

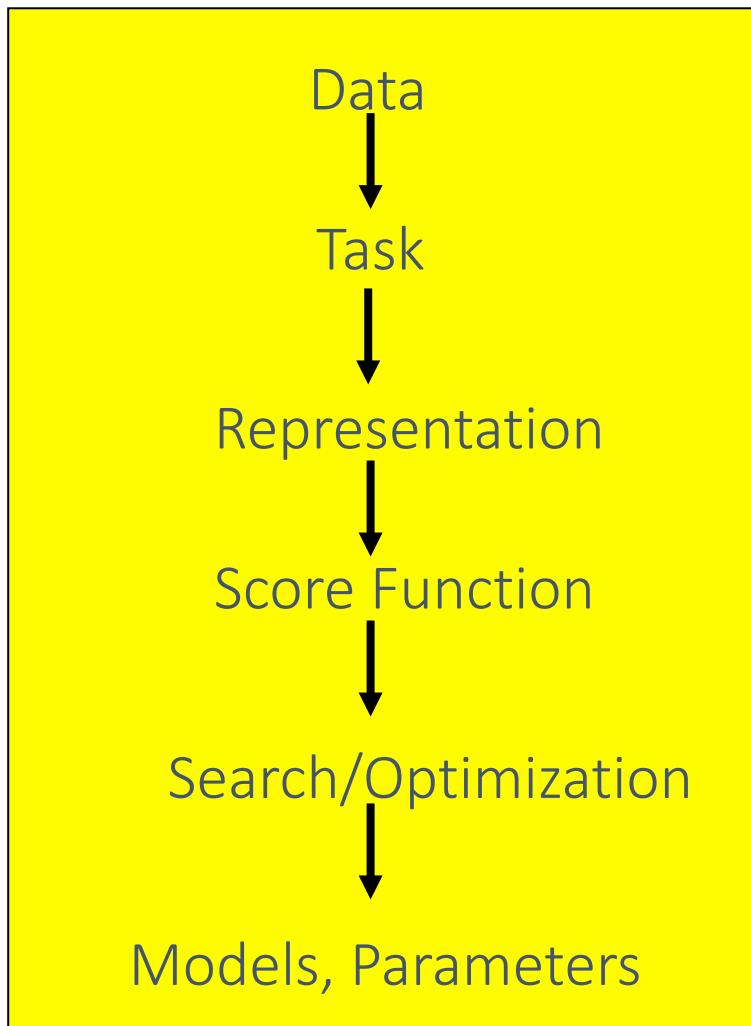
UVA CS 4774: Machine Learning

Lecture 3: Linear Regression Basics

Dr. Yanjun Qi

University of Virginia
Department of Computer Science

Machine Learning in a Nutshell



ML grew out of work in AI

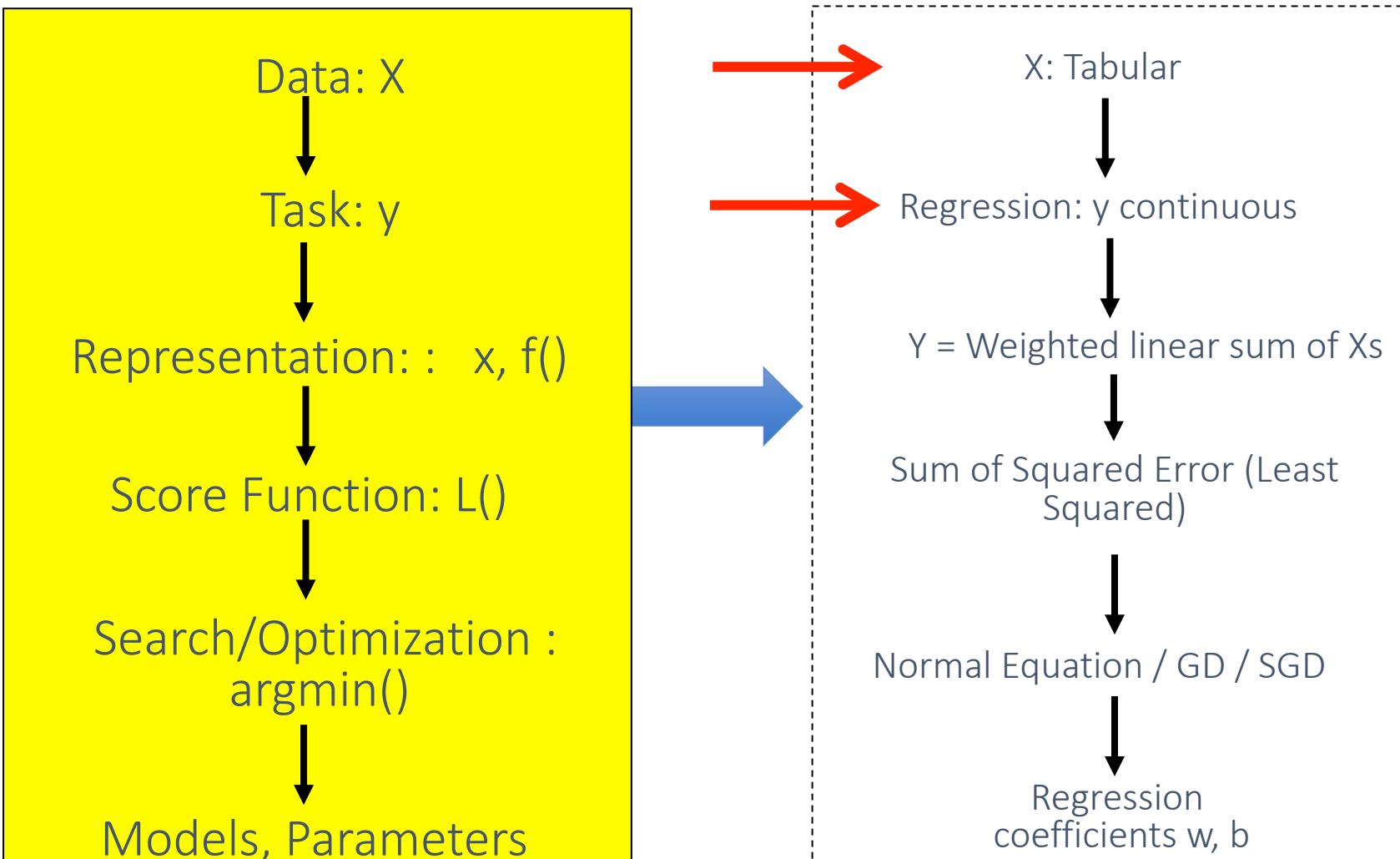
Optimize a performance criterion using example data or past experience,

Aiming to generalize to unseen data

Rough Sectioning of this Course

- 
- 1. Basic Supervised Regression + on Tabular Data
 - 2. Basic Deep Learning + on 2D Imaging Data
 - 3. Advanced Supervised learning + on Tabular Data
 - 4. Generative and Deep + on 1D Sequence Text Data
 - 5. Not Supervised + Mostly on Tabular Data

Today: Multivariate Linear Regression **in a Nutshell**



X_1	X_2	X_3	Y

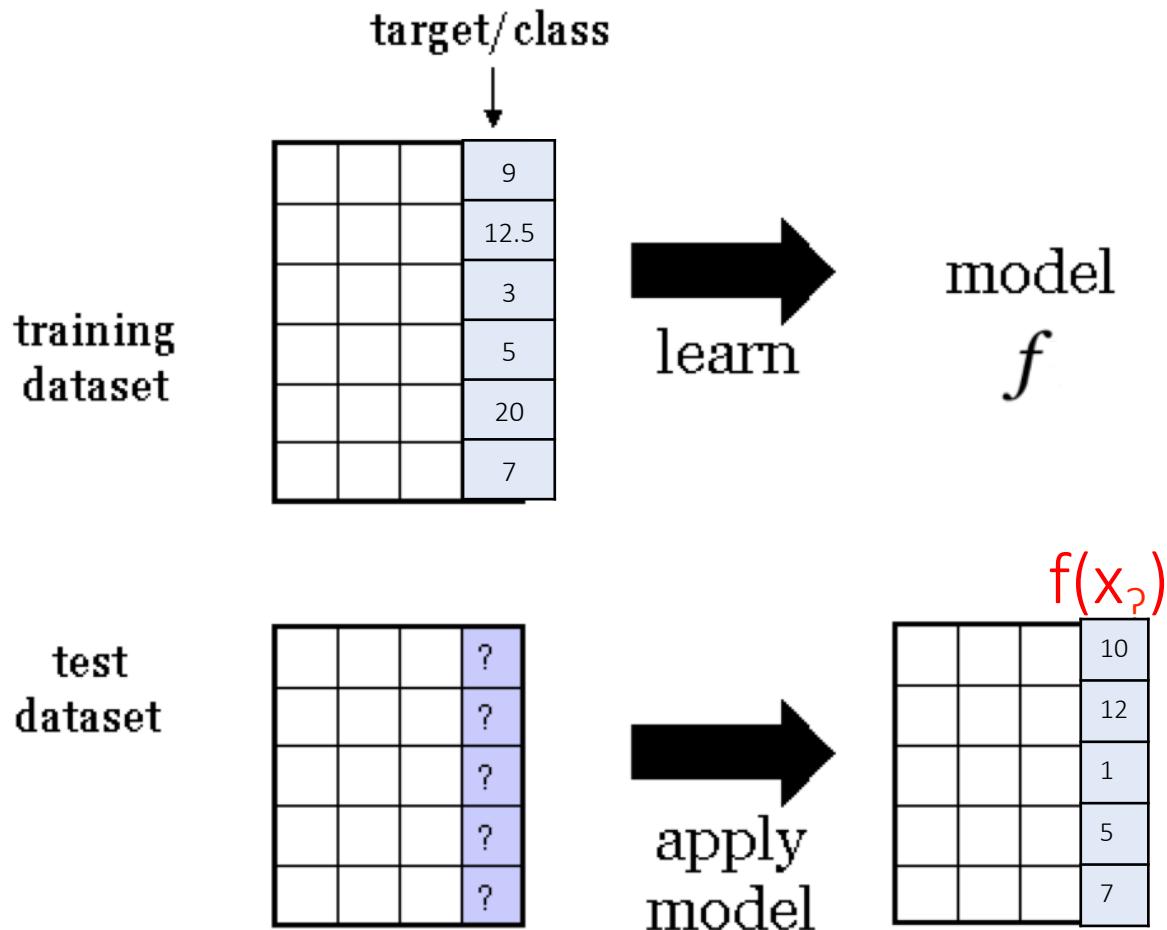
Tabular Dataset for regression

$$f : [X] \longrightarrow [Y]$$

continuous
valued
variable

- Data/*points/instances/examples/samples/records*: [[rows](#)]
- Features/*attributes/dimensions/independent variables/covariates/predictors/regressors*: [[columns, except the last](#)]
- Target/*outcome/response/label/dependent variable*: special column to be predicted [[last column](#)]

SUPERVISED Regression



Training dataset
consists of
input-output
pairs

- Target Y: **continuous** target variable

$$\mathbf{x} \downarrow train \uparrow = [\mathbf{x} \downarrow 1 \uparrow T \& \dots \& \mathbf{x} \downarrow 2 \uparrow T \& \dots \& \dots \& \dots \& \mathbf{x} \downarrow n \uparrow T \& \dots]$$

$$\vec{y}_{train} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

training dataset

$$\mathbf{x} \downarrow test \uparrow = [\mathbf{x} \downarrow n+1 \uparrow T \& \dots \& \mathbf{x} \downarrow n+2 \uparrow T \\ \& \dots \& \dots \& \dots \& \mathbf{x} \downarrow n+m \uparrow T \& \dots]$$

$$\vec{y}_{test} = \begin{bmatrix} y_{n+1} \\ y_{n+2} \\ \vdots \\ y_{n+m} \end{bmatrix}$$

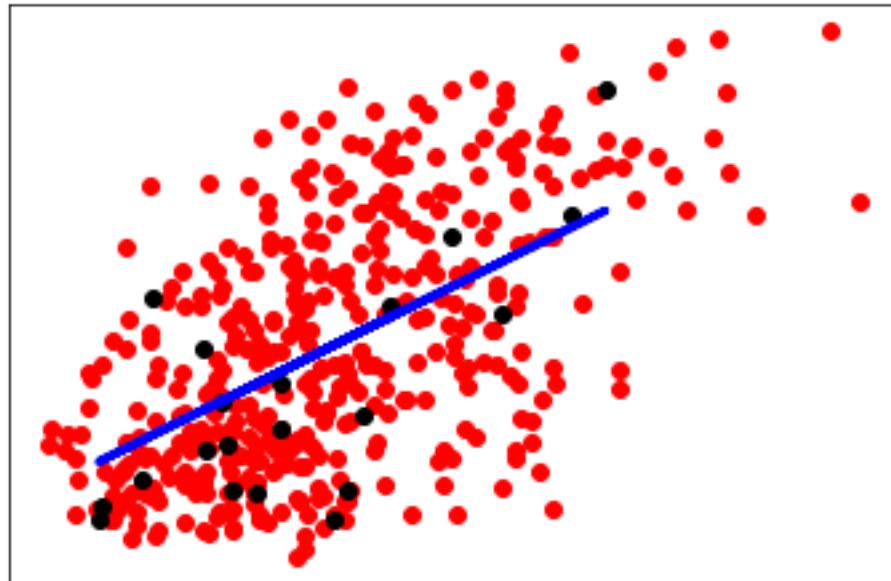
test dataset

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

<https://colab.research.google.com/drive/1mI-PFB2J1UVIp6JjLhQKEPirPZIQnDH7?usp=sharing>

I will code-run through: Scikit-Learn Linear Regression Example

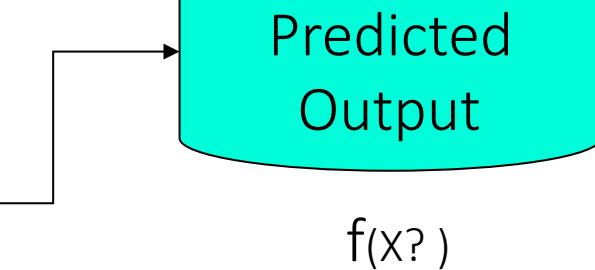
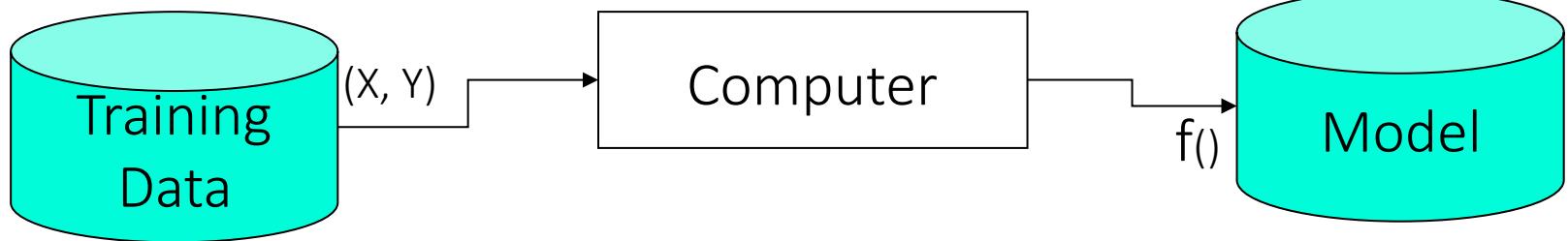
Adapted from: https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html



Two Modes of Machine Learning

Consists of **input-output** pairs

Training



Testing

Non-Linear Regression: Example on COVID19

<https://www.kaggle.com/notebooks>



The Story of COVID-19 in India EDA and Prediction

Python notebook using data from [multiple data sources](#) · 23,372 views · 4mo ago · beginner, data visualization, exploratory data analysis, +1 more



284

<https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge>

I will code-run through: The Story of COVID-19 in India EDA and Prediction
Adapted from:

<https://www.kaggle.com/nitishabharathi/the-story-of-covid-19-in-india-eda-and-prediction>

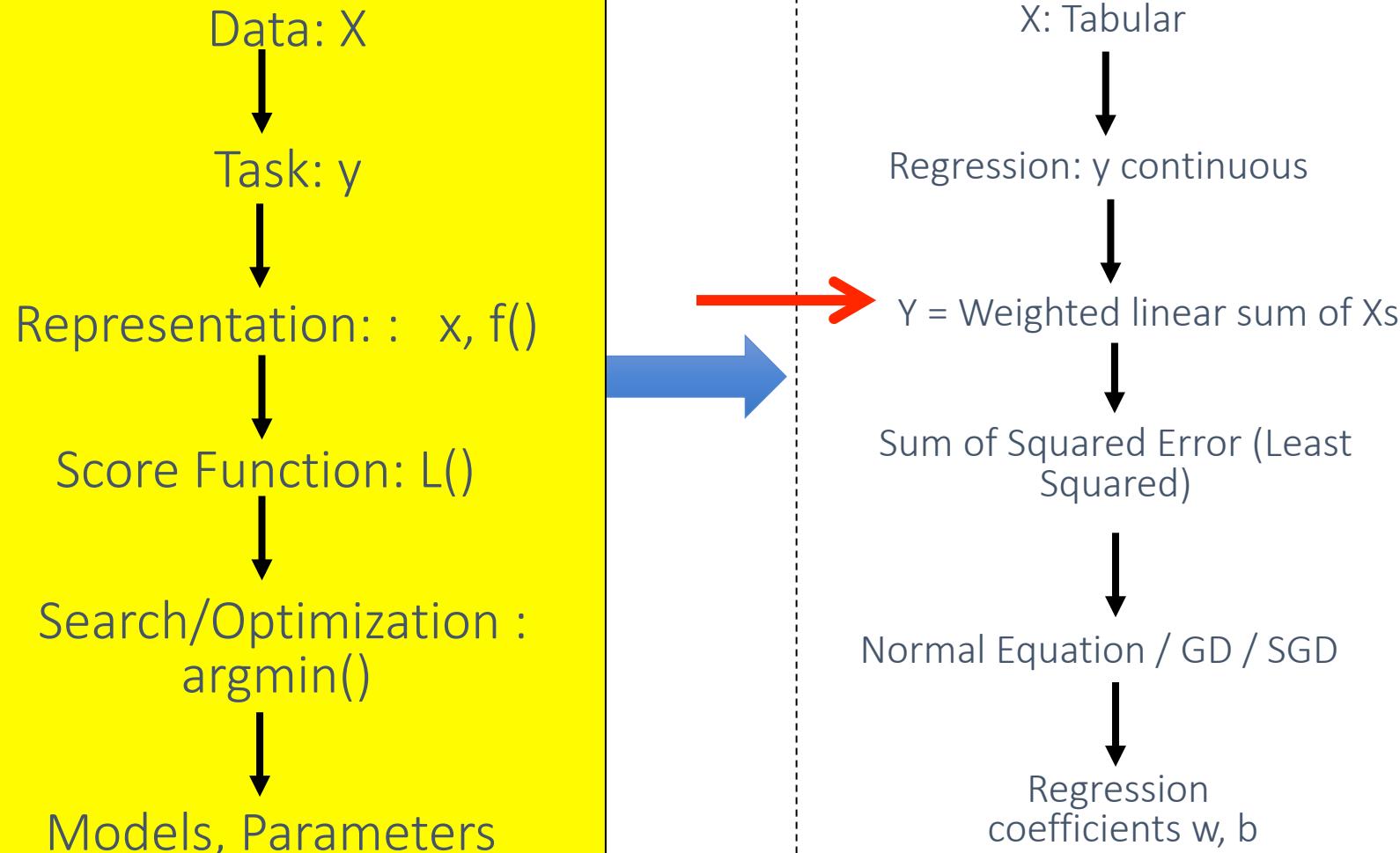
Table of Contents

- [World Updates](#)
- [India Updates](#)
- [Age Group/Gender Analysis](#)
- [The Spike in India](#)
- [Testing in India](#)
- [Statewise Insights](#)
- [Statewise Testing and Healthcare Insights](#)
- [Economic Impact of COVID-19 on India](#)
- [Prediction](#)
 - [Growth Factor](#)
 - [Prophet Model](#)
 - [ARIMA \(5,2,0\)](#)
 - [LightGBM](#)
 - [Random Forest Regressor](#)
 - [XGBoost Regressor](#)
- [Interesting Insights from other Sources](#)

Thank You



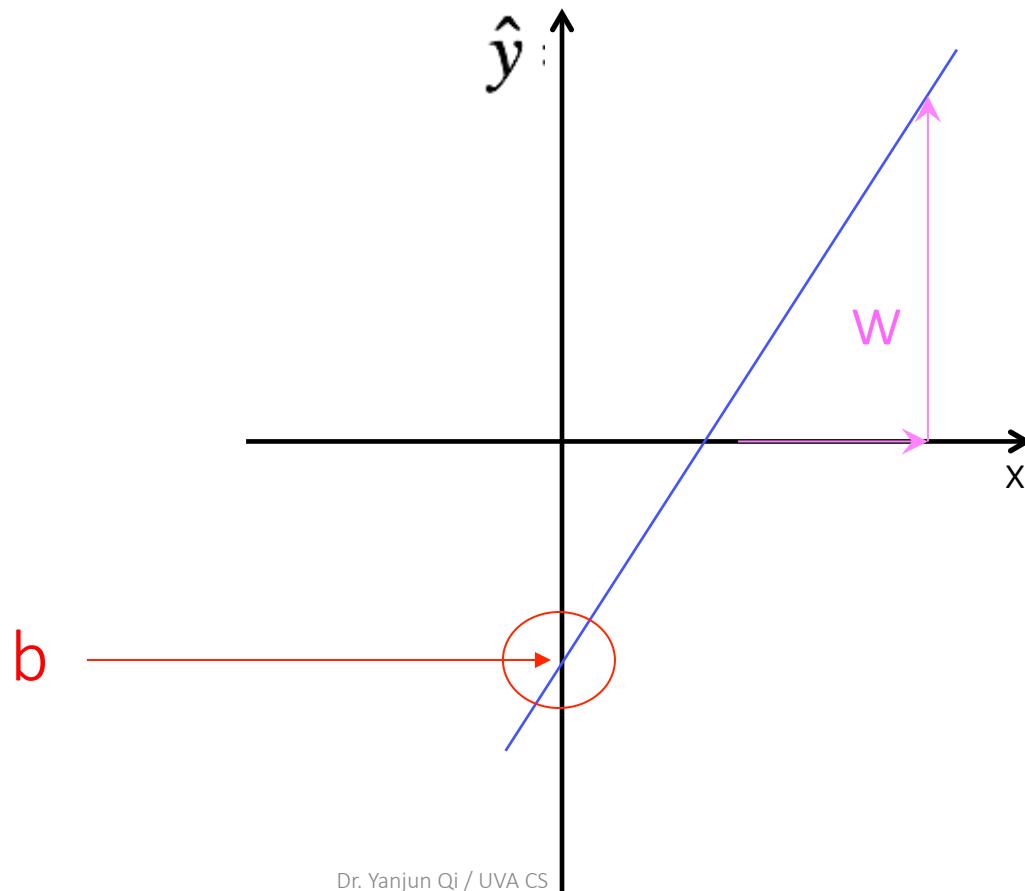
Today: Multivariate Linear Regression **in a Nutshell**



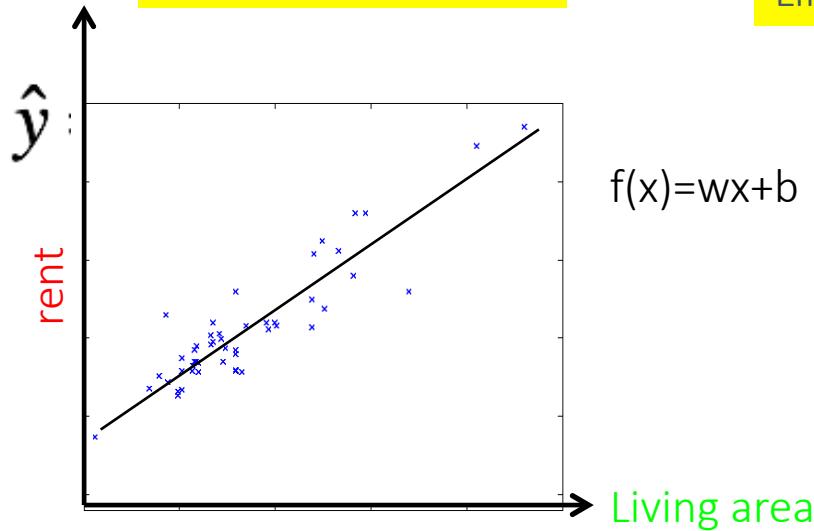
Review: $f(x)$ is Linear when X is single variable

- $f(x)=wx+b?$

A slope of 2 (i.e. $w=2$) means that every 1-unit change in X yields a 2-unit change in Y.



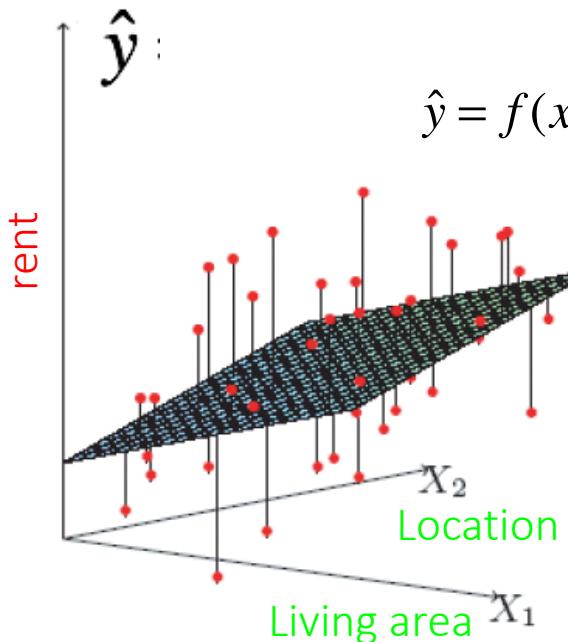
e.g., Living area as X



Linear Regression: Y as Weighted linear sum of Xs

1D case ($\mathcal{X} = \mathbb{R}$): a line

e.g., (Living area, Location) as X



$\mathcal{X} = \mathbb{R}^2$: a plane

Linear Supervised Regression

$$f: X \rightarrow Y$$

Linear Regression Models

Linear Regression: Y as Weighted linear sum of Xs

$$\hat{y} = f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

=> Features x_i :

e.g., Living area, distance to campus, # bedroom ...

=> Target y :

e.g., Rent (a continuous variable)

A Concise Notation: via Vector/Matrix Product

- Represent each data sample \mathbf{x} as a column vector, plus a pseudo feature
 - We add a pseudo "feature" $x_0=1$ (this is the intercept term), and RE-define the feature vector to be:
- The parameter vector $\boldsymbol{\theta}$ is also a column vector

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} \quad \rightarrow \quad \hat{y} = f(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\theta} = \boldsymbol{\theta}^T \mathbf{x}$$

$$\hat{y} = f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

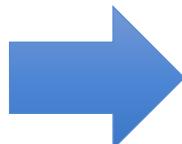
A Concise Notation: via Vector/Matrix Product

- Represent each data sample \mathbf{x} as a column vector, plus a pseudo feature
 - We add a pseudo "feature" $x_0=1$ (this is the intercept term), and RE-define the feature vector to be:

$$\mathbf{x}^T = [(x_0=1), x_1, x_2, \dots, x_p]$$

- The parameter vector θ is also a column vector

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}$$



$$\hat{y} = f(\mathbf{x})$$

$$= \mathbf{x}^T \theta = \theta^T \mathbf{x}$$

$$\begin{aligned} \hat{y} = f(x) &= \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_p x_p \\ &= \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_p x_p \end{aligned}$$

$$\hat{y} = f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

Review:

- Dot (or Inner) Product of two vectors $\langle \mathbf{x}, \mathbf{a} \rangle$

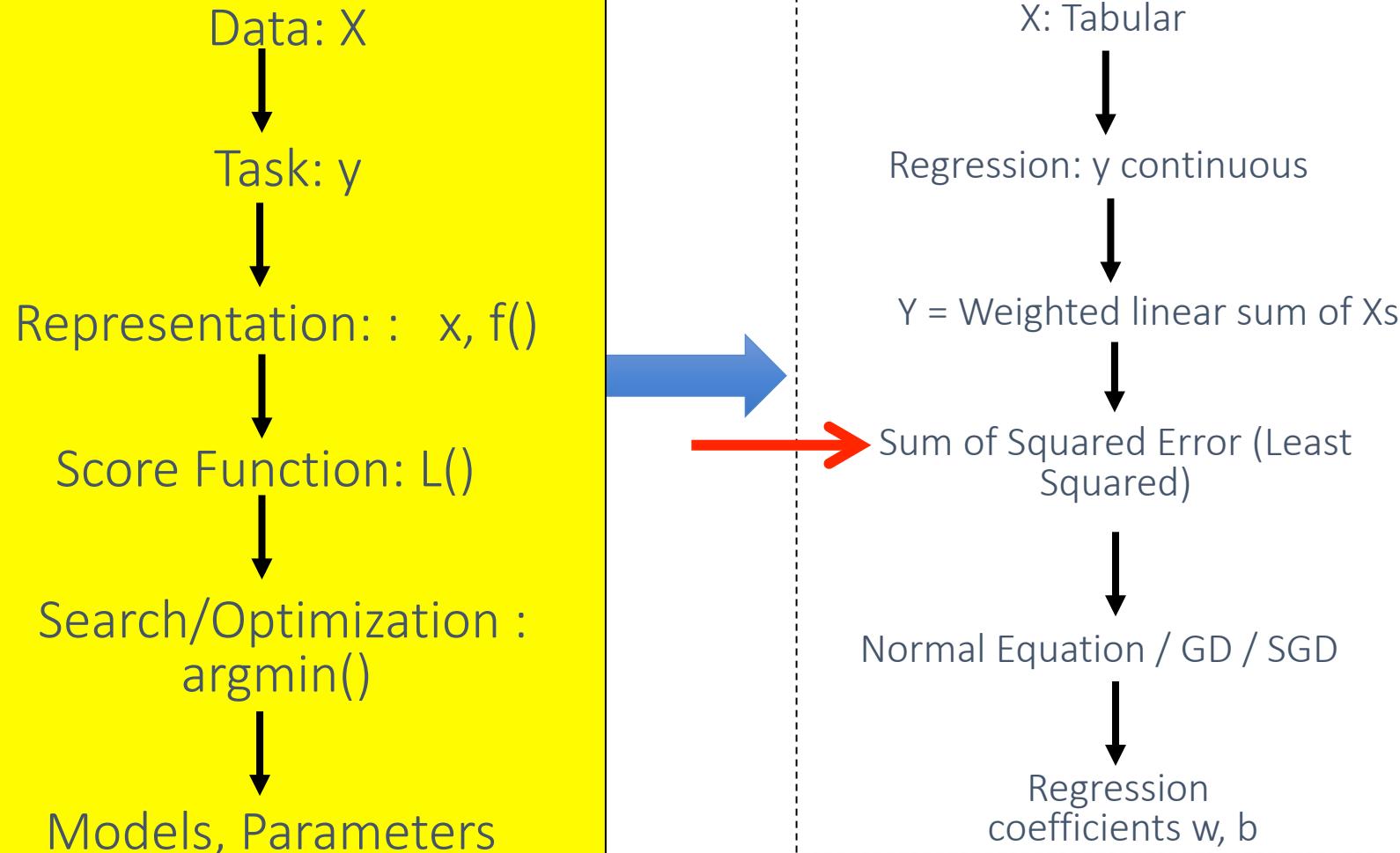
is the sum of products of elements in similar positions for the two vectors

$$\langle \mathbf{x}, \Theta \rangle = \langle \Theta, \mathbf{x} \rangle$$

Linear Regression: Y as Weighted linear sum of Xs

$$\begin{aligned}\hat{y} &= f(\mathbf{x}) \\ &= \mathbf{x}^T \theta = \theta^T \mathbf{x}\end{aligned}$$

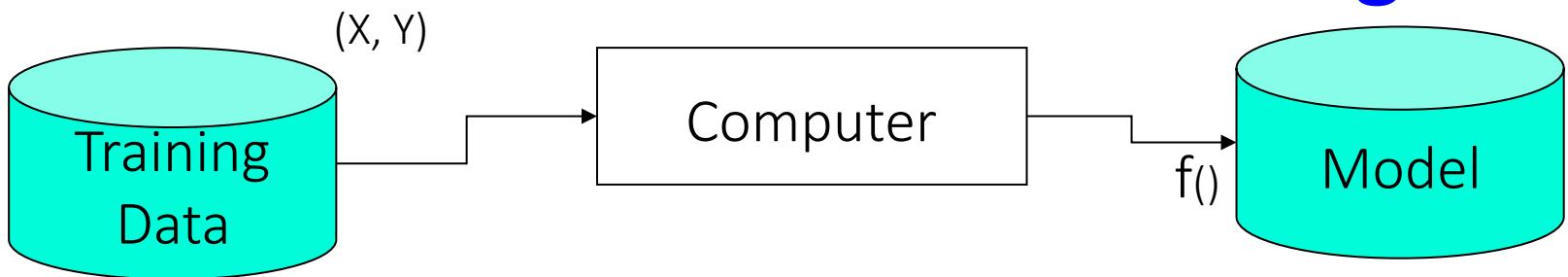
Today: Multivariate Linear Regression **in a Nutshell**



Training Modes of Machine Learning

Consists of **input-output** pairs

Training



```
# Split the data into training/testing sets
diabetes_X_train = diabetes_X[:-20]
diabetes_X_test = diabetes_X[-20:]

# Split the targets into training/testing sets
diabetes_y_train = diabetes_y[:-20]
diabetes_y_test = diabetes_y[-20:]

# Create linear regression object
regr = linear_model.LinearRegression()

# Train the model using the training sets
regr.fit(diabetes_X_train, diabetes_y_train)
```

Basic Concepts

- Training (i.e. learning parameters w, b)
 - Training set includes
 - available examples x_1, \dots, x_n
 - available corresponding labels y_1, \dots, y_n
 - Find (w, b) by minimizing loss / Cost function $L()$
 - (i.e. difference between y and $f(x)$ on available examples in training set)

$$(w, b) = \underset{w, b}{\operatorname{argmin}} \sum_{i=1}^L \ell(f(x_i), y_i)$$

Loss/Cost function for Regression L()

SSE: Sum of squared error

- Our goal is to search for the optimal θ that minimize the following lost / cost function:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2$$

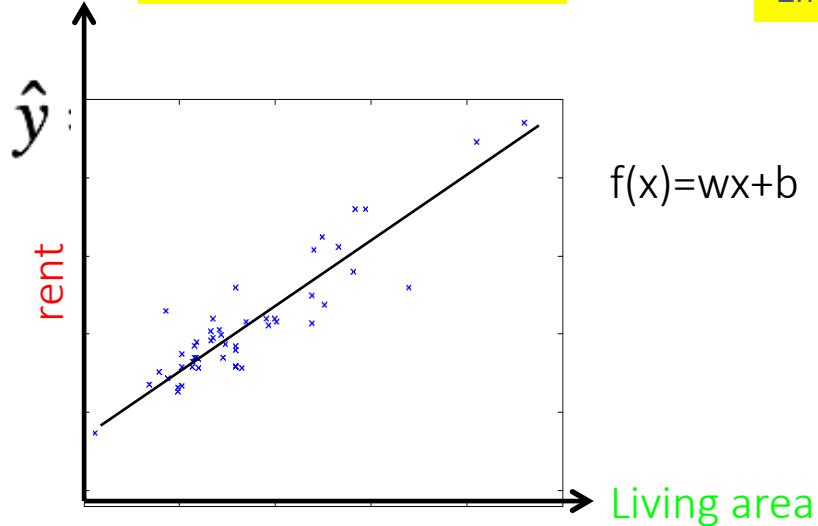
$$\hat{y} = f(\mathbf{x})$$

$$= \mathbf{x}^T \theta = \theta^T \mathbf{x}$$

One concrete example

One concrete example

e.g., Living area as X

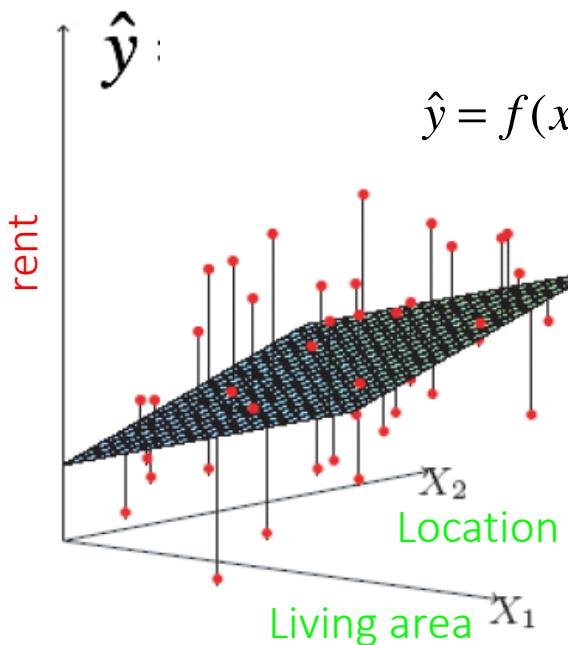


Linear Regression: Y as Weighted linear sum of Xs

$$J(\theta) = \frac{1}{2} \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2$$

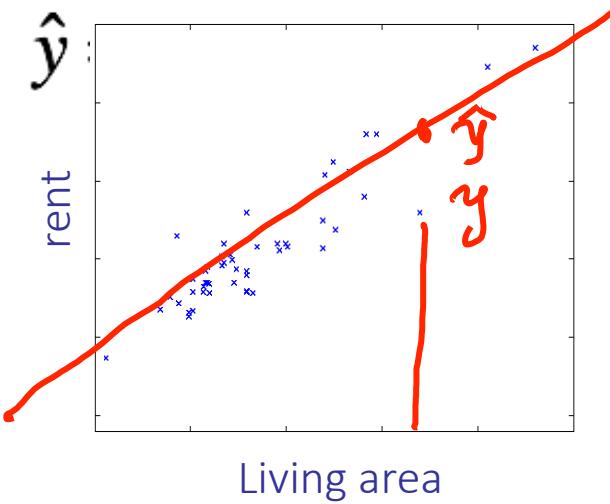
1D case ($\mathcal{X} = \mathbb{R}$): a line

e.g., (Living area, Location) as X



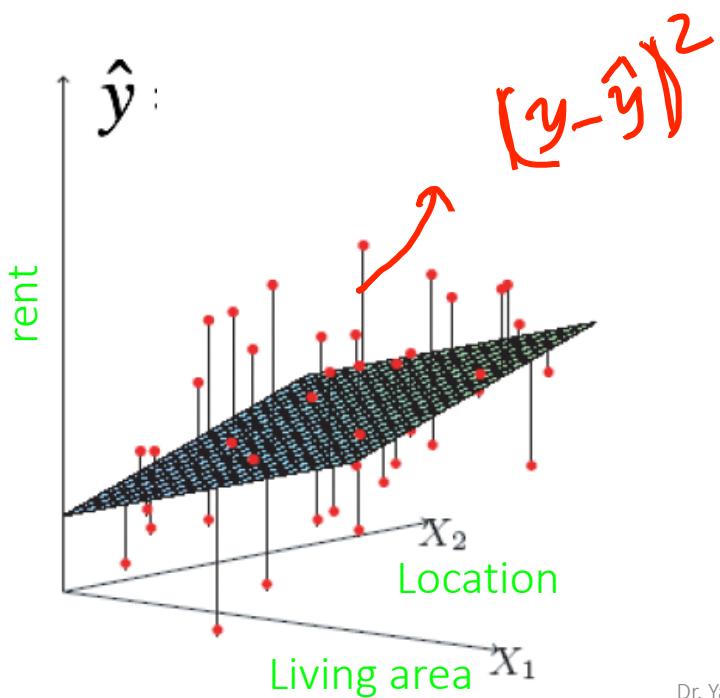
$$J(\theta) = \frac{1}{2} \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2$$

$\mathcal{X} = \mathbb{R}^2$: a plane



$$J(\theta) = \frac{1}{2} \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2$$

1D case ($\mathcal{X} = \mathbb{R}$): a line



$$J(\theta) = \frac{1}{2} \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2$$

$\mathcal{X} = \mathbb{R}^2$: a plane

Now the loss function: via A Concise Notation

- Using matrix form, we get the following general representation of the linear regression function:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2$$

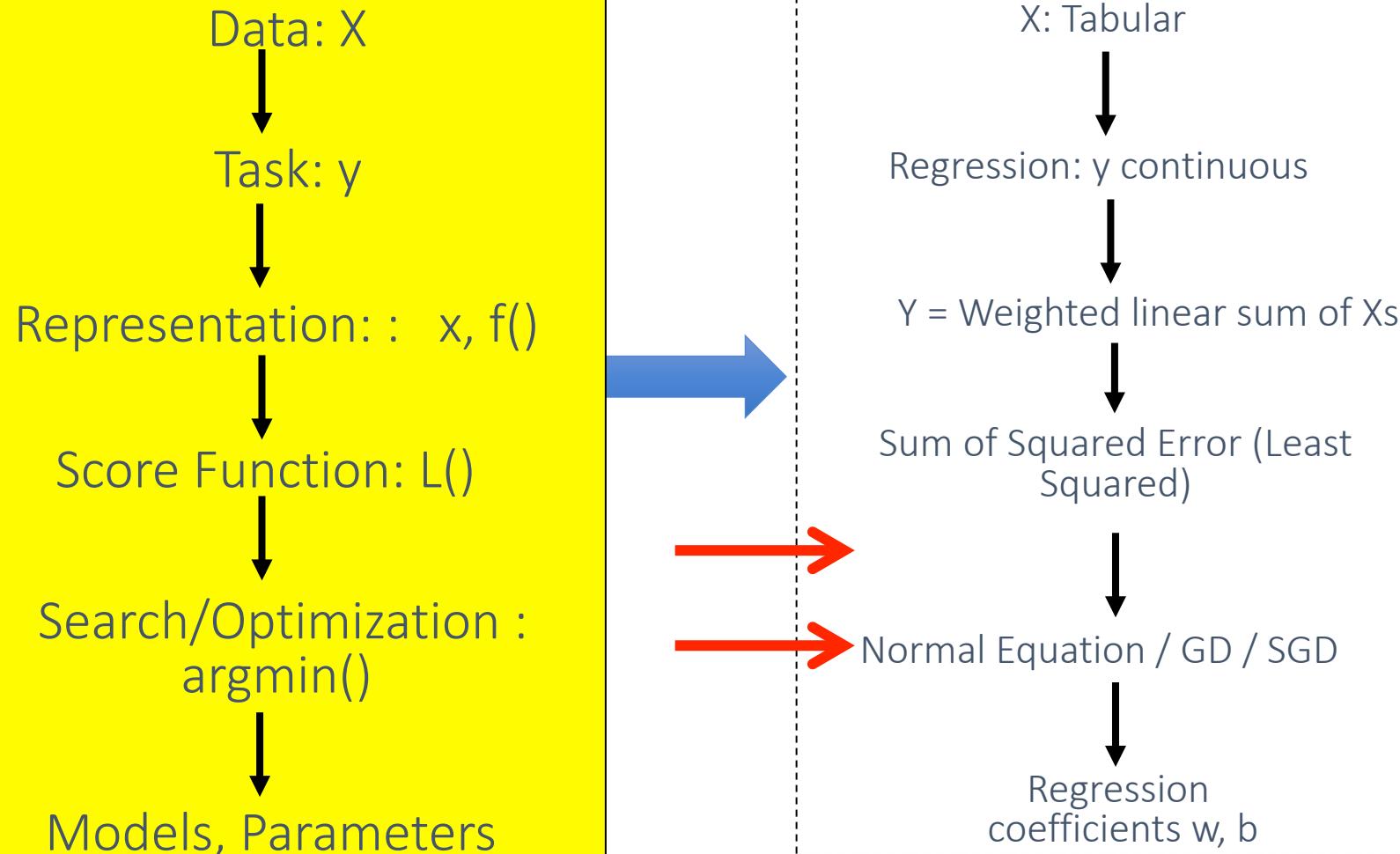
$$\mathbf{X} = \begin{bmatrix} \cdots & \mathbf{x}_1^T & \cdots \\ \cdots & \mathbf{x}_2^T & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \mathbf{x}_n^T & \cdots \end{bmatrix}$$

$$\begin{aligned} J(\theta) &= \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i^T \theta - y_i)^2 \\ &= \frac{1}{2} (\mathbf{X}\theta - \bar{\mathbf{y}})^T (\mathbf{X}\theta - \bar{\mathbf{y}}) \\ &= \frac{1}{2} (\theta^T \mathbf{X}^T \mathbf{X}\theta - \theta^T \mathbf{X}^T \bar{\mathbf{y}} - \bar{\mathbf{y}}^T \mathbf{X}\theta + \bar{\mathbf{y}}^T \bar{\mathbf{y}}) \end{aligned}$$

Thank You



Today: Multivariate Linear Regression **in a Nutshell**



Default Vector Form is the Column Form

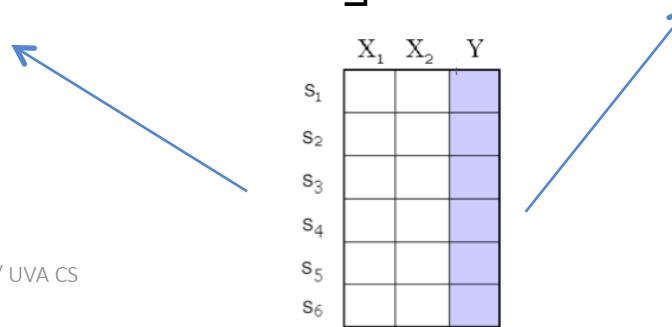
$$\vec{x}_i = \begin{bmatrix} 1 \\ x_{i,1} \\ x_{i,2} \\ x_{i,3} \\ \vdots \\ x_{i,p} \end{bmatrix}$$

$$\vec{x}_i^T = [1, x_{i,1}, x_{i,2}, \dots, x_{i,p}]$$

Training Set in Matrix Form

- the whole Training set (with n samples) as matrix form :

$$\mathbf{X} = \begin{bmatrix} \cdots & \mathbf{x}_1^T & \cdots \\ \cdots & \mathbf{x}_2^T & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \mathbf{x}_n^T & \cdots \end{bmatrix} = \begin{bmatrix} x_{1,0} & x_{1,1} & \cdots & x_{1,p} \\ x_{2,0} & x_{2,1} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,0} & x_{n,1} & \cdots & x_{n,p} \end{bmatrix} \quad \bar{\mathbf{y}}_{train} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$



	X_1	X_2	Y
s_1			
s_2			
s_3			
s_4			
s_5			
s_6			

Training Set in Matrix Form

- the whole Training set (with n samples) as matrix form :

$$\mathbf{X}\theta = \begin{bmatrix} \cdots & \mathbf{x}_1^T & \cdots \\ \cdots & \mathbf{x}_2^T & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \mathbf{x}_n^T & \cdots \end{bmatrix} = \begin{bmatrix} \theta \end{bmatrix} = \begin{bmatrix} x_{1,0} & x_{1,1} & \cdots & x_{1,p} \\ x_{2,0} & x_{2,1} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,0} & x_{n,1} & \cdots & x_{n,p} \end{bmatrix} \quad \bar{y}_{train} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$\cancel{\mathbf{X}}\theta = \begin{bmatrix} x_1^T \theta \\ \vdots \\ x_n^T \theta \end{bmatrix}$

	X ₁	X ₂	Y
s ₁			
s ₂			
s ₃			
s ₄			
s ₅			
s ₆			

Cont: Loss function: via A Concise Notation

- Using matrix form, we get the following general representation of the linear regression function:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i^T \theta - y_i)^2$$

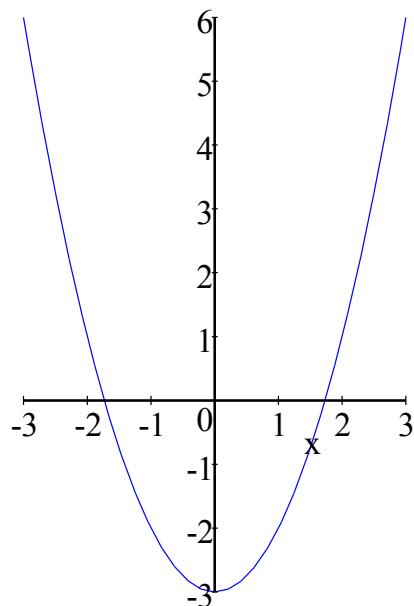
$$= \frac{1}{2} (\mathbf{X}\theta - \bar{\mathbf{y}})^T (\mathbf{X}\theta - \bar{\mathbf{y}})$$

$$= \frac{1}{2} (\theta^T \mathbf{X}^T \mathbf{X}\theta - \theta^T \mathbf{X}^T \bar{\mathbf{y}} - \bar{\mathbf{y}}^T \mathbf{X}\theta + \bar{\mathbf{y}}^T \bar{\mathbf{y}})$$

$$\mathbf{X} = \begin{bmatrix} \cdots & \mathbf{x}_1^T & \cdots \\ \cdots & \mathbf{x}_2^T & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \mathbf{x}_n^T & \cdots \end{bmatrix} \theta$$

$$\mathbf{a} = (\mathbf{X}\theta - \bar{\mathbf{y}}) = \underbrace{\mathbf{X}\theta}_{n \times p} - \underbrace{\bar{\mathbf{y}}}_{p \times 1} = \begin{bmatrix} \mathbf{x}_1^T \theta - y_1 \\ \mathbf{x}_2^T \theta - y_2 \\ \vdots \\ \mathbf{x}_n^T \theta - y_n \end{bmatrix}$$

$$\mathbf{a}^T \mathbf{a} = \sum_{i=1}^n a_i^2$$



Review (IV): Derivative of a Quadratic Function

$$L(w) = w^2 - 3$$

$$L'(w) = 2w$$

$$\operatorname{argmin}_w L(w) = 0$$

This quadrative (convex) function is minimized @ the unique point whose derivative (slope) is zero.

→ When we find zeros of the derivative of this function, we also find the minima (or maxima) of that function.

Critical Points

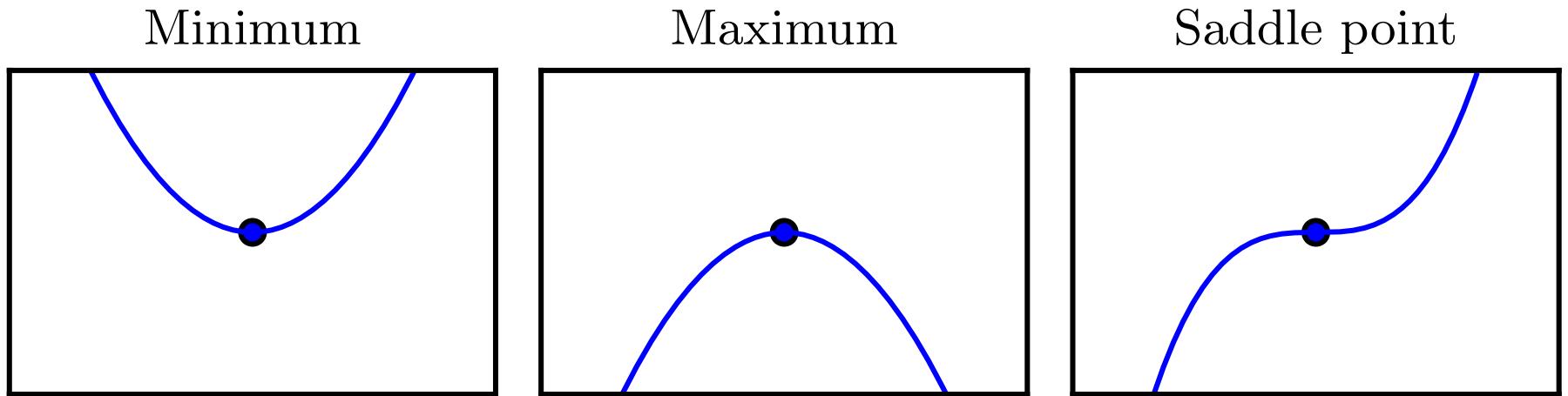


Figure 4.2

Find best θ via Solving $\nabla J(\theta) = 0$

- Write the cost function in matrix form:

$$\begin{aligned} J(\theta) &= \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i^T \theta - y_i)^2 \\ &= \frac{1}{2} (\mathbf{X}\theta - \bar{\mathbf{y}})^T (\mathbf{X}\theta - \bar{\mathbf{y}}) \\ &= \frac{1}{2} (\theta^T \mathbf{X}^T \mathbf{X}\theta - \theta^T \mathbf{X}^T \bar{\mathbf{y}} - \bar{\mathbf{y}}^T \mathbf{X}\theta + \bar{\mathbf{y}}^T \bar{\mathbf{y}}) \end{aligned}$$

To minimize $J(\theta)$, take derivative (gradient) and set to zero:

$$\nabla J(\theta) = 0$$

$$\Rightarrow \boxed{\mathbf{X}^T \mathbf{X}\theta = \mathbf{X}^T \bar{\mathbf{y}}}$$

The normal equations

Closed form
solution

$$\theta^* = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \bar{\mathbf{y}}$$

Method I: normal equations to minimize the loss

- Write the cost function in matrix form:

$$\begin{aligned} J(\theta) &= \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i^\top \theta - y_i)^2 & \nabla \downarrow \theta J(\theta) = \frac{\partial J(\theta)}{\partial \theta} &= \frac{1}{2} (2\mathbf{x}^\top \theta \\ &= \frac{1}{2} (\mathbf{X}\theta - \bar{y})^\top (\mathbf{X}\theta - \bar{y}) & \theta^\top \mathbf{x}^\top \bar{y} \\ &= \frac{1}{2} (\theta^\top \mathbf{X}^\top \mathbf{X}\theta - \theta^\top \mathbf{X}^\top \bar{y} - \bar{y}^\top \mathbf{X}\theta + \bar{y}^\top \bar{y}) & \cancel{2\mathbf{x}^\top \theta} & \cancel{\mathbf{1} \times \mathbf{n}} \cancel{\mathbf{n} \times \mathbf{1}} \end{aligned}$$

To minimize $J(\theta)$, take derivative and set to zero:

$$\nabla \downarrow \theta J(\theta) = 0$$

$$\Rightarrow$$

$$\mathbf{X}^\top \mathbf{X}\theta = \mathbf{X}^\top \bar{y}$$

The normal equations

Closed form
solution

$$\theta^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \bar{y}$$



Detailed
Derivation is
Optional /
Extra

See handout 4.1 + 4.3 \Rightarrow matrix calculus, partial deri \rightarrow Gradient

$$\nabla_{\theta} (\theta^T \mathbf{x}^T \mathbf{x} \theta) = 2 \mathbf{x}^T \mathbf{x} \theta \quad (\text{P24})$$

$$\nabla_{\theta} (-2 \theta^T \mathbf{x}^T \mathbf{y}) = -2 \mathbf{x}^T \mathbf{y} \quad (\text{P24})$$

$$\nabla_{\theta} (\mathbf{y}^T \mathbf{y}) = 0$$

$$\Rightarrow \nabla_{\theta} J(\theta) = \boxed{\mathbf{x}^T \mathbf{x} \theta - \mathbf{x}^T \mathbf{y}}$$

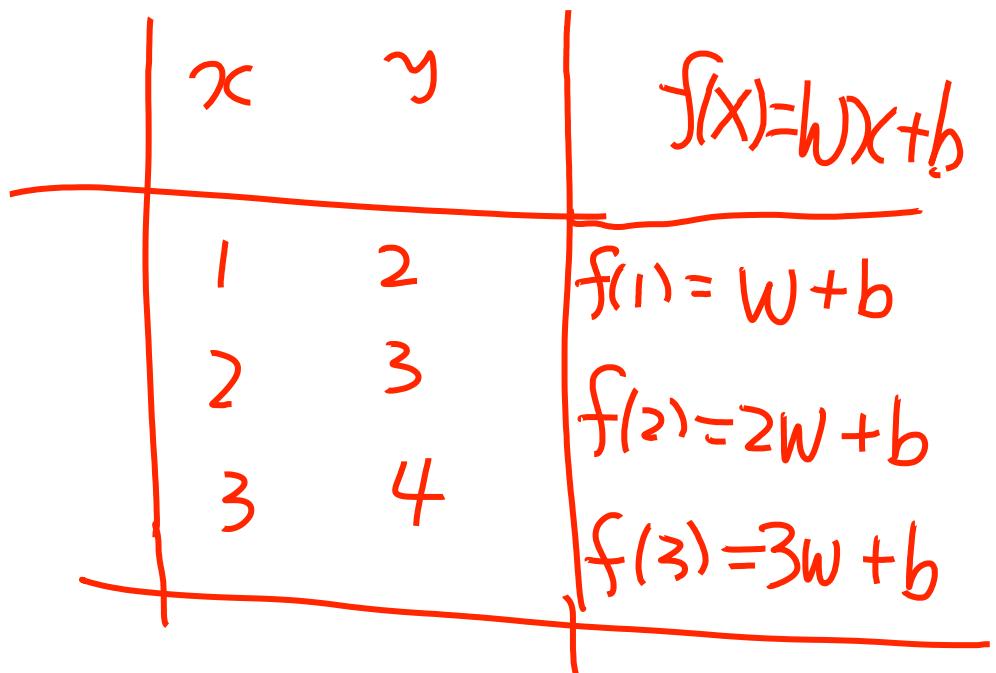
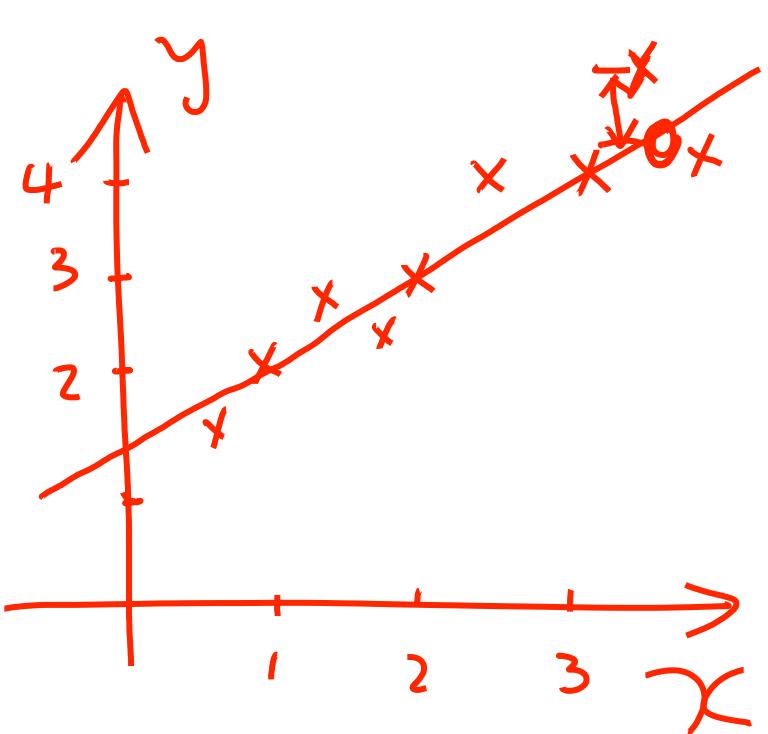
This Slide
is
Optional /
Extra

$$\Rightarrow \text{Hessian } H(J(\theta)) = \frac{\partial \nabla_{\theta} J(\theta)}{\partial \theta} = \frac{\partial (\mathbf{x}^T \mathbf{x} \theta)}{\partial \theta} =$$

Gram matrix $\Sigma^T \Sigma$

Advanced / Optional / More in
Extra Slides

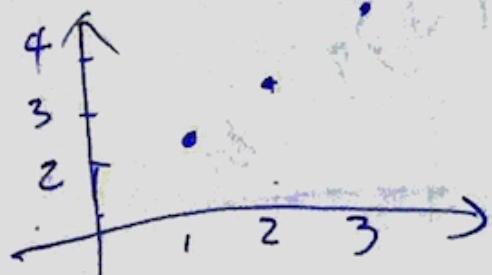
One concrete example



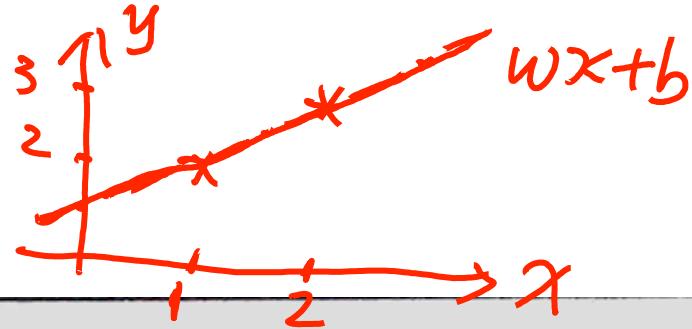
$$\text{loss}(w, b) = \left(w+b-2 \right)^2 + \left(2w+b-3 \right)^2 + \left(3w+b-4 \right)^2 \quad \left. \Rightarrow \right. \begin{array}{l} (w, b) \\ \arg\min_{w, b} \text{loss}() \end{array}$$

One concrete example

$p=1$	x_i	y_i	$f(x) = w x + b$	$(f(x) - y)^2$
$n=3$	1	2	$w + b$	$(w + b - 2)^2$
	2	3	$2w + b$	$(2w + b - 3)^2$
	3	4	$3w + b$	$(3w + b - 4)^2$



One concrete example



$$J(w, b) \quad [a]$$

$$(w^*, b^*) = \underset{w, b}{\operatorname{arg\,min}} J(w, b) \quad [b]$$

[c]

$$\begin{aligned} &= (w+b-2)^2 \\ &+ (2w+b-3)^2 \\ &\quad \cancel{3w} + 18b - 48 = 0 \\ \{ &10w + 6b - 16 = 0 \\ &\frac{6w + 4b - 10}{3w} + 20b - 50 = 0 \end{aligned}$$

[D]

$$\Rightarrow 2b - 2 = 0 \Rightarrow b = 1$$

$$\Rightarrow w = 1$$

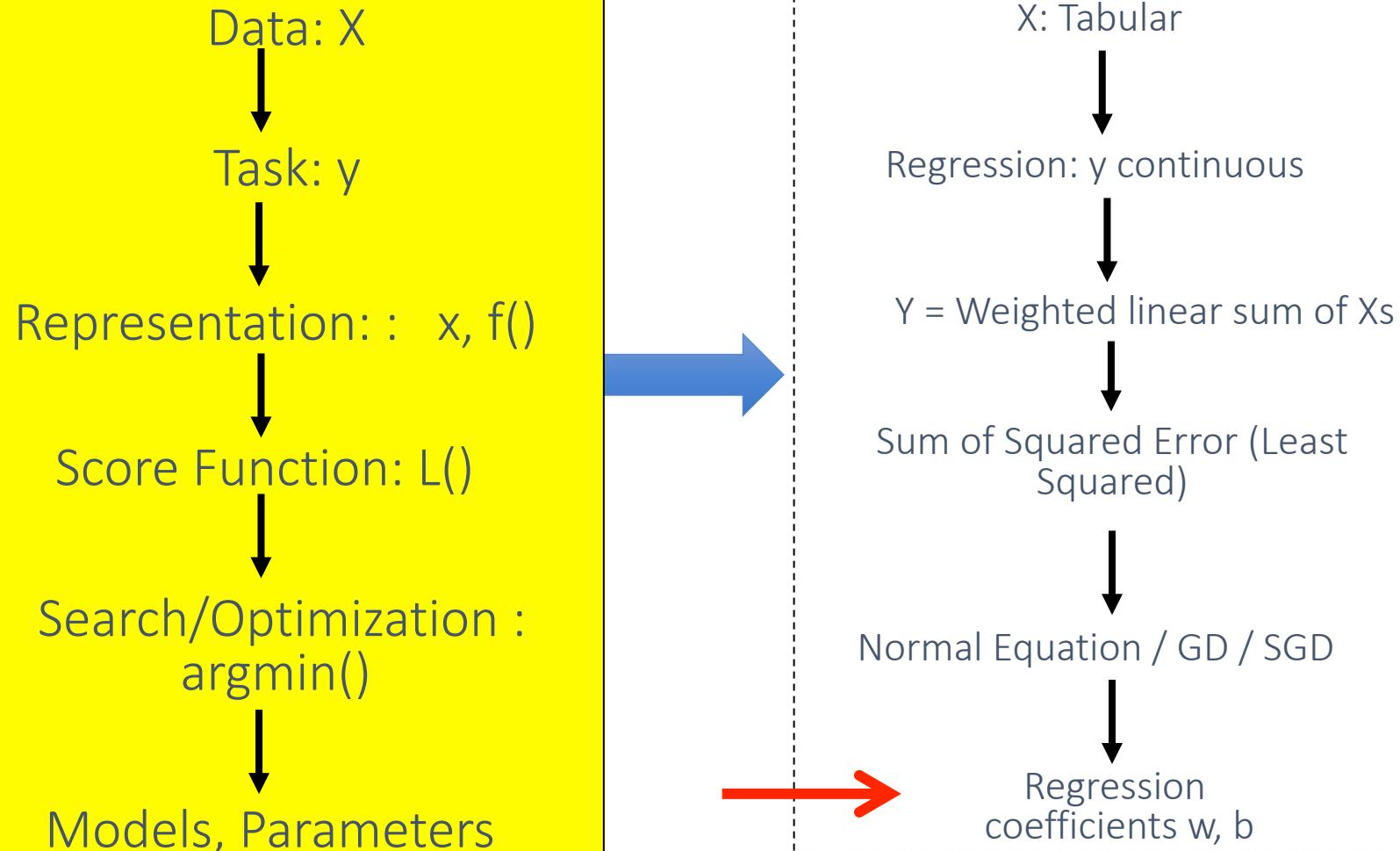
$$\begin{aligned} \frac{\partial J(w, b)}{\partial w} &= 2(w+b-2) \\ &+ 4(2w+b-3) = 0 \\ \frac{\partial J(w, b)}{\partial b} &= 2(w+b-2) \\ &+ 2(2w+b-3) = 0 \end{aligned}$$

In Step [D], we solve the matrix equation via Gaussian Elimination

Thank You



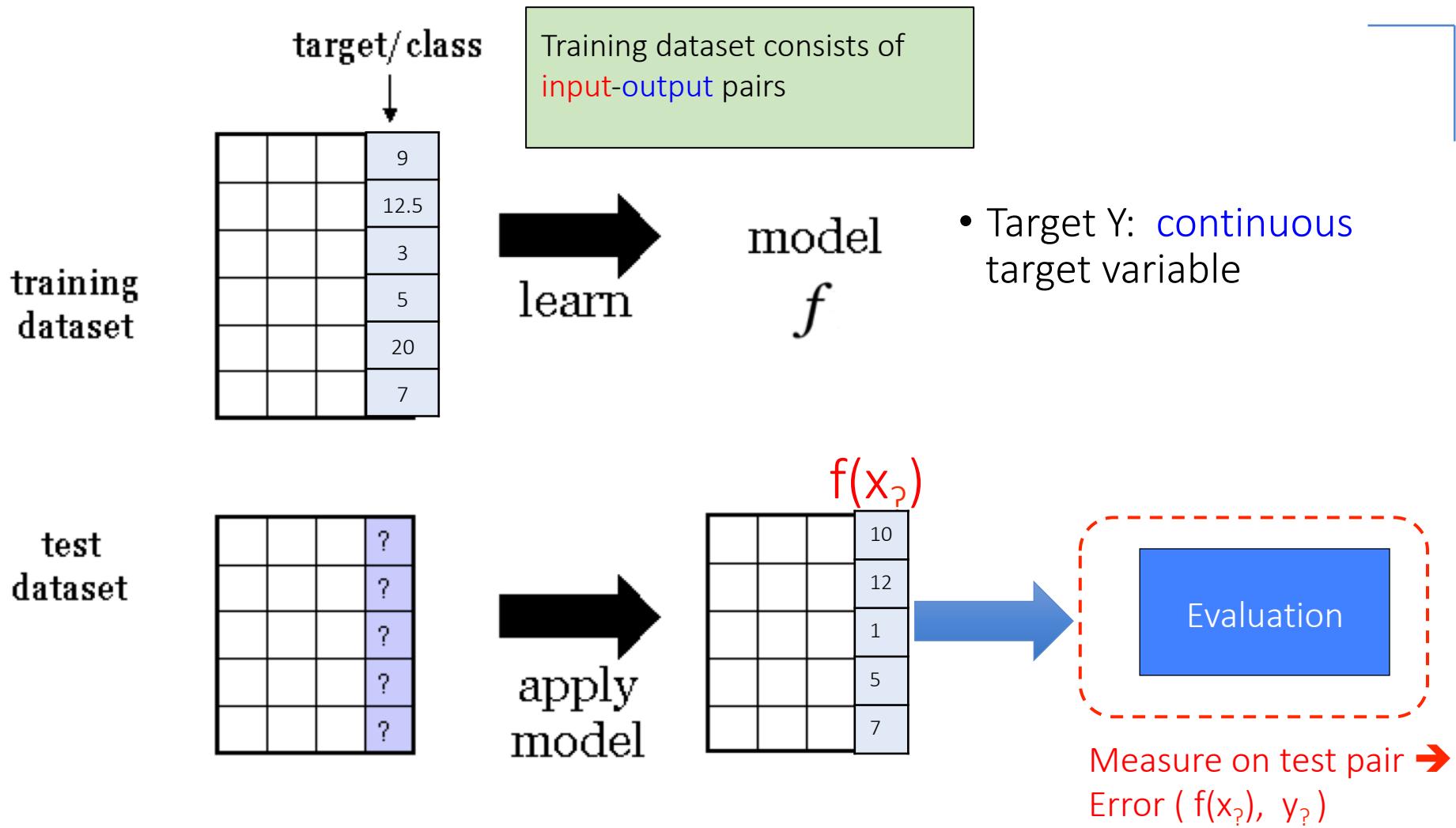
Today: Multivariate Linear Regression **in a Nutshell**



We aim to make the learned model

- 
- 1. Generalize Well
 - 2. Computational Scalable and Efficient
 - 3. Robust / Trustworthy / Interpretable
 - Especially for some domains, this is about trust!

How to know the regression program works well: Metrics on Prediction on Test Inputs



Ways for Measuring Regression Predictions: e.g., Correlation Coefficient

- Pearson correlation coefficient

$$r(x,y) = \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^m (x_i - \bar{x})^2 \times \sum_{i=1}^m (y_i - \bar{y})^2}}$$

where $\bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$ and $\bar{y} = \frac{1}{m} \sum_{i=1}^m y_i$.

$$|r(x,y)| \leq 1$$

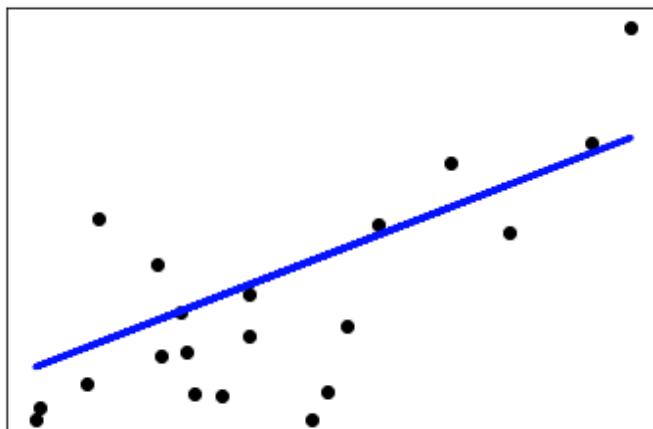
- Measuring the **linear correlation** between two sequences, x and y ,
- Generating a value between +1 and -1 inclusive, where 1 is total positive **correlation**, 0 is no **correlation**, and -1 is total negative **correlation**.

- For regression: $r(\vec{y}_{\text{predicted}}, \vec{y}_{\text{known}})$

```
Mean squared error: 2548.07  
Coefficient of determination: 0.47
```

```
# Train the model using the training sets  
regr.fit(diabetes_X_train, diabetes_y_train)  
  
# Make predictions using the testing set  
diabetes_y_pred = regr.predict(diabetes_X_test)  
  
# The coefficients  
print('Coefficients: \n', regr.coef_)  
# The mean squared error  
print('Mean squared error: %.2f'  
      % mean_squared_error(diabetes_y_test, diabetes_y_pred))  
# The coefficient of determination: 1 is perfect prediction  
print('Coefficient of determination: %.2f'  
      % r2_score(diabetes_y_test, diabetes_y_pred))
```

```
# Plot outputs  
plt.scatter(diabetes_X_test, diabetes_y_test, color='black')  
plt.plot(diabetes_X_test, diabetes_y_pred, color='blue', linewidth=3)  
  
plt.xticks(())  
plt.yticks(())  
  
plt.show()
```



We aim to make the learned model

- 1. Generalize Well
- • 2. Computational Scalable and Efficient
- 3. Robust / Trustworthy / Interpretable
 - Especially for some domains, this is about trust!

Why we Prefer Concise Vector/Matrix Form?

Training: Closed form solution

$$\theta^* = (X_{train}^T X_{train})^{-1} (X_{train}^T \vec{y}_{train})$$

Testing: on multiple Test Inputs

$$\hat{\vec{y}}_{test} = X_{test} \theta^*$$

Extra: Naïve Matrix Multiply

{implements $C = C + A * B$ }

for $i = 1$ to n

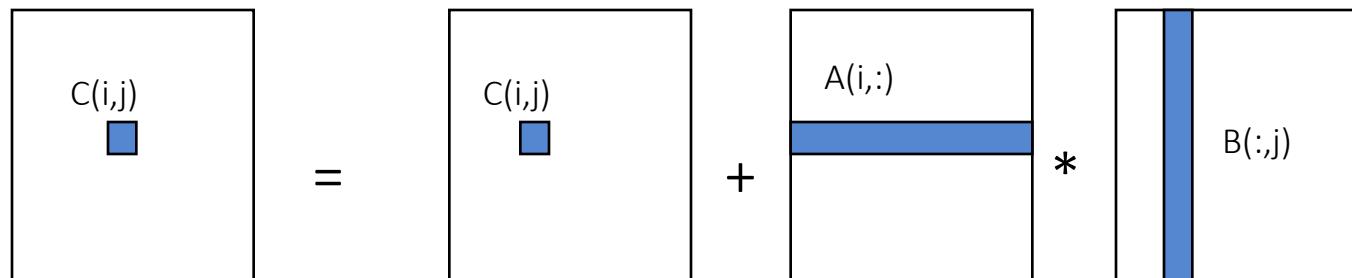
 for $j = 1$ to n

 for $k = 1$ to n

$$C(i,j) = C(i,j) + A(i,k) * B(k,j)$$

Algorithm has $2*n^3 = O(n^3)$ Flops and
operates on $3*n^2$ words of memory

Extra:



$O(n)$

Computational Cost (Naïve Way)

$$\hat{\theta}^* = \left(\begin{matrix} X^T \\ n \times p \end{matrix} \begin{matrix} X \\ p \times n \end{matrix} \right)^{-1} X^T Y$$

$X_i^T X_j$
 $1 \times n \quad n \times 1$
 $\cancel{p^2}$

$O(p^2 n)$

$$X^T X : O(p^2 n)$$

$$(X^T X)^{-1} : O(p^3)$$

$O(np^2 + p^3)$

Extra:

When $n \gg p$,

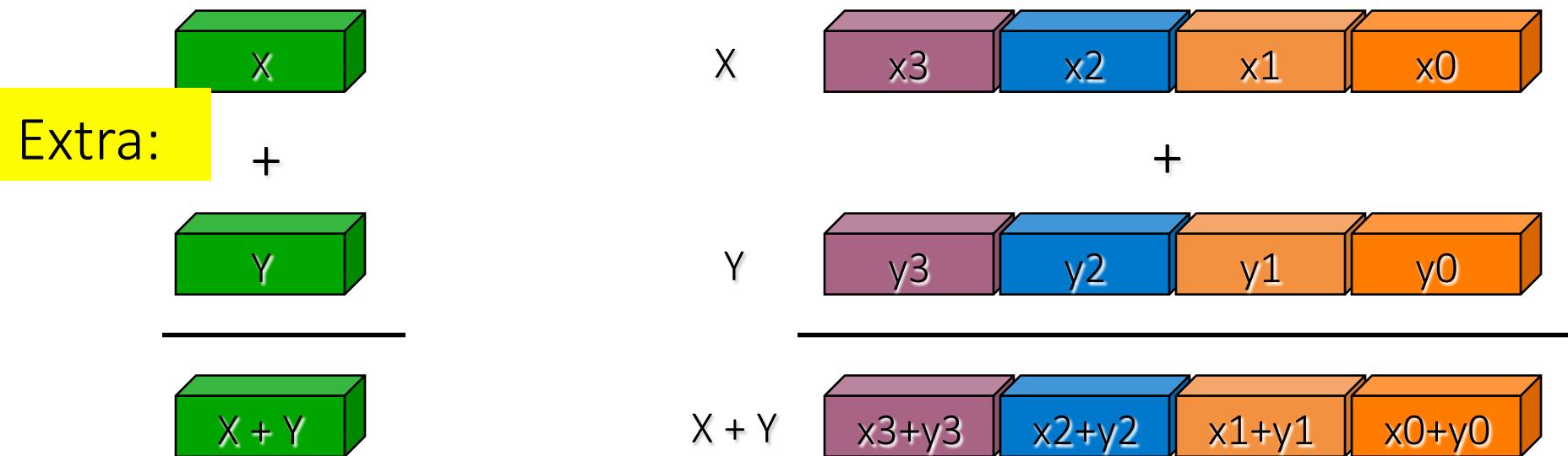
matrix mult:
slower than inversion

Many architecture details and Algorithm details to consider

- Data parallelization through CPU SIMD / Multithreading/ GPU parallelization /
- Memory hierarchical / locality
- Better algorithms, like Strassen's Matrix Multiply and many others

SIMD: Single Instruction, Multiple Data

- Scalar processing
 - traditional mode
 - one operation produces one result
- SIMD processing
 - with SSE / SSE2
 - SSE = streaming SIMD extensions
 - one operation produces multiple results

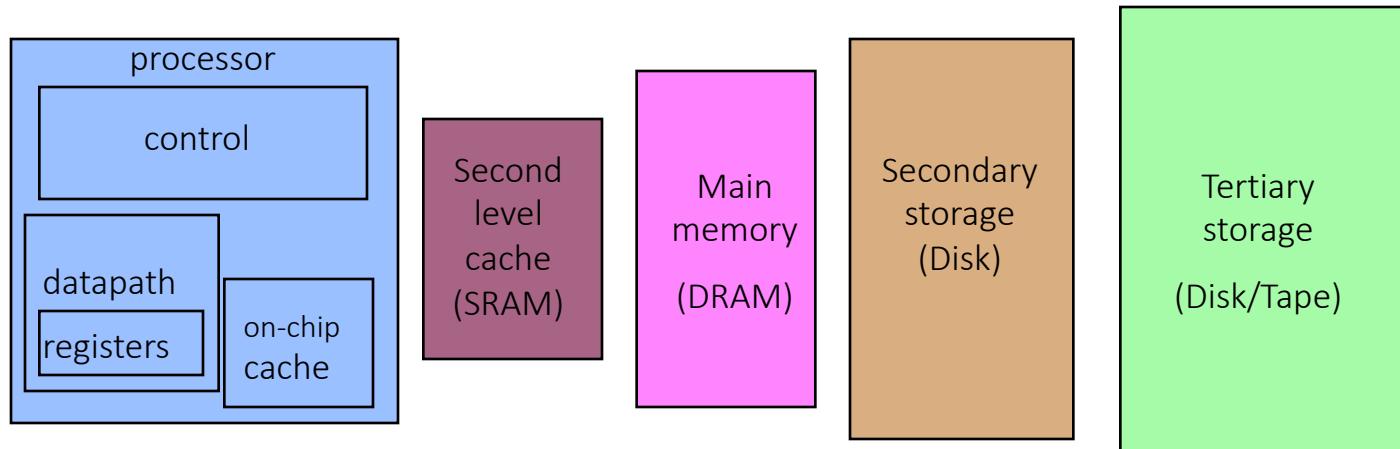


Slide Source: Alex Klimovitski & Dean Macri, Intel Corporation

Memory Hierarchy

- Most programs have a high degree of **locality** in their accesses
 - **spatial locality**: accessing things nearby previous accesses
 - **temporal locality**: reusing an item that was previously accessed
- Memory hierarchy tries to exploit locality to improve average

Extra:



Speed	1ns	10ns	100ns	10ms	10sec
Size	KB	MB	GB	TB	PB

The following complexity figures assume that arithmetic with individual elements has complexity $O(1)$, as is the case with fixed-precision operations on a finite field.

Operation	Input	Output	Algorithm	Complexity
Matrix multiplication	Two $n \times n$ matrices	One $n \times n$ matrix	Schoolbook matrix multiplication	$O(n^3)$
			Strassen algorithm	$O(n^{2.807})$
			Coppersmith–Winograd algorithm	$O(n^{2.376})$
			Optimized CW-like algorithms ^{[14][15][16]}	$O(n^{2.373})$
Matrix multiplication	One $n \times m$ matrix & one $m \times p$ matrix	One $n \times p$ matrix	Schoolbook matrix multiplication	$O(nmp)$
Matrix inversion*	One $n \times n$ matrix	One $n \times n$ matrix	Gauss–Jordan elimination	$O(n^3)$
			Strassen algorithm	$O(n^{2.807})$
			Coppersmith–Winograd algorithm	$O(n^{2.376})$
			Optimized CW-like algorithms	$O(n^{2.373})$
Singular value decomposition	One $m \times n$ matrix	One $m \times m$ matrix, one $m \times n$ matrix, & one $n \times n$ matrix		$O(mn^2)$ ($m \leq n$)
		One $m \times r$ matrix, one $r \times r$ matrix, & one $n \times r$ matrix		
Determinant	One $n \times n$ matrix	One number	Laplace expansion	$O(n!)$
			Division-free algorithm ^[17]	$O(n^4)$
			LU decomposition	$O(n^3)$
			Bareiss algorithm	$O(n^3)$
			Fast matrix multiplication ^[18]	$O(n^{2.373})$
Back substitution	Triangular matrix	n solutions	Back substitution ^[19]	$O(n^2)$

Extra:

Basic Linear Algebra Subroutines (BLAS) →

numpy: a wrapper library of BLAS

- Industry standard interface (evolving)
 - www.netlib.org/blas, www.netlib.org/blas/blast-forum
- Vendors, others supply optimized implementations
- History
 - BLAS1 (1970s):
 - vector operations: dot product, saxpy ($y=a*x+y$), etc
 - $m=2*n$, $f=2*n$, $q = f/m$ = computational intensity ~ 1 or less
 - BLAS2 (mid 1980s)
 - matrix-vector operations: matrix vector multiply, etc
 - $m=n^2$, $f=2*n^2$, $q \sim 2$, less overhead
 - somewhat faster than BLAS1
 - BLAS3 (late 1980s)
 - matrix-matrix operations: matrix matrix multiply, etc
 - $m \leq 3n^2$, $f=O(n^3)$, so $q=f/m$ can possibly be as large as n , so BLAS3 is potentially much faster than BLAS2
- Good algorithms use BLAS3 when possible (LAPACK & ScaLAPACK)
 - See www.netlib.org/{lapack,scalapack}

Functionality [edit]

BLAS functionality is categorized into three sets of routines called "levels", which correspond to both the chronological order of definition and publication, as well as the degree of the polynomial in the complexities of algorithms; Level 1 BLAS operations typically take linear time, $O(n)$, Level 2 operations quadratic time and Level 3 operations cubic time.^[18] Modern BLAS implementations typically provide all three levels.

Level 1 [edit]

This level consists of all the routines described in the original presentation of BLAS (1979),^[1] which defined only *vector operations* on strided arrays: dot products, vector norms, a generalized vector addition of the form

$$\mathbf{y} \leftarrow \alpha \mathbf{x} + \mathbf{y}$$

(called "axpy") and several other operations.

Versions of BLAS compared: BLAS library from the Netlib Repository, ATLAS library, Intel-MKL library, AMD ACML Library and Goto BLAS.

Level 2 [edit]

This level contains *matrix-vector operations* including, among other things, matrix-vector multiplication (gemv):

$$\mathbf{y} \leftarrow \alpha \mathbf{A} \mathbf{x} + \beta \mathbf{y}$$

as well as a solver for \mathbf{x} in the linear equation

$$\mathbf{T} \mathbf{x} = \mathbf{y}$$

with \mathbf{T} being triangular. Design of the Level 2 BLAS started in 1984, with release in 1988.^[19] The Level 2 subroutines are especially intended to improve performance of programs using BLAS on *vector processors*, where Level 1 BLAS are suboptimal because they hide the matrix-vector nature of the operations from the compiler.

Level 3 [edit]

This level, formally published in 1990,^[18] contains *matrix-matrix operations* "general matrix multiplication" (gemm), of the form

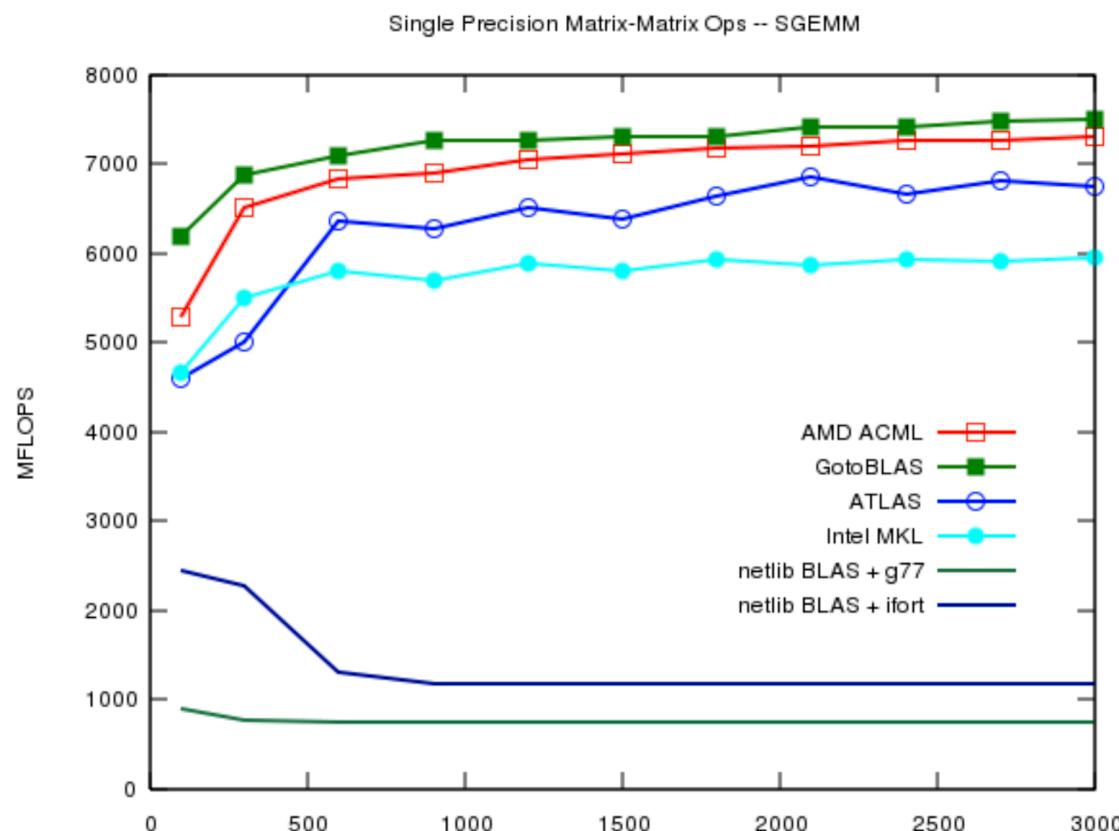
$$\mathbf{C} \leftarrow \alpha \mathbf{A} \mathbf{B} + \beta \mathbf{C}$$

where \mathbf{A} and \mathbf{B} can optionally be transposed or hermitian-conjugated inside and all three matrices may be strided. The ordinary matrix multiplication $\mathbf{A} \mathbf{B}$ is performed by setting α to one and \mathbf{C} to an all-zeros matrix of the appropriate size.

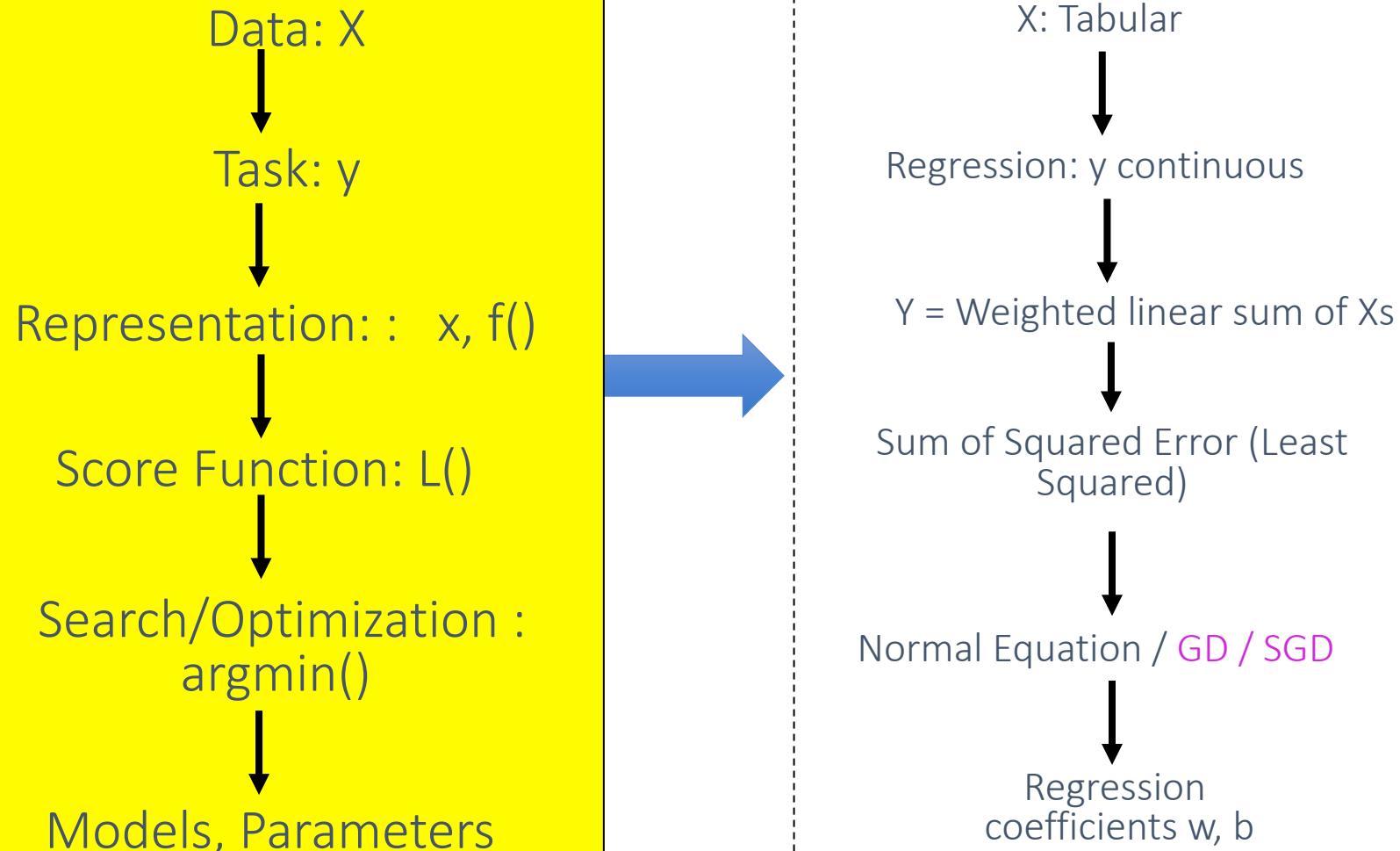
Also included in Level 3 are routines for solving

$$\mathbf{B} \leftarrow \alpha \mathbf{T}^{-1} \mathbf{B}$$

where \mathbf{T} is a triangular matrix, among other functionality.



Recap : Multivariate Linear Regression **in a Nutshell**



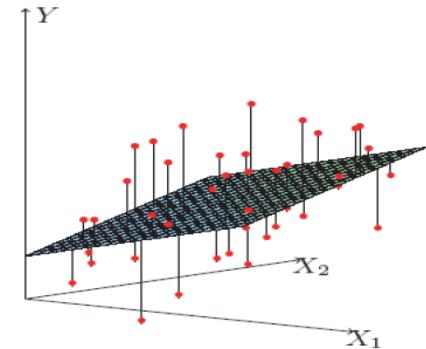
References

- Big thanks to Prof. Eric Xing @ CMU for allowing me to reuse some of his slides
 - <http://www.cs.cmu.edu/~zkolter/course/15-884/linalg-review.pdf>
 - Prof. Alexander Gray's slides

EXTRA

In Case you are interested in more advanced details!

Probabilistic Interpretation of Linear Regression (Extra)



- Let us assume that the target variable and the inputs are related by the equation:

$$y_i = \theta^T \mathbf{x}_i + \varepsilon_i$$

Error data on each point

where ε is an error term of unmodeled effects or random noise

- Now assume that ε follows a Gaussian $N(0, \sigma)$, then we have:

$$p(y_i | x_i; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \theta^T \mathbf{x}_i)^2}{2\sigma^2}\right)$$

- By iid (among samples) assumption:

Many more variations of LinearR from this perspective, e.g. binomial / poisson (LATER)

$$L(\theta) = \prod_{i=1}^n p(y_i | x_i; \theta) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp\left(-\frac{\sum_{i=1}^n (y_i - \theta^T \mathbf{x}_i)^2}{2\sigma^2}\right)$$

Review (I):

- Sum the Squared Elements of a Vector equals
Vector dot product to itself

$$\vec{a} = \sum \theta - y$$

$$\mathbf{a} = \begin{bmatrix} 5 \\ 2 \\ 8 \end{bmatrix}$$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i^T \theta - y_i)^2$$

$$\mathbf{a}^T = \begin{bmatrix} 5 & 2 & 8 \end{bmatrix}$$

$$\vec{a}^T \vec{a} = \sum_{i=1}^n a_i^2$$

$$\mathbf{a}^T \mathbf{a} = \begin{bmatrix} 5 & 2 & 8 \end{bmatrix} \begin{bmatrix} 5 \\ 2 \\ 8 \end{bmatrix} = 5^2 + 2^2 + 8^2 = 93$$

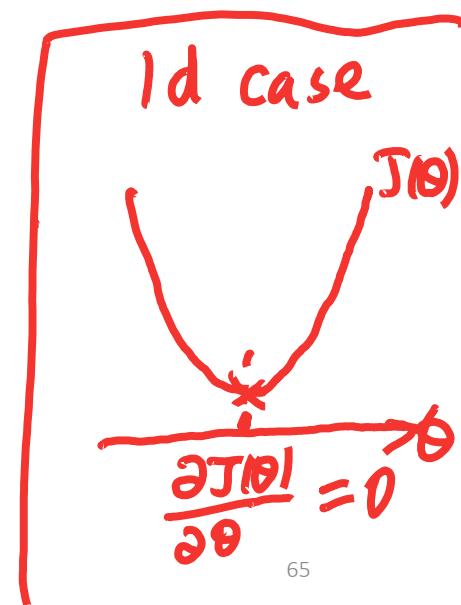
Review(I) :

$$a = \begin{bmatrix} \mathbf{x}_1^T \theta - y_1 \\ \mathbf{x}_2^T \theta - y_2 \\ \vdots \\ \mathbf{x}_n^T \theta - y_n \end{bmatrix} = X\theta - \vec{y}$$

$$\mathbf{a}^\top \mathbf{a} = 2J(\theta) = \sum_{i=1}^n (\mathbf{x}_i^T \theta - y_i)^2$$

$$\begin{aligned}
 J(\theta) &= (\mathbf{x}\theta - \mathbf{y})^T (\mathbf{x}\theta - \mathbf{y}) \frac{1}{2} \\
 &= ((\mathbf{x}\theta)^T - \mathbf{y}^T)(\mathbf{x}\theta - \mathbf{y}) \frac{1}{2} \\
 &= (\theta^T \mathbf{x}^T - \mathbf{y}^T)(\mathbf{x}\theta - \mathbf{y}) \frac{1}{2} \\
 &= \underbrace{(\theta^T \mathbf{x}^T \mathbf{x}\theta - \theta^T \mathbf{x}^T \mathbf{y} - \mathbf{y}^T \mathbf{x}\theta + \mathbf{y}^T \mathbf{y})}_{\text{since } \theta^T \mathbf{x}^T \mathbf{y} = \mathbf{y}^T \mathbf{x}\theta} \frac{1}{2} \\
 &\quad \langle \mathbf{x}\theta, \mathbf{y} \rangle \quad \langle \mathbf{y}, \mathbf{x}\theta \rangle \\
 &= \left(\underbrace{\theta^T \mathbf{x}^T \mathbf{x}\theta}_{\text{1d case}} - 2 \underbrace{\theta^T \mathbf{x}^T \mathbf{y}}_{\frac{\partial J(\theta)}{\partial \theta} = 0} + \mathbf{y}^T \mathbf{y} \right) \frac{1}{2}
 \end{aligned}$$

$\Rightarrow J(\theta)$ quadratic func of θ ;



Review (II): gradient of linear form

$$\frac{\partial(\theta^T X^T y)}{\partial \theta} = X^T y$$

One
Concrete
Example

$$f(w) = w^T a = [w_1, w_2, w_3] \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = w_1 + 2w_2 + 3w_3$$

$$\frac{\partial f}{\partial w_1} = 1$$

$$\frac{\partial f}{\partial w_2} = 2$$

$$\frac{\partial f}{\partial w_3} = 3$$



$$\frac{\partial f}{\partial w} = \frac{\partial w^T a}{\partial w} = a = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

Review (III): Gradient of Quadratic Form

- See L2-note.pdf -> Page 17, Page 23-24
- See white board

$$\frac{\partial(\theta^T X^T X \theta)}{\partial \theta} = \frac{\partial(\theta^T G \theta)}{\partial \theta} = 2G\theta = 2X^T X \theta$$

Review (III): Single Var-Func to Multivariate

Single Var-Function	Multivariate Calculus
Derivative	Partial Derivative
Second-order derivative	Gradient
	Directional Partial Derivative
	Vector Field
	Contour map of a function
	Surface map of a function
	Hessian matrix
	Jacobian matrix (vector in / vector out)

Review (IV) : Definitions of gradient (Matrix_calculus / Scalar-by-vector)

- Size of gradient is always the same as the size of variable

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n \quad \text{if } x \in \mathbb{R}^n$$

In principle, gradients are a natural extension of partial derivatives to functions of multiple variables.

Review (V): Rank of a Matrix

- $\text{rank}(A)$ (the rank of a m -by- n matrix A) is
 - = The maximal number of linearly independent columns
 - = The maximal number of linearly independent rows

- If A is n by m , then
 - $\text{rank}(A) \leq \min(m, n)$
 - If $n = \text{rank}(A)$, then A has full row rank
 - If $m = \text{rank}(A)$, then A has full column rank

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix}$$

Rank=? Rank=?

If A is $n \times n$, $\text{rank}(A) = n$ iff A is invertible

$\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$

Extra: Loss $J()$ is Convex

$$\Rightarrow J(\theta) = \frac{1}{2} (\theta^T X^T X \theta - 2\theta^T X^T y + y^T y)$$

$$\Rightarrow \text{Hessian } (J(\theta)) = \nabla^T \nabla$$

Gram matrix
[PSD]

$J(\theta)$ is convex

If $\nabla J(\theta^*) = 0$, $J(\theta)$ is minimized @ θ^*

Review: Hessian Matrix

Derivatives and Second Derivatives

Cost function

$$J(\boldsymbol{\theta})$$

Gradient

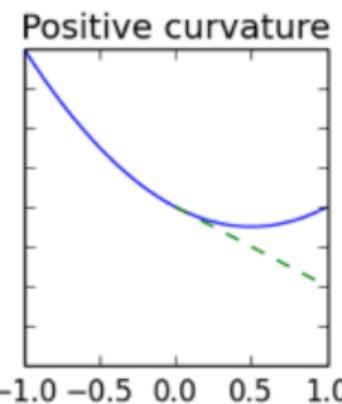
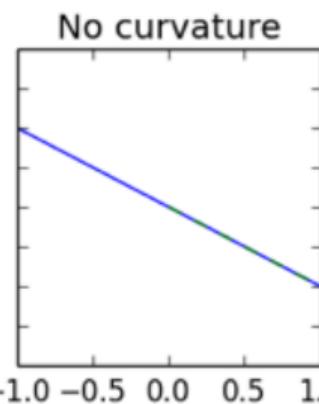
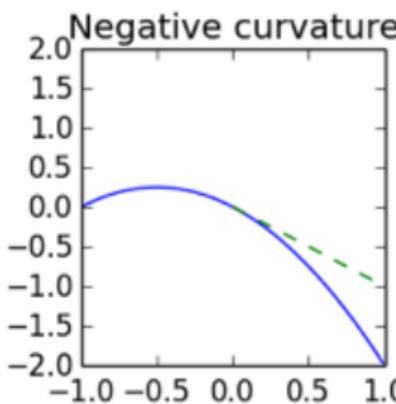
$$\mathbf{g} = \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

$$g_i = \frac{\partial}{\partial \theta_i} J(\boldsymbol{\theta})$$

Hessian

$$\mathbf{H}$$

$$H_{i,j} = \frac{\partial}{\partial \theta_j} g_i$$



$\mathbf{H} \text{ PD}$
for positive
curvature

Positive Definite Hessian

Extra: Convex function

- Intuitively, a convex function (1D case) has a single point at which the derivative goes to zero, and this point is a minimum.
- Intuitively, a function f (1D case) is convex on the range $[a,b]$ if a function's second derivative is positive every-where in that range.
- Intuitively, if a multivariate function's Hessians is pd (positive definite!), this (multivariate) function is Convex
 - Intuitively, we can think “Positive definite” matrices as analogy to positive numbers in matrix case

Our loss function $J()$'s Hessian is
Positive Semi-definite - PSD

$H = \mathbf{X}^T \mathbf{X}$
is always
PSD

Intuitively, a convex function with PD hessian is
minimized @ point whose

- ✓ derivative (slope) is zero
- ✓ gradient is zero vector (multivariate case)

Gram Matrix $H = X^T X$

when X is full rank, H is positive definite!

Extra: Gram Matrix $H = X^T X$
is always positive semi-definite!

Because for any vector a

$$a^T X^T X a = \|Xa\|_2^2 \geq 0$$



Besides, when X is full rank,
 H is Positive Definite (PD) and invertible

Later: Comments on the normal equation

When \mathbf{X} full rank $\theta^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{\mathbf{y}}$

- In most situations of practical interest, the number of data points n is larger than the dimensionality p of the input space and the matrix \mathbf{X} is of full column rank. If this condition holds, then it is easy to verify that $\mathbf{X}^T \mathbf{X}$ is necessarily invertible.

$$n \gg p$$

- The assumption that $\mathbf{X}^T \mathbf{X}$ is invertible implies that it is positive definite, thus the critical point (by solving gradient to zero) we have found is a minimum.
- What if \mathbf{X} has less than full column rank? \rightarrow regularization (later).

when \mathbf{X} not full rank, e.g.

e.g. when $p > n \Rightarrow \text{rank}(\mathbf{X}) < p \Rightarrow \text{rank}(\underbrace{\mathbf{X}^T \mathbf{X}}_{p \times p}) \leq \min\{\text{rank}(\mathbf{X})\} < p \Rightarrow \text{not invertible}$

Extra: positive semi-definite!

$$A \in \mathbb{R}^{n \times n}, \forall x \in \mathbb{R}^n$$

L2-Note: Page 17

If $x^T A x \geq 0$
 $(x \in \mathbb{R}^n)$

$\Rightarrow A$ is positive semi-definite (PSD)

If $x^T A x > 0$

$\Rightarrow A$ is PD \Rightarrow full rank / invertible

See proof on L2-Note: Page 18

Extra: Scalability to big data?

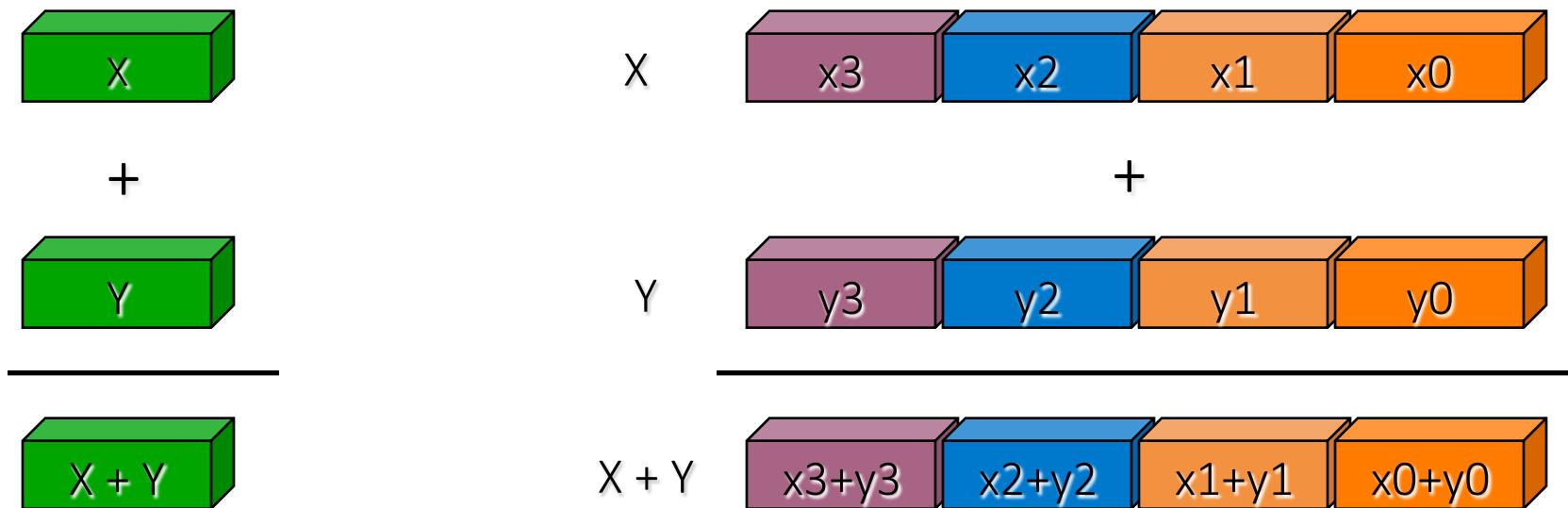
- Traditional CS view: Polynomial time algorithm, Wow!
- Large-scale learning: Sometimes even $O(n)$ is bad! => Many state-of-the-art solutions (e.g., low rank, sparse, hardware, sampling, randomized...)

Simple example: Matrix multiplication

$$\begin{matrix} n & \begin{matrix} \text{[Blue square]} \end{matrix} & \times & \begin{matrix} \text{[Blue square]} \end{matrix} & = & \begin{matrix} \text{[Blue square]} \end{matrix} & O(n^3)! \\ & \text{n} & & & & & \end{matrix}$$

SIMD: Single Instruction, Multiple Data

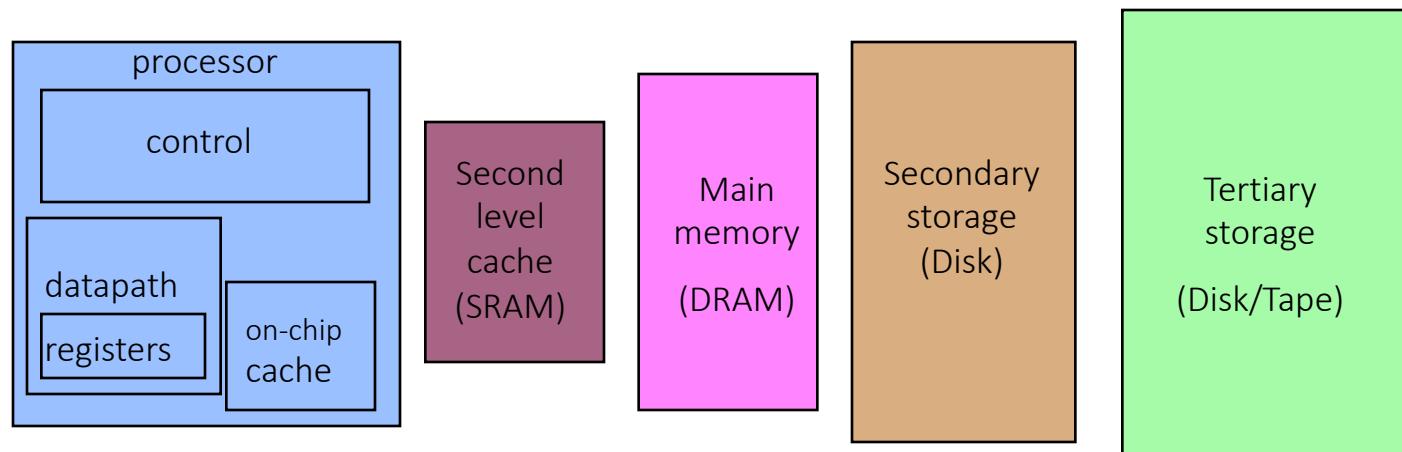
- Scalar processing
 - traditional mode
 - one operation produces one result
- SIMD processing
 - with SSE / SSE2
 - SSE = streaming SIMD extensions
 - one operation produces multiple results



Slide Source: Alex Klimovitski & Dean Macri, Intel Corporation

Memory Hierarchy

- Most programs have a high degree of **locality** in their accesses
 - **spatial locality**: accessing things nearby previous accesses
 - **temporal locality**: reusing an item that was previously accessed
- Memory hierarchy tries to exploit locality to improve average



Speed	1ns	10ns	100ns	10ms	10sec
Size	KB	MB	GB	TB	PB

Note on Matrix Storage

- A matrix is a 2-D array of elements, but memory addresses are “1-D”
- Conventions for matrix layout
 - by column, or “column major” (Fortran default); $A(i,j)$ at $A+i+j*n$
 - by row, or “row major” (C default) $A(i,j)$ at $A+i*n+j$
 - recursive (later)

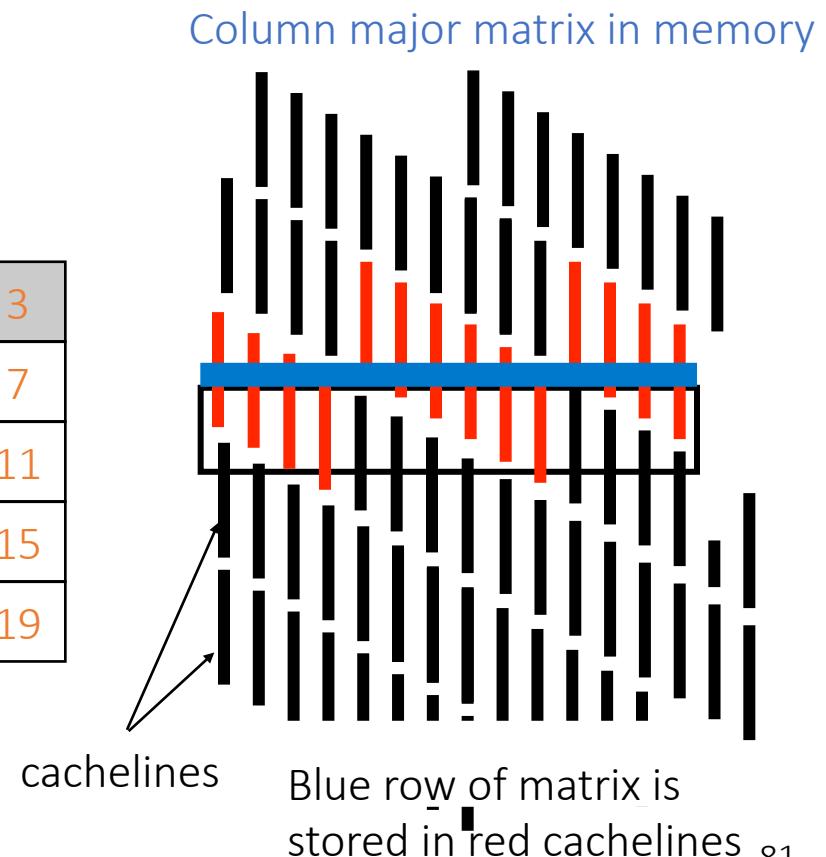
Column major

0	5	10	15
1	6	11	16
2	7	12	17
3	8	13	18
4	9	14	19

Row major

0	1	2	3
4	5	6	7
8	9	10	11
12	13	14	15
16	17	18	19

- Column major (for now)



Strassen's Matrix Multiply

- The traditional algorithm (with or without tiling) has $O(n^3)$ flops
- Strassen discovered an algorithm with asymptotically lower flops
 - $O(n^{2.81})$
- Consider a 2×2 matrix multiply, normally takes 8 multiplies, 4 adds
 - Strassen does it with 7 multiplies and 18 adds

Let $M = \begin{matrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{matrix} = \begin{matrix} a_{11} & a_{12} & b_{11} & b_{12} \\ a_{21} & a_{22} & b_{21} & b_{22} \end{matrix}$

$$\begin{aligned} p_1 &= (a_{12} - a_{22}) * (b_{21} + b_{22}) & p_5 &= a_{11} * (b_{12} - b_{22}) \\ p_2 &= (a_{11} + a_{22}) * (b_{11} + b_{22}) & p_6 &= a_{22} * (b_{21} - b_{11}) \\ p_3 &= (a_{11} - a_{21}) * (b_{11} + b_{12}) & p_7 &= (a_{21} + a_{22}) * b_{11} \\ p_4 &= (a_{11} + a_{12}) * b_{22} \end{aligned}$$

$$\text{Then } m_{11} = p_1 + p_2 - p_4 + p_6$$

$$m_{12} = p_4 + p_5$$

$$m_{21} = p_6 + p_7$$

$$m_{22} = p_2 - p_3 + p_5 - p_7$$

Extends to $n \times n$ by divide&conquer

Strassen (continued)

$$\begin{aligned} T(n) &= \text{Cost of multiplying } nxn \text{ matrices} \\ &= 7*T(n/2) + 18*(n/2)^2 \\ &= O(n \log_2 7) \\ &= O(n^{2.81}) \end{aligned}$$

- Asymptotically faster
 - Several times faster for large n in practice
 - Cross-over depends on machine
 - “Tuning Strassen's Matrix Multiplication for Memory Efficiency”, M. S. Thottethodi, S. Chatterjee, and A. Lebeck, in Proceedings of Supercomputing '98
- Possible to extend communication lower bound to Strassen
 - #words moved between fast and slow memory $= \Omega(n^{\log 2})$
 $\frac{7}{M} / M^{(\log 2)/2 - 1} \sim \Omega(n^{2.81} / M^{0.4})$
 - (Ballard, D., Holtz, Schwartz, 2011)
 - Attainable too