Lecture 1

Introduction, Calculus

History of Data Science, Winter 2022 @ UC San Diego

Suraj Rampure

Hi (again)! 👋

Suraj Rampure (call me Suraj, pronounced "soo-rudge")

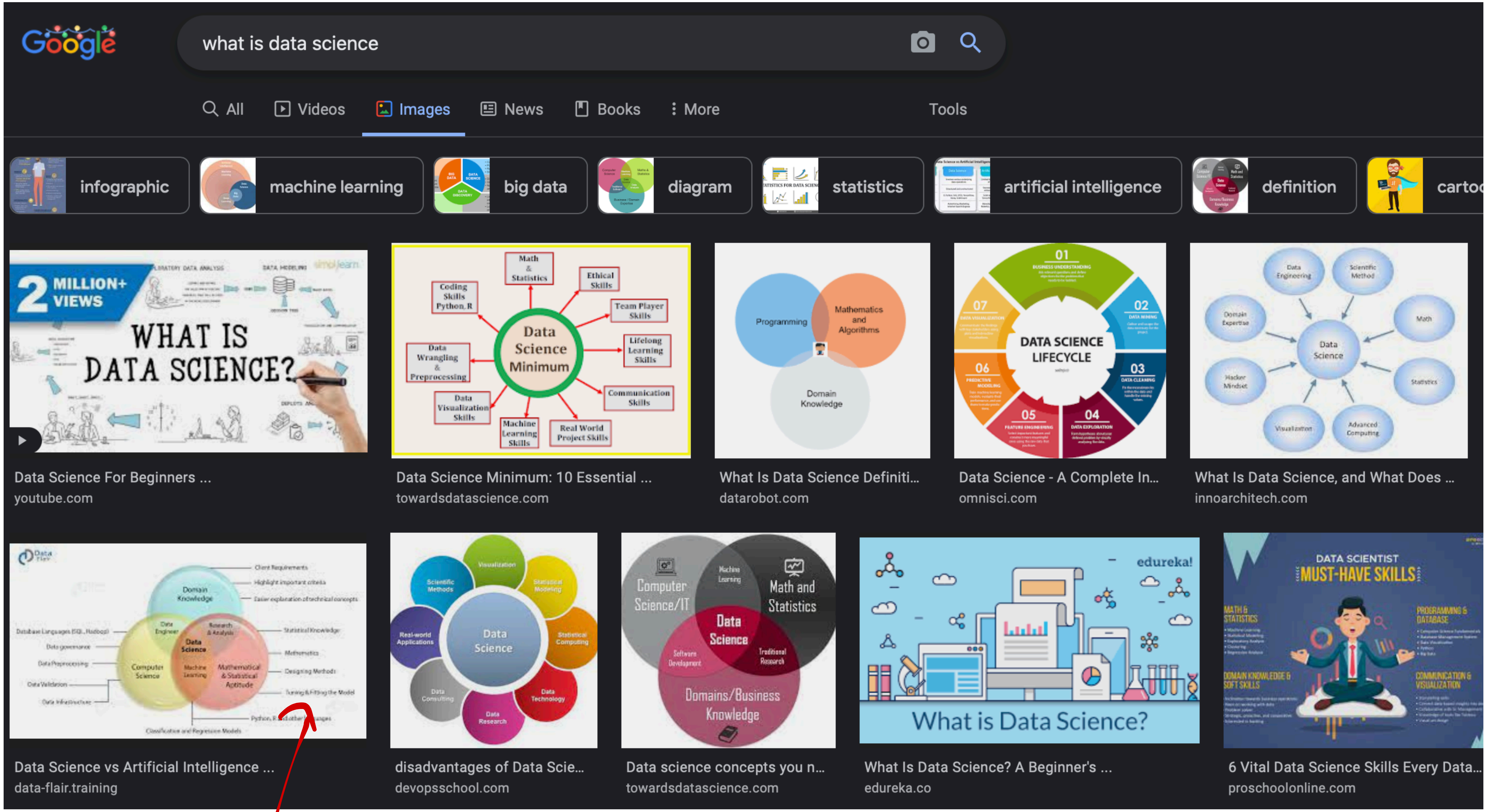
- Originally from Windsor, ON, Canada 🇨🇦.
- BS ('20) and MS ('21) in EECS from UC Berkeley 🐻.
- Second quarter teaching at UCSD 🌴.
    - Teaching DSC 10 this quarter; taught 10 and 40A last quarter.
- Outside the classroom 🧑‍🏫: watching basketball, traveling, eating, watching TikTok, etc.

Agenda

already covered

- Who am I?
- What is this course about?
- Why is calculus relevant?
- Key figures and ideas in math in Ancient Greece.

What is this course about?



what is data science

Data Science For Beginners ...
youtube.com

Data Science Minimum: 10 Essential ...
towardsdatascience.com

What Is Data Science Definiti...
datarobot.com

Data Science - A Complete In...
omnisci.com

What Is Data Science, and What Does ...
innoarchitech.com

Data Science vs Artificial Intelligence ...
data-flair.training

disadvantages of Data Scie...
devopsschool.com

Data science concepts you n...
towardsdatascience.com

What Is Data Science? A Beginner's ...
edureka.co

6 Vital Data Science Skills Every Data...
proschoolonline.com

Venn diagrams

Origins of the term "data science"

- John Tukey, the originator of many ideas in modern data science, wrote "The Future of Data Analysis" in 1962[1,2], in which he said:

> *"For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt… All in all, I have come to feel that my central interest is in data analysis, which I take to include,* ***among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data****"*

- In 1974[3], Peter Naur defined "data science" as being:

Turing award

> *"The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences."*

[1] Tukey, "The Future of Data Analysis"
[2] Donoho, "50 years of Data Science"
[3] Naur, "Concise Survey of Computer Methods"

What is “data science”?

- Nowadays, everyone seems to have their own definition of data science.
  - In DSC 10, we said “data science is **about drawing useful conclusions from data using computation**”.
- Regardless of how you define it, it is clear that data science relies on tools from a variety of disciplines (computer science, mathematics, statistics, cognitive science, etc).
- The term “data science” is relatively new, but the foundations of the field are centuries old.

A timeline spanning centuries

200 BC
0
2022

- **200s BC:** Archimedes approximates the value of π using the method of exhaustion, an early precursor to limits (and, hence, calculus).
- **Late 1600s:** Newton and Leibniz establish many of the ideas central to calculus.

didn't have $f(x)$

- **Late 1700s:** Playfair develops line graphs, pie charts, and bar charts[1].
- **Early 1800s:** Gauss and Legendre develop the method of least squares.
- **Late 1800s**: Galton, Darwin's cousin, develops the theory of regression and correlation.
- **1943:** ENIAC, the first programmable computer, was built by the US during WW2.
- **Also 1943**: McCulloch and Pitts describe the first neural network based on a model of neurons[2].

neuroscience

[1] Playfair, William. Encyclopedia of Mathematics.
[2] History: The 1940's to the 1970's

Why study history?

- We will study the origins of several key ideas in data science.
  - Only DSC 10 and Math 20AB will be assumed.
- When visiting an idea, we'll look at who contributed to its development and what their motivations were.
  - We'll **read** excerpts from original papers, work on problems that use these ideas, and discuss how these ideas play a role in modern data science.
- **In doing so, we'll develop an appreciation for the methods we are using today, and better understand how to use them[1].**

[1] https://rohanalexander.com/history_of_the_data_sciences.html

Logistics

Technology

- All content (lecture slides, recordings, readings, homeworks) will be posted on the **course website**, historyofdsc.com.
- We will use **Campuswire** for communication.
  - We encourage you to discuss the material and post homework questions there.
- Homeworks will be submitted to **Gradescope**.
  - If you weren't added to either Campuswire or Gradescope, send me an email.

Logistics

n unit $\leq 3n$ hours / week

- **Monday 6PM:** class meeting.
    - Remote for at least the first two weeks; in-person (but with the option to be remote) afterwards.
    - **Each class, we'll introduce a new topic.**
    - **We won't actually meet for 3 hours.** We'll end lecture at 7:30PM; the remaining time will be an open working period (i.e. I will be available until 8:50PM)*.
    - Recordings will be posted, though attendance is part of your grade.
- **Throughout the week:** complete readings and homework.
    - Homeworks will contain a mix of reading questions and technical problems.
    - Expect to spend ~3 hours per week on the course outside of class, mostly reading.
- **Sunday, 11:59PM:** homework due.

Evaluation

- This course is 2 units, graded P/NP.
- To pass, you'll need to:

1. **Attend** and **participate** in at least 7/8 class sessions.
    - Let me know in advance if you can't make a particular class.
2. Complete at least 7/8 **homework assignments** satisfactorily.
    - First homework will be released by noon tomorrow and is due this Sunday at 11:59PM.

honest attempt at all problems

Relevance of calculus

average value of $(\text{actual}-\text{predicted})^2$

Why is calculus relevant?

- In modern data science, we're often concerned with making "the best" predictions. To do this, we often need to solve some sort of **optimization problem**.
- Example: **linear regression** (DSC 10, DSC 40A, and upper-div ML).
    - In linear regression, we make predictions using $y = a + bx$.
    - To find the **best slope** (b) and **best intercept** (a), we minimize **mean squared error:**

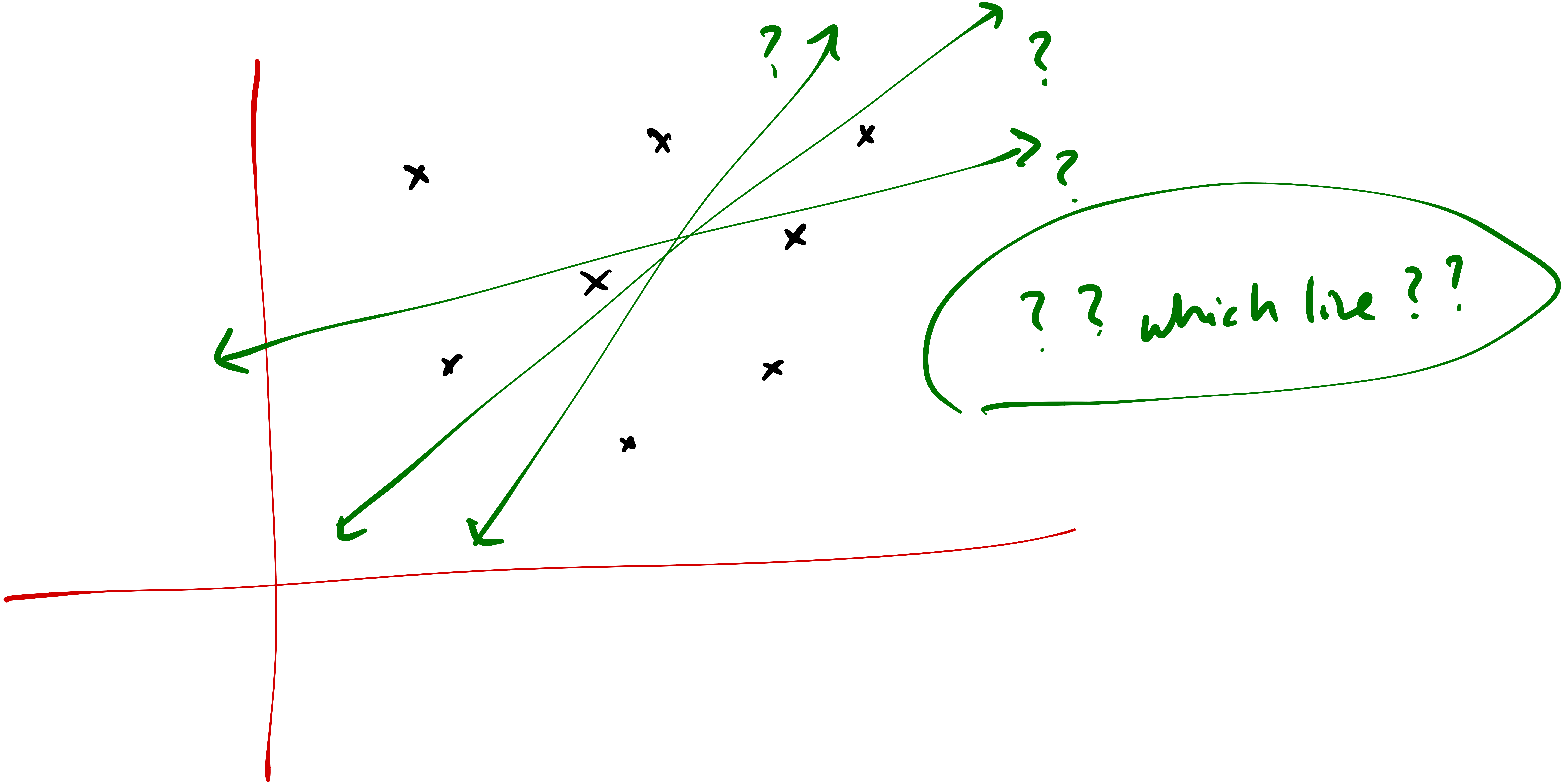
$$\text{mean squared error} = \frac{1}{n}\sum_{i=1}^{n}\Big(\underbrace{y_i}_{\text{actual}} - \underbrace{(a + bx_i)}_{\text{predicted}}\Big)^2$$

    - Using **calculus**, we can minimize mean squared error to find:

$$\boxed{b = r \cdot \frac{\text{SD of } y}{\text{SD of } x}, \quad a = (\text{mean of } y) - b \cdot (\text{mean of } x)}$$

best slope

best intercept


?
?
?
??which line??

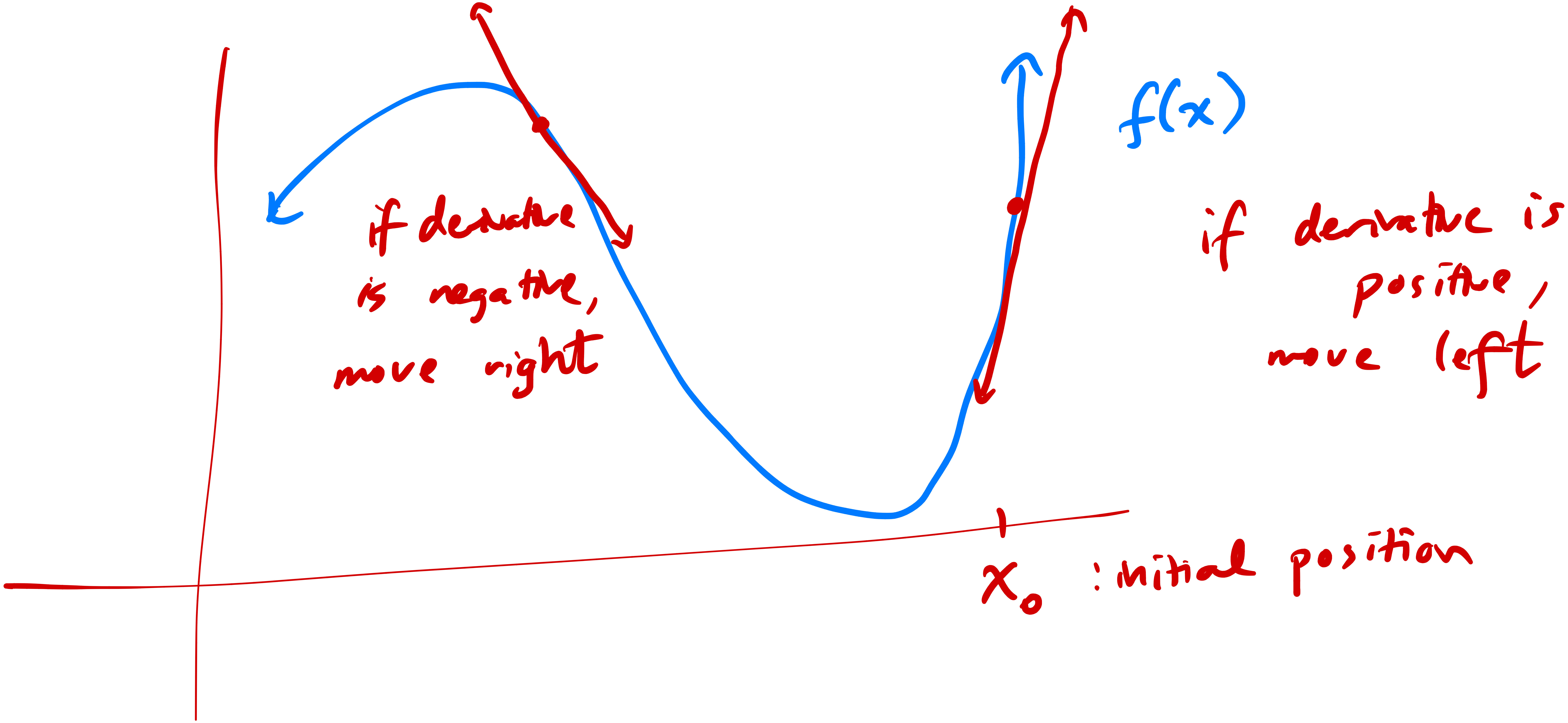
Why is calculus relevant?

- Example: **gradient descent** (DSC 40A and upper-div ML).
  - Gradient descent is a technique used to **minimize** functions.
    - Intuitive idea: to get to the bottom of the hill, travel downhill.

      $f(x,y) = 2x^2 - 3xy$

    - "**Gradient**" is the multivariate version of "**derivative**".
  - It is an iterative method. To minimize $f(x)$, we start with an initial guess x_0 and choose a "step size" α, and use the **update rule**:

$$x_{i+1} = x_i \underbrace{-}_{\text{opposite the direction of derivative}} \alpha \cdot \frac{d}{dx} f(x_i)$$



f(x)
if derivative is negative, move right
if derivative is positive, move left
x_0 : initial position

Preface

- One of the goals of this class is to show you how ideas were developed over time.

- As such, we’re not going to skip straight to “who invented calculus”. There is no one answer to that question, as field of calculus itself was developed over thousands of years.

- We’ll start with some relevant results by mathematicians in Ancient Greece and work our way from there.

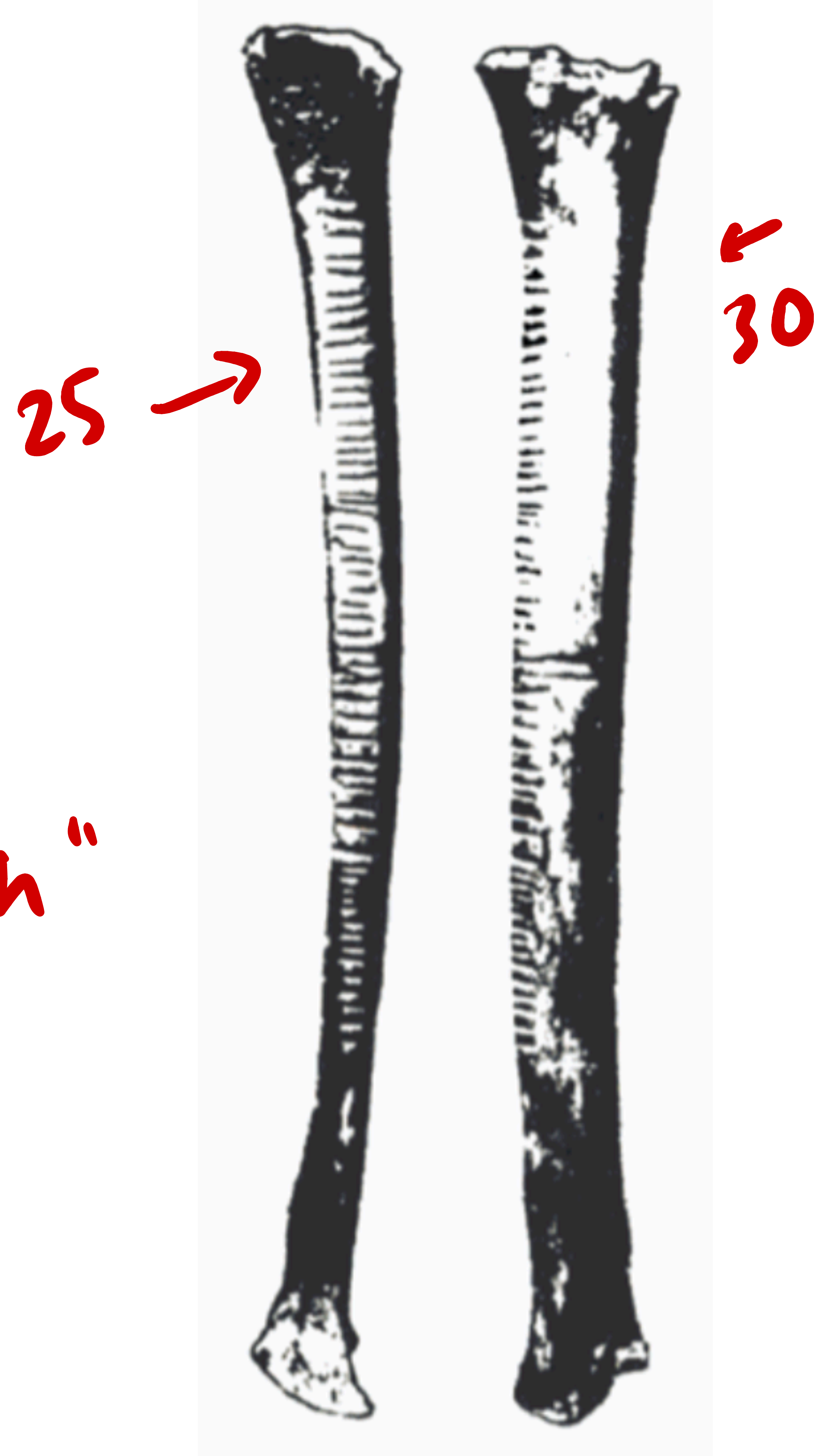
Mathematics in Ancient Greece

The origins of mathematics

- Even before Ancient Greece, there were numbers.
- Humans knew how to count as far back as the Stone Age.
- There is evidence that some civilizations first learned to count to two before they learned to count above two.
  - In English (and most European languages), “two” and “second” have different root words, while “three” and “third”, “four” and “fourth”, etc. are related.

    “twoth”

    ordinal numbers
  - The Piraha tribe in the Amazon only has words for “one”, “two”, and “many” to this day[1].



A stone age tally stick. The tibia (shin) of a wolf with two long incisions in the center, and two series of 25 and 30 marks. Found in Věstonice, Moravia (Czechoslovakia) in 1937.[2]

[1] https://www.nature.com/articles/news040816-10

Pythagoras

Supposedly they proved that $\sqrt{2}$ irrational but hid this!!!

- Pythagoras was one of the first prominent mathematicians in Ancient Greece.
  - He was a mathematician and philosopher who lived from 570 BC to 500-490BC[1].
  - He traveled throughout (modern-day) Europe and Asia and settled in Italy, and developed a school of thought known as **Pythagoreanism**.
    - One belief: reincarnation[2].
- His followers, the Pythagoreans, searched for connections between numbers and nature.
  - Much of what is attributed to him was likely developed by his followers.

1. https://www.britannica.com/biography/Pythagoras
2. https://plato.stanford.edu/entries/pythagoreanism/

The Pythagorean theorem

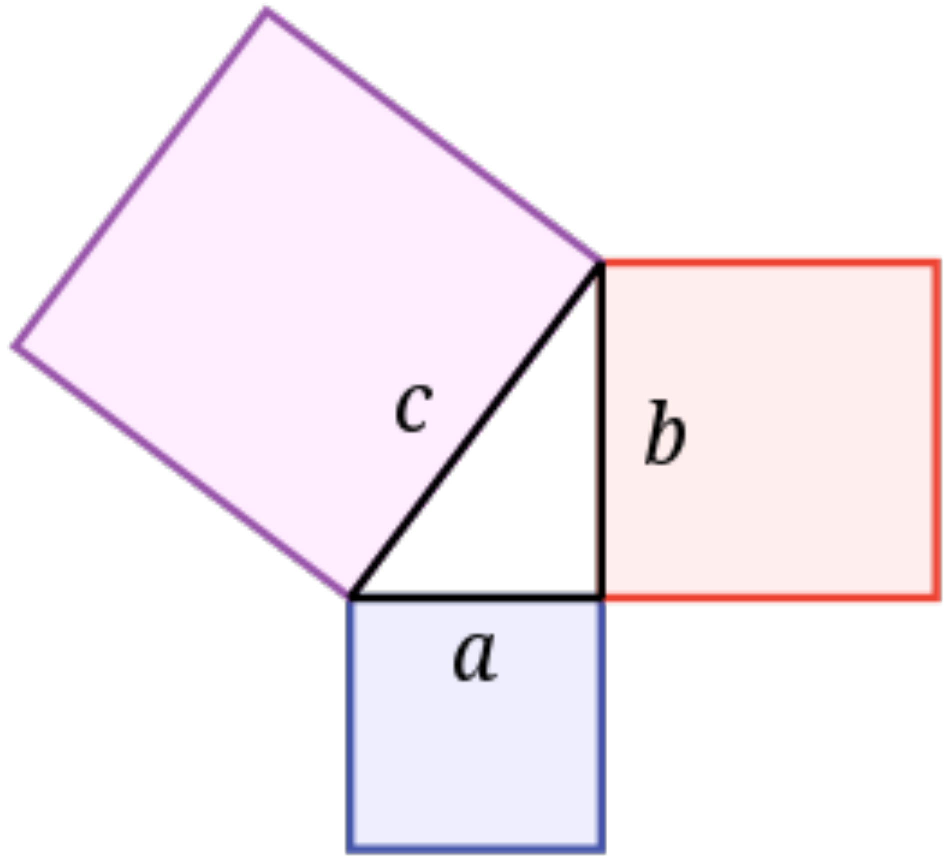
Babylon ≈ Middle East (Mesopotamia)

- Pythagoras is credited for the **Pythagorean theorem**, $\underline{c^2 = a^2 + b^2}$, however he was not the first to discover it.
- There is evidence that the **Babylonians** knew about the theorem more than 1000 years before Pythagoras lived. See Figure 1 for a translation of a Babylonian tablet (that is now archived in the British Museum)[1].
- The Pythagorean theorem appears in many data science-relevant contexts.

  $\text{length}^2 = v_1^2 + v_2^2 + \dots + v_n^2$

  "2-norm"

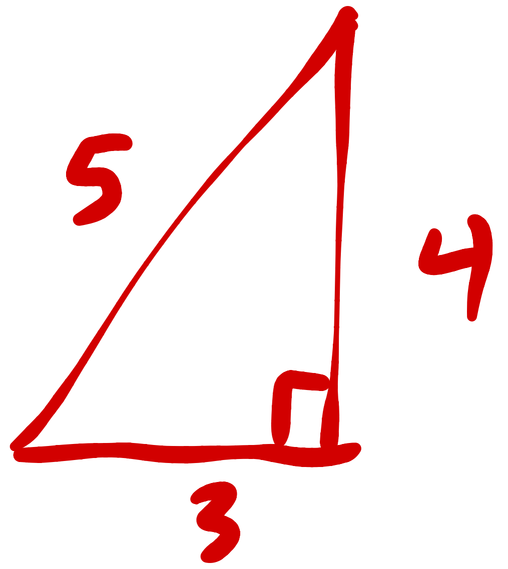
  - The length of a vector, $\|v\|_2 = \sqrt{v_1^2 + v_2^2 + \ldots + v_n^2}$, is an extension of it.
  - In DSC 40A, you'll see that mean squared error can be written in terms of the squared length of a vector, i.e. using the Pythagorean theorem.



4 is the length and 5 the diagonal.
What is the breadth?
Its size is not known.
4 times 4 is 16.
5 times 5 is 25.
You take 16 from 25 and there remains 9.
What times what shall I take in order to get 9?
3 times 3 is 9.
3 is the breadth.

Figure 1

$3^2 + 4^2 = 5^2$



[1] https://mathshistory.st-andrews.ac.uk/HistTopics/Babylonian_Pythagoras/

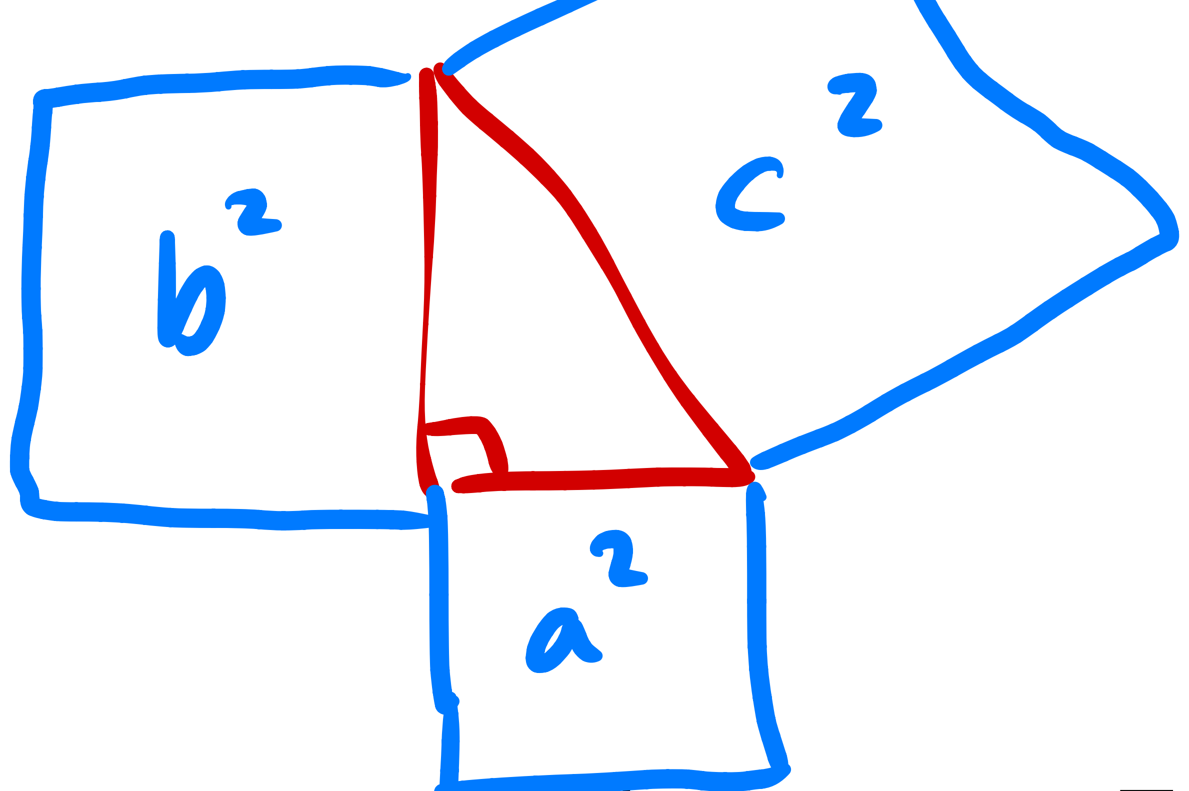
Perfect fifths

- Pythagoras is also credited with discovering connections between frequencies in **music** and counting numbers.
  - The "**perfect fifth**" interval consists of two notes whose frequencies make the ratio 3:2.
    - Middle C: 261.36 Hz[1].
    - G above middle C: 392 Hz.
    - Ratio = 392 / 261.36 ~ 1.499. → 3/2
  - In modern music, perfect fifth intervals span 7 semitones, and sound very pleasing ("consonant").
    - Try it out at this virtual piano.

An Ancient Greek lyre.

[1] https://pages.mtu.edu/~suits/notefreqs.html

Euclid



Egypt

- **Euclid** (~300 BC)[1], sometimes known as **Euclid of Alexandria**, is another prominent figure in the development of mathematics.
- His primary contribution is a series of 13 books, known as "Euclid's Elements", that documented much of what was known about geometry and number theory at the time, including proofs.
  - Euclid's Elements has been referred to as the "most influential textbook of all time"[1].
  - A digital version of the books can be found **here**.
  - Book 1, Proposition 47: "In right-angled triangles the square on the side subtending the right angle is equal to the squares on the sides containing the right angle."

1. https://www.britannica.com/biography/Euclid-Greek-mathematician
2. https://archive.org/details/historyofmathema00boye/page/n21/mode/2up

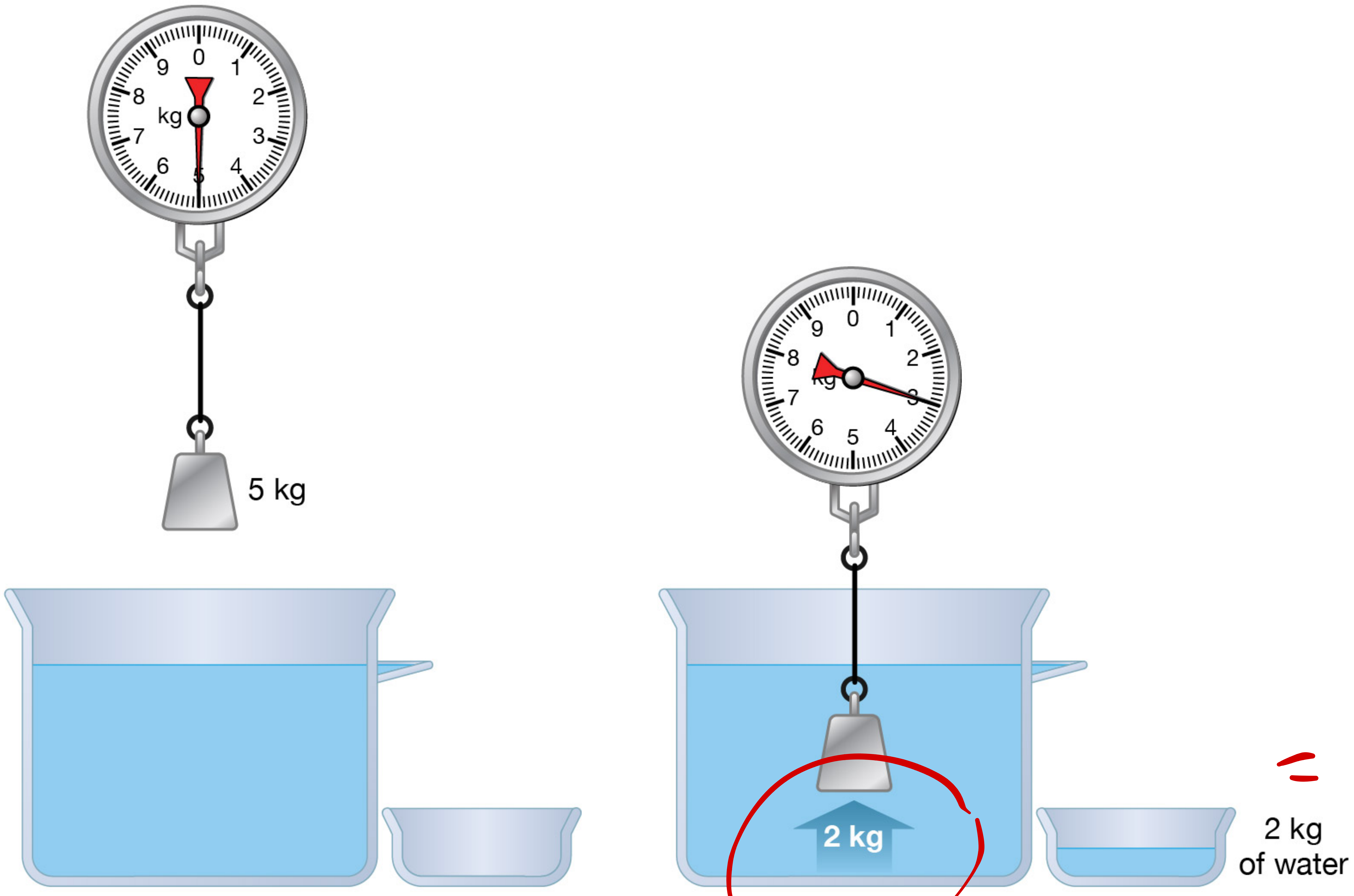
Archimedes

- **Archimedes** (287 BC-212/211 BC)[1], also known as Archimedes of Syracuse, is thought to be one of the most influential figures in the history of science in mathematics.
    - At the time, Syracuse was part of a Greek settlement in Sicily (modern-day Italy).
    - He was supposedly killed when the Romans captured Sicily.
- He is most known for his principle of buoyancy ("**Archimedes' Principle**") and **Archimedes' screw**.

source

[1] https://www.britannica.com/biography/Archimedes

Archimedes’ principle



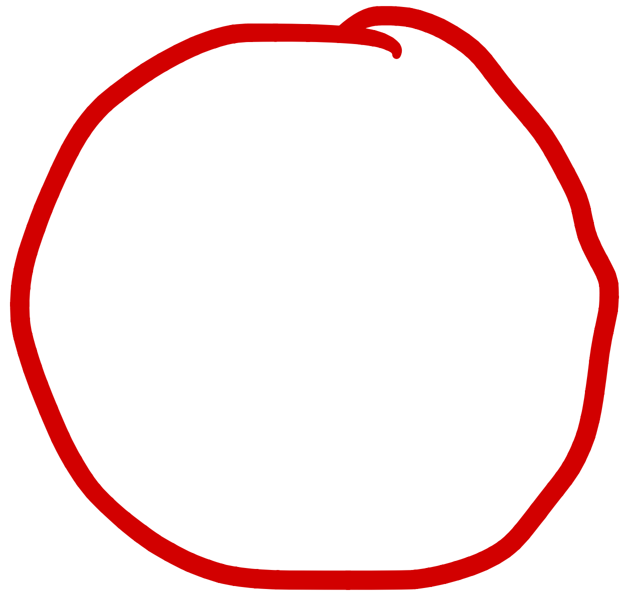
© 2012 Encyclopædia Britannica, Inc.

source

buoyant force

Method of exhaustion

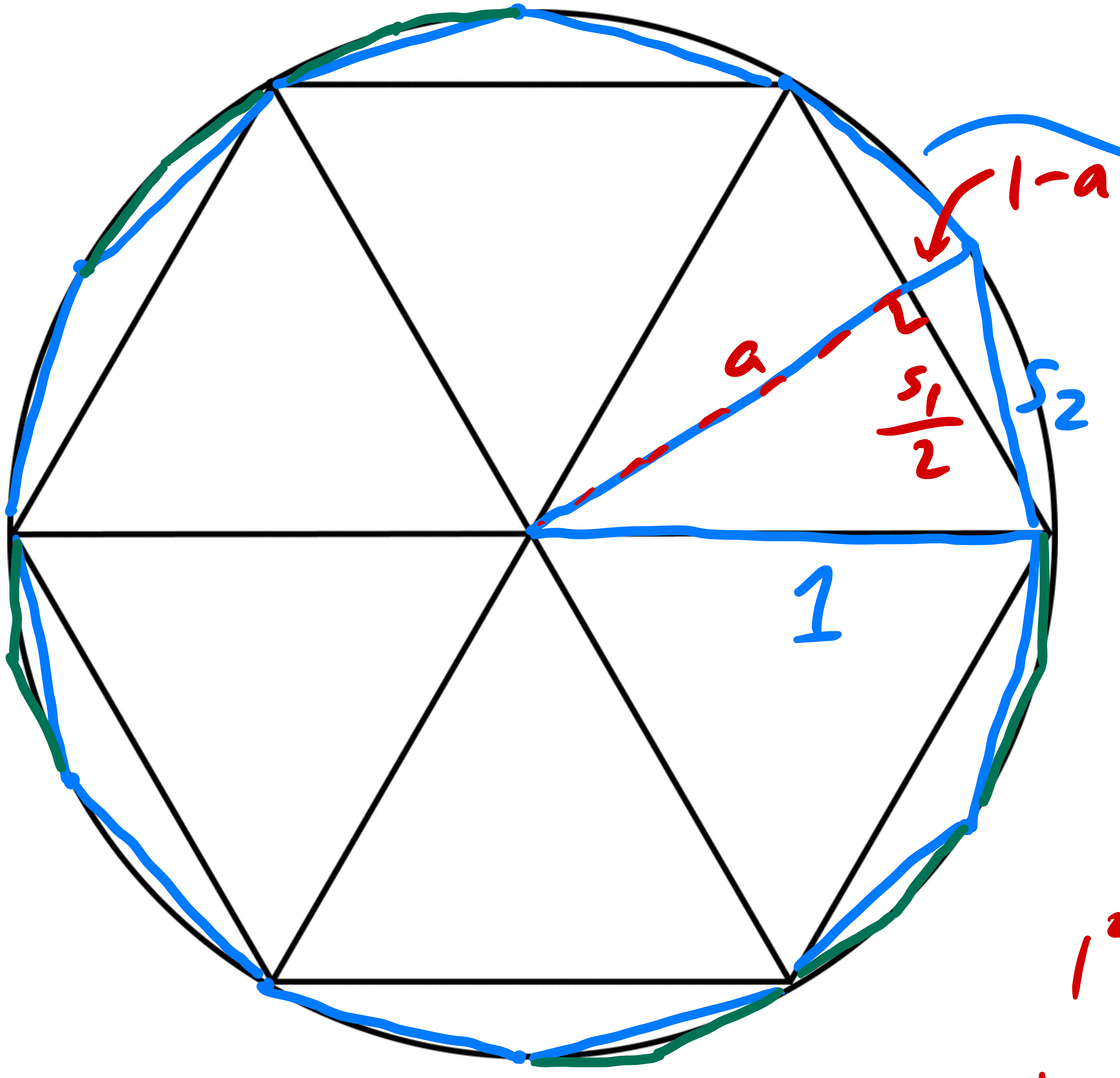
$$\frac{\text{circumference}}{\text{diameter}}$$



$$\frac{C}{2r} = \pi$$

- Archimedes used the **method of exhaustion** to come up with an estimate for the **ratio of the circumference of a circle to the circle's diameter**.
    - We now know this number as π, but this symbol only came into use in the 1700s.
    - In August, Swiss researchers found the value of π to 62.8 **trillion** decimal places![1]
- Main idea behind the method of exhaustion: approximate the circumference of a circle with the perimeter of a regular n-sided polygon, and increase the number of sides.
    - Early pre-cursor to **integral calculus**.
    - Let's try it ourselves!

[1]: https://www.livescience.com/record-number-of-pi-digits.html



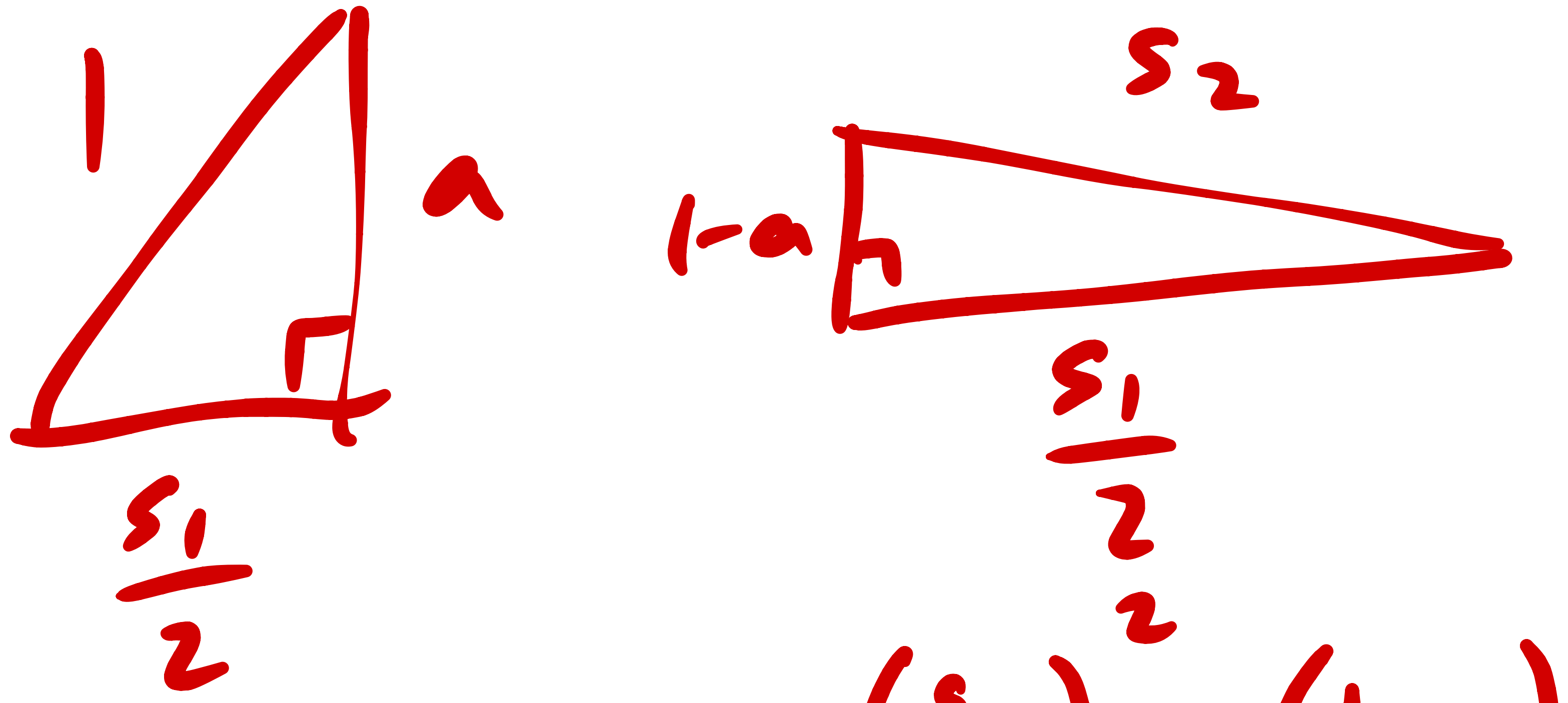
hexagon perimeter = 6

diameter = 2

$\pi \approx \frac{6}{2} = 3$

dodecagon (12 sides)

need to find sidelength, s_2



$$1^2 = \left(\frac{s_1}{2}\right)^2 + a^2$$

$$1 - \left(\frac{s_1}{2}\right)^2 = a^2$$

$$\sqrt{1 - \left(\frac{s_1}{2}\right)^2} = a$$

$$\left(\frac{s_1}{2}\right)^2 + (1-a)^2 = s_2^2$$

$$\sqrt{\left(\frac{s_1}{2}\right)^2 + \left(1 - \sqrt{1 - \left(\frac{s_1}{2}\right)^2}\right)^2} = s_2$$

Notebook demo

- Throughout the quarter, we'll intersperse some Jupyter Notebook demos, which you can access by clicking the "code" link in the week's webpage.
- Let's use our manual calculations to come up with a sequence of estimates of π using the method of exhaustion.
  - Archimedes did this for 6, 12, 24, 48, and 96-sided regular polygons.
  - We can continue the process much further!

Summary, next time

Summary, next time

- Greek mathematicians Pythagoras, Euclid, and Archimedes each contributed substantially to the field of mathematics from 570-200 BC.
  - In particular, Archimedes used techniques that resemble modern calculus (limits and integration).
- However, modern calculus did not fully start to develop until the Scientific Revolution, in the mid-1500s.
- Next time, we will start with studying some of the key figures in the development of calculus.
  - Isaac Newton and Gottfried Wilhelm Leibniz are thought of as being the "founders of calculus", but their work builds on the work of many of their contemporaries.
- After finishing our discussion of calculus, we'll begin discussing the idea of **aggregation**.