

Session 1-1

Course Overview

PSYC 560: Machine Learning Tools in Psychology

Winter 2023
Heungsun Hwang



Course Overview

- Time: Tuesday & Thursday 1:05 – 2:25 pm
- Location: 2001 McGill College, Room 461
- Prerequisite: PSYC 305 or equivalent



Course Overview

- Instructor: Heungsun Hwang
 - Office: 2001 McGill College, Room 712
 - Office Hour: Tuesday 2:30 – 3:30 pm
 - Email: [heungsun.hwang\[at\]mcgill.ca](mailto:heungsun.hwang[at]mcgill.ca)



Course Overview

- Teaching Assistant: Gyeongcheol Cho
 - Office Hour: Friday 10:00 – 11:00 am
(2001 McGill College #464)
 - Email: [gyeongcheol.cho\[at\]mail.mcgill.ca](mailto:gyeongcheol.cho[at]mail.mcgill.ca)





Focus of Course

- This course provides an overview of various methods and algorithms for understanding and predicting data, which have been developed in statistics and machine learning.
- It introduces a set of important techniques for supervised and unsupervised learning.



Course Material

- This course is composed of lectures and labs.
 - Lecture (T) – Heungsun Hwang
 - Lab (R) – Gyeongcheol Cho
- No mandatory textbook.
- Statistical software:   RStudio®

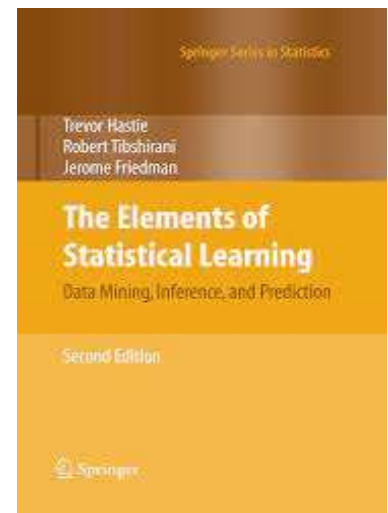
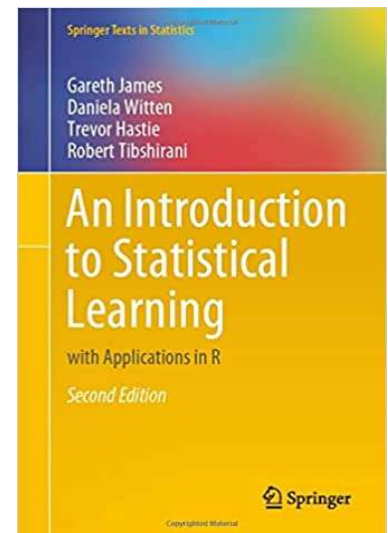
Course Material

- References:

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning*. 2nd Edition. Springer.

- <https://www.statlearning.com>

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Data Mining, Inference, and Prediction. 2nd Edition. Springer.





Evaluation Criteria

- Individual assignments (7) 70 %
 - 10% per assignment
- Group project presentation 25 %
- Class participation 5 %



Individual Assignments

1. Linear regression
2. Classification
3. Shrinkage and data reduction methods
4. Tree-based methods
5. Support vector machines
6. Clustering methods
7. Neural network models



Group Project Presentation

- Students are to organize themselves into groups of **three or four** by the third session.
- Each group first needs to find an actual dataset.
- The group then needs to
 - formulate a research problem based on the data
 - apply a machine learning method(s) covered in the class to resolve the problem efficiently.



Group Project Presentation

- This presentation will serve to describe research objectives, statistical analyses and results.
- The presentation will be evaluated based on
 - Clarity of research questions/objectives
 - Adequacy of statistical analyses
 - Effectiveness in answering questions
- The presentation should not exceed 15 minutes.
- Each group is expected to prepare and submit a PowerPoint presentation.



How to Find Data?

- There are a lot of online databases (e.g., Kaggle, UCI Machine Learning Repository, CMU Libraries, etc.). For example, check out the following blogs for a list of such databases.
 - <https://serokell.io/blog/best-machine-learning-datasets>
 - <https://towardsdatascience.com/top-sources-for-machine-learning-datasets-bb6d0dc3378b>
 - github : <https://github.com/awesomedata/awesome-public-datasets>
 - https://www.finereport.com/kr/%eb%8d%b0%ec%9d%b4%ed%84%b0-%eb%b6%84%ec%84%9d%ec%97%90-%ed%95%84%ec%9a%94%ed%95%9c-%ec%98%a4%ed%94%88-%ea%b3%b5%ea%b3%b5%eb%8d%b0%ec%9d%b4%ed%84%b0-%ec%86%8c%ec%8a%a4-%ed%8f%ac%ed%84%b8-20%ea%b0%80/?utm_campaign=article&utm_medium=facebook&utm_source=media&fbclid=IwAR2qLZpaS9-ju15hfS_dIUdTWs-SsoP4iE5RBg_HjWPhStwQ9bP2jSJE-6l



Class Participation

- Students are expected to actively participate in class, in particular, during group presentations.

For more details, refer to
the course syllabus

Session 1-2

An Overview of Machine Learning

PSYC 560

Heungsun Hwang



Machine Learning

- A vast set of statistical tools for **explaining** and/or **predicting** data.
- Supervised learning
 - Both inputs and output(s)
- Unsupervised learning
 - Only inputs
 - Explain or summarize the inherent structure of input data



Two Cultures of Statistical Modeling: To Explain or to Predict?

- **Explanation:** Statistical models/methods are used for understanding or describing associations between variables or testing hypotheses about parameters.
 - Which predictors are associated with the response?
 - What is the relationship between the response and each predictor?
 - How are many different variables associated with one another?



Example: Use of Linear Regression for Explanation

- The children's antisocial behaviour data: Part of the National Longitudinal Survey of Youth (NLSY) reported in Curran (1998).
 - Response (output/DV) = The antisocial behaviour of children measured at the first time point (0-12).
 - Predictors (inputs):
 - Gender (female = 0 and male = 1)
 - Cognitive stimulation for children at home (0-14)
 - Emotional support for children at home (0-13)



Example: Use of Linear Regression for Explanation

Model Summary

| Model | R | R ² | Adjusted R ² | RMSE |
|-------|-------|----------------|-------------------------|-------|
| 1 | 0.285 | 0.082 | 0.069 | 1.485 |

ANOVA

| Model | | Sum of Squares | df | Mean Square | F | p |
|-------|------------|----------------|-----|-------------|-------|--------|
| 1 | Regression | 42.481 | 3 | 14.160 | 6.418 | < .001 |
| | Residual | 478.758 | 217 | 2.206 | | |
| | Total | 521.240 | 220 | | | |



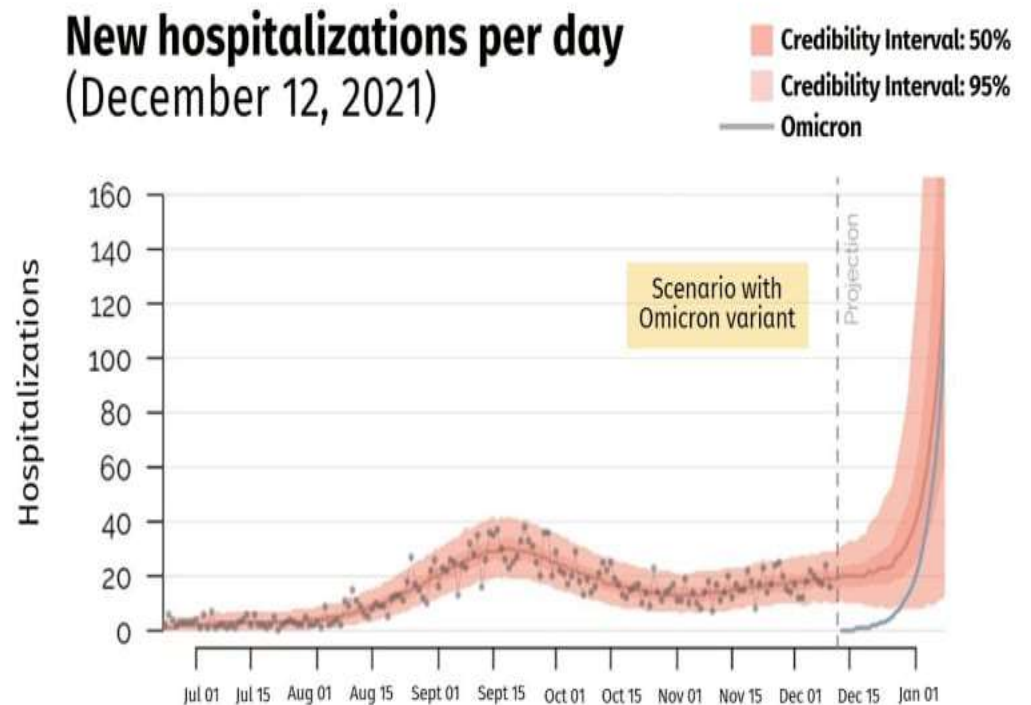
Example: Use of Linear Regression for Explanation

Coefficients

| Model | | Unstandardized | Standard Error | Standardized | t | p |
|-------|-------------|----------------|----------------|--------------|--------|--------|
| 1 | (Intercept) | 2.448 | 0.500 | | 4.896 | < .001 |
| | GENDER | 0.634 | 0.201 | 0.206 | 3.155 | 0.002 |
| | COGSTM | 0.013 | 0.043 | 0.020 | 0.295 | 0.769 |
| | EMOTSUP | -0.151 | 0.047 | -0.222 | -3.216 | 0.001 |

Two Cultures of Statistical Modeling: To Explain or to Predict?

- **Prediction:**
Statistical models/methods are used for predicting new or future observations.



Statistical Modeling: The Two Cultures

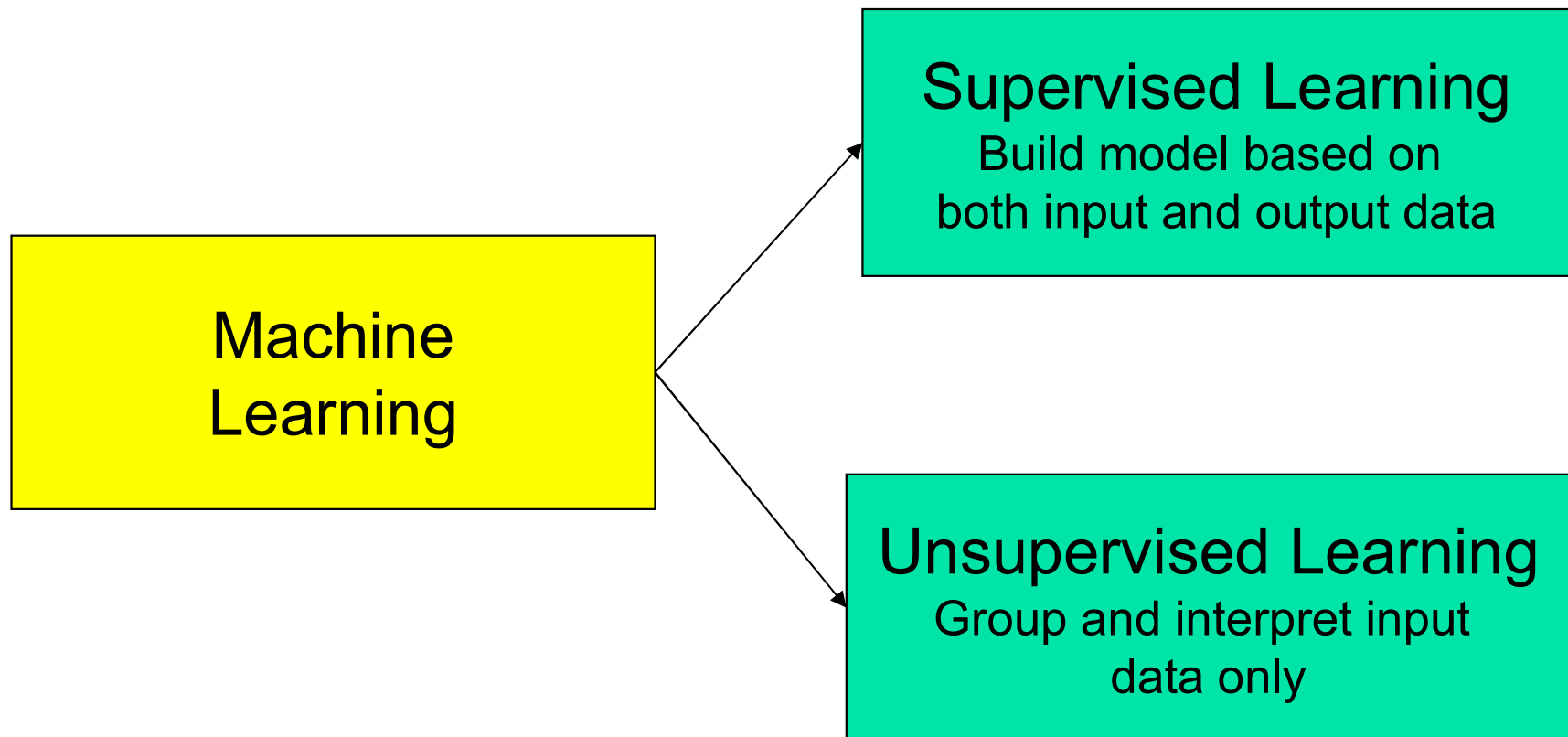
Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Also see Shmueli, G. (2010). To Explain or to Predict?, *Statistical Science*, 25, 289–310

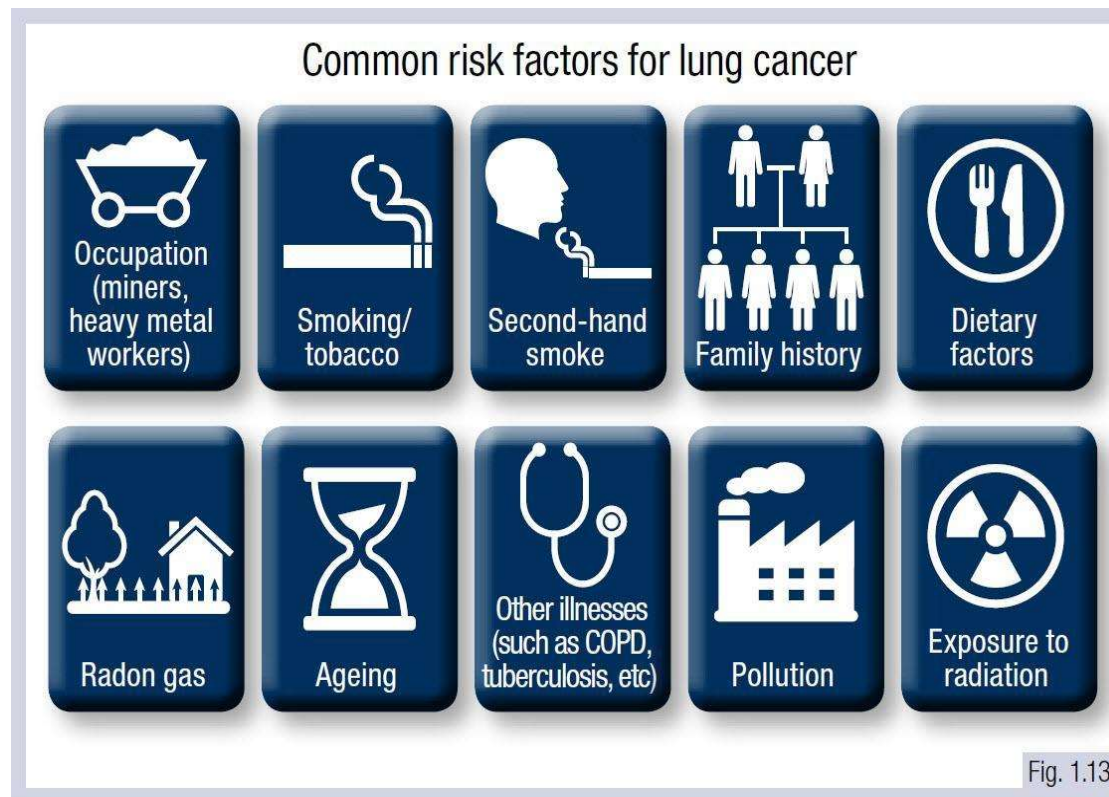


Machine Learning



Supervised Learning Problems

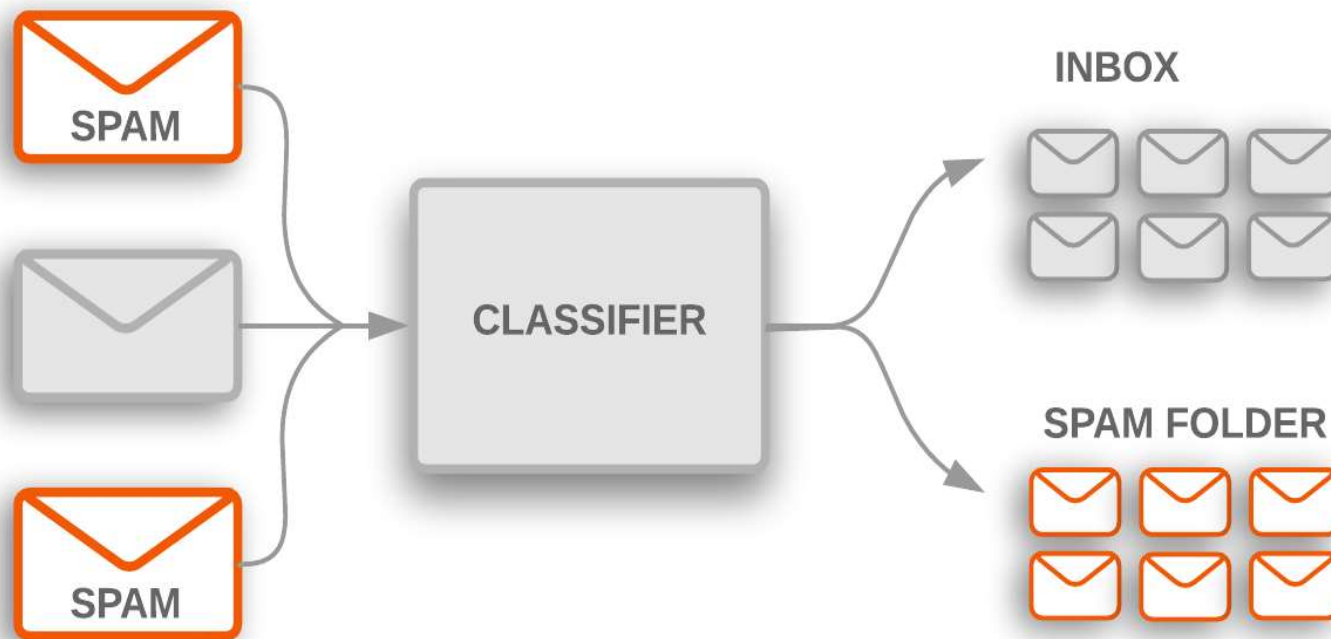
- Identify the risk factors for lung cancer.



COPD, Chronic obstructive pulmonary disease

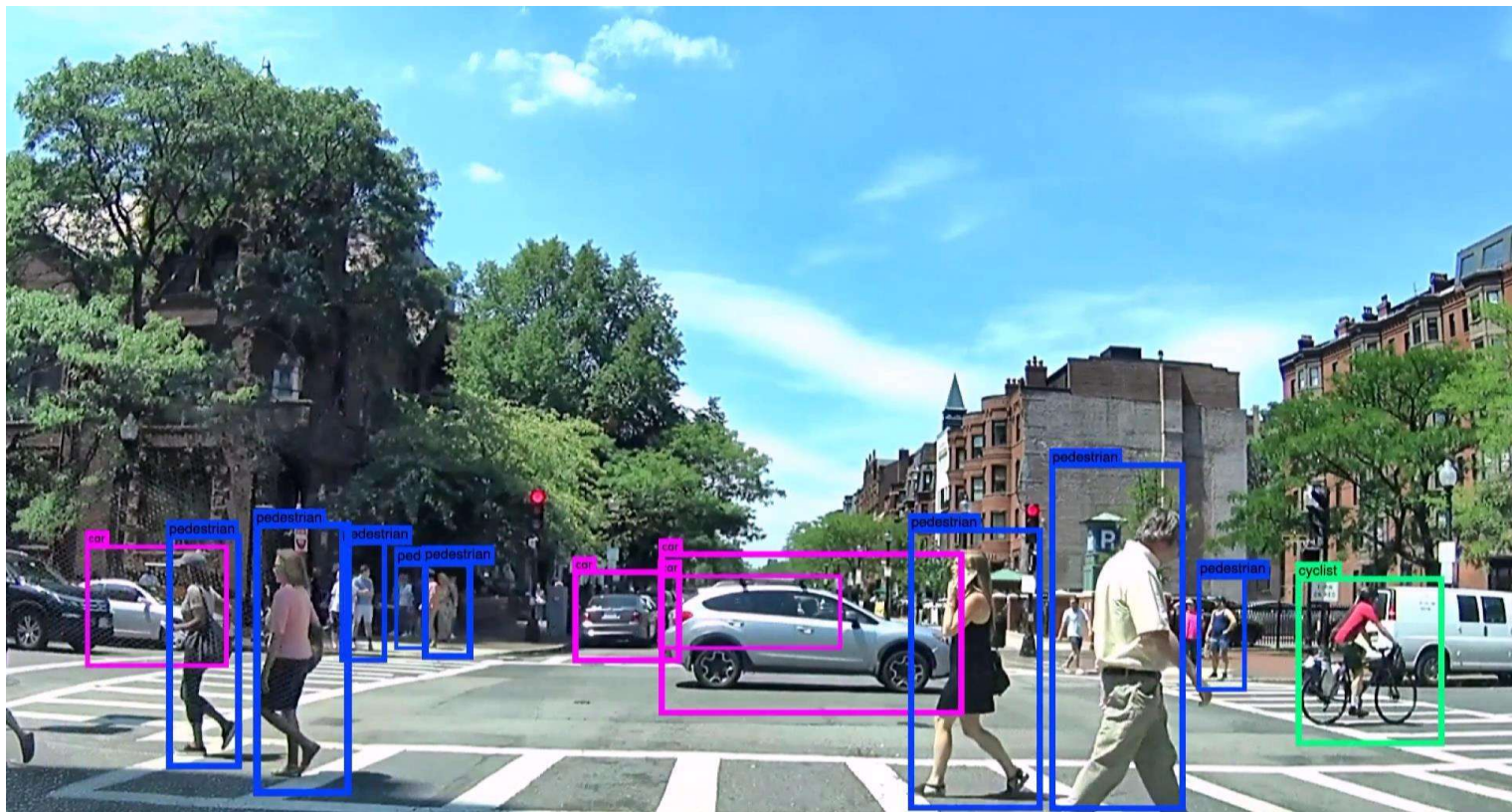
Supervised Learning Problems

- Distinguish spam emails from important messages.



Supervised Learning Problems

- Train a self-driving car to detect and classify objects based on the data gathered by the vehicle's different sensors.





Supervised Learning

| Inputs | Output |
|-----------------------|--------------------|
| Predictors | Response |
| Independent variables | Dependent variable |
| X | Y |

$$Y = f(X) + e$$

Supervised learning refers to a set of approaches for estimating f for explanation and/or prediction



Supervised Learning

$$Y = f(X) + e$$

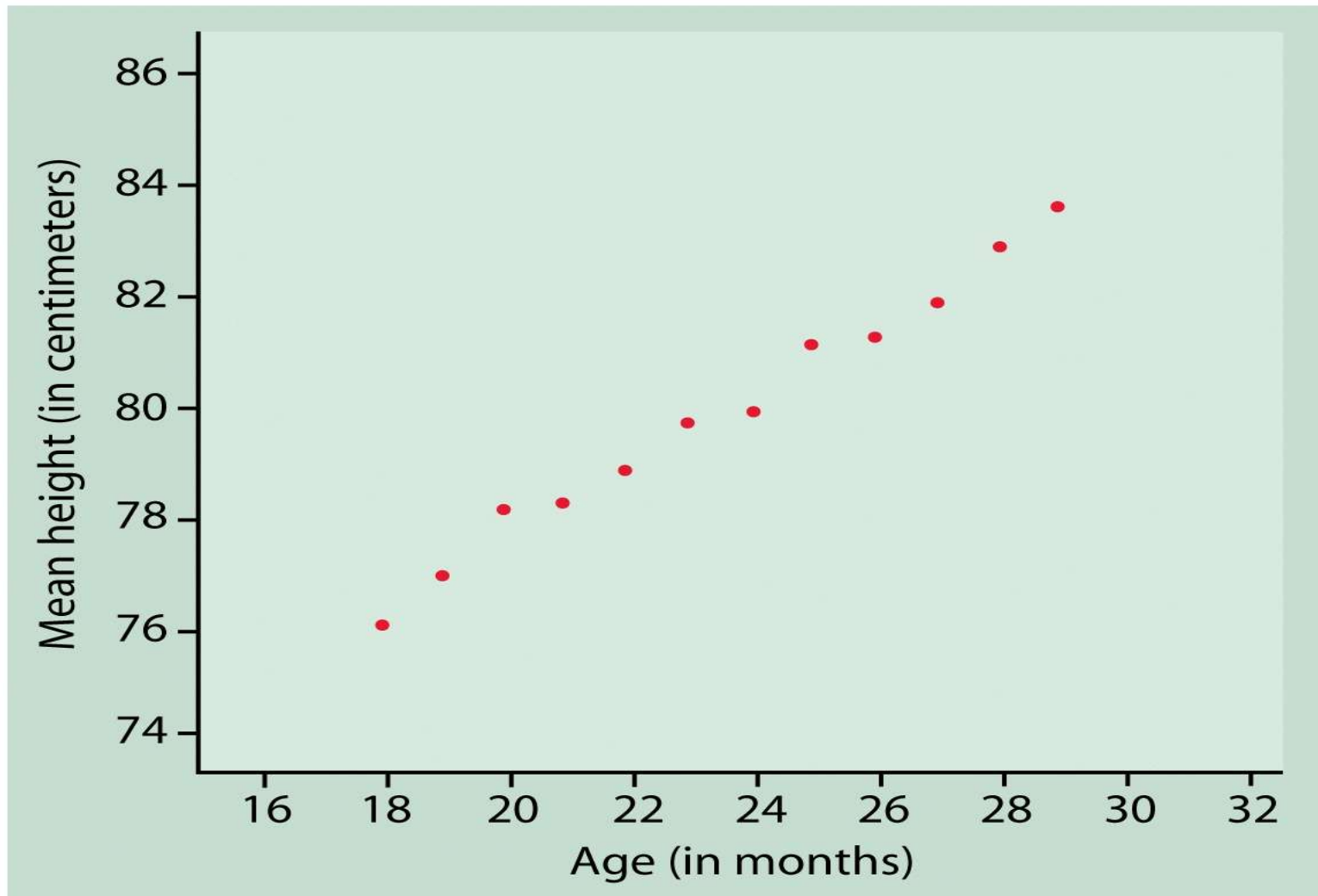
- f can be a linear or non-linear function.
- **Parametric** – we make an assumption about the form of f . For example,

$$f(X) = \beta_0 + \beta X$$

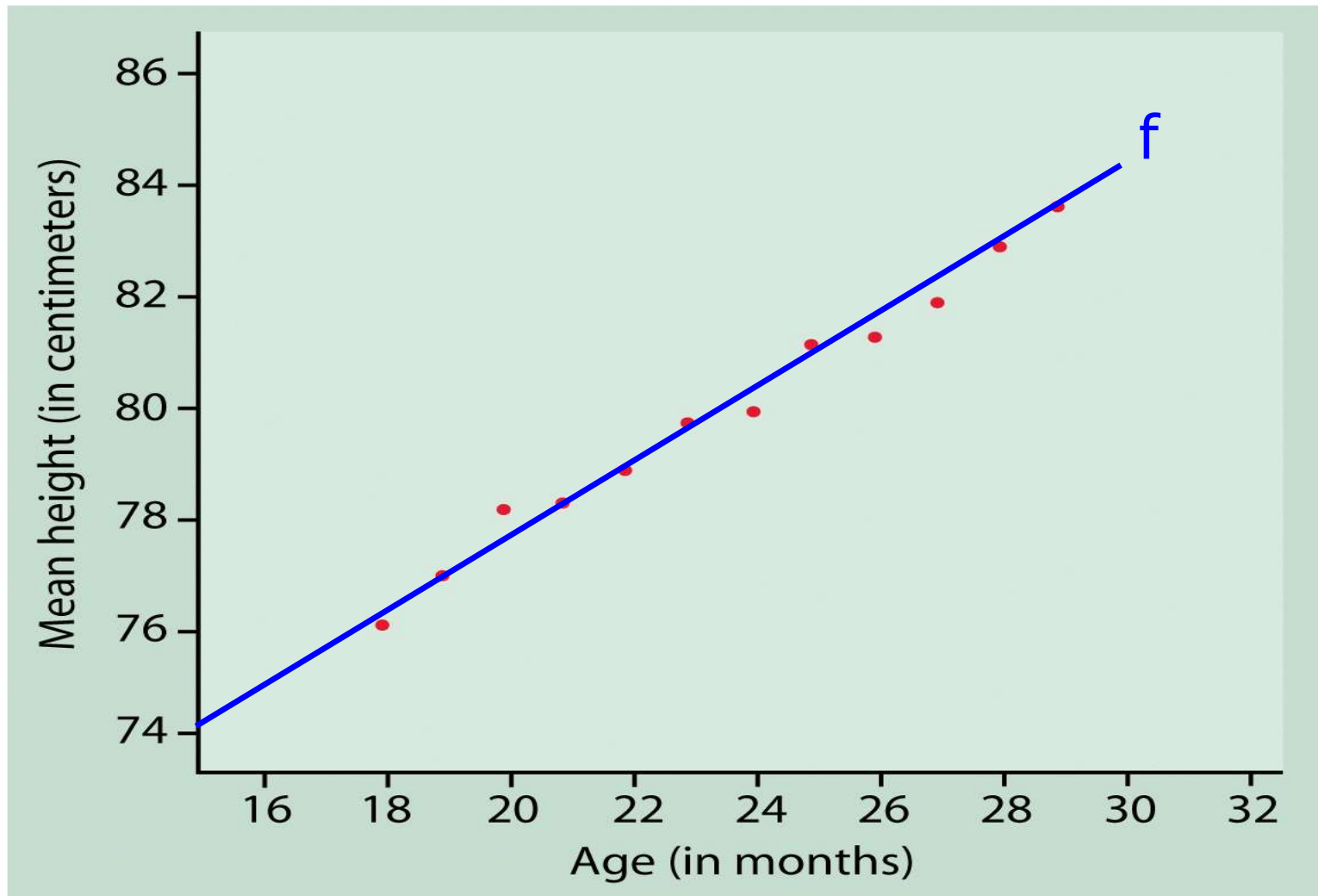
- **Non-parametric** – we do not make explicit assumptions about the form of f .

Supervised Learning

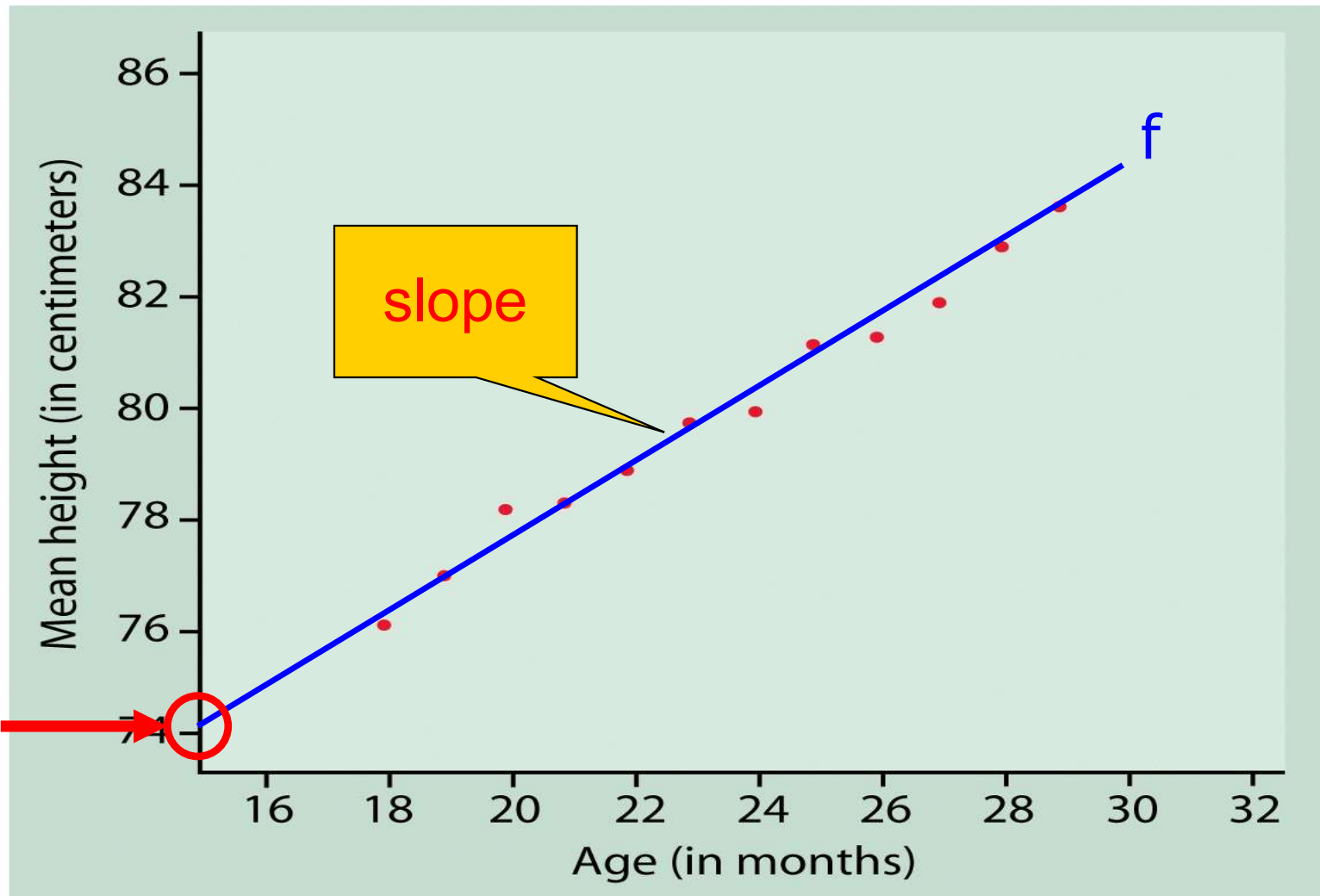
Kalama,
Egypt



Supervised Learning



Supervised Learning



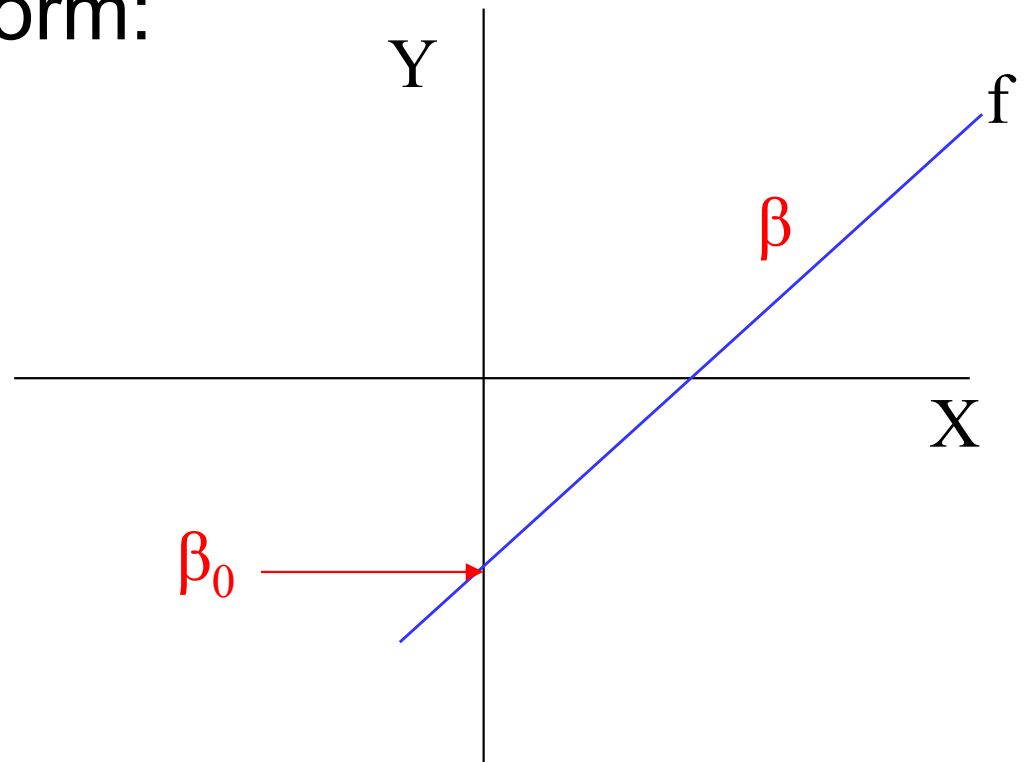
Supervised Learning

- A straight-line relating Y to X has the following form:

- $f(X) = \beta_0 + \beta X$

- β_0 = intercept
- β = slope

$$\begin{aligned} Y &= f(X) + e \\ &= \beta_0 + \beta X + e \end{aligned}$$



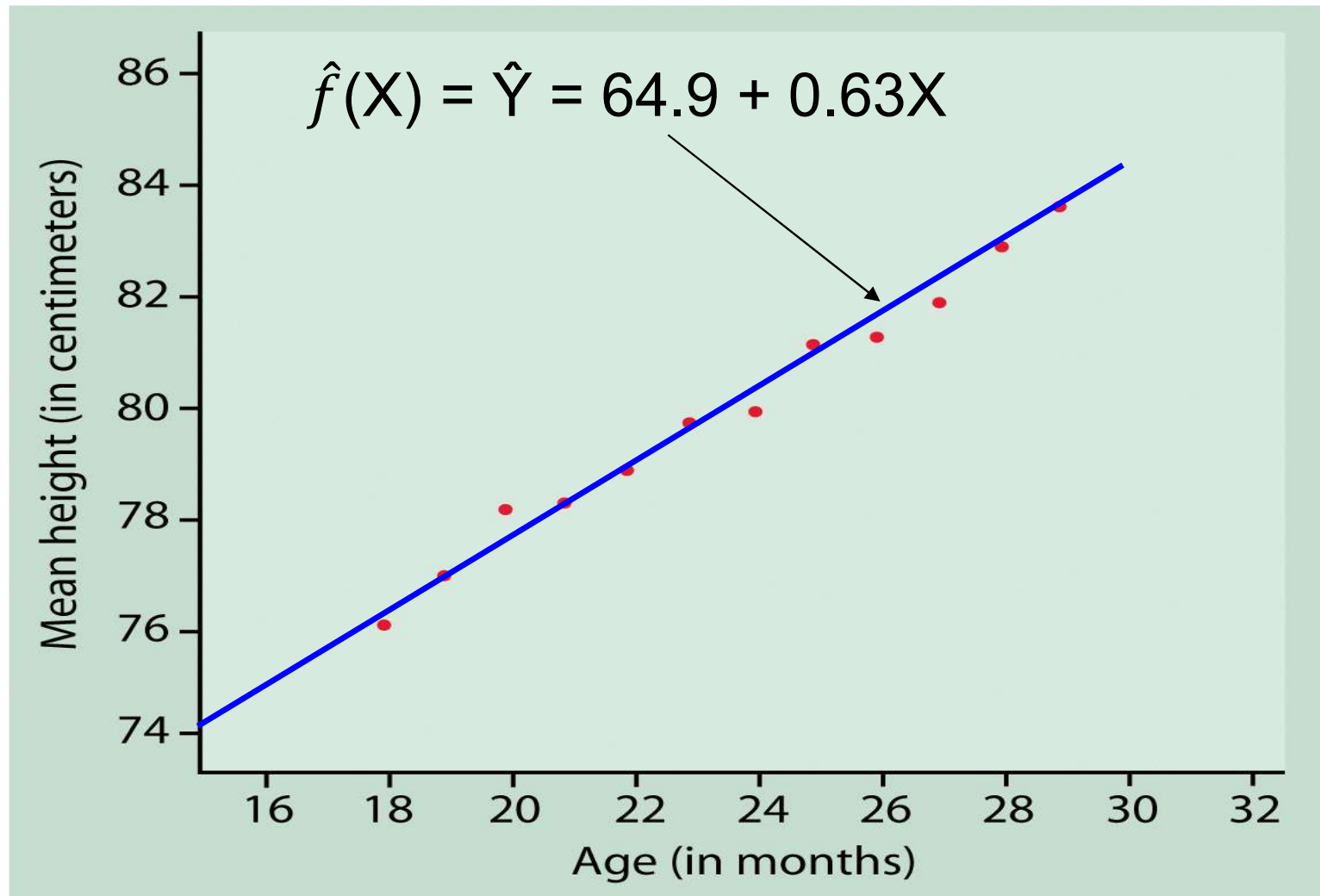


Training Data

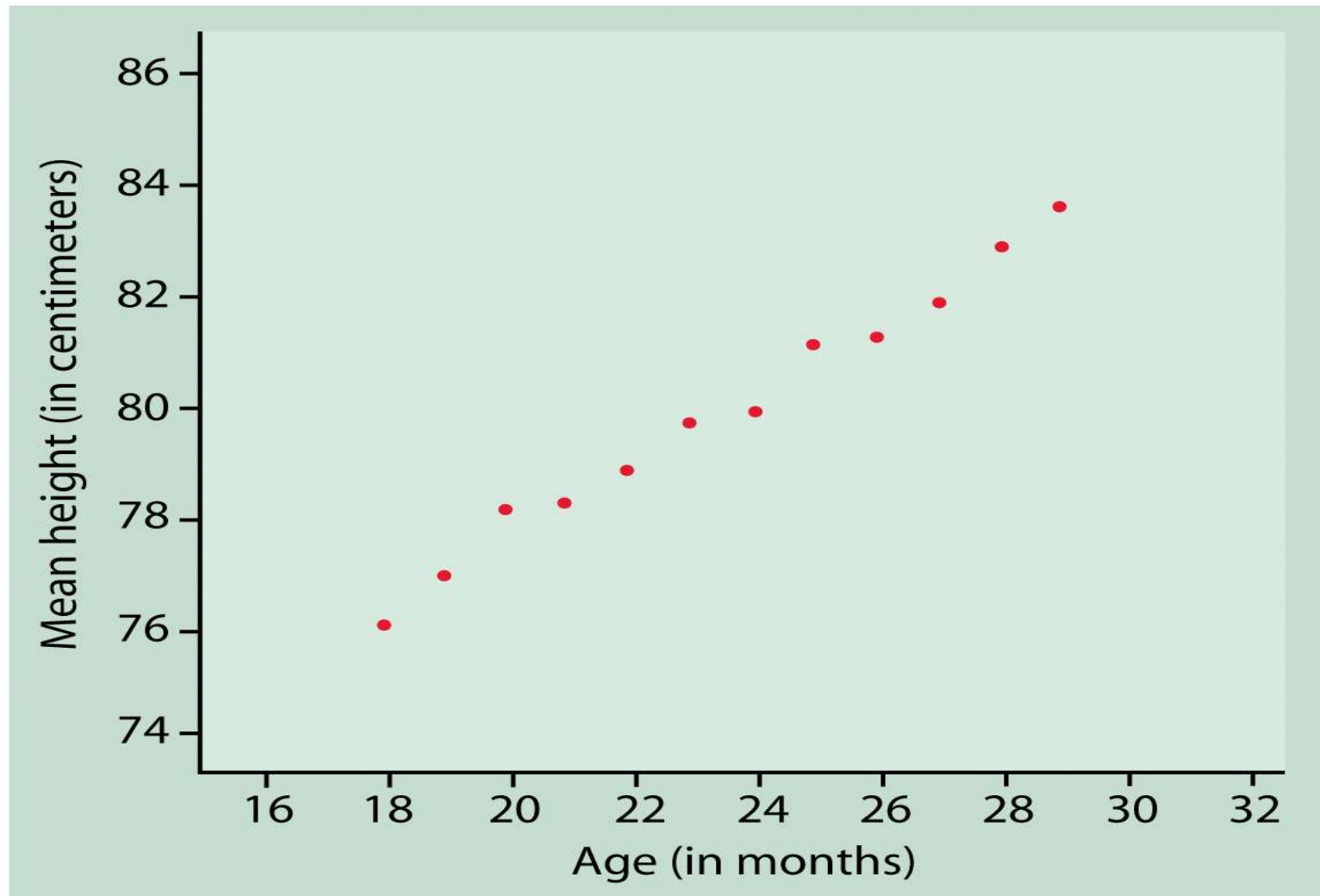
| Age X (month) | Height Y (cm) |
|------------------|------------------|
| 18 | 76.1 |
| 19 | 77.0 |
| 20 | 78.1 |
| 21 | 78.2 |
| 22 | 78.8 |
| 23 | 79.7 |
| 24 | 79.9 |
| 25 | 81.1 |
| 26 | 81.2 |
| 27 | 81.8 |
| 28 | 82.8 |
| 29 | 83.5 |

- These observations are called the **training data** because they are used to train or teach our method how to estimate the unknown function f .

Supervised Learning

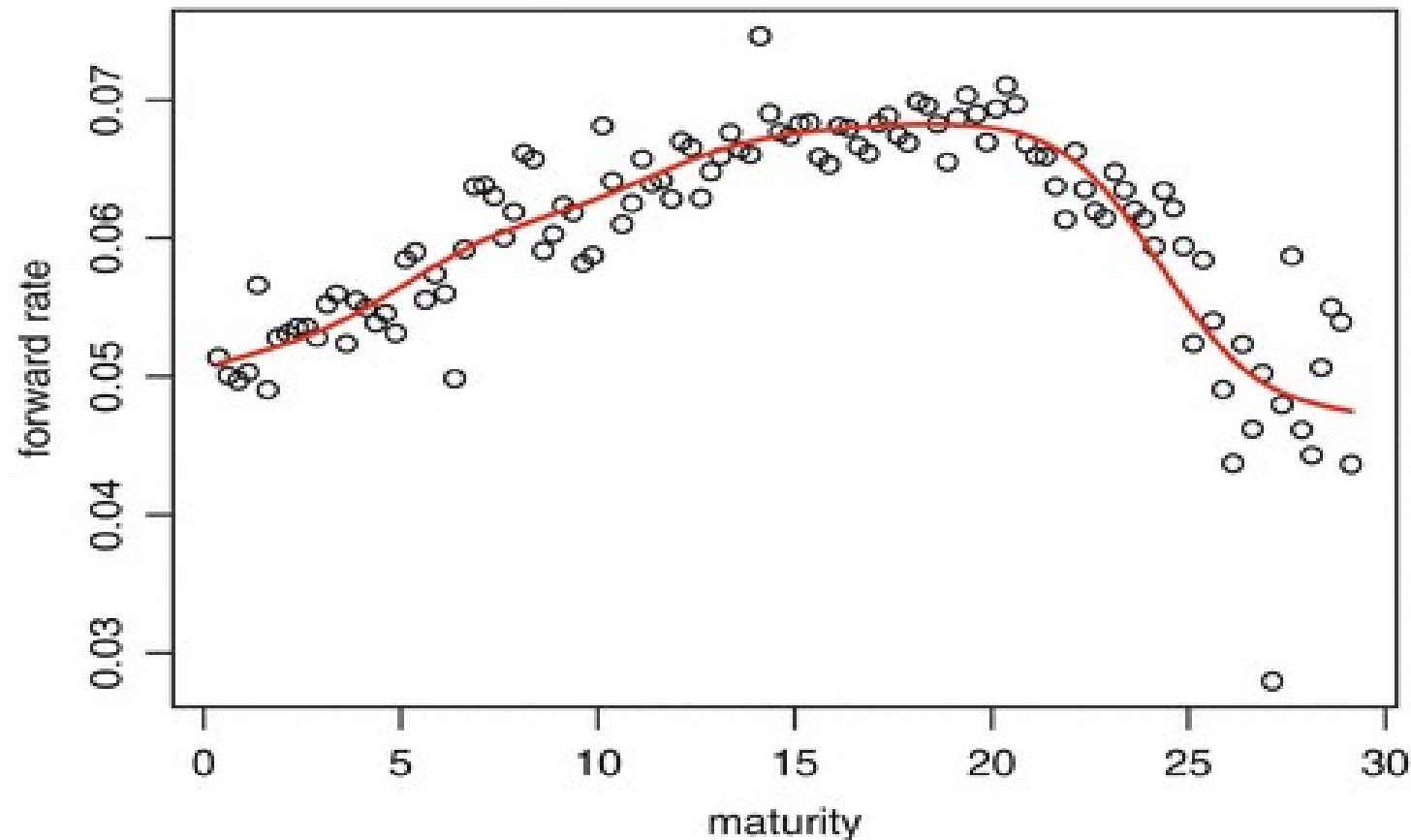


Supervised Learning



Moore & McCabe (1998, p.136)

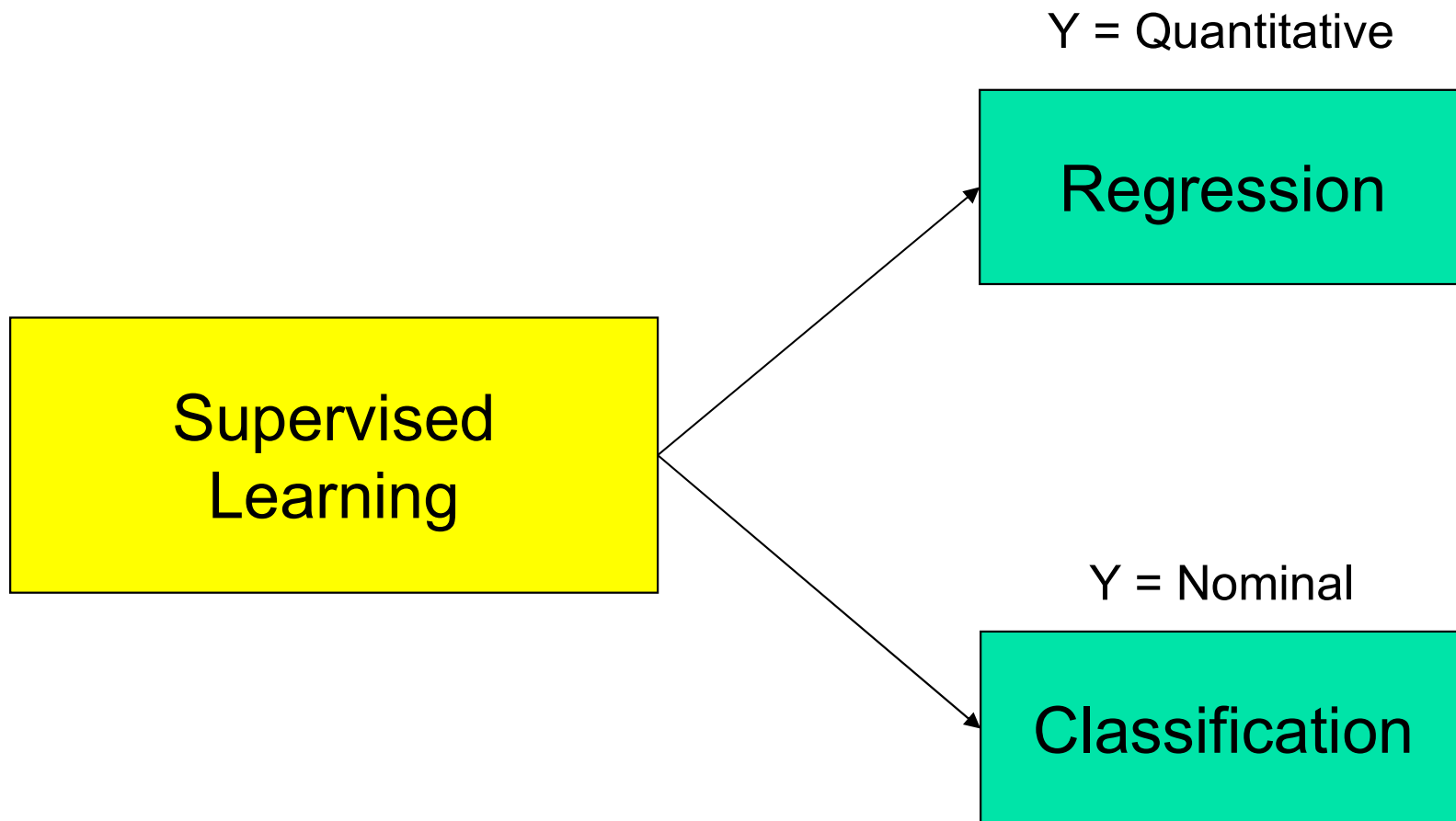
Supervised Learning



Ruppert D. & Matteson D.S. (2015) Nonparametric Regression and Splines. In: Statistics and Data Analysis for Financial Engineering. Springer Texts in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-4939-2614-5_21



Supervised Learning



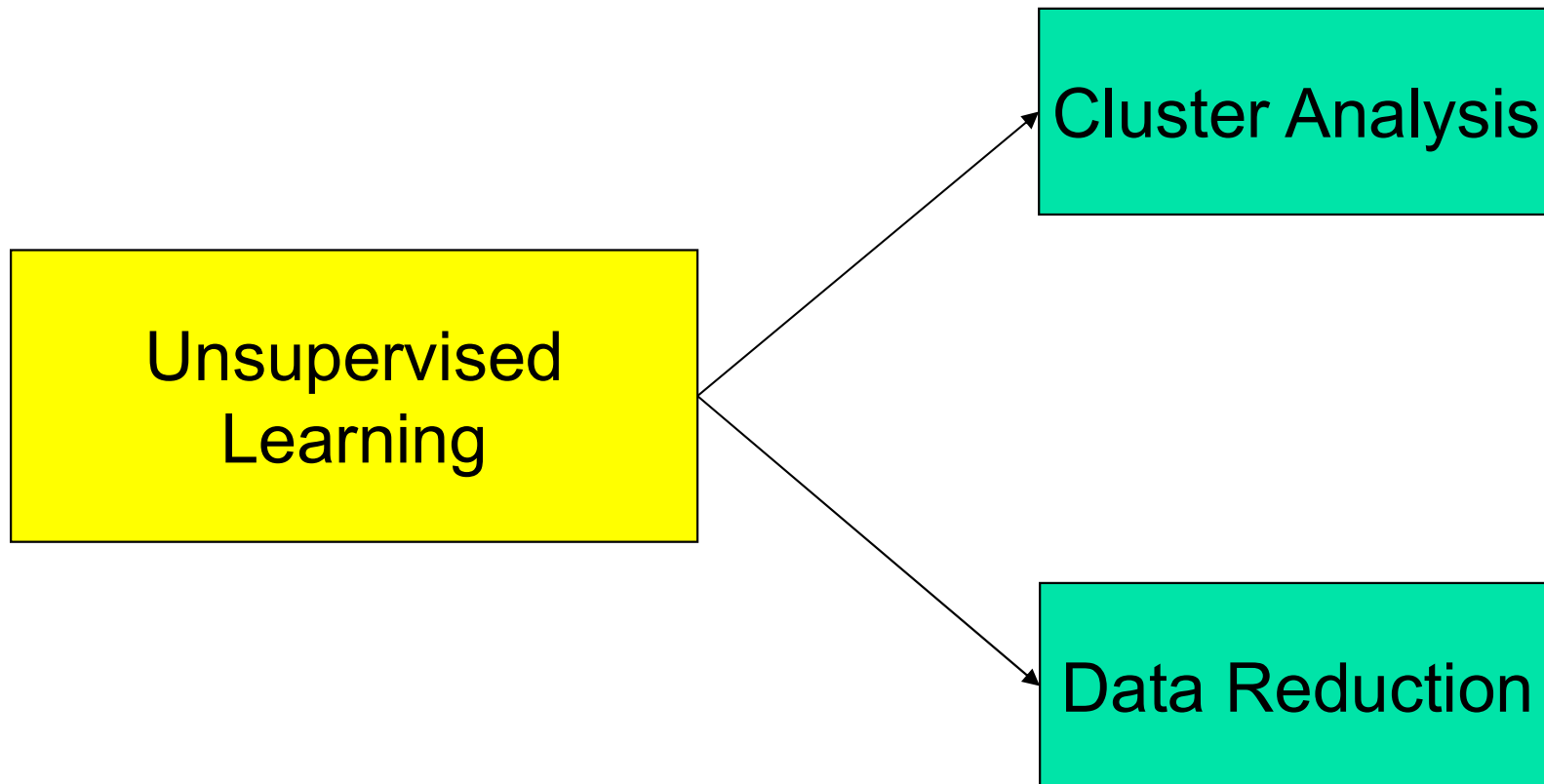


Supervised Learning Methods

- Linear Regression
- K-Nearest Neighbors Regression
- Logistic Regression
- Discriminant Analysis
- Naïve Bayes
- K-Nearest Nearest Neighbors
- Regularized Regression
- Data Reduction Regression
- Tree-Based Methods
- Support Vector Machines



Unsupervised Learning





Unsupervised Learning

- No outcome variable, just a set of variables (features) measured on a set of samples.
- The objective is to find groups of samples that behave similarly, find features that behave similarly, or find linear combinations of features with the most variation.
- It can be useful as a pre-processing step for supervised learning.



Unsupervised Learning Methods

- Clustering Methods
 - K-means Clustering
 - Hierarchical Clustering
 - Finite Mixture Models
- Data Reduction Methods
 - Principal Components Analysis
 - Factor Analysis
 - Canonical Correlation Analysis

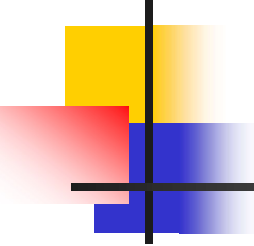
Unsupervised Learning Problems - Clustering



sample



Cluster/group



Unsupervised Learning Problems - Data Reduction: PCA

- Buchala, Davey, Gale, & Frank (2005)
 - A subset of FERET (Facial Recognition Technology) database
 - 2670 grey scale frontal face images

| Property | No. Categories | Categories | No. Faces |
|-----------|----------------|-------------------------------------|-----------|
| Gender | 2 | Male | 1603 |
| | | Female | 1067 |
| Ethnicity | 3 | Caucasian | 1758 |
| | | African | 320 |
| | | East Asian | 363 |
| Age | 5 | 20 – 29 | 665 |
| | | 30 – 39 | 1264 |
| | | 40 – 49 | 429 |
| | | 50 – 59 | 206 |
| | | 60+ | 106 |
| Identity | 358 | Individuals with 3 or more examples | 1161 |

Unsupervised Learning Problems - Data Reduction: PCA

- Each image is pre-processed to a **65 X 75** resolution
- Aligned based on eye locations
- Cropped such that little or no hair information is available





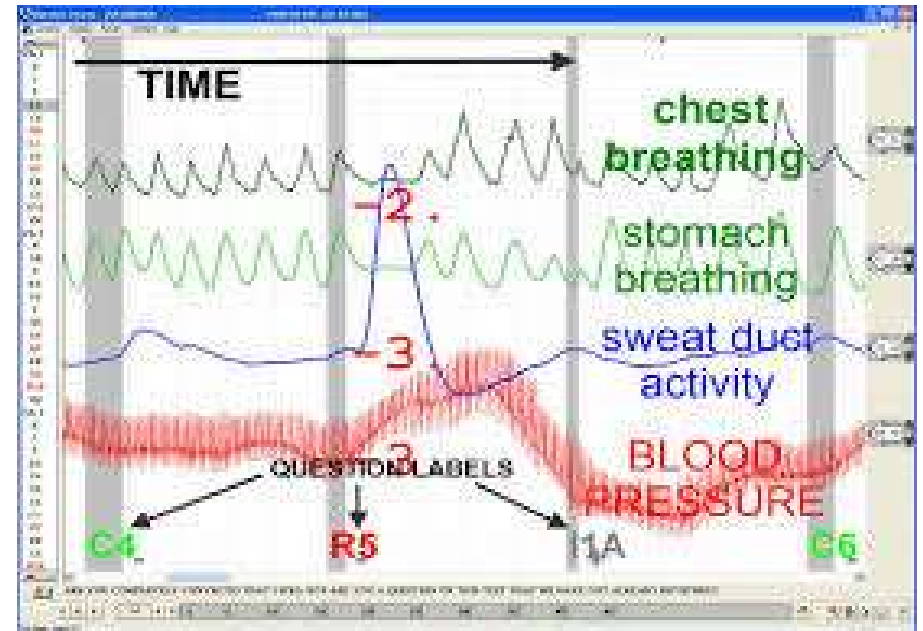
Unsupervised Learning Problems - Data Reduction: PCA

- Images of 65×75 resolution leads to a dimensionality of **4875** (= num of variables)
- The first **350** components accounted for 90% variance of the data
- Each face is thus represented using 350 components instead of 4875 dimensions
- In this example, $D = 350 \ll P = 4875$

Unsupervised Learning Problems - Data Reduction: PCA



- By a single component (3rd) – related to gender



The Lie Detection Example

- We examine how the components of two intensive longitudinal predictors “skin conductance level” and “pulse” affect a binary outcome (1 = lies and 0 = truth).

skin conductance level



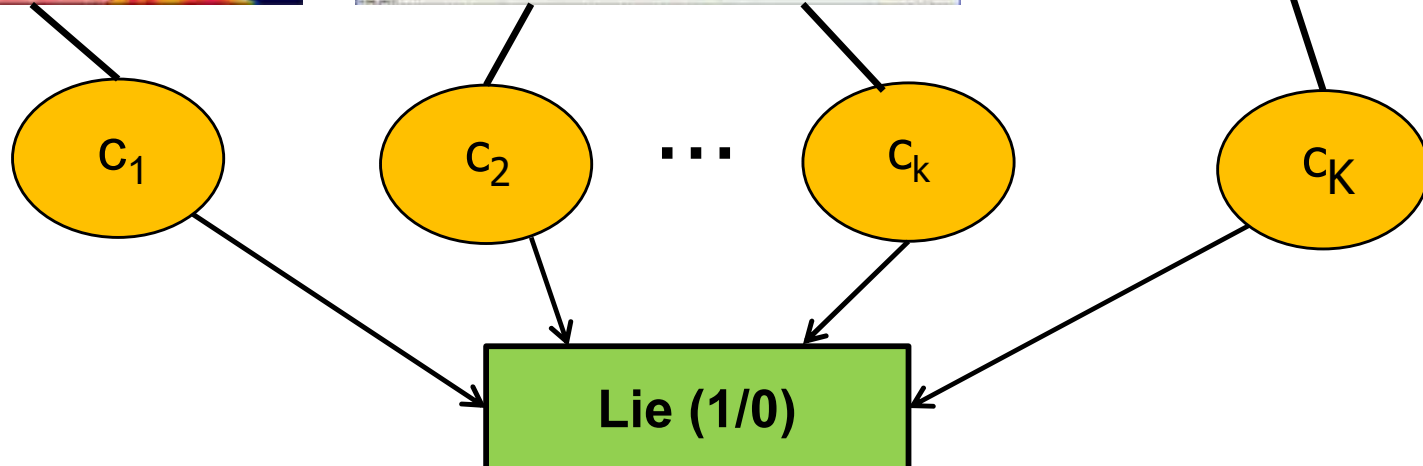
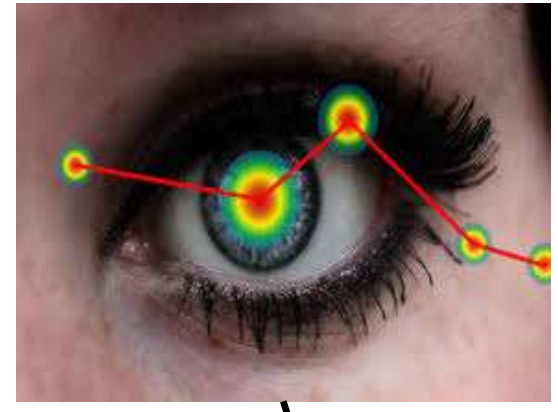
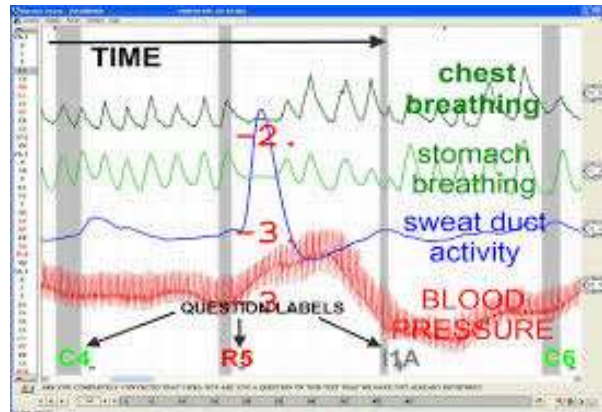
pulse



Device



Data





<https://www.r-project.org/>



Studio[®]

<https://www.rstudio.com/>