Mathematics of Big Data, I

Lecture 1: Introduction of Big Data & Overview of Big Data Analytics

Weiqing Gu
Professor of Mathematics
Director of the Mathematics Clinic

Harvey Mudd College
Spring 2019
https://math189sp19.github.io/syllabus.html

https://math189sp19.github.io/

Office: Shan 3420

Email: gu@g.hmc.edu

Phone: ext. 1-8929

Office Hours: Wednesday 1:00pm - 2:00pm

Course Meeting:

M, 7:00pm - 9:45pm

Course Location:

Harvey Mudd College Shan 3485

TA/Tutoring Hours:

Name: Joseph (Joe) Nunez

Email: jbnunez@g.hmc.edu

Tutoring Hours: (Optional)

Monday: TBD in Shan 3485 (discuss in class)

Name: Jonathan (Jon) Hayase

Email: jhayase@g.hmc.edu

Name: Ricky Shapley

Email: rshapley@g.hmc.edu

Tutoring Hours: (Optional)

Sunday: 7:00pm - 9:00pm in Shan 3485

Textbook:

All members of the class will be required to obtain the following text:

Kevin Patrick Murphy, *Machine Learning: a Probabilistic Perspective* . MIT Press, 2012.

Grading:

- 5% Reading Summary
- 35% Homework
- 20% Midterm (Project or Exam, TBD)
- 40% Final Project
- [Up to 5% Extra Credit]

Course Requirements and Evaluation:

- *Reading Presentations*

All readings are compulsory, but some are more compulsory than others.

To encourage the goal of reading active research in the field, we will assign each non-Murphy reading to a group of two students who will write a summary of 1-2 pages to be turned in at the start of class. Each student will do approxiamately two summaries in total. They must be clear and demonstrate that you have read the paper with a high degree of confidence. Credit will be given on a 0-10 scale for each summary. Your summary should be done at a high level, and should focus on the main point of the readings (i.e. avoid complicated math). As long as your summary is reasonable, you will be given full credit.

Homework!

- ***Homework***

The homework is due every week at the beginning of each lecture. There will be two parts for each assignment: math and coding. The homework is split approximately evenly between mathematical analysis and extension of our course material and application of algorithms to real world data.

For coding: You are highly recommended to use Python3. For each problem, the starter code and the sample solution are implemented in Python3. All the results and graphs for the sample solutions were produced under Python 3.5.2 under macOS Sierra; different versions of Python or system environment may produce different results. You are also welcome to use Jupyter Notebooks, but the starter code is not provided in notebook format.

Numpy and Pandas are two important python libraries to know for coding assignment for this course. You might also want to look at Matplotlib for generating plots. If you never used these libraries before, make sure you check out the tutorials online before starting the first assignment.

- At the end of each lecture, the head grader will give you some instruction on how to start to write your code and what would be some of the expected challenges for the next coding assignment.

Note:

1) When doing the coding problem for each homework set, you are not allowed to use any machine learning algorithms implemented by external libraries, such as LinearRegression in sklearn. However, you may use these algorithms in your final project.

2) Each homework has both pdf and tex versions. To have the tex files successfully compiled, make sure that you have downloaded both macros.tex and hmcpset.cls and put them and the hw tex file under same folder.

If you have any questions with regard to the compilation of the tex files, feel free to ask the grutors for help.

3) For each coding problem, please submit your code to GitHub; please print out any graph or printing statements and submit them with the written part.

Exams

- ***Midterm***

The midterm will either be a take-home exam covering all topics seen in the first week of the course or a project where you will apply the methods learned in the first half of the course (TBD).

- ***Final Project***

The final project is the largest component of the course. Each student will discover, explore, and attack a real world problem of your choosing. The detailed description and requirements for the final project can be found under the "Final Project" tab.

- ***GitHub***

As we stated in the course overview, students are expected to become comfortable with Github. Hence, each student is required to create a Github account for coding assignment submission and final project submission. If you already have a Github account, that's perfect. If not, please create a personal Github account and go over the tutorials online.

Note: **Please make sure to send the username of your Github account to TA for homework grading.**

Classroom Policies:

- ***Attendence***

Attendence for each lecture is mandatory and is expected of all class members. if you're going to miss a lecture, it is neccessary for you to inform the instructor as soon as possible. You are also responsible for obtaining notes from another class member.

- ***Devices***

You are welcome to use your computer or tablet for note-taking (the PowerPoint slides will also be posted shortly after the lecture for your convenience).

A Big Picture of Mathematics of Big Data, I

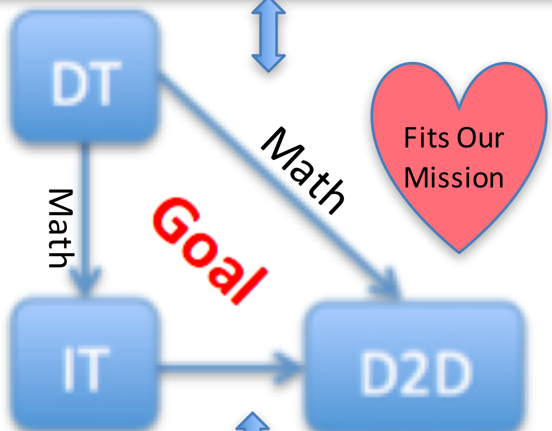
D2D Basic Steps:
- Data Visualization
- Mathematical Modeling
- Computation Methods
- Validation & Verification

Machine Learning
- Supervised
- Unsupervised

Modeling Approaches:
- *Statistical calculus*
- *Geometric analytic*
- *Probabilistic*

Each has its own merit

Keys:
Metrics for define similarities
Strategies of Effective
Optimization and computation



Discrete vs. Smooth
Local vs. Global

Unify Math background:
- Linear algebra
- Statistics & Probability
- Multivariable Calculus
- Geometry & discrete M.

Hadoop *(for large data)*
HDF *(fault resilience)*
MapReduce *(divide & conquer)*
Spark*(fast in memory comput'n)*
Zookeeper (orchestration)

- Utilize the power of computers
- Utilize the power of mathematics

Big Data Introduction
- ***Where does big data come from?***
- ***Different ways to describe big data***
- ***Data could be structure, semi-structured, or unstructured***
- ***Data challenges (e.g. "dirty" data)***

Data Processing
ETL
Data Structuring
Dim Reduction

Data Representations:
Discrete
Real valued
Vector valued
Matrix valued
Manifold valued
Mixed…

Today's Lecture

- **Frist: Big data introduction (answer first two questions)**

Big Data Introduction
- ***Where does big data come from?***
- ***Different ways to describe big data***

- **Second: Use linear regression as an example to give an overview of big data analytics**

Modeling Approaches:
- *Statistical calculus*
- *Geometric analytic*
- *Probabilistic*

Each has its own merit

- Note:

Mathematics of Big Data (in academic) ==

Big Data Analytics (in industry).

First: Introduction of Big Data

- *Where does big data come from?*

Organizations

Machines

People

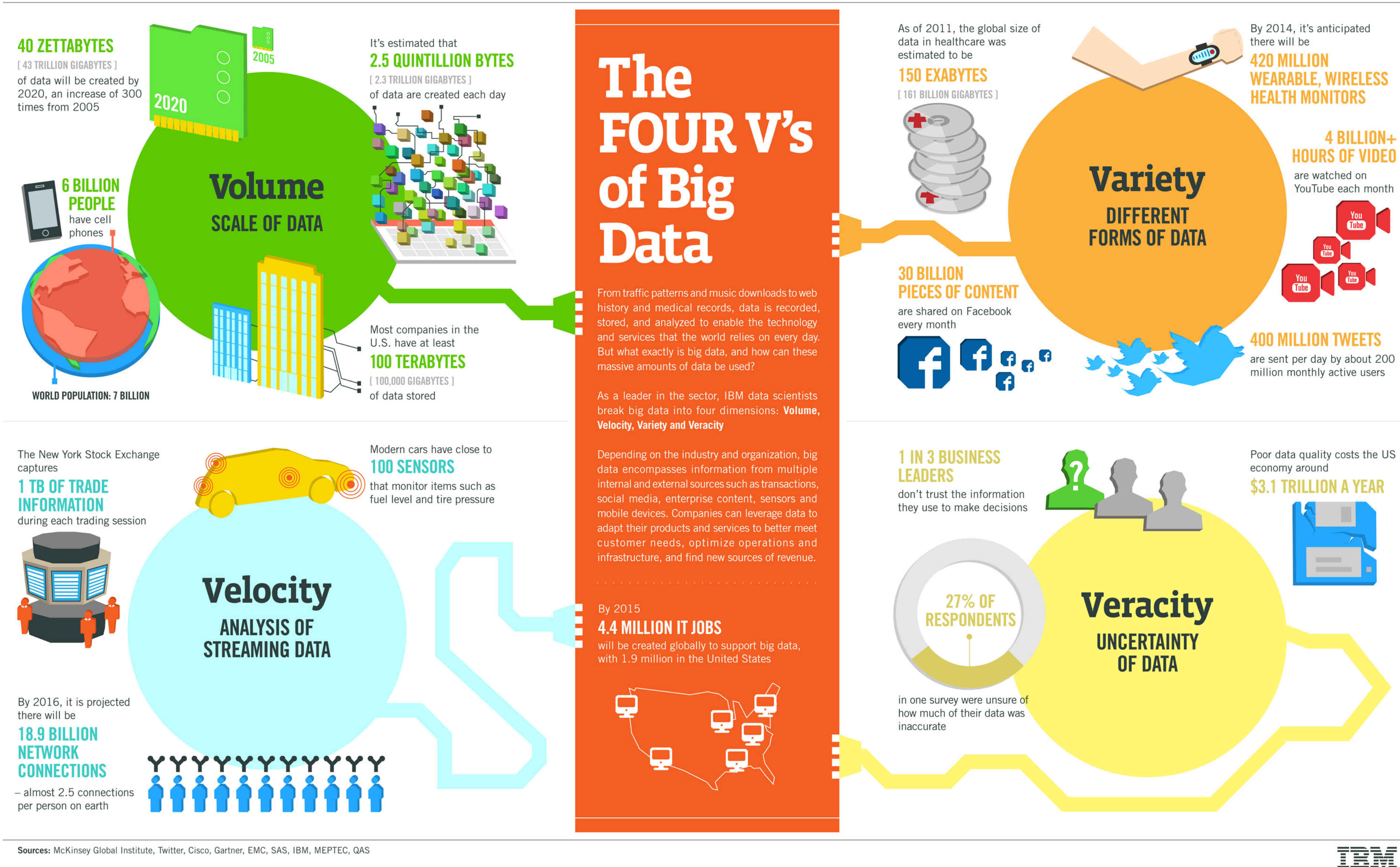
Data is not new. But the scale has been changed!
The way how people using data has been transformed!

Types of big data

1. Structured data (e.g. often Generated by organizations)
2. Semi-structured data (e.g. Generated by machine with manual records)
3. Unstructured data (often Generated by people)

- **What exactly is big data?**

- Does “big" here mean “big volume”?
- In fact, there are 5 “V”s to describe big data.
  - **Volume (Size)**
  - **Velocity (Speed)**
  - **Variety (Types)**
  - **Veracity (Quality)**
  - **Valence (Relationships)**



The FOUR V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
Volume
SCALE OF DATA
40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
Variety
DIFFERENT FORMS OF DATA
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
Velocity
ANALYSIS OF STREAMING DATA
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
Veracity
UNCERTAINTY OF DATA
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM

Data to Decision (D2D)

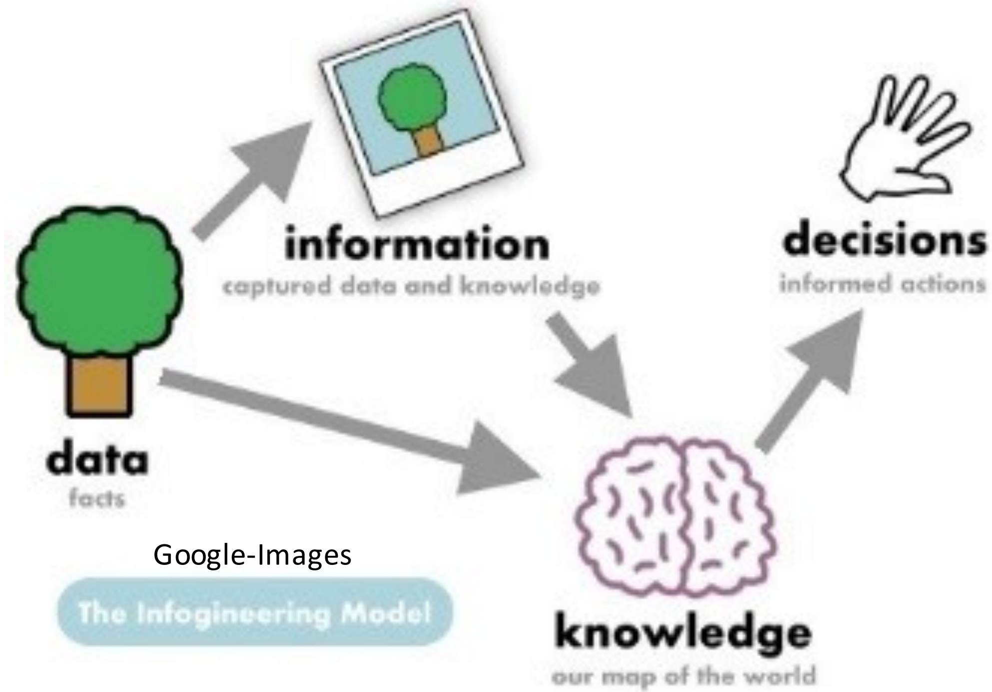
Volume

Velocity

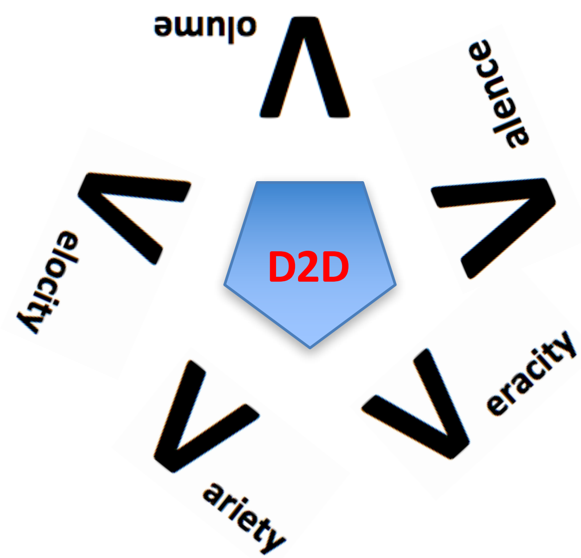
Variety

Veracity

Valence



Google-Images



Second for today: Analytic Approaches

- Use "linear regression" as an example to give an overview of big data analytics

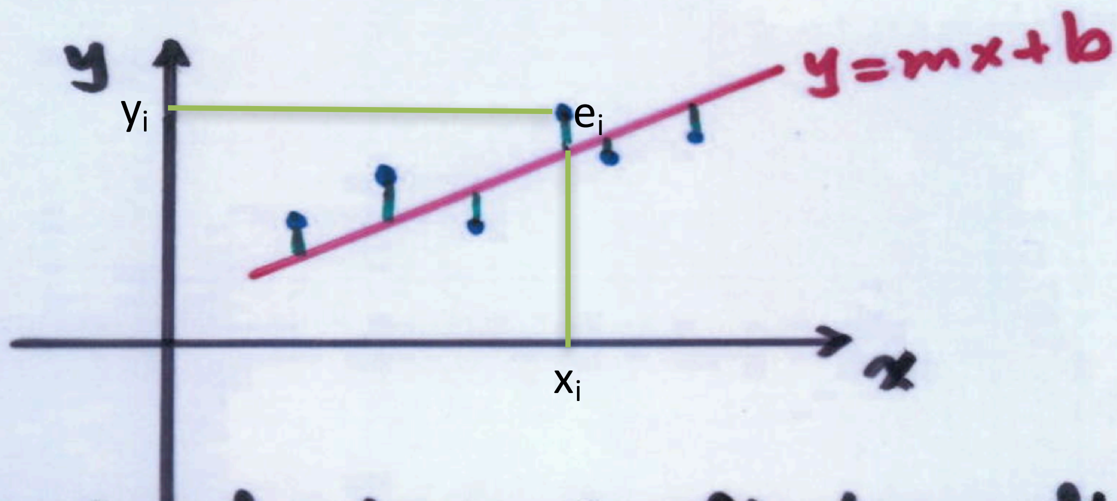
Modeling Approaches:

- *Statistical calculus*
- *Geometric analytic*
- *Probabilistic*

Each has its own merit

1. Statistical Calculus Approach
(Classical Least Square Approximation)

Spse we have data pts (x_i, y_i) and want to find the line $y = mx + b$ which best describes the data.



The problem boils down to find m & b.

The error between one point and the line is

$$e_i = y_i - (mx_i + b)$$

Our objective is minimizing the total error.

- However, the errors e_i, some could be positive and some could be negative. A simple sum of the errors would not work well.
- Can you think about an example why not working well?
- How to fix this problem?
- Instead we consider the following **objective or cost function**:

L_2 *norm*

- $\boxed{J(m,b) = \sum (e_i)^2 = \sum (y_i - mx_i - b)^2}$

- Can we use $\sum |e_i|$ instead?

L_1 *norm*

Goal: Find m and b to minimize the cost function J

- How?
- Set all partials equal to zero!
- Work out the details with the students on the board.

Obtained solution using Cramer's rule

- Give a linear system: $\begin{cases} a_1 x + b_1 y &= c_1 \\ a_2 x + b_2 y &= c_2 \end{cases}$
- Write it into matrix form: $\begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$

Assume the coefficient matrix is invertible, i.e. the det = $a_1b_2 - b_1a_2$ is nonzero. Then

$$x = \frac{\begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}} = \frac{c_1 b_2 - b_1 c_2}{a_1 b_2 - b_1 a_2}, \quad y = \frac{\begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}} = \frac{a_1 c_2 - c_1 a_2}{a_1 b_2 - b_1 a_2}.$$

Close formula for Least Square Approximation

Using Cramer's rule, we get solution for m, b:

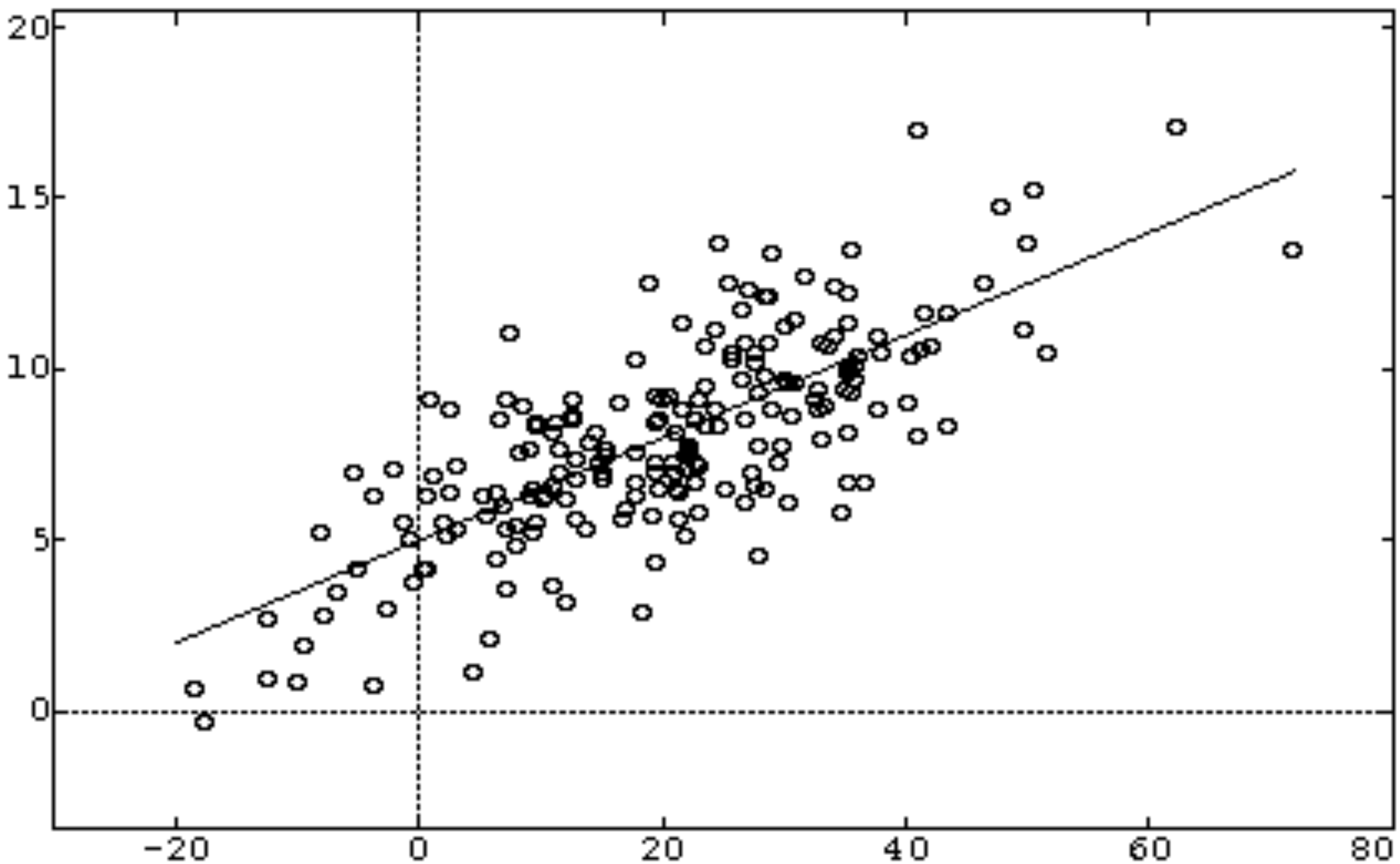
$$m = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

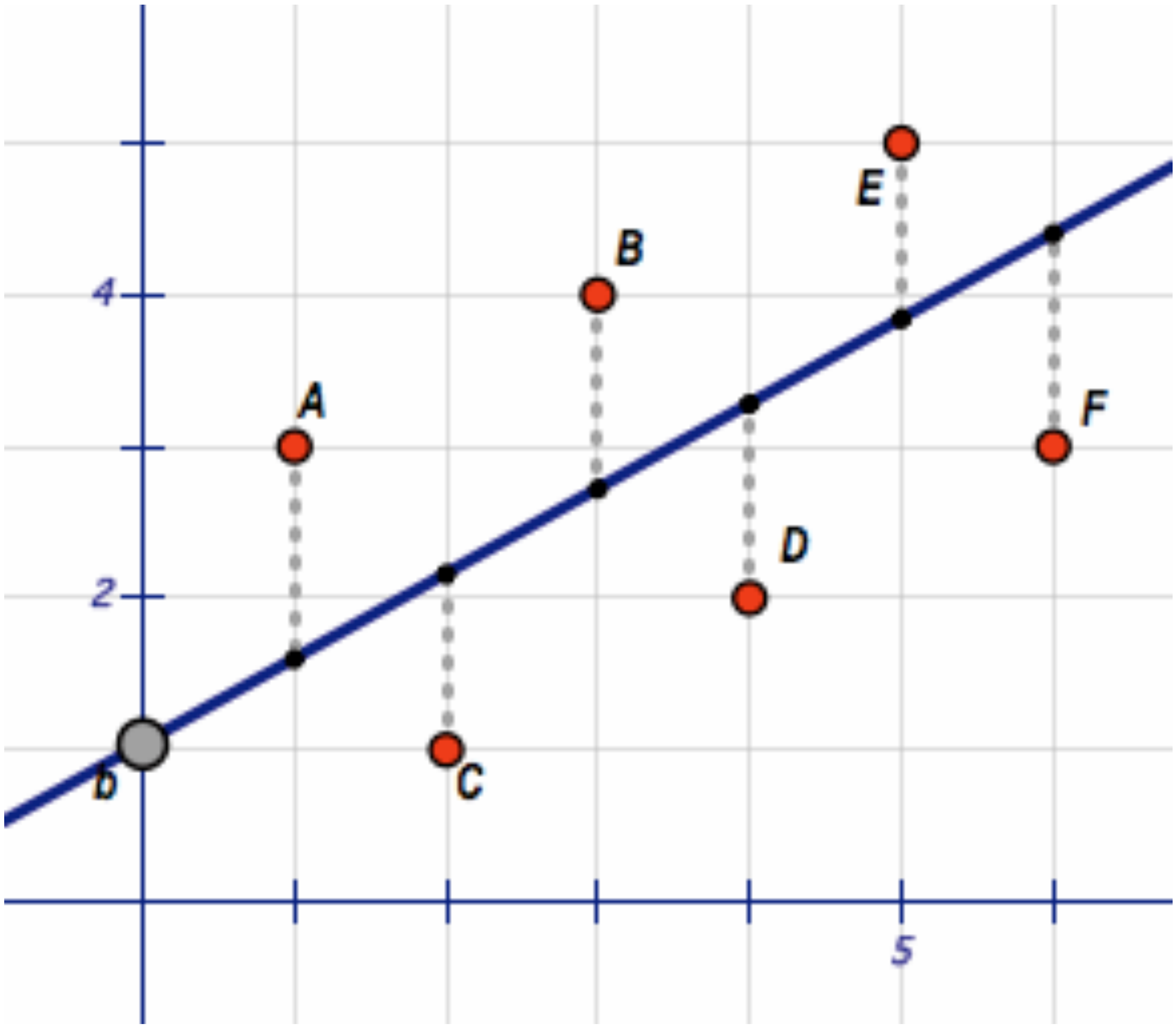
$$b = \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

But the formula is massy. Next we'll find a compact form of this formula.

Linear Regression

Given some data: $D = \{x_i, y_i\}$





A
B
C
D
E
F
b
4
2
5

Normal Equation for Least Square Approximation

- i.e. Representing the Least Square Solution in Matrix Form
- Work out the details with the students on the board.
- Recall the product rule:
- f, g: R→R: $(f \cdot g)' = f' \cdot g + f \cdot g'$
- f, g: R^n→R: $\nabla(f \cdot g) = \nabla f \cdot g + f \cdot \nabla g$
- **f**, **g**: R^n→R^n: $(\mathbf{f} \cdot \mathbf{g})' = \mathbf{f}\,' \cdot \mathbf{g} + \mathbf{f} \cdot \mathbf{g}\,'$

$$\boxed{\theta = (X^T X)^{-1} X^T \vec{y}.}$$

Homework problem

- Given 4 points as below:

$(0, 1), (2, 3), (3, 6), (4, 8)$

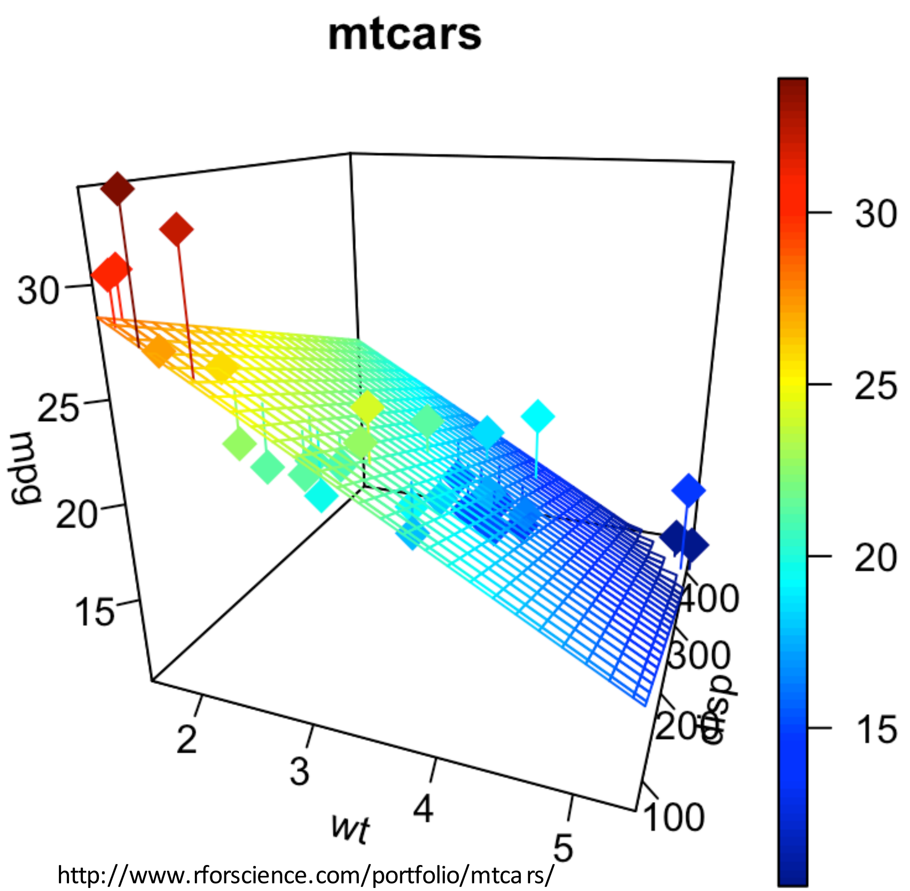
a) Find y = mx + b based on Cramer's rule.

- Hint:

x_i	y_i	x_i^2	$x_i y_i$
0	1	0	0
2	3	4	6
3	6	9	18
4	8	16	32
$\Sigma x_i = 9$	$\Sigma y_i = 18$	$\Sigma x_i^2 = 29$	$\Sigma x_i y_i = 56$

- b) Use the normal formula to find the solution and compare it with that of a).
- c) Plot the data points, and draw y = mx +b.
- d) (All by coding) Find another 100 points near the line y = mx + b. Then find the least square approxim'n again & plot both the data points & the new line.

How about fit data by a plane?



http://www.rforscience.com/portfolio/mtcars/

Get the same close solution by normal equation!

- Can you imagine what other cases you would get the same kind of solution?

2. Geometric Analytic Approach (Geometric Least Square)

- Work out the details with the students on the board.

Assume a linear model

$$\begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix}$$

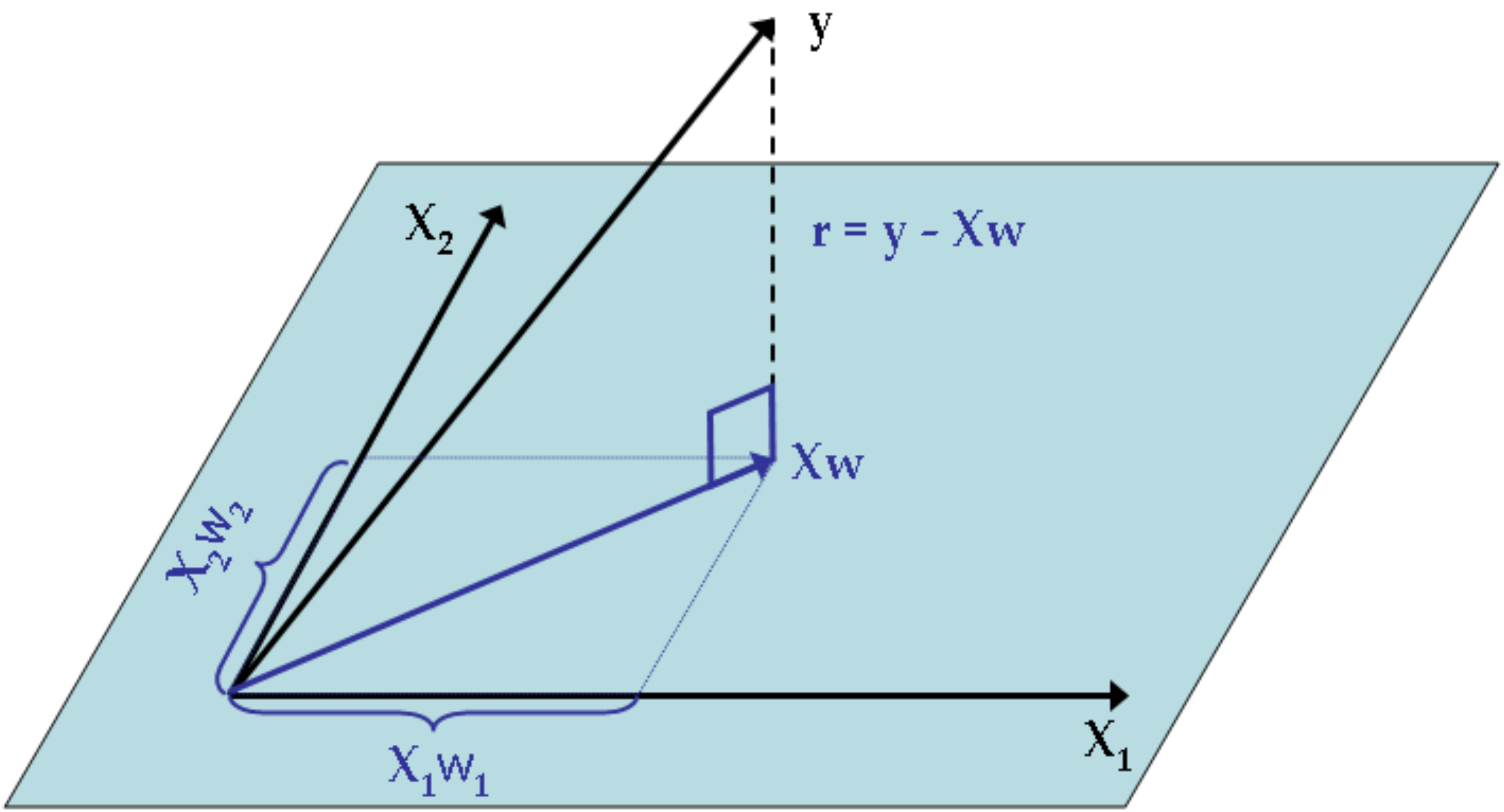
$$\Longrightarrow \qquad \mathbf{r} = (\mathbf{y} - \mathbf{X}\mathbf{w})$$

This is equivalent to

$$y_i = \sum_j w_j x_{ij} + \mathcal{N}(0, \sigma^2) = \mathbf{x}_i \mathbf{w} + \mathcal{N}(0, \sigma^2)$$

Key in *Geometric* Least Square Approximation

Geometrically you can see the solution!



$$\mathbf{w}_{MLE} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Again we get the same solution!

$$\theta = (X^T X)^{-1} X^T \vec{y}.$$

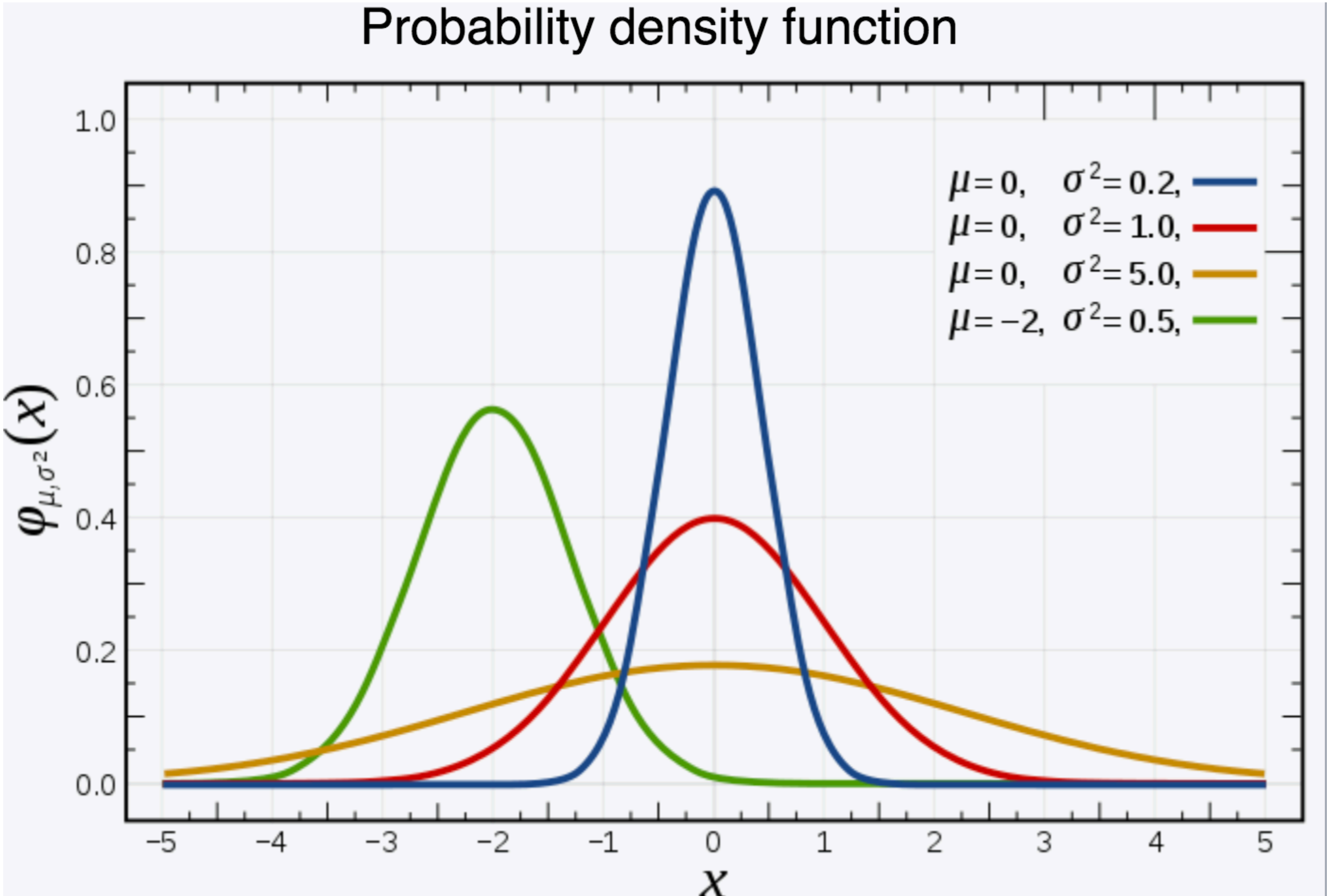
Q: But what's wrong if we use Cramer's rule to solve it?

Or directly use the formula by finding the inverse $X^T X$?

3. Probabilistic Approach (Maximal Likelihood)

- Work out the details with the students on the board.

Recall Gaussian distribution



The red curve is the *standard normal distribution*

Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 > 0$ = variance (squared scale)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Expected Value

Let X be a random variable with a finite number of finite outcomes x_1, x_2, ..., x_k occurring with probabilities p_1, p_2, ..., p_k, respectively. The **expectation** of X is defined as

$$\mathrm{E}[X] = \sum_{i=1}^{k} x_i\, p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k .$$

Since all probabilities p_i add up to 1 ($p_1 + p_2 + \cdots + p_k = 1$), the expected value is the weighted average, with p_i's being the weights.

Special case: Average

If all outcomes x_i are equiprobable (that is, $p_1 = p_2 = \cdots = p_k$), then the weighted average turns into the simple average. This is intuitive: the expected value of a random variable is the average of all values it can take; thus the expected value is what one expects to happen *on average*.

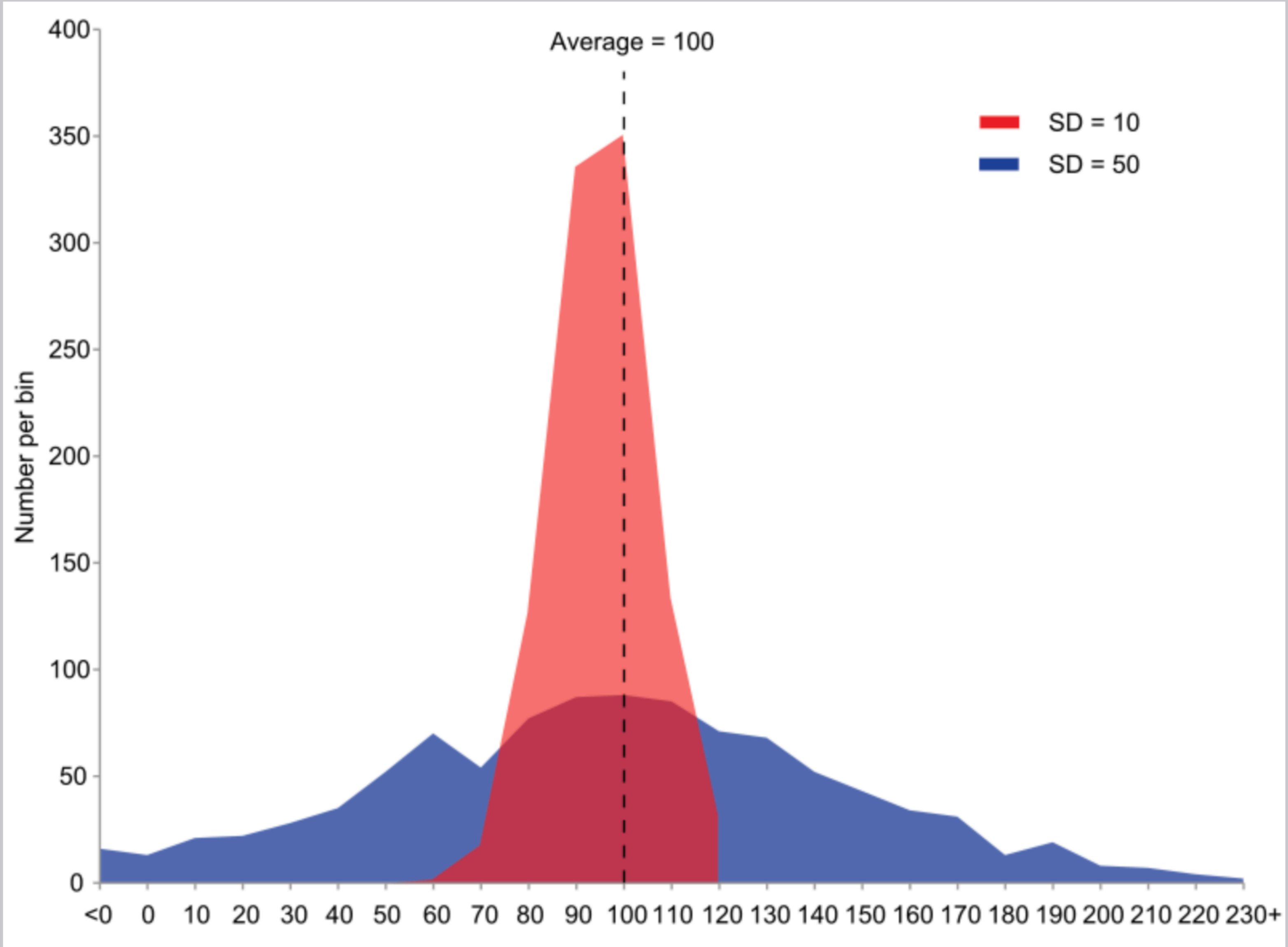
Continuous case

If X is a random variable whose cumulative distribution function admits a density $f(x)$, then the expected value is defined as the following Lebesgue integral:

$$\mathrm{E}[X] = \int_{\mathbb{R}} x f(x)\, dx.$$

The variance of a random variable X is the expected value of the squared deviation from the mean of X, $\mu = \mathbf{E}[X]$:

$$\operatorname{Var}(X) = \mathbf{E}\left[(X - \mu)^2\right].$$



Example of samples from two populations with the same mean but different variances. The red population has mean 100 and variance 100 (SD=10) while the blue population has mean 100 and variance 2500 (SD=50).

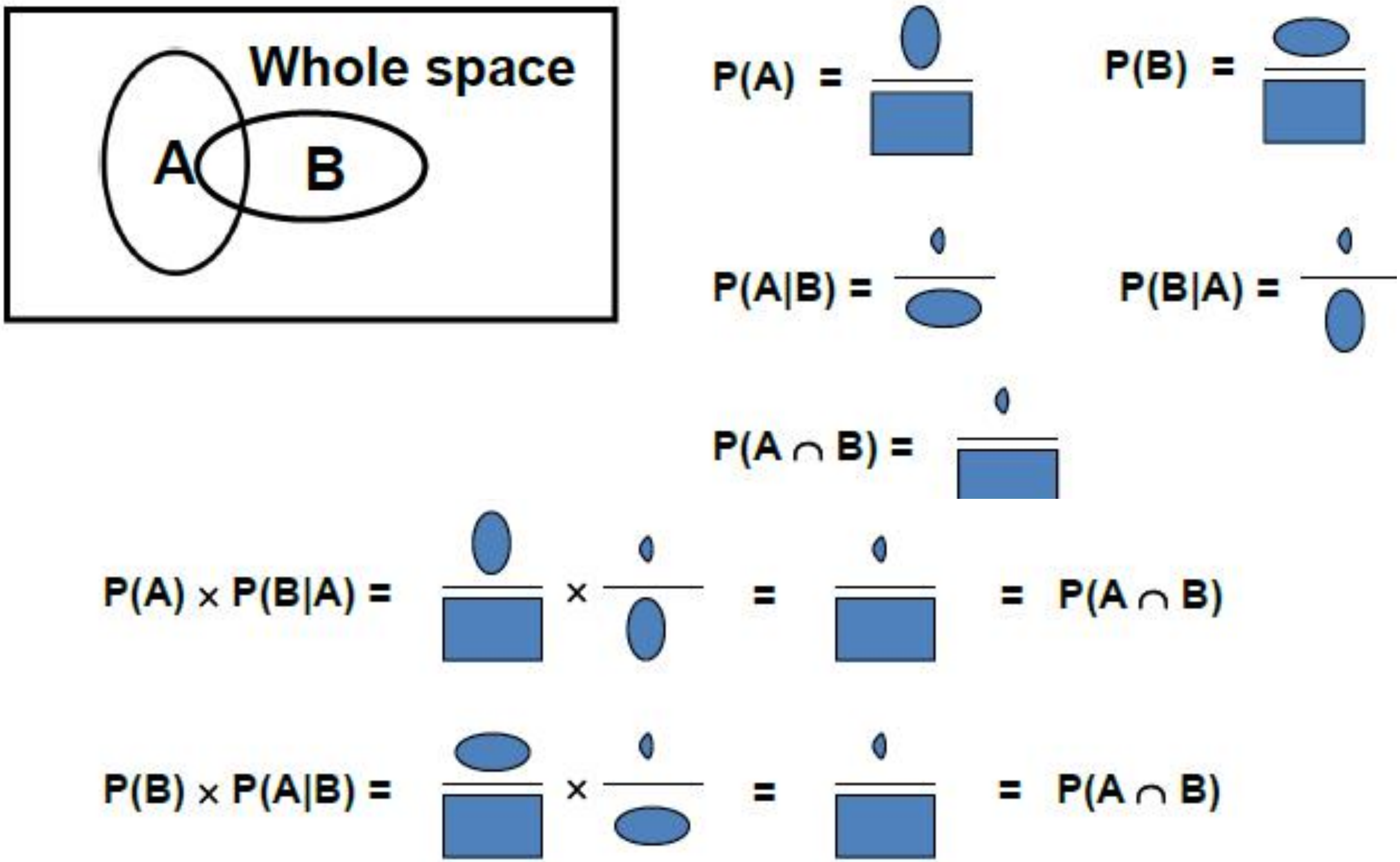
$$
\begin{aligned}
\operatorname{Var}(X) &= \mathbf{E}\left[(X - \mathbf{E}[X])^2\right] \\
&= \mathbf{E}\left[X^2 - 2X\,\mathbf{E}[X] + \mathbf{E}[X]^2\right] \\
&= \mathbf{E}\left[X^2\right] - 2\,\mathbf{E}[X]\,\mathbf{E}[X] + \mathbf{E}[X]^2 \\
&= \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2
\end{aligned}
$$

Continuous case

If the random variable X represents samples generated by a continuous distribution with probability density function $f(x)$,

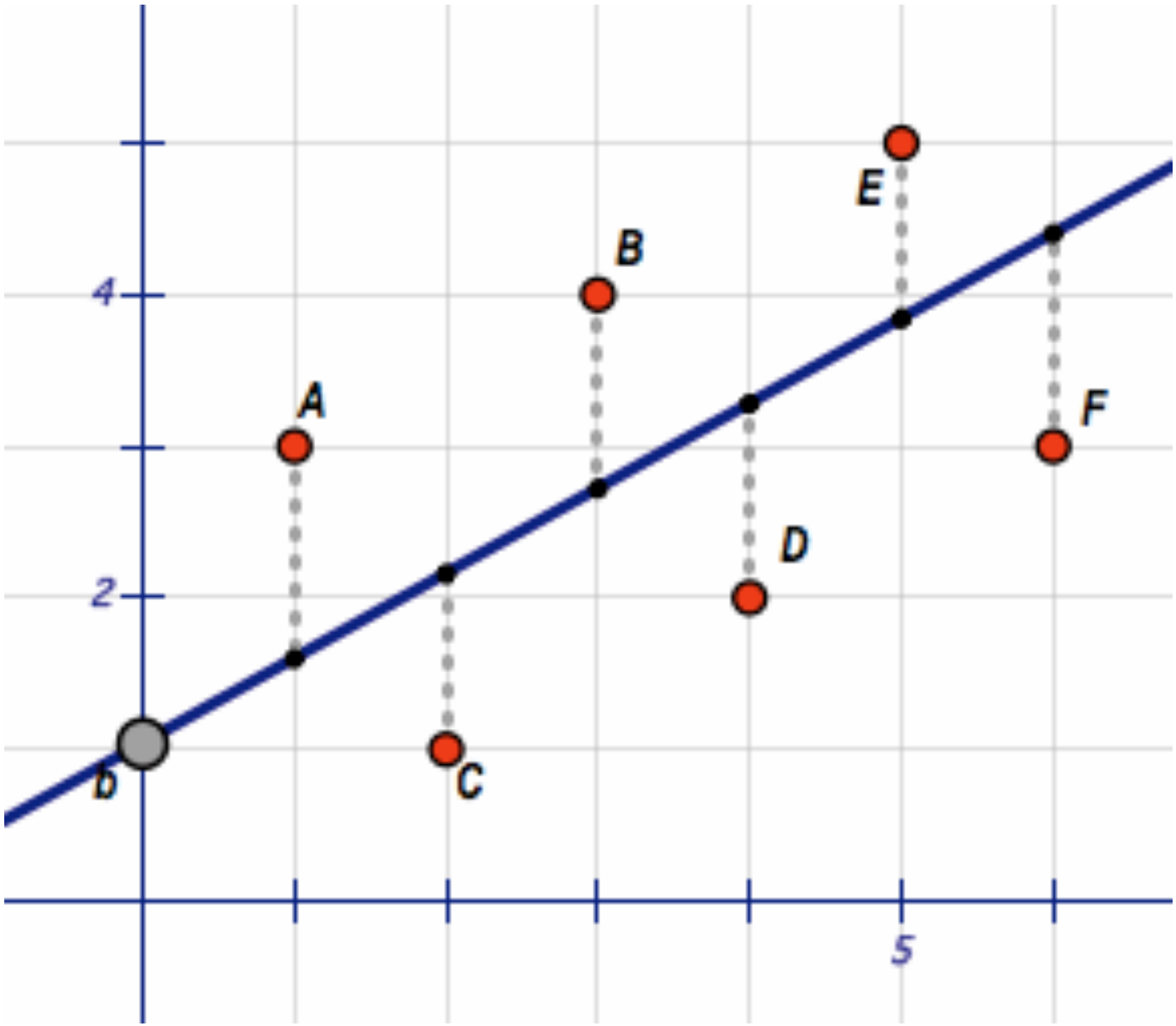
$$\operatorname{Var}(X) = \sigma^2 = \int (x-\mu)^2 f(x)\, dx$$

Visualize Bayes' Theorem



$$\Rightarrow P(B|A) = P(A|B) \times P(B) / P(A)$$

Viewed as a function of \theta.



- Back up slides

Taking Partial Derivatives for different Types of functions

Type 1: $\mathbb{R} \to \mathbb{R}$ (one-to-one)

$x \mapsto f(x)$

$$\frac{\partial f}{\partial x} = \frac{df}{dx}$$

☆☆ **Type 2**: $\mathbb{R}^n \to \mathbb{R}$ (Many-to-one)

$(x_1, x_2, \cdots, x_n) \mapsto f(x_1, \cdots, x_n)$

$$\left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n}\right) \overset{\Delta}{\underset{\text{denoted}}{=}} \nabla f$$

$$\nabla f(\vec{a}) = \left(\frac{\partial f}{\partial x_1}\Big|_{\vec{a}}, \frac{\partial f}{\partial x_2}\Big|_{\vec{a}}, \cdots, \frac{\partial f}{\partial x_n}\Big|_{\vec{a}}\right)$$

is called the gradient of f at $\vec{a}$.

Type 3: $\mathbb{R} \to \mathbb{R}^m$ (one-to-many)

$t \mapsto (f_1(t), \cdots, f_m(t)) \overset{\Delta}{=} f(t)$

$$\begin{bmatrix} \frac{\partial f_1}{\partial t} \\ \vdots \\ \frac{\partial f_m}{\partial t} \end{bmatrix} = \begin{bmatrix} \frac{df_1}{dt} \\ \vdots \\ \frac{df_m}{dt} \end{bmatrix} \overset{\Delta}{=} f'(t)$$

Key Technique: Treat each component function as many-to-one function!

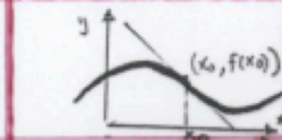
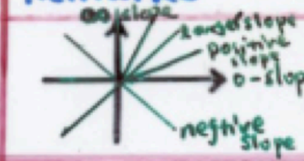
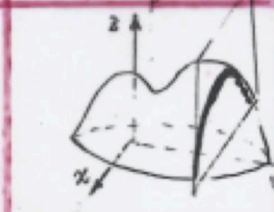
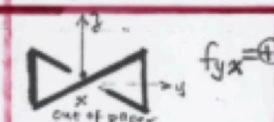
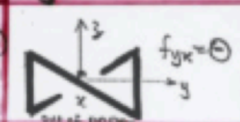
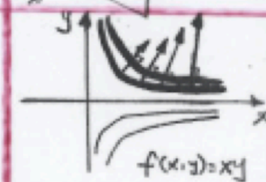
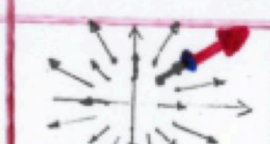
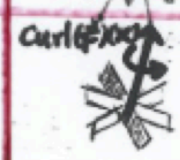
☆☆ **Type 4**: $\mathbb{R}^n \to \mathbb{R}^m$ (many-to-many)

$\underbrace{(x_1, \cdots, x_n)}_{\vec{x}} \mapsto (f_1(\vec{x}), \cdots, f_m(\vec{x}))$

$$\underbrace{Df(\underbrace{x_1, x_2, \ldots, x_n}_{\vec{x}}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}}_{\text{Derivative matrix}} \begin{matrix} \leftarrow \nabla f_1(\vec{x}) \\ \leftarrow \nabla f_2(\vec{x}) \\ \\ \leftarrow \nabla f_m(\vec{x}) \end{matrix}$$

You must keep your mind clear what type of function your are dealing with!

A Big Picture of Derivatives (By Prof. Gu)

Type of funcs	Type of Derivatives	Notations	Pictures	Meanings	Remarks	Formulas
$f: \mathbb{R} \to \mathbb{R}$	Derivative of $f(x)$ at x_0	$\frac{df}{dx}$ or $f'(x)$		Slope at $(x_0, f(x_0))$ of the curve		The tangent line at $(x_0, f(x_0))$: $\boxed{y = f(x_0) + f'(x_0)(x - x_0)}$
$f: \mathbb{R}^2 \to \mathbb{R}$ or $\mathbb{R}^n \to \mathbb{R}$	Partial derivatives w.r.t x or y	$\frac{\partial f}{\partial x} = f_x$, $\frac{\partial f}{\partial y} = f_y$		f_x = slope of graph in x-direction	Similarly, f_y = slope of graph in y-direction	(If exists) The tangent plane at $(a, b, f(a,b))$: $\boxed{z = f(a,b) + f_x(a,b)(x-a) + f_y(a,b)(y-b)}$
$f: \mathbb{R}^2 \to \mathbb{R}$ or $\mathbb{R}^n \to \mathbb{R}$	Higher order partials: Here: 2^{nd} partials	$\frac{\partial^2 f}{\partial x^2} = f_{xx}$, $\frac{\partial^2 f}{\partial y^2} = f_{yy}$		f_{xx} = Concavity in x direction; $(f_y)_x$ = rate change of f_y as x increases	$f_{xx} > 0$: $f_{yy} < 0$:	Laplace's eqn: $\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = 0$
$f: \mathbb{R}^2 \to \mathbb{R}$ or $\mathbb{R}^n \to \mathbb{R}$	Mixed partials	$\frac{\partial^2 f}{\partial x \partial y} = f_{yx}$			In the first picture, as x increases f_y increases from neg to positive $\Rightarrow f_{yx} = \oplus$	Spse $f(x,y)$ has continuous 1^{st}- and 2^{nd} partial deriv's $\Rightarrow \boxed{f_{xy} = f_{yx}}$
$f: \mathbb{R}^n \to \mathbb{R}$	Directional derivative of $\mathbb{R}$-valued func of 2-variable	$D_{\vec{u}} f(\vec{a})$		Rate change of f in the direction of v.	Any directional derivative is completely determined by the direction and gradient.	$\boxed{D_{\vec{u}} f(\vec{a}) = \nabla f(\vec{a}) \cdot \vec{u}}$ $\vec{u}$ (a unit vector!)
$f: \mathbb{R}^n \to \mathbb{R}$	Gradient of real valued function of 2 or 3 variables	∇f $\nabla f(\vec{a})$		$\frac{\nabla f(a,b)}{\lVert \nabla f(a,b) \rVert} \triangleq \vec{u}_a$ direction of steepest ascent, i.e. where $D_{\vec{u}} f(a)$ is maximized	If S is a surface given by: $f(x,y,z) = c$ Then an equation for tangent plane to S at $\vec{x}_0$: $\nabla f(\vec{x}_0) \cdot (\vec{x} - \vec{x}_0) = 0$	$\nabla = \vec{i} \frac{\partial}{\partial x} + \vec{j} \frac{\partial}{\partial y} + \vec{k} \frac{\partial}{\partial z}$ $\nabla f = \vec{i} \frac{\partial f}{\partial x} + \vec{j} \frac{\partial f}{\partial y} + \vec{k} \frac{\partial f}{\partial z}$
$\vec{F}: \mathbb{R}^n \to \mathbb{R}^n$	Divergence of a vector field	$\text{div}(\vec{F}) = \nabla \cdot \vec{F}$		Measurement of the "net mass flow" of $\vec{F}$ in or out at a pt.	or div $\vec{F}$ = rate of expansion per unit volume.	$\text{div} \vec{F} = \nabla \cdot \vec{F} = \frac{\partial F_1}{\partial x_1} + \frac{\partial F_2}{\partial x_2} + \cdots + \frac{\partial F_n}{\partial x_n}$
$\vec{F}: \mathbb{R}^3 \to \mathbb{R}^3$	Curl of a vector field	$\text{curl}(\vec{F}) = \nabla \times \vec{F}$		A tiny twig or Paddle at $x \in \mathbb{R}^3$ will spin around the axis in dir^n of vector curl $\vec{F}$ obeying R-H rule w/ angular velocity $\frac{1}{2}$ curl $\vec{F}(\vec{x})$ and angular speed $\frac{1}{2} \lvert \text{curl} \vec{F}(\vec{x}) \rvert$ rdians/sec.		$\text{Curl}(\vec{F}) = \nabla \times \vec{F} = \begin{vmatrix} \vec{i} & \vec{j} & \vec{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \end{vmatrix}$

Normal Equation for Least Square Approximation

- i.e. Representing the Least Square Solution in Matrix Form
- Work out the details with the students on the board.
- Recall the product rule:
- f, g: R→R: $(f \cdot g)' = f' \cdot g + f \cdot g'$
- f, g: R^n→R: $\nabla(f \cdot g) = \nabla f \cdot g + f \cdot \nabla g$
- **f**, **g**: R^n→R^n: $(\mathbf{f} \cdot \mathbf{g})' = \mathbf{f}\,' \cdot \mathbf{g} + \mathbf{f} \cdot \mathbf{g}\,'$

$$\boxed{\theta = (X^T X)^{-1} X^T \vec{y}.}$$

Mean

- Arithmetic Mean

The *arithmetic mean* (or simply "mean") of a sample $x_1, x_2, \ldots, x_n$, usually denoted by $\bar{x}$, is the sum of the sampled values divided by the number of items in the example

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- Expected Value:

The mean of a probability distribution is the long-run arithmetic average value of a random variable having that distribution. In this context, it is also known as the expected value. For a discrete probability distribution, the mean is given by $\sum xP(x)$, where the sum is taken over all possible values of the random variable and $P(x)$ is the probability mass function.

Mean of a probability Distribution (Expected Value)

The mean of a probability distribution is the long-run arithmetic average value of a random variable having that distribution. In this context, it is also known as the expected value. For a discrete probability distribution, the mean is given by $\sum xP(x)$, where the sum is taken over all possible values of the random variable and $P(x)$ is the probability mass function.

For a continuous distribution, the mean is $\int_{-\infty}^{\infty} x f(x)\, dx$, where $f(x)$ is the probability density function.

If the entries in the column vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

are random variables, each with finite variance, then the covariance matrix Σ is the matrix whose (*i*, *j*) entry is the covariance

$$\Sigma_{ij} = \mathrm{cov}(X_i, X_j) = \mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)] = \mathrm{E}[X_i X_j] - \mu_i \mu_j$$

where the operator E denotes the expected (mean) value of its argument, and

$$\mu_i = \mathrm{E}(X_i)$$

is the expected value of the *i* th entry in the vector **X**.

Covariance Matrix

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1-\mu_1)(X_1-\mu_1)] & \mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathrm{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix}.$$

Note: The covariance matrix is a symmetric matrix.

In fact, a covariant matrix is also positive semi-definite.

The inverse of this matrix, Σ^{-1}, if it exists, is the inverse covariance matrix, also known as the concentration matrix or precision matrix.[1]