# Unit 1 Lecture 1: Intro to Modern Data Mining

## STAT 471

August 31, 2021

# What is data mining?

The automated extraction of actionable information from data.

Example: Clinical Decision Support

A patient comes into the emergency room with stroke symptoms. Based on her CT scan, is the stroke ischemic or hemorrhagic?

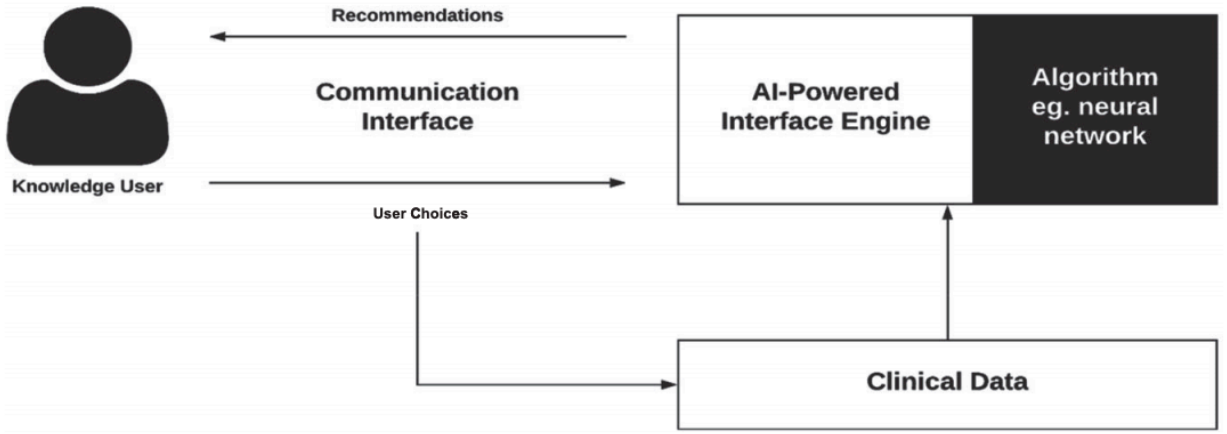


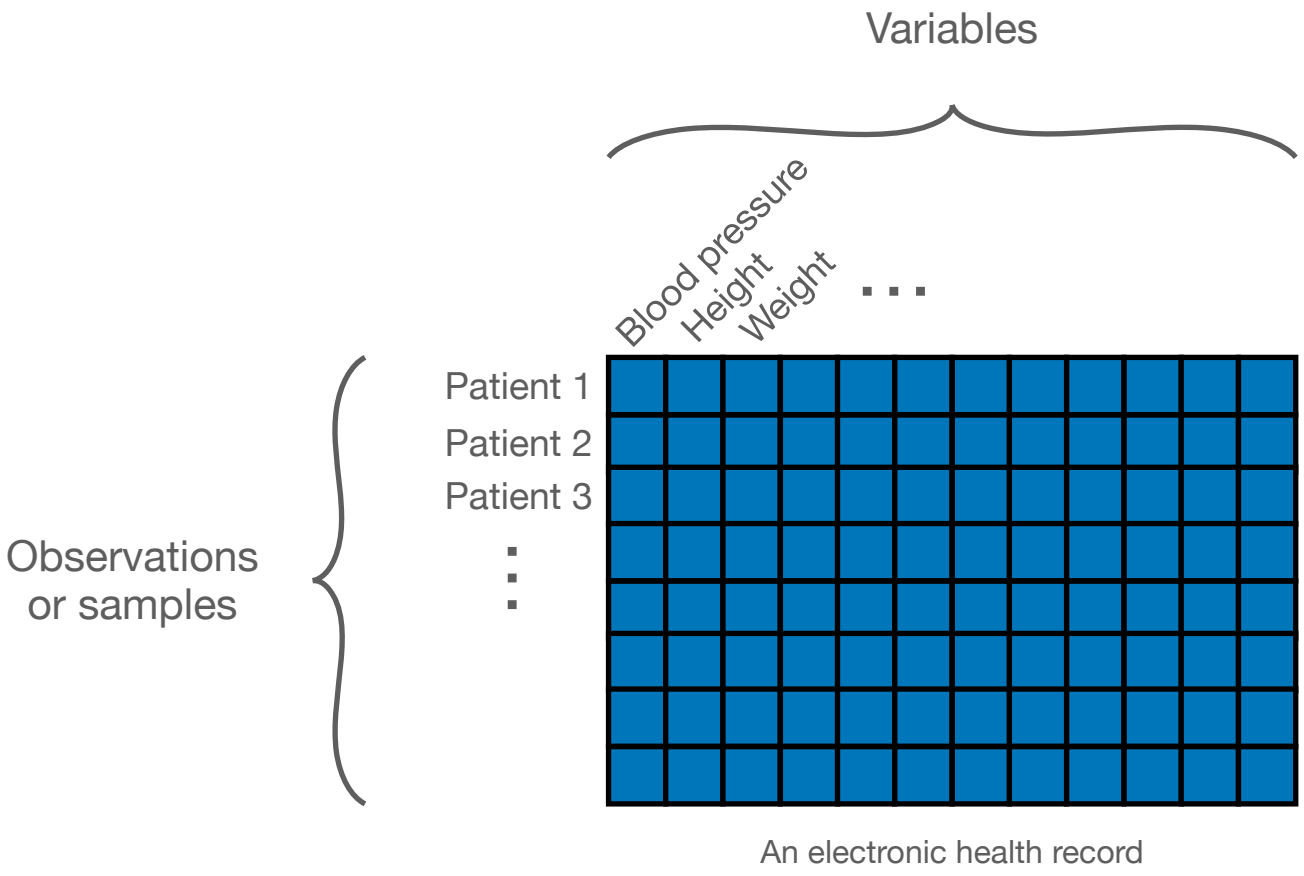

Image source: Sutton et al. 2020 (npj Digit. Med.)

# What questions can data answer?

| Field | Key question |
|---|---|
| Healthcare | Clinical decision support |
| Business | Market segmentation |
| Finance | Stock market prediction |
| Technology | Machine translation |
| Science | Protein folding |
| Politics | Election forecasting |

Data mining has revolutionized each of these fields.

# The structure of a database




| Field | Key question |
|---|---|
| Healthcare | Clinical decision support |
| Business | Market segmentation |
| Finance | Stock market prediction |
| Technology | Machine translation |
| Science | Protein folding |
| Politics | Election forecasting |

What might databases in other fields look like?

# Two kinds of data mining problems

Data mining problems come in two kinds:

- Supervised learning: Learn the relationship between one variable of special interest (*response*) and the others (*features*).
- Unsupervised learning: Find patterns among observations (e.g. clusters) based on all variables.

In this class we will focus on supervised learning.

| Field | Key question |
|---|---|
| Healthcare | Clinical decision support |
| Business | Market segmentation |
| Finance | Stock market prediction |
| Technology | Machine translation |
| Science | Protein folding |
| Politics | Election forecasting |

# Types of response variables

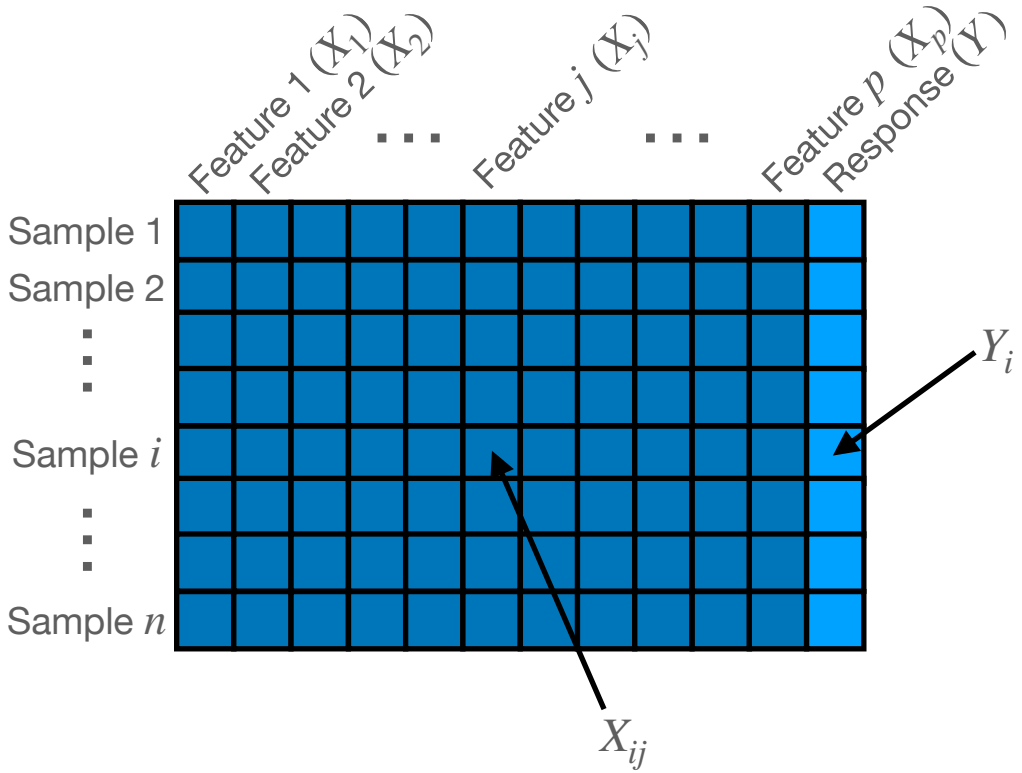
- Continuous: Response takes real values (regression problem).
- Categorical: Response takes discrete values, e.g. binary (classification problem).
- Structured: Response is an "object" (e.g. a sentence or an image).

| Field | Key question |
|---|---|
| Healthcare | Clinical decision support |
| Business | Market segmentation |
| Finance | Stock market prediction |
| Technology | Machine translation |
| Science | Protein folding |
| Politics | Election forecasting |

The type of response dictates the statistical learning methodology.

# Supervised learning notation

- Samples are indexed by $i$.
- Features are indexed by $j$.
- Feature $j$ denoted by $X_j$ (in abstract); $i$th observation of feature $j$ is $X_{ij}$.
- Features collectively referred to as $X = (X_1, \ldots, X_p)$.
- All features for observation $i$ are denoted $X_i = (X_{i1}, \ldots, X_{ip})$.
- Response denoted by $Y$ (in abstract); $i$th observation of response is $Y_i$.



# Supervised learning

- **Data:** We are given training observations $(X_i, Y_i)$ for $i = 1, \ldots, n$ .
- **Learning:** We "learn from" the training data to estimate a function $\hat{f}$ that inputs $X$ and outputs a guess for $Y$, i.e. $\widehat{Y} = \hat{f}(X)$.
- **Goal:** Given separate set of $N$ test observations $(X_i^{\text{test}}, Y_i^{\text{test}})$, want to guess responses based on features as accurately as possible, i.e. $Y_i^{\text{test}} \approx \hat{f}(X_i^{\text{test}})$.

How to measure success? Evaluation metric depends on response type. For continuous responses, typically use the mean-squared error (MSE):

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (Y_i^{\text{test}} - \hat{f}(X_i^{\text{test}}))^2.$$

# Two objectives of supervised learning

- Goal 1: Prediction. You want to predict an unknown outcome (e.g. in the future, or hard to measure) based on a set of features available to you.
- Goal 2: Insight. You want insight into the true relationship between a set of features and a response.

  These are related, but distinct goals.

  Consider: trying to earn money by predicting tomorrow's stock price versus trying to learn what variables impact stock price.

  In a given application, it is good to know which of these two goals is most important.

| Field | Key question |
|---|---|
| Healthcare | Clinical decision support |
| Finance | Stock market prediction |
| Technology | Machine translation |
| Science | Protein folding |
| Politics | Election forecasting |

# On insight

Consider a continuous response $Y$, which is related to features $X$ via

$$Y = f(X) + \epsilon, \text{ where } \mathbb{E}[\epsilon] = 0 \text{ and } \text{Var}[\epsilon] = \sigma^2 .$$

Insight into relationship between $Y$ and $X \iff$ learning about $f$.

MSE is connected to how well $\hat{f}$ approximates $f$:

$$\begin{aligned}\mathbb{E}[\text{MSE}] &= \mathbb{E}[(Y^{\text{test}} - \hat{f}(X^{\text{test}}))] \\ &= \mathbb{E}[(f(X^{\text{test}}) + \epsilon - \hat{f}(X^{\text{test}}))] = \mathbb{E}[(f(X^{\text{test}}) - \hat{f}(X^{\text{test}}))] + \sigma^2 .\end{aligned}$$

Therefore, best predictions $\iff$ closest approximation to $f$.

# On insight (continued)

The estimate $\hat{f}$ is only a noisy approximation to $f$, so the insights based on $\hat{f}$ about $f$ may or may not be reliable.

How do we quantify the uncertainty in $\hat{f}$? How do we give precise probabilistic guarantees about $f$ based on limited data?

This is the domain of statistical inference (e.g. STAT 431), and is beyond the scope of this course.

In STAT 471, we will still examine the learned function $\hat{f}$ for insights, but we must take these with a grain of salt.

# Data mining in practice

Modeling is just one part of data mining; several other skills are needed.

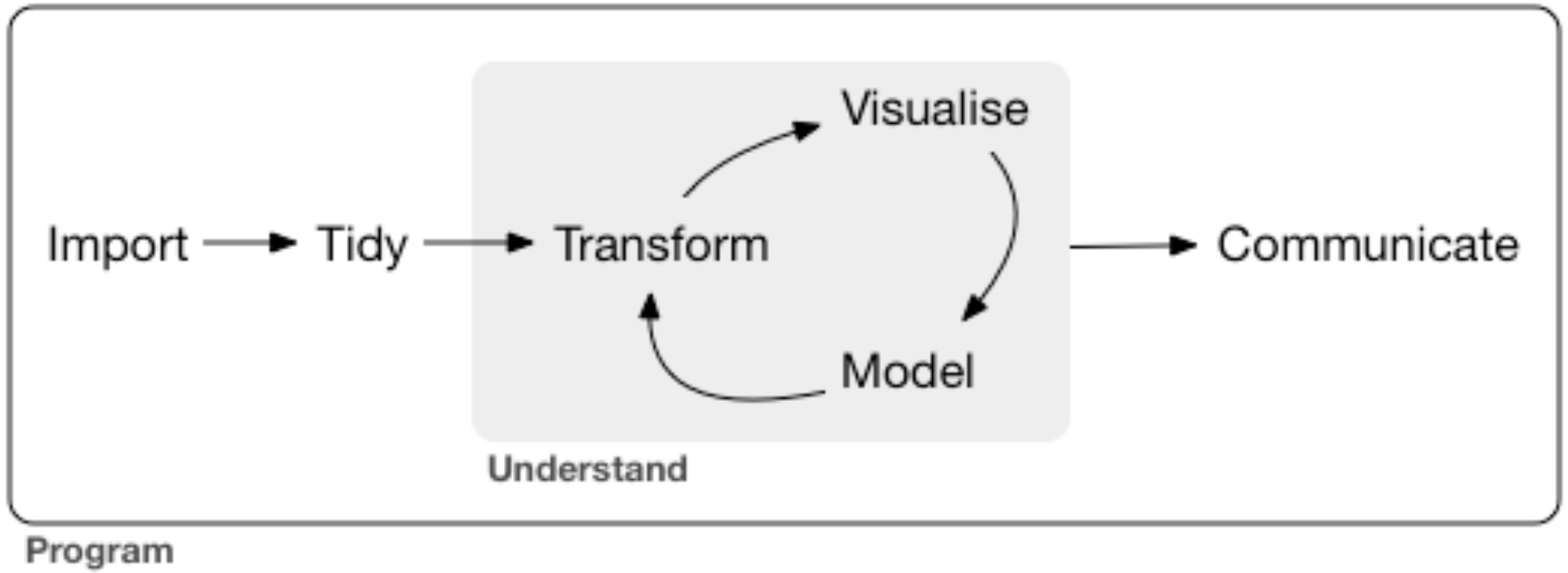



Image source: R for Data Science

# Tools of the trade





# Looking ahead

**Unit 1:** Intro to modern data mining

**Unit 2:** Tuning predictive models

**Unit 3:** Regression-based methods

**Unit 4:** Tree-based methods

**Unit 5:** Deep learning

**Lecture 1:** Intro to modern data mining

**Lecture 2:** Data wrangling

**Lecture 3:** Exploratory data analysis

**Lecture 4:** Linear regression

**Lecture 5:** Unit review and quiz in class

Homework 1 due the following day.