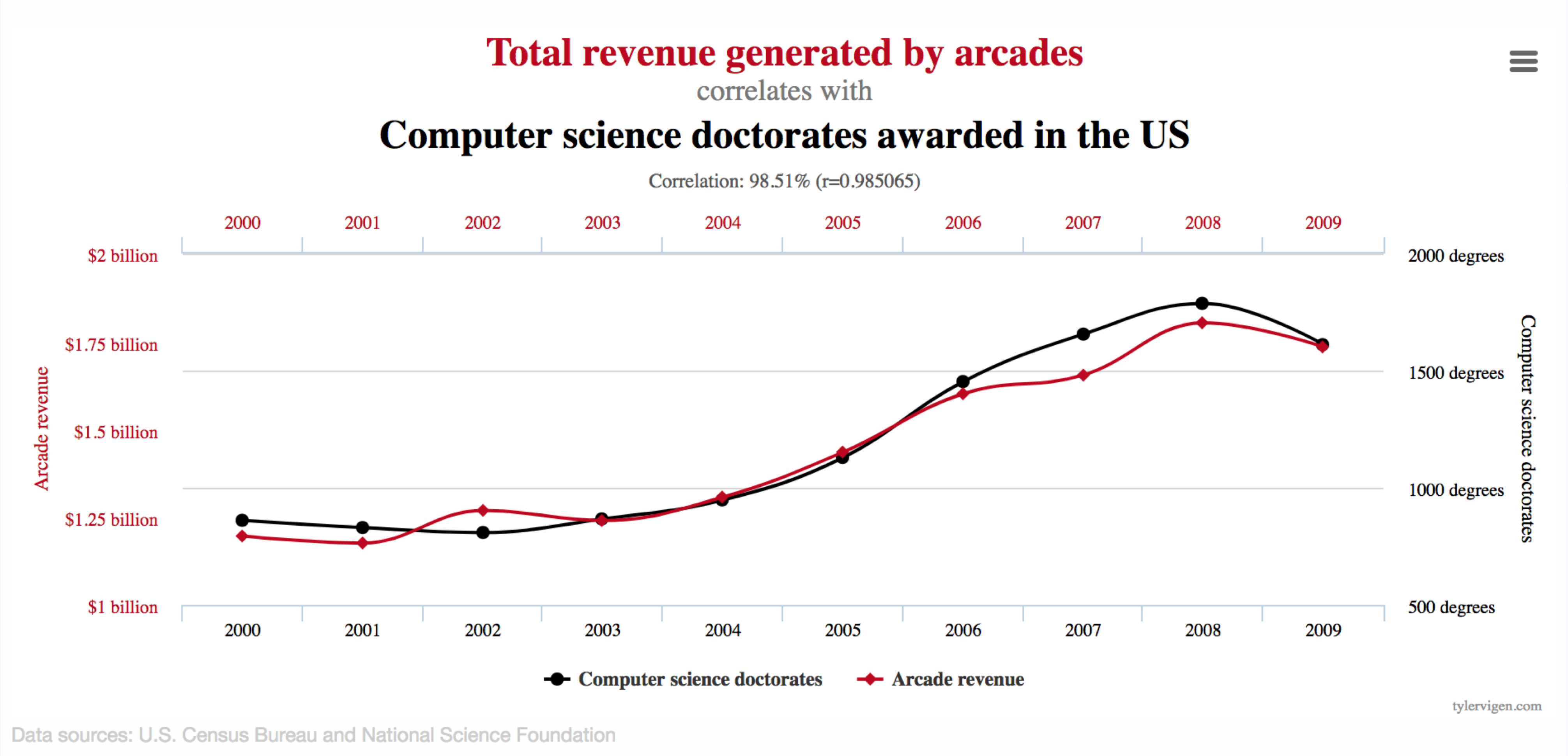
MATH FOUNDATIONS

8 JULY 2019 - WWW.SEA-MLS.COM

CHENG SOON ONG

ong-home.my

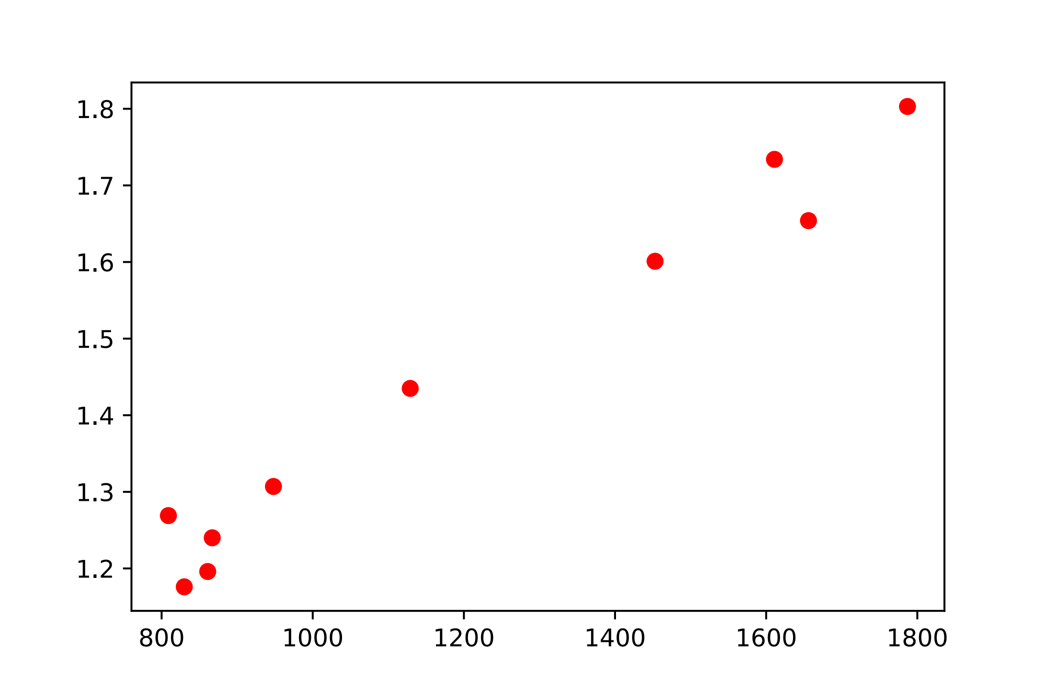
PREDICT ARCADE REVENUE FROM CS PHD



tylervigen.com

PREDICT ARCADE REVENUE FROM CS PHD

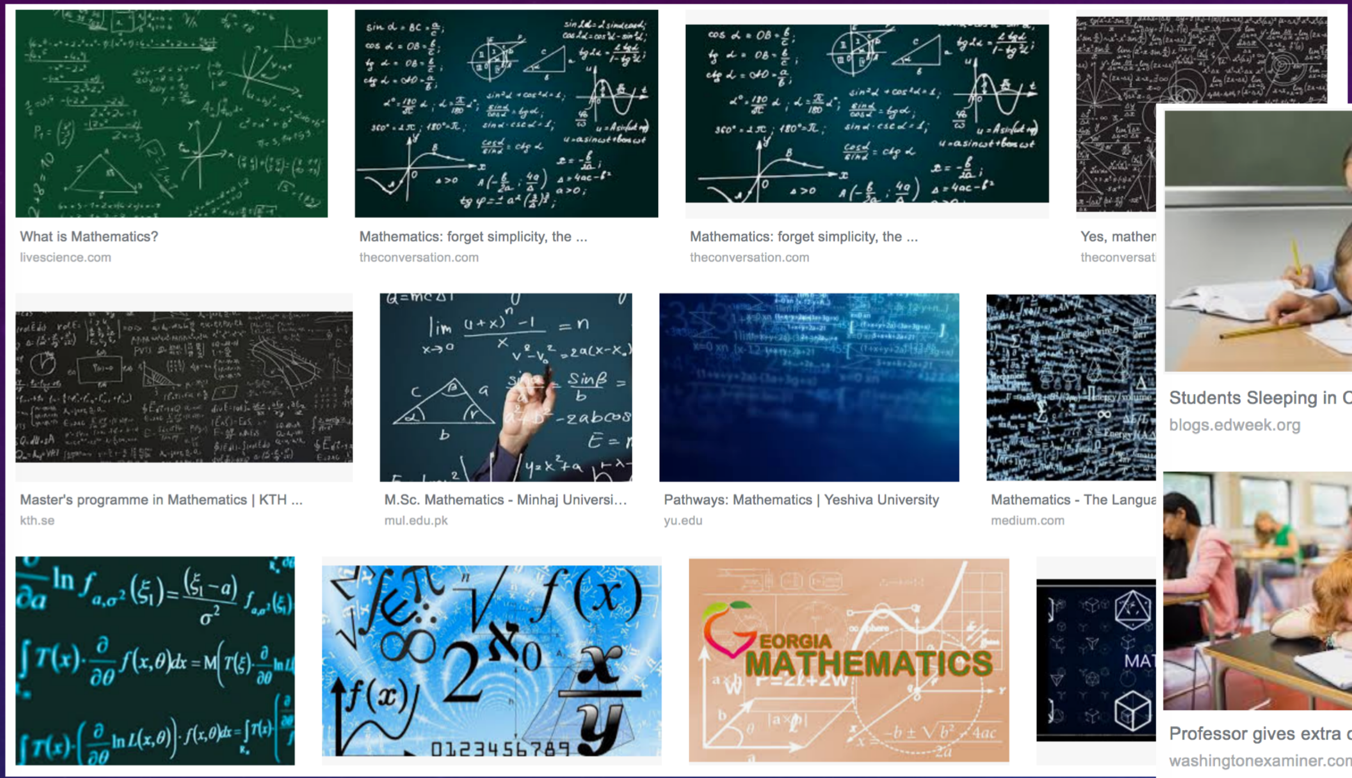
Year	CS PhD in USA	Arcade ($billions)
2000	861	1.196
2001	830	1.176
2002	809	1.269
2003	867	1.240
2004	948	1.307
2005	1129	1.435
2006	1453	1.601
2007	1656	1.654
2008	1787	1.803
2009	1611	1.734

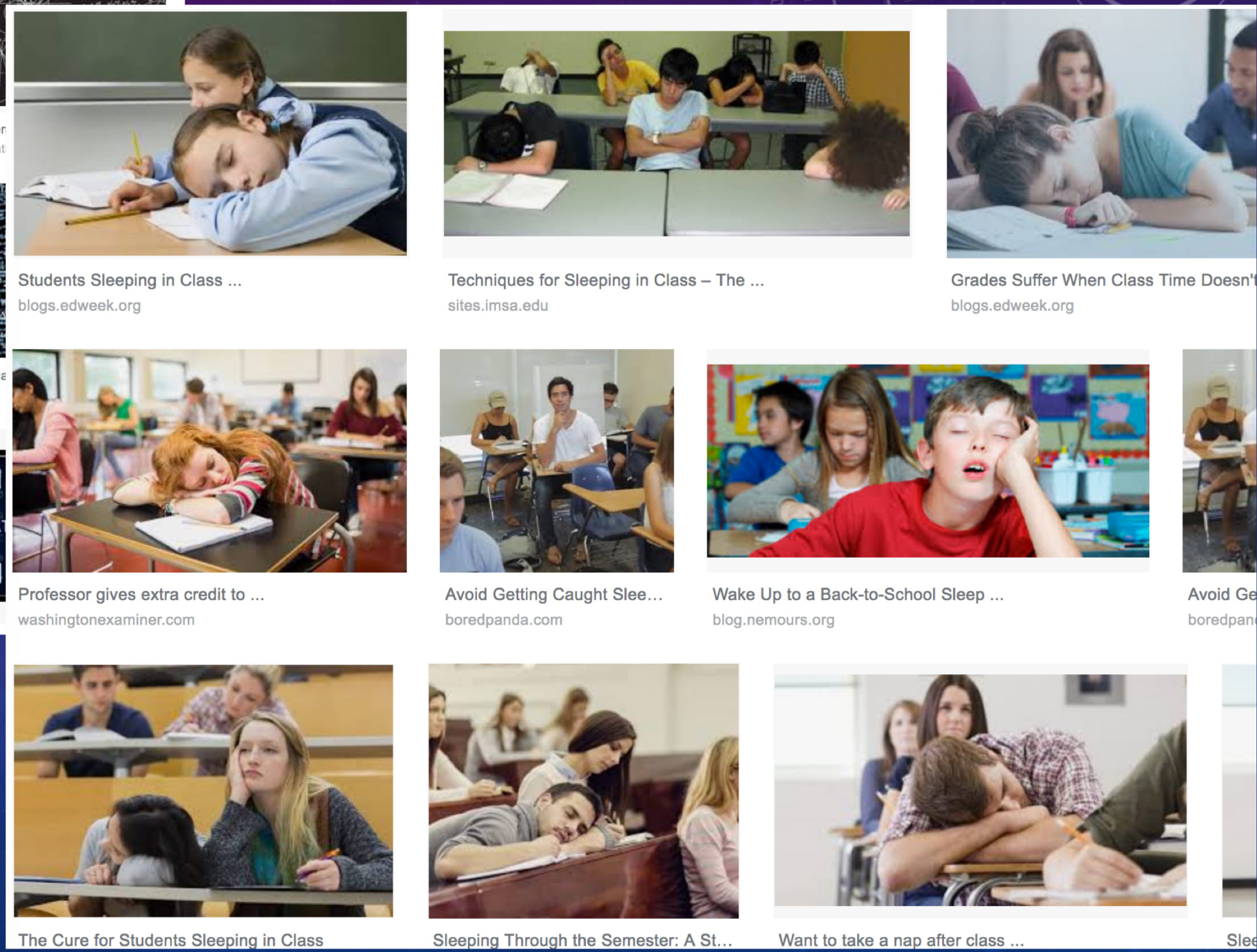


tylervigen.com

Looks like machine learning, but before we can learn that we need to do some maths

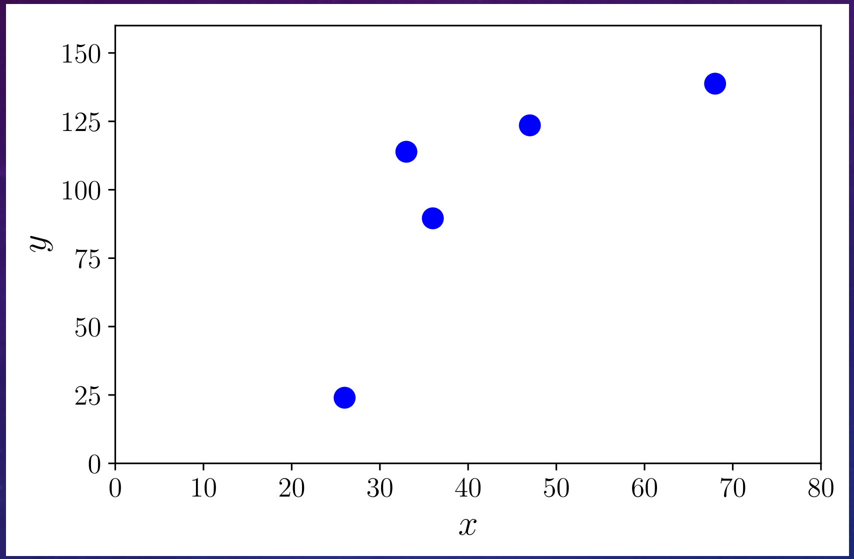
MATHEMATICS? THIS IS BORING…



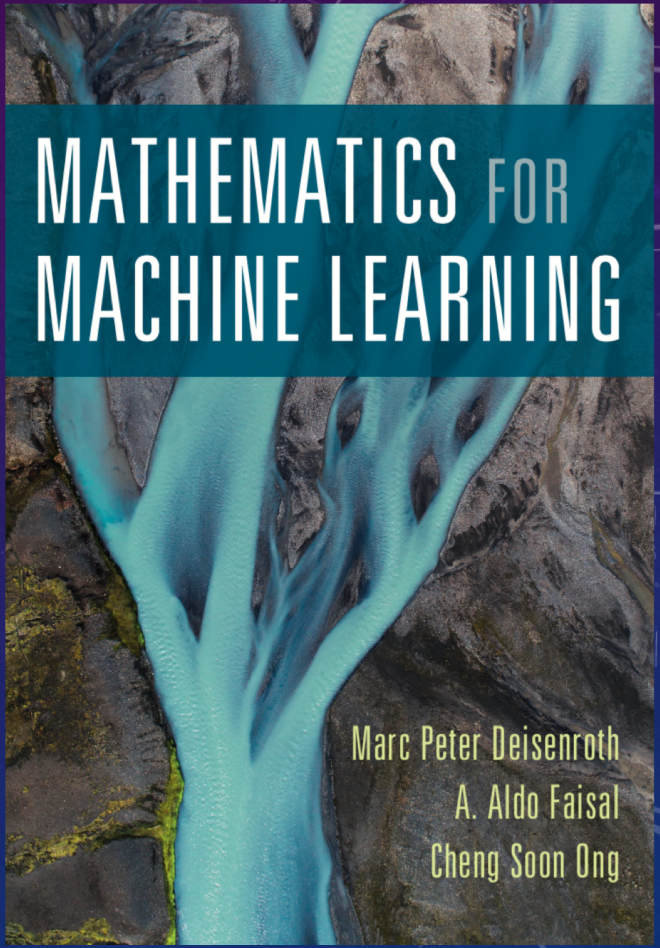


SHAMELESS PLUG: MATH FOR ML

Age	Annual Salary (in thousands)
36	89.563
47	123.543
26	23.989
68	138.769
33	113.888



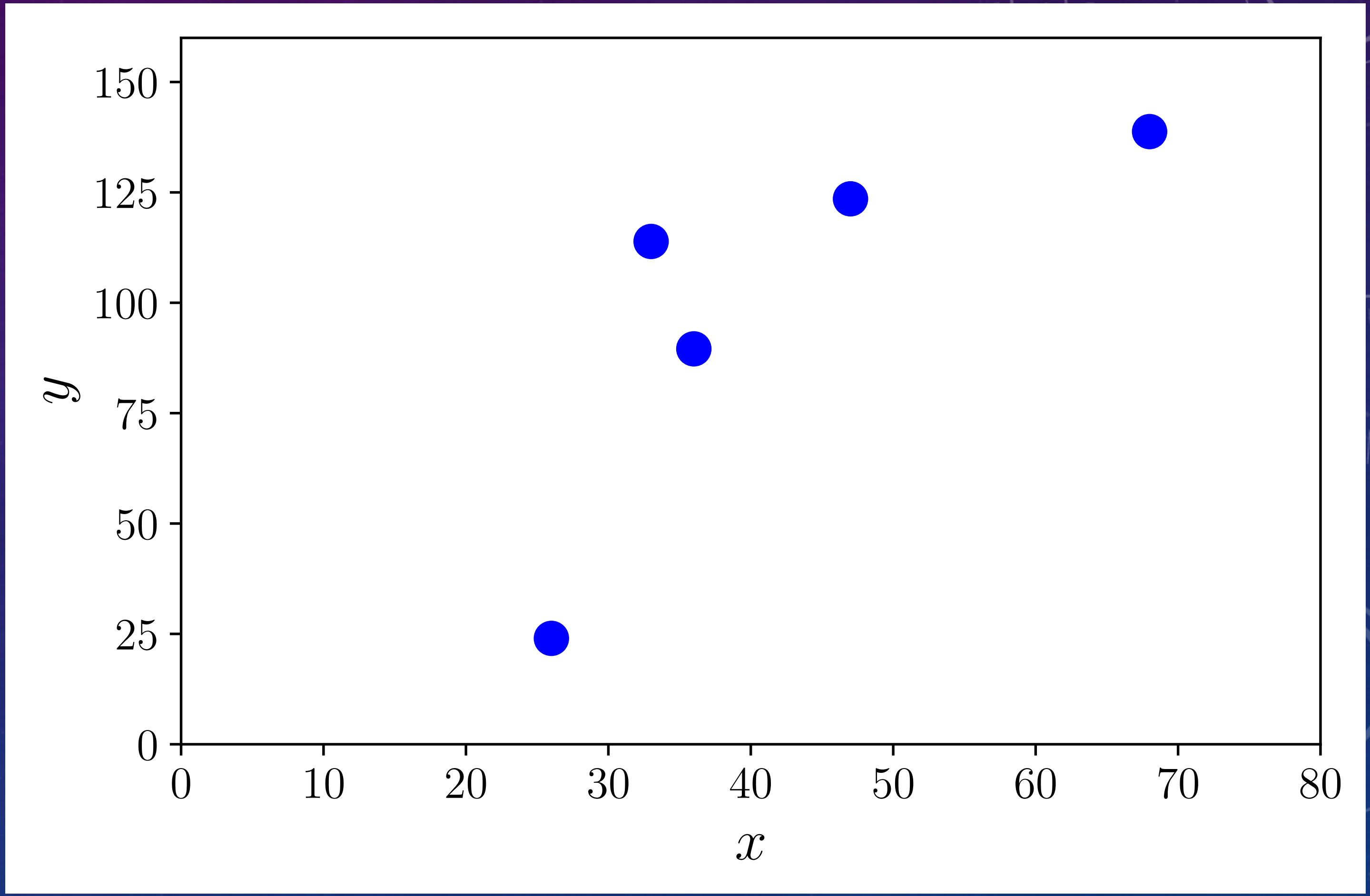
Predict salary (y) from age (x)



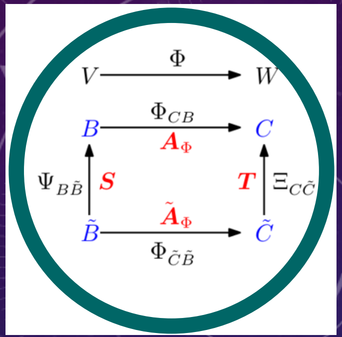
WHAT DO WE MEAN WHEN WE DRAW DOTS?

- Three views of a vector
  - (CS) array of numbers
  - (physics) magnitude and direction
  - (math) satisfies + and x

Age	Annual Salary (in thousands)
36	89.563
47	123.543
26	23.989
68	138.769
33	113.888



LINEAR ALGEBRA – WHAT IS A VECTOR?

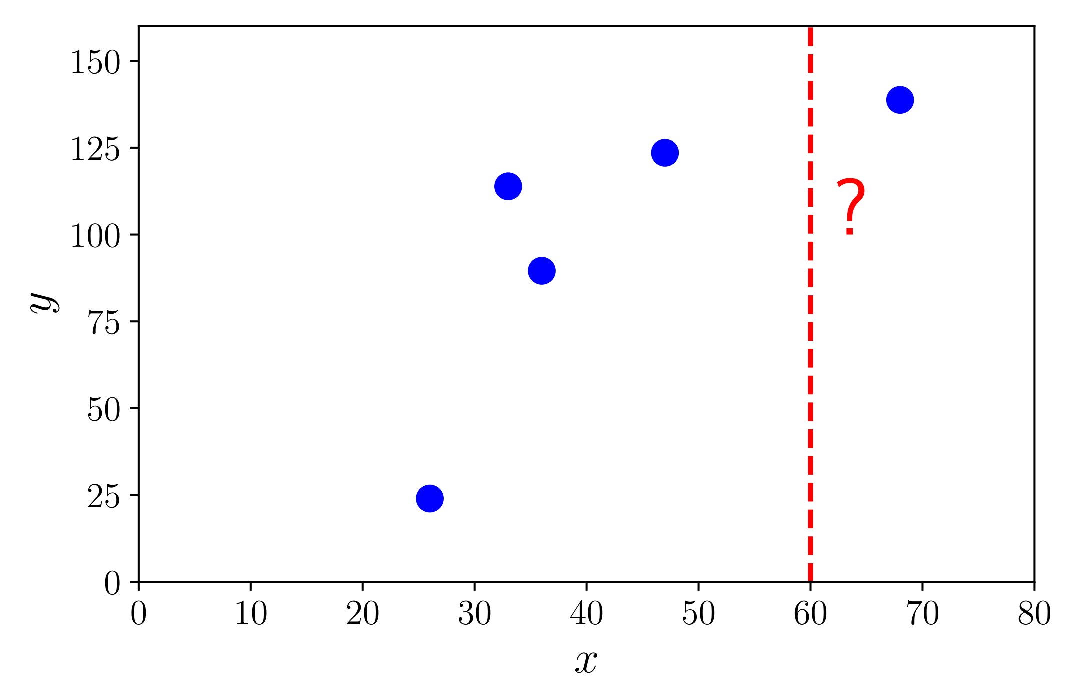


- Algebra: Set of objects and set of rules to manipulate them
  - Objects: vectors x and y
  - Rules: + and x, as well as defining a zero.
- Linear: ax + by
  - distributivity
  - associativity
- Vector space:
  - Closure: adding and scaling vectors keeps things in the vector space

MACHINE LEARNING IS ABOUT PREDICTION

Age	Annual Salary (in thousands)
36	89.563
47	123.543
26	23.989
68	138.769
33	113.888

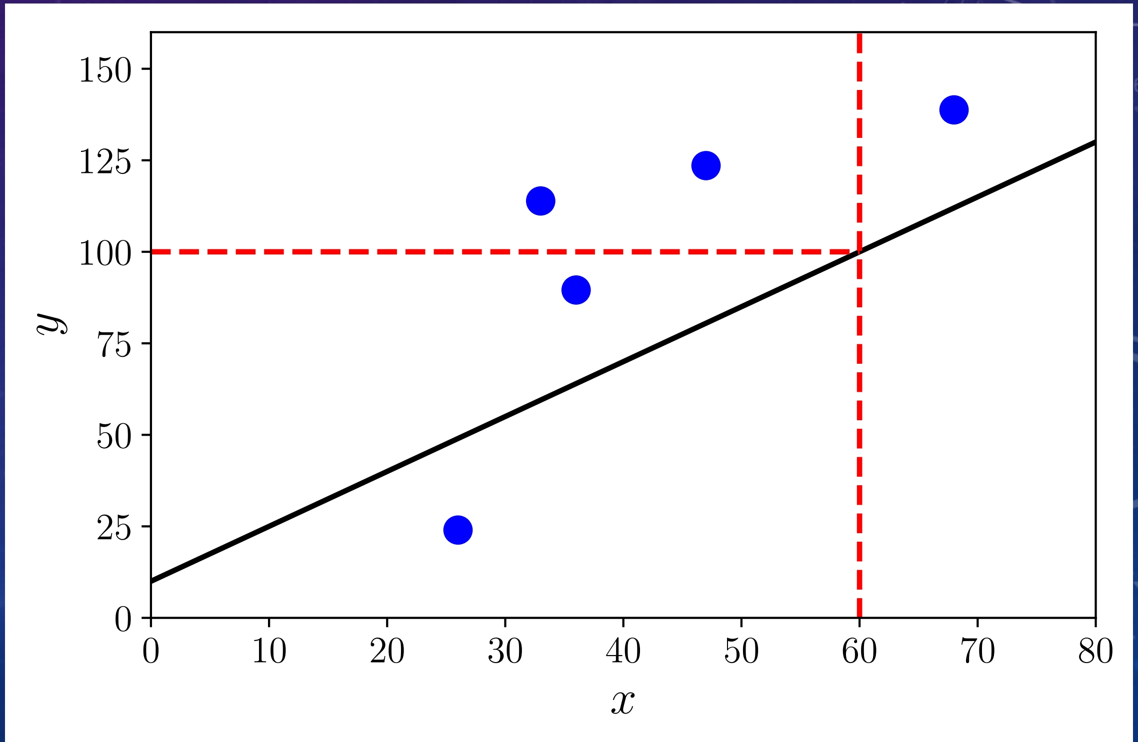
- The values of y for the training data is not the main focus
- We are interested in generalization error:
  - What is the error we make on unseen data?
- Do not train on the test set



FITTING A LINE

Age	Annual Salary (in thousands)
36	89.563
47	123.543
26	23.989
68	138.769
33	113.888

- The values of y for the training data is not the main focus
- We are interested in generalization error:
  - What is the error we make on unseen data?
- Do not train on the test set

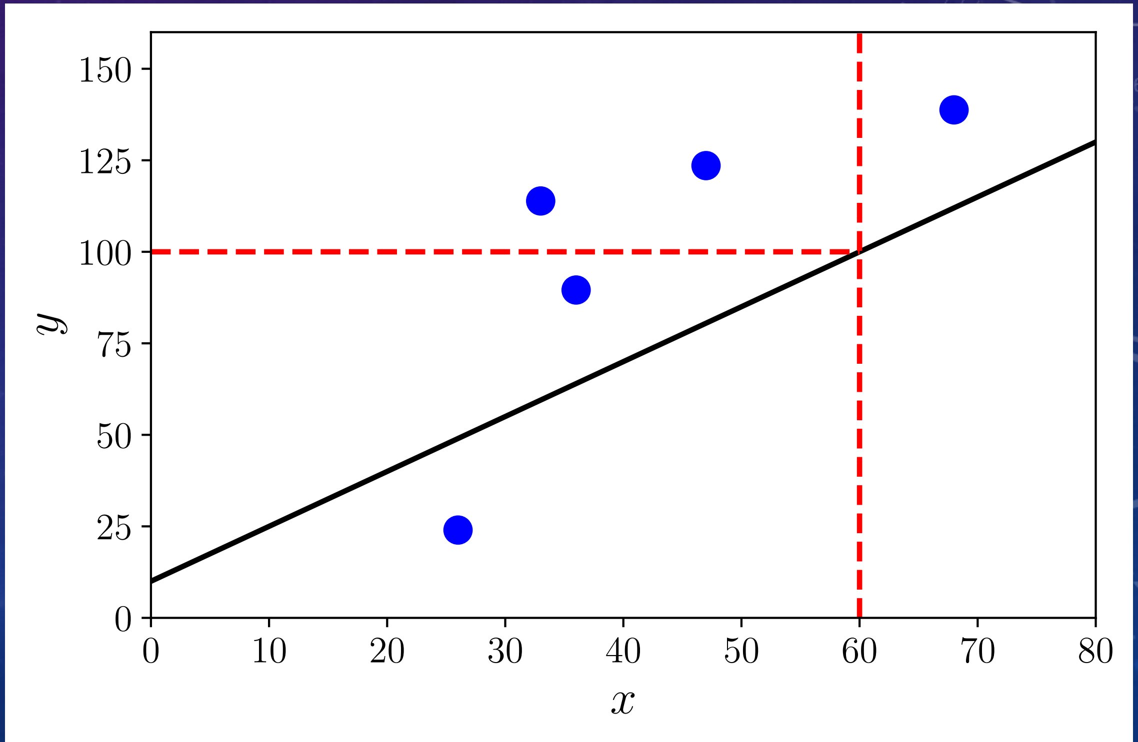


FITTING A LINE - NOTATION

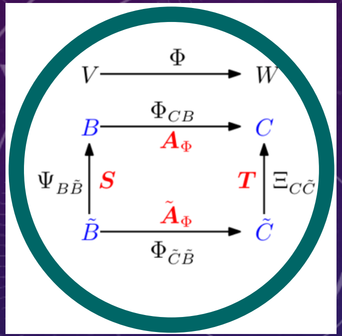
Age	Annual Salary (in thousands)
36	89.563
47	123.543
26	23.989
68	138.769
33	113.888

- $N = 5$
- $\mathbf{x} = [x_1, \dots, x_N]^T$
- $\mathbf{y} = [y_1, \dots, y_N]^T$
- x_n is a real number
- y_n is a real number
- $f(x) = w_1 x + w_0$
- $X = [\mathbf{x}\ 1]$

- Find the best line that fits the data



LINEAR ALGEBRA - MATRIX



- Want to solve X**w** = **y**

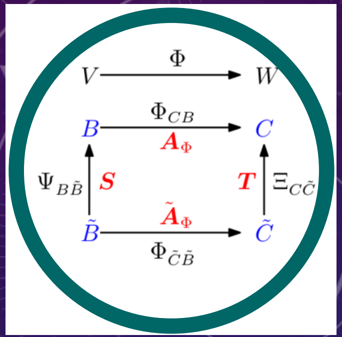
- Solutions of linear equations
- Inverse and transpose

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{A}^{-1} &= \boldsymbol{I} = \boldsymbol{A}^{-1}\boldsymbol{A} \\
(\boldsymbol{A}\boldsymbol{B})^{-1} &= \boldsymbol{B}^{-1}\boldsymbol{A}^{-1} \\
(\boldsymbol{A}+\boldsymbol{B})^{-1} &\neq \boldsymbol{A}^{-1}+\boldsymbol{B}^{-1} \\
(\boldsymbol{A}^\top)^\top &= \boldsymbol{A} \\
(\boldsymbol{A}+\boldsymbol{B})^\top &= \boldsymbol{A}^\top+\boldsymbol{B}^\top \\
(\boldsymbol{A}\boldsymbol{B})^\top &= \boldsymbol{B}^\top\boldsymbol{A}^\top
\end{aligned}$$

Definition 2.3 (Inverse). Consider a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Let matrix $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ have the property that $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{I}_n = \boldsymbol{B}\boldsymbol{A}$. $\boldsymbol{B}$ is called the *inverse* of $\boldsymbol{A}$ and denoted by $\boldsymbol{A}^{-1}$.

Definition 2.4 (Transpose). For $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ the matrix $\boldsymbol{B} \in \mathbb{R}^{n\times m}$ with $b_{ij} = a_{ji}$ is called the *transpose* of $\boldsymbol{A}$. We write $\boldsymbol{B} = \boldsymbol{A}^\top$.

LINEAR ALGEBRA



- Linear independence

Definition 2.11 (Linear Combination). Consider a vector space V and a finite number of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in V$. Then, every $\boldsymbol{v} \in V$ of the form

$$\boldsymbol{v} = \lambda_1 \boldsymbol{x}_1 + \cdots + \lambda_k \boldsymbol{x}_k = \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i \in V \tag{2.65}$$

with $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$ is a *linear combination* of the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$.

- Basis and rank

Definition 2.14 (Basis). Consider a vector space $V = (\mathcal{V}, +, \cdot)$ and $\mathcal{A} \subseteq \mathcal{V}$. A generating set $\mathcal{A}$ of V is called *minimal* if there exists no smaller set $\tilde{\mathcal{A}} \subseteq \mathcal{A} \subseteq \mathcal{V}$ that spans V. Every linearly independent generating set of V is minimal and is called a *basis* of V.

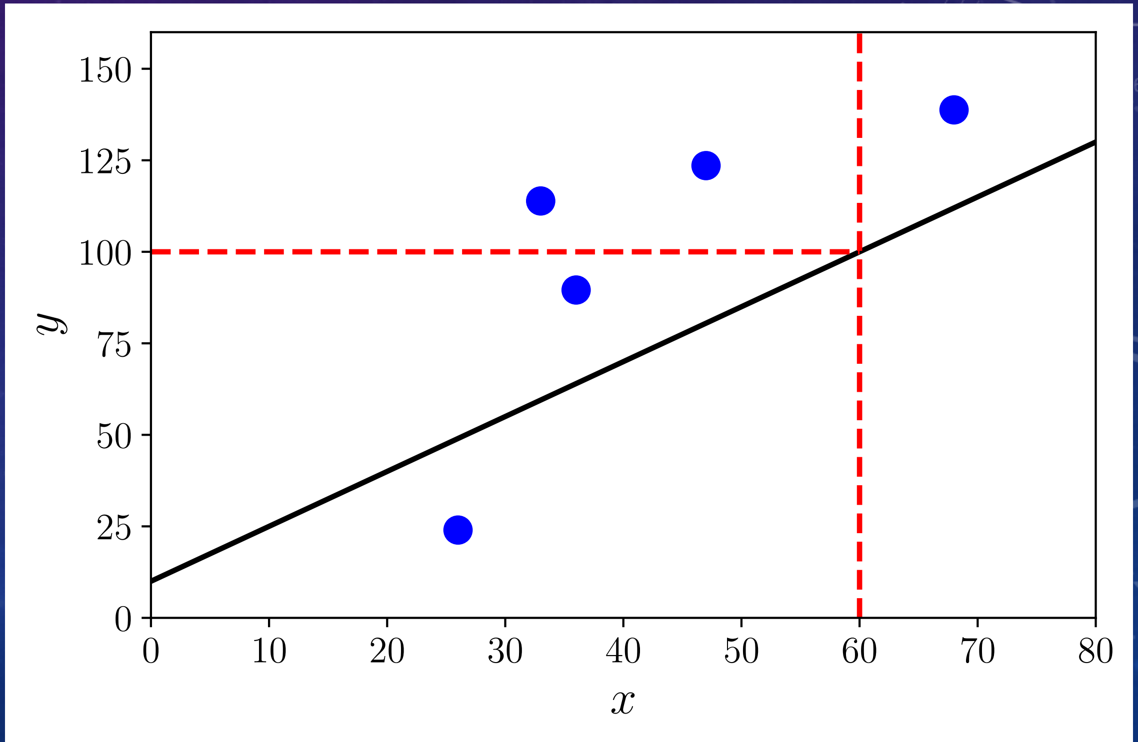
- Matrix: represent data vs represent transformations
- Linear vs Affine space: what is a linear regressor?

FITTING A LINE – LINEAR ALGEBRA

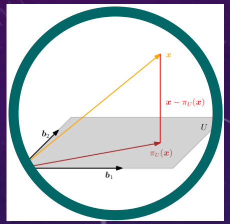
- Want to solve X**w** = **y**
- If points don't fall perfectly on the line, no solution
- Find a point z that lies in the column space of X and is closest to **y**

Age	Annual Salary (in thousands)
36	89.563
47	123.543
26	23.989
68	138.769
33	113.888

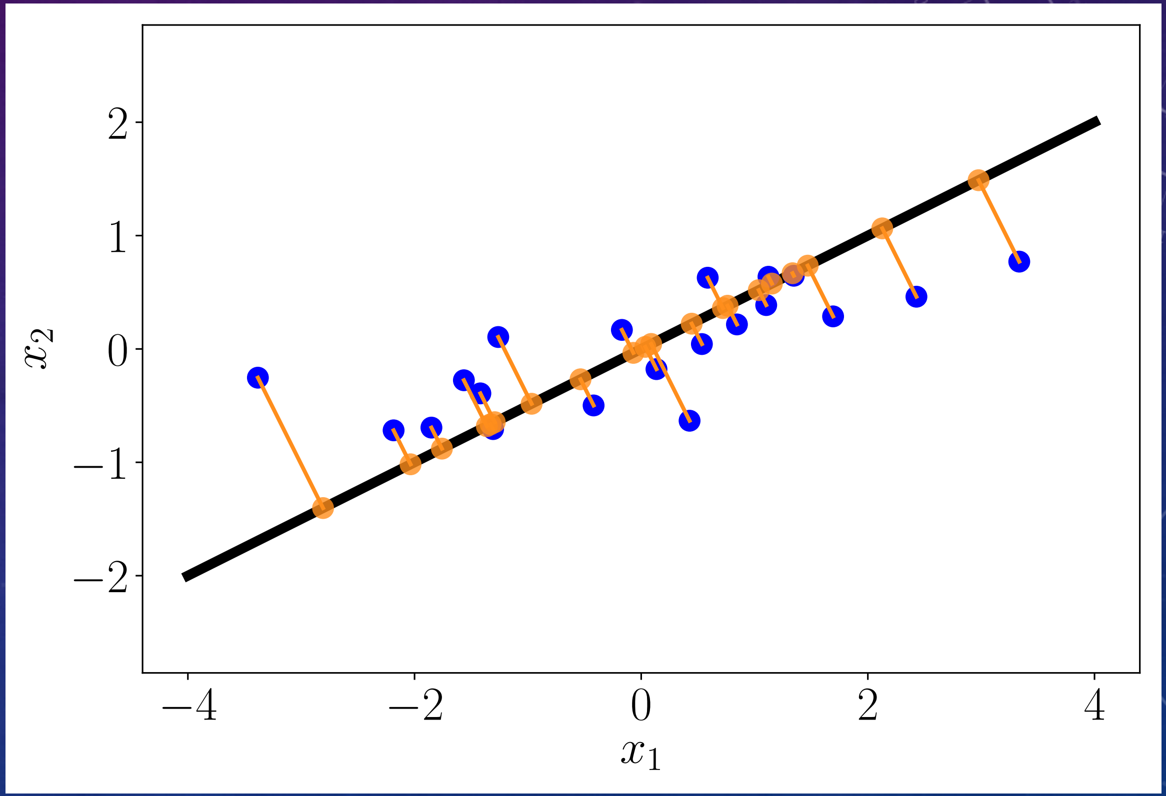
Closest?



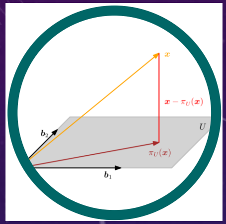
ANALYTIC GEOMETRY



- Inner products
- Distances
- Orthogonality
- Orthogonal projection



ANALYTIC GEOMETRY



- Inner products

Definition 3.2. Let V be a vector space and $\Omega : V \times V \to \mathbb{R}$ be a bilinear mapping that takes two vectors and maps them onto a real number. Then

- Ω is called *symmetric* if $\Omega(\boldsymbol{x}, \boldsymbol{y}) = \Omega(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in V$, i.e., the order of the arguments does not matter.
- Ω is called *positive definite* if

$$\forall \boldsymbol{x} \in V \setminus \{\boldsymbol{0}\} : \Omega(\boldsymbol{x}, \boldsymbol{x}) > 0 \,, \quad \Omega(\boldsymbol{0}, \boldsymbol{0}) = 0 \,. \tag{3.8}$$

A positive definite, symmetric bilinear mapping $\Omega : V \times V \to \mathbb{R}$ is called an *inner product* on V. We typically write $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ instead of $\Omega(\boldsymbol{x}, \boldsymbol{y})$.

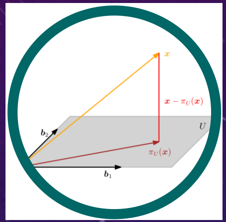
- Distances

$$\|\boldsymbol{x}\| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$$

Definition 3.6 (Distance and Metric)**.** Consider an inner product space $(V, \langle \cdot, \cdot \rangle)$. Then

$$d(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} - \boldsymbol{y}\| = \sqrt{\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y} \rangle} \tag{3.21}$$

ANALYTIC GEOMETRY



- Orthogonality

Definition 3.7 (Orthogonality). Two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are *orthogonal* if and only if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$, and we write $\boldsymbol{x} \perp \boldsymbol{y}$. If additionally $\|\boldsymbol{x}\| = 1 = \|\boldsymbol{y}\|$, i.e., the vectors are unit vectors, then $\boldsymbol{x}$ and $\boldsymbol{y}$ are *orthonormal*.

- Orthogonal projection (recall linear mapping = transformation matrix)

Definition 3.10 (Projection). Let V be a vector space and $U \subseteq V$ a subspace of V. A linear mapping $\pi : V \to U$ is called a *projection* if $\pi^2 = \pi \circ \pi = \pi$.

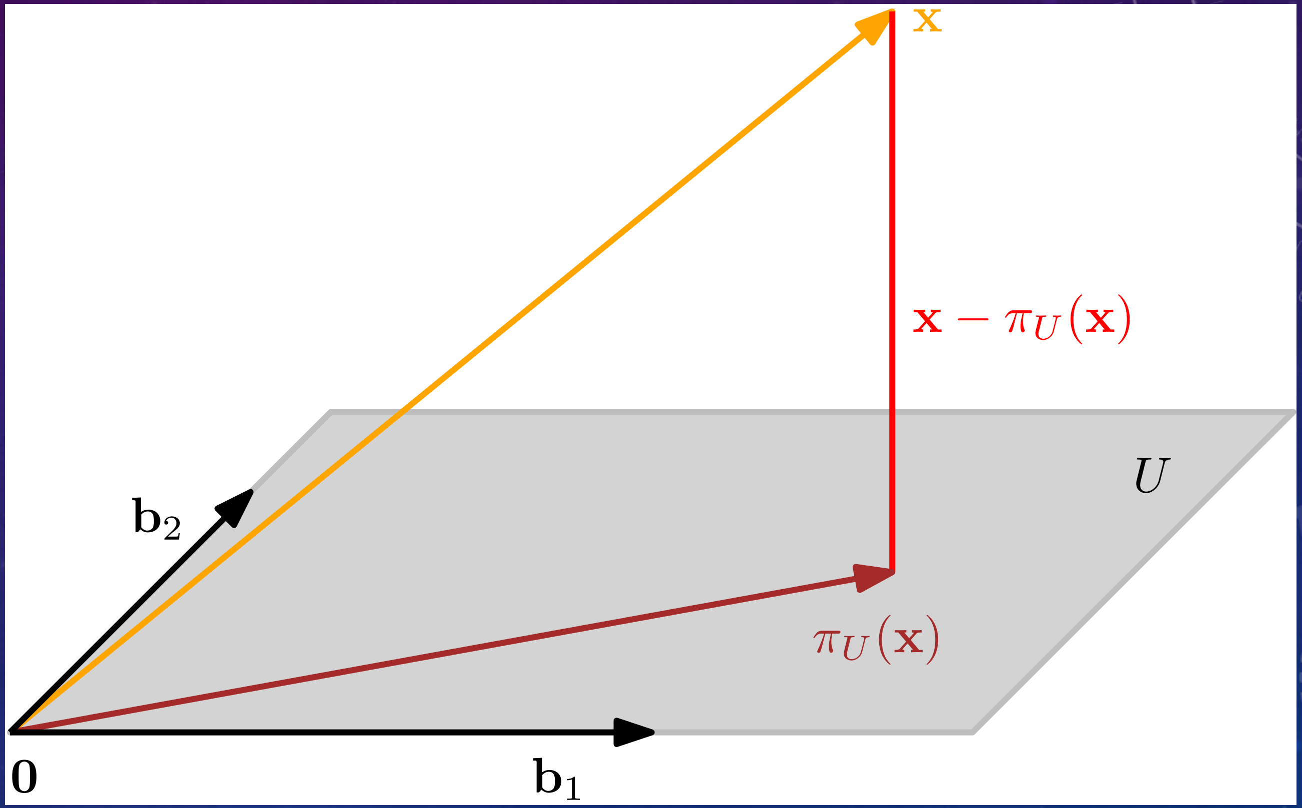
The projection $\pi_U(\boldsymbol{x})$ is closest to $\boldsymbol{x}$, where "closest" implies that the distance $\|\boldsymbol{x} - \pi_U(\boldsymbol{x})\|$ is minimal. It follows that the segment $\pi_U(\boldsymbol{x}) - \boldsymbol{x}$ from $\pi_U(\boldsymbol{x})$ to $\boldsymbol{x}$ is orthogonal to U, and therefore the basis vector $\boldsymbol{b}$ of U. The orthogonality condition yields $\langle \pi_U(\boldsymbol{x}) - \boldsymbol{x}, \boldsymbol{b} \rangle = 0$ since angles between vectors are defined via the inner product.

The projection $\pi_U(\boldsymbol{x})$ of $\boldsymbol{x}$ onto U must be an element of U and, therefore, a multiple of the basis vector $\boldsymbol{b}$ that spans U. Hence, $\pi_U(\boldsymbol{x}) = \lambda \boldsymbol{b}$, for some $\lambda \in \mathbb{R}$.

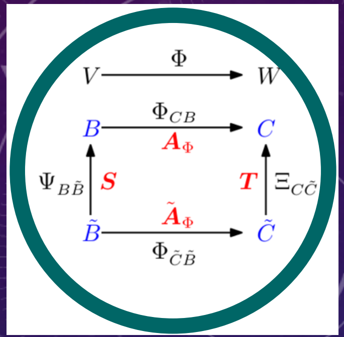
ANALYTIC GEOMETRY

- Want to solve X**w** = **y**

- Find a point z that lies in the column space of X and is closest to **y**
- z is found by the orthogonal projection of **y** onto the column space of X



REMINDER OF MATRIX OPERATIONS



- Want to solve X**w** = **y**

- Solutions of linear equations
- Inverse and transpose

Definition 2.3 (Inverse). Consider a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Let matrix $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ have the property that $\boldsymbol{AB} = \boldsymbol{I}_n = \boldsymbol{BA}$. $\boldsymbol{B}$ is called the *inverse* of $\boldsymbol{A}$ and denoted by $\boldsymbol{A}^{-1}$.

Definition 2.4 (Transpose). For $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ the matrix $\boldsymbol{B} \in \mathbb{R}^{n\times m}$ with $b_{ij} = a_{ji}$ is called the *transpose* of $\boldsymbol{A}$. We write $\boldsymbol{B} = \boldsymbol{A}^\top$.

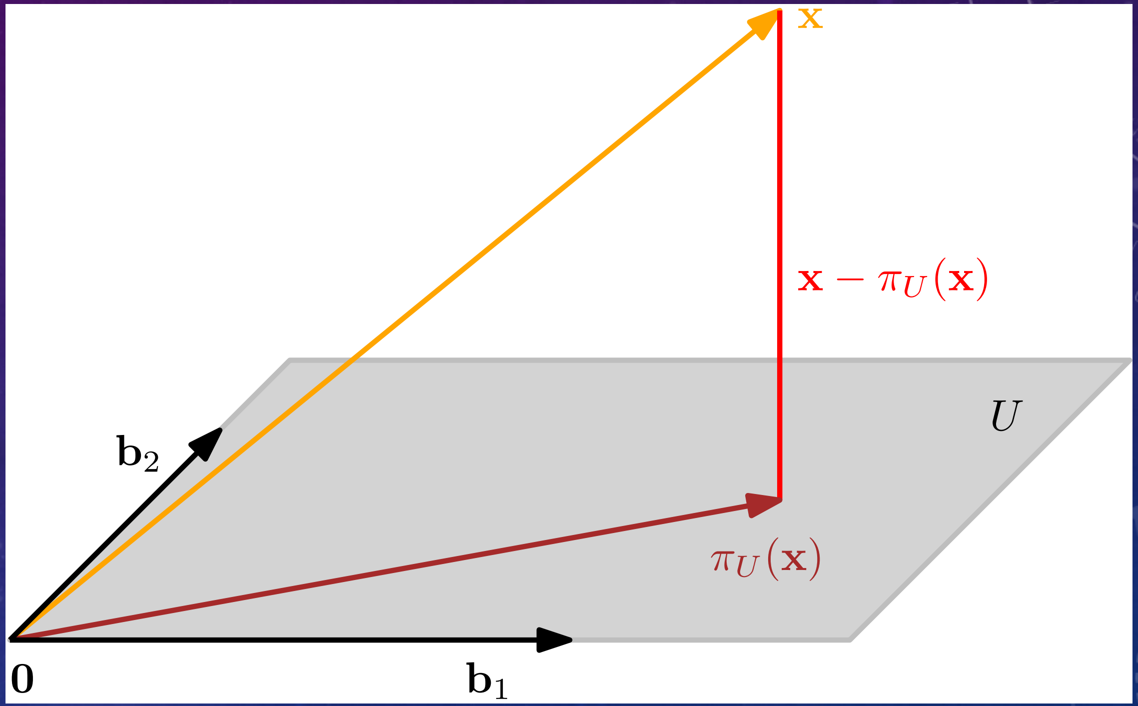
$$
\begin{aligned}
\boldsymbol{AA}^{-1} &= \boldsymbol{I} = \boldsymbol{A}^{-1}\boldsymbol{A} \\
(\boldsymbol{AB})^{-1} &= \boldsymbol{B}^{-1}\boldsymbol{A}^{-1} \\
(\boldsymbol{A}+\boldsymbol{B})^{-1} &\neq \boldsymbol{A}^{-1} + \boldsymbol{B}^{-1} \\
(\boldsymbol{A}^\top)^\top &= \boldsymbol{A} \\
(\boldsymbol{A}+\boldsymbol{B})^\top &= \boldsymbol{A}^\top + \boldsymbol{B}^\top \\
(\boldsymbol{AB})^\top &= \boldsymbol{B}^\top\boldsymbol{A}^\top
\end{aligned}
$$

ANALYTIC GEOMETRY

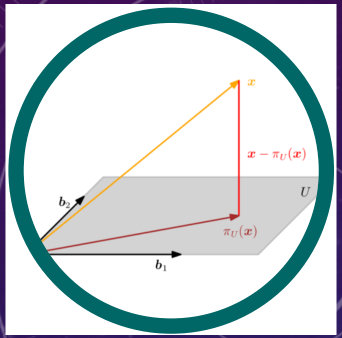
- z is found by the orthogonal projection of **y** onto the column space of X
- Column space of X is spanned by {x,1}, and hence we need to find coordinates w_1 and w_0 of the projection, such that the linear combination Xw is closest to y.
- Closest means that the vector connecting z to y is orthogonal to the column space of X.

  $X^T(y - z) = 0 \rightarrow X^T(y\text{-}Xw) = 0$

- Solving gives $w = (X^TX)^{-1} X^Ty$



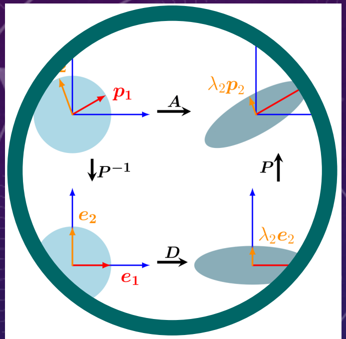
ANALYTIC GEOMETRY



- Want to solve X**w** = **y**
- Find a point z that lies in the column space of X and is closest to **y**
- z is found by the orthogonal projection of **y** onto the column space of X

- N = 5
- $\mathbf{x} = [x_1, \ldots, x_N]^T$
- $\mathbf{y} = [y_1, \ldots, y_N]^T$
- x_n is a real number
- y_n is a real number
- $f(x) = w_1 x + w_0$
- $X = [\mathbf{x}\ 1]$
- $\mathbf{w}^* = (X^TX)^{-1} X^T\mathbf{y}$

MATRIX DECOMPOSITIONS



- How do we compute the inverse of a matrix?

$$A^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

$$\det(A) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

- But for matrices that are larger, we do not have a closed form rule.
- Recall that linear mappings have an associated transformation matrix
- Disentangle different parts by an eigenvalue decomposition (inverse of a diagonal matrix is easy)

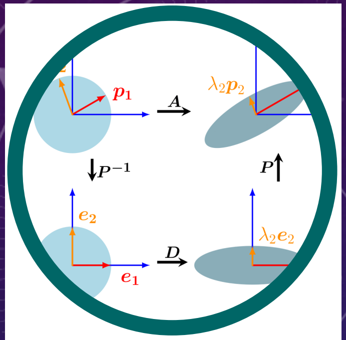
Theorem 4.20 (Eigendecomposition)**.** *A square matrix* $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ *can be factored into*

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}, \quad (4.55)$$

where $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ *and* $\boldsymbol{D}$ *is a diagonal matrix whose diagonal entries are the eigenvalues of* $\boldsymbol{A}$*, if and only if the eigenvectors of* $\boldsymbol{A}$ *form a basis of* $\mathbb{R}^n$*.*

According to the Abel–Ruffini theorem, there is in general no algebraic solution for polynomials of degree 5 or more (Abel, 1826).

OTHER MATRIX DECOMPOSITIONS



- For positive definite matrices

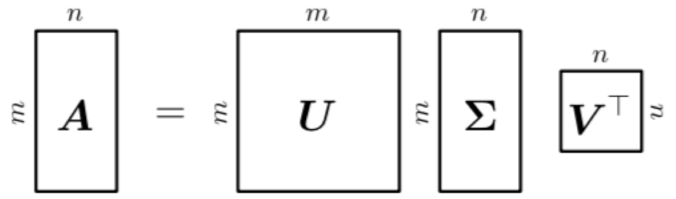
$$\forall \boldsymbol{x} \in V \backslash \{\boldsymbol{0}\} : \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} > 0 .$$

  we have the Cholesky decomposition

Theorem 4.18 (Cholesky Decomposition). *A symmetric, positive definite matrix $\boldsymbol{A}$ can be factorized into a product $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$, where $\boldsymbol{L}$ is a lower-triangular matrix with positive diagonal elements:*

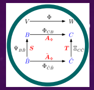
- For non-square matrices we have the singular value decomposition

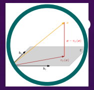
Theorem 4.22 (SVD Theorem). *Let $\boldsymbol{A}^{m \times n}$ be a rectangular matrix of rank $r \in [0, \min(m, n)]$. The SVD of $\boldsymbol{A}$ is a decomposition of the form*

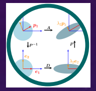


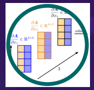
(4.64)

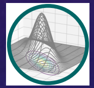
SUMMARY: MATH FOR ML

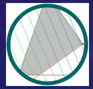

Chap. 2: Linear Algebra
vector space, linear maps, affine space

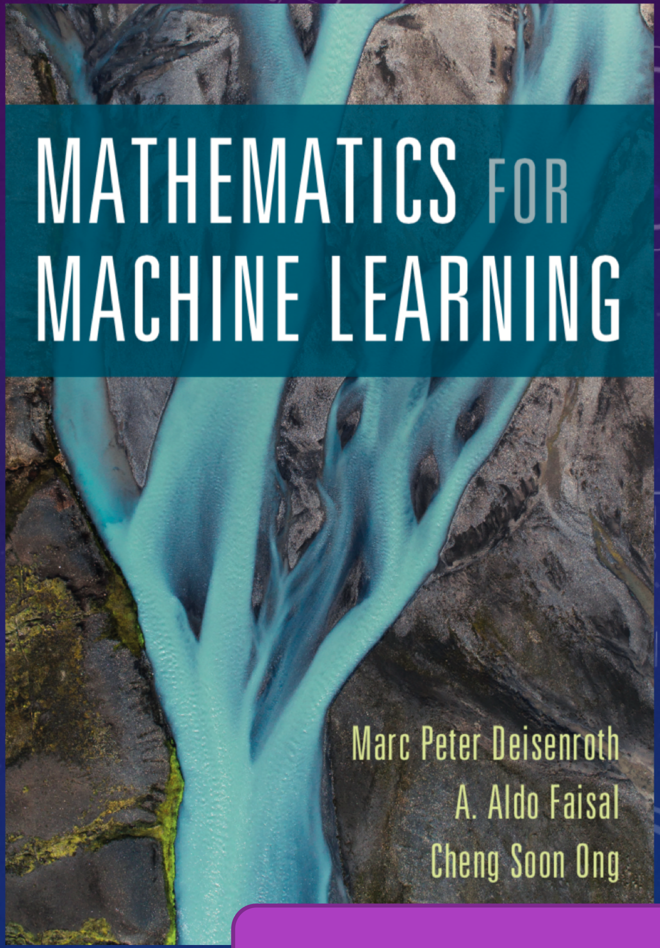

Chap. 3: Analytic Geometry
inner products, distances, orthogonality


Chap. 4: Matrix Decompositions
Cholesky, eigenvalue decomposition, singular value decomposition









mml-book.com

ong-home.my

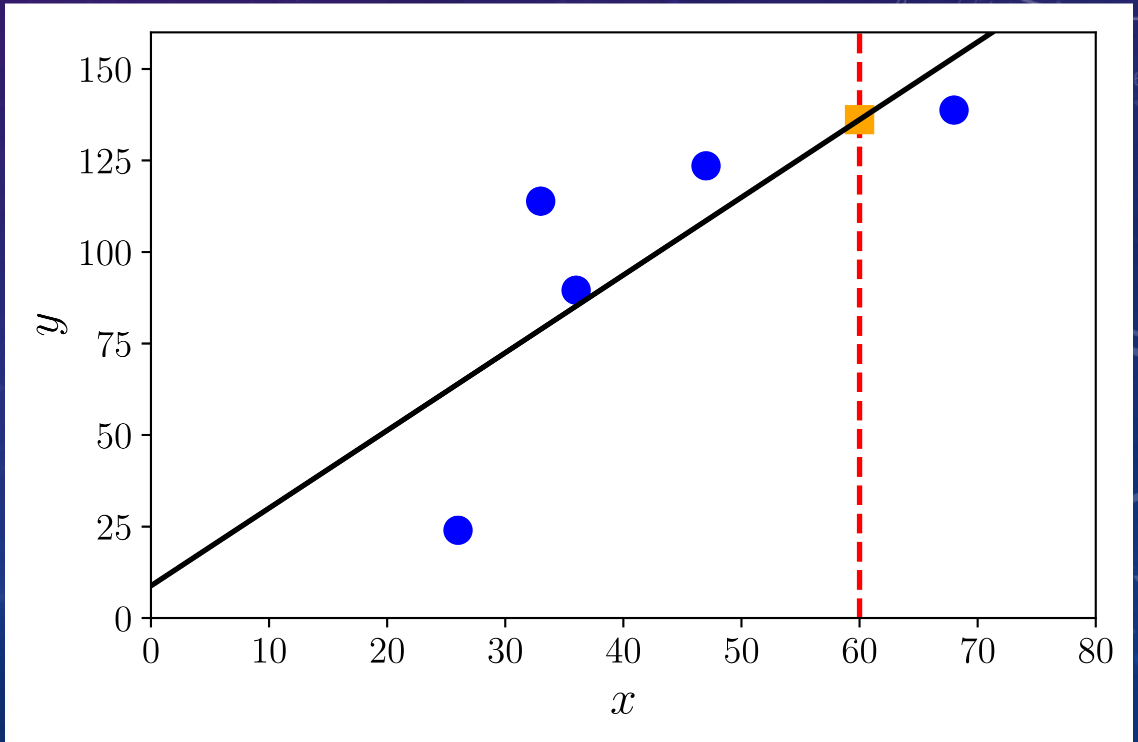
FITTING A LINE – OPTIMIZATION

- Want to solve $X\mathbf{w} = \mathbf{y}$
- If points don't fall perfectly on the line, no solution
- Instead, find the closest approximate solution

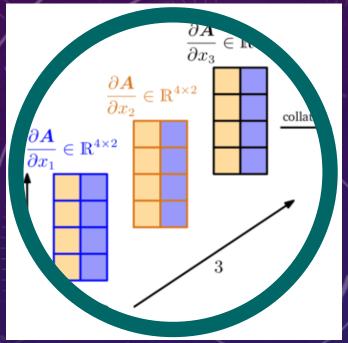
$$\min_w || Xw - y ||^2$$

- Solve for a minimum by taking the gradient and setting to zero.

Age	Annual Salary (in thousands)
36	89.563
47	123.543
26	23.989
68	138.769
33	113.888



VECTOR CALCULUS - GRADIENT



$$\ell(x) = x^4 + 7x^3 + 5x^2 - 17x + 3\,,$$

Univariate calculus

$$\frac{\mathrm{d}\ell(x)}{\mathrm{d}x} = 4x^3 + 21x^2 + 10x - 17\,.$$

Definition 5.5 (Partial Derivative)**.** For a function $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{x} \mapsto f(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^n$ of n variables $x_1, \ldots, x_n$ we define the *partial derivatives* as

$$\begin{aligned} \frac{\partial f}{\partial x_1} &= \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(\boldsymbol{x})}{h} \\ &\vdots \\ \frac{\partial f}{\partial x_n} &= \lim_{h \to 0} \frac{f(x_1, \ldots, x_{n-1}, x_n + h) - f(\boldsymbol{x})}{h} \end{aligned} \tag{5.39}$$

and collect them in the row vector

$$\nabla_{\boldsymbol{x}} f = \mathrm{grad} f = \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} & \frac{\partial f(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{1 \times n}, \tag{5.40}$$

row vector

FITTING A LINE – OPTIMIZATION

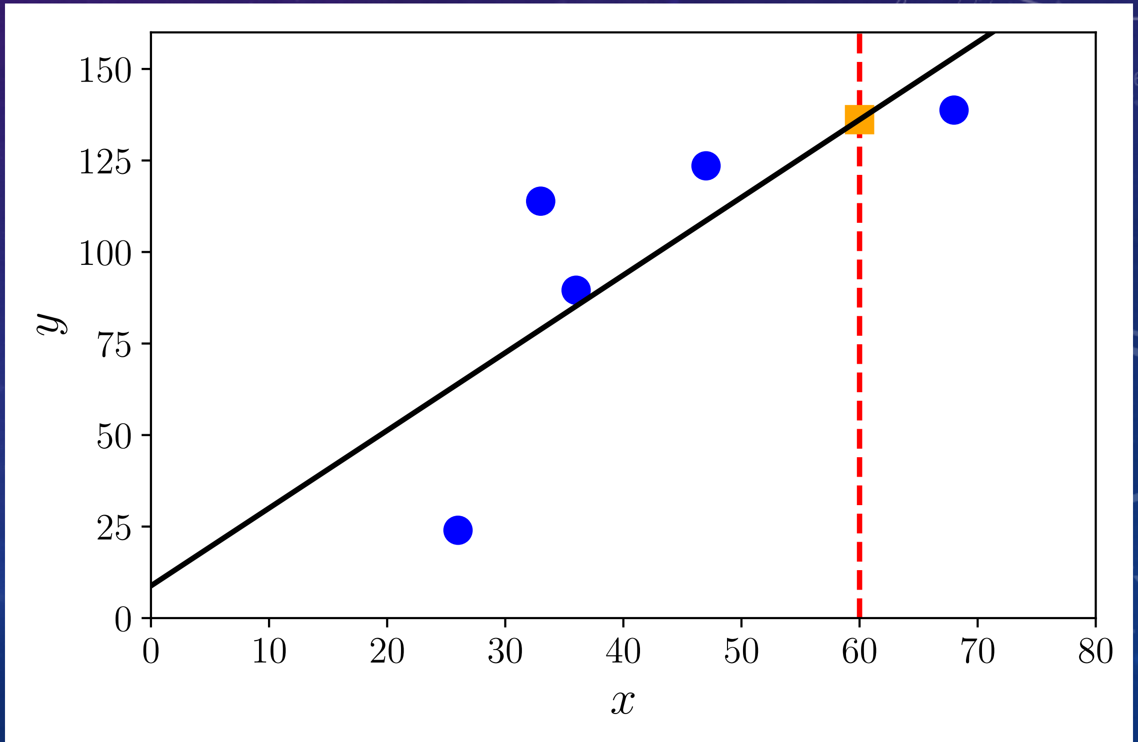
- Find the closest approximate solution

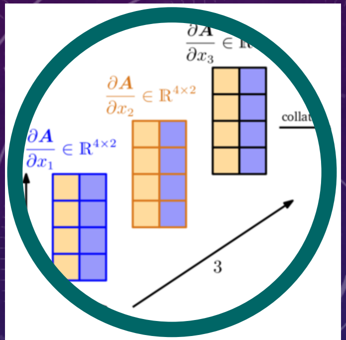
  $\min_w || Xw - y ||^2$

- Solve for a minimum by taking the gradient and setting to zero.
- Gradient (wrt w) is $2 (Xw - y)^T X$
- Solving for stationary point gives

  $X^T X w = X^T y$

Age	Annual Salary (in thousands)
36	89.563
47	123.543
26	23.989
68	138.769
33	113.888





VECTOR CALCULUS - JACOBIAN

Vector valued functions

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^m .$$

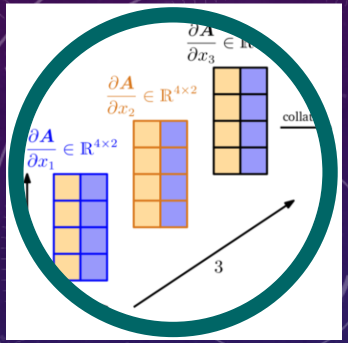
$$\frac{\partial \boldsymbol{f}}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix}$$

Definition 5.6 (Jacobian). The collection of all first-order partial derivatives of a vector-valued function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ is called the *Jacobian*. The Jacobian $\boldsymbol{J}$ is an $m \times n$ matrix, which we define and arrange as follows:

$$\boldsymbol{J} = \nabla_{\boldsymbol{x}} \boldsymbol{f} = \frac{\mathrm{d}\boldsymbol{f}(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \qquad (5.57)$$

The gradient of a function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ is a matrix of size $m \times n$.

SUM RULE, PRODUCT RULE, CHAIN RULE



Product rule:

$$(fg)' = f'g + fg',$$

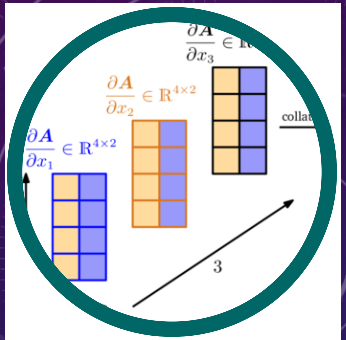
Sum rule:

$$(f+g)' = f' + g',$$

Chain rule:

$$(g(f))' = g'(f)f'$$

SUM RULE, PRODUCT RULE, CHAIN RULE

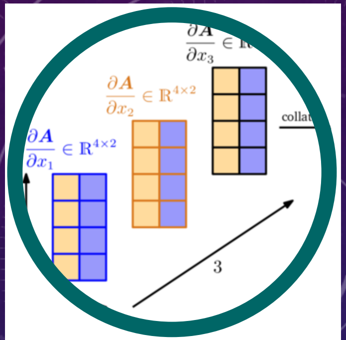


$$\text{Product rule:} \quad \frac{\partial}{\partial \boldsymbol{x}}\big(f(\boldsymbol{x})g(\boldsymbol{x})\big) = \frac{\partial f}{\partial \boldsymbol{x}} g(\boldsymbol{x}) + f(\boldsymbol{x}) \frac{\partial g}{\partial \boldsymbol{x}} \tag{5.46}$$

$$\text{Sum rule:} \quad \frac{\partial}{\partial \boldsymbol{x}}\big(f(\boldsymbol{x}) + g(\boldsymbol{x})\big) = \frac{\partial f}{\partial \boldsymbol{x}} + \frac{\partial g}{\partial \boldsymbol{x}} \tag{5.47}$$

Product rule:
$(fg)' = f'g + fg'$,
Sum rule:
$(f+g)' = f' + g'$,
Chain rule:
$(g(f))' = g'(f)f'$

SUM RULE, PRODUCT RULE, CHAIN RULE



Product rule: $$\frac{\partial}{\partial \boldsymbol{x}}\big(f(\boldsymbol{x})g(\boldsymbol{x})\big) = \frac{\partial f}{\partial \boldsymbol{x}} g(\boldsymbol{x}) + f(\boldsymbol{x})\frac{\partial g}{\partial \boldsymbol{x}} \tag{5.46}$$

Sum rule: $$\frac{\partial}{\partial \boldsymbol{x}}\big(f(\boldsymbol{x}) + g(\boldsymbol{x})\big) = \frac{\partial f}{\partial \boldsymbol{x}} + \frac{\partial g}{\partial \boldsymbol{x}} \tag{5.47}$$

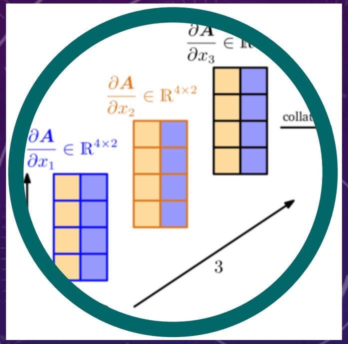
Product rule:
$(fg)' = f'g + fg'$,
Sum rule:
$(f+g)' = f' + g'$,
Chain rule:
$(g(f))' = g'(f)f'$

If $f(x_1, x_2)$ is a function of x_1 and x_2, where $x_1(s,t)$ and $x_2(s,t)$ themselves functions of two variables s and t, the chain rule yields

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s},$$
$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t},$$

$$\frac{\mathrm{d}f}{\mathrm{d}(s,t)} = \frac{\partial f}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial (s,t)} = \underbrace{\begin{bmatrix}\frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2}\end{bmatrix}}_{=\frac{\partial f}{\partial \boldsymbol{x}}} \underbrace{\begin{bmatrix}\frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t}\end{bmatrix}}_{=\frac{\partial \boldsymbol{x}}{\partial (s,t)}}$$

CHAIN RULE



If $f(x_1, x_2)$ is a function of x_1 and x_2, where $x_1(s,t)$ and $x_2(s,t)$ themselves functions of two variables s and t, the chain rule yields

$$\frac{\mathrm{d}f}{\mathrm{d}(s,t)} = \frac{\partial f}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{x}}{\partial(s,t)} = \underbrace{\begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix}}_{= \frac{\partial f}{\partial \boldsymbol{x}}} \underbrace{\begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t} \end{bmatrix}}_{= \frac{\partial \boldsymbol{x}}{\partial(s,t)}}$$

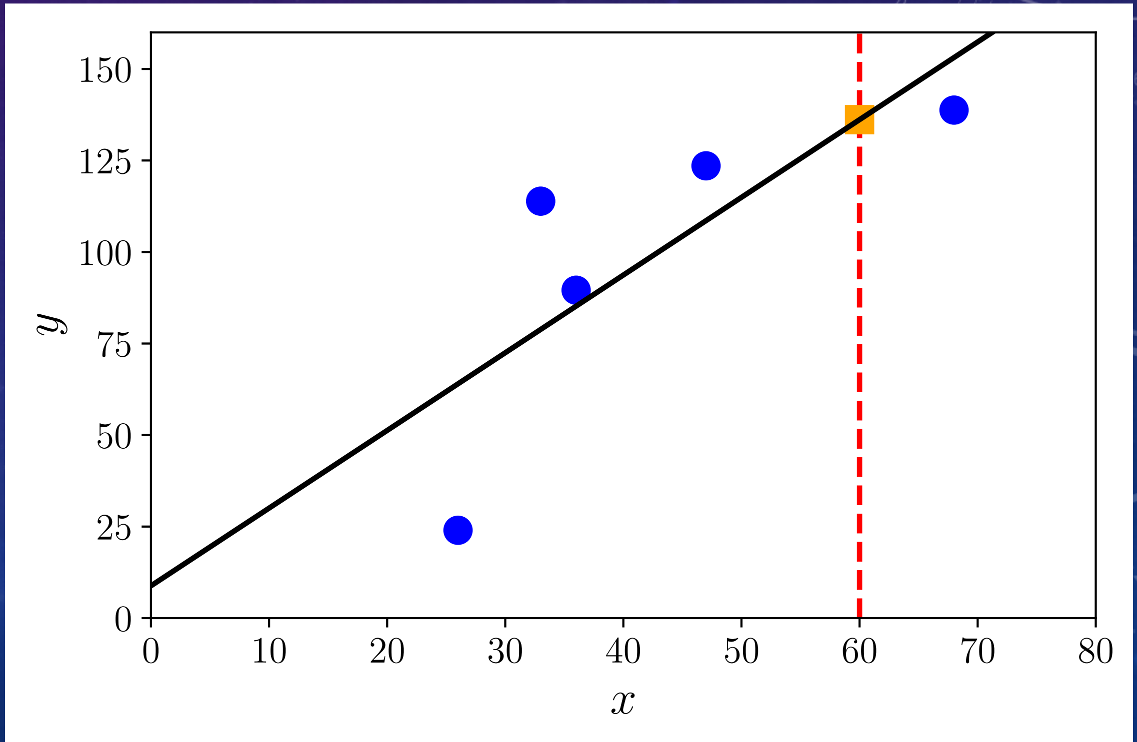
FITTING A LINE – OPTIMIZATION

- Want to solve X**w** = **y**
- If points don't fall perfectly on the line, no solution
- Instead, find the closest approximate solution

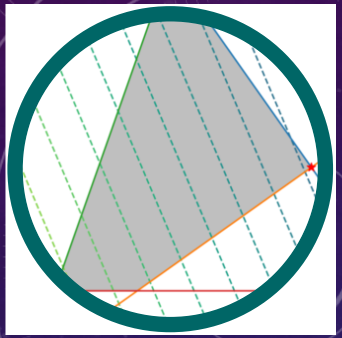
  $\min_w \quad || Xw - y ||^2$

- For some functions, we may not have a closed form solution for the minimum. Find minimum numerically.

Age	Annual Salary (in thousands)
36	89.563
47	123.543
26	23.989
68	138.769
33	113.888



CONTINUOUS OPTIMIZATION



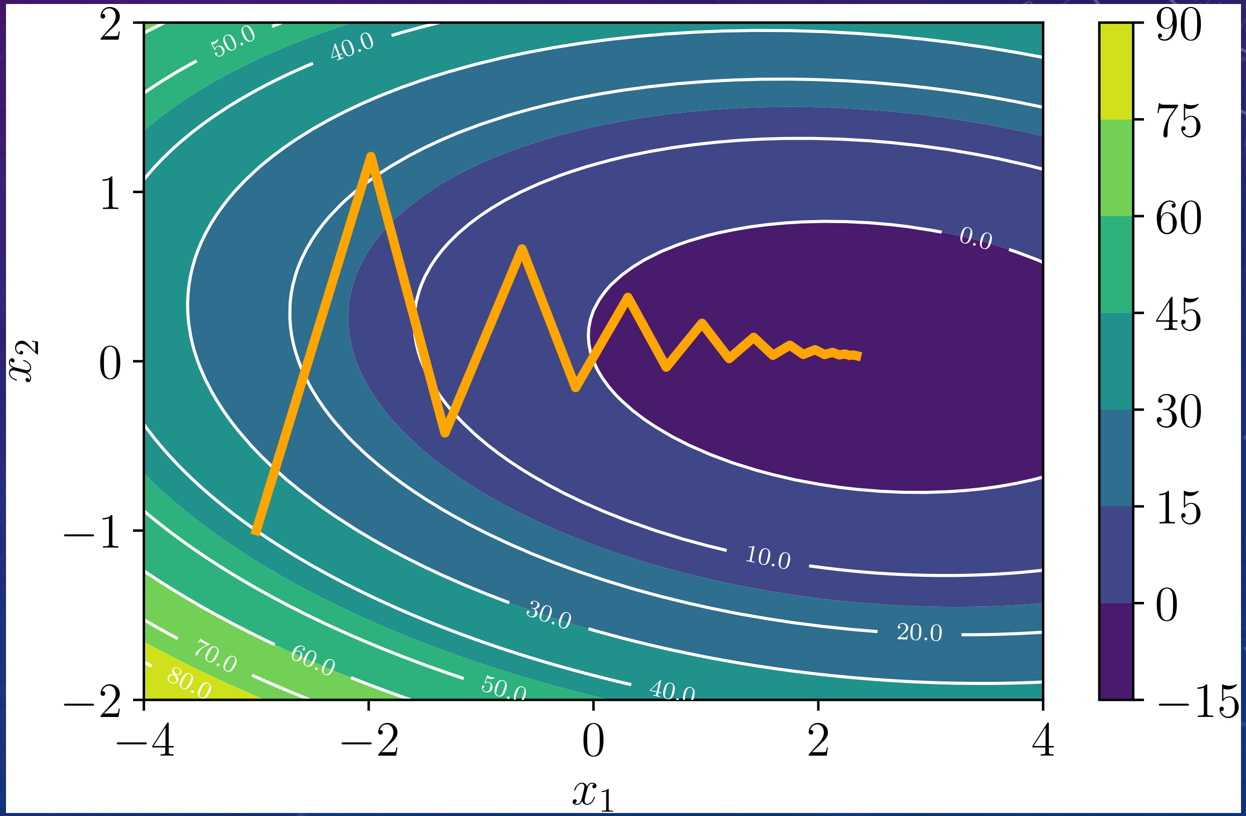
- Objective function

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

- Gradient Descent

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 - \gamma((\nabla f)(\boldsymbol{x}_0))^\top$$

- Gradient ∇f
- Step-size γ



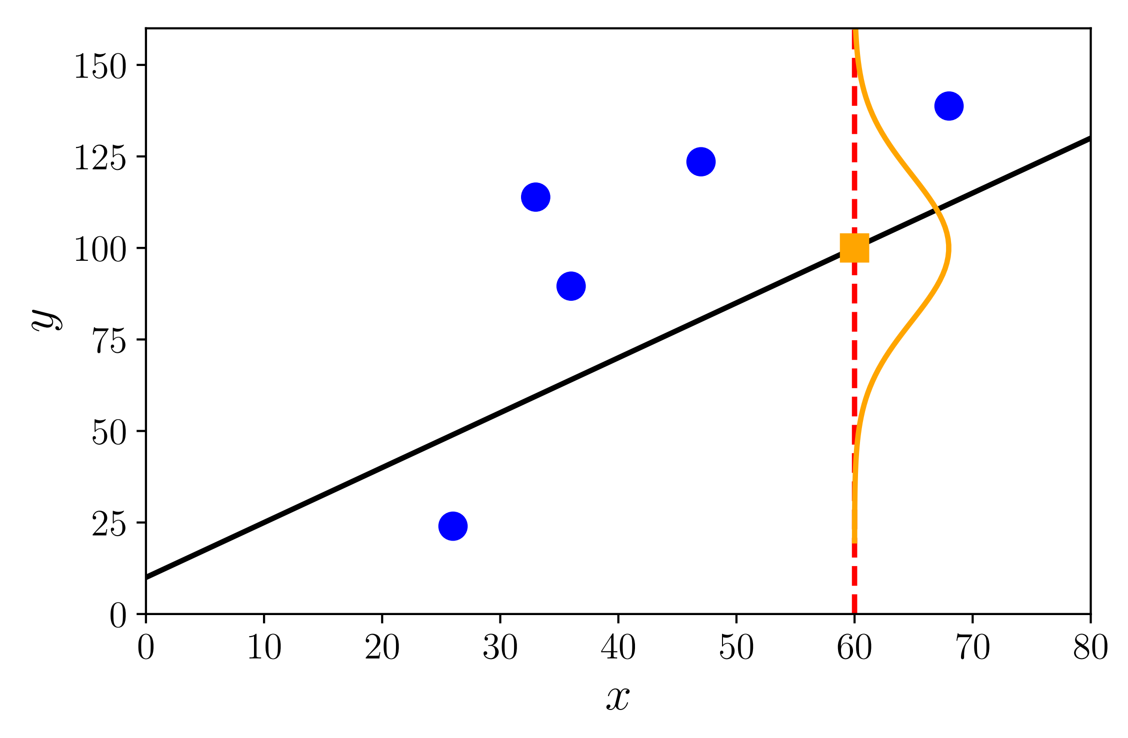
FITTING A LINE – MAXIMUM LIKELIHOOD

Age	Annual Salary (in thousands)
36	89.563
47	123.543
26	23.989
68	138.769
33	113.888

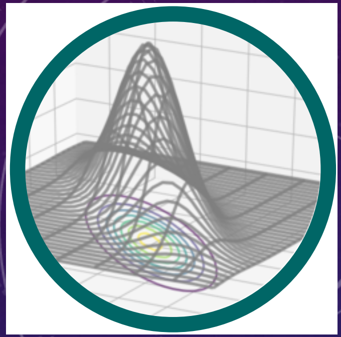
- Want to solve X**w** = **y**
- If points don't fall perfectly on the line, no solution
- Assume data (X, y) is represented by random variables
- And for a given family of probability densities, compute the maximum likelihood

$$\max_w \; p(y \mid X, w)$$

- What is the noise model?
- What is the prior?
- What is the predictive uncertainty?



PROBABILITY AND DISTRIBUTIONS



- Probability space
  - Sample space (Ω), e.g {hh, ht, th, tt}
  - Event space, e.g. one head = {ht, th}
  - Probability space, e.g. P(one head) = 0.5
- Random variable
  - Target space, e.g. discrete or real
  - Random variable is a function X

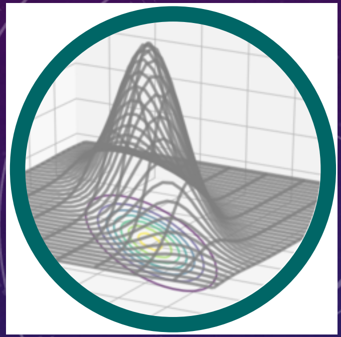
$$X : \Omega \to \mathcal{T}$$

$$S \subseteq \mathcal{T}$$

$$P_X(S) = P(X \in S) = P(X^{-1}(S)) = P(\{\omega \in \Omega : X(\omega) \in S\})$$

Type	"Point probability"	"Interval probability"
Discrete	$P(X = x)$ Probability mass function	Not applicable
Continuous	$p(x)$ Probability density function	$P(X \leqslant x)$ Cumulative distribution function

PROBABILITY AND DISTRIBUTIONS



Definition 6.1 (Probability Density Function). A function $f : \mathbb{R}^D \to \mathbb{R}$ is called a *probability density function* (*pdf*) if

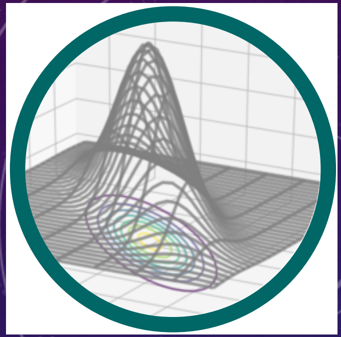
1. $\forall \boldsymbol{x} \in \mathbb{R}^D : f(\boldsymbol{x}) \geqslant 0$
2. Its integral exists and

$$\int_{\mathbb{R}^D} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1 \,. \tag{6.15}$$

- Distribution (or law) of the random variable $X : \Omega \to \mathcal{T}$

$$P(a \leqslant X \leqslant b) = \int_a^b f(x) \mathrm{d}x \,,$$

RULES OF PROBABILITY



- Sum rule

$$p(\boldsymbol{x}) = \begin{cases} \displaystyle\sum_{\boldsymbol{y} \in \mathcal{Y}} p(\boldsymbol{x}, \boldsymbol{y}) & \text{if } \boldsymbol{y} \text{ is discrete} \\ \displaystyle\int_{\mathcal{Y}} p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y} & \text{if } \boldsymbol{y} \text{ is continuous} \end{cases}$$

- Product rule

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x})$$

- Bayes' Theorem

$$\underbrace{p(\boldsymbol{x} \mid \boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y} \mid \boldsymbol{x})}^{\text{likelihood}} \overbrace{p(\boldsymbol{x})}^{\text{prior}}}{\underbrace{p(\boldsymbol{y})}_{\text{evidence}}}$$

Definition 6.3 (Expected Value). The *expected value* of a function $g : \mathbb{R} \to \mathbb{R}$ of a univariate continuous random variable $X \sim p(x)$ is given by

$$\mathbb{E}_X[g(x)] = \int_{\mathcal{X}} g(x) p(x) \mathrm{d}x \,. \tag{6.28}$$

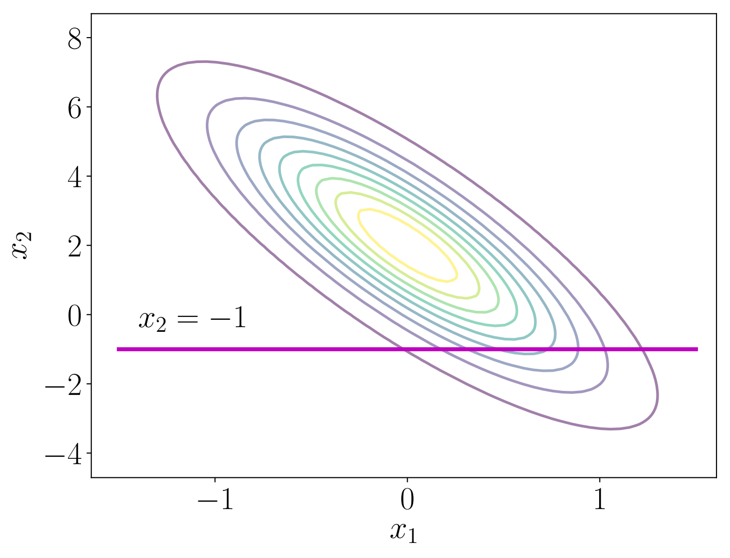
Definition 6.6 (Covariance (Multivariate)). If we consider two multivariate random variables X and Y with states $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{y} \in \mathbb{R}^E$ respectively, the *covariance* between X and Y is defined as

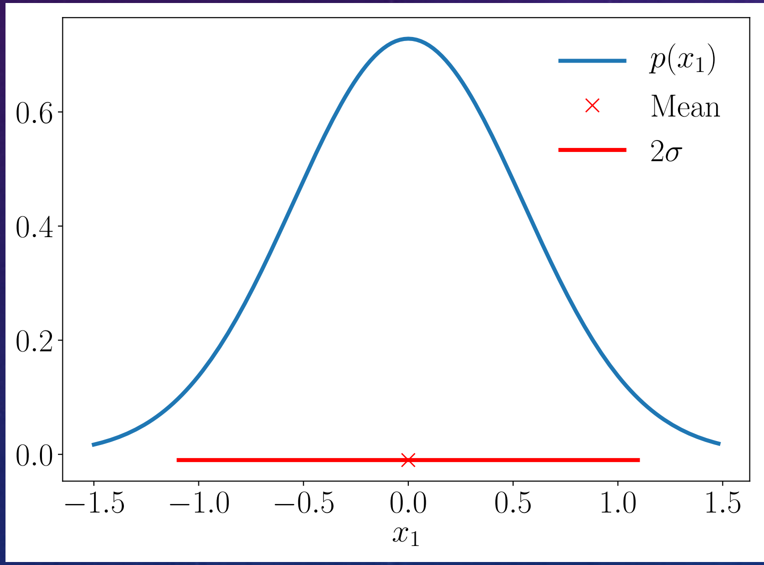
$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \mathbb{E}[\boldsymbol{x}\boldsymbol{y}^\top] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}]^\top = \mathrm{Cov}[\boldsymbol{y}, \boldsymbol{x}]^\top \in \mathbb{R}^{D \times E} \,. \tag{6.37}$$

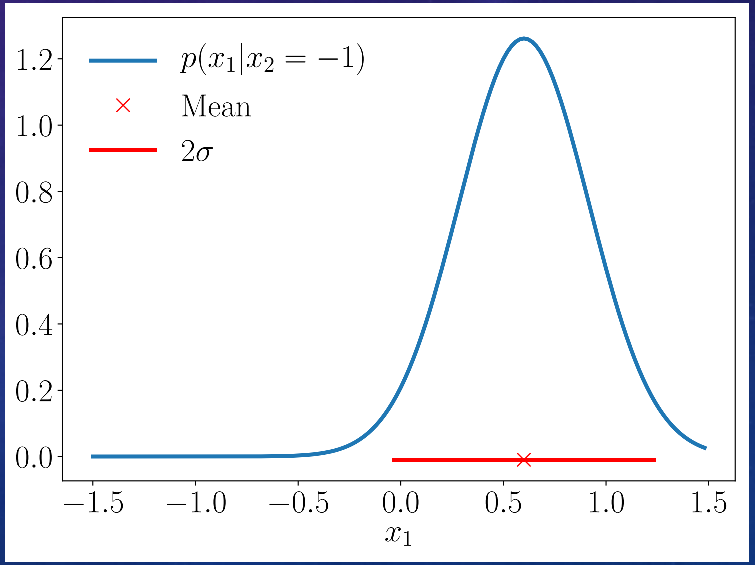
GAUSSIAN DISTRIBUTION

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$







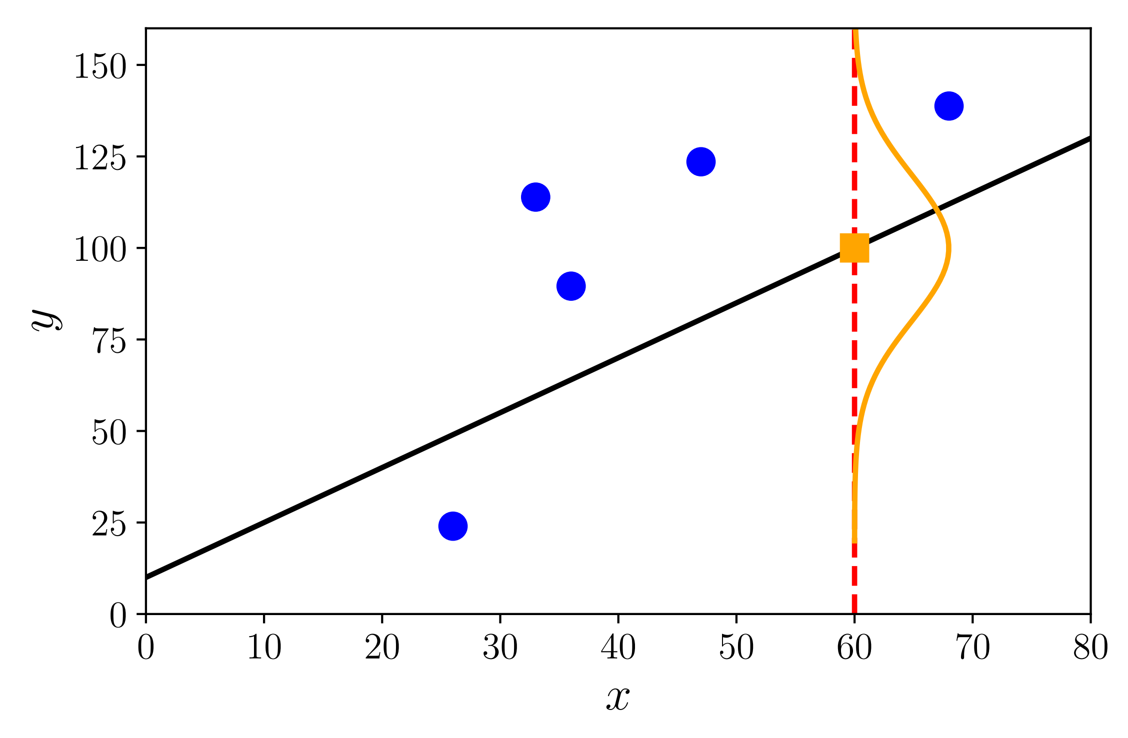
FITTING A LINE – MAXIMUM LIKELIHOOD

Age	Annual Salary (in thousands)
36	89.563
47	123.543
26	23.989
68	138.769
33	113.888

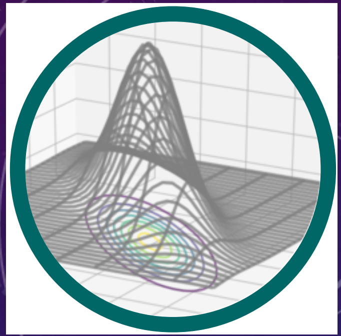
- Want to solve X**w** = **y**
- If points don't fall perfectly on the line, no solution
- Assume data (X, y) is represented by random variables
- And for a given family of probability densities, compute the maximum likelihood

  $\max_w \ p(y \mid X, w)$

- What is the noise model?
  We assume Gaussian noise



CONJUGACY AND EXPONENTIAL FAMILY



$$\underbrace{p(\boldsymbol{x} \mid \boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y} \mid \boldsymbol{x})}^{\text{likelihood}} \overbrace{p(\boldsymbol{x})}^{\text{prior}}}{\underbrace{p(\boldsymbol{y})}_{\text{evidence}}}$$

Bayes' Theorem

Definition 6.13 (Conjugate Prior)**.** A prior is *conjugate* for the likelihood function if the posterior is of the same form/type as the prior.

An *exponential family* is a family of probability distributions, parameterized by $\boldsymbol{\theta} \in \mathbb{R}^D$, of the form

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta})\right) , \qquad (6.107)$$

- Natural parameters θ
- Sufficient statistics $\phi(x)$
- Log partition function $A(\theta)$

Theorem 6.14 (Fisher-Neyman)**.** *[Theorem 6.5 in Lehmann and Casella (1998)] Let X have probability density function $p(x \mid \theta)$. Then the statistics $\phi(x)$ are sufficient for θ if and only if $p(x \mid \theta)$ can be written in the form*

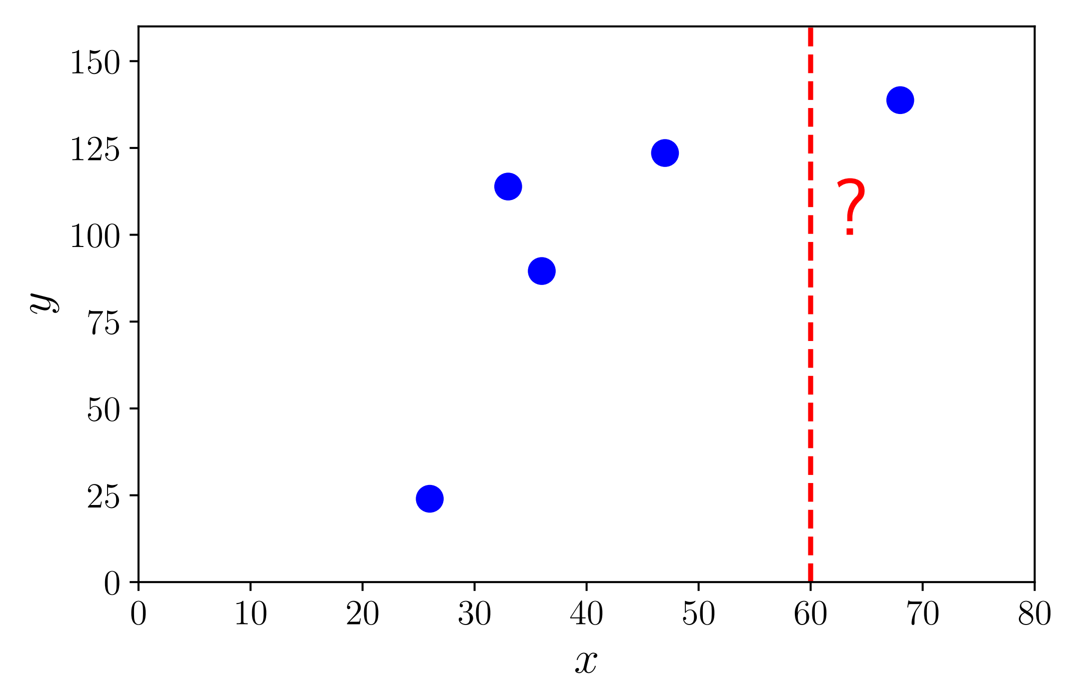
$$p(x \mid \theta) = h(x) g_\theta(\phi(x)) , \qquad (6.106)$$

where $h(x)$ is a distribution independent of θ and g_θ captures all the dependence on θ via sufficient statistics $\phi(x)$.

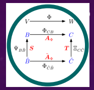
MACHINE LEARNING IS ABOUT PREDICTION

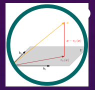
Age	Annual Salary (in thousands)
36	89.563
47	123.543
26	23.989
68	138.769
33	113.888

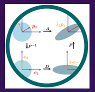
- Predict salary (y) from age (x)
- The values of y for the training data is not the main focus
- We are interested in generalization error:
    - What is the error we make on unseen data?
- Do not train on the test set

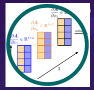


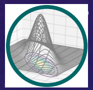
SUMMARY: MATH FOR ML

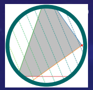

Chap. 2: Linear Algebra
vector space, linear maps, affine space


Chap. 3: Analytic Geometry
inner products, distances, orthogonality

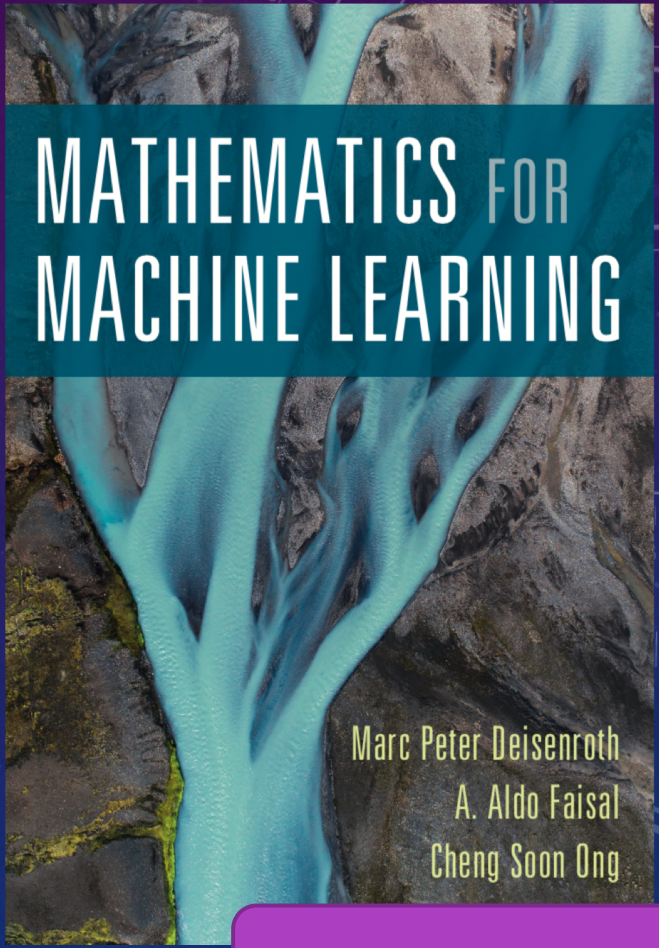

Chap. 4: Matrix Decompositions
Cholesky, eigenvalue decomposition, singular value decomposition


Chap. 5: Vector Calculus
gradient is a row vector, chain rule


Chap. 6: Probability and Distributions
random variable, distribution, Bayes rule, expectation


Chap. 7: Continuous Optimization
gradient descent, convex duality

… and 5 chapters of ML



mml-book.com

ong-home.my