# 1. Introduction

- course goal
- data types
- math settings
- essence of statistical learning
- required tools

# Course goal

in many applications,

we would like to construct a **model** that explains a pattern of **data**

**examples:**

1. kids from many regions in thailand face the same obesity problem or not ?
   - pattern: the trend of BMI can represent the obesity pattern
   - data: height, weight of the kids from many regions
2. how does the USD exchange rate change over time ?
   - pattern: monotonicity, or rate change
   - data: exchange rates collected from various years

a **model** is an description of the variable of interest chosen by the user

this course provide tools for constructing **statistical** models

# Basic concept

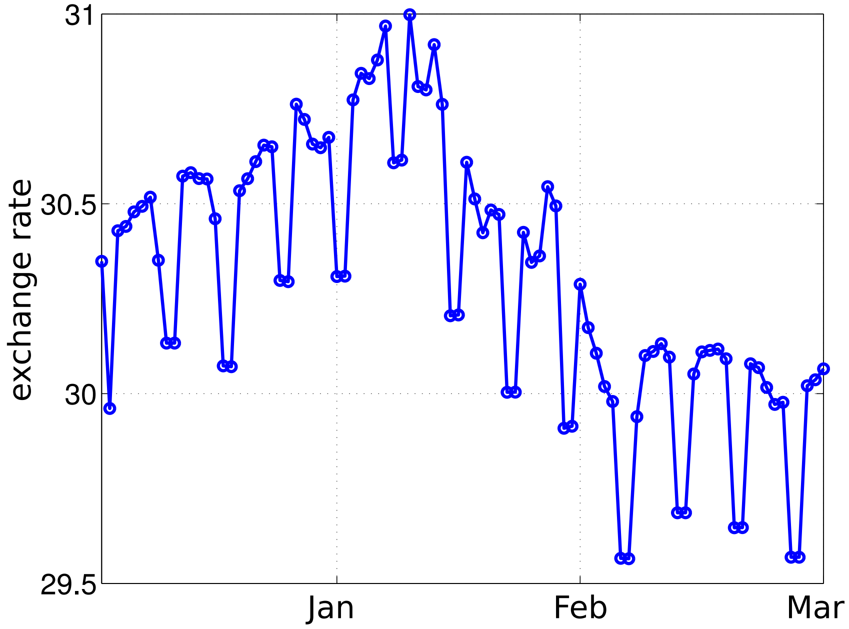
**objective**: how to draw a pattern or conclusive results in a quantative way from data using statistical concepts

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 structure of stock prices
  (data = stock indices of interest)

# Prediction

example: forecast the Thai Baht in Apr, May,... ?

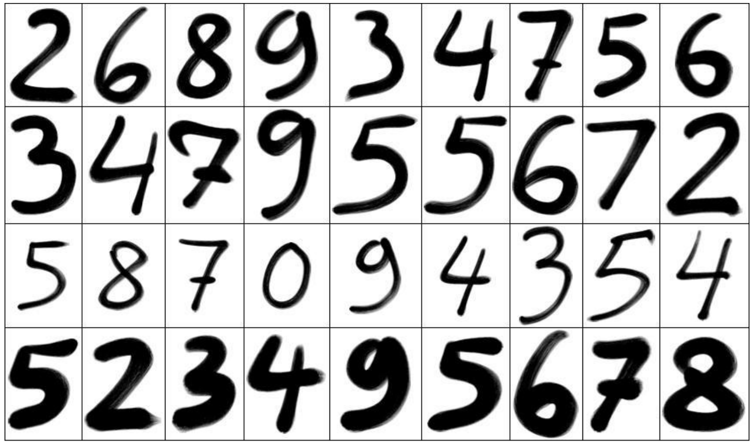



- data = historical records of exchange rate time series and economic variables
- need a model for prediction, e.g.

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

# Classification

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



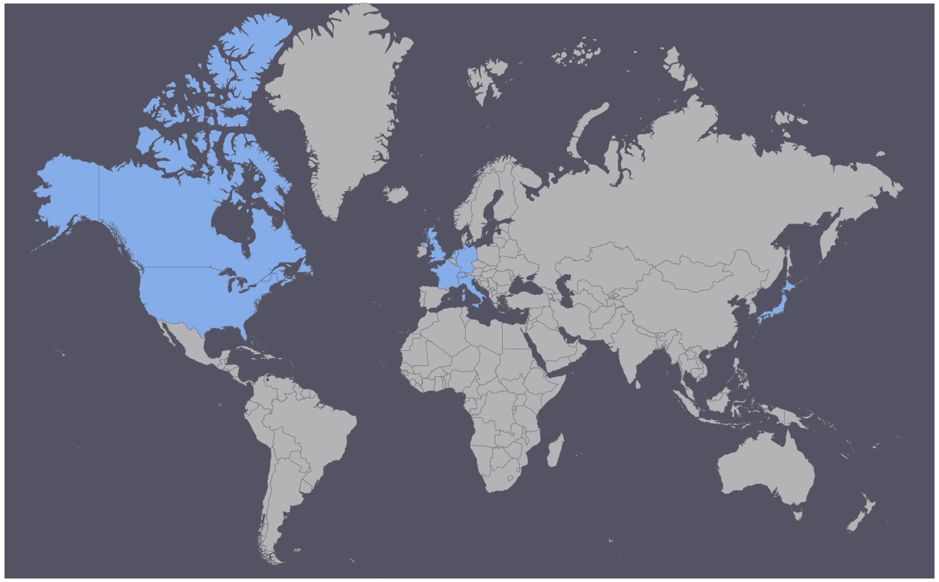
data $=$ images of handwritten digits of the same size and orientation

example 2: classify credit risk into categories

data $=$ credit score ratings from agencies

# Inference

example: learn dependency stuctures among stock prices and oil prices[1]



data = stock prices from CA, FR, GE, HK, IT, JP, NE, SW, UK, US and oil prices from Brent and OPEC

[1] K. Sukcharoen et.al., Interdependence of oil prices and stock market indices: A copula approach, Energy Economics, 2014

# 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** is a process of obtaining model parameters based on a data set

# Data types

**quantitative data**

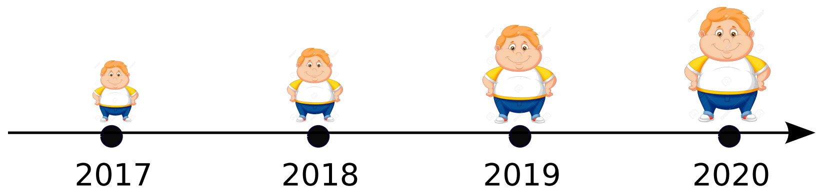
- cross-section data
- time series data
- panel (longitudinal) data
- repeated (pooled) cross-section data

| **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 |

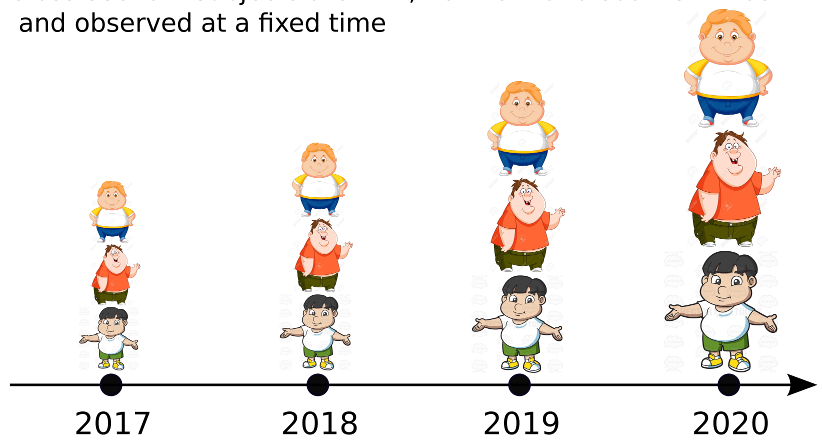
**example:** study about kid obesity by measuring height, weight, etc

BKK kids | Northern kids | Southern kids

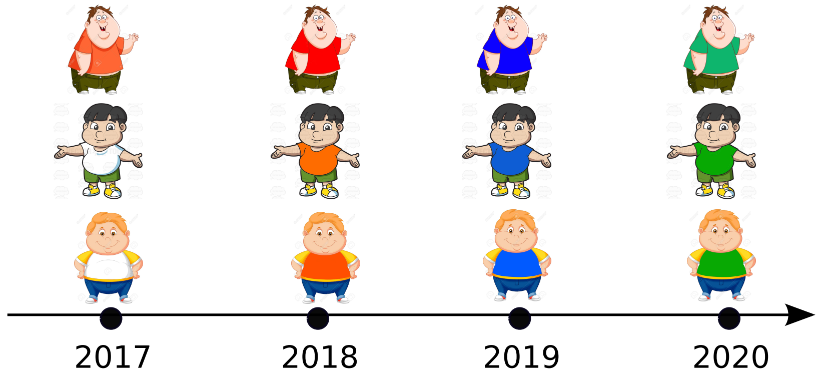

cross-section: subjects are BKK, northern and southern kids
and observed at a fixed time




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 $\}$

# 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 $Y$
- $\epsilon$ 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

# 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 financial statement fraud; feature $X$ can be debt, total assets, gross profit, primary business income, cash and deposits, accounts receivable, etc.[2]

[2] P.RavisankaraV et.al, Detection of financial statement fraud and feature selection using data mining techniques, Decision Support Systems, 2011

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

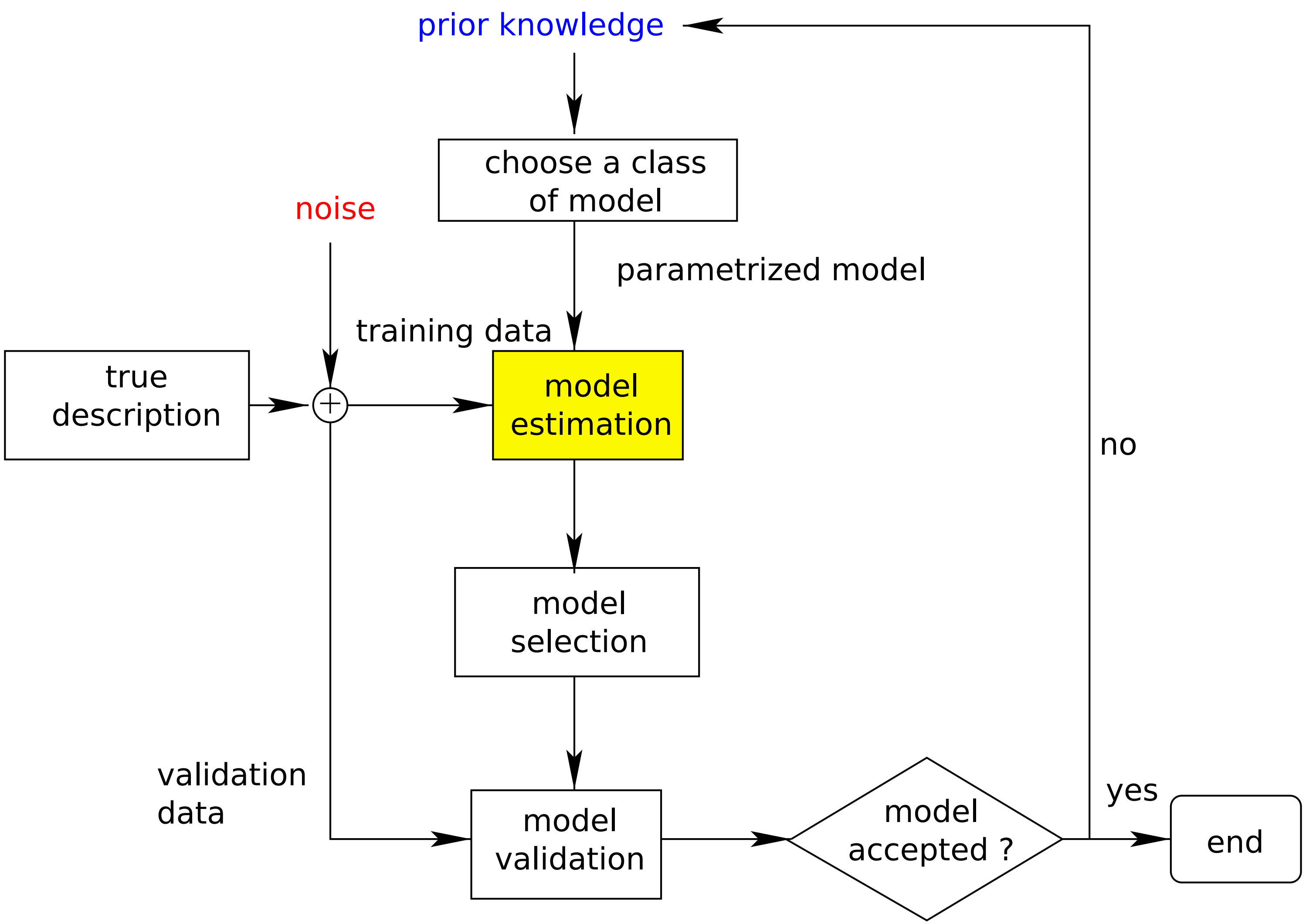
  estimating $f$ then becomes the problem of estimating parameters in $\hat{f}$

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

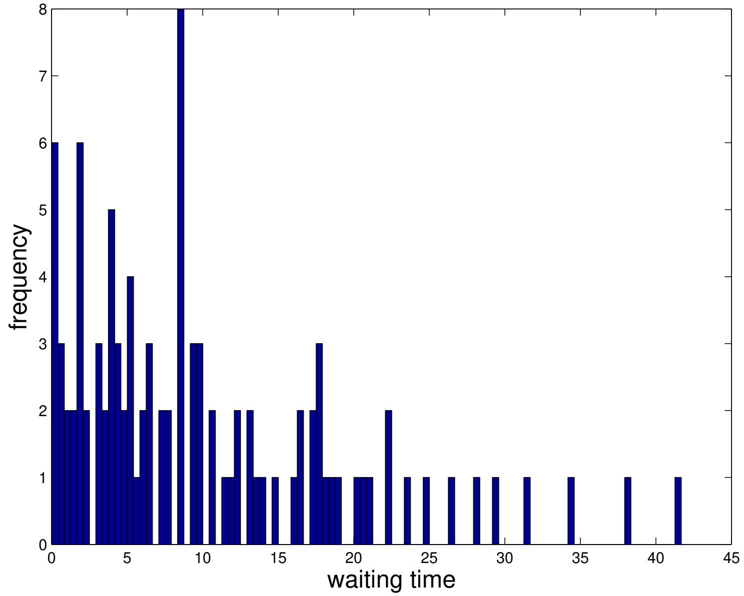
# Procedures in Statistical Learning

- 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

# Procedures in learning from data

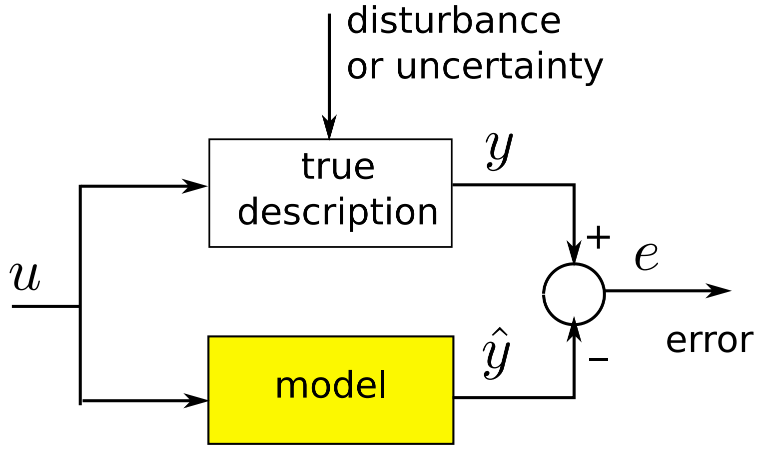


**example:** the characteristics of the waiting times in a bank




- **data randomness**: the waiting times ($T$) always change when we recollect
- model: choose a probablilistic model (pdf) to explain the data
- **prior knowledge**: the waiting time is nonnegative
- chosen model: pdf of exponential random variable $f(T) = \lambda e^{-\lambda T}$
- **model estimation**: determine the best value of $\lambda$ (best in some sense)

# 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

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

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)
   - 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

# Required tools

this class focuses on

- techniques used in model estimation
- analysis of model/estimator properties

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 estimation process