Introduction to Web Science

Tutorial (Assignment 6)

Olga Zagovora

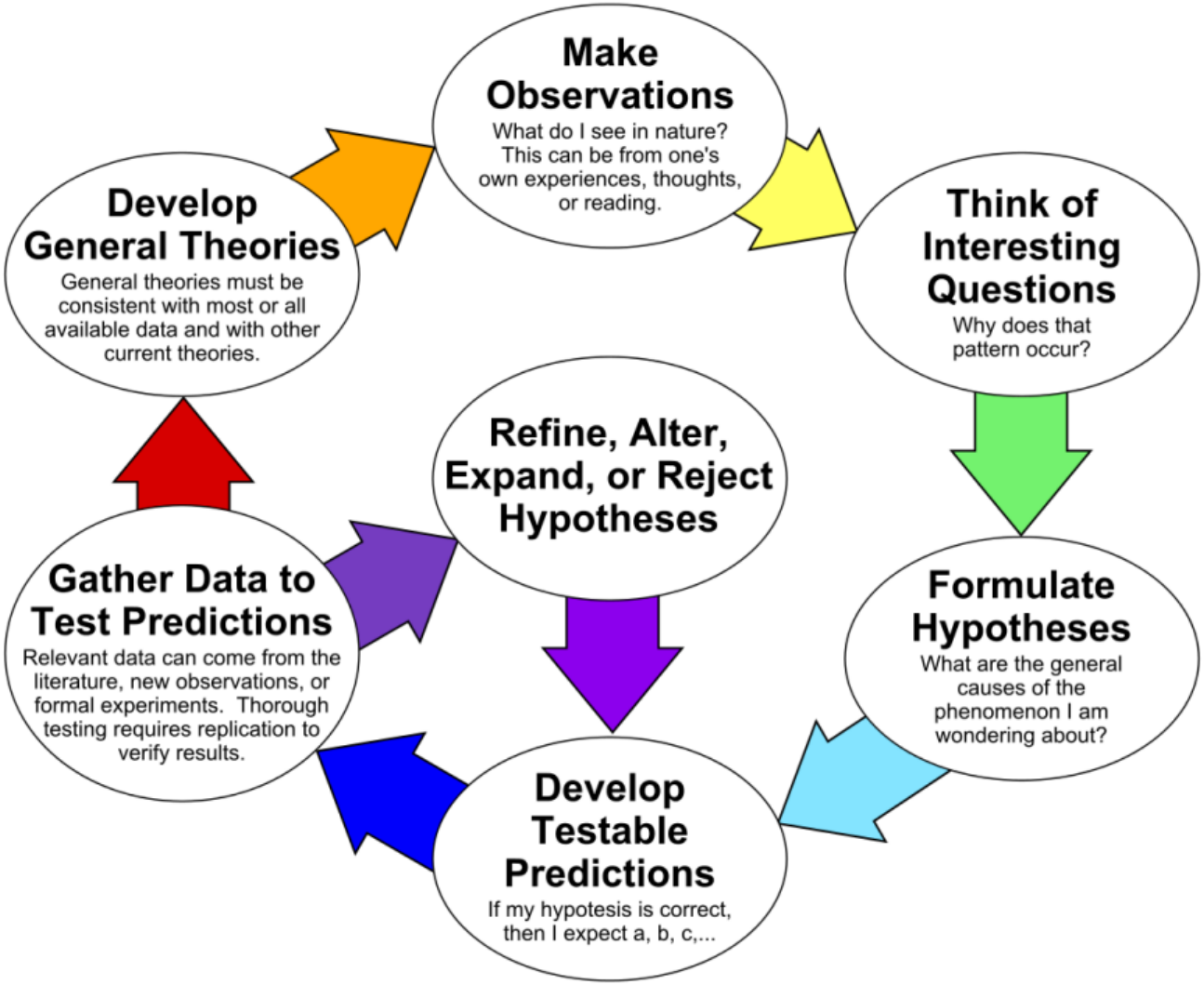
WeST
People and Knowledge Networks

Exercise 1

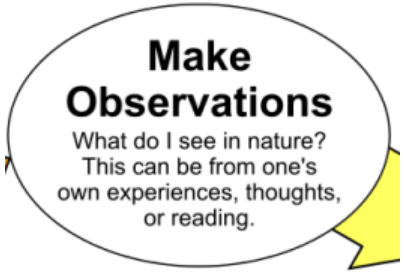
-> Blackboard

Exercise 2

The Scientific Method as an Ongoing Process



Exercise 2. Step 1:



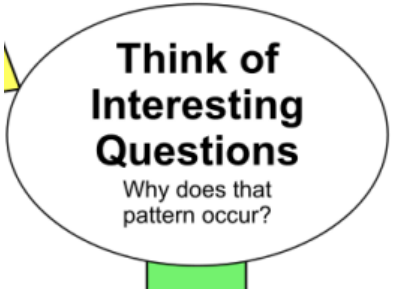
Example:

QUEBEC

4 out of 5 SEW articles have less than 500 words.

4 out of 5 SEW articles have less than 30 sentences.

Exercise 2. Step 2:

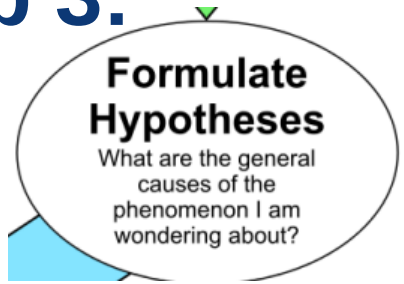


Example:

QUEBEC

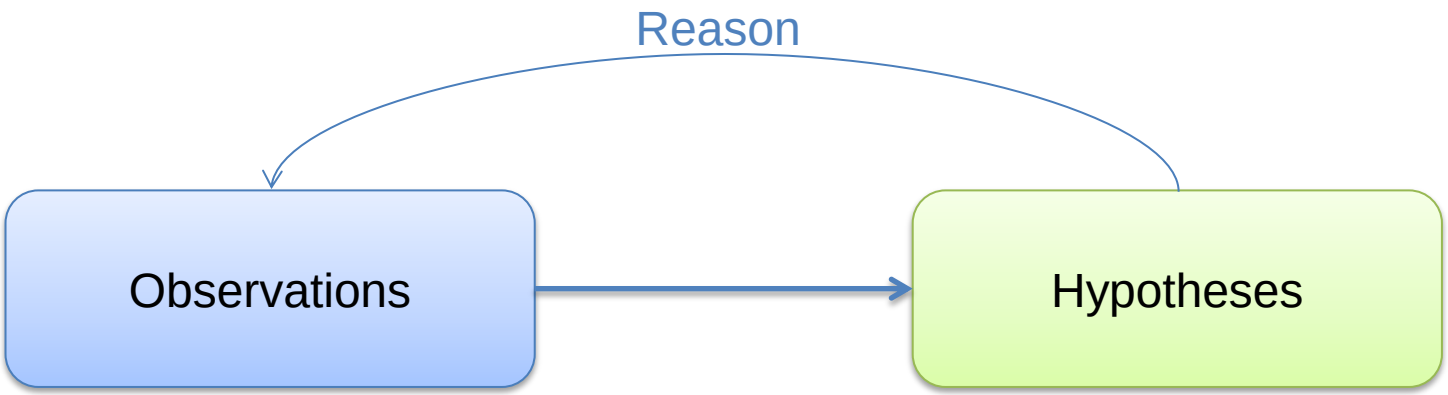
The third observation makes us the most curious, because this pattern could point that articles are short and could be read fast. Are the articles of SEW meant to be short?

Exercise 2. Step 3:

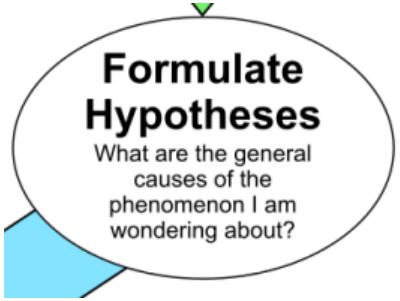


Hypothesis:

- clear and short
- concise
- to the point



Exercise 2. Step 3:



Example:

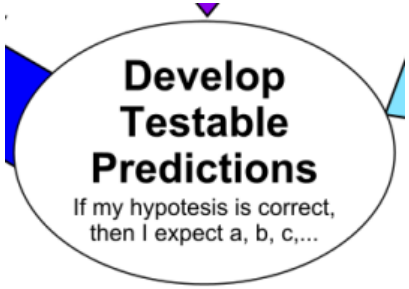
QUEBEC

80% of articles on SEW are short.

We define every article with under 30 sentences as short.

80% of articles on SEW are shorter than 30 sentences.

Exercise 2. Step 4:



Example:

We will count number of sentences in each article of SEW. Then we will estimate percentage of articles with less than 30 sentences.
If the hypothesis is correct, we expect to encounter about 80% of all articles with less than 30 sentences.

Exercise 2. Step 5:

-> Exercise 3

5. Explain how you would like to use the data set to test the prediction by means of descriptive statistics. Also explain how you would expect your outcome.

Example:

We will count number of sentences in each article of SEW using corresponding python library (nltk). Then we will estimate percentage of articles with less than 30 sentences.

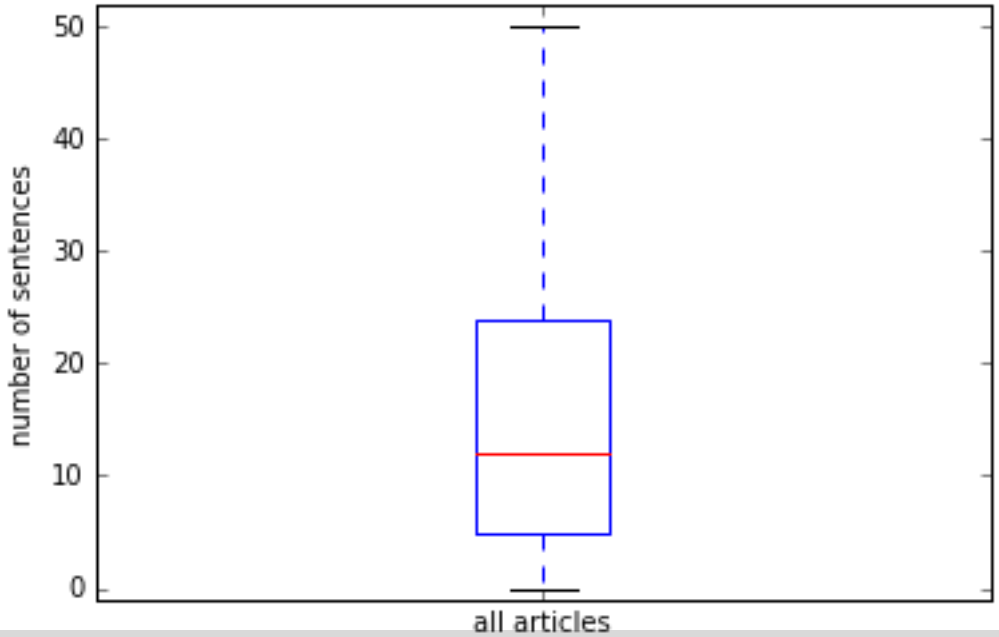
We use boxplot in order to visualize distribution of sentences in articles .

Exercise 3

Example:

Percentage of articles with less than 30 sentences:

$$\frac{numberOfArticlesWithLessthan30sentences}{numberOfArticles} = 0.82$$



Multithreading in ipython

Demo -> ipython notebook

Questions?


zagovora@uni-koblenz.de