AMIA Joint Summits 2017

Data Science Training Workshop: Using Jupyter Notebook and R (with a little Spark)

Leslie D. McIntosh, PhD, MPH
Connie Zabarovskaya, MITM
Lorinette S. Wirth, MPH

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

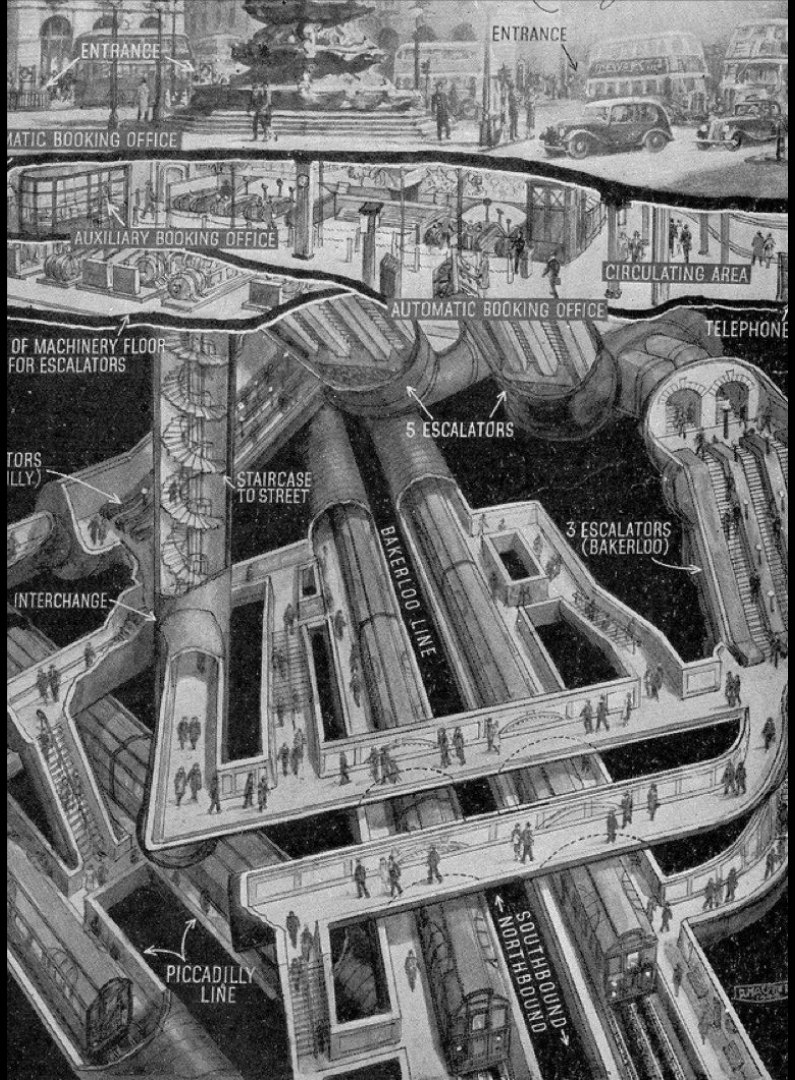
Special Thanks

Michael Yingling

Statistical Data Analyst

Washington University in St. Louis

An Informatics Perspective

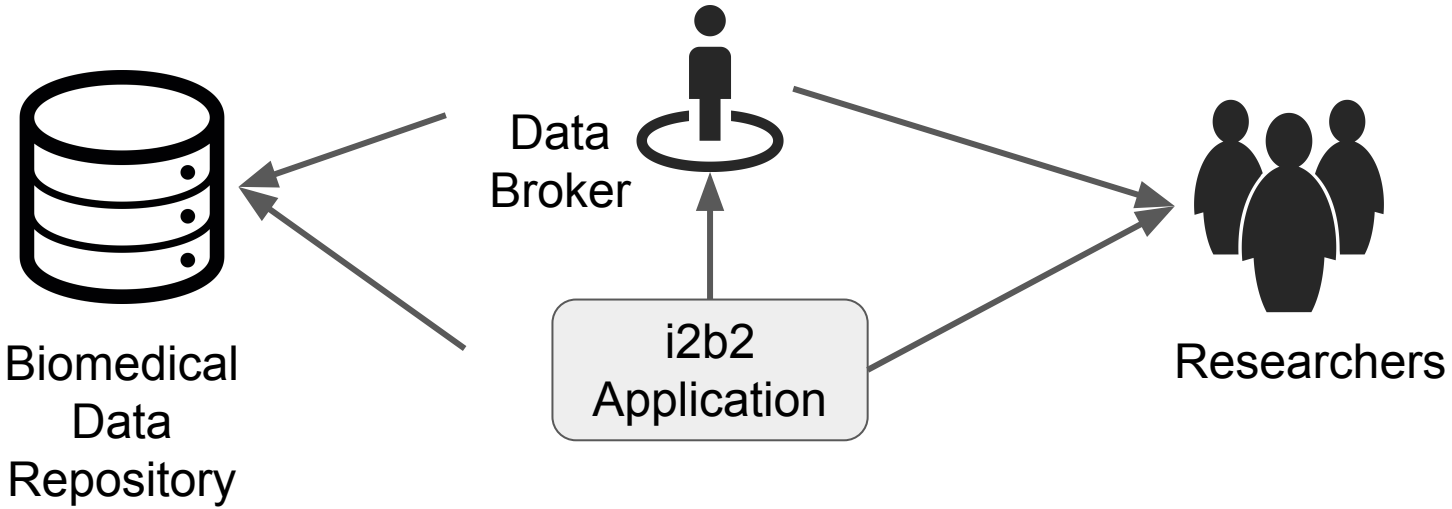

ENTRANCE
ENTRANCE
ATIC BOOKING OFFICE
AUXILIARY BOOKING OFFICE
CIRCULATING AREA
AUTOMATIC BOOKING OFFICE
TELEPHONE
OF MACHINERY FLOOR
FOR ESCALATORS
5 ESCALATORS
TORS
ILLY)
STAIRCASE TO STREET
3 ESCALATORS (BAKERLOO)
INTERCHANGE
BAKERLOO LINE
SOUTHBOUND
NORTHBOUND
PICCADILLY LINE

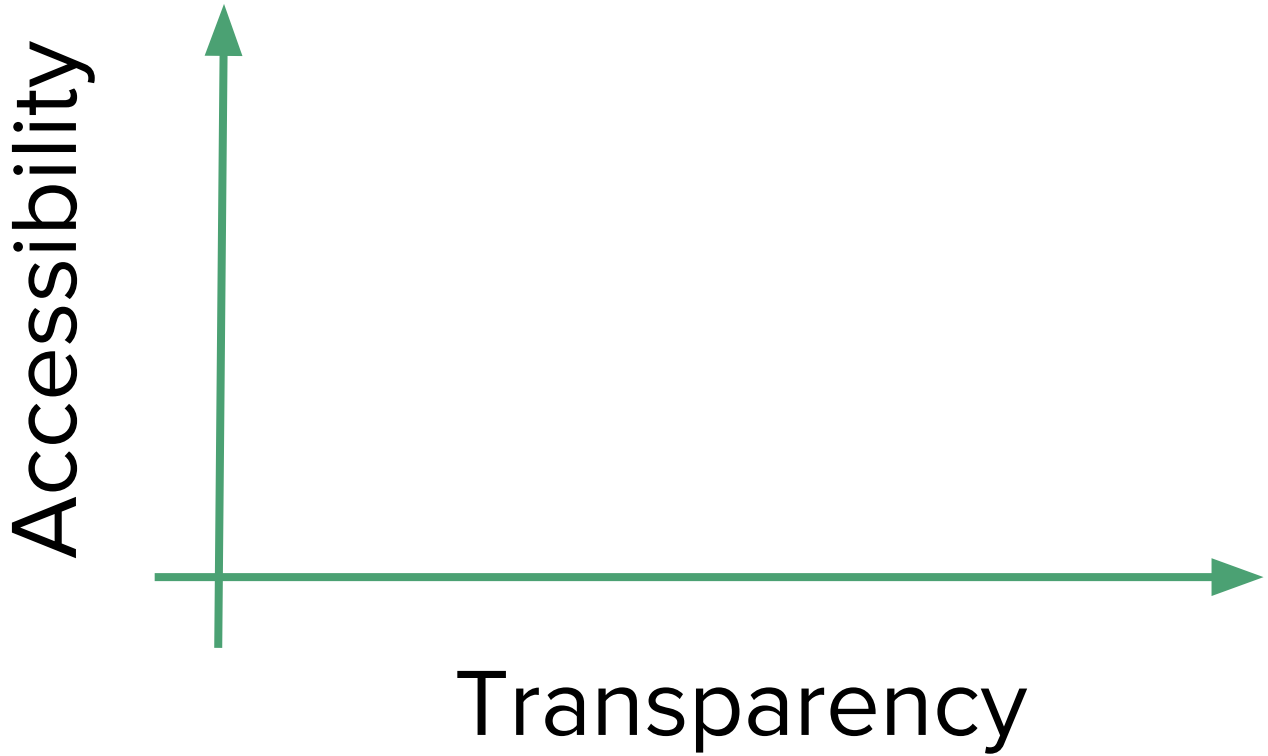

Harrow-on-the-Hill
Northwick Park
South Kenton
North Wembley
Wembley Central
Stonebridge Park
Harlesden
Willesden Junction
Preston Road
Wembley Park
Neasden
Dollis Hill
Willesden Green
Kilburn
Brondesbury Park
West Hampstead
Kensal Rise
Brondesbury
Kensal Green
Queen's Park
Golders Green
Hampstead
Hampstead Heath
Gospel Oak
Finchley Road & Frognal
Belsize Park
Chalk Farm
Camden Town
Finchley Road
Swiss Cottage
St. John's Wood
Archway
Tufnell Park
Kentish Town West
Kentish Town
Mornington Crescent
Camden Road
Manor House
Arsenal
Holloway Road
Caledonian Road
Caledonian Road & Barnsbury
Tottenham Hale
Walthamstow Central
Finsbury Park
Canonbury
Hackney Central
Hackney Wick
Highbury & Islington
Dalston Kingsland
Homerton
King's Cross St. Pancras
Euston
Great Portland Street
Baker Street
Angel
Farringdon
Barbican
Old Street
Liverpool Street
Bethnal Green
Mile End
Shoreditch
Bow Road
Bow Church
Stepney Green
Whitechapel
Devons Road
All Saints
Aldgate East
Aldgate
Shadwell
Westferry
Poplar
Limehouse
Wapping
West India Quay
Canary Wharf
Heron Quays
South Quay
Crossharbour & London Arena
Mudchute
Island Gardens
Cutty Sark for Maritime Greenwich
Deptford Bridge
Elverson Road
Lewisham
New Cross
New Cross Gate
Surrey Quays
Rotherhithe
London Bridge
Bermondsey
Canada Water
Tower Hill
Tower Gateway
Monument
Fenchurch Street
Bank
Cannon Street
Mansion House
Blackfriars
Temple
Embankment
Charing Cross
Waterloo
Southwark
Borough
Lambeth North
Elephant & Castle
Vauxhall
Kennington
Oval
Stockwell
Clapham North
Clapham South
Clapham Common
Brixton
Balham
Pimlico
Victoria
Westminster
St. James's Park
Sloane Square
South Kensington
Gloucester Road
Earl's Court
West Kensington
Barons Court
Hammersmith
Goldhawk Road
Shepherd's Bush
Kensington (Olympia)
Ravenscourt Park
Stamford Brook
Acton Central
East Acton
White City
Latimer Road
Ladbroke Grove
Westbourne Park
Royal Oak
Warwick Avenue
Kilburn Park
Maida Vale
Paddington
Edgware Road
Marylebone
Bayswater
Holland Park
Notting Hill Gate
Queensway
Lancaster Gate
Marble Arch
Bond Street
Oxford Circus
Tottenham Court Road
Warren Street
Regent's Park
Euston Square
Goodge Street
Holborn
Russell Square
Chancery Lane
St. Paul's
Moorgate
Covent Garden
Leicester Square
Piccadilly Circus
Green Park
Hyde Park Corner
High Street Kensington
Knightsbridge
Fulham Broadway
West Brompton
Parsons Green
Putney Bridge
East Putney
Southfields
Wimbledon Park
Wimbledon
River Thames
Interchange stations
Connections with National Rail
Connections with riverboat services
Airport interchange
Explanation of zones
Points
Bakerloo
Central
Circle
District

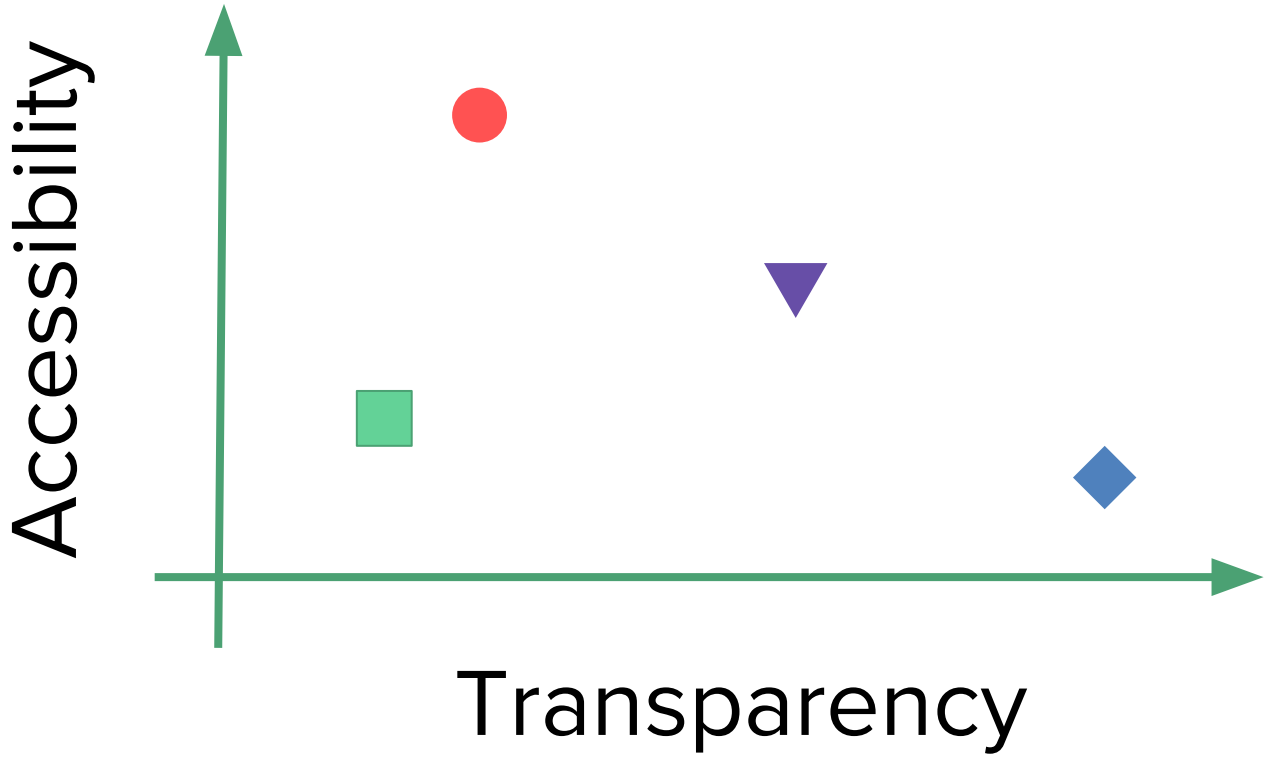

WAY OUT

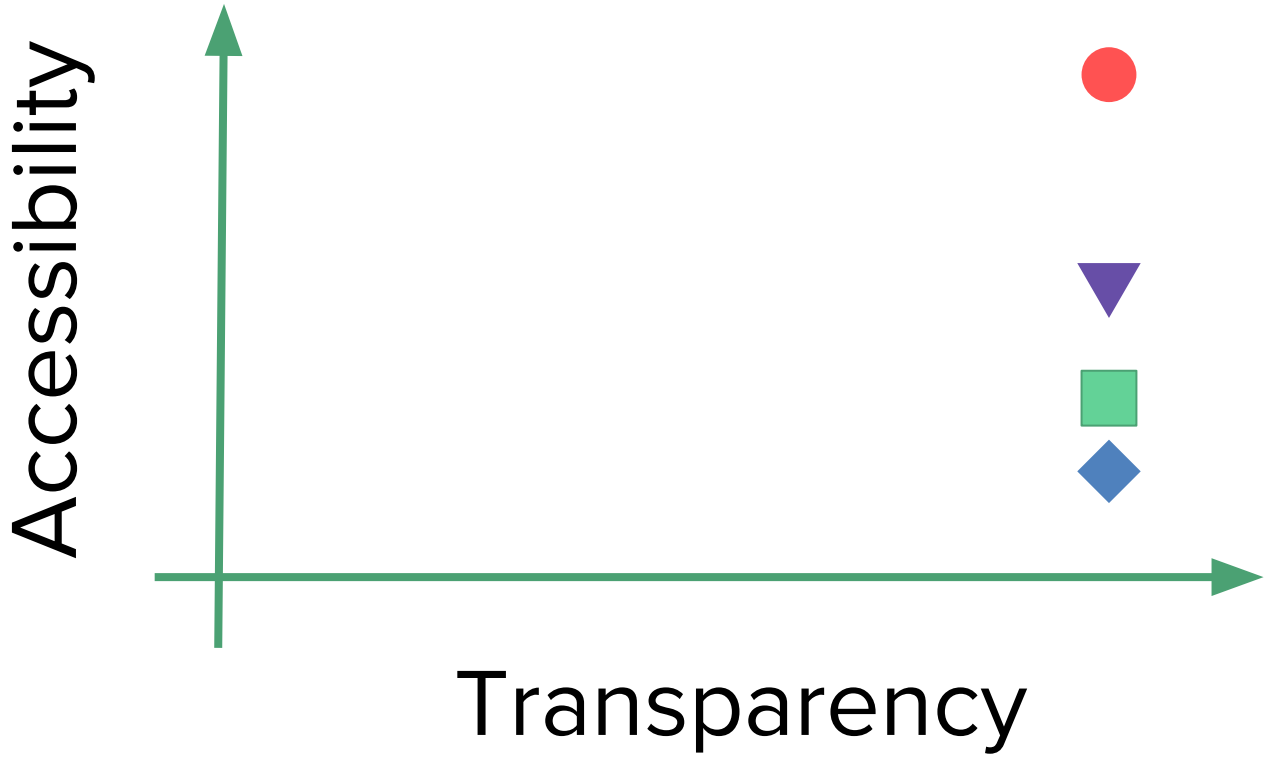
BDaaS

Biomedical Data as a Service




Accessibility
Transparency


Accessibility
Transparency


Accessibility
Transparency

Challenges

Clinical recommendations based upon restricted use data pose challenges for research transparency and accessibility.

Given the protected nature of much biomedical research, what is needed to call our research reproducible?

How do we as biomedical researchers facilitate research reproducibility?

What *is* reproducibility research anyway?

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/

Why do we care?

Climategate

2009

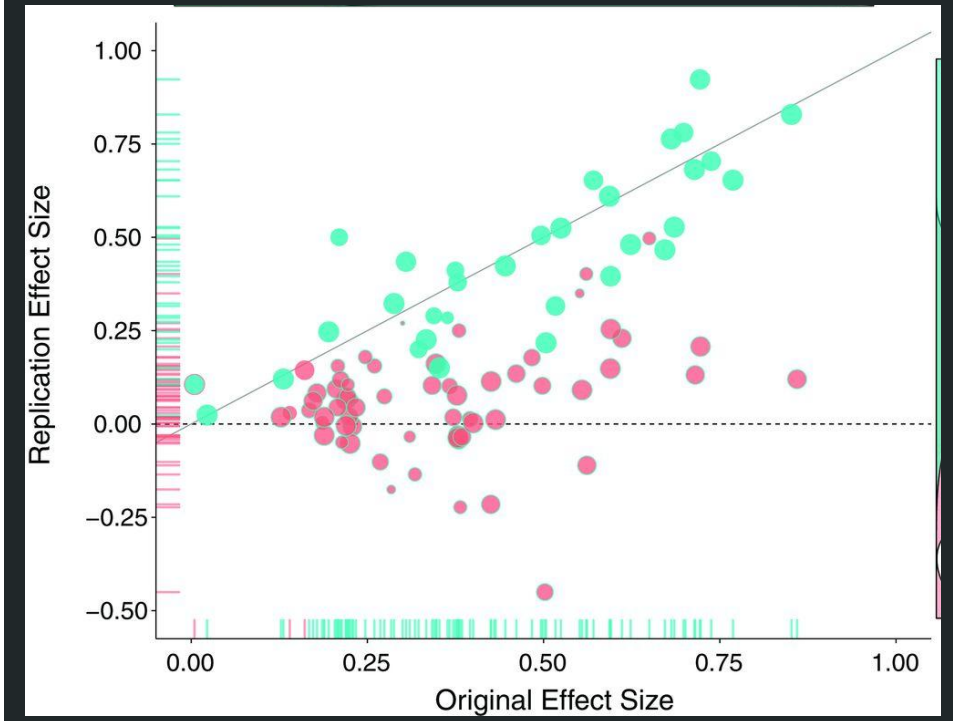
Duke's Precision Medicine Bust

2007 - 2011+

(Lack of)

Reproducibility of Psychological Science

2015



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

Paper	Conclusion	Focus of key experiment	Replication results	Citations
Sirota, M. *et al. Sci. Transl. Med.* **3,** 96ra77 (2011)	Public gene expression data can identify unintuitive uses for old drugs	Growth of tumours treated with an anti-ulcer drug	**Substantially reproduced**	334
Sugahara, K. N. *et al. Science* **328,** 1031–1035 (2010)	A tumour-penetrating peptide enhances the effects of cancer drugs	Growth of peptide-treated tumours	**Not reproduced**	495
Willingham, S. B. *et al. Proc. Natl Acad. Sci. USA* **109,** 6662–6667 (2012)	Blocking contact between CD47 and another protein inhibits tumour	Growth and metastasis of treated tumours	**Uninterpretable**	290
Delmore, J. E. *et al. Cell* **146,** 904–917 (2011)	Blocking a protein sequence damps down pro-cancer genes	Gene expression in treated cells; growth of treated tumours	**Substantially reproduced**	1059
Berger, M. F. *et al. Nature* **485,** 502–506 (2012)	Sequencing reveals gene that is frequently mutated in melanoma and accelerates growth	Tumour formation in cells carrying mutations	**Uninterpretable**	428

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!

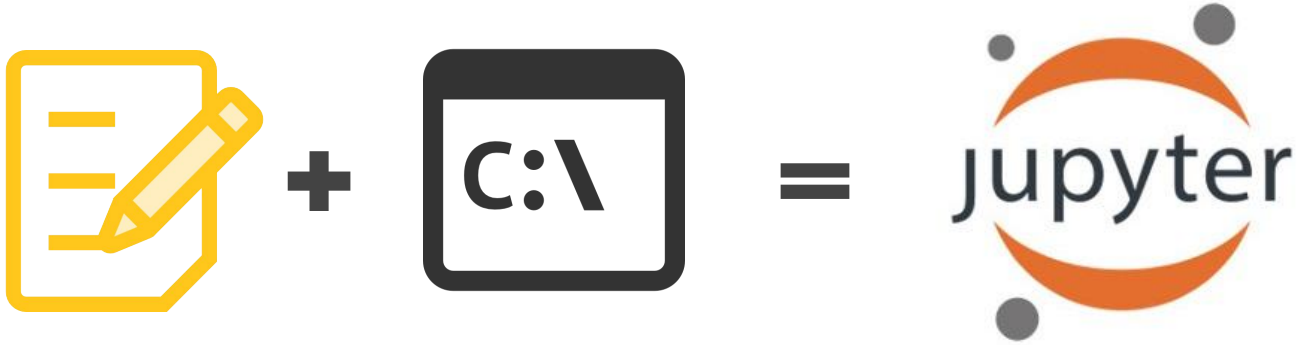
Export document as .pdf, .doc, .html with ease

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

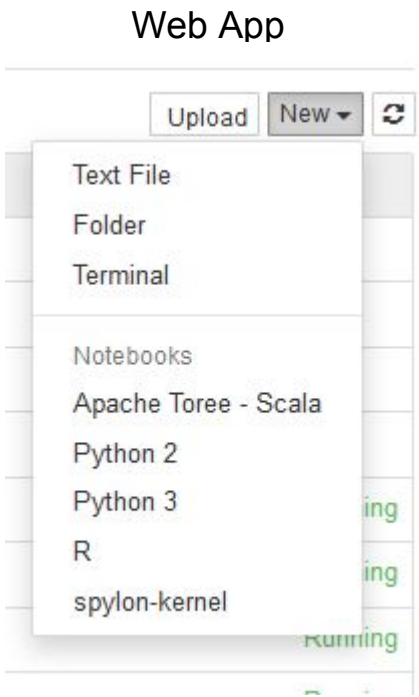
Easy to attach as supplement for manuscript publication

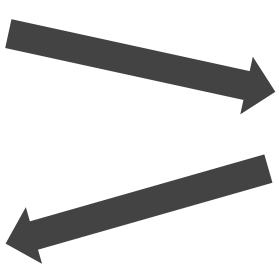
- Transparency

Jupyter Notebooks Overview



How Does it Work?





Kernel

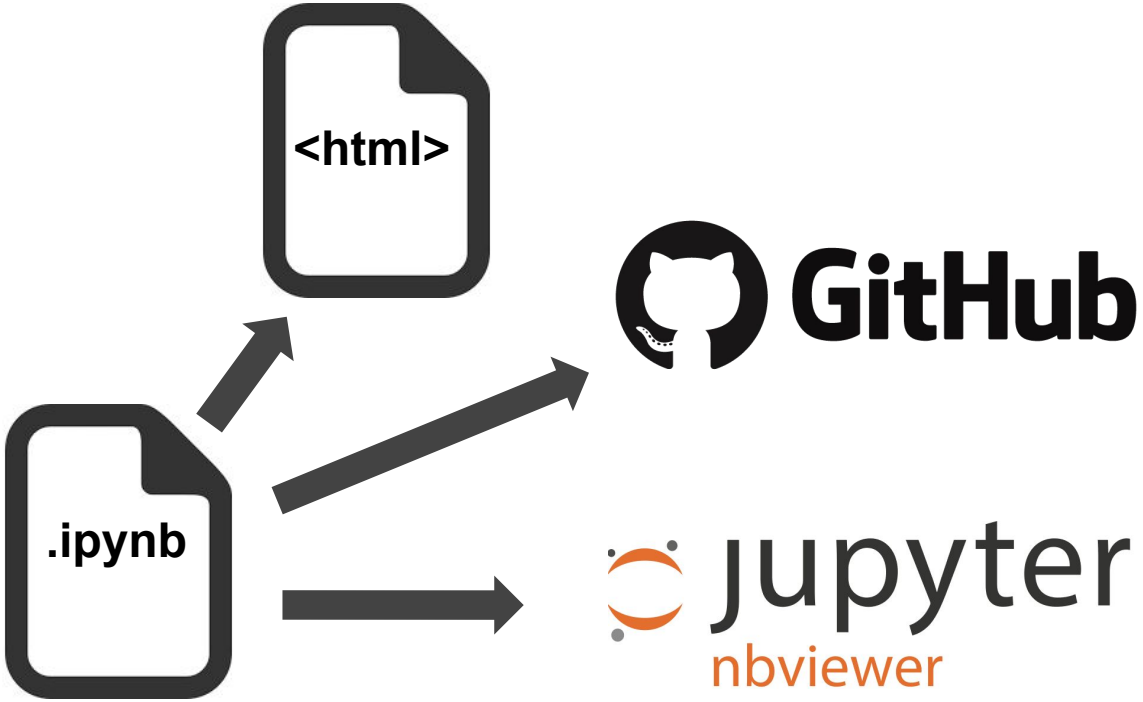
julia

Scala

jupyter

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)

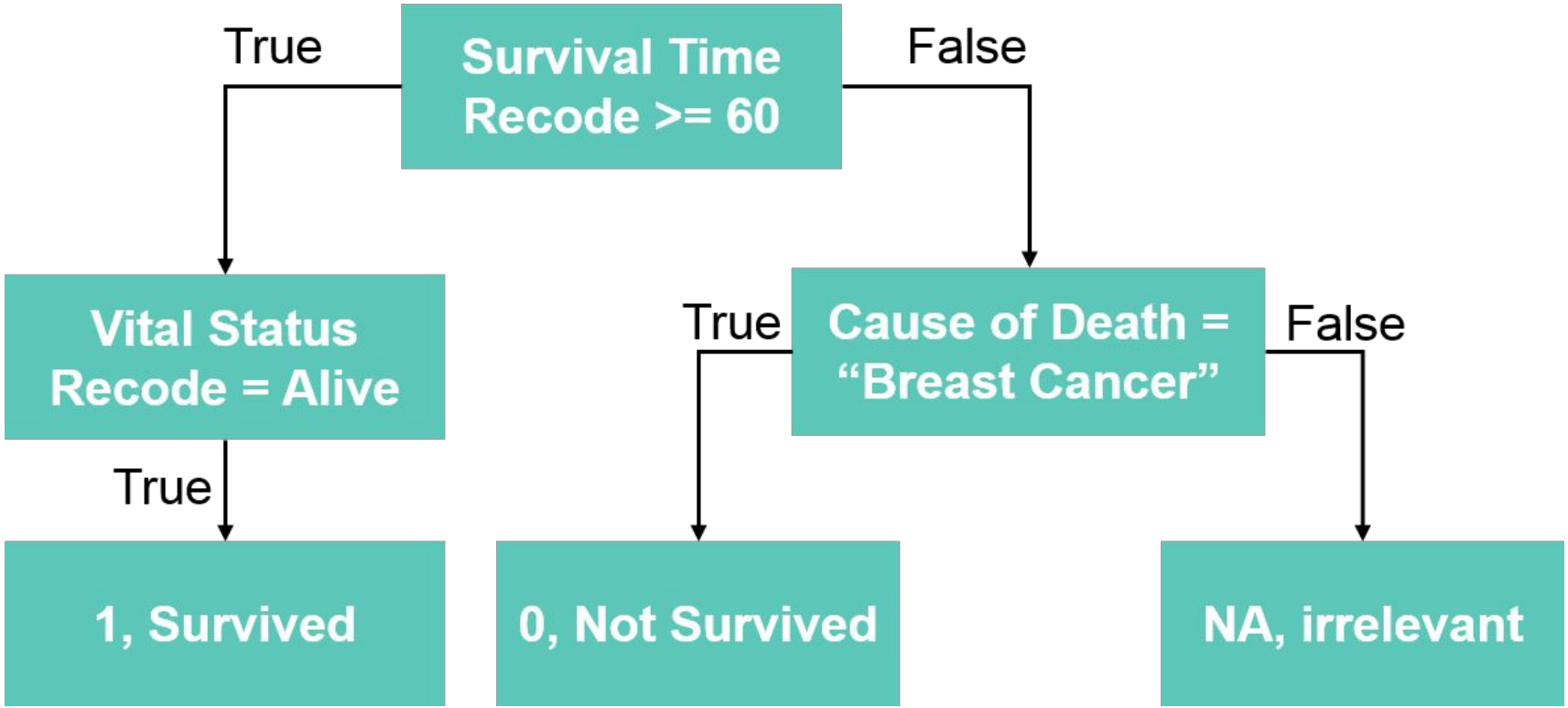
"Breast.txt" Ascii file, 1973-2002

Methodology (Bellaachia & Guven, 2006)

Covariates

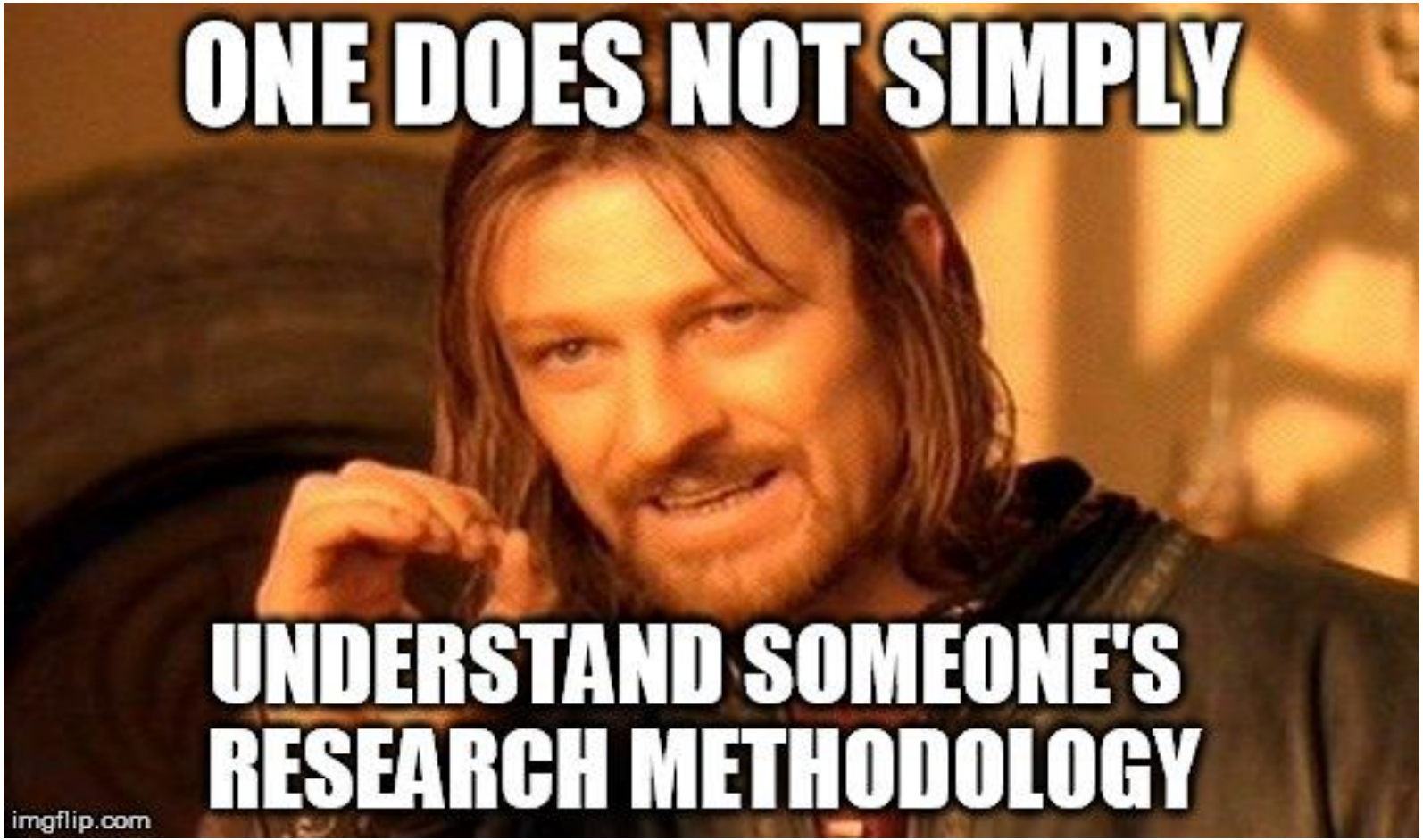
- Nominal: Race, Marital Status, Primary site code, Histologic type, Behavior code, Grade, Extension of tumor, Lymph node involvement, Site specific surgery code, Radiation, Stage of Cancer
- Numeric: Age, Tumor size, # of Positive nodes, # of Nodes, # of primaries

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?

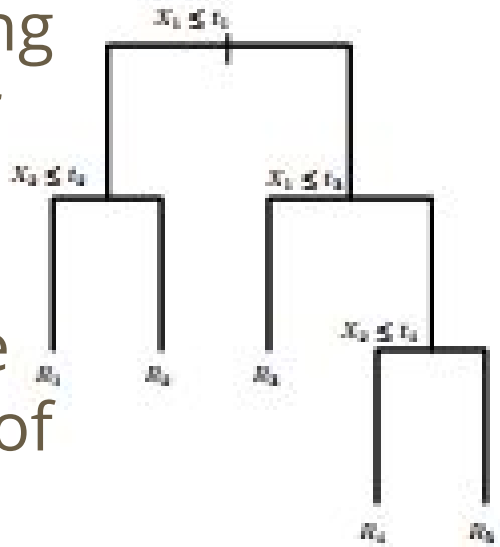

ONE DOES NOT SIMPLY
UNDERSTAND SOMEONE'S RESEARCH METHODOLOGY
imgflip.com

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 gives the acknowledge on 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/2mLNGCW
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 “amia_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%

Why Jupyter Notebooks?

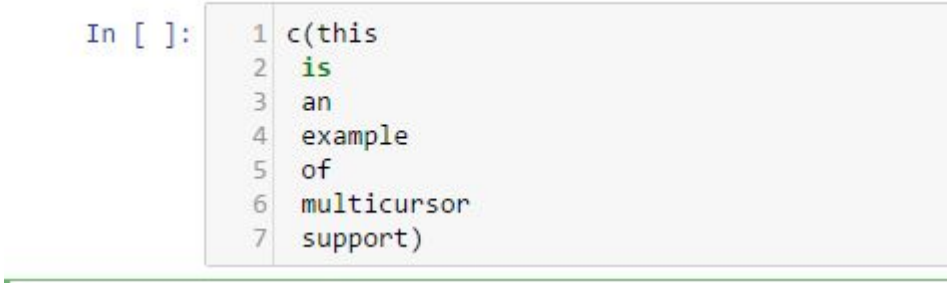
THE PURPOSE OF THIS PRESENTATION IS NOT TO ENDORSE A PARTICULAR NOTEBOOK

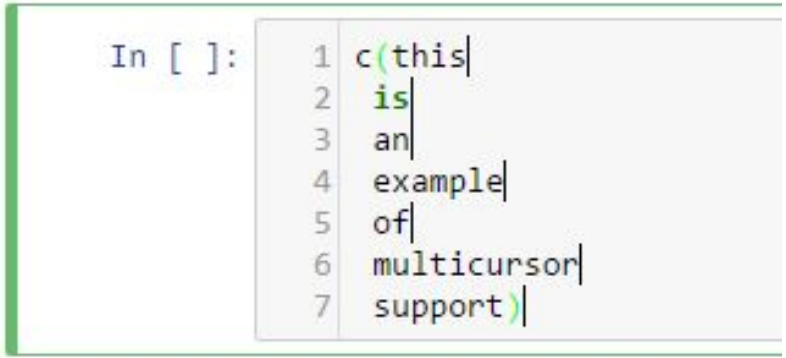
- All notebooks have their pros and cons
- Explore for yourself!

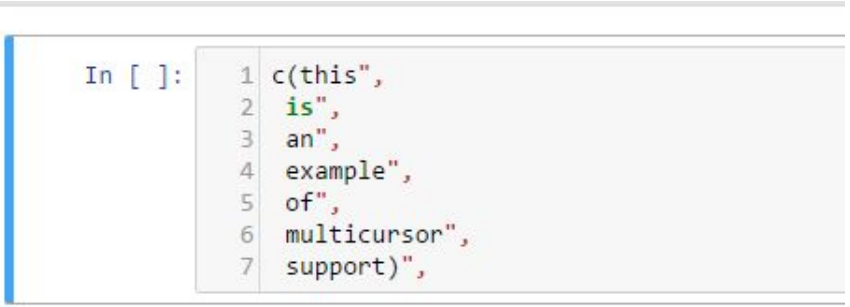
Why Jupyter Notebook?

- **Strong community support** and publicly available documentation compared to other Notebook software (e.g. Zeppelin, R Notebooks)
- Run as many as **40 programming languages**

Multicursor Support
Alt+mouse selection; Ctrl+ mouse clicks





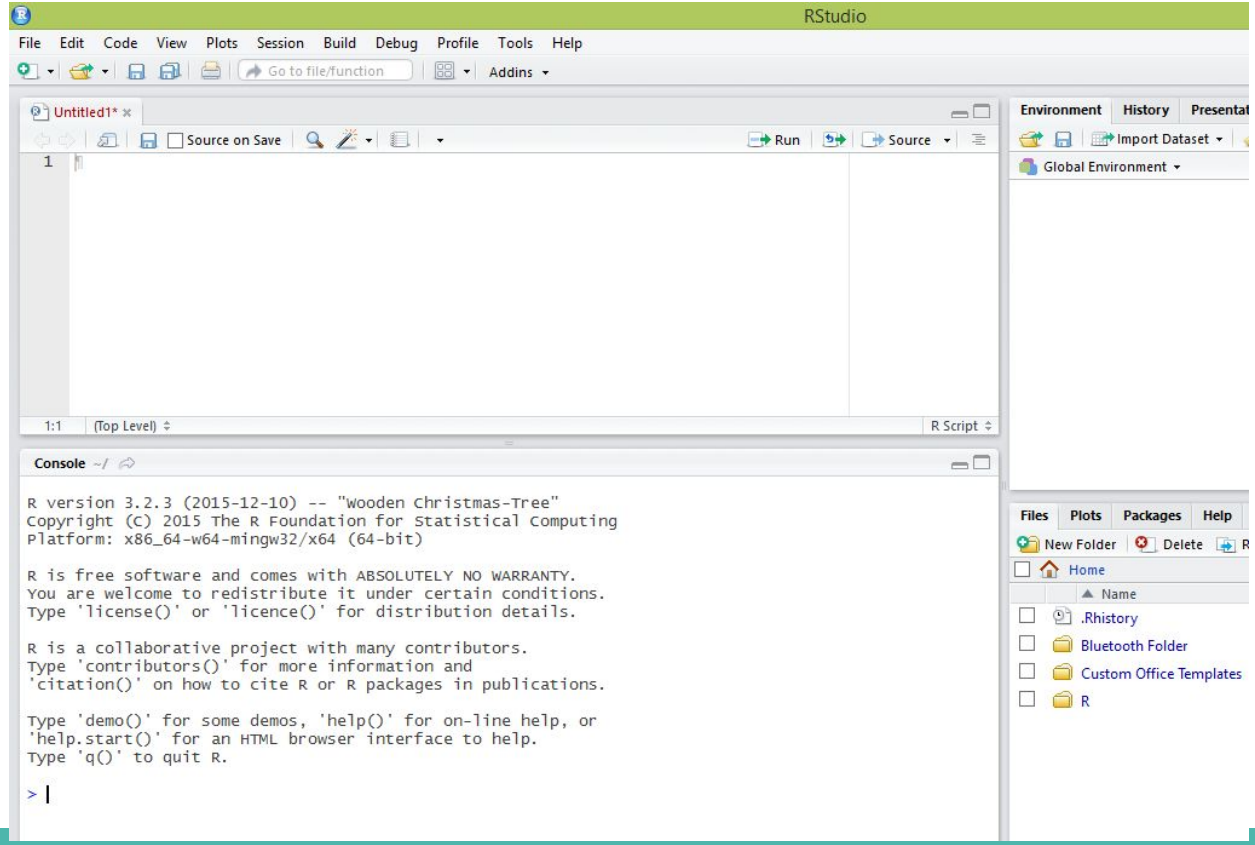


Rpy2

Use both R and Python programming languages simultaneously within the Python Kernel

See http://stackoverflow.com/search?q=rpy2

No Console


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Untitled1*
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Console ~/
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Environment
History
Import Dataset
Global Environment
Files
Plots
Packages
Help
New Folder
Delete
Home
Name
.Rhistory
Bluetooth Folder
Custom Office Templates
R

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:

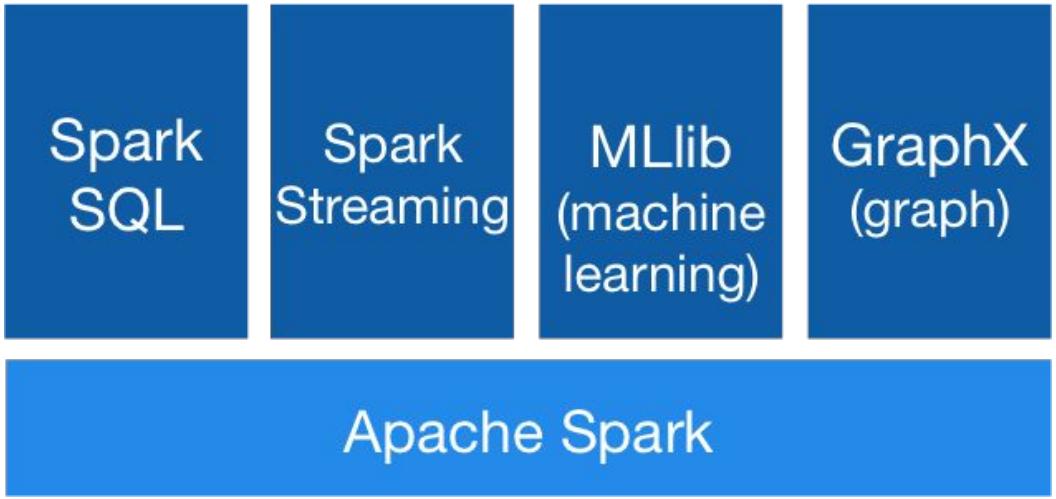
```
png("resources/image1.png", width = 4, height = 4, units =
'in', res = 300)

plot(data)

graphics.off()
```

How to Cite Jupyter Notebooks?

Option 1: Attach notebook file as supplement to electronically published paper.

Option 2: Include link to notebook file and data shared on public hosting service as part of electronically published paper.

Spark … Finally



Courtesy of http://spark.apache.org/

Spark with Jupyter Notebooks

Communicate with Spark through backend process.

Options to connect to Spark:

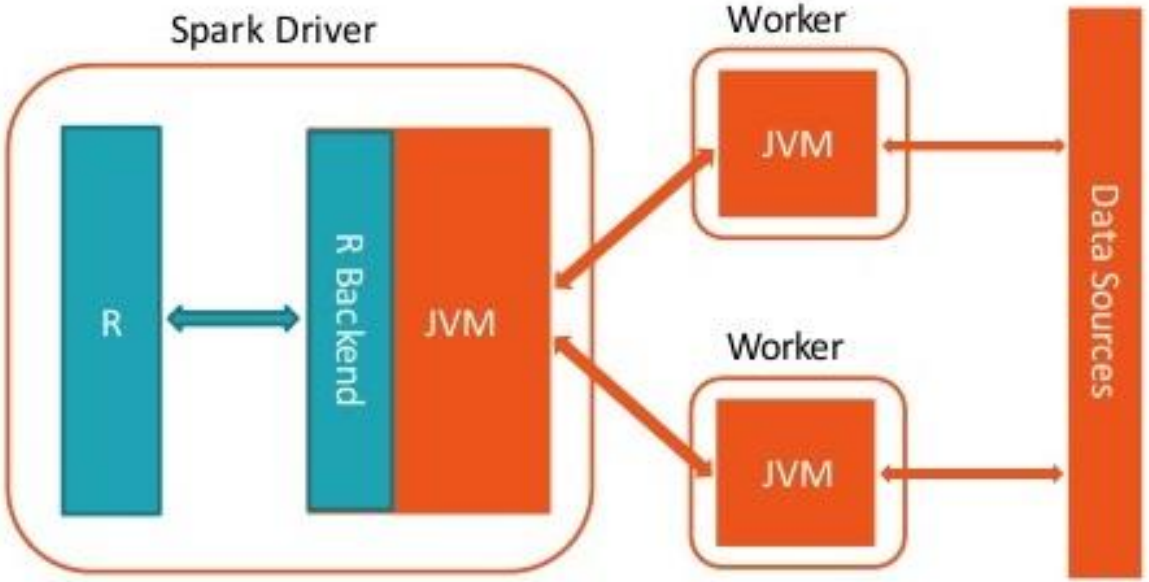
1.

- **Install Docker**
- **Pull image with Spark and Jupyter Notebooks**
- **Use some library to start Spark context (e.g. SparkR)**

2.

- **Install Anaconda and Spark**
- **Link Spark with Jupyter**
- **Use pyspark to run scripts**

SparkR architecture



databricks

Our Review of SparkR

Great for “computing” map, reduce, filter, aggregate operations

Not all common R functions have SparkR equivalents

Common R functions may take longer on SparkR DataFrames

Use magrittr package to perform data wrangling with friendly pipe-symbol syntax

THANK YOU!
Questions?