Quantitative epistemology

Readings for today

- Dretske, F. I. (1983). Précis of Knowledge and the Flow of Information. Behavioral and Brain Sciences, 6(1), 55-63.
- Vlastelica M. (2019). Learning Theory: Empirical Risk Minimization. Towards Data Science.

Topics

1. What is data science?

2. Information flow & knowledge

3. Data science as epistemology

4. Class overview

What is data science?

The story of data

$$x_i$$

The story of data

$$y_i \leftarrow x_i$$

The story of data

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \leftarrow \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ x_{2,1} & \cdots & x_{2,p} \\ \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$$

The story of data

$$Y \leftarrow X$$

The story of data

$$Y = f(X)$$

The story of data

Truth

Experience

Concept Class: A set of true function f that describe the structure of X (and its relationship to Y)

$$Y = f(X)$$

The story of data

Truth

Concept Class: A set of true function f that describe the structure of X (and its relationship to Y)

$$Y = h(X)$$

Experience

Hypothesis Class: A set of candidate functions h that describe the structure of X (and its relationship to Y)

The story of data

Truth

Concept Class: A set of true function f that describe the structure of X (and its relationship to Y)

$$Y = h(X) \rightarrow f(X)$$

Experience

Hypothesis Class: A set of candidate functions h that describe the structure of X (and its relationship to Y)

What is data science?

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.

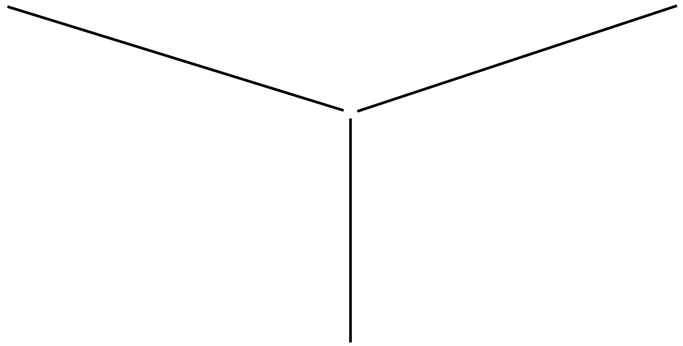
https://en.wikipedia.org/wiki/Data_science

What can I know from my data?

What is data science?

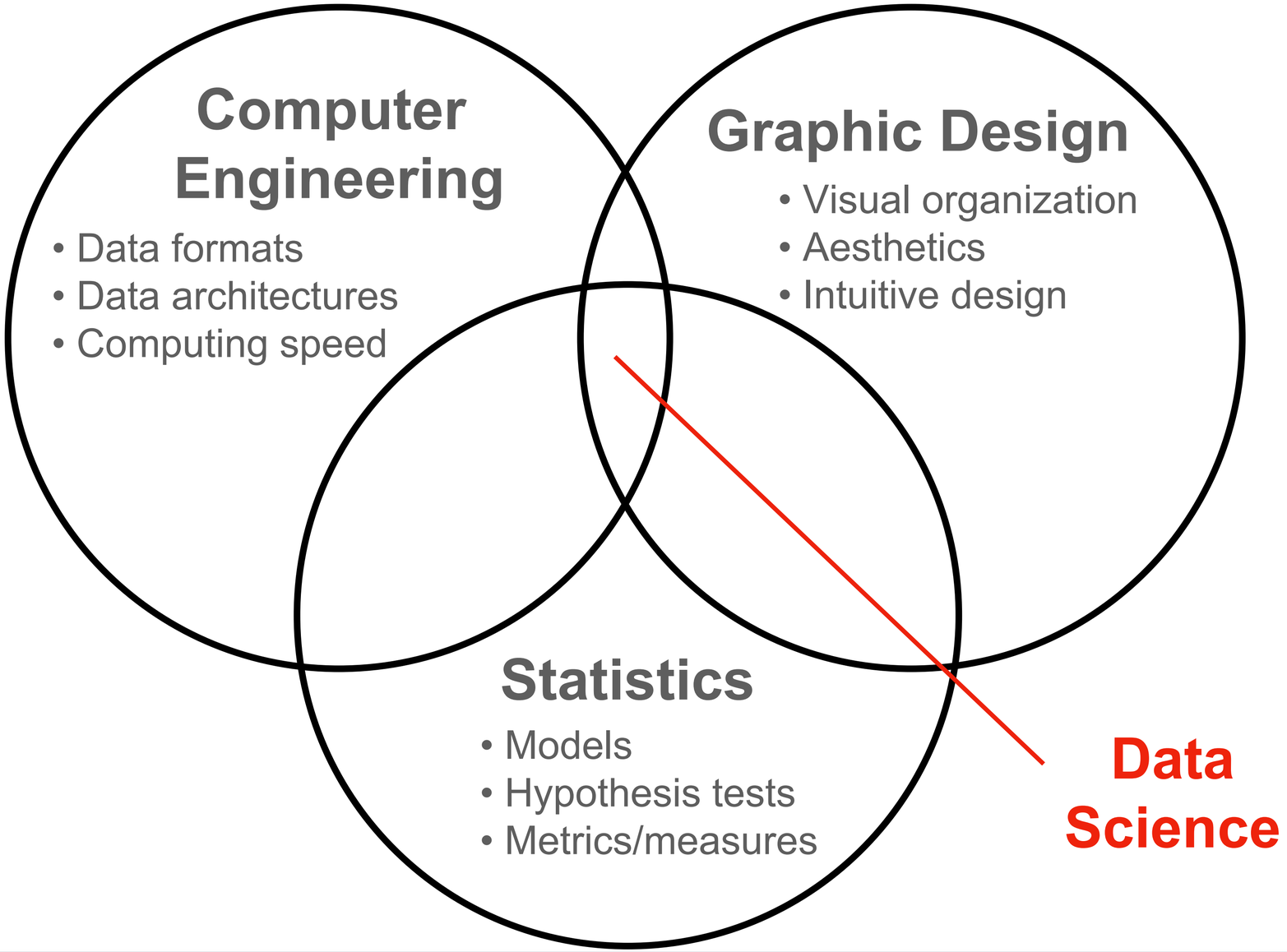
Engineering:
To build

Art & Design:
To communicate

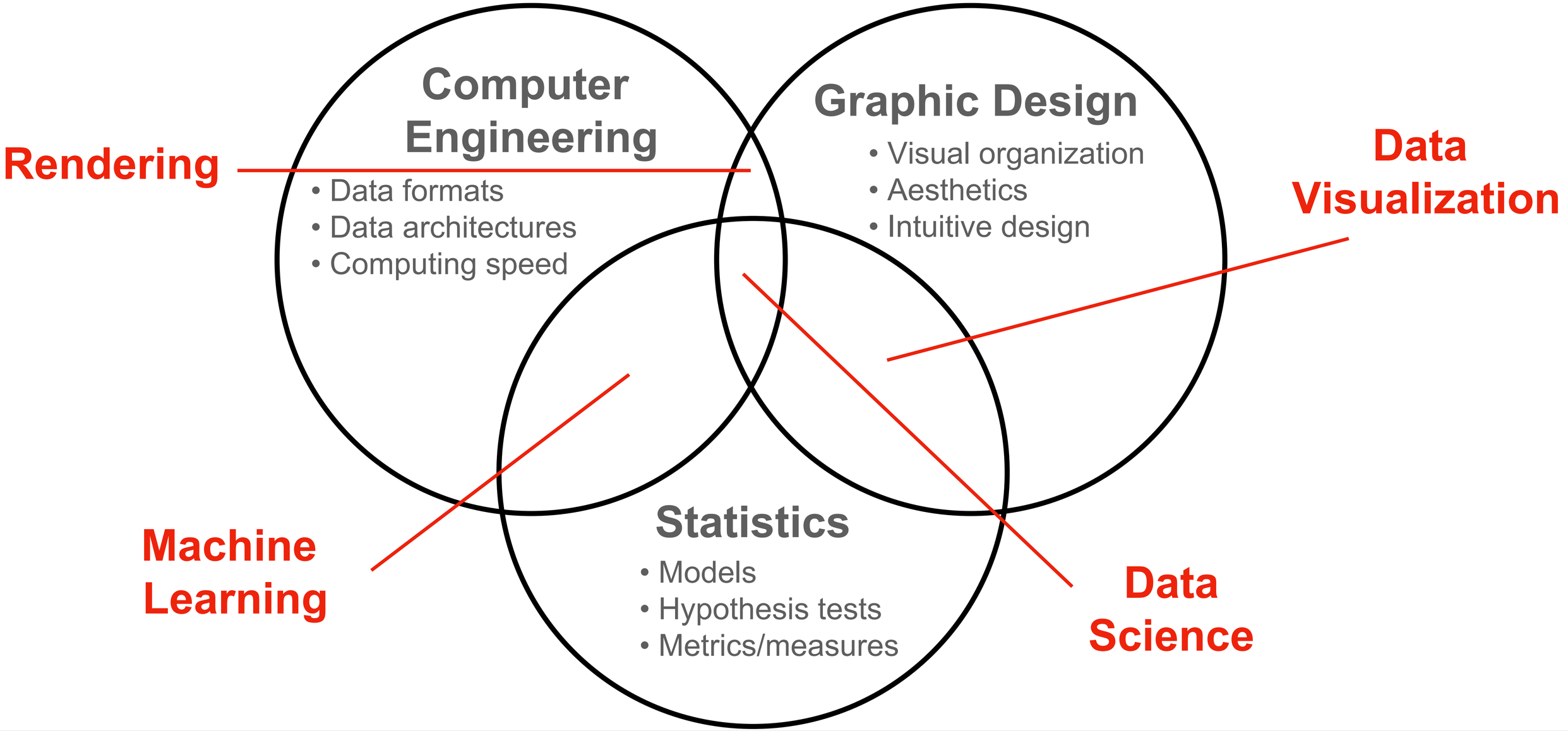


Science:
To understand

What is data science?



What is data science?



Data science as epistemology

Information → Knowledge → Understanding

Information How do we learn the structure embedded in our data?

Knowledge How does the structure in our data predict observations?

Understanding How does our knowledge generalize to new contexts?

Risk

$$R(h) = \ell(h(X), \hat{y})$$

Continuous

$$\Sigma(\hat{y} - y)^2$$

Categorical

$$I(\hat{y} = y)$$

(Vlastelica 2019; Vapnik 1991)

Empirical risk minimization

Expected Risk

$$E_{\text{risk}}\left(h, n, P\right) = \underbrace{\int_{(\mathbf{X},\mathbf{Y})}}_{\text{train}} \underbrace{R(h)}_{\text{risk}} \underbrace{dP_{(X,Y)}}_{\text{distribution}}$$

(Vlastelica 2019; Vapnik 1991)

Empirical risk minimization

Expected Risk

$$E_{\text{risk}}\left(h, n, P\right) = \underbrace{\int_{(\mathbf{X},\mathbf{Y})}}_{\text{train}} \underbrace{R(h)}_{\text{risk}} \underbrace{dP_{(X,Y)}}_{\text{distribution}}$$

$$= \underbrace{\int_{(\mathbf{X},\mathbf{Y})}}_{\text{test}} \underbrace{\int_{(\mathbf{X},\mathbf{Y})}}_{\text{train}} \underbrace{R(h)}_{\text{risk}} \underbrace{dP_{X,Y}}_{\text{distribution}} \underbrace{dP_{(X,Y)}}_{\text{distribution}}$$

Assumption: Both the training and test data come from the same distribution.

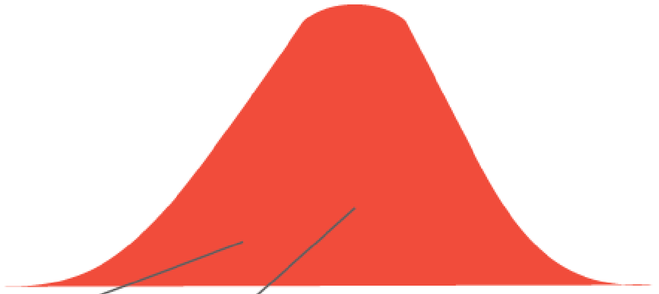
(Vlastelica 2019; Vapnik 1991)

Empirical risk minimization

Expected Risk

$$E_{\text{risk}}(h, n, P) = \underbrace{\int_{(\mathbf{X},\mathbf{Y})}}_{\text{train}} \underbrace{R(h)}_{\text{risk}} \underbrace{dP_{(X,Y)}}_{\text{distribution}}$$

$$= \underbrace{\int_{(\mathbf{X},\mathbf{Y})}}_{\text{test}} \underbrace{\int_{(\mathbf{X},\mathbf{Y})}}_{\text{train}} \underbrace{R(h)}_{\text{risk}} \underbrace{dP_{X,Y}}_{\text{distribution}} \underbrace{dP_{(X,Y)}}_{\text{distribution}}$$



Assumption: Both the training and test data come from the same distribution.

(Vlastelica 2019; Vapnik 1991)

Information → Knowledge → Understanding

Information

$$E_{risk}(h, n, P) = \int_{(\mathbf{X},\mathbf{Y})} R(h) \quad dP_{X,Y}$$

$\widetilde{train}$ $\widetilde{risk}$ $\widetilde{distribution}$

Knowledge

$$E_{risk}(h, n, P) = \int_{(\mathbf{X},\mathbf{Y})} \int_{(\mathbf{X},\mathbf{Y})} R(h) \quad dP_{X,Y} \quad dP_{(X,Y)}$$

$\widetilde{test}$ $\widetilde{train}$ $\widetilde{risk}$ $\widetilde{distribution}$ $\widetilde{distribution}$

Understanding

$$E_{risk}(h, n, P) = \int_{(\mathbf{X},\mathbf{Y})_n} \int_{(\mathbf{X},\mathbf{Y})} R(h) \quad dP_{X,Y} \quad dP_{(X,Y)_n}$$

$\widetilde{new}$ $\widetilde{train}$ $\widetilde{risk}$ $\widetilde{distribution}$ $\widetilde{distribution}$

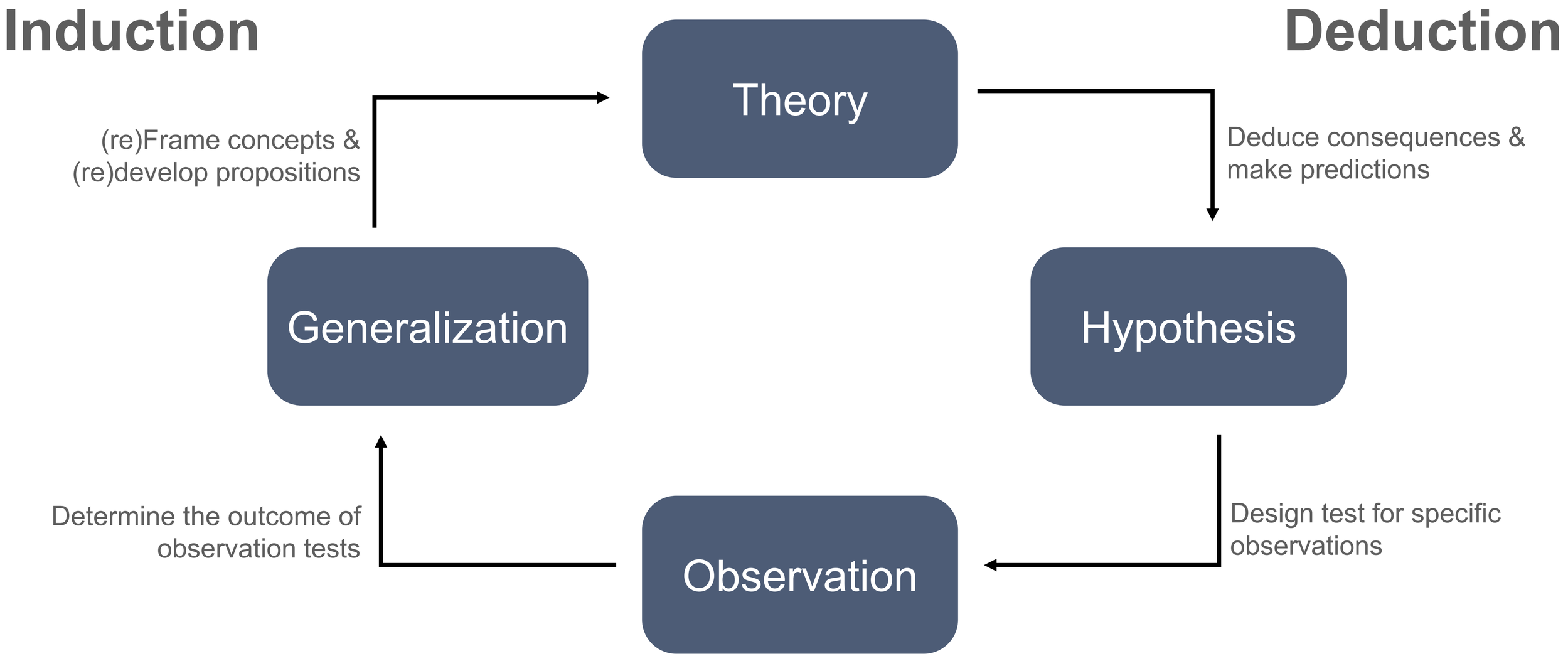
Information → Knowledge → Understanding

Information How do we learn the structure embedded in our data?

Knowledge How does the structure in our data predict observations?

Understanding How does our knowledge generalize to new contexts?

Hypothetico-deductive model of science



Wallace, W. L. (1971). *The Logic of Science in Sociology*.

(Hempel 1966; Popper 1935, 1963)

Class overview

Goal of the class

Show how data science approaches can be useful for revealing information and knowledge from observational data.

Learning objectives for this class

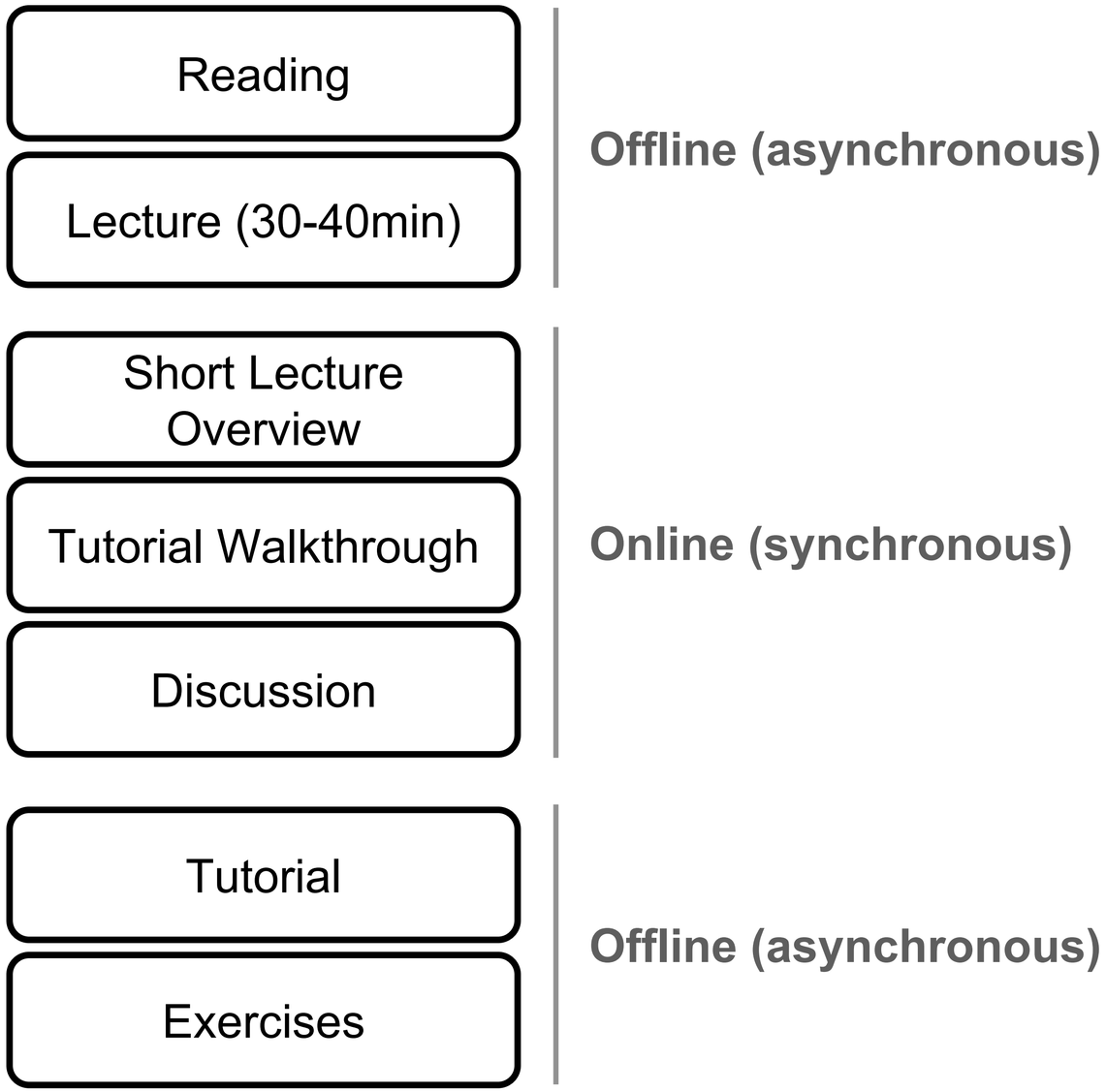
1. Understand basic principles of statistical theory, measurement, and experimental design;
2. Be able to clean and organize data effectively;
3. Be well versed the execution and interpretation of data analysis;
4. Use information resources to find appropriate data science tools;
5. Communicate statistical results effectively in multiple modalities;
6. Be a critical consumer of data science techniques and their application in empirical research.

Prior knowledge

1. Introductory level understanding of probability theory and statistics (CMU 36-309, 86-309, or equivalent)

2. Basic familiarity with R or similar functional data analysis languages.

Class structure

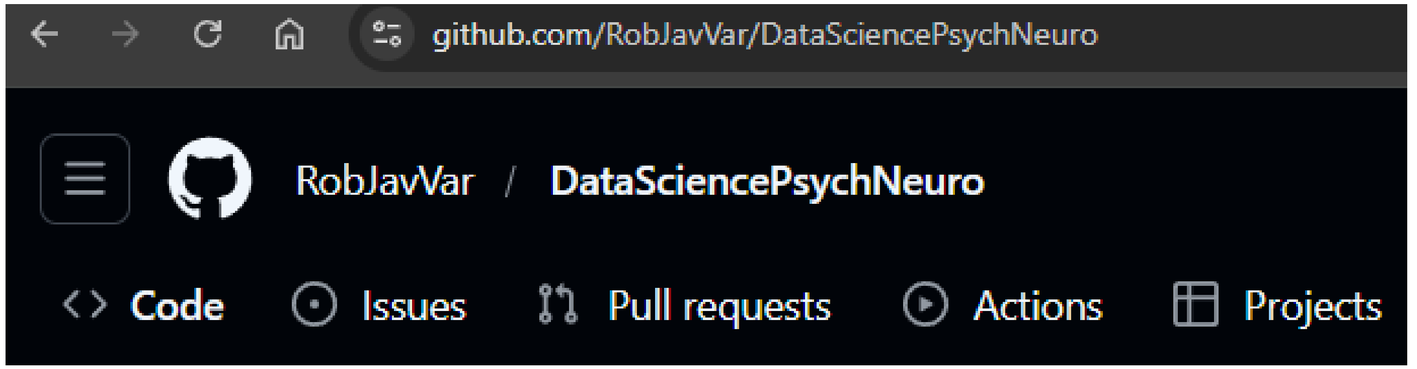


Goal:

- Content knowledge (crystalized) prior to class.
- Dynamic discussion (fluid) during class.

Resources

1. Texts:
   - Jupyter Book: Data Explorations (https://robjavvar.github.io/DSPN_CourseNotebook/intro.html)
   - Textbook: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R (2nd edition). New York: springer. (http://www.statlearning.com).
   - Auxiliary readings will be posted on Canvas/Github for class sections covering material not in the main textbook.
2. Github Repository: https://github.com/RobJavVar/DataSciencePsychNeuro



Take home message

- Data science can be seen as a branch of epistemology revealing how meaning and knowledge can be determined from information.
- These approaches fit into a larger process of scientific discovery that links abstract theories to empirical data.