UT Southwestern Medical Center
Lyda Hill Department of Bioinformatics

BioHPC

Machine Learning using R

[web] portal.biohpc.swmed.edu
[email] biohpc-help@utsouthwestern.edu

2021-08-18

Outline

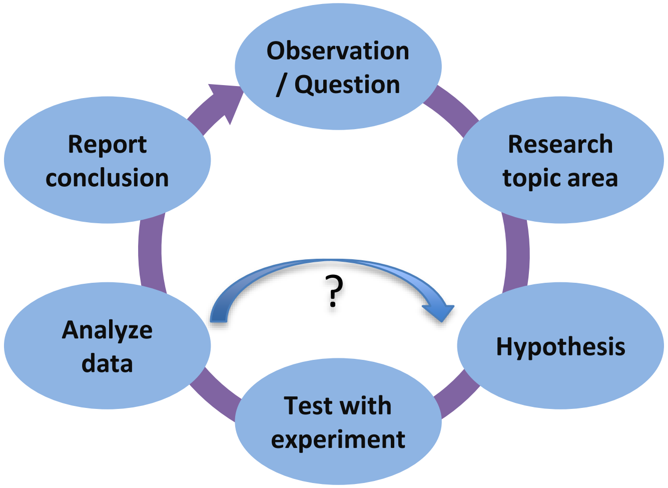
A primer on ML

- ML for hypothesis search & discovery

Hand-on in R

- Assessing available data
- Picking algorithms/model types
  - Descriptive / Predictive models
- Pre-processing of data
- Cross validation
- Assessing results

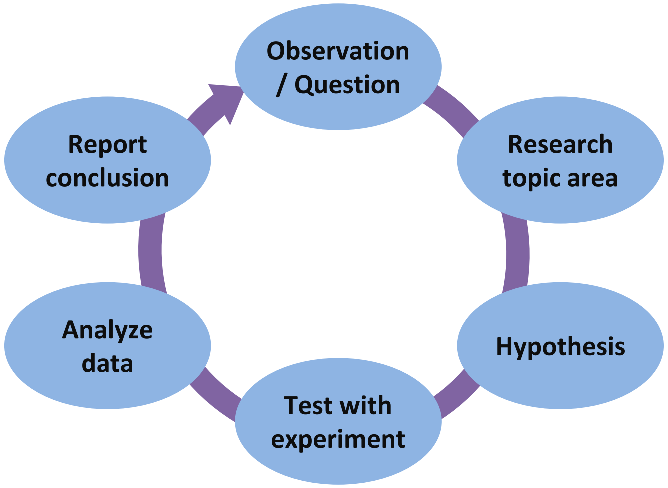
ML for hypothesis search & discovery



The Scientific Method

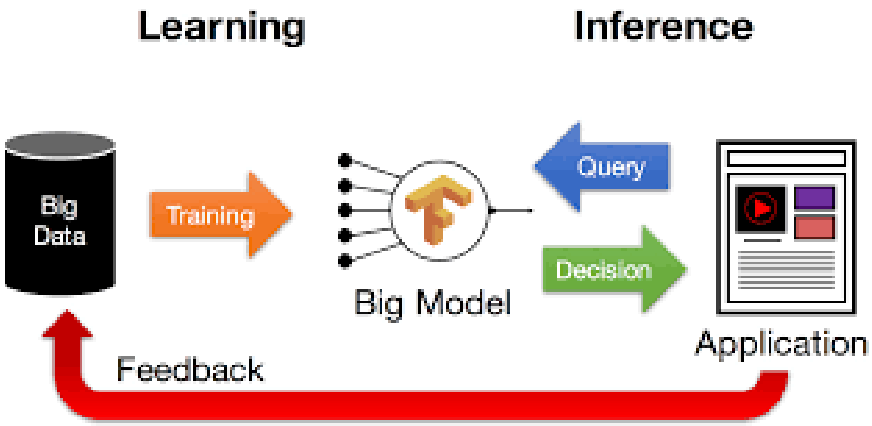
- Universal-Orthodox
  all sciences, all scientists
- Hypothesis-centric
  Focus on accepting/dismissing hypothesis
- Cyclic-open
  Fail test > re-pose question: inner cycle

ML for hypothesis search & discovery



The Scientific Method

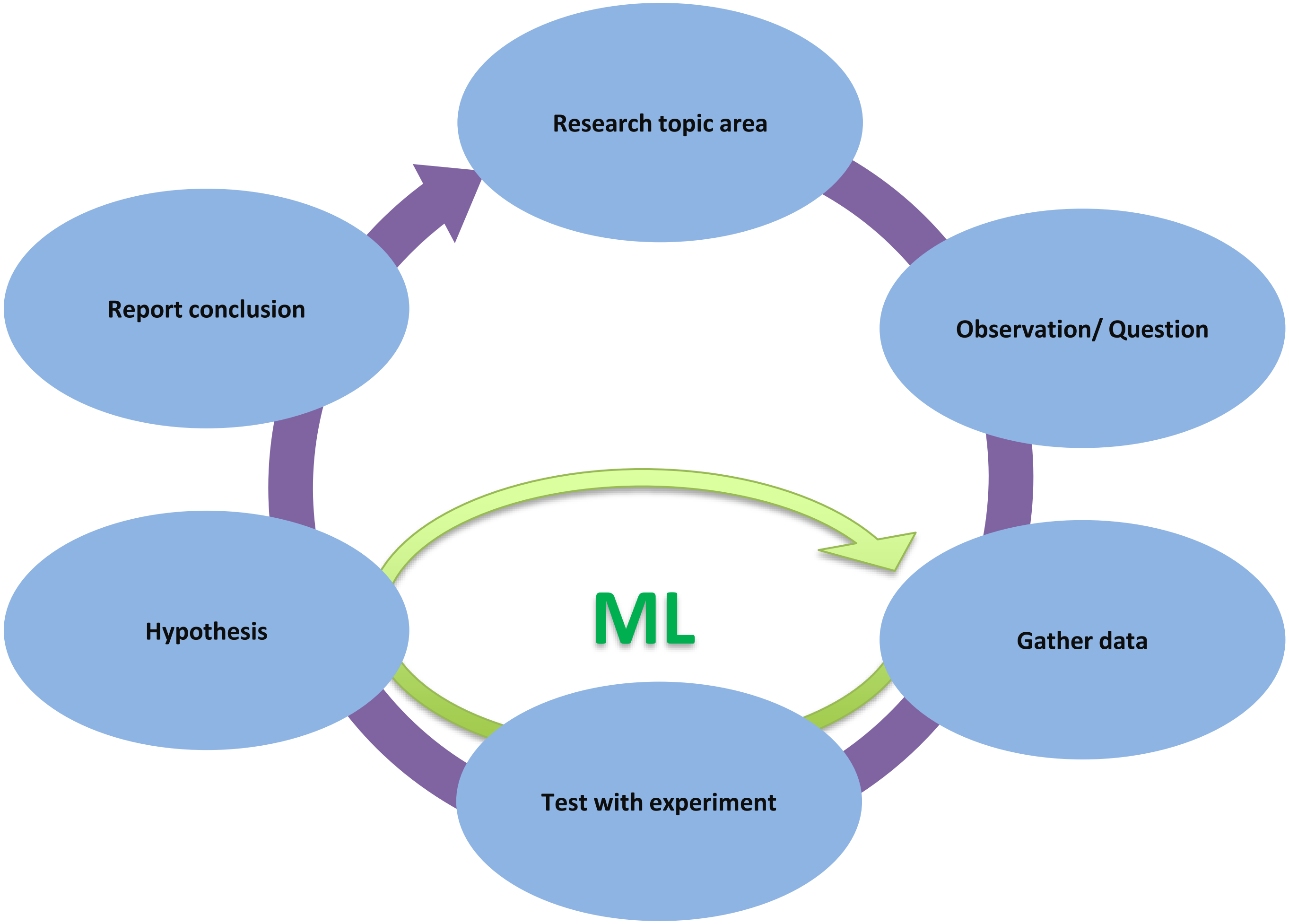
- Universal-Orthodox
  all sciences, all scientists
- Hypothesis-centric
  Focus on accepting/dismissing hypothesis
- Cyclic-open
  Fail test > re-pose question: inner cycle



The ML Method

- Domain-Unorthodox
  Results outweigh Method
- Data-centric
  Improve data rather than question the observation
- Causality/Predictability
  correct prediction ?= accept hypothesis

ML for hypothesis search & discovery



Data
anything around us

Information
data recorded for
a purpose

Knowledge
correct interpretation of
integrated information

Hands-on in R

In this tutorial we will:

- Explore some unknown dataset
- Establish a few questions we can learn the answer from the data
- Preprocess/sample/reshape data if necessary
- Run a few different ML models to answer our questions
- Assess the model results
- Present our findings in a compact format

You will need the following:

RStudio:

- Personal computer
- BioHPC workstation
- Nucleus node

Sample data:
https://archive.ics.uci.edu/ml/datasets/wine+quality

Accessing available data

Data properties

The data provided contains measurements of Portuguese vino verde.
Each row represents measurements of a specific wine label.
Measurements are values of physio-chemical properties of the wine.
Each row contains a ‘quality’ indicator as scored by tasters.
Data is available for Red and White whines.

Possible questions:

- Can we determine if wine is Red or White depending on physiochemical attributes?
- Can we determine perceived quality of the wine based on physiochemical attributes?
    - Quality as label (classification)
    - Quality as value (regression)

RStudio demo

Q&A time

Thank you