Mathematics of Big Data, I

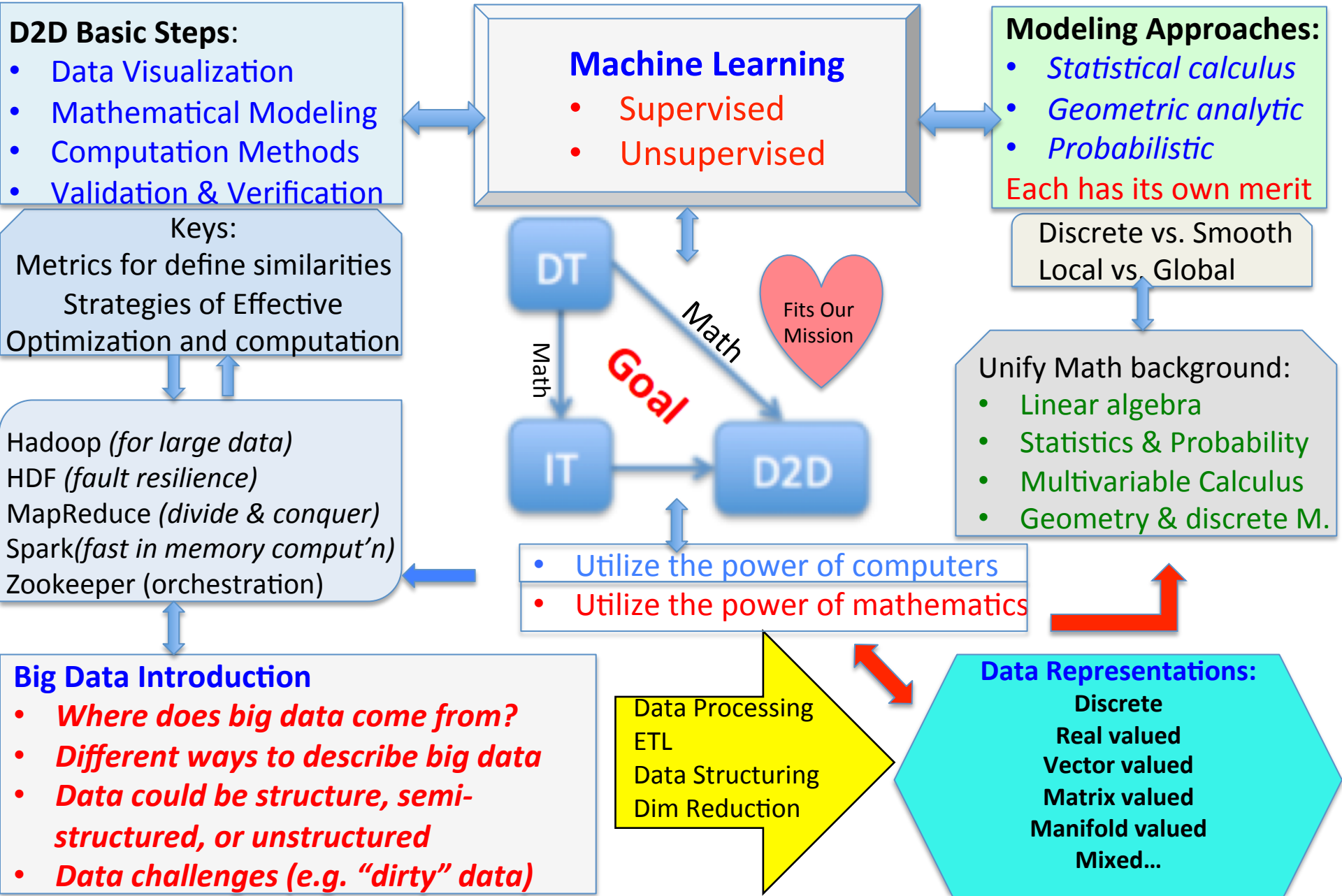
Lecture 1: Introduction of Big Data & Overview of Big Data Analytics

Weiqing Gu
Professor of Mathematics
Director of the Mathematics Clinic

Harvey Mudd College
Summer 2017

https://math189su17.github.io/project.html
https://github.com/math189su17/math189su17.github.io

A Big Picture of Mathematics of Big Data, I



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

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

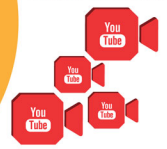
Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

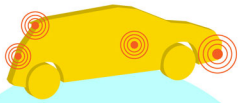
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month



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

in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Data to Decision (D2D)

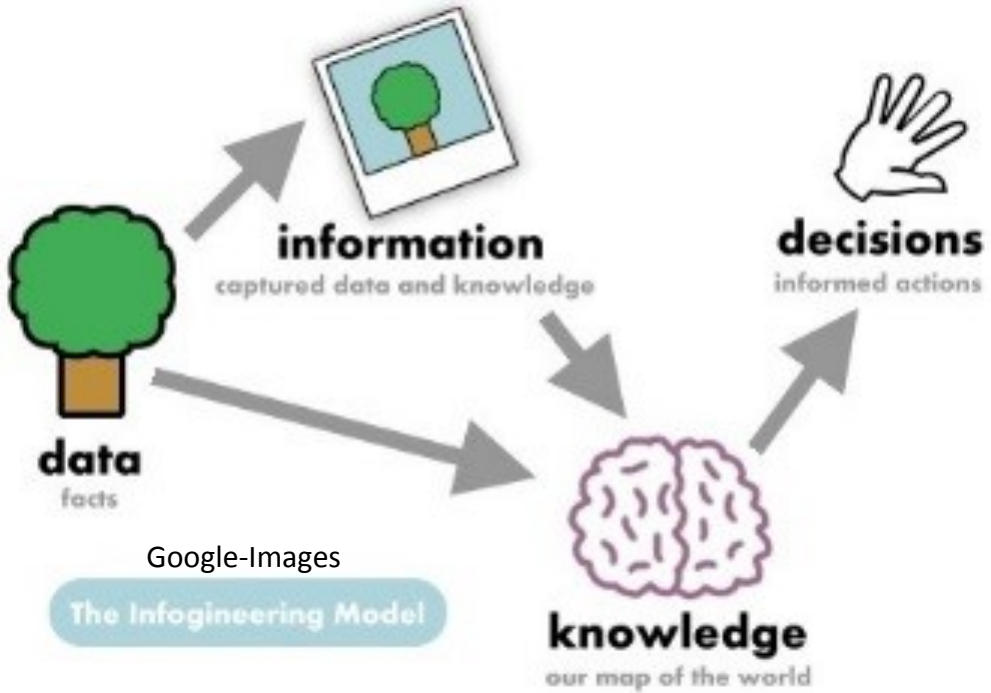
Volume

Velocity

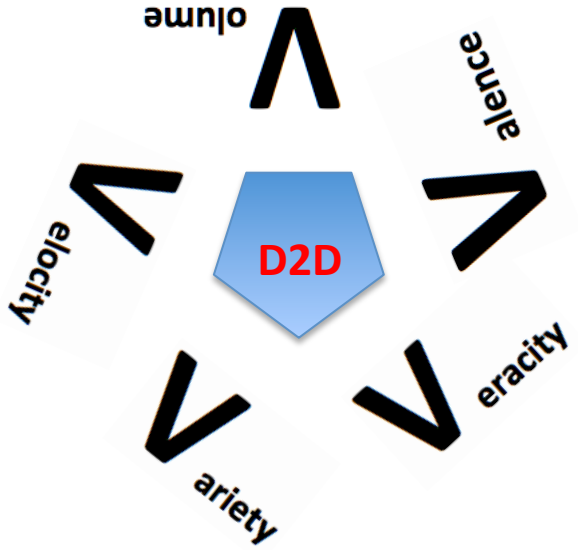
Variety

Veracity

Valence



Google-Images



Second for today: Analytic Approaches

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

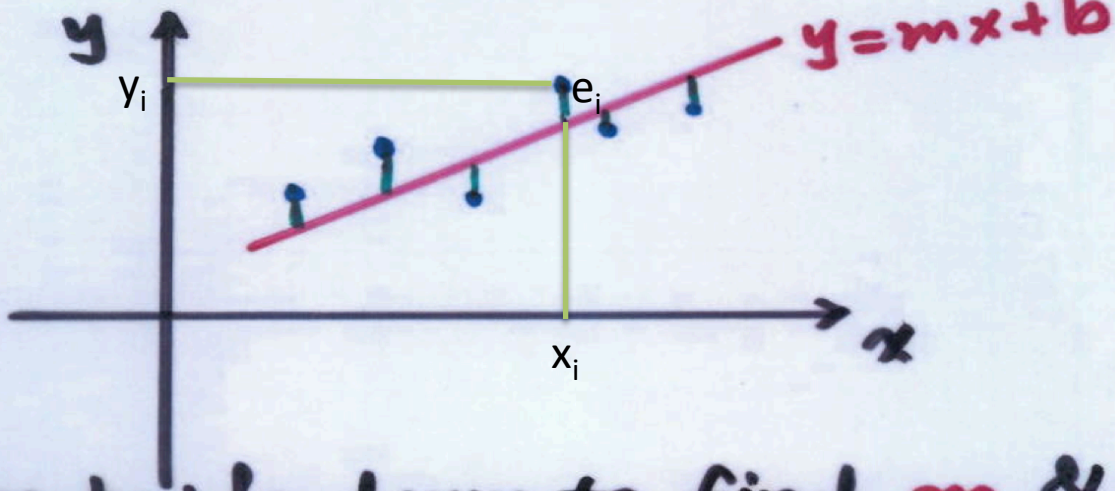
Modeling Approaches:

- *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**:
- $J(m,b) = \sum (e_i)^2 = \sum (y_i - mx_i - b)^2$

L_2 norm

- 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:

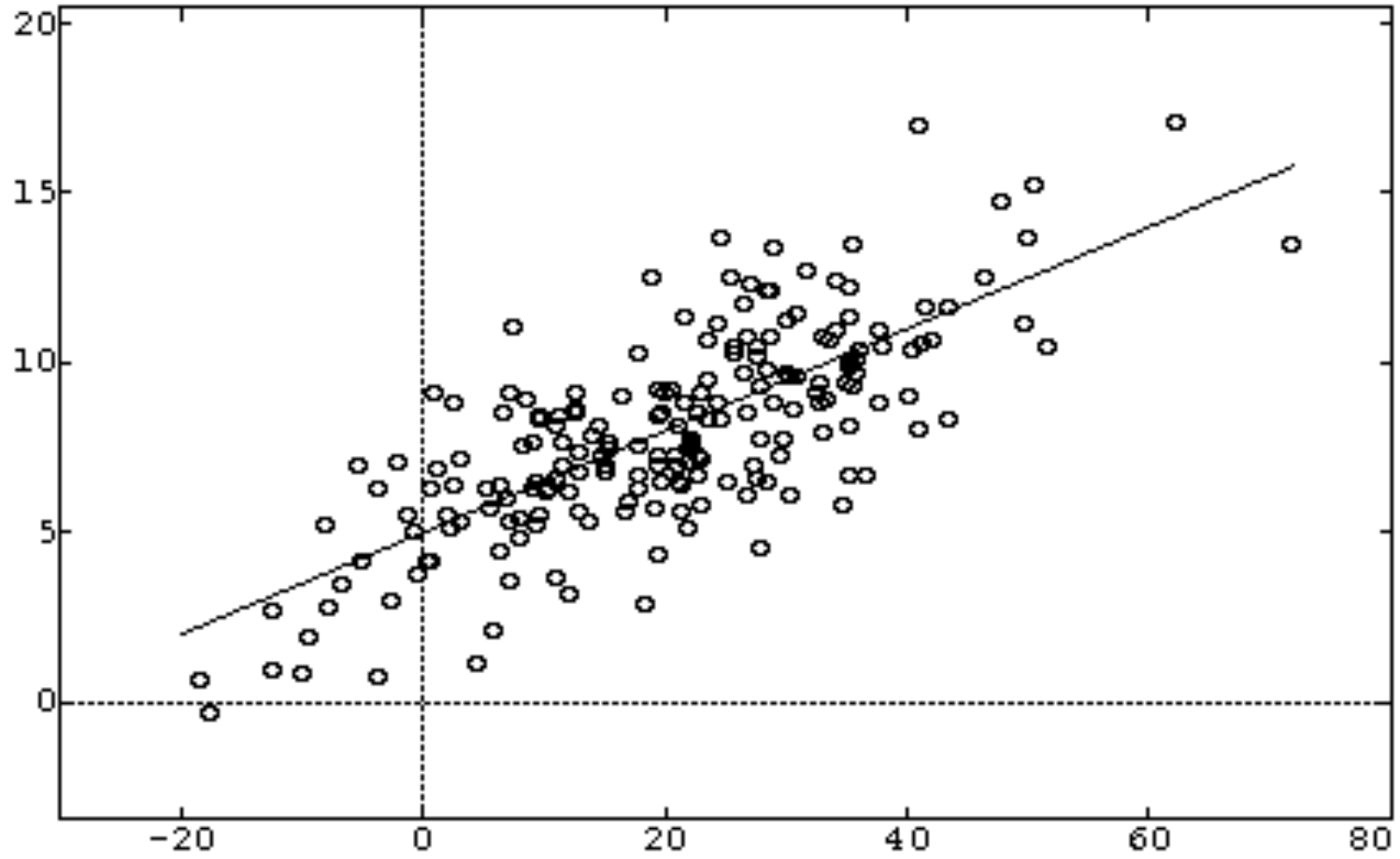
$$m = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

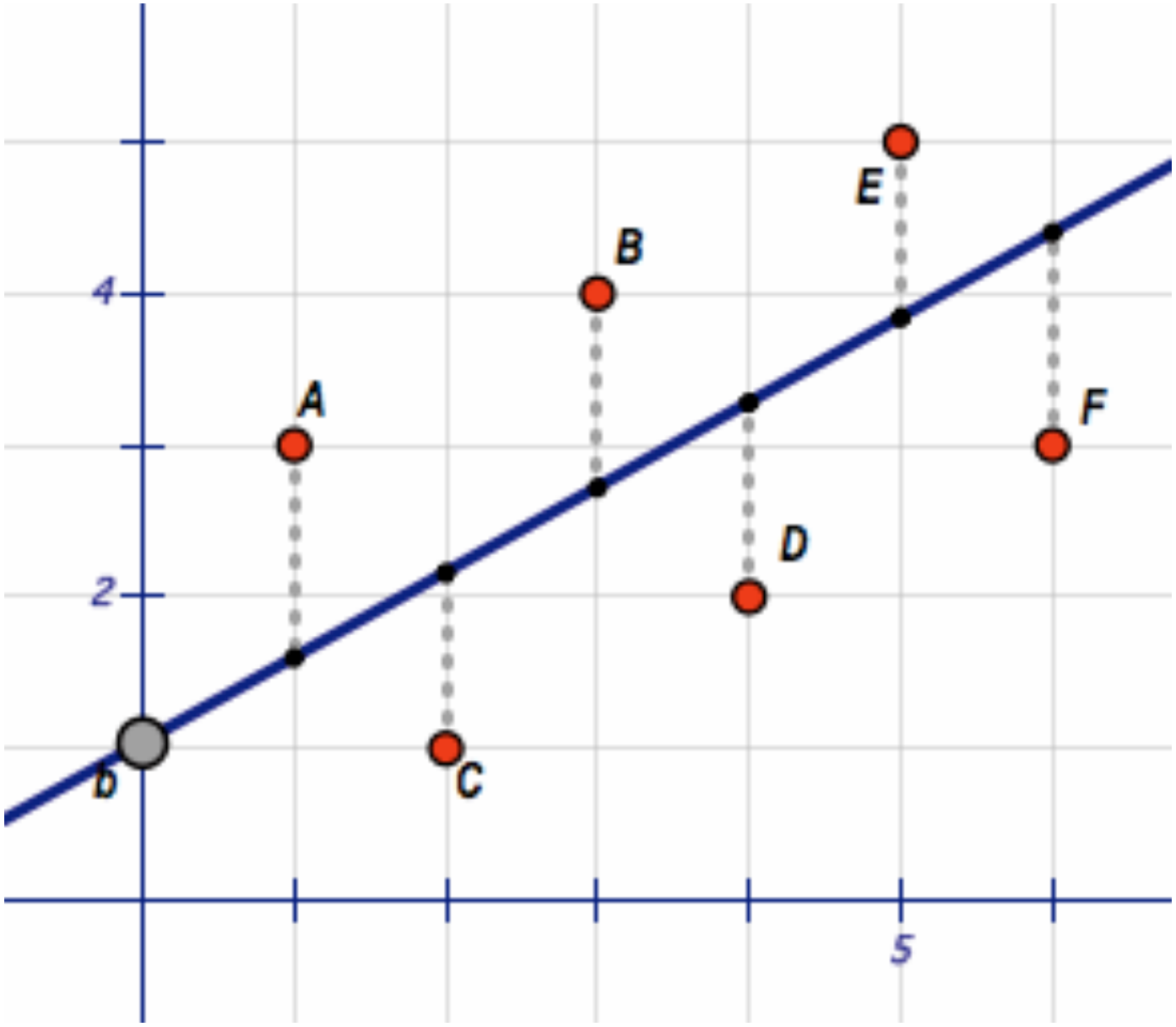
$$b = \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

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

Linear Regression

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




E
B
4
A
F
D
2
b
C
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: $(\mathbf{f} \cdot \mathbf{g})' = \mathbf{f}' \cdot \mathbf{g} + \mathbf{f} \cdot \mathbf{g}'$

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

Homework problem

- Given 4 points as below:

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

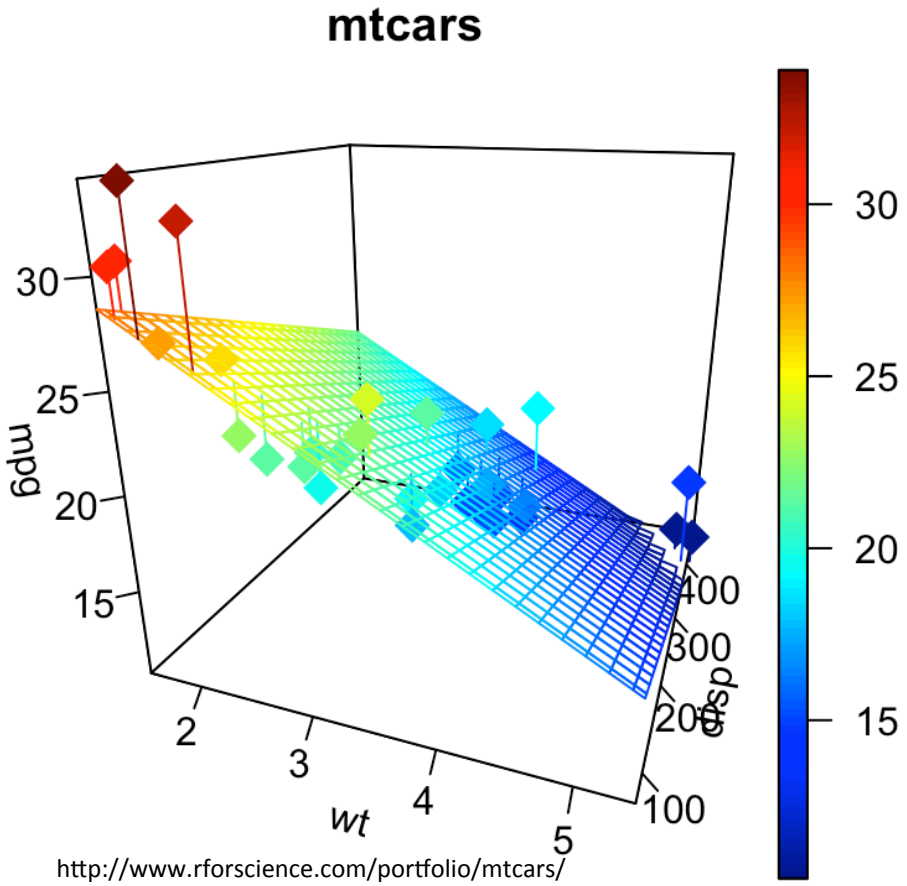
a) Find $y = mx + b$ based on Cramer's rule.

- 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

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

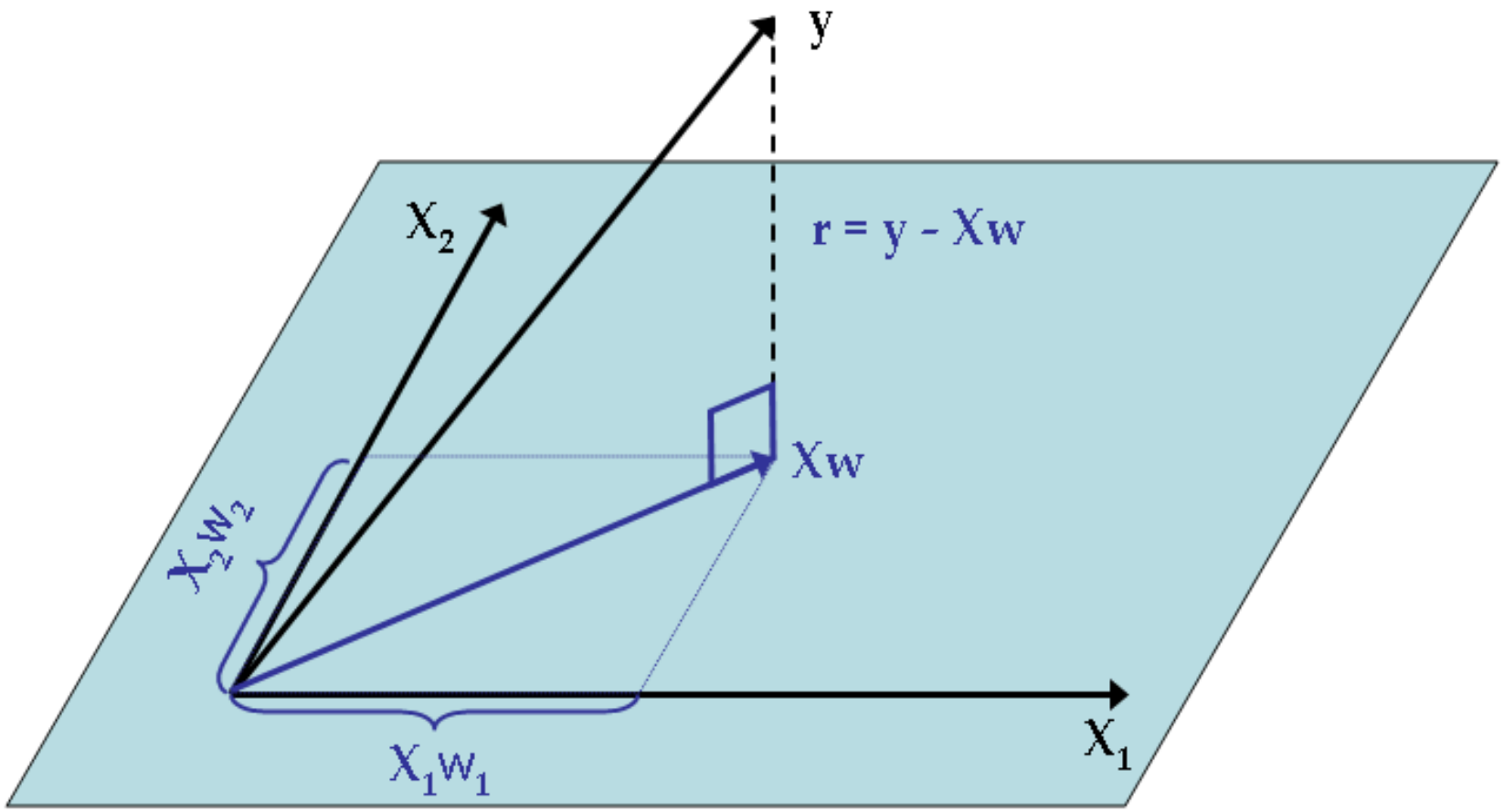
$$\rightarrow \quad \mathbf{r} = (\mathbf{y} - \mathbf{X}\mathbf{w})$$

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!



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

3. Probabilistic Approach (Maximal Likelihood)

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

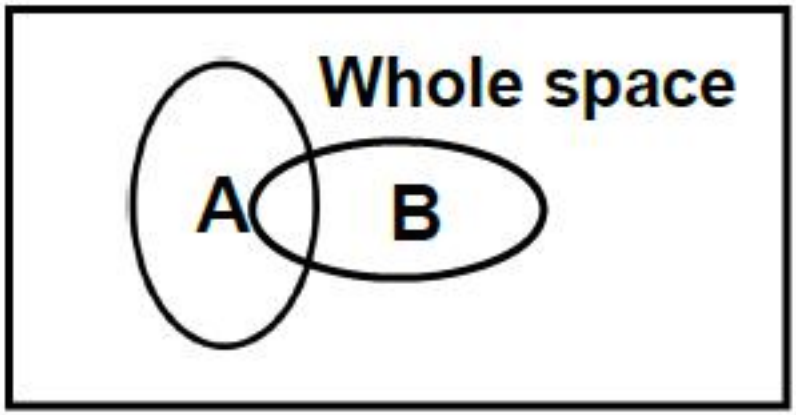
Again we get the same solution!

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

Q: But what's wrong if we use Cramer's rule to solve it?
Or directly use the formula by finding the inverse $X^T X$?

- Back up slides

Visualize Bayes' Theorem



$$P(A) = \frac{\text{A}}{\text{Whole space}} \qquad P(B) = \frac{\text{B}}{\text{Whole space}}$$

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

$$P(A \cap B) = \frac{\text{A} \cap \text{B}}{\text{Whole space}}$$

$$P(A) \times P(B|A) = \frac{\text{A}}{\text{Whole space}} \times \frac{\text{A} \cap \text{B}}{\text{A}} = \frac{\text{A} \cap \text{B}}{\text{Whole space}} = P(A \cap B)$$

$$P(B) \times P(A|B) = \frac{\text{B}}{\text{Whole space}} \times \frac{\text{A} \cap \text{B}}{\text{B}} = \frac{\text{A} \cap \text{B}}{\text{Whole space}} = P(A \cap B)$$

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

Taking Partial Derivatives for different Types of functions

Type 1 : $\mathbb{R} \to \mathbb{R}$, $x \mapsto f(x)$ (one-to-one)

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

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

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

$$\left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n}\right) \triangleq \nabla f \quad \text{(denoted)}$$

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

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

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

$t \mapsto (f_1(t), \cdots, f_m(t)) \triangleq f(t)$

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

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

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

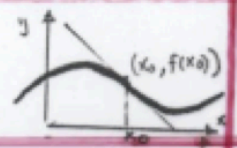
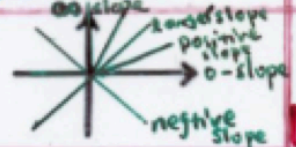
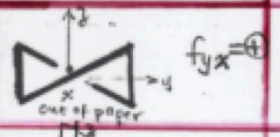
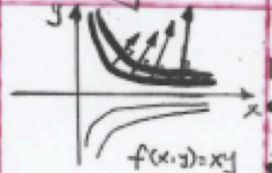
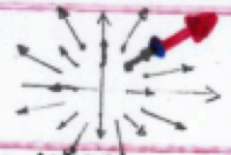
$(x_1, \cdots, x_n) \mapsto (f_1(\vec{x}), \cdots, f_m(\vec{x}))$

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

Derivative 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 functions	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 $(x_0, f(x_0))$: $y=f(x_0)+f'(x_0)(x-x_0)$
$f:\mathbb{R}^2\to\mathbb{R}$ 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 $(a,b,f(a,b))$: $z=f(a,b)+f_x(a,b)(x-a)+f_y(a,b)(y-a)$
$f:\mathbb{R}^2\to\mathbb{R}$ or $\mathbb{R}^n\to\mathbb{R}$	Higher order partials: Here: 2nd 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 of 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 2nd- and 2nd partial deriv's $f_{xy}=f_{yx}$
$f:\mathbb{R}^n\to\mathbb{R}$	Directional derivative of $\mathbb{R}$-valued funcn of 2-variable	$D_{\vec{u}}f(\vec{a})$		Rate change of f in the direction of $\vec{v}$.	Any directional derivative is completely determined by the direction and gradient.	$D_{\vec{u}}f(\vec{a})=\nabla f(\vec{a})\cdot\vec{u}$ ($\vec{u}$ a unit vector!)
$f:\mathbb{R}^n\to\mathbb{R}$	Gradient of real valued function of 2 or 3 variables	∇f $\nabla f(\vec{a})$		$\frac{\nabla f(a,b)}{\|\nabla f(a,b)\|}=\vec{u}$, 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 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	$\operatorname{div}(\vec{F})=\nabla\cdot\vec{F}$		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.	$\operatorname{div}\vec{F}=\nabla\cdot\vec{F}$ $=\frac{\partial F_1}{\partial x_1}+\frac{\partial F_2}{\partial x_2}+\cdots+\frac{\partial F_n}{\partial x_n}$
$\vec{F}:\mathbb{R}^3\to\mathbb{R}^3$	Curl of a vector field	$\operatorname{curl}(\vec{F})=\nabla\times\vec{F}$		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 $\frac{1}{2}\|\operatorname{curl}\vec{F}(\vec{x})\|$ rdians/sec.	$\operatorname{Curl}(\vec{F})=\nabla\times\vec{F}$ $=\begin{vmatrix}\vec{i}&\vec{j}&\vec{k}\\ \frac{\partial}{\partial x}&\frac{\partial}{\partial y}&\frac{\partial}{\partial z}\\ \cdots&\cdots&\cdots\end{vmatrix}$

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: $(\mathbf{f} \cdot \mathbf{g})' = \mathbf{f}' \cdot \mathbf{g} + \mathbf{f} \cdot \mathbf{g}'$

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