Lesson2:
Modelling the Web with Simple Statistical Descriptive Text Models
Unit3:
Formulating a research hypothesis and finding evidence for it

Rene Pickhardt

Introduction to Web Science Part 2

Emerging Web Properties

WeST
People and Knowledge Networks

Institute for Web Science and Technologies · University of Koblenz-Landau, Germany

Completing this unit you should

- Understand the ongoing, cyclic process of research
- Know what falsifiable means and why every research hypothesis needs to be falsifiable
- Be able to formulate your own research hypothesis

First: Start with an observation

- There is English Wikipedia
- There is Simple English
- The purpose of Simple English Wikipedia is to be easier to understand and therefore more accessible than English Wikipedia

Second: Be critical and curious

- The purpose of Simple English Wikipedia is to be easier to understand and therefor more accessible than English Wikipedia

- Ask yourself: Is this really true?
  - Of course, the purpose is true

- But what about the goal?
  - Is it achieved?
  - Is it really easier to understand?

Third: Transform your question and observations into an hypothesis

- Research - Hypothesis:

  Simple English Wikipedia is easier to understand than English Wikipedia!

Some thoughts on scientific methodology

- Recall our Research – Hypothesis:
  Simple English Wikipedia is easier to understand than English Wikipedia!

- This hypothesis is **falsifiable**

- Once we find a hint why this hypothesis is not true it is falsified

- Every sound research hypothesis has this property of being falsifiable

- C.f. Karl Popper

Fourth: Develop Testable Predictions

- This is most probably the point where modeling comes into play

Testable Predictions:

- Less words are needed to understand a larger fraction of Simple English Wikipedia than English Wikipedia
  - This is a simple counting exercise

- Overall the sentences in Simple English Wikipedia are shorter and use shorter words than the ones in English Wikipedia
  - Another simple counting exercise

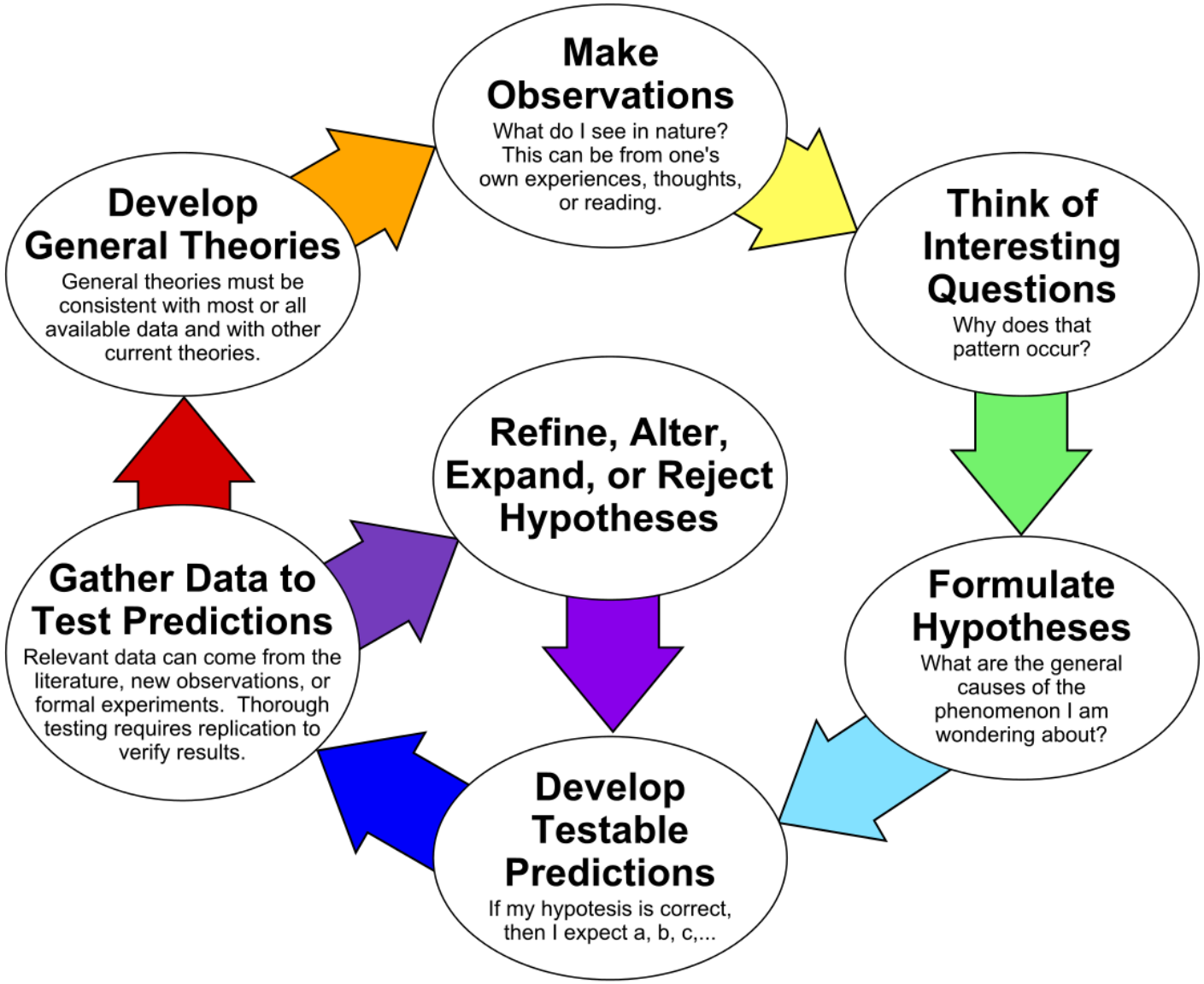
Fifth: Gather data to test predictions

- Often very difficult for the following reasons:
- Data might be in “silos” if private companies own it
  - Interesting research questions could be answered on Facebook data but it is not accessible
- Data needs to be created by asking people
  - To participate in a user study
  - Fill out questionnaires
- One of the reasons we work with Wikipedia
  - The data is available and open
  - It is just an awesome playground for research
  - It is limited since it is not used by everybody

Now we probably have to make some choice

- Either
  - Refine alter expend or reject the hypothesis
  - Go back to step 3 / 4
- Or
  - Go forward in trying to develop a general theory
  - It must be consistent with other theories and all available data
  - Often you make new observations and start over at step 1

The Scientific Method as an Ongoing Process



Roadmap for the next two units

- Analyze each of our two testable predictions
- Check if less words are needed to understand a larger fraction of Simple English Wikipedia
- See if sentences and words are really shorter
- Interpret the results and discuss them critically

Thank you for your attention!

Contact:

Rene Pickhardt
Institute for Web Science and Technologies
Universität Koblenz-Landau
rpickhardt@uni-koblenz.de

WeST
People and Knowledge Networks

Copyright: