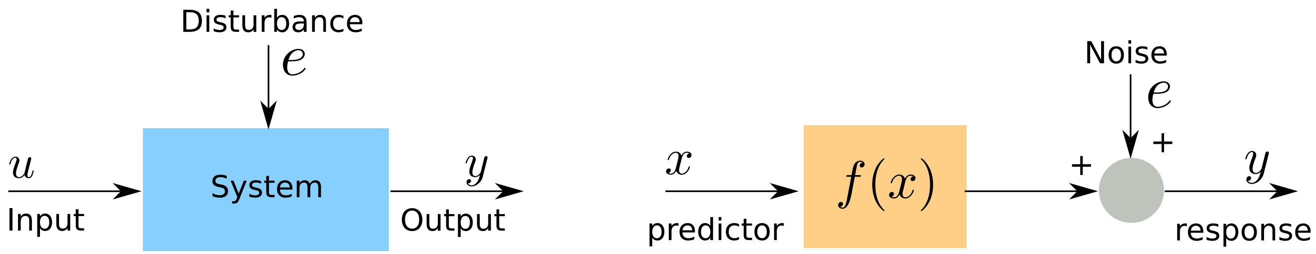
1. Introduction

- basic concept
- statistical learning methods
- procedures in statistical learning
- covered topics

Basic concept

objective: how to build a statistical model that explains a response variable from measurements



when we talk about a model

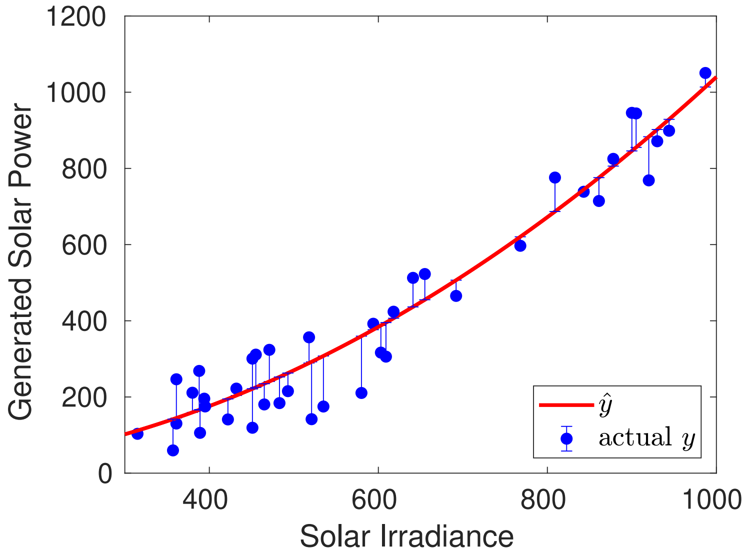
- a dynamical model with input u and output y: $y = Gu$
- a statistical model with predictor x and response y: $y = f(x)$

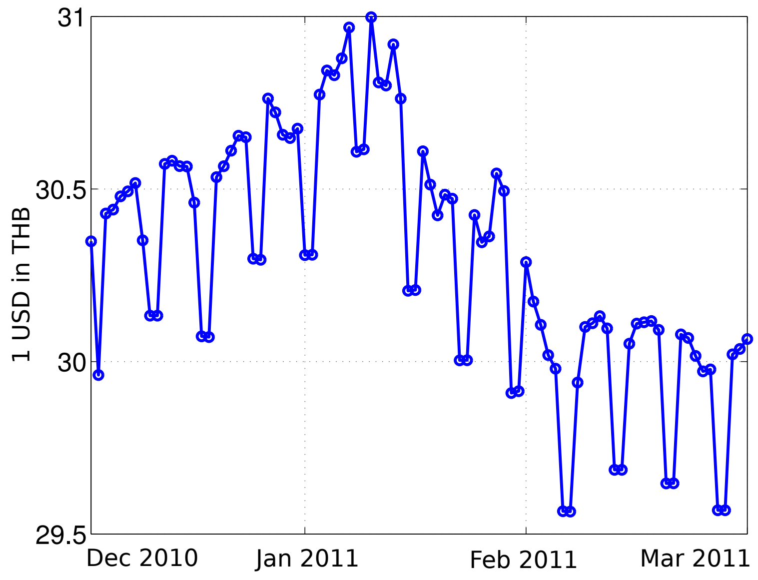
due to uncertainty of measurement or unexplained phenomenon

the output is assumed to be corrupted by noise example of learning problems:

- prediction: whether a patient due to a heart attack will have a second one
  (data = demographic, diet, clinical measurements of patients)
- prediction: forecast stock price of 1 week from now
  (data = company performance measures and economic data)
- classification: filter spam emails
  (data = relevant emails and spam emails)
- estimation: wages of population in a region
  (data = gender, age, education, year)
- inference: learn dependency structures among variables
  (data = stock prices and oil prices

Prediction





- left: estimate generated solar power from measurements of solar irradiance

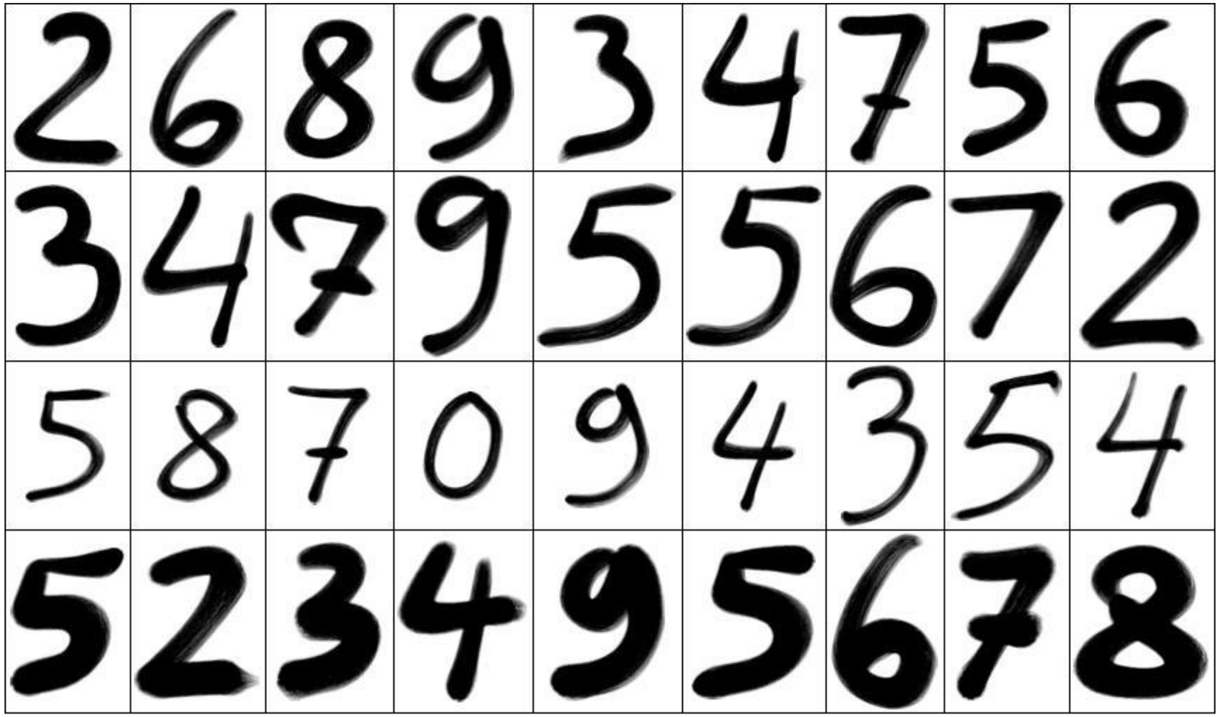
$$\text{solar power} = f(\text{solar irradiance}) \approx \beta_0 + \beta_1 I + \beta_2 I^2 + \cdots + \beta_n I^n$$

- right: forecast the Thai Baht in Apr, May,... ? need a model for prediction

$$\hat{x}_{\mathrm{Apr}} = a_1 x_{\mathrm{Mar}} + a_2 x_{\mathrm{Feb}}$$

Classification

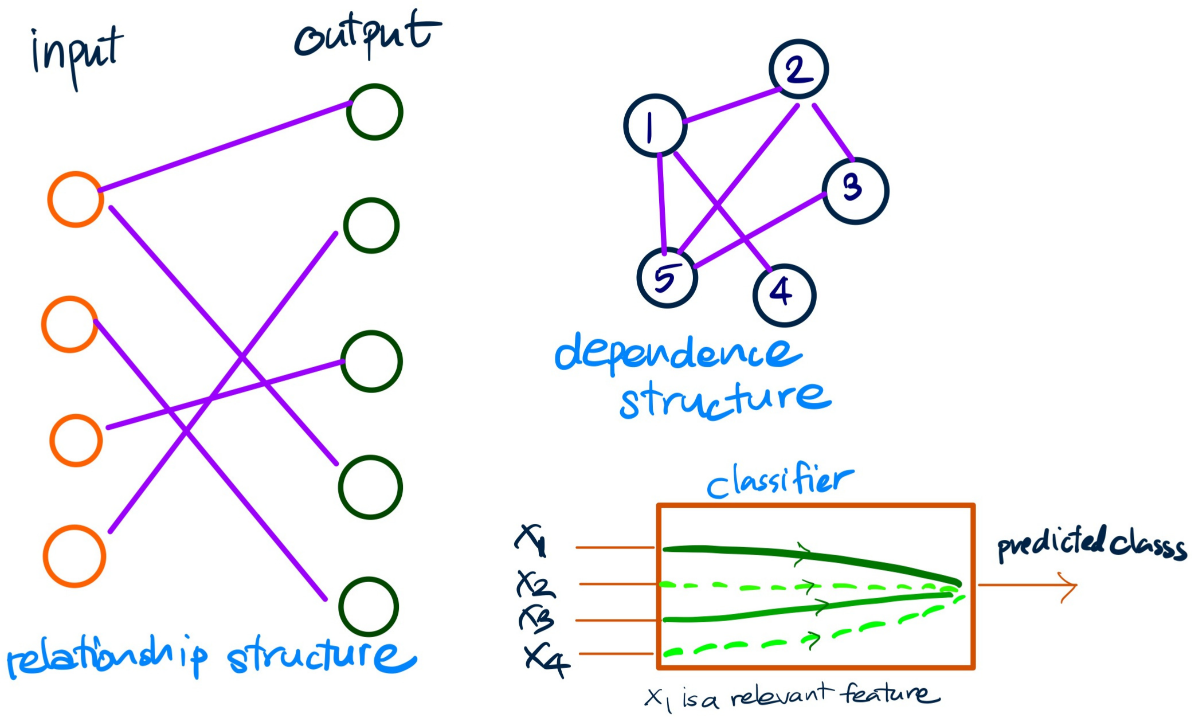
example: classify handwritten numbers from images into each number in $\{0, 1, \ldots, 9\}$



data $=$ images of handwritten digits of the same size and orientation

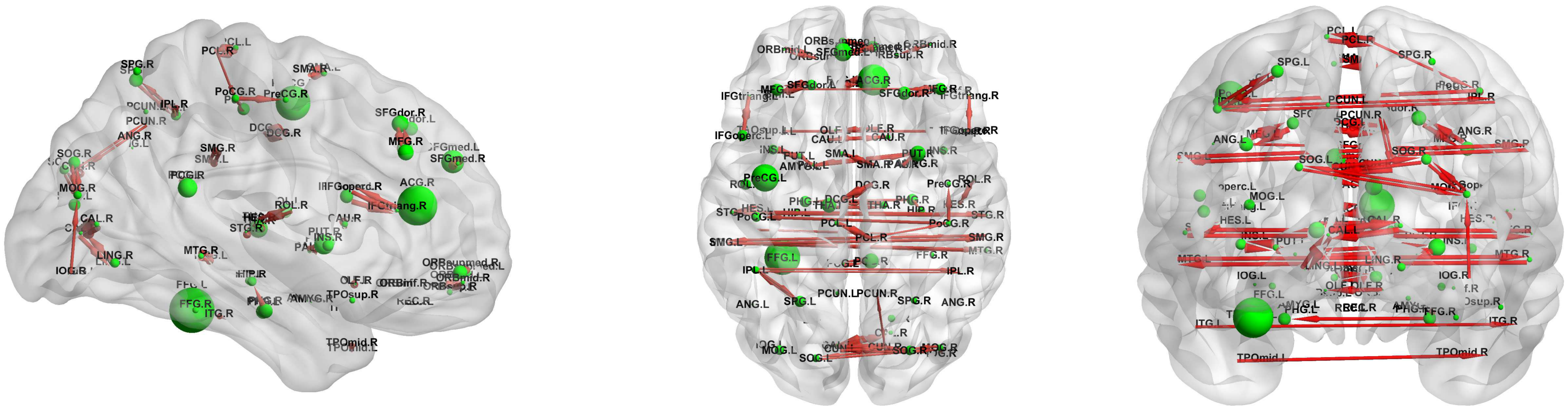
Model inference

model parameters (or its function) can *infer* some pattern of data



- interconnection structure between (y, u) or among the variables
- relevancy of using a set of features to explain the response variables

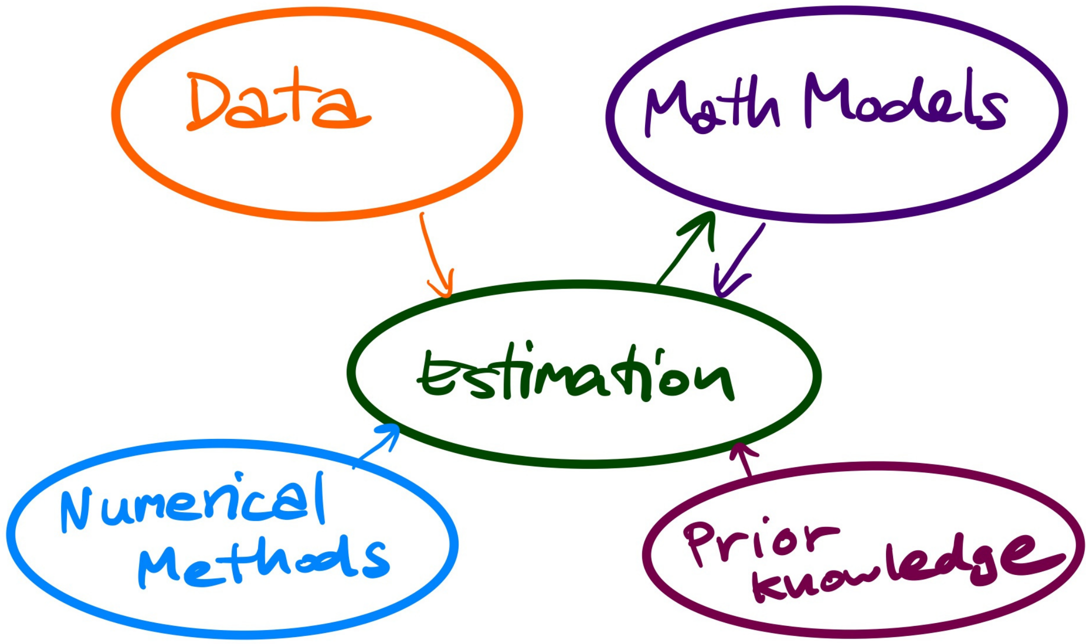
example: learn a connectivity pattern of brain regions (called brain network)



data = fMRI images (brain signals)

Essential elements

users develop a math model to explain data using prior knowledge of applications



- applicable estimation techniques depend on a selected model
- most model estimation problems require numerical methods to get a numerical solution

Mathematical setting

in most statistical learning problems, we seek for an association between

input variables (X) and output variables (Y)

- X: predictors, independent variables, features
- Y: response, dependent variables, target

a relationship between X and Y is presented in a general form

$$Y = f(X) + \epsilon$$

- f is some fixed but unknown function that represents *systematic* information that X provides about Y
- ϵ is a random **error term** which is independent of X

statistical learning refers to approaches for **estimating** f

Importance of estimating f

- classification: Y represent class labels; we can classify data once new X is obtained
- prediction: we can predict the outcome: $\hat{Y} = \hat{f}(X)$ where
  - $\hat{f}$ as a black box or explicit form that yields a good accuracy of approximating f
  - example: wage $= f($education, age, gender , year $)$ and f is linear
- inference: we can understand how Y change as a function of X; example of questions
  - which predictors are associated with the responses?
  - what is the relationship between the response and each predictor?
  - e.g., which advertising channel affect most of the sales?, which brain region is mostly-activated?

  for inference problem, an exact form of $\hat{f}$ must be provided

Approaches of estimating f

goal: apply a method to estimate the unknown function f such that

$$Y \approx \hat{f}(X)$$

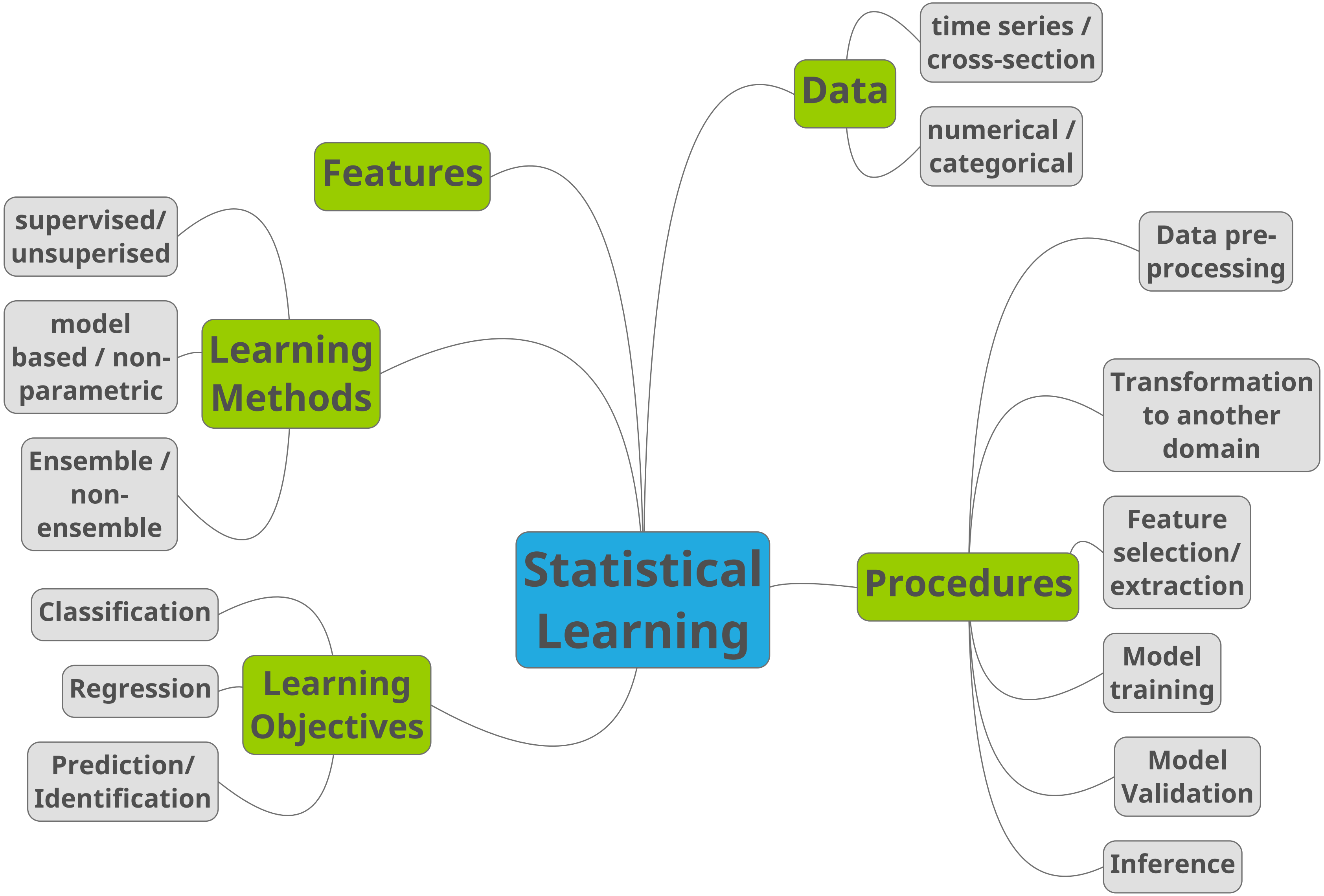
most methods for this task can be characterized as

- **parametric** (model-based) approach
  - $\hat{f}(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
  - $\hat{f}(X) = \frac{1}{1+e^{-\beta^T X}}$

  estimating f then becomes the problem of estimating parameters in $\hat{f}$

- **non-parametric** approach: do not make explicit assumptions about the form of $\hat{f}$

Elements in Statistical Learning



Data types

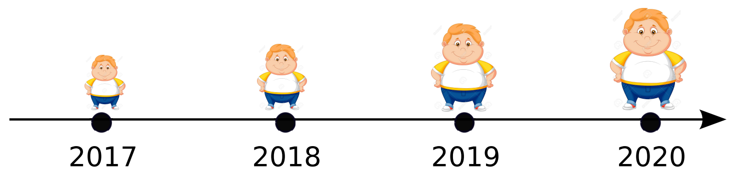
quantitative data

- cross-section data
- time series data
- panel (longitudinal) data
- repeated (pooled) cross-section data

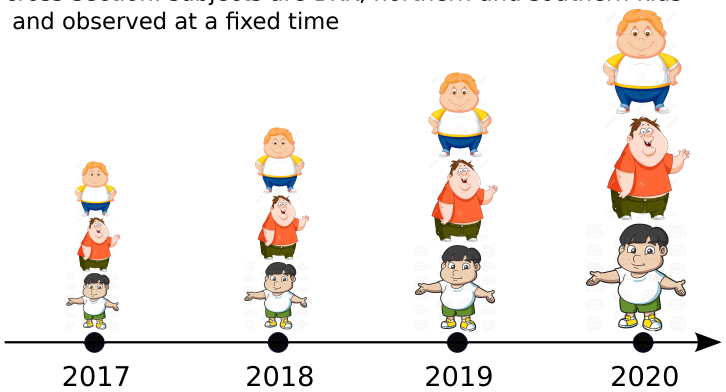
data type	**brief description**
cross-section	collected from several subjects at the *same* point of time
time series	a certain entity is observed at *various* points in time
panel	combine both cross-section and time series data
repeated cross-section	observe different subjects at different points of time

example: study about kid obesity by measuring height, weight, etc

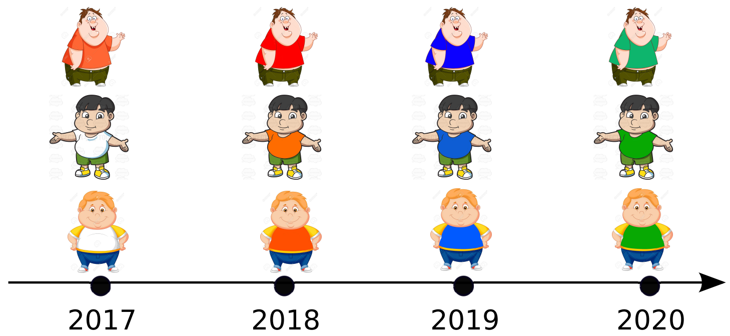
cross-section: subjects are BKK, northern and southern kids and observed at a fixed time



time series: BKK kids are observed over time



panel: kids from three groups are observed over time



repeated cross-section: kids from each group but different individual are observed at different times

Data types

qualitative data

- non-numerical and often assumed to be in a finite set
- examples: 3-class labels of states as { BKK, Chiangmai, Phuket }, patient condition as { negative, positive }
- also referred to as **categorical, discrete** variables or **factors**
- can be represented by numerical *codes*

ordered categorical data

- qualitative data with some ordering but no metric notion is appropriate
- example: { small, medium, large }

Features

a feature is an input variable that is informative for the response variable

in many cases, raw data may not be relavant or redundant to the output variable, so we need

- feature selection: select X that mostly explain Y
- feature extraction: transform raw data into another domain

methods in feature extraction/selection include subset selection, principal component analysis (PCA) or independent component analysis (ICA)

for example: Y is the state of seizure (on/off) and EEG signals are raw data; feature X can be signal energy in a low-frequency band (computed in frequency-domain, or in wavelet-domain)

Models

a description of the system, or a relationshop among observed data

a model should capture the essential information about the system

types of models

- mathematical models, e.g., algebraic, differential or difference equations

$$y = Ax, \quad \dot{y}(t) = Ay(t), \quad y(t+1) = Ay(t)$$

- probablilistic models, e.g, probability density function

$$y \sim \mathcal{N}(\mu, \sigma^2)$$

estimation (or model training) is a process of obtaining model parameters based on a data set

Statistical learning methods

categorized based on how a model is used

- **model-based (or parametric) approach**
  - an explicit form of model is made, $\hat{f}$
  - given training data set, estimate model parameters as model complexity varies
  - advantage: it reduces the problem of estimating f to a small number of model parameters
  - disadvantage: if the assumed functional form of f is very different from the true f, the model will not fit well with the data
- **non-parametric approach**
  - no assumption about the form of f is made
  - major advantage: it does not reduce the problem of estimating f to a small number of parameters, a very large number of observations is required to accurately estimate f

Statistical learning methods

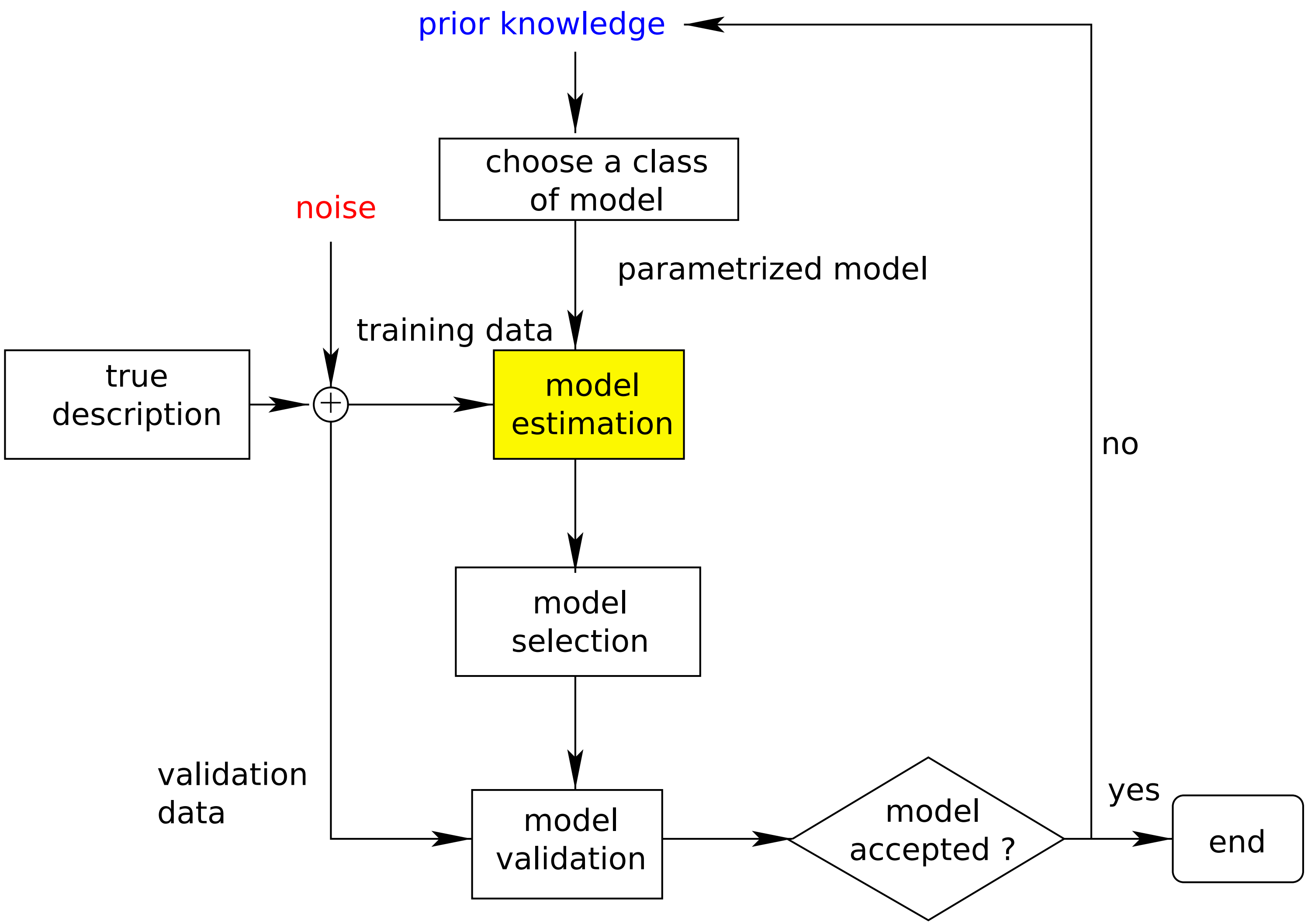
categorized based on how to guide the learning process

- **supervised learning**
  - the presence of outcome variable is used to guide the learning process
  - examples are regression, support vector machine, neural network
- **unsupervised learning**
  - we observe only the features (no measurements of outcome) and describe how the data are clustered
  - examples are k-means clustering, k-nearest-neighbor, principal components analysis

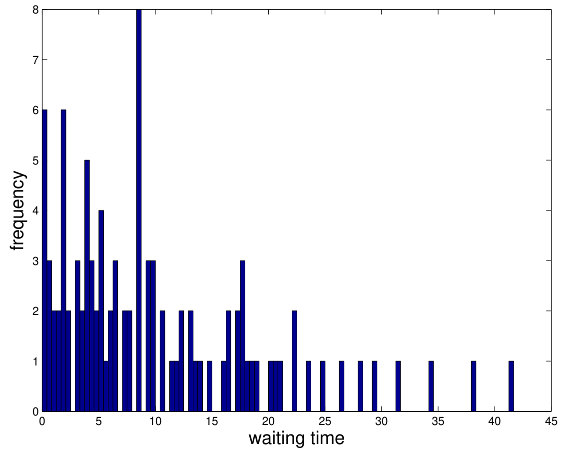
Procedures in Statistical Learning

- data pre-processing: missing-value imputation, removing artifacts, normalization, preparation of data sets for experiments
- feature selection/extraction: to choose relavant input variables for the output
- model training: this is to estimate f from (X, Y) data
  - this steps involve varying complexity of models
  - one obtain many candidate models in this step
- model validation: compare candidate model performance evaluated on unseen data (validation set)
  - example of methods: leave-one-out cross-validation, k-fold cross-validation, residual analysis, white-ness test
- inference: use the selected model to further infer about the learning goal

Flow chart of training and validation process

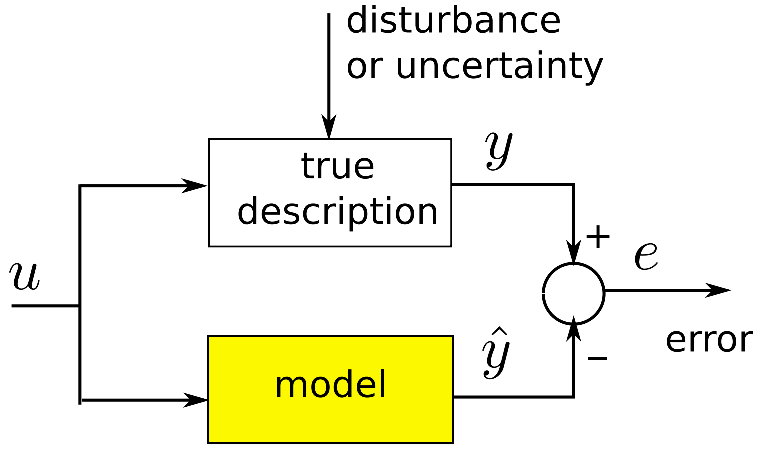


example: the characteristics of the waiting times (T) in a bank



- **data randomness**: the waiting times (T) always change when we recollect
- model: choose a probablilistic model (pdf) to explain the data
- **prior knowledge**: T is nonnegative and varies with date and hour of operation, x
- chosen model: pdf of exponential random variable $f(T) = \lambda e^{-\lambda T}$ and choose $\lambda = e^{x^T \beta}$ (the distribution parameter is linked with predictors)
- **model estimation**: determine an optimal value of λ (or β)

Model estimation



- errors are from i) model mismatch and ii) part of noise characteristics the model can't explain
- measured quantitatively by some metric, *e.g.*, sum of square, likelihood
- having a lowest error is a way to judge if a model is good (goodness of fit)
- the process of obtaining model parameters that lead to an optimal model
- model estimation is often an optimization problem (variable = model parameter)

Essense of model accuracy

a given estimate $\hat{f}$ that yields $\hat{Y} = \hat{f}(X)$ follows

$$\mathbf{E}[(Y - \hat{Y})^2] = \mathbf{E}[(f(X) + \epsilon - \hat{f}(X))^2] = \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{reducible}} + \underbrace{\mathbf{var}(\epsilon)}_{\text{irreducible}}$$

the accuracy of $\hat{Y}$ (here mean squared error) depends on two quantities

- reducible error: depends on the choice of $\hat{f}$
- irreducible error: how much measurement data are corrupted by noise

important notes:

- several statistical methods aim to minimize the reducible error
- the irreducible error is always a lower bound of the estimation error (but this bound is almost unknown in practice)

Essense of model selection/validation

objective of model selection: obtain a good model at a low cost

1. **quality of the model:** defined by a measure of the goodness, e.g., the mean-squared error

    - MSE consists of a *bias* and a *variance* contribution
    - to reduce the bias, one has to use more flexible model structures (requiring more parameters)
    - the variance typically increases with the number of estimated parameters
    - the best model structure is therefore a trade-off between *flexibility* and *parsimony*

2. **price of the model:** an estimation method (which typically results in an optimization problem) highly depends on the model structures, which influences:

   - algorithm complexity
   - properties of the loss function

3. intended use of the model, *e.g.*,

   - summarize the main features of a complex reality
   - predict some outcome
   - test some important hypothesis

Bias-variance decomposition

assume that the observation Y obeys

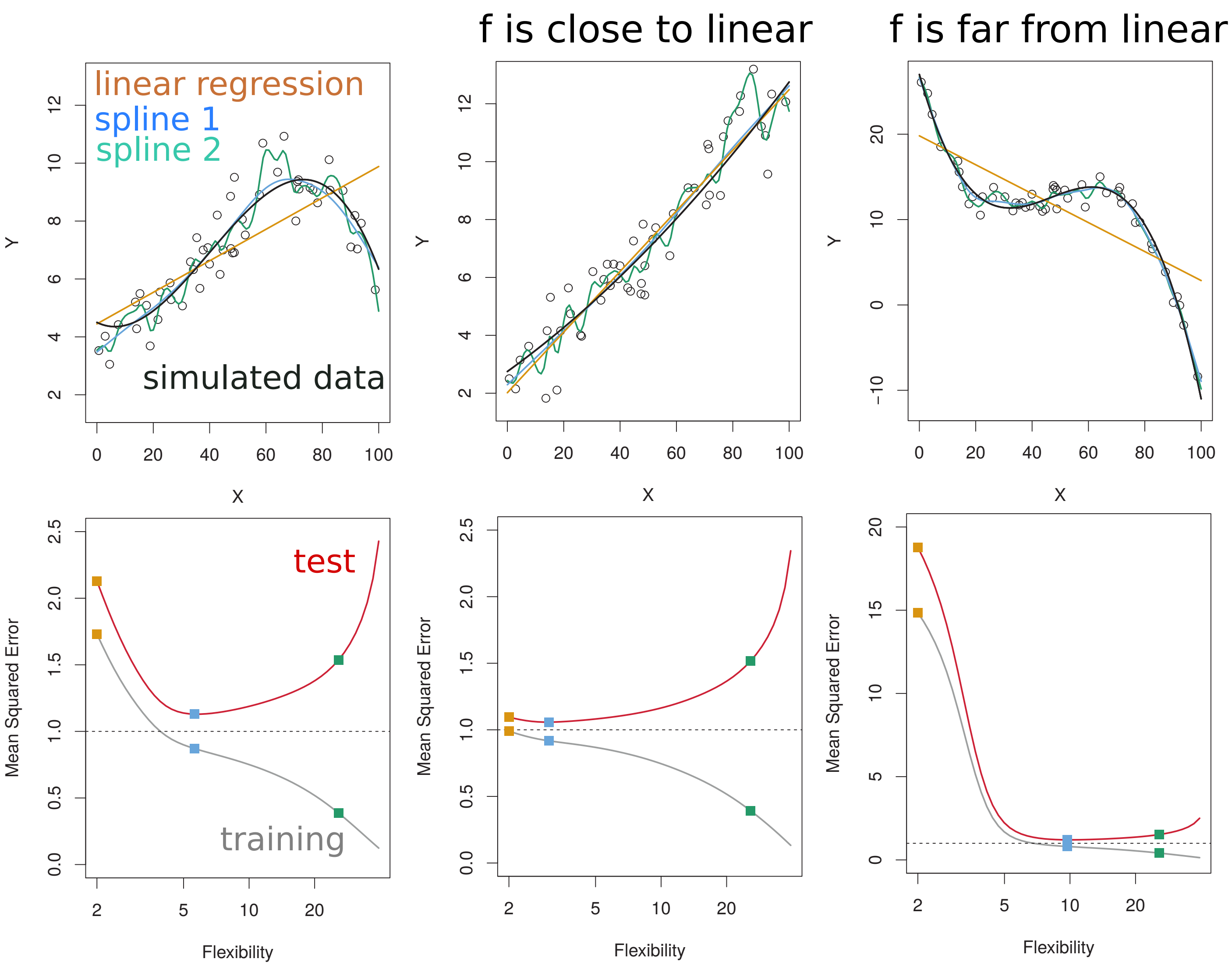
$$Y = f(X) + \nu, \quad \mathbf{E}\nu = 0, \quad \mathbf{cov}(\nu) = \sigma^2$$

the mean-squared error of a regression fit $\hat{f}(X)$ at $X = x$ is

$$\begin{aligned}
\text{MSE} &= \mathbf{E}[(Y - \hat{f}(X))^2 | X = x] \\
&= \sigma^2 + [\mathbf{E}\hat{f}(X) - f(X) | X = x]^2 + \mathbf{E}[\hat{f}(X) - \mathbf{E}\hat{f}(X) | X = x]^2 \\
&= \sigma^2 + \text{Bias}^2(\hat{f}(x)) + \text{Var}(\hat{f}(x))
\end{aligned}$$

- this relation is known as **bias-variance decomposition**
- no matter how well we estimate $f(x)$, σ^2 represents *irreducible error*
- typically, the more complex we make model $\hat{f}$, the lower the bias, but the higher the variance

figures taken from G. James, D. Witten, T. Hastie and R. Tibshirani book page 31-34



- left: more flexible models yield lower MSE in training but could have higher MSE in test data (overfitting)
- middle: when the true f is close to linear, linear regression provides a comparably good fit to the data
- right: when the true f is far from linear, linear regression gives a poor fit

how the bias and variance are varied ? depends on the choice of $\hat{f}$

- model bias: high if a model is simple
- model variance: high if a model is flexible or complex (in general)

the variance is changed when we use different training data sets proof of bias-variance decomposition: note that

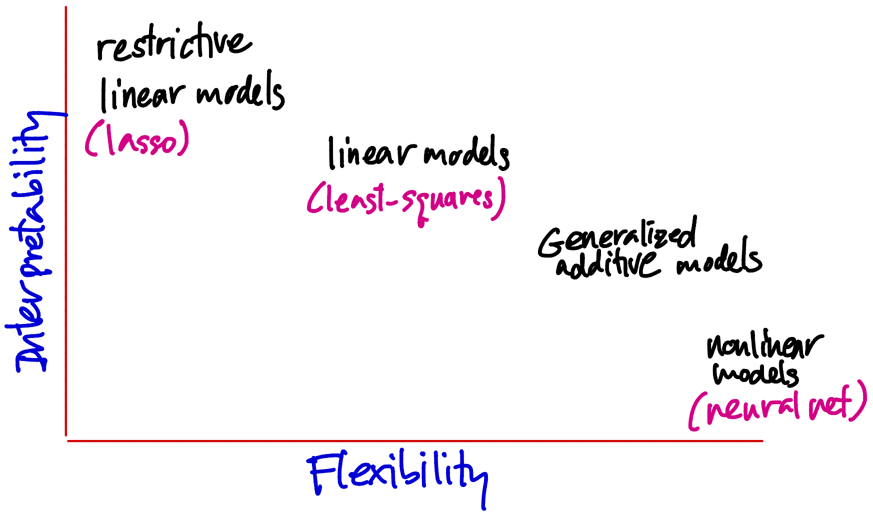
- the true f is deterministic
- $\mathbf{var}(Y|X=x) = \sigma^2$ and $\mathbf{E}[Y|X=x] = f(x)$
- $\hat{f}(x)$ is random

we will omit the notation of conditioning on $X = x$

$$
\begin{aligned}
\mathbf{E}[(Y - \hat{f}(X))^2] &= \mathbf{E}[Y^2] + \mathbf{E}[\hat{f}(x)^2] - \mathbf{E}[2Y\hat{f}(x)] \\
&= \mathbf{var}(Y) + \mathbf{E}[Y]^2 + \mathbf{var}\,\hat{f}(x) + \mathbf{E}[\hat{f}(x)]^2 - 2f(x)\mathbf{E}[\hat{f}(x)] \\
&= \mathbf{var}(Y) + f(x)^2 + \mathbf{var}\,\hat{f}(x) + \mathbf{E}[\hat{f}(x)]^2 - 2f(x)\mathbf{E}[\hat{f}(x)] \\
&= \sigma^2 + \mathbf{var}\,\hat{f}(x) + (f(x) - \mathbf{E}[\hat{f}(x)])^2 \\
&= \sigma^2 + \mathbf{var}\,\hat{f}(x) + (\mathbf{E}[f(x) - \hat{f}(x)])^2 \\
&= \sigma^2 + \mathbf{var}\,\hat{f}(x) + [\mathrm{Bias}(\hat{f}(x))]^2
\end{aligned}
$$

bias-variance dilemma can also be considered joinly with a trade-off between

prediction accuracy VS model interpretability



- prediction is the goal: more flexible model is preferred (but not always)
- inference is the goal: simple and restrictive model is favoured in order to intepret relationships between predictors and response

Selected topics

- bias and variance dilemma
- linear regression and robust regression
- resampling methods
- model selection and model assessment
- regularization techniques
- nonlinear regression model
- Bayes' theorem for classification
- logistic regression
- k-nearest negihbor
- linear and quadratic discrimination analysis (LDA,QDA)

- support vector machine
- tree-based methods (decision trees, bagging, random forest)
- principal component analysis (PCA)
- k-mean clustering
- Gaussian mixture models and EM
- self-organizing maps (SOM)

Required tools

this class focuses on

- selected techniques used in supervised and unsupervised learning
- analysis of statistical properties of models

for these reasons, we require skills on

- statistics: to analyze all random quantities
- mathematics: linear algebra, differential equations, calculus
  - to formulate a model
  - to analyze properties of model and its parameters
- optimization: in training process

References

Chapter 1,2 in
T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, Second edition, 2009

Chapter 1-3 in
G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Application in R*, Springer, 2013