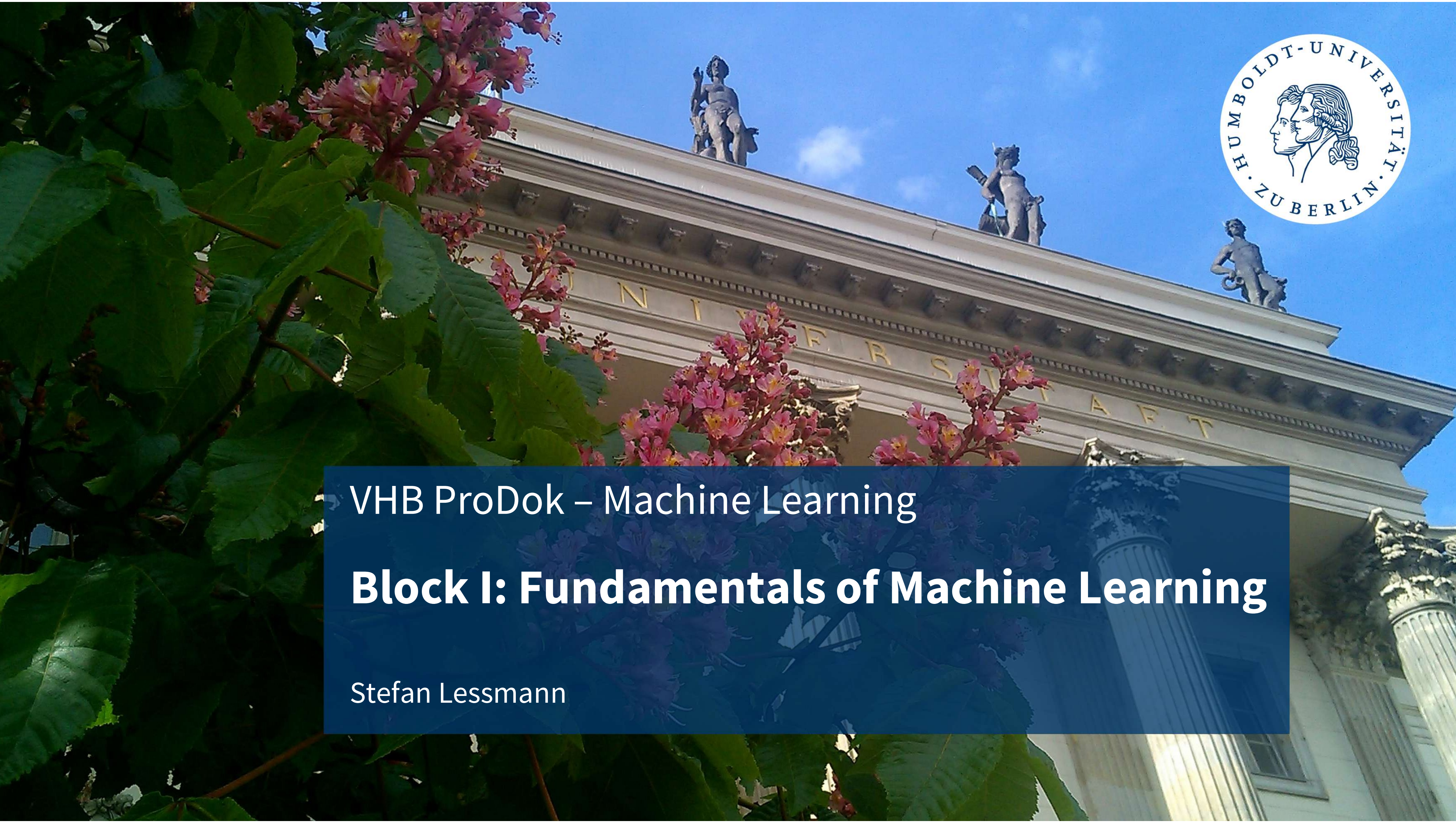
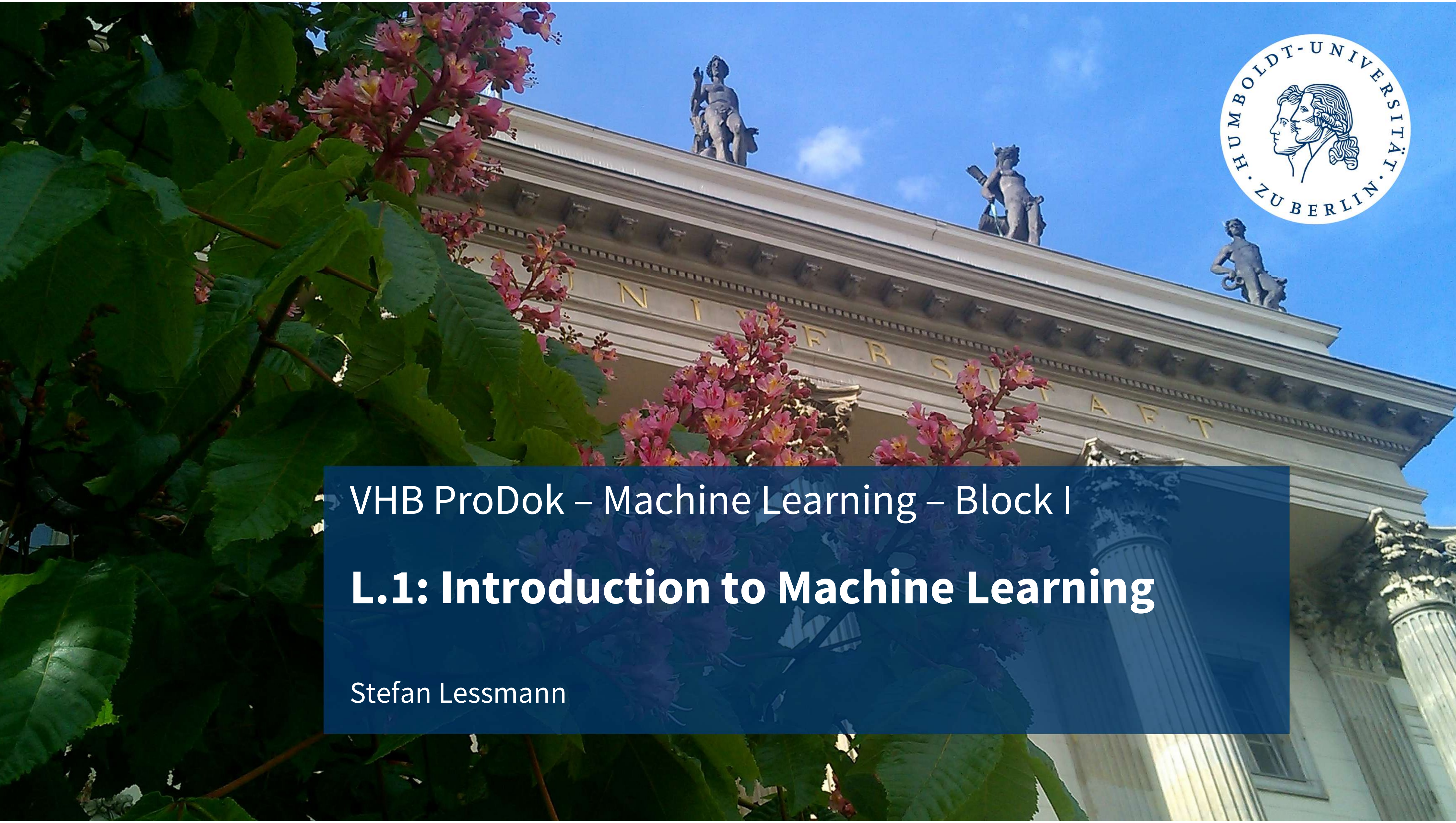


VHB ProDok – Machine Learning

Block I: Fundamentals of Machine Learning

Stefan Lessmann



VHB ProDok – Machine Learning – Block I

L.1: Introduction to Machine Learning

Stefan Lessmann

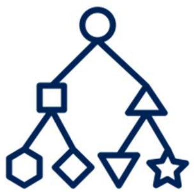
Machine Learning (ML) and Artificial Intelligence (AI) Delineated

Artificial Intelligence

Enable computers to mimic human behavior

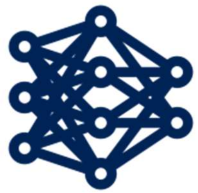
Machine Learning

Ability to learn without explicitly being programmed



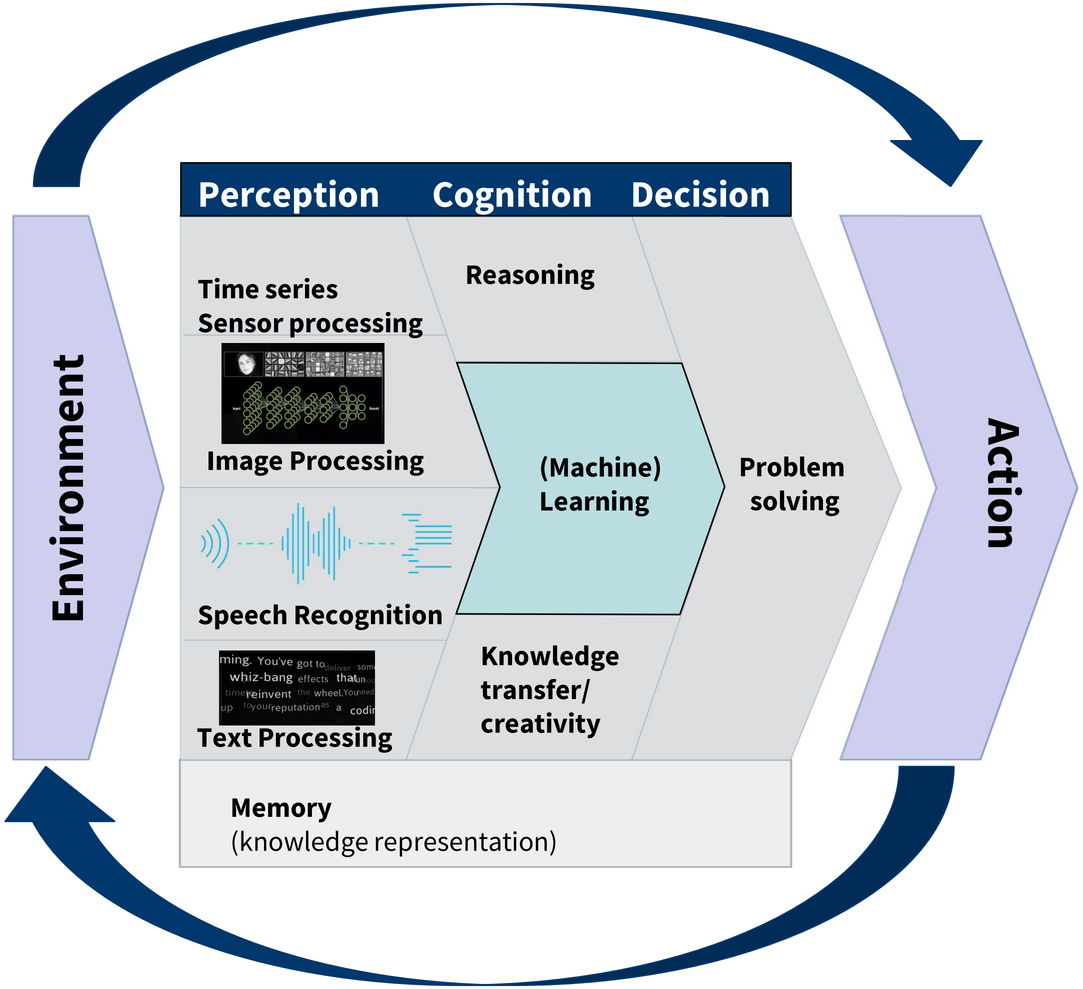
Deep Learning

Ability to automatically extract features from data using artificial neural networks



Machine Learning (ML) and Artificial Intelligence (AI)

AI and ML refer to different concepts and should be distinguished. "ML/AI" is a misnomer.



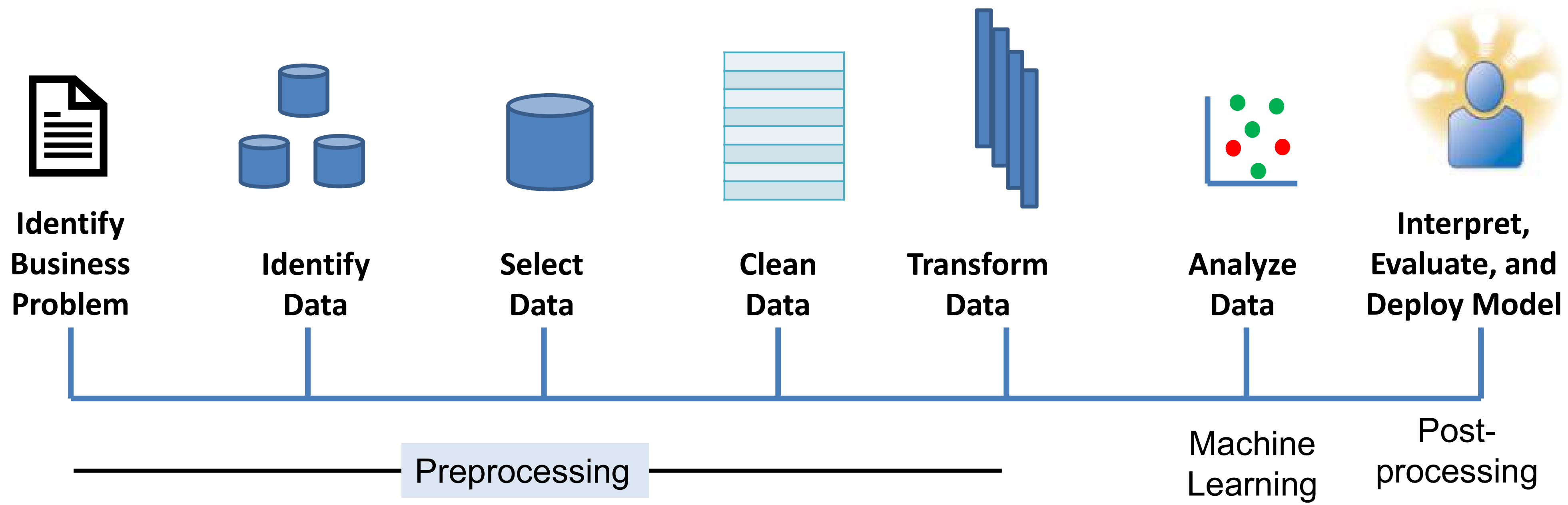
Famous definitions

- "*AI is the science and engineering of making computers behave in ways that, until recently, we thought required human intelligence.*" (McCarthy 1955, 2007; Moore, 2017)
- "*A computer program is said to learn from experience $\boldsymbol{E}$ with respect to some class of tasks $\boldsymbol{T}$ and performance measure $\boldsymbol{P}$, if its performance at tasks in $\boldsymbol{T}$, as measured by $\boldsymbol{P}$, improves with experience $\boldsymbol{E}$.*" (Mitchell, 1997)

Intuitive delineation

- AI is goal-oriented (i.e., intelligent behavior)
  - Contemporary understanding shaped by LLMs
  - Systems that understand and generate texts
- ML is method-oriented
  - Achieving *intelligent* behavior by learning from data
  - (as opposed to explicitly programming task-specific rules)

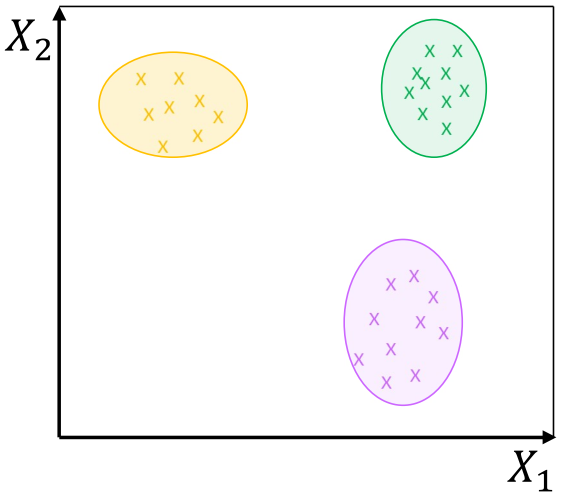
A Process Perspective Toward Machine Learning

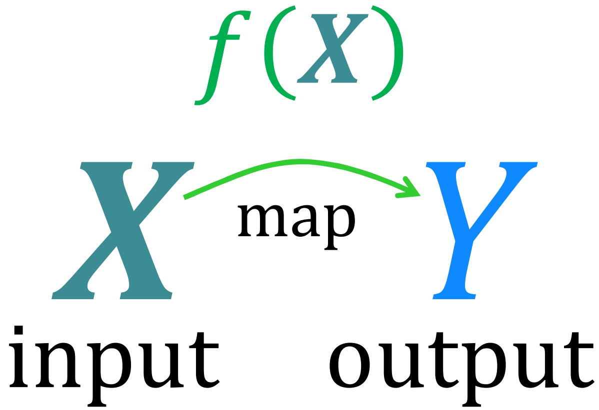


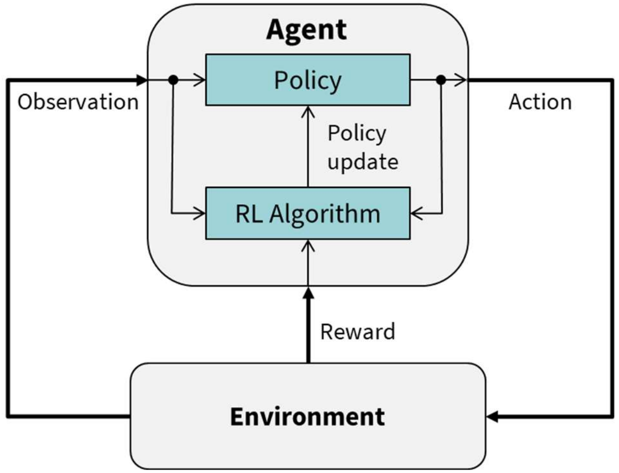
Types of Machine Learning

Learning from static data (un/supervised) or interactions

Unsupervised Learning	Supervised Learning	Reinforcement Learning







Starting Point for Machine Learning

Un/supervised learning algorithms learn from static tabular datasets

- **Tabular data: rows & columns represent individual subjects & features characterizing these subjects**
  - Denote by $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im}) \in \mathbb{R}^m$ and individual subject
  - And by $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^{n}$ the matrix of feature value (aka the data table)
- **Supervised learning setting further requires a target variable** $y \in \mathbb{R}$

i	X_1	X_2	X_3	$\ldots$	X_m
1	x_{11}	x_{12}	x_{13}	$\ldots.$	x_{1m}
2	x_{21}	x_{22}	x_{23}	$\ldots.$	x_{2m}
3	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
4	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
5	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
n	x_{n1}	x_{n2}	x_{n3}	$\ldots.$	x_{nm}

Y
y_1
y_2
$\ldots$
$\ldots$
$\ldots$
$\ldots$
y_m

Machine Learning Lingo

Many terms carry slightly different meaning

Features, attributes, characteristics, covariates, predictors, (independent) variables

Target, outcome, label, response (variable), dependent (variable)

Observations, cases, examples, data items, subjects

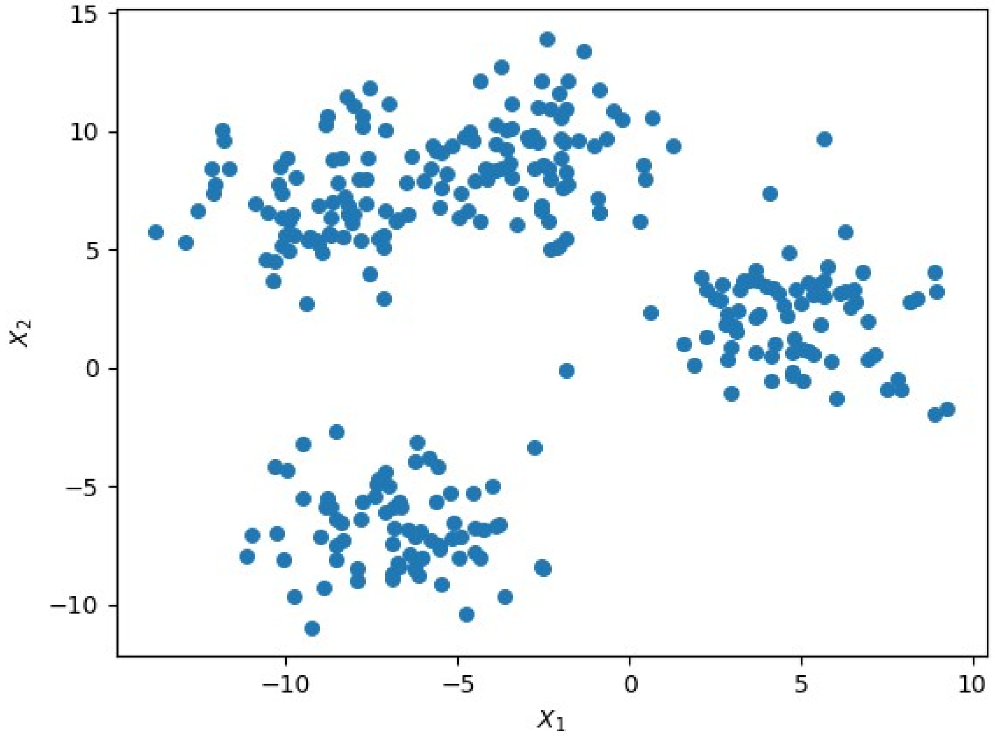
i	X_1	X_2	X_3	$\dots$	X_m
1	x_{11}	x_{12}	x_{13}	$\dots.$	x_{1m}
2	x_{21}	x_{22}	x_{23}	$\dots.$	x_{2m}
3	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
4	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
5	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
n	x_{n1}	x_{n2}	x_{n3}	$\dots.$	x_{nm}

Y
y_1
y_2
$\dots$
$\dots$
$\dots$
$\dots$
y_{m}

Visual Intuition of Machine Learning

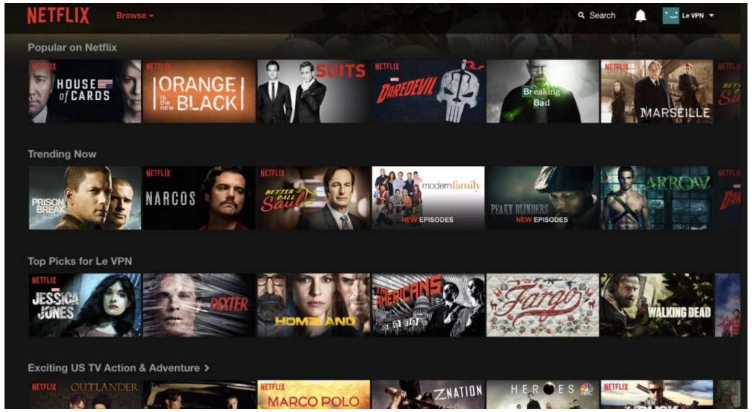
ML algorithms process data in high-dimensional feature spaces

i	X_1	X_2	$\dots$	X_m
1	x_{11}	x_{12}	$\dots.$	x_{1m}
2	x_{21}	x_{22}	$\dots.$	x_{2m}
3	$\dots$	$\dots$	$\dots$	$\dots$
4	$\dots$	$\dots$	$\dots$	$\dots$
5	$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
n	x_{n1}	x_{n2}	$\dots.$	x_{nm}



Real Life Example: Streaming Services

Data structure for unsupervised learning



- **Various attributes characterize clients**
  - Demographic information (e.g., Age, marital status, …)
  - Socio-demographic data (e.g., address, avg. income in neighborhood, …)
  - Behavioral/interaction data (e.g., avg. time on platform, movies watched, ...)
  - Attitudinal data (e.g., likes, reviews, …)
- **Observed & approximated attributes**
- **Numerical & categorical attributes**
- **Attribute values represent clients for the sake of the ML exercise**
  - Each client represented as an array of values
  - E.g., client 4: (33, “Premium”, 9, …, “Drama”)

i	AGE	SUBSCRIPTION	AVG HOURS/ WEEK	…	GENRE MODE
1	28	Premium	12	….	Romantic
2	18	Premium	6	….	Action
3	41	Standard (ads)	4	…	Family
4	33	Premium	9	…	Drama
5	37	Standard (ads)	3	…	Action
…	…		…	…	…
n	53	Standard (no ads)	6	…	SciFi

Common Forms of Unsupervised Learning

Discover patterns in raw data

i	AGE	SUBSCRIPTION	AVG HOURS/ WEEK	…	GENRE MODE
1	28	Premium	12	….	Romantic
…	…		…	…	…
n	53	Standard (no ads)	6	…	SciFi

- **Different approaches sharing the underlying data structure**
- **Dimensionality reduction**
  - Reduce number of attributes w/o losing (much) information → extract latent features
  - Useful to facilitate visual inspection of the data
- **Association rule mining**
  - Detect co-occurrences in transactions (e.g., movie often watched together)
  - Useful to plan shop layout or build recommendation systems
- **Clustering**
  - Detect subgroups (i.e., clusters) with similar attribute values
  - Useful to devise tailormade treatments for homogeneous subgroups
- **Outlier detection**

Discussion Break:

Suggest an application of regression in research or practice

Regression Example: Leasing Business
Supporting decision making using regression

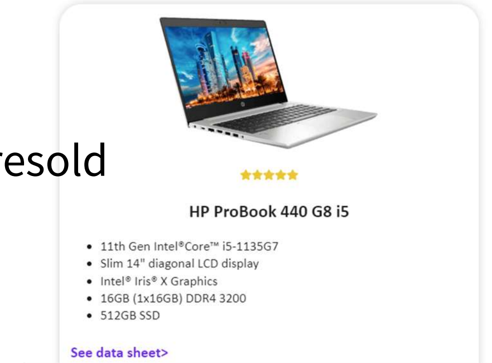
- **Business model**
  - Clients use equipment for agreed period, paying a monthly fee
  - Lessor is responsible for re-marketing the used item afterwards
- **Tabular data**
  - Notebook leasing example
  - Facilitates all forms of unsupervised learning discussed above

i	PRODUCT	LIST PRICE [$]	AGE [month]	CLIENT INDUSTRY	…
1	Dell XPS 15'	2,500	36	Mining	…
2	Dell XPS 15'	2,500	24	Health	
3	Dell XPS 17'	3,000	36	Manufacturing	
4	HP Envy 17'	1,300	24	Office	
5	HP EliteBook 850	1,900	36	Manufacturing	
…	…				
n	Lenovo Yoga 11'	799	12	Office	…

Regression Example: Leasing Business

Resale price forecasting use case

- **Lessor's pricing problem: Decide on the leasing rate**
  - Leasing rate must cover all relevant costs including depreciation
  - Item's residual value is unknown when signing the contract
- **We can use regression to *model* residual value/resale price**
  - Attribute values are observed before signing the contract
  - Resale prices are unobservable until after the contract expires and the item is resold



Regression Example: Leasing Business

Resale price forecasting use case

- **Lessor's pricing problem: Decide on the leasing rate**
  - Leasing rate must cover all relevant costs including depreciation
  - Item's residual value is unknown when signing the contract
- **We can use regression to *model* residual value/resale price**
  - Attribute values are observed before signing the contract
  - Resale prices are unobservable until after the contract expires and the item is resold

i	PRODUCT	LIST PRICE [$]	AGE [month]	CLIENT INDUSTRY	…	**RESALE PRICE [$]**
1	Dell XPS 15'	2,500	36	Mining		347
2	Dell XPS 15'	2,500	24	Health		416
3	Dell XPS 17'	3,000	36	Manufacturing	…	538
4	HP Envy 17'	1,300	24	Office		121
5	HP EliteBook 850	1,900	36	Manufacturing		172
…	…					…
n	Lenovo Yoga 11'	799	12	Office	…	88

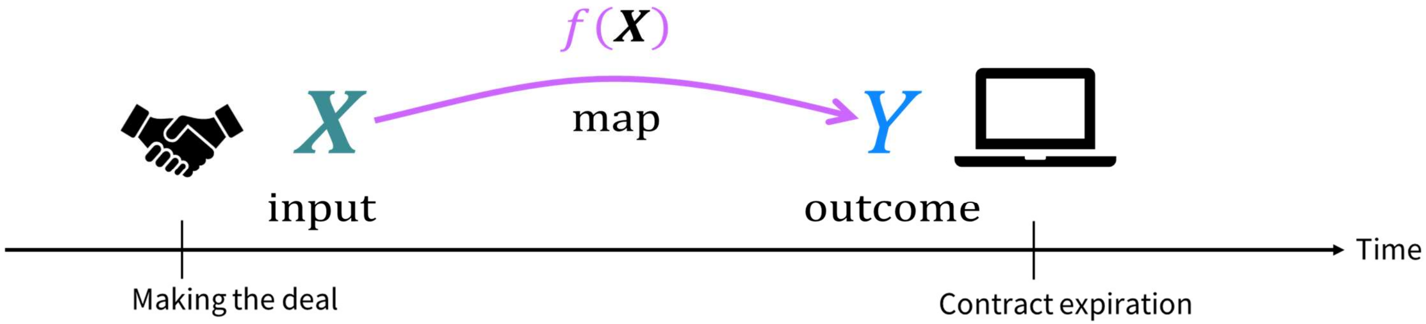
Regression Example: Leasing Business

Resale price forecasting use case

- **Lessor's pricing problem: Decide on the leasing rate**
  - Leasing rate must cover all relevant costs including depreciation
  - Item's residual value is unknown when signing the contract

- **We can use regression to *model* residual value/resale price**
  - Attribute values are observed before signing the contract
  - Resale prices are unobservable until after the contract expires and the item is resold

- **Regression setting with a dependent and independent variables**



The Two Faces of Linear Regression

Linear regression supports both, *explanatory* and *predictive* modeling

- **Regression function explains variation in RESALE PRICES by AGE**
- **Due to this ability, linear regression is an *explanatory model***
  - Clarifies relationship between features and target
  - Can work out the strength of a feature's effect
  - Can calculate elasticities, i.e., how a 1% change in AGE will change RESALE PRICES
- **Linear regression also facilitates prediction**
  - Given an AGE value, we can predict the corresponding RESALE PRICE using the estimated coefficients
  - Just evaluate regression equation
  - $\text{RESALE PRICE FORECAST} = bias + w_1\text{AGE}$

Linear Regression in a Nutshell

- **Model specification**
  - ☐ Continuous target variable y
  - ☐ Expectation of y given x is linear
  - ☐ Random variation $\epsilon \sim \mathcal{N}(0, \sigma^2)$

$$y = \mathbb{E}(y|\boldsymbol{x}) + \epsilon$$

$$y = b + \boldsymbol{w}^\top \boldsymbol{x} + \epsilon$$

- **Model estimation**
  - ☐ Determine free parameter $\boldsymbol{w}$
  - ☐ Set $\widehat{\boldsymbol{w}}$ to maximize model fit
  - ☐ Minimize least-squares loss

$$\widehat{\boldsymbol{w}} \leftarrow \min_{\boldsymbol{w}, b} \left(\sum_i (y_i - \hat{y}_i)^2 \right),$$

where $\hat{y}_i = b + \boldsymbol{w}^\top \boldsymbol{x}_i$

- **The final model**
  - ☐ Is given by the estimated coefficients $\widehat{\boldsymbol{w}}$
  - ☐ Facilitates explanation
  - ☐ Facilitates prediction

$$\hat{y} = \hat{b} + \widehat{\boldsymbol{w}}^\top x$$

Training data incl. Y

i	Y	X_1	X_2	…	X_m
1	…	…	…	…	…
…	…	…	…	…	…
n	…	…	…	…	…

Learning Algorithm

Model $\widehat{\boldsymbol{w}}$

Insight into $\mathbb{E}(y|\boldsymbol{x})$ from linear equation.

e.g., elasticity:
1% increase in x increases y by ? %

New data w/o Y

i	X_1	X_2	…	X_m
$n+1$	…	…	…	…
…	…	…	…	…
N	…	…	…	…

Forecasts of Y

i	$\hat{Y}$
$n+1$	…
$n+2$	…
…	…
N	…

Different Perspectives and Cultures in Data-Oriented Disciplines

Athey and Imbens (2019)

Machine learning

- **Focus on prediction**
  - Map from covariates to outcomes
  - Cope with high dimensionality
- **Practices to test predictive quality**
  - Cross-validation
  - Model selection (e.g., regularization)
- **Few assumptions**
  - Independent observations
  - Stability of the joint distribution of (Y, X)
- **More data-driven**

Econometrics

- **Understand structural properties**
  - Estimate parameters of interest
  - Cause-effect relationships are one example
- **Apply linear models to all data**
  - Explanatory modeling
  - Data is typically low-dimensional
- **Key interests**
  - Unbiasedness of parameters
  - Efficiency and convergence rates
- **Many assumption**
- **More theory-driven**

General Implications for Business/ECON Research

- **Confirmatory research needs more than vanilla ML**
  - No statistical tests to verify hypotheses
  - No way to proof ML-induced patterns (e.g., functional dependencies)
  - BUT:
    - Demonstrating the superiority of an advanced ML model over a classical (linear) model hints at a *theory gap*
    - In combination with XAI (see course block III) ML models can verify theories or inform theory development
- **ML can unlock new data modalities (text, image, …) to shed light on longstanding research questions and topics**
  - Financial market efficiency, corporate disclosure quality and tone, brand perception and sentiment, …
  - Behavioral biases, political communication, labor market skills, innovation, climate risk, …
- **ML is a tool to approach advanced optimization problems**
  - Routing & last-mile delivery optimization, energy load balancing & trading, marketing budget allocation, dynamic pricing & revenue management, portfolio optimization under complex constraints, …
  - Advanced ML paradigms (e.g., reinforcement learning, causal ML, policy learning)

Summary

Learning goals

- Difference between ML and AI
- Overview of ML approaches and characteristics

Findings

- ML is a component of *what people refer to* as AI
- Detecting patterns in heterogenous dataset
  - Our focus for now is static tabular data
- Forms of unsupervised ML
- Supervised ML approximates functional dependencies
- Linear regression exemplifies supervised ML
- Common use cases of ML in business research

What next

- Python exercise on data handling
- Supervised ML principles and methods

A

Appendix

Materials are not discussed in course and are provided for self-study

The linear regression model

Parametric model, loss function and residuals, model estimation

Regression Example: Leasing Business
Resale price forecasting use case

- **Lessor's pricing problem: Decide on the leasing rate**
  - Leasing rate must cover all relevant costs including depreciation
  - Item's residual value is unknown when signing the contract

- **We can use regression to *model* residual value/resale price**
  - Attribute values are observed before signing the contract
  - Resale prices are unobservable until after the contract expires and the item is resold

i	PRODUCT	LIST PRICE [$]	AGE [month]	CLIENT INDUSTRY	…	**RESALE PRICE [$]**
1	Dell XPS 15'	2,500	36	Mining		347
2	Dell XPS 15'	2,500	24	Health		416
3	Dell XPS 17'	3,000	36	Manufacturing		538
4	HP Envy 17'	1,300	24	Office	…	121
5	HP EliteBook 850	1,900	36	Manufacturing		172
…	…					…
n	Lenovo Yoga 11'	799	12	Office	…	88

Linear Regression Model

Postulates a linear, additive feature-to-target relationship

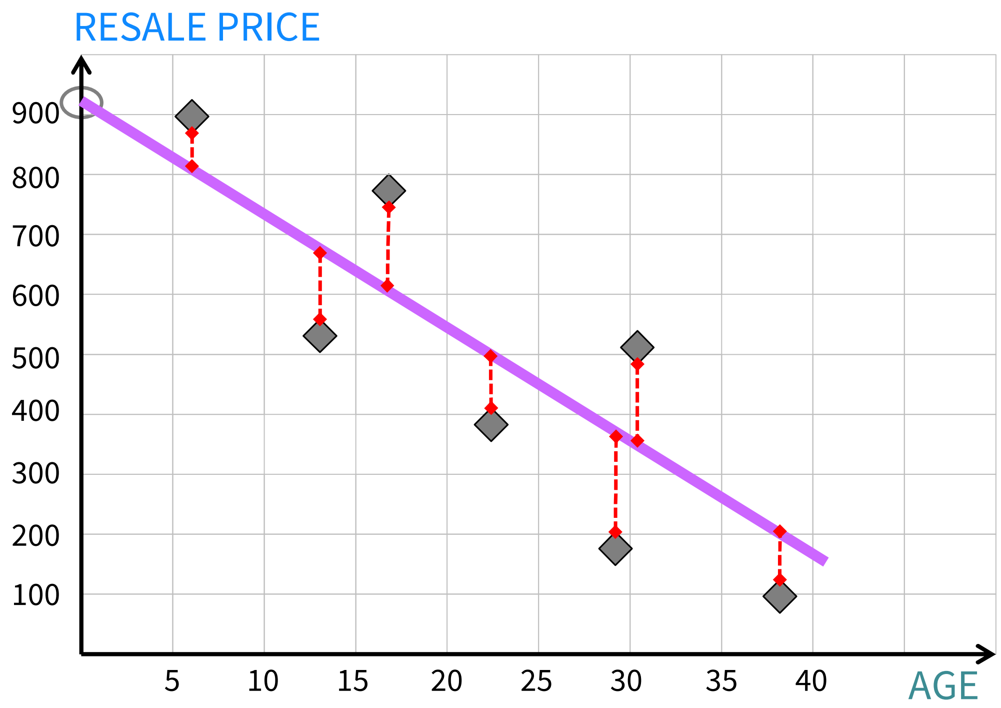
- **Famous regression equation in a resale price forecasting context**

$$\text{RESALE PRICE} = bias + w_1\text{LIST PRICE} + w_2\text{AGE} + \cdots + w_m\text{INDUSTRY} + residual$$

- **Simplification for plotting:**

$$\text{RESALE PRICE} = bias + w_1\text{AGE} + \epsilon$$

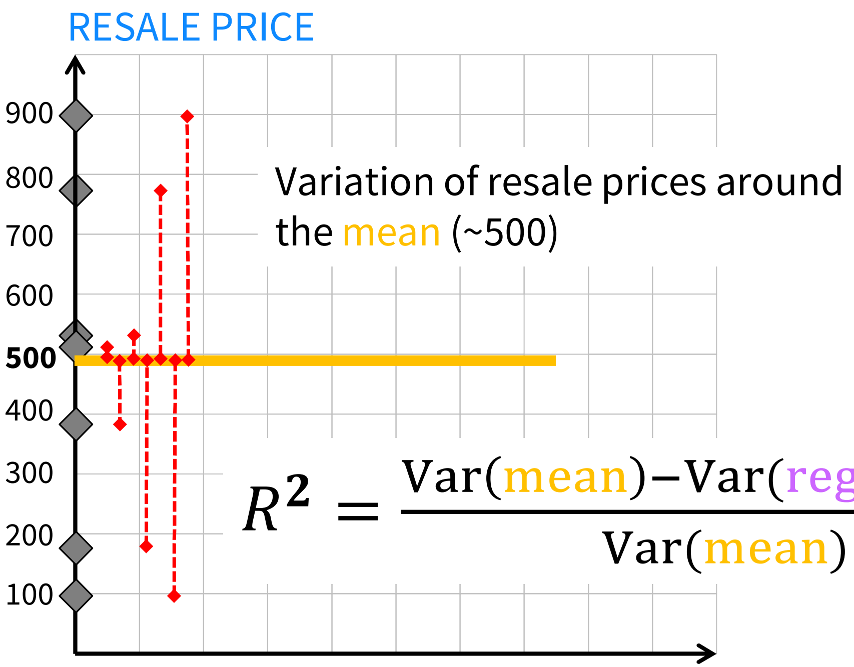
…	AGE [MONTH]	…	RESALE PRICE [$]
…	6	…	900
…	13	…	515
…	17	…	890
…	23	…	395
…	29	…	180
…	31	…	501
…	38	…	100
…	…	…	…



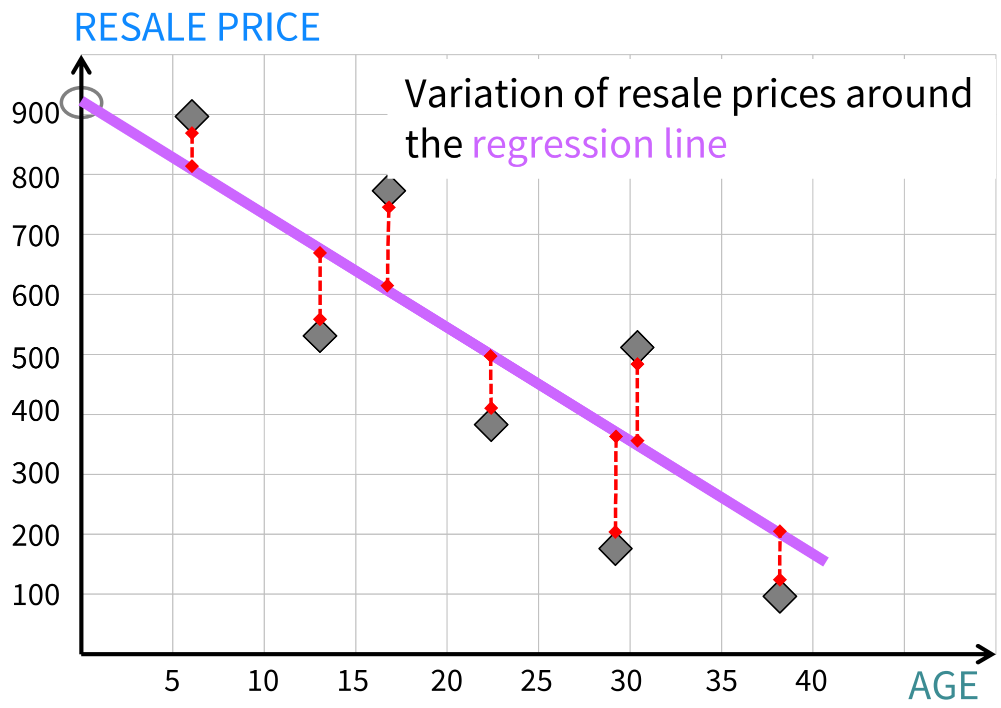
Popular Regression Statistics

The maybe most famous statistic is the coefficient of determination, R^2

- **Regression model explains variation in resale prices (target) by differences in the Age (feature values) of resold items**
- **R^2 statistic captures how much of the variation in resale prices the regression explains**



$$R^2 = \frac{\mathrm{Var}(\text{mean}) - \mathrm{Var}(\text{regression})}{\mathrm{Var}(\text{mean})}$$



Regression Model Estimation

Determine the free parameters of the regression function

- **Linear regression belongs to the family of parametric models**
  - We assume we know the true dependency of the target variable and the features
  - We specify a function expressing the assumed relationship (e.g., linear and additive)
  - We incorporate free parameters that govern the shape of the function
- **Formally**

$$\text{RESALE PRICE} = bias + w_1 \text{LIST PRICE} + w_2 \text{AGE} + \cdots + w_m \text{INDUSTRY} + residual$$

$$Y = b + w_1 X_1 + w_2 X_2 + \cdots + w_m X_m + \epsilon$$

- **Model estimation (aka training, fitting, development)**
  - Find *suitable* values for the free parameters (here denoted by w)
  - Introduce a measure that captures what is meant by *suitable*
  - Set parameters such this measure signals an optimal fit of the model to the data

Regression Model Estimation

Loss functions measure the quality of a model (using the true outcomes)

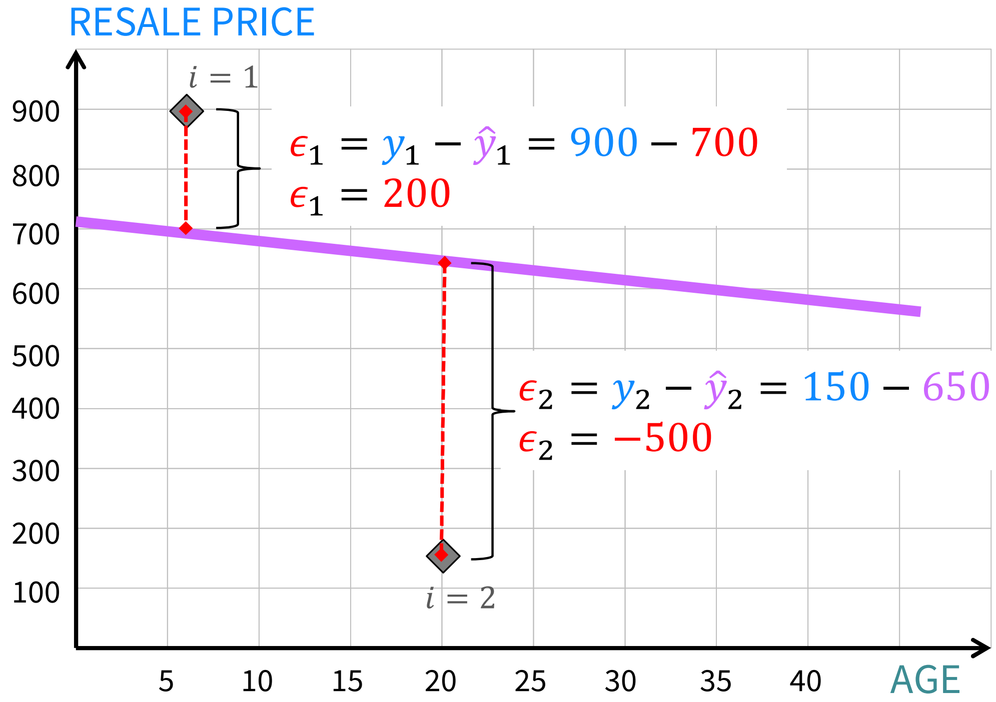
- **A loss function J measures how much model outputs $\hat{Y}$ agree with the true, actually observed, value of the target Y**
- **Squared-error loss**

$$J^{LS} = \sum_{i=1}^{n} (\epsilon_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- **Where model outputs depends on the free parameters w**

$$\hat{y}_i = b + w_1 x_{i1} + w_2 x_{i2} + \cdots + w_m x_{im}$$

$$\hat{y}_i = b + \sum_{j=1}^{m} w_j x_{ij}$$



So here, the value of the loss function J is:

$$\sum_{i=1}^{n} (\epsilon_i)^2 = (\epsilon_1)^2 + (\epsilon_2)^2 = 200^2 + (-500)^2 = 290{,}000$$

Regression Model Estimation

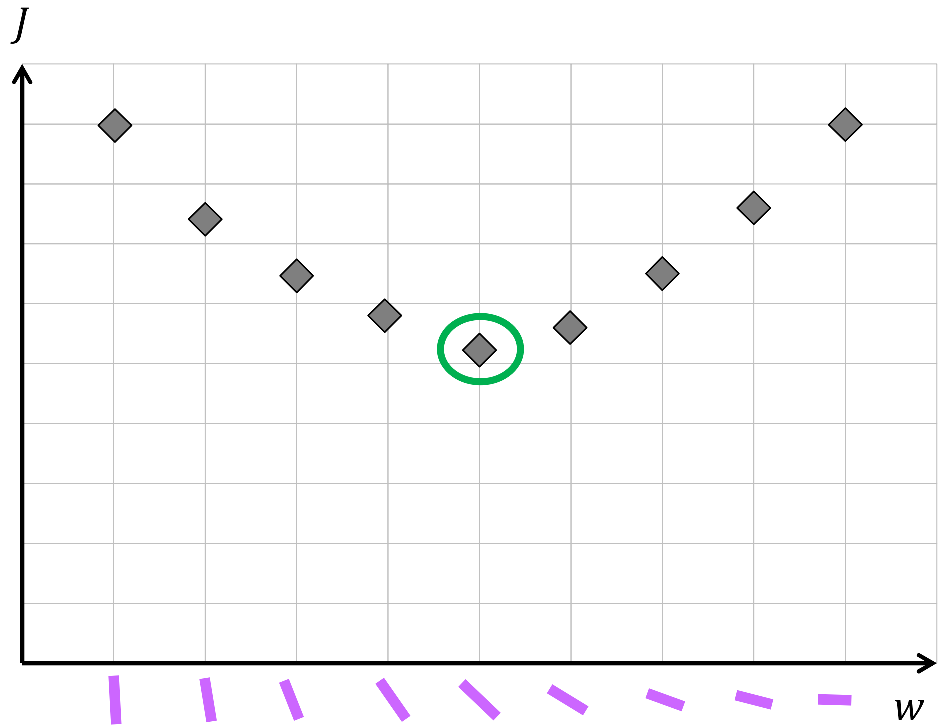
Maximizing model fit through minimizing a loss function

- **Given the model output depends on the free parameters w and the bias b through**

$$\hat{y}_i = b + \sum_{j=1}^{m} w_j x_{ij}$$

1

- **We can adjust the model output by adjusting w (or b)**
  - Lets focus, for simplicity, on w
  - Changing w will change the slope of the regression line
- **For each slope we can calculate the loss (e.g., sum of squared residuals)**
- **Eventually, we know which slope (i.e., value of w) gave the lowest loss; our *least-squares solution***



Regression Model Estimation

Minimizing empirical loss, that is the loss over our data

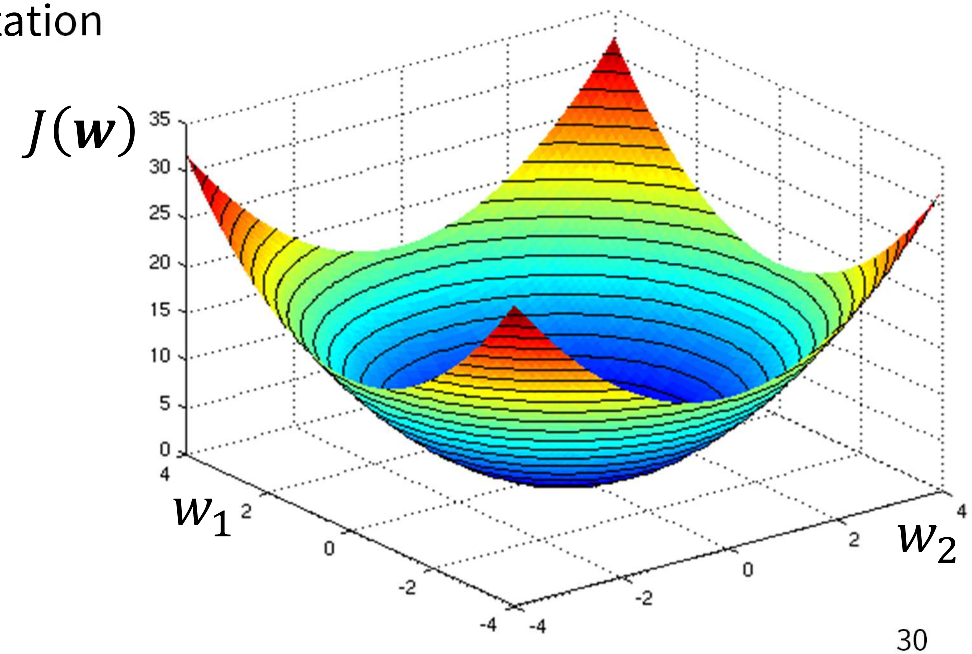
- **Squared-error loss**

$$J(\boldsymbol{w}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} w_j x_{ij} \right)^2 = \sum_{i=1}^{n} (y_i - \boldsymbol{w}\boldsymbol{X}_i)^2 = (\boldsymbol{y} - \boldsymbol{w}\boldsymbol{X})^\top (\boldsymbol{y} - \boldsymbol{w}\boldsymbol{X})$$

  - Note that we have dropped the bias to simplify the notation
  - Mathematically, we can add a constant to the data $\boldsymbol{X}$

- **To fit the regression model, we set w such that the error is minimal**

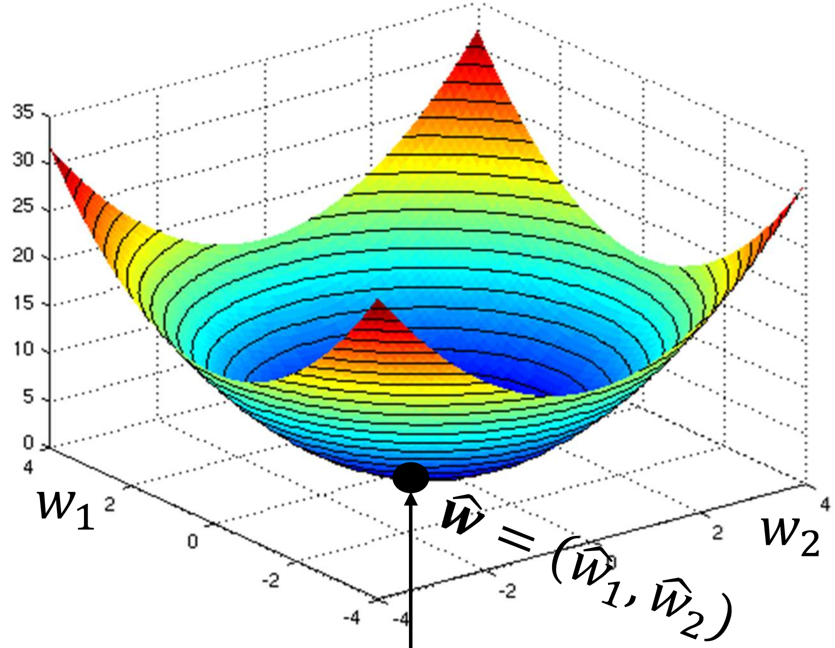
  - Mathematically, we seek the minimum of the loss function over the model parameters
  - $\hat{\boldsymbol{w}} \leftarrow \text{argmin}_{\boldsymbol{w}} \; J(\boldsymbol{w})$



Regression Model Estimation

Finding the optimal solution

- **Formalization of estimation (training) task**
  - $J(\boldsymbol{w}) = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} w_j x_{ij}\right)^2$
  - $\hat{\boldsymbol{w}} \leftarrow \operatorname{argmin}_{\boldsymbol{w}}\ J(\boldsymbol{w})$
- **Solution**
  - Apply calculus principle
  - Calculate partial derivatives of $J(\boldsymbol{w})$ with respect to each w_j and set to zero



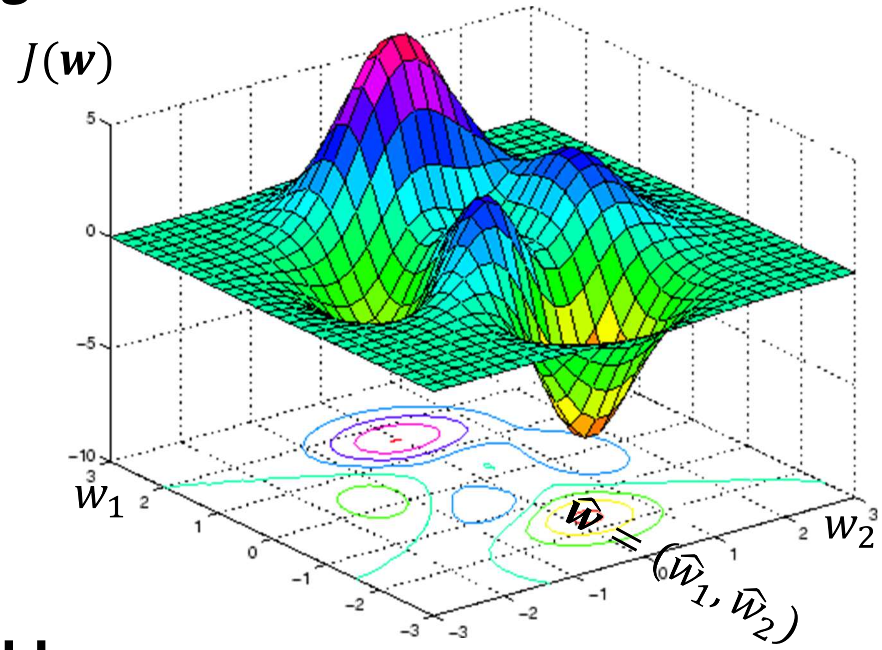
$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} = 0 \quad \left\{ \begin{array}{l} \dfrac{\partial J(\boldsymbol{w})}{\partial w_1} = 0 \\[2ex] \dfrac{\partial J(\boldsymbol{w})}{\partial w_2} = 0 \end{array} \right. \quad \Rightarrow \hat{\boldsymbol{w}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

For the special case of linear regression with squared error loss, we obtain an analytical solution.

Regression Model Estimation
Generalization

- **Minimization problem remains (largely) unchanged**
  - $J(\boldsymbol{w}) = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} w_j x_{ij}\right)^2$
  - $\widehat{\boldsymbol{w}} \leftarrow \operatorname{argmin}_{\boldsymbol{w}} J(\boldsymbol{w})$

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} = 0 \quad \begin{cases} \dfrac{\partial J(\boldsymbol{w})}{\partial w_1} = 0 \\[2ex] \dfrac{\partial J(\boldsymbol{w})}{\partial w_2} = 0 \end{cases} \Rightarrow \widehat{\boldsymbol{w}} = ?$$



- **Computing analytical solution typically impossible**
- **Use iterative numerical algorithms to find $\widehat{\boldsymbol{w}}$**

The Two Faces of Linear Regression – Explanatory vs. Predictive Models

Linear regression supports both, explanatory and predictive modeling

- **Regression function explains variation in resale prices by age**
- **Due to this ability, linear regression is an *explanatory model***
  - ☐ Clarifies relationship between features and target
  - ☐ Can work out the strength of a feature's effect
  - ☐ Can calculate elasticities, i.e., how a 1% change in age will change resale prices
- **Linear regression also facilitates prediction**
  - ☐ Given an age value, we can predict the corresponding resale price using the estimated coefficients
  - ☐ Just evaluate regression equation
  - ☐ $\text{Resale Price Forecast} = bias + w_1\text{Age}$

Linear Regression Summary

- **Model specification**
  - Continuous target variable Y
  - Expectation of Y given X is linear
  - Random variation $\epsilon \sim \mathcal{N}(0, \sigma^2)$

$$Y = \mathrm{E}(Y|\boldsymbol{X}) + \epsilon$$

$$Y = b + \boldsymbol{w}\boldsymbol{X} + \epsilon$$

- **Model estimation**
  - Determine free parameter $\boldsymbol{w}$
  - Set $\hat{\boldsymbol{w}}$ to maximize model fit
  - Minimize least-squares loss

$$\hat{\boldsymbol{w}} \leftarrow \min_{\boldsymbol{w},b}\left(\sum_i \left(Y_i - \hat{Y}_i\right)^2\right),$$

where $\hat{Y}_i = b + \boldsymbol{w}\boldsymbol{X}_i$

- **The final model**
  - Is given by the estimated coefficients $\hat{\boldsymbol{w}}$
  - Facilitates explanation
  - Facilitates prediction

$$\hat{Y} = \hat{b} + \hat{\boldsymbol{w}}\boldsymbol{X}$$

Training data incl. Y

i	Y	X_1	X_2	...	X_m
1	…	…	...	...	...
…	…	...	...	...	...
n	…	...	...	...	...

Learning Algorithm

Model
$\hat{\boldsymbol{w}}$

Insight into $\mathrm{E}(Y|X)$ from linear equation.

e.g., elasticity:
1% increase in X_1 increases Y by ? %

New data w/o Y

i	X_1	X_2	...	X_m
$n+1$	…	...	...	...
…	...	...	...	...
N	...	...	...	...

Forecasts of Y

i	$\hat{Y}$
$n+1$	…
$n+2$	...
…	...
N	...

Contact

Stefan Lessmann

Chair of Information Systems
School of Business and Economics
Humboldt-University of Berlin, Germany

Tel.   +49.30.2093.5742
Fax.   +49.30.2093.5741

stefan.lessmann@hu-berlin.de
http://bit.ly/lessmann

www.hu-berlin.de