Data Science Training Workshop: Using Jupyter Notebook and R

Leslie D. McIntosh, PhD, MPH @mcintold

Connie Zabarovskaya, MITM @connie_zest

Lorinette S. Wirth, MPH @lorinettew

Funding Support

- Washington University Institute of Clinical and Translational Sciences: NIH CTSA Grant Number UL1TR000448 and UL1TR000448-09S1
- Saint Louis University Center for Health Outcomes Research (SLUCOR)

Washington University in St. Louis
School of Medicine

Saint Louis University
Center for Health Outcomes Research

Special Thanks

Michael Yingling

Statistical Data Analyst

Washington University in St. Louis

An Informatics Perspective

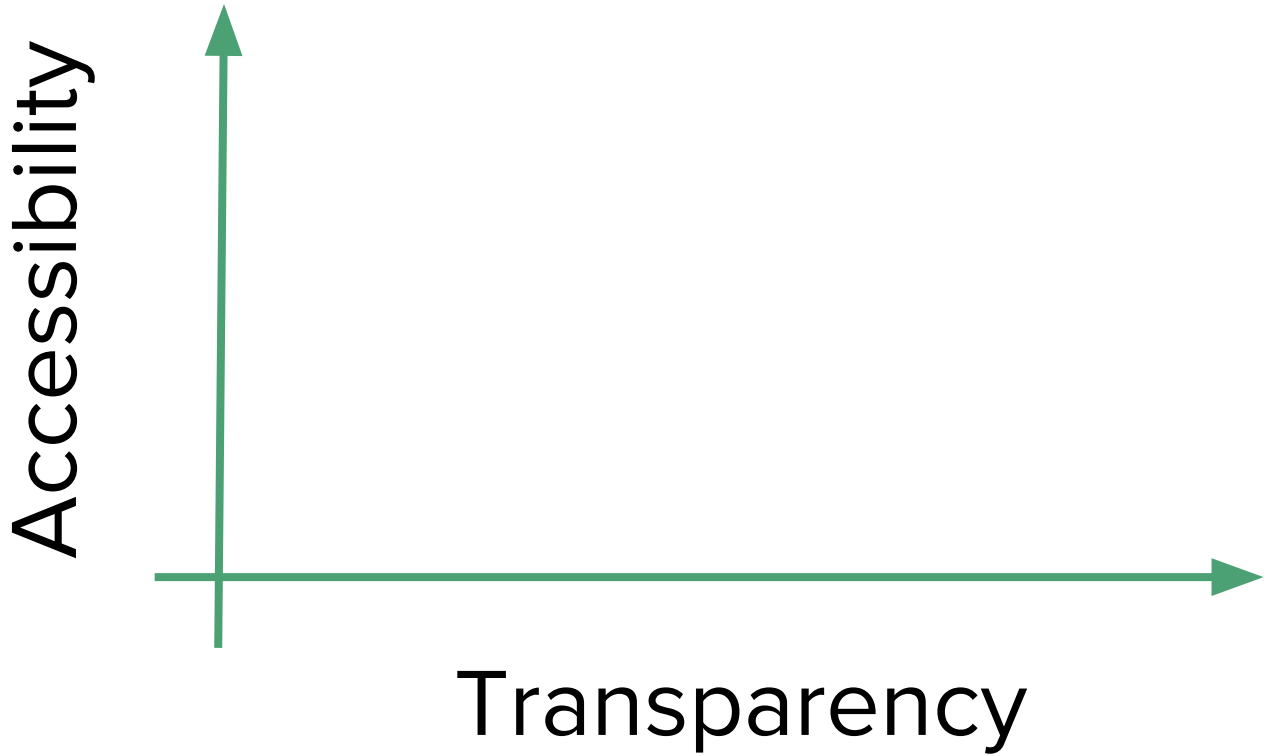
What is reproducible research?

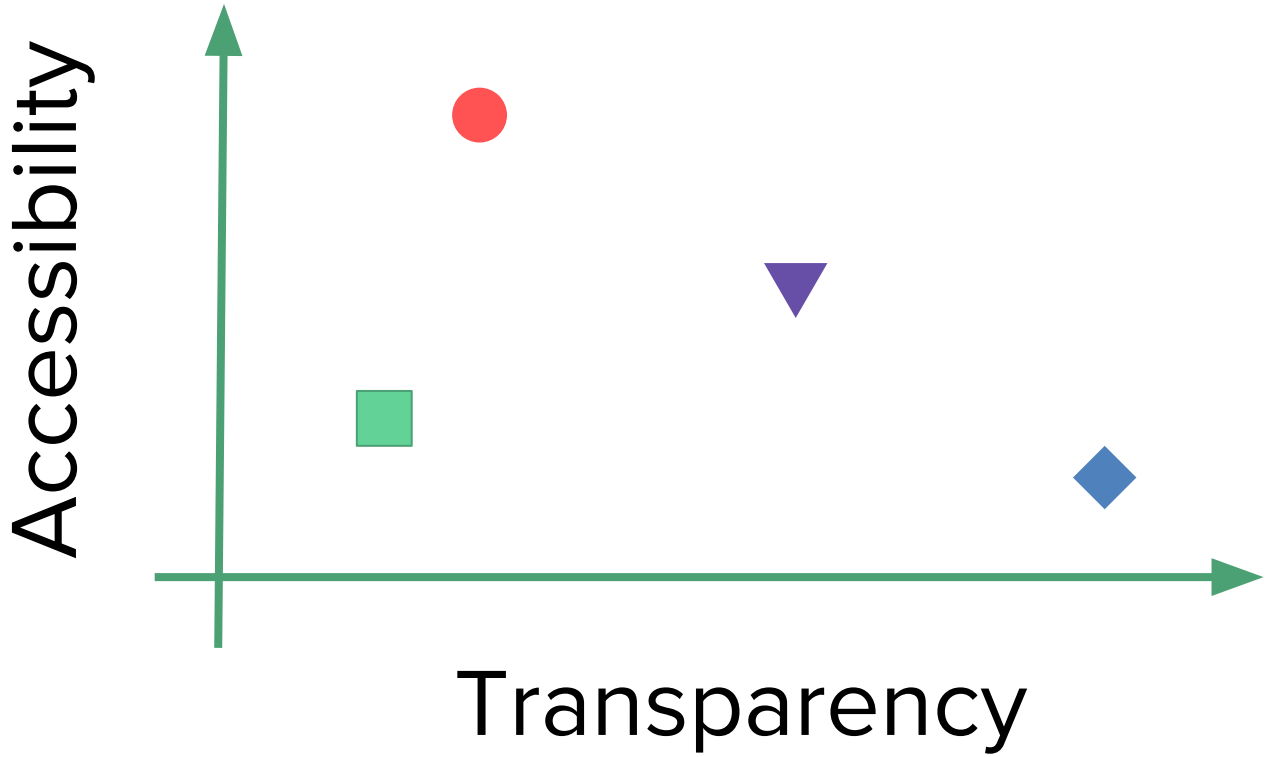
Definitions

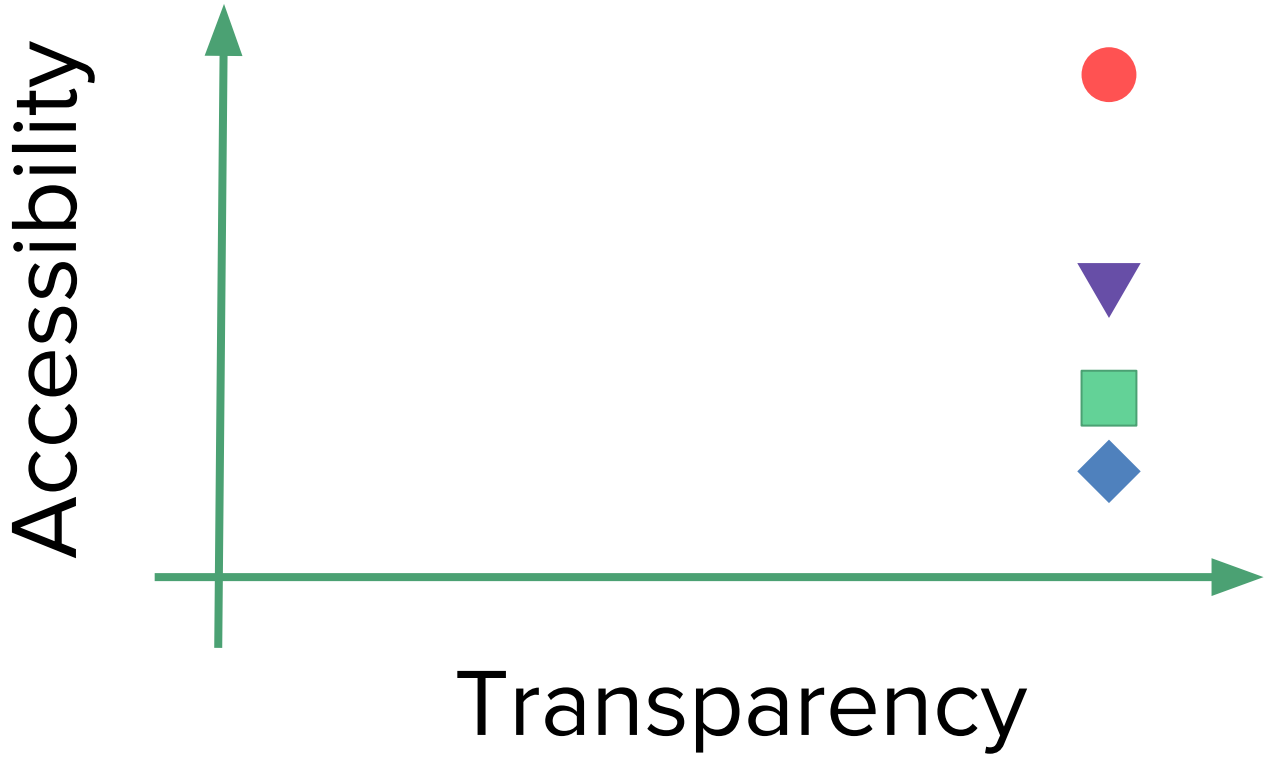
Replicable - independent people, collecting new data, and using same methods

Reproducible - independent people analyzing the same data

V. Stodden, “Trust Your Science? Open Your Data and Code,” Amstat News, 1 July 2011; http://magazine. amstat.org/blog/2011/07/01/trust-your-science/


Accessibility
Transparency


Accessibility
Transparency


Accessibility
Transparency

Why do we care?

Duke's Precision Medicine Bust

2007 - 2011+

Reproducibility of Cancer Biology

2017

https://cos.io/our-services/research/rpcb-overview/?imm_mid=0eceb8&cmp=em-data-na-na-newsltr_20170201

https://elifesciences.org/collections/reproducibility-project-cancer-biology

How does this training address any of these problems?

Center for Open Science. (2013). *Open Science Framework.* image retrieved from https://osf.io/

Notebooks in Reproducible Research

Interweave and save notes, codes, outputs, and graphics as one document

- Data Management
- Pick-up where you left off months later (R&Rs)...Effortlessly!

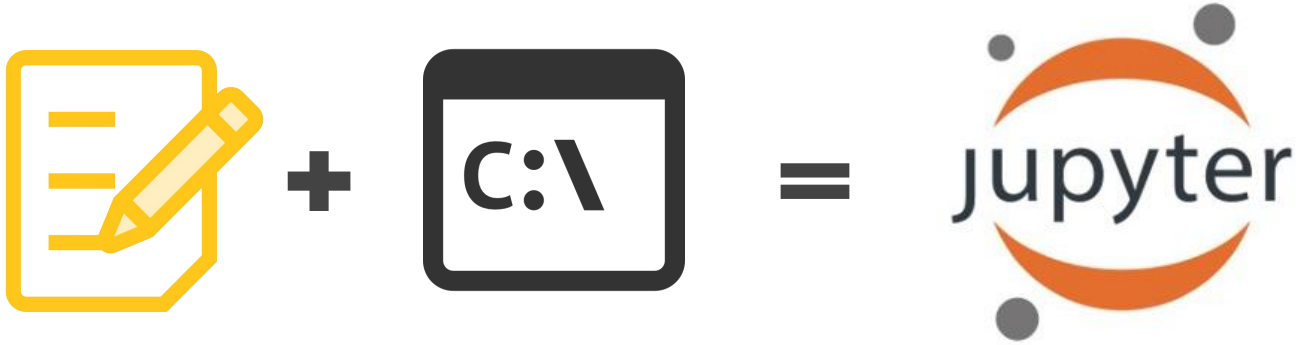
Export document as .pdf, .doc, .html with ease

- Great for consulting!
- More efficient than copying/pasting into Word

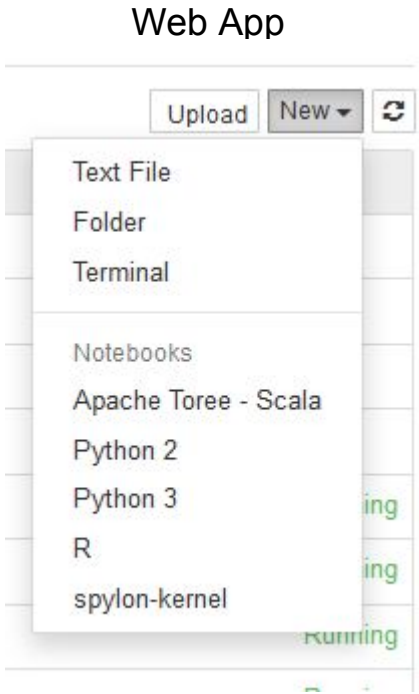
Easy to attach as supplement for manuscript publication

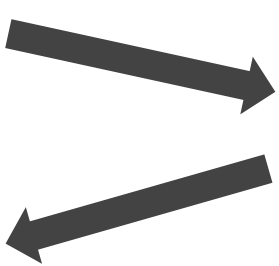
- Transparency

Jupyter Notebooks Overview



How Does it Work?

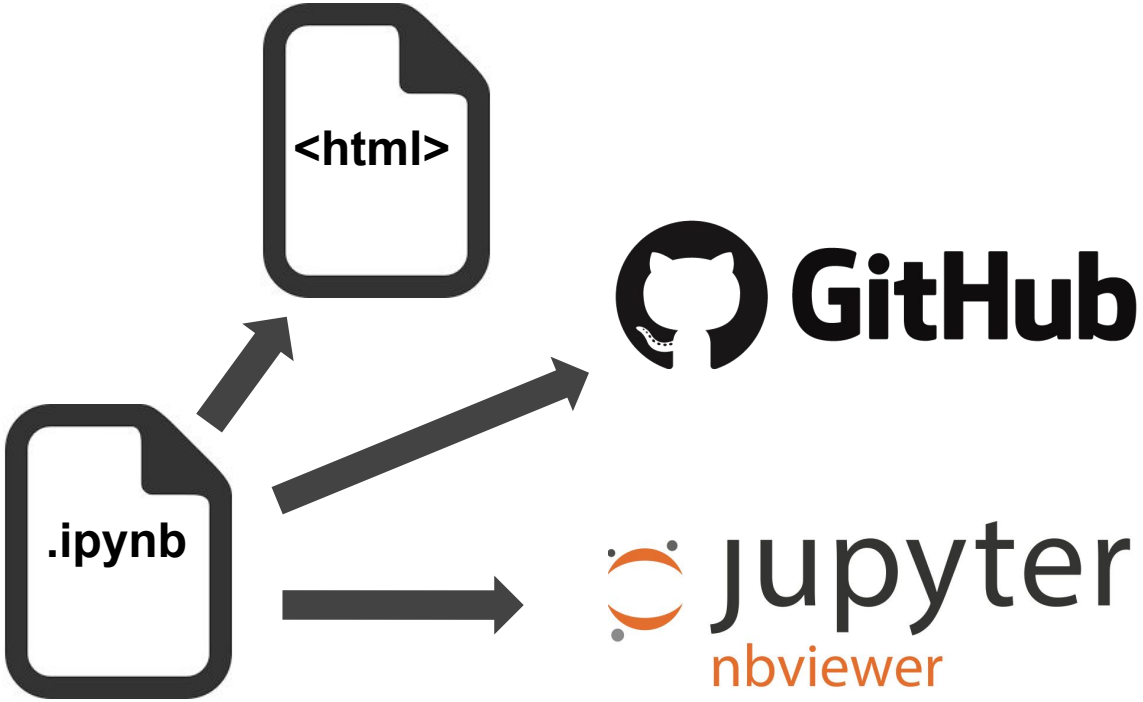




Kernel

https://github.com/jupyter/jupyter/wiki/Jupyter-kernels

How do you share Jupyter Notebooks?



How do you install Jupyter Notebooks?

Installation Options:

Install Python and R
Install Anaconda
>conda install jupyter
>conda install -c r r-irkernel=0.7.1
>conda install -c r r-essentials
>jupyter notebook

or

Install Docker
Pull Image
Start Jupyter

Case Study

Bellaachia, A., & Guven, E. (2006). Predicting Breast Cancer Survivability Using Data Mining Techniques.Society for Industrial and Applied Mathematics Conference on Data Mining 2006.

Case Study

1. Evaluate reproducibility of case study
2. Demonstrate Jupyter Notebook as tool to boost reproducibility
3. Learn Data Mining techniques

Predicting Breast Cancer Survivability Using Data Mining Techniques

Abdelghani Bellaachia, Erhan Guven

Department of Computer Science
The George Washington University
Washington DC 20052
{bell, eguven}@gwu.edu

Abstract

In this paper we present an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques. The data used is the SEER Public-Use Data. The preprocessed data set consists of 151,886 records, which have all the available 16 fields from the SEER database. We have investigated three data mining techniques: the Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. Several experiments were conducted using these algorithms. The achieved prediction performances are comparable to existing techniques. However, we found out that C4.5 algorithm has a much better performance than the other two techniques.

Keywords: Breast cancer survivability, data mining, SEER, Weka.

relationship of the association. Data driven statistical research is becoming a common complement to many scientific areas like medicine and biotechnology. This trend is becoming more and more visible as in the studies of Houston et al. [5] and Cios et al. [6].

In this paper, we present data mining techniques to predict the survivability rate of breast cancer patients. In our study, we have used the SEER data and have introduced a pre-classification approach that take into account three variables: Survival Time Recode (STR), Vital Status Recode (VSR), and Cause of Death (COD).

This paper is organized as follows. The next section reviews related work. Section 3 gives the methodology used to conduct the prediction analysis. Experimental results are presented in Section 4. Conclusion and future work are given in the last section.

Methodology (Bellaachia & Guven, 2006)

Objective: Compare data mining techniques' ability to predict breast cancer survivability

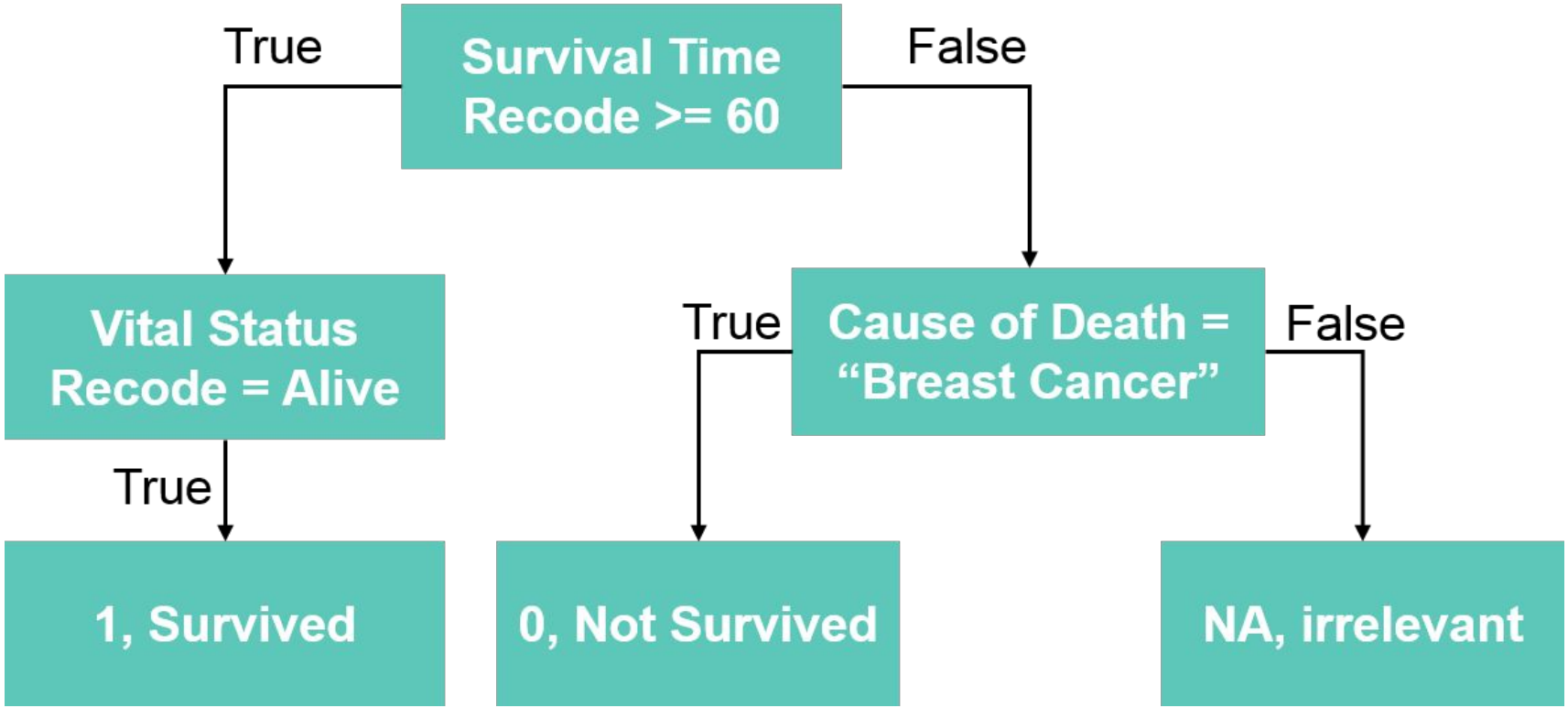
Dataset: Surveillance, Epidemiology, and End Results (SEER)

"Breast.txt" Ascii file, 1973-2002

Methodology (Bellaachia & Guven, 2006)

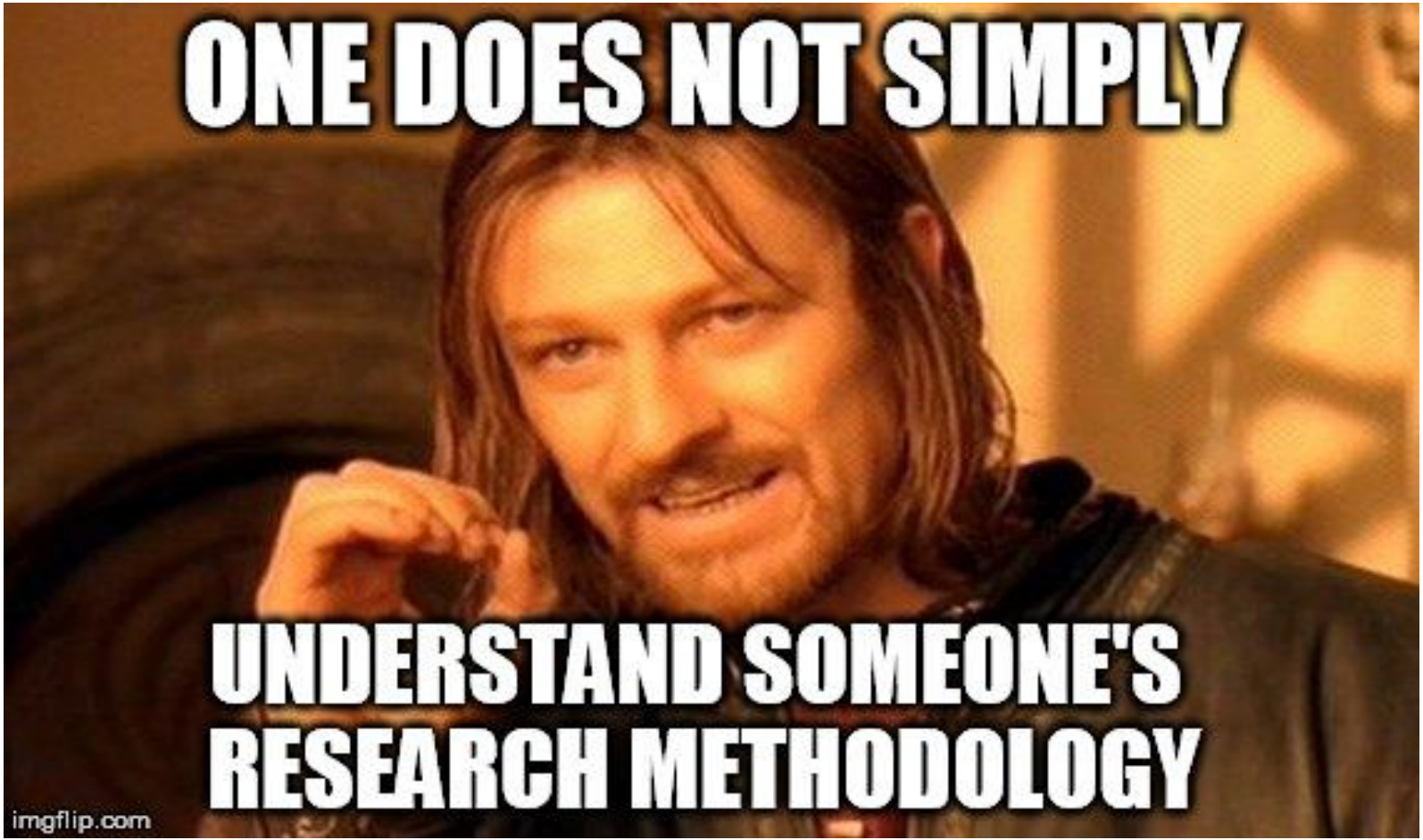
Nominal	- Race, Marital Status, Primary site code, Histologic type, Behavior code, Grade, Extension of tumor, Lymph node involvement, Site specific surgery code, Radiation, Stage of Cancer
Numeric	- Age, Tumor size, # of Positive nodes, # of Nodes, # of primaries

Outcome: Breast Cancer Survival



Methodology (Bellaachia & Guven, 2006)

- Data Cleaning
  - Extension of tumor and Site Specific Surgery fields had missing for ~50% of records
    - data gathered prior to 1988 were heavily missing
      - Removed records from the test data set
  - SEER 1998+ for Site Specific Surgery are coded differently compared to pre-1998
    - Split information across 5 fields to account for variation in coding?

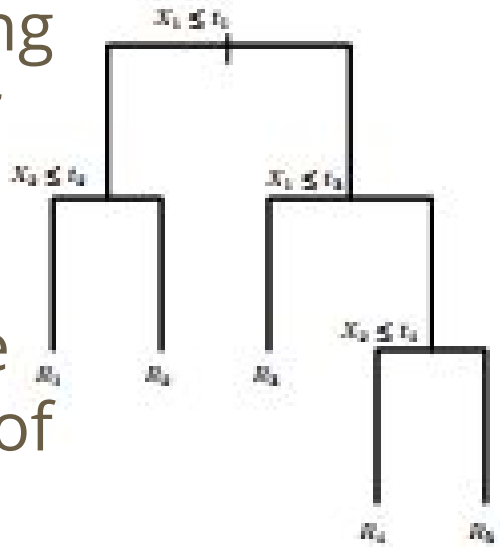

ONE DOES NOT SIMPLY
UNDERSTAND SOMEONE'S RESEARCH METHODOLOGY
imgflip.com

Methodology (Bellaachia & Guven, 2006)

- Data Mining Techniques
  - Naïve Bayes
  - Artificial neural networks
  - C4.5 decision-tree generating algorithm
  - Cross Validation; k=10 folds
  - Testing:Training Ratio ?
- Software
  - Weka toolkit

Decision Tree Algorithm

1. Use **recursive binary splitting to grow a large tree on the training data**, stopping only when each terminal node has fewer than some minimum number of observations.
2. Apply **cost complexity pruning** to the large tree in order to obtain a sequence of best subtrees, as a function of α.
   - Use **K-fold cross-validation** to choose α.



Our Methodology: Data Cleaning

- Dataset: Surveillance, Epidemiology, and End Results (SEER)
  - “Breast.txt” for 1973 - 2013
  - Filtered to 1973 - 2002

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2013), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2016, based on the November 2015 submission.

Our Methodology: Data Cleaning

- Covariates
  - Mostly same as Bellaachia & Guven, 2006
  - Combined 2 Site Specific Surgery variables into one
  - Derived binary variable for Presence of Positive Nodes
- Outcome: Breast Cancer Survival (Bellaachia & Guven, 2006)
- Data Cleaning
  - Recoded missing codes to NA, but didn't drop records

Our Methodology: Data Analysis

- C5.0 decision-tree generating algorithm
  - Extended version of C4.5
  - Solves problem of over fitting and error pruning
  - C5.0 algorithm deals with noise and missing data
  - Uses recursive binary splitting to grow a large tree on the training data
    - Stops at *node purity*

Our Methodology: Data Analysis

Statistical Software

- Python
  - create dataset from SEER ascii file
- R
  - Package “C50”
    - Training data: 80%

Break
Setup of Jupyter

Prep for Hands-On Jupyter Demo

1) Install Jupyter as explained here: http://bit.ly/2okeMU7
2) Download from same link (go to notebooks folder):
    a) data files from GitHub
    b) notebook files from GitHub
3) Start up Jupyter Notebooks application
4) Open “STLDataScience_data_cleaning_training.ipynb”

Walkthrough of Methodology in Jupyter Notebooks

Comparing Models

	Bellaachia et al.	**McIntosh et al.**
Data period	1973—2002	1973—2002
Original Count of records	482,052 records	483,489 records
Original Variables	Most likely around 80 variables	Current SEER database has over 130 variables
Final Count of records	151,886 records	Full Dataset - 238,457 Complete Cases Data - 92,518
Survivability Breakdown	23% not survived, 77% survived	Full Dataset: 28% not survived, 72% survived

	Bellaachia et al.	**McIntosh et al.**
Final Variables	17 variables (16 predictor variables and 1 dependent variable)	18 variables (17 predictor variables and 1 dependent variable)
Pre-classification base	Survival Time Recode (STR); Vital Status Recode (VSR); Cause of Death (COD)	Survival Time Recode (STR); Vital Status Recode (VSR); Cause of Death (COD)

	Bellaachia et al	**McIntosh et al**
Accuracy of Predictions	C4.5 Decision Tree with 86.7% Accuracy	C5.0 Decision Tree with 86.1% Accuracy
Sensitivity/Recall	96%	94.2%
Precision	88%	87.4%
Information Gain (Top 4)	Extension, Stage, LN Involvement, Site Specific Surgery	Behavior, Site Specific Surgery, Age, Positive Node Presence
Analysis Tools Used	Weka (Open-source Java tool)	R; Python; Jupyter Notebooks

Performance Metrics Summary

Full dataset accuracy - 86.1%

Complete-cases dataset accuracy - 88.5%

Comparing Full Dataset Models with Different Confidence Levels

Technique	Accuracy	Error_Rate	Sensitivity	Specificity
CF 0.10	85.9%	14.1%	94.3%	64.7%
CF 0.15	86%	14%	94.4%	64.6%
CF 0.25	**86.1%**	**13.9%**	**94.2%**	**65.4%**
CF 0.30	86.1%	13.9%	94.2%	65.4%

Comparing Full Dataset Models with Different Confidence Levels (Cont'd)

Technique	Prevalence	PPV	NPV	Kappa
CF 0.10	71.7%	87.1%	81.6%	62.9%
CF 0.15	71.7%	87.1%	82%	63.1%
CF 0.25	**71.7%**	**87.4%**	**81.6%**	**63.4%**
CF 0.30	71.7%	87.4%	81.7%	63.5%

Graphics in Jupyter Notebooks

Use an abundance of graphics packages, static and interactive

Render tables, just by calling the data.frame or table object

Save images to use in publications

Graphics in Jupyter Notebooks

R example:

```
png("resources/image1.png", width = 4, height = 4, units =
'in', res = 300)

plot(data)

graphics.off()
```

“Easter Egg”

You can write R and Python code in one notebook if you use Beaker Notebooks

http://beakernotebook.com/

Available for Mac/Windows/Linux

Runs natively or through Docker

THANK YOU!
Questions?