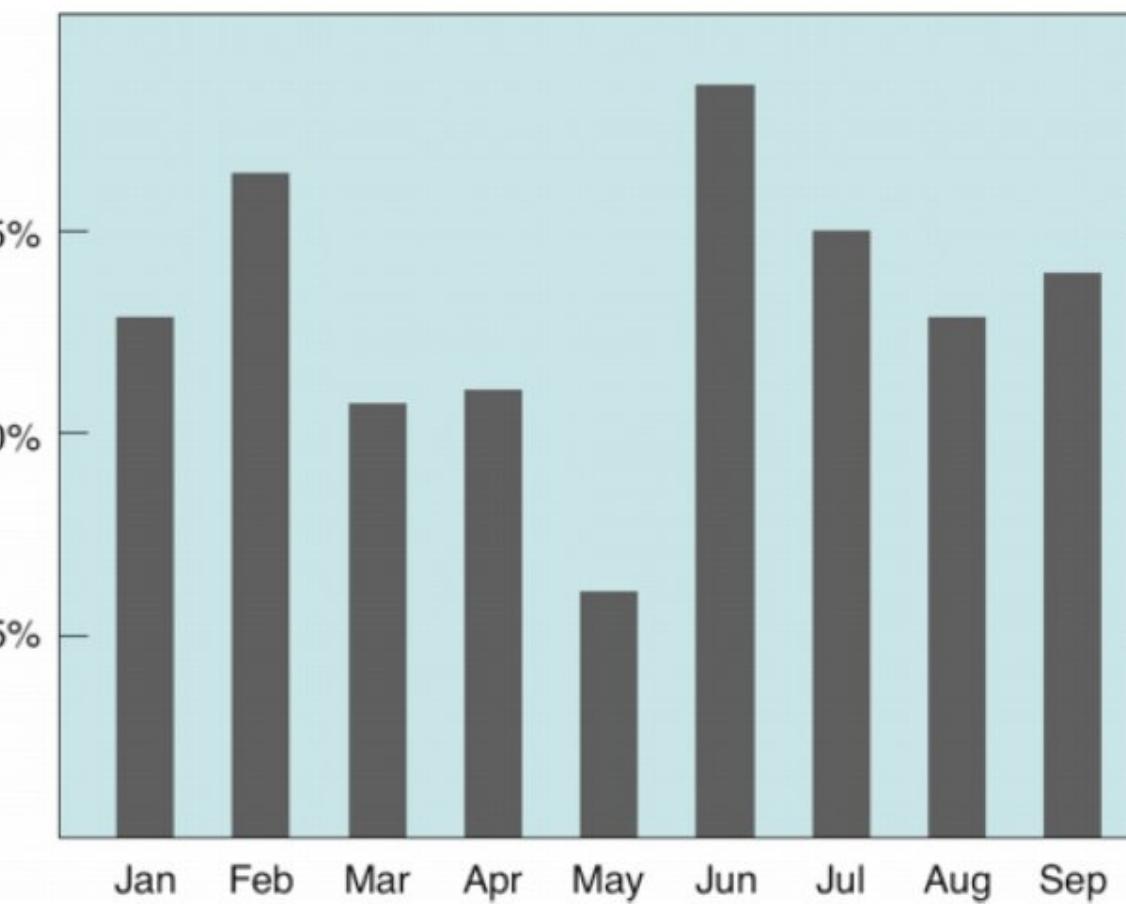
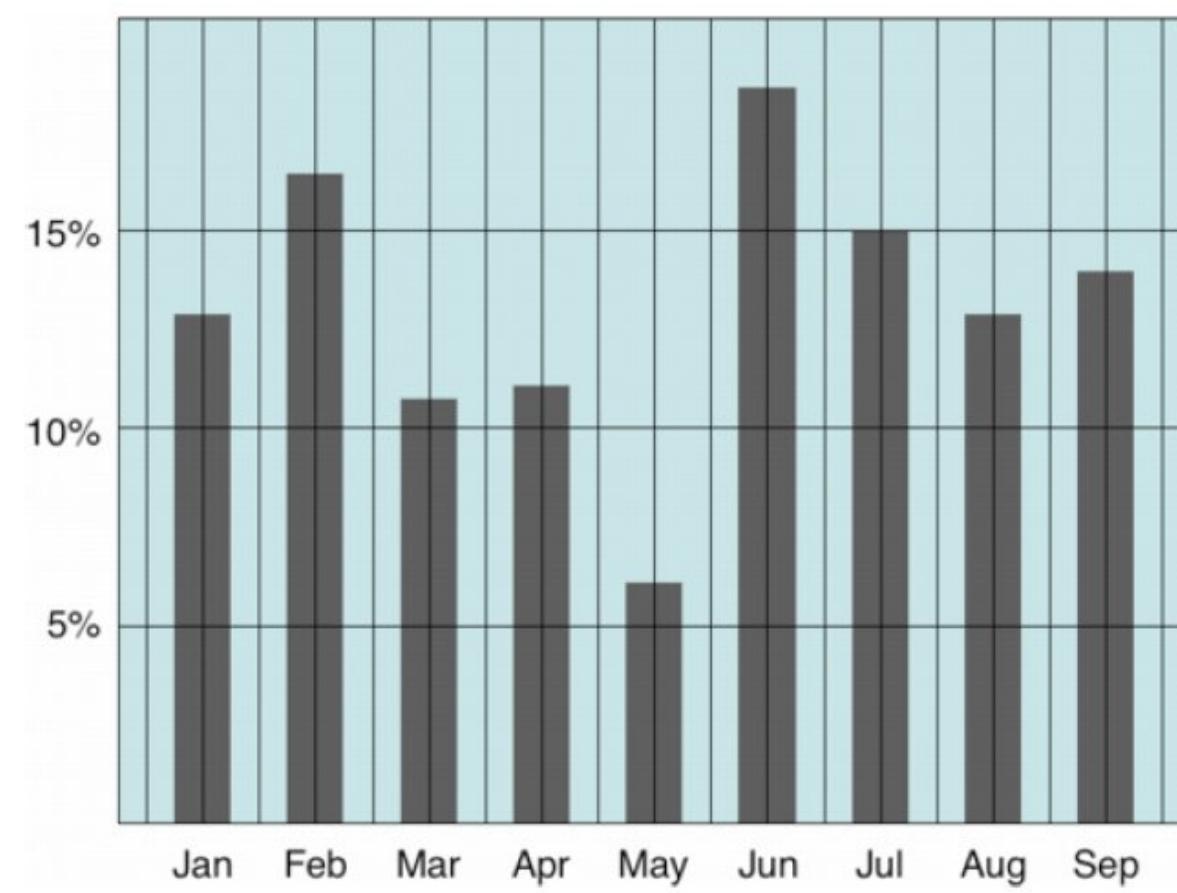


Do we need grid lines?

PRINCIPLES OF DATA VISUALIZATION

Occam's Razor: If visualization with less ink possible, do so.

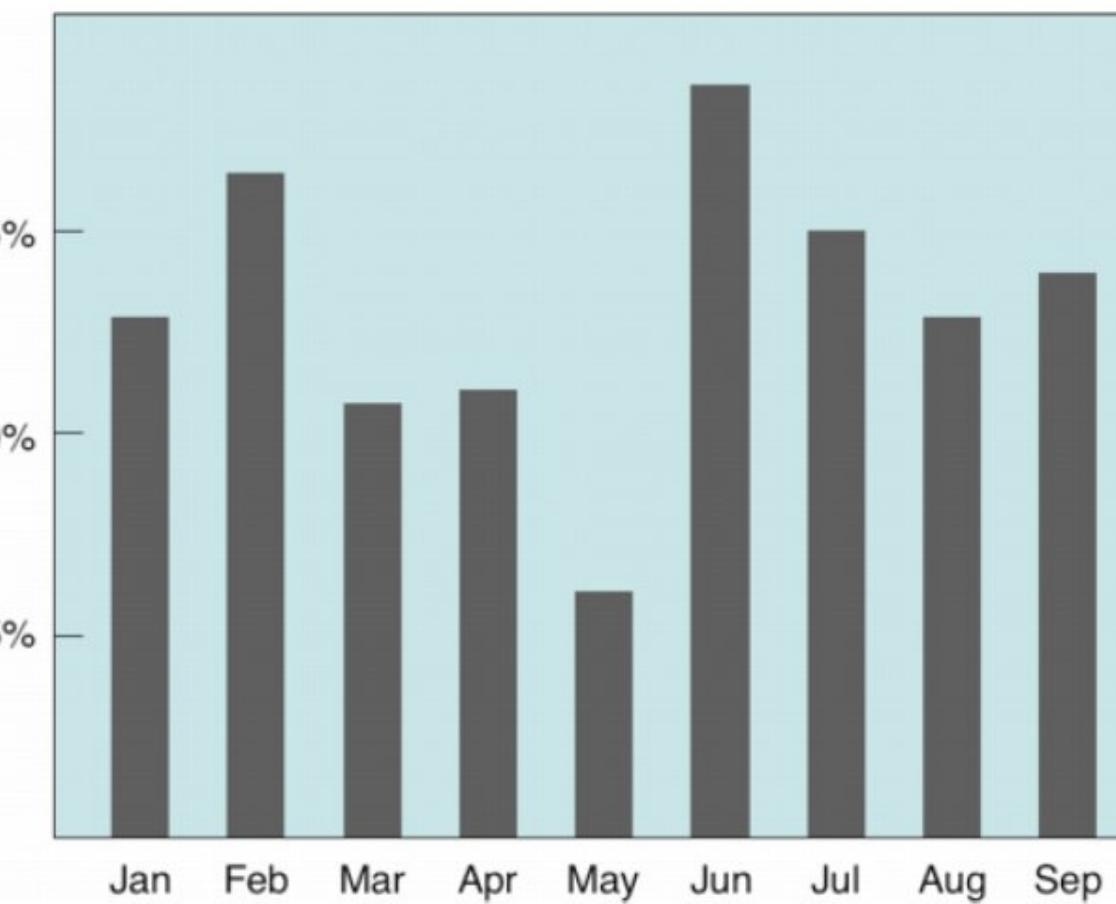
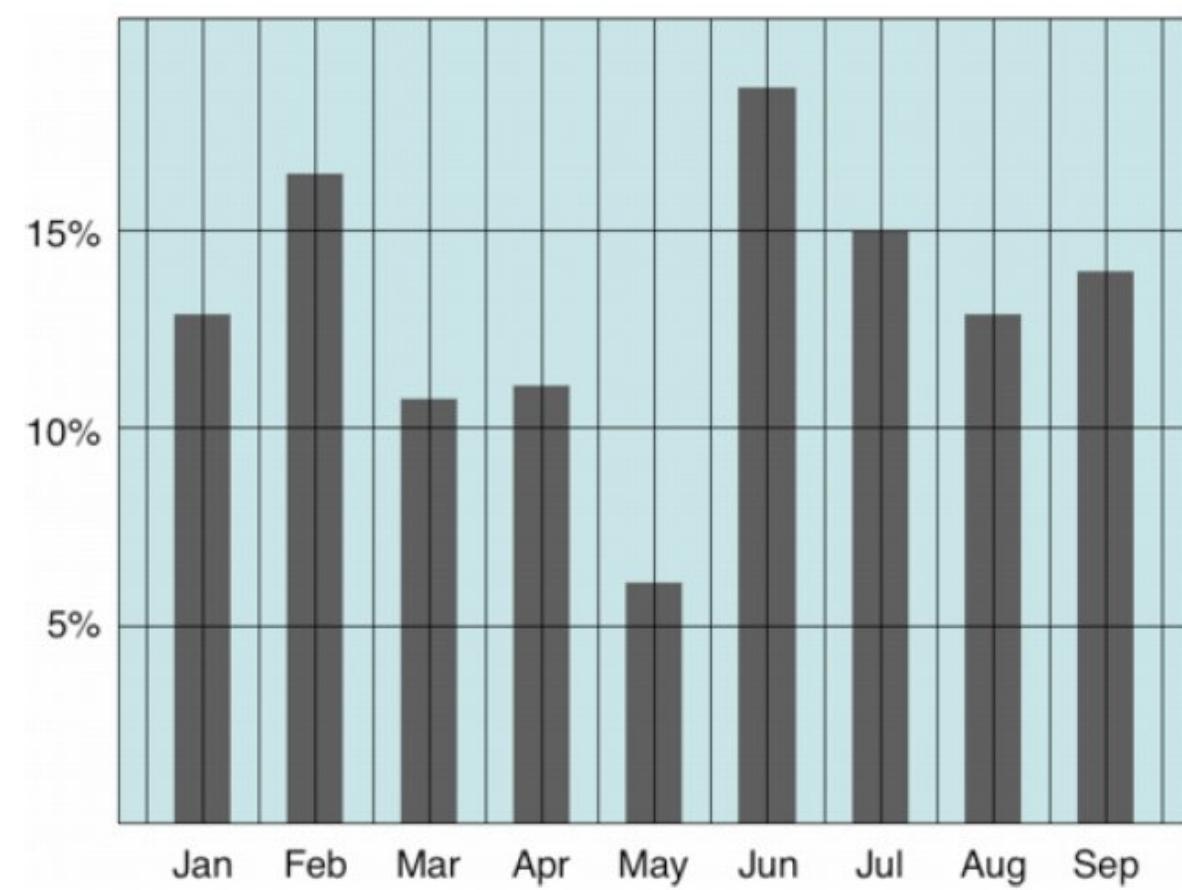
Further examples: Erase non-data ink.



PRINCIPLES OF DATA VISUALIZATION

Occam's Razor: If visualization with less ink possible, do so.

Further examples: Erase non-data ink.

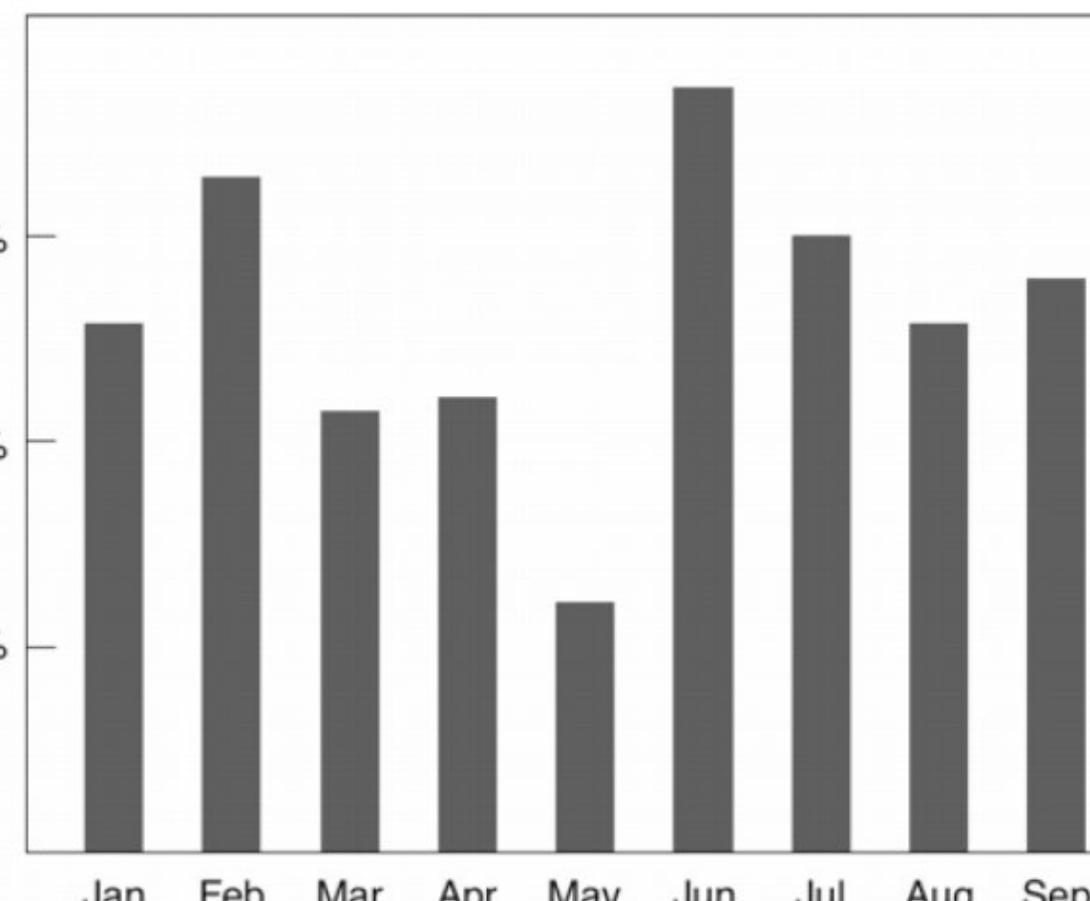
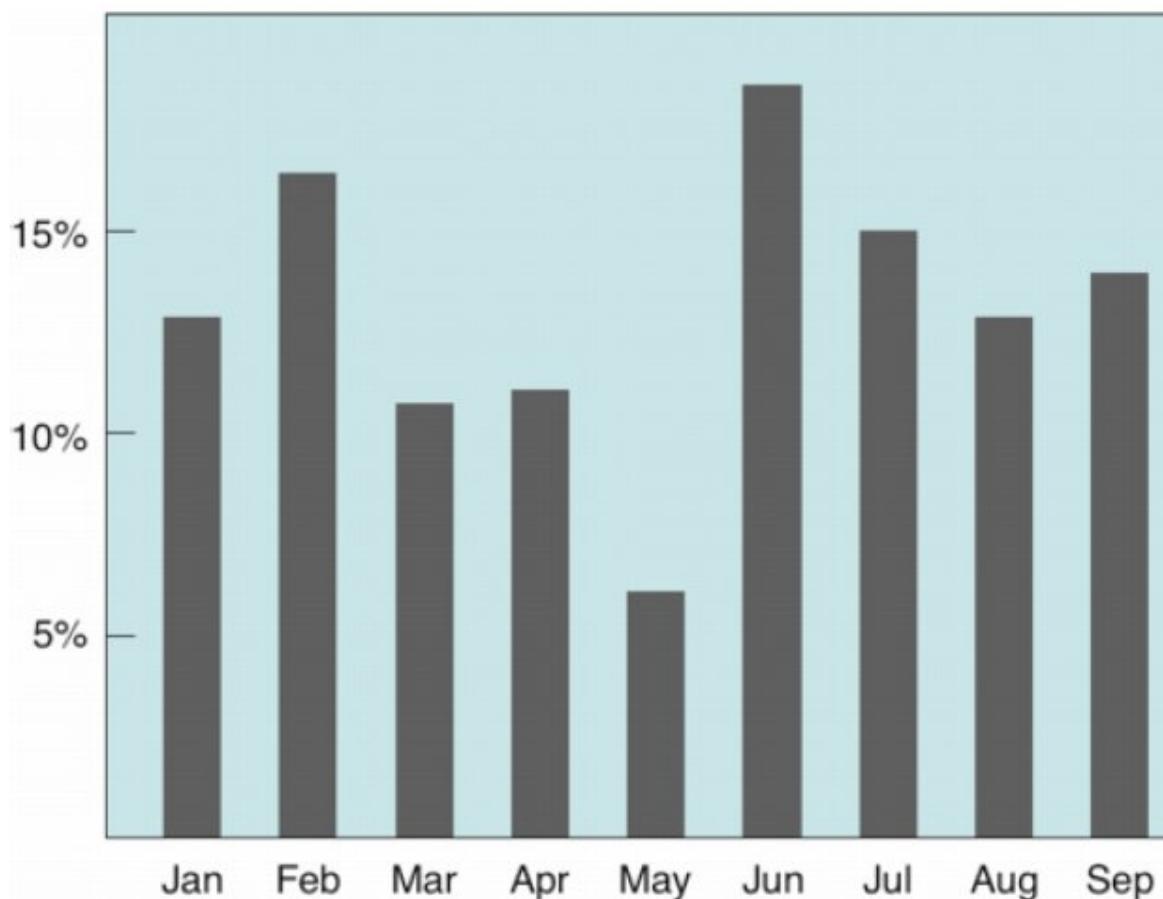
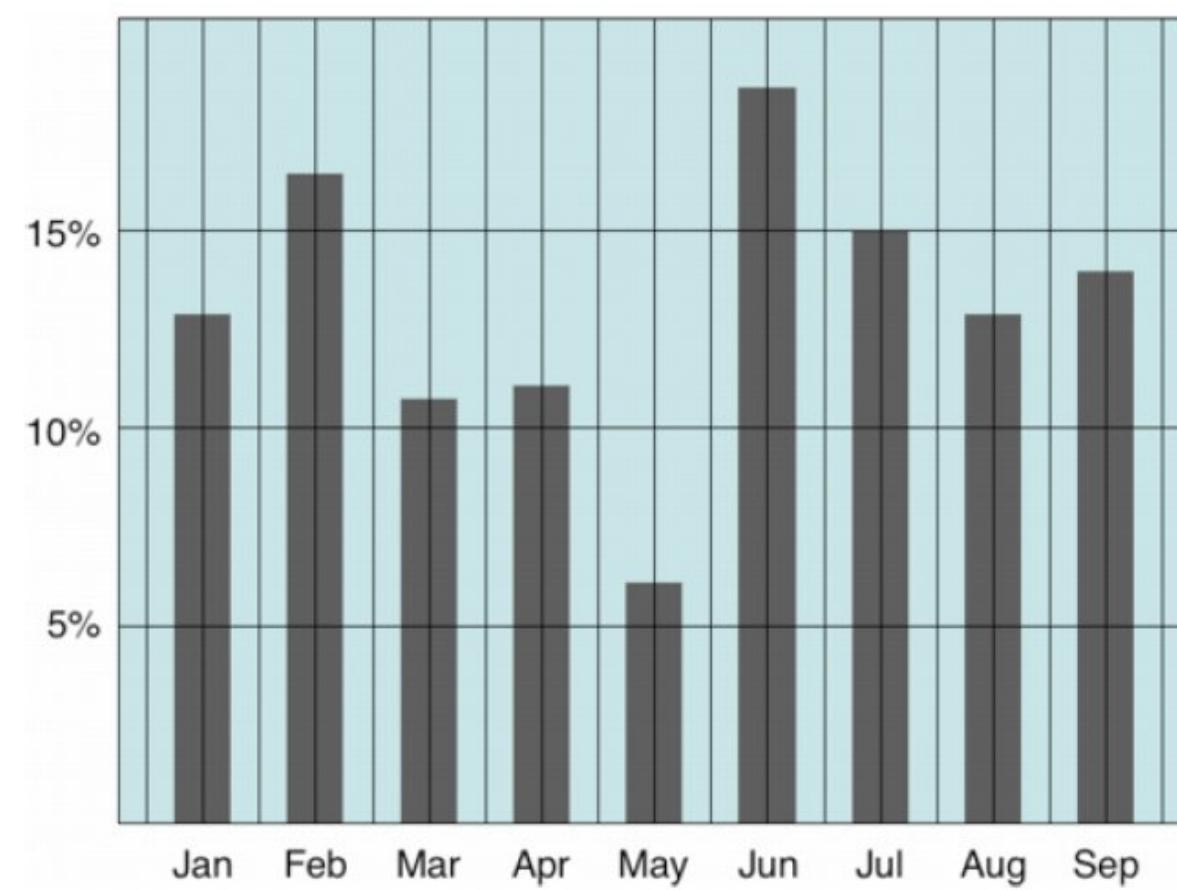


Do we need color?

PRINCIPLES OF DATA VISUALIZATION

Occam's Razor: If visualization with less ink possible, do so.

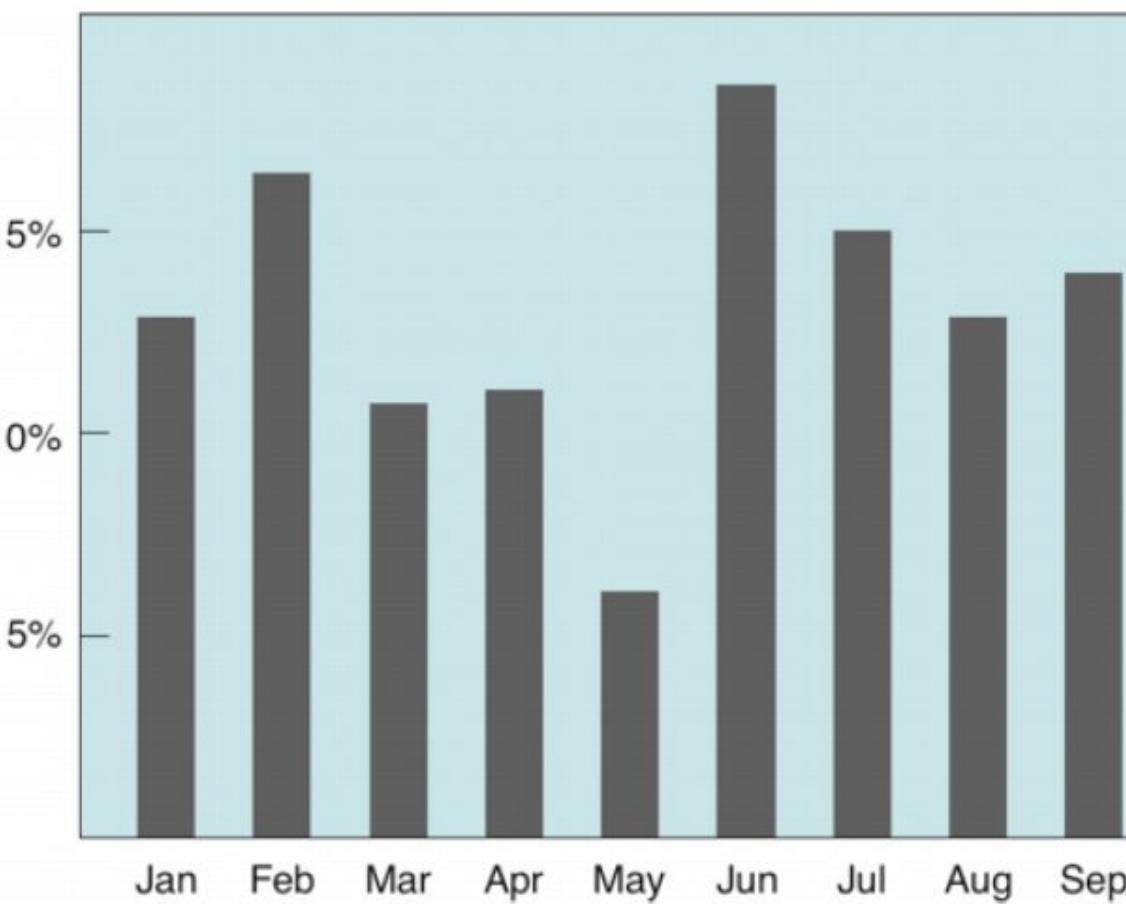
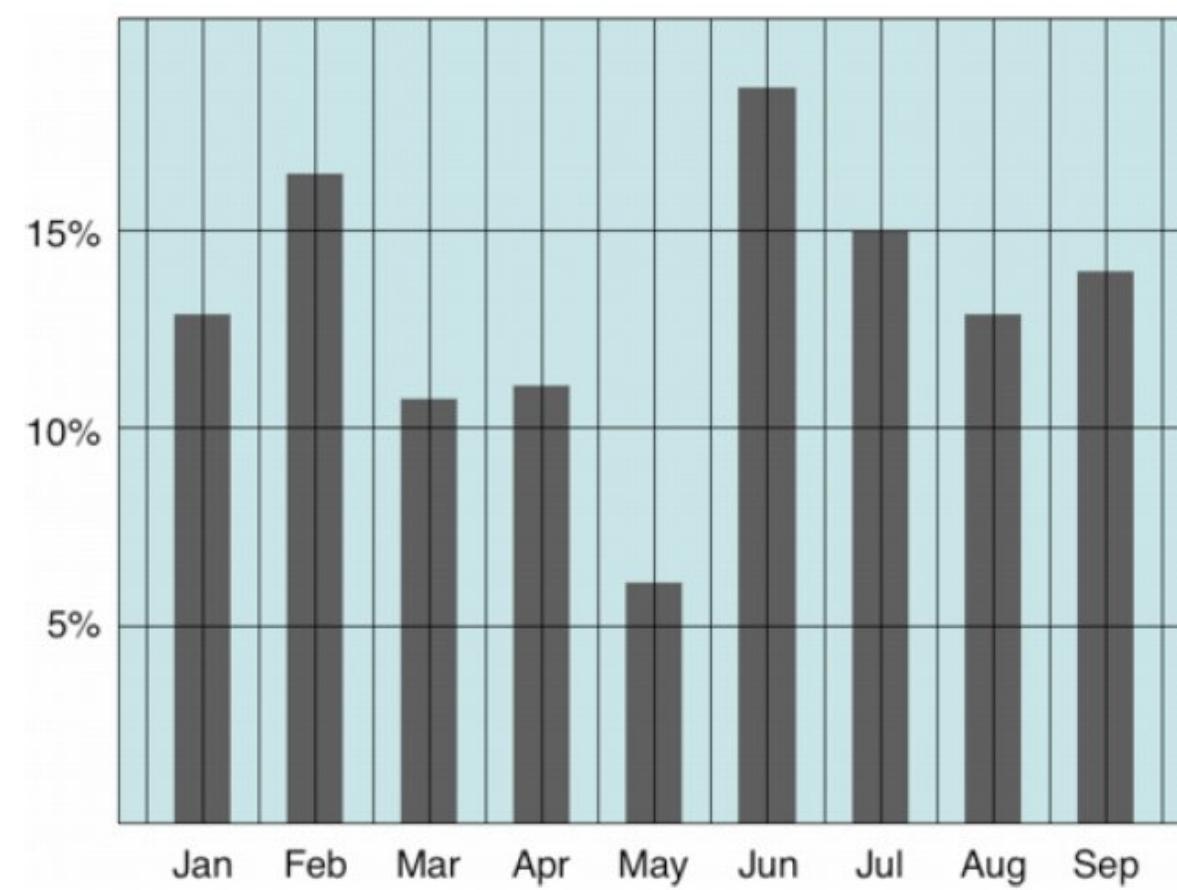
Further examples: Erase non-data ink.



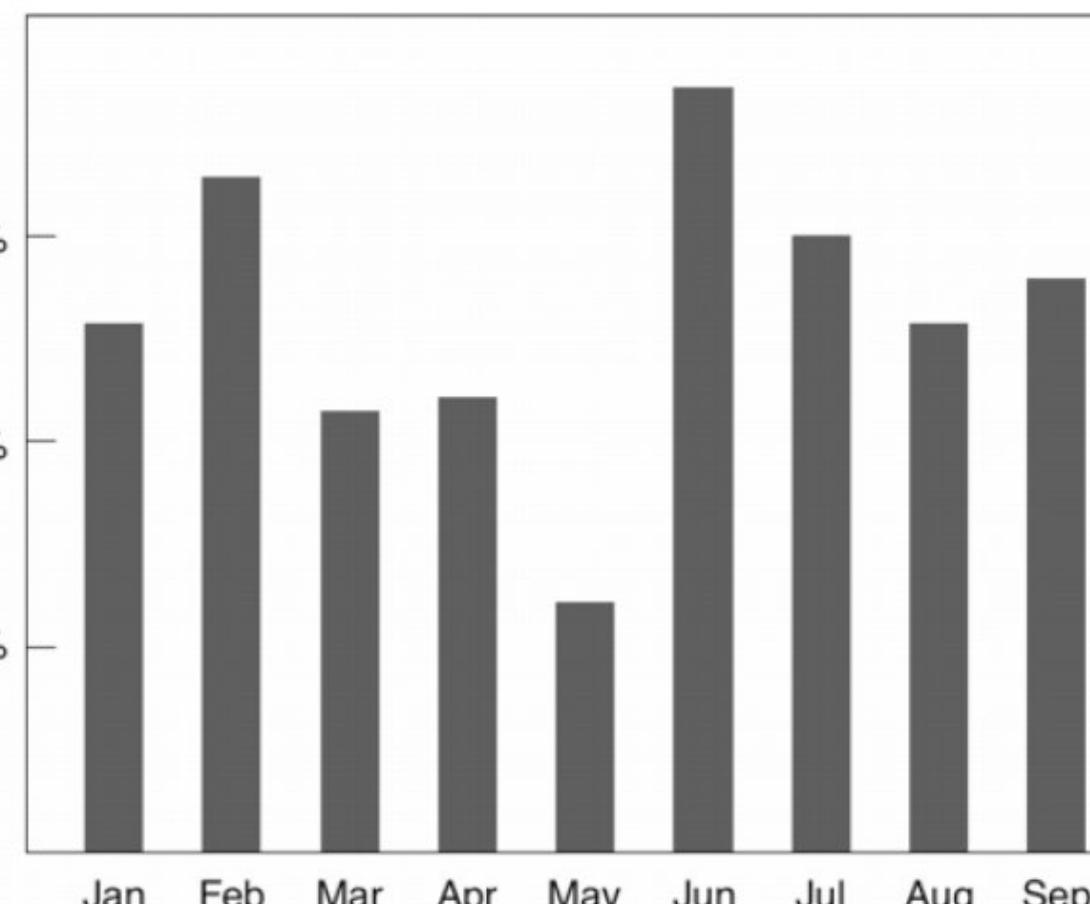
PRINCIPLES OF DATA VISUALIZATION

Occam's Razor: If visualization with less ink possible, do so.

Further examples: Erase non-data ink.



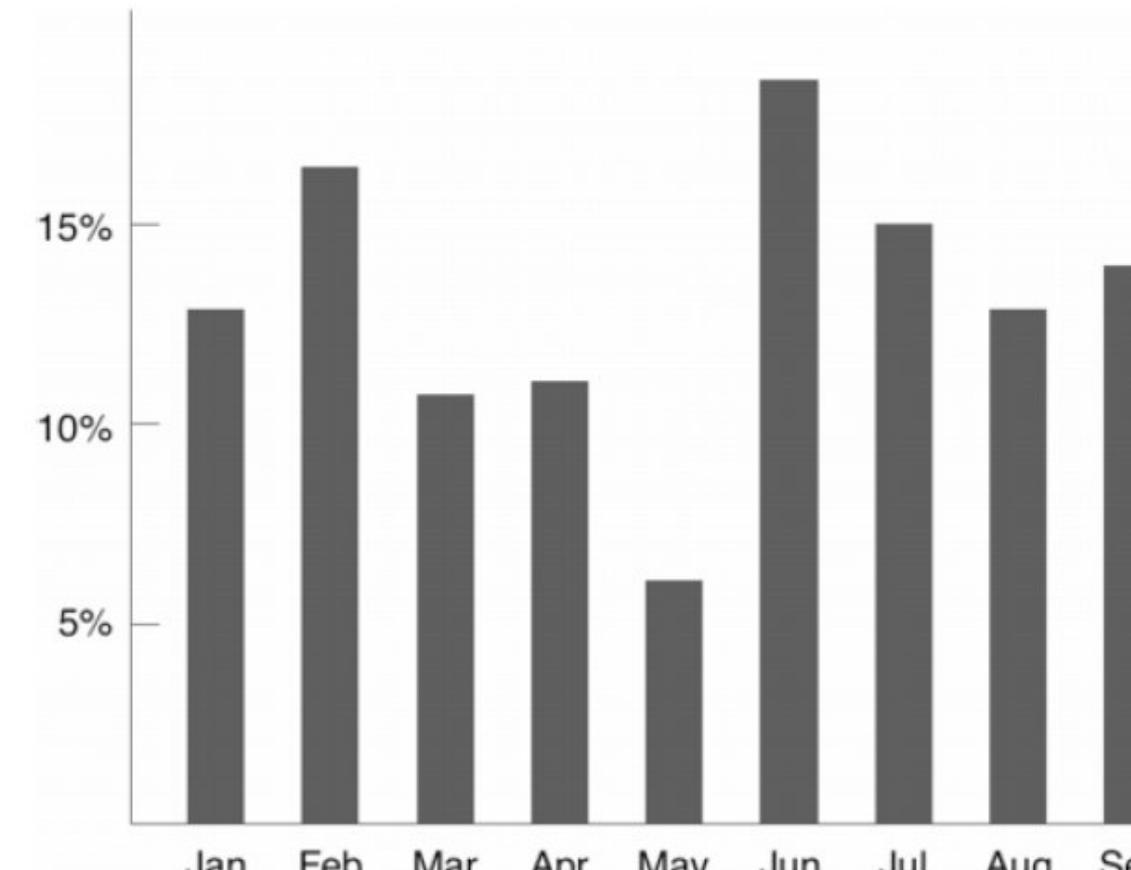
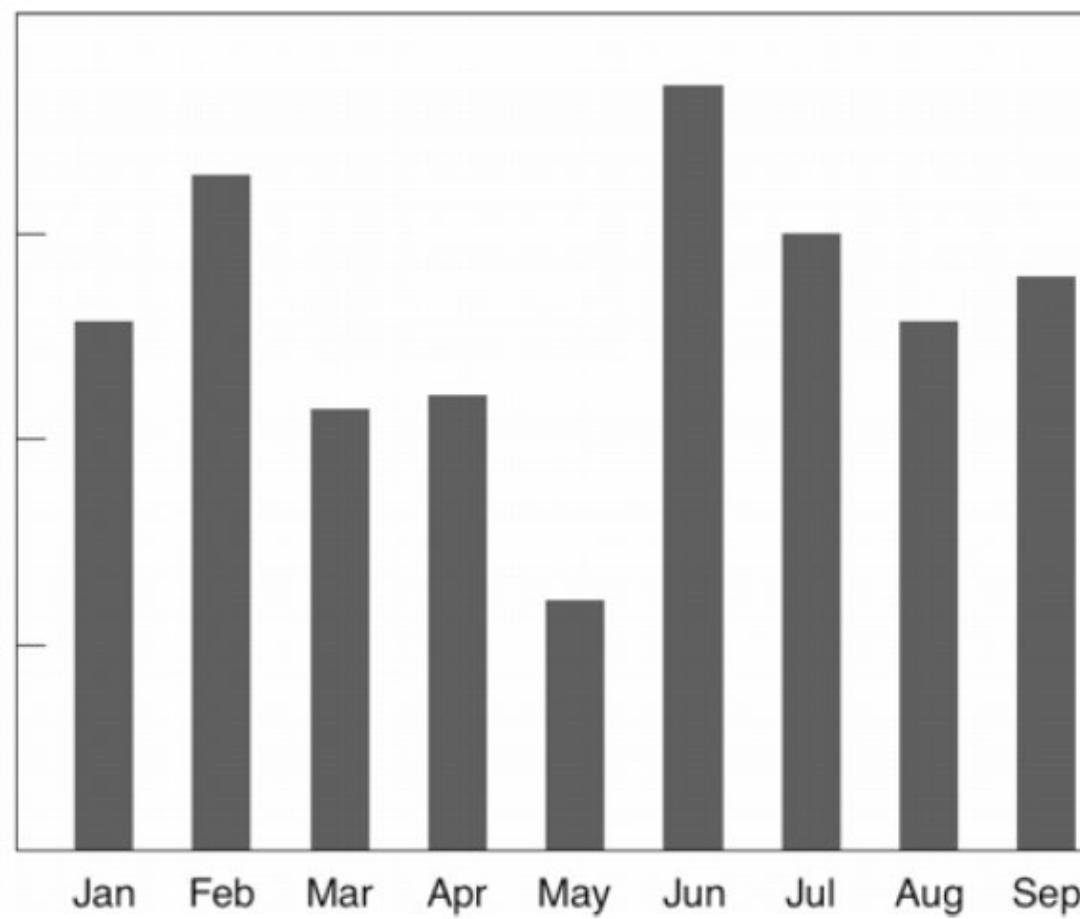
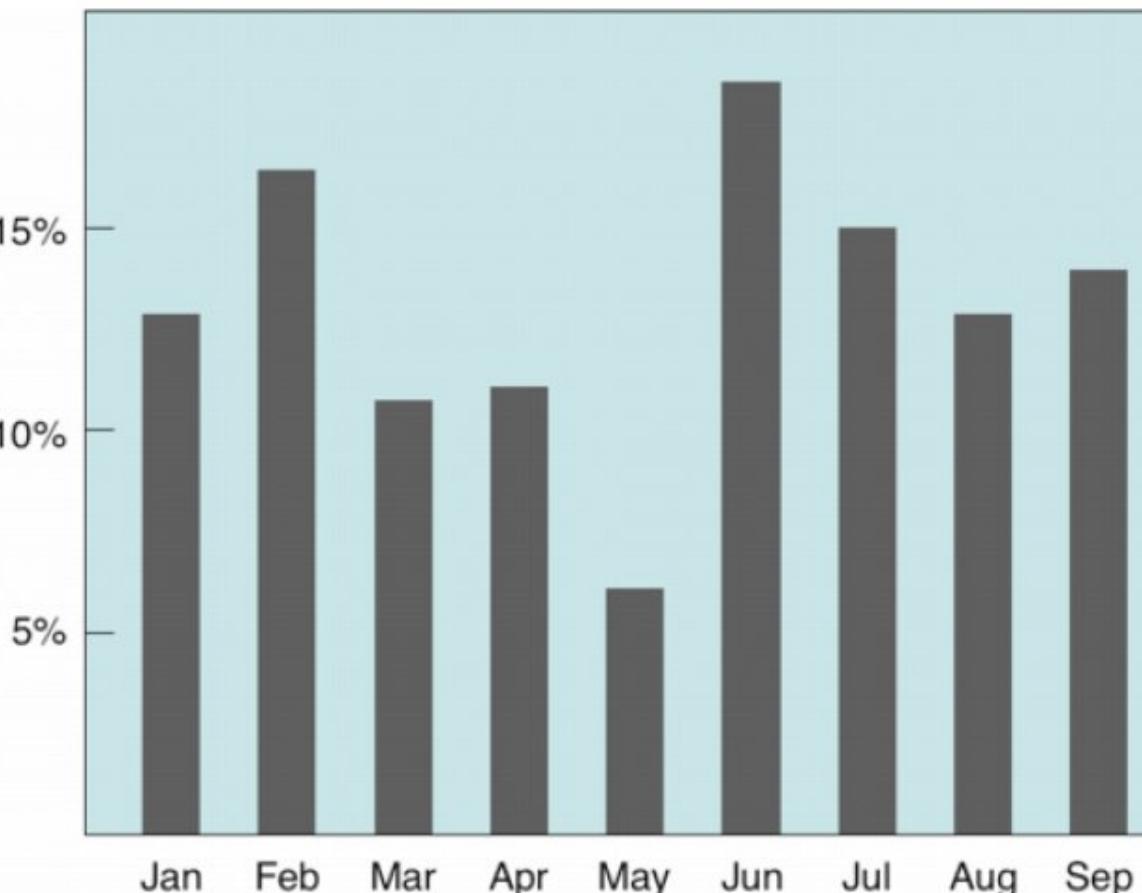
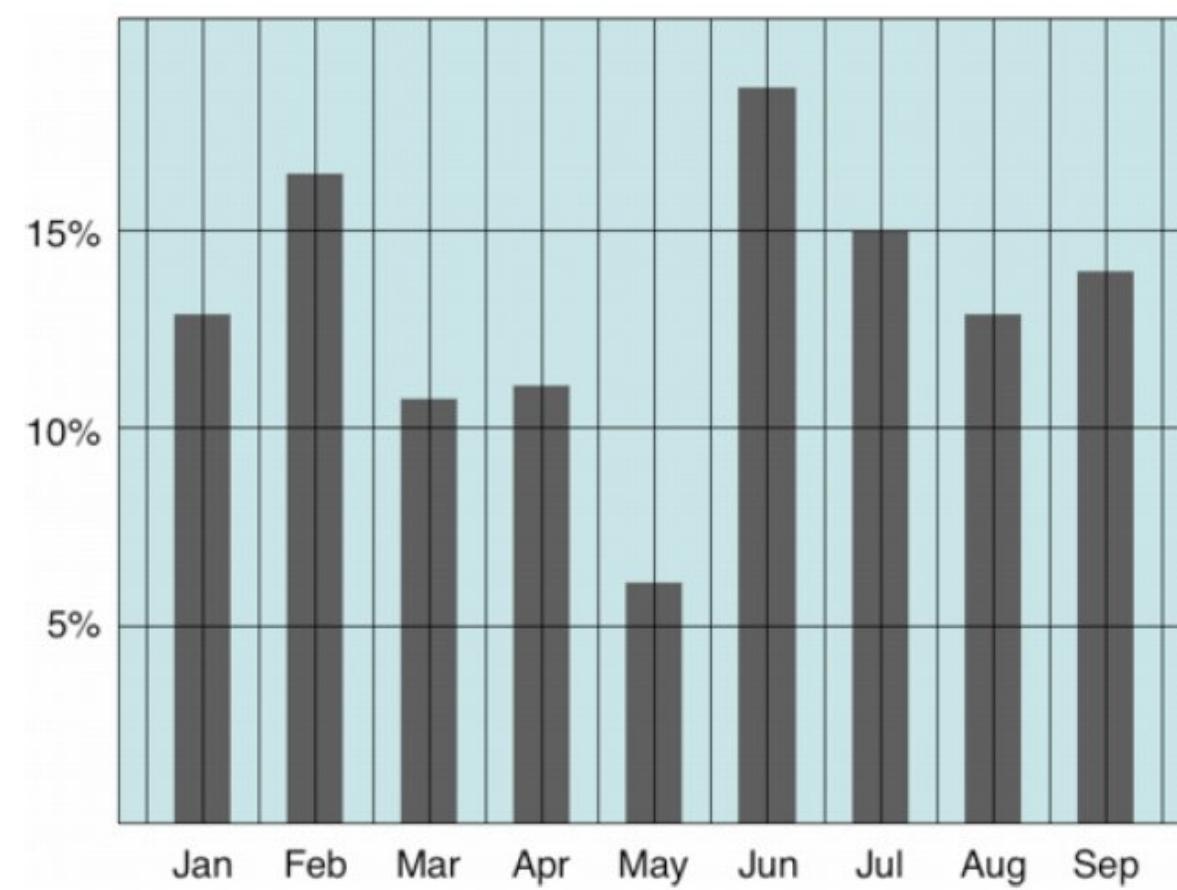
Do we need the full box?



PRINCIPLES OF DATA VISUALIZATION

Occam's Razor: If visualization with less ink possible, do so.

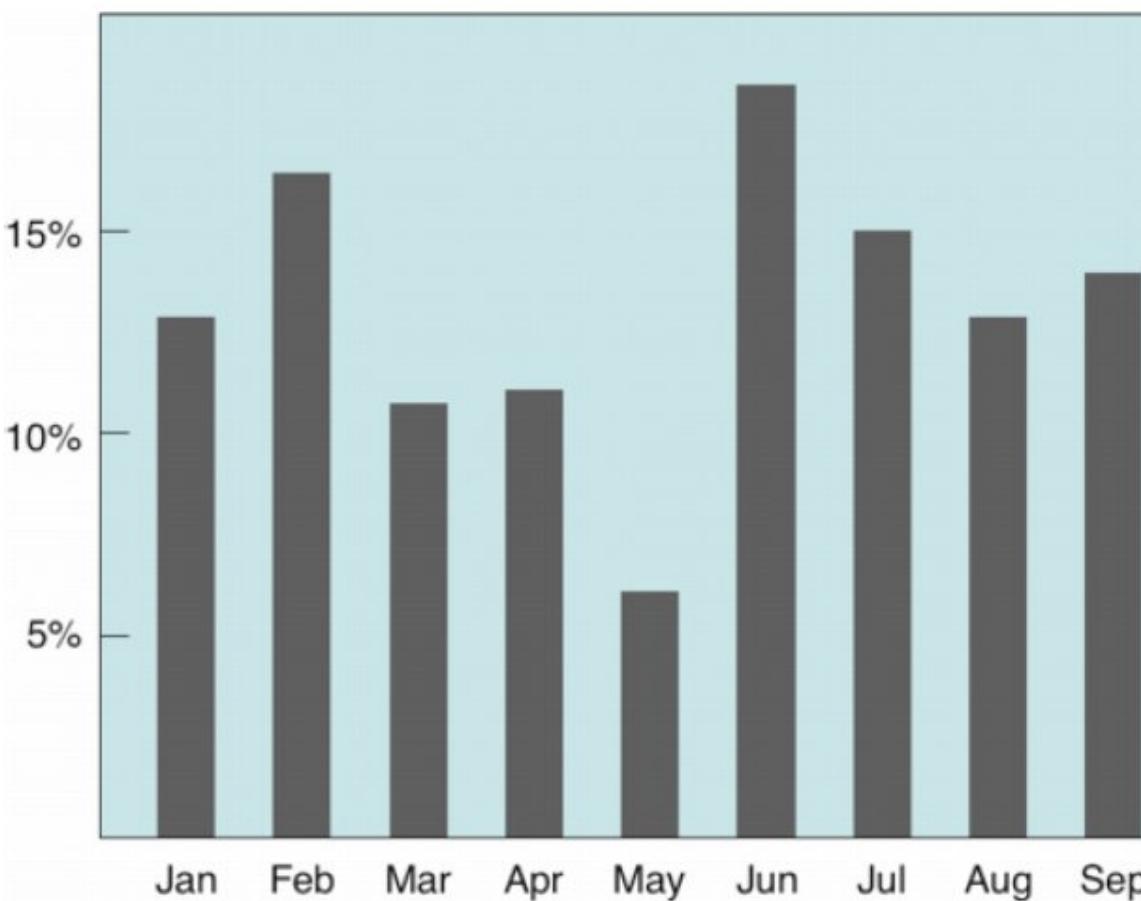
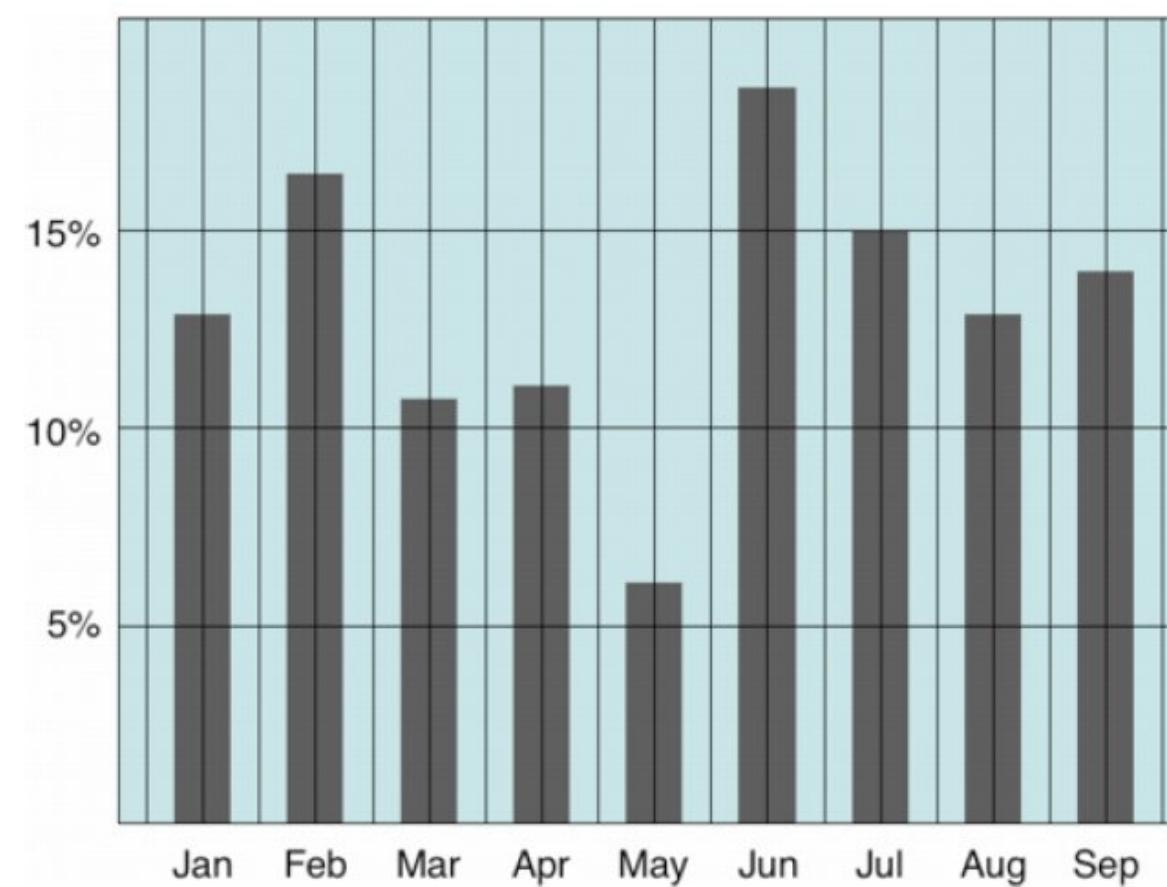
Further examples: Erase non-data ink.



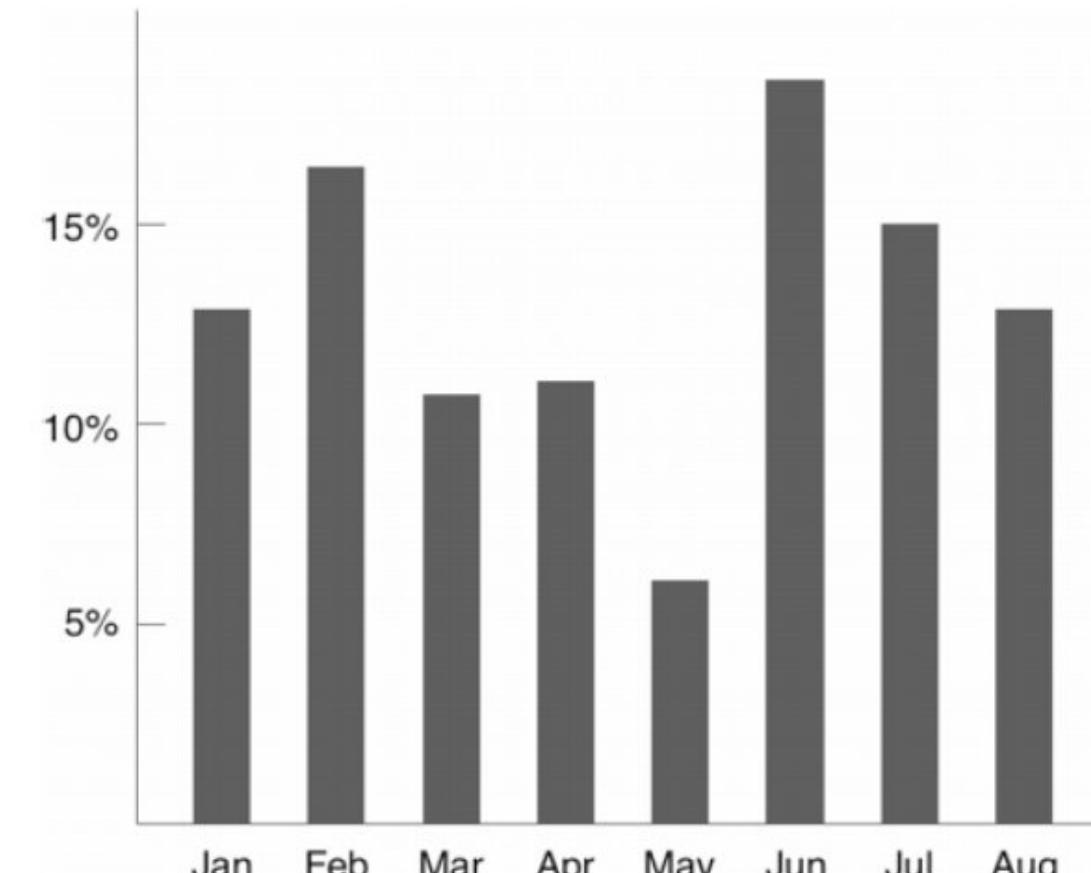
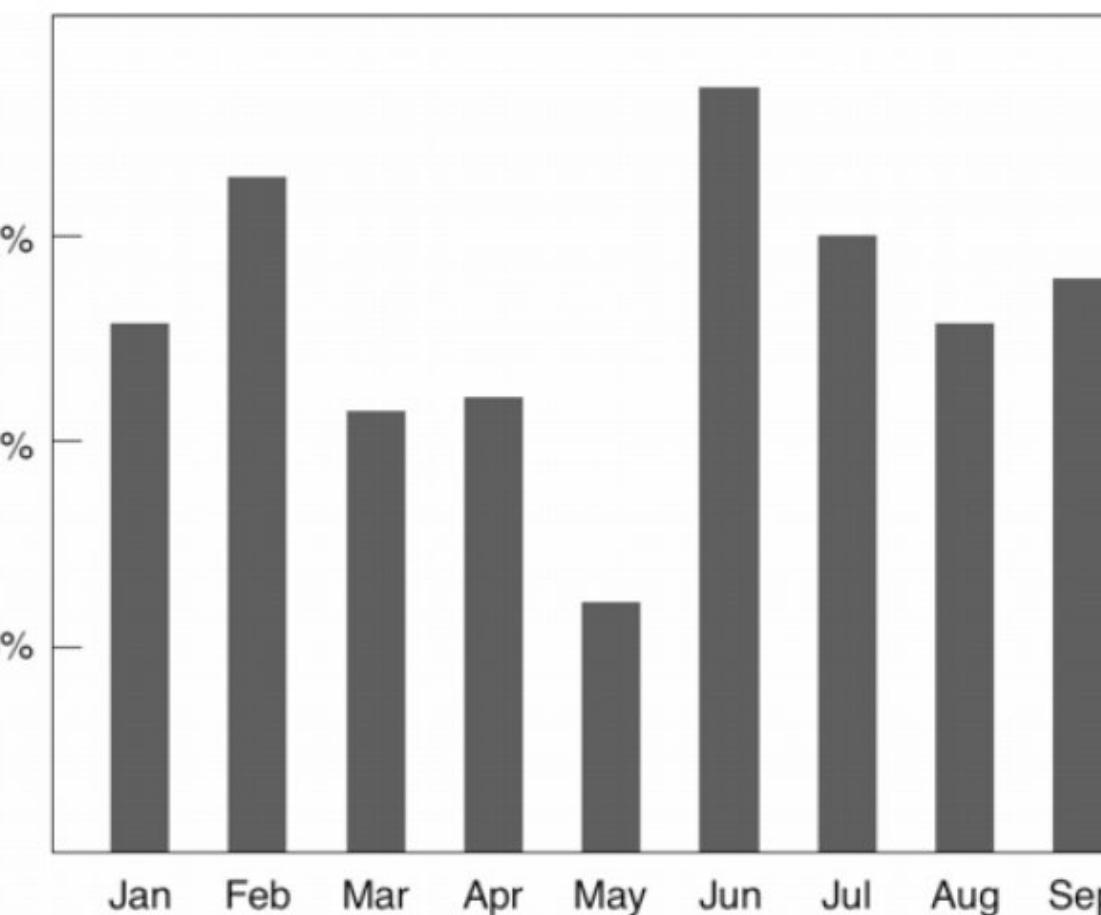
PRINCIPLES OF DATA VISUALIZATION

Occam's Razor: If visualization with less ink possible, do so.

Further examples: Erase non-data ink.



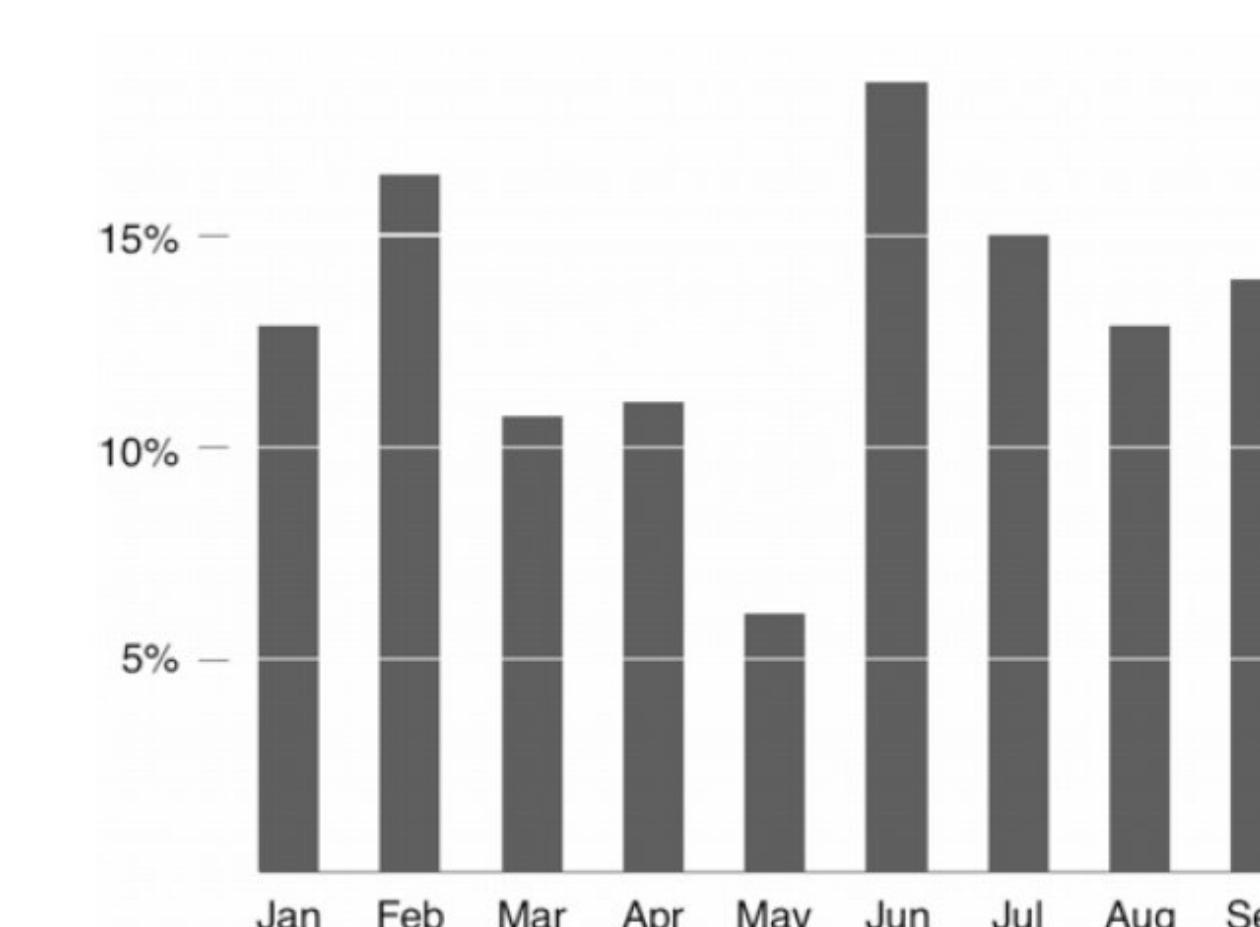
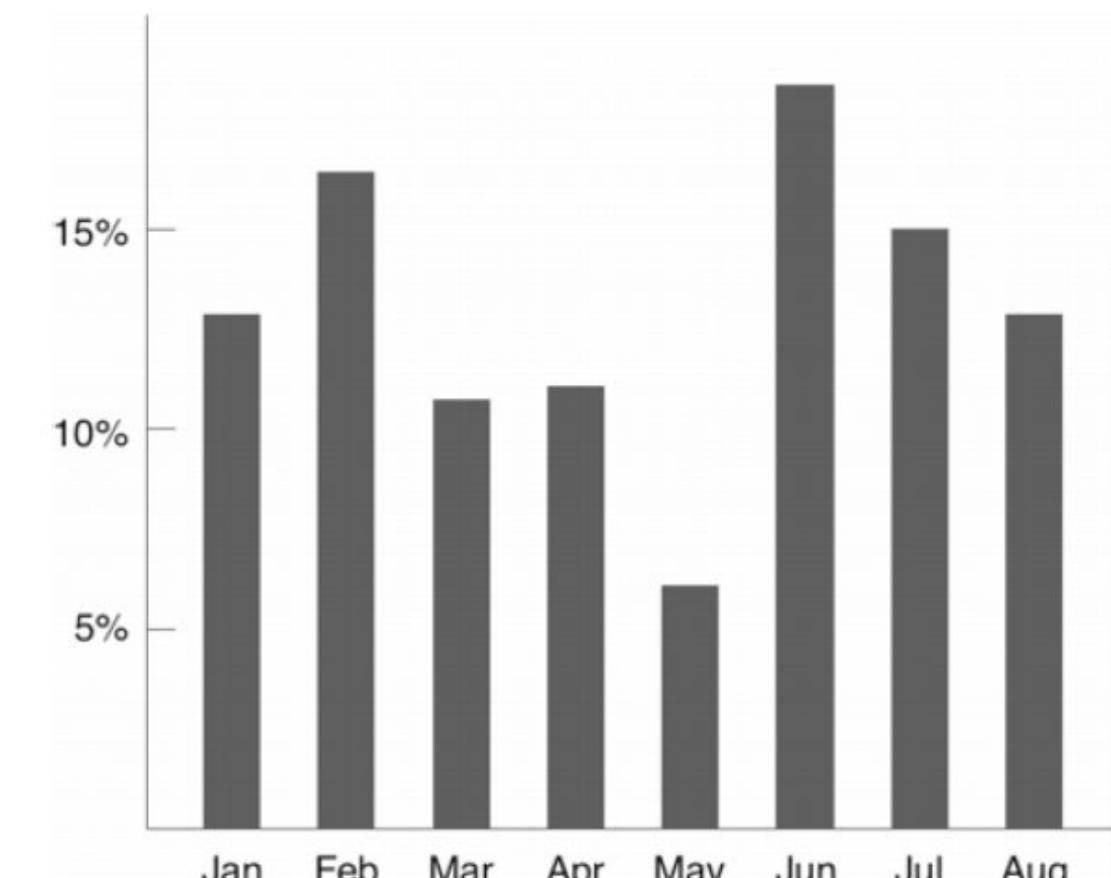
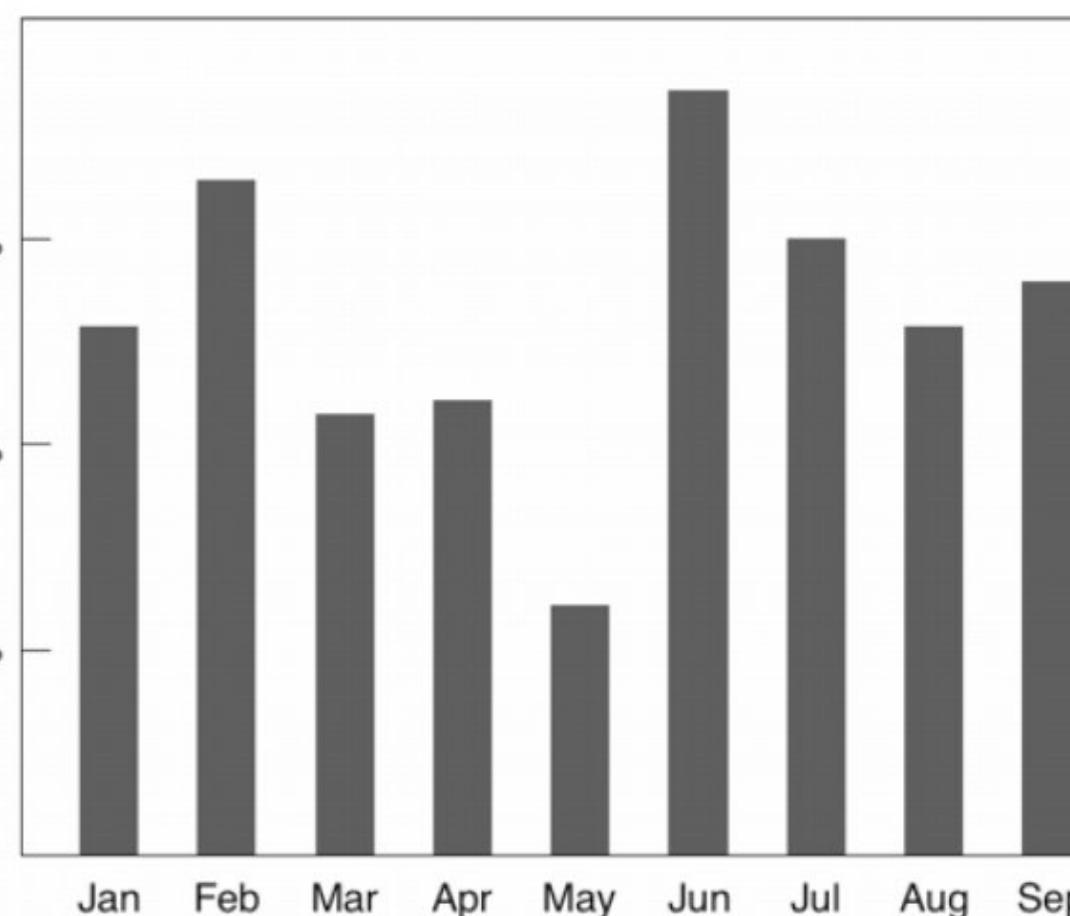
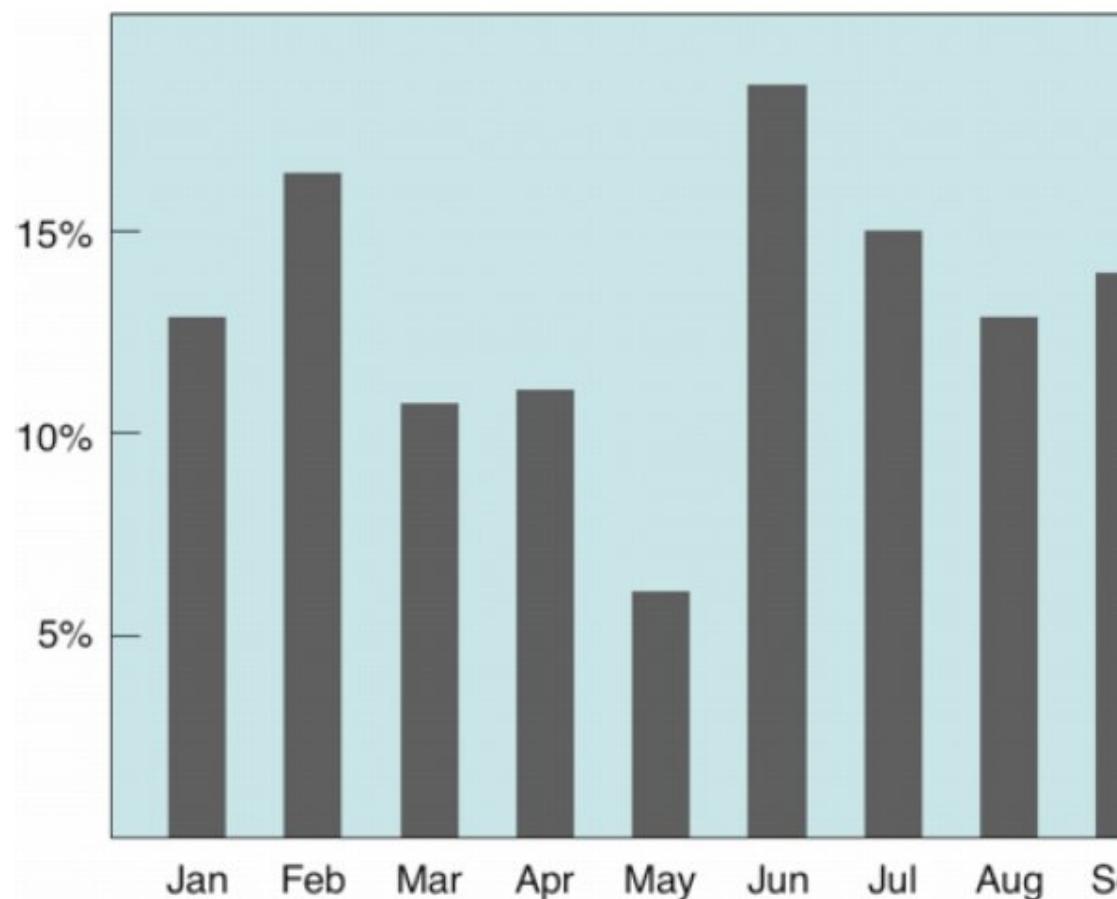
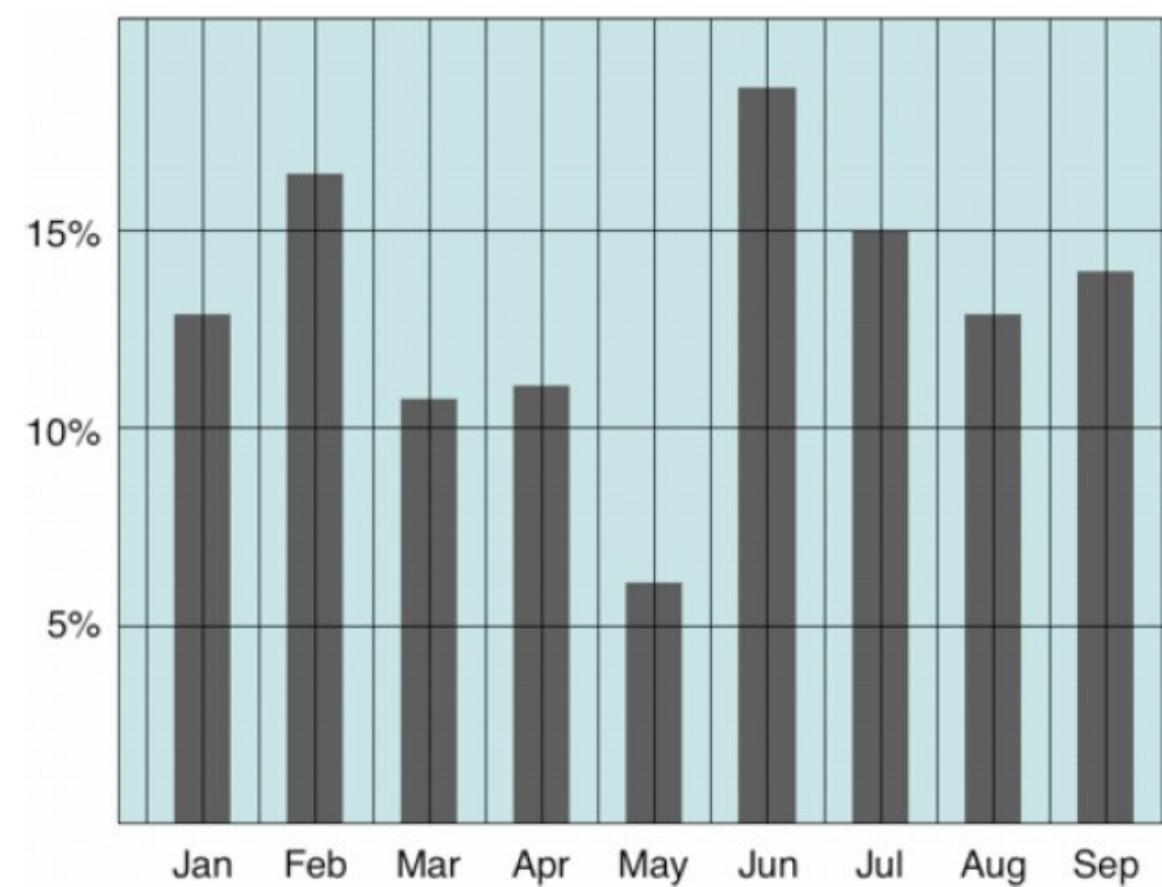
Can less be more?



PRINCIPLES OF DATA VISUALIZATION

Occam's Razor: If visualization with less ink possible, do so.

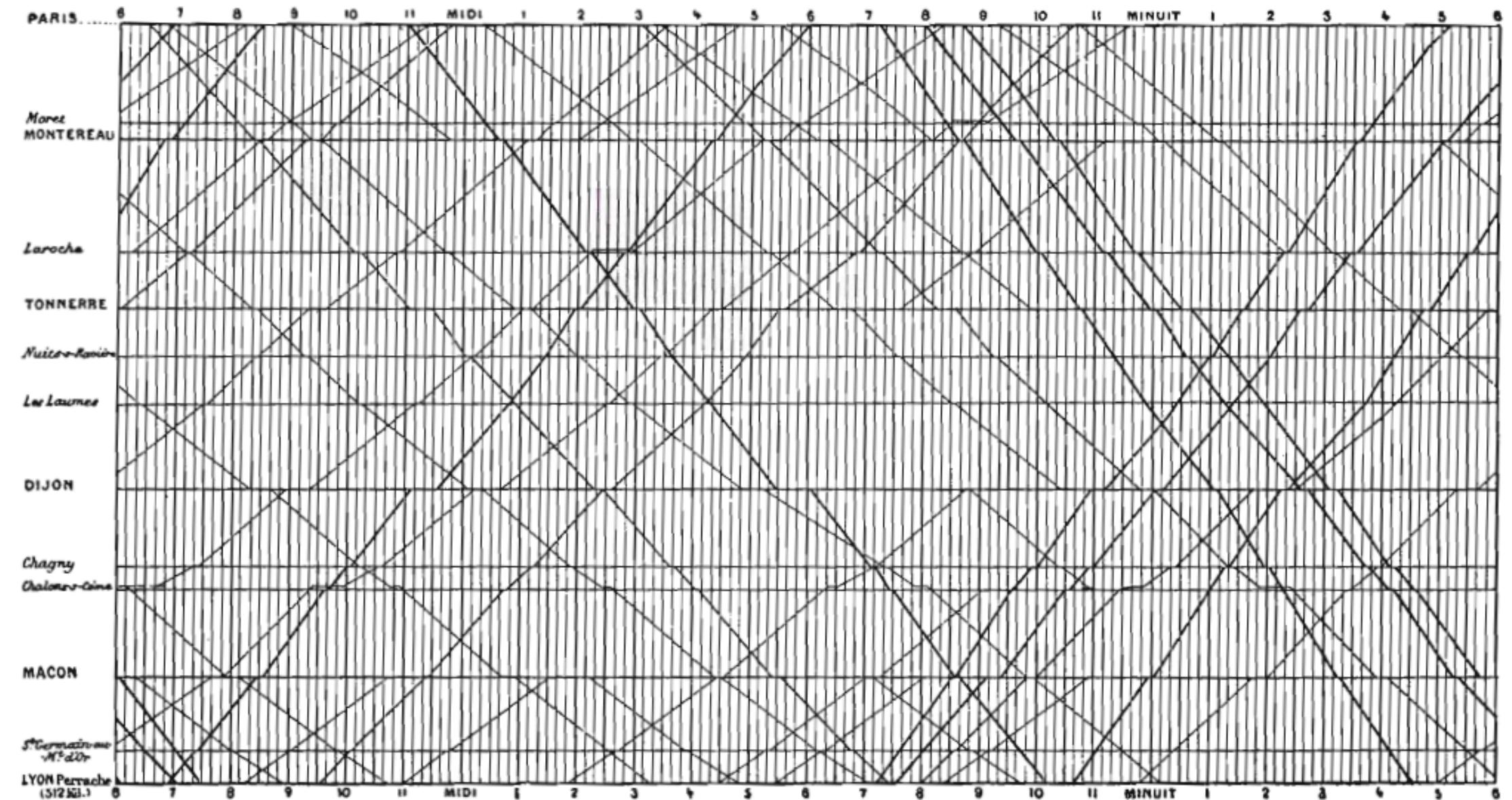
Further examples: Erase non-data ink.



PRINCIPLES OF DATA VISUALIZATION

Occam's Razor: If visualization with less ink possible, do so.

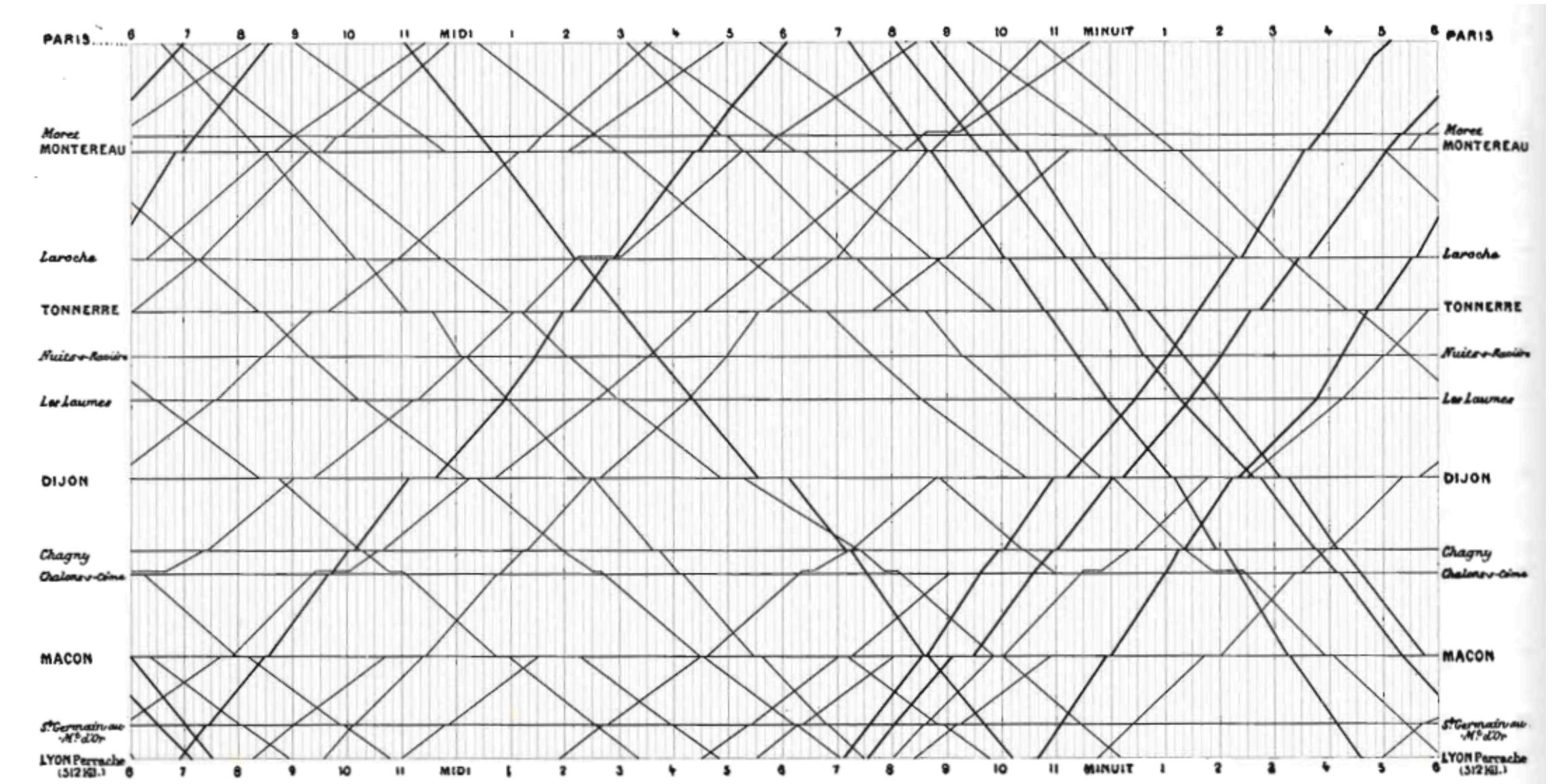
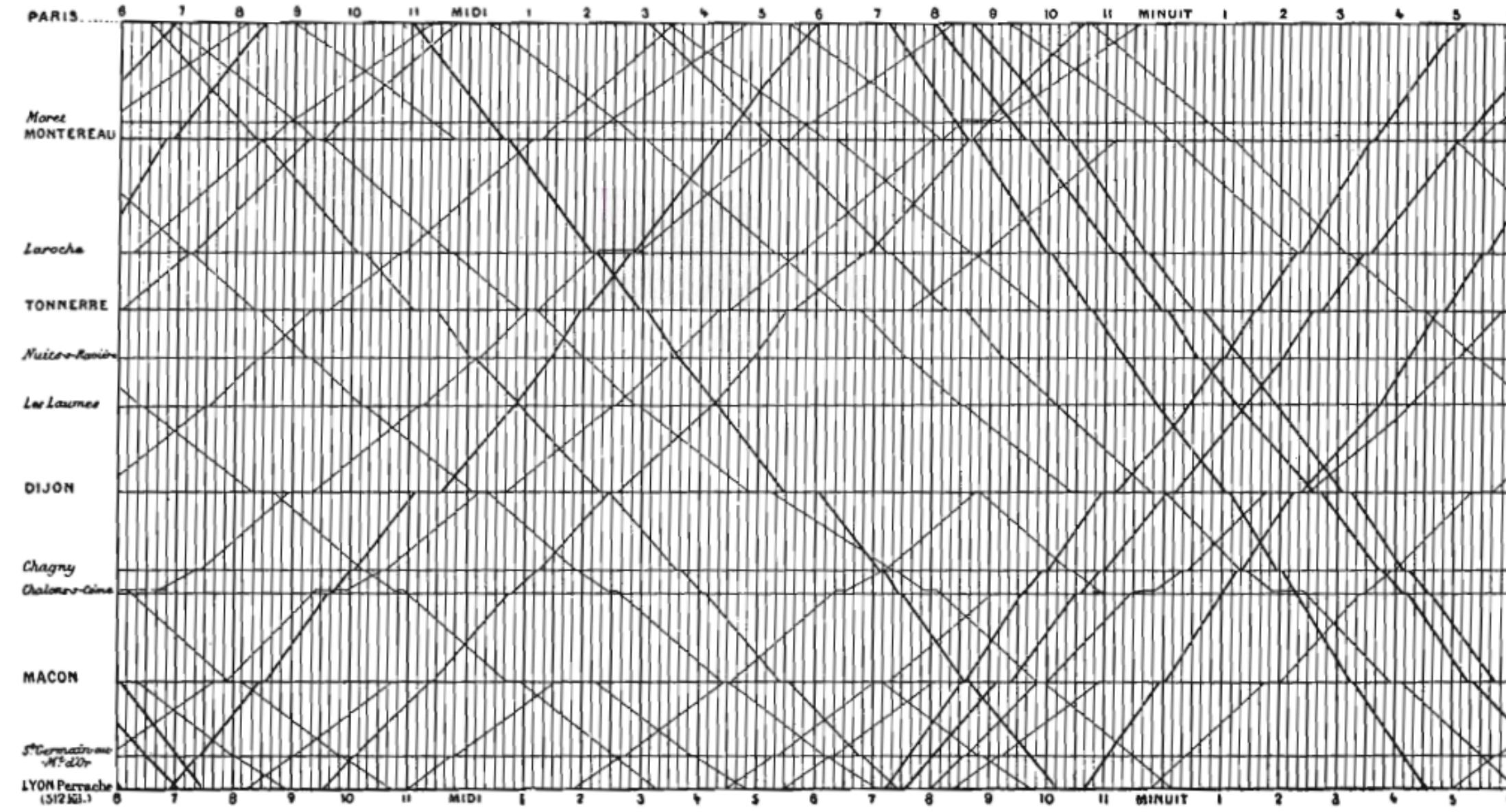
Further examples: Marey train schedule



PRINCIPLES OF DATA VISUALIZATION

Occam's Razor: If visualization with less ink possible, do so.

Further examples: Marey train schedule



PRINCIPLES OF DATA VISUALIZATION

Things to Remember

Tabular Data

Country	Area	Density	Birthrate	Population	Mortality	GDP
Russia	17075200	8.37	99.6	142893540	15.39	8900.0
Mexico	1972550	54.47	92.2	107449525	20.91	9000.0
Japan	377835	337.35	99.0	127463611	3.26	28200.0
United Kingdom	244820	247.57	99.0	60609153	5.16	27700.0
New Zealand	268680	15.17	99.0	4076140	5.85	21600.0
Afghanistan	647500	47.96	36.0	31056997	163.07	700.0
Israel	20770	305.83	95.4	6352117	7.03	19800.0
United States	9631420	30.99	97.0	298444215	6.5	37800.0
China	9596960	136.92	90.9	1313973713	24.18	5000.0
Tajikistan	143100	51.16	99.4	7320815	110.76	1000.0
Burma	678500	69.83	85.3	47382633	67.24	1800.0
Tanzania	945087	39.62	78.2	37445392	98.54	600.0
Tonga	748	153.33	98.5	114689	12.62	2200.0
Germany	357021	230.86	99.0	82422299	4.16	27600.0
Australia	7686850	2.64	100.0	20264082	4.69	29000.0

What improvements can be made?

PRINCIPLES OF DATA VISUALIZATION

Things to Remember

Tabular Data

Country	Area	Density	Birthrate	Population	Mortality	GDP
Russia	17075200	8.37	99.6	142893540	15.39	8900.0
Mexico	1972550	54.47	92.2	107449525	20.91	9000.0
Japan	377835	337.35	99.0	127463611	3.26	28200.0
United Kingdom	244820	247.57	99.0	60609153	5.16	27700.0
New Zealand	268680	15.17	99.0	4076140	5.85	21600.0
Afghanistan	647500	47.96	36.0	31056997	163.07	700.0
Israel	20770	305.83	95.4	6352117	7.03	19800.0
United States	9631420	30.99	97.0	298444215	6.5	37800.0
China	9596960	136.92	90.9	1313973713	24.18	5000.0
Tajikistan	143100	51.16	99.4	7320815	110.76	1000.0
Burma	678500	69.83	85.3	47382633	67.24	1800.0
Tanzania	945087	39.62	78.2	37445392	98.54	600.0
Tonga	748	153.33	98.5	114689	12.62	2200.0
Germany	357021	230.86	99.0	82422299	4.16	27600.0
Australia	7686850	2.64	100.0	20264082	4.69	29000.0

Country	Population	Area	Density	Mortality	GDP	Birth Rate
Afghanistan	31,056,997	647,500	47.96	163.07	700	36.0
Australia	20,264,082	7,686,850	2.64	4.69	29,000	100.0
Burma	47,382,633	678,500	69.83	67.24	1,800	85.3
China	1,313,973,713	9,596,960	136.92	24.18	5,000	90.9
Germany	82,422,299	357,021	230.86	4.16	27,600	99.0
Israel	6,352,117	20,770	305.83	7.03	19,800	95.4
Japan	127,463,611	377,835	337.35	3.26	28,200	99.0
Mexico	107,449,525	1,972,550	54.47	20.91	9,000	92.2
New Zealand	4,076,140	268,680	15.17	5.85	21,600	99.0
Russia	142,893,540	17,075,200	8.37	15.39	8,900	99.6
Tajikistan	7,320,815	143,100	51.16	110.76	1,000	99.4
Tanzania	37,445,392	945,087	39.62	98.54	600	78.2
Tonga	114,689	748	153.33	12.62	2,200	98.5
United Kingdom	60,609,153	244,820	247.57	5.16	27,700	99.0
United States	298,444,215	9,631,420	30.99	6.50	37,800	97.0

What improvements can be made?

Order rows/columns

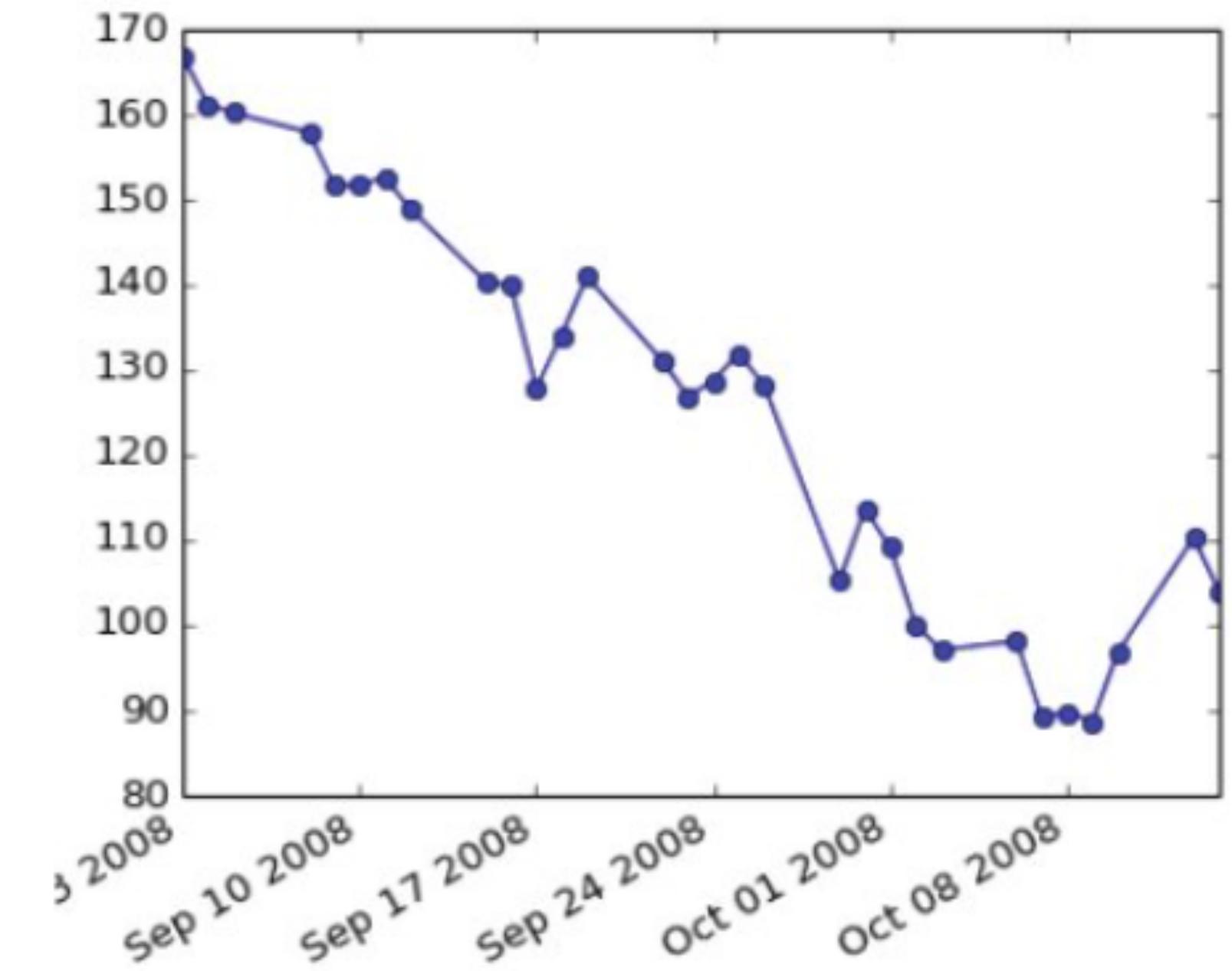
Remove uninformative digits, right justify numbers, add commas

Highlight biggest values

PRINCIPLES OF DATA VISUALIZATION

Things to Remember

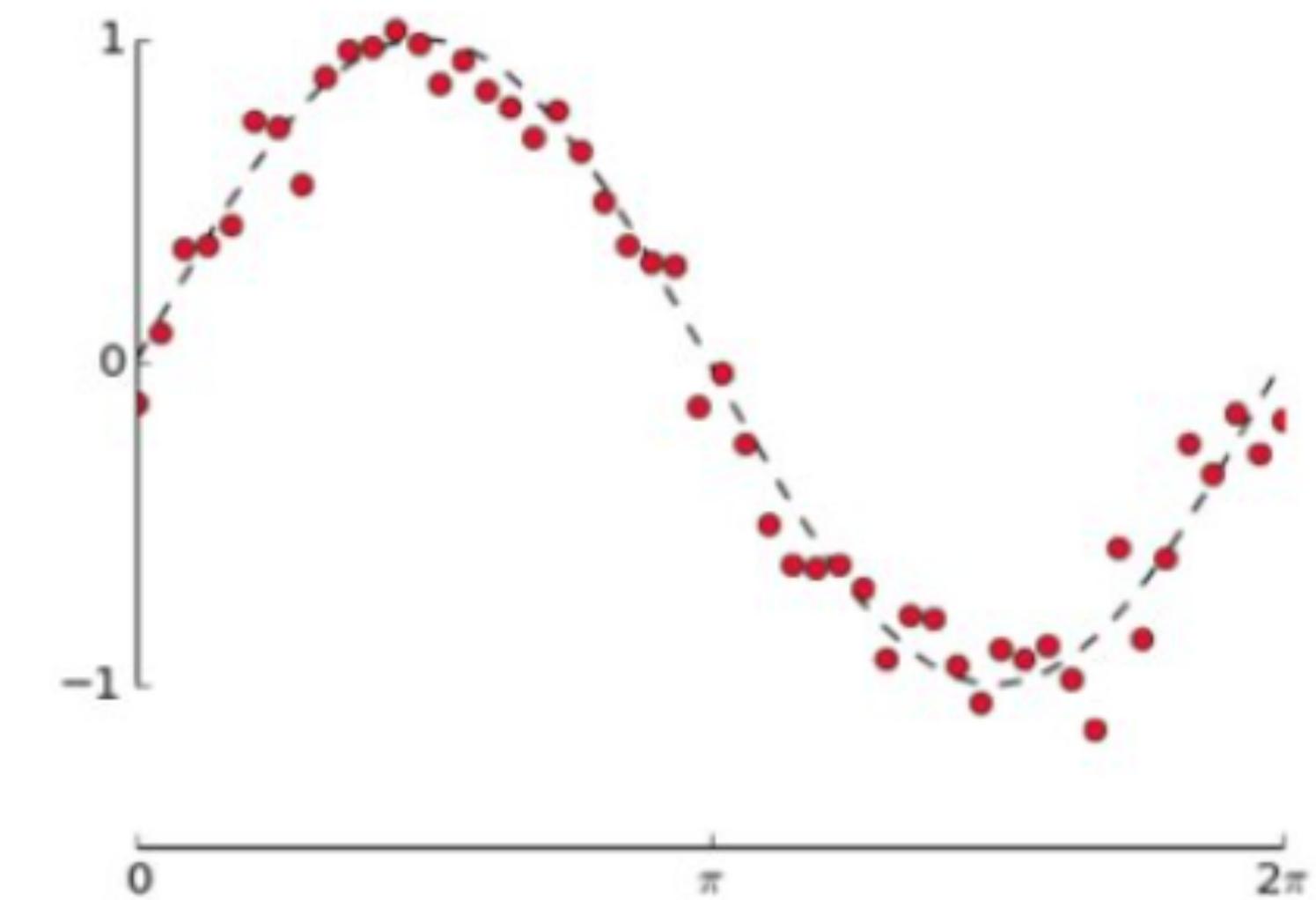
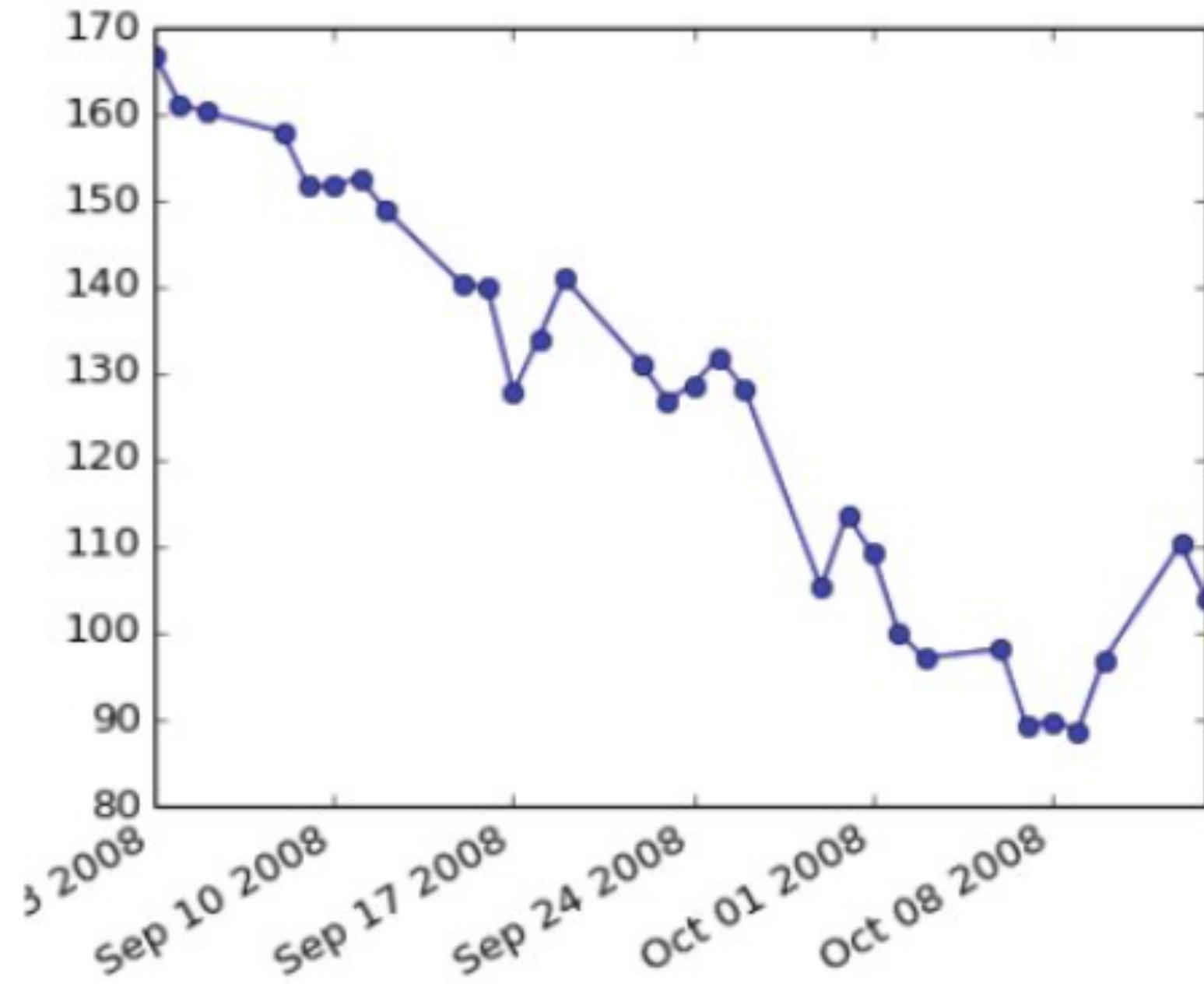
Line Charts: Series of data points connected with lines.



PRINCIPLES OF DATA VISUALIZATION

Things to Remember

Line Charts: Series of data points connected with lines.



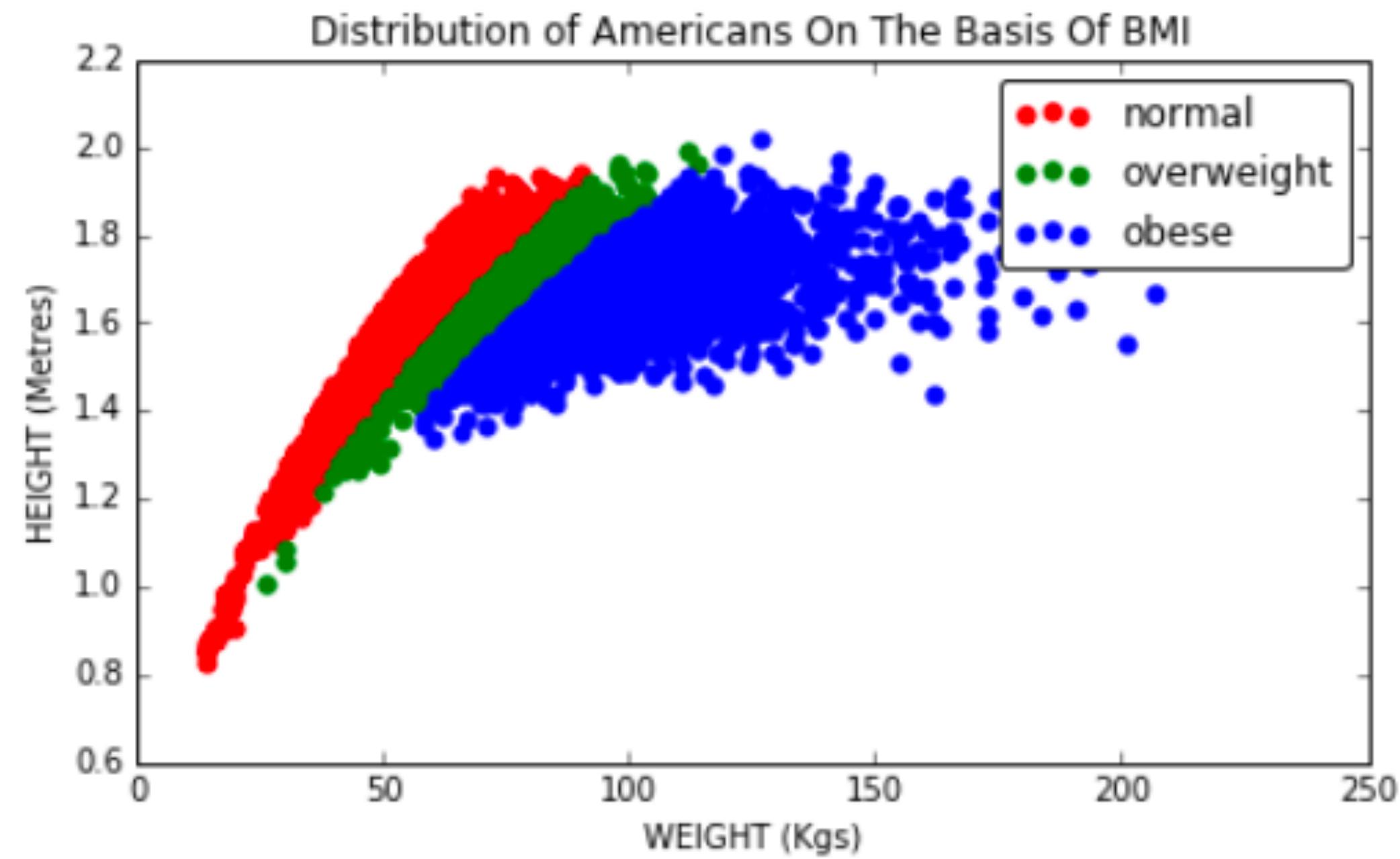
May make sense with time series data.

If no meaningful adjacency between data points, don't us line chart, instead overlay fitted curve.

PRINCIPLES OF DATA VISUALIZATION

Things to Remember

Scatter plots: Dot size

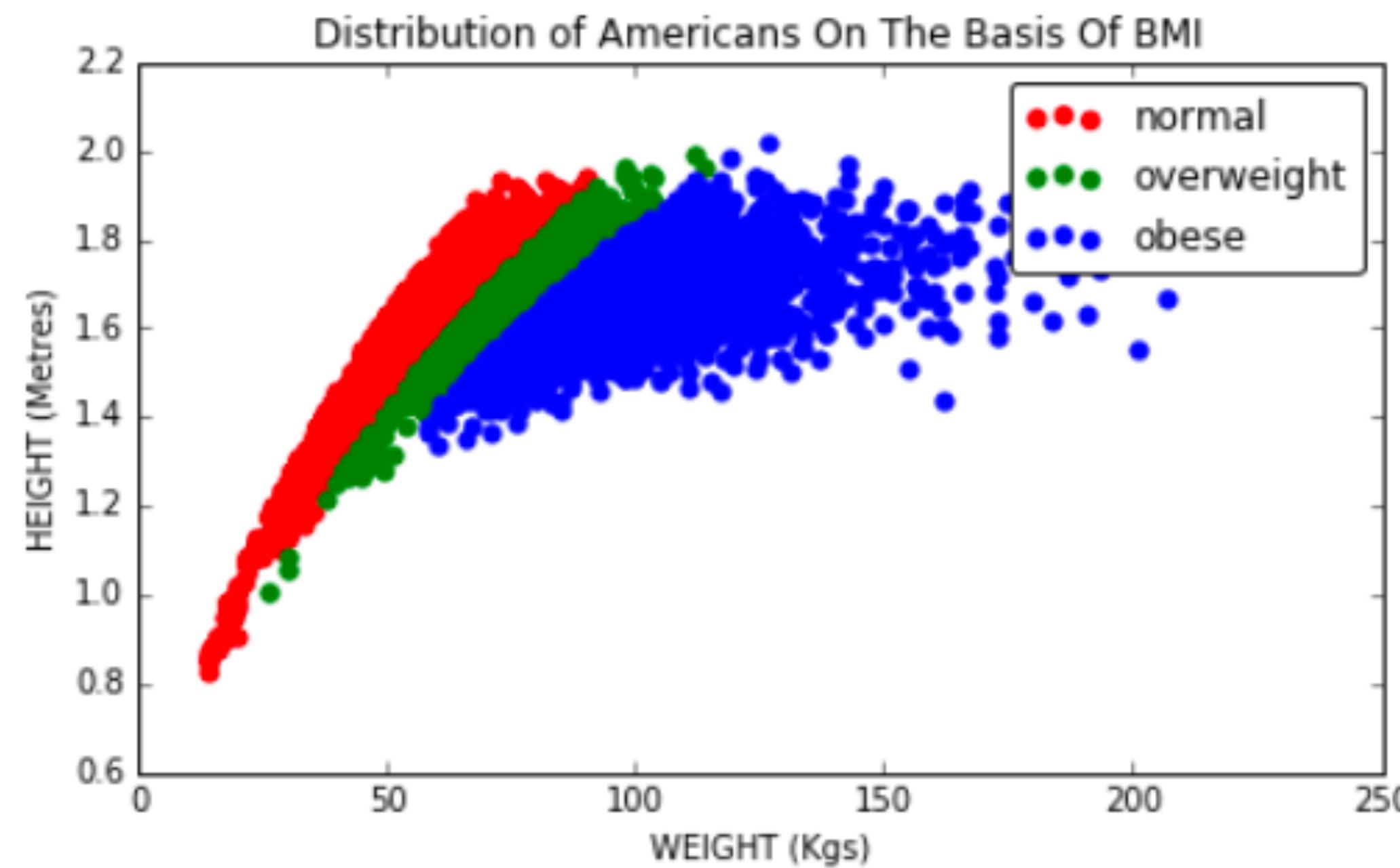


Large dot size: more
overlaps and obscuring

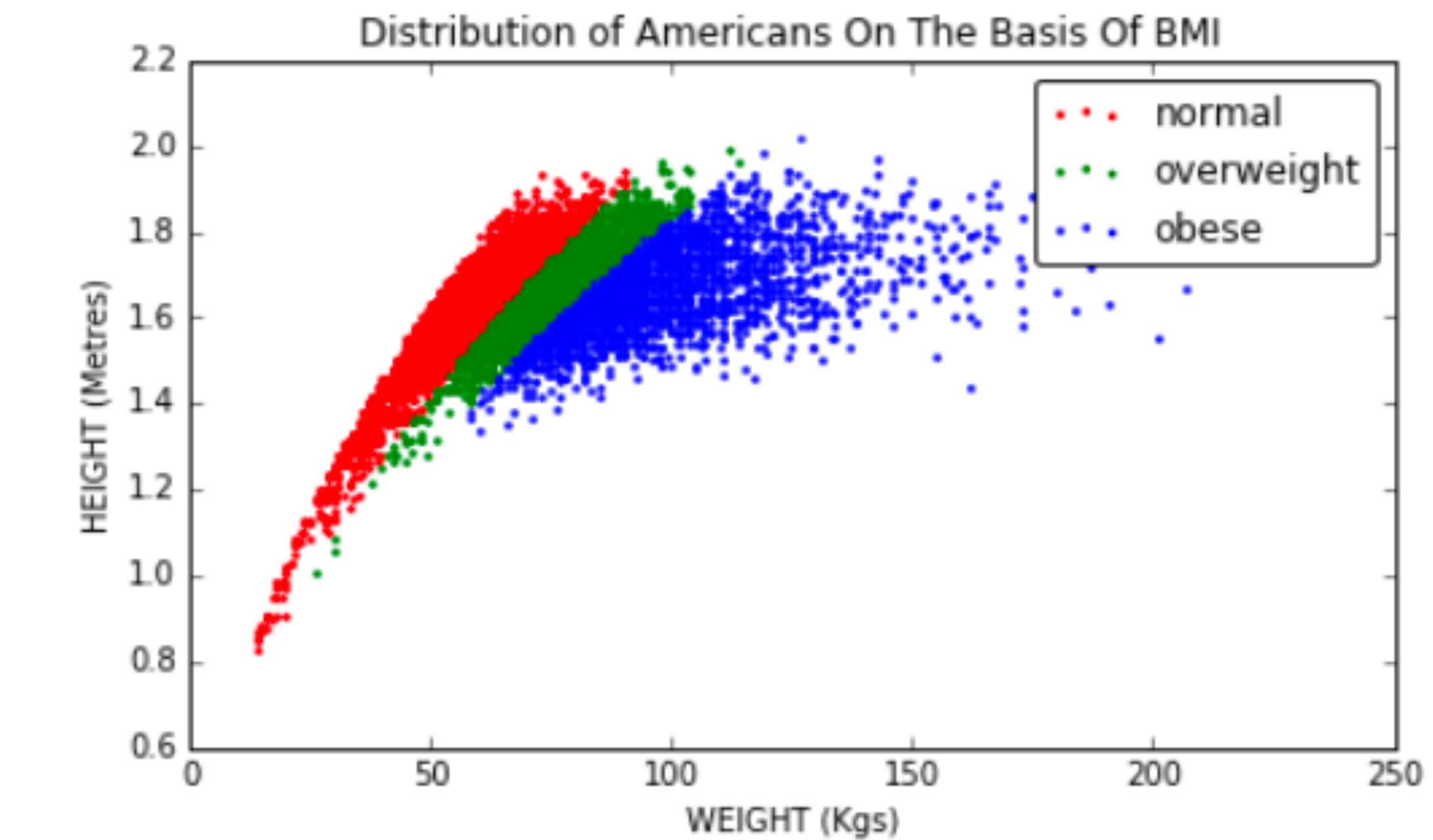
PRINCIPLES OF DATA VISUALIZATION

Things to Remember

Scatter plots: Dot size



Large dot size: more overlaps and obscuring

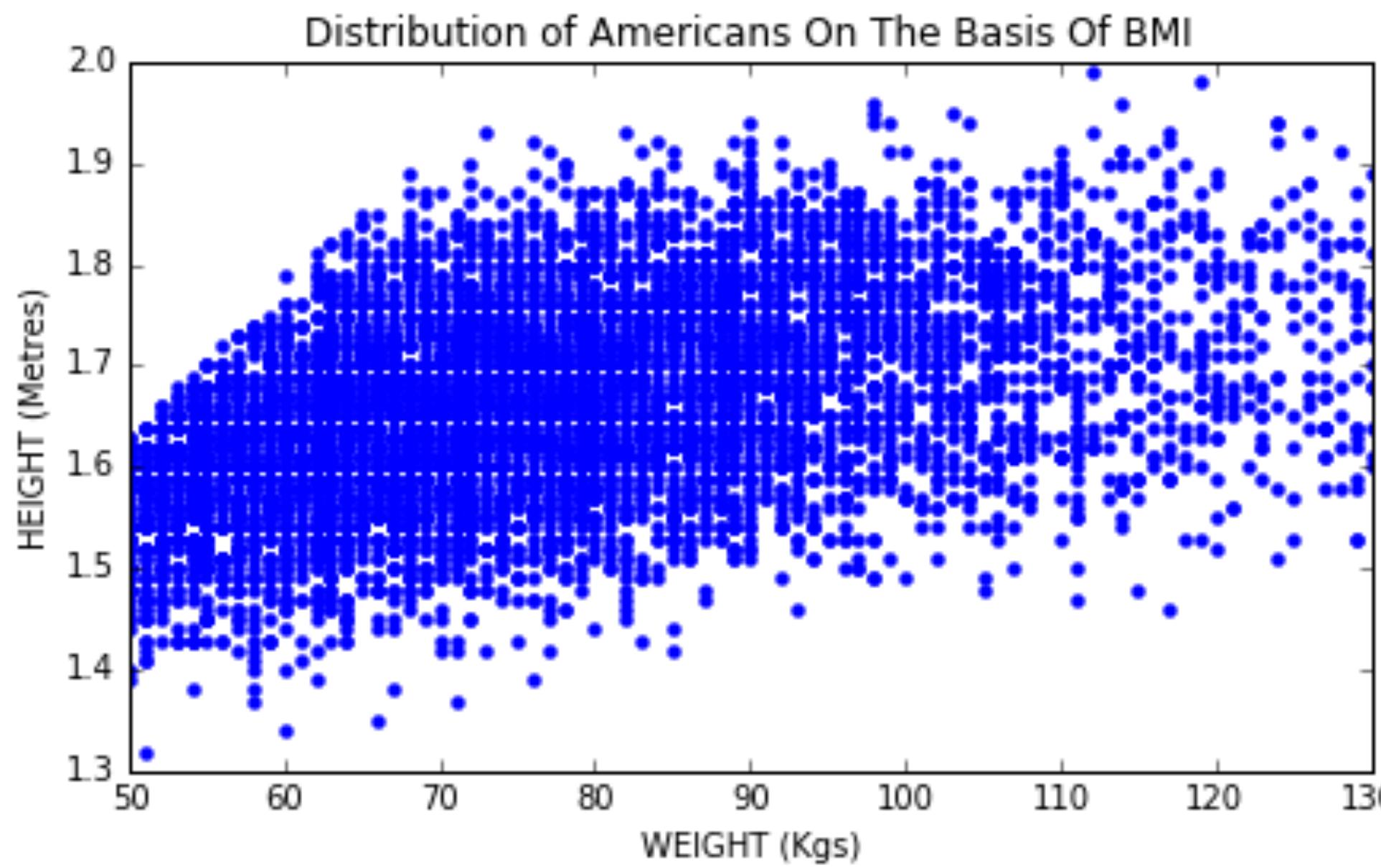


Smaller dots for this size is better

PRINCIPLES OF DATA VISUALIZATION

Things to Remember

Scatter plots: Heat map for better frequency visualization

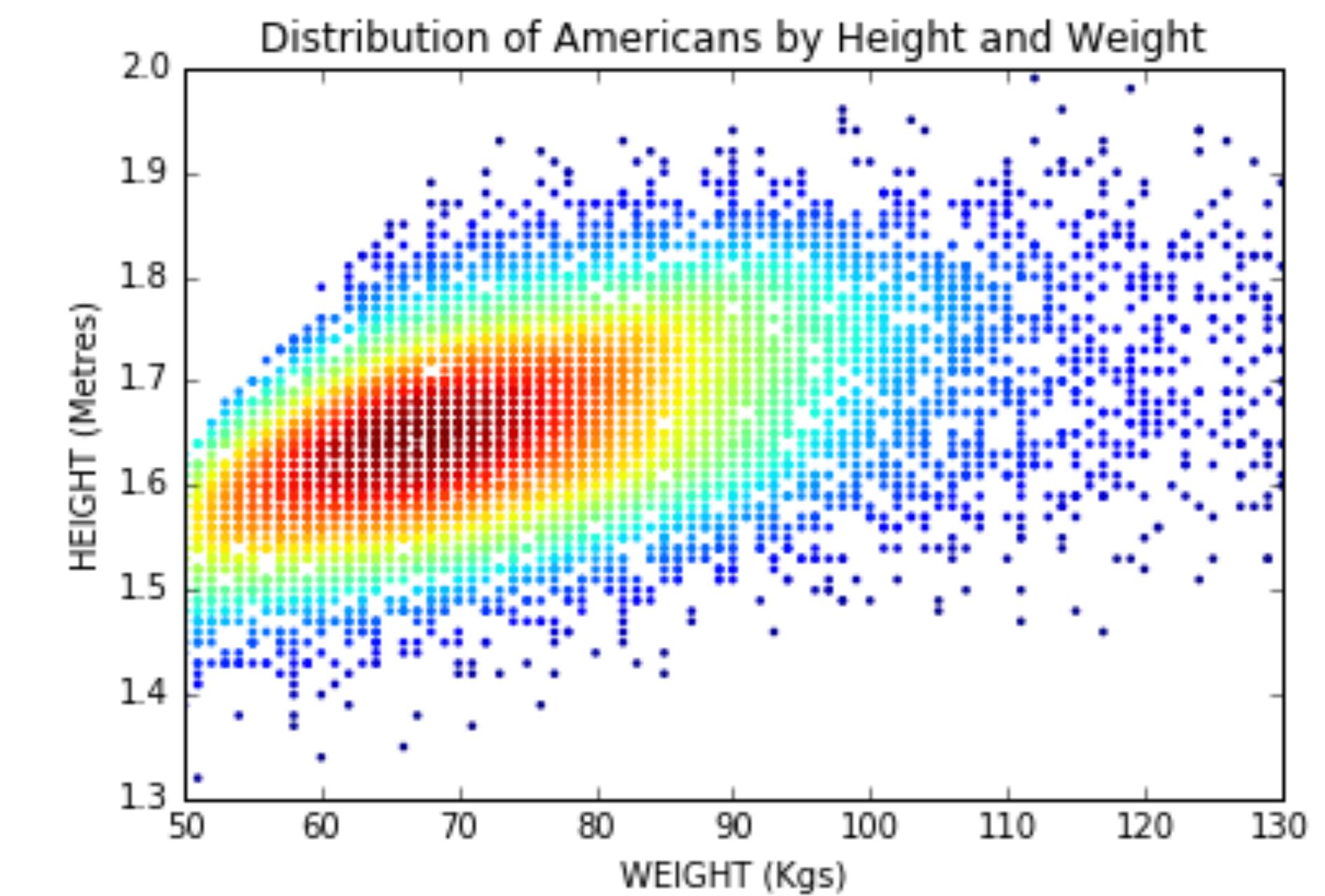
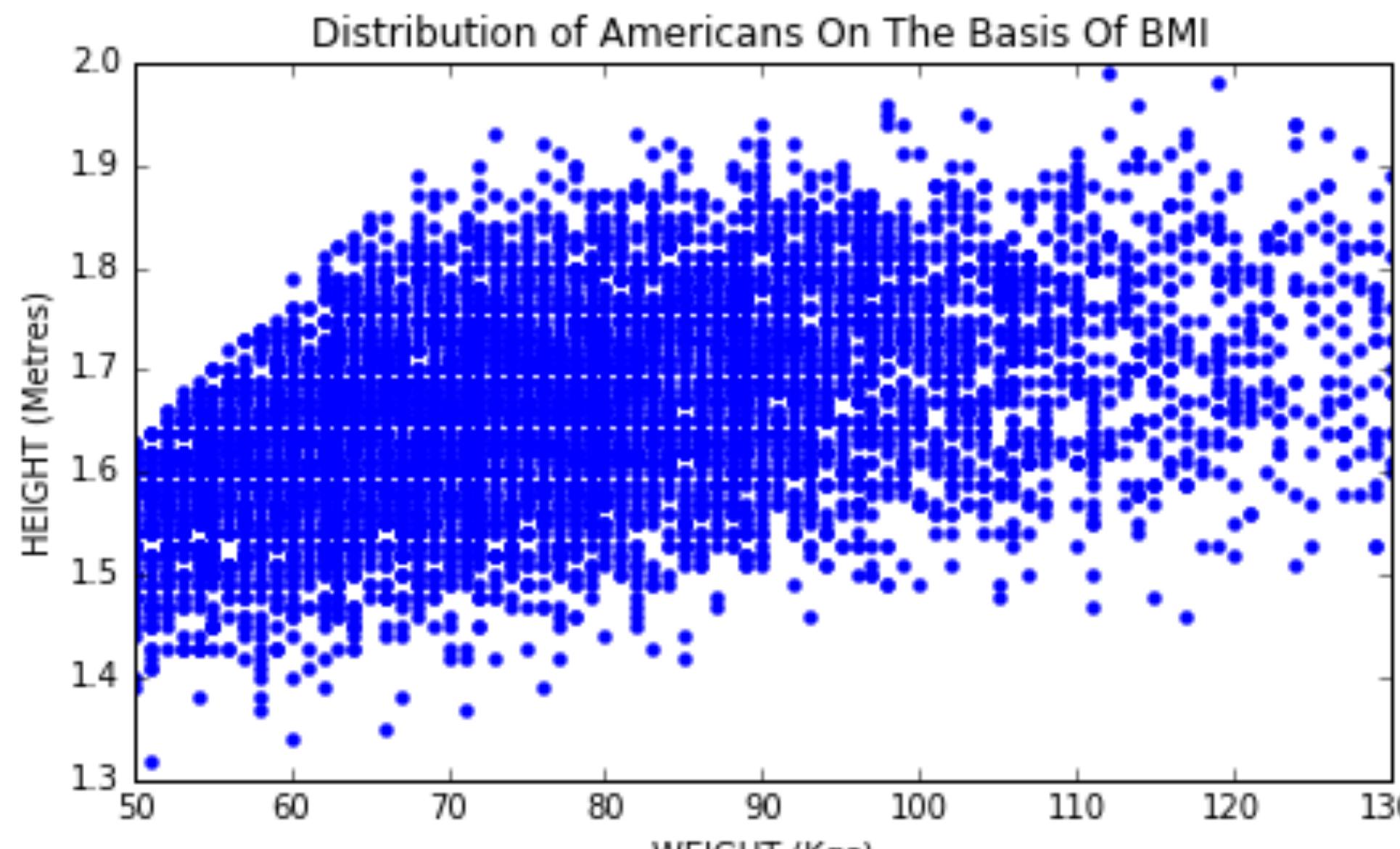


How to use color to
our advantage?

PRINCIPLES OF DATA VISUALIZATION

Things to Remember

Scatter plots: Heat map for better frequency visualization



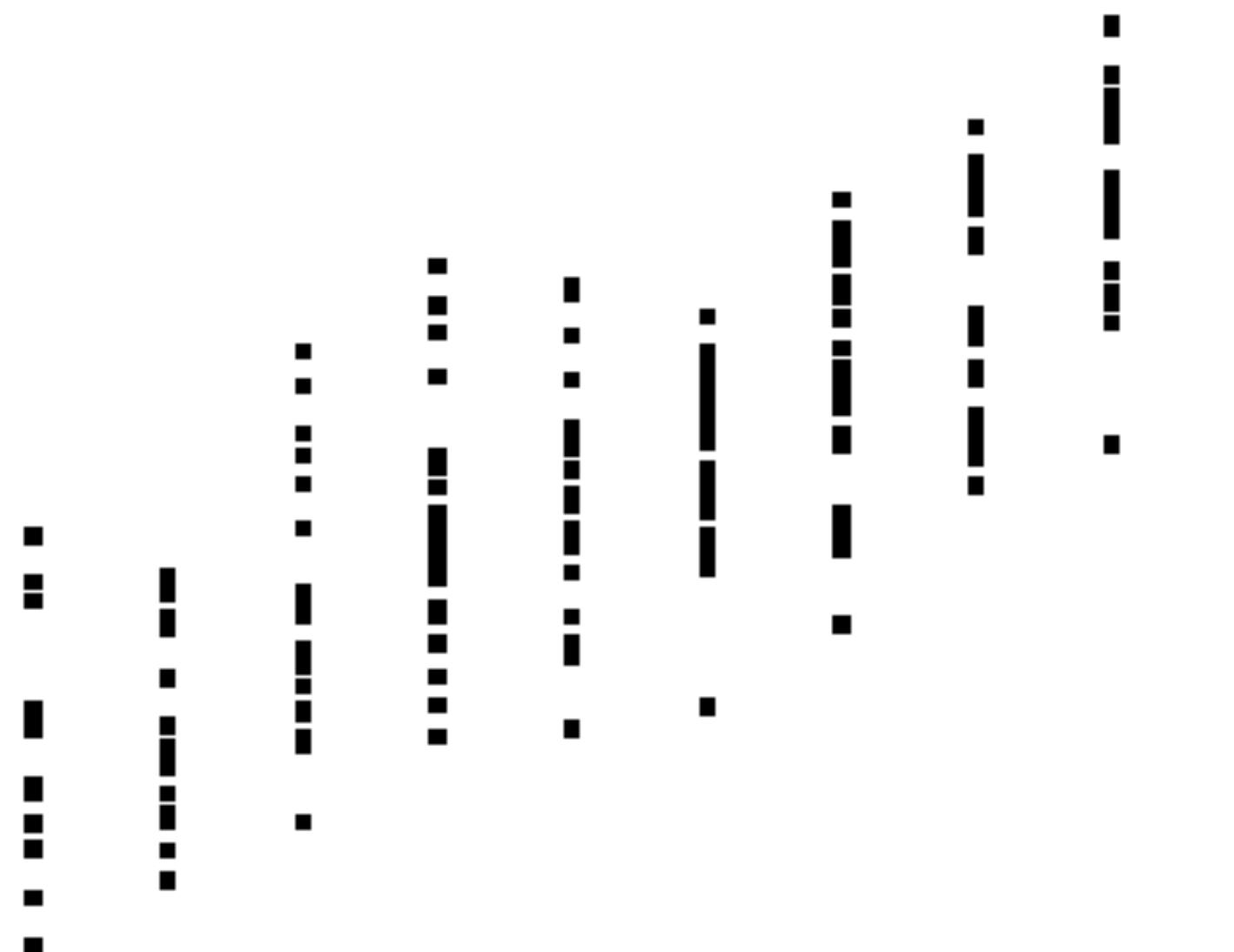
How to use color to
our advantage?

Colored heat map reveals the
distribution better.

PRINCIPLES OF DATA VISUALIZATION

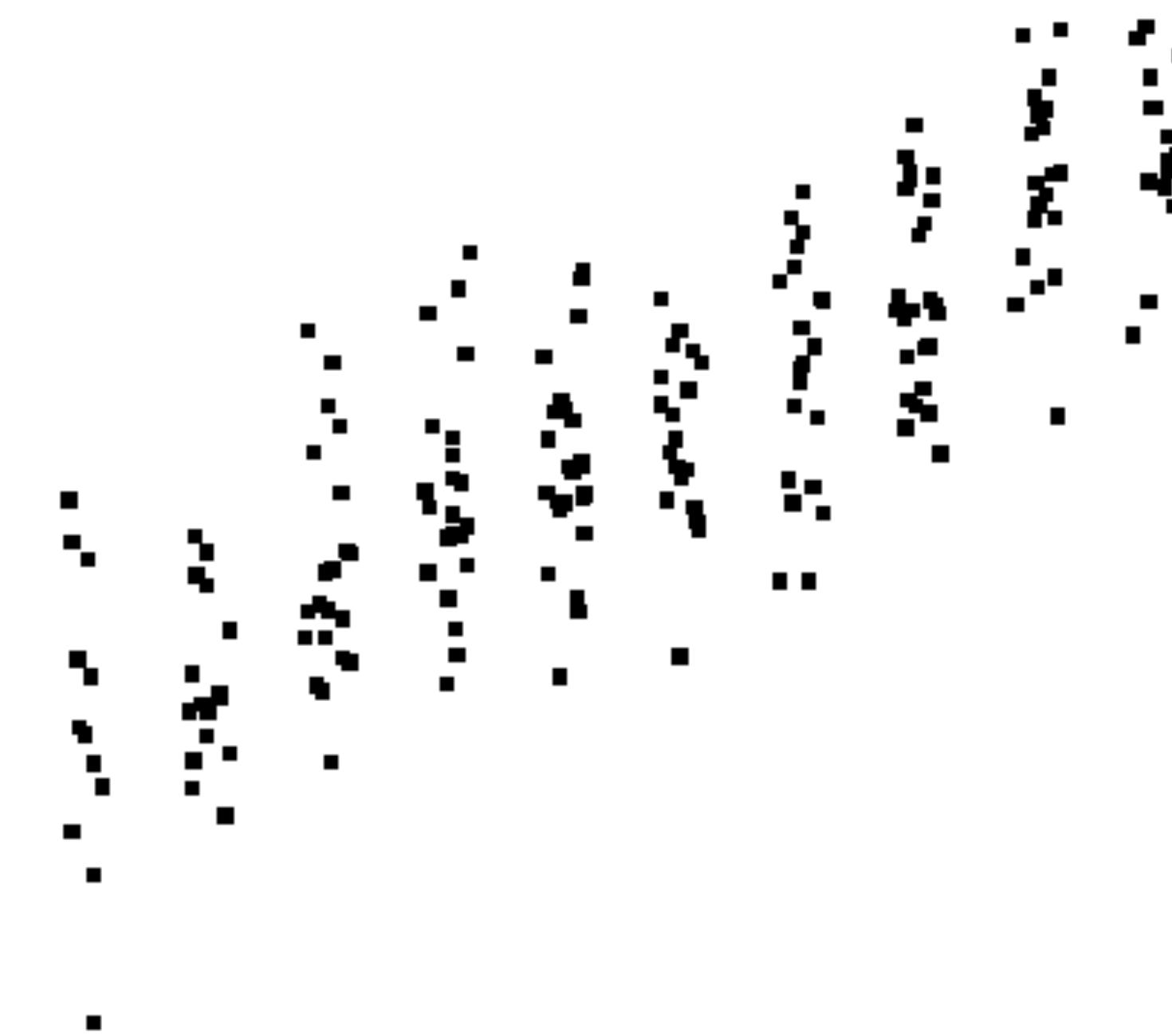
Things to Remember

Scatter plots: Jittering: add random noise ϵ to x and/or y coordinate.



- thomasleeper.com/Rcourse/Tutorials

Difficult to see the distribution.

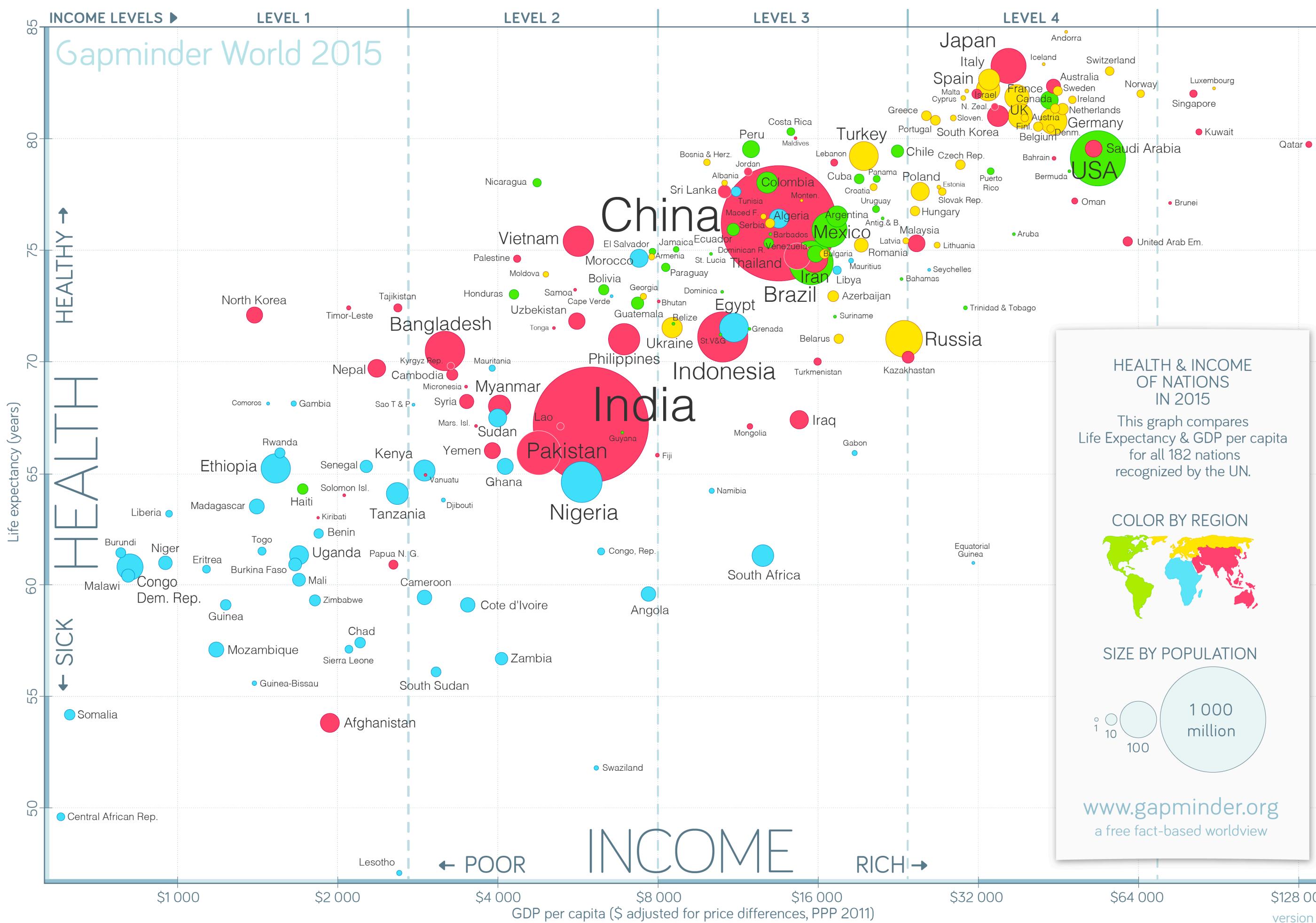


Much easier to see the distribution after jittering of x-coordinates.

PRINCIPLES OF DATA VISUALIZATION

Things to Remember

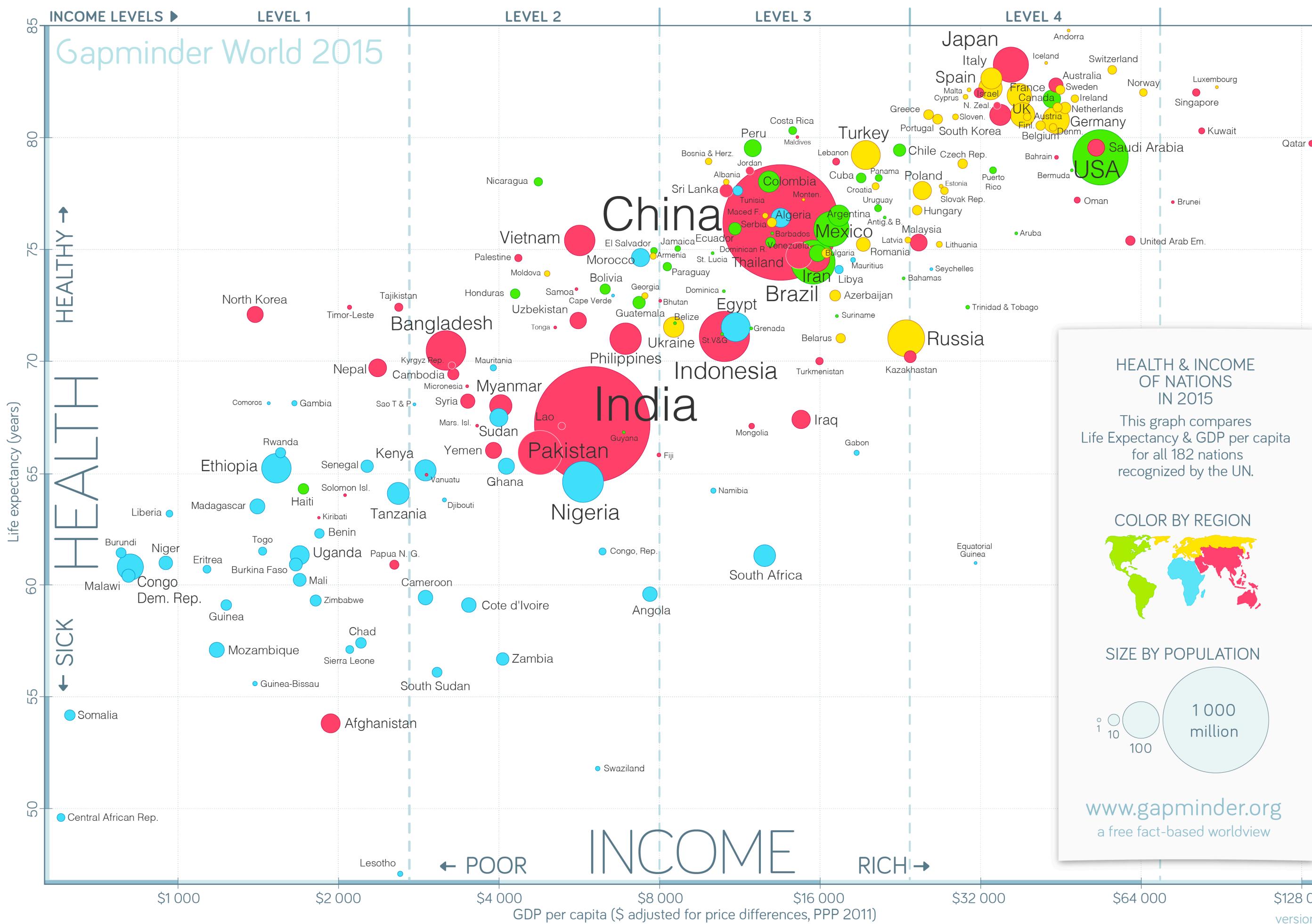
Scatter plots: Visualize >2 variables: extra dimensions (size and color)



PRINCIPLES OF DATA VISUALIZATION

Things to Remember

Scatter plots: Visualize >2 variables: extra dimensions (size and color)



Size: Population

Color: Region

y-coordinate: Life expectancy

x-coordinate: GDP per person
(Log transformed)

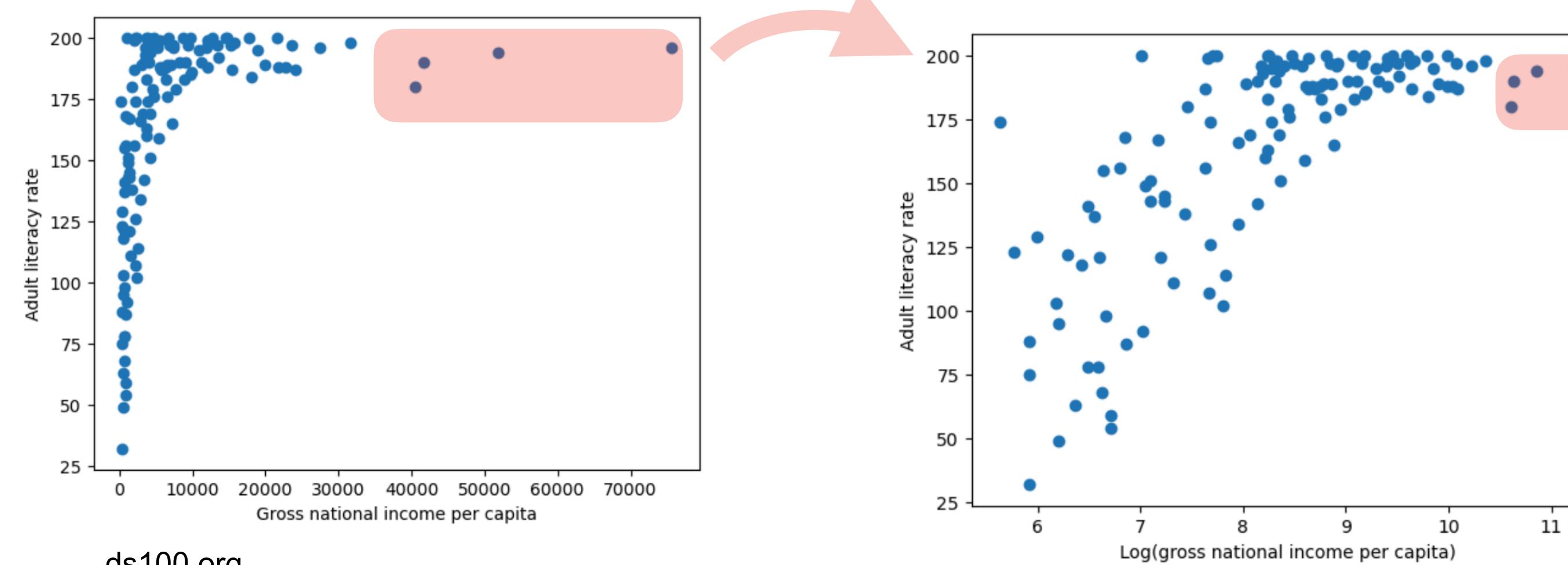
Why do we have
transformations?

Gapminder World

PRINCIPLES OF DATA VISUALIZATION

Things to Remember

Transformations: better visualize/understand relationship bw variables



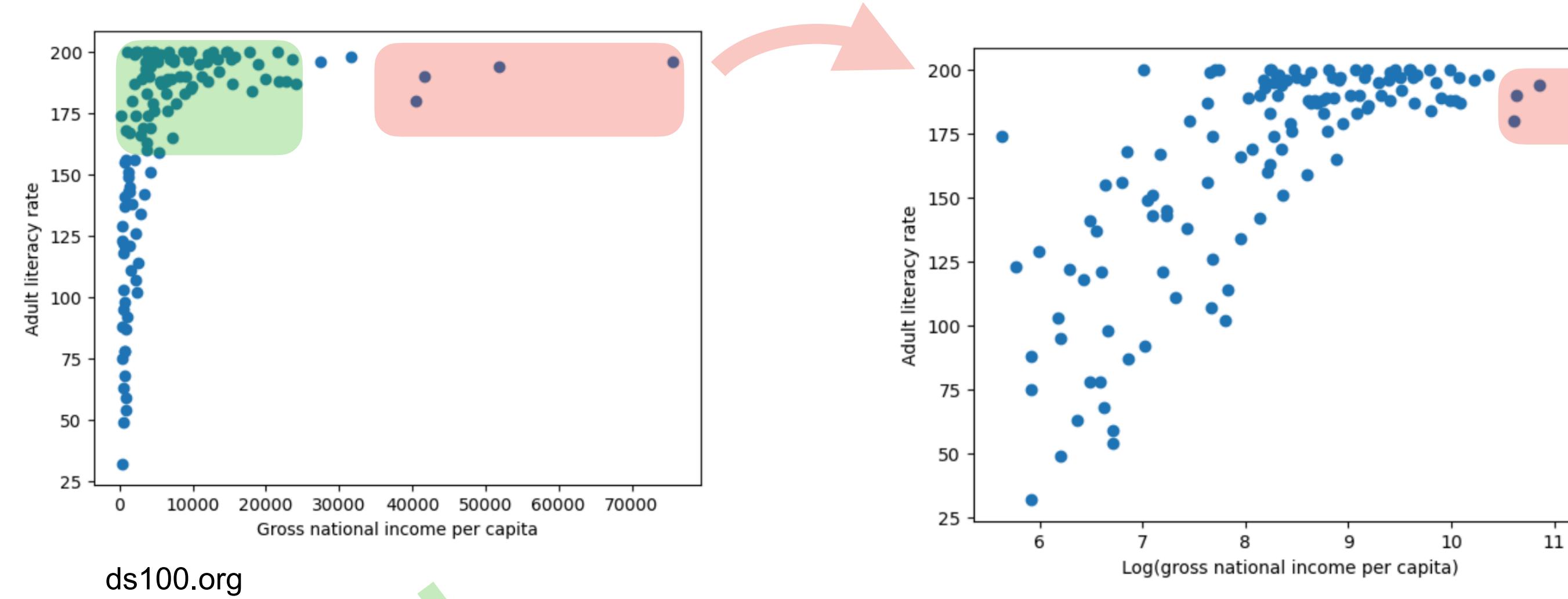
ds100.org

Few outliers at large x:
Log transform x-coordinates.

PRINCIPLES OF DATA VISUALIZATION

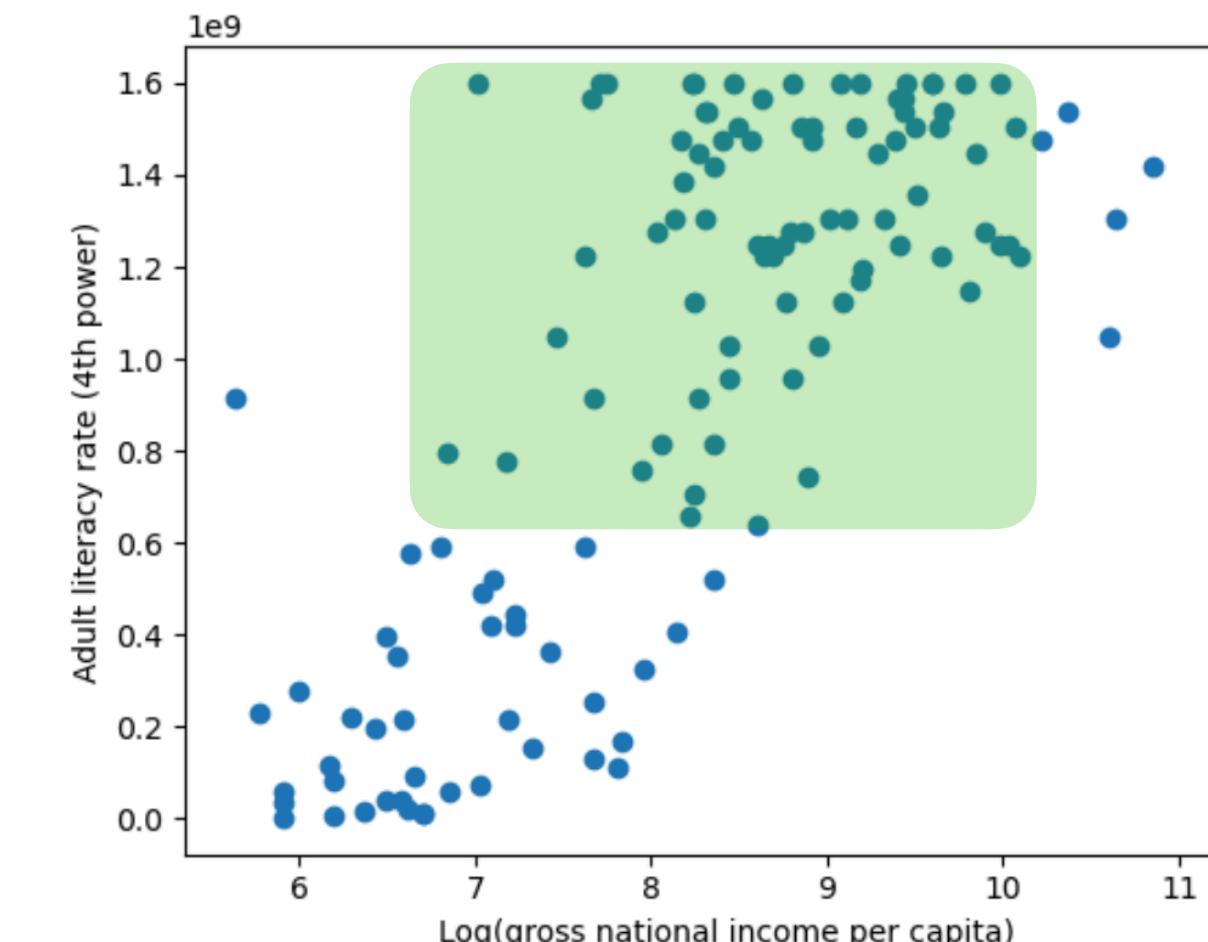
Things to Remember

Transformations: better visualize/understand relationship bw variables



ds100.org

Few outliers at large x:
Log transform x-coordinates.

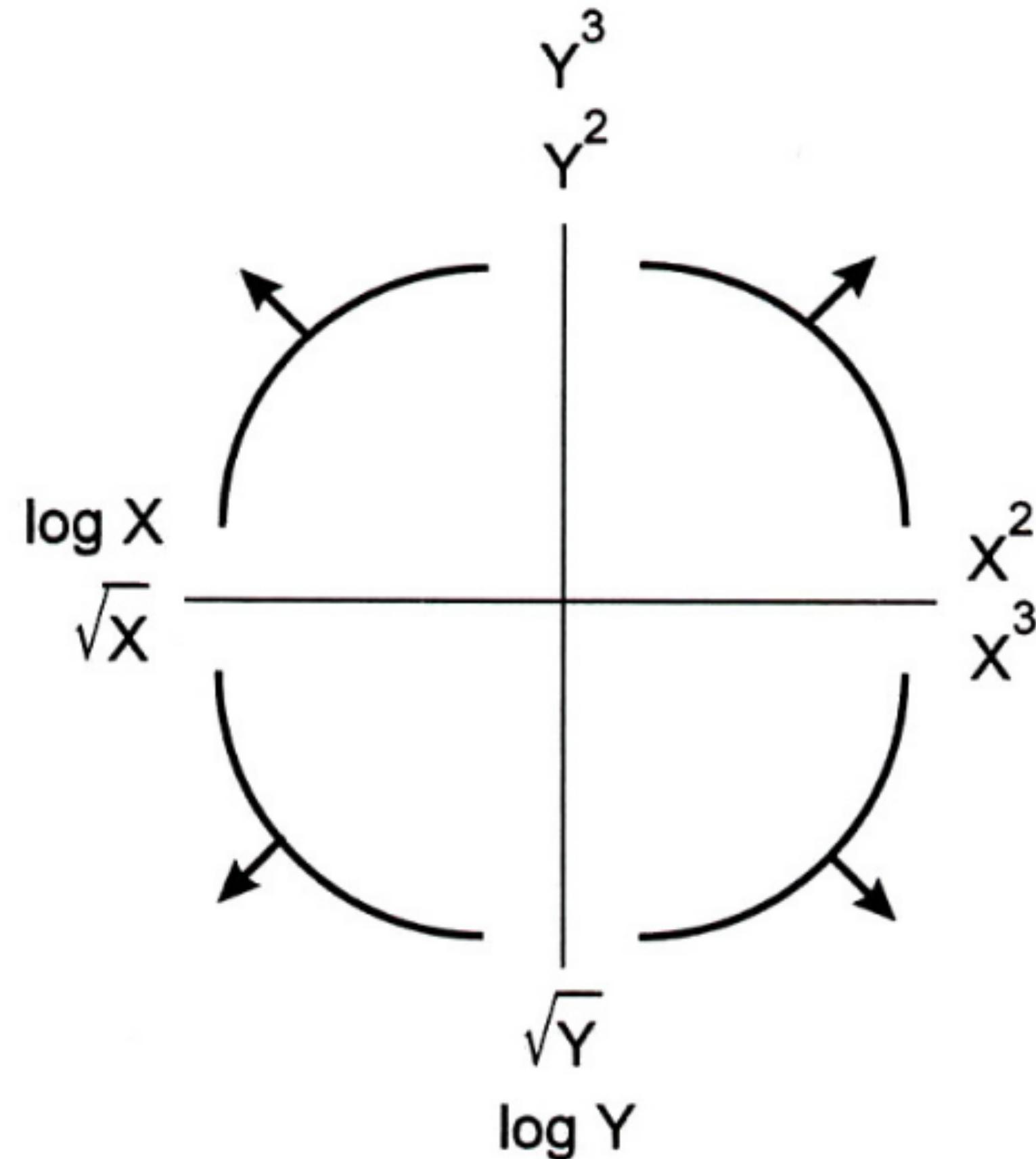


Heavy density at large y:
Power transform y-coordinates.

PRINCIPLES OF DATA VISUALIZATION

Things to Remember

Transformations: better visualize/understand relationship bw variables



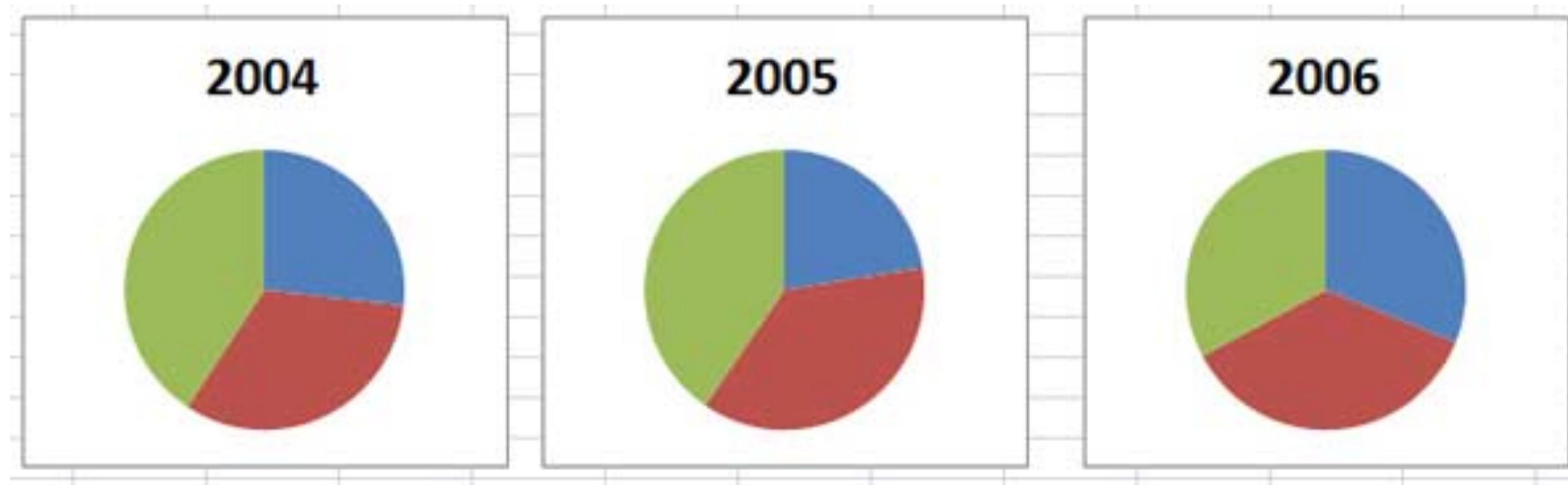
Use:

Tukey-Mosteller Bulge Diagram
as a guide for finding suitable
transformation.

PRINCIPLES OF DATA VISUALIZATION

Things to Remember

Pie charts vs bar plots:

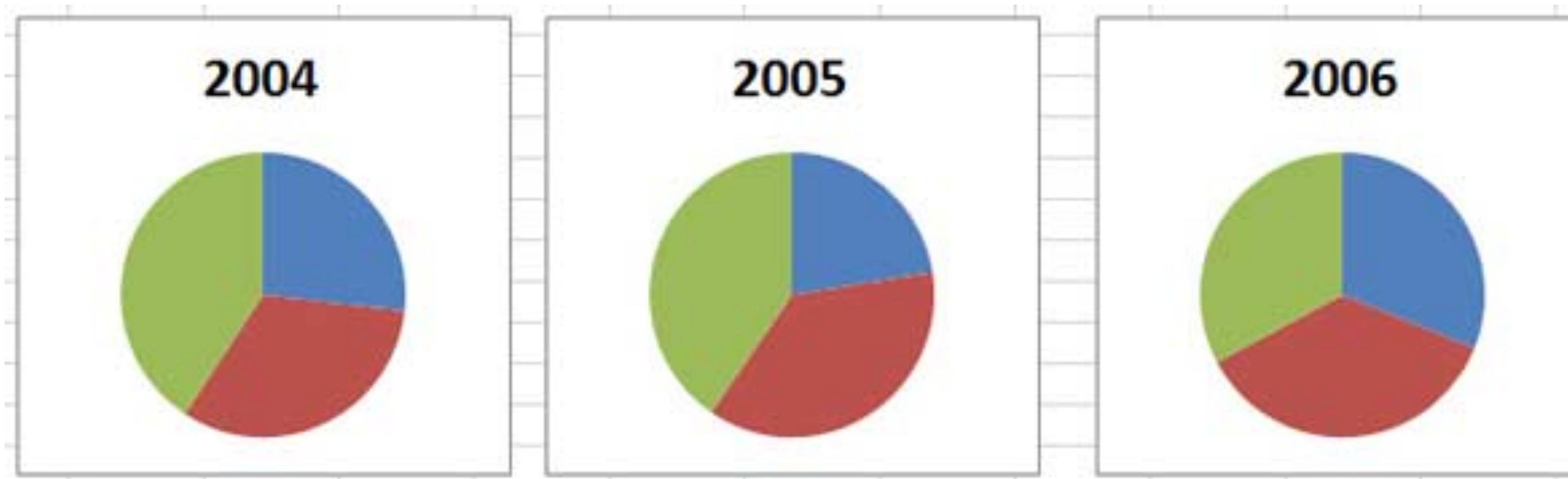


Hard to read & see differences
Maybe good for percentages

PRINCIPLES OF DATA VISUALIZATION

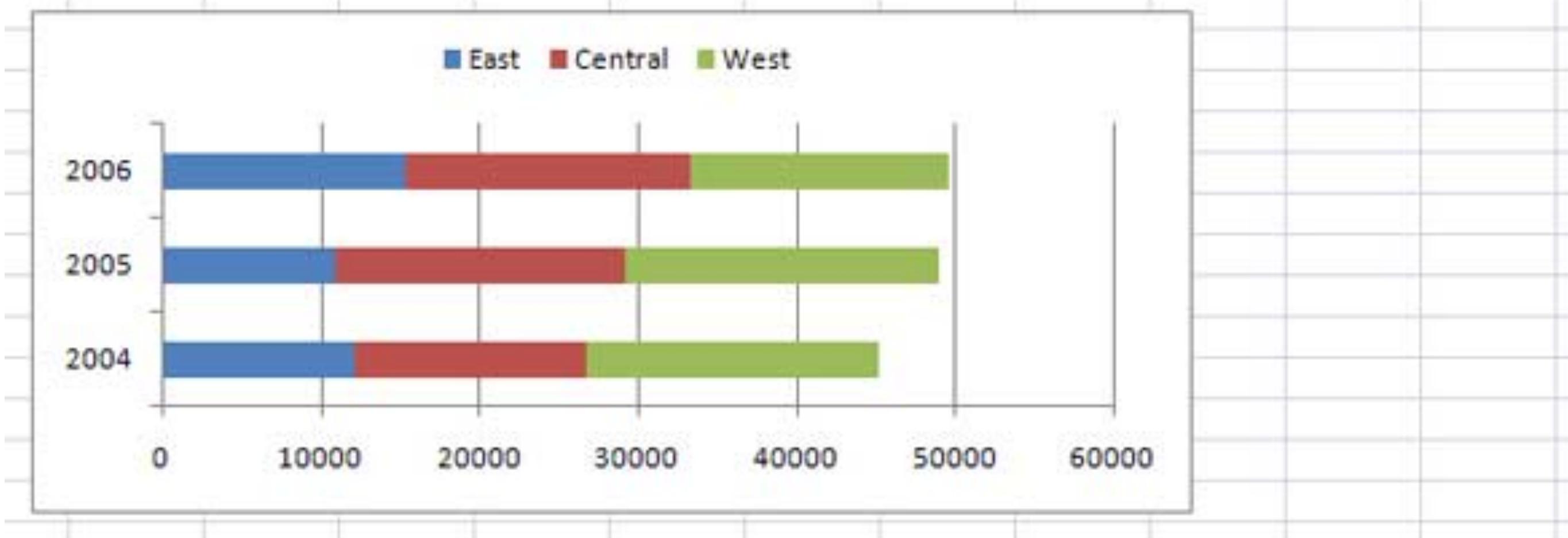
Things to Remember

Pie charts vs bar plots:



Hard to read & see differences

Maybe good for percentages

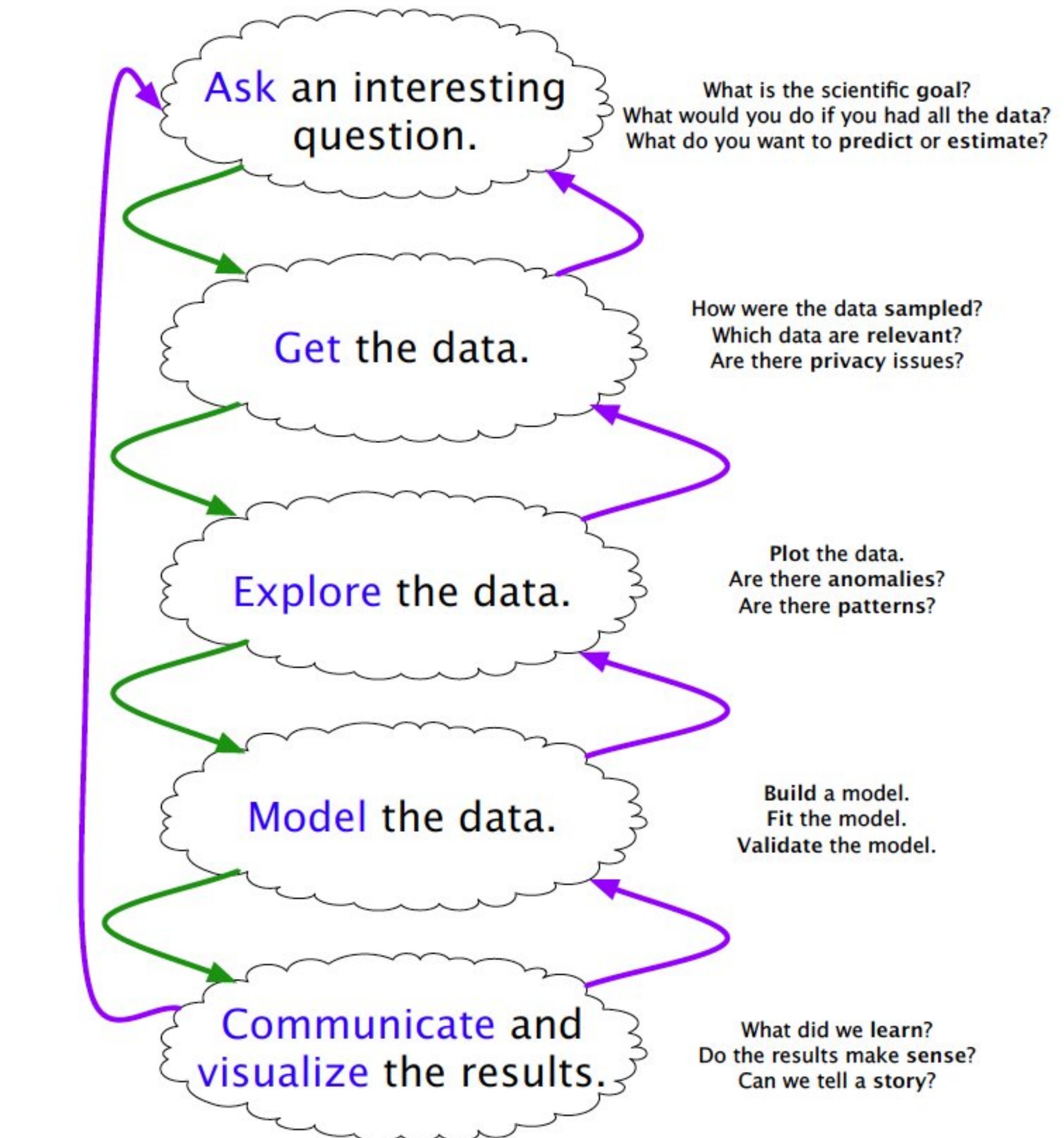


Lengths are easier to distinguish than angles.

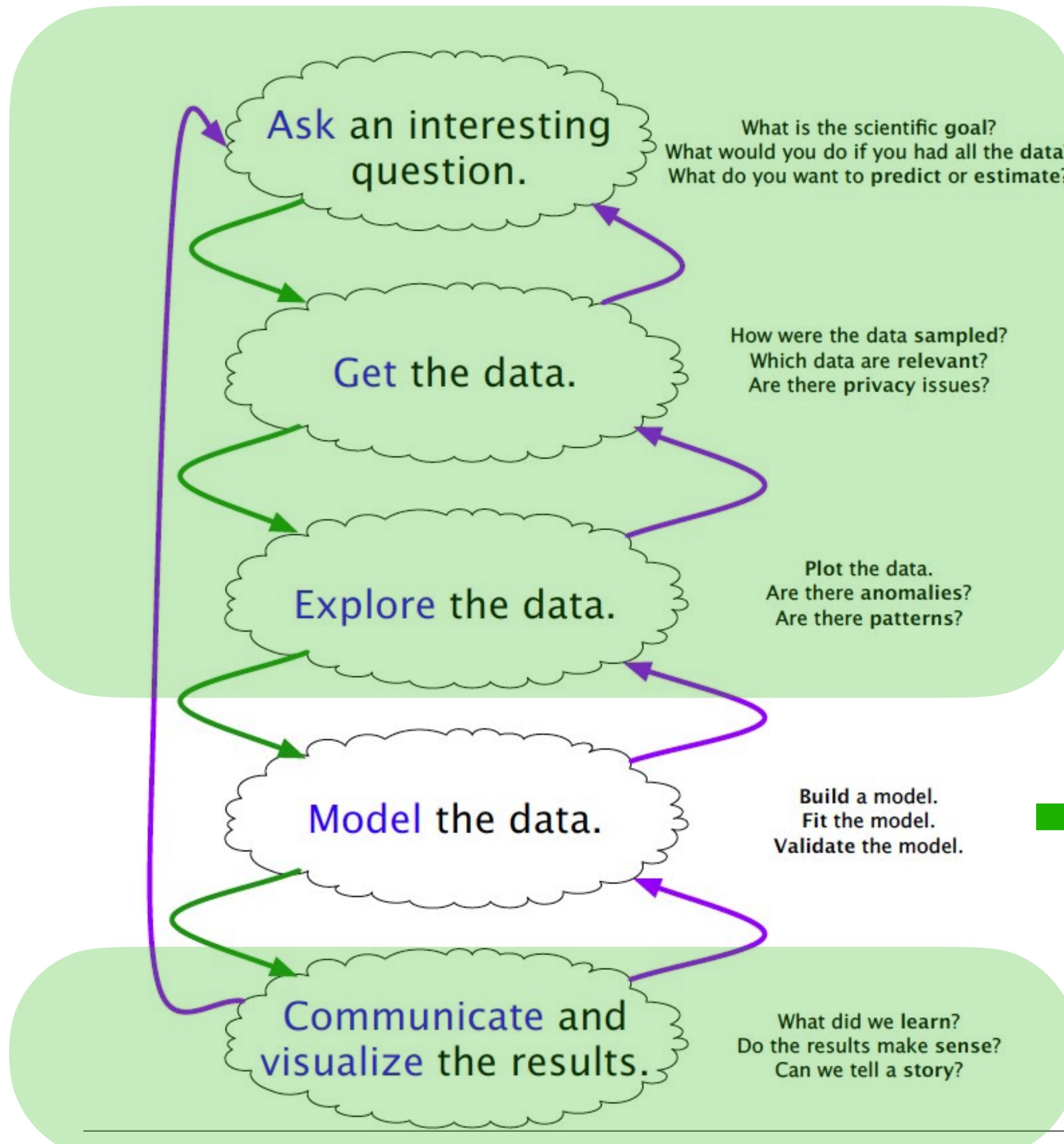
Bar plots better for comparisons, detecting differences.

MODELING

ROLE OF MODELING IN DATA SCIENCE



ROLE OF MODELING IN DATA SCIENCE



Covered

For most of the remainder
of the semester

Covered

WHAT IS A MODEL?

Modeling: Process of encapsulating information into a tool which can make forecasts/predictions.

A model is an **idealized representation** of a system.

“All models are wrong, but some models are useful.”

– George Box (1919-2013)

First-Principle vs. Data-Driven Models

Newton’s Law of Gravitation vs predicting flight cancellations

PRINCIPLES OF MODELING

Occam's Razor: The simplest explanation is the best explanation

Minimize the number of model parameters.

Inherent trade-off between accuracy and simplicity

Bias-Variance Tradeoff:

Bias: Error from incorrect assumptions built into the model

E.g. interpolating function linear rather than a higher-order curve.

(Underfitting)

Variance: Error from sensitivity to small fluctuations in the training set

(Overfitting)

First-principle models may have bias, data-driven models may overfit.

MODELING PIPELINE

1. Choose a model

How should we represent the world?

2. Choose a loss function

How do we quantify prediction error?

3. Fit the model

How do we choose the best parameters of our model given our data?

4. Evaluate model performance

How do we evaluate whether this process gave rise to a good model?

CLASSIFICATION VS PREDICTION

Features				Output
x_{11}	x_{12}	\dots	x_{1p}	y_1
x_{21}	x_{22}	\dots	x_{2p}	y_2
\vdots	\vdots	\ddots	\vdots	\vdots
x_{n1}	x_{n2}	\dots	x_{np}	y_n

Classification:
Output is a categorical
variable

Prediction:
Output is a numerical
variable

CLASSIFICATION VS PREDICTION

Features				Output
x_{11}	x_{12}	\dots	x_{1p}	y_1
x_{21}	x_{22}	\dots	x_{2p}	y_2
\vdots	\vdots	\ddots	\vdots	\vdots
x_{n1}	x_{n2}	\dots	x_{np}	y_n

Classification:
Output is a categorical variable

Prediction:
Output is a numerical variable

Other names:	Independent variables Explanatory variables Predictors Inputs ...	Outcome Response Dependent variable ...
--------------	---	--

PROJECT IS OUT: DUE DECEMBER 10

Suggestions:

Read the project guidelines carefully.

Find project partners soon (next few days).

Start with the “data” portion **now** (spend 2-3 weeks):

- Spend a day to see if other relevant datasets in CTDH (Data Hub) or somewhere else
- Get datasets, turn them into data frames, do EDA/visualizations
- Decide on interesting questions and tasks (prediction/classification)
- You may need to do a lot of data cleaning/preprocessing:
 - Merging datasets (unify location etc.), normalizations etc.
 - Retain only data you need, handle missing values etc.
 - If classification: form new relevant columns (e.g. crime index: 1-5)

Start with the “modeling” portion **in 10-15 days** (spend 2-3 weeks):

- Choose your algorithms (at least 2), Partition data (train/test split, CV etc.)
- Do the evaluations. Get resulting plots.

Start writing the report **one week before** the deadline:

- Include your findings from the “data” portion (statistics/correlations/nice visualizations etc.)
- Include your findings from modeling (tables/plots for presenting results, discussions etc.)

MODELING PIPELINE

1. Choose a model

How should we represent the world?

2. Choose a loss function

How do we quantify prediction error?

3. Fit the model

How do we choose the best parameters of our model given our data?

4. Evaluate model performance

How do we evaluate whether this process gave rise to a good model?

MODELING PIPELINE

1. Choose a model

How should we represent the world?

2. Choose a loss function

How do we quantify prediction error?

3. Fit the model

How do we choose the best parameters of our model given our data?

4. Evaluate model performance

How do we evaluate whether this process gave rise to a good model?

Let's start from the end and first focus here.

EVALUATING MODEL PERFORMANCE

What to compare against?

Baseline Models: Simplest reasonable models to compare against

EVALUATING MODEL PERFORMANCE

Possible Baselines for Classification:

EVALUATING MODEL PERFORMANCE

Possible Baselines for Classification:

- Uniform classifier: assign one label consistently.
e.g. assign “yes” for any query in binary classification (yes or no)

EVALUATING MODEL PERFORMANCE

Possible Baselines for Classification:

- Uniform classifier: assign one label consistently.

e.g. assign “yes” for any query in binary classification (yes or no)

Features				Output
x_{11}	x_{12}	\dots	x_{1p}	Yes
x_{21}	x_{22}	\dots	x_{2p}	No
\vdots	\vdots	\ddots	\vdots	\vdots
x_{n1}	x_{n2}	\dots	x_{np}	Yes

Don't care about
the training data.

Assign ‘Yes’ to
each row’s label.

EVALUATING MODEL PERFORMANCE

Possible Baselines for Classification:

- Blind classifier: assign random labels or classes.
e.g. assign “yes” or “no” randomly for each row.

EVALUATING MODEL PERFORMANCE

Possible Baselines for Classification:

- Blind classifier: assign random labels or classes.
e.g. assign “yes” or “no” randomly for each row.

Features				Output
x_{11}	x_{12}	\dots	x_{1p}	Yes
x_{21}	x_{22}	\dots	x_{2p}	No
\vdots	\vdots	\ddots	\vdots	\vdots
x_{n1}	x_{n2}	\dots	x_{np}	Yes

Don't care about
the training data.

Assign ‘Yes’ or
‘No’ randomly to
each row’s label.

EVALUATING MODEL PERFORMANCE

Possible Baselines for Classification:

- The most common label in the training data.

EVALUATING MODEL PERFORMANCE

Possible Baselines for Classification:

- The most common label in the training data.

Features				Output
x_{11}	x_{12}	\dots	x_{1p}	Yes
x_{21}	x_{22}	\dots	x_{2p}	No
\vdots	\vdots	\ddots	\vdots	\vdots
x_{n1}	x_{n2}	\dots	x_{np}	No

Find most common label. Say it is ‘No’

Assign ‘No’ to each row’s label

EVALUATING MODEL PERFORMANCE

Possible Baselines for Classification:

- Other baselines:
 - The best performing single-variable model.
 - Same label as the previous point in time:
Time series forecasting. e.g. did it rain yesterday?
 - A model built by someone else

EVALUATING MODEL PERFORMANCE

Possible Baselines for Prediction (Regression):

EVALUATING MODEL PERFORMANCE

Possible Baselines for Prediction (Regression):

- Mean or median of the target in training data

EVALUATING MODEL PERFORMANCE

Possible Baselines for Prediction (Regression):

- Mean or median of the target in training data

Features				Output
x_{11}	x_{12}	\dots	x_{1p}	3.2
x_{21}	x_{22}	\dots	x_{2p}	4.0
\vdots	\vdots	\ddots	\vdots	\vdots
x_{n1}	x_{n2}	\dots	x_{np}	3.6

Training
data

Test
data

Find mean or median

Ignore features,
assign mean or
median to each row

EVALUATING MODEL PERFORMANCE

Possible Baselines for Prediction (Regression):

- Other baselines:
 - Linear Regression: Details next week.
 - Same value as the previous point in time:
Time series data: e.g. What was temperature yesterday?

EVALUATING CLASSIFIERS

Confusion matrix (contingency table) in a binary classifier:

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

True positives (TP): + is labeled +

True negative (TN): - is labeled -

False positives (FP): - is labeled +

False negatives (FN): + is labeled -

EVALUATING CLASSIFIERS

Confusion matrix (contingency table) in a binary classifier:

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

True positives (TP): + is labeled +

True negative (TN): - is labeled -

False positives (FP): - is labeled +

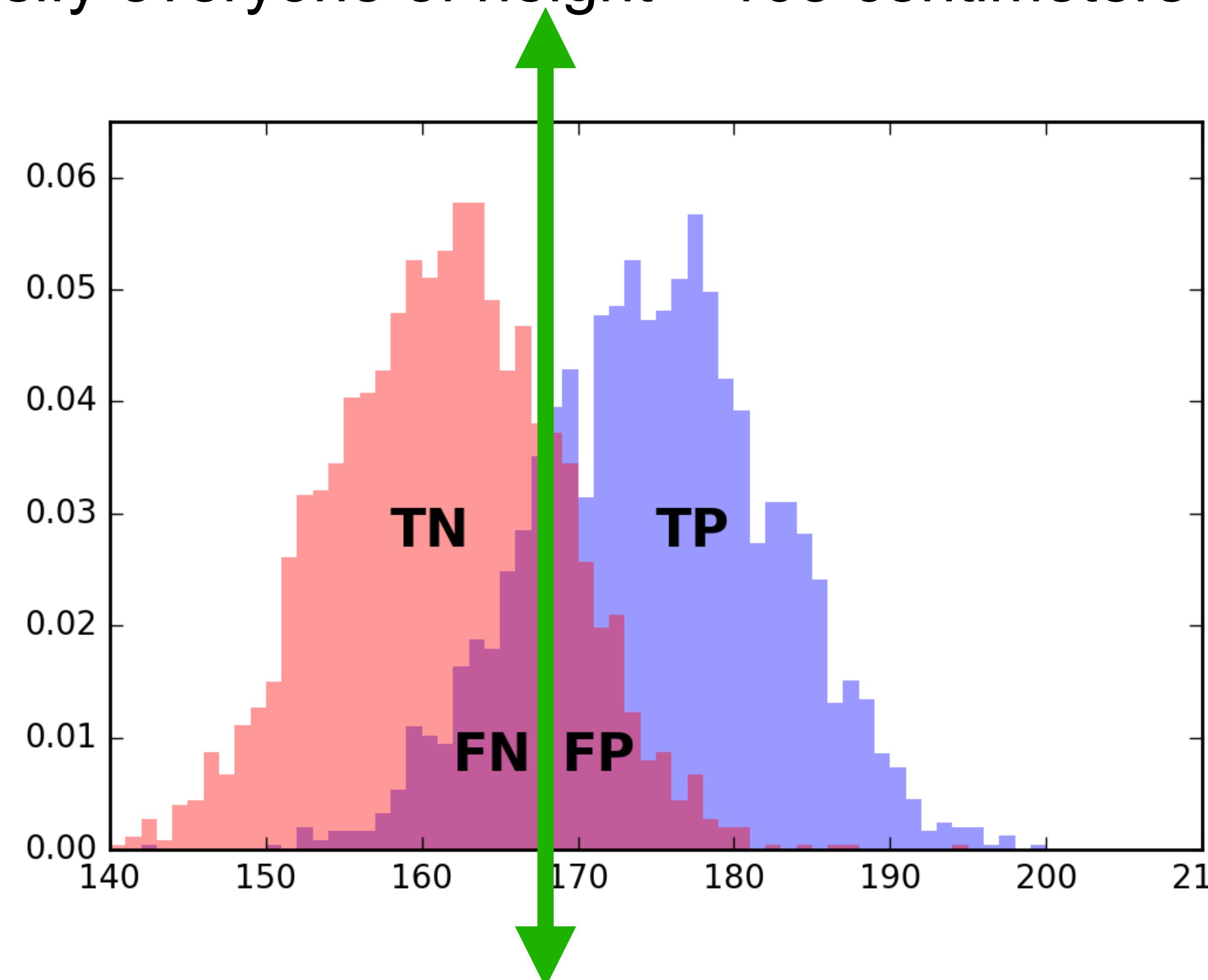
False negatives (FN): + is labeled -

Choose more interesting class as **positive**.

Example: Spam classification \Rightarrow Spam positive, non-spam negative.

EVALUATING CLASSIFIERS

Example: Classify everyone of height ≥ 168 centimeters as male



MEASURES FOR EVALUATING CLASSIFIERS

Accuracy: Ratio of correct predictions over all predictions.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

MEASURES FOR EVALUATING CLASSIFIERS

Accuracy: Ratio of correct predictions over all predictions.

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

What happens when the classes are balanced?

Blind classifier accuracy: 50 %

Most common label classifier accuracy: 50 %

MEASURES FOR EVALUATING CLASSIFIERS

Accuracy: Ratio of correct predictions over all predictions.

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

What happens when the classes are balanced?

Blind classifier accuracy: 50 %

Most common label classifier accuracy: 50 %

What happens when the classes are unbalanced?

Example: Cancer diagnosis. Positive class (cancer) 5 % .

Blind classifier accuracy: 50 %

Most common label classifier accuracy: 95 %

MEASURES FOR EVALUATING CLASSIFIERS

Accuracy: Ratio of correct predictions over all predictions.

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Accuracy can be misleading if class imbalance.

We need measures sensitive to getting more TP.

MEASURES FOR EVALUATING CLASSIFIERS

Precision: Ratio of correct positive predictions over positive predictions.

$$precision = \frac{TP}{TP + FP}$$

MEASURES FOR EVALUATING CLASSIFIERS

Precision: Ratio of correct positive predictions over positive predictions.

$$precision = \frac{TP}{TP + FP}$$

The case with unbalanced classes (problematic for accuracy)

Example: Cancer diagnosis. Positive class (cancer) 5 % .

Blind classifier precision: 5 %

Most common label classifier precision: undefined

Least common label classifier precision: 5 %

MEASURES FOR EVALUATING CLASSIFIERS

Recall: Ratio of correct positive predictions over all positive instances.

$$\text{recall} = \frac{TP}{P} = \frac{TP}{TP + FN}$$

MEASURES FOR EVALUATING CLASSIFIERS

Recall: Ratio of correct positive predictions over all positive instances.

$$\text{recall} = \frac{TP}{P} = \frac{TP}{TP + FN}$$

The case with unbalanced classes (problematic for accuracy)

Example: Cancer diagnosis. Positive class (cancer) 5 % .

Blind classifier recall: 50 %

Most common label classifier recall: 0 %

Least common label classifier recall: 100 %

MEASURES FOR EVALUATING CLASSIFIERS

Precision vs Recall:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{P} = \frac{TP}{TP + FN}$$

MEASURES FOR EVALUATING CLASSIFIERS

Precision vs Recall:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{P} = \frac{TP}{TP + FN}$$

Are False Positives more important or False Negatives?

Preliminary diagnoses tests for cancer:

Might be ok to have FPs but not ok to have FNs.

Spam detection:

Might be ok to have FNs but not ok to have FPs.

MEASURES FOR EVALUATING CLASSIFIERS

F-score: Harmonic mean of precision and recall.

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

For a good F-score both precision and recall must be good.

MEASURES FOR EVALUATING CLASSIFIERS

Summary:

- Class sizes substantially different \Rightarrow accuracy is misleading.
- High precision is hard to achieve in unbalanced class sizes:
If classifier makes many positive predictions, most will miss the mark
If classifier makes few positive predictions, few will catch rare positives
- F-score does the best job of any single statistic, but all four work together to describe the performance of a classifier.

MEASURES FOR EVALUATING CLASSIFIERS

ROC (Receiver Operating Characteristic) Curve:

Class prediction usually via some ‘score’ reflecting **in classness**.

Ex: Say height is our score for classifying female (+), male (-):

MEASURES FOR EVALUATING CLASSIFIERS

ROC (Receiver Operating Characteristic) Curve:

Class prediction usually via some ‘score’ reflecting **in classness**.

Ex: Say height is our score for classifying female (+), male (-):

$F, F, M, F, \dots, F, M, M$

Increasing height

MEASURES FOR EVALUATING CLASSIFIERS

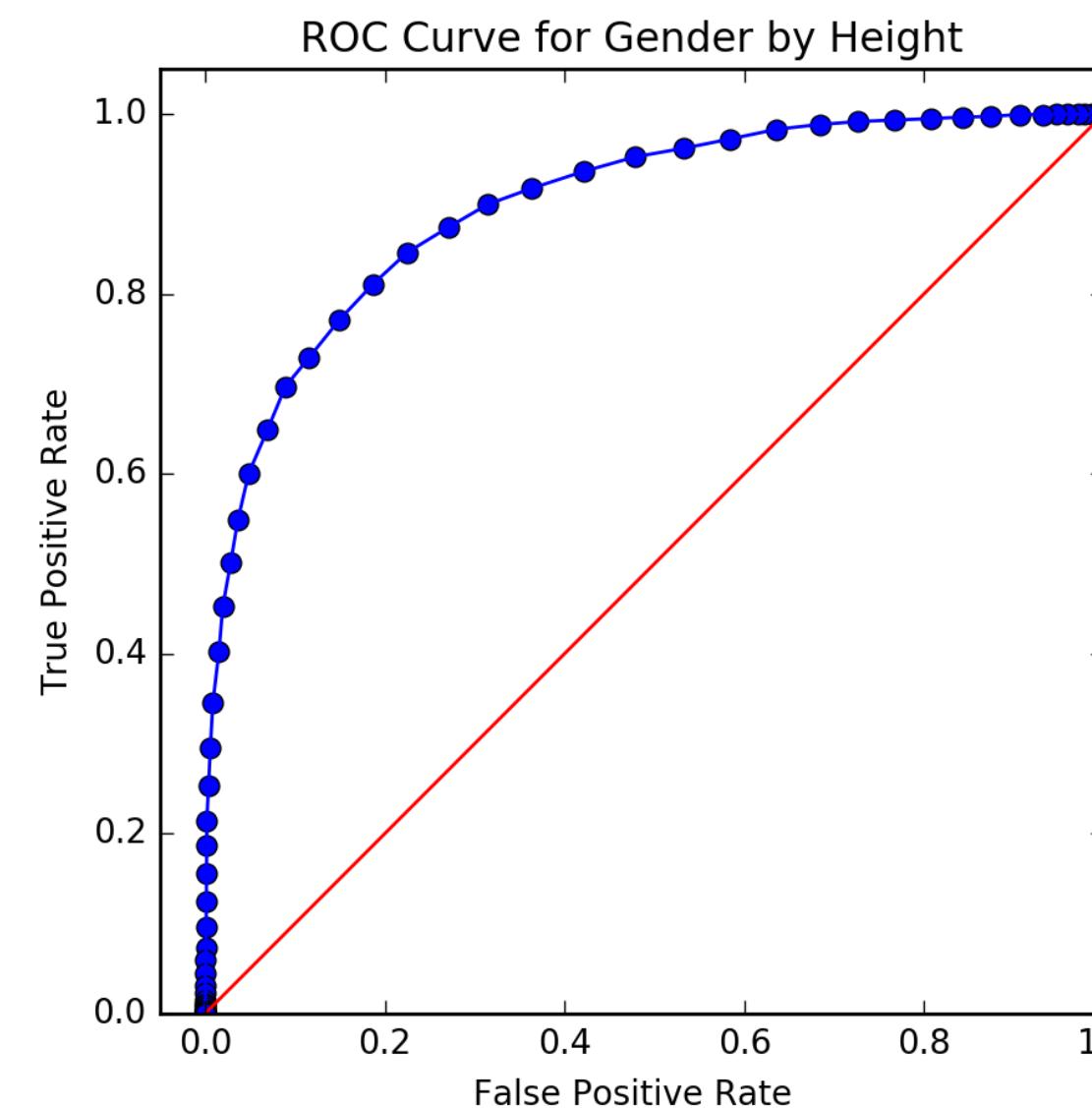
ROC (Receiver Operating Characteristic) Curve:

Class prediction usually via some ‘score’ reflecting **in classness**.

Ex: Say height is our score for classifying female (+), male (-):

$F, F, M, F, \dots, F, M, M$

Increasing height →



For increasing values of height threshold:

$$TPR = Recall = \frac{TP}{P} \quad FPR = \frac{FP}{N}$$

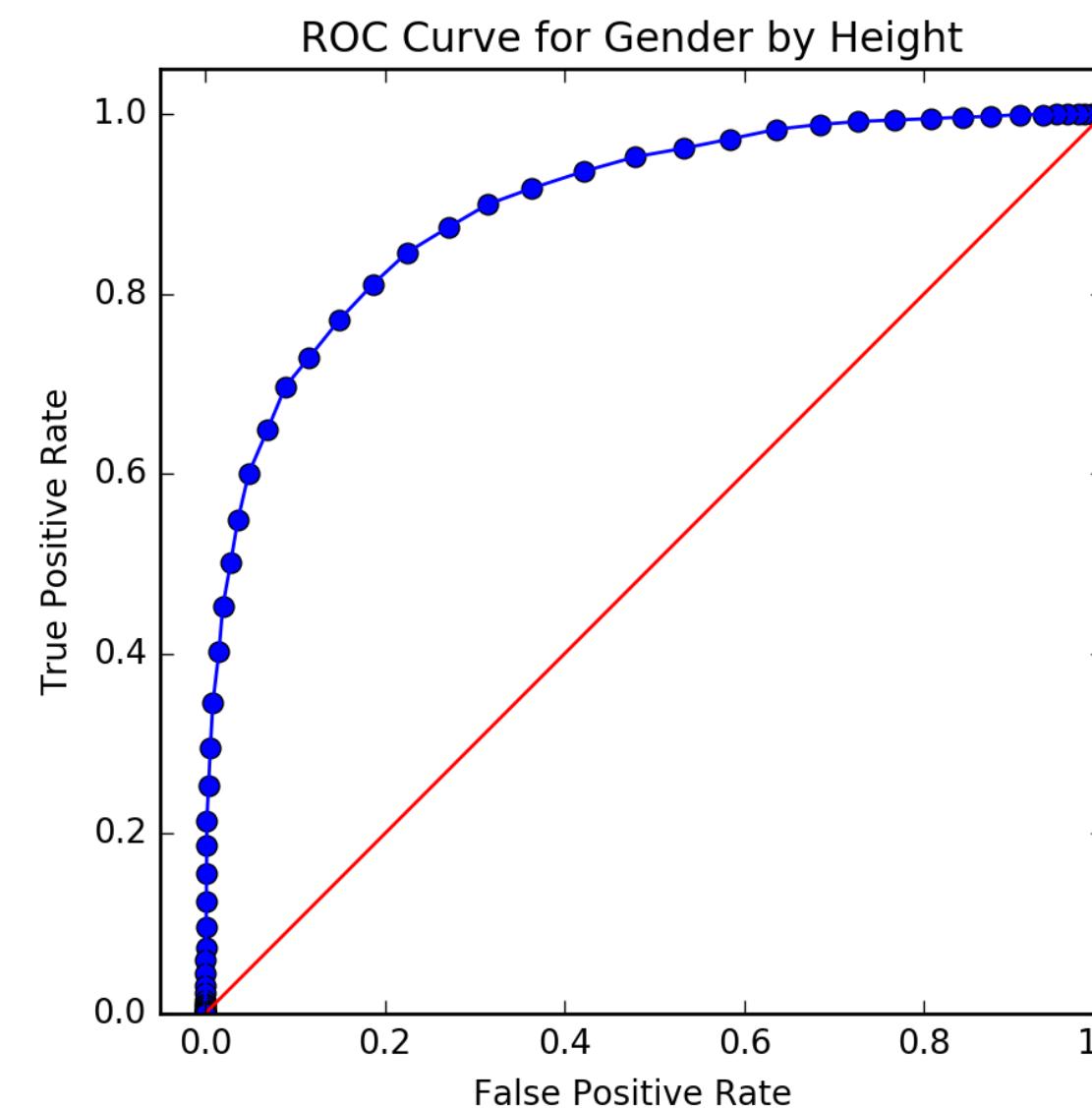
MEASURES FOR EVALUATING CLASSIFIERS

ROC (Receiver Operating Characteristic) Curve:

Class prediction usually via some ‘score’ reflecting **in classness**.

Ex: Say height is our score for classifying female (+), male (-):

$F, F, M, F, \dots, F, M, M$
→
Increasing height



Area under ROC (AUROC):
How good is the score defining the classification.
At most 1.

For increasing values of height threshold:

$$TPR = Recall = \frac{TP}{P} \quad FPR = \frac{FP}{N}$$

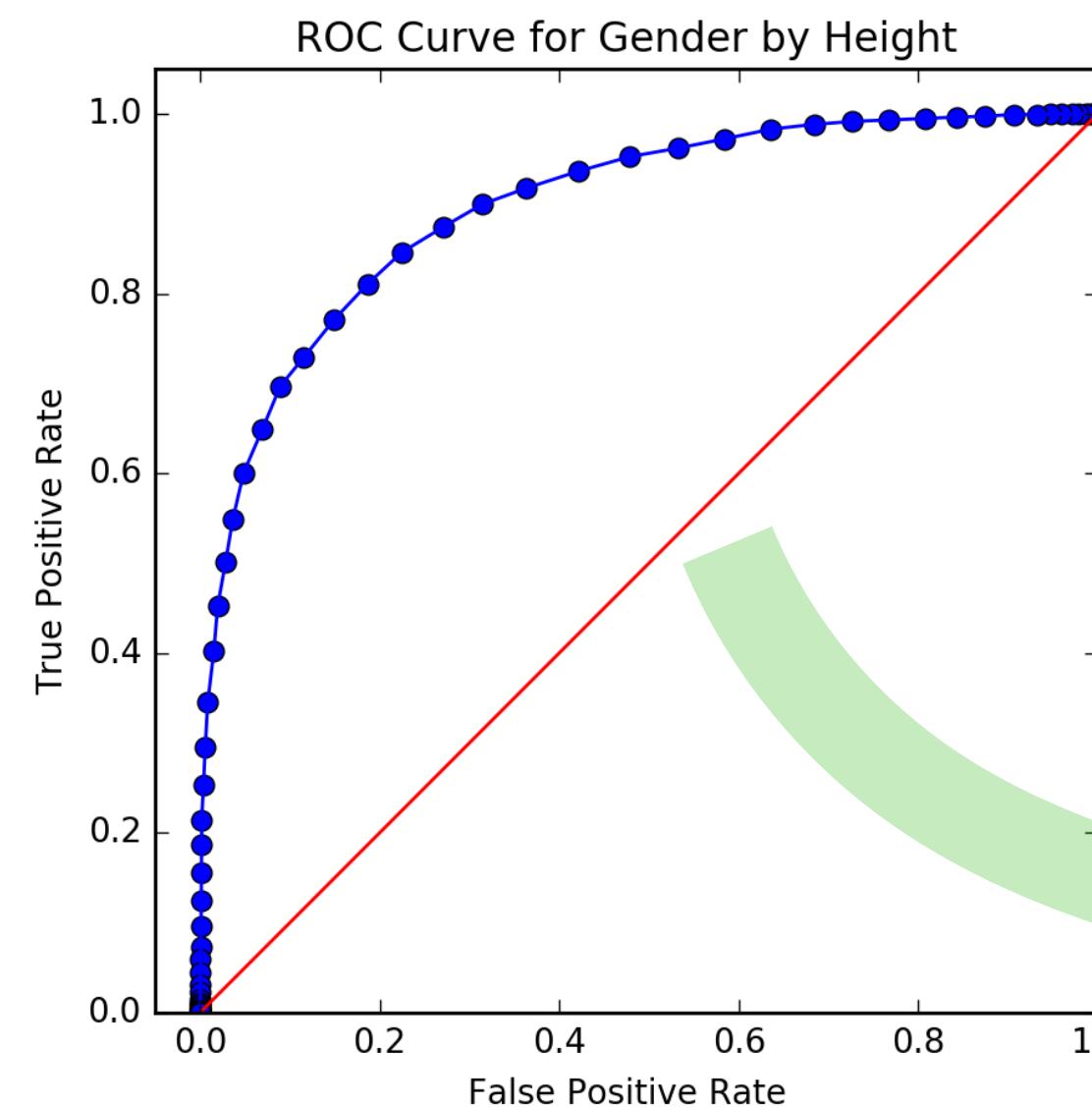
MEASURES FOR EVALUATING CLASSIFIERS

ROC (Receiver Operating Characteristic) Curve:

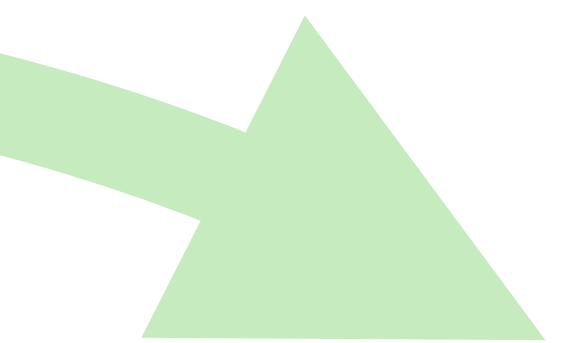
Class prediction usually via some ‘score’ reflecting **in classness**.

Ex: Say height is our score for classifying female (+), male (-):

$F, F, M, F, \dots, F, M, M$
→
Increasing height



Area under ROC (AUROC):
How good is the score defining the classification.
At most 1.



For increasing values of height threshold:

$$TPR = Recall = \frac{TP}{P}$$

$$FPR = \frac{FP}{N}$$

Diagonal: Blind classifier.

REVIEW OF PREVIOUS LECTURE

Measures for evaluating classifiers:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$F = 2 \times \frac{precision \times recall}{precision + recall}$$

To evaluate a ‘score’ for in-classness:

Compute AUROC: ROC curve plots TPR, FPR for each threshold.

MEASURES FOR EVALUATING CLASSIFIERS

Multiclass systems: Accuracy easy! How to generalize precision/recall?

Ex: Confusion matrix for OCR of single digits

Actual Digits	Predicted									
	0	1	2	3	4	5	6	7	8	9
0	351	0	5	4	2	7	2	1	6	0
1	0	254	0	0	2	0	0	1	1	2
2	1	1	166	4	5	1	3	2	2	1
3	1	2	4	142	0	5	0	1	4	0
4	3	3	8	1	180	3	2	5	4	4
5	0	0	3	11	0	140	3	0	7	1
6	0	2	2	0	4	0	158	0	1	0
7	0	0	2	2	1	0	0	132	2	1
8	2	1	8	0	0	0	2	1	137	1
9	1	1	0	2	6	4	0	4	2	167

MEASURES FOR EVALUATING CLASSIFIERS

Multiclass systems: Accuracy easy! How to generalize precision/recall?

Ex: Confusion matrix for OCR of single digits

Actual Digits	Predicted									
	0	1	2	3	4	5	6	7	8	9
0	351	0	5	4	2	7	2	1	6	0
1	0	254	0	0	2	0	0	1	1	2
2	1	1	166	4	5	1	3	2	2	1
3	1	2	4	142	0	5	0	1	4	0
4	3	3	8	1	180	3	2	5	4	4
5	0	0	3	11	0	140	3	0	7	1
6	0	2	2	0	4	0	158	0	1	0
7	0	0	2	2	1	0	0	132	2	1
8	2	1	8	0	0	0	2	1	137	1
9	1	1	0	2	6	4	0	4	2	167

$$\text{precision}_i = C[i, i] / \sum_{j=1}^d C[j, i]$$

E.g. $\text{precision}_0 = 351/359$

$$\text{recall}_i = C[i, i] / \sum_{j=1}^d C[i, j]$$

E.g. $\text{recall}_0 = 351/378$

EVALUATING VALUE PREDICTION MODELS

Features				Output
x_{11}	x_{12}	\dots	x_{1p}	3.2
x_{21}	x_{22}	\dots	x_{2p}	4.0
\vdots	\vdots	\ddots	\vdots	\vdots
x_{n1}	x_{n2}	\dots	x_{np}	3.5

Prediction:
Output is a numerical
variable

We use x to predict y . Let the predictions be $\hat{y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}$.

EVALUATING VALUE PREDICTION MODELS

Actual values $y = \{y_1, y_2, \dots, y_n\}$. Predictions $\hat{y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}$.

Absolute Error: $\Delta = \hat{y}_i - y_i$

The sign distinguishes $\hat{y} > y$ from $\hat{y} < y$.

Can't aggregate directly (offsetting errors with different signs).

EVALUATING VALUE PREDICTION MODELS

Actual values $y = \{y_1, y_2, \dots, y_n\}$. Predictions $\hat{y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}$.

Absolute Error: $\Delta = \hat{y}_i - y_i$

The sign distinguishes $\hat{y} > y$ from $\hat{y} < y$.

Can't aggregate directly (offsetting errors with different signs).

Relative Error: $(\hat{y}_i - y_i)/y_i$

Unit-less quantity as compared to absolute error.

EVALUATING VALUE PREDICTION MODELS

Actual values $y = \{y_1, y_2, \dots, y_n\}$. Predictions $\hat{y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}$.

Absolute Error: $\Delta_i = \hat{y}_i - y_i$

The sign distinguishes $\hat{y} > y$ from $\hat{y} < y$.

Can't aggregate directly (offsetting errors with different signs).

Relative Error: $(\hat{y}_i - y_i)/y_i$

Unit-less quantity as compared to absolute error.

Squared Error: $\Delta^2 = (\hat{y}_i - y_i)^2$

Values can be meaningfully summed.

Large errors contribute more \Rightarrow outliers may dominate error.

EVALUATING VALUE PREDICTION MODELS

Good idea to plot the distribution of absolute error values.

What is expected from a good predictor?

Symmetric distribution centered around zero.

Bell-shaped (small errors more common than large ones)

Extreme outliers rare.

EVALUATING VALUE PREDICTION MODELS

Good idea to plot the distribution of absolute error values.

What is expected from a good predictor?

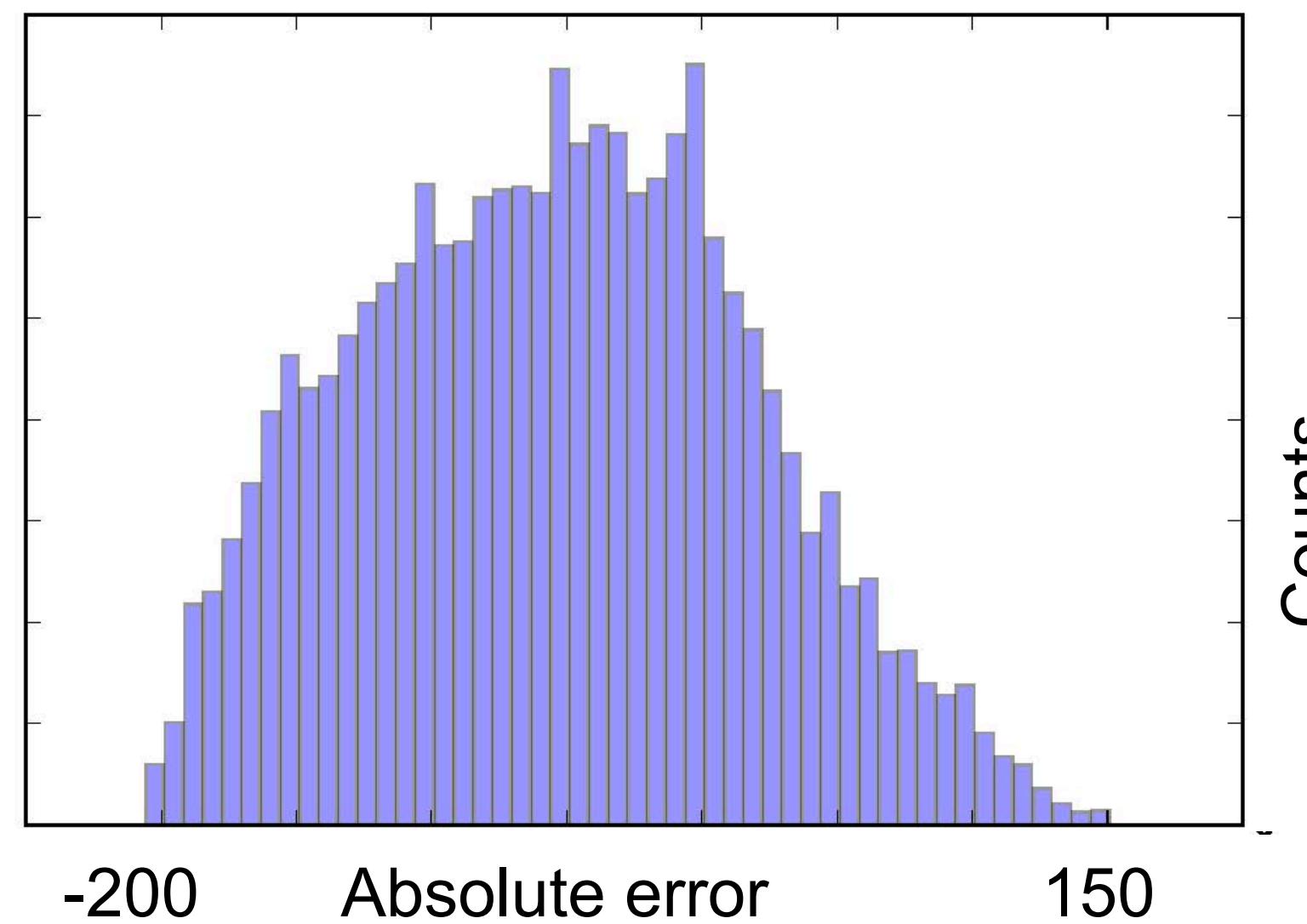
Symmetric distribution centered around zero.

Bell-shaped (small errors more common than large ones)

Extreme outliers rare.

Ex: Document date (1800-2005) predictor from word usage.

Random
model



Why is the left tail fat?

EVALUATING VALUE PREDICTION MODELS

Good idea to plot the distribution of absolute error values.

What is expected from a good predictor?

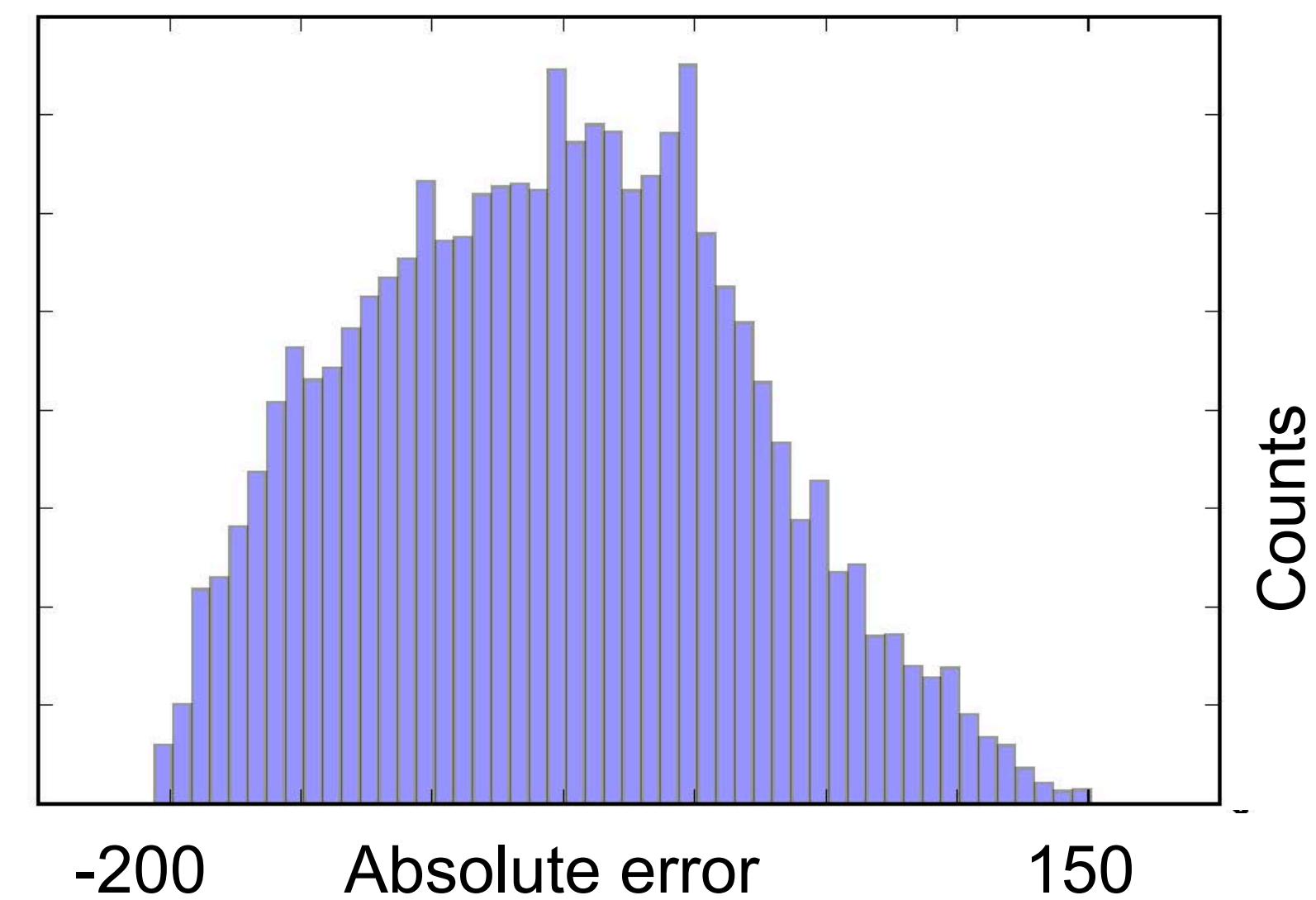
Symmetric distribution centered around zero.

Bell-shaped (small errors more common than large ones)

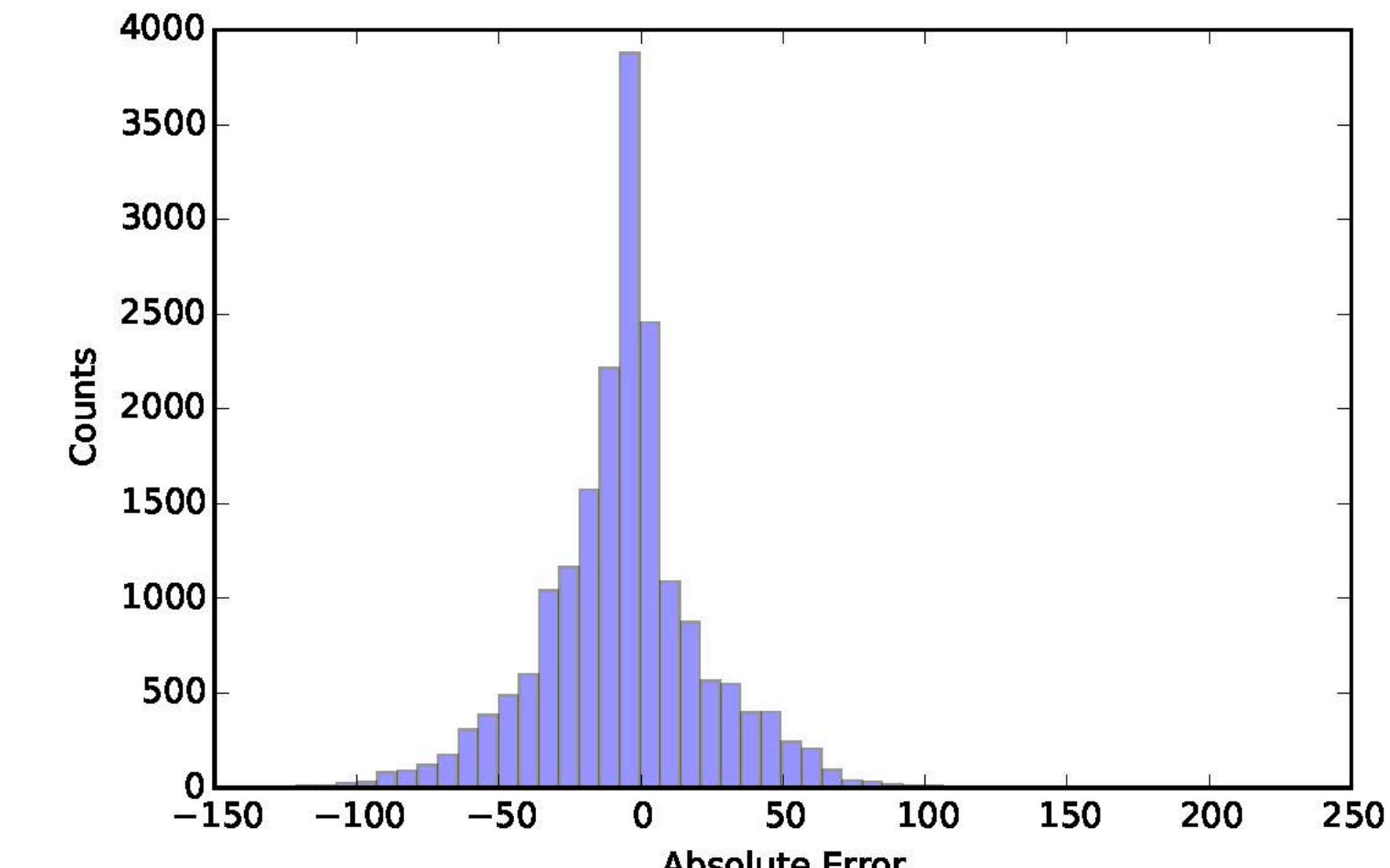
Extreme outliers rare.

Ex: Document date (1800-2005) predictor from word usage.

Random
model



Naive
Bayes
model



EVALUATING VALUE PREDICTION MODELS

We need aggregate summary statistics as well.

Mean Squared Error: $MSE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$

Outliers have disproportionate effect.

Median squared error better for noisy instances.

Root Mean Squared Error: $RMSD(y, \hat{y}) = \sqrt{MSE(y, \hat{y})}$

Magnitude same scale as original values.

Outliers may still be problematic.

PARTITIONING THE DATA

The best way to assess models involve out-of-sample predictions.

PARTITIONING THE DATA

The best way to assess models involve **out-of-sample predictions**.

Partition the data:



80 % training set

Study the domain and determine parameters of the model.

20 % test set

Never leaks into the model until the end.

TRAINING ERROR VS VALIDATION ERROR

```
from sklearn.datasets import load_iris
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

iris = load_iris()

X = iris.data
y = iris.target

clf = RandomForestClassifier(n_estimators=2, random_state=0)
# X is the training data
clf.fit(X, y)

# Training error (train and validate on the same data)
y_pred = clf.predict(X)
acc = accuracy_score(y, y_pred)

print(f'Accuracy: {acc:.2f}')
```

Evaluating on
training data in
terms of accuracy

If performance is
not good ⇒
underfitting

Accuracy: 0.97

TRAINING ERROR VS VALIDATION ERROR

```
from sklearn.model_selection import train_test_split  
  
# split into training/validation sets: 70-30  
# random_state set to an int for reproducibility  
X_train, X_validate, y_train, y_validate = \  
    train_test_split(X, y, train_size=0.7, random_state=0)  
  
clf = RandomForestClassifier(n_estimators=2, random_state=0)  
clf.fit(X_train, y_train)  
  
# validate with unseen data  
y_pred = clf.predict(X_validate)  
acc = accuracy_score(y_validate, y_pred)  
  
print(f'Accuracy: {acc:.2f}')  
  
Accuracy: 0.91
```

Split into training and test sets.
Evaluate on validation data in terms of accuracy

If performance is good on training data but not on test data ⇒ overfitting

K-FOLD CROSS VALIDATION

By random chance, may have selected a validation set that was not representative of other unseen data that the model might encounter.

Validation error from training/validation split #1: 0.82

Validation error from training/validation split #2: 0.97

...

K-FOLD CROSS VALIDATION

By random chance, may have selected a validation set that was not representative of other unseen data that the model might encounter.

Validation error from training/validation split #1: 0.82

Validation error from training/validation split #2: 0.97

...

Ideally: Assess model performance on several different validation sets before touching the test set.

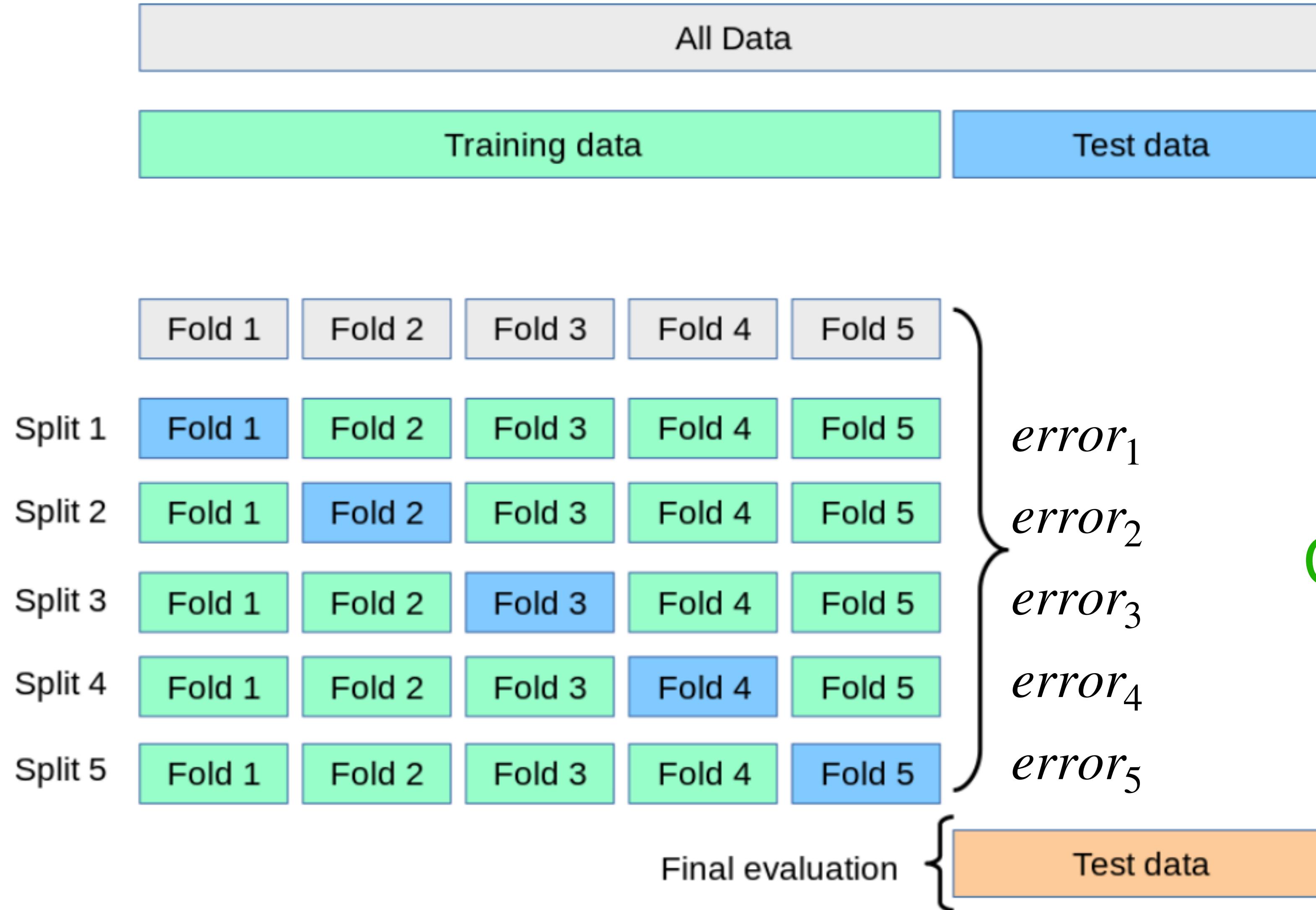
Especially useful for:

Small datasets.

Hyperparameter tuning.

K-FOLD CROSS VALIDATION

Example: 5-fold cross validation



$$CV \text{ error} = 1/5 \sum_{i=1}^5 error_i$$

K-FOLD CROSS VALIDATION: EXAMPLE

```
from sklearn.model_selection import cross_validate
cv = cross_validate(clf, X, y, scoring='accuracy', cv=3)
cv["test_score"]

array([0.98, 0.92, 0.96])
```

3-fold cross validation, find error (accuracy).

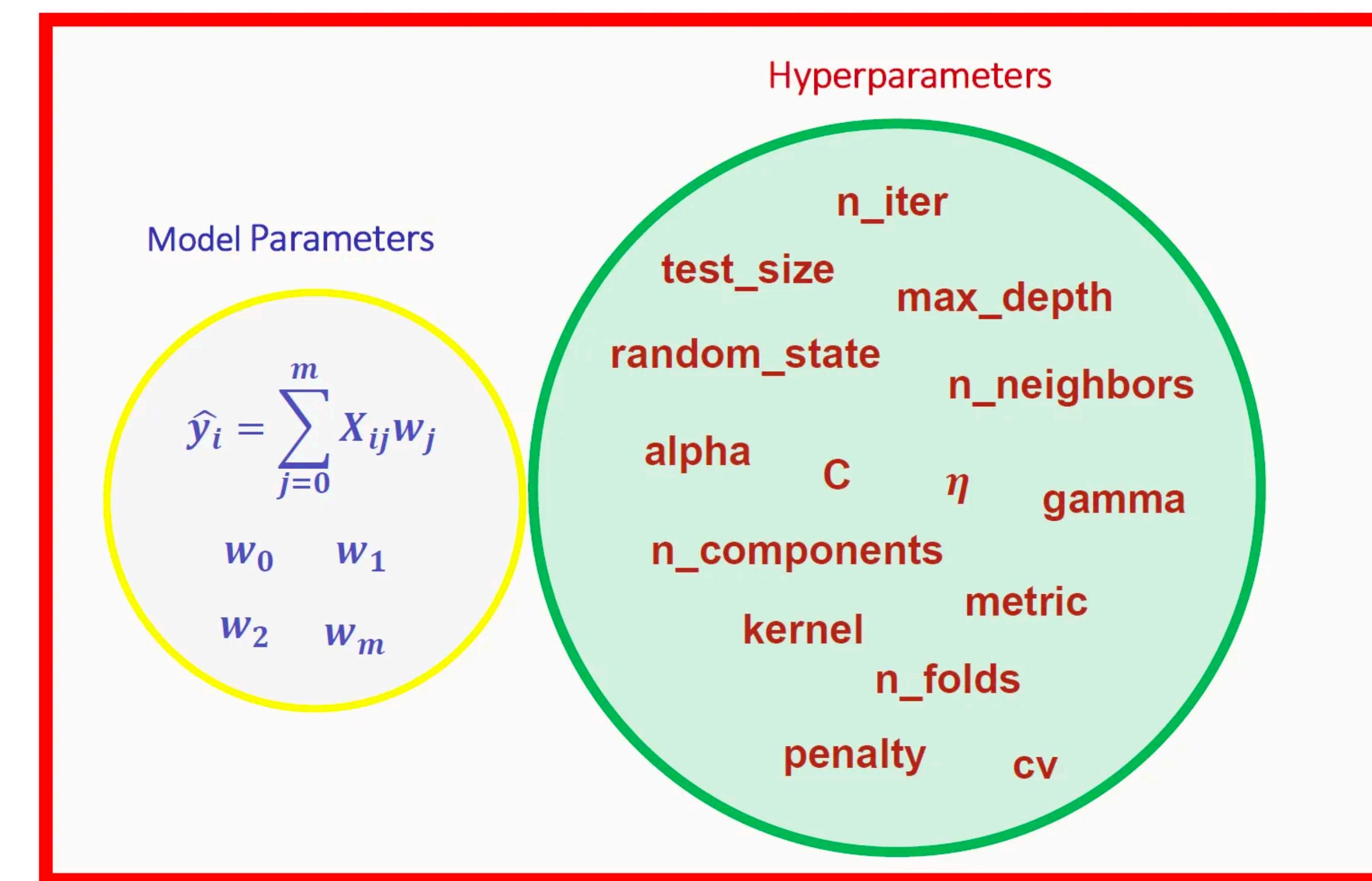
```
cv["test_score"].mean()
```

```
0.9533333333333333
```

PARAMETERS VS HYPERPARAMETERS

Parameters: Learned during training. Specific to the model being used. E.g., in linear regression, coefficients assigned to each feature.

Hyperparameters: Set by the user before training. Not trained by ML algorithms to find the optimum settings. We must choose it ourselves.



HYPERPARAMETER TUNING

For each hyperparameter setting $h_i \in \{h_0, h_1, \dots, h_r\}$
run cross-validation to compute the CV error for h_i

Select h_i with lowest CV error.

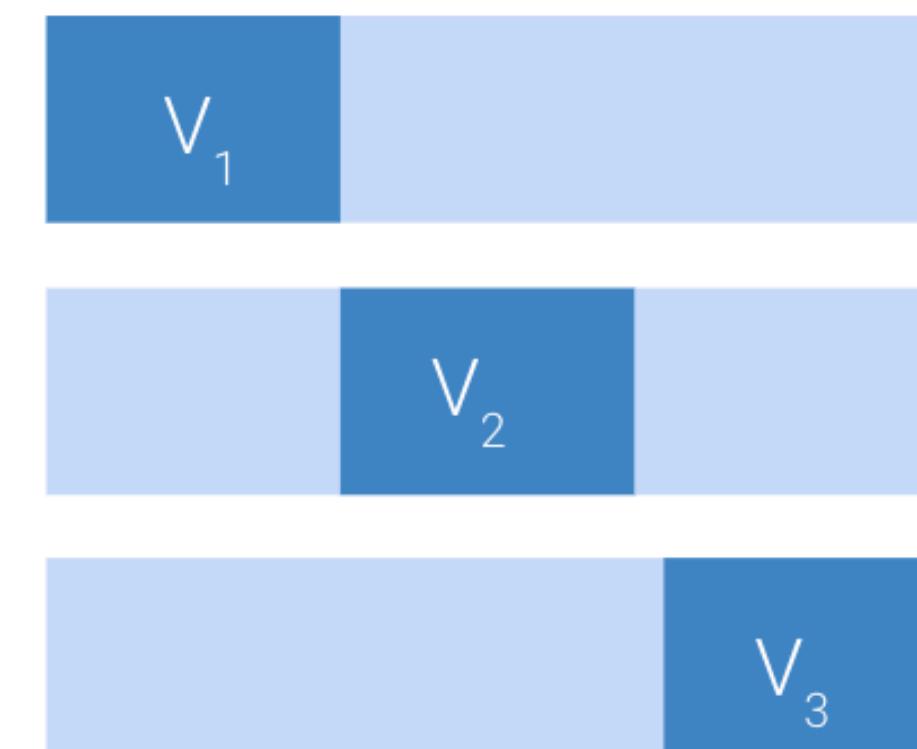
HYPERPARAMETER TUNING

For each hyperparameter setting $h_i \in \{h_0, h_1, \dots, h_r\}$
run cross-validation to compute the CV error for h_i

Select h_i with lowest CV error.

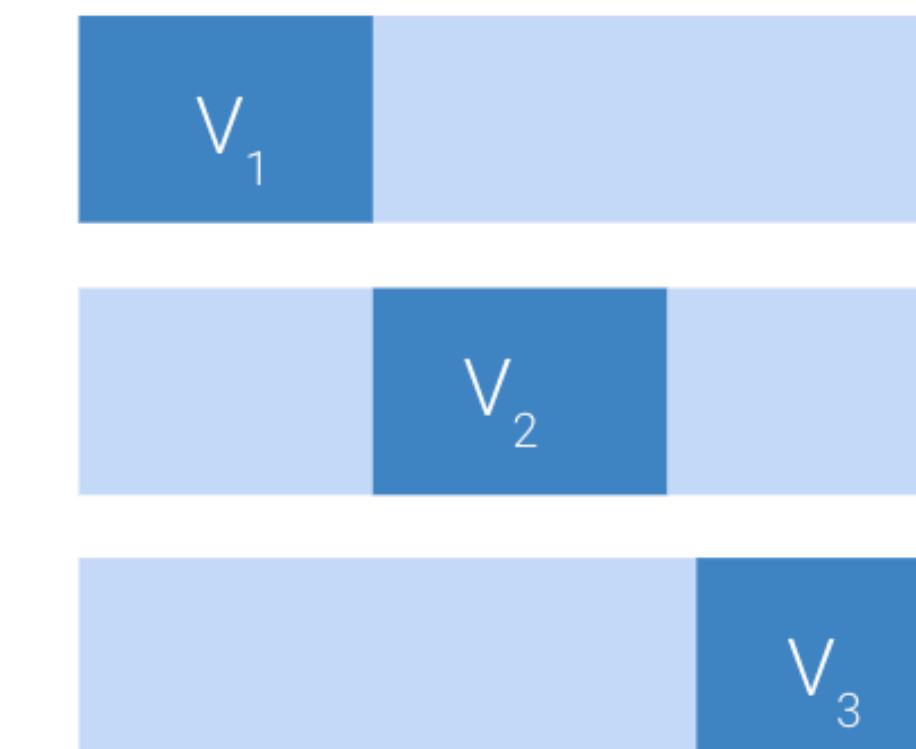
Example: Hyperparameter tuning for $n_estimators$ at settings of 10, 20, 50 with a 3-fold cross validation.

$$n_estimators = 10$$



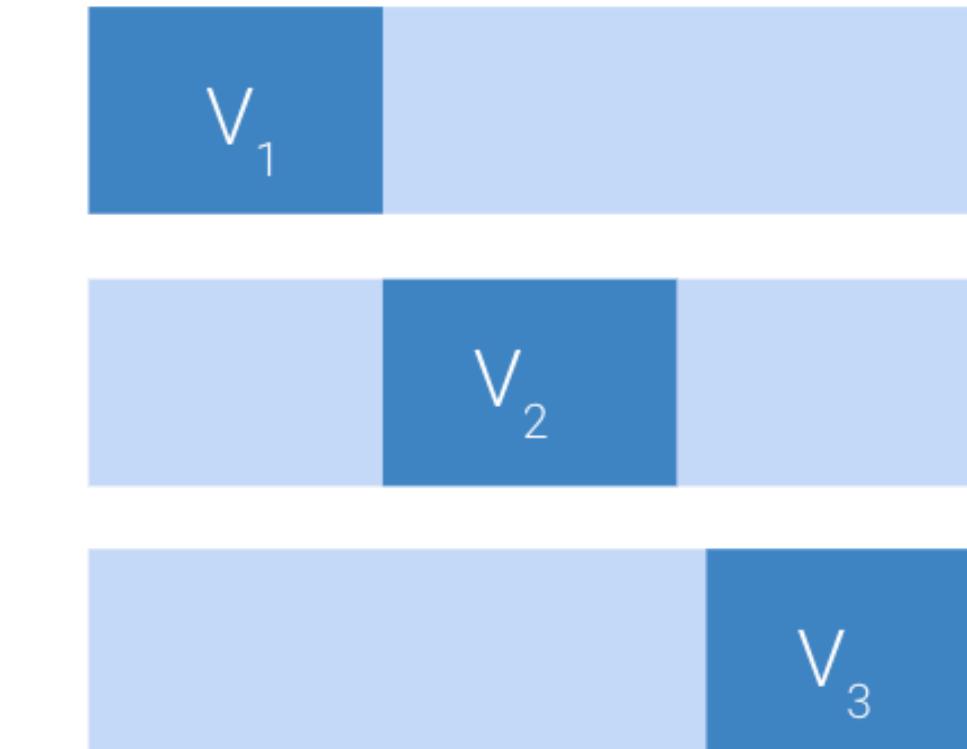
$$CV\ error = 6.56$$

$$= 20$$



$$CV\ error = 4.22$$

$$= 50$$



$$CV\ error = 3.15$$

LINEAR ALGEBRA REVIEW

Why?

- Typical data in Data Science:
All the relevant information in one or more data matrices
The rows \Rightarrow items (examples), the columns \Rightarrow features (attributes).
- Many ML algorithms are best understood through linear algebra
Ex: Linear regression reduces to a single formula.

Most of you must have taken a linear algebra course already.
This will just be a review.

LINEAR ALGEBRA REVIEW

Matrix representations of important objects

- Data:

Features (dimensions)			
Items (instances)	x_{11}	x_{12}	\dots
	x_{21}	x_{22}	\dots
\vdots	\vdots	\vdots	\ddots
	x_{n1}	x_{n2}	\dots
			x_{np}

LINEAR ALGEBRA REVIEW

Matrix representations of important objects

- System of equations:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n. \end{aligned}$$

$$\xrightarrow{\hspace{1cm}} \begin{matrix} A & x & b \\ \left(\begin{array}{cccc} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{array} \right) & \left(\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \right) & = \left(\begin{array}{c} b_1 \\ b_2 \\ \vdots \\ b_n \end{array} \right) \end{matrix}$$

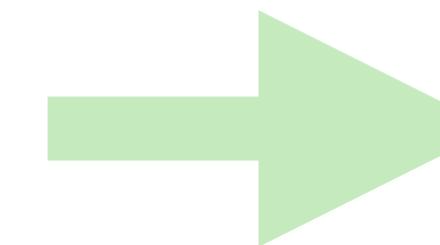
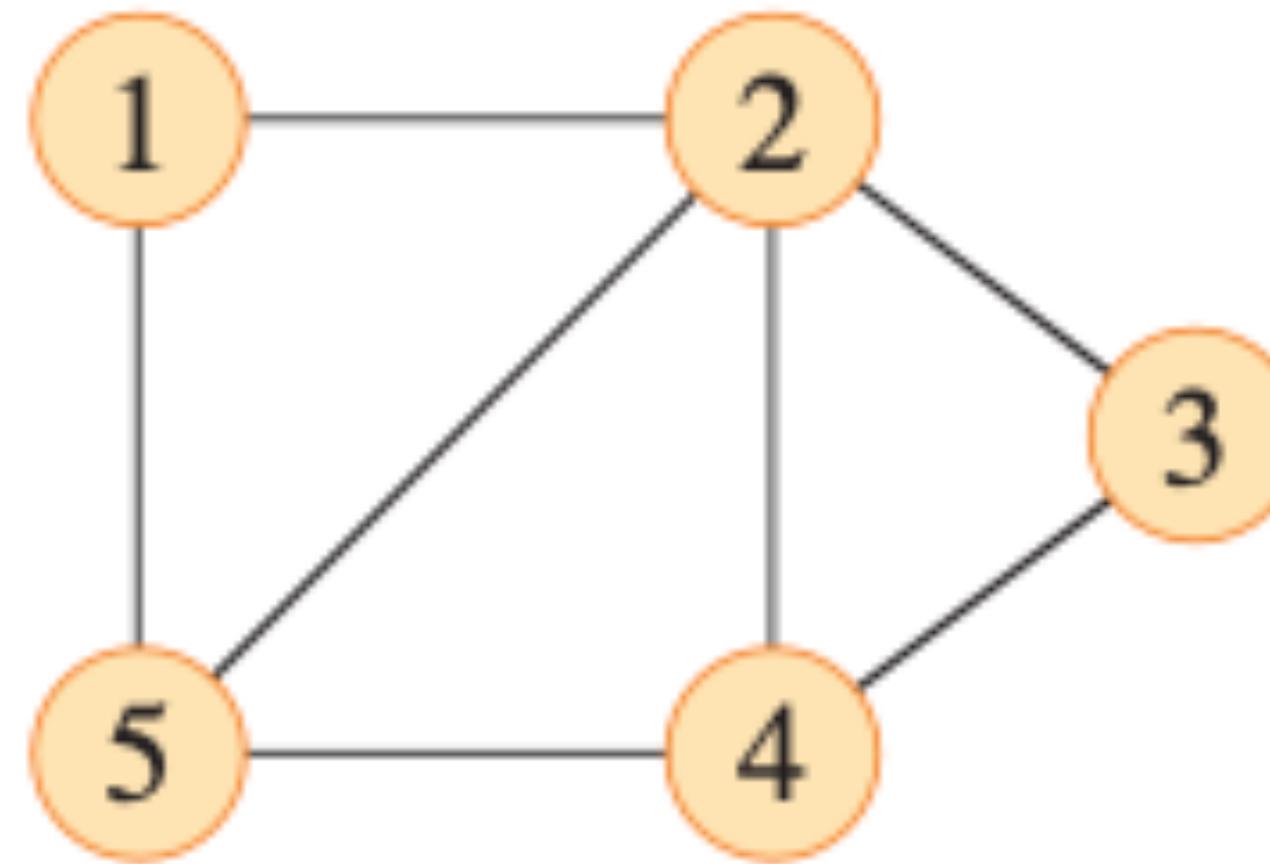
Row: an equation

Column: coefficients of a variable

LINEAR ALGEBRA REVIEW

Matrix representations of important objects

- Graphs and networks



	1	2	3	4	5
1	0	1	0	0	1
2	1	0	1	1	1
3	0	1	0	1	0
4	0	1	1	0	1
5	1	1	0	1	0

Adjacency matrix

LINEAR ALGEBRA REVIEW

Matrix representations of important objects

- Geometric points sets: Geometry and vectors

$n \times m$ matrix \Rightarrow Cloud of n points (rows) in m dimensions (columns)

LINEAR ALGEBRA REVIEW

Matrix representations of important objects

- Geometric points sets: Geometry and vectors

$n \times m$ matrix \Rightarrow Cloud of n points (rows) in m dimensions (columns)

Dot product of two vectors:

$$A \cdot B = \sum_{i=1}^n A_i B_i$$

LINEAR ALGEBRA REVIEW

Matrix representations of important objects

- Geometric points sets: Geometry and vectors

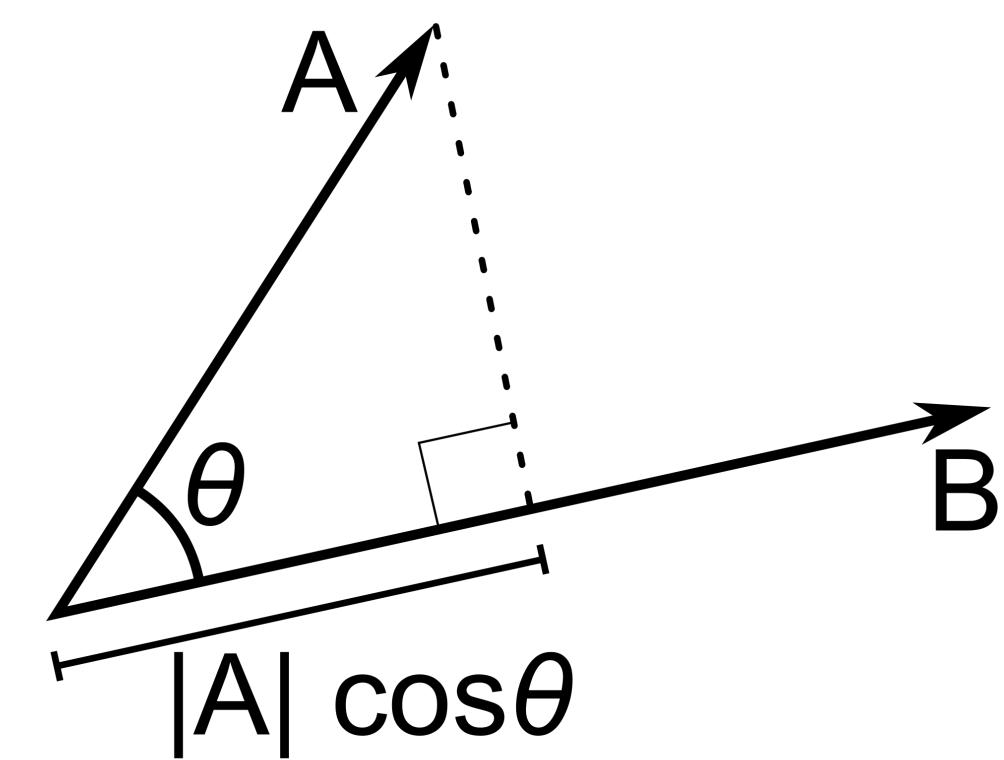
$n \times m$ matrix \Rightarrow Cloud of n points (rows) in m dimensions (columns)

Dot product of two vectors:

$$A \cdot B = \sum_{i=1}^n A_i B_i$$

Can compute the angle:

$$\cos(\Theta) = \frac{A \cdot B}{\|A\| \|B\|}$$



where $\|A\| = \sqrt{\sum_{i=1}^n A_i^2}$ is the magnitude (length).

LINEAR ALGEBRA REVIEW

Matrix representations of important objects

- Geometric points sets: Geometry and vectors

$n \times m$ matrix \Rightarrow Cloud of n points (rows) in m dimensions (columns)

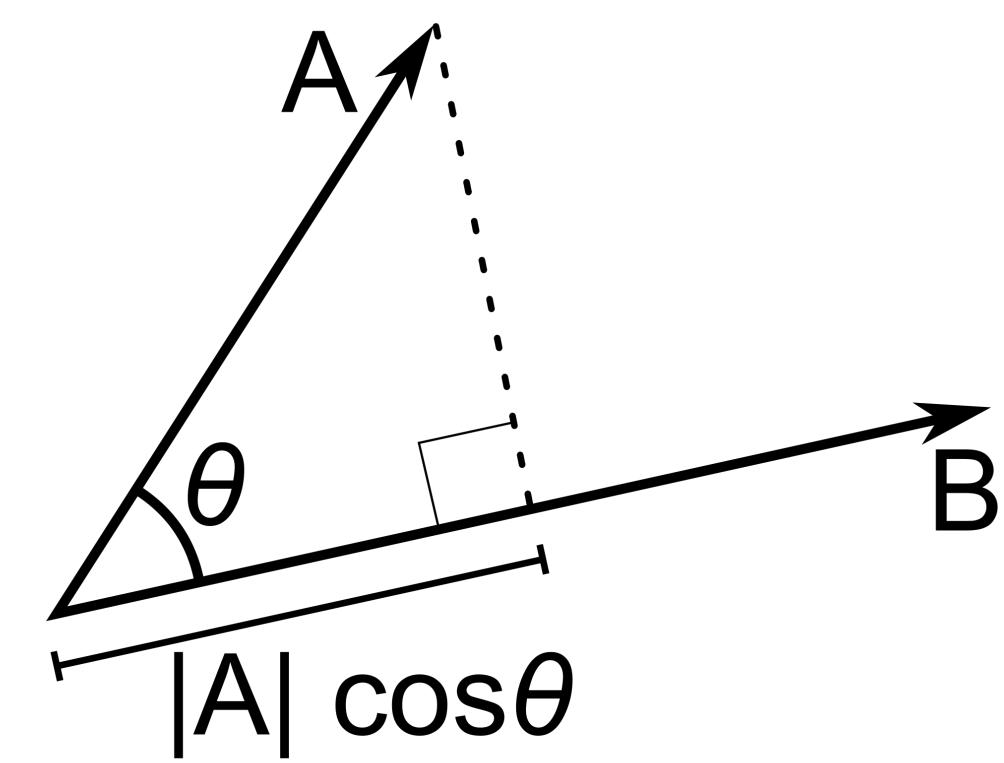
Dot product of two vectors:

$$A \cdot B = \sum_{i=1}^n A_i B_i$$

Familiar?

Can compute the angle:

$$\cos(\Theta) = \frac{A \cdot B}{\|A\| \|B\|}$$



where $\|A\| = \sqrt{\sum_{i=1}^n A_i^2}$ is the magnitude (length).

LINEAR ALGEBRA REVIEW

Matrix representations of important objects

- Geometric points sets: Geometry and vectors

$n \times m$ matrix \Rightarrow Cloud of n points (rows) in m dimensions (columns)

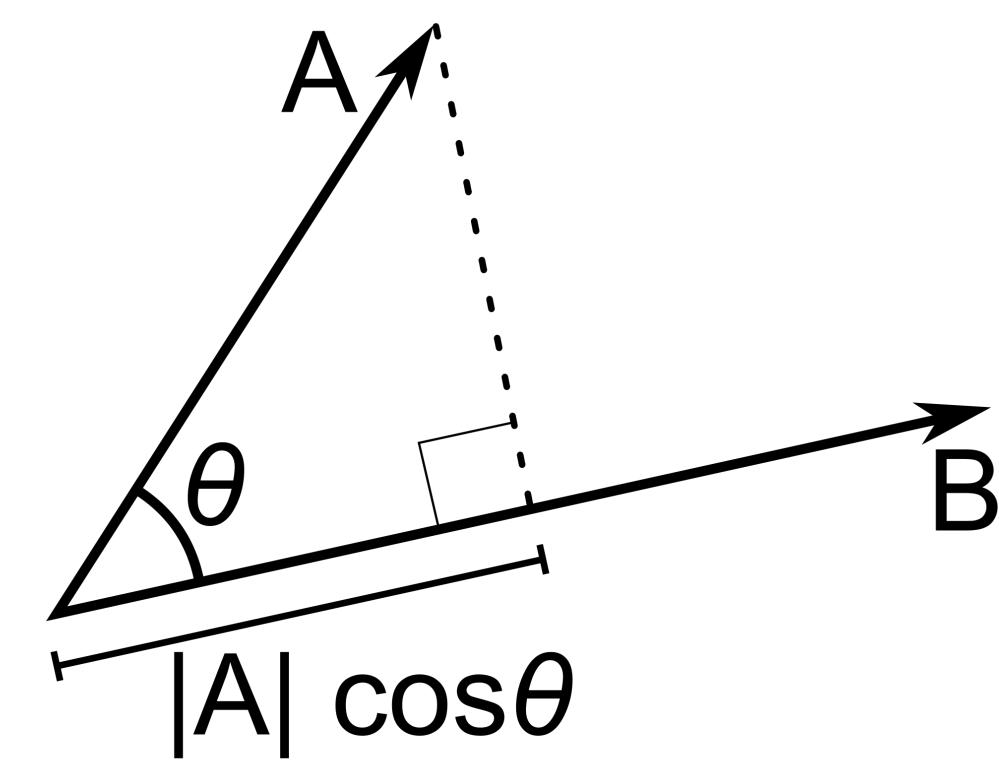
Dot product of two vectors:

$$A \cdot B = \sum_{i=1}^n A_i B_i$$

Familiar?

Can compute the angle:

$$\cos(\Theta) = \frac{A \cdot B}{\|A\| \|B\|}$$



Cosine similarity: in $[-1, 1]$
(Similar to Pearson correlation coefficient - exactly the same when means are 0.)

where $\|A\| = \sqrt{\sum_{i=1}^n A_i^2}$ is the magnitude (length).

MATRIX OPERATIONS

Matrix addition: $C = A + B$

$$C_{ij} = A_{ij} + B_{ij}, \quad \forall 1 \leq i \leq n, 1 \leq j \leq m$$

Scalar multiplication: $A' = rA$

$$A'_{ij} = rA_{ij}, \quad \forall 1 \leq i \leq n, 1 \leq j \leq m$$

MATRIX OPERATIONS

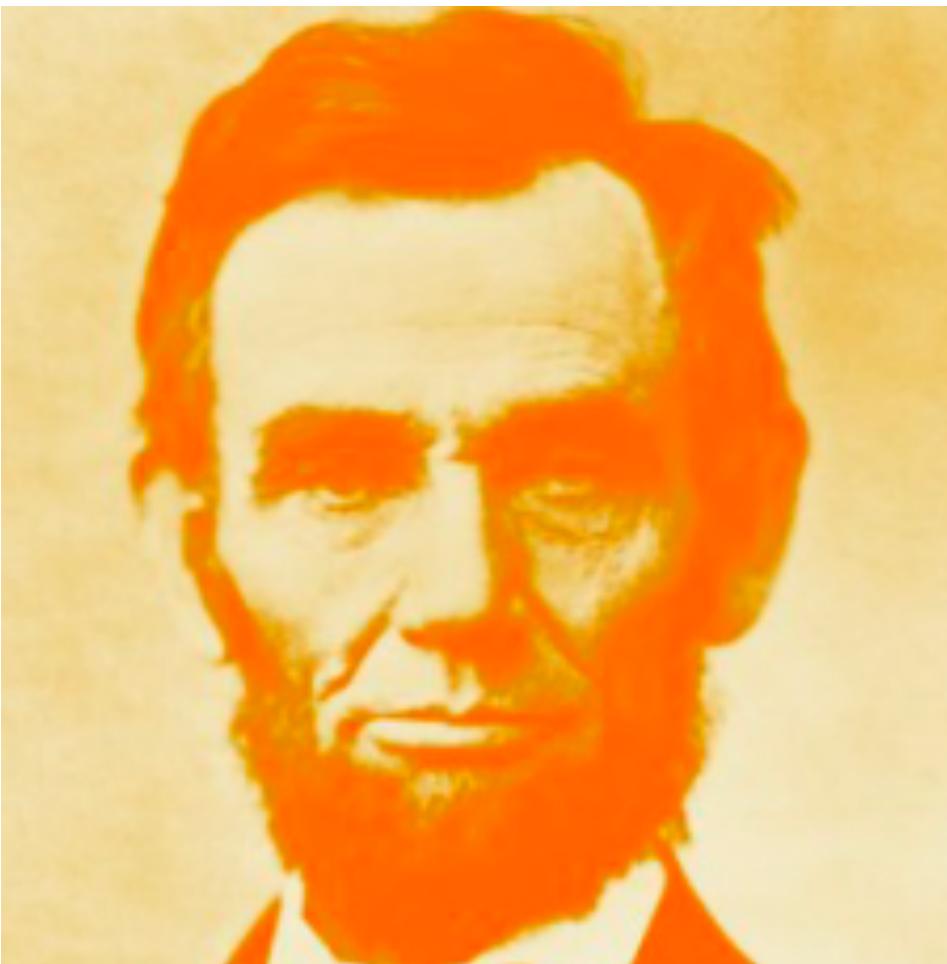
Matrix addition: $C = A + B$

$$C_{ij} = A_{ij} + B_{ij}, \quad \forall 1 \leq i \leq n, 1 \leq j \leq m$$

Scalar multiplication: $A' = rA$

$$A'_{ij} = rA_{ij}, \quad \forall 1 \leq i \leq n, 1 \leq j \leq m$$

Linear combinations: $\alpha A + (1 - \alpha)B$



for $\alpha = 0.5$

$$\alpha L + (1 - \alpha)M = \text{Lincoln\&Memorial}$$

MATRIX OPERATIONS

Matrix transpose: A^T

$$A_{ij}^T = A_{ji}, \quad \forall 1 \leq i \leq n, 1 \leq j \leq m$$

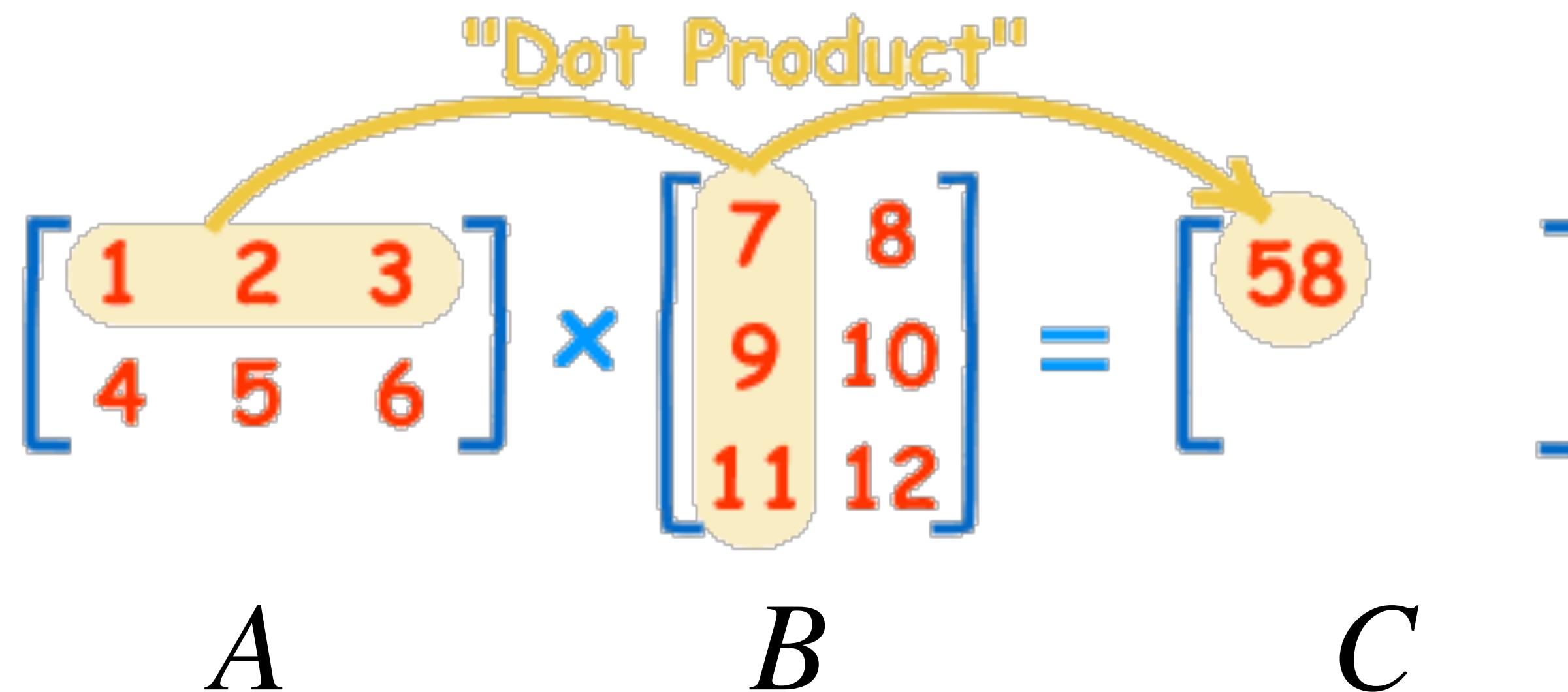
MATRIX OPERATIONS

Matrix transpose: A^T

$$A_{ij}^T = A_{ji}, \quad \forall 1 \leq i \leq n, 1 \leq j \leq m$$

Matrix multiplication: $C = AB$ (A is $n \times k$, B is $k \times m$)

$$C_{ij} = A_i \cdot B_j, \quad \forall 1 \leq i \leq n, 1 \leq j \leq m$$



Note that, Dot products measure how “in sync” the two vectors are (like covariance or correlation).

MULTIPLYING A MATRIX WITH ITS TRANSPOSE

$A A^T$ measures “in-syncness” among rows (items/points).

$A^T A$ measures “in-syncness” among columns (features/dimensions).

MULTIPLYING A MATRIX WITH ITS TRANSPOSE

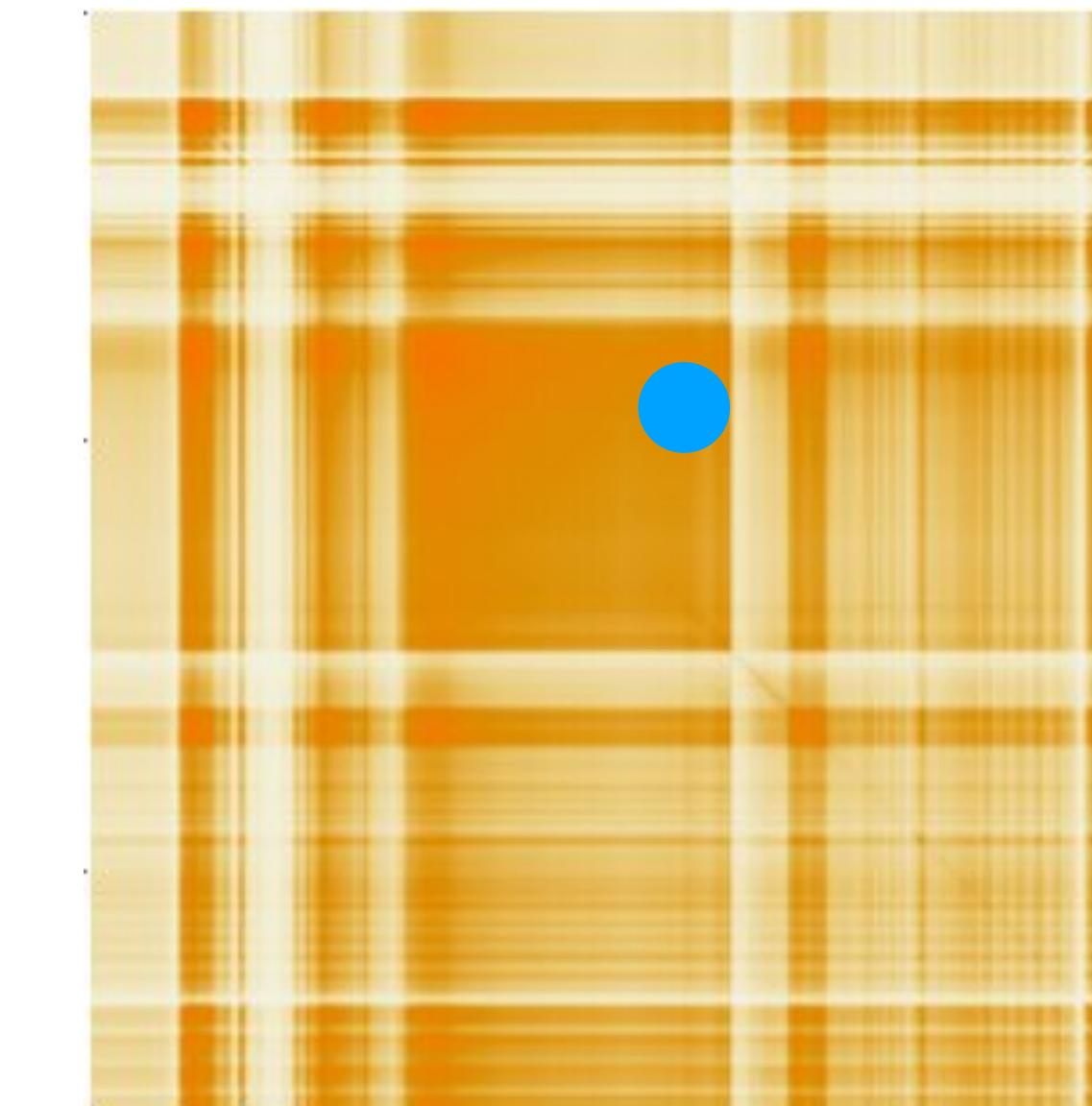
$A A^T$ measures “in-syncness” among rows (items/points).

$A^T A$ measures “in-syncness” among columns (features/dimensions).

M



MM^T



MULTIPLYING A MATRIX WITH ITS TRANSPOSE

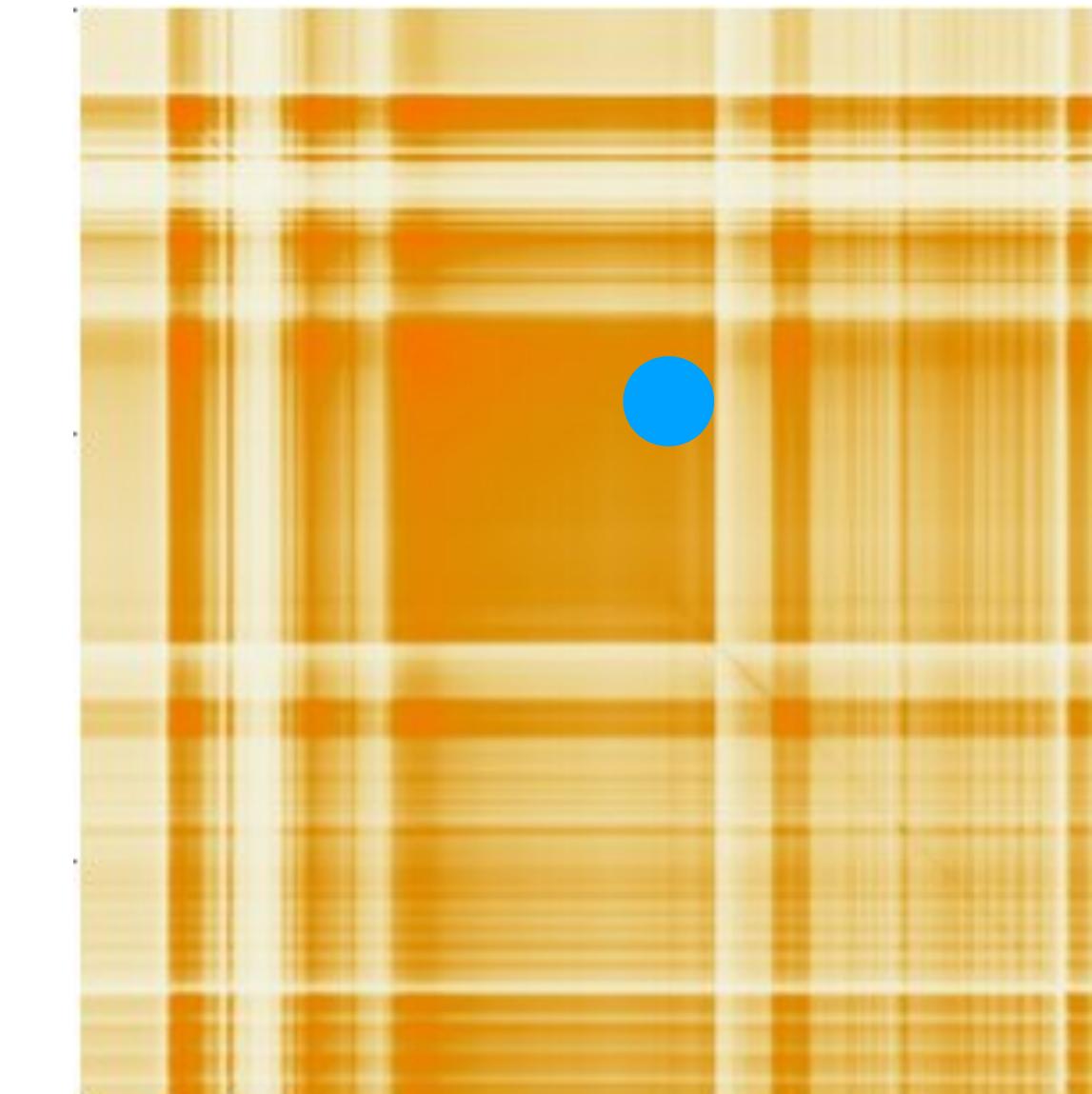
$A A^T$ measures “in-syncness” among rows (items/points).

$A^T A$ measures “in-syncness” among columns (features/dimensions).

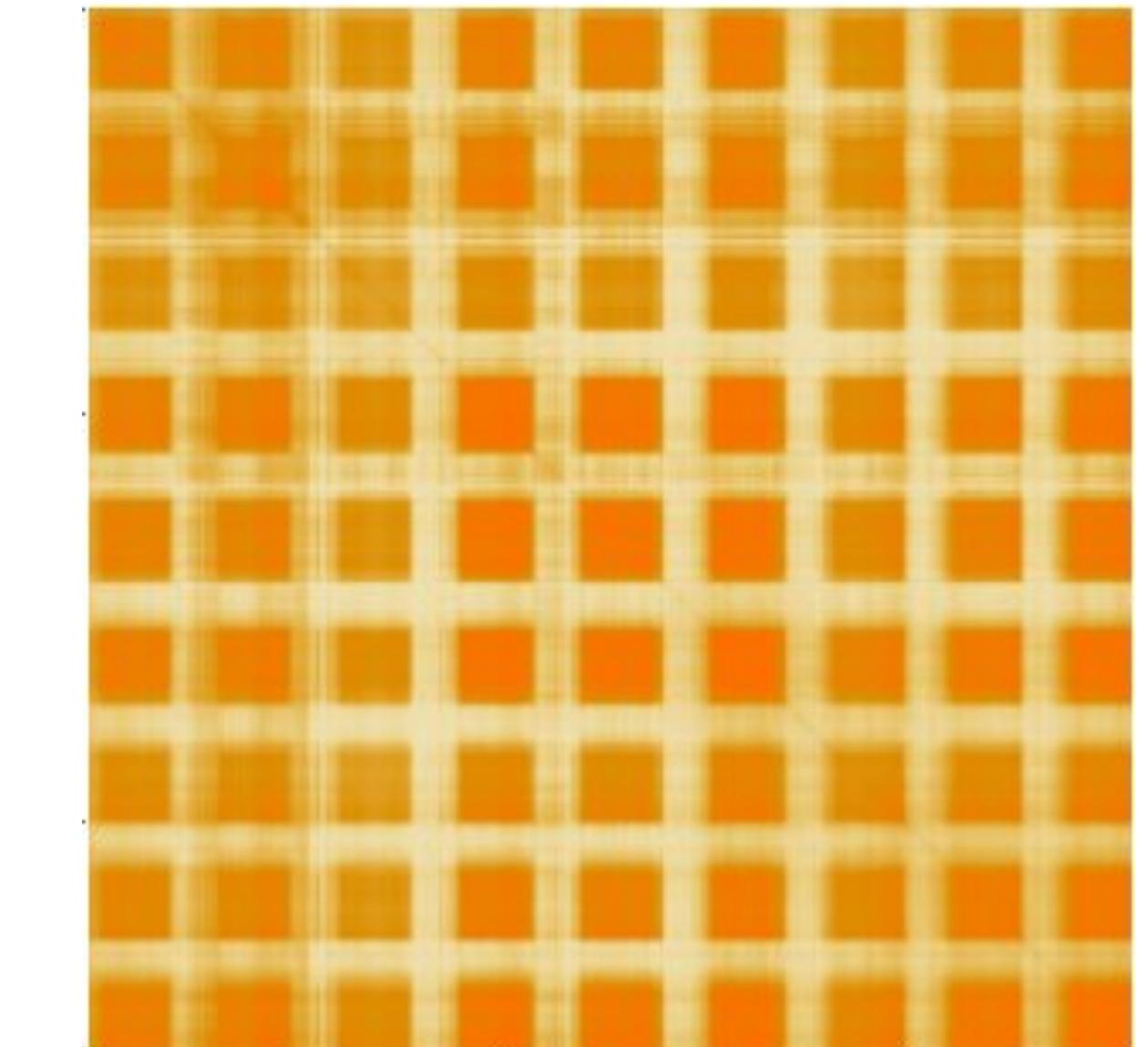
M



MM^T



$M^T M$



QUIZ

What is k-fold cross validation and why is it useful?

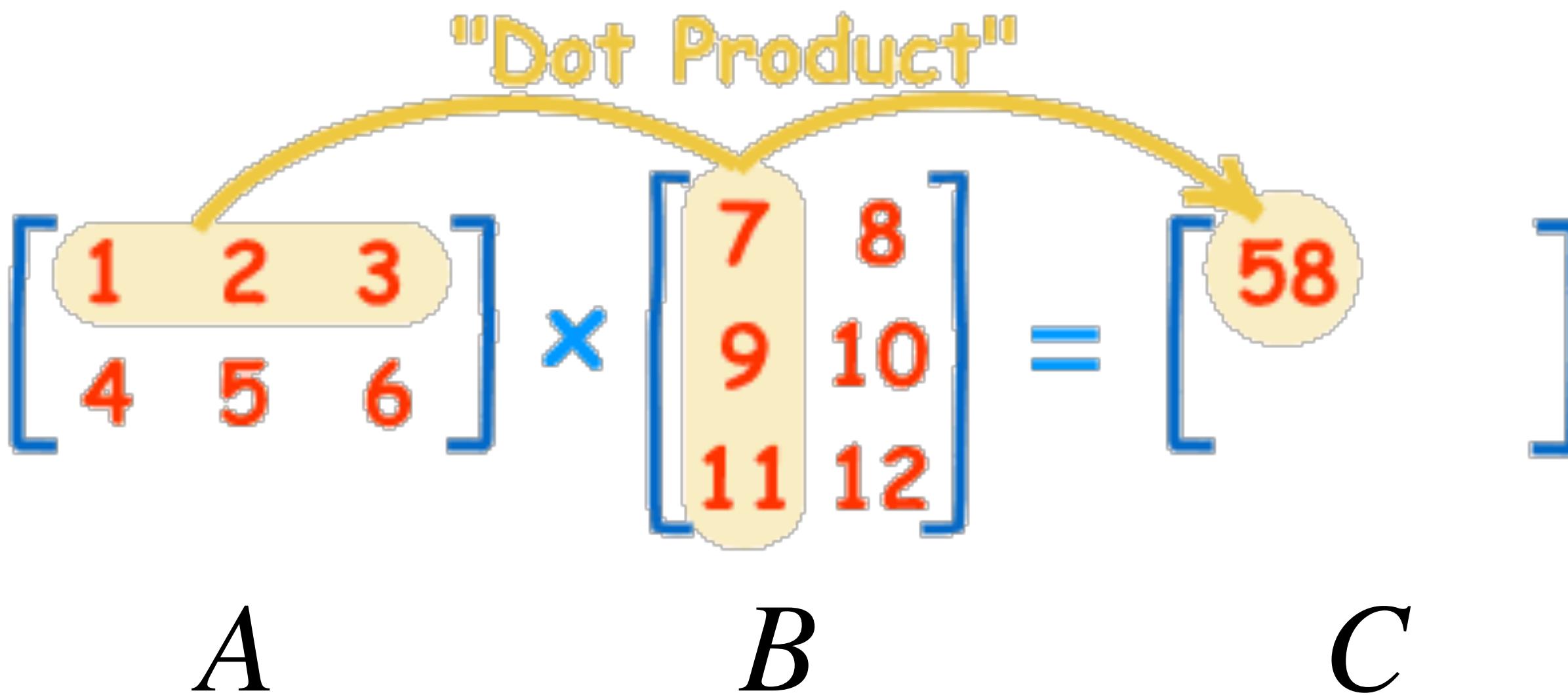
REVIEW OF (THE LAST PART OF) PREVIOUS LECTURE

Matrix transpose: A^T

$$A_{ij}^T = A_{ji}, \quad \forall 1 \leq i \leq n, 1 \leq j \leq m$$

Matrix multiplication: $C = AB$ (A is $n \times k$, B is $k \times m$)

$$C_{ij} = A_i \cdot B_j, \quad \forall 1 \leq i \leq n, 1 \leq j \leq m$$



Note that, Dot products measure how “in sync” the two vectors are (like covariance or correlation).

REVIEW OF (THE LAST PART OF) PREVIOUS LECTURE

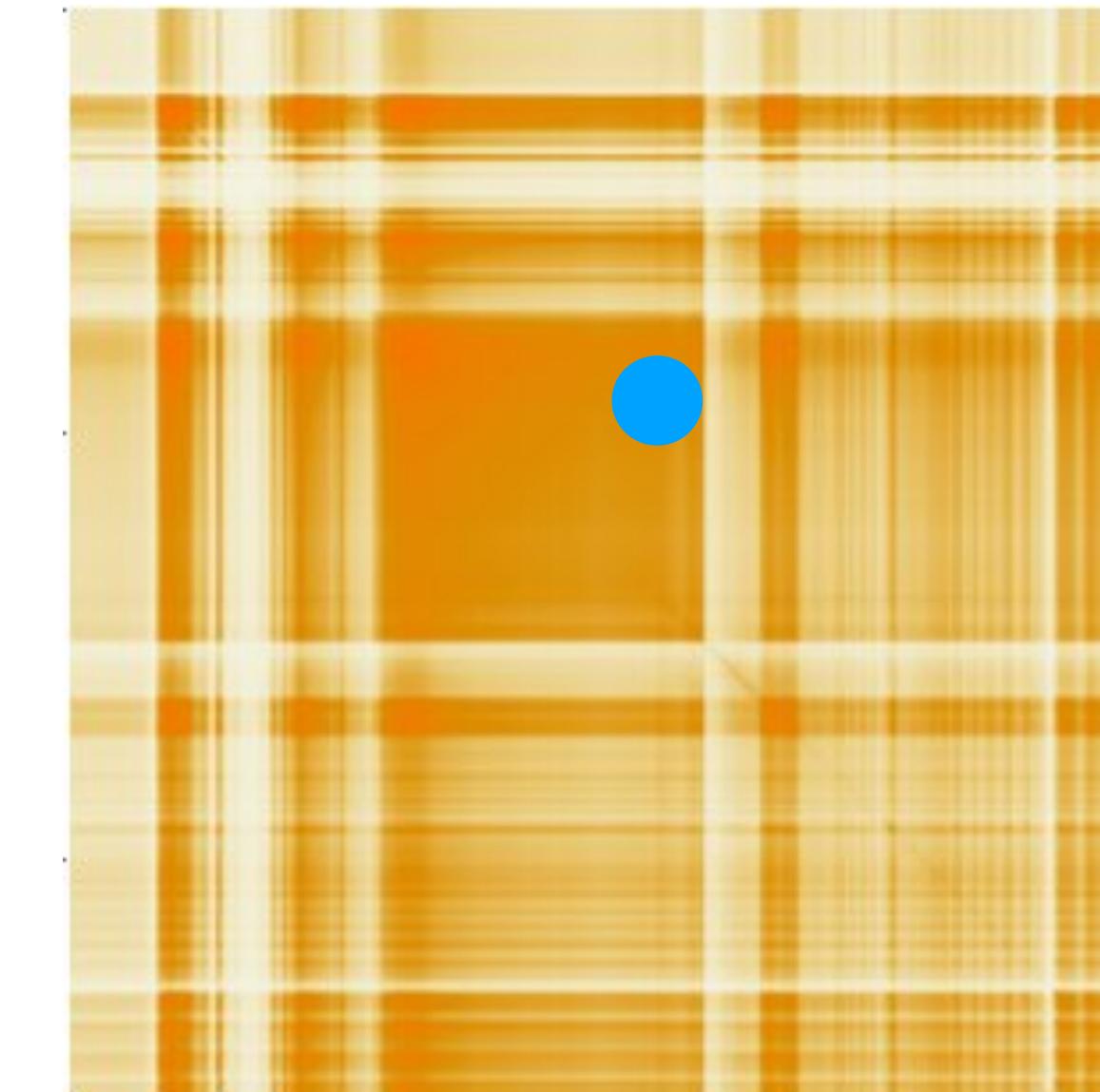
AA^T measures “in-syncness” among rows (items/points).

$A^T A$ measures “in-syncness” among columns (features/dimensions).

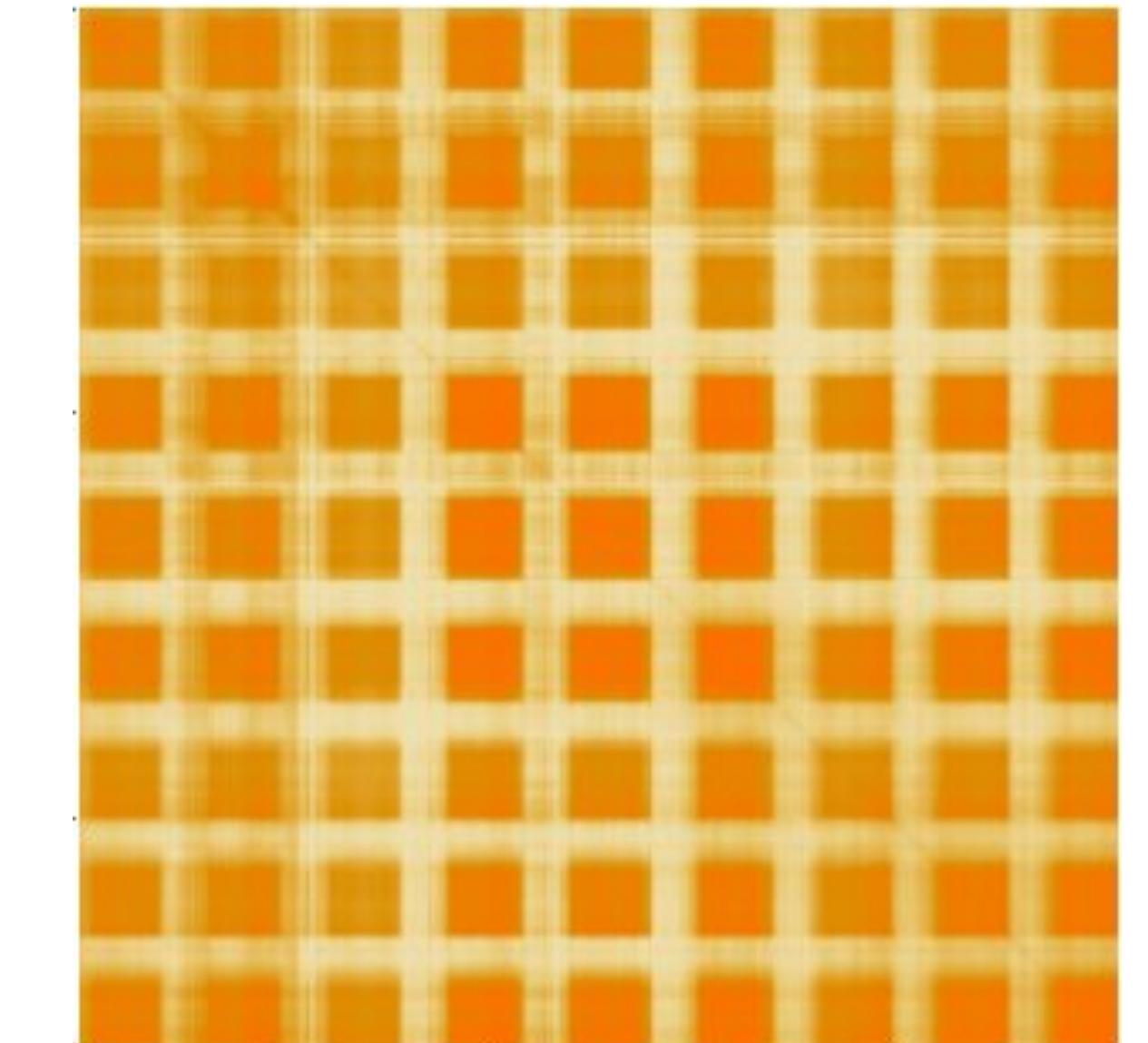
M



MM^T



$M^T M$



ALGORITHMS FOR MODELING

ALGORITHMS FOR MODELING

This lecture:

- Simple Linear regression
- Multiple linear regression
- Goodness of fit

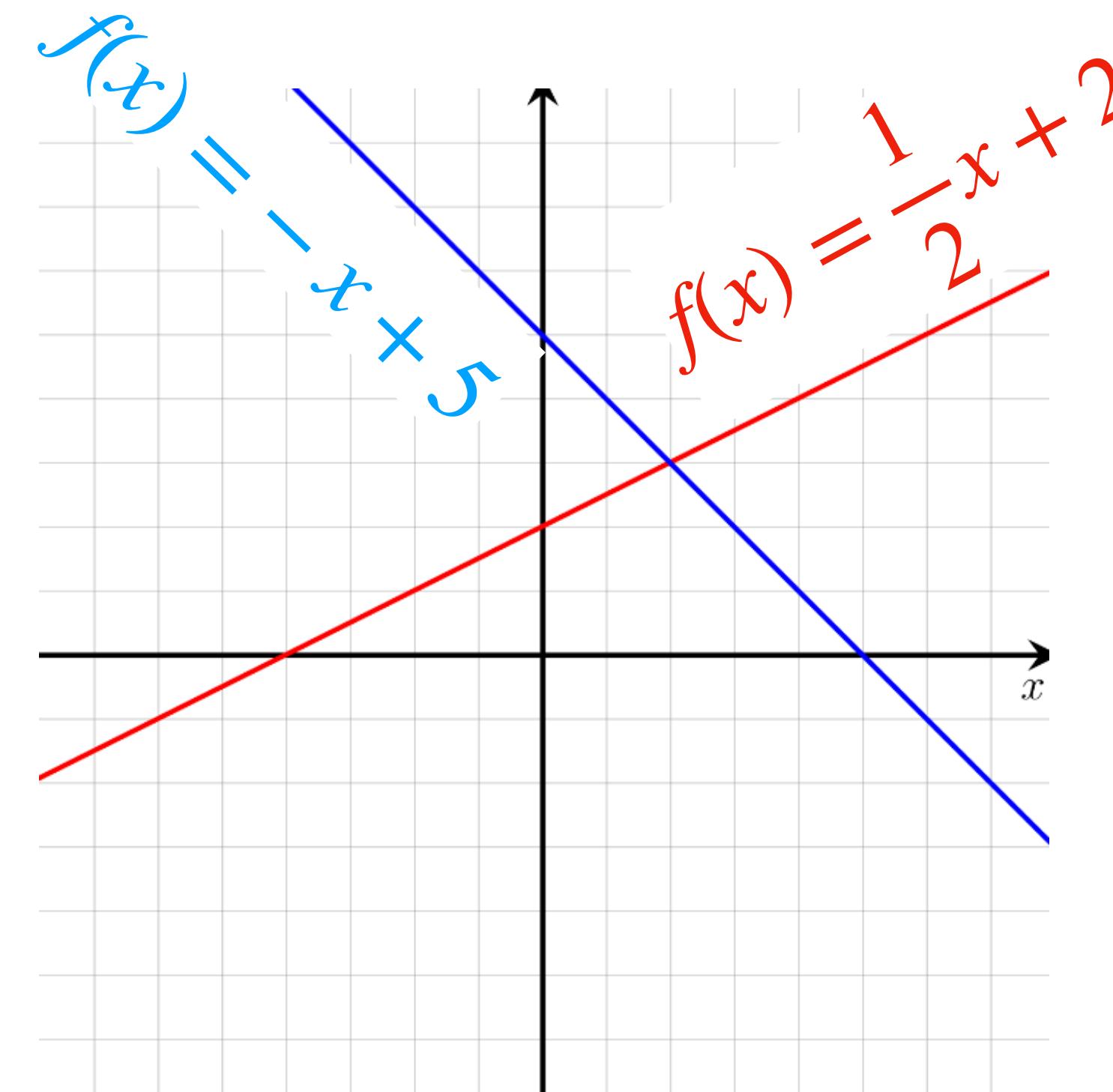
Readings:

- Ch3 from ISLP, James et al. textbook.
- 9.1-9.5 from Skiena's textbook.

LINEAR REGRESSION

Line equation: $f(x) = b + ax$

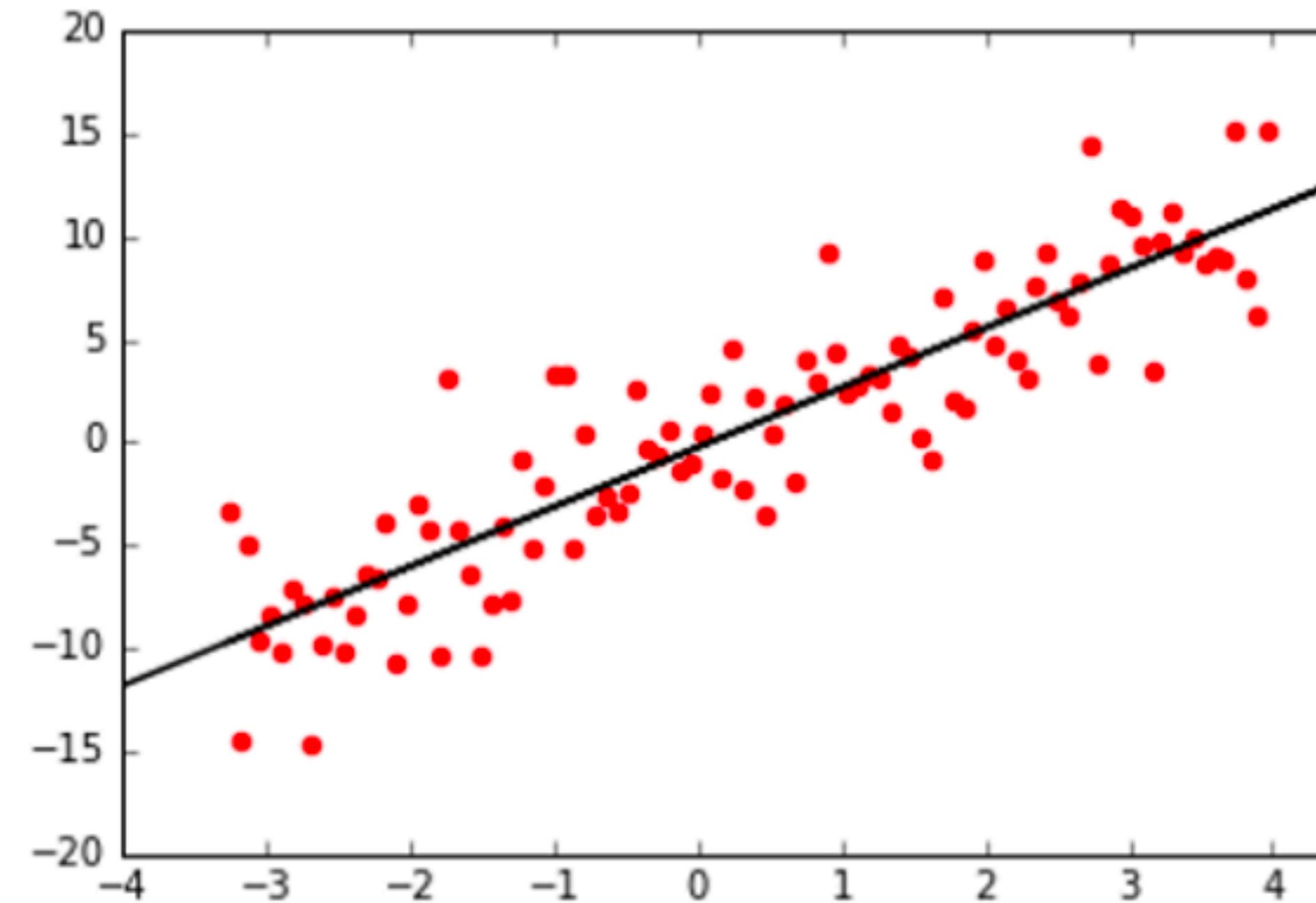
Intercept
Slope



LINEAR REGRESSION

Linear regression in 2D:

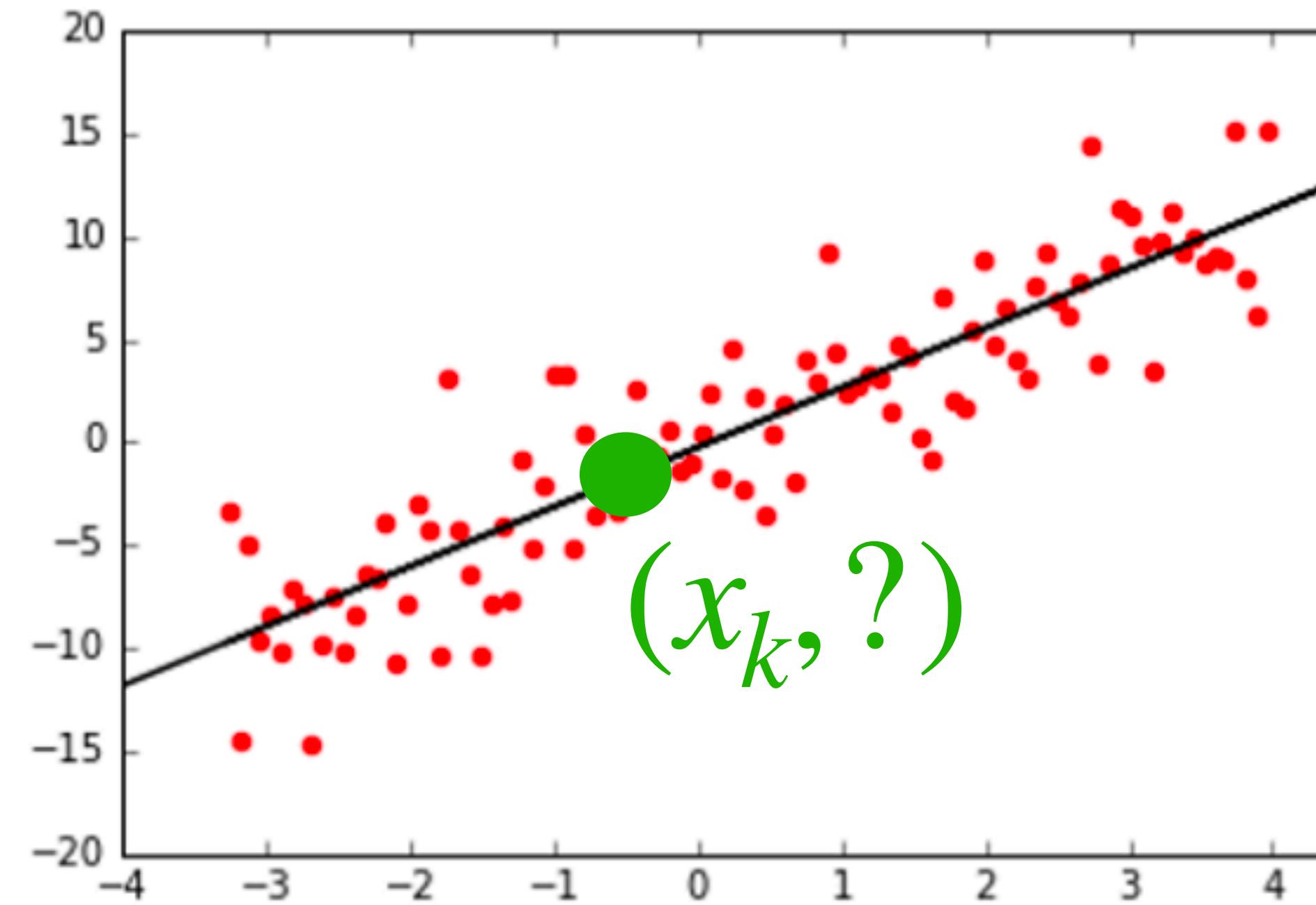
Find $\hat{y} = \hat{w}_0 + \hat{w}_1 x$ that **best** fits the set of points.



LINEAR REGRESSION

Linear regression in 2D:

Find $\hat{y} = \hat{w}_0 + \hat{w}_1 x$ that **best** fits the set of points.



Given x_k of a new point p_k , we predict its y-coord by $\hat{y}_k = \hat{w}_0 + \hat{w}_1 x_k$.

LINEAR REGRESSION

Linear regression in 2D:

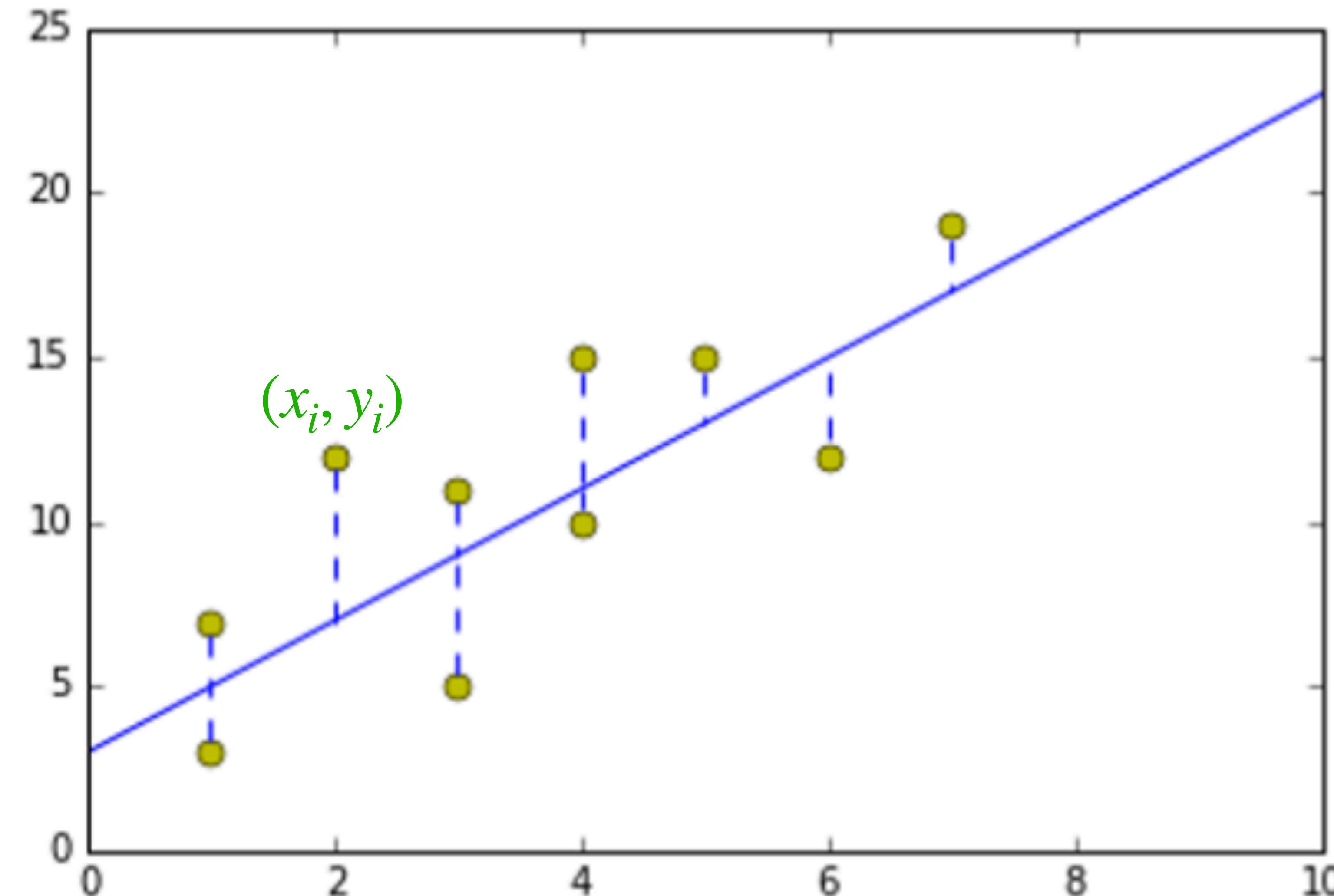
Find $\hat{y} = \hat{w}_0 + \hat{w}_1 x$ that **best** fits the set of points.

How to define **best**?

LINEAR REGRESSION

Residual of data point (x_i, y_i) : $y_i - \hat{y}_i$

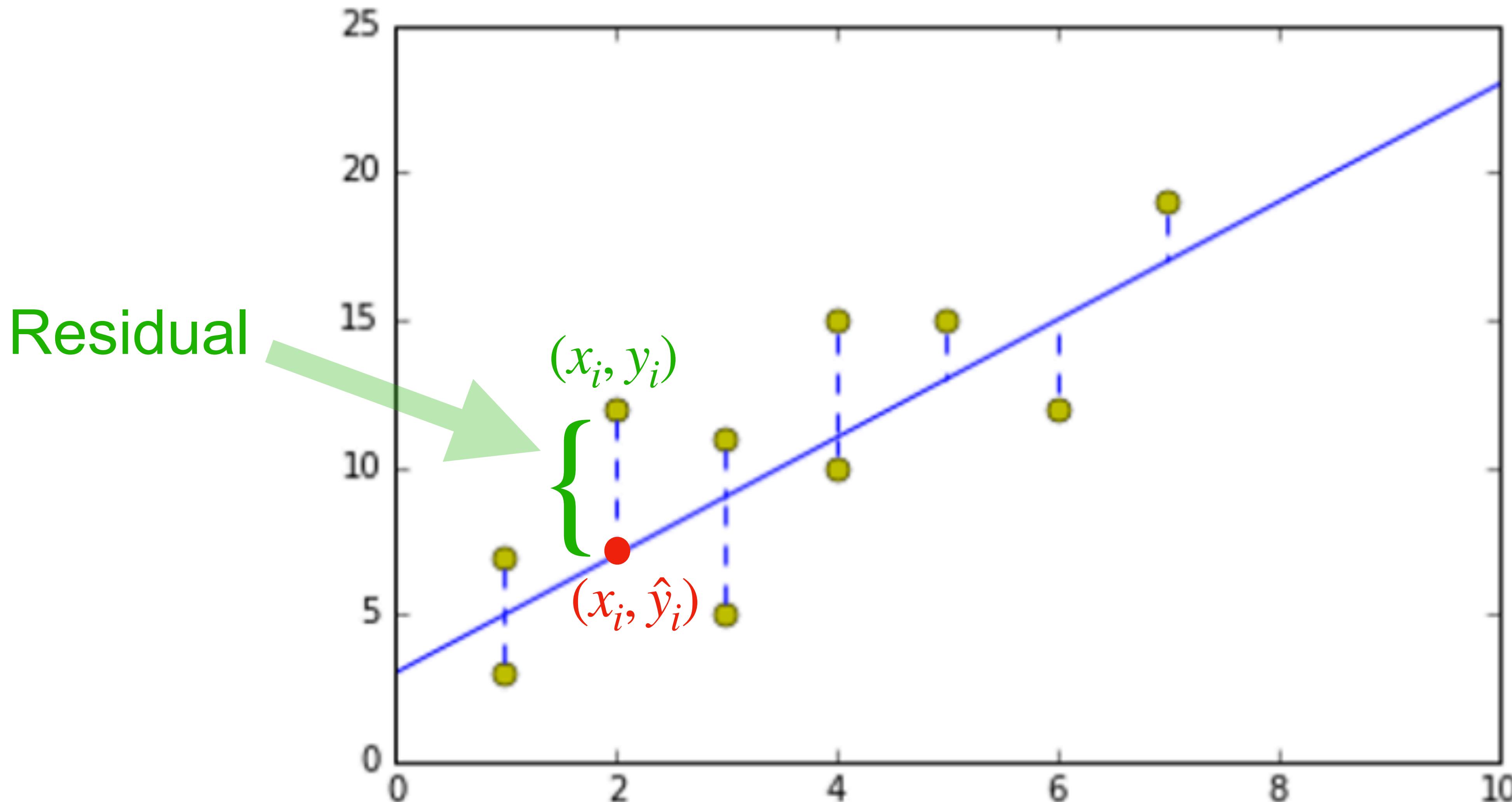
$$(\hat{y}_i = \hat{w}_0 + \hat{w}_1 x_i)$$



LINEAR REGRESSION

Residual of data point (x_i, y_i) : $y_i - \hat{y}_i$

$$(\hat{y}_i = \hat{w}_0 + \hat{w}_1 x_i)$$



LINEAR REGRESSION

Residual of data point (x_i, y_i) : $y_i - \hat{y}_i$

$$(\hat{y}_i = \hat{w}_0 + \hat{w}_1 x_i)$$

Residual sum of squares:
(RSS) $\sum_{i=1}^n (y_i - \hat{y}_i)^2$

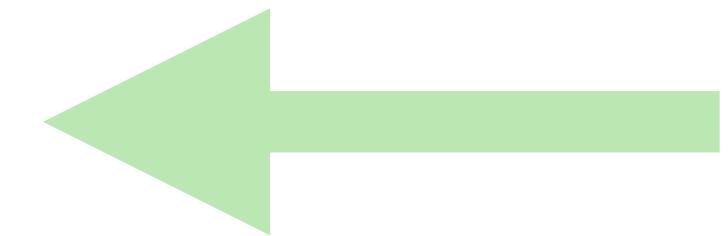
LINEAR REGRESSION

Residual of data point (x_i, y_i) : $y_i - \hat{y}_i$

$$(\hat{y}_i = \hat{w}_0 + \hat{w}_1 x_i)$$

Residual sum of squares:
(RSS)

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$



Find line that
minimizes RSS.

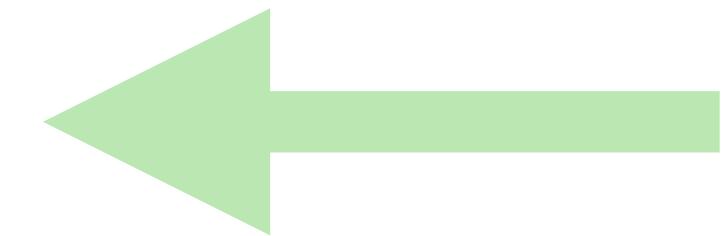
LINEAR REGRESSION

Residual of data point (x_i, y_i) : $y_i - \hat{y}_i$

$$(\hat{y}_i = \hat{w}_0 + \hat{w}_1 x_i)$$

Residual sum of squares:
(RSS)

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$



Find line that
minimizes RSS.

With one variable, X :

$$\hat{w}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{w}_0 = \bar{y} - \hat{w}_1 \bar{x}$$

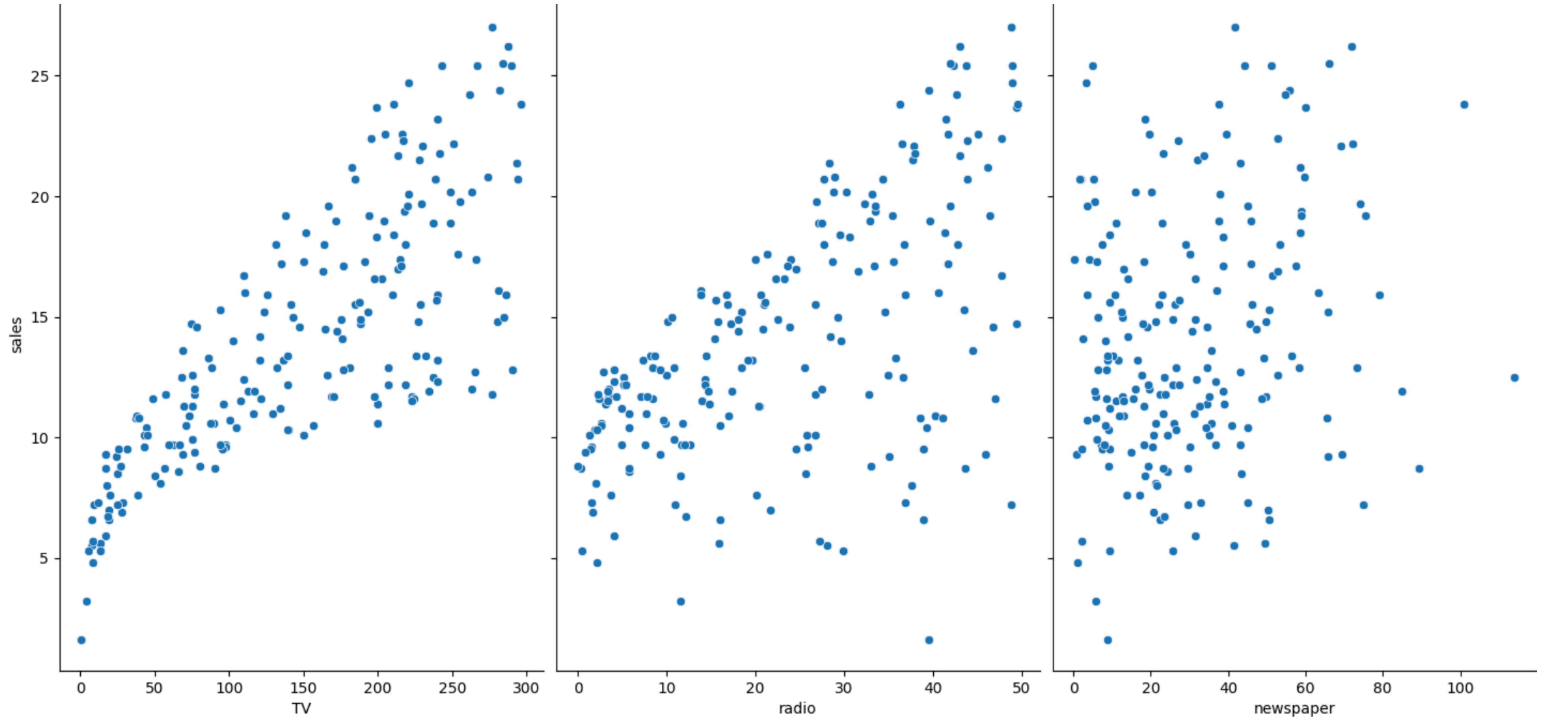
LINEAR REGRESSION

Ex: Advertising data set.

```
df.head()
```

	TV	radio	newspaper	sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

LINEAR REGRESSION



LINEAR REGRESSION

Linear regression only using TV feature.

	x	y
	TV	sales
0	230.1	22.1
1	44.5	10.4
2	17.2	9.3
3	151.5	18.5
4	180.8	12.9

Fit: Find \hat{w}_0, \hat{w}_1 such that $\sum_{i=1}^n (y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$ is minimum.

LINEAR REGRESSION

```
X = df['TV'].values.reshape(-1, 1)
y = df.sales

X_train,X_test, y_train, y_test = \
    train_test_split(X,y, train_size=0.8, random_state=0)

# Fit one feature
lr = linear_model.LinearRegression()
lr.fit(X_train, y_train)

# Print coefficients
print(lr.intercept_)
print(lr.coef_)
```

7.292493773559364
[0.04600779]

LINEAR REGRESSION

```
x = df['TV'].values.reshape(-1, 1)
y = df.sales

X_train,X_test, y_train, y_test = \
    train_test_split(X,y, train_size=0.8, random_state=0)

# Fit one feature
lr = linear_model.LinearRegression()
lr.fit(X_train, y_train)

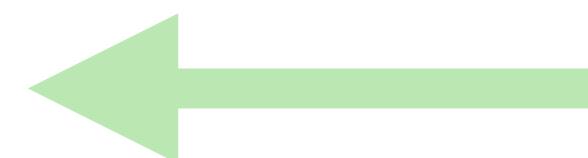
# Print coefficients
print(lr.intercept_)
print(lr.coef_)
```

7.292493773559364 ← \hat{w}_0 Assuming no advertising, will sell approximately 7292 units.
[0.04600779] ← \hat{w}_1 Every additional \$1000 spent on TV advertising is associated with selling approximately 46 additional units.

LINEAR REGRESSION

More on coefficients: Accuracy of coefficient estimates

```
import statsmodels.api as sm  
X_train_sm = sm.add_constant(X_train)  
lr_sm = sm.OLS(y_train, X_train_sm)  
lr_sm_result = lr_sm.fit()  
lr_sm_result.summary()
```



With statsmodels we need to add intercept explicitly.

LINEAR REGRESSION

More on coefficients: Accuracy of coefficient estimates

```
import statsmodels.api as sm
X_train_sm = sm.add_constant(X_train)
lr_sm = sm.OLS(y_train, X_train_sm)
lr_sm_result = lr_sm.fit()
lr_sm_result.summary()
```

	coef	std err	t	P> t	[0.025	0.975]
const	7.2925	0.529	13.795	0.000	6.248	8.337
x1	0.0460	0.003	15.031	0.000	0.040	0.052

LINEAR REGRESSION

More on coefficients: Accuracy of coefficient estimates

```
import statsmodels.api as sm
X_train_sm = sm.add_constant(X_train)
lr_sm = sm.OLS(y_train, X_train_sm)
lr_sm_result = lr_sm.fit()
lr_sm_result.summary()
```

	coef	std err	t	P> t	[0.025 0.975]
const	7.2925	0.529	13.795	0.000	6.248 8.337
x1	0.0460	0.003	15.031	0.000	0.040 0.052

Recall that:

$$\mu \in \bar{X}_n \pm k \times SE(\hat{\mu})$$

($k \approx 2$ for 95 % confidence interval)

LINEAR REGRESSION

More on coefficients: Accuracy of coefficient estimates

```
import statsmodels.api as sm
X_train_sm = sm.add_constant(X_train)
lr_sm = sm.OLS(y_train, X_train_sm)
lr_sm_result = lr_sm.fit()
lr_sm_result.summary()
```

	coef	std err	t	P> t	[0.025 0.975]
const	7.2925	0.529	13.795	0.000	6.248 8.337
x1	0.0460	0.003	15.031	0.000	0.040 0.052

Recall that:

$$\mu \in \bar{X}_n \pm k \times SE(\hat{\mu})$$

($k \approx 2$ for 95 % confidence interval)

We can apply to \hat{w}_0, \hat{w}_1 :

95 % confidence interval for w_0 is $\hat{w}_0 \pm 2 \times SE(\hat{w}_0)$

In our example (6.248,8.337)

LINEAR REGRESSION

Accuracy of the model: RSE measure

```
import statsmodels.api as sm
X_train_sm = sm.add_constant(X_train)
lr_sm = sm.OLS(y_train, X_train_sm)
lr_sm_result = lr_sm.fit()
lr_sm_result.summary()
```

LINEAR REGRESSION

Accuracy of the model: RSE measure

```
import statsmodels.api as sm
X_train_sm = sm.add_constant(X_train)
lr_sm = sm.OLS(y_train, X_train_sm)
lr_sm_result = lr_sm.fit()
lr_sm_result.summary()
```

```
# Residual standard error
np.sqrt(lr_sm_result.scale)
```

3.2789684108159634

Residual standard error -
Average amount that the response will deviate from true regression line:

$$RSE = \sqrt{\frac{RSS}{n - 2}}$$

LINEAR REGRESSION

Accuracy of the model: R^2 statistic

```
X = df['TV'].values.reshape(-1, 1)
y = df.sales

X_train,X_test, y_train, y_test = \
    train_test_split(X,y, train_size=0.8, random_state=0)

# Fit one feature
lr = linear_model.LinearRegression()
lr.fit(X_train, y_train)
```

Back to lr model using
scikit learn

LINEAR REGRESSION

Accuracy of the model: R^2 statistic

```
X = df['TV'].values.reshape(-1, 1)
y = df.sales

X_train,X_test, y_train, y_test = \
    train_test_split(X,y, train_size=0.8, random_state=0)

# Fit one feature
lr = linear_model.LinearRegression()
lr.fit(X_train, y_train)

r2 = lr.score(X_train, y_train)
print('R-squared: ', r2)
```

R-squared: 0.5884742462828709

Fraction of Variability of Y
explained using X .

$$R^2 = 1 - \frac{\sum_i (y_i - f(x_i))^2}{\sum_i (y_i - \bar{y})^2}$$

Fraction of
Variance
Unexplained

LINEAR REGRESSION

Accuracy of the model: R^2 statistic

```
X = df['TV'].values.reshape(-1, 1)
y = df.sales

X_train,X_test, y_train, y_test = \
    train_test_split(X,y, train_size=0.8, random_state=0)
```

```
# Fit one feature
lr = linear_model.LinearRegression()
lr.fit(X_train, y_train)
```

```
r2 = lr.score(X_train, y_train)
print('R-squared: ', r2)
```

R-squared: 0.5884742462828709

```
np.corrcoef(X_train[:,0], y_train.to_numpy())
```

```
array([[1.          , 0.76712075],
       [0.76712075, 1.        ]])
```

R^2 same as square of correlation coefficient in simple linear regression.

MULTIPLE LINEAR REGRESSION

Not just one feature x , but instead m features: x_1, x_2, \dots, x_m

Ex: Predicting **sales** from 2 features and 3 features.

MULTIPLE LINEAR REGRESSION

Not just one feature x , but instead m features: x_1, x_2, \dots, x_m

Ex: Predicting sales from 2 features and 3 features.

df.head()

	TV	radio	newspaper	sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

Let's start with 3 features.

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2 + \hat{w}_3 x_3$$

Predicted sales Intercept TV radio newspaper

The diagram illustrates the multiple linear regression equation $\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2 + \hat{w}_3 x_3$. A green arrow points from the term \hat{w}_0 to the 'Predicted sales' label. A red arrow points from the term $\hat{w}_1 x_1$ to the 'Intercept' label. Green arrows point from the terms $\hat{w}_2 x_2$ and $\hat{w}_3 x_3$ to the 'TV', 'radio', and 'newspaper' columns respectively in the dataset table.

MULTIPLE LINEAR REGRESSION

Not just one feature x , but instead m features: x_1, x_2, \dots, x_m

Ex: Predicting sales from 2 features and 3 features.

df.head()

	TV	radio	newspaper	sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

Let's start with 3 features.

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2 + \hat{w}_3 x_3$$

↑ ↗ ↗ ↗
Predicted sales Intercept TV radio newspaper

Ex: Say row 4 is not in training data and we want to predict it. $\hat{y}_4 = \hat{w}_0 + \hat{w}_1 180.8 + \hat{w}_2 10.8 + \hat{w}_3 58.4$ where $y_4 = 12.9$

MULTIPLE LINEAR REGRESSION

Apply 5-fold cross validation and predict from all 3 features.

```
# 5-fold cross validation
X = df[['TV', 'radio', 'newspaper']]
y = df.sales
X_train,X_test, y_train, y_test = \
    train_test_split(X,y, train_size=0.8, random_state=0)
cvs = cross_val_score(linear_model.LinearRegression(),
                      X_train, y_train, cv = 5, scoring='r2')
print('Mean R2: ', cvs.mean())
```

Mean R2: 0.8985145014718592

MULTIPLE LINEAR REGRESSION

Apply 5-fold cross validation and predict from all 3 features.

```
# 5-fold cross validation
X = df[['TV', 'radio', 'newspaper']]
y = df.sales
X_train,X_test, y_train, y_test = \
    train_test_split(X,y, train_size=0.8, random_state=0)
cvs = cross_val_score(linear_model.LinearRegression(),
                      X_train, y_train, cv = 5, scoring='r2')
print('Mean R2: ', cvs.mean())
```

Mean R2: 0.8985145014718592

Better R^2 than only using TV feature.

How about 2 features?

MULTIPLE LINEAR REGRESSION

Apply 5-fold cross validation and predict from TV and radio.

```
X = df[['TV', 'radio']]  
y = df.sales  
X_train,X_test, y_train, y_test = \  
    train_test_split(X,y, train_size=0.8, random_state=0)  
cvs = cross_val_score(linear_model.LinearRegression(),  
                      X_train, y_train, cv = 5, scoring='r2')  
print('Mean R2: ', cvs.mean())
```

Mean R2: 0.9010773153077544

MULTIPLE LINEAR REGRESSION

Apply 5-fold cross validation and predict from TV and radio.

```
X = df[['TV', 'radio']]  
y = df.sales  
X_train,X_test, y_train, y_test = \  
    train_test_split(X,y, train_size=0.8, random_state=0)  
cvs = cross_val_score(linear_model.LinearRegression(),  
                      X_train, y_train, cv = 5, scoring='r2')  
print('Mean R2: ', cvs.mean())
```

Mean R2: 0.9010773153077544

Better R^2 than using all 3 features.

Fit using these 2 features and evaluate on the test set with mean RSS .

MULTIPLE LINEAR REGRESSION

Fit using TV and radio. Evaluate on the test set with mean *RSS*.

```
X = df[['TV', 'radio']]
y = df.sales
X_train,X_test, y_train, y_test = \
    train_test_split(X,y, train_size=0.8, random_state=0)
lr_tr = linear_model.LinearRegression()
lr_tr.fit(X_train, y_train)
lr_tr_predict = lr_tr.predict(X_test)
# mse
lr_tr_mse = mean_squared_error(y_test,lr_tr_predict)
print('MSE : ',lr_tr_mse)
```

MSE : 4.391429763581883

MULTIPLE LINEAR REGRESSION

Dataset

x_{11}	x_{12}	\cdots	x_{1m}	y_1
x_{21}	x_{22}		x_{2m}	y_2
		\vdots		
x_{n1}	x_{n2}		x_{nm}	y_n

How is multiple linear regression formulated and solved in general?

Ex

```
df.head()
```

	TV	radio	newspaper	sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

MULTIPLE LINEAR REGRESSION

Dataset

x_{11}	x_{12}	\cdots	x_{1m}	y_1
x_{21}	x_{22}		x_{2m}	y_2
		\vdots		
x_{n1}	x_{n2}		x_{nm}	y_n

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_m x_m$$

$$\hat{y}_1 = w_0 + w_1 x_{11} + w_2 x_{12} + \cdots + w_m x_{1m}$$

$$\hat{y}_2 = w_0 + w_1 x_{21} + w_2 x_{22} + \cdots + w_m x_{2m}$$

\vdots

$$\hat{y}_n = w_0 + w_1 x_{n1} + w_2 x_{n2} + \cdots + w_m x_{nm}$$

MULTIPLE LINEAR REGRESSION

Dataset

x_{11}	x_{12}	\cdots	x_{1m}	y_1
x_{21}	x_{22}		x_{2m}	y_2
		\vdots		
x_{n1}	x_{n2}		x_{nm}	y_n

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_m x_m$$

$$\hat{y}_1 = w_0 + w_1 x_{11} + w_2 x_{12} + \cdots + w_m x_{1m}$$

$$\hat{y}_2 = w_0 + w_1 x_{21} + w_2 x_{22} + \cdots + w_m x_{2m}$$

\vdots

$$\hat{y}_n = w_0 + w_1 x_{n1} + w_2 x_{n2} + \cdots + w_m x_{nm}$$

How to turn it into nice matrix form?

MULTIPLE LINEAR REGRESSION

\hat{y}	X					w
\hat{y}_1	1	x_{11}	x_{12}	\cdots	x_{1m}	w_0
\hat{y}_2	1	x_{21}	x_{22}		x_{2m}	w_1
	1			\vdots		
\hat{y}_n	1	x_{n1}	x_{n2}		x_{nm}	w_m

$$\hat{y} = w_0 + w_1x_1 + w_2x_2 + \cdots + w_mx_m$$

$$\hat{y}_1 = w_0 + w_1x_{11} + w_2x_{12} + \cdots + w_mx_{1m}$$

$$\hat{y}_2 = w_0 + w_1x_{21} + w_2x_{22} + \cdots + w_mx_{2m}$$

\vdots

$$\hat{y}_n = w_0 + w_1x_{n1} + w_2x_{n2} + \cdots + w_mx_{nm}$$

How to turn it into nice matrix form?

MULTIPLE LINEAR REGRESSION

$$\hat{y} = \begin{matrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{matrix} = \begin{matrix} X \\ \vdots \\ X \end{matrix} \begin{matrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{matrix}$$

	1	x_{11}	x_{12}	\cdots	x_{1m}	w_0
1	x_{21}	x_{22}			x_{2m}	w_1
1			\vdots			
1	x_{n1}	x_{n2}			x_{nm}	w_m

Note these are
the actual values

$$y = \begin{matrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{matrix}$$

MULTIPLE LINEAR REGRESSION

$$\hat{y} = \begin{matrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{matrix} = \begin{matrix} X \\ \vdots \\ X \end{matrix} \begin{matrix} w \\ \vdots \\ w \end{matrix}$$

The matrix X is defined as:

	1	x_{11}	x_{12}	\cdots	x_{1m}	w_0
1	1	x_{21}	x_{22}		x_{2m}	w_1
	1			\vdots		
1	1	x_{n1}	x_{n2}		x_{nm}	w_m

Note these are
the actual values

$$y = \begin{matrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{matrix}$$

We know X, y . How to find w minimizing:

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

MULTIPLE LINEAR REGRESSION

Method 1: Algebraic closed form formula.

Theorem: Vector w minimizing mean squared error is

$$(X^T X)^{-1} X^T y$$

Might be problematic in practice: Matrix inversion slow, prone to numerical stability. Hard to generalize to other optimization problems.

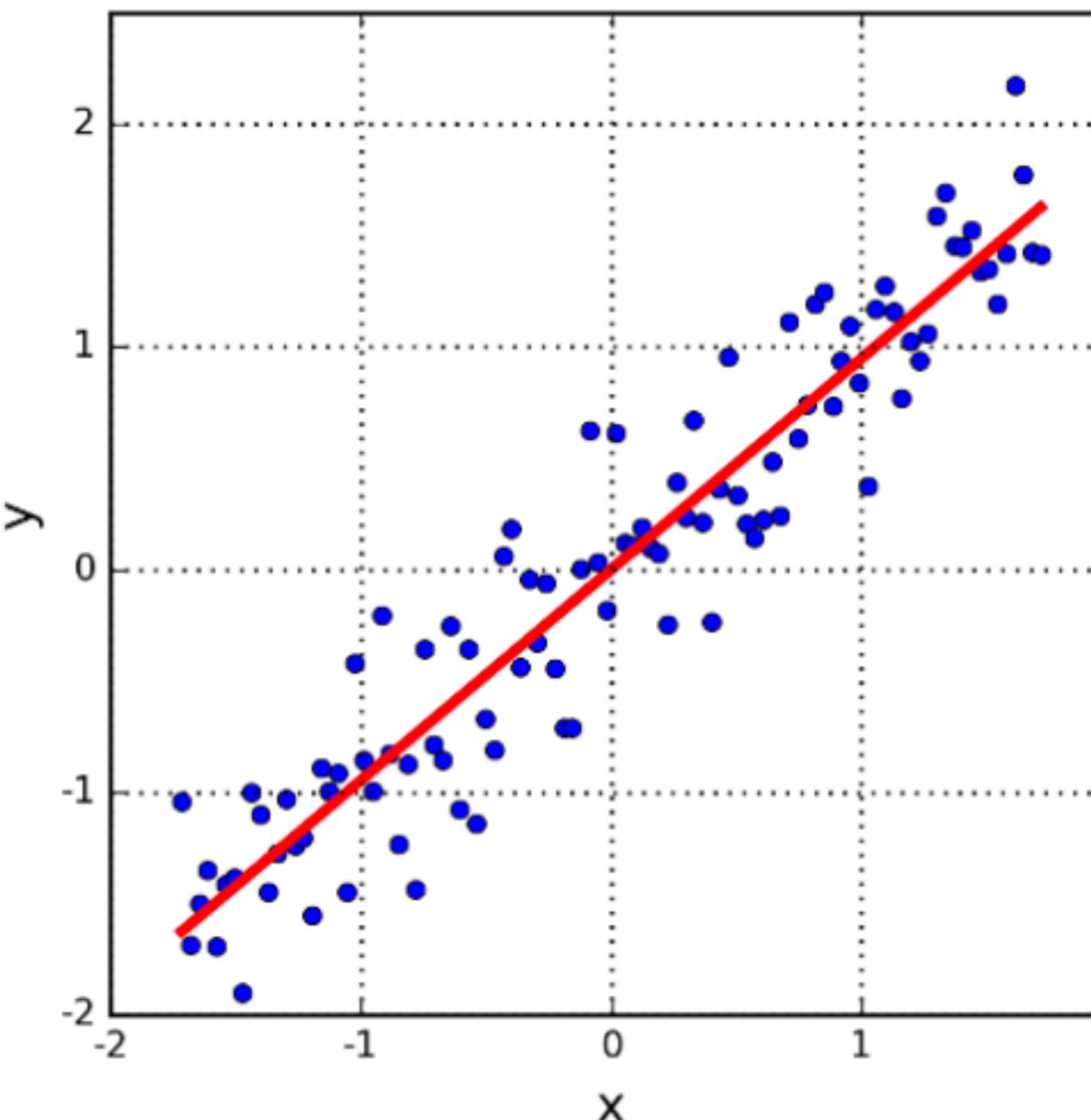
MULTIPLE LINEAR REGRESSION

Method 2: Parameter fitting via Gradient Descent.

Simple setting: Just one variable and intercept is 0: $\hat{y} = w_1x$

What does the squared error function

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$
 look like?



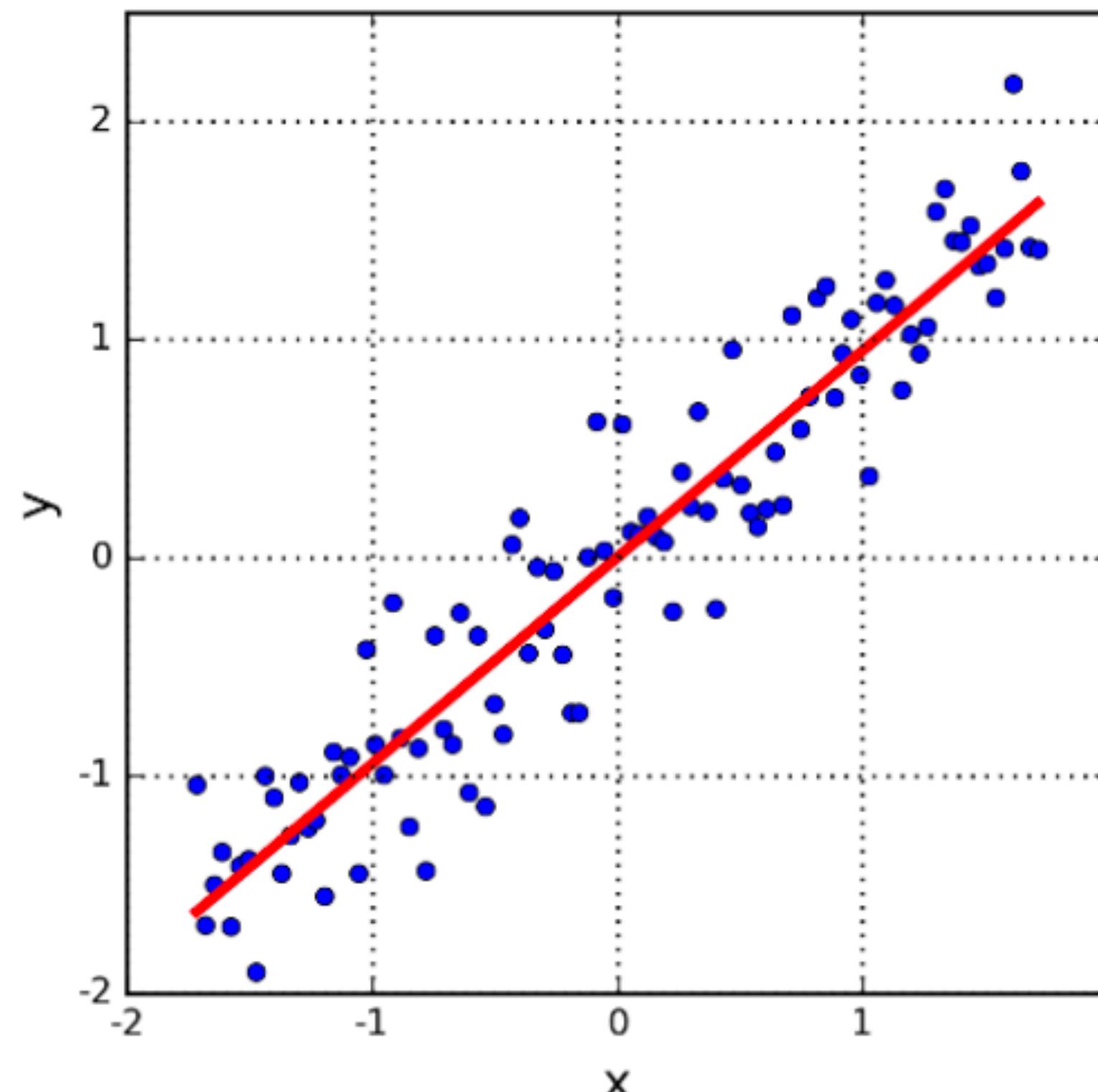
MULTIPLE LINEAR REGRESSION

Method 2: Parameter fitting via Gradient Descent.

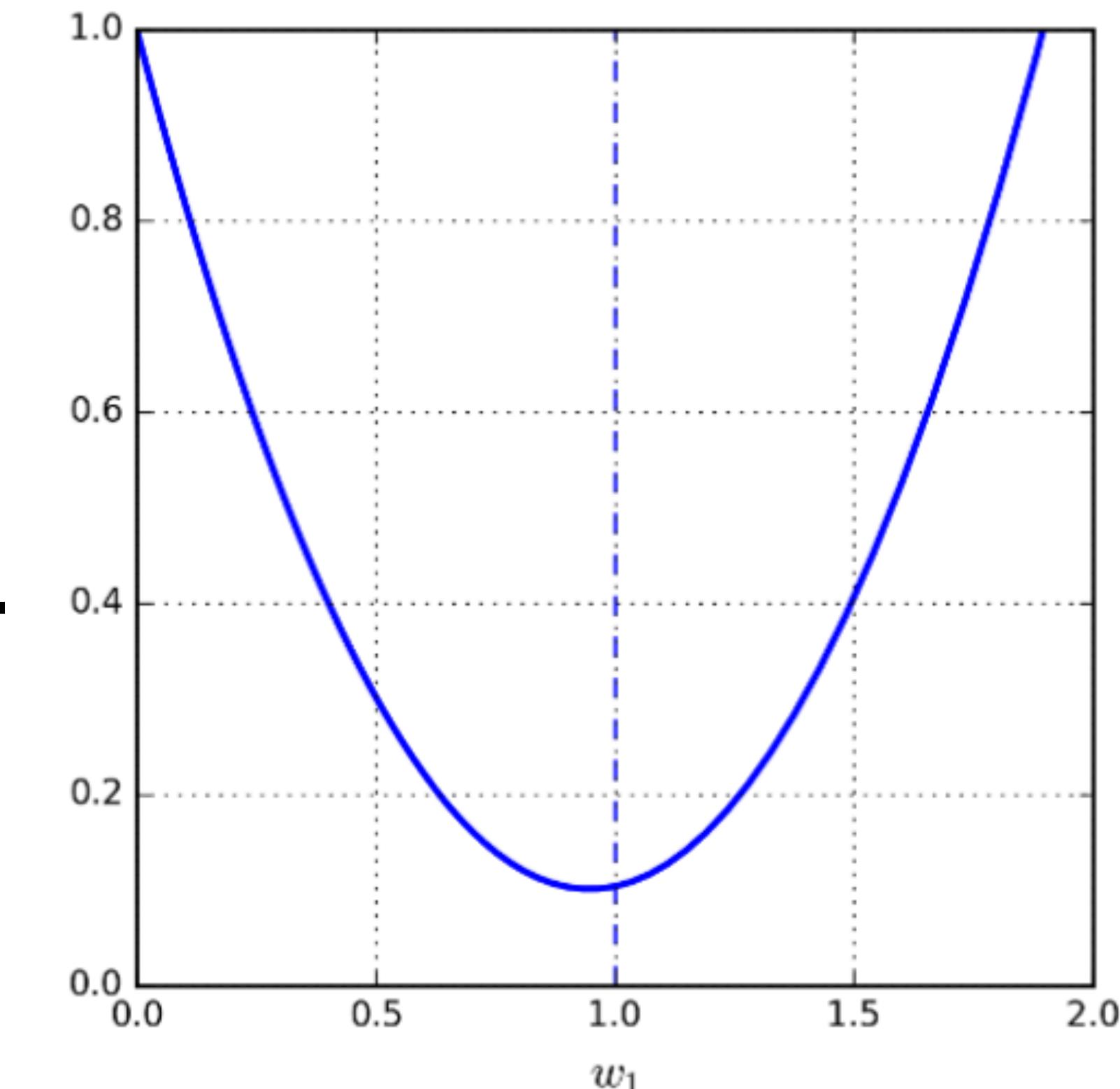
Simple setting: Just one variable and intercept is 0: $\hat{y} = w_1x$

What does the squared error function

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

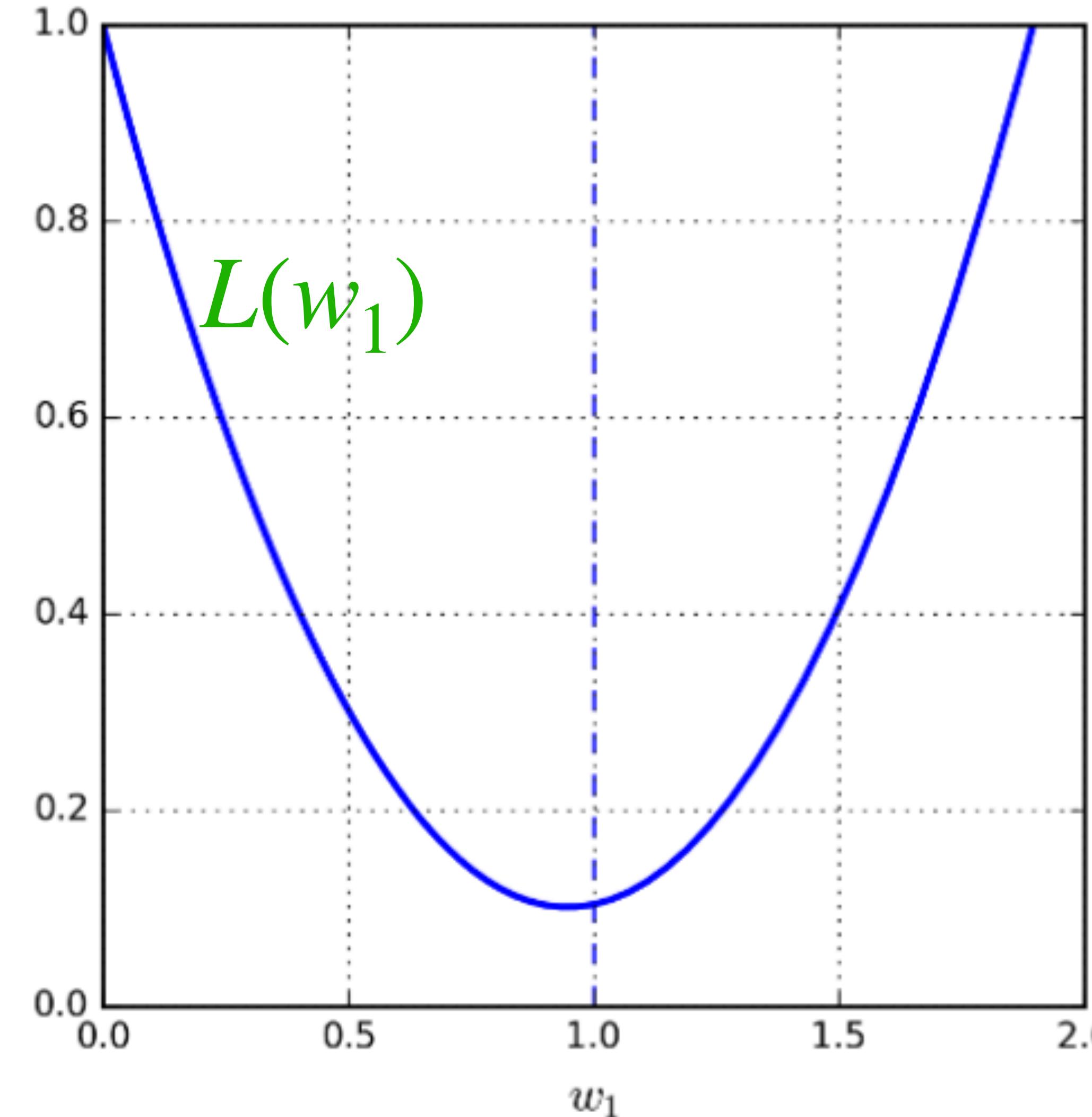


Quadratic function
of w_1 . Call it $L(w_1)$.
Easy to find the
minimum point.



MULTIPLE LINEAR REGRESSION

Method 2: Parameter fitting via Gradient Descent.

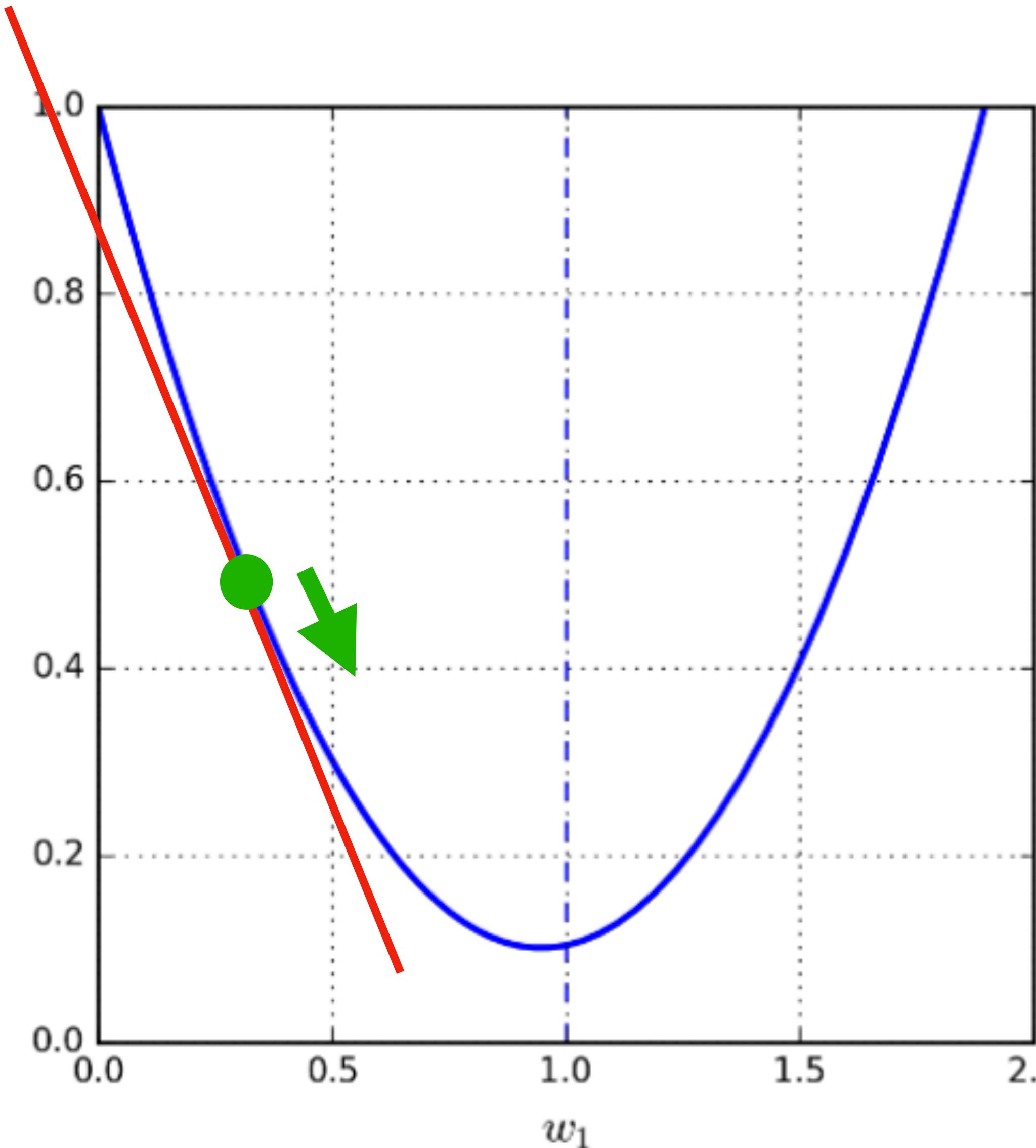


Start somewhere
move wrt to the
slope (derivative)

MULTIPLE LINEAR REGRESSION

Method 2: Parameter fitting via Gradient Descent.

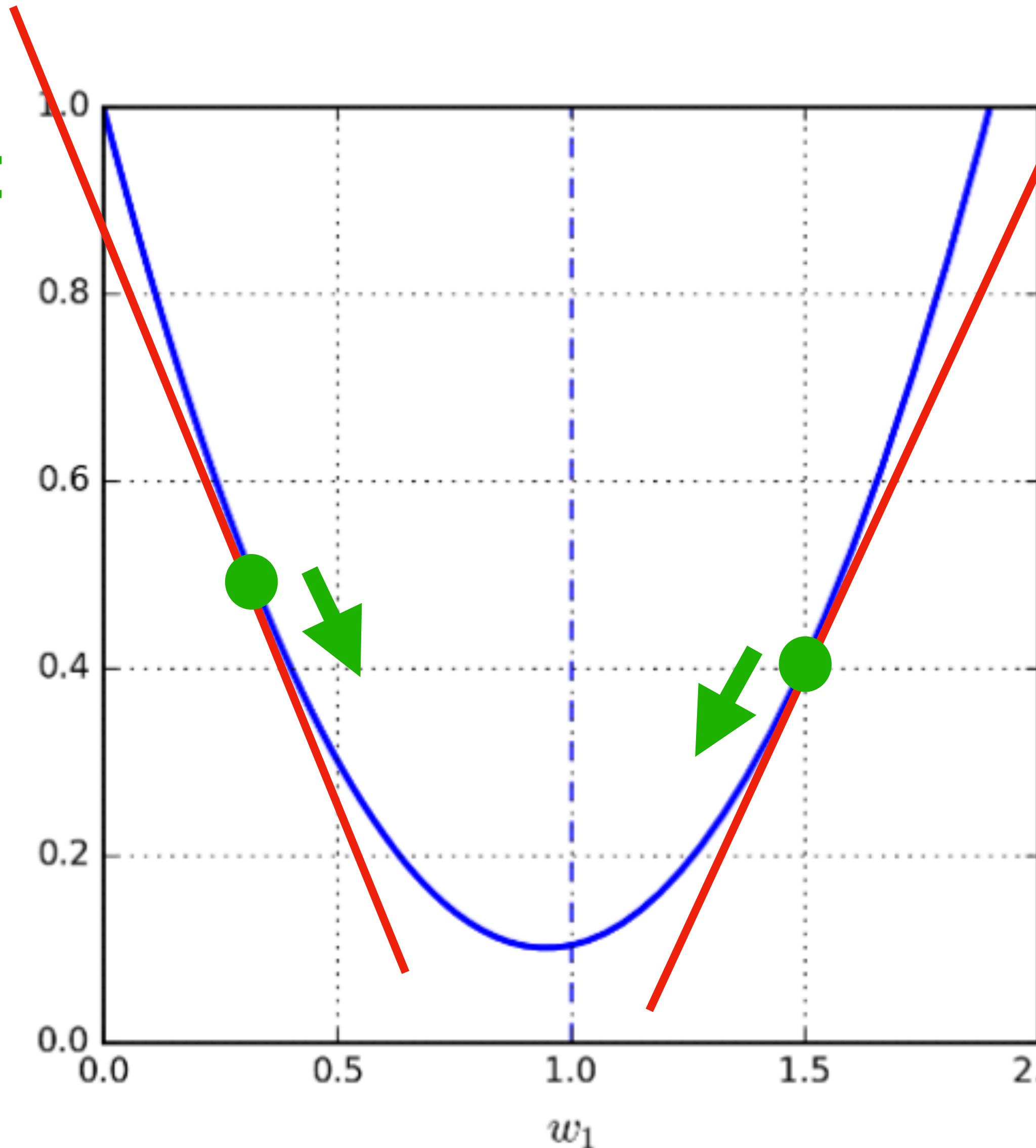
Negative slope:
Move in the
positive
direction.



MULTIPLE LINEAR REGRESSION

Method 2: Parameter fitting via Gradient Descent.

Negative slope:
Move in the
positive
direction.



Positive slope:
Move in the
negative
direction.

MULTIPLE LINEAR REGRESSION

Method 2: Parameter fitting via Gradient Descent.

Repeat until convergence:

$$w_1^{t+1} = w_1^t - \frac{d}{dw_1} L(w_1)$$

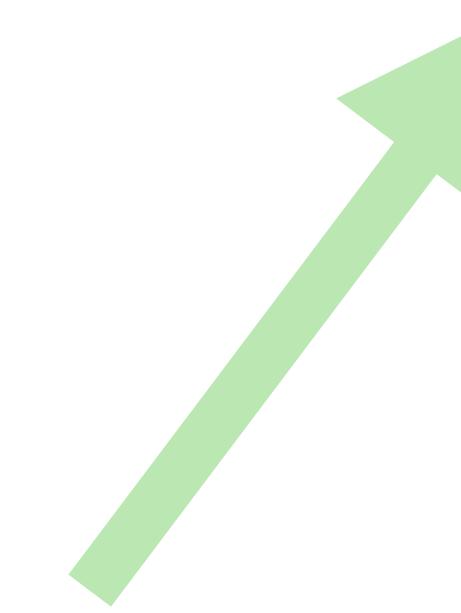
Not power! Value of
 w_1 after t iterations.

MULTIPLE LINEAR REGRESSION

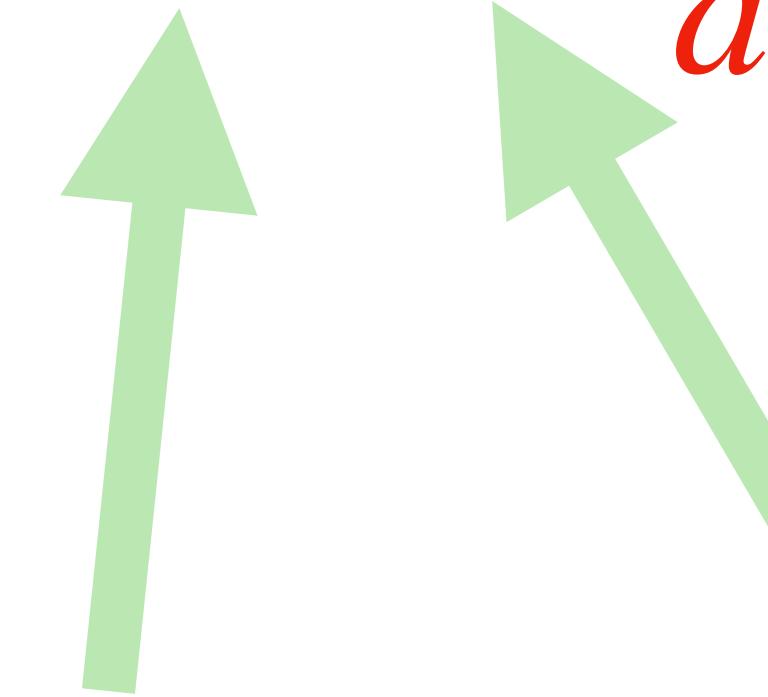
Method 2: Parameter fitting via Gradient Descent.

Repeat until convergence:

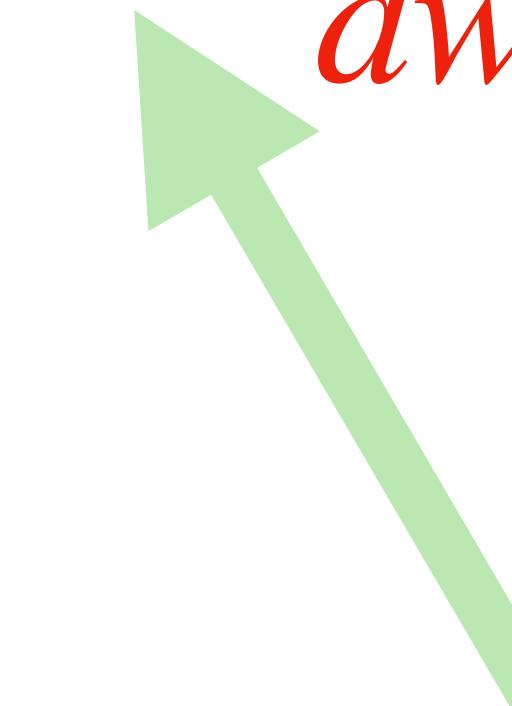
$$w_1^{t+1} = w_1^t - \frac{d}{dw_1} L(w_1)$$



Next guess for w



Current w



Move opposite to the slope

MULTIPLE LINEAR REGRESSION

Method 2: Parameter fitting via Gradient Descent.

Repeat until convergence:

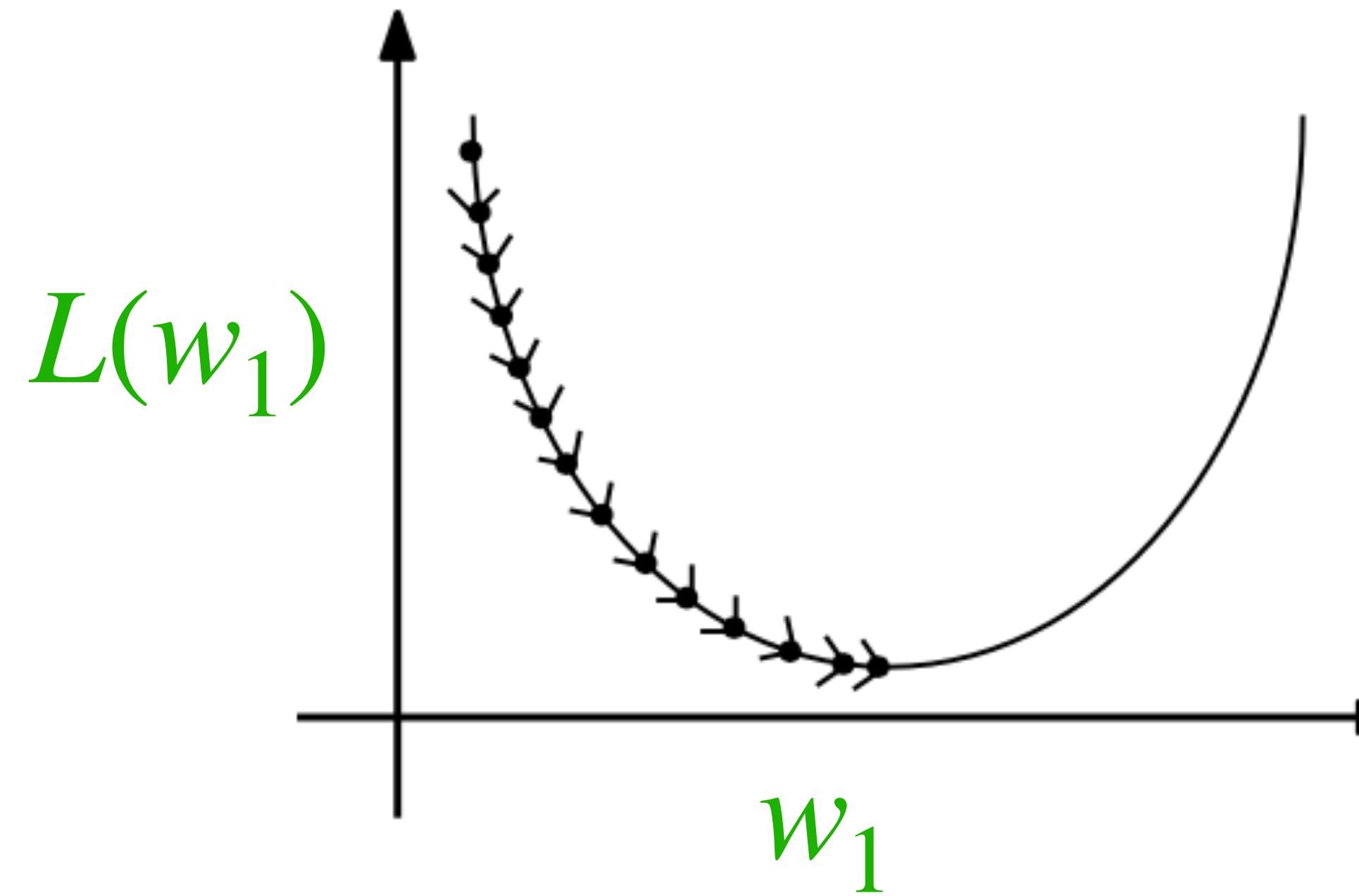
$$w_1^{t+1} = w_1^t - \alpha \frac{d}{dw_1} L(w_1)$$



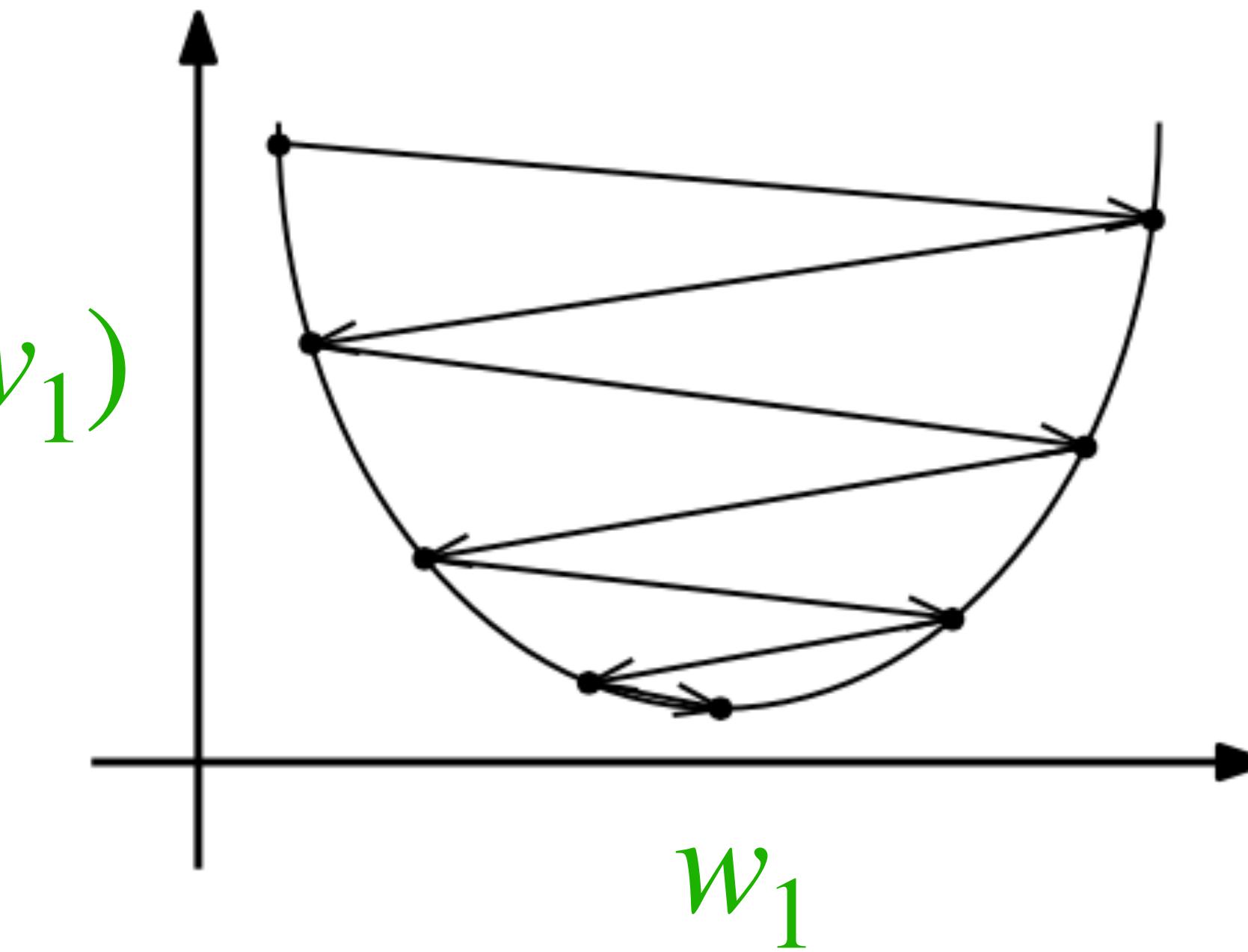
How large steps? Introduce α , learning rate, for step size.

MULTIPLE LINEAR REGRESSION

Method 2: Parameter fitting via Gradient Descent.



Small α may converge slowly.



Large α may overshoot.

MULTIPLE LINEAR REGRESSION

Method 2: Parameter fitting via Gradient Descent.

What happens with multiple dimensions (more than one)?

MULTIPLE LINEAR REGRESSION

Method 2: Parameter fitting via Gradient Descent.

What happens with multiple dimensions (more than one)?

Repeat until convergence:

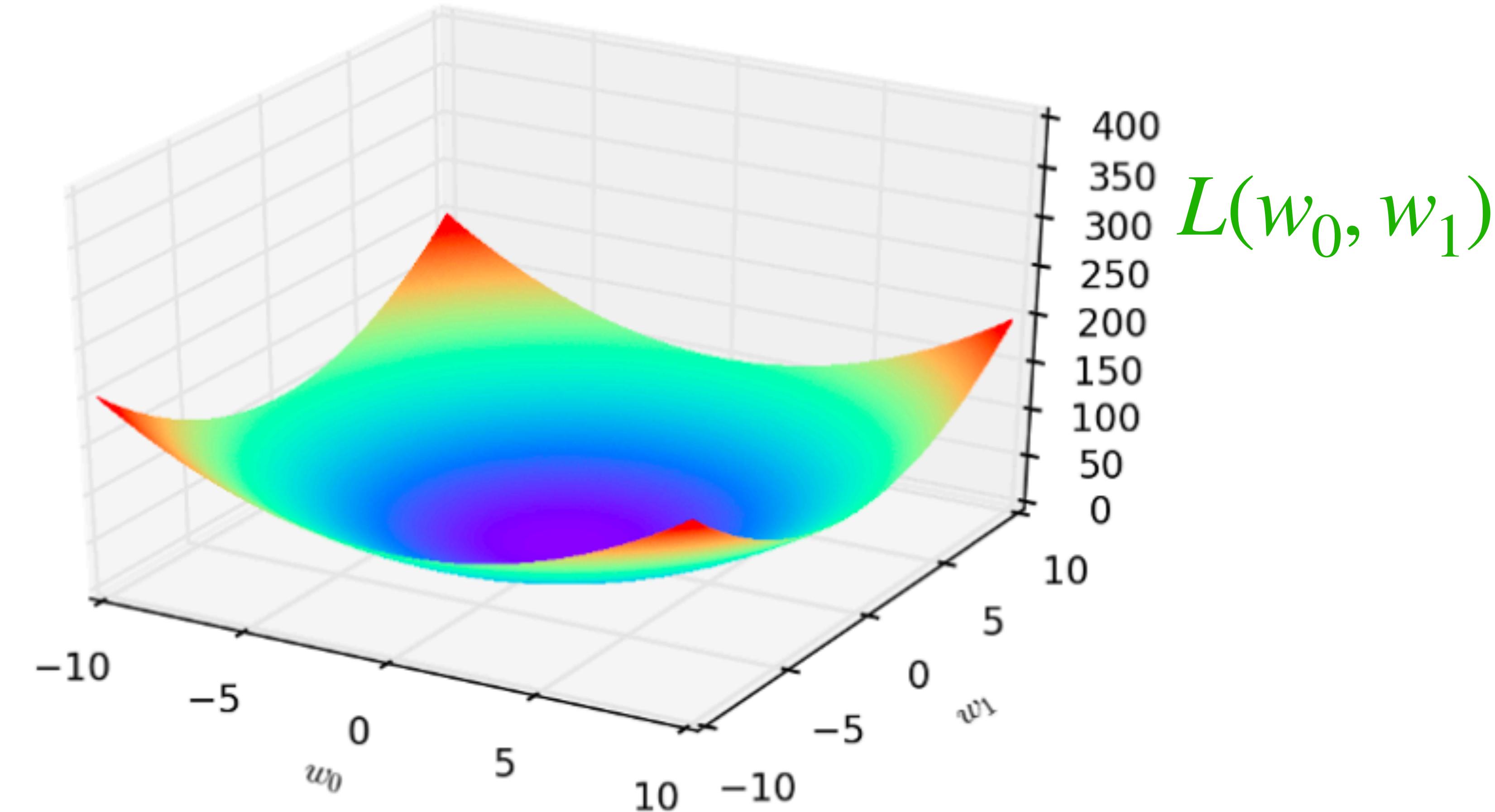
$$\begin{bmatrix} w_0^{t+1} \\ w_1^{t+1} \\ \vdots \\ w_m^{t+1} \end{bmatrix} = \begin{bmatrix} w_0^t \\ w_1^t \\ \vdots \\ w_m^t \end{bmatrix} - \alpha \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_m} \end{bmatrix}$$

Why does it still work?

MULTIPLE LINEAR REGRESSION

Method 2: Parameter fitting via Gradient Descent.

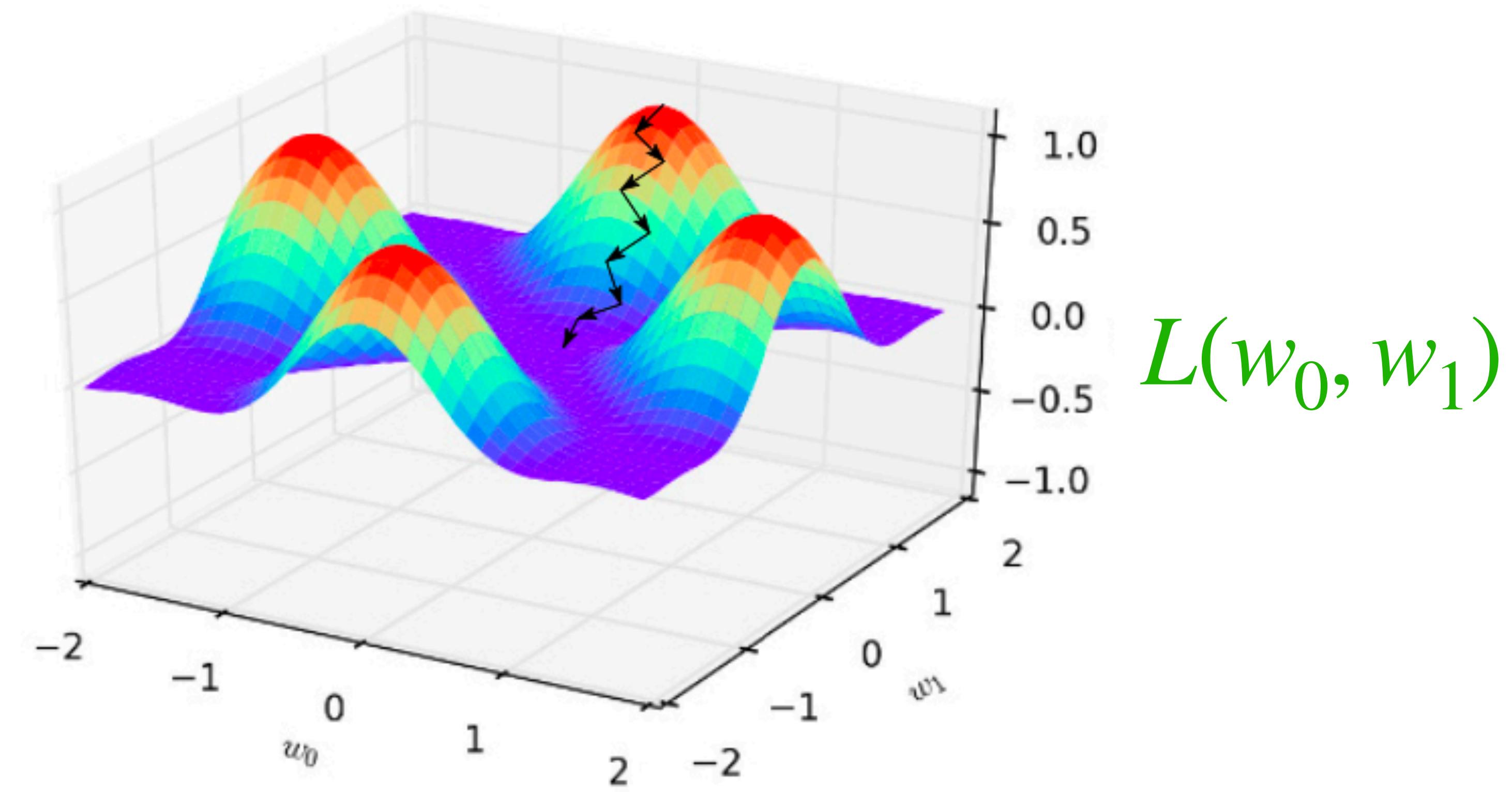
What happens with multiple dimensions (more than one)?



Why does it still work? Loss function remains **convex** no matter what m .

MULTIPLE LINEAR REGRESSION

Method 2: Parameter fitting via Gradient Descent.



If the function to optimize is not convex? Gradient descent may not find the global optimum, but will find some local minima.

MINI BATCH AND STOCHASTIC GRADIENT DESCENT

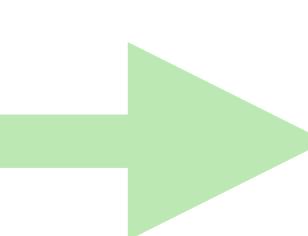
Problem with Method 2, Gradient Descent:

$$\begin{bmatrix} w_0^{t+1} \\ w_1^{t+1} \\ \vdots \\ \vdots \\ w_m^{t+1} \end{bmatrix} = \begin{bmatrix} w_0^t \\ w_1^t \\ \vdots \\ \vdots \\ w_m^t \end{bmatrix} - \alpha \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \\ \vdots \\ \vdots \\ \frac{\partial L}{\partial w_m} \end{bmatrix}$$

MINI BATCH AND STOCHASTIC GRADIENT DESCENT

Problem with Method 2, Gradient Descent:

$$\begin{bmatrix} w_0^{t+1} \\ w_1^{t+1} \\ \vdots \\ w_m^{t+1} \end{bmatrix} = \begin{bmatrix} w_0^t \\ w_1^t \\ \vdots \\ w_m^t \end{bmatrix} - \alpha \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_m} \end{bmatrix}$$



Each partial derivative hides a lot of computations in the form of a sum:

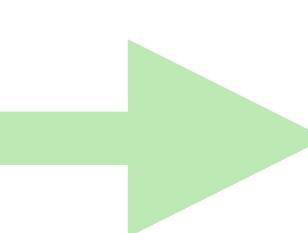
$$\frac{\partial L}{\partial w_j} = \frac{\partial}{\partial w_j} \sum_{i=1}^n (y_i - (w_0 + w_1x_{i1} + w_2x_{i2} + \cdots + w_mx_{im}))^2$$

For each dimension $j \Rightarrow$ go through all n training points.

MINI BATCH AND STOCHASTIC GRADIENT DESCENT

Problem with Method 2, Gradient Descent:

$$\begin{bmatrix} w_0^{t+1} \\ w_1^{t+1} \\ \vdots \\ w_m^{t+1} \end{bmatrix} = \begin{bmatrix} w_0^t \\ w_1^t \\ \vdots \\ w_m^t \end{bmatrix} - \alpha \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_m} \end{bmatrix}$$



Each partial derivative hides a lot of computations in the form of a sum:

$$\frac{\partial L}{\partial w_j} = \frac{\partial}{\partial w_j} \sum_{i=1}^n (y_i - (w_0 + w_1x_{i1} + w_2x_{i2} + \cdots + w_mx_{im}))^2$$

For each dimension $j \Rightarrow$ go through all n training points.

Computationally expensive if too many training points.

Solution: Mini-batch or stochastic gradient descent.

MINI BATCH AND STOCHASTIC GRADIENT DESCENT

Mini-batch gradient descent

Use a subset of the data at each update.

MINI BATCH AND STOCHASTIC GRADIENT DESCENT

Mini-batch gradient descent

Use a subset of the data at each update.

One training epoch:

Compute gradient on the first x% of the data. Update w_j , for $\forall w_j$

Compute gradient on the next x% of the data. Update w_j , for $\forall w_j$

...

Compute gradient on the last x% of the data. Update w_j , for $\forall w_j$

MINI BATCH AND STOCHASTIC GRADIENT DESCENT

Mini-batch gradient descent

Use a subset of the data at each update.

One training epoch:

Compute gradient on the first x% of the data. Update w_j , for $\forall w_j$

Compute gradient on the next x% of the data. Update w_j , for $\forall w_j$

...

Compute gradient on the last x% of the data. Update w_j , for $\forall w_j$

In a single training epoch, we use every datapoint in the data once.

Several training epochs are performed as necessary.

MINI BATCH AND STOCHASTIC GRADIENT DESCENT

Stochastic gradient descent

Extreme case of mini-batch: Each subset is a single datapoint.

MINI BATCH AND STOCHASTIC GRADIENT DESCENT

Stochastic gradient descent

Extreme case of mini-batch: Each subset is a single datapoint.

One training epoch:

Compute gradient on the first datapoint. Update w_j , for $\forall w_j$

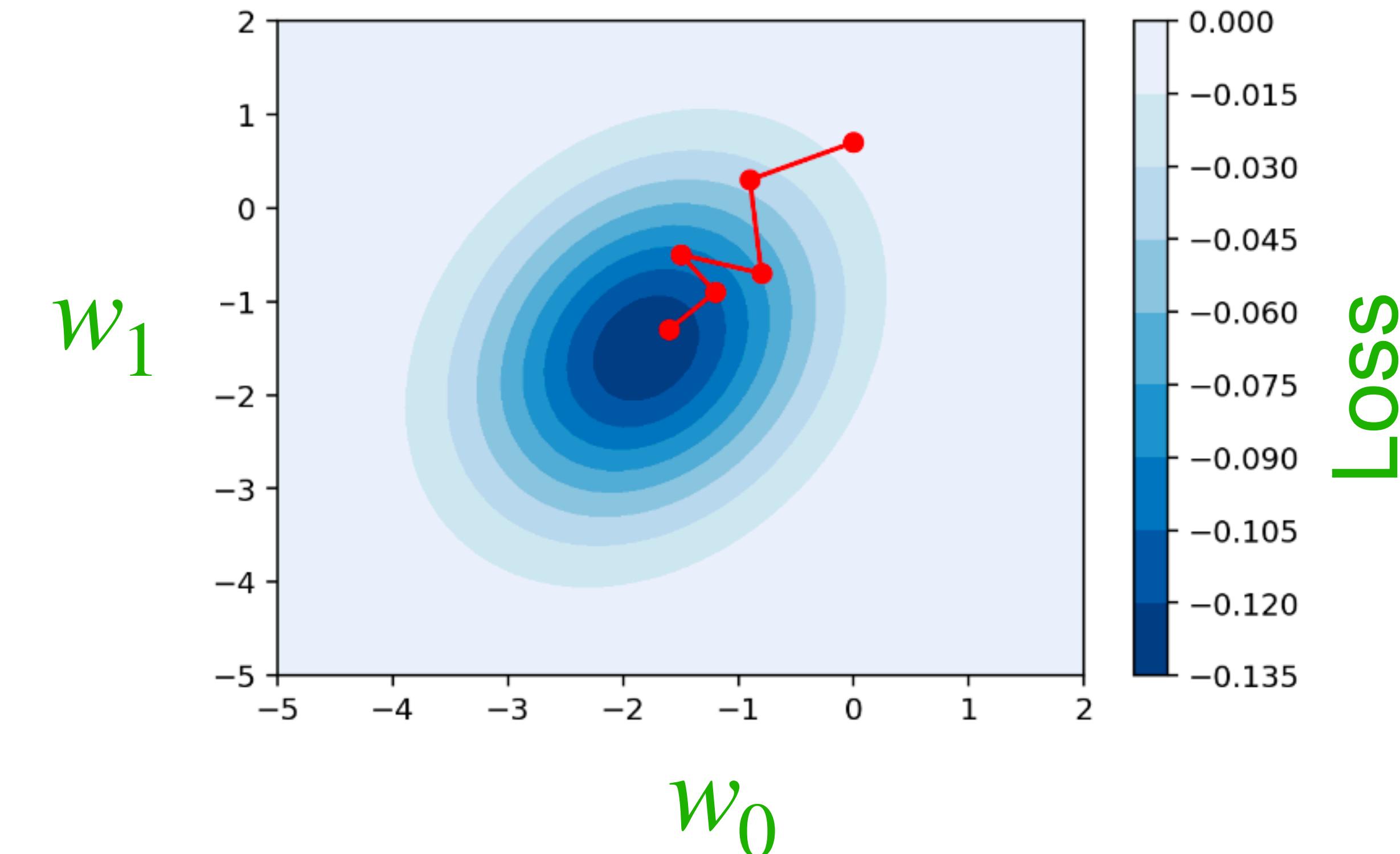
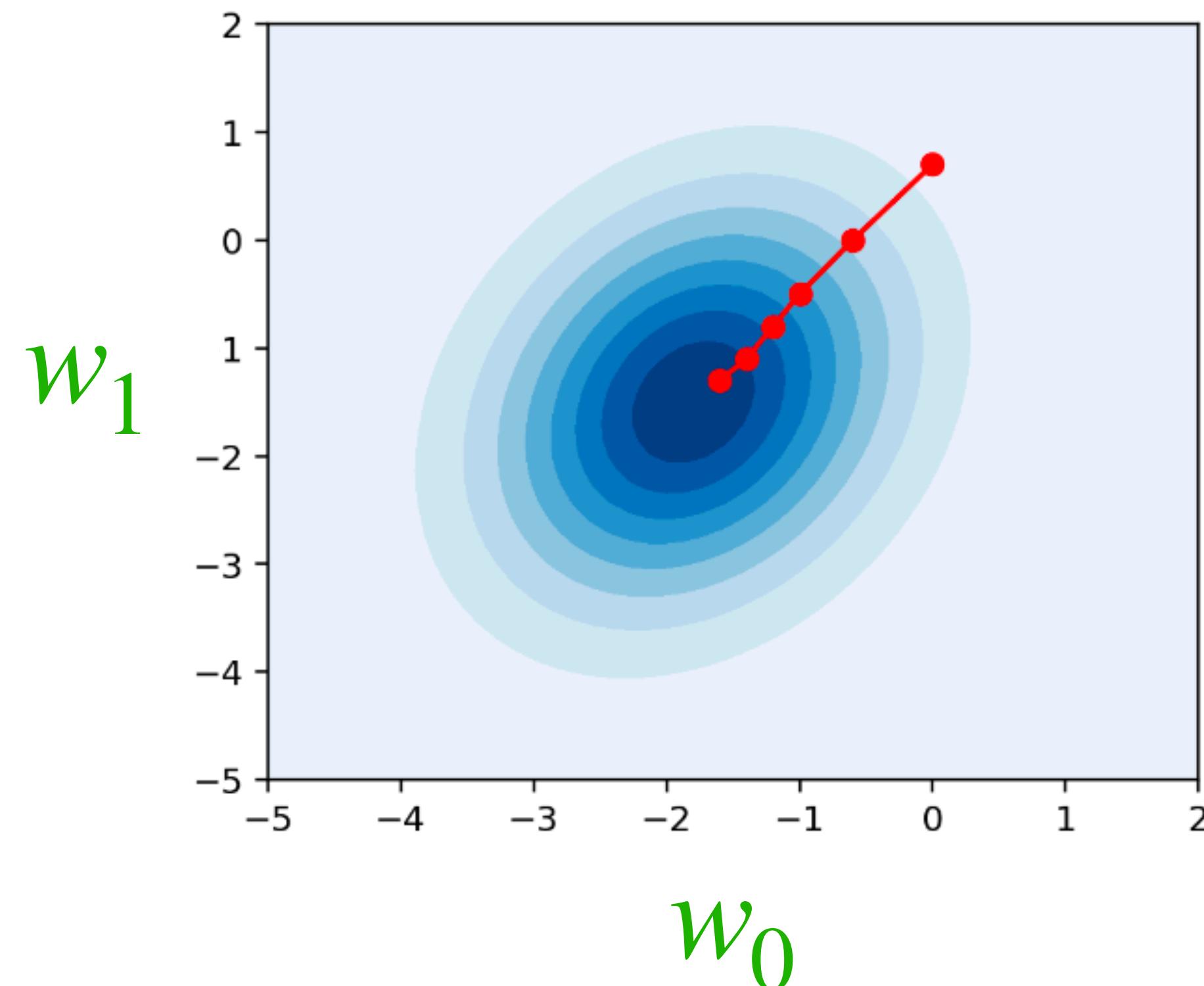
Compute gradient on the next datapoint. Update w_j , for $\forall w_j$

...

Compute gradient on the last datapoint. Update w_j , for $\forall w_j$

Several training epochs are performed as necessary.

MINI BATCH AND STOCHASTIC GRADIENT DESCENT



Gradient descent:

- True gradient
- Always descends toward true minimum loss.

Mini-batch/stochastic gradient descent:

- Approximation of true gradient
- May not descend toward true minimum loss with each update.

QUIZ

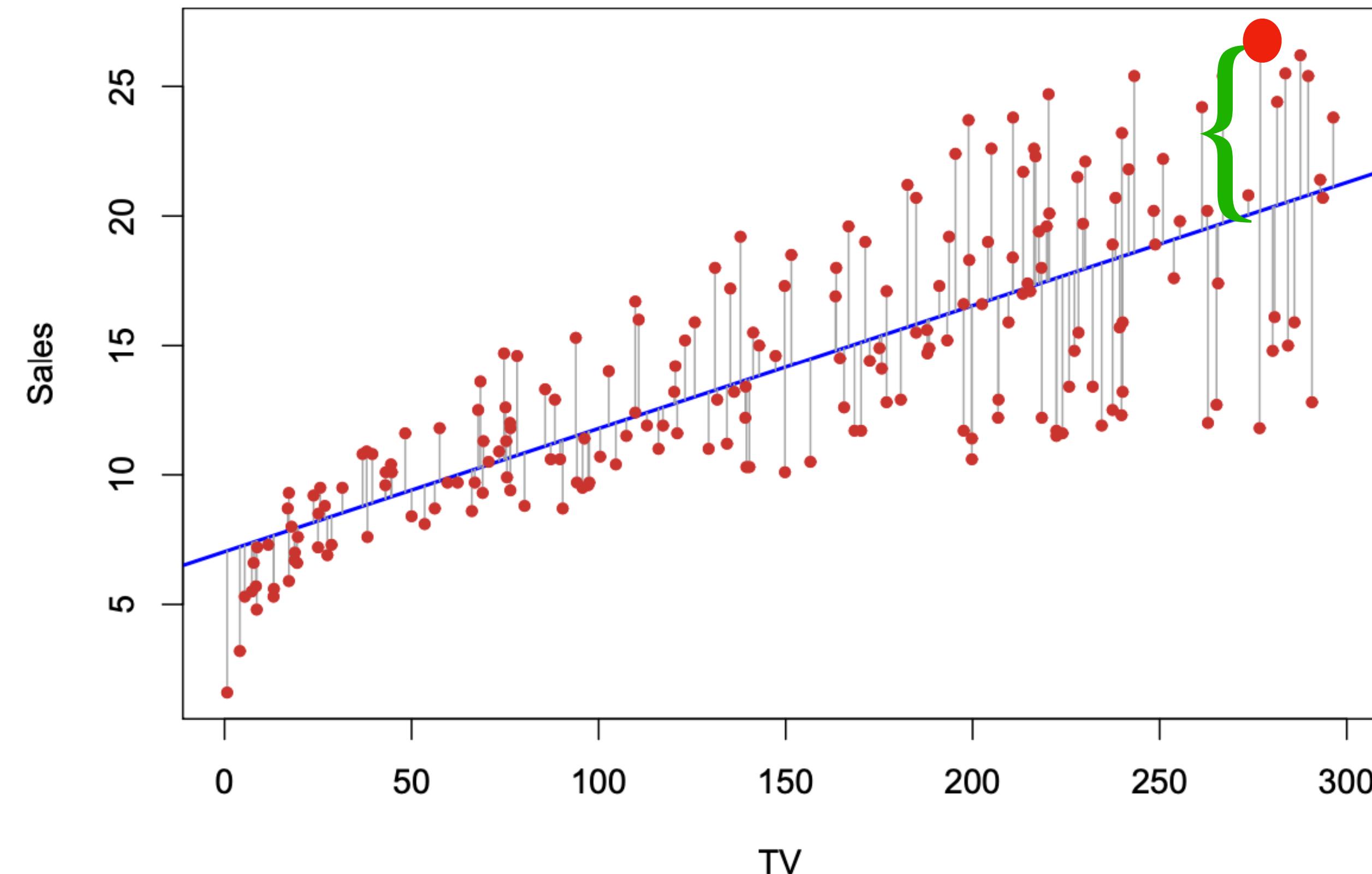
What is the objective function to minimize in linear regression?

REVIEW OF THE PREVIOUS LECTURE

Simple Linear Regression

'Best' fitting $\hat{y} = \hat{w}_0 + \hat{w}_1 x$:

$$\text{Minimize } RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



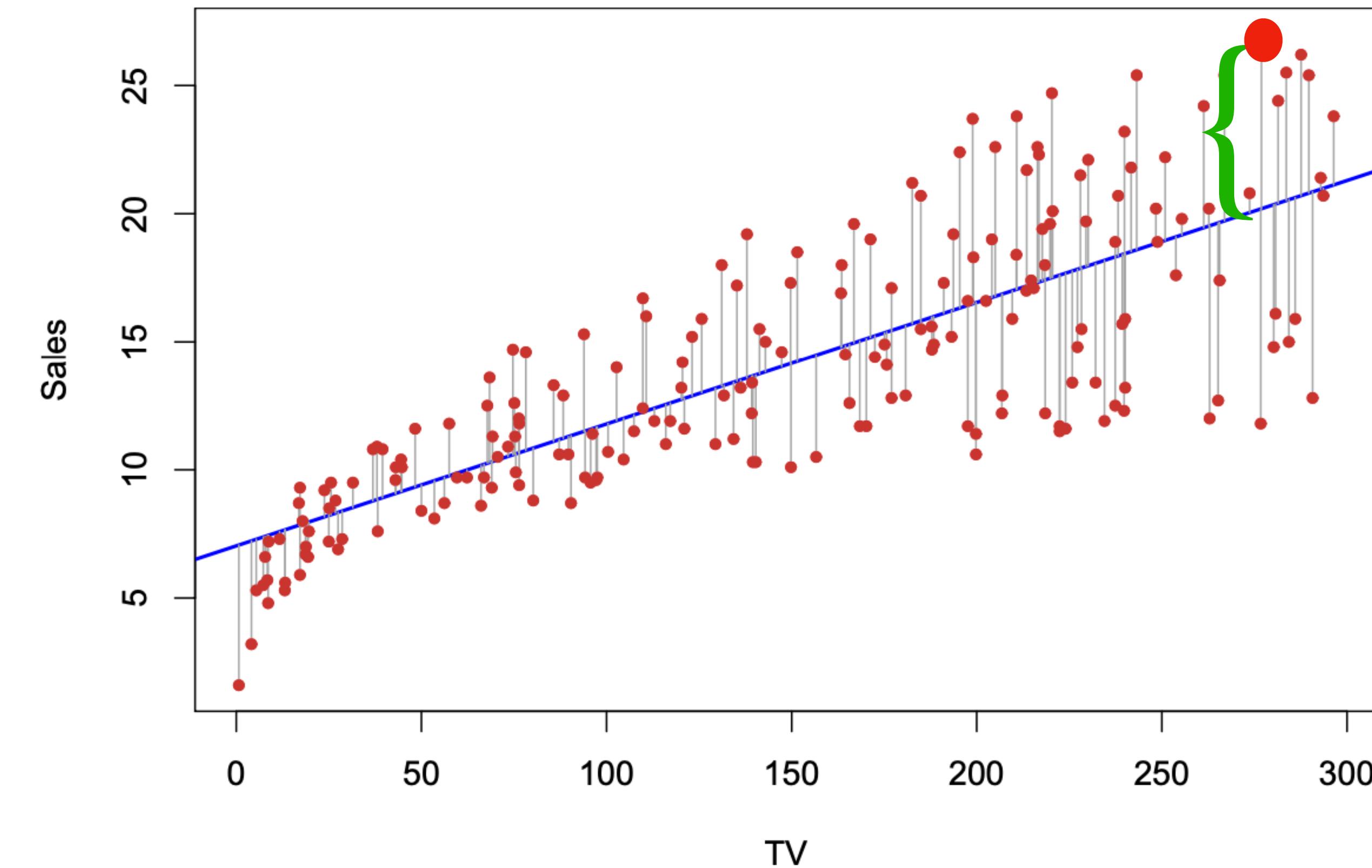
REVIEW OF THE PREVIOUS LECTURE

Simple Linear Regression

'Best' fitting $\hat{y} = \hat{w}_0 + \hat{w}_1 x$:

$$\text{Minimize } RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Accuracy of the model RSE, R^2



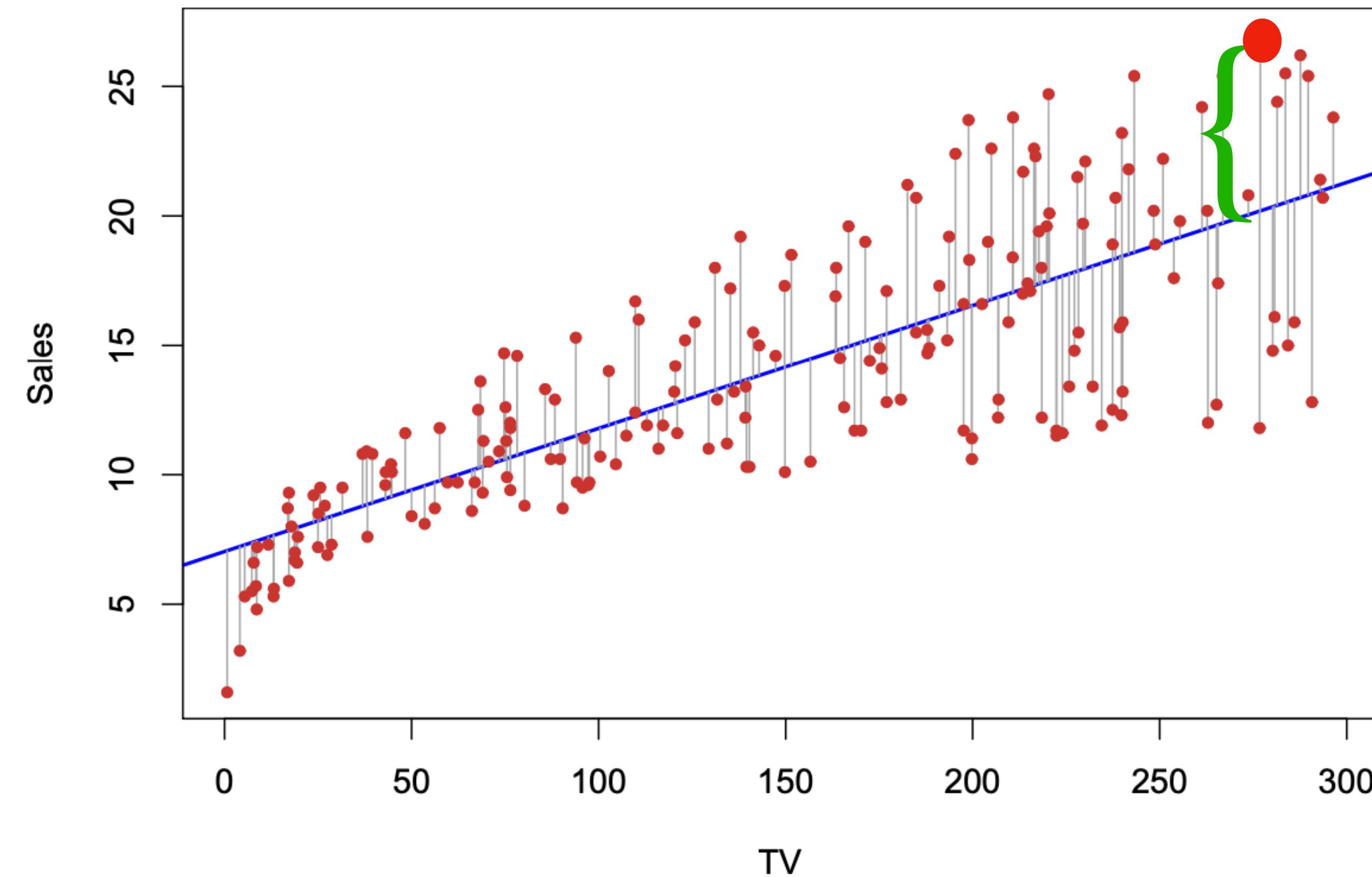
REVIEW OF THE PREVIOUS LECTURE

Simple Linear Regression

'Best' fitting $\hat{y} = \hat{w}_0 + \hat{w}_1 x$:

$$\text{Minimize } RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Accuracy of the model RSE, R^2



Multiple Linear Regression

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2 + \cdots + \hat{w}_m x_m$$

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Synergy/Interaction effect:

$$\widehat{\text{sales}} = \hat{w}_0 + \hat{w}_1 TV \quad R^2 \approx 0.58$$

$$\widehat{\text{sales}} = \hat{w}_0 + \hat{w}_1 TV + \hat{w}_2 radio \quad R^2 \approx 0.90$$

One unit increase in TV increases sales by \hat{w}_1 units.

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Synergy/Interaction effect:

$$\widehat{\text{sales}} = \hat{w}_0 + \hat{w}_1 TV \quad R^2 \approx 0.58$$

$$\widehat{\text{sales}} = \hat{w}_0 + \hat{w}_1 TV + \hat{w}_2 radio \quad R^2 \approx 0.90$$

One unit increase in TV increases sales by \hat{w}_1 units.

What if more $radio$ ads increase effectiveness of TV advertising?

Splitting \$100 between TV , $radio$ may be better than all on TV or all on $radio$.

Unit increase in TV affects sales in amount dependent also on $radio$.

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Synergy/Interaction effect:

How to achieve it?

Introduce a new interaction term as a product of the two.

$$\widehat{sales} = \hat{w}_0 + \hat{w}_1 \times TV + \hat{w}_2 \times radio + \hat{w}_3 \times TV \times radio$$

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Synergy/Interaction effect:

How to achieve it?

Introduce a new interaction term as a product of the two.

$$\widehat{\text{sales}} = \hat{w}_0 + \hat{w}_1 \times TV + \hat{w}_2 \times radio + \hat{w}_3 \times TV \times radio$$

$$= \hat{w}_0 + (\hat{w}_1 + \hat{w}_3 \times radio) \times TV + \hat{w}_2 \times radio$$

$$= \hat{w}_0 + \hat{w}'_1 \times TV + \hat{w}_2 \times radio$$

where $\hat{w}'_1 = \hat{w}_1 + \hat{w}_3 \times radio$

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Ex: Back to the Advertising data set.

```
# add new feature tv*radio to dataframe  
df['TV_radio'] = df['TV'] * df['radio']  
df.head()
```

	TV	radio	newspaper	sales	TV_radio
0	230.1	37.8	69.2	22.1	8697.78
1	44.5	39.3	45.1	10.4	1748.85
2	17.2	45.9	69.3	9.3	789.48
3	151.5	41.3	58.5	18.5	6256.95
4	180.8	10.8	58.4	12.9	1952.64



OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Train with TV, Radio, TV_radio. Measure R^2 .

```
# train/test with TV, radio, and new feature|  
X_new = df[['TV','radio', 'TV_radio']]  
X_train_new,X_test_new, _, _ = \  
    train_test_split(X_new,y, train_size=0.8, random_state=0)  
  
# linear regression using TV, radio, TV*radio  
lr_new = linear_model.LinearRegression()  
lr_new.fit(X_train_new, y_train)
```

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Train with TV, Radio, TV_radio. Measure R^2 .

```
# train/test with TV, radio, and new feature|  
X_new = df[['TV','radio', 'TV_radio']]  
X_train_new,X_test_new, _, _ = \  
    train_test_split(X_new,y, train_size=0.8, random_state=0)  
  
# linear regression using TV, radio, TV*radio  
lr_new = linear_model.LinearRegression()  
lr_new.fit(X_train_new, y_train)
```

w_0: 6.652712826674431

w_1, w_2, w_3: [0.0194162 0.03790871 0.0010505]

R-squared: 0.9732546974049714

Considerable
improvement!

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Non-linear Relationships:

True relationship between response and predictors may be non-linear.

How to extend? **Polynomial regression.**

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

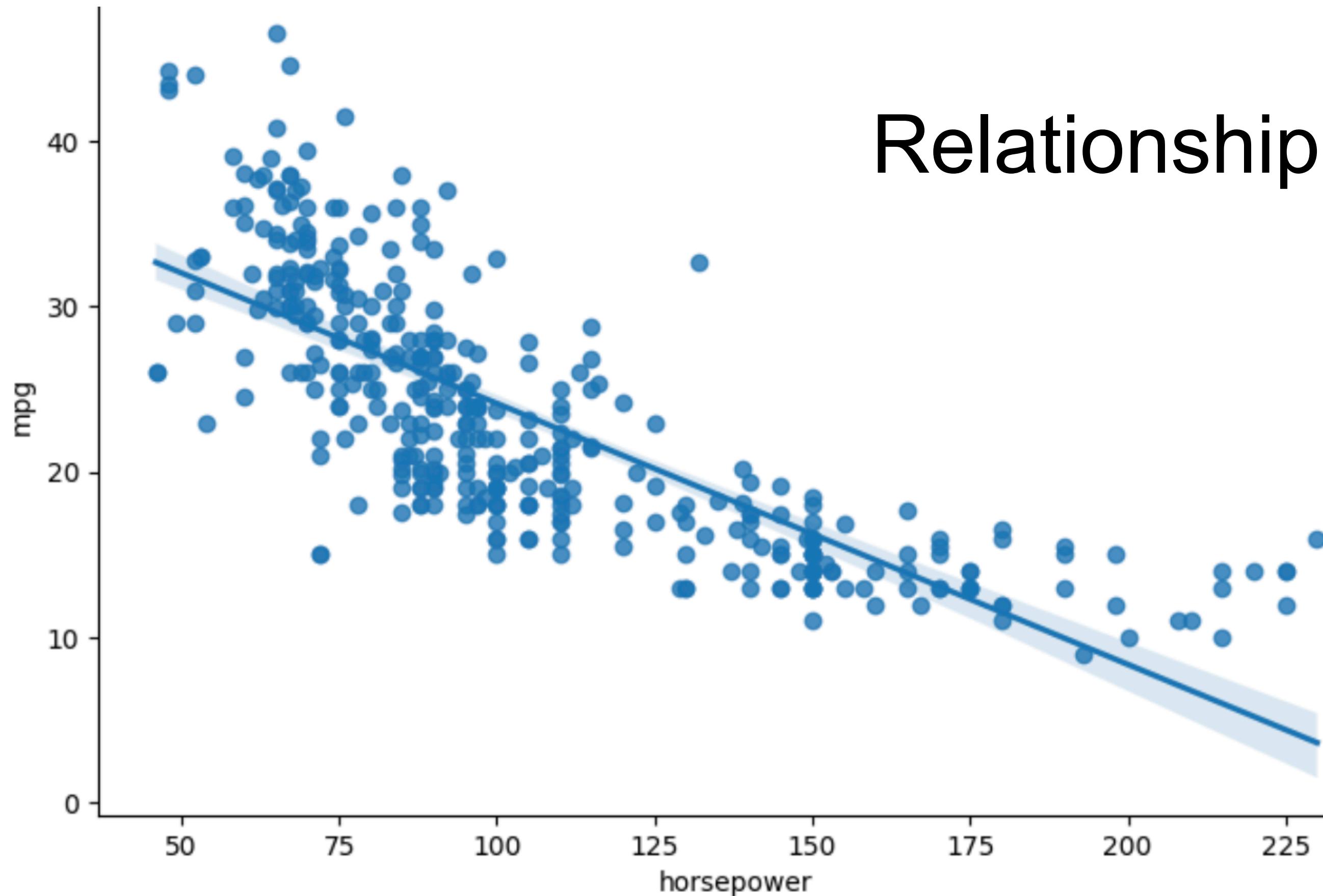
Ex: Auto data set.

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin		name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu	
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320	
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite	
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst	
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino	

Want to predict *mpg* from *horsepower*.

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

```
sns.pairplot(df, x_vars=['horsepower'], y_vars='mpg', height=5,  
             aspect =1.5, kind='reg')
```



Relationship looks quadratic rather than linear

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Train with *horsepower* only.

```
X = df[['horsepower']]
y = df.mpg
X_train,X_test, y_train, y_test = \
    train_test_split(X,y, train_size=0.8, random_state=0)
# linear regression using TV, radio
lr = linear_model.LinearRegression()
lr.fit(X_train, y_train)
```

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Train with *horsepower* only.

```
X = df[['horsepower']]
y = df.mpg
X_train,X_test, y_train, y_test = \
    train_test_split(X,y, train_size=0.8, random_state=0)
# linear regression using TV, radio
lr = linear_model.LinearRegression()
lr.fit(X_train, y_train)
```

w_0: 39.96223257893908

w_1: [-0.15844506]

R-squared: 0.5955625946891978

Doesn't look bad, but depends on the application area.

Possible improvement based on our previous observation?

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Add new feature $horsepower^2$ and train using the two.

```
# add new feature horsepower^2
df['horsepower2'] = df['horsepower']**2
# train/test with new feature, using same rows
X_2 = df[['horsepower', 'horsepower2']]
X_train_2,X_test_2, _, _ =
    train_test_split(X_2,y, train_size=0.8, random_state=0)

# linear regression using TV, radio, TV*radio
lr_2 = linear_model.LinearRegression()
lr_2.fit(X_train_2, y_train)
```

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Add new feature $horsepower^2$ and train using the two.

```
# add new feature horsepower^2
df['horsepower2'] = df['horsepower']**2
# train/test with new feature, using same rows
X_2 = df[['horsepower', 'horsepower2']]
X_train_2,X_test_2, _, _ =
    train_test_split(X_2,y, train_size=0.8, random_state=0)

# linear regression using TV, radio, TV*radio
lr_2 = linear_model.LinearRegression()
lr_2.fit(X_train_2, y_train)
```

$$mpg = \hat{w}_0 + \hat{w}_1 horsepower + \hat{w}_2 \times horsepower^2$$

(Polynomial regression)

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Add new feature $horsepower^2$ and train using the two.

```
# add new feature horsepower^2
df['horsepower2'] = df['horsepower']**2
# train/test with new feature, using same rows
X_2 = df[['horsepower', 'horsepower2']]
X_train_2, X_test_2, _, _ = \
    train_test_split(X_2, y, train_size=0.8, random_state=0)

# linear regression using TV, radio, TV*radio
lr_2 = linear_model.LinearRegression()
lr_2.fit(X_train_2, y_train)
```

w_0: 56.280426279112035

w_1, w_2: [-0.4546145 0.00118246]

R-squared: 0.673124476641667

Considerable improvement!

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Qualitative Predictors

So far all features in our linear regression were quantitative.

How to extend in case of quantitative features? One hot encoding

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Ex: Carseats data set.

	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc
0	9.50	138	73	11	276	120	Bad
1	11.22	111	48	16	260	83	Good
2	10.06	113	35	10	269	80	Medium
3	7.40	117	100	4	466	97	Medium
4	4.15	141	64	3	340	128	Bad

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Ex: Carseats data set.

	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc
0	9.50	138	73	11	276	120	Bad
1	11.22	111	48	16	260	83	Good
2	10.06	113	35	10	269	80	Medium
3	7.40	117	100	4	466	97	Medium
4	4.15	141	64	3	340	128	Bad

Want to predict *Sales*. All features are numeric other than *ShelveLoc*.

How to use *ShelveLoc* in our prediction?

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Introduce dummy features as necessary.

```
from sklearn.preprocessing import OneHotEncoder
enc = OneHotEncoder(drop='first') # drop the first dummy variable
enc_df = pd.DataFrame(enc.fit_transform(df[['ShelveLoc']]).toarray())
enc_df.columns = ['Good', 'Medium']
df = df.join(enc_df)
df.head()
```

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Introduce dummy features as necessary.

```
from sklearn.preprocessing import OneHotEncoder
enc = OneHotEncoder(drop='first') # drop the first dummy variable
enc_df = pd.DataFrame(enc.fit_transform(df[['ShelveLoc']]).toarray())
enc_df.columns = ['Good', 'Medium']
df = df.join(enc_df)
df.head()
```

	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Good	Medium
0	9.50	138	73	11	276	120	Bad	0.0	0.0
1	11.22	111	48	16	260	83	Good	1.0	0.0
2	10.06	113	35	10	269	80	Medium	0.0	1.0
3	7.40	117	100	4	466	97	Medium	0.0	1.0
4	4.15	141	64	3	340	128	Bad	0.0	0.0

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Introduce dummy features as necessary.

```
from sklearn.preprocessing import OneHotEncoder  
enc = OneHotEncoder(drop='first') # drop the first dummy variable  
enc_df = pd.DataFrame(enc.fit_transform(df[['ShelveLoc']]).toarray())  
enc_df.columns = ['Good', 'Medium']  
df = df.join(enc_df)  
df.head()
```

Why just *Good, Medium*?

	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Good	Medium
0	9.50	138	73	11	276	120	Bad	0.0	0.0
1	11.22	111	48	16	260	83	Good	1.0	0.0
2	10.06	113	35	10	269	80	Medium	0.0	1.0
3	7.40	117	100	4	466	97	Medium	0.0	1.0
4	4.15	141	64	3	340	128	Bad	0.0	0.0

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Introduce dummy features as necessary.

```
from sklearn.preprocessing import OneHotEncoder  
enc = OneHotEncoder(drop='first') # drop the first dummy variable  
enc_df = pd.DataFrame(enc.fit_transform(df[['ShelveLoc']]).toarray())  
enc_df.columns = ['Good', 'Medium']  
df = df.join(enc_df)  
df.head()
```

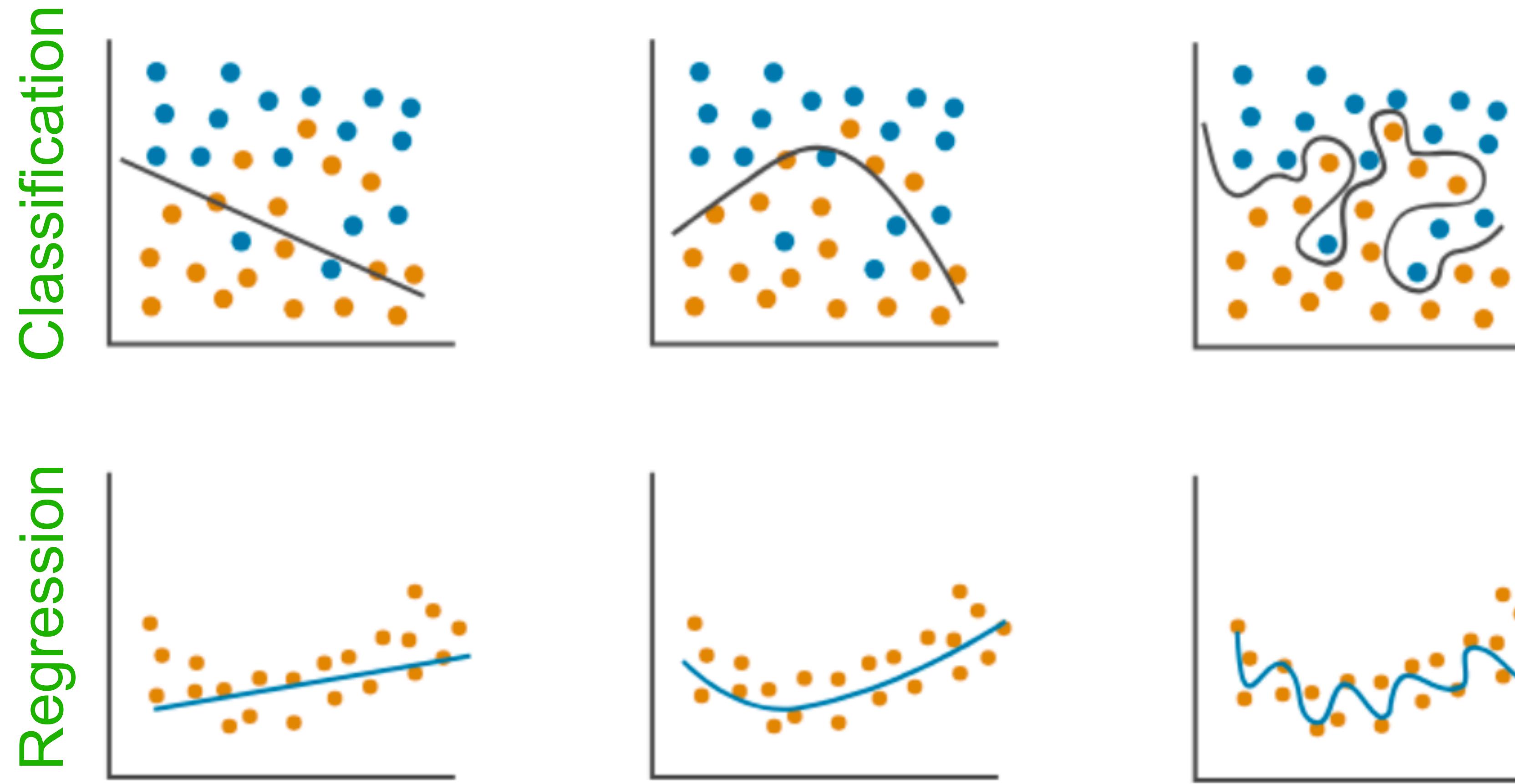
Why just *Good, Medium*?

Intercept would be linear combination of dummy features.

	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Good	Medium
0	9.50	138	73	11	276	120	Bad	0.0	0.0
1	11.22	111	48	16	260	83	Good	1.0	0.0
2	10.06	113	35	10	269	80	Medium	0.0	1.0
3	7.40	117	100	4	466	97	Medium	0.0	1.0
4	4.15	141	64	3	340	128	Bad	0.0	0.0

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Restricting Model Complexity

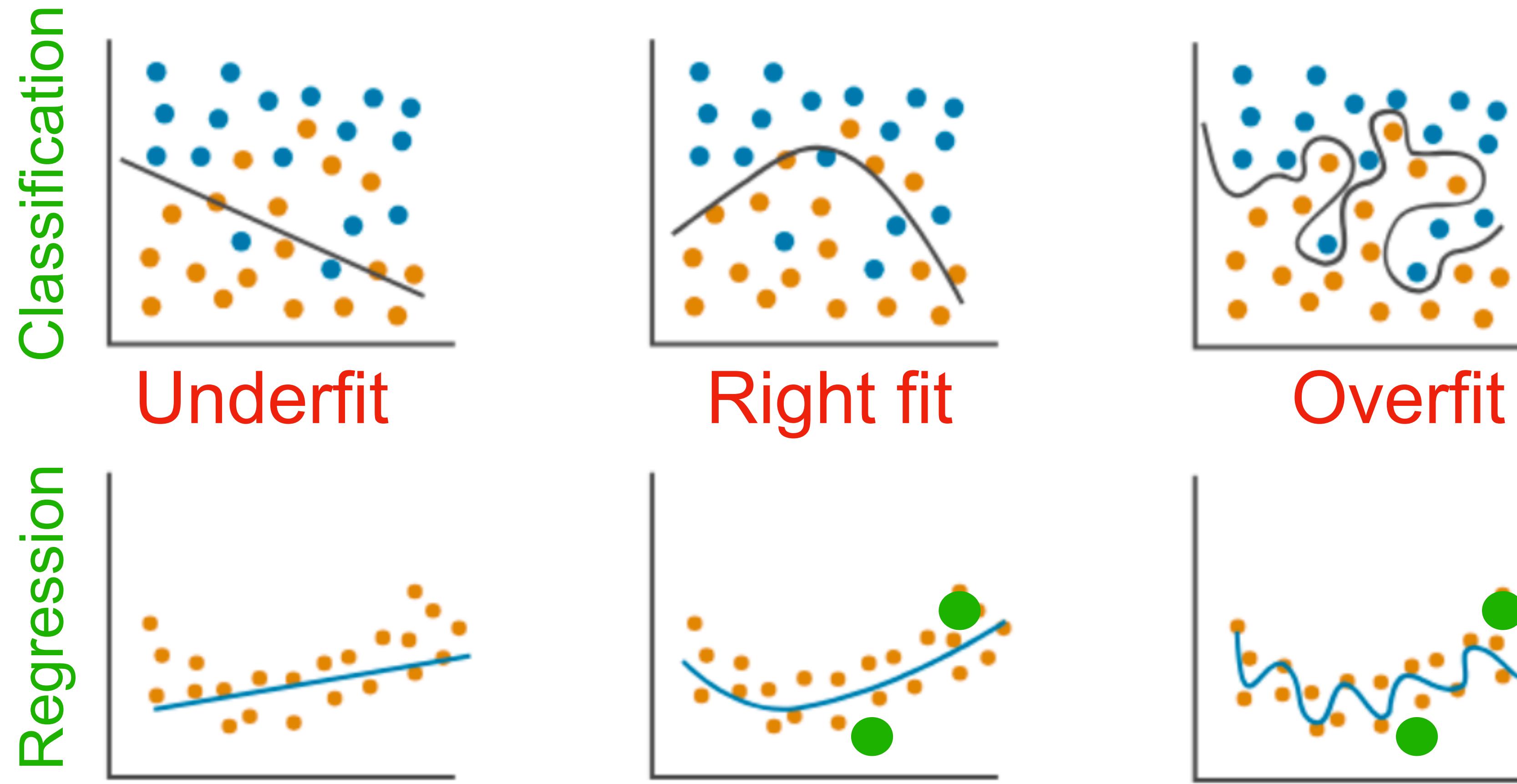


Which plots correspond to underfitting, right fit, and overfitting?

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Restricting Model Complexity

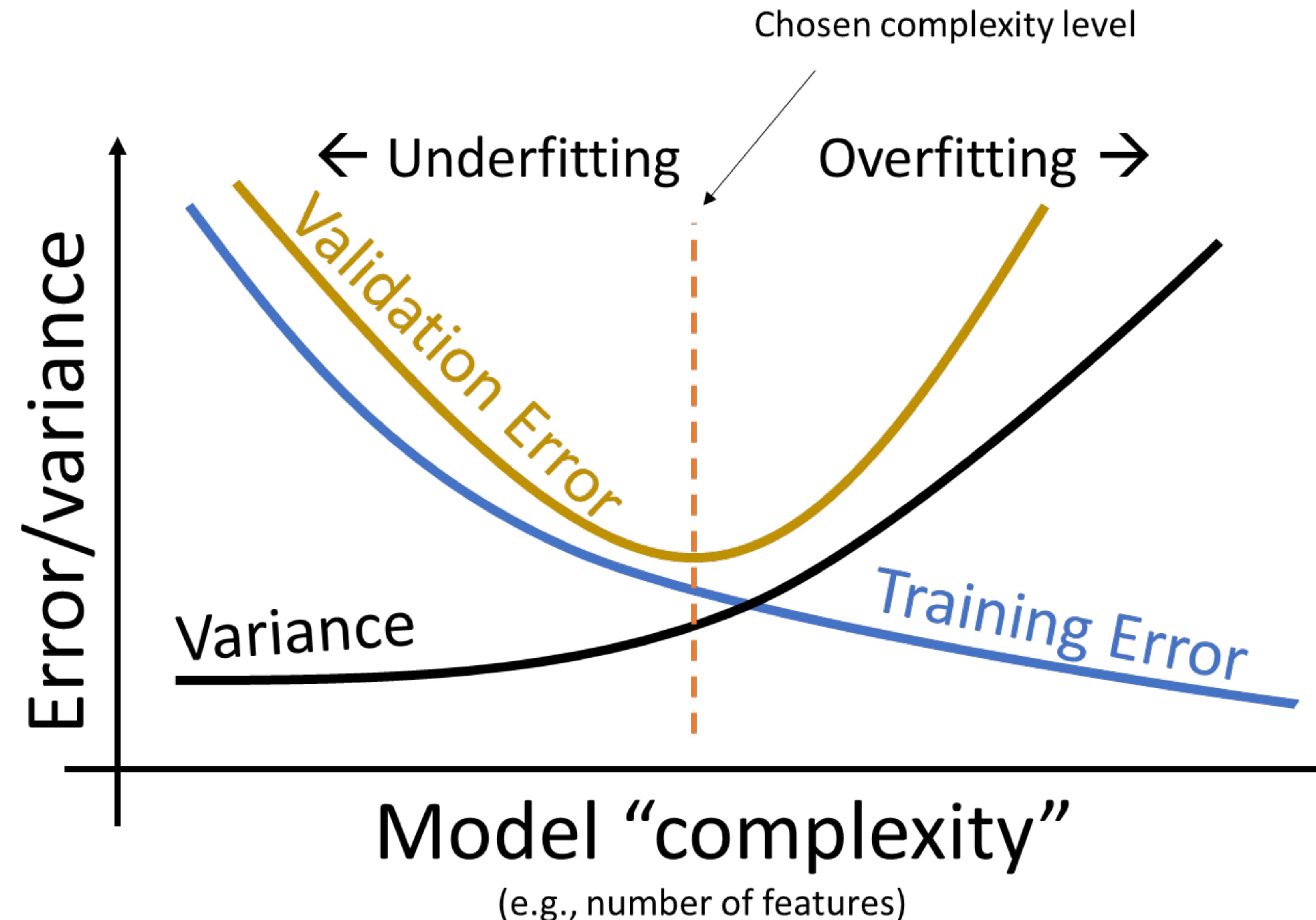
Bias/variance tradeoff resulting in underfitting/overfitting.



Consider green test points. Overfit model doesn't perform well!

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Restricting Model Complexity



RESTRICTING MODEL COMPLEXITY: REGULARIZATION

Loss function to minimize:

$$RSS = \sum_{i=1}^n (y_i - w_0^2 - (w_1 x_{i1} + w_2 x_{i2} + \cdots + w_m x_{im}))^2$$

RESTRICTING MODEL COMPLEXITY: REGULARIZATION

Loss function to minimize:

$$RSS = \sum_{i=1}^n (y_i - w_0^2 - (w_1 x_{i1} + w_2 x_{i2} + \cdots + w_m x_{im}))^2$$

To simplify, add constraints:

Alternative 1: $\sum_{i=1}^m |w_i| \leq t$ L1 Regularization (LASSO)

RESTRICTING MODEL COMPLEXITY: REGULARIZATION

Loss function to minimize:

$$RSS = \sum_{i=1}^n (y_i - w_0^2 - (w_1 x_{i1} + w_2 x_{i2} + \cdots + w_m x_{im}))^2$$

To simplify, add constraints:

Alternative 1: $\sum_{i=1}^m |w_i| \leq t$ L1 Regularization (LASSO)

Alternative 2: $\sum_{i=1}^m w_i^2 \leq t$ L2 Regularization (Ridge)

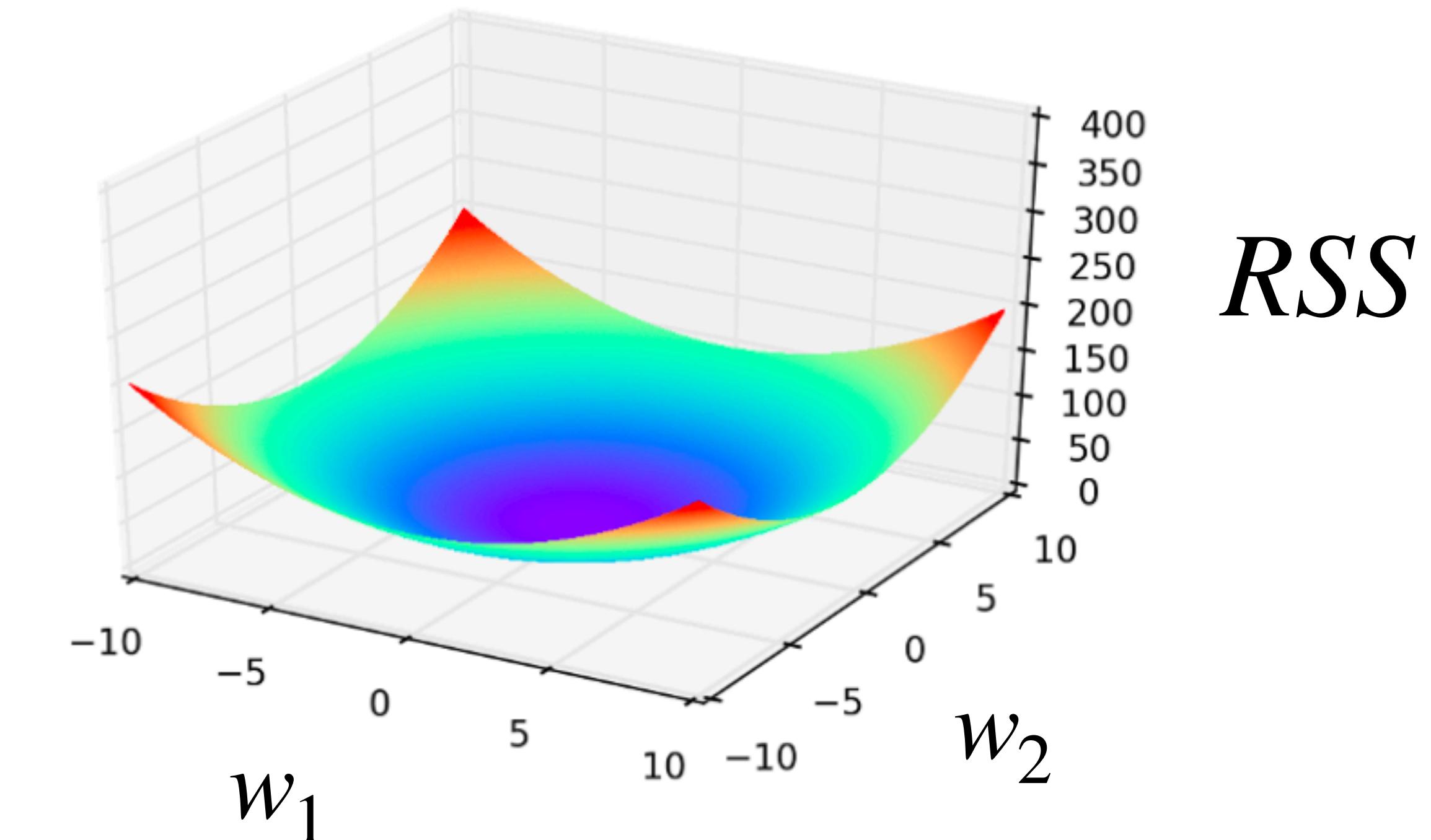
Note: No intercept term in constraints.

RESTRICTING MODEL COMPLEXITY: REGULARIZATION

For better intuition, consider $w_0 = 0$ and 2 coefficients w_1, w_2 :

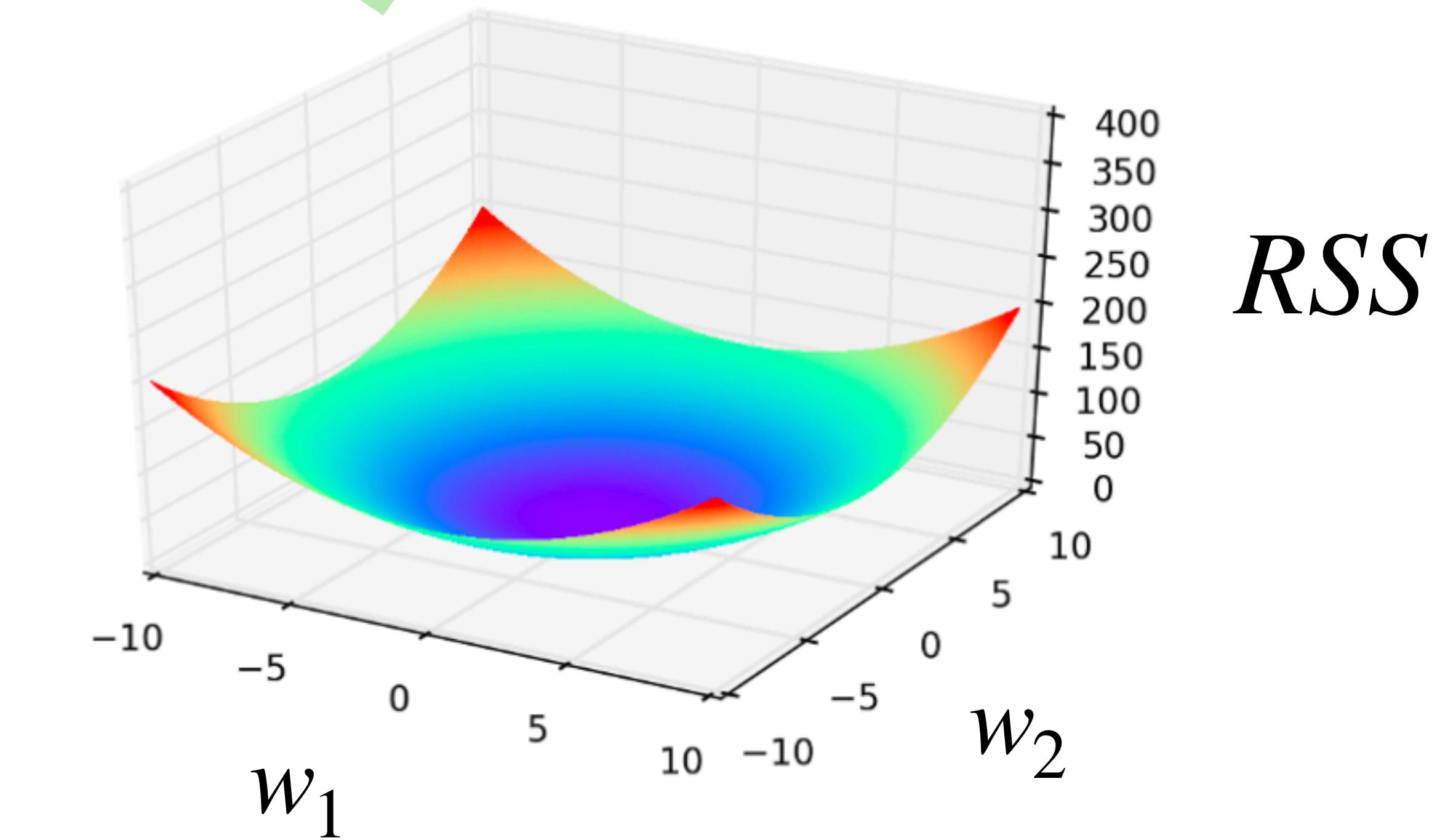
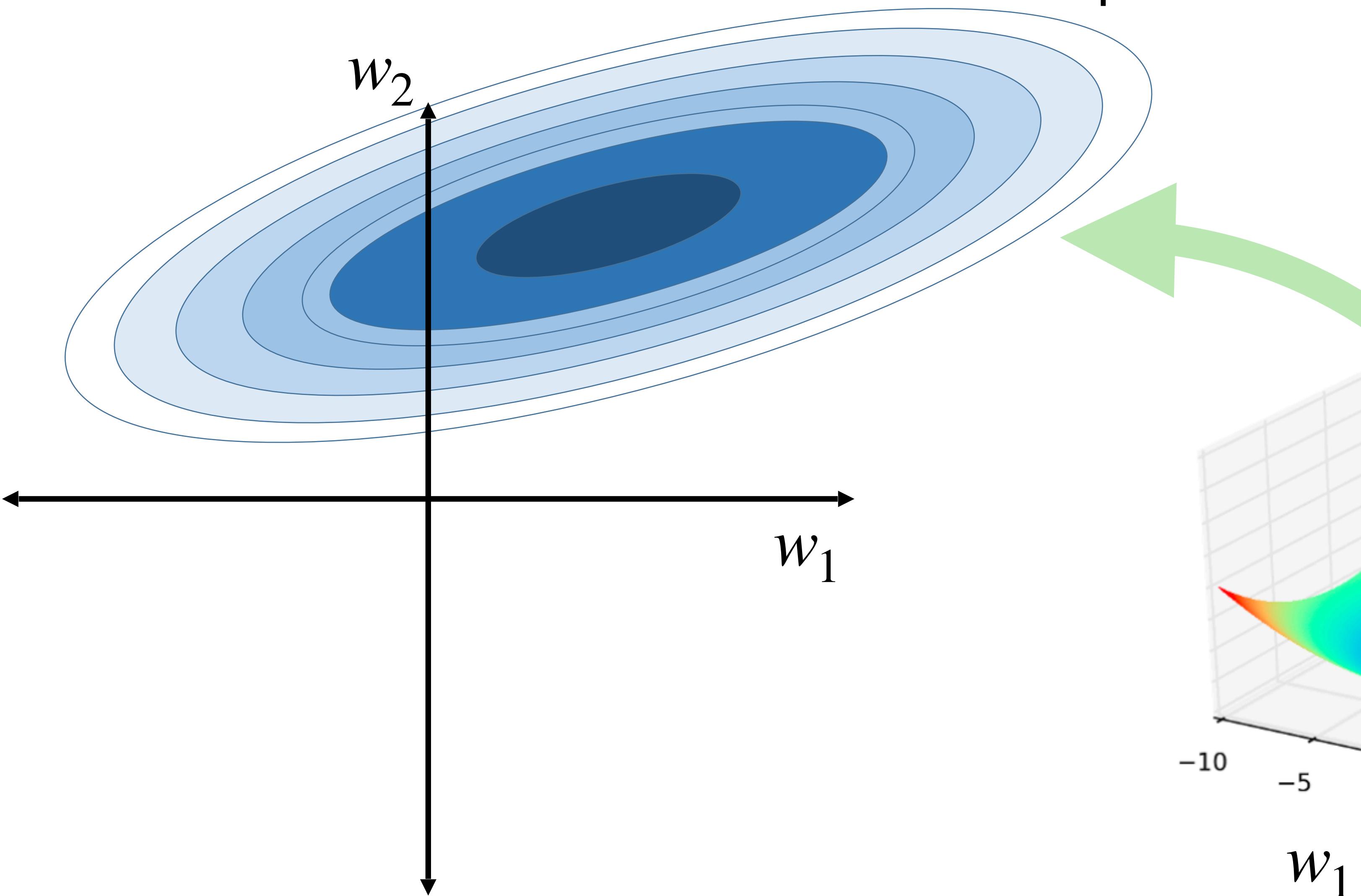
Minimize

$$RSS = \sum_{i=1}^n (y_i - (w_1 x_{i1} + w_2 x_{i2}))^2$$

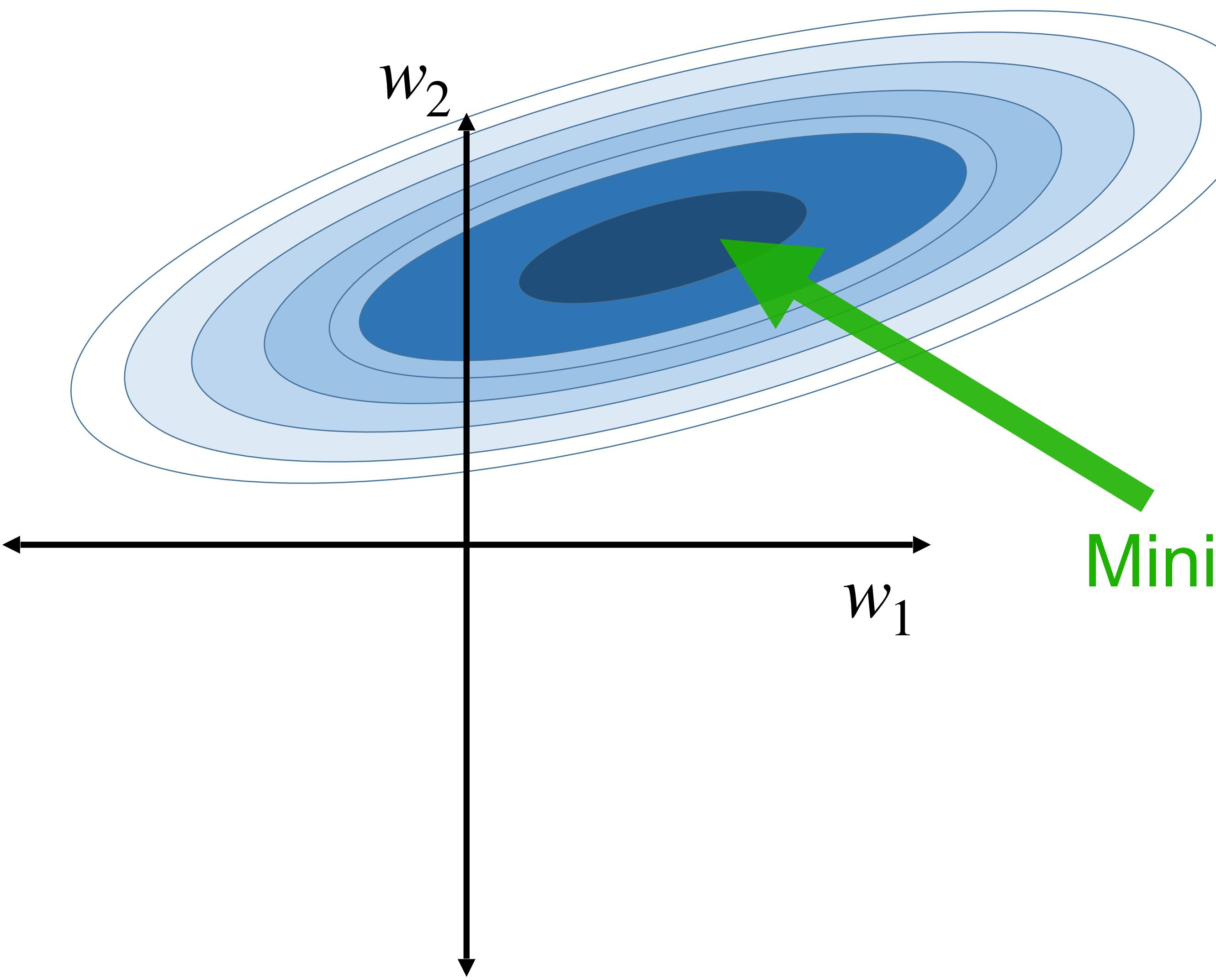


RESTRICTING MODEL COMPLEXITY: REGULARIZATION

Loss surface as contour map



RESTRICTING MODEL COMPLEXITY: REGULARIZATION

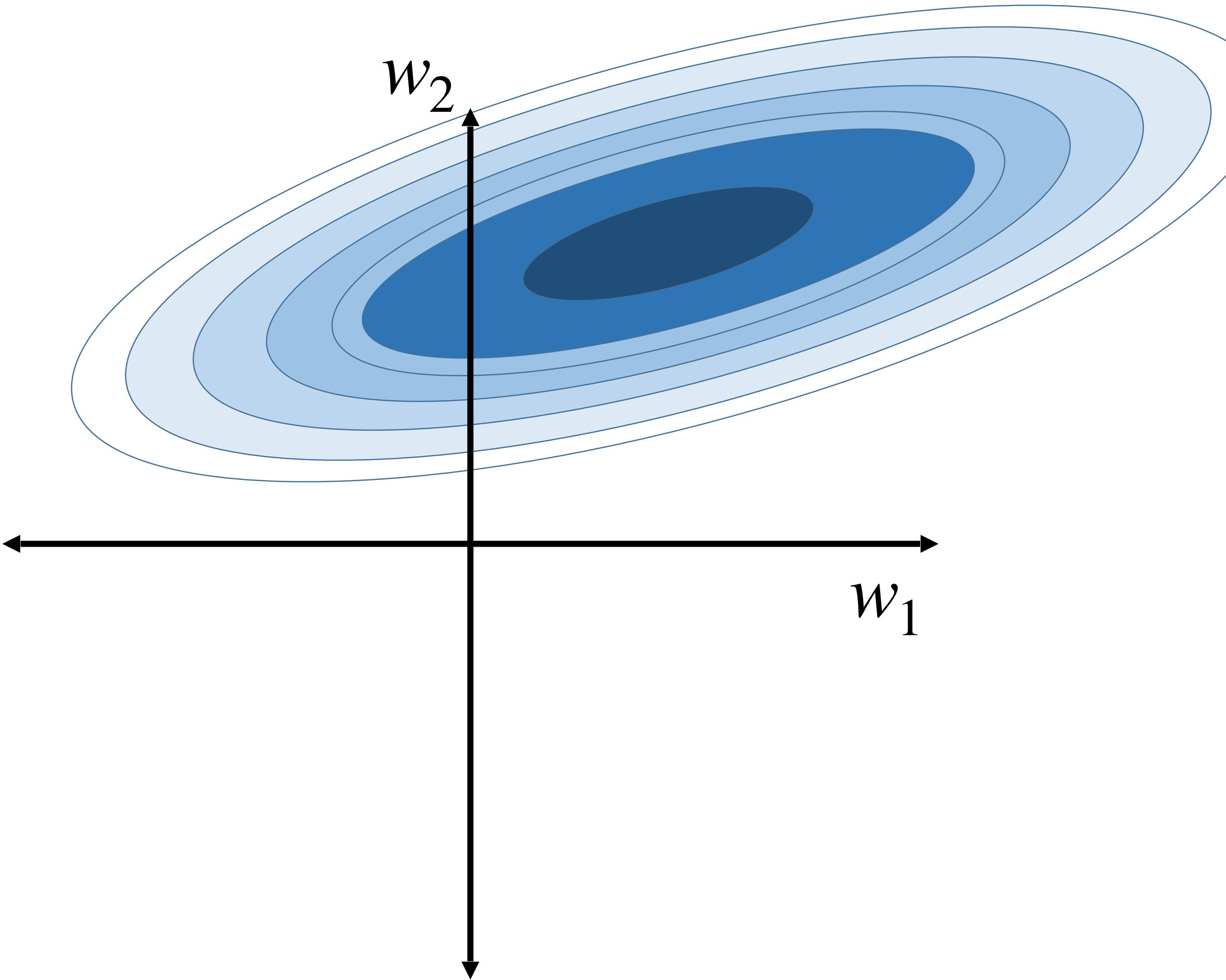


Minimize

$$RSS = \sum_{i=1}^n (y_i - (w_1 x_{i1} + w_2 x_{i2}))^2$$

Minimum loss achieved here.

RESTRICTING MODEL COMPLEXITY: REGULARIZATION



Minimize

$$RSS = \sum_{i=1}^n (y_i - (w_1 x_{i1} + w_2 x_{i2}))^2$$

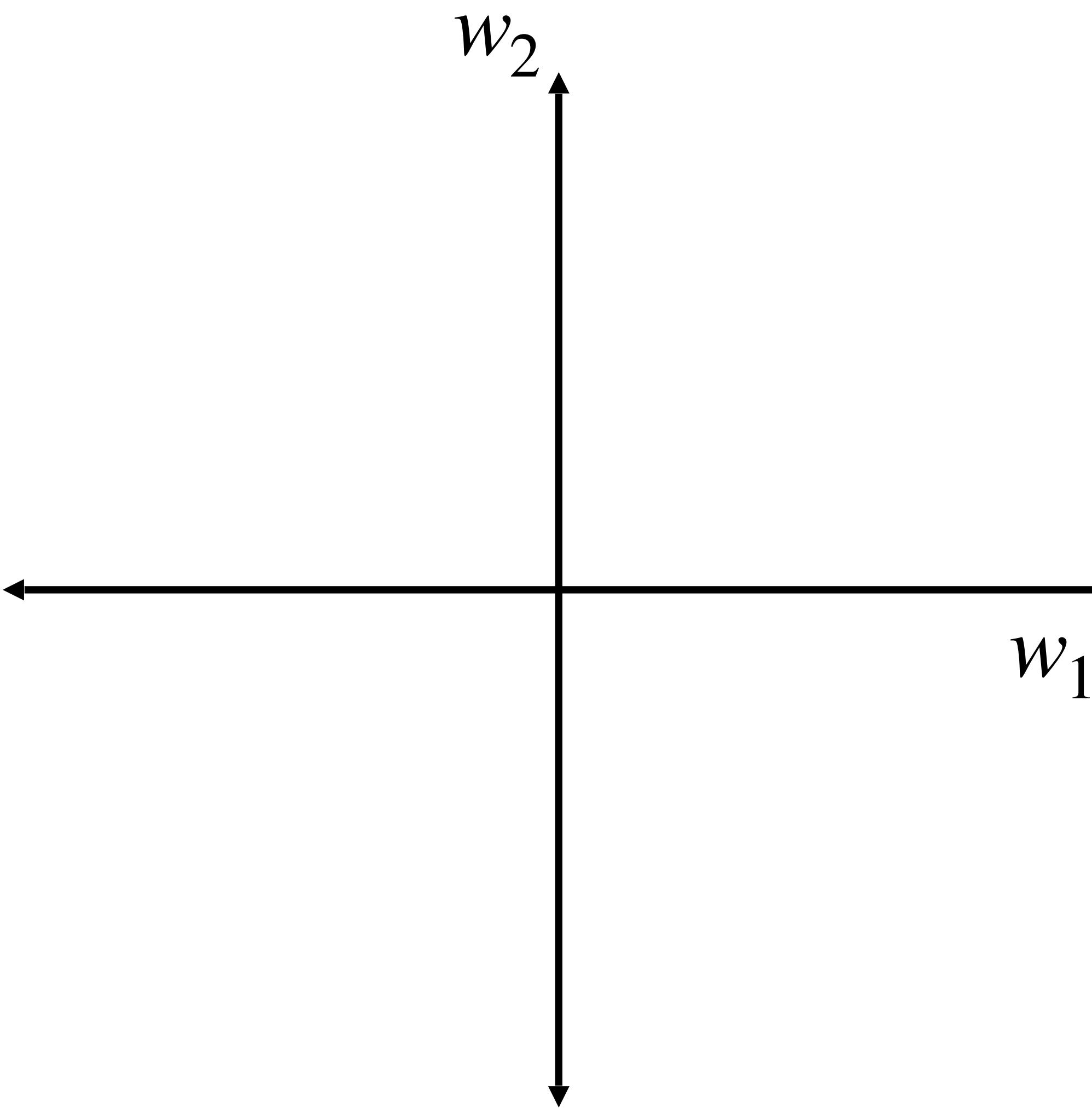
Subject to constraint:

$$\sum_{i=1}^2 |w_i| \leq t$$

L1 REGULARIZATION (LASSO)



RESTRICTING MODEL COMPLEXITY: REGULARIZATION

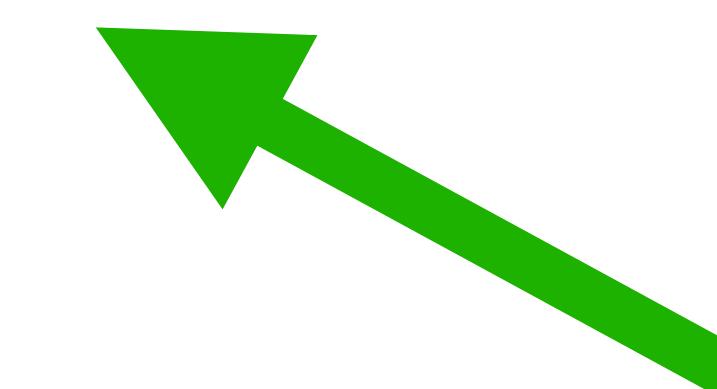


Minimize

$$RSS = \sum_{i=1}^n (y_i - (w_1 x_{i1} + w_2 x_{i2}))^2$$

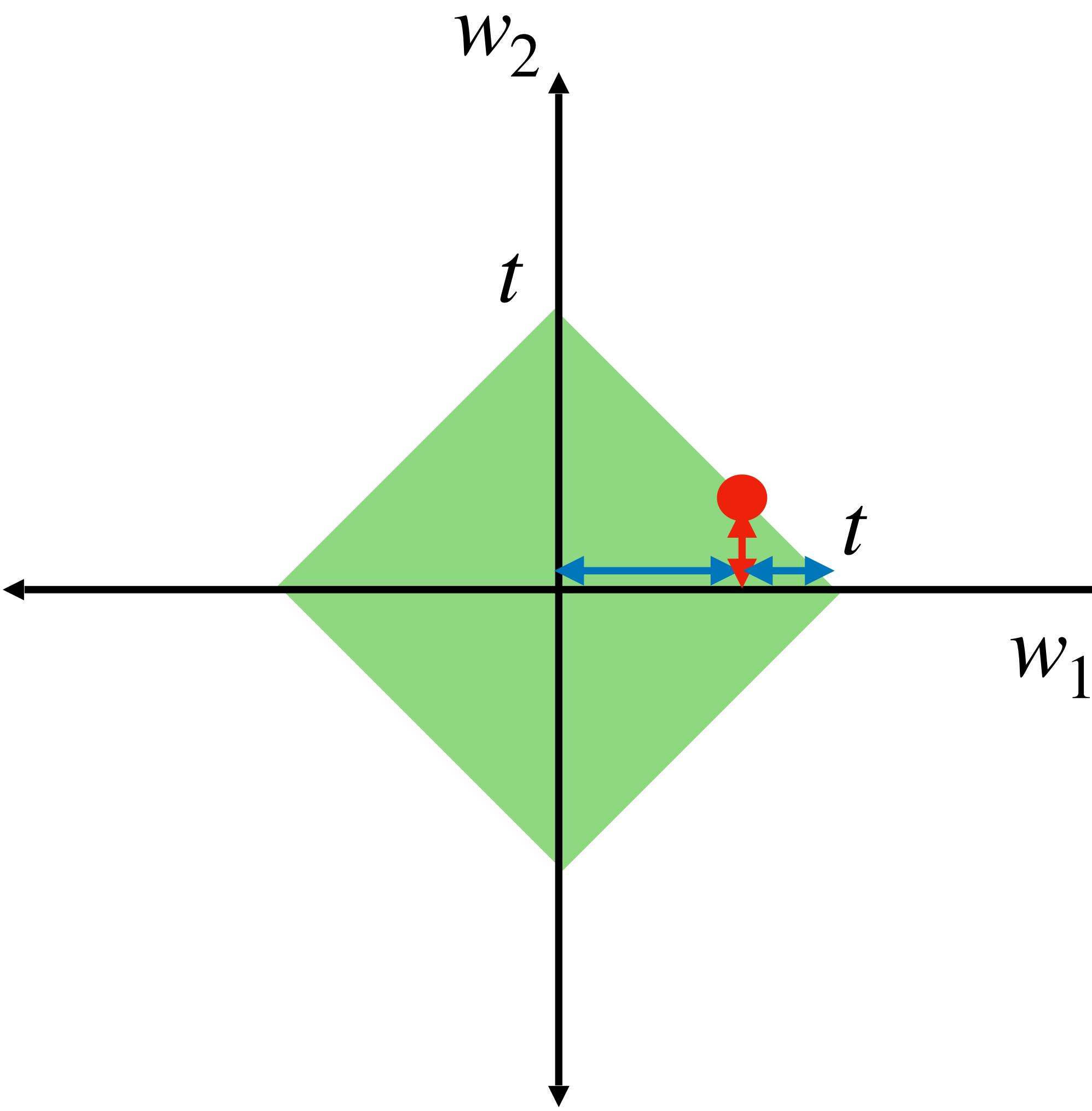
Subject to constraint:

$$\sum_{i=1}^2 |w_i| \leq t$$



What does
the region
satisfying this
look like?

RESTRICTING MODEL COMPLEXITY: REGULARIZATION

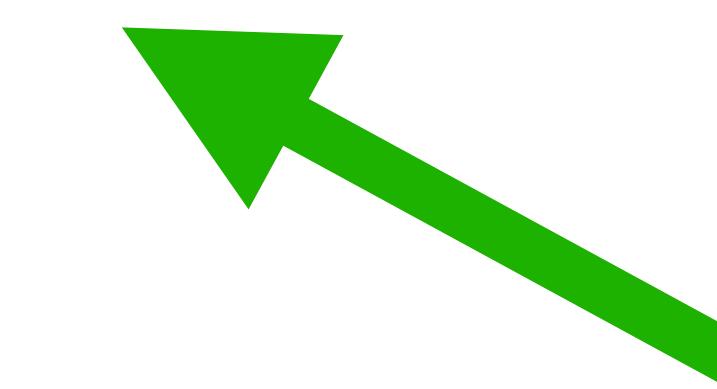


Minimize

$$RSS = \sum_{i=1}^n (y_i - (w_1 x_{i1} + w_2 x_{i2}))^2$$

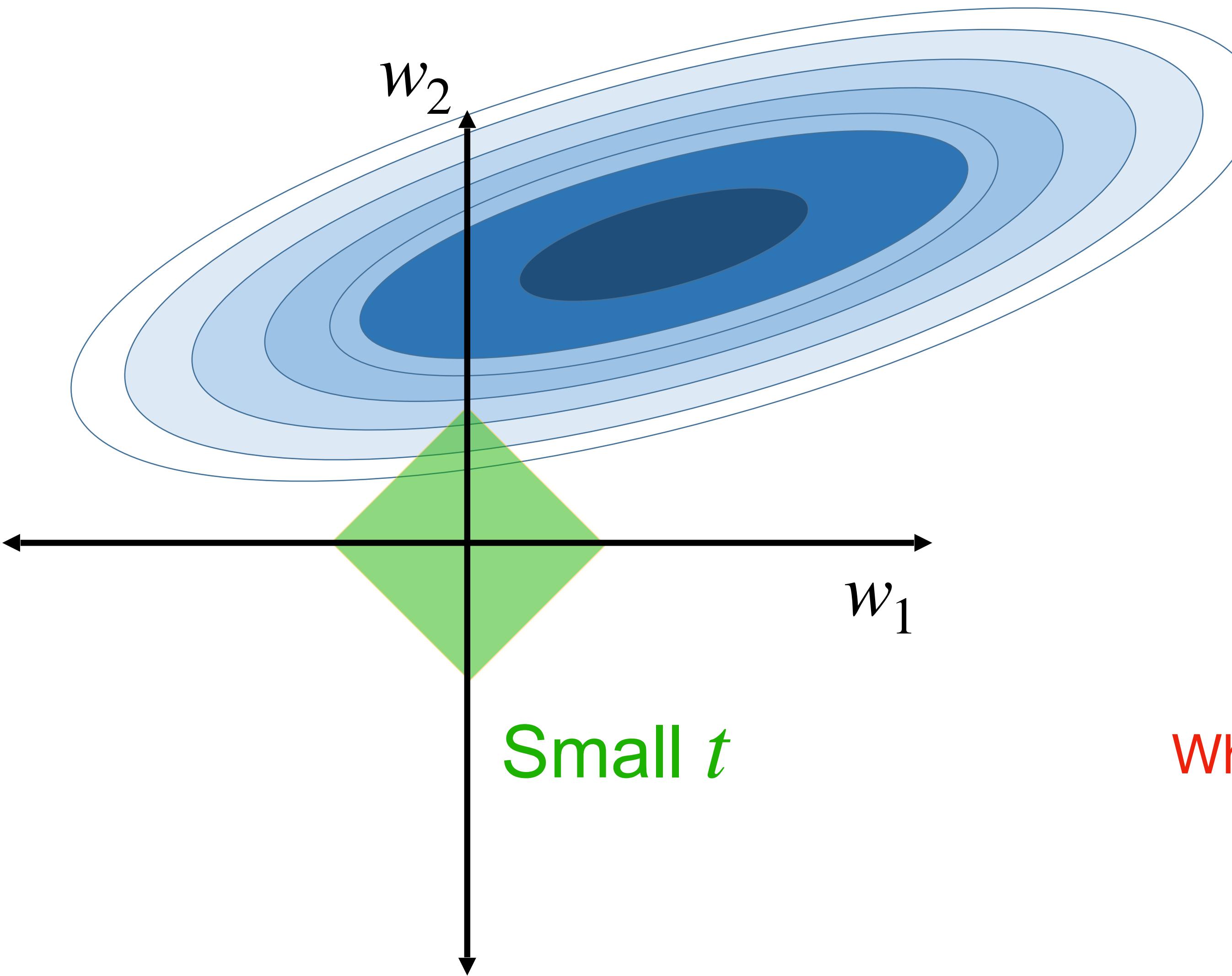
Subject to constraint:

$$\sum_{i=1}^2 |w_i| \leq t$$



Satisfied
anywhere
inside square

RESTRICTING MODEL COMPLEXITY: REGULARIZATION



Minimize

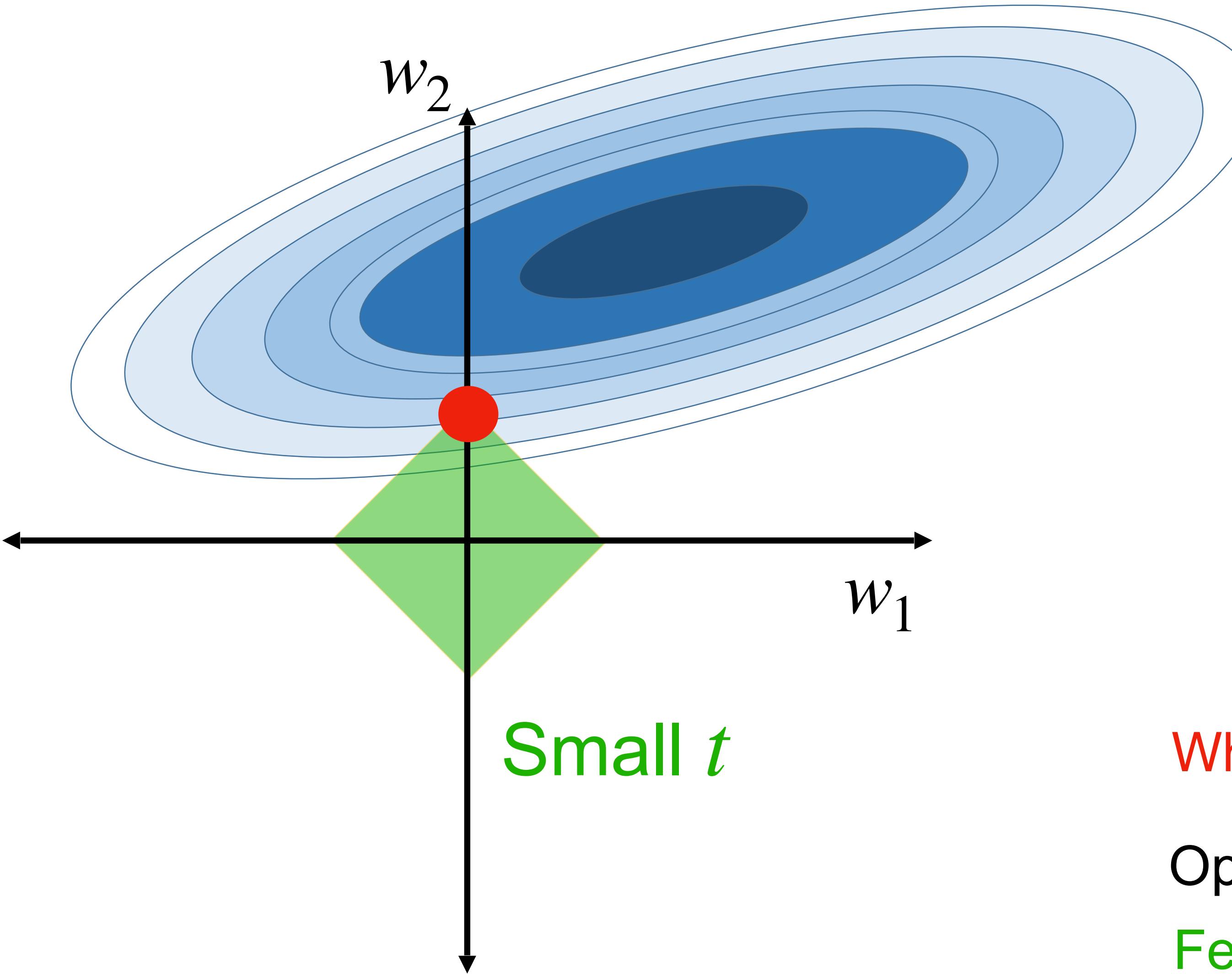
$$RSS = \sum_{i=1}^n (y_i - (w_1 x_{i1} + w_2 x_{i2}))^2$$

Subject to constraint:

$$\sum_{i=1}^2 |w_i| \leq t$$

Where is the optimum coefficient setting?

RESTRICTING MODEL COMPLEXITY: REGULARIZATION



Minimize

$$RSS = \sum_{i=1}^n (y_i - (w_1 x_{i1} + w_2 x_{i2}))^2$$

Subject to constraint:

$$\sum_{i=1}^2 |w_i| \leq t$$

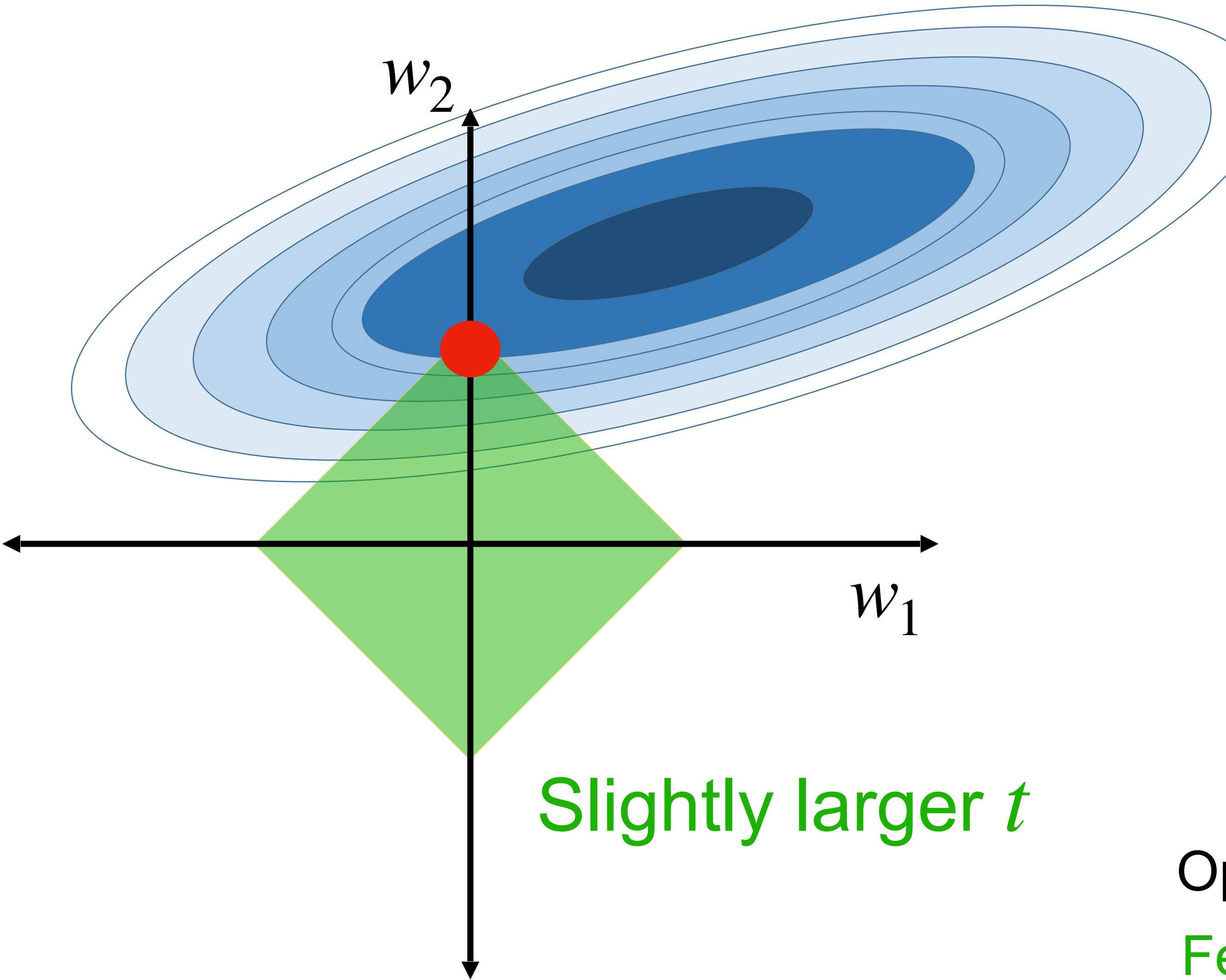
Small t

Where is the optimum coefficient setting?

Optimum coefficient setting at Red point.

Feature w_1 contributes none, model is simple.

RESTRICTING MODEL COMPLEXITY: REGULARIZATION



Minimize

$$RSS = \sum_{i=1}^n (y_i - (w_1 x_{i1} + w_2 x_{i2}))^2$$

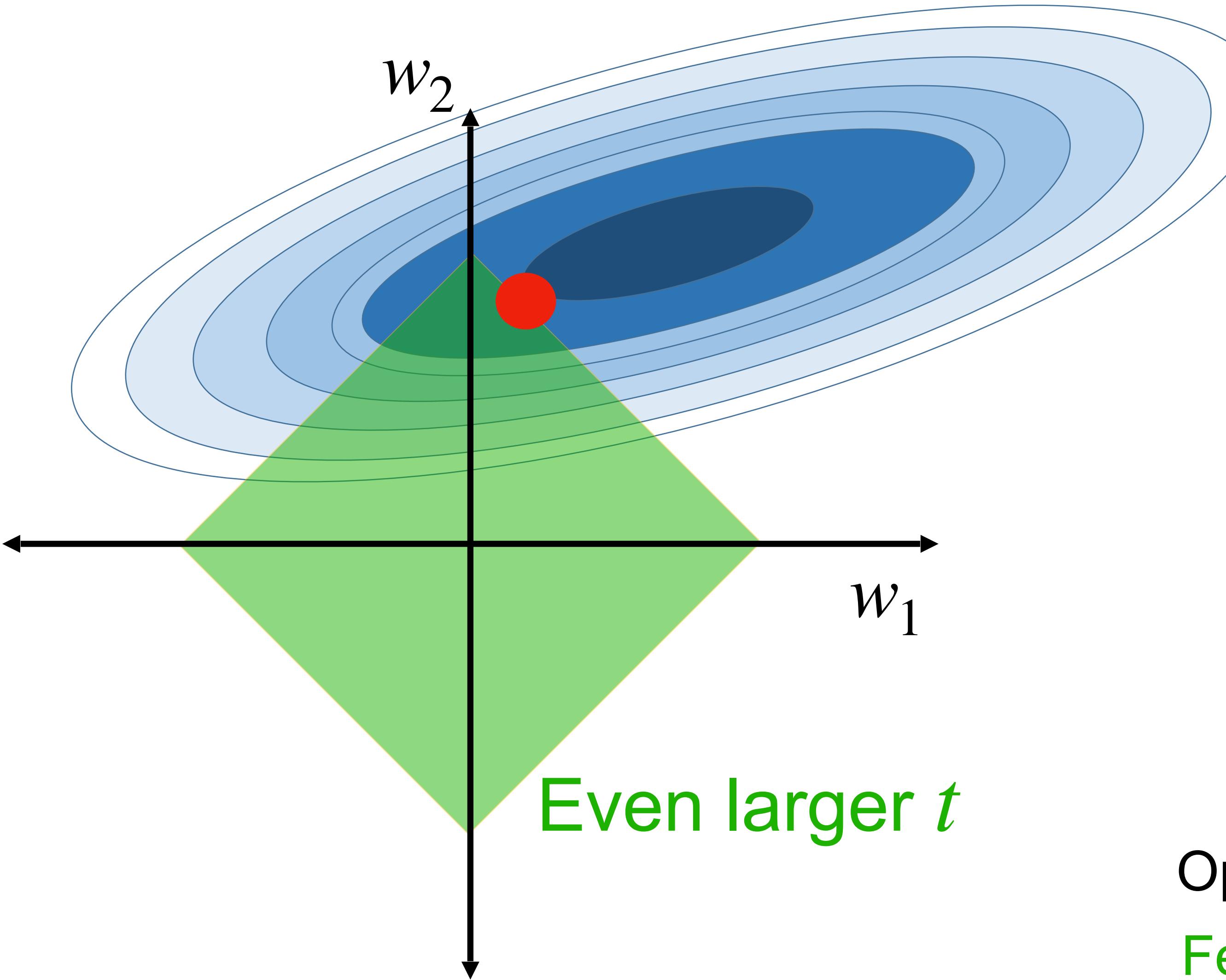
Subject to constraint:

$$\sum_{i=1}^2 |w_i| \leq t$$

Optimum coefficient setting at Red point.

Feature w_1 contributes none, model is simple.

RESTRICTING MODEL COMPLEXITY: REGULARIZATION



Minimize

$$RSS = \sum_{i=1}^n (y_i - (w_1 x_{i1} + w_2 x_{i2}))^2$$

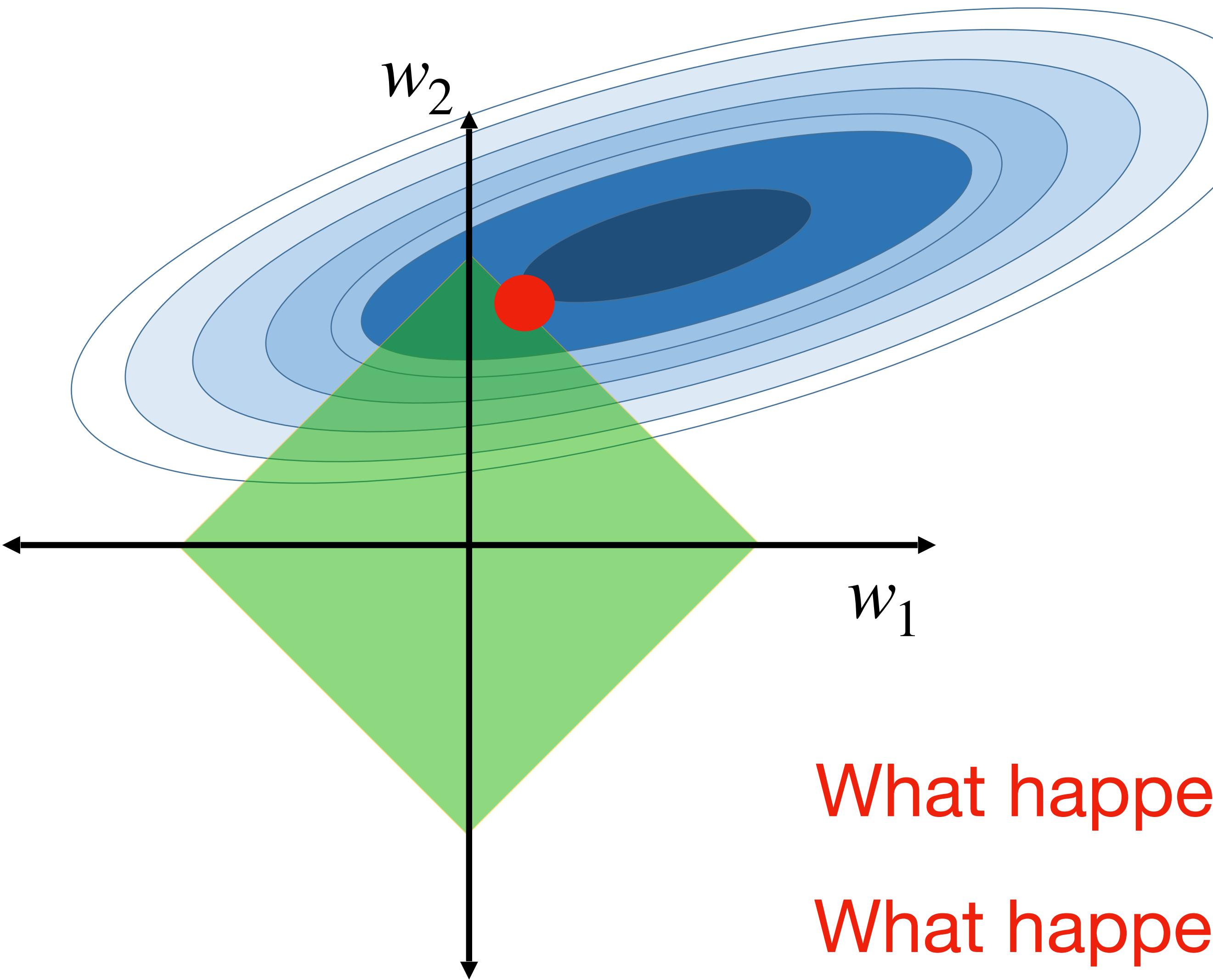
Subject to constraint:

$$\sum_{i=1}^2 |w_i| \leq t$$

Optimum coefficient setting at Red point.

Feature w_1 contributes, model more complex.

RESTRICTING MODEL COMPLEXITY: REGULARIZATION



Minimize

$$RSS = \sum_{i=1}^n (y_i - (w_1 x_{i1} + w_2 x_{i2}))^2$$

Subject to constraint:

$$\sum_{i=1}^2 |w_i| \leq t$$

What happens when t very large, say $t \rightarrow \infty$?

What happens when t very small, say $t \approx 0$?

RESTRICTING MODEL COMPLEXITY: REGULARIZATION

Lasso: Optimization Formulation

$$\text{Minimize} \sum_{i=1}^n (y_i - (w_0 + w_1x_{i1} + w_2x_{i2} + \cdots + w_mx_{im}))^2 \text{ subject to } \sum_{i=1}^m |w_i| \leq t$$

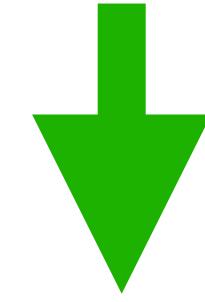
RESTRICTING MODEL COMPLEXITY: REGULARIZATION

Lasso: Optimization Formulation

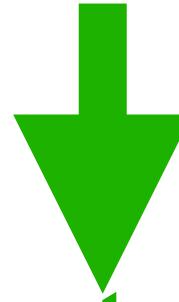
$$\text{Minimize} \sum_{i=1}^n (y_i - (w_0 + w_1x_{i1} + w_2x_{i2} + \cdots + w_mx_{im}))^2 \text{ subject to} \sum_{i=1}^m |w_i| \leq t$$

≡

$$\text{Minimize} \sum_{i=1}^n (y_i - (w_0 + w_1x_{i1} + w_2x_{i2} + \cdots + w_mx_{im}))^2 + \lambda \sum_{i=1}^m |w_i|$$



Keep MSE small



Keep coefficients small.
Simpler model.

LASSO (Least absolute shrinkage and selection operator)

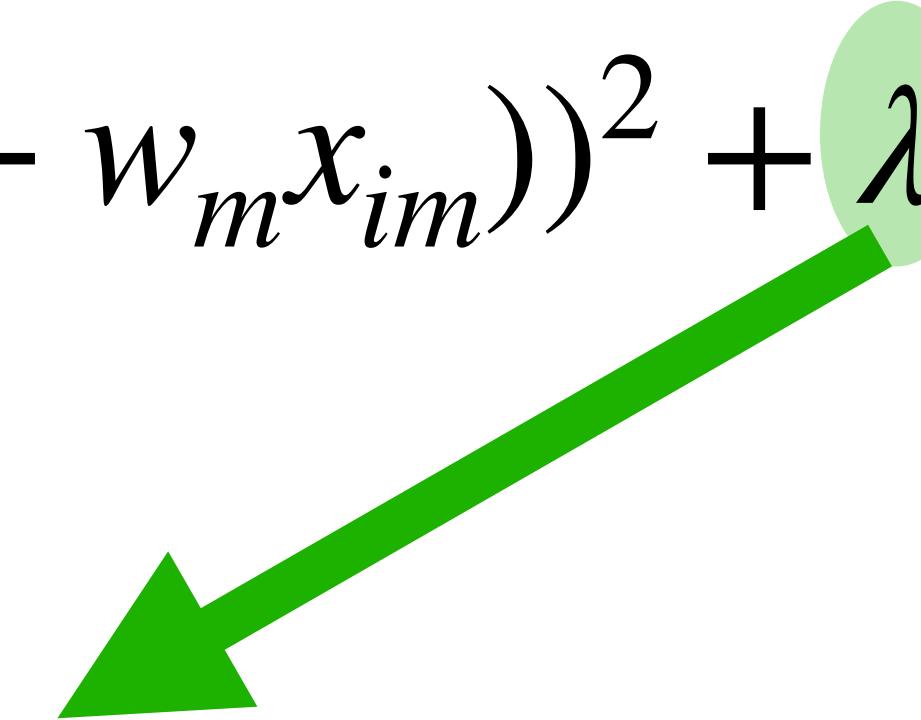
RESTRICTING MODEL COMPLEXITY: REGULARIZATION

Lasso: Optimization Formulation

$$\text{Minimize} \sum_{i=1}^n (y_i - (w_0 + w_1x_{i1} + w_2x_{i2} + \cdots + w_mx_{im}))^2 \text{ subject to} \sum_{i=1}^m |w_i| \leq t$$

≡

$$\text{Minimize} \sum_{i=1}^n (y_i - (w_0 + w_1x_{i1} + w_2x_{i2} + \cdots + w_mx_{im}))^2 + \lambda \sum_{i=1}^m |w_i|$$



Regularization penalty:
Determine good value using
cross validation

RESTRICTING MODEL COMPLEXITY: REGULARIZATION

Minimize

$$RSS = \sum_{i=1}^n (y_i - (w_1x_{i1} + w_2x_{i2}))^2$$

Subject to constraint:

$$\sum_{i=1}^m w_i^2 \leq t$$

L2 REGULARIZATION (RIDGE)

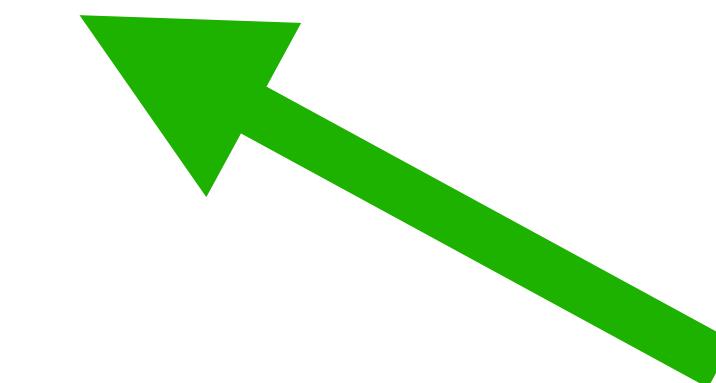
RESTRICTING MODEL COMPLEXITY: REGULARIZATION

Minimize

$$RSS = \sum_{i=1}^n (y_i - (w_1 x_{i1} + w_2 x_{i2}))^2$$

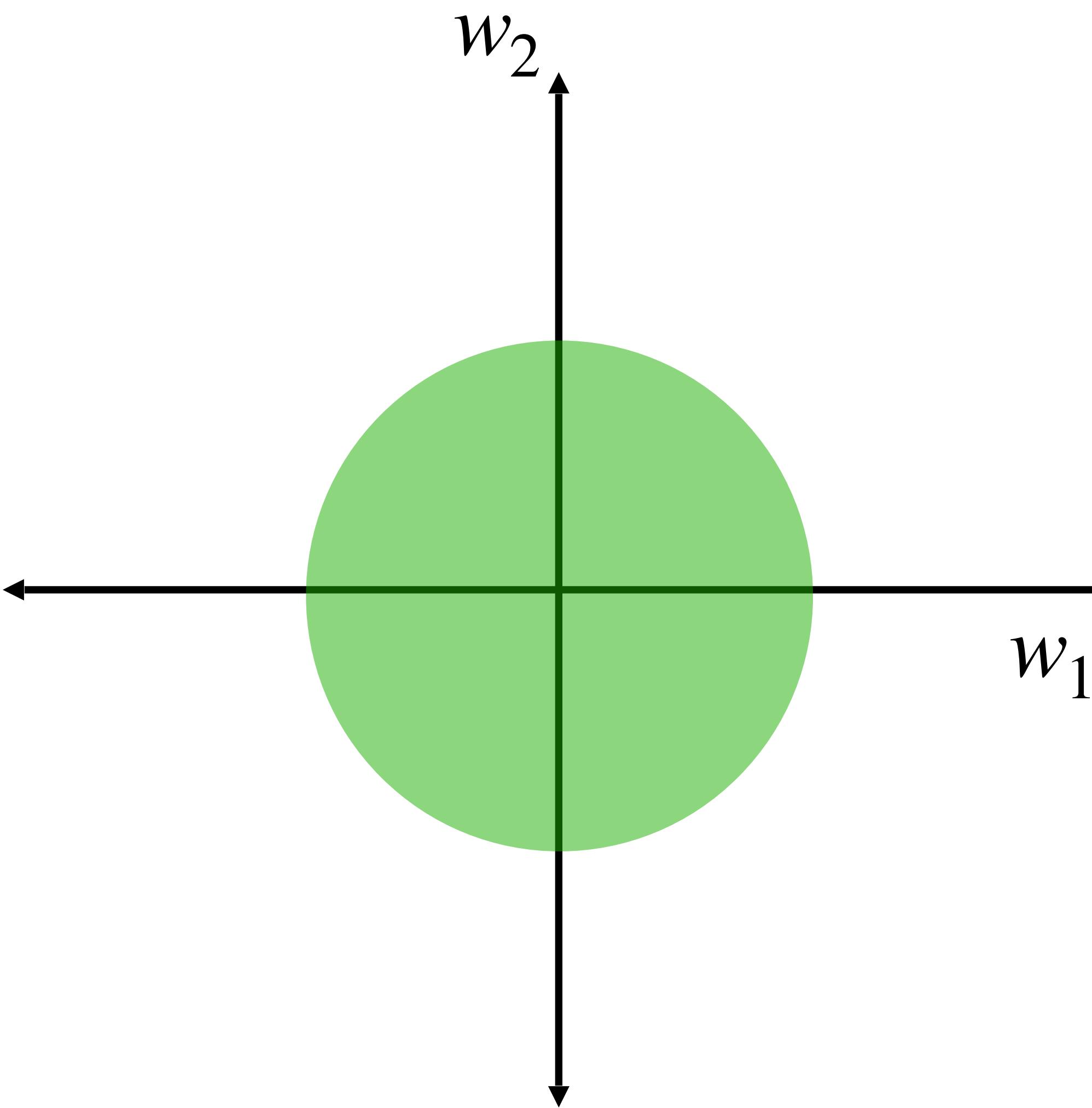
Subject to constraint:

$$\sum_{i=1}^m w_i^2 \leq t$$



What does
the region
satisfying this
look like?

RESTRICTING MODEL COMPLEXITY: REGULARIZATION

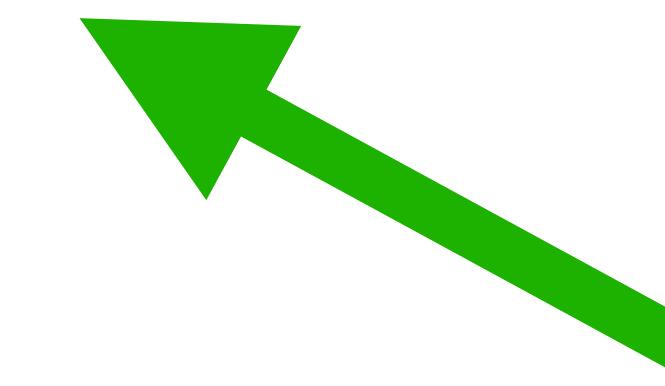


Minimize

$$RSS = \sum_{i=1}^n (y_i - (w_1 x_{i1} + w_2 x_{i2}))^2$$

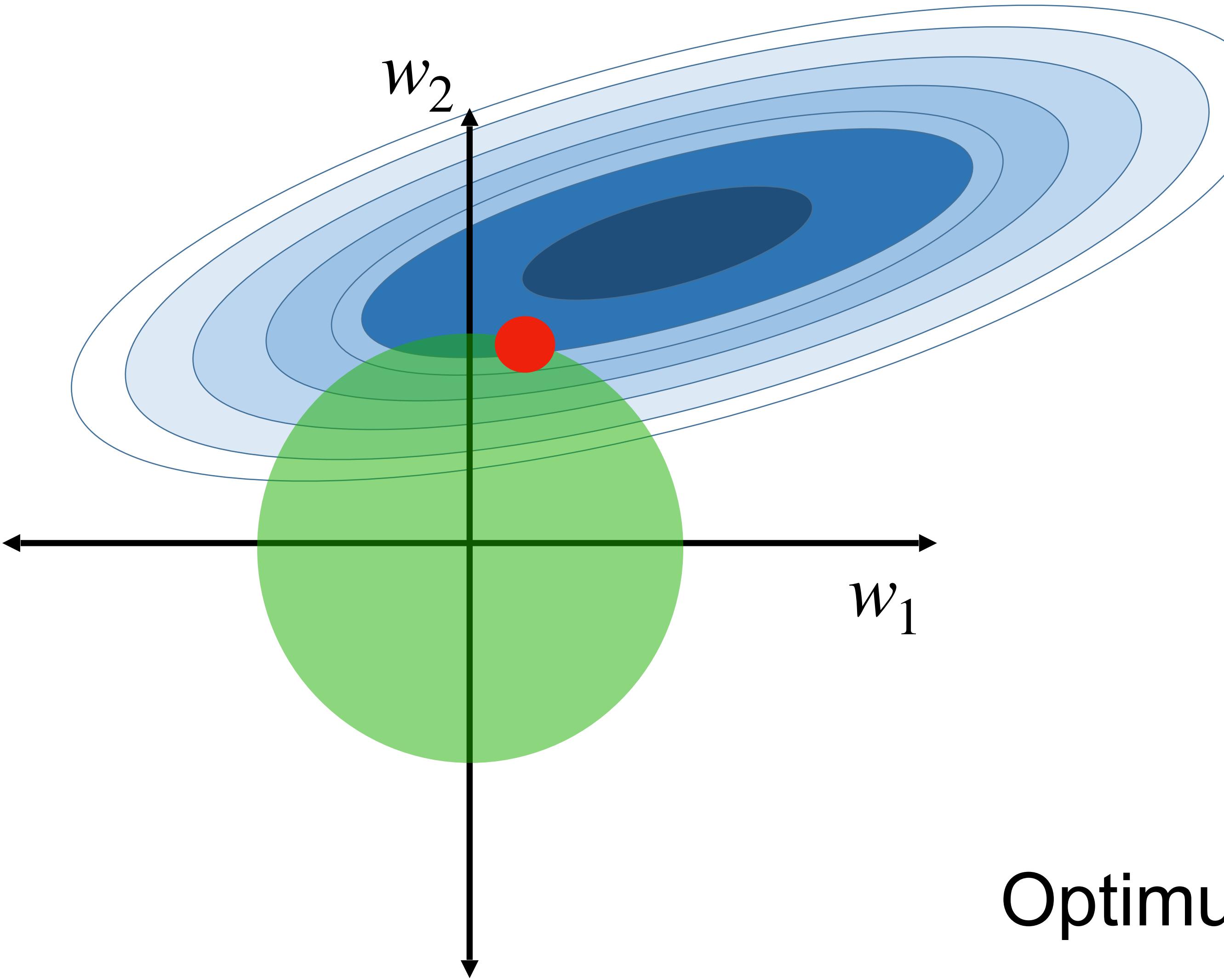
Subject to constraint:

$$\sum_{i=1}^m w_i^2 \leq t$$



What does
the region
satisfying this
look like?

RESTRICTING MODEL COMPLEXITY: REGULARIZATION



Minimize

$$RSS = \sum_{i=1}^n (y_i - (w_1 x_{i1} + w_2 x_{i2}))^2$$

Subject to constraint:

$$\sum_{i=1}^m w_i^2 \leq t$$

Optimum coefficient setting at Red point.

RESTRICTING MODEL COMPLEXITY: REGULARIZATION

Ridge: Optimization Formulation

$$\text{Minimize} \sum_{i=1}^n (y_i - (w_0 + w_1x_{i1} + w_2x_{i2} + \cdots + w_mx_{im}))^2 \text{ subject to} \sum_{i=1}^m w_i^2 \leq t$$

====

$$\text{Minimize} \sum_{i=1}^n (y_i - (w_0 + w_1x_{i1} + w_2x_{i2} + \cdots + w_mx_{im}))^2 + \lambda \sum_{i=1}^m w_i^2$$

RESTRICTING MODEL COMPLEXITY: REGULARIZATION

Ex: Hitters data set.

...	Years	CAtBat	CHits	CHmRun	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	Salary
	1	293	66	1	30	29	14	A	E	446	33	20	NaN
	14	3449	835	69	321	414	375	N	W	632	43	10	475.0
	3	1624	457	63	224	266	263	A	W	880	82	14	480.0
	11	5628	1575	225	828	838	354	N	E	200	11	3	500.0
	2	396	101	12	48	46	33	N	E	805	40	4	91.5

RESTRICTING MODEL COMPLEXITY: REGULARIZATION

Scale the features

```
# in regularization it is important to scale features
from sklearn.preprocessing import StandardScaler

# we do scaling after train/test split to avoid data leakage
X_train, X_test, y_train, y_test = \
    train_test_split(X, y, train_size=0.8, random_state=0)

scaler = StandardScaler().fit(X_train)
X_train[X.columns] = scaler.transform(X_train[X.columns])
X_test[X.columns] = scaler.transform(X_test[X.columns])
```

Years	CAtBat	CHits	CHmRun	CRuns	CRBI	CWalks	PutOuts	Assists	Errors
-0.464631	-1.001763	-0.980581	-0.897367	-0.957761	-0.945831	-0.884467	-0.613848	0.967475	0.443495
-0.249713	0.279628	0.353806	0.995644	0.567252	0.535436	0.249045	-0.189638	2.410126	0.592988
1.469625	1.857587	2.086141	2.618225	1.970388	2.626043	0.967858	-0.058581	-0.793506	-0.004983
-0.894465	-0.894980	-0.851090	-0.498839	-0.814009	-0.683242	-0.706738	-0.475894	-0.807052	-0.154476
3.618797	5.144516	5.610819	1.379940	5.679792	3.370043	5.197794	0.786390	-0.563224	-0.453462

Scaling after train/
test split to avoid
data leakage

RESTRICTING MODEL COMPLEXITY: REGULARIZATION

Apply Lasso

```
# apply lasso with arbitrarily set alpha=1
from sklearn.linear_model import Lasso

reg = Lasso(alpha=1)
reg.fit(X_train, y_train)
reg.score(X_train, y_train)
```

0.49169521567171304

alpha here is our λ .

```
# find mse on the train and test data
predict_train = reg.predict(X_train)
mse_train = mean_squared_error(y_train,predict_train)
print('training MSE : ',mse_train)

predict_test = reg.predict(X_test)
mse_test = mean_squared_error(y_test,predict_test)
print('test MSE : ',mse_test)
```

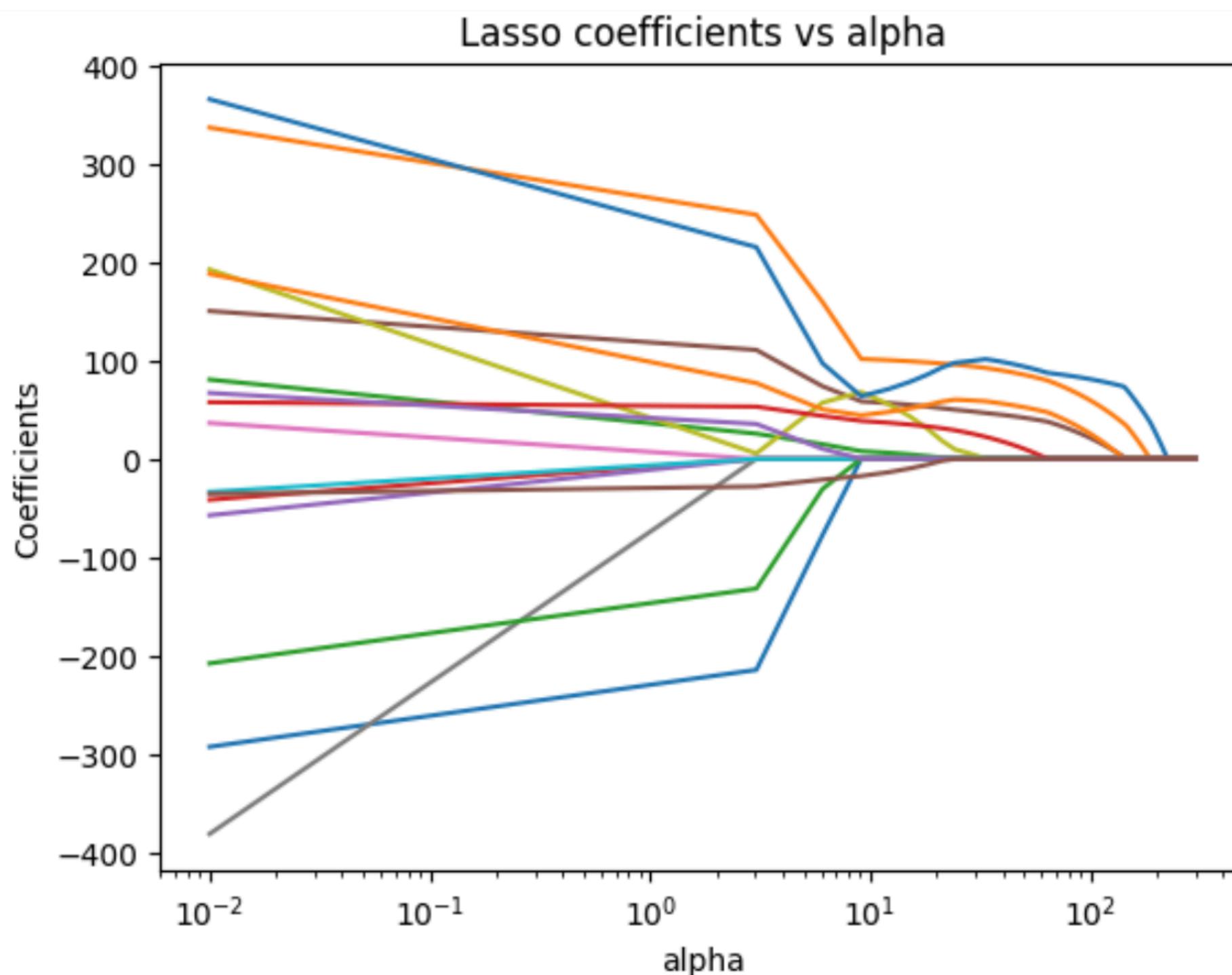
training MSE : 88524.3983554207
test MSE : 117768.90478161338

RESTRICTING MODEL COMPLEXITY: REGULARIZATION

Try different alpha values and plot mse values.

```
alphas = np.linspace(0.01,300,100)
coefs = []
for alpha in alphas:
    lasso_reg = Lasso(alpha=alpha)
    lasso_reg.fit(X_train, y_train)
    coefs.append(lasso_reg.coef_)
```

For the homework log space



QUIZ

How can we apply linear regression if we have qualitative (categorical) features (predictors)?

REVIEW OF THE PREVIOUS LECTURE

Interaction effects: Given features X_1, X_2 , introduce $X_1 \times X_2$ as well.

Polynomial regression: Given feature X_1 , introduce X_1^2 as new feature.

Qualitative features: One hot encoding.

REVIEW OF THE PREVIOUS LECTURE

Interaction effects: Given features X_1, X_2 , introduce $X_1 \times X_2$ as well.

Polynomial regression: Given feature X_1 , introduce X_1^2 as new feature.

Qualitative features: One hot encoding.

Restricting model complexity through regularization:

LASSO: Minimize $RSS + \lambda \sum_{i=1}^m |w_i|$

Ridge: Minimize $RSS + \lambda \sum_{i=1}^m w_i^2$

Notes: λ determined via cross validation.

Scaling is important if we do regularization.

OTHER CONSIDERATIONS: OUTLIERS

Remember the goal is to minimize the **quadratic** loss function,

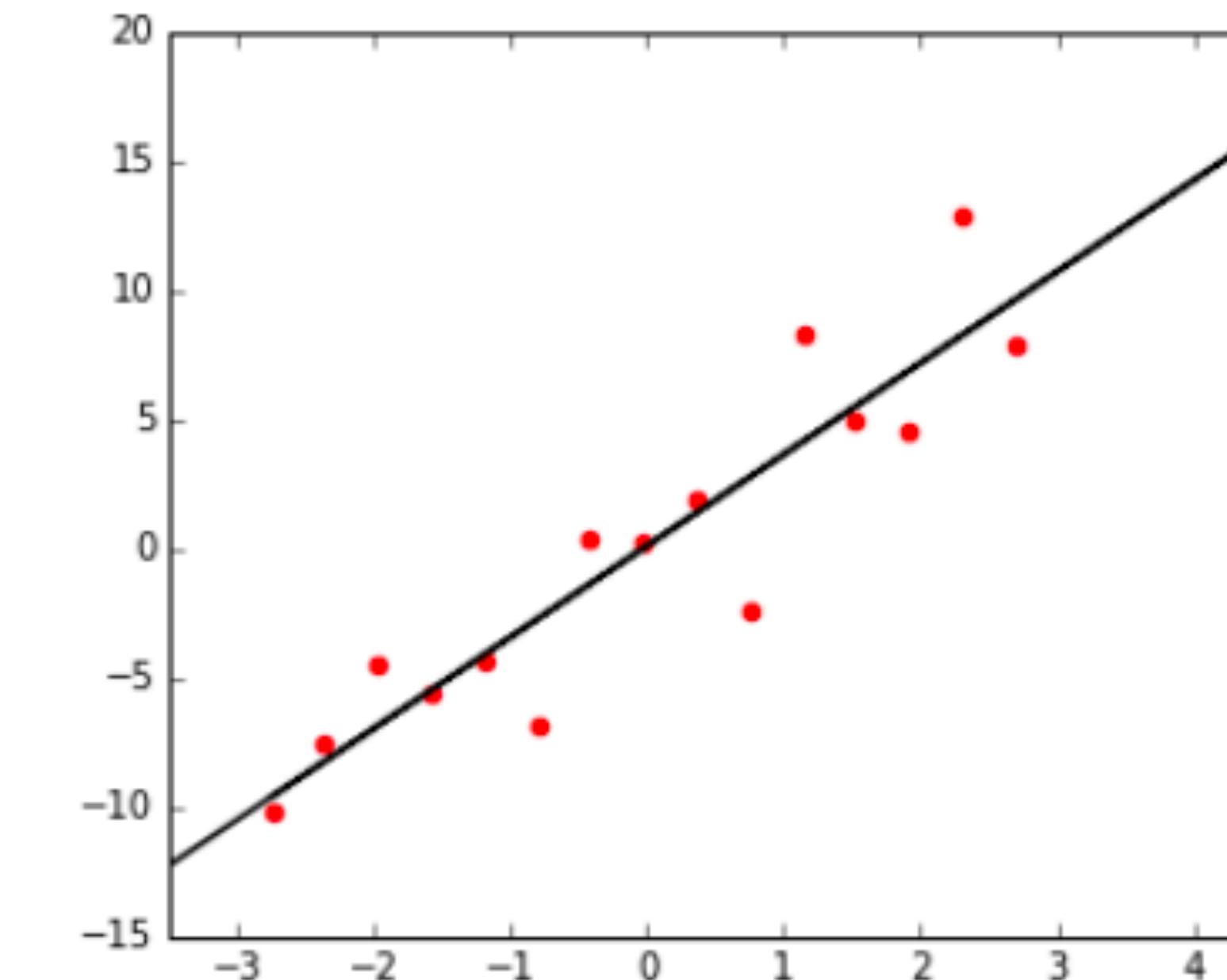
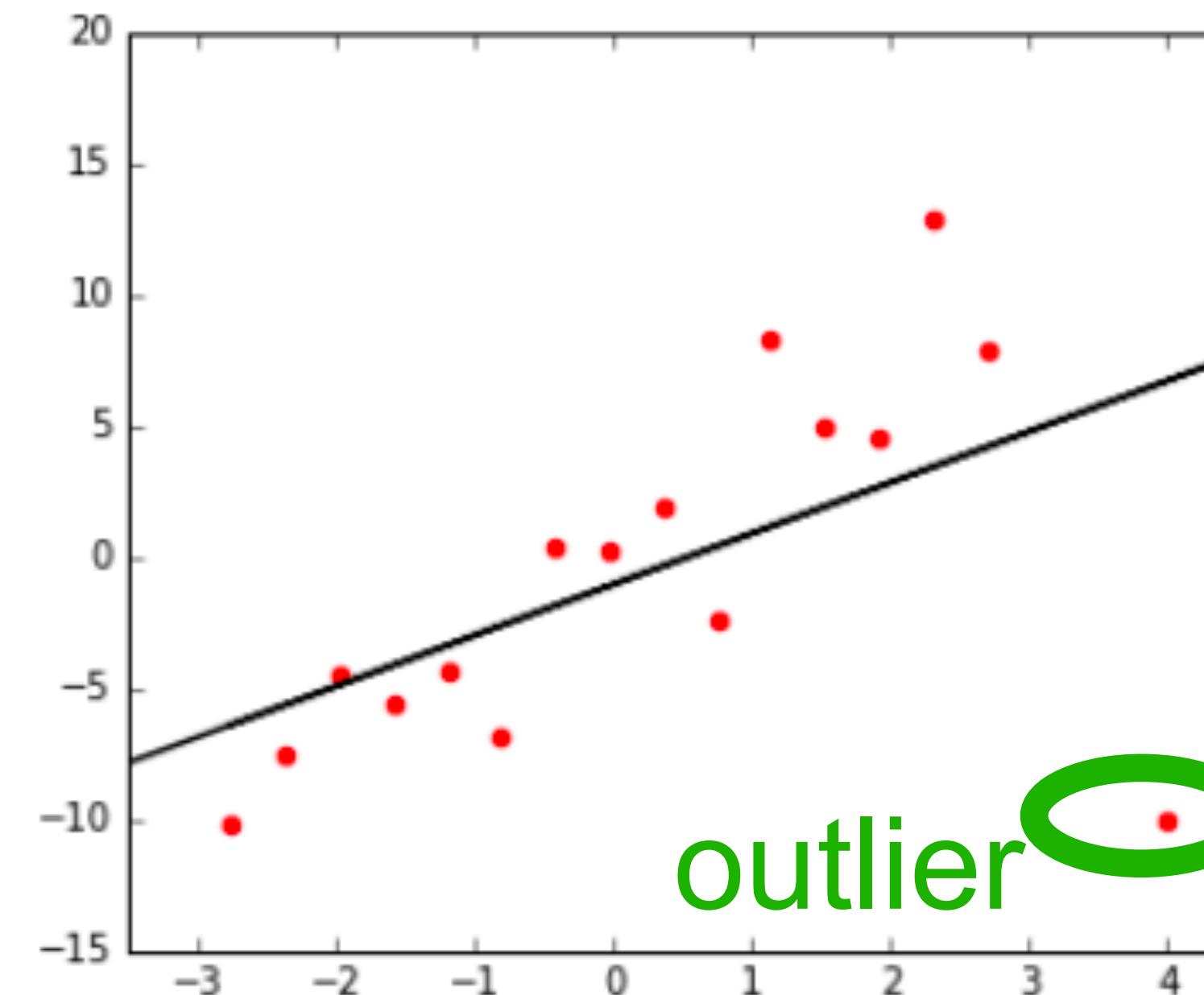
$$RSS = \sum_{i=1}^n (y_i - (w_0 + w_1x_{i1} + w_2x_{i2} + \dots + w_mx_{im}))^2$$

OTHER CONSIDERATIONS: OUTLIERS

Remember the goal is to minimize the **quadratic** loss function,

$$RSS = \sum_{i=1}^n (y_i - (w_0 + w_1x_{i1} + w_2x_{i2} + \dots + w_mx_{im}))^2$$

With quadratic error functions **the effect of the outliers is strong:**



Remove outliers, if you believe they reflect noise rather than signal.

SCALING THE DATA

Features vary over a wide range \Rightarrow Coefficients vary over wide range

Ex:

$$GDP = \$10,000x_1 + 10,000,000,000,000x_2$$

The equation $GDP = \$10,000x_1 + 10,000,000,000,000x_2$ illustrates the concept of scaling data features. The term x_1 is associated with 'population' and the term x_2 is associated with 'literacy rate'. This shows that while the population feature has a coefficient of \$10,000, the literacy rate feature has an extremely large coefficient of 10,000,000,000,000. This wide range of coefficients is a result of the wide range of values for the features themselves.

population

literacy rate

SCALING THE DATA

Features vary over a wide range \Rightarrow Coefficients vary over wide ranges

Ex:

$$GDP = \$10,000x_1 + 10,000,000,000,000x_2$$

population

literacy rate

Problems:

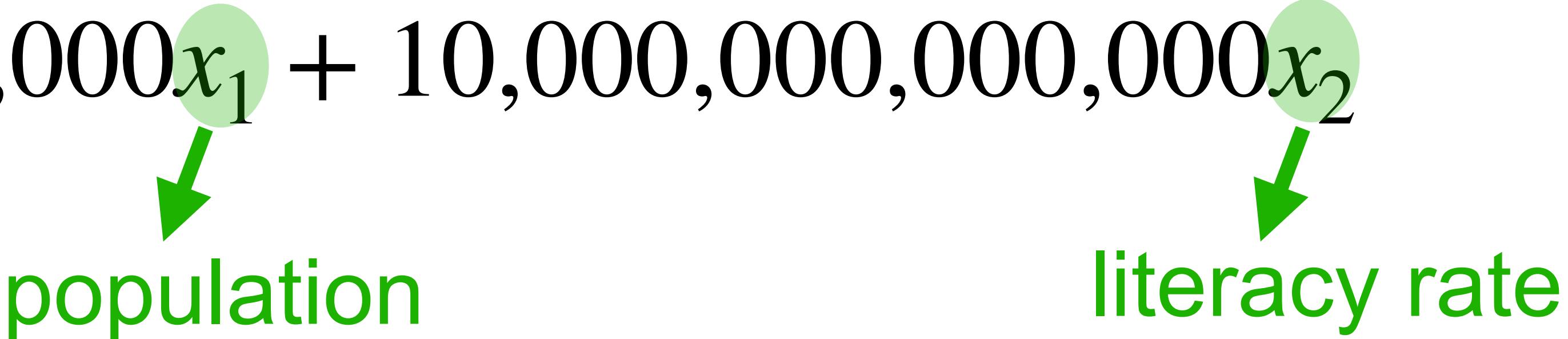
Interpretability: *Which feature is more important?*

Parameters of ML algorithms: Constants that must hold for all variables (Step size in gradient descent with respect to which variable?)

SCALING THE DATA

Features vary over a wide range \Rightarrow Coefficients vary over wide range

Ex:

$$GDP = \$10,000x_1 + 10,000,000,000,000x_2$$


The equation shows GDP as a function of two features, x_1 and x_2 . Two green circles are placed around the variables x_1 and x_2 . Green arrows point from each circle to the corresponding feature name below the equation: 'population' under x_1 and 'literacy rate' under x_2 .

Problems:

Interpretability: *Which feature is more important?*

Parameters of ML algorithms: *Constants that must hold for all variables
(Step size in gradient descent with respect to which variable?)*

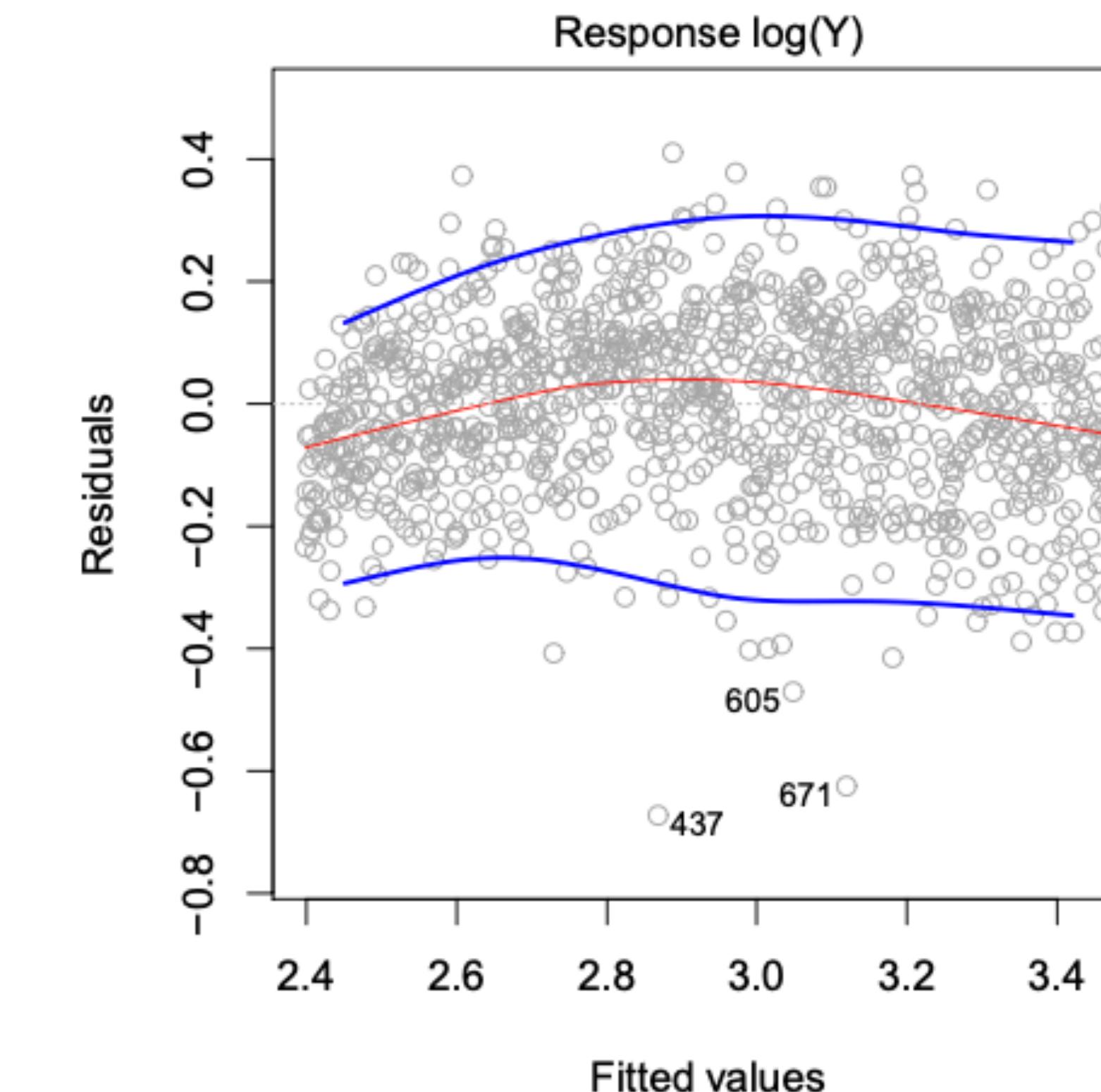
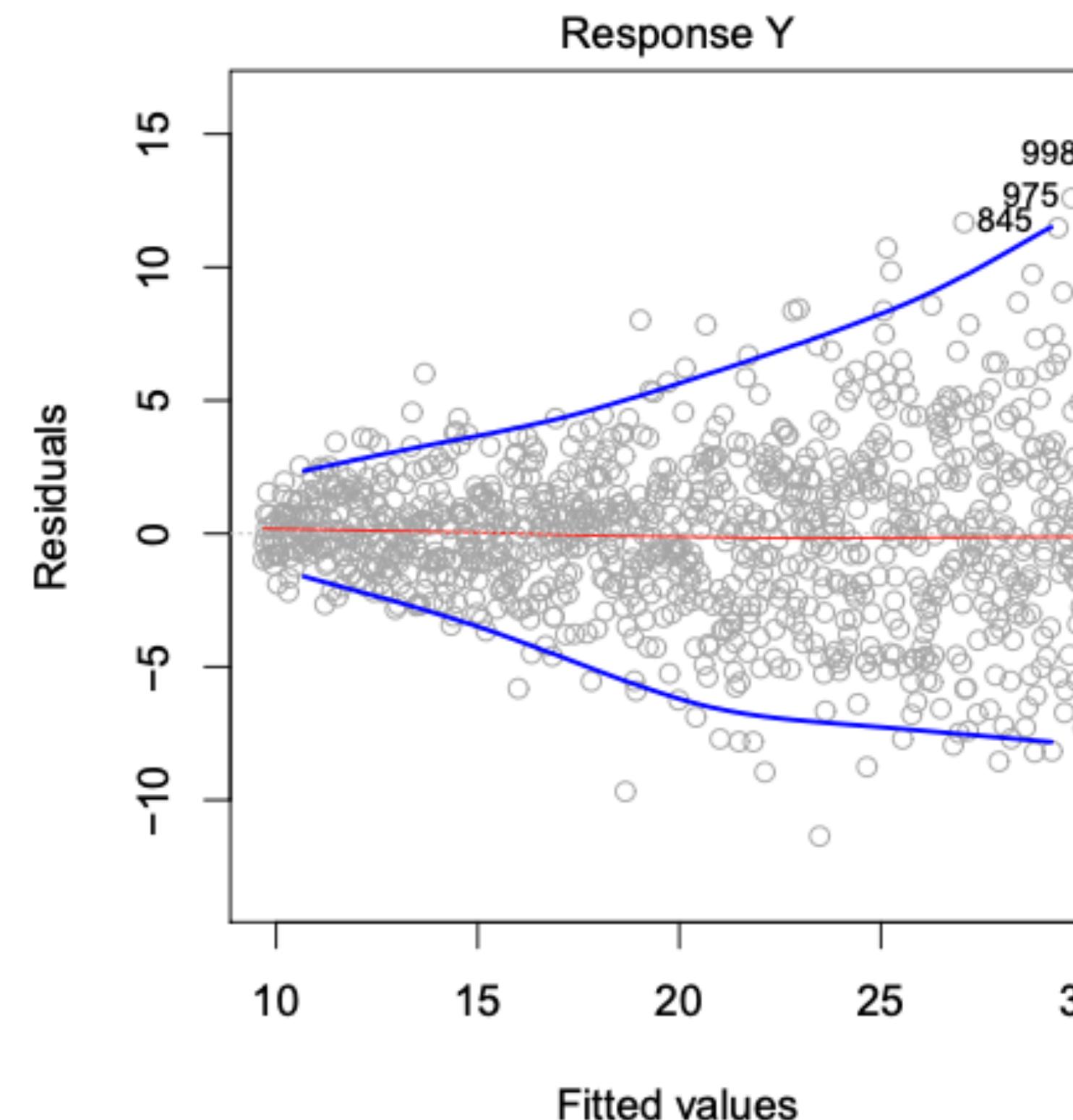
Solution: Bring them to the same scale by standardization \Rightarrow Z-score

(For feature j , $z_{ij} = \frac{x_{ij} - \bar{X}_j}{\hat{\sigma}_j}$ where \bar{X}_j is mean, $\hat{\sigma}_j$ is standard deviation of feature j .)

HETEROSCEDASTICITY

SE, Confidence interval etc. rely on constant variance of error terms.

Sometimes variance of error terms not constant:



Transform response with
 $\log Y$ or \sqrt{Y} .

DEALING WITH HIGHLY CORRELATED FEATURES

Having multiple features that are highly correlated is problematic.

Ex: 2 features: Annual income and net worth.

DEALING WITH HIGHLY CORRELATED FEATURES

Having multiple features that are highly correlated is problematic.

Ex: 2 features: Annual income and net worth.

Highly correlated features: not just neutral but harmful.

- Good models only on x_1 , or only on x_2 , or on any linear combination.
How to interpret the coefficients?
- Hypothesis tests for $w_i = 0 \Rightarrow$ different results based on which specific linear combination used in the model.

DEALING WITH HIGHLY CORRELATED FEATURES

Having multiple features that are highly correlated is problematic.

Ex: 2 features: Annual income and net worth.

Highly correlated features: not just neutral but harmful.

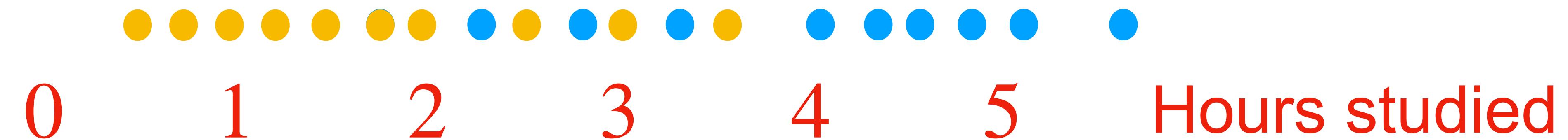
- Good models only on x_1 , or only on x_2 , or on any linear combination.
How to interpret the coefficients?
- Hypothesis tests for $w_i = 0 \Rightarrow$ different results based on which specific linear combination used in the model.

Solution:

- Check covariance matrix to find correlated features, eliminate one.
- OR combine into a single variable.

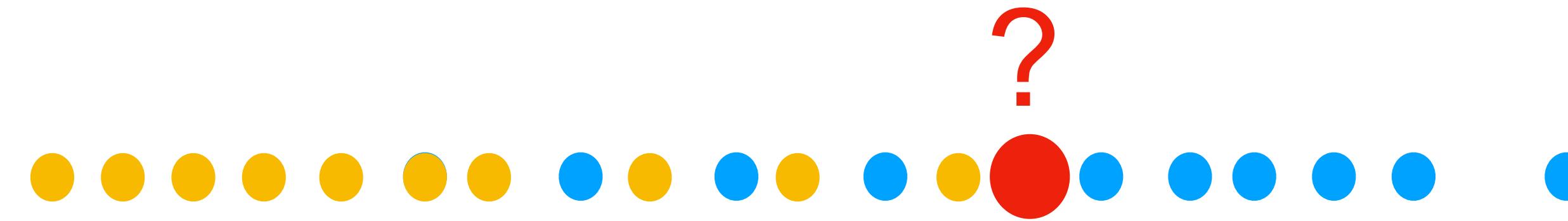
CLASSIFICATION WITH LOGISTIC REGRESSION

Training data: Number of hours studied for some course. We also have Pass (1) and Fail (0) response variable for the data points.



CLASSIFICATION WITH LOGISTIC REGRESSION

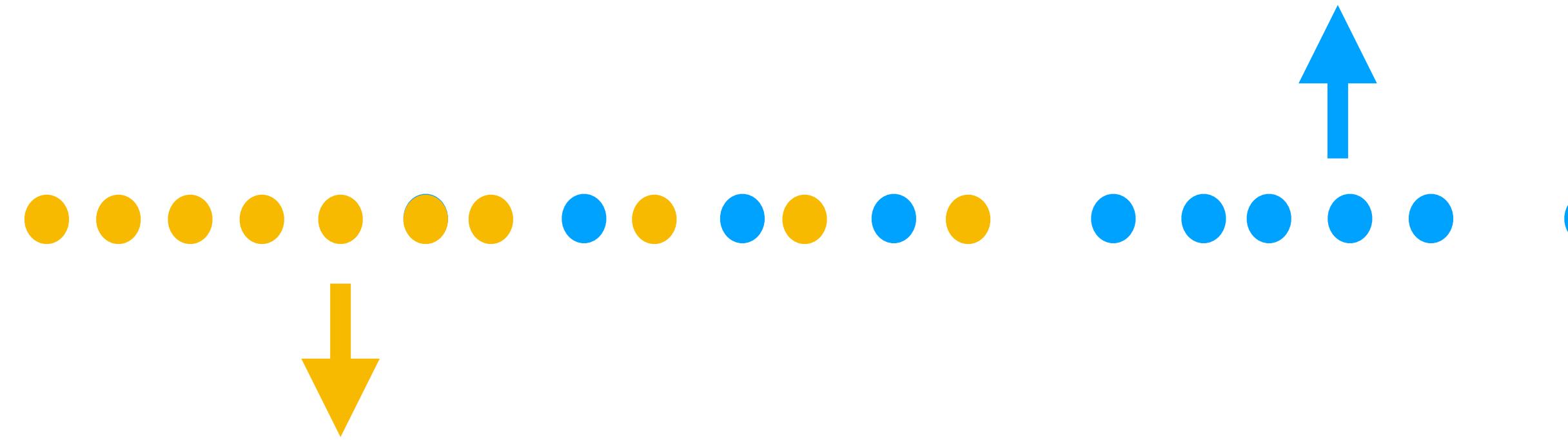
Training data: Number of hours studied for some course. We also have Pass (1) and Fail (0) response variable for the data points.



Can we train a model so that given a **new data point (red)** we can predict whether it corresponds to pass or fail?

Different from value prediction: we now try to classify data point.

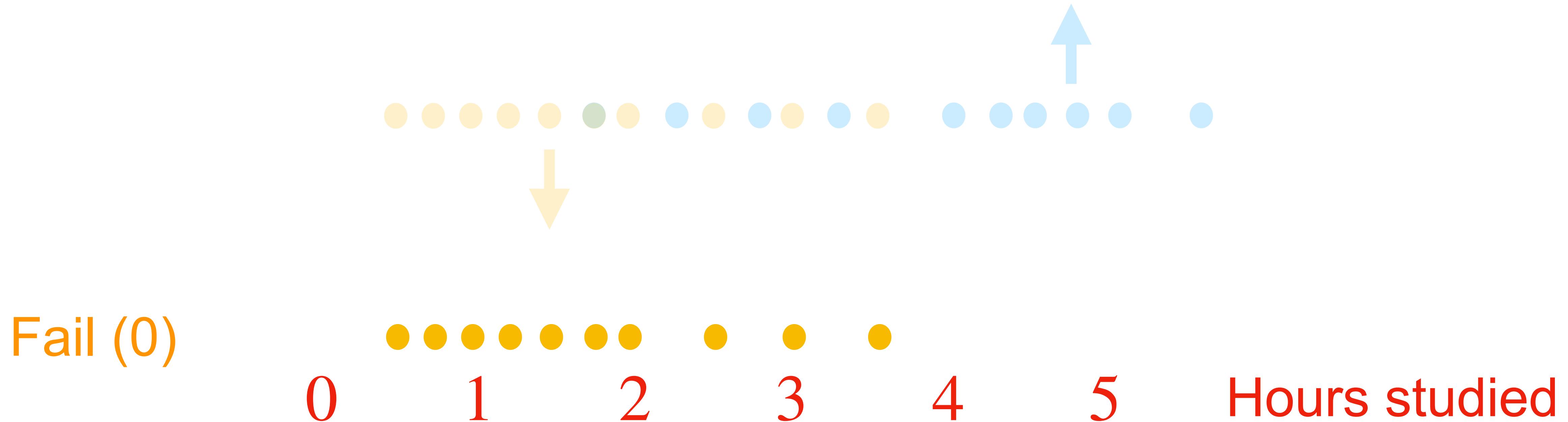
CLASSIFICATION WITH LOGISTIC REGRESSION



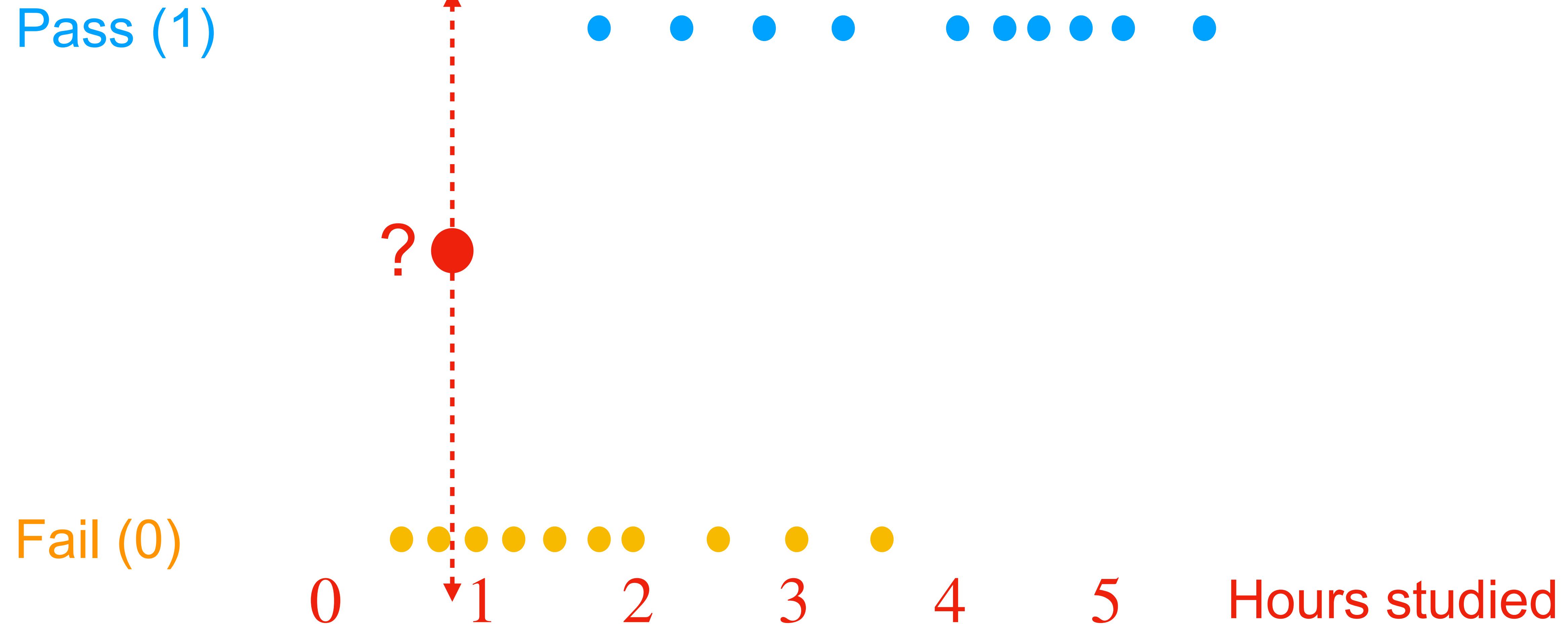
Let's first separate them preserving X values.

CLASSIFICATION WITH LOGISTIC REGRESSION

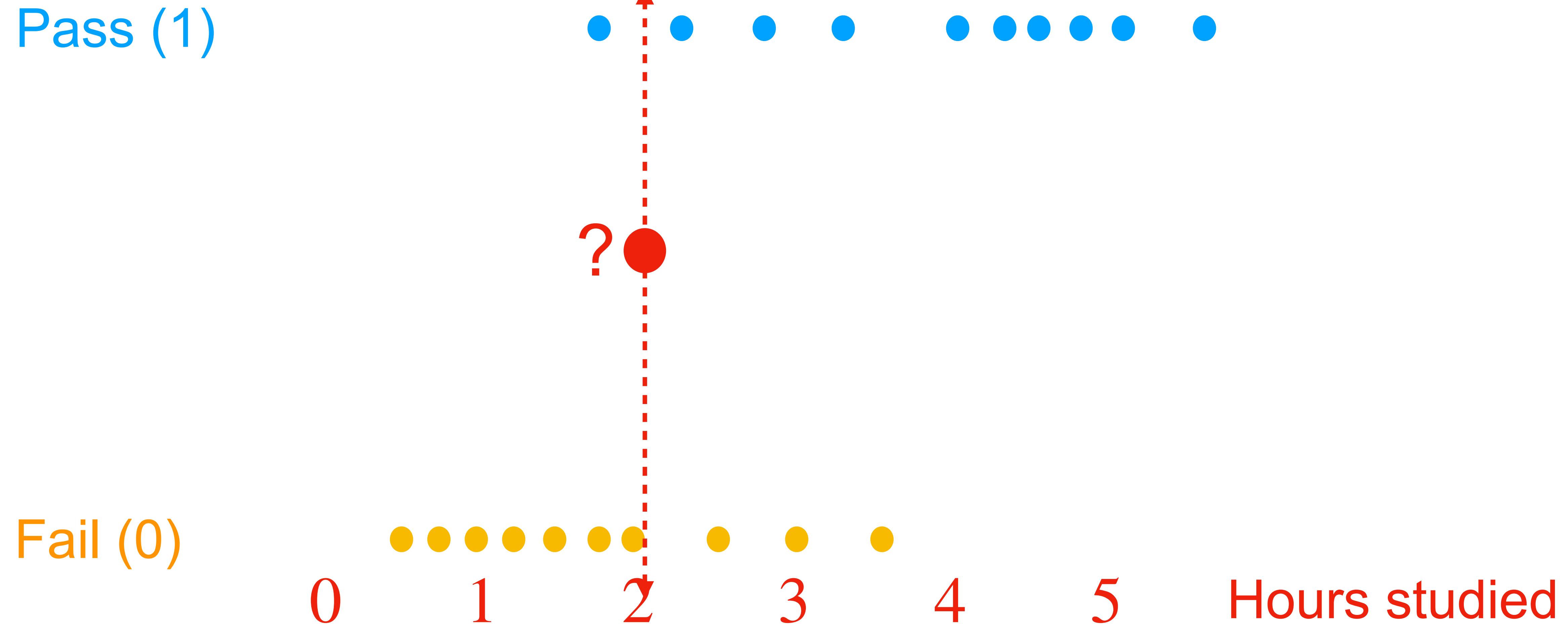
Pass (1)



CLASSIFICATION WITH LOGISTIC REGRESSION



CLASSIFICATION WITH LOGISTIC REGRESSION



CLASSIFICATION WITH LOGISTIC REGRESSION

Pass (1)

Fail (0)

0

1

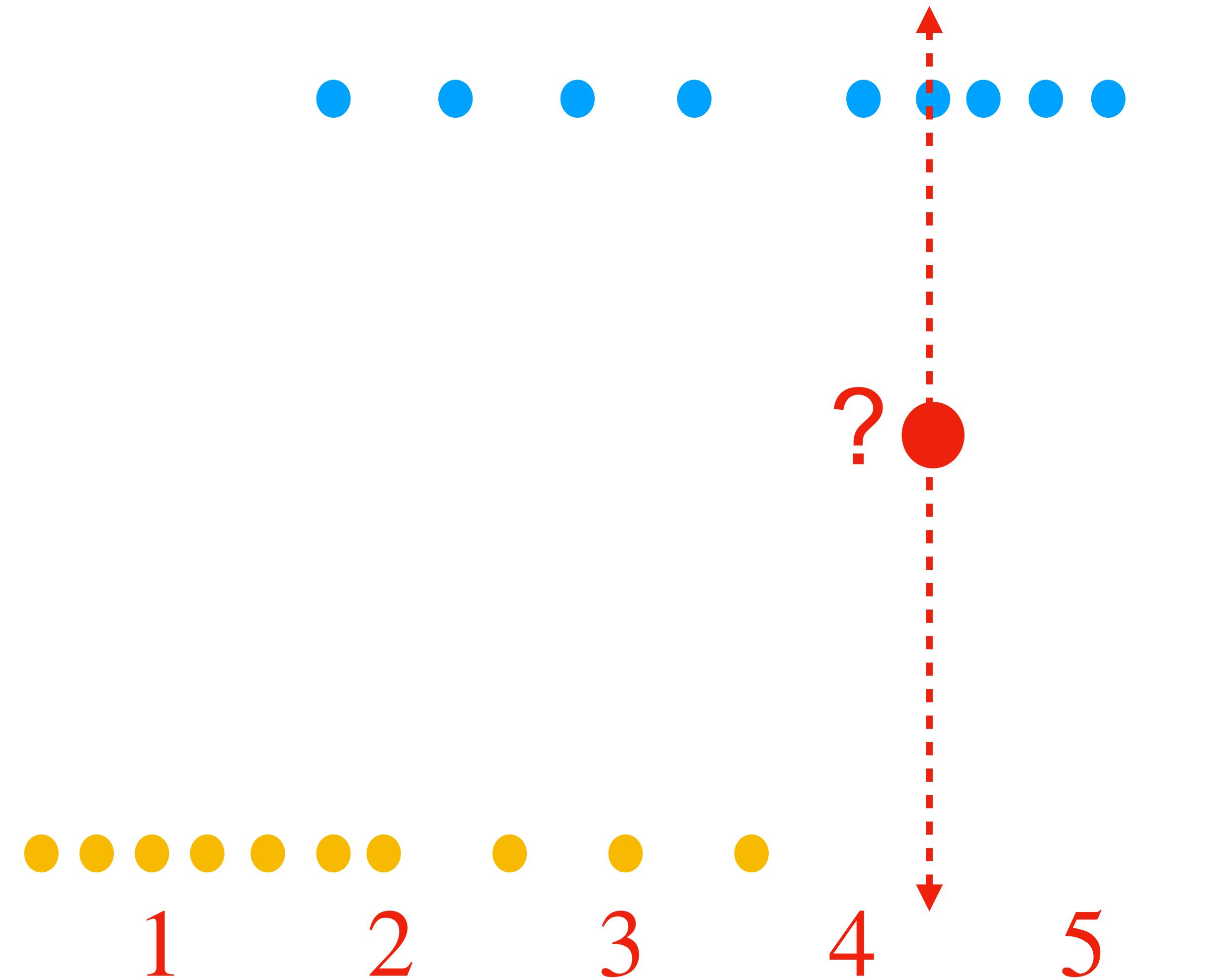
2

3

4

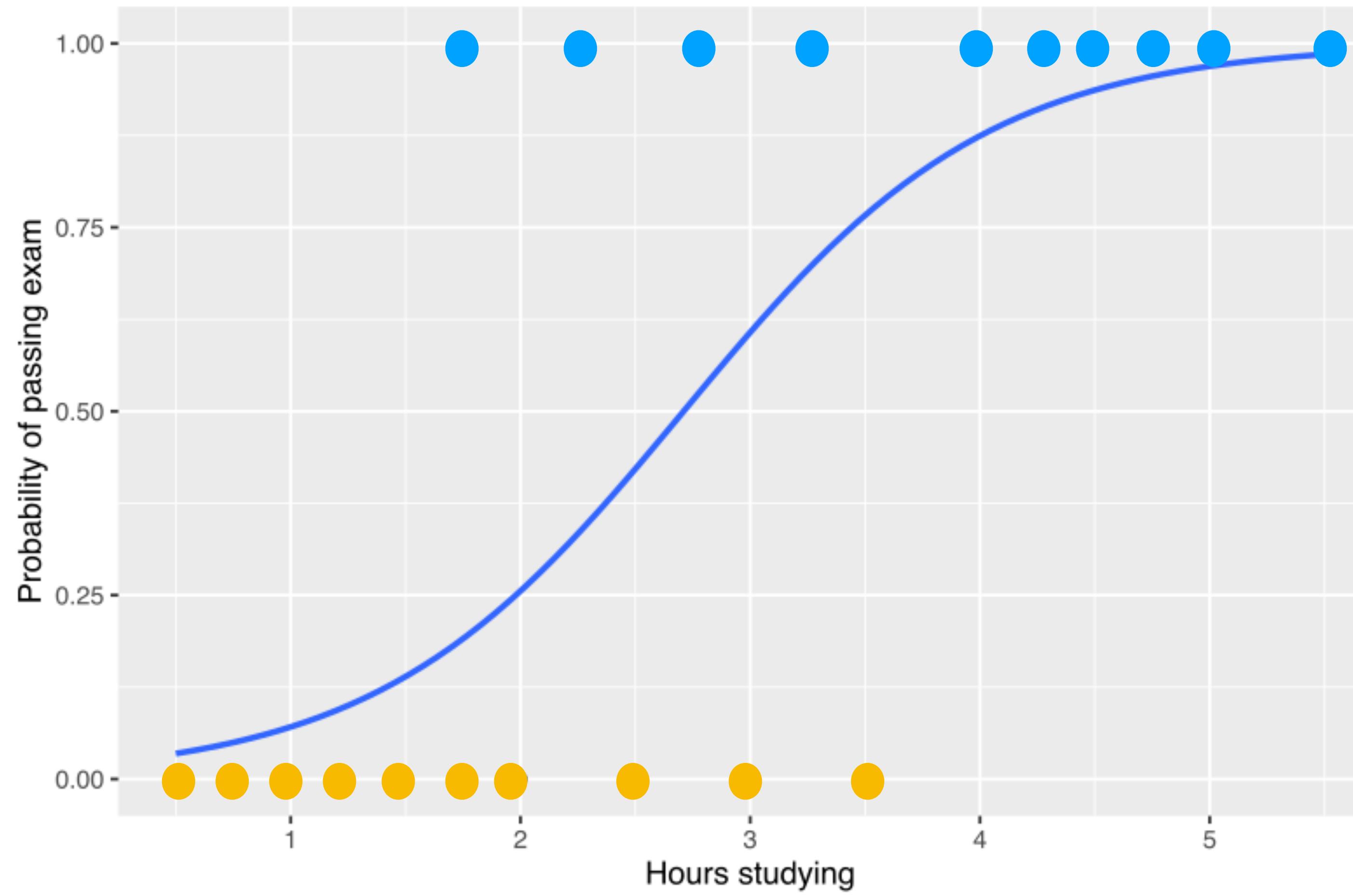
5

Hours studied



CLASSIFICATION WITH LOGISTIC REGRESSION

$Y = 1$



$Y = 0$

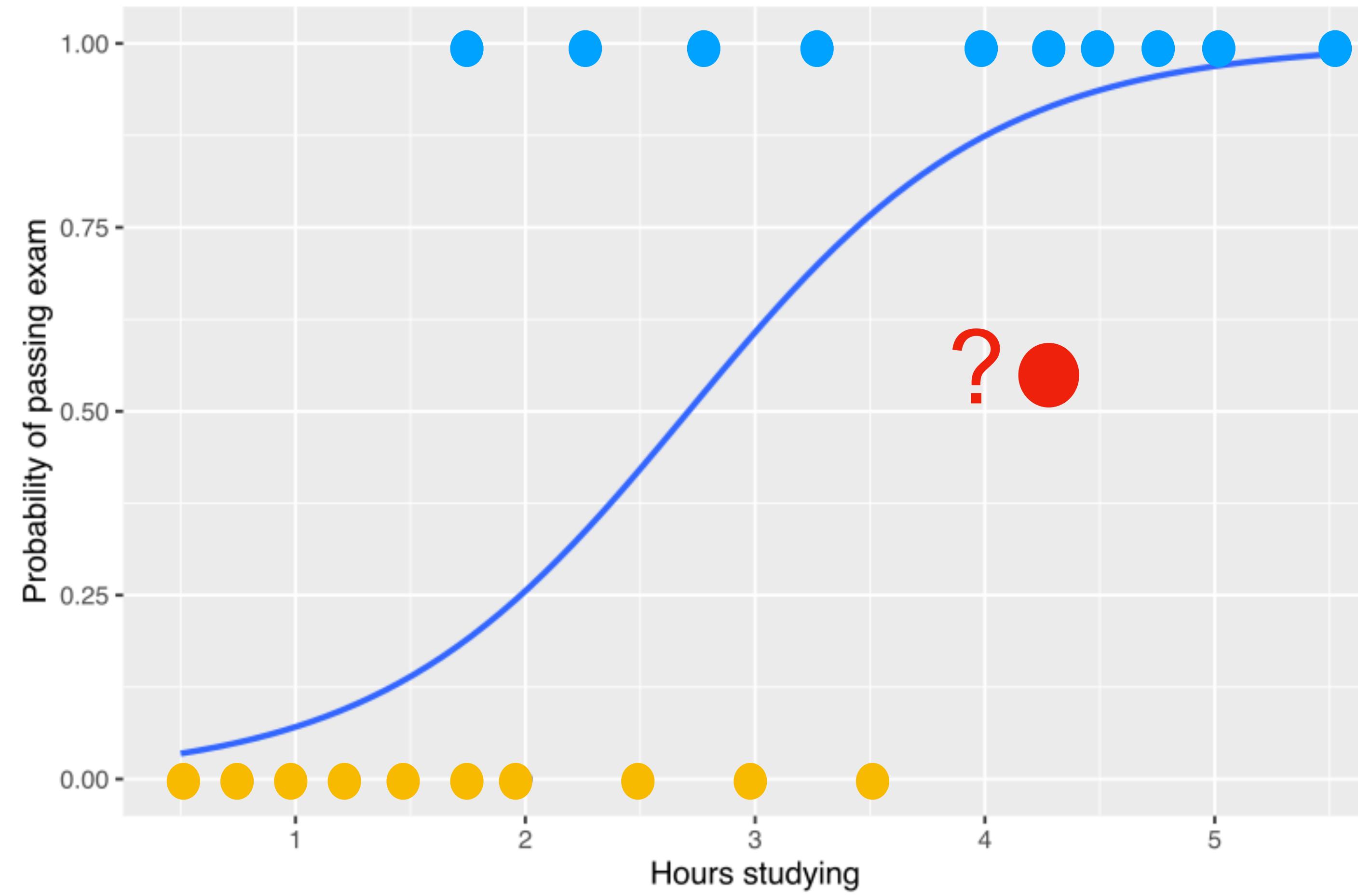
X : predictor (hours studying)
 Y : response (pass or fail)

Blue curve plots:
 $P(Y = 1 | X)$

CLASSIFICATION WITH LOGISTIC REGRESSION

$Y = 1$

$Y = 0$

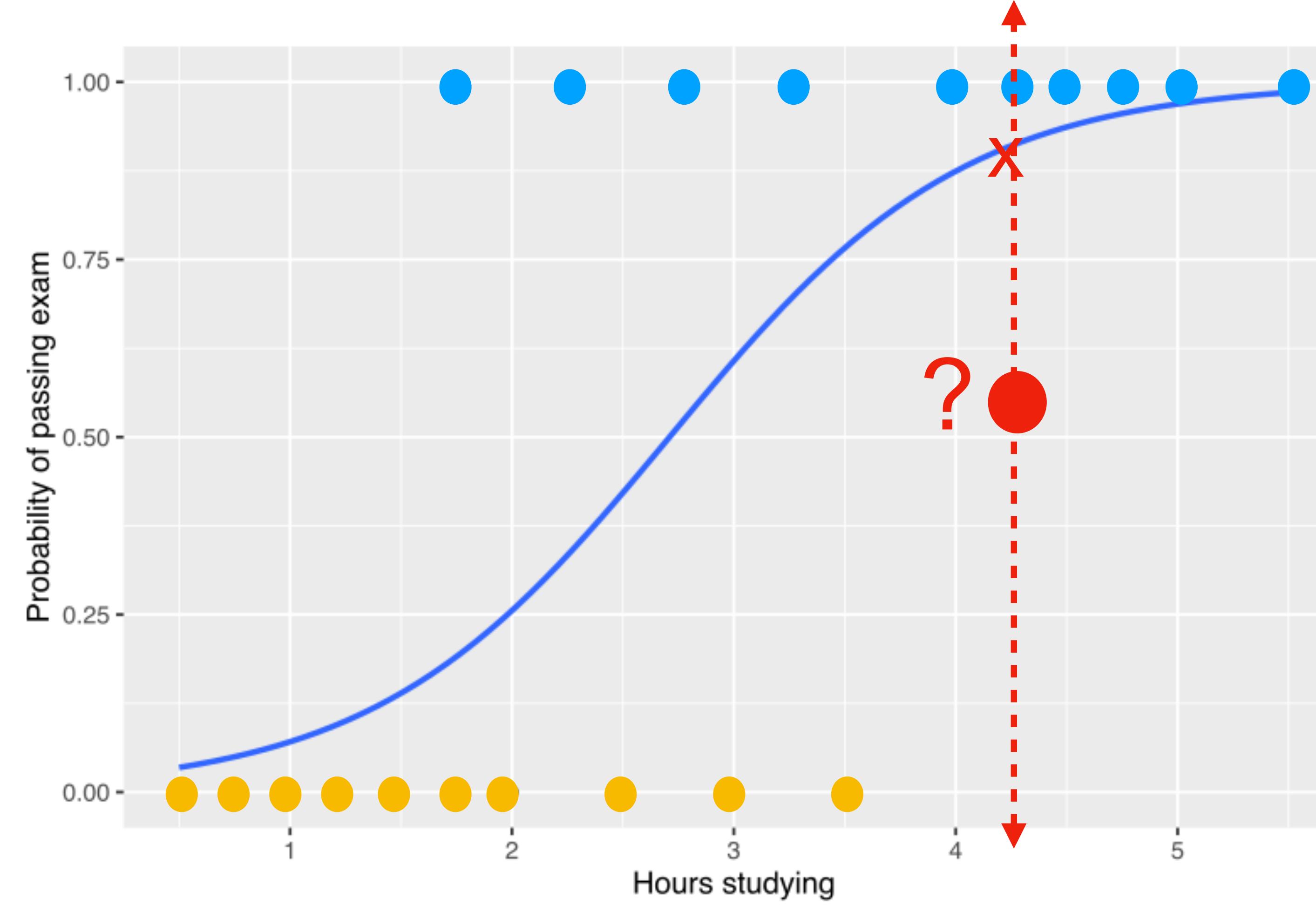


We can predict the class of red point based on blue curve:
if probability < 0.5
predict fail (0)
else
predict pass (1)

CLASSIFICATION WITH LOGISTIC REGRESSION

$Y = 1$

$Y = 0$

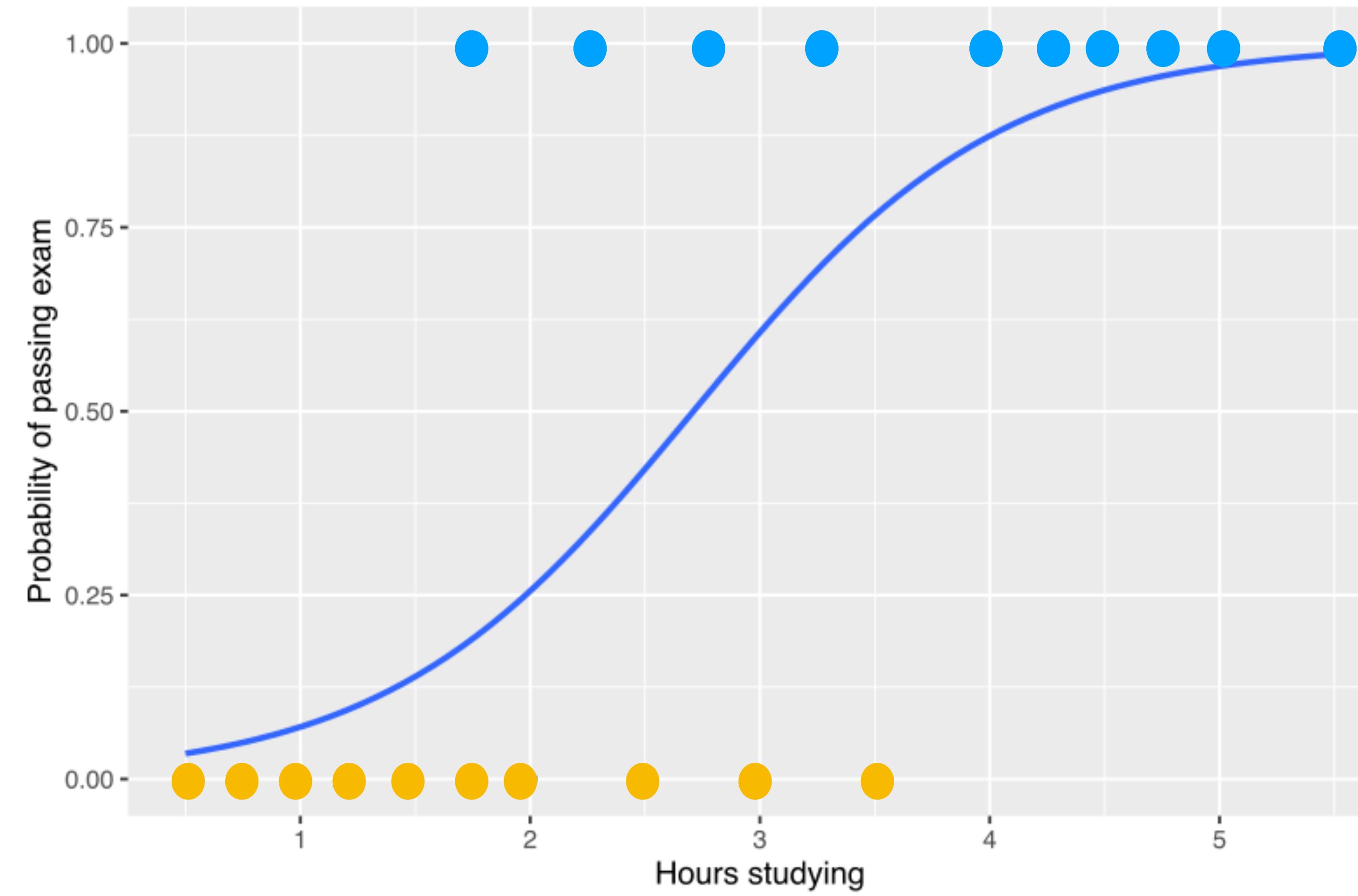


We can predict the class of red point based on blue curve:
if probability < 0.5
predict fail (0)
else
predict pass (1)

CLASSIFICATION WITH LOGISTIC REGRESSION

$Y = 1$

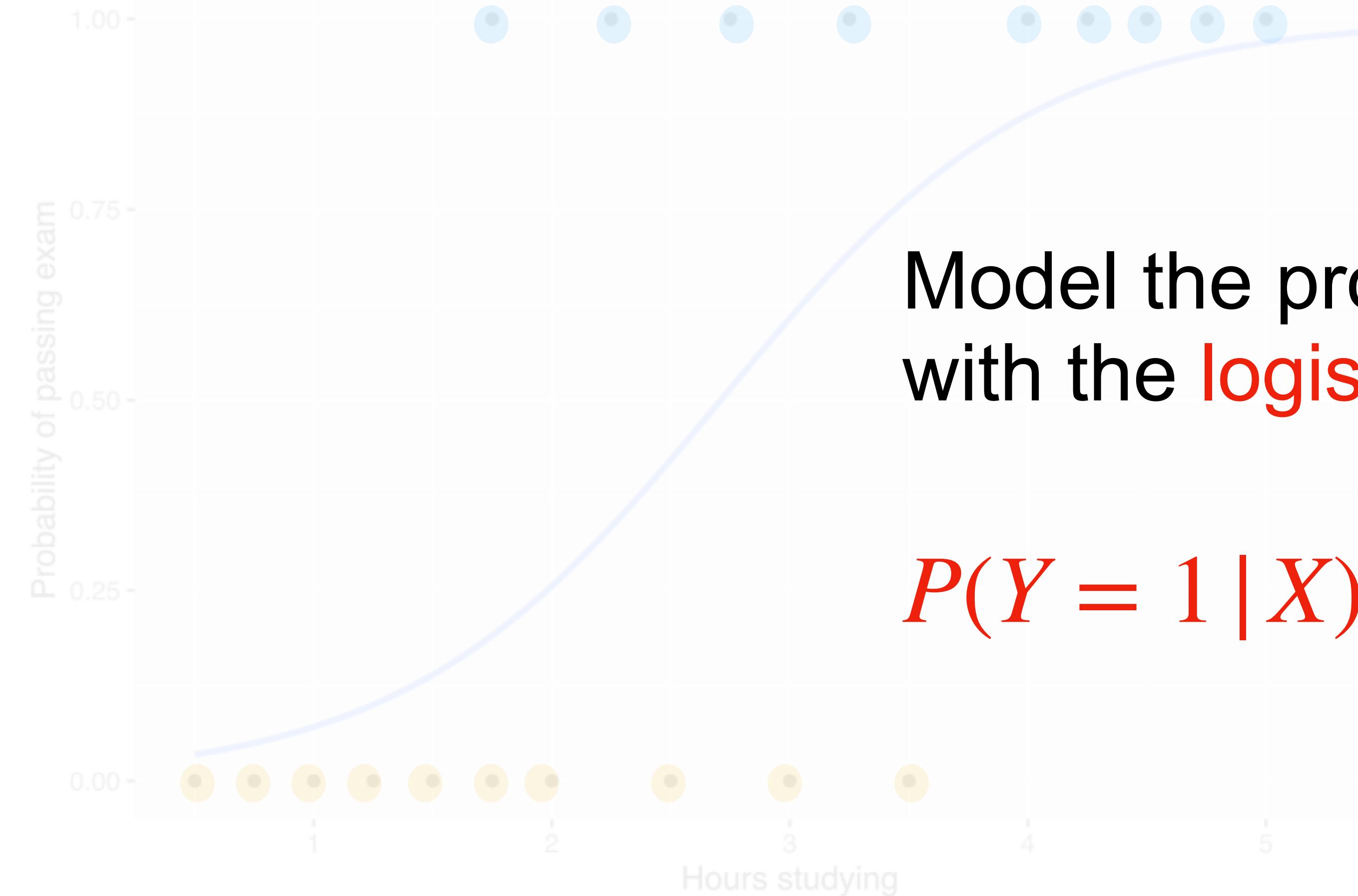
$Y = 0$



How to model the blue probability curve?

CLASSIFICATION WITH LOGISTIC REGRESSION

$Y = 1$



Model the probability curve
with the logistic function:

$$P(Y = 1 | X) = \frac{1}{1 + e^{-(w_0 + w_1 x)}}$$

$Y = 0$

CLASSIFICATION WITH LOGISTIC REGRESSION

Where does the logistic function come from?

CLASSIFICATION WITH LOGISTIC REGRESSION

Where does the **logistic function** come from?

From linear regression we know to work with **unbounded** functions.
A good **unbounded** function to predict?

CLASSIFICATION WITH LOGISTIC REGRESSION

Where does the **logistic function** come from?

From linear regression we know to work with **unbounded** functions.
A good **unbounded** function to predict?

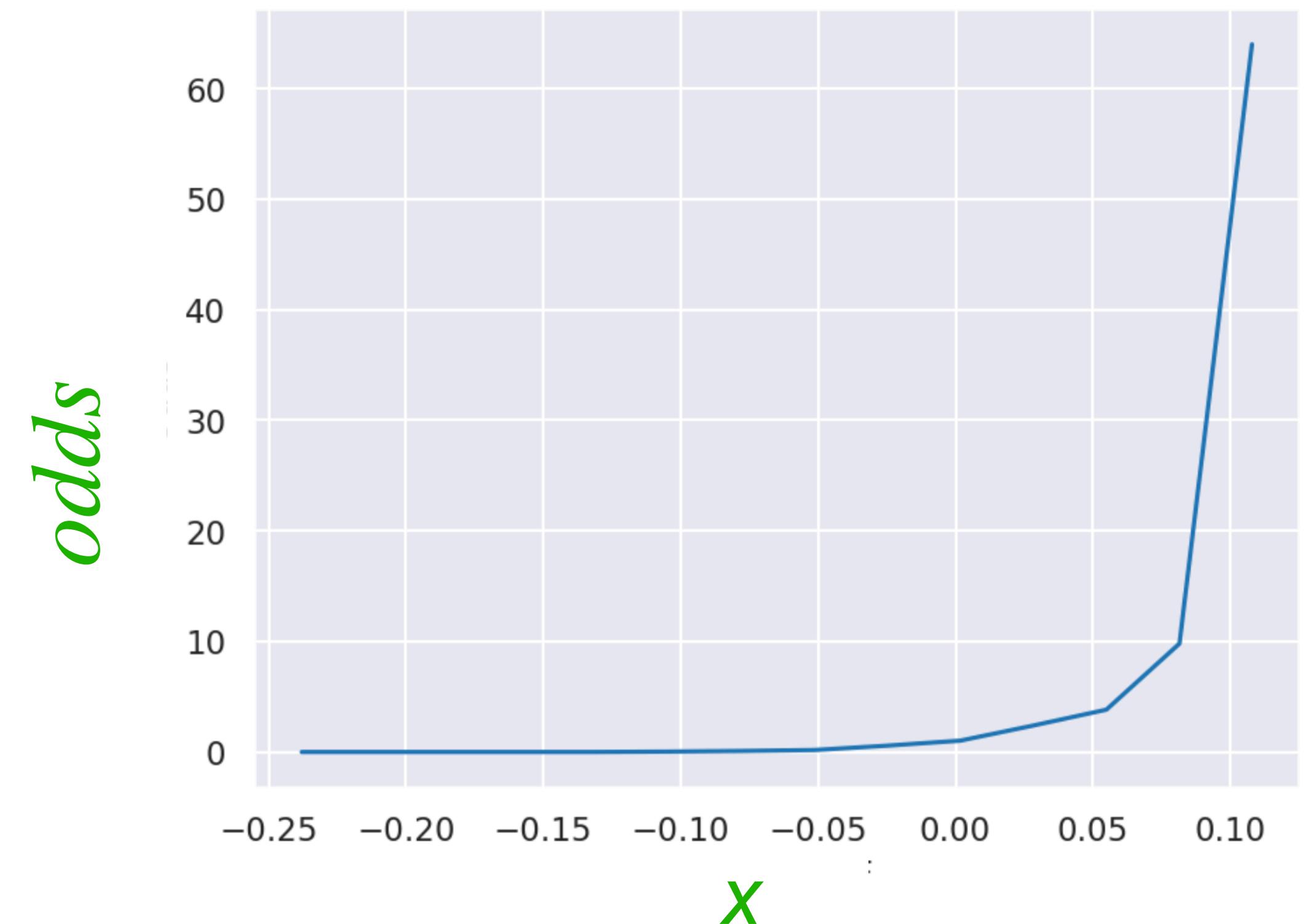
$$odds = \frac{P(Y = 1 | x)}{P(Y = 0 | x)} = \frac{p}{1 - p}$$

CLASSIFICATION WITH LOGISTIC REGRESSION

Where does the **logistic function** come from?

From linear regression we know to work with **unbounded** functions.
A good **unbounded** function to predict?

$$odds = \frac{P(Y = 1 | x)}{P(Y = 0 | x)} = \frac{p}{1 - p}$$



Intuitively, $0 \leq odds < 1$? $odds = 1$? $odds > 1$?

CLASSIFICATION WITH LOGISTIC REGRESSION

Where does the **logistic function** come from?

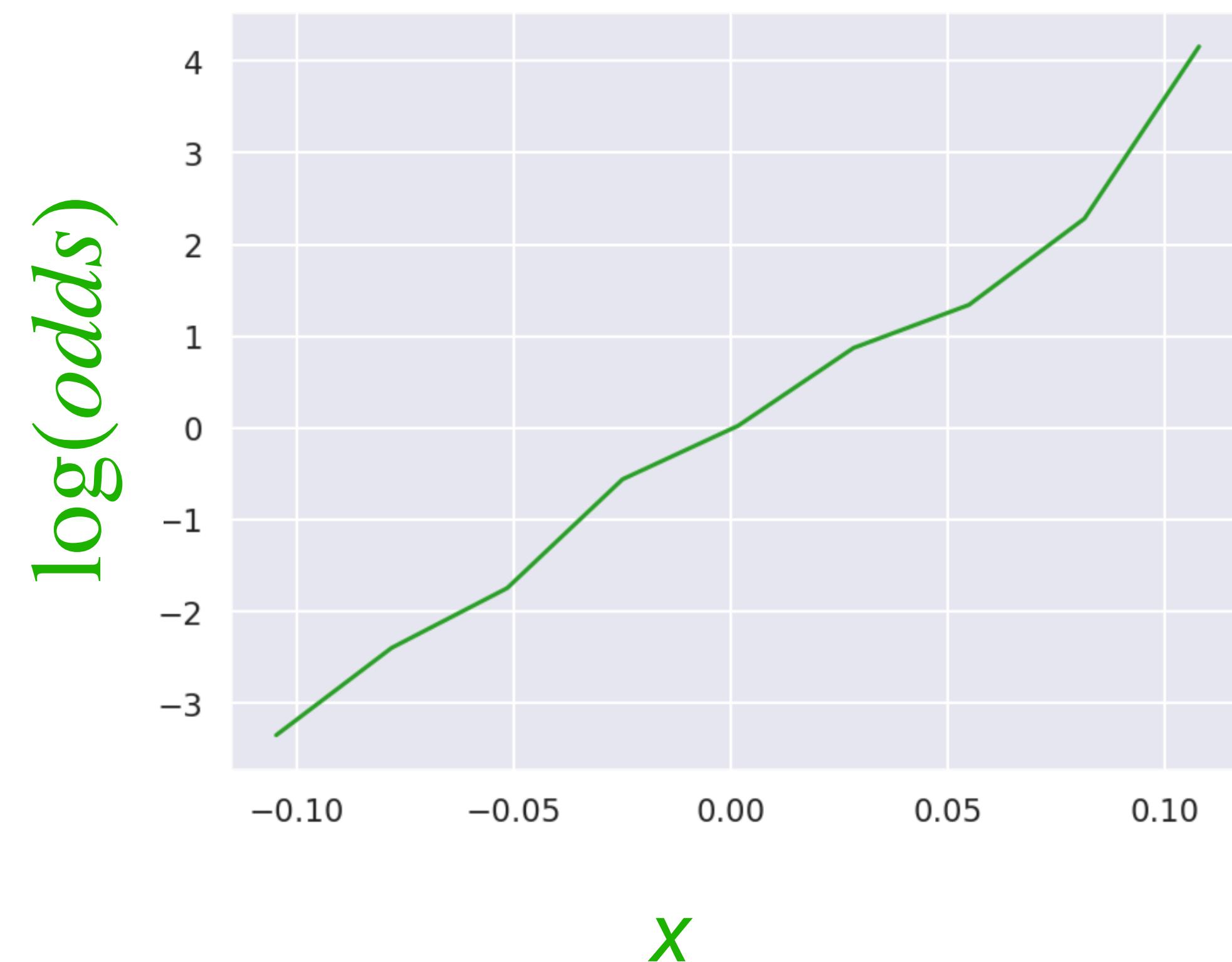
From linear regression we know to work with **linear** functions.
A good **unbounded and linear** function to predict?

CLASSIFICATION WITH LOGISTIC REGRESSION

Where does the **logistic function** come from?

From linear regression we know to work with **linear** functions.
A good **unbounded and linear** function to predict?

$$\log(\text{odds}) = \log\left(\frac{p}{1-p}\right)$$

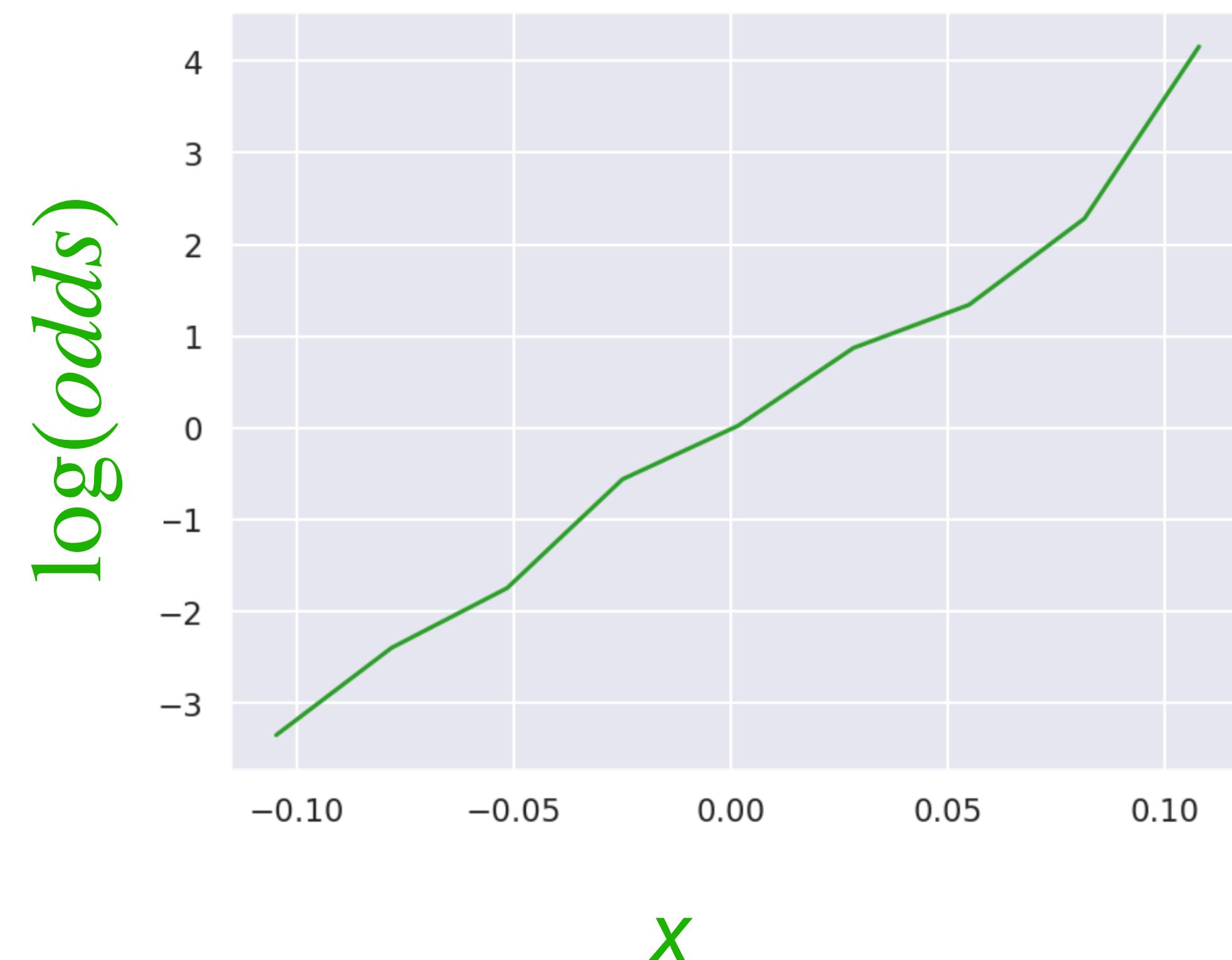


CLASSIFICATION WITH LOGISTIC REGRESSION

Where does the **logistic function** come from?

From linear regression we know to work with **linear** functions.
A good **unbounded and linear** function to predict?

$$\log(\text{odds}) = \log\left(\frac{p}{1-p}\right)$$



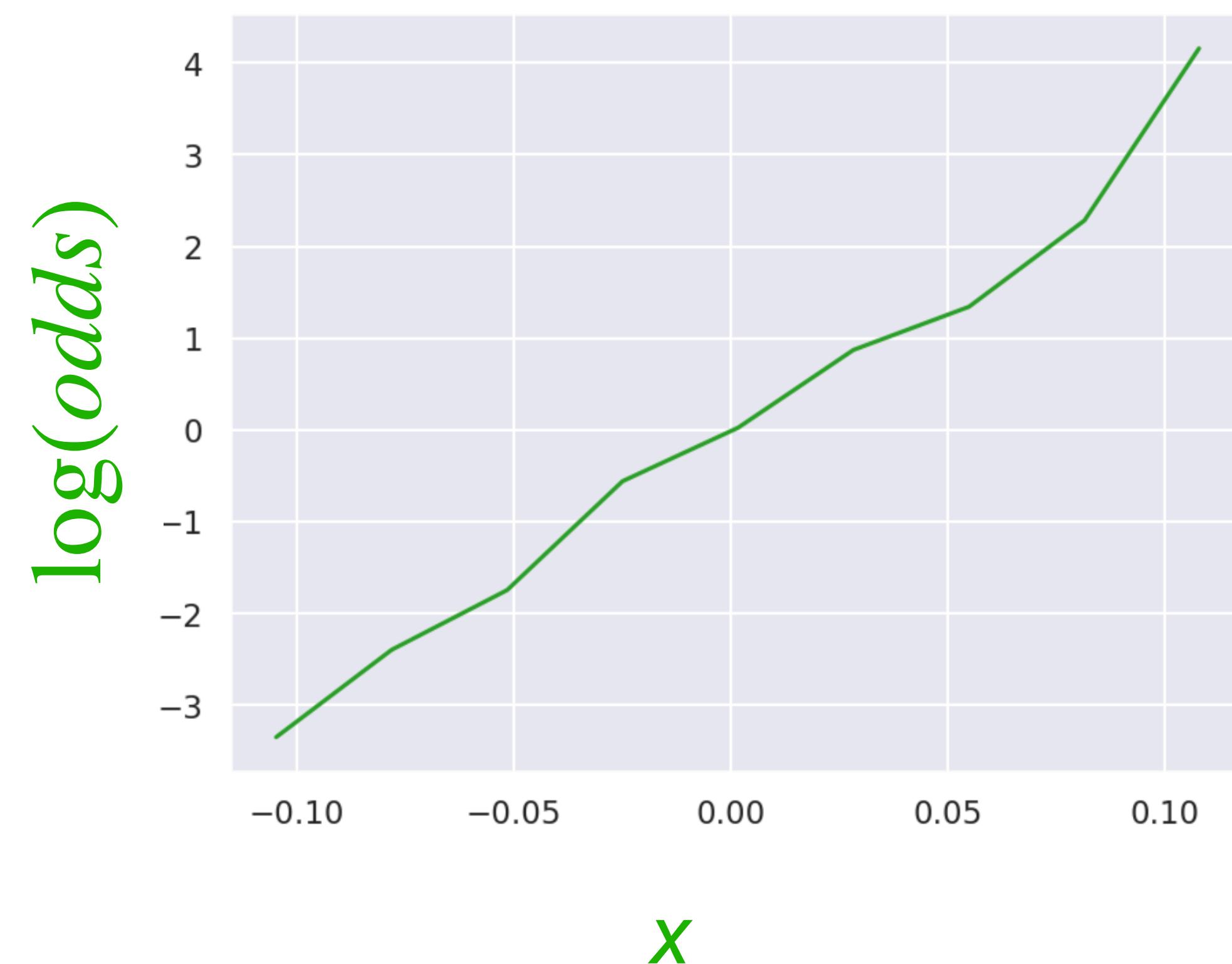
$$\log\left(\frac{p}{1-p}\right) = w_0 + w_1 x_1$$

CLASSIFICATION WITH LOGISTIC REGRESSION

Where does the **logistic function** come from?

From linear regression we know to work with **linear** functions.
A good **unbounded and linear** function to predict?

$$\log(\text{odds}) = \log\left(\frac{p}{1-p}\right)$$



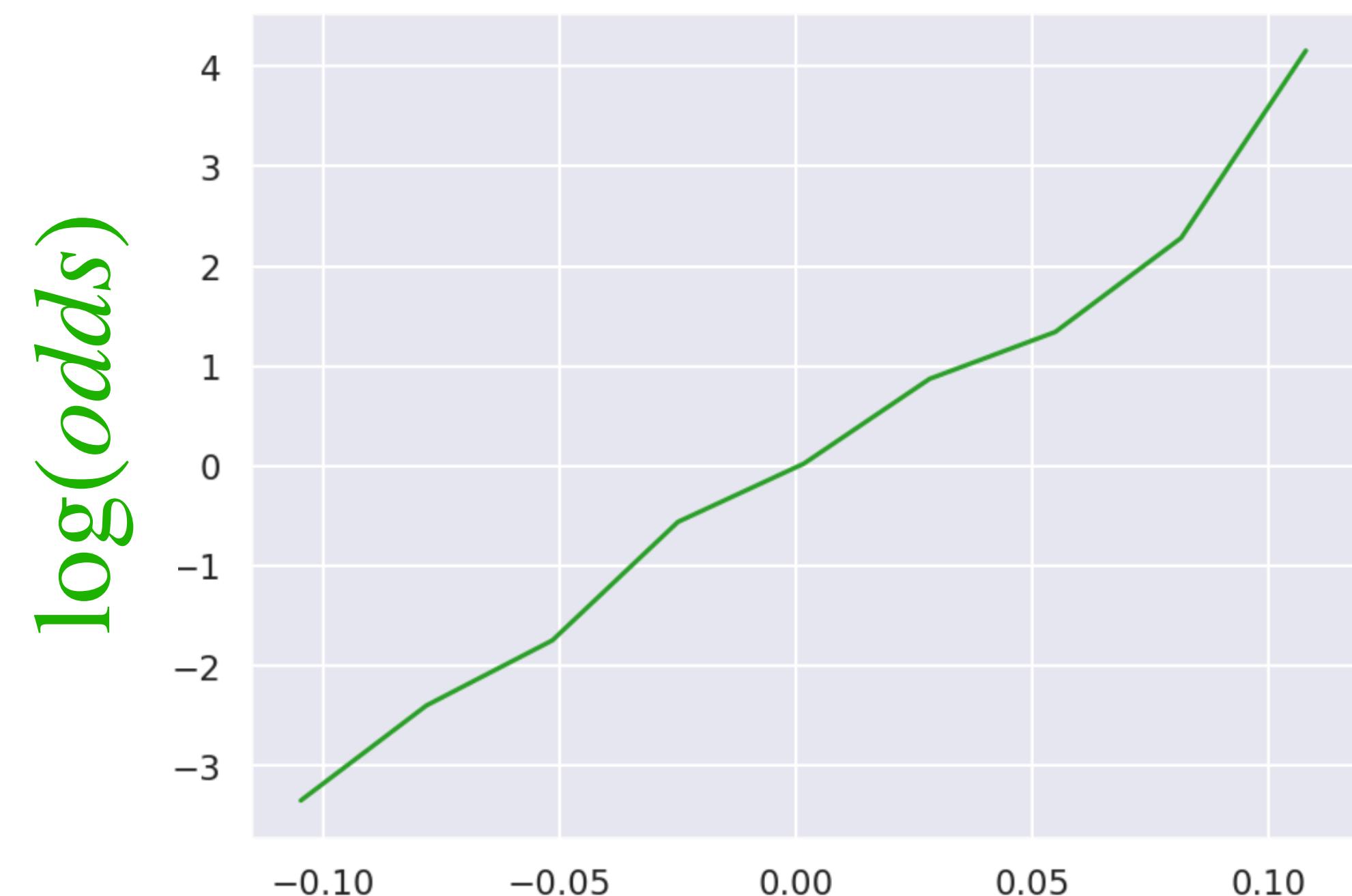
$$\log\left(\frac{p}{1-p}\right) = w_0 + w_1 x_1 \quad \Rightarrow \quad \frac{p}{1-p} = e^{w_0 + w_1 x_1}$$

CLASSIFICATION WITH LOGISTIC REGRESSION

Where does the **logistic function** come from?

From linear regression we know to work with **linear** functions.
A good **unbounded and linear** function to predict?

$$\log(\text{odds}) = \log\left(\frac{p}{1-p}\right)$$



$$\log\left(\frac{p}{1-p}\right) = w_0 + w_1 x_1 \quad \Rightarrow \quad \frac{p}{1-p} = e^{w_0 + w_1 x_1} \quad \Rightarrow \quad p = \frac{1}{1 + e^{-(w_0 + w_1 x_1)}}$$

CLASSIFICATION WITH LOGISTIC REGRESSION

Ex: Say we fit the model and got $w_0 = -0.15, w_1 = 0.575$.

What is the probability that a new student who has studied for $x_1 = 2$ hours passes?

CLASSIFICATION WITH LOGISTIC REGRESSION

Ex: Say we fit the model and got $w_0 = -0.15, w_1 = 0.575$.

What is the probability that a new student who has studied for $x_1 = 2$ hours passes?

$$P(Y_1 = 1 | x_1) = \frac{1}{1 + e^{-(w_0 + w_1 x_1)}} = \frac{1}{1 + e^{-(-0.15 + 0.575 \times 2)}} = \frac{1}{1 + e^{-1}} \approx 0.73$$

CLASSIFICATION WITH LOGISTIC REGRESSION

Loss function for linear regression was

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

where $\hat{y}_i = w_0 + w_1 x_i$

CLASSIFICATION WITH LOGISTIC REGRESSION

How about for logistic regression?

Loss function to minimize:

$$-\sum_{i=1}^n (y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

where

$$p_i = P(y_i = 1) = \frac{1}{1 + e^{-(w_0 + w_1 x)}}$$

Cross-entropy loss (*CE*)

CLASSIFICATION WITH LOGISTIC REGRESSION

How about for logistic regression?

Loss function to minimize:

$$-\sum_{i=1}^n (y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

where

$$p_i = P(y_i = 1) = \frac{1}{1 + e^{-(w_0 + w_1 x)}}$$

Cross-entropy loss (*CE*)

Intuitively, what is the error when:

$y_i = 1$ and $P(y_i = 1) \approx 1$?

$y_i = 1$ and $P(y_i = 1) \approx 0$?

$y_i = 0$ and $P(y_i = 1) \approx 1$?

$y_i = 0$ and $P(y_i = 1) \approx 0$?

CLASSIFICATION WITH LOGISTIC REGRESSION

Where does the loss function come from?

Reminder:

Likelihood: Probability of observed data viewed as a function of the parameters of a statistical model.

$$\mathcal{L}(\theta | x) = P_{\theta}(X = x)$$

Maximum Likelihood Estimation: Estimate parameters that makes observed data most probable.

CLASSIFICATION WITH LOGISTIC REGRESSION

Where does the loss function come from?

First some simplification:

$$P(Y_i = 1) = p_i \text{ and } P(Y_i = 0) = 1 - p_i$$

Combine them into a single equation:

$$P(Y_i = y_i) = p_i^{y_i} (1 - p_i)^{1-y_i}$$

CLASSIFICATION WITH LOGISTIC REGRESSION

Where does the loss function come from?

Maximize $\prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}$ Likelihood of data

CLASSIFICATION WITH LOGISTIC REGRESSION

Where does the loss function come from?

$$\text{Maximize} \quad \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}$$

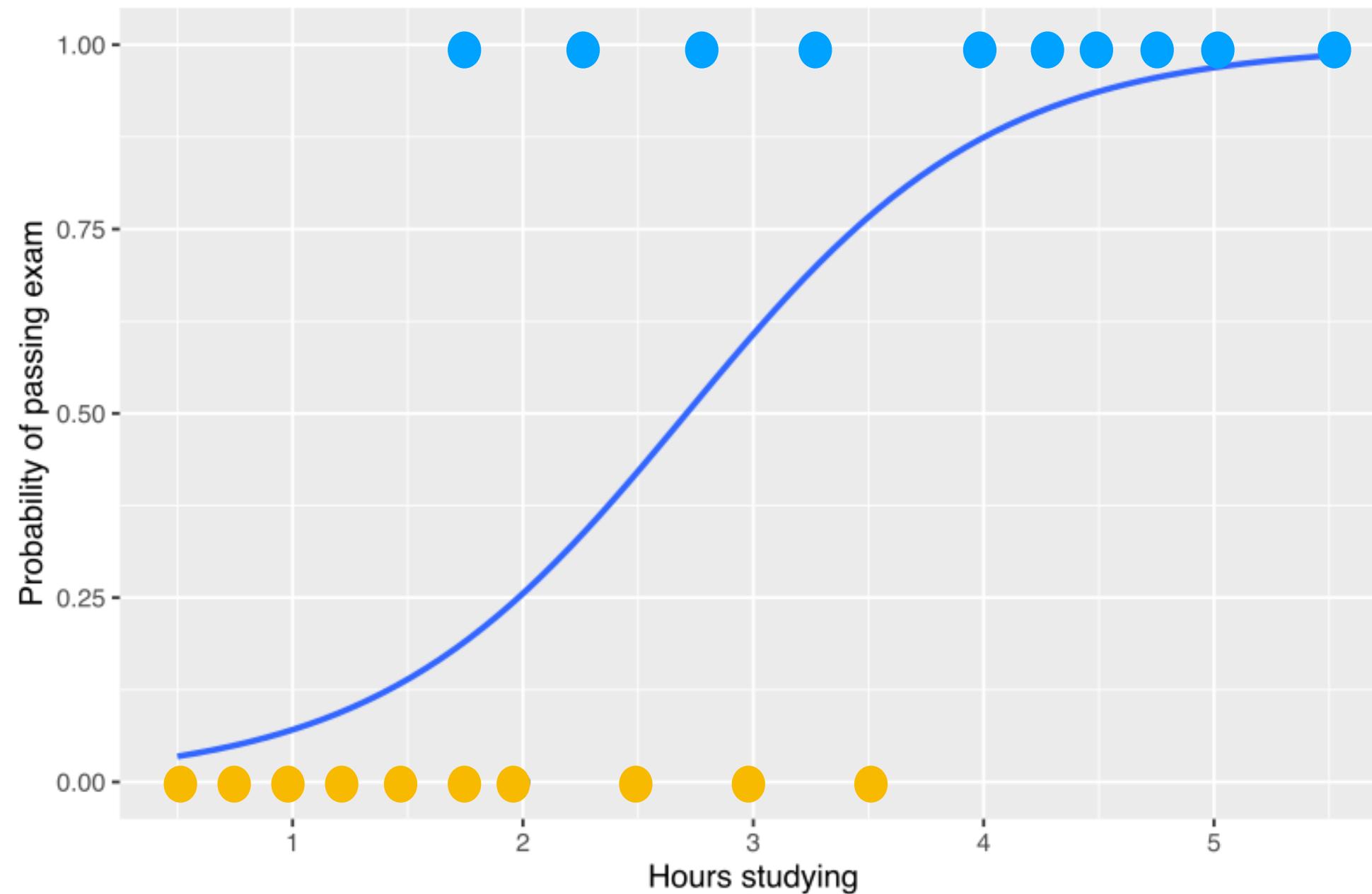
Likelihood of data

Maximize

$$\log\left(\prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}\right) = \sum_{i=1}^n \log(p_i^{y_i} (1 - p_i)^{1-y_i}) = \sum_{i=1}^n y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$$

$$\text{Minimize} \quad - \sum_{i=1}^n (y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

REVIEW OF BASICS OF LOGISTIC REGRESSION



Model the probability curve with the **logistic function**:

$$P(Y = 1 | X) = \frac{1}{1 + e^{-(w_0 + w_1 x)}}$$

Find w_0, w_1 that minimizes **cross-entropy loss**:

$$-\sum_{i=1}^n (y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

$$\text{where } p_i = \frac{1}{1 + e^{-(w_0 + w_1 x)}}$$

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Multiple logistic regression: What happens with multiple predictors?

Extends naturally. Logistic function:

$$P(Y = 1 | X) = \frac{1}{1 + e^{-(w_0 + w_1x_1 + \dots + w_mx_m)}}$$

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Multi-class Classification

Ex: Movie classification into “drama”, “action”, “documentary”, etc.

How to handle it?

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Multi-class Classification

Ex: Movie classification into “drama”, “action”, “documentary”, etc.

How to handle it?

Bad Idea: Assign 0,1,2,... to the classes, apply linear regression.

- What if classes can't be ordered naturally (data not ordinal)?
- Ex: drama:0, action:1, documentary=2, ...

drama < action < documentary means what?

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Multi-class Classification

Ex: Movie classification into “drama”, “action”, “documentary”, etc.

How to handle it?

Bad Idea: Assign 0,1,2,... to the classes, apply linear regression.

- What if classes can't be ordered naturally (data not ordinal)?
- Ex: drama:0, action:1, documentary=2, ...

drama < action < documentary means what?

May work: For classes defined by *Likert scales*:

Completely Agree \leftrightarrow Mostly Agree \leftrightarrow Neutral \leftrightarrow Mostly Disagree \leftrightarrow Completely Disagree

Four stars \leftrightarrow Three stars \leftrightarrow Two stars \leftrightarrow One stars \leftrightarrow Zero stars

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Multi-class Classification

Ex: Movie classification into “drama”, “action”, “documentary”, etc.

How to handle it?

Better Idea: Multiple *one-vs-rest* classifiers

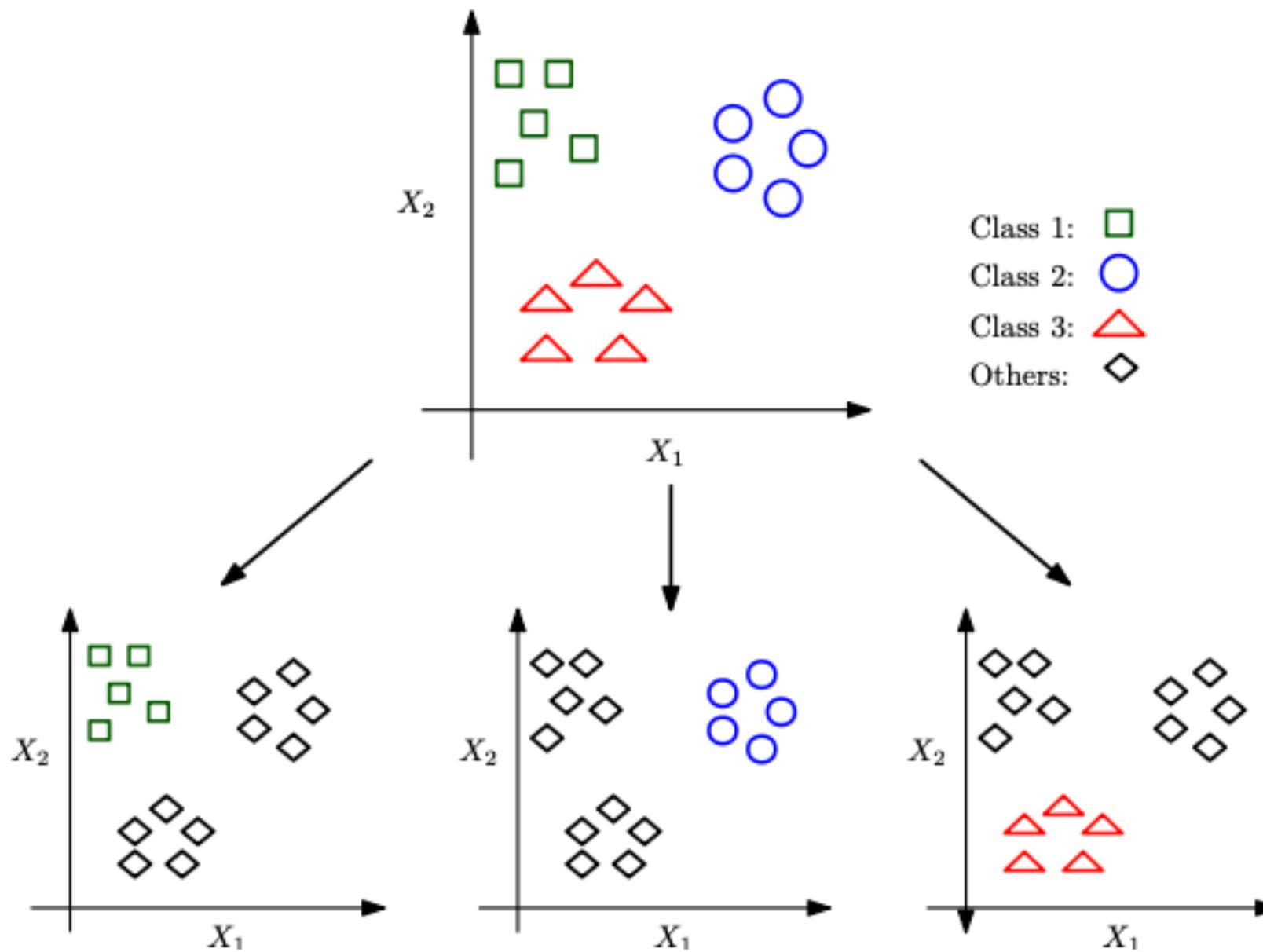
OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Multi-class Classification

Ex: Movie classification into “drama”, “action”, “documentary”, etc.

How to handle it?

Better Idea: Multiple *one-vs-rest* classifiers



Train multiple logistic classifiers:

For each C_i : C_i against the rest

To identify the class of a new x :

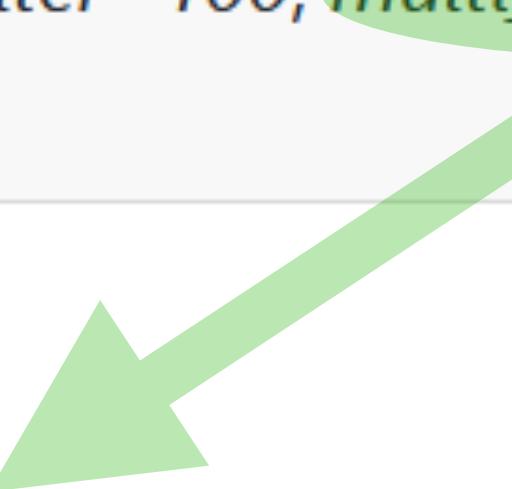
Test against each classifier.

Pick class with highest probability.

Another Option: Multinomial regression (Probabilities now sum to 1)

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

`sklearn.linear_model.LogisticRegression`

```
class sklearn.linear_model.LogisticRegression(penalty='l2', *, dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1,  
class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False,  
n_jobs=None, l1_ratio=None)  [source]
```

`multi_class='ovr'` implements one-vs-rest

`multi_class='multinomial'` implements multinomial

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Ex: Digits data from OCR

```
x, y = load_digits(return_X_y=True)
x.shape
```

```
(1797, 64)
```

```
x_train, x_test, y_train, y_test =\
    train_test_split(x, y, train_size=0.8, random_state=0)

scaler = StandardScaler()
x_train = scaler.fit_transform(x_train)

model = LogisticRegression(multi_class='ovr', random_state=0)
model.fit(x_train, y_train)

x_test = scaler.transform(x_test)
y_pred = model.predict(x_test)
print(f"Train accuracy: {model.score(x_train, y_train)}")
print(f"Test accuracy: {model.score(x_test, y_test)}")
```

```
Train accuracy: 0.9930410577592206
```

```
Test accuracy: 0.9583333333333334
```

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Ex: Digits data from OCR

```
x, y = load_digits(return_X_y=True)
x.shape
(1797, 64)

x_train, x_test, y_train, y_test =\
    train_test_split(x, y, train_size=0.8, random_state=0)
scaler = StandardScaler()
x_train = scaler.fit_transform(x_train)

model = LogisticRegression(multi_class='ovr', random_state=0)
model.fit(x_train, y_train)

x_test = scaler.transform(x_test)
y_pred = model.predict(x_test)
print(f"Train accuracy: {model.score(x_train, y_train)}")
print(f"Test accuracy: {model.score(x_test, y_test)}")

Train accuracy: 0.9930410577592206
Test accuracy: 0.9583333333333334
```

```
confusion_matrix(y_test, y_pred)

array([[27,  0,  0,  0,  0,  0,  0,  0,  0,  0],
       [ 0, 33,  0,  0,  0,  0,  1,  0,  1,  0],
       [ 0,  0, 35,  1,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0, 29,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0, 29,  0,  0,  1,  0,  0],
       [ 0,  1,  0,  0,  0, 39,  0,  0,  0,  0],
       [ 0,  1,  0,  0,  0,  0, 43,  0,  0,  0],
       [ 0,  0,  0,  0,  2,  0,  0, 37,  0,  0],
       [ 0,  3,  1,  0,  0,  0,  0,  0, 35,  0],
       [ 0,  0,  0,  1,  0,  1,  0,  1,  0, 38]])
```

```
print(classification_report(y_test, y_pred))

          precision    recall  f1-score

           0         1.00      1.00      1.00
           1         0.87      0.94      0.90
           2         0.97      0.97      0.97
           3         0.94      1.00      0.97
           4         0.94      0.97      0.95
           5         0.97      0.97      0.97
           6         0.98      0.98      0.98
           7         0.95      0.95      0.95
           8         0.97      0.90      0.93
           9         1.00      0.93      0.96
```

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Class Imbalance: Had already seen the issues related to it.

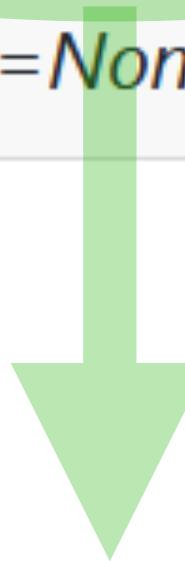
How to resolve?

- Discard members of the larger class to enforce balance.
- Replicate items from small class (with some random noise)
- Weigh the smaller class instances more in the loss function.

LOGISTIC REGRESSION

`sklearn.linear_model.LogisticRegression`

```
class sklearn.linear_model.LogisticRegression(penalty='l2', *, dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1,  
class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False,  
n_jobs=None, l1_ratio=None) \[source\]
```



Can provide a dictionary of class & weight associations

OR

`class_weight='balanced'`: values of y used to automatically adjust weights inversely proportional to class frequencies in the input data

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Regularization

Large weights are possible (especially with linearly separable data)

To avoid them and reduce model complexity apply regularization:

- (Like linear regression) Make sure features are standardized
- Minimize

$$Cross\ Entropy + \lambda \sum_{j=1}^m w_j^2$$

OTHER CONSIDERATIONS IN THE REGRESSION MODEL

Regularization

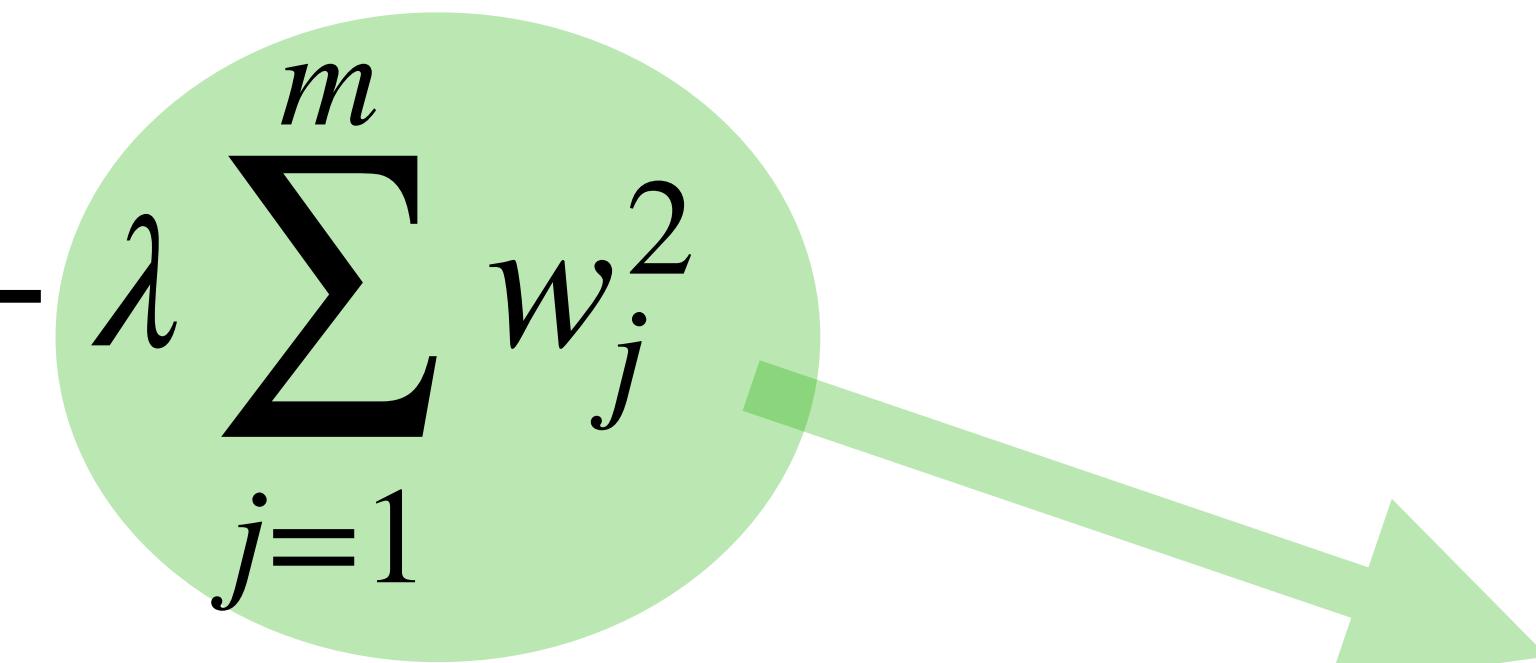
Large weights are possible (especially with linearly separable data)

To avoid them and reduce model complexity apply regularization:

- (Like linear regression) Make sure features are standardized
- Minimize

$$\text{Cross Entropy} + \lambda \sum_{j=1}^m w_j^2$$

Can replace with: $\lambda \sum_{j=1}^m |w_j|$ for LASSO.



OTHER CONSIDERATIONS IN THE REGRESSION MODEL

`sklearn.linear_model.LogisticRegression`

```
class sklearn.linear_model.LogisticRegression(penalty='l2', *, dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1,  
class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False,  
n_jobs=None, l1_ratio=None) 
```

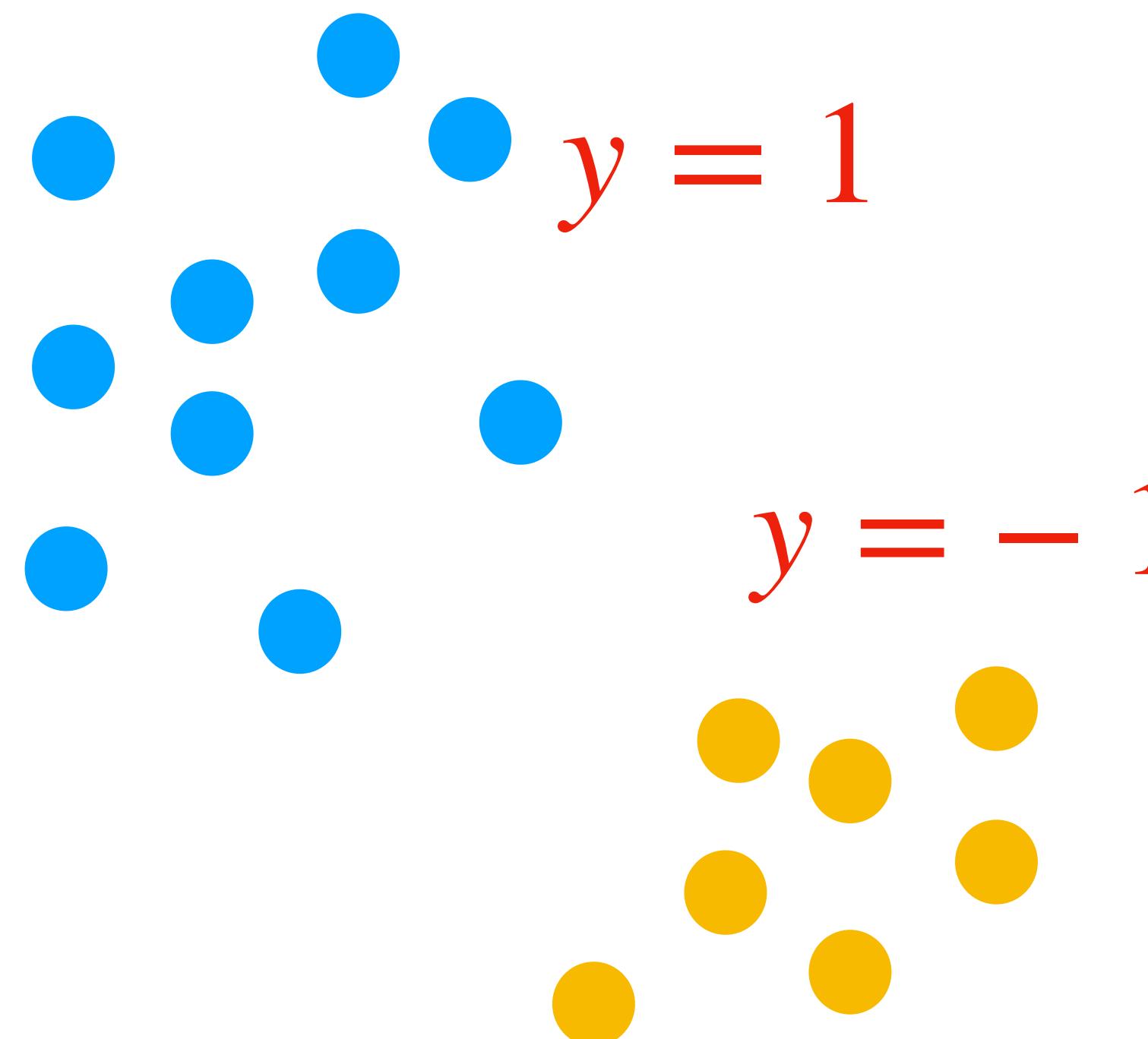
[source]

Make it 'l1' for LASSO:
may have to assign specific solver.

$C = 1/\lambda$: For aggressive regularization set to a small value.

SUPPORT VECTOR MACHINES (SVM)

SVM Summary:



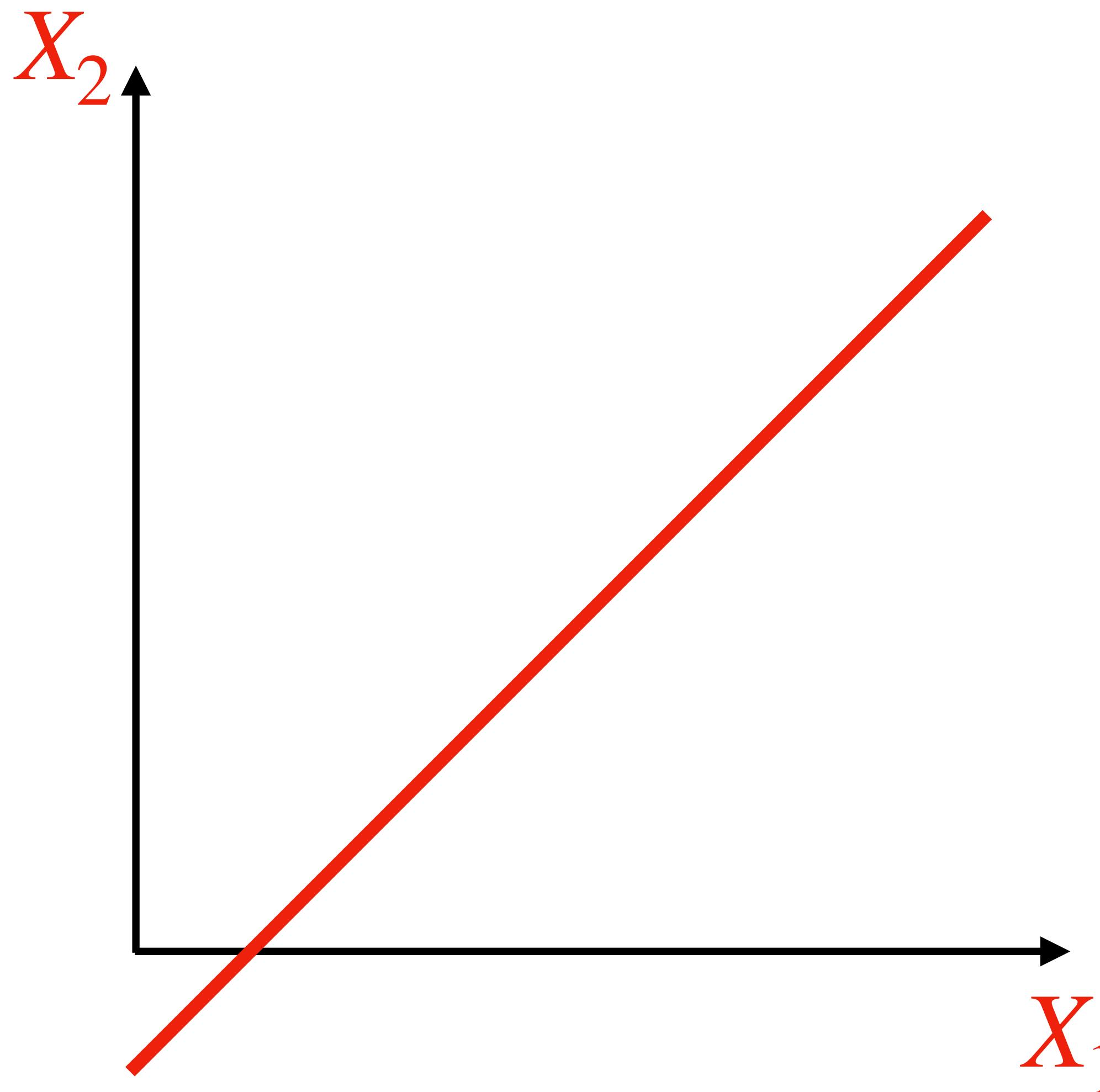
Let y denote the target variable.

For logistic regression we had labelled the classes 0,1.

For SVM let's label them with 1, -1 .

SUPPORT VECTOR MACHINES (SVM)

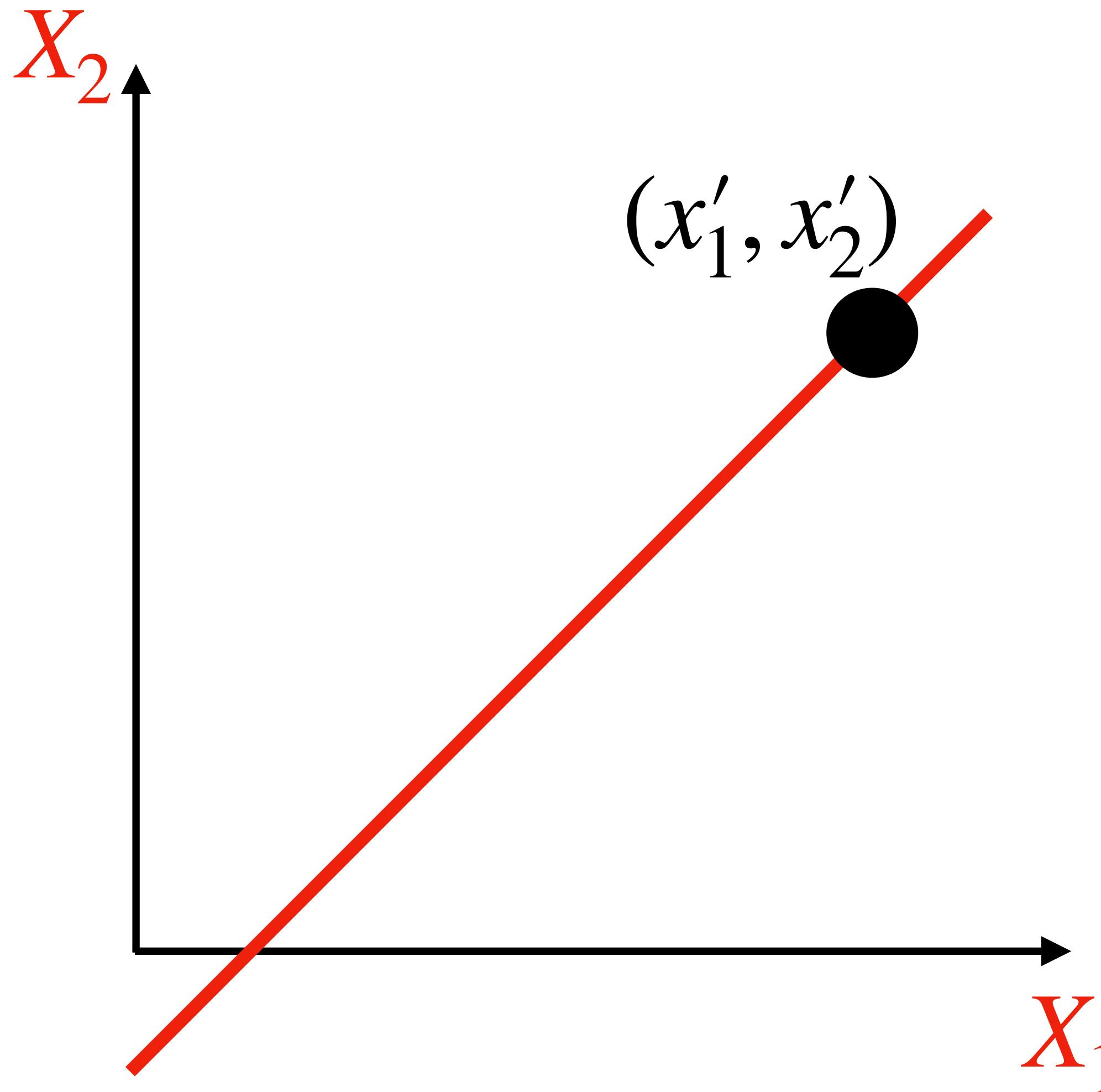
Hyperplane: Start with a hyperplane in 2D, which is a line.



$$w_0 + w_1X_1 + w_2X_2 = 0$$

SUPPORT VECTOR MACHINES (SVM)

Hyperplane: Start with a hyperplane in 2D, which is a line.



$$w_0 + w_1 X_1 + w_2 X_2 = 0$$

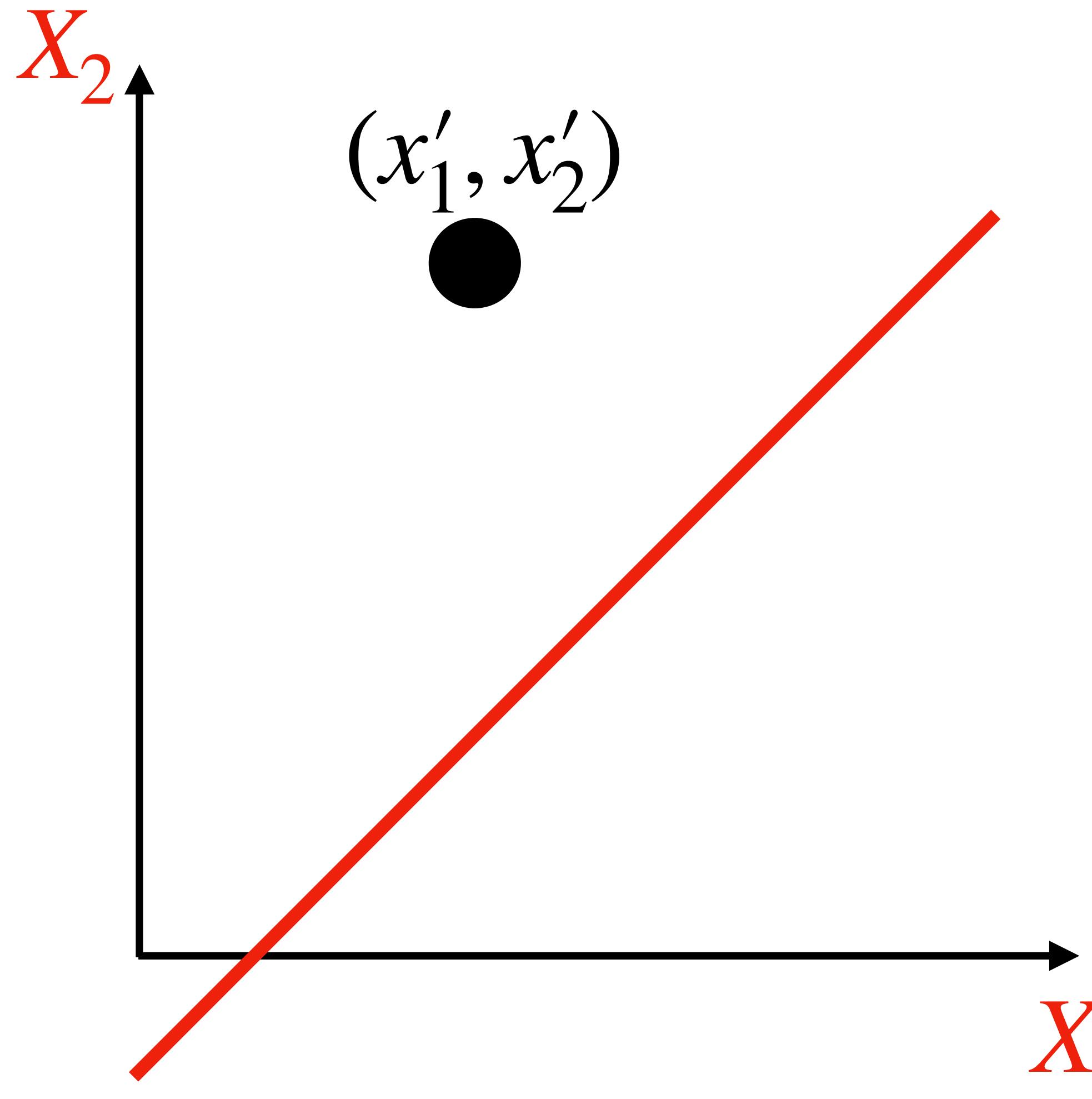
For a point (x'_1, x'_2) :

If $w_0 + w_1 x'_1 + w_2 x'_2 = 0$

then it is **on** the line.

SUPPORT VECTOR MACHINES (SVM)

Hyperplane: Start with a hyperplane in 2D, which is a line.



$$w_0 + w_1 X_1 + w_2 X_2 = 0$$

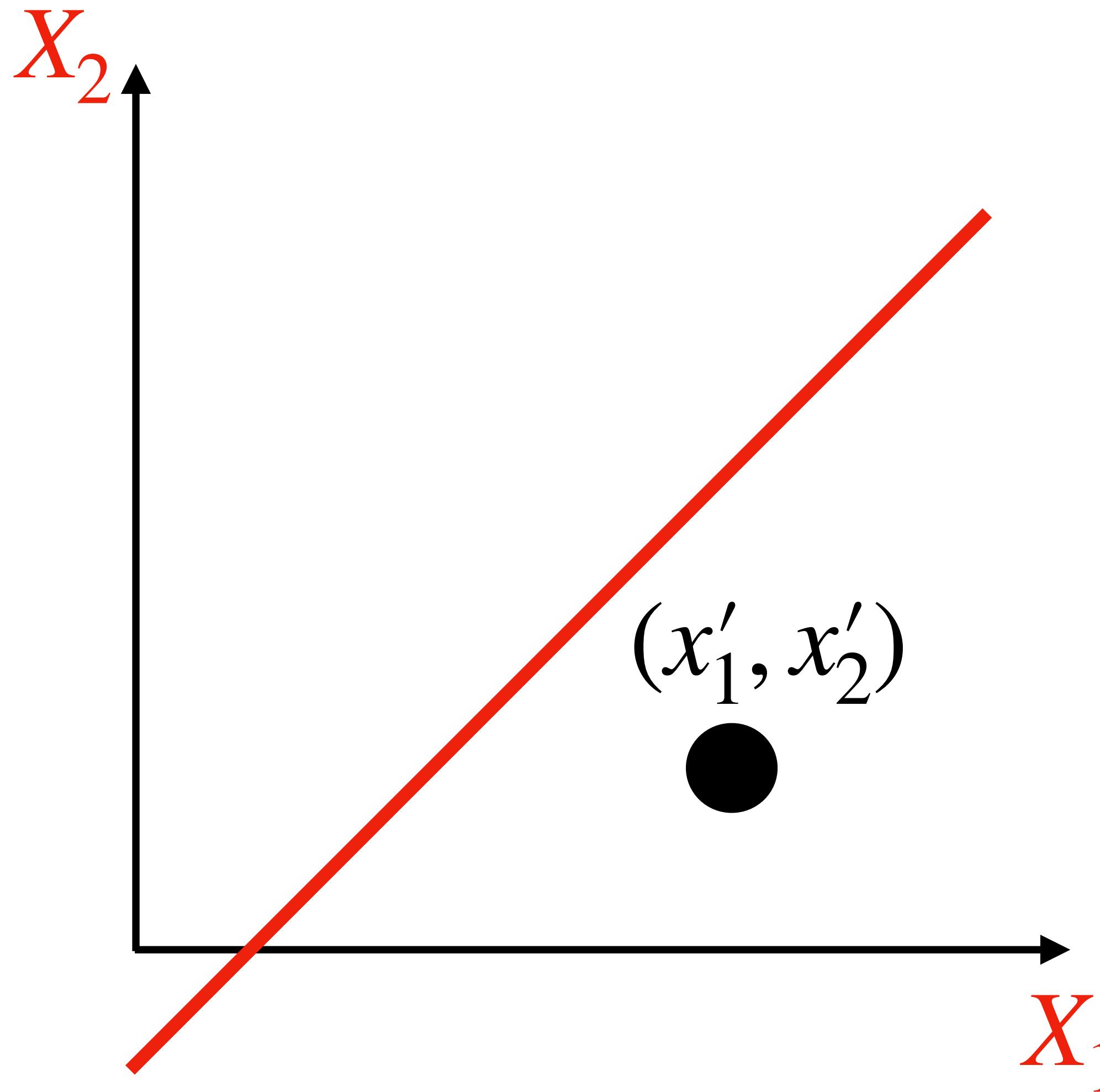
For a point (x'_1, x'_2) :

If $w_0 + w_1 x'_1 + w_2 x'_2 > 0$

then it is **above** the line.

SUPPORT VECTOR MACHINES (SVM)

Hyperplane: Start with a hyperplane in 2D, which is a line.



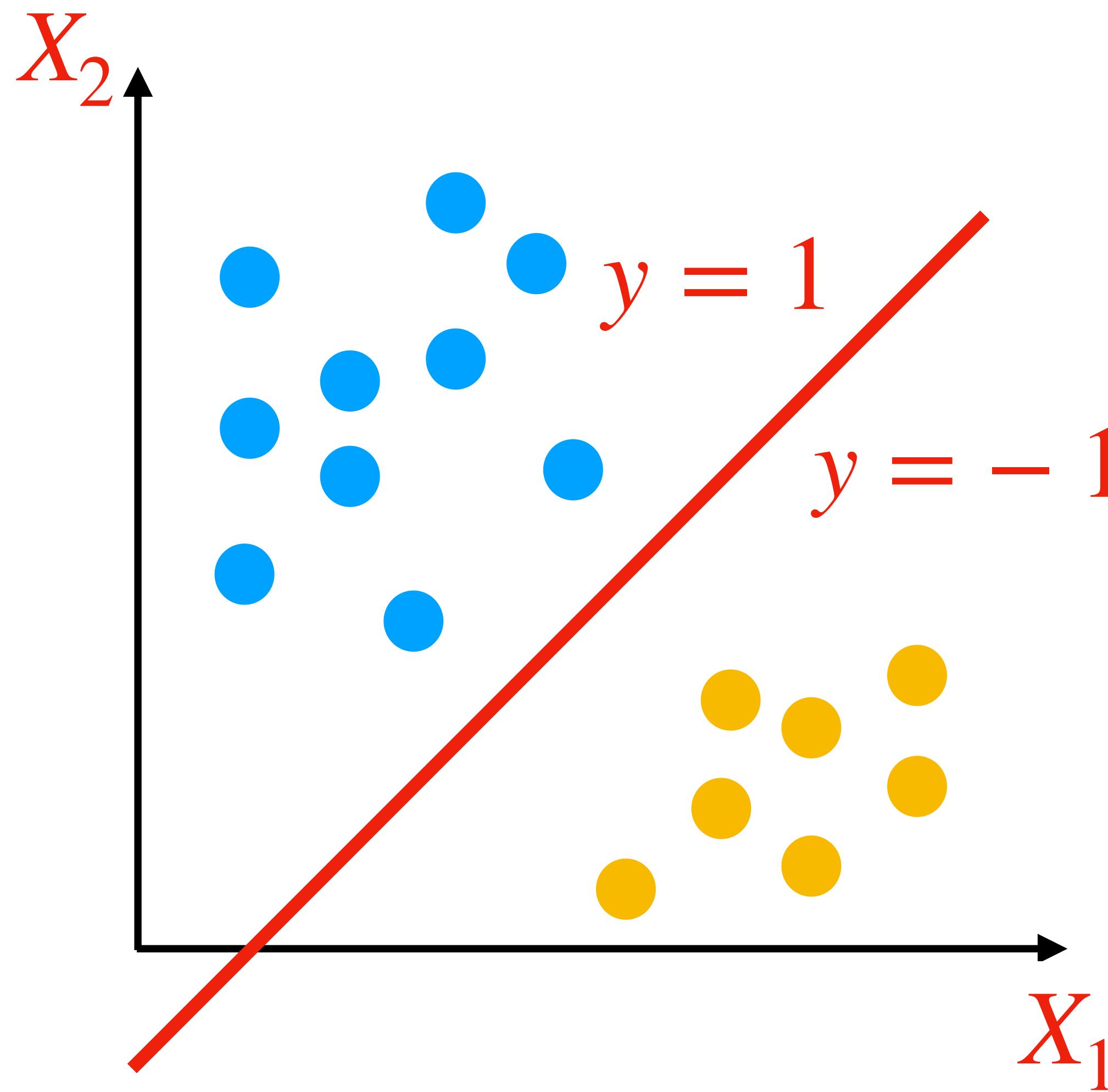
$$w_0 + w_1 X_1 + w_2 X_2 = 0$$

For a point (x'_1, x'_2) :

If $w_0 + w_1 x'_1 + w_2 x'_2 < 0$
then it is **below** the line.

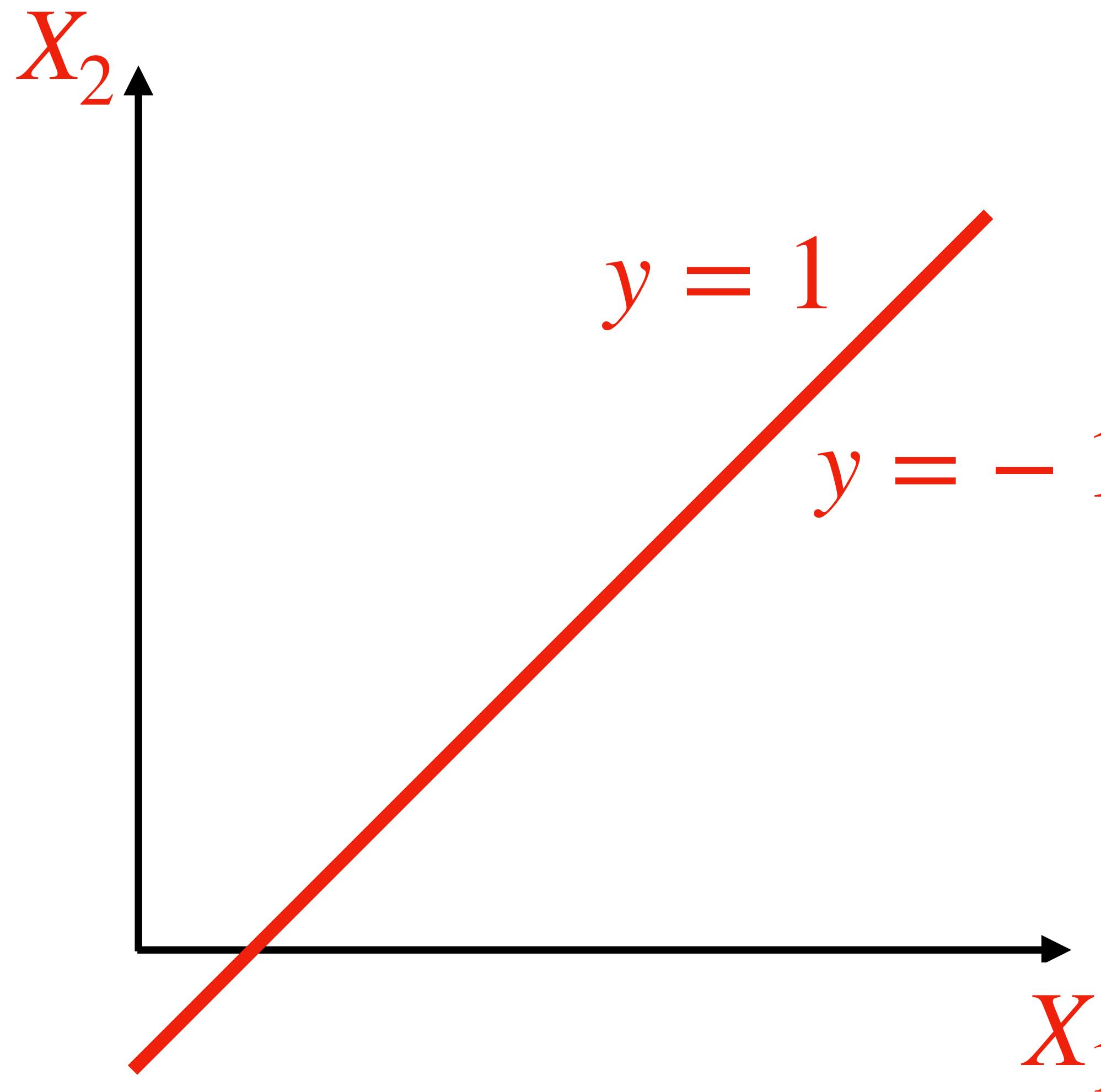
SUPPORT VECTOR MACHINES (SVM)

We want to find the hyperplane that
'separates' the classes 'well'.



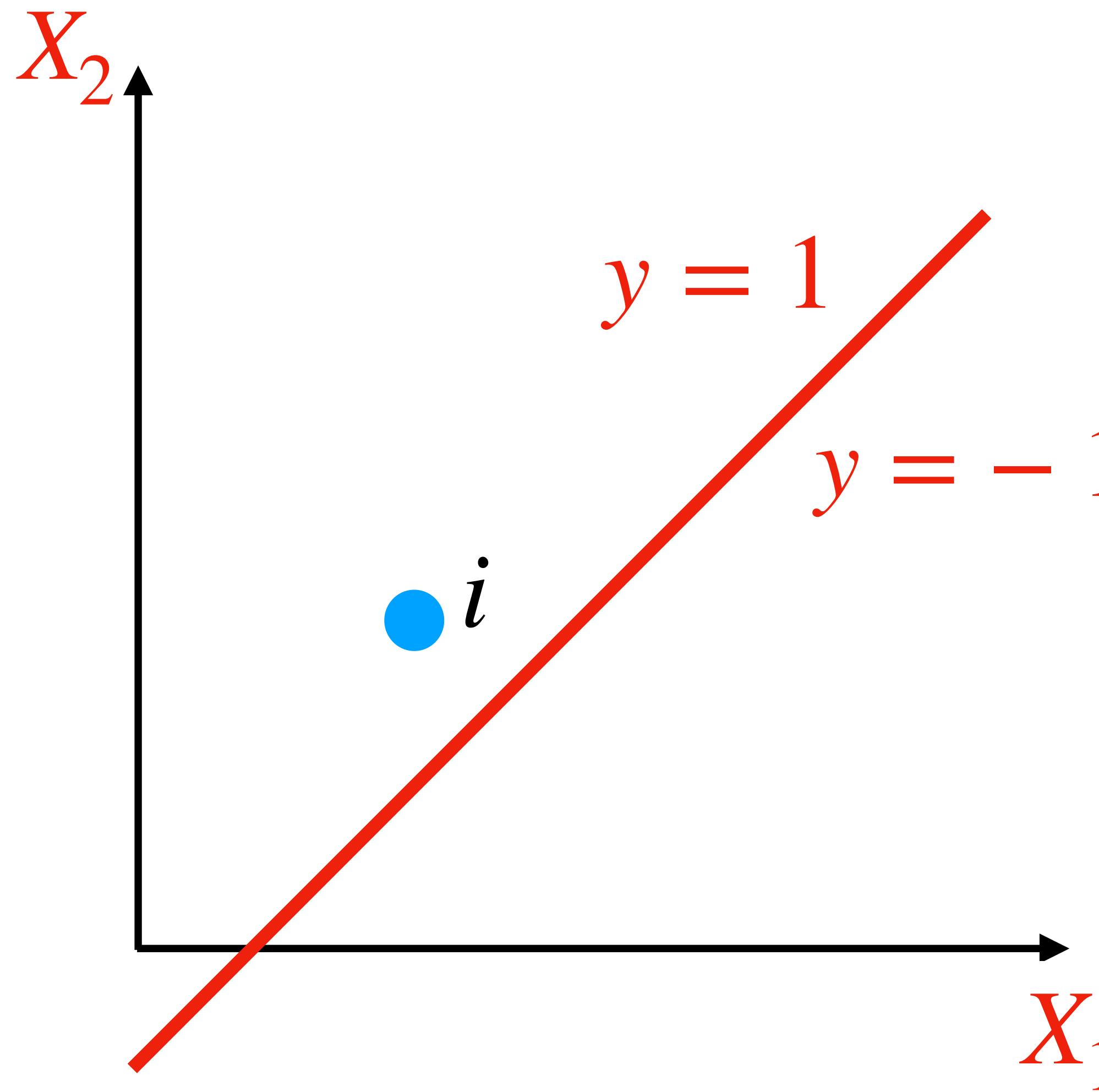
SUPPORT VECTOR MACHINES (SVM)

We want to find the hyperplane that
‘separates’ the classes ‘well’.



Meaning of ‘separates classes’:

SUPPORT VECTOR MACHINES (SVM)



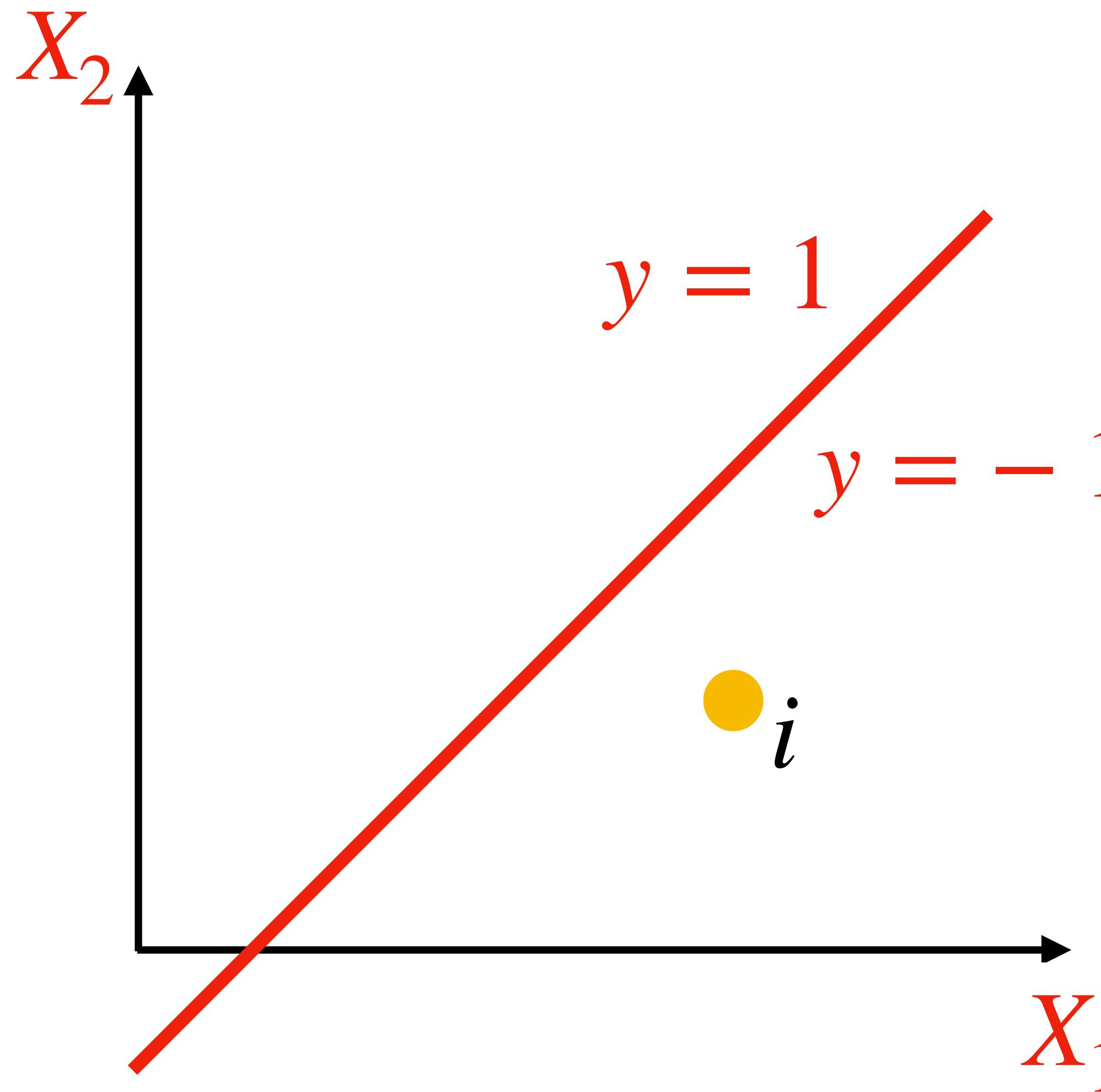
We want to find the hyperplane that
‘separates’ the classes ‘well’.

Meaning of ‘separates classes’:

For a training point i with $y_i = 1$:

$$w_0 + w_1x_{i1} + w_2x_{i2} > 0$$

SUPPORT VECTOR MACHINES (SVM)



We want to find the hyperplane that ‘separates’ the classes ‘well’.

Meaning of ‘separates classes’:

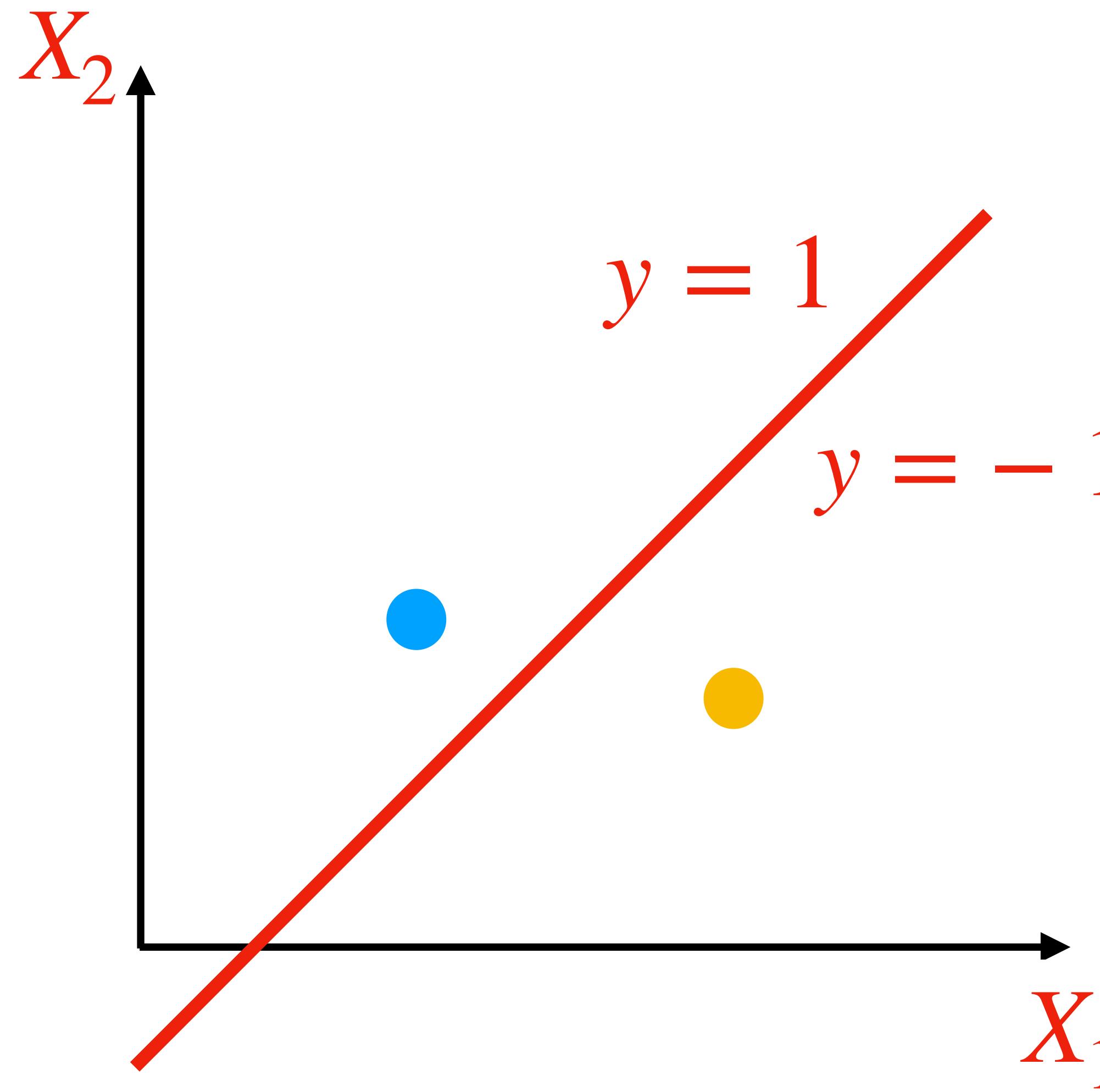
For a training point i with $y_i = 1$:

$$w_0 + w_1 x_{i1} + w_2 x_{i2} > 0$$

For a training point i with $y_i = -1$:

$$w_0 + w_1 x_{i1} + w_2 x_{i2} < 0$$

SUPPORT VECTOR MACHINES (SVM)



We want to find the hyperplane that 'separates' the classes 'well'.

Meaning of 'separates classes':

For a training point i with $y_i = 1$:

$$w_0 + w_1x_{i1} + w_2x_{i2} > 0$$

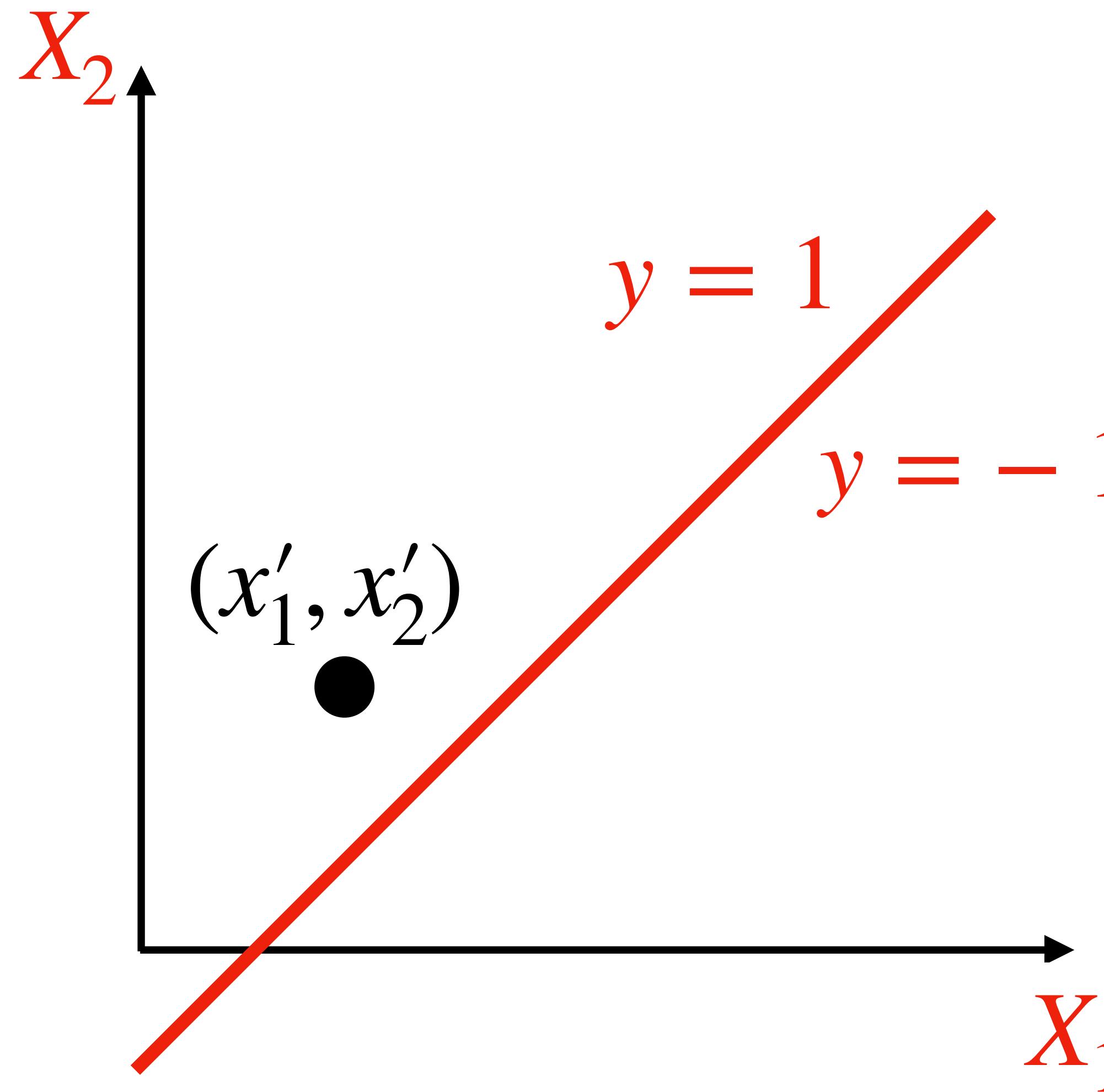
For a training point i with $y_i = -1$:

$$w_0 + w_1x_{i1} + w_2x_{i2} < 0$$

Can combine them in one inequality:

$$y_i(w_0 + w_1x_{i1} + w_2x_{i2}) > 0$$

SUPPORT VECTOR MACHINES (SVM)

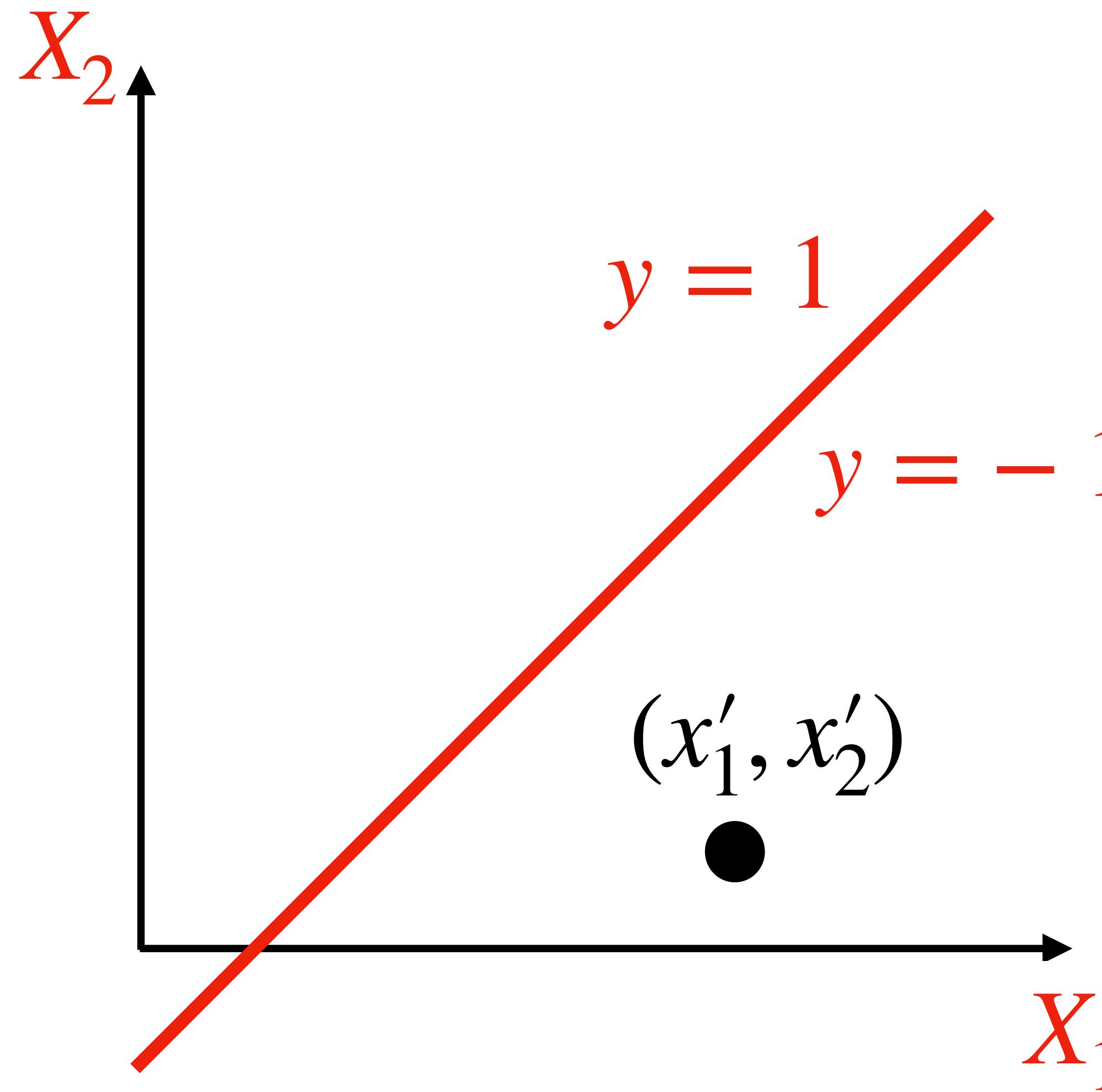


Given a test point (x'_1, x'_2) :

If $w_0 + w_1x'_1 + w_2x'_2 > 0$

Assign it to Class 1.

SUPPORT VECTOR MACHINES (SVM)



Given a test point (x'_1, x'_2) :

If $w_0 + w_1x'_1 + w_2x'_2 > 0$

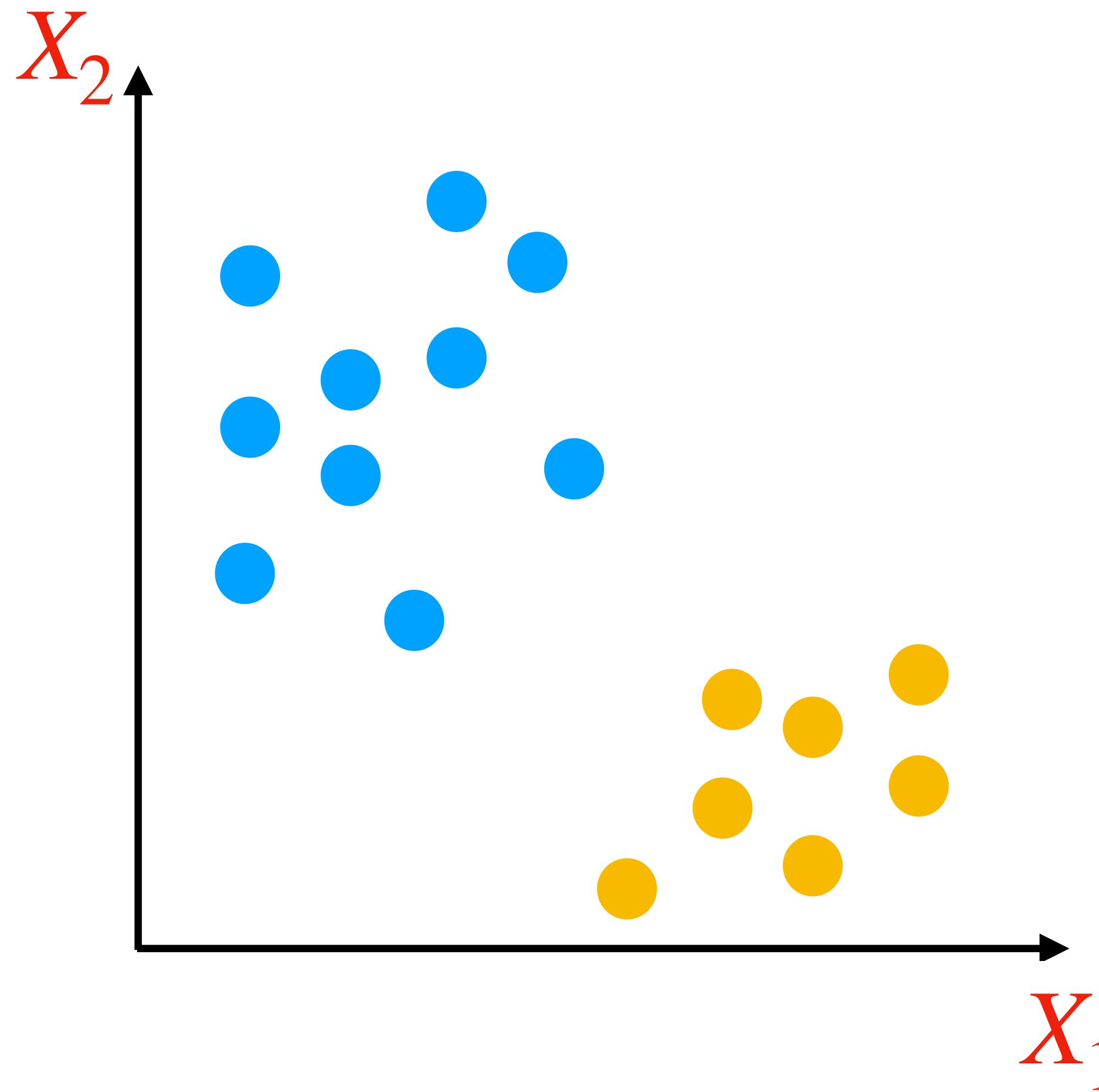
Assign it to Class 1.

If $w_0 + w_1x'_1 + w_2x'_2 < 0$

Assign it to Class -1.

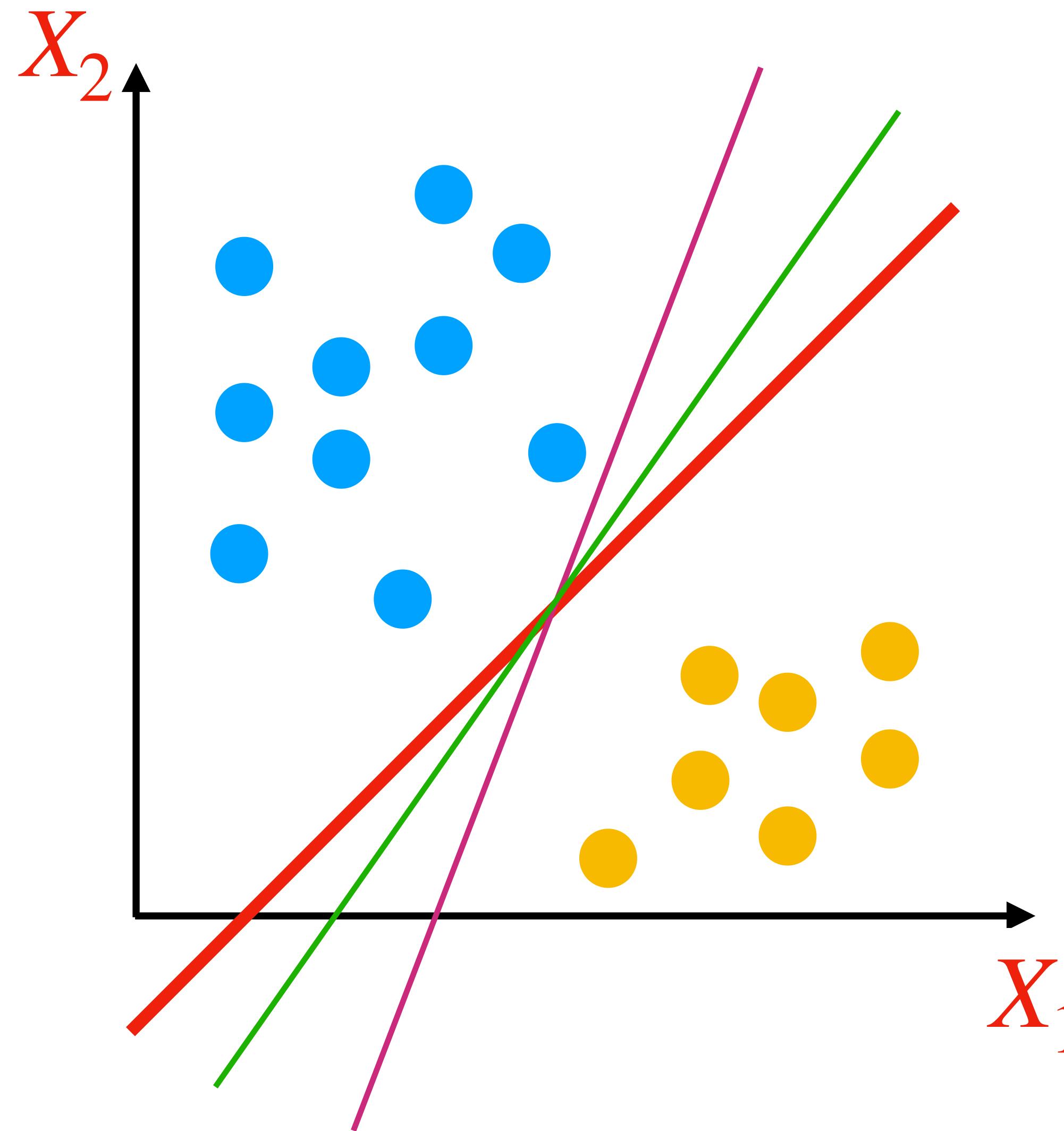
SUPPORT VECTOR MACHINES (SVM)

We want to find the hyperplane that
'separates' the classes 'well'.



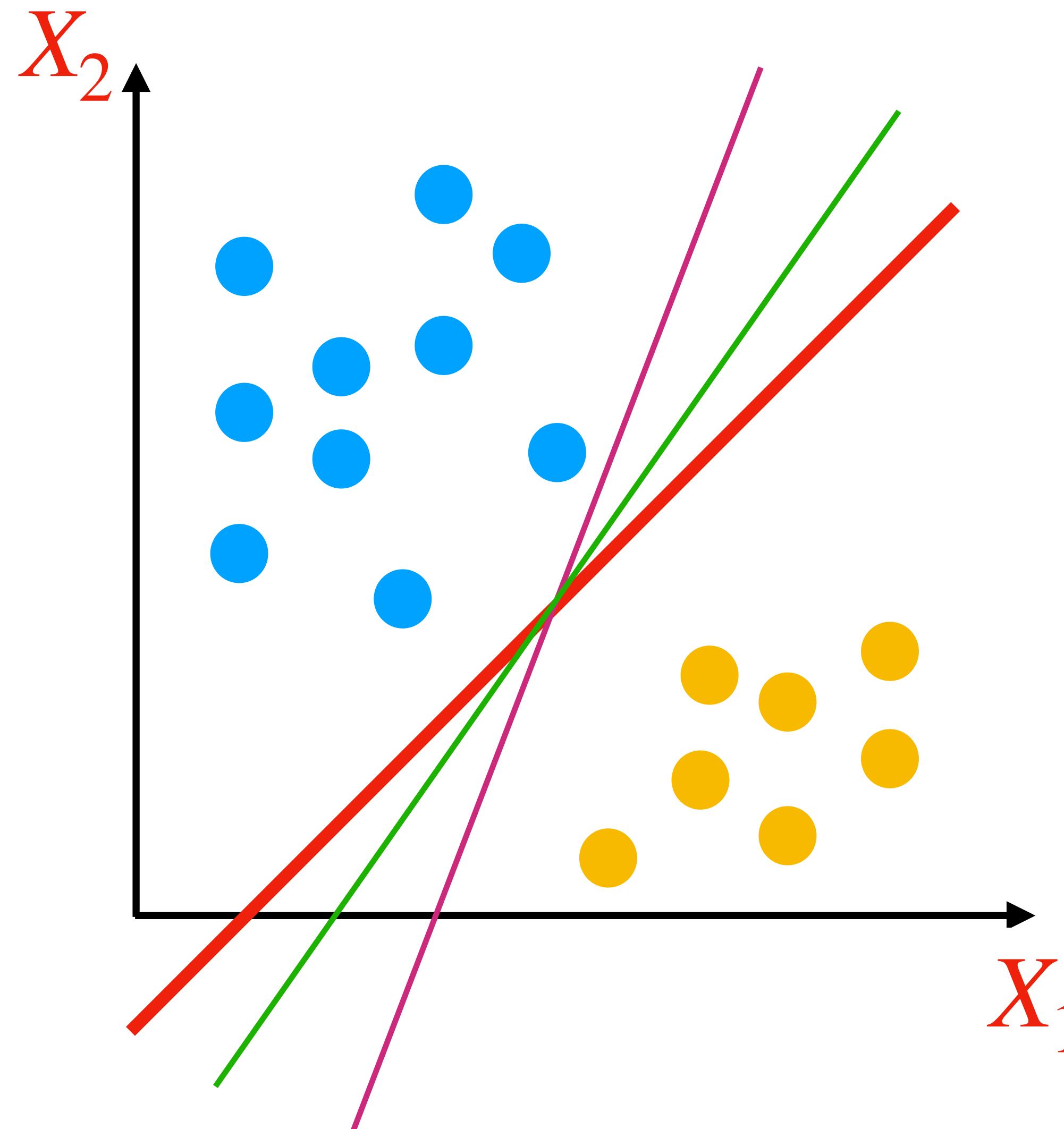
How about 'separates classes well':

SUPPORT VECTOR MACHINES (SVM)



How about ‘separates classes well’:
Among red, green, purple:
red line separates the classes best.

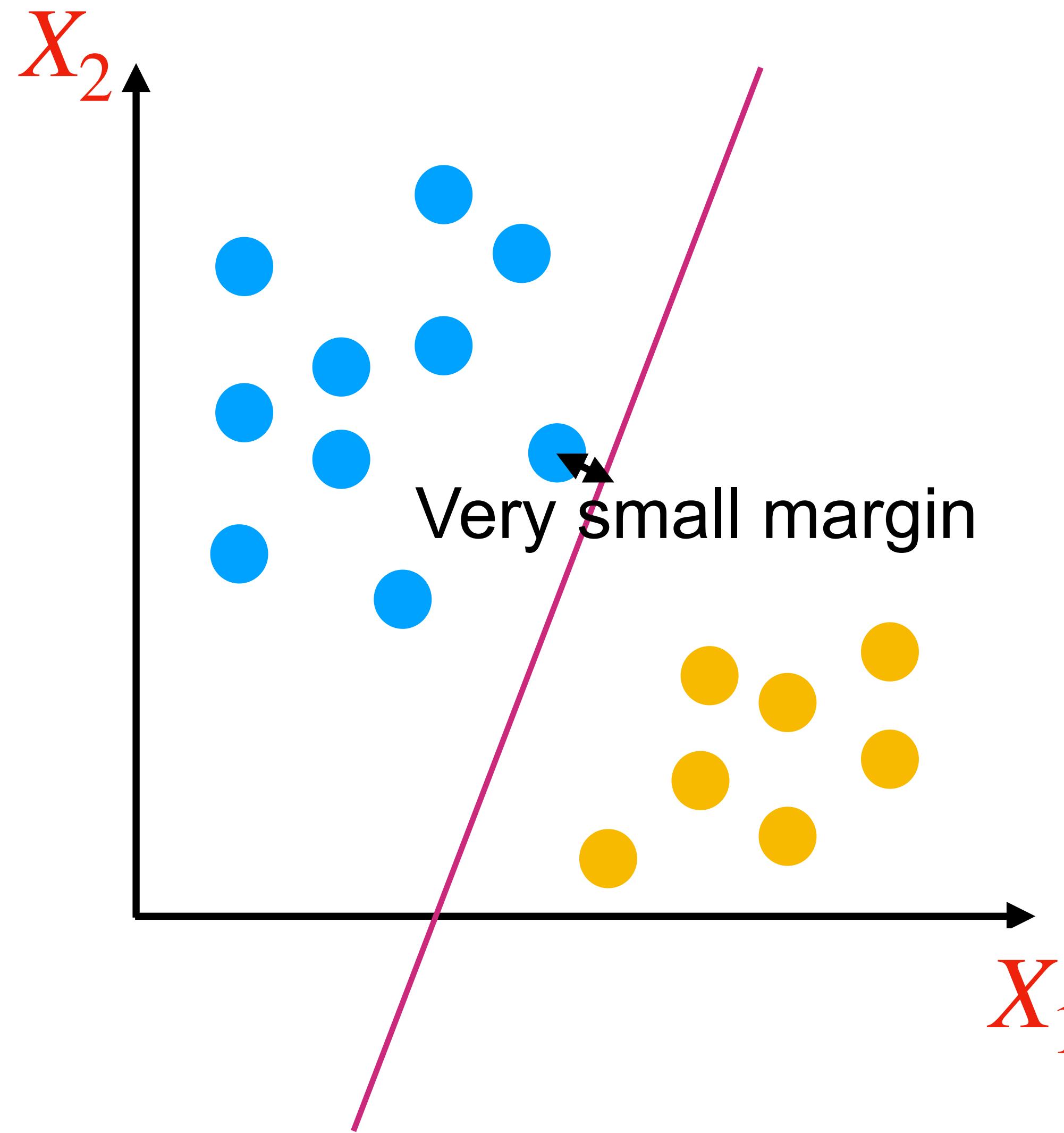
SUPPORT VECTOR MACHINES (SVM)



How about ‘separates classes well’: Among red, green, purple:
red line separates the classes best.

Margin: minimum distance from any training point to the hyperplane.

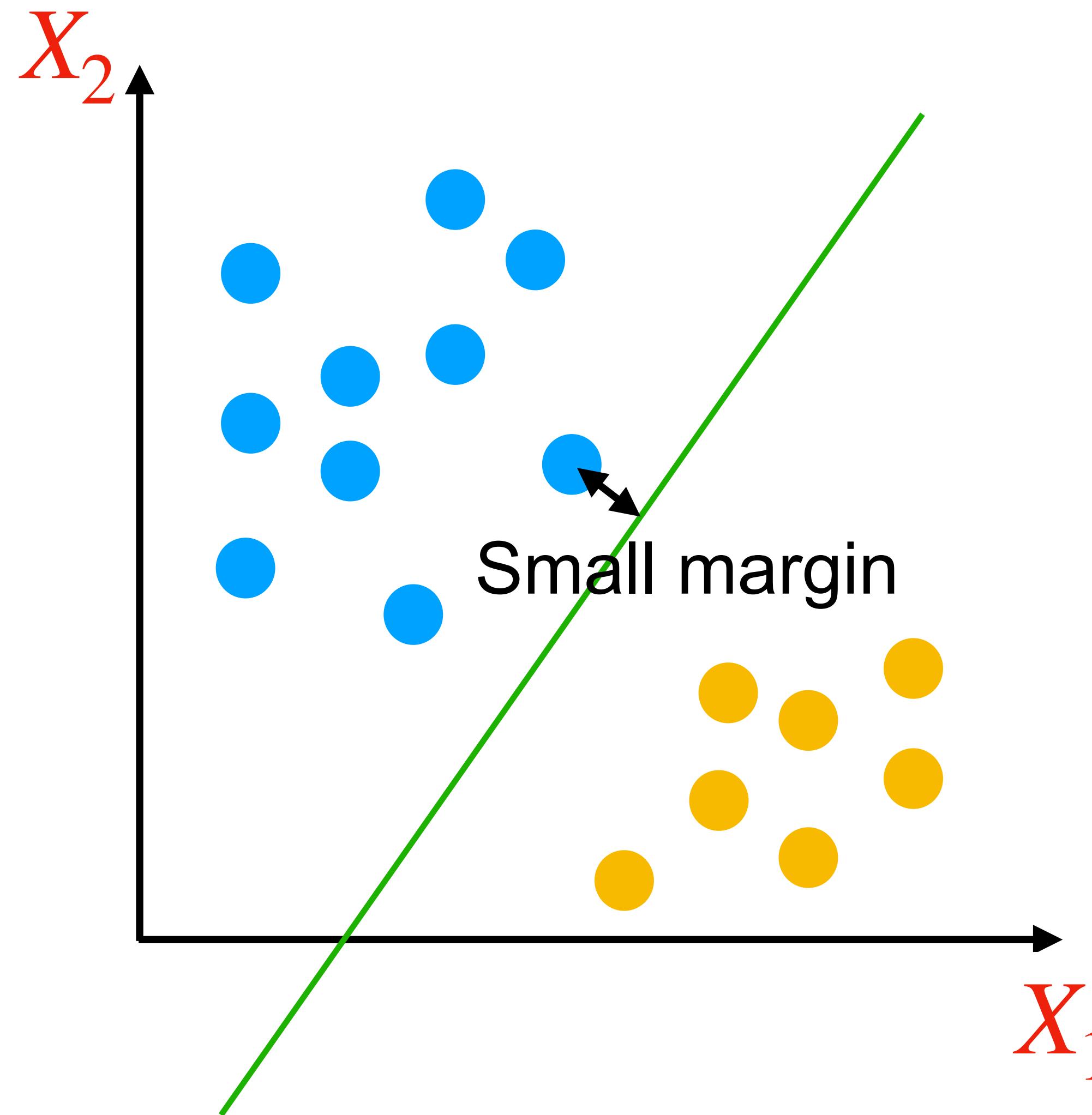
SUPPORT VECTOR MACHINES (SVM)



How about ‘separates classes well’: Among red, green, purple:
red line separates the classes best.

Margin: minimum distance from any training point to the hyperplane.

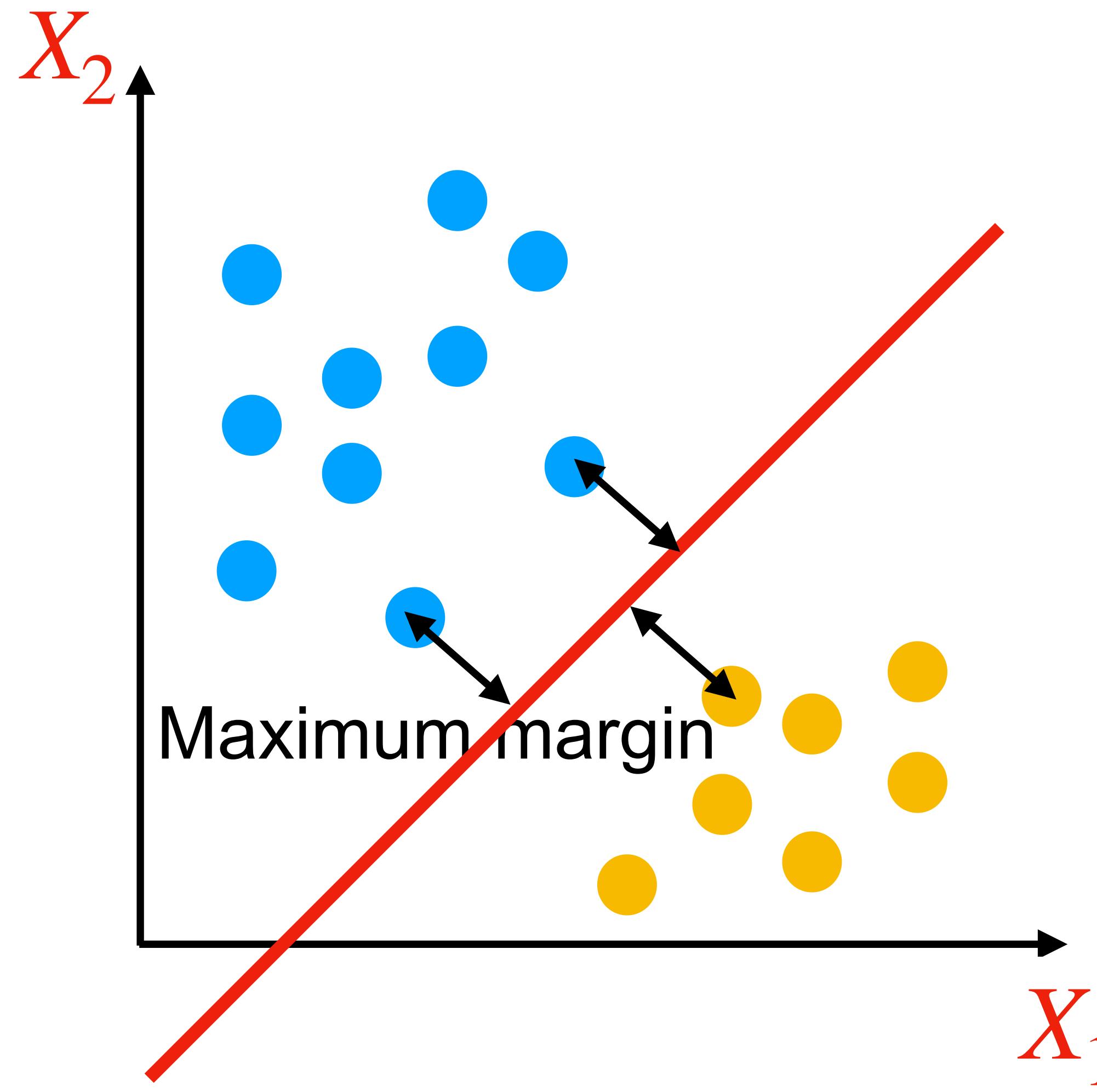
SUPPORT VECTOR MACHINES (SVM)



How about ‘separates classes well’: Among red, green, purple:
red line separates the classes best.

Margin: minimum distance from any training point to the hyperplane.

SUPPORT VECTOR MACHINES (SVM)

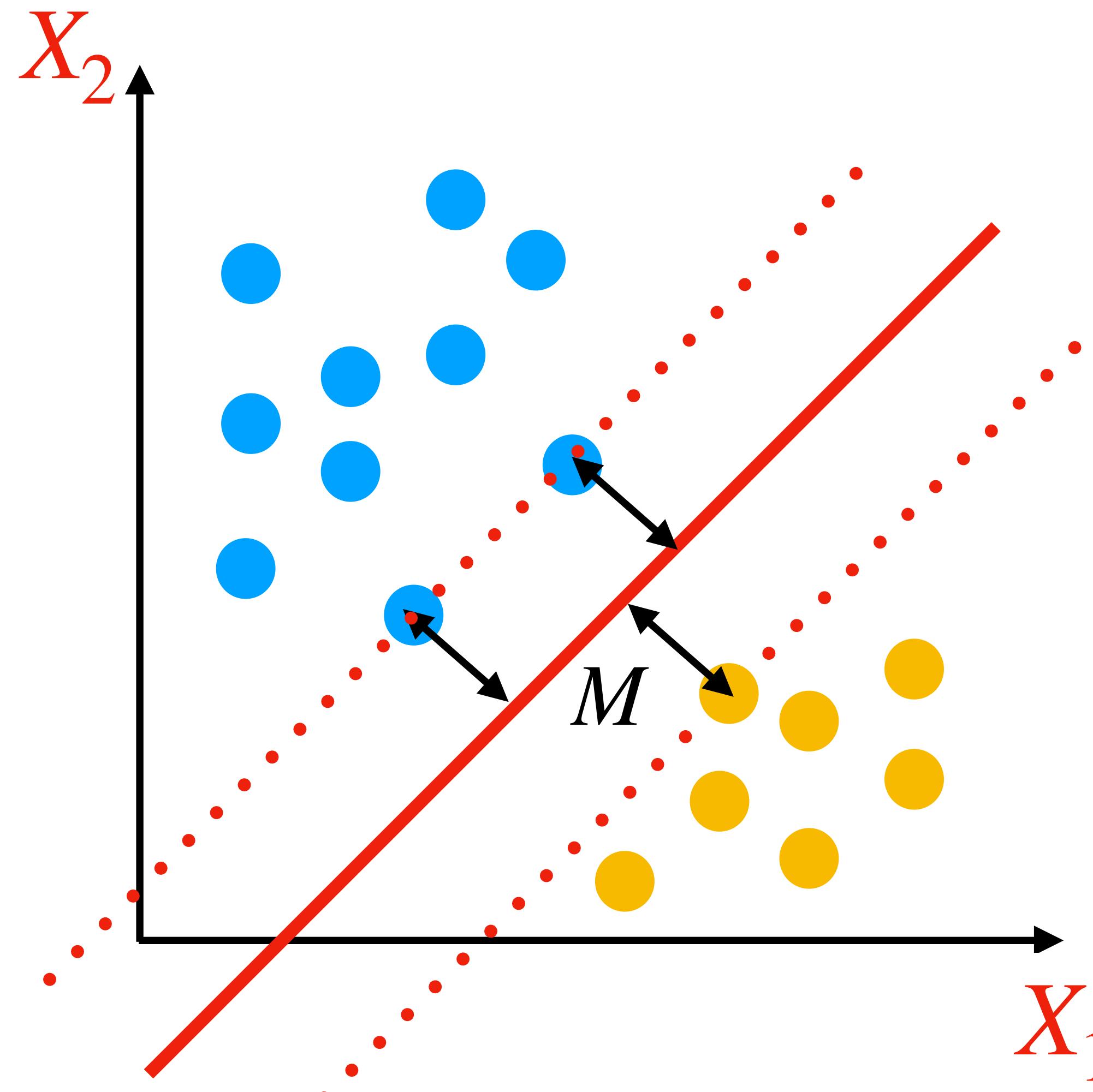


How about ‘separates classes well’: Among red, green, purple:
red line separates the classes best.

Margin: minimum distance from any training point to the hyperplane.

Red line has the maximum margin.

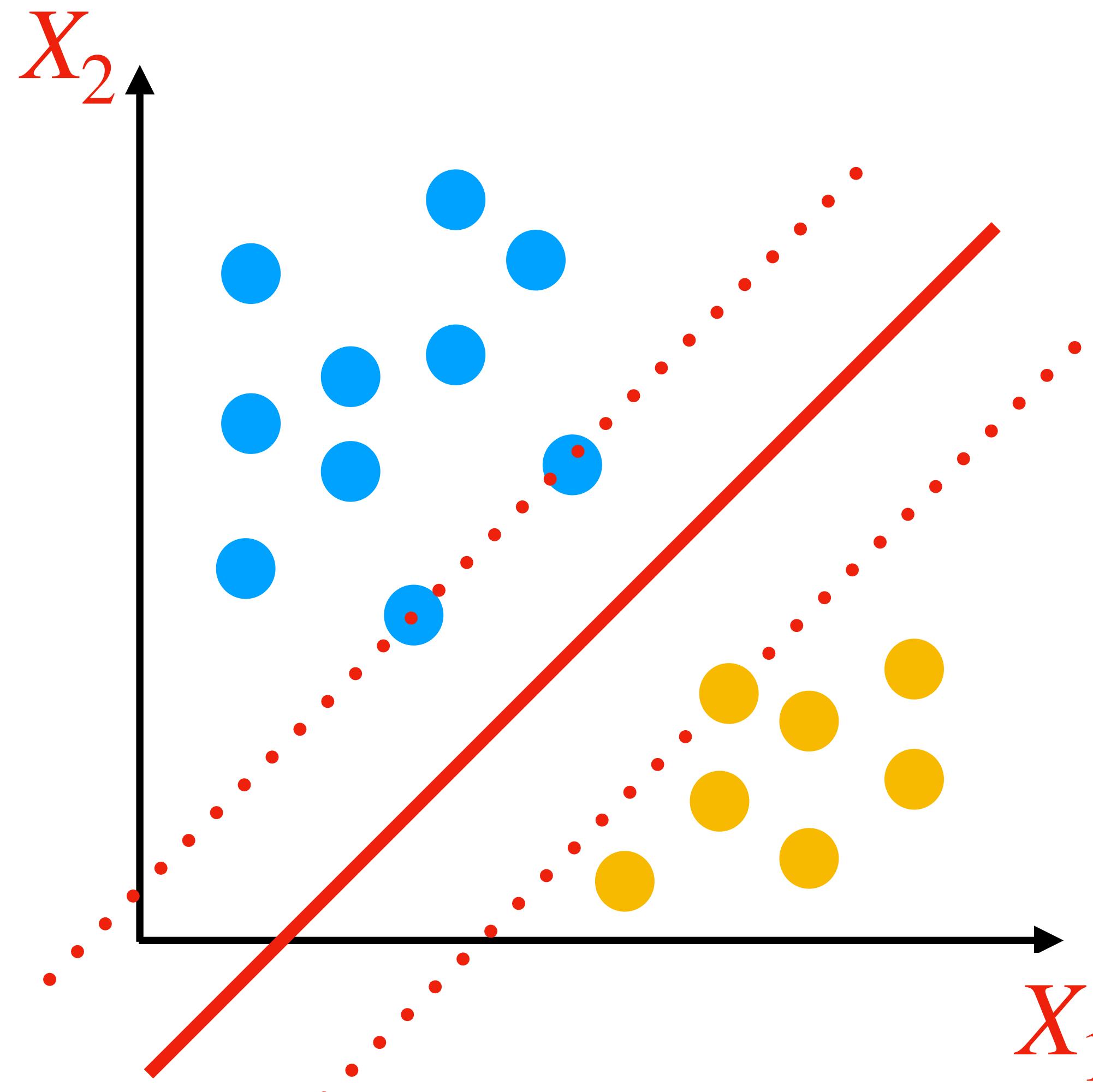
SUPPORT VECTOR MACHINES (SVM)



The points with maximum margin M distance from separating hyperplane are called the **support vectors**.

In this example we have 3 of them.

SUPPORT VECTOR MACHINES (SVM)

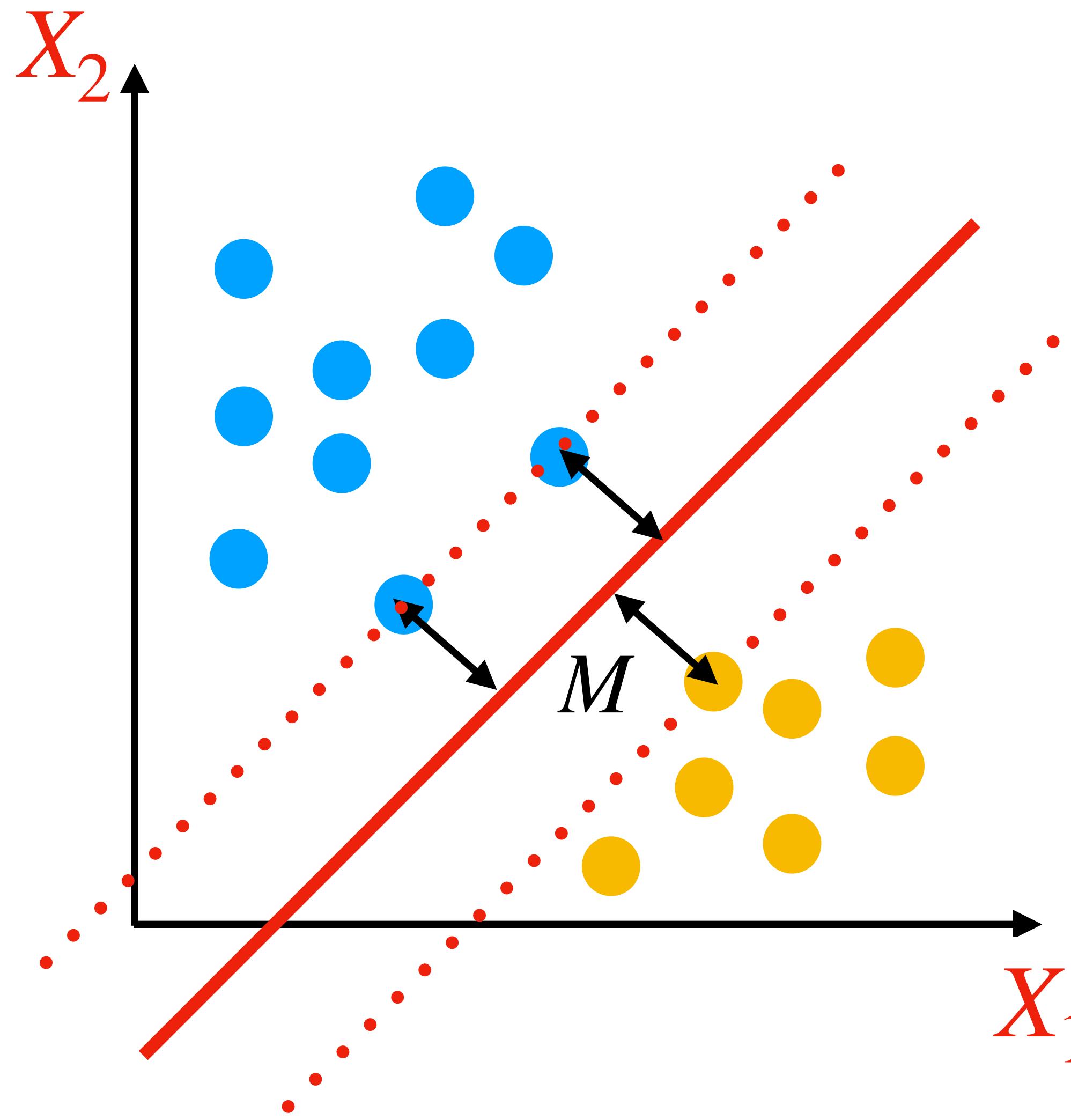


The points with maximum margin M distance from separating hyperplane are called the **support vectors**.

In this example we have 3 of them.

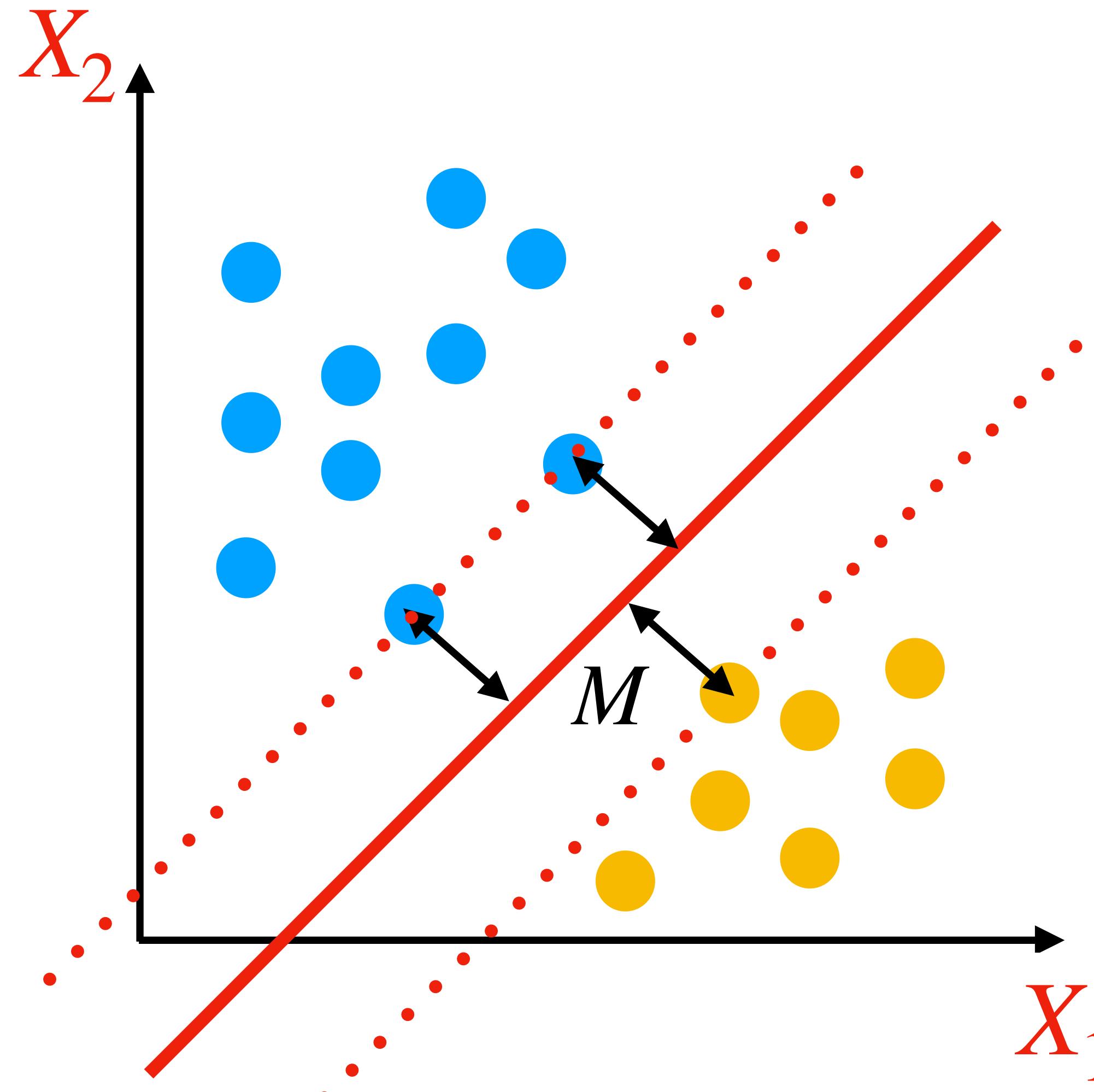
No point lies inside the margin.

SUPPORT VECTOR MACHINES (SVM)



Informally the goal is to:
Find hyperplane that provides
maximum margin M .

SUPPORT VECTOR MACHINES (SVM)



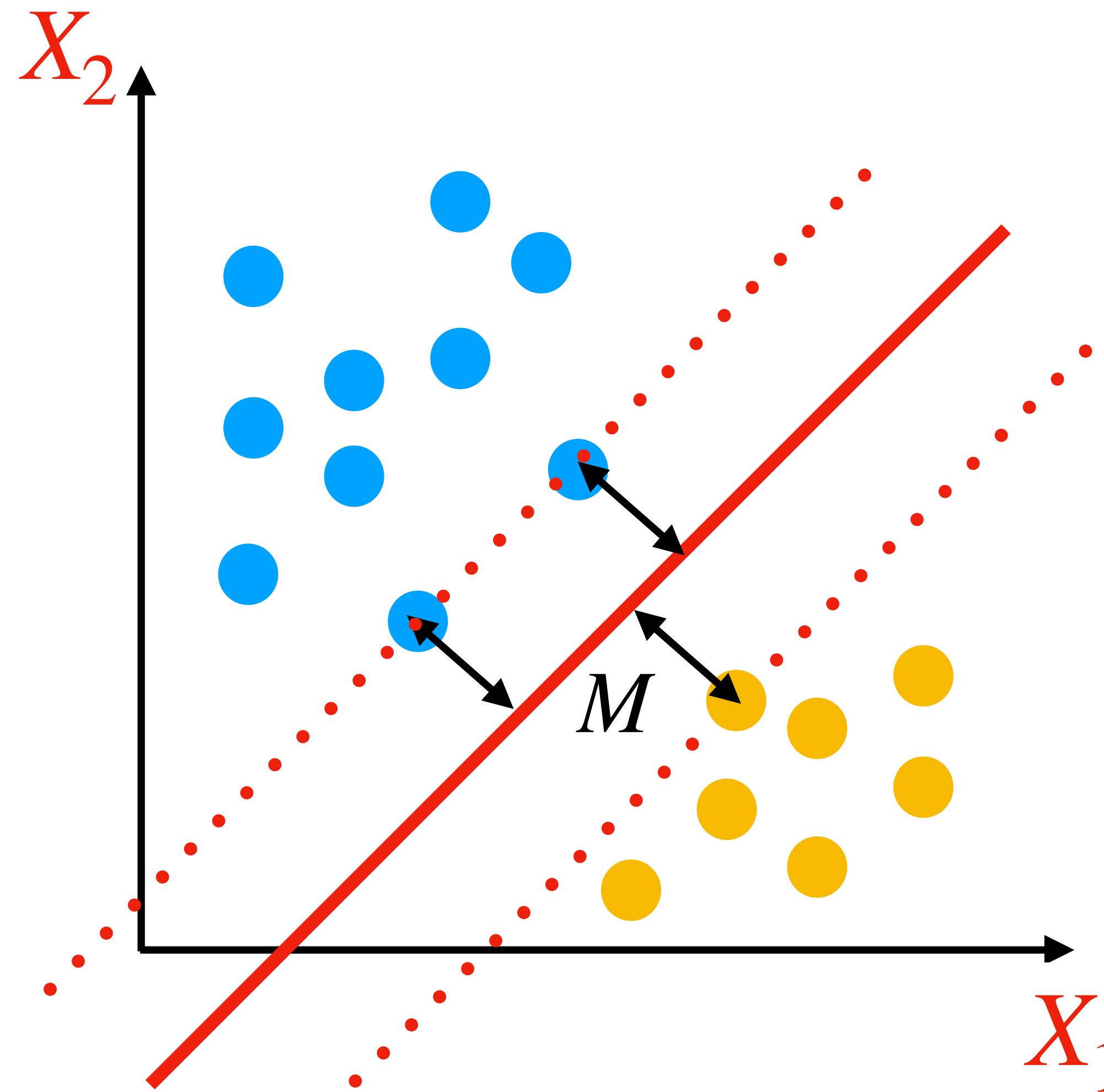
Formally the goal is to:
Find w_0, w_1, w_2 that
maximizes M
with the constraints

$$\mathcal{C}_1 \quad w_1^2 + w_2^2 = 1$$

$$\mathcal{C}_2 \quad y_i(w_0 + w_1x_{i1} + w_2x_{i2}) \geq M$$

for any training point i

SUPPORT VECTOR MACHINES (SVM)



Formally the goal is to:
Find w_0, w_1, w_2 that
maximizes M
with the constraints

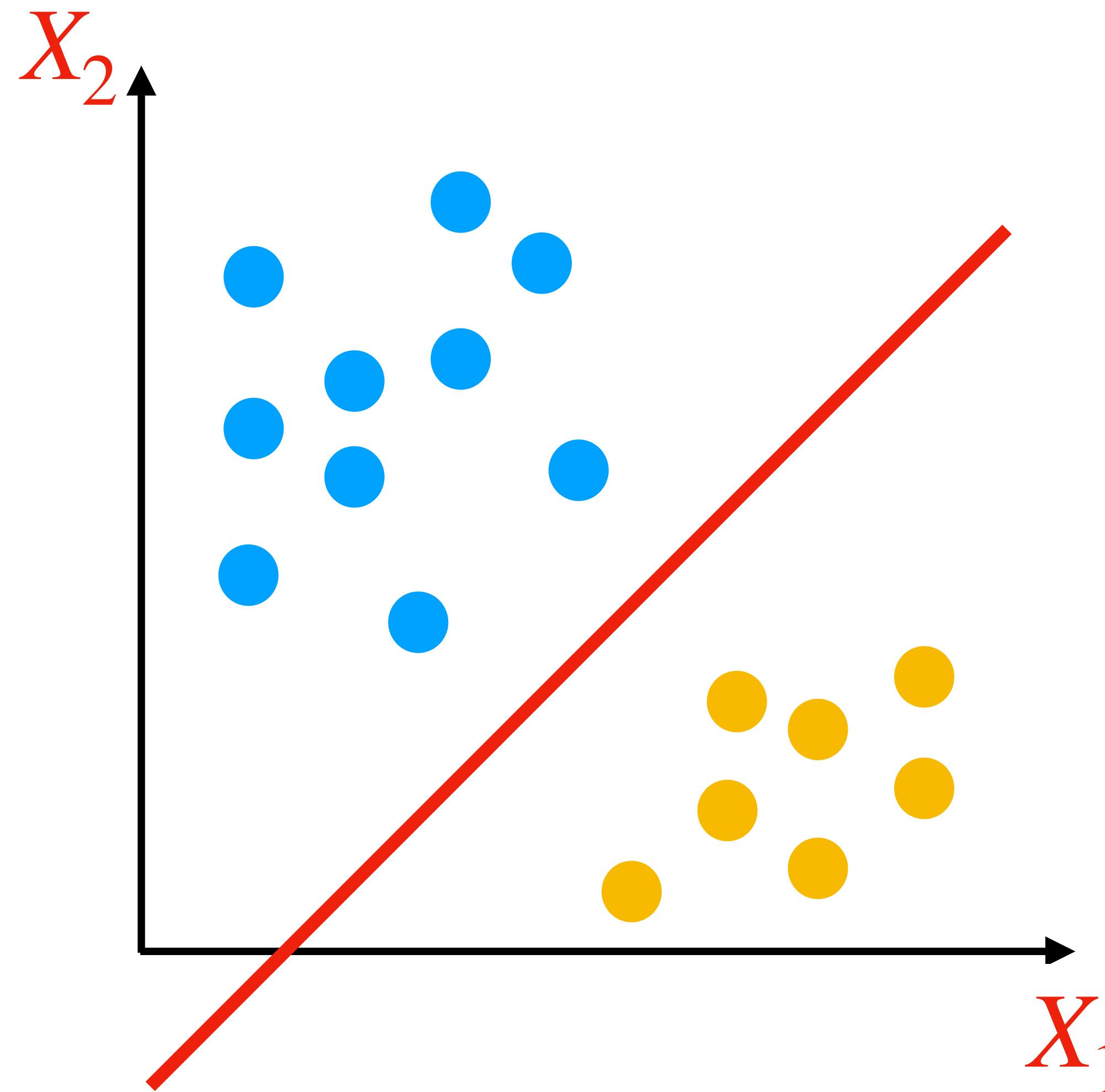
$$\mathcal{C}_1 \quad w_1^2 + w_2^2 = 1$$

$$\mathcal{C}_2 \quad y_i(w_0 + w_1x_{i1} + w_2x_{i2}) \geq M$$

for any training point i

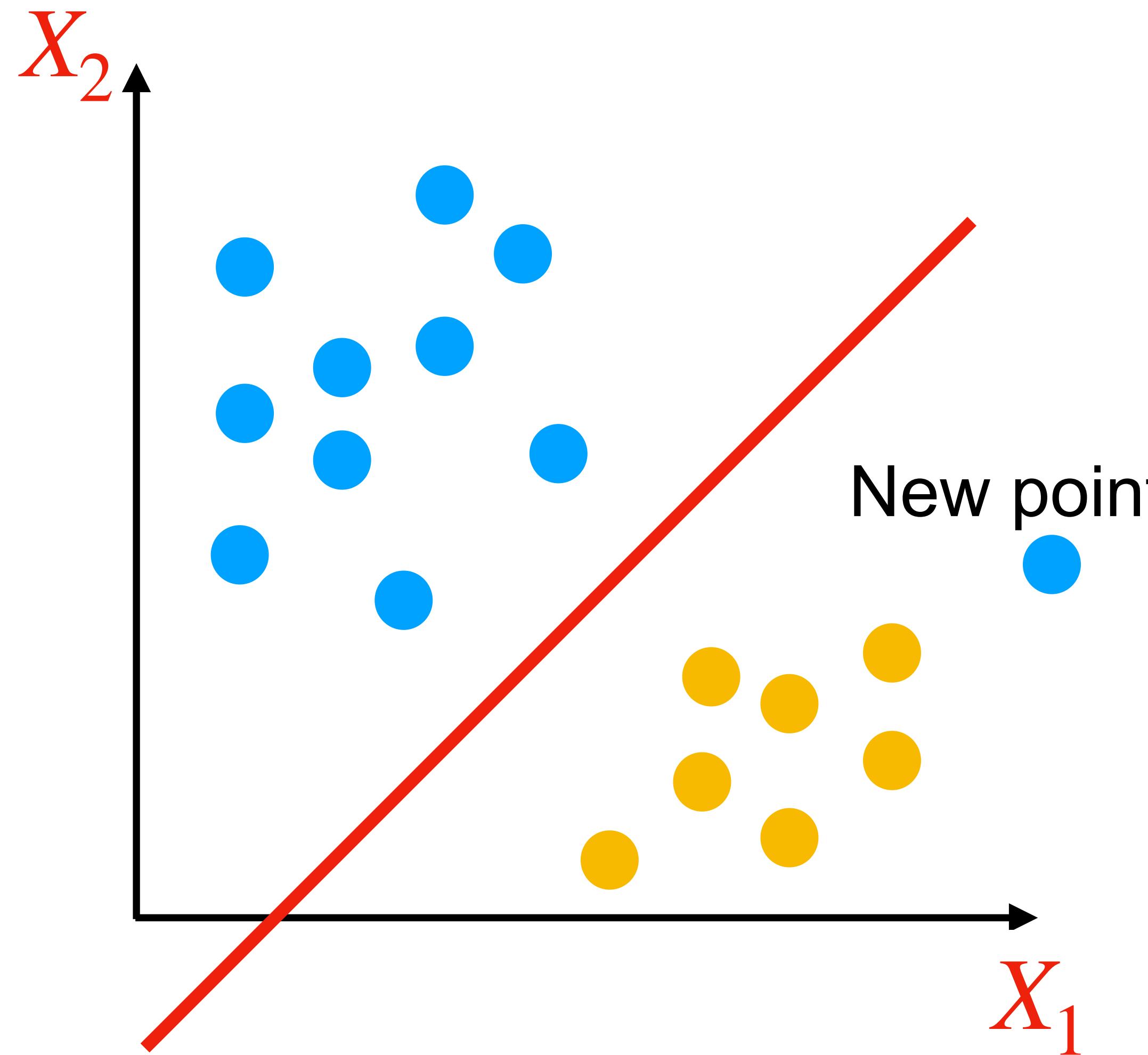
Idea: Due to \mathcal{C}_1 , $y_i(w_0 + w_1x_{i1} + w_2x_{i2})$ is
distance of point i to hyperplane.

SUPPORT VECTOR MACHINES (SVM)



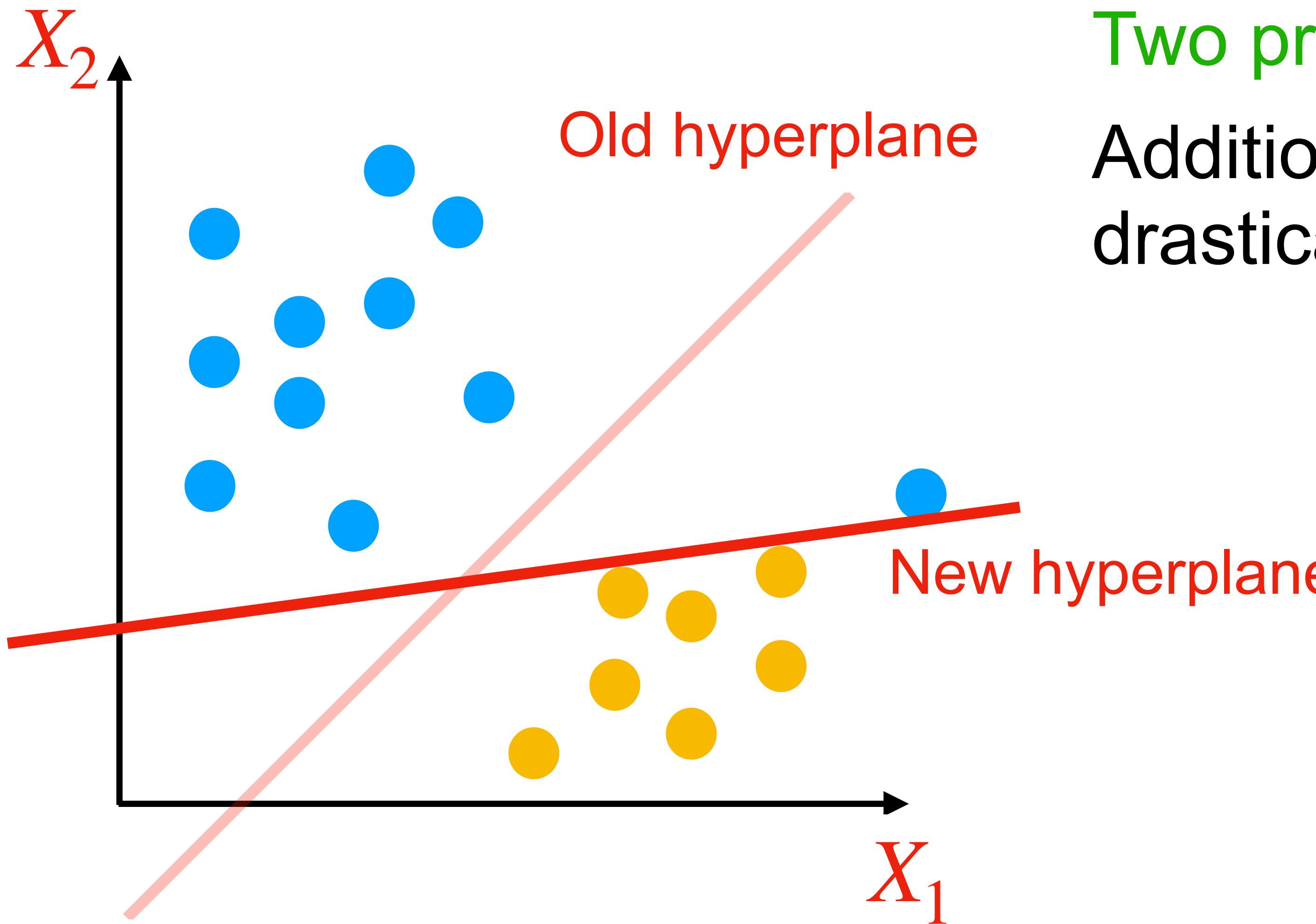
Two problems with the same solution:
Addition of a single point might
drastically change the hyperplane.

SUPPORT VECTOR MACHINES (SVM)



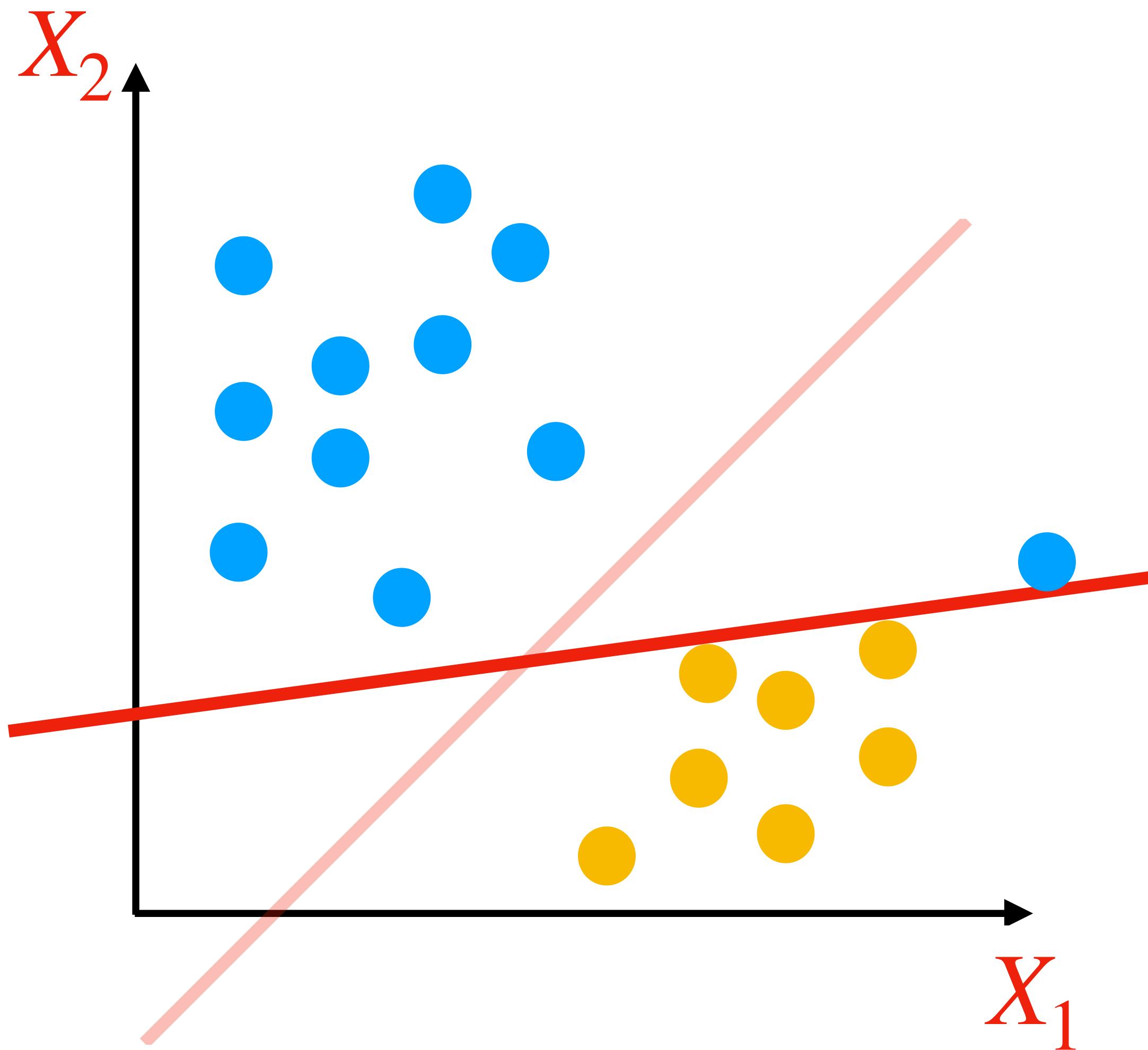
Two problems with the same solution:
Addition of a single point might
drastically change the hyperplane.

SUPPORT VECTOR MACHINES (SVM)



Two problems with the same solution:
Addition of a single point might
drastically change the hyperplane.

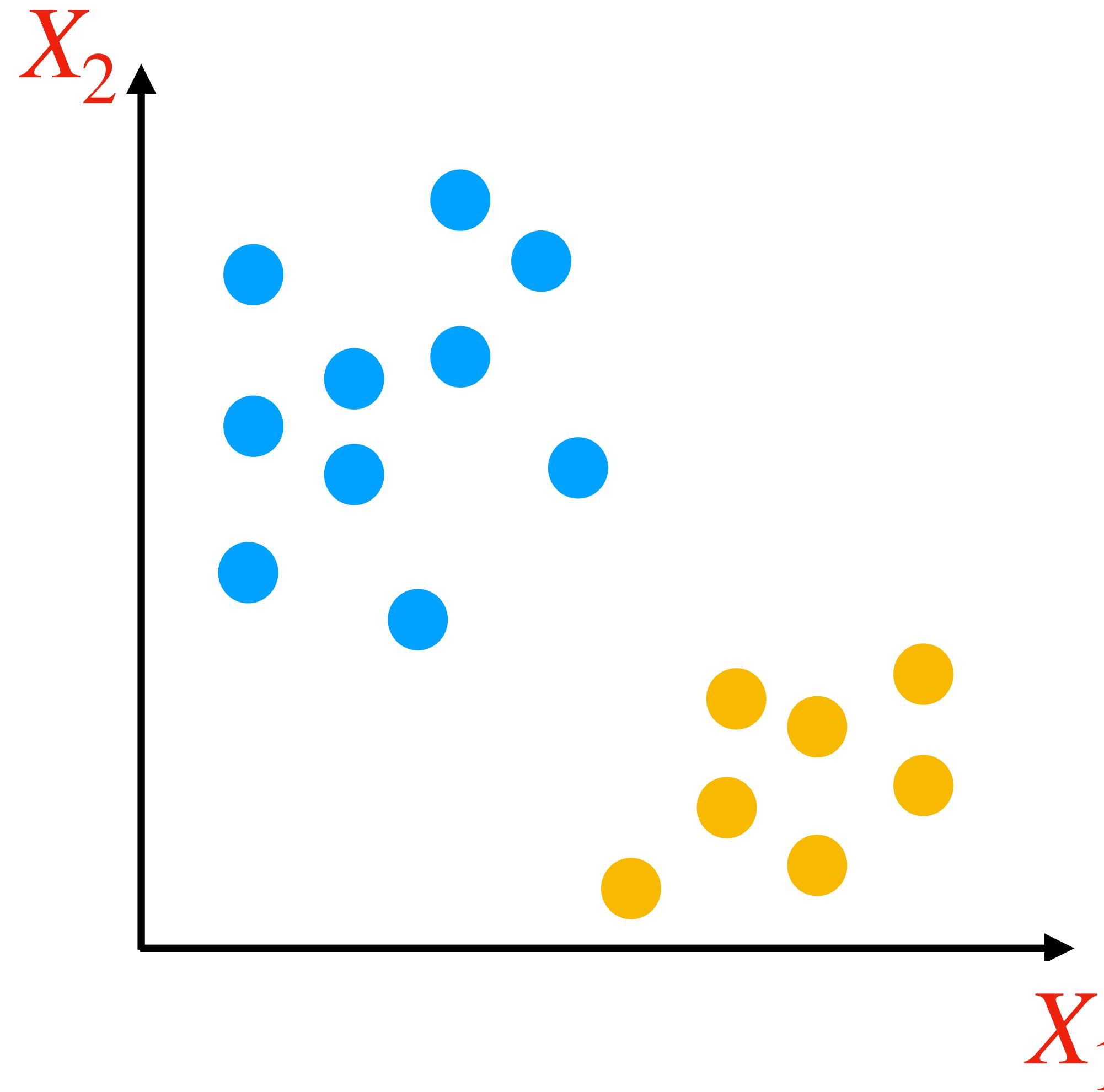
SUPPORT VECTOR MACHINES (SVM)



Two problems with the same solution:
Addition of a single point might
drastically change the hyperplane.

Insisting on perfect separation may
cause overfitting!

SUPPORT VECTOR MACHINES (SVM)



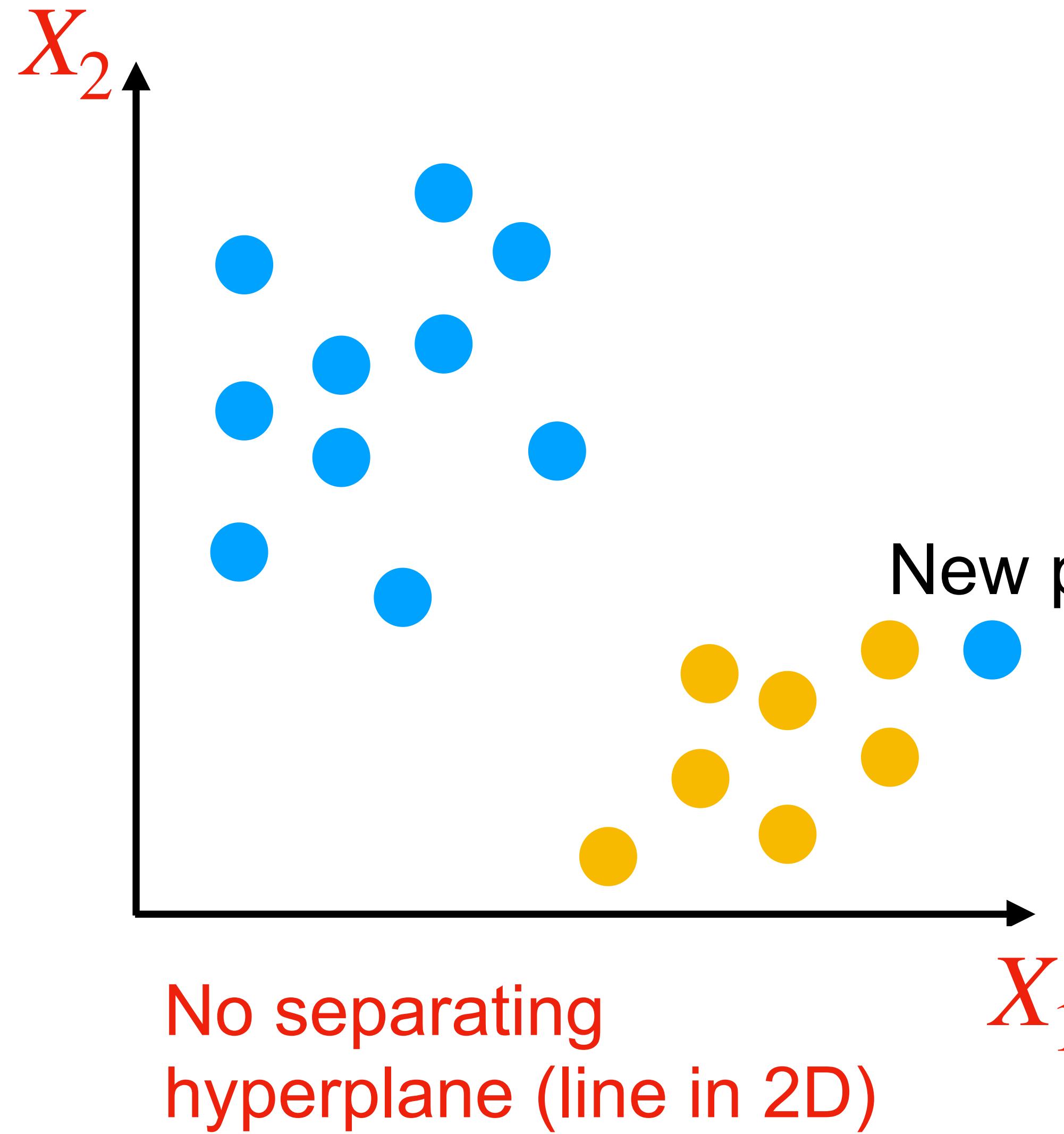
Two problems with the same solution:

Addition of a single point might
drastically change the hyperplane.

Insisting on perfect separation may
cause overfitting!

It may not even be possible to find a
separating hyperplane (line in 2D).

SUPPORT VECTOR MACHINES (SVM)



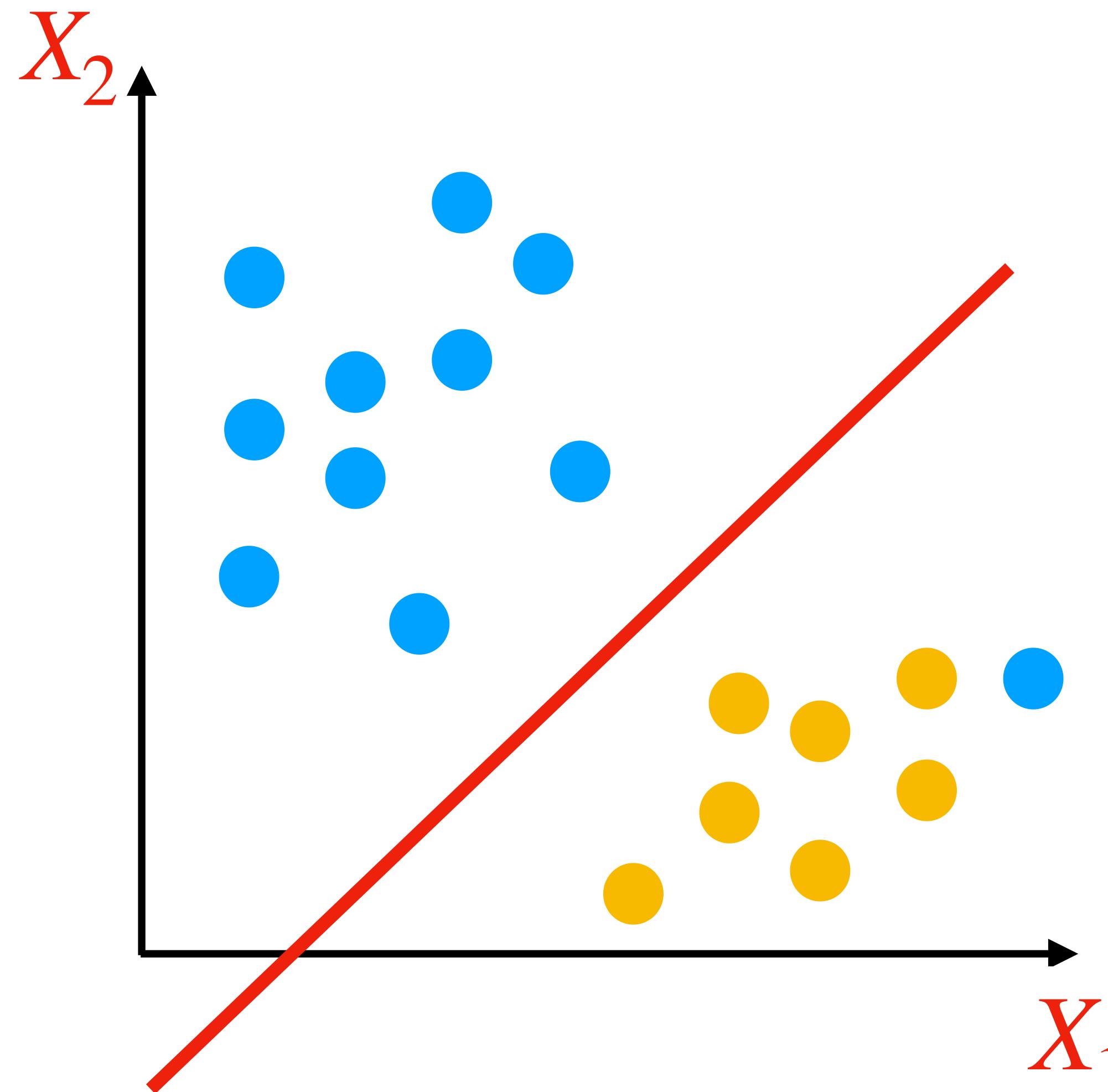
Two problems with the same solution:

Addition of a single point might drastically change the hyperplane.

Insisting on perfect separation may cause overfitting!

It may not even be possible to find a separating hyperplane (line in 2D).

SUPPORT VECTOR MACHINES (SVM)



Two problems with the same solution:

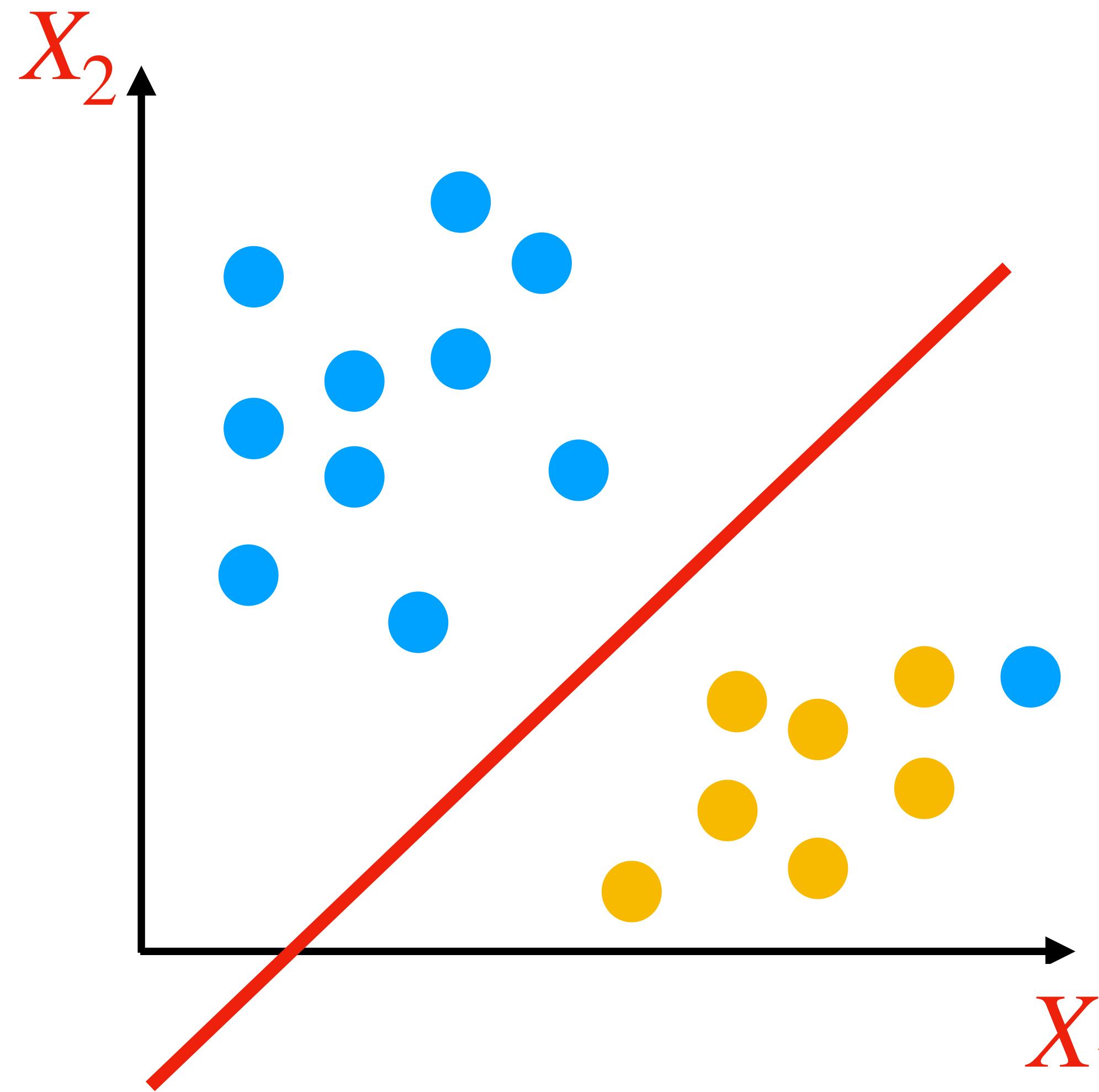
Addition of a single point might
drastically change the hyperplane.

Insisting on perfect separation may
cause overfitting!

It may not even be possible to find a
separating hyperplane (line in 2D).

Solution is to use Soft Margin:
Add flexibility; some observations can
be on the wrong side.

SUPPORT VECTOR MACHINES (SVM)



Soft Margin:

Find w_0, w_1, w_2 that
maximizes M
with the constraints

$$\mathcal{C}_1 \quad w_1^2 + w_2^2 = 1$$

$$\mathcal{C}_2 \quad y_i(w_0 + w_1x_{i1} + w_2x_{i2}) \geq M(1 - \epsilon_i) \quad \text{for any training point } i$$

$$\mathcal{C}_3 \quad \epsilon_i \geq 0, \sum_{i=1}^n \epsilon_i \leq C$$

SUPPORT VECTOR MACHINES (SVM)

What does all this mean?

As before due to \mathcal{C}_1

$y_i(w_0 + w_1x_{i1} + w_2x_{i2})$ is

distance of point i to hyperplane

$$\mathcal{C}_1 \quad w_1^2 + w_2^2 = 1$$

SUPPORT VECTOR MACHINES (SVM)

What does all this mean?

Don't insist distance be at least M for all training points.

Each training point i has some slack ϵ_i .

$\epsilon_i > 0$ allows point i to be inside the margin.

$\epsilon_i > 1$ allows point i to be on the wrong side of the hyperplane.

\mathcal{C}_1

$$w_1^2 + w_2^2 = 1$$

\mathcal{C}_2

$$y_i(w_0 + w_1x_{i1} + w_2x_{i2}) \geq M(1 - \epsilon_i)$$

for any training point i

SUPPORT VECTOR MACHINES (SVM)

What does all this mean?

C : total budget for all the slacks.
(hyperparameter tuned via cross-validation)

Controls bias-variance tradeoff:

Small $C \Rightarrow$ narrow margins that should not be violated (low bias, high variance).

Large $C \Rightarrow$ wide margins, points allowed on the wrong side of hyperplane (high bias, low variance).

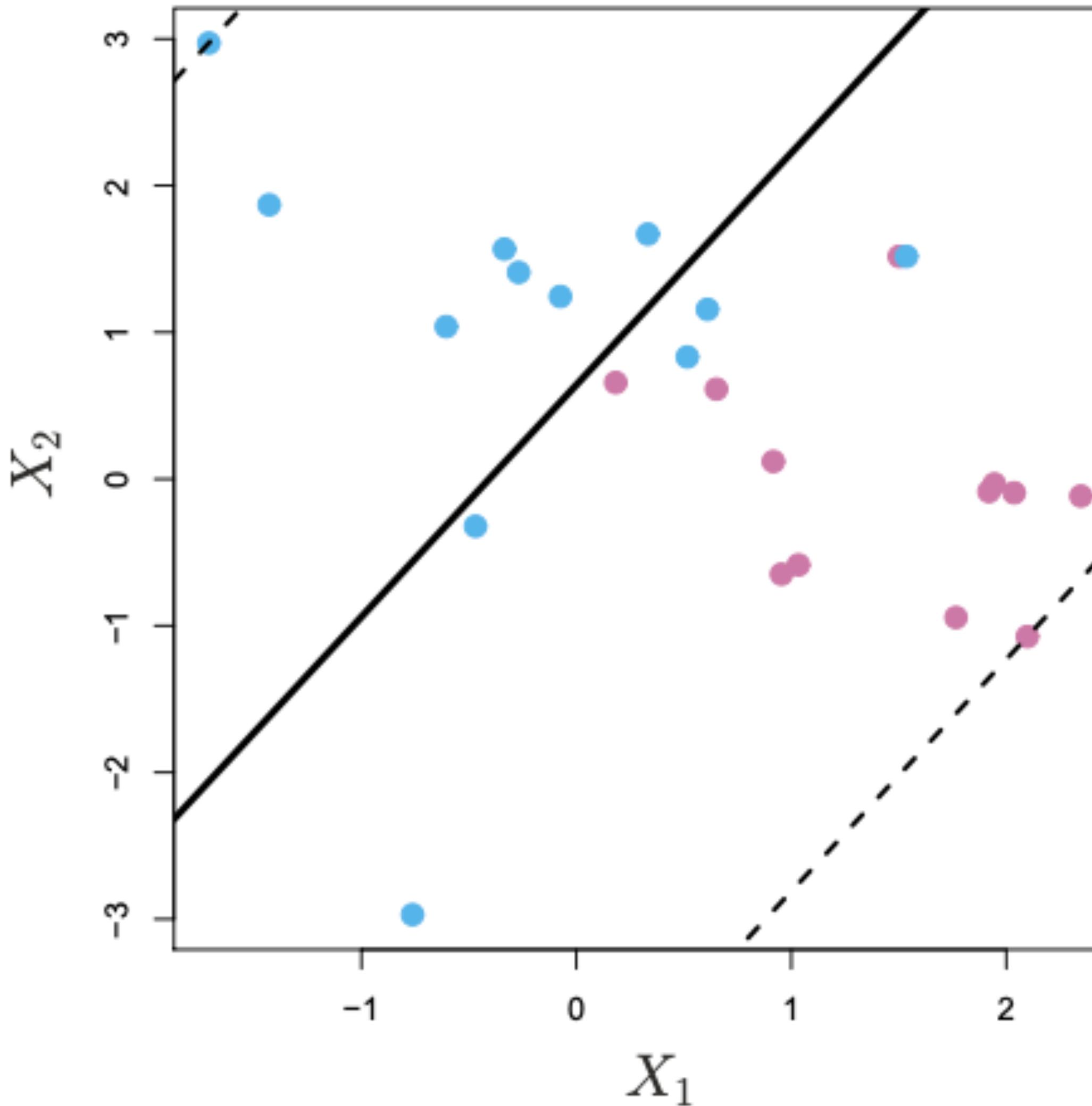
$$\mathcal{C}_1 \quad w_1^2 + w_2^2 = 1$$

$$\mathcal{C}_2 \quad y_i(w_0 + w_1x_{i1} + w_2x_{i2}) \geq M(1 - \epsilon_i) \quad \text{for any training point } i$$

$$\mathcal{C}_3 \quad \epsilon_i \geq 0, \sum_{i=1}^n \epsilon_i \leq C$$

SUPPORT VECTOR MACHINES (SVM)

What does all this mean?

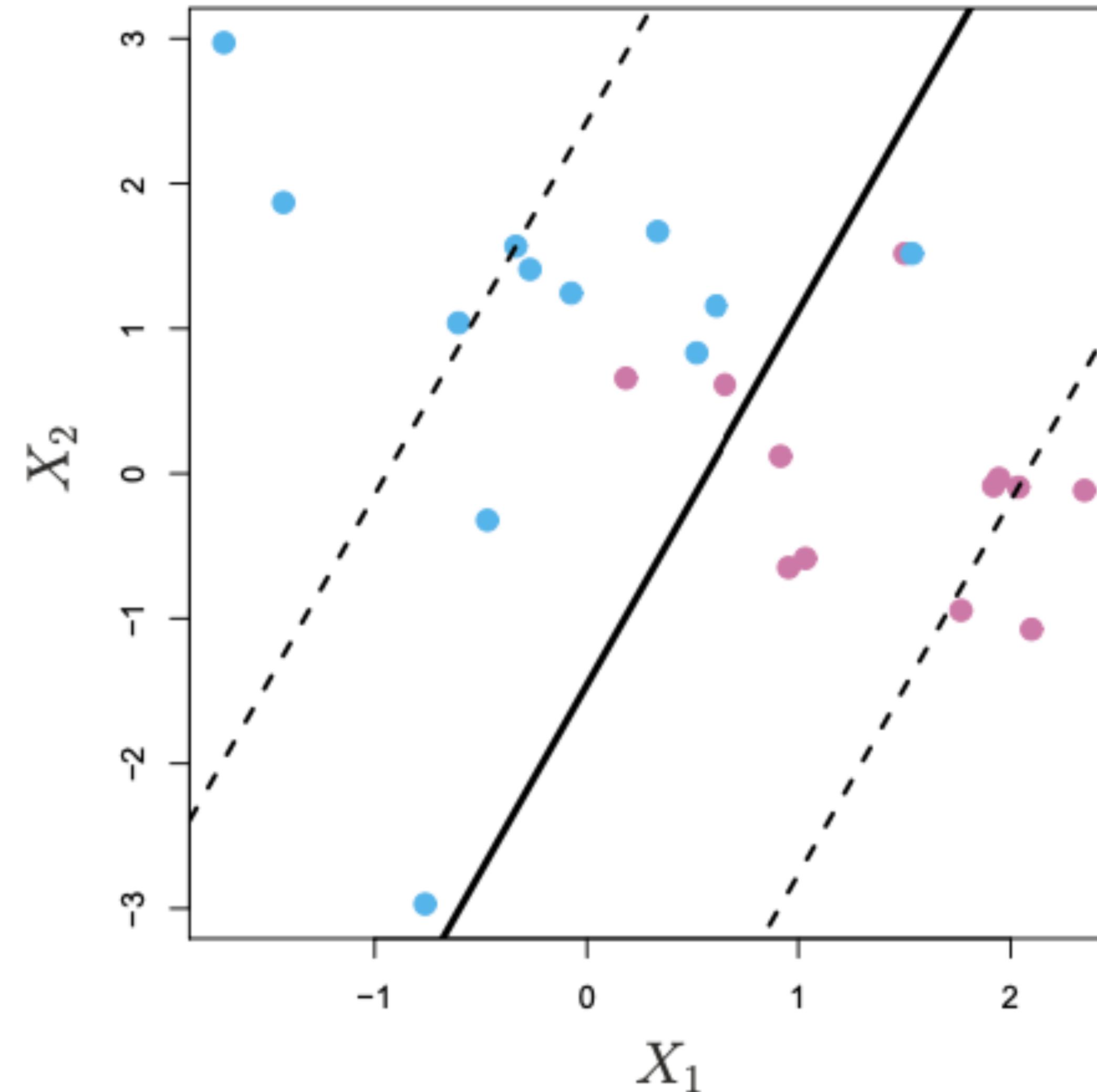


Very large C .

(Many points inside margin, many points on the wrong side of the line)

SUPPORT VECTOR MACHINES (SVM)

What does all this mean?

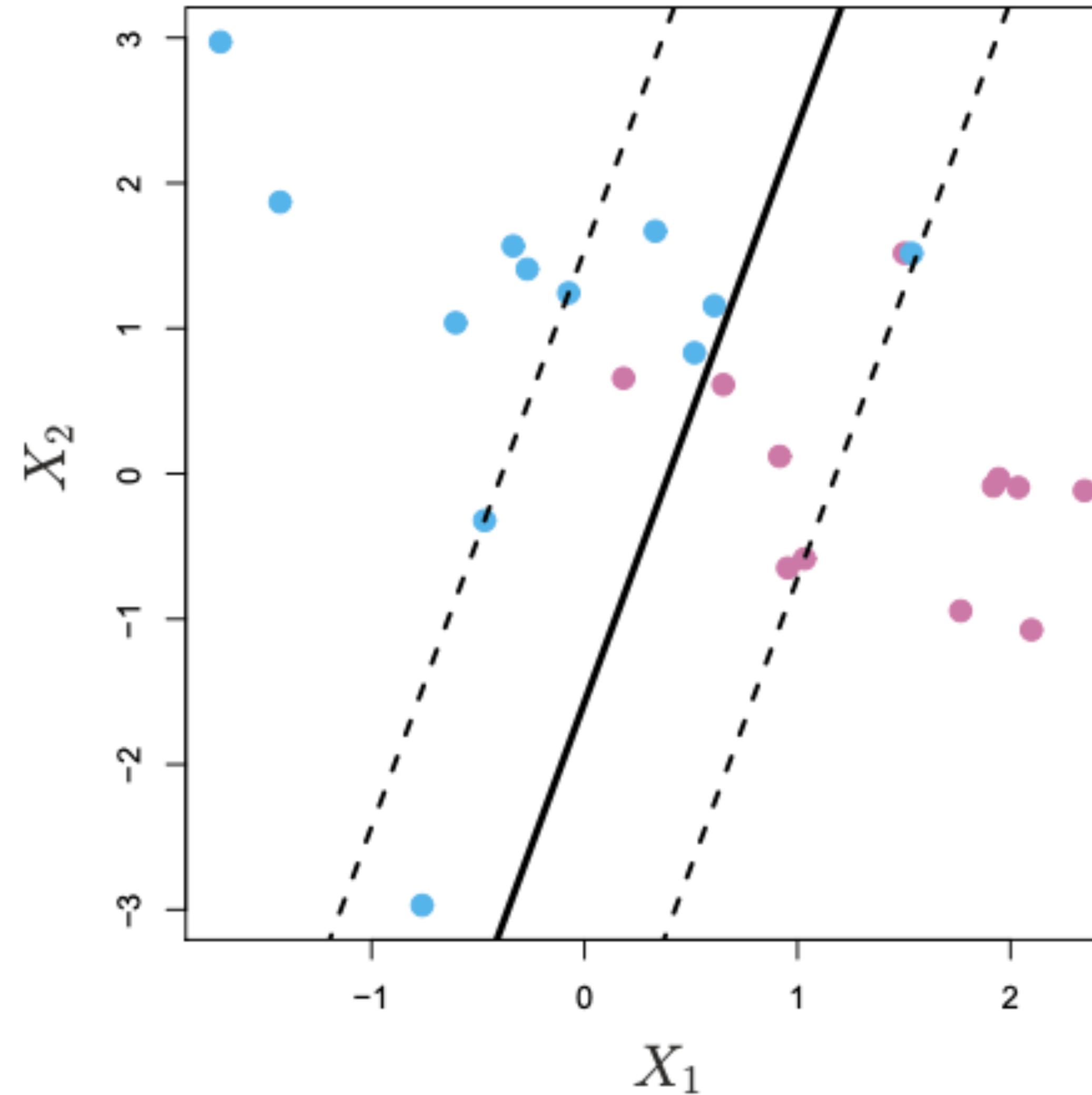


Smaller C .

(Fewer points inside margin, fewer points on the wrong side of the line)

SUPPORT VECTOR MACHINES (SVM)

What does all this mean?

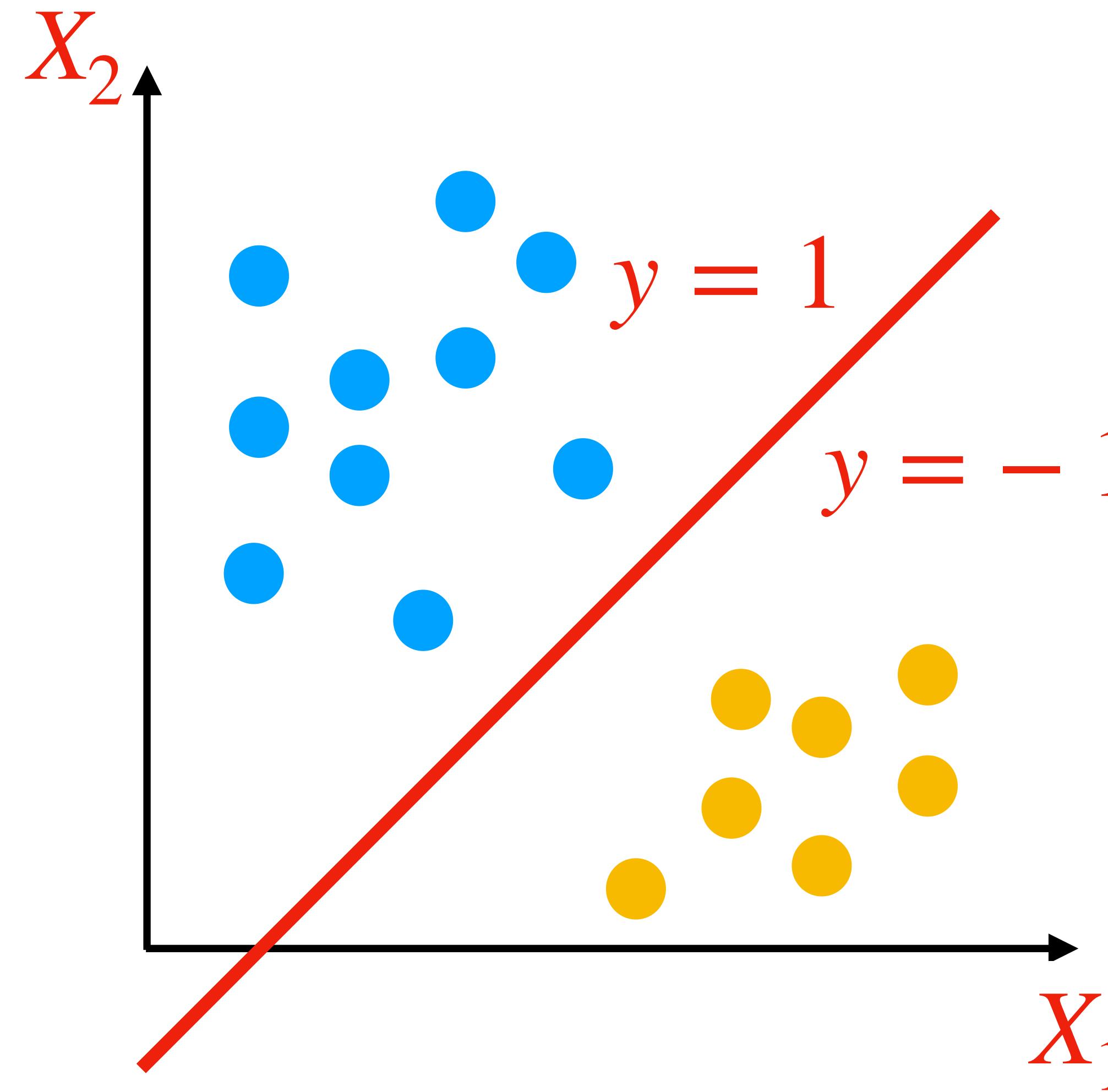


Even Smaller C .

(Even fewer points inside margin, very few points on the wrong side of the line)

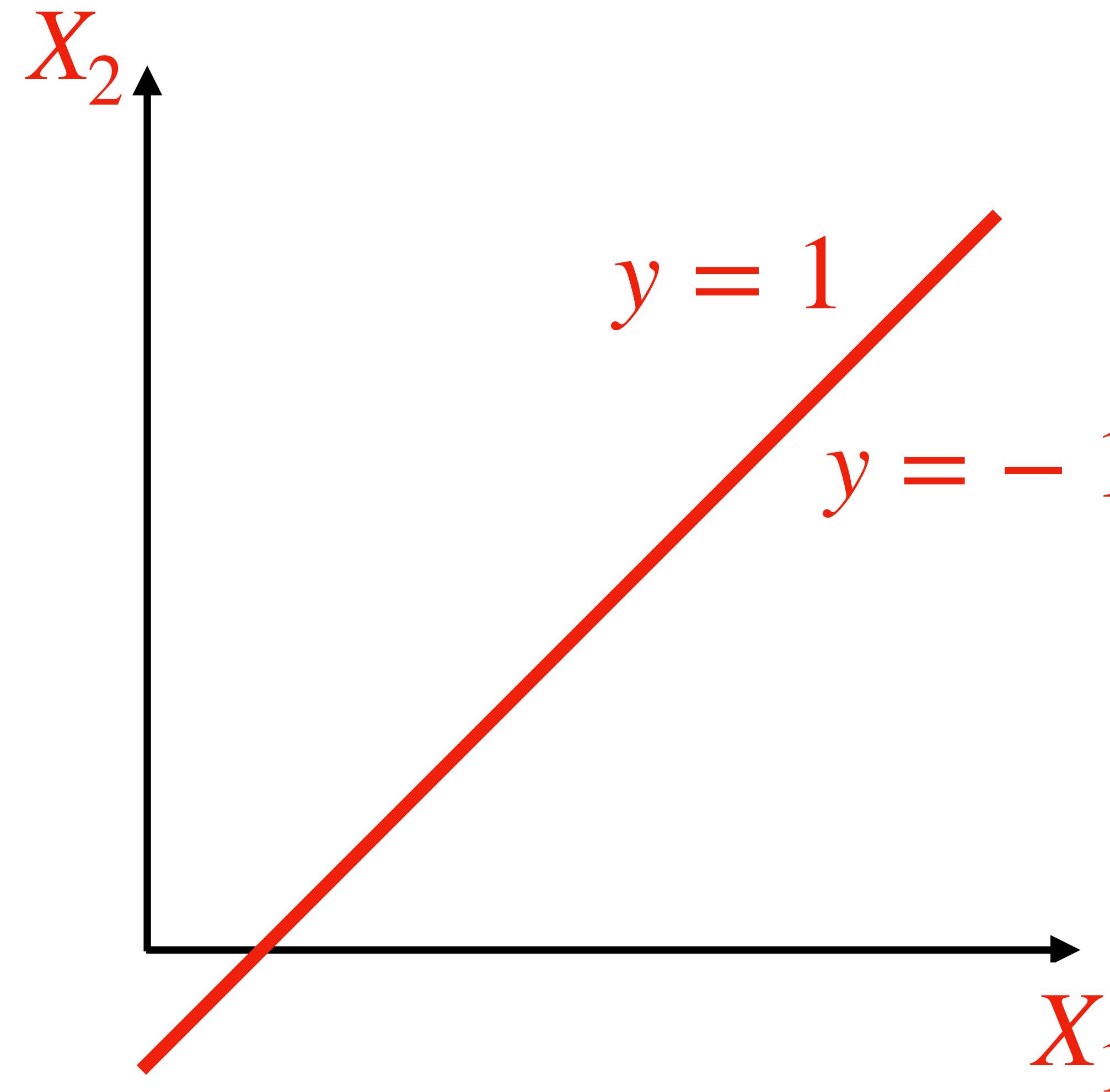
REVIEW OF PREVIOUS LECTURE: SUPPORT VECTOR MACHINES

We want to find the hyperplane that
'separates' the classes 'well'.



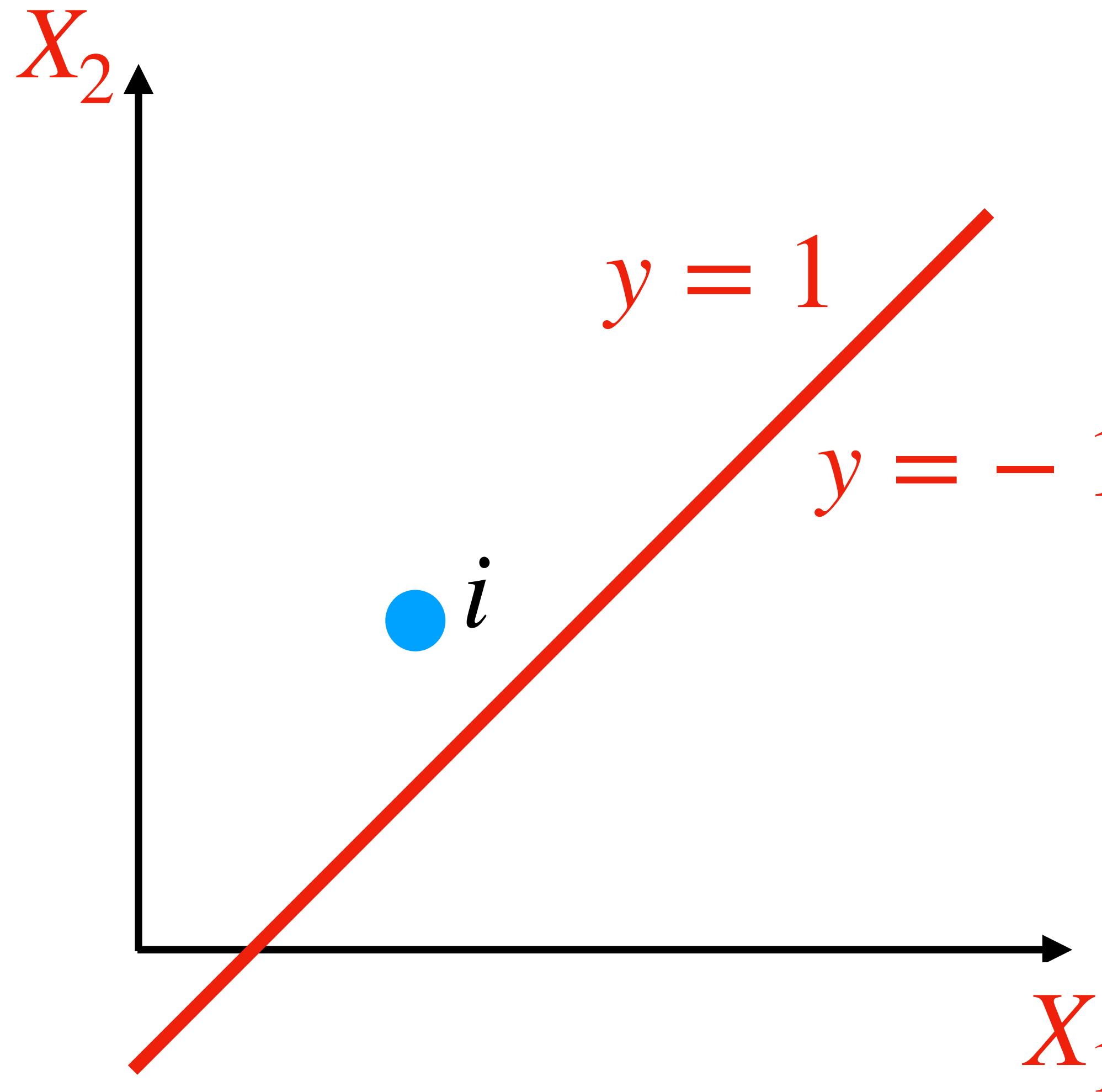
REVIEW OF PREVIOUS LECTURE: SUPPORT VECTOR MACHINES

We want to find the hyperplane that
'separates' the classes 'well'.



Meaning of 'separates classes':

REVIEW OF PREVIOUS LECTURE: SUPPORT VECTOR MACHINES



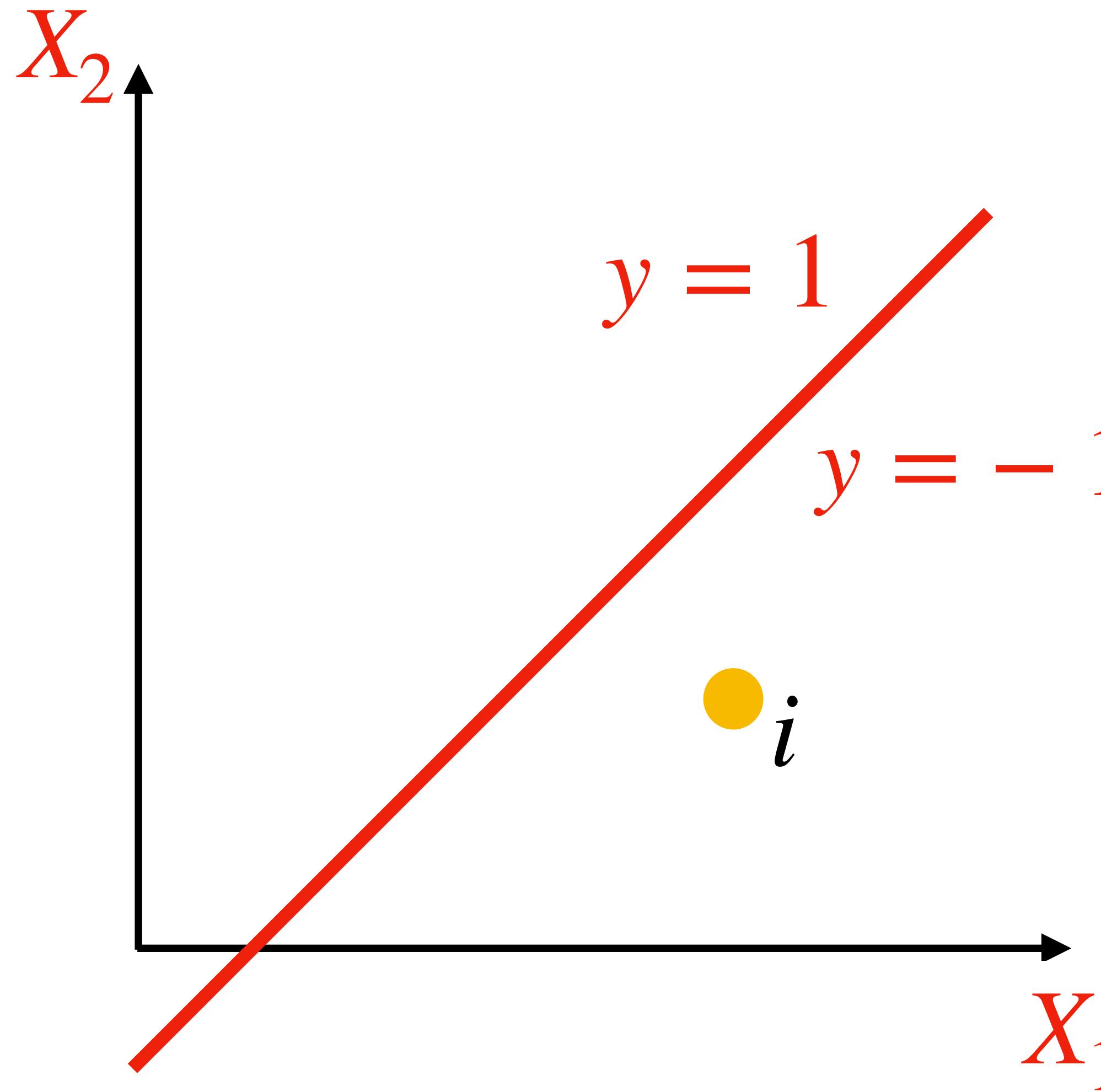
We want to find the hyperplane that
'separates' the classes 'well'.

Meaning of 'separates classes':

For a training point i with $y_i = 1$:

$$w_0 + w_1 x_{i1} + w_2 x_{i2} > 0$$

REVIEW OF PREVIOUS LECTURE: SUPPORT VECTOR MACHINES



We want to find the hyperplane that
'separates' the classes 'well'.

Meaning of 'separates classes':

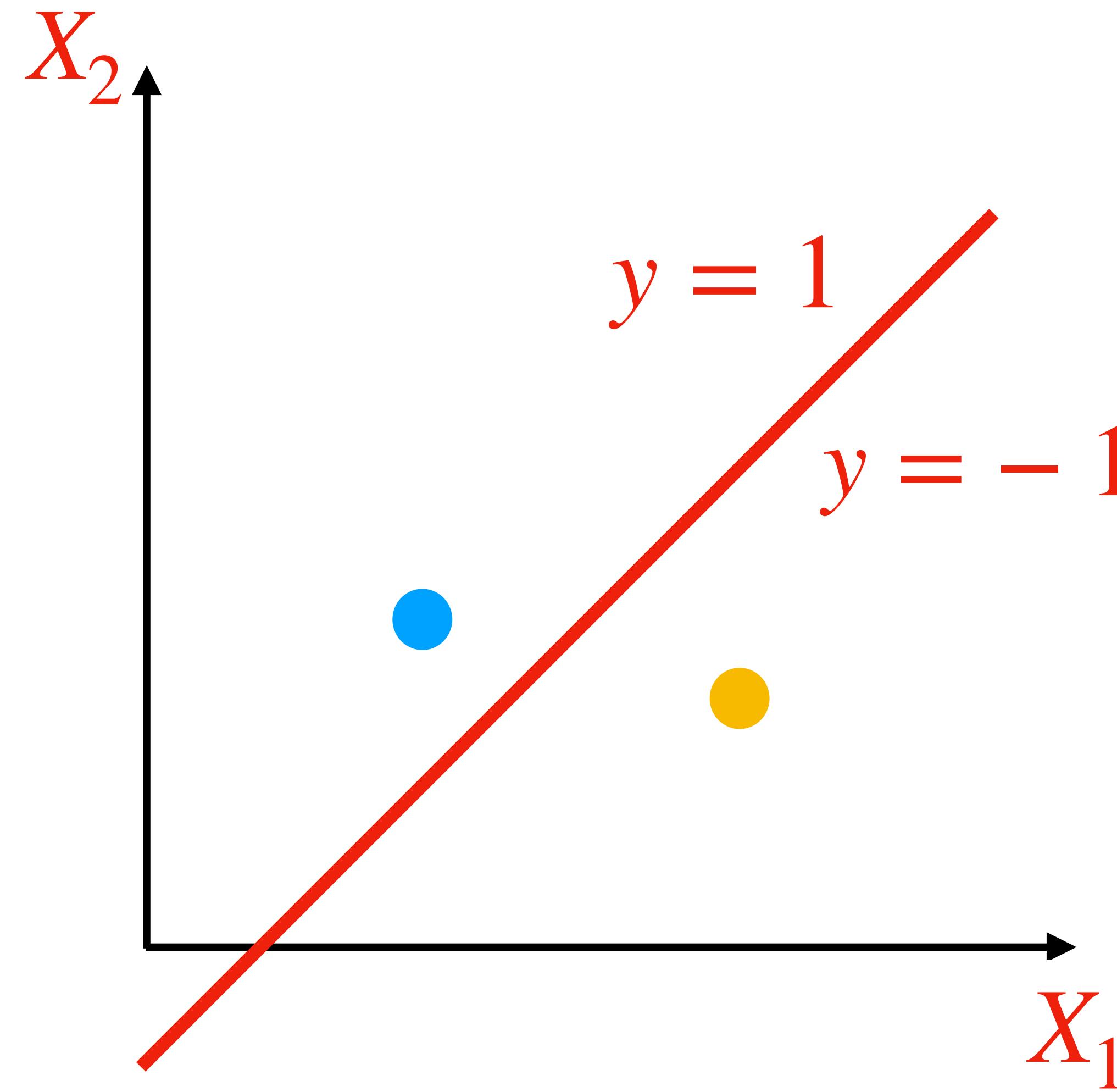
For a training point i with $y_i = 1$:

$$w_0 + w_1 x_{i1} + w_2 x_{i2} > 0$$

For a training point i with $y_i = -1$:

$$w_0 + w_1 x_{i1} + w_2 x_{i2} < 0$$

REVIEW OF PREVIOUS LECTURE: SUPPORT VECTOR MACHINES



We want to find the hyperplane that
'separates' the classes 'well'.

Meaning of 'separates classes':

For a training point i with $y_i = 1$:

$$w_0 + w_1 x_{i1} + w_2 x_{i2} > 0$$

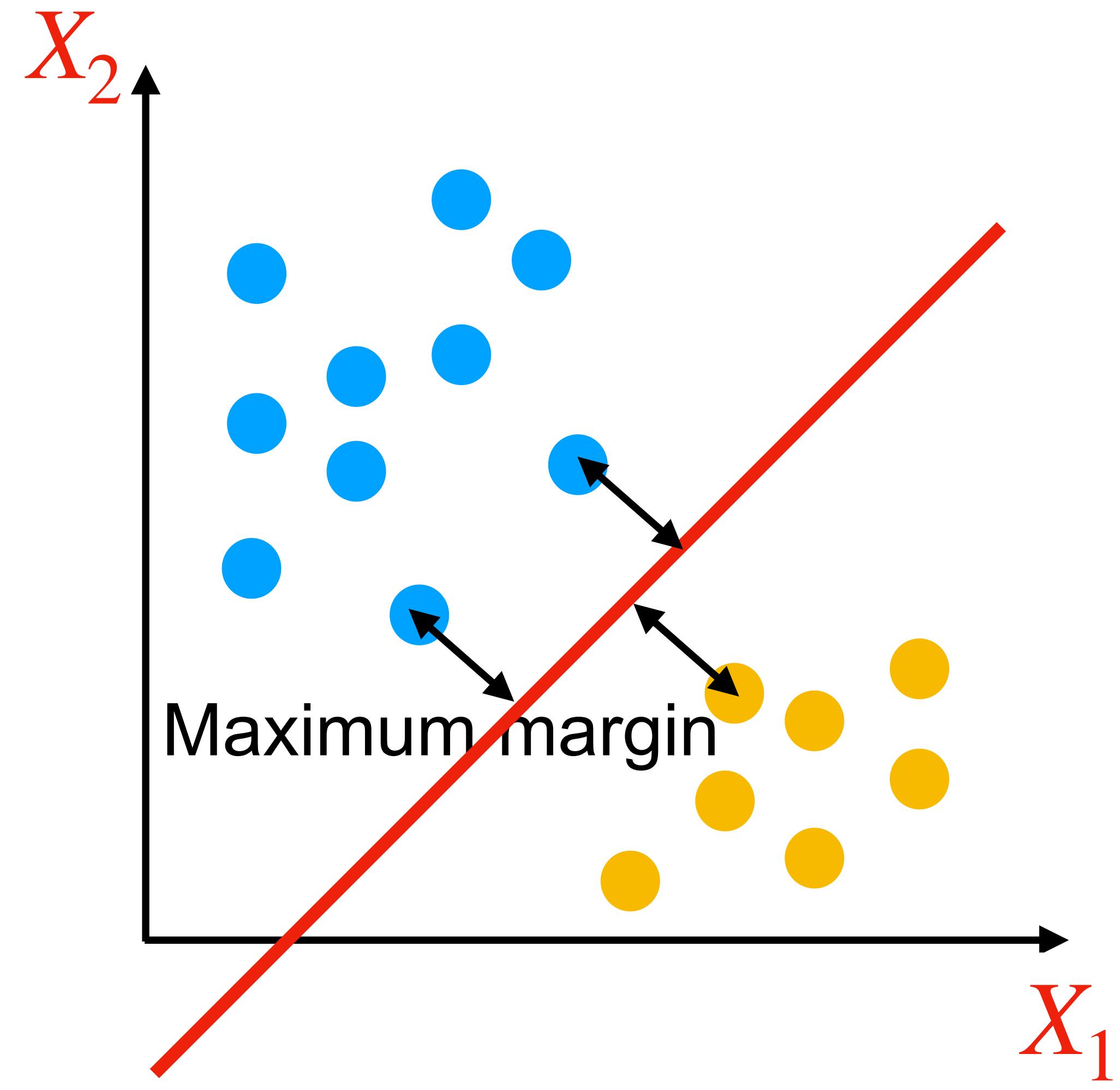
For a training point i with $y_i = -1$:

$$w_0 + w_1 x_{i1} + w_2 x_{i2} < 0$$

Can combine them in one inequality:

$$y_i(w_0 + w_1 x_{i1} + w_2 x_{i2}) > 0$$

REVIEW OF PREVIOUS LECTURE: SUPPORT VECTOR MACHINES

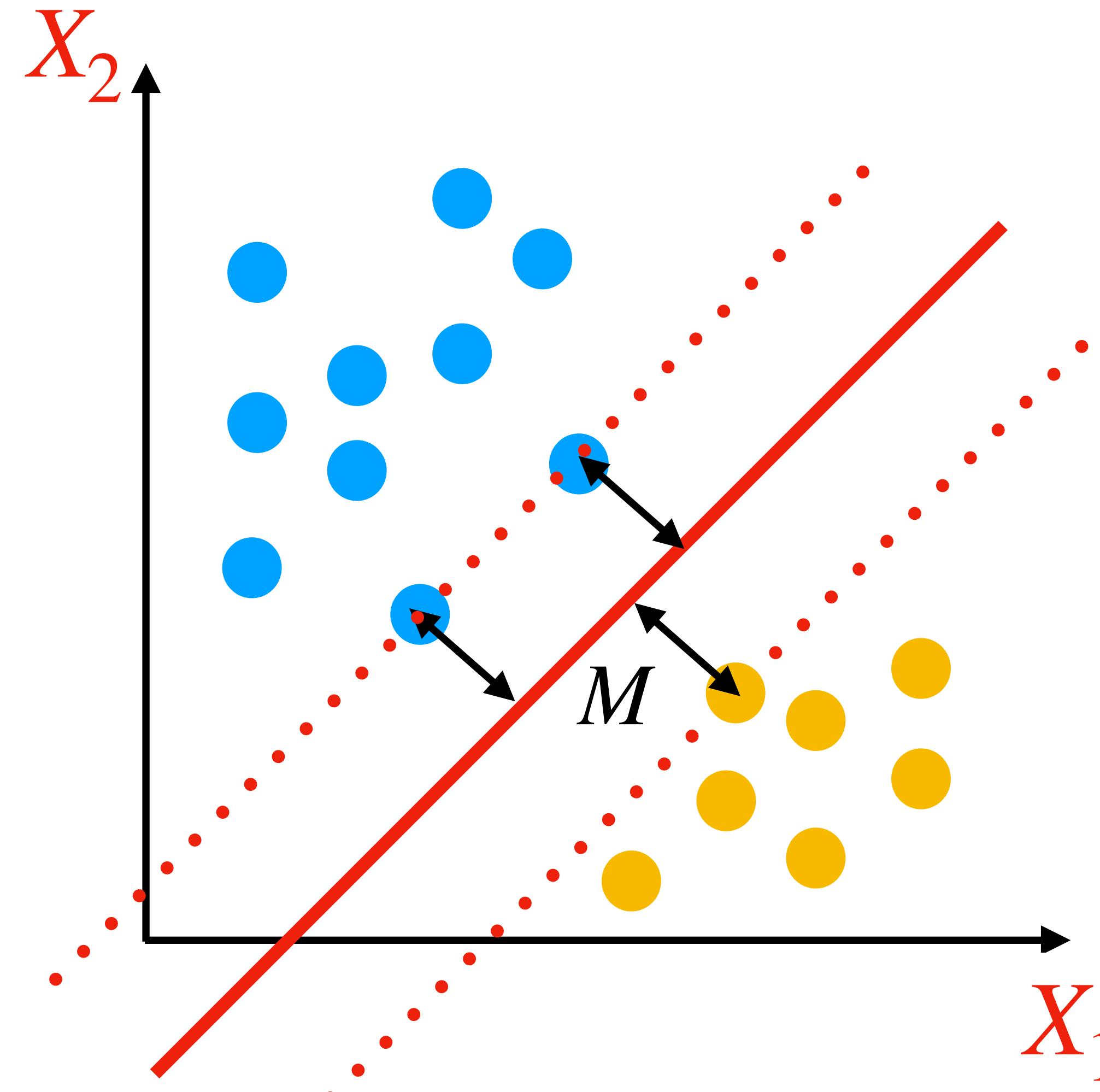


How about ‘separates classes well’: Among red, green, purple:
red line separates the classes best.

Margin: minimum distance from any training point to the hyperplane.

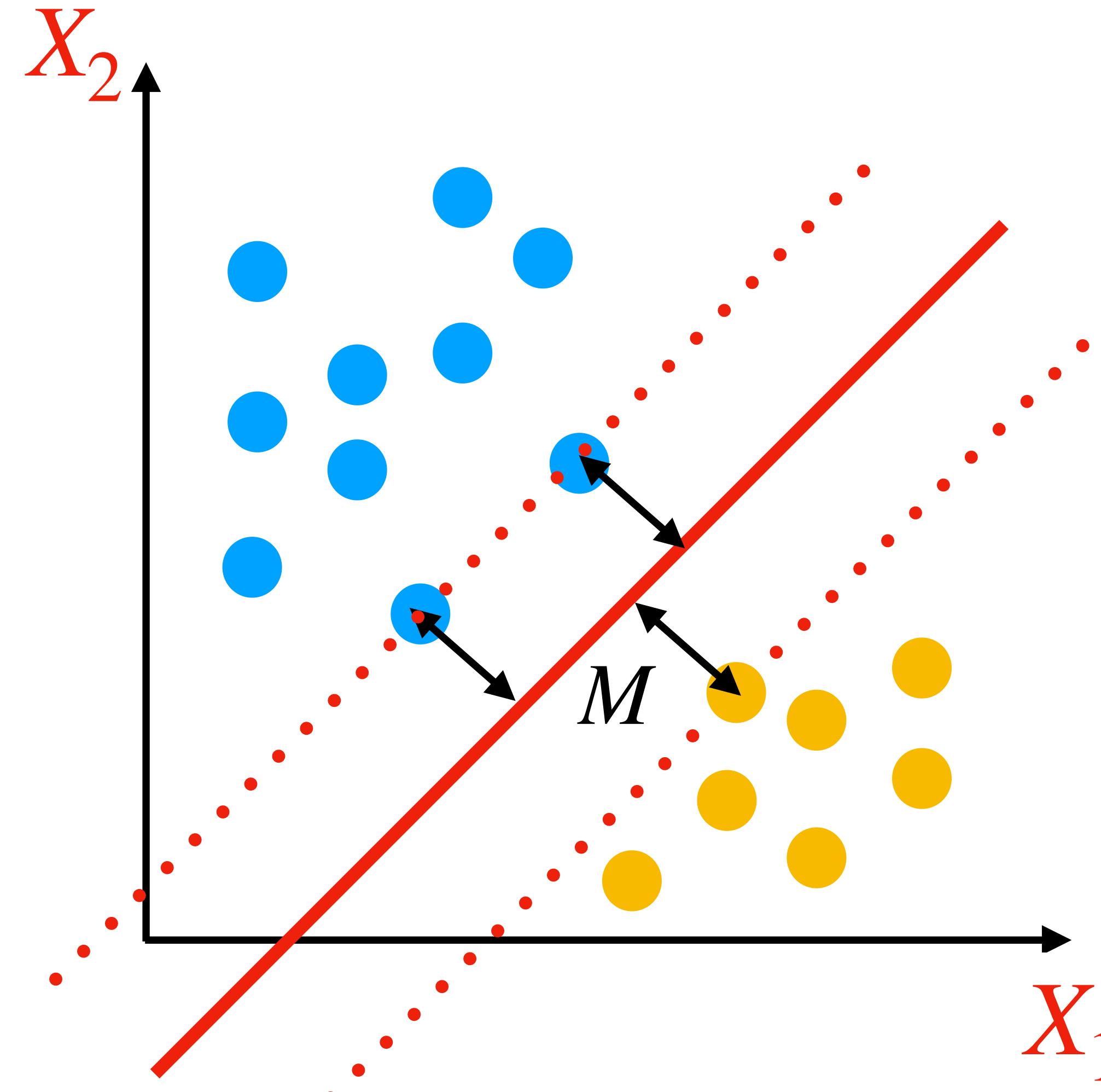
Red line has the maximum margin.

REVIEW OF PREVIOUS LECTURE: SUPPORT VECTOR MACHINES



Informally the goal is to:
Find hyperplane that provides
maximum margin M .

REVIEW OF PREVIOUS LECTURE: SUPPORT VECTOR MACHINES



Formally the goal is to:

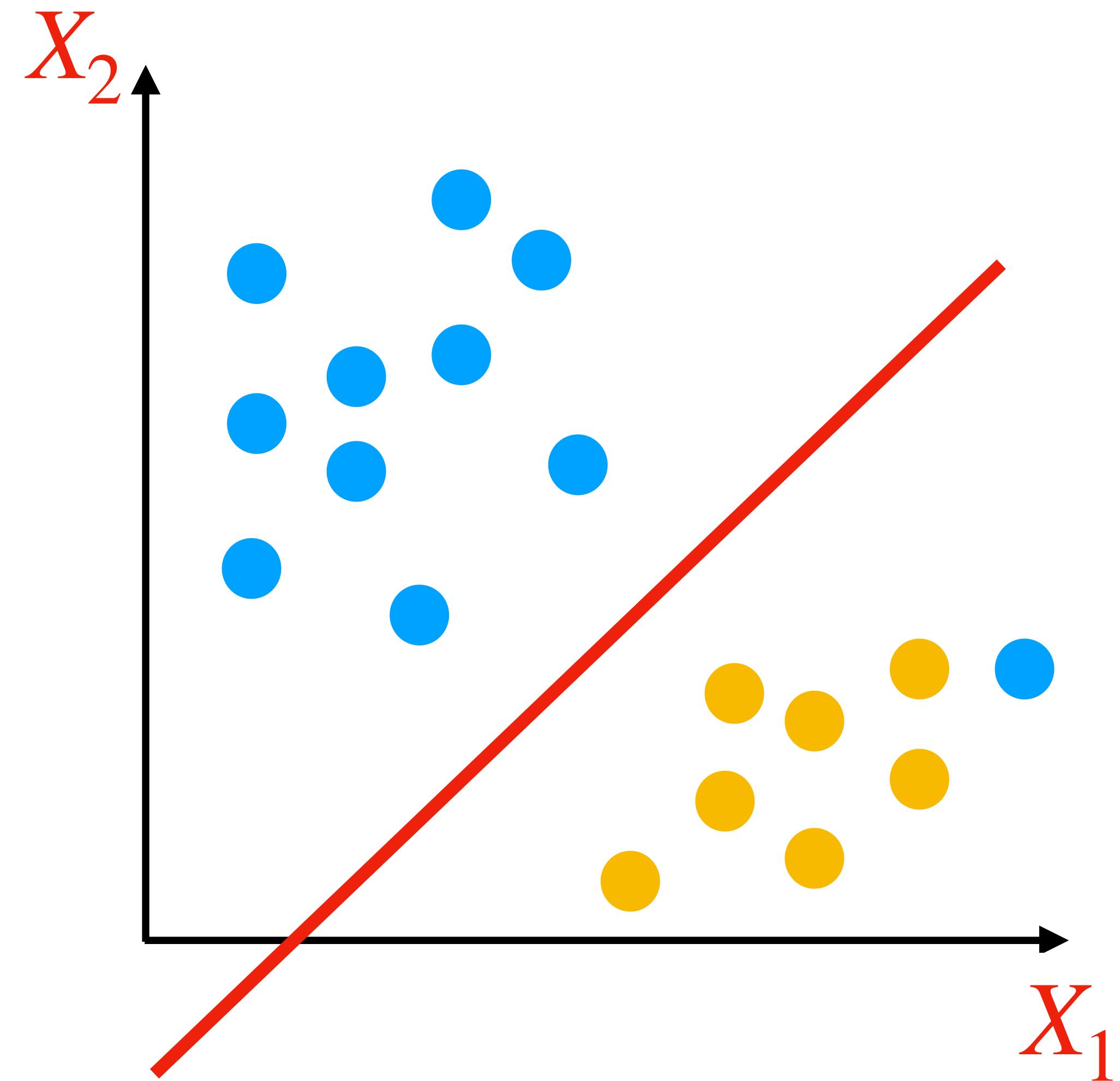
Find w_0, w_1, w_2 that
maximizes M
with the constraints

$$\mathcal{C}_1 \quad w_1^2 + w_2^2 = 1$$

$$\mathcal{C}_2 \quad y_i(w_0 + w_1x_{i1} + w_2x_{i2}) \geq M$$

for any training point i

REVIEW OF PREVIOUS LECTURE: SUPPORT VECTOR MACHINES



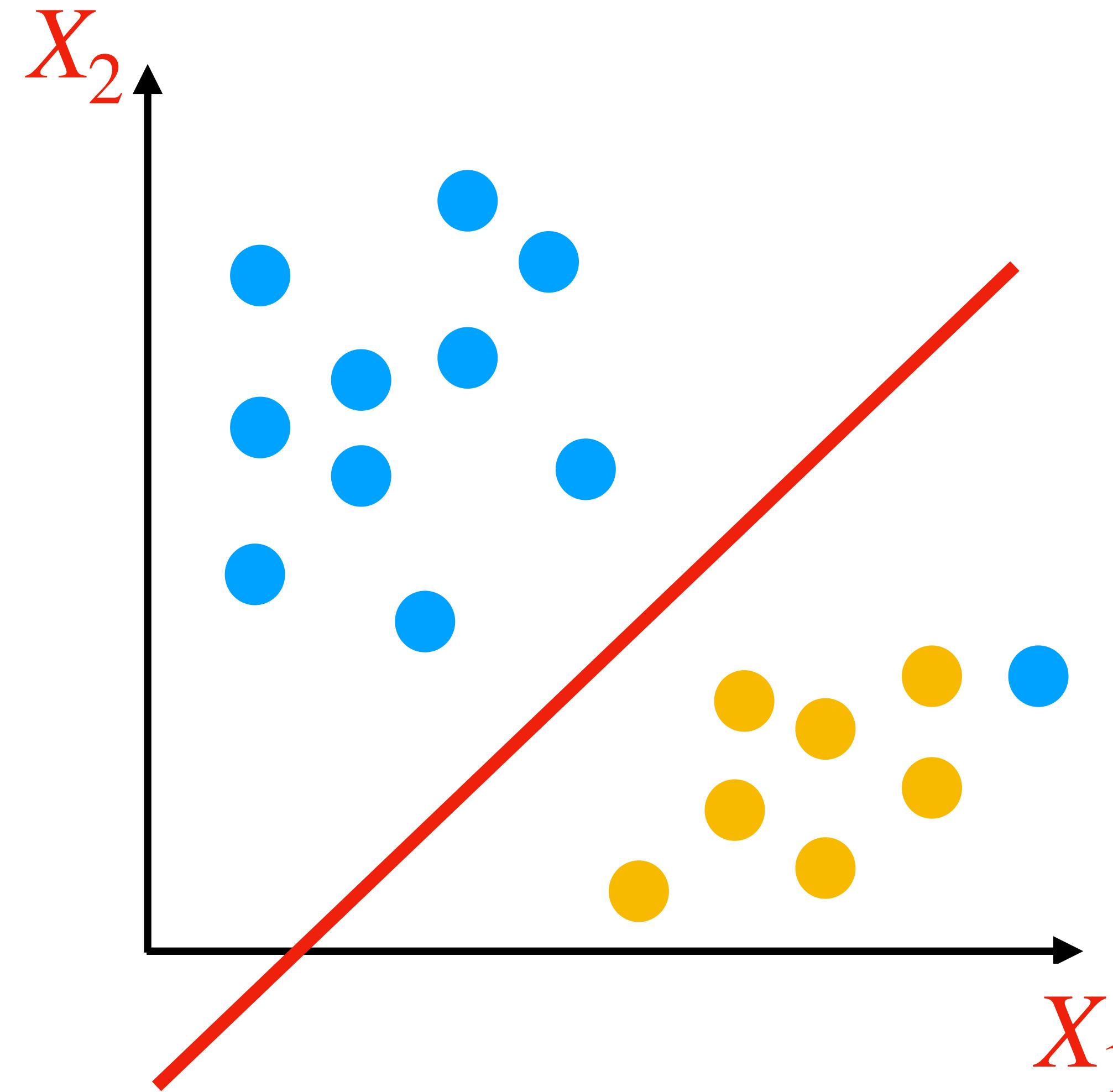
Two problems with the same solution:
Addition of a single point might
drastically change the hyperplane.

Insisting on perfect separation may
cause overfitting!

It may not even be possible to find a
separating hyperplane (line in 2D).

Solution is to use **Soft Margin**:
Add flexibility; some observations can
be on the wrong side.

REVIEW OF PREVIOUS LECTURE: SUPPORT VECTOR MACHINES



Soft Margin:

Find w_0, w_1, w_2 that
maximizes M
with the constraints

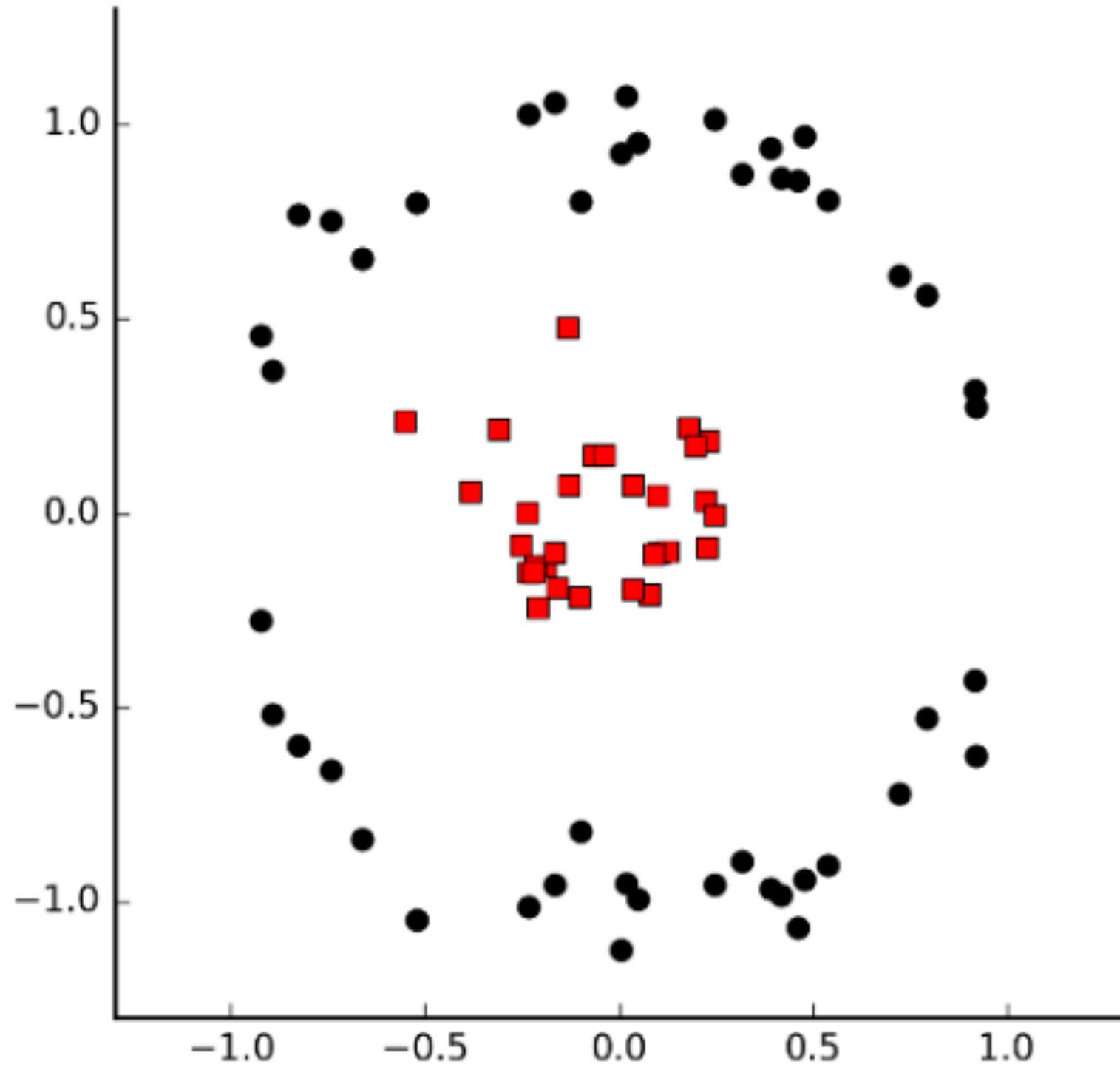
$$\mathcal{C}_1 \quad w_1^2 + w_2^2 = 1$$

$$\mathcal{C}_2 \quad y_i(w_0 + w_1x_{i1} + w_2x_{i2}) \geq M(1 - \epsilon_i) \quad \text{for any training point } i$$

$$\mathcal{C}_3 \quad \epsilon_i \geq 0, \sum_{i=1}^n \epsilon_i \leq C$$

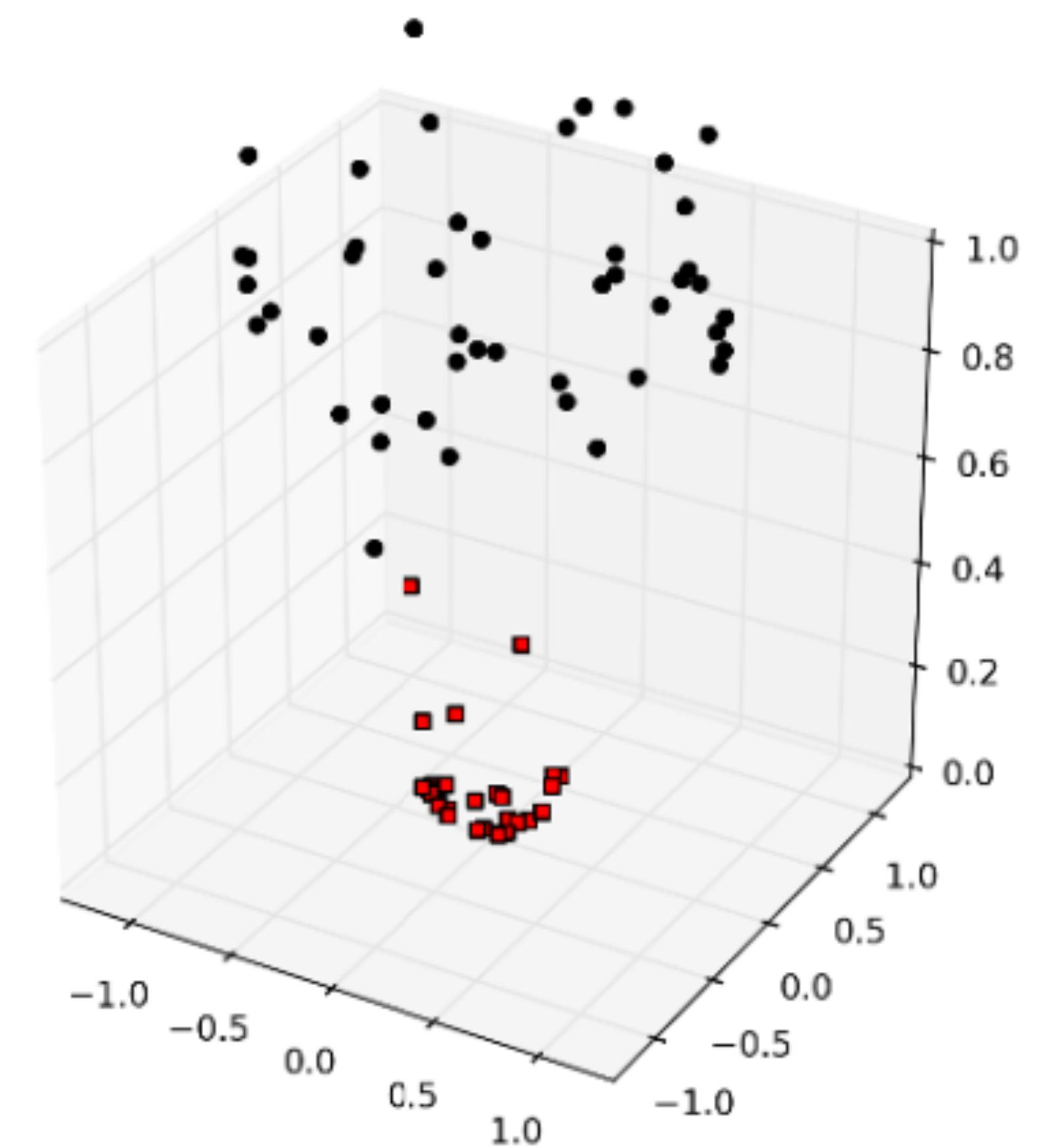
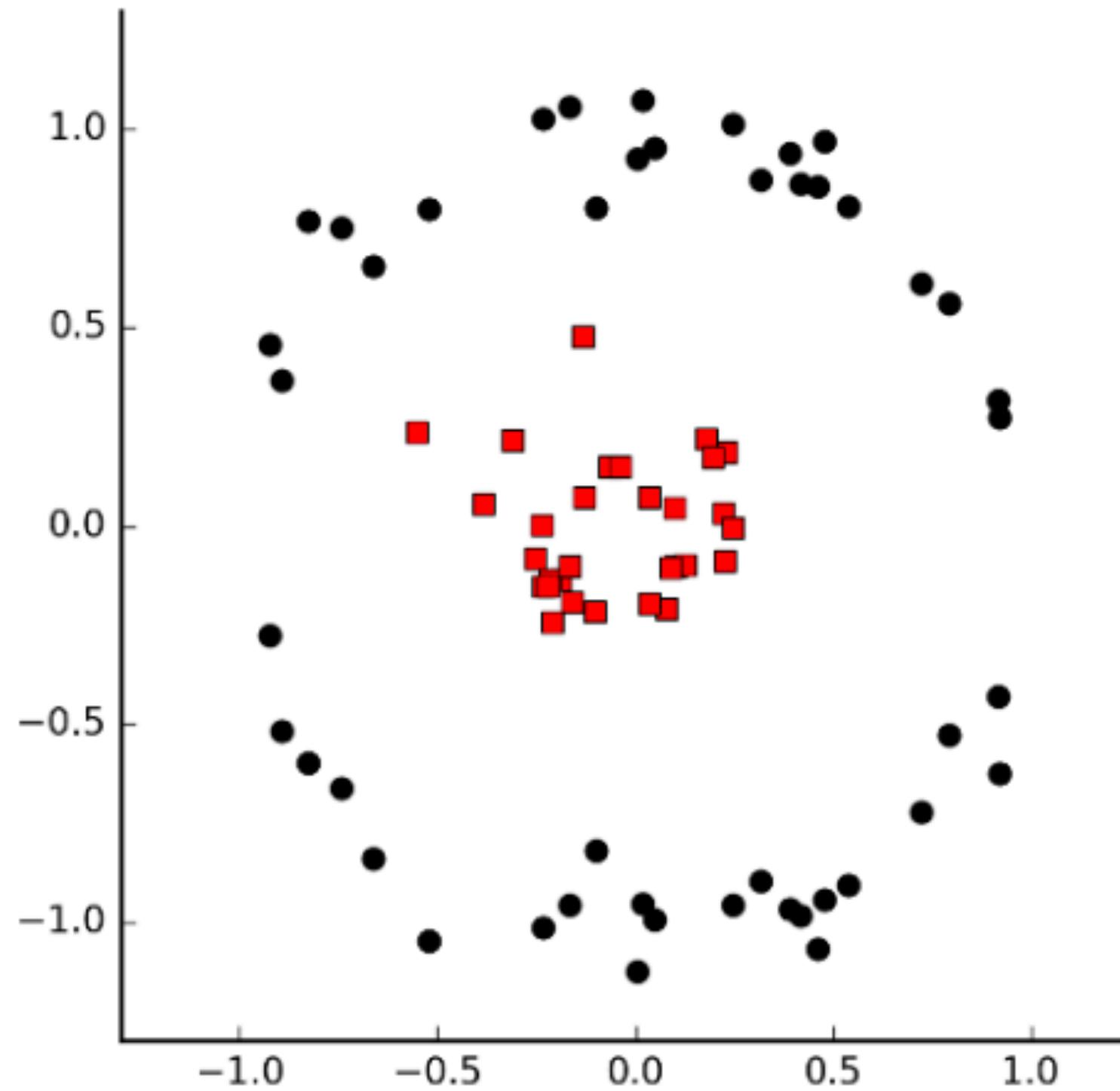
SUPPORT VECTOR MACHINES (SVM)

What if we need nonlinear separation?



SUPPORT VECTOR MACHINES (SVM)

What if we need nonlinear separation?

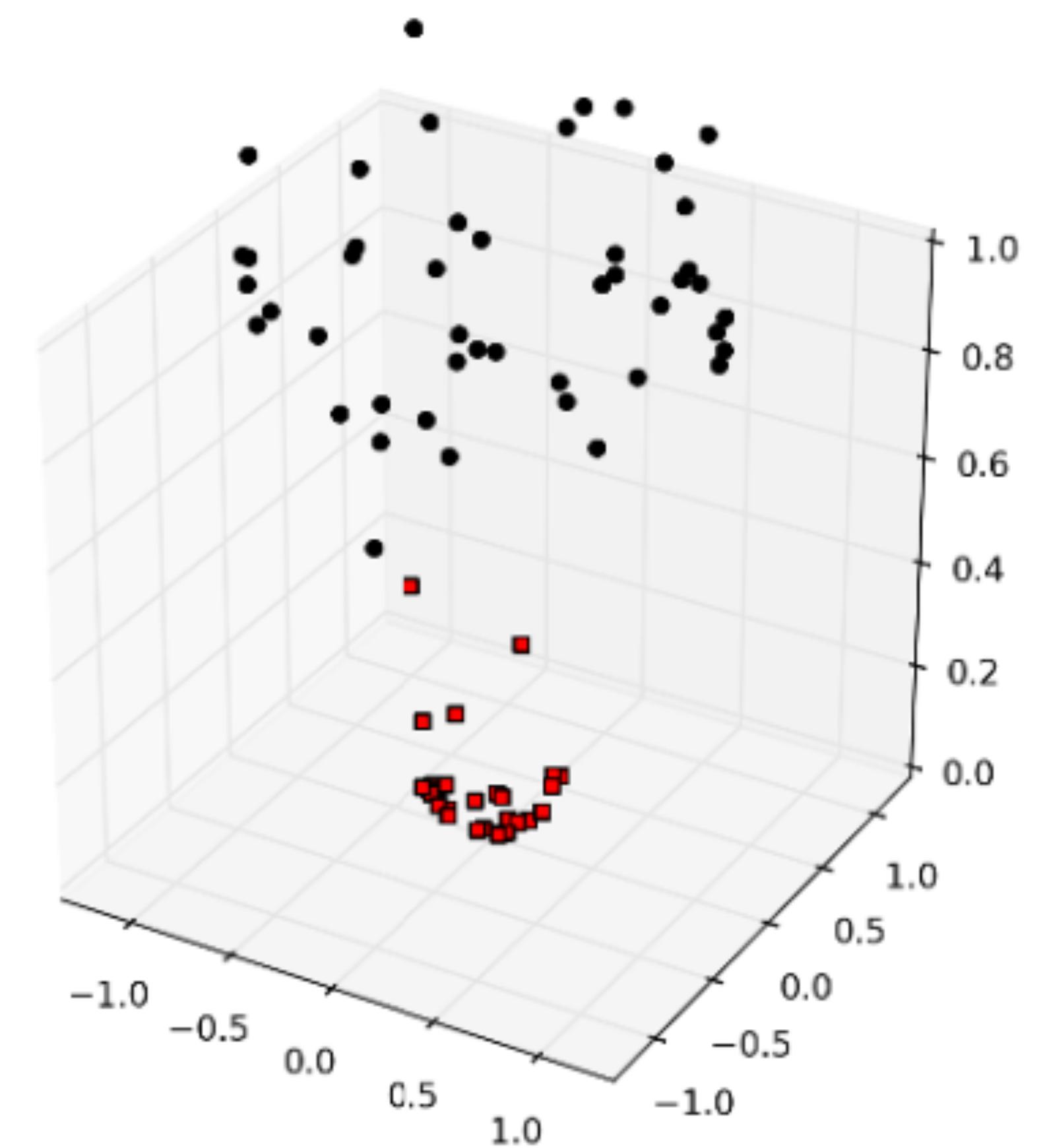
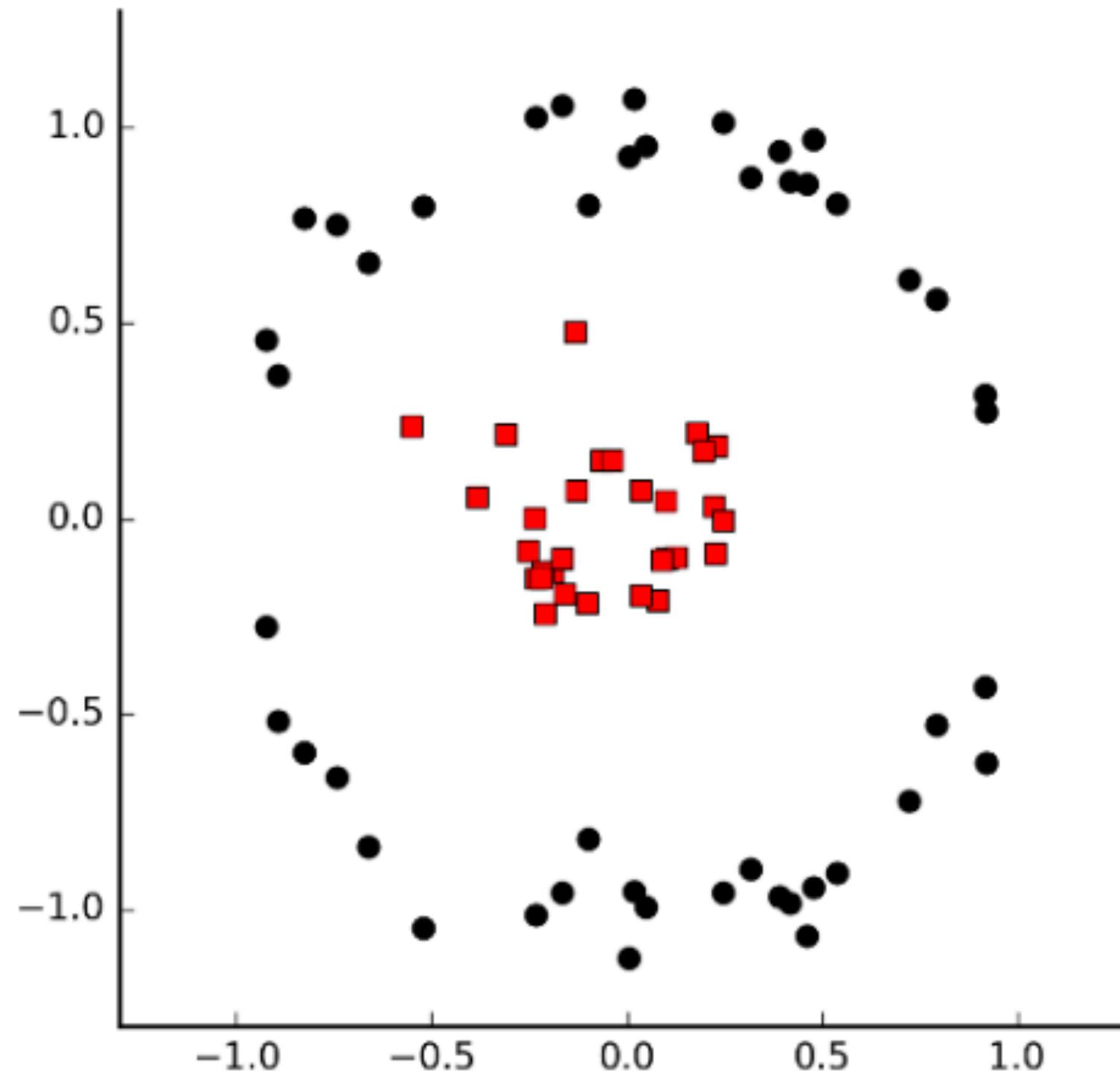


Project m dimensional points to higher dimensions.

Ex: $(x, y) \rightarrow (x, y, x^2 + y^2)$

SUPPORT VECTOR MACHINES (SVM)

What if we need nonlinear separation?



How to project points each of dimension m into points of dimension n ?

Kernel functions

SUPPORT VECTOR MACHINES (SVM)

Kernel Functions

Normally for a training point (x_1, x_2) we mainly check the function

$$w_0 + w_1 x_1 + w_2 x_2 \quad (\text{If positive class 1, otherwise class -1})$$

SUPPORT VECTOR MACHINES (SVM)

Kernel Functions

Normally for a training point (x_1, x_2) we mainly check the function

$$w_0 + w_1 x_1 + w_2 x_2 \quad (\text{If positive class 1, otherwise class } -1)$$

This is equivalent to checking the function:

$$w_0 + \sum_{i=1}^n \alpha_i (x \cdot x_i)$$

Dot product

Recall dot product measures similarity (linear relationship).

SUPPORT VECTOR MACHINES (SVM)

Kernel Functions

Normally for a training point (x_1, x_2) we mainly check the function

$$w_0 + w_1 x_1 + w_2 x_2 \quad (\text{If positive class 1, otherwise class } -1)$$

This is equivalent to checking the function:

$$w_0 + \sum_{i=1}^n \alpha_i (x \cdot x_i)$$

Dot product

Recall dot product measures similarity (linear relationship).

Generalize: Replace dot product with any appropriate similarity measure, called kernel

$$w_0 + \sum_{i=1}^n \alpha_i K(x, x_i)$$

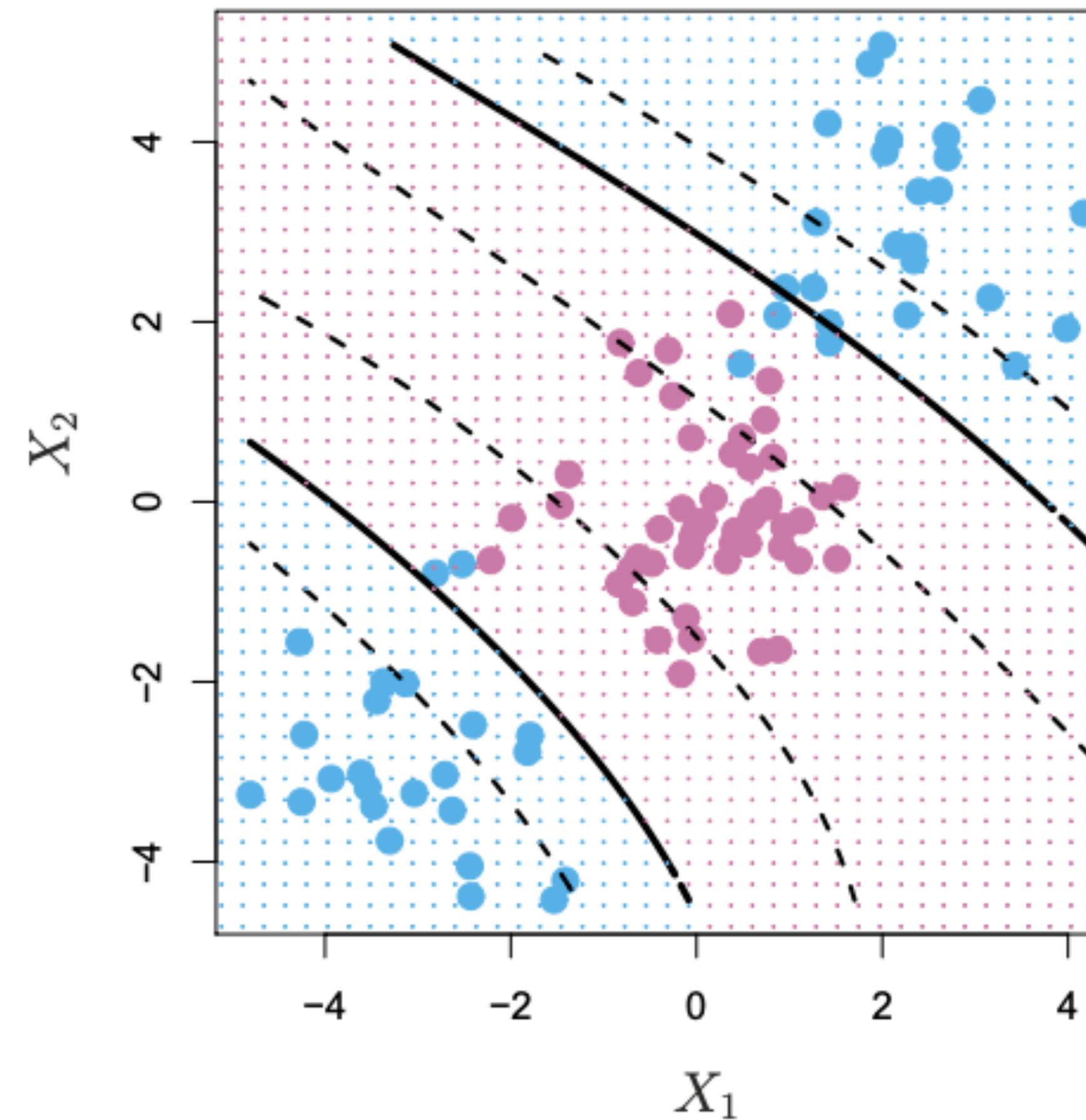
Kernel

Popular non-linear kernels:
Polynomial, radial

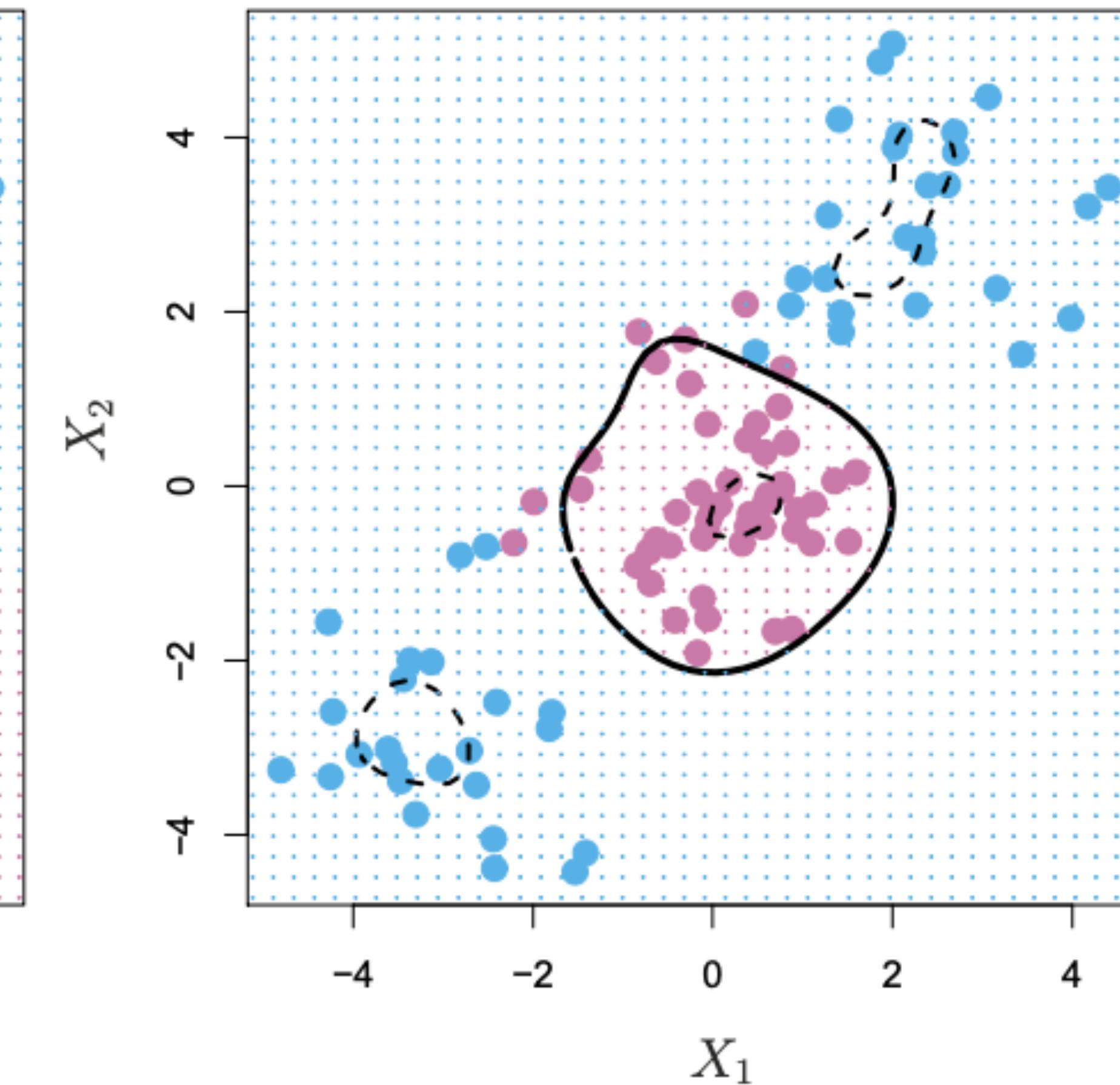
SUPPORT VECTOR MACHINES (SVM)

Kernel Functions

Example: Non-linear data



Polynomial ($d = 3$) kernel



Radial kernel

SUPPORT VECTOR MACHINES (SVM)

SVM in scikit learn:

```
class sklearn.svm.SVC(*, C=1.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0,  
shrinking=True, probability=False, tol=0.001, cache_size=200, class_weight=None,  
verbose=False, max_iter=-1, decision_function_shape='ovr', break_ties=False,  
random_state=None)
```

Regularization parameter.

Note: strength of regularization is inversely proportional to C .
(penalty of the error term. small C implies wider margin - may misclassify more points)

Kernel type:

linear, poly, rbf (radial), etc.

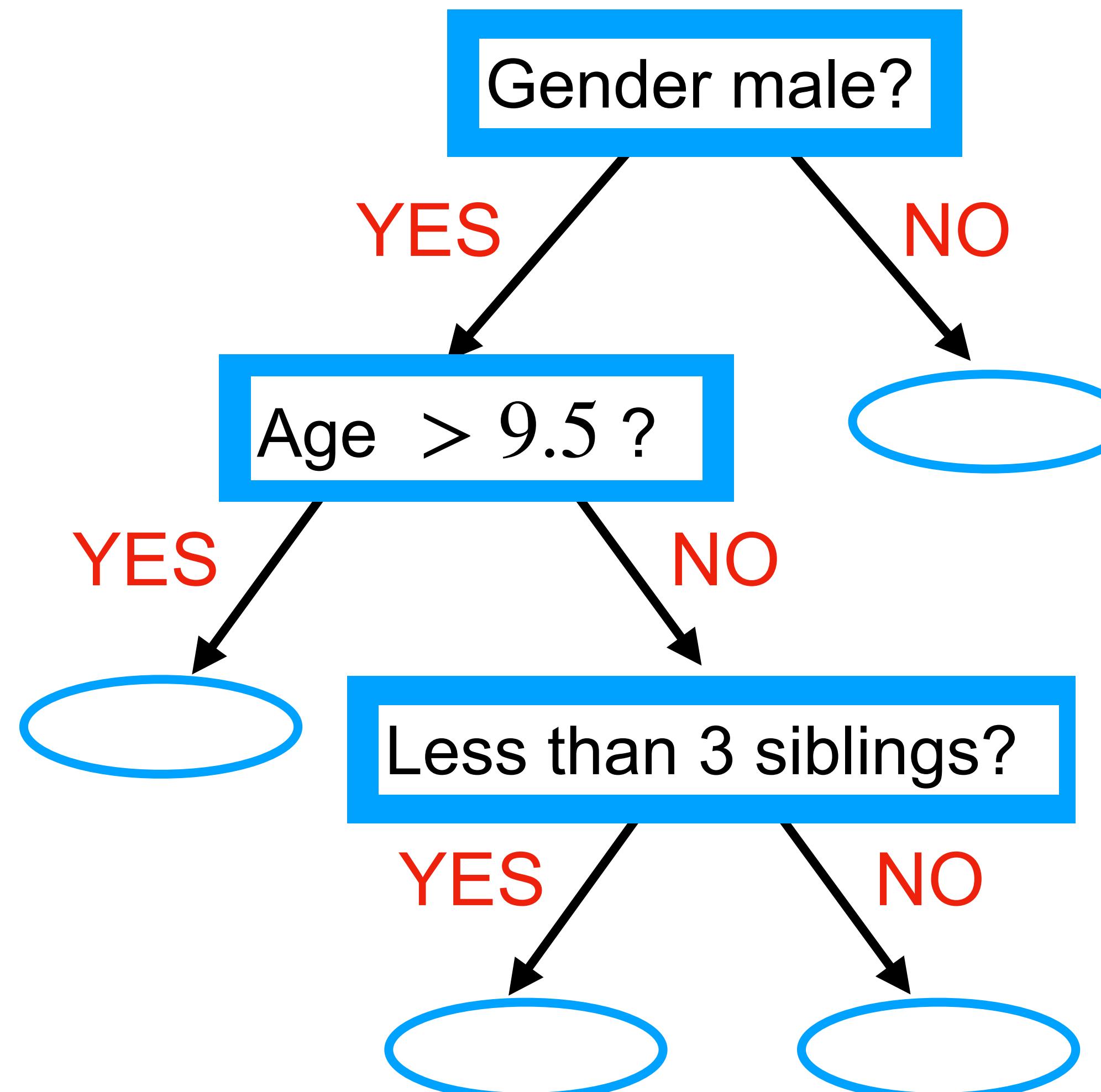
DECISION TREES

Ex: Simple decision tree for predicting survival in Titanic

Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2833	C85	C
1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

DECISION TREES

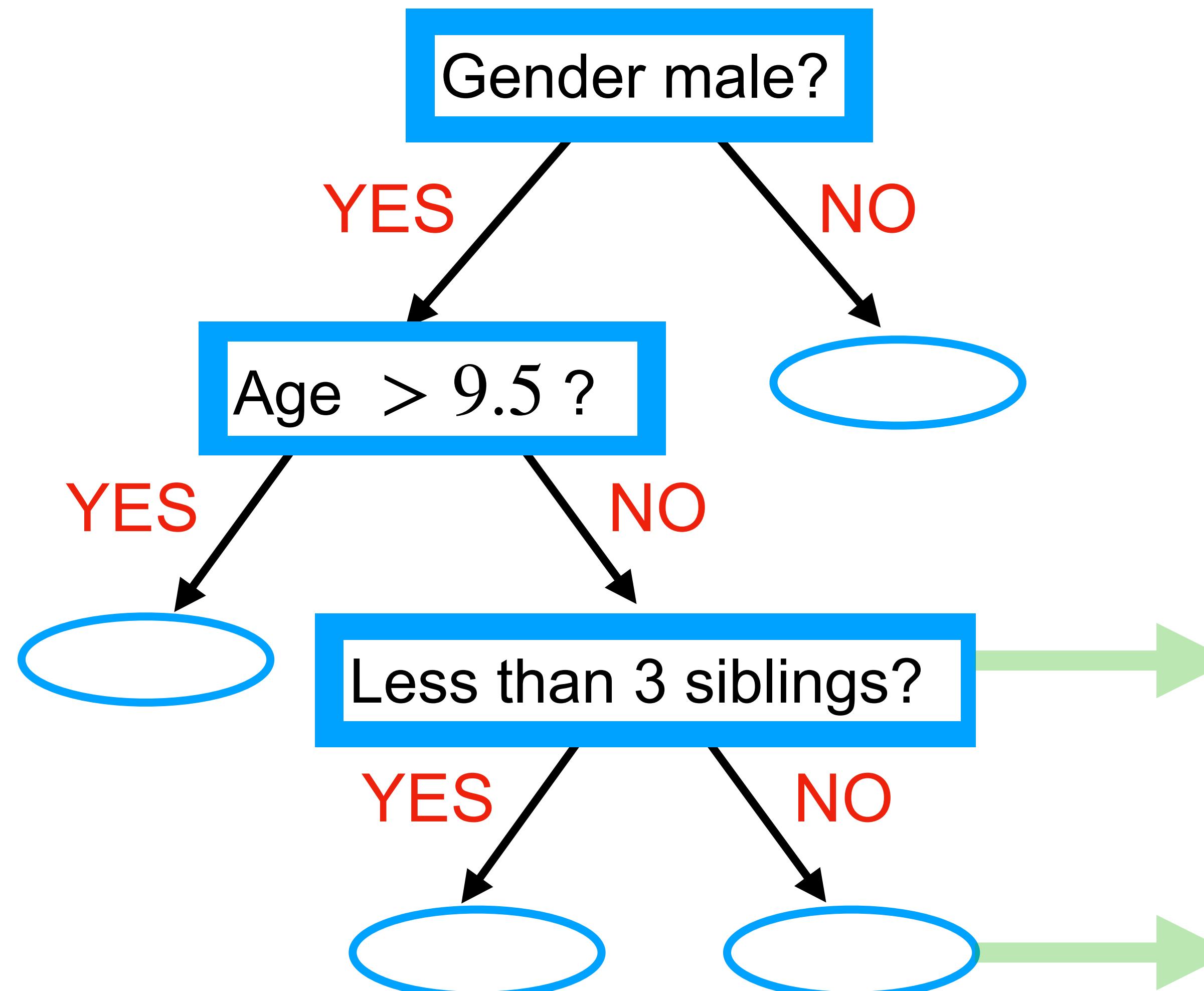
Ex: Simple decision tree for predicting survival in Titanic



Each node corresponds to a group of training data points.

DECISION TREES

Ex: Simple decision tree for predicting survival in Titanic



Each node corresponds to a group of training data points.

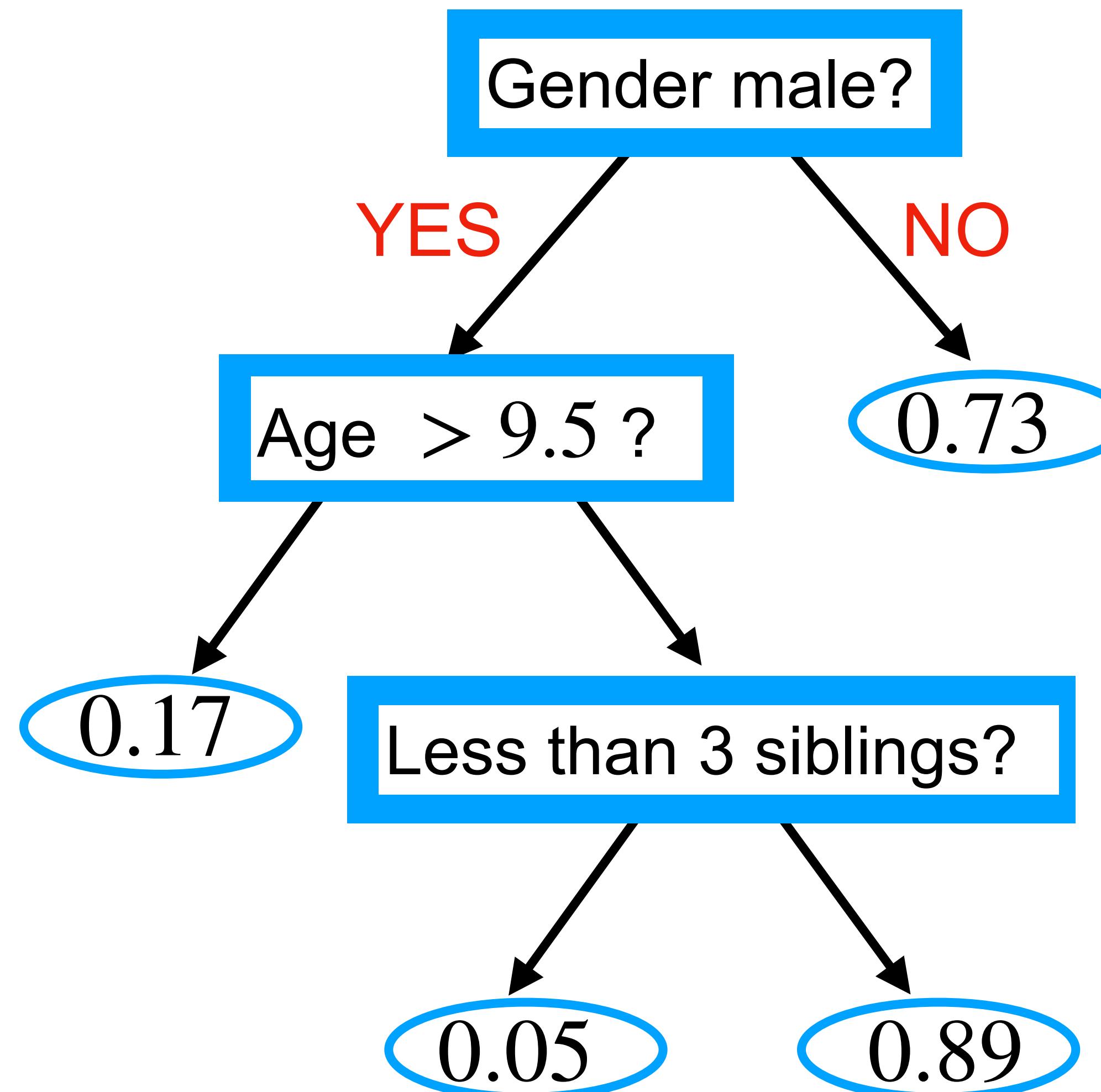
Example:

Passengers who are male and ≤ 9.5 years old.

Passengers who are male and ≤ 9.5 years old and have more than 2 siblings.

DECISION TREES

Ex: Simple decision tree for predicting survival in Titanic



For classifying a new test point:

Start at the root of the tree.

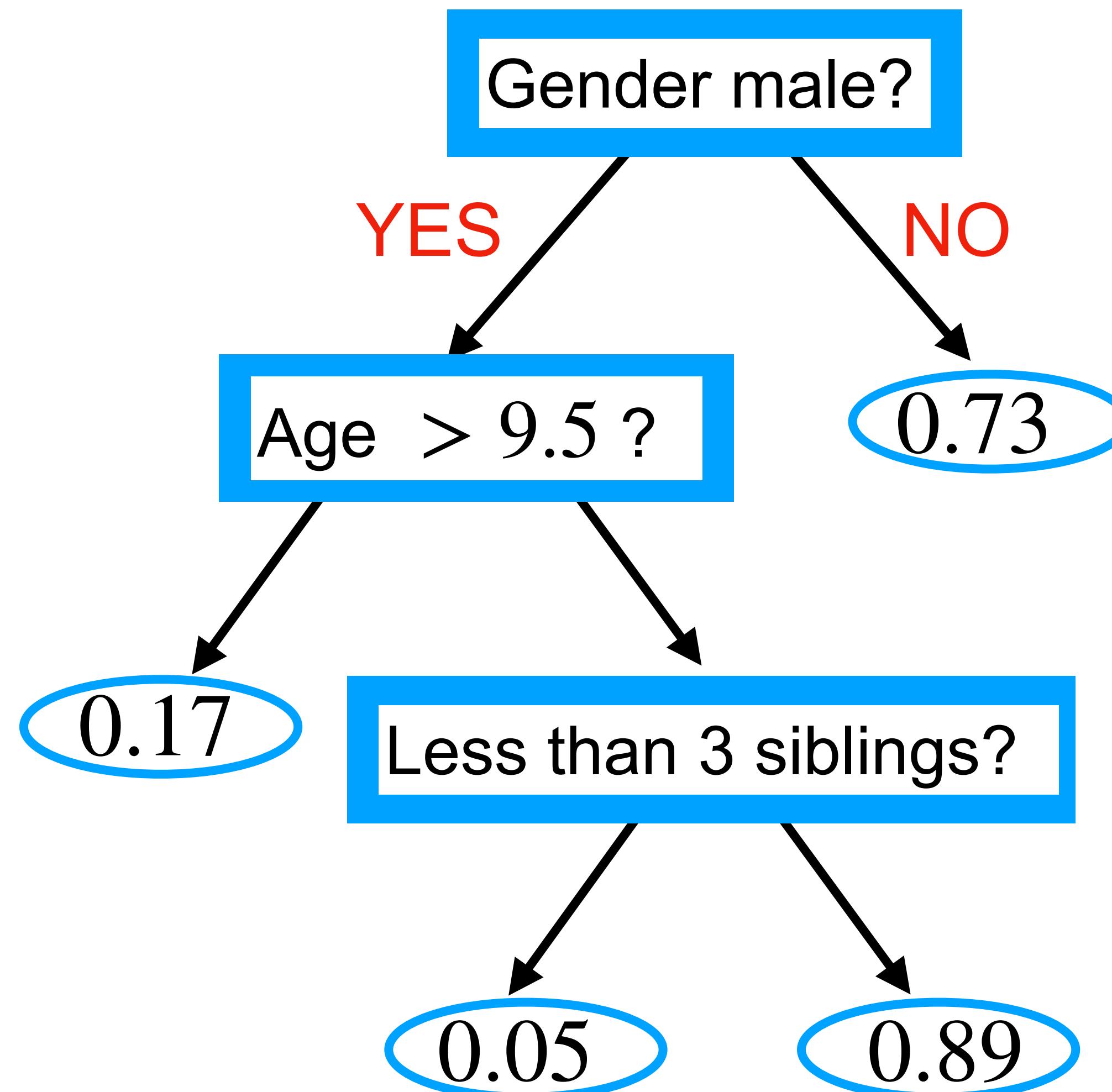
Traverse the tree until a terminal node.

Classify with the majority class.

(Number at a terminal node is the fraction of **survivors** in that group)

DECISION TREES

Ex: Simple decision tree for predicting survival in Titanic



For classifying a new test point:

Start at the root of the tree.

Traverse the tree until a terminal node.

Classify with the majority class.

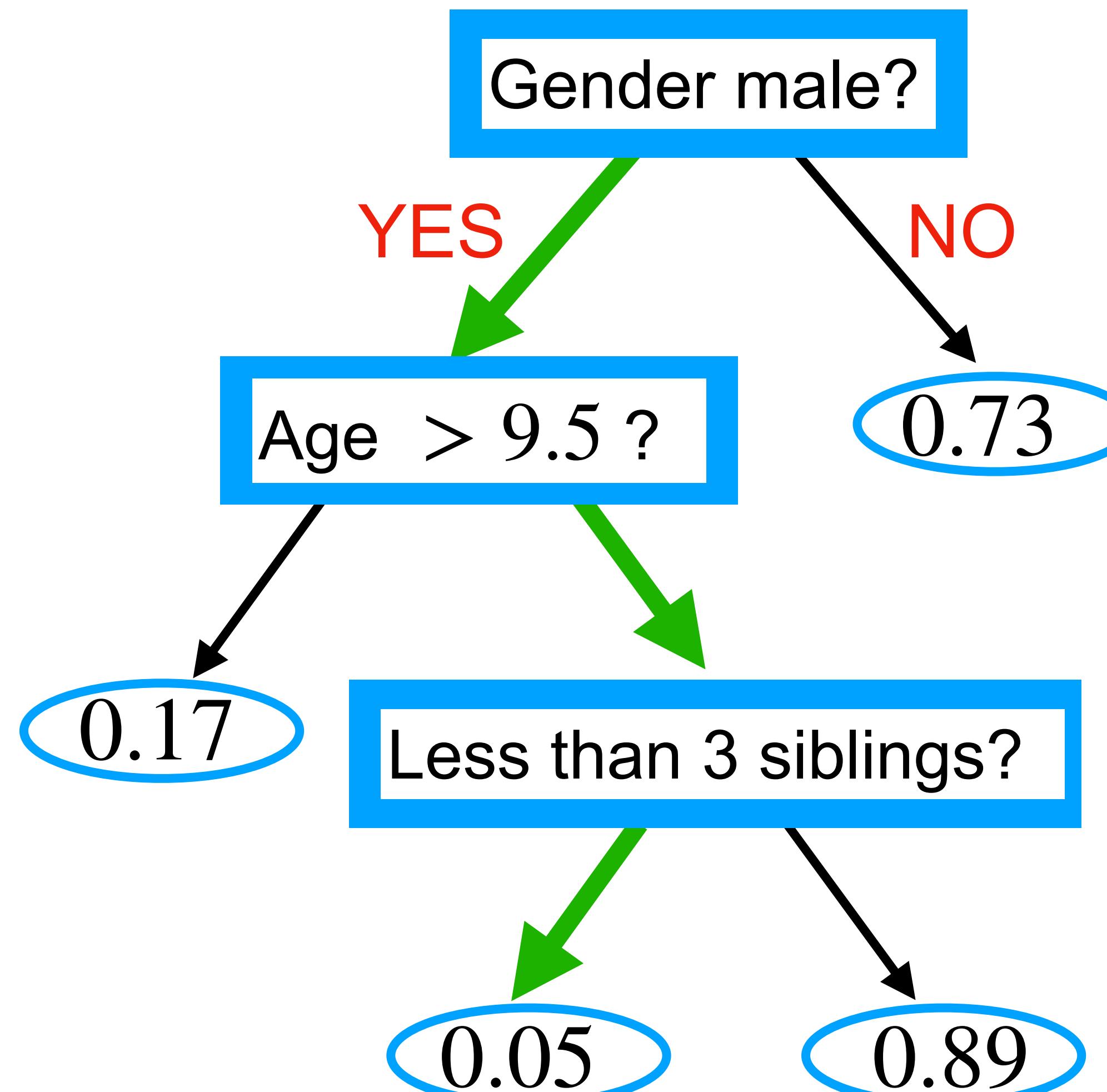
(Number at a terminal node is the fraction of **survivors** in that group)

Example:

Male, 8 years old, with one sibling?

DECISION TREES

Ex: Simple decision tree for predicting survival in Titanic



Majority class in group
is 'Not survived'.

For classifying a new test point:

Start at the root of the tree.

Traverse the tree until a terminal node.

Classify with the majority class.

(Number at a terminal node is the fraction of **survivors** in that group)

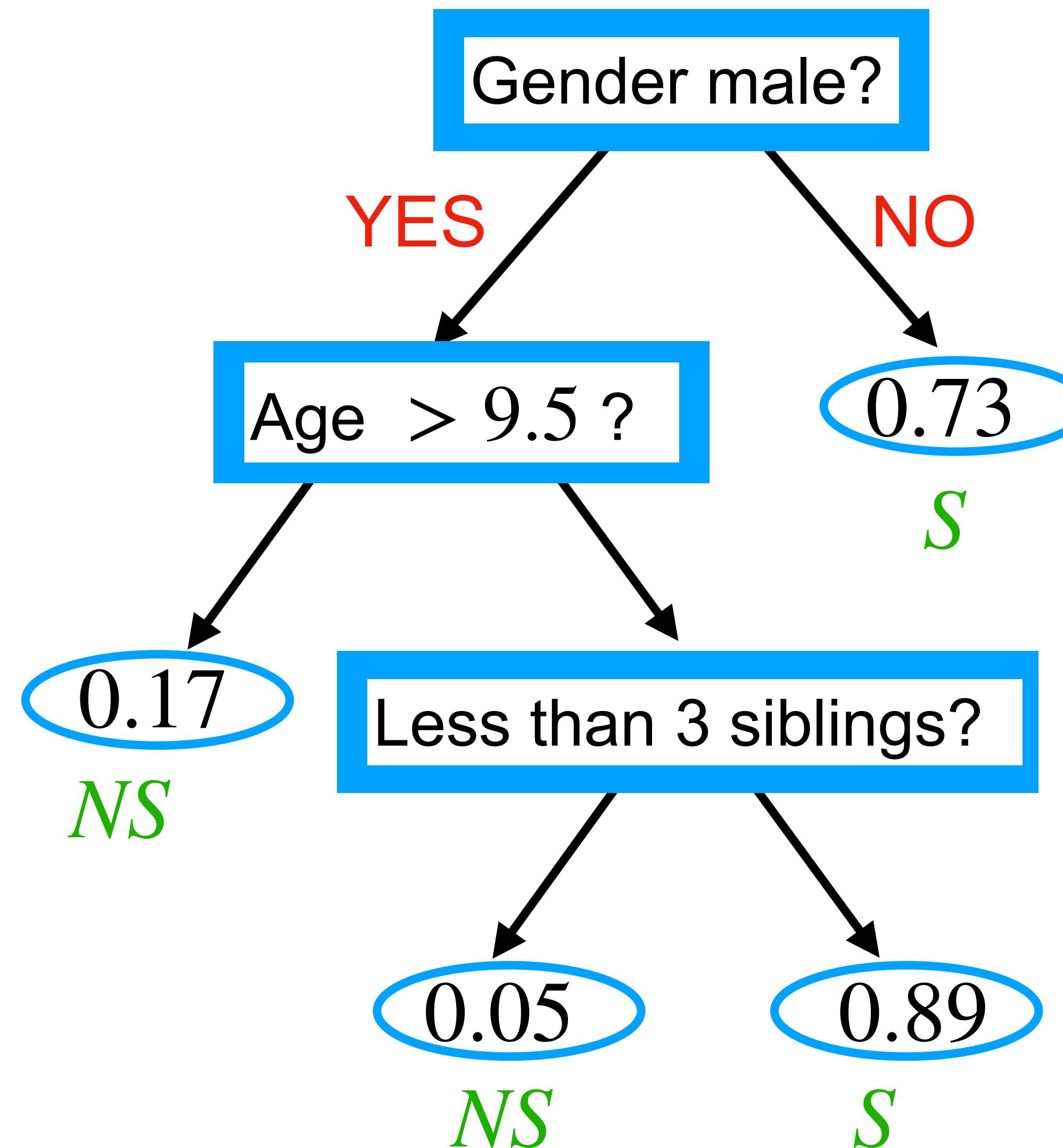
Example:

Male, 8 years old, with one sibling?

Not survived.

DECISION TREES

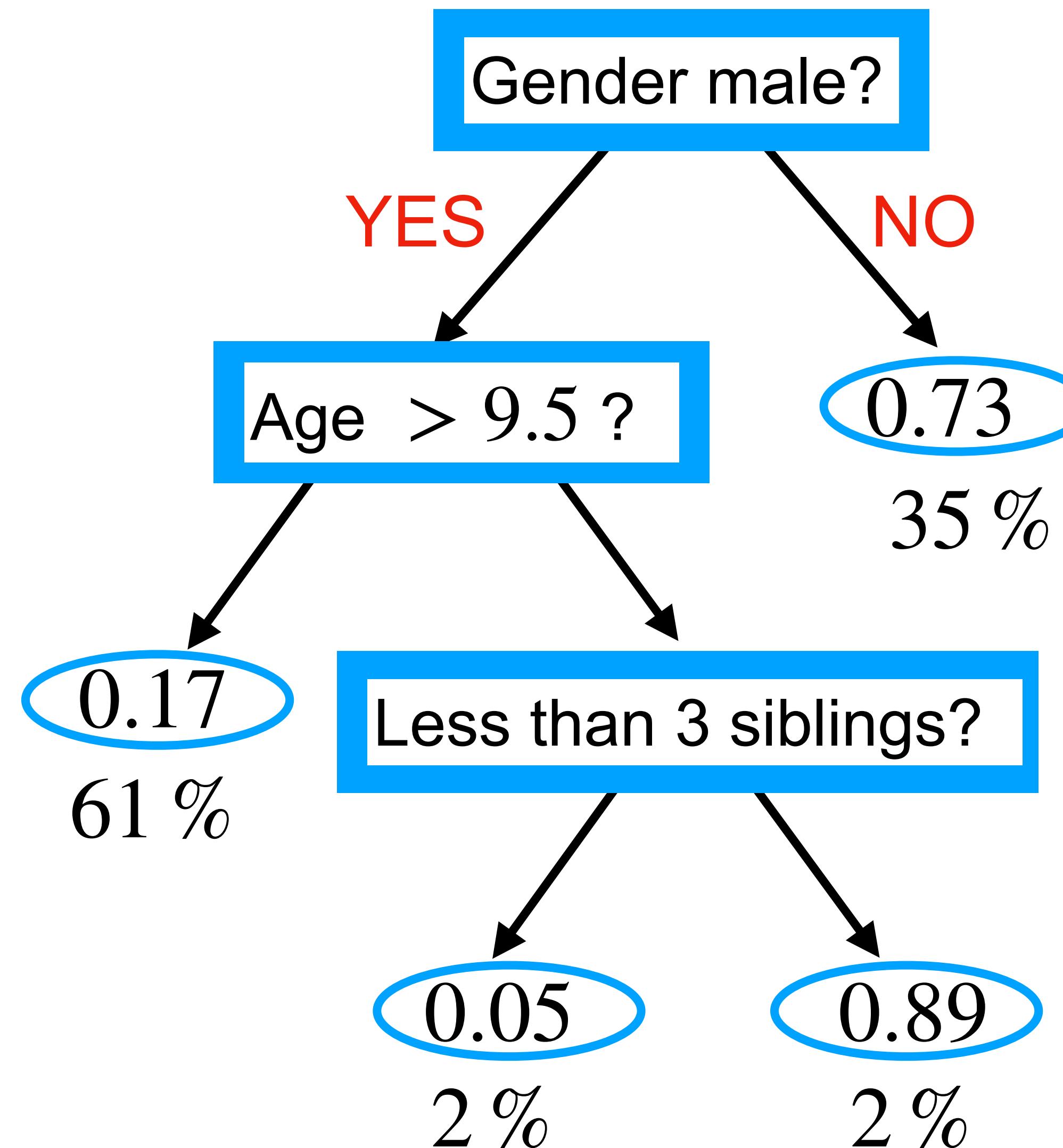
Ex: Simple decision tree for predicting survival in Titanic



Classification of each point falling into a terminal node would be as shown.
(S: Survived, NS: not survived)

DECISION TREES

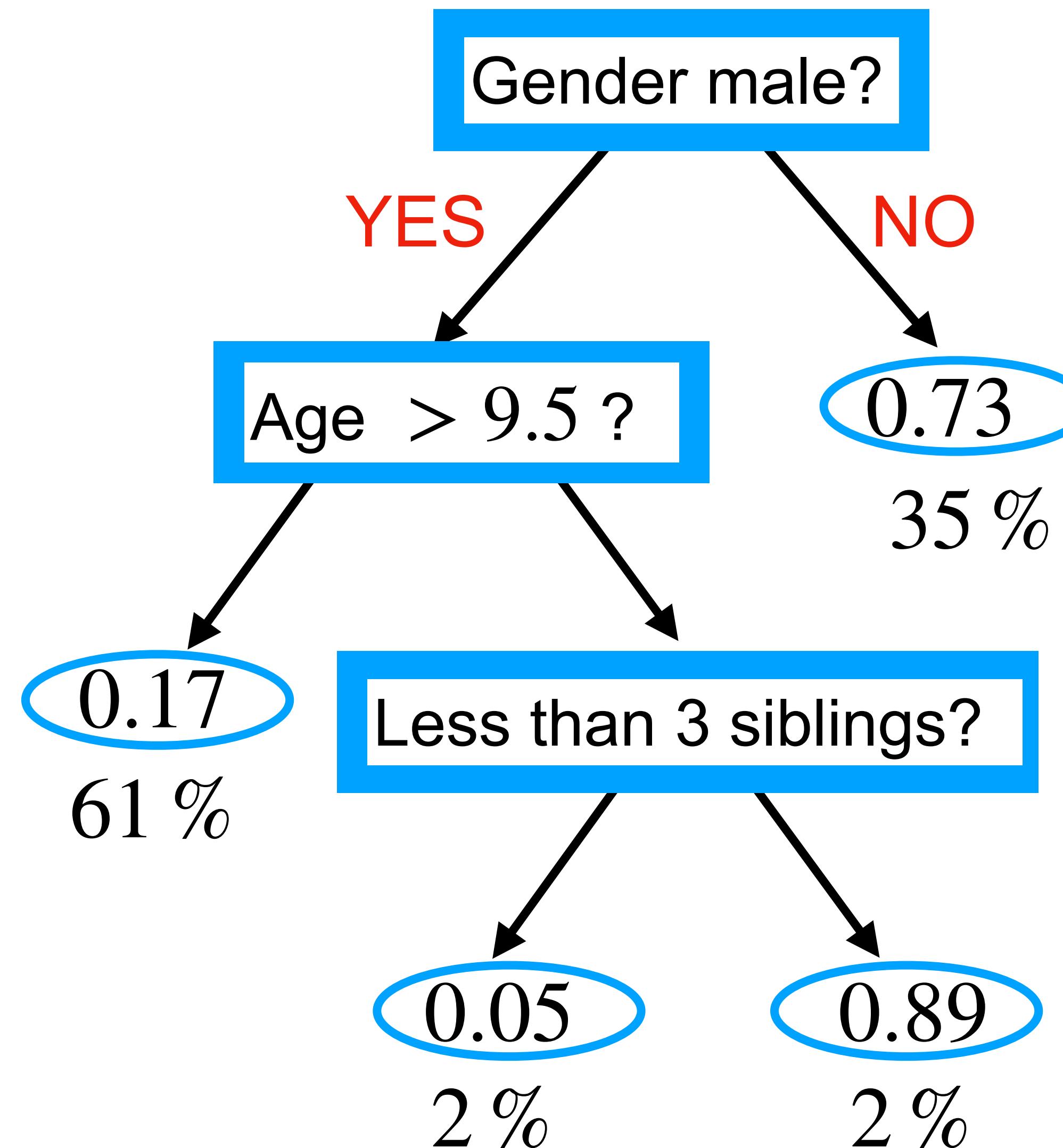
Ex: Simple decision tree for predicting survival in Titanic



Training accuracy of the decision tree?
(Percentage at a terminal node is the percentage of the group in Titanic.)

DECISION TREES

Ex: Simple decision tree for predicting survival in Titanic

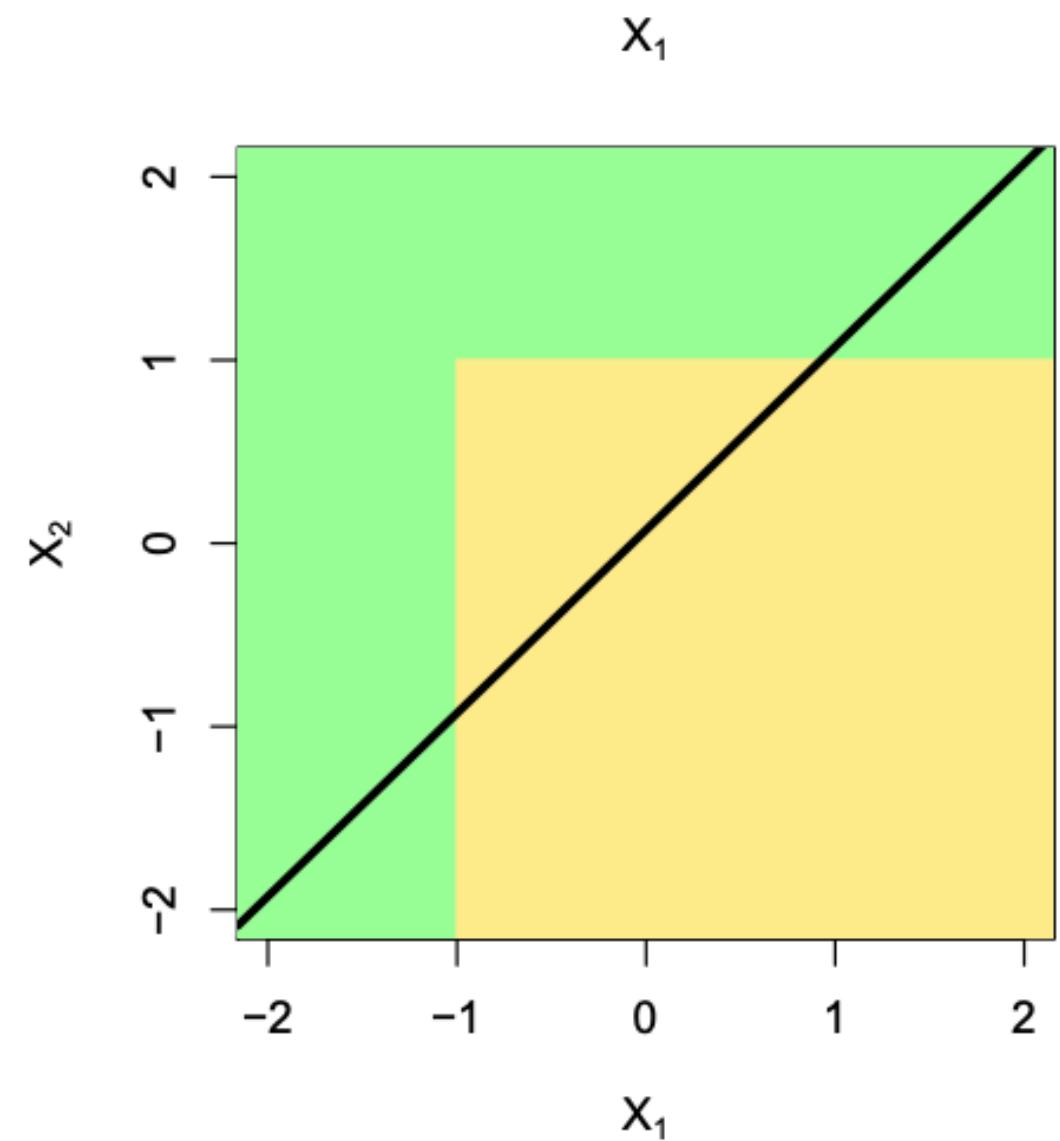


Training accuracy of the decision tree?
(Percentage at a terminal node is the percentage of the group in Titanic.)

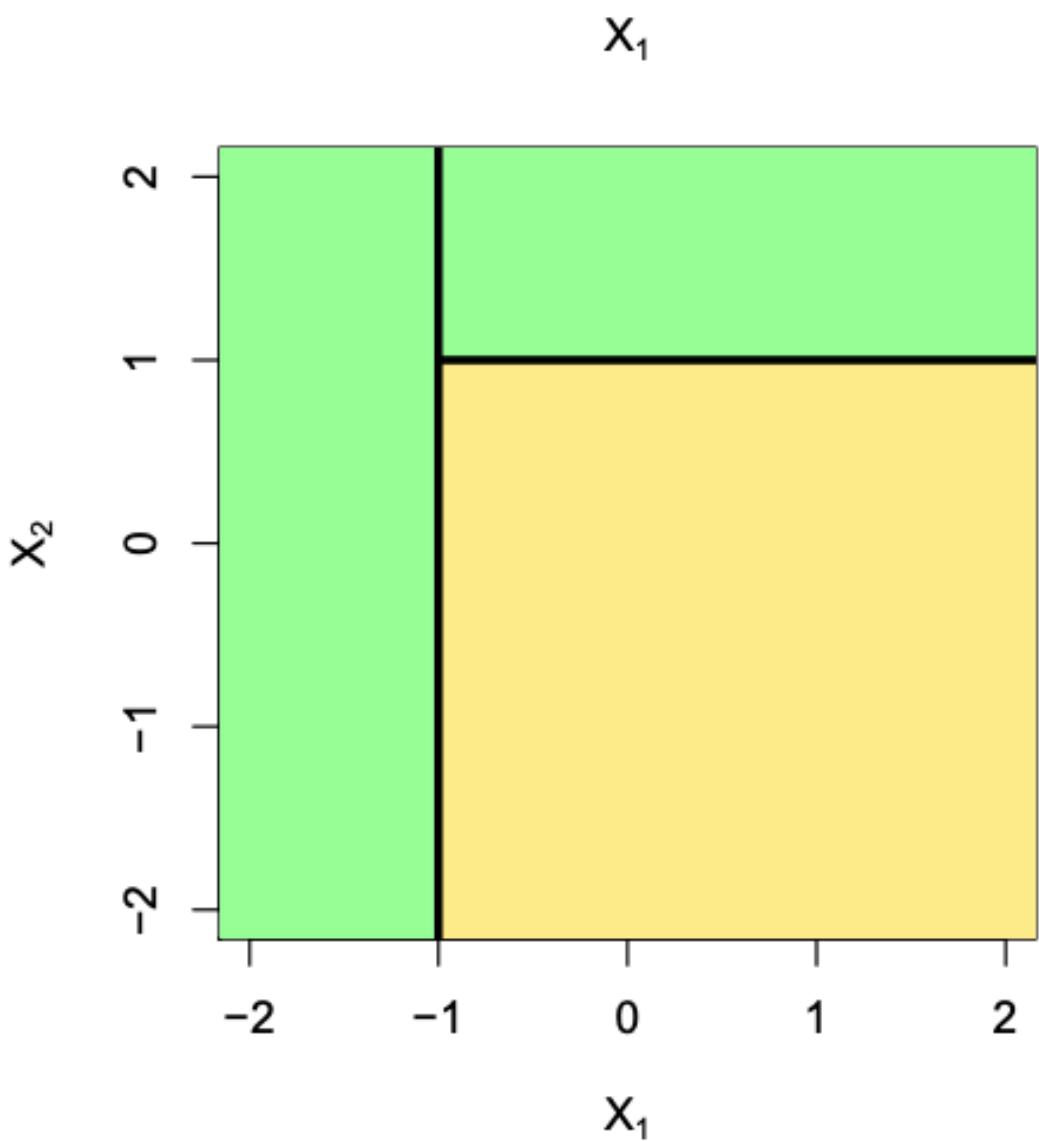
$$0.73 \times 0.35 + 0.83 \times 0.61 + 0.95 \times 0.02 + 0.89 \times 0.02$$

ADVANTAGES OF DECISION TREES

Non-linearity:



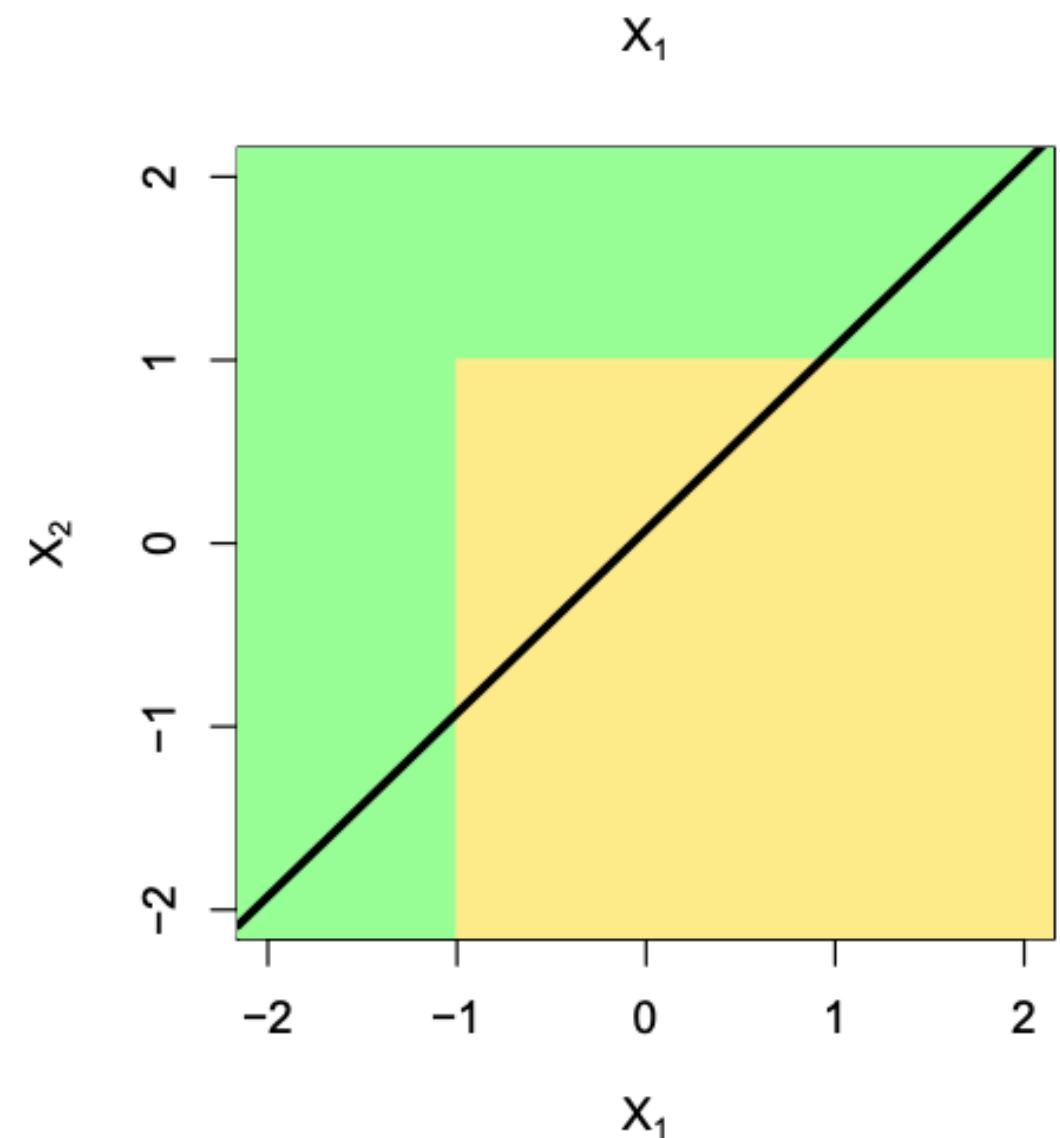
Linear model



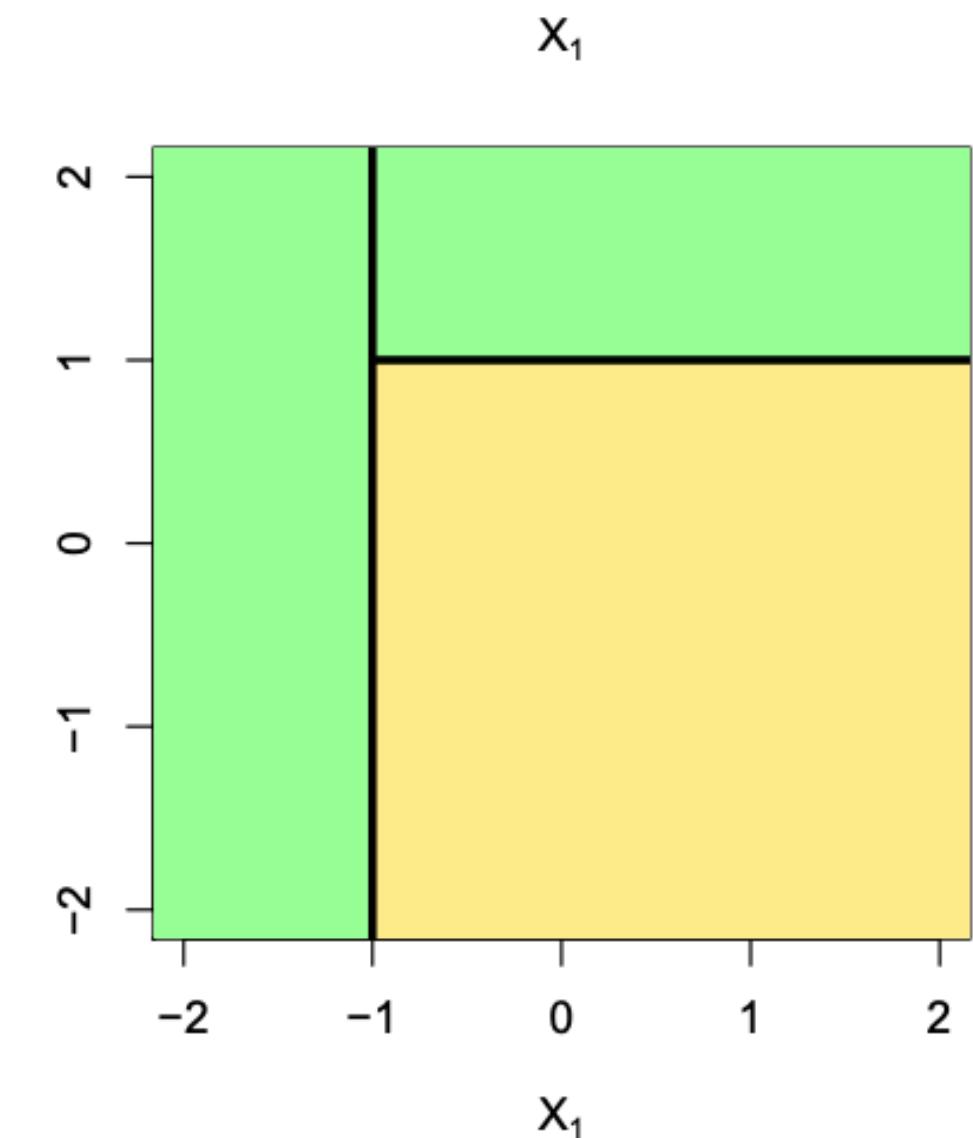
Decision tree

ADVANTAGES OF DECISION TREES

Non-linearity:



Linear model



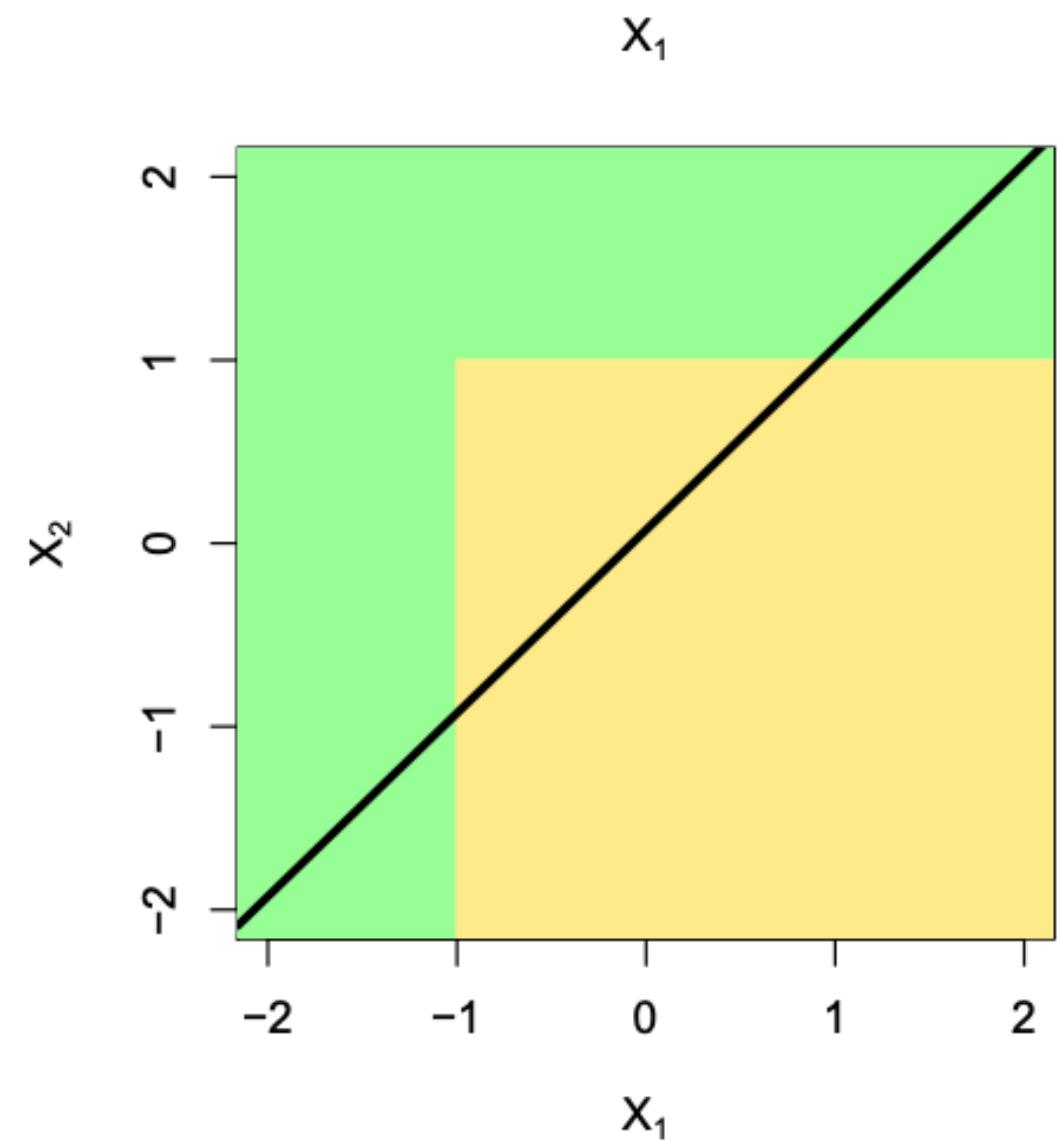
Decision tree

Support for categorical variables (hair=red):

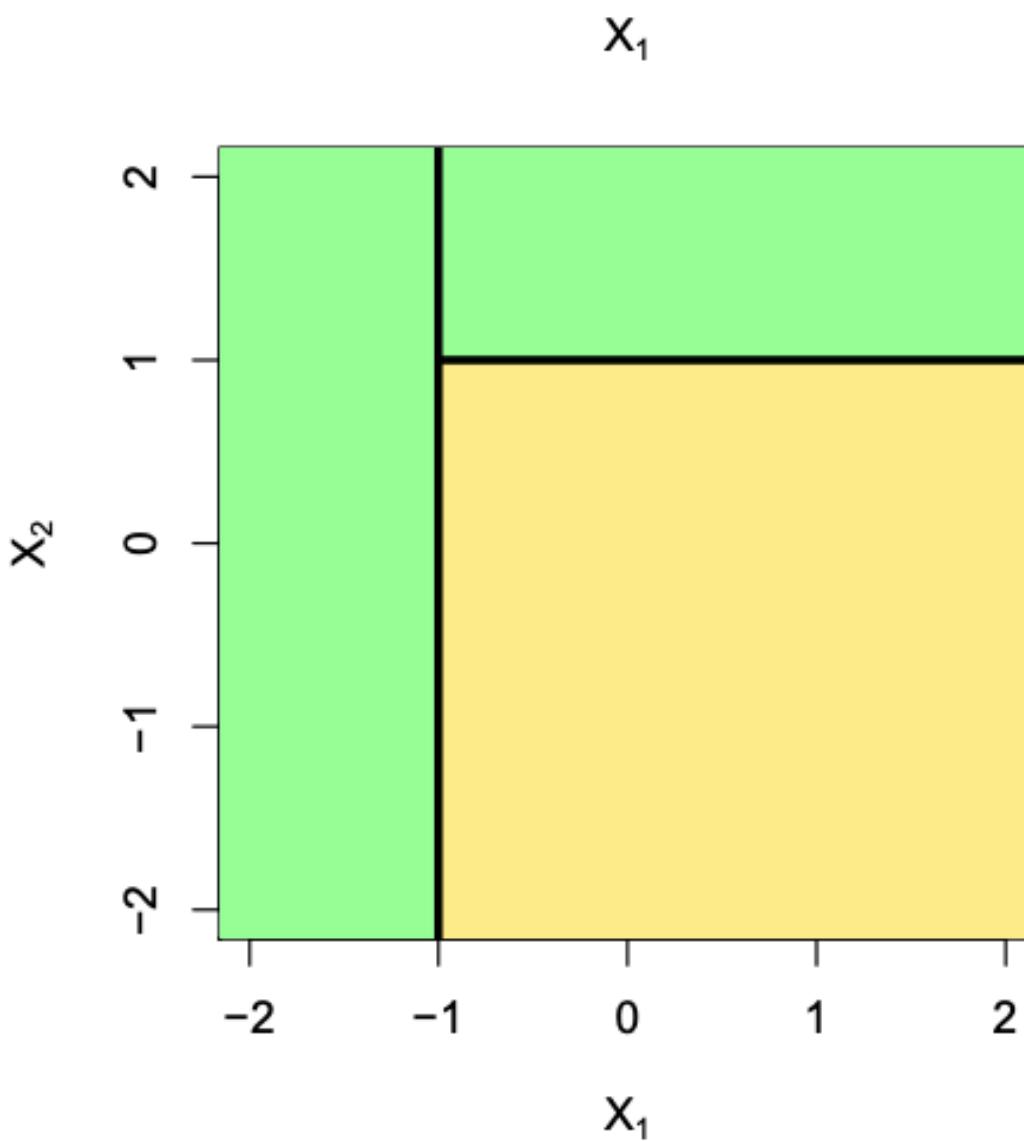
The splits of the tree can be done based on categorical values without the need for one hot encoding.

ADVANTAGES OF DECISION TREES

Non-linearity:



Linear model



Decision tree

Support for categorical variables (hair=red):

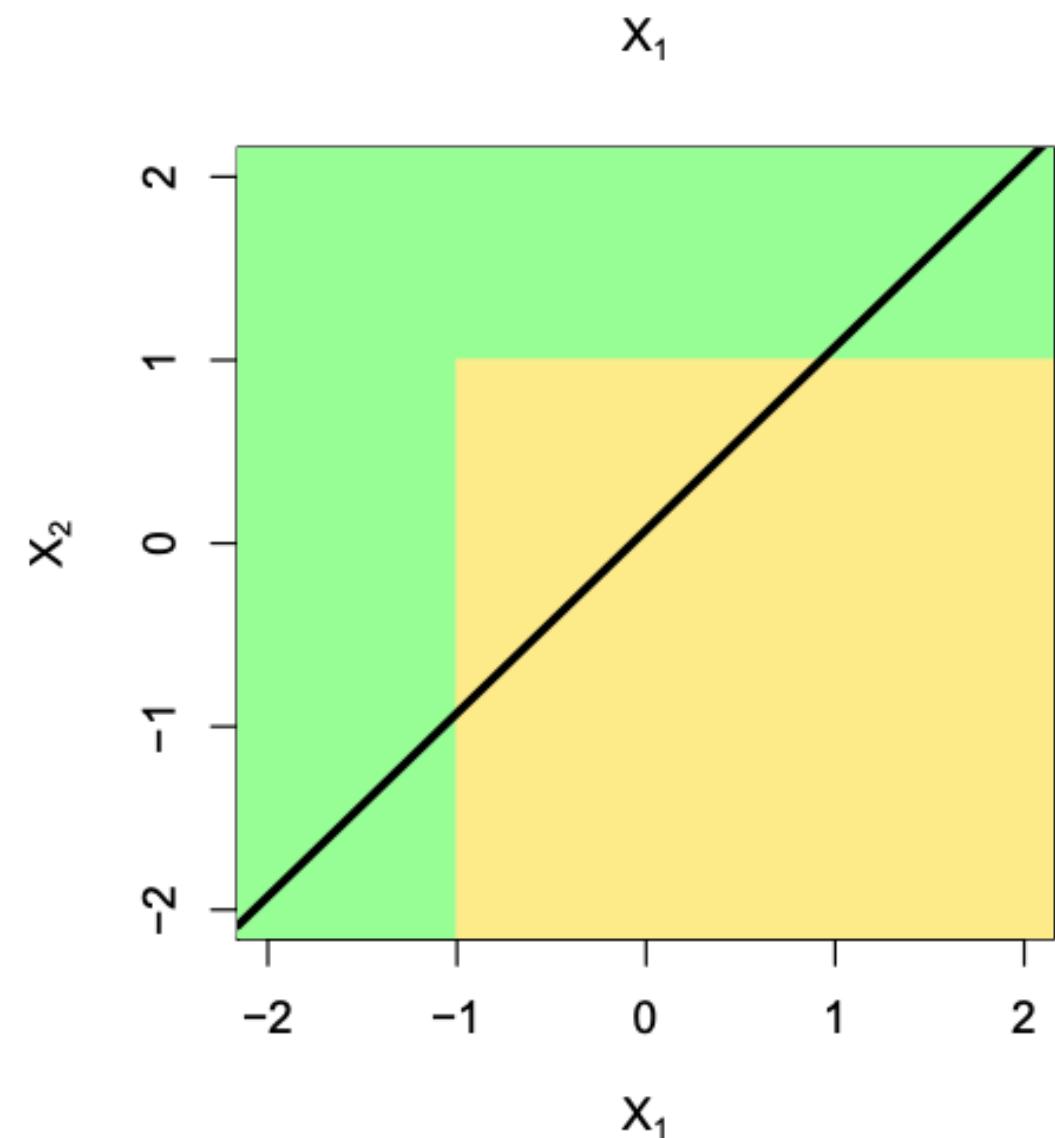
The splits of the tree can be done based on categorical values without the need for one hot encoding.

Interpretability:

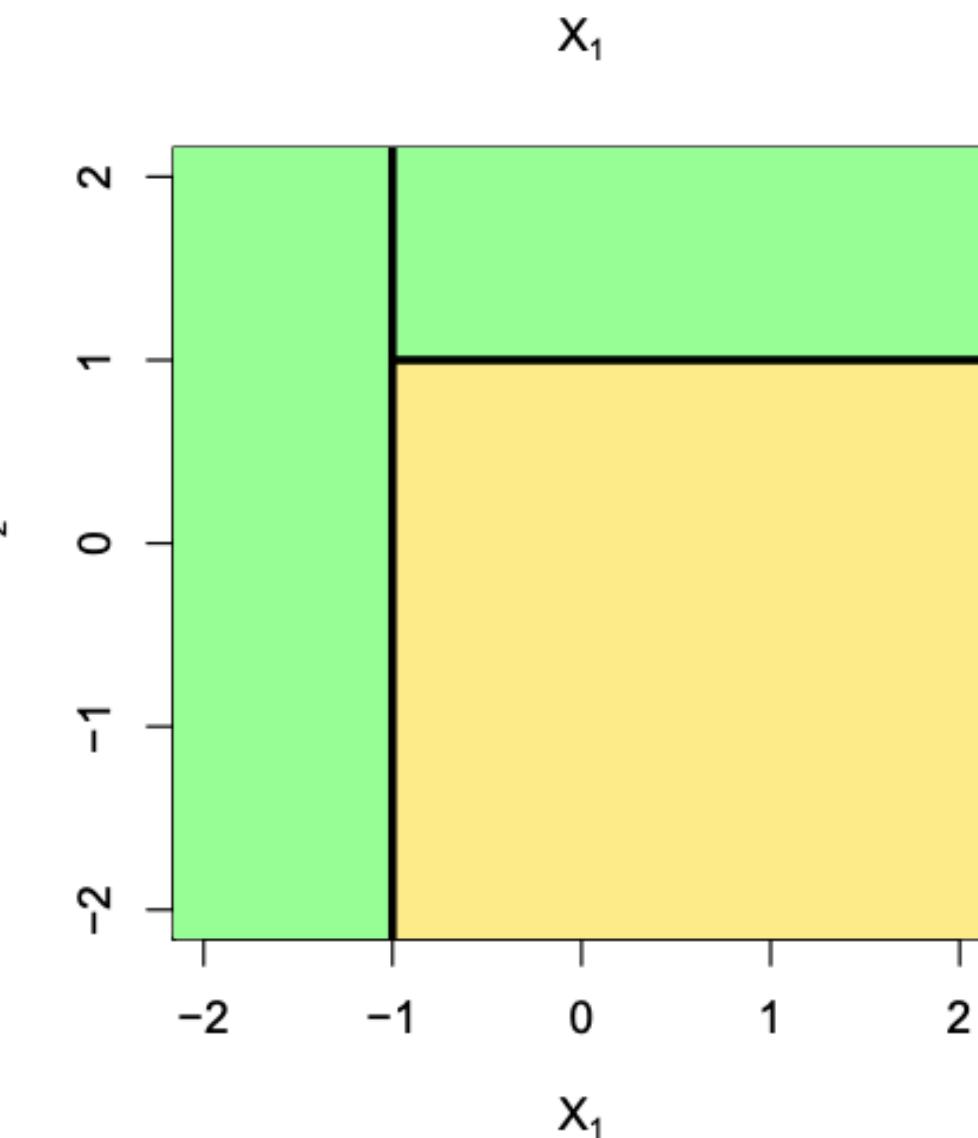
Features closest to root are important.

ADVANTAGES OF DECISION TREES

Non-linearity:



Linear model



Decision tree

Support for categorical variables (hair=red):

The splits of the tree can be done based on categorical values without the need for one hot encoding.

Interpretability:

Features closest to root are important.

Applies to regression:

Take the mean of the group at the terminal node.

HOW IS A DECISION TREE CONSTRUCTED?

Recursive binary splitting: Top-down, greedy.

- Each node: a binary predicate derived from a predictor..
- Discrete valued features: $\text{is } x_i = v?$ where v : a possible value
Numerical features: $\text{is } x_i \geq t?$ where t : a threshold

HOW IS A DECISION TREE CONSTRUCTED?

Recursive binary splitting: Top-down, greedy.

- Each node: a binary predicate derived from a predictor..
- Discrete valued features: $\text{is } x_i = v?$ where v : a possible value
Numerical features: $\text{is } x_i \geq t?$ where t : a threshold

Important question:

Which predictor x_i and threshold t (or value v) to split at some node?

Try all predictors and all possible thresholds (basic decision tree).

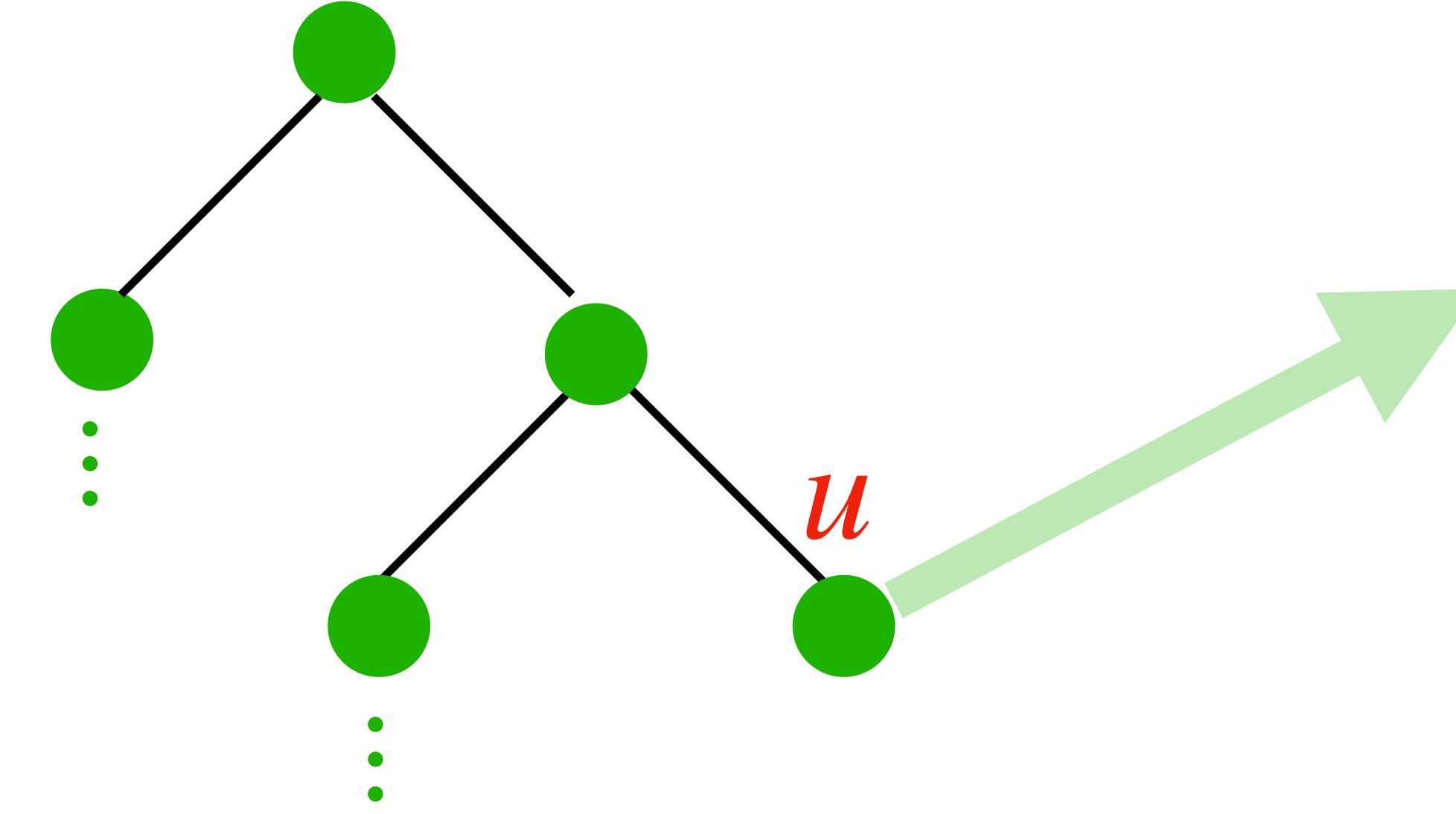
Find the ‘**best**’ one.

HOW IS A DECISION TREE CONSTRUCTED?

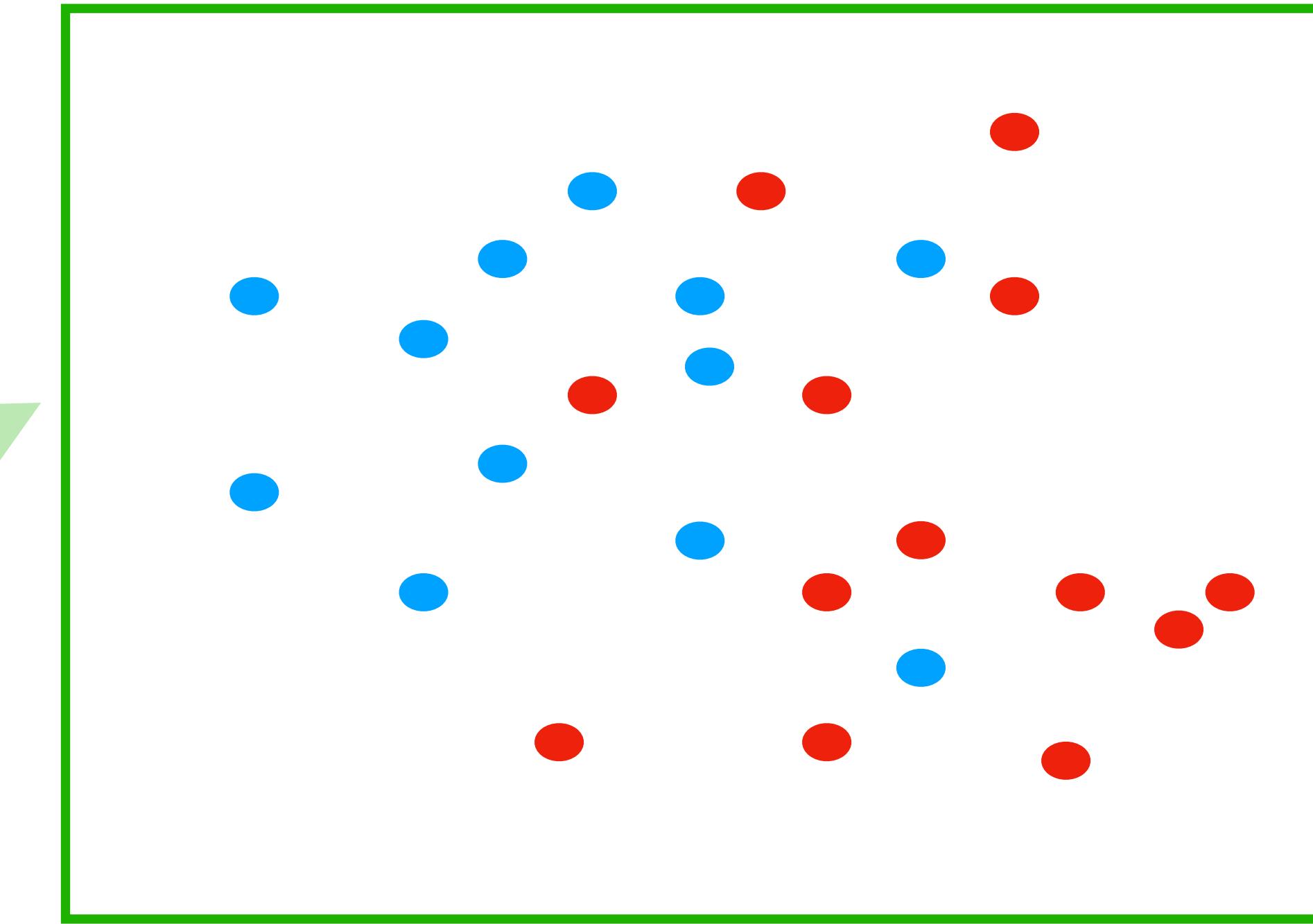
Example: Two classes: Red or blue. Two predictors, x_1 : numerical, x_2 : categorical.

HOW IS A DECISION TREE CONSTRUCTED?

Example: Two classes: Red or blue. Two predictors, x_1 : numerical, x_2 : categorical.



Decision tree constructed up to now
and we are trying to do split at node u



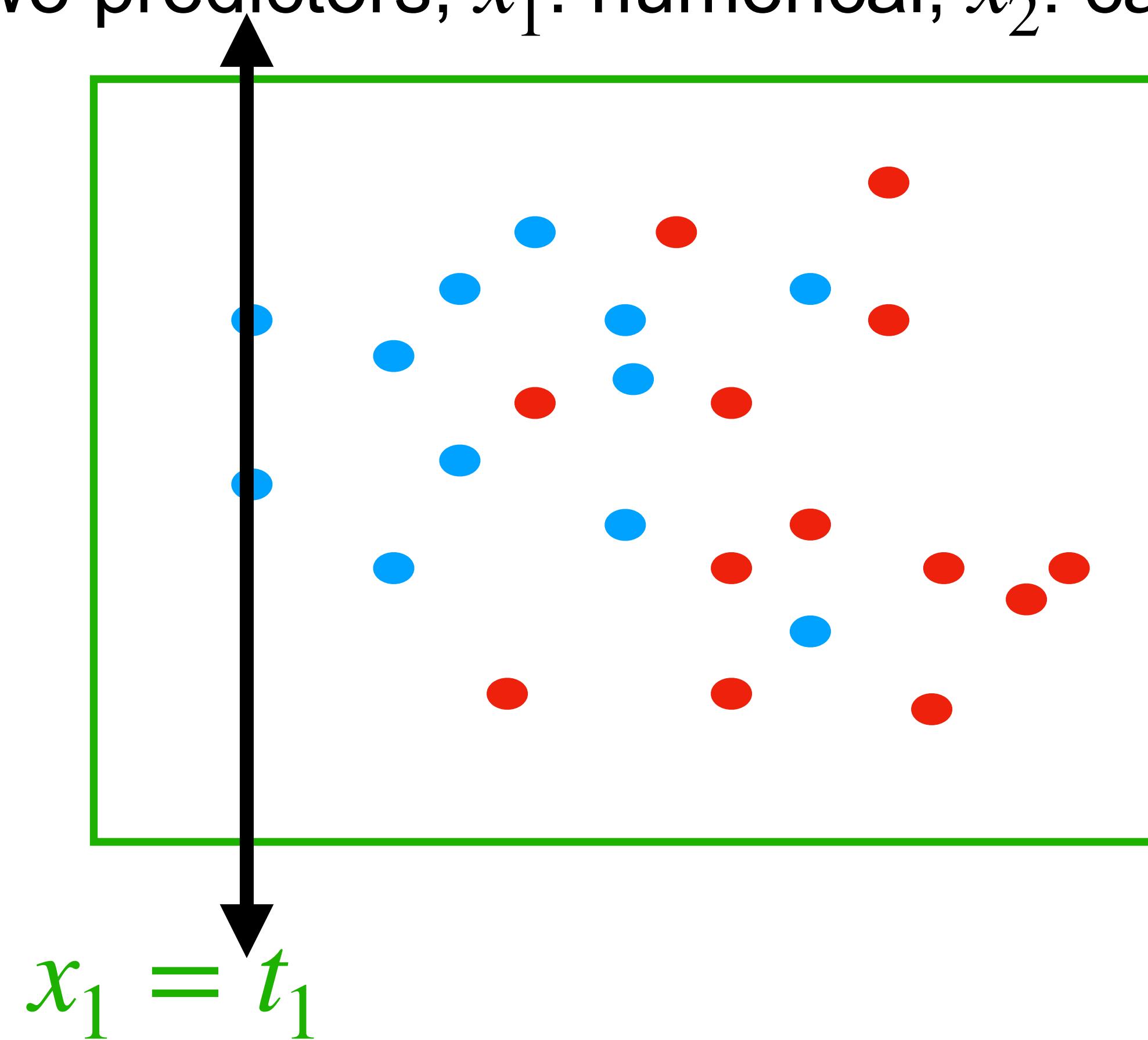
Training points at node u

HOW IS A DECISION TREE CONSTRUCTED?

Example: Two classes: Red or blue. Two predictors, x_1 : numerical, x_2 : categorical.

Boolean predicate at u ?

Try all possible thresholds
based on predictor x_1 and
record ‘quality of split’ for each.



Example thresholds:

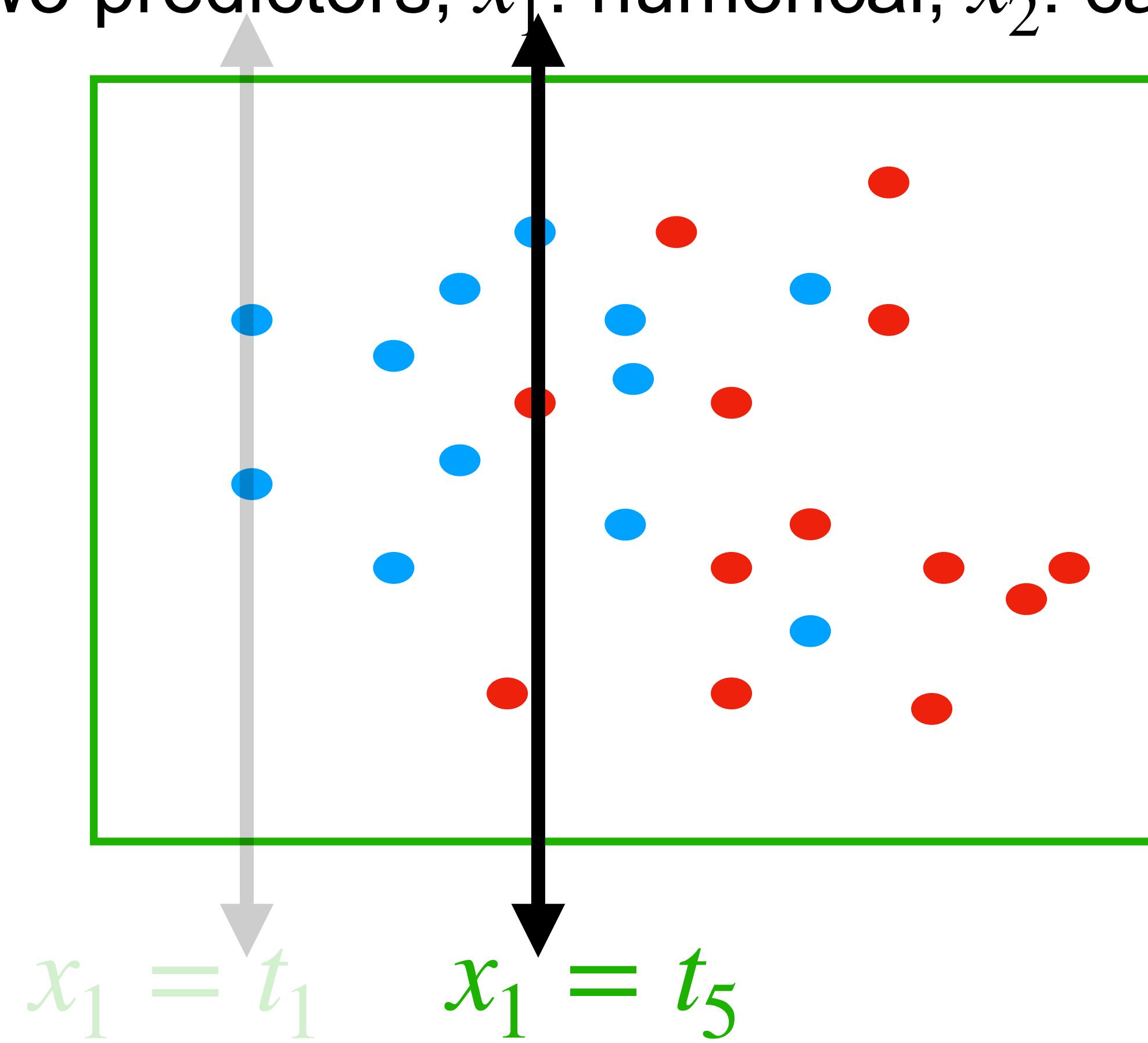
$x_1 \leq t_1$ vs $x_1 > t_1$ OR

HOW IS A DECISION TREE CONSTRUCTED?

Example: Two classes: Red or blue. Two predictors, x_1 : numerical, x_2 : categorical.

Boolean predicate at u ?

Try all possible thresholds
based on predictor x_1 and
record ‘quality of split’ for each.



Example thresholds:

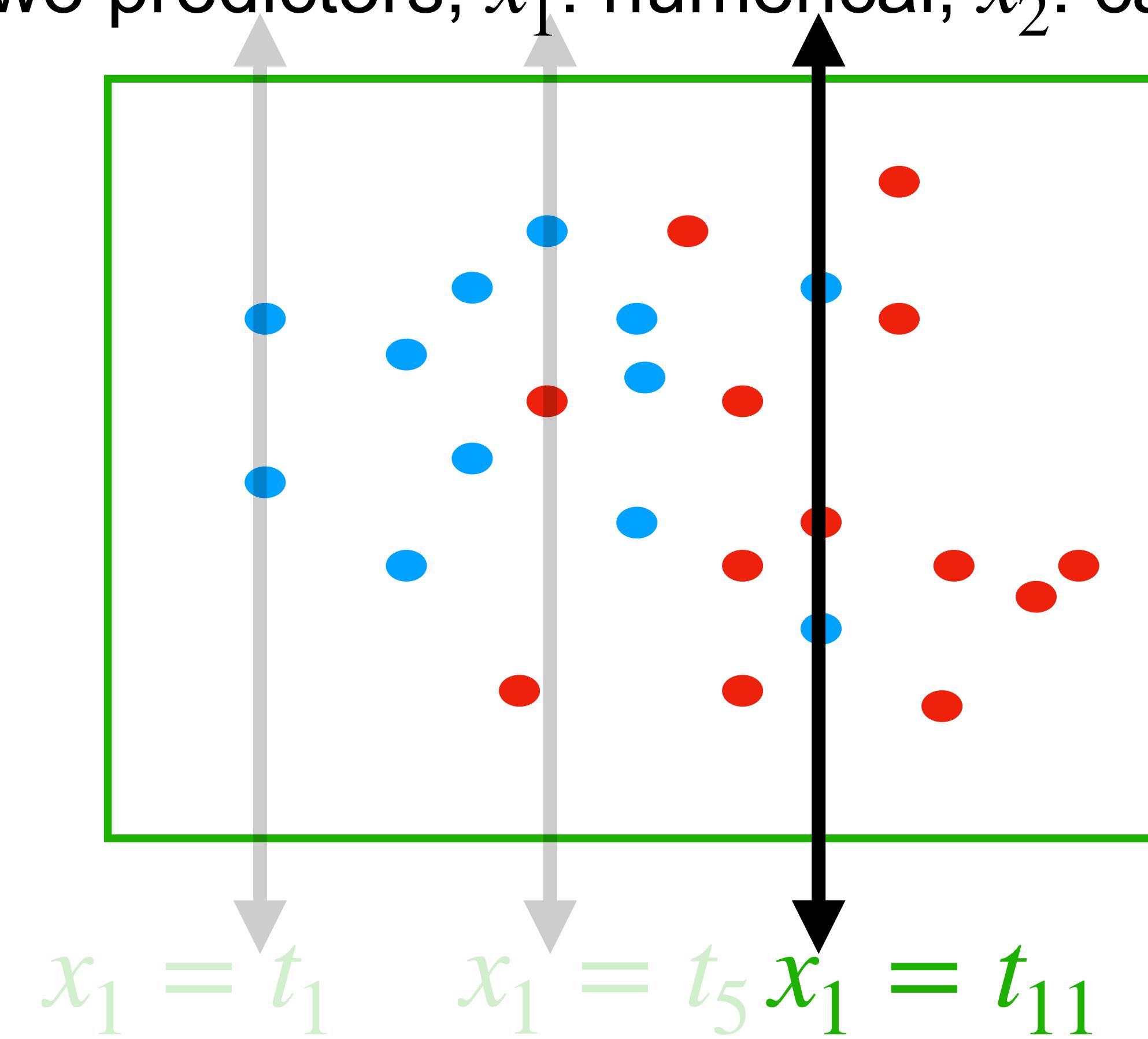
$x_1 \leq t_1$ vs $x_1 > t_1$ OR $x_1 \leq t_5$ vs $x_1 > t_5$ OR

HOW IS A DECISION TREE CONSTRUCTED?

Example: Two classes: Red or blue. Two predictors, x_1 : numerical, x_2 : categorical.

Boolean predicate at u ?

Try all possible thresholds
based on predictor x_1 and
record ‘quality of split’ for each.



Example thresholds:

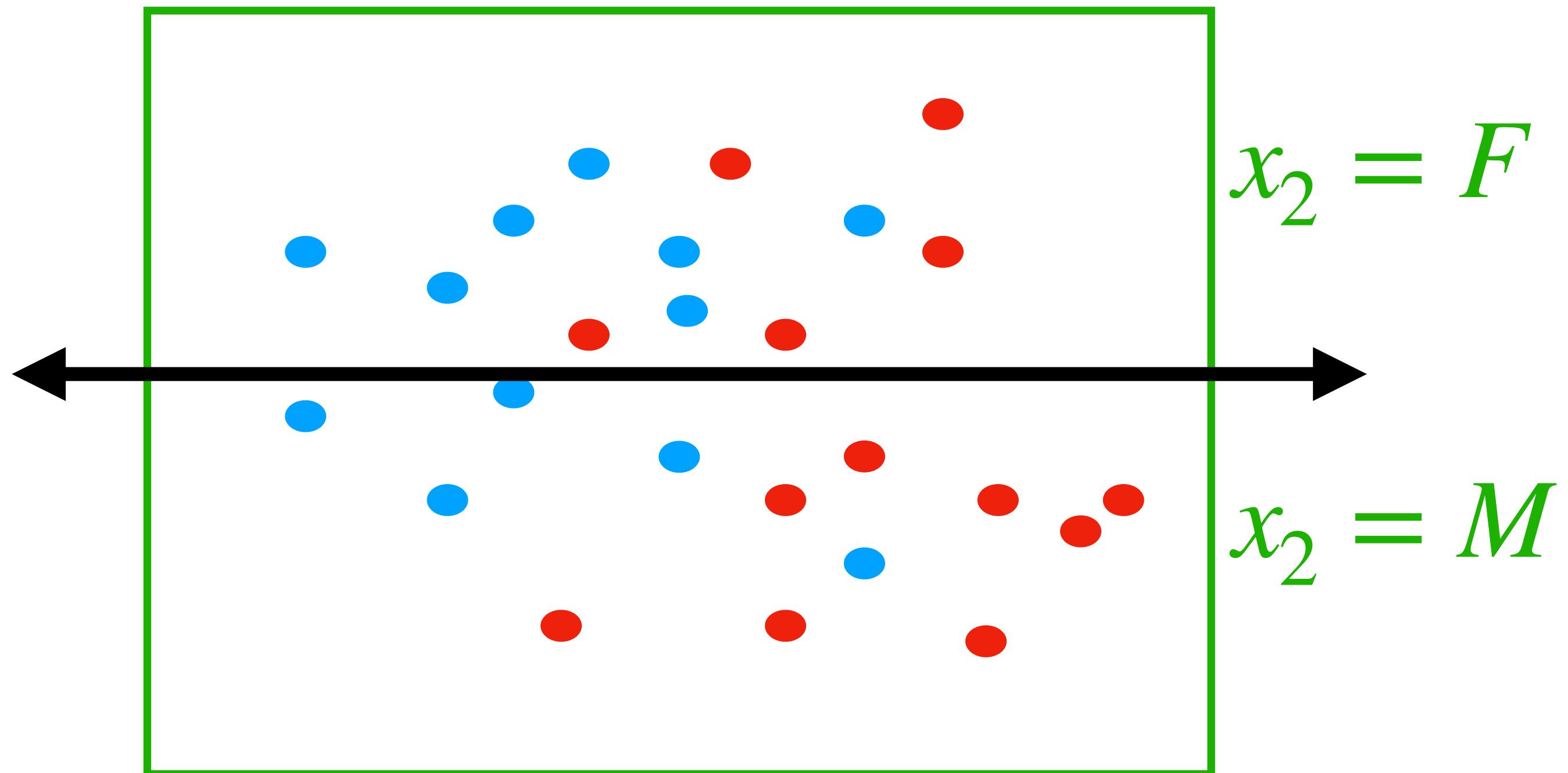
$x_1 \leq t_1$ vs $x_1 > t_1$ OR $x_1 \leq t_5$ vs $x_1 > t_5$ OR $x_1 \leq t_{11}$ vs $x_1 > t_{11}$...

HOW IS A DECISION TREE CONSTRUCTED?

Example: Two classes: Red or blue. Two predictors, x_1 : numerical, x_2 : categorical.

Boolean predicate at u ?

Try all possible values based
on predictor x_2 and record
'quality of split' for each.



HOW IS A DECISION TREE CONSTRUCTED?

How to measure quality of split?

Alternative 1:

$$Gini\ index = G = \sum_{j=1}^K p_j(1 - p_j) = 1 - \sum_{j=1}^K p_j^2$$

K : number of classes

p_j : proportion of training
data points from class j .

HOW IS A DECISION TREE CONSTRUCTED?

How to measure quality of split?

Alternative 1:

$$\text{Gini index} = G = \sum_{j=1}^K p_j(1 - p_j) = 1 - \sum_{j=1}^K p_j^2$$

K : number of classes

p_j : proportion of training data points from class j .

Intuitively:

All proportions near 0 or 1 \Rightarrow Small Gini index, large purity (one class dominant).

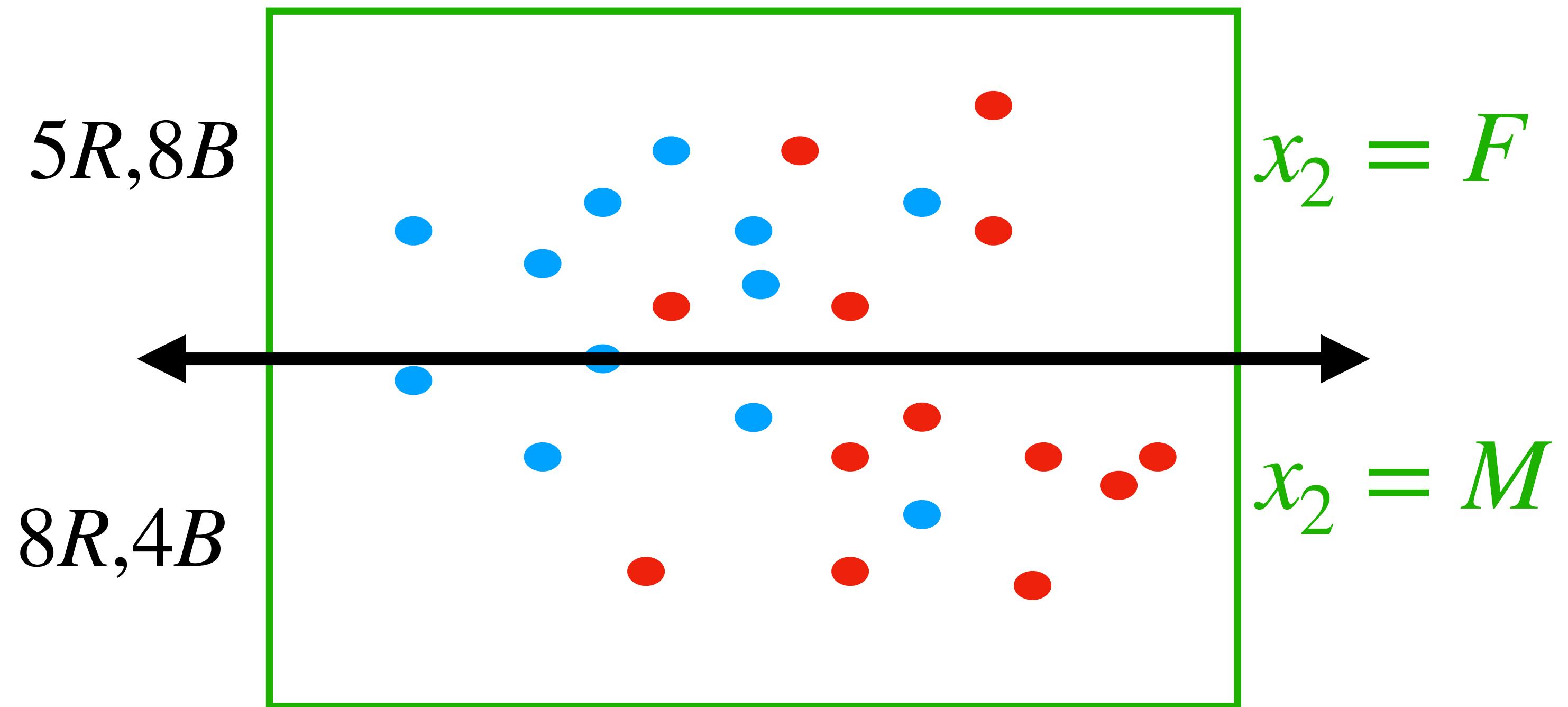
Good split: Minimizes **Gini index of the split**

Weighted-sum of Gini indices of the two partitions

↓
Proportion of data points in the partition.

HOW IS A DECISION TREE CONSTRUCTED?

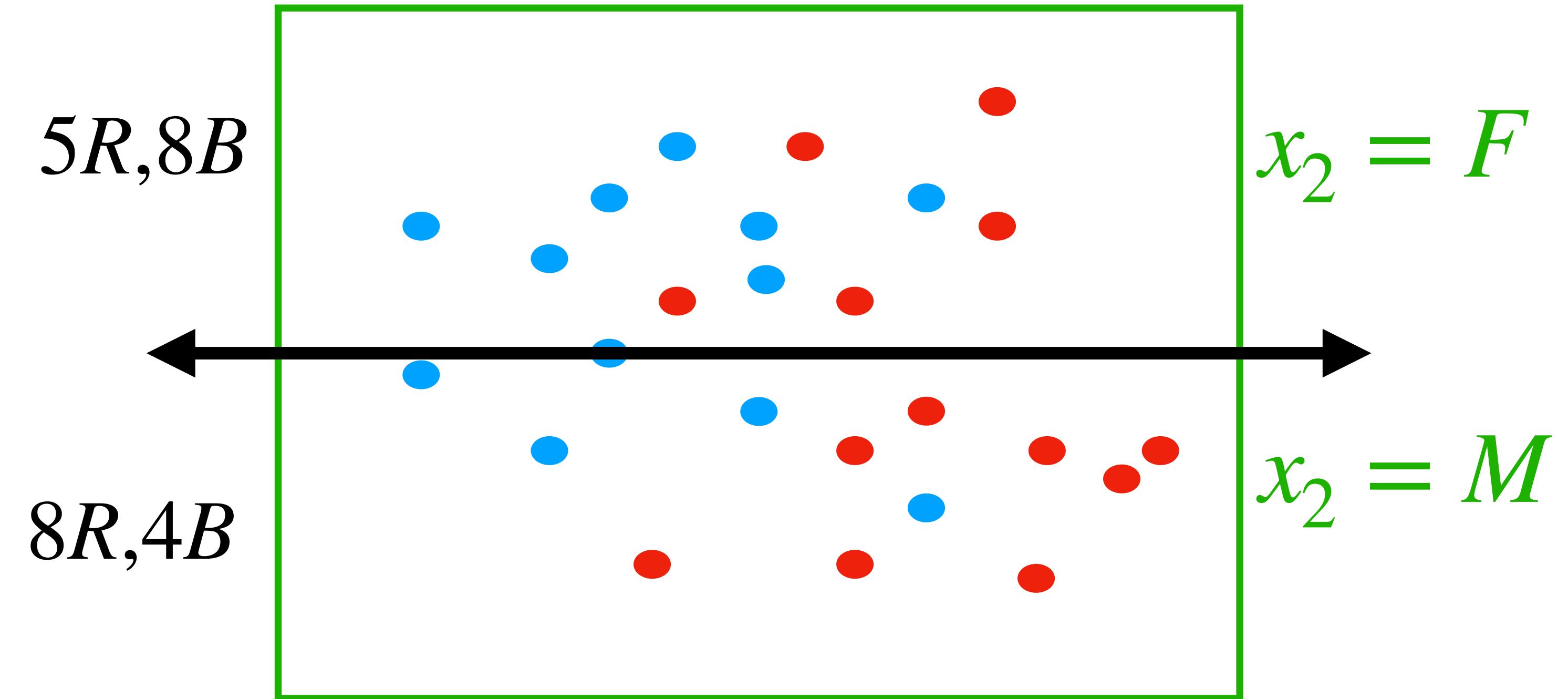
How to measure quality of split?



Example: What is the Gini index of the split?

HOW IS A DECISION TREE CONSTRUCTED?

How to measure quality of split?



Example: What is the Gini index of the split?

$$\frac{13}{25} \left[1 - \left(\left(\frac{5}{13} \right)^2 + \left(\frac{8}{13} \right)^2 \right) \right] + \frac{12}{25} \left[1 - \left(\left(\frac{8}{12} \right)^2 + \left(\frac{4}{12} \right)^2 \right) \right]$$