

# **DATA SCIENCE: HOW DO YOU DO IT?**

**MIKE GELBART, PH.D. & TIFFANY TIMBERS, PH.D.**

# Outline

---

- The data science software stack
- Choosing the "best" software tool or language for the job
- What is the (statistical) question?
- A vignette: Asking and answering a predictive question
- Where can I learn more data science?
- Where can I access compute resources?

# The data science software stack

---

## Type of tools needed:

- Programming language
- Code editor
- Version control software

## Examples:



# Choosing the "best" tool for the job

---

As long as the tools allow you to meet the following criteria for your analysis, they will be suitable for data science:

1. Reproducible and auditable
2. Accurate
3. Collaborative

Parker, H. (2017), Opinionated Analysis Development.  
*PeerJ*, doi: [10.7287/peerj.preprints.3210v1](https://doi.org/10.7287/peerj.preprints.3210v1)

# Reproducible and auditable

---

Data science requires reproducible and auditable analyses.

This means that you should be able to hand over your analysis and the raw data and they should be able to generate the exact same results as you. And see how this was done!

Performing analysis with code and version tracking can make your analysis reproducible and auditable!

*Parker, H. (2017), Opinionated Analysis Development.  
PeerJ, doi: [10.7287/peerj.preprints.3210v1](https://doi.org/10.7287/peerj.preprints.3210v1)*

# Reproducible and auditable

[https://github.com/ttimbers/breast\\_cancer\\_predictor](https://github.com/ttimbers/breast_cancer_predictor)

```
1 ---
2 title: "Exploratory data analysis of the Wisconsin Breast Cancer data set"
3 output: github_document
4 bibliography: ../doc/breast_cancer_refs.bib
5 ---
6
7 ```{r setup, include=FALSE}
8 knitr::opts_chunk$set(echo = FALSE, warning = FA
9 library(feather)
10 library(tidyverse)
11 library(knitr)
12 library(caret)
13 library(ggthemes)
14 library(ggthemes)
15 theme_set(theme_minimal())
16 set.seed(2020)
17 ```
18
19 # Summary of the data set
20
21 ```{r load data}
22 bc_data <- read_feather("../data/raw/wdbc.feather")
23 colnames(bc_data) <- c("id",
24   "class",
25   "mean_radius",
26   "mean_texture",
27   "mean_perimeter",
28   "mean_area",
29   .. .. .
```

## Exploratory data analysis of the Wisconsin Breast Cancer data set

### Summary of the data set

The data set used in this project is of digitized breast cancer image features created by Dr. William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian at the University of Wisconsin, Madison (Street, Wolberg, and Mangasarian 1993). It was sourced from the UCI Machine Learning Repository (Dua and Graff 2017) and can be found [here](#), specifically [this file](#). Each row in the data set represents summary statistics from measurements of an image of a tumour sample, including the diagnosis (benign or malignant) and several other measurements (e.g., nucleus texture, perimeter, area, etc.). Diagnosis for each image was conducted by physicians. There are 569 observations in the data set, and 31 features. There are 0 observations with missing values in the data set. Below we show the number of each observations for each of the classes in the data set.

Benign cases	Malignant cases
357	212

Table 1. Counts of observation for each class.

### Partition the data set into training and test sets

Before proceeding further, we will split the data such that 75% of observations are in the in the test set. Below we list the counts of observations for each class:

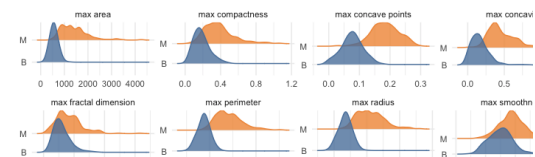
Data partition	Benign cases	Malignant cases
Training	268	159
Test	89	53

Table 2. Counts of observation for each class for each data partition.

There is a minor class imbalance, but it is not so great that we will plan to immediately st under-sampling. If during initial tuning, there are indicators that it may in fact be a great confusion matrix indicates that the model makes a lot more mistakes on the minority class only then start to explore whether employing techniques to address class imbalance may performance in regards to predicting the minority class.

### Exploratory analysis on the training data set

To look at whether each of the predictors might be useful to predict the tumour class, we predictor from the training data set and coloured the distribution by class (benign: blue & malignant: orange). We see that class distributions for all of the mean and max predictors for all the measure quite a difference in their centres and spreads. This is less so for the standard error (se) errors of fractal dimension, smoothness, symmetry and texture look very similar in both T. Thus, we might choose to omit these from our model.



Commits on Feb 5, 2020

updated usage instructions for running analysis with Docker

tttimbers committed on Feb 5 ✓

a06418c

using results created from Docker container

tttimbers committed on Feb 5 ✓

9cfa83b

Added a dockerfile

tttimbers committed on Feb 5 ✓

a60cd07

Commits on Jan 28, 2020

added Makefile

tttimbers committed on Jan 28 ✓

Verified 58eb063

Commits on Jan 25, 2020

removed commented out code

tttimbers committed on Jan 25 ✓

Verified 997b7a7

added a working reticulate script

tttimbers committed on Jan 25 ✓

5acb7b8

draft of R script the uses reticulate to do ML with sklearn

tttimbers committed on Jan 25 ✓

3bbe8ae

Commits on Jan 24, 2020

added make to the usage instructions and dependencies

tttimbers committed on Jan 24 ✓

b91f85e

added Makefile

8a2b461

# Accurate

---

For analysis to be trustworthy, it also has to be accurate.


Writing your analysis with modular and testable code helps make your analysis accurate.

As does having your analysis code reviewed by others.

Parker, H. (2017), Opinionated Analysis Development.  
*PeerJ*, doi: [10.7287/peerj.preprints.3210v1](https://doi.org/10.7287/peerj.preprints.3210v1)


# Accurate

```
1 import numpy as np
2 import pandas as pd
3 import pytest
4 from sklearn.datasets import load_iris
5
6 from feature_selection import variance_thresholding
7
8 iris = pd.DataFrame(load_iris().data)
9
10
11 def test_1d_array_support():
12     """
13     Test with 1d array
14     """
15     result = variance_thresholding([1, 2, 3, 4, 5])
16     assert np.array_equal(result, [0])
17
18
19 def test_2d_array_support():
20     """
21     Test with 2d array
22     """
23     result = variance_thresholding(
24         [[1, 6, 0, 5], [1, 2, 4, 5], [1, 7, 8, 5]]
25     )
26     assert np.array_equal(result, [1, 2])
27
28
29 def test_df_support():
30     """
31     Test DataFrame support
32     """
33     iris_copy = pd.DataFrame.copy(iris)
34     iris_copy['fake_num'] = np.zeros(iris_copy.shape[0])
35     iris_copy['fake_categorical'] = 'abcde'
```

 Reworked ex2 solution to make it easier to understand #65

Changes from all commits ▾ File filter... ▾ Jump to... ▾ ⚙ ▾

0 / 1 files viewed ⓘ Review changes ▾

34 source/lab2/lab2.Rmd 

Viewed ⋮

95 - mutate(ci = map(party, ~ci\_prop(polls, party, .))) %>%

96 - left\_join(props) %>%

97 - unnest(ci) %>%

98 - mutate(party = fct\_reorder(party, prop, .desc = TRUE))

86 + est <- polls %>%

87 + rep\_sample\_n(size = nrow(polls), reps = 15000, replace = TRUE) %>%

88 + group\_by(replicate, party) %>%

89 + summarize(stat = n() / nrow(polls)) %>%

90 + ungroup() %>%


91 + select(-replicate) %>%

92 + group\_by(party) %>%

93 + nest() %>%

94 + summarize(ci = map(data, ~get\_confidence\_interval(., level = 0.95, type = 'percentile')))

Write Preview


 AA B i “ < > ↺ ⋮ ≡ ≡ @ 📌 ↶

We should be using 90% confidence intervals, not 95%

'''suggestion

summarize(ci = map(data, ~get\_confidence\_interval(., level = 0.90, type = 'percentile')))

'''

Attach files by dragging & dropping, selecting or pasting them. 

Cancel Add single comment Start a review

95 + unnest(ci) %>%

96 + left\_join(props)

99 97



# Collaborative

---


An analysis is not useful unless it is used/viewed by others.

Most analysis for solving real-world problems are quite complex and sophisticated. Thus multiple experts are often required to complete such an analysis.

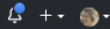
Collaborating on writing and code can be challenging if care is not taken, and versions are not managed.

Parker, H. (2017), Opinionated Analysis Development.  
*PeerJ*, doi: [10.7287/peerj.preprints.3210v1](https://doi.org/10.7287/peerj.preprints.3210v1)

# Collaborative



[Pull requests](#) [Issues](#) [Marketplace](#) [Explore](#)



[LerouxLab / Celegans\\_wild\\_isolate\\_behaviour](#) Private

Unwatch 5 Star 0 Fork 0

[Code](#) [Issues](#) 10 [Pull requests](#)

master

2 branches

tttimbers Update README

Posters

bin

data

manuscript

presentations

.gitignore

Celegans\_wild\_isolate\_behaviour

LICENSE.md

Makefile

README.md


README.md

Caenorhabditis

Authors: Tiffany Tir

## Data preparation for Choreography #1


Open tttimbers opened this issue on Sep 16, 2015 · 3 comments




tttimbers commented on Sep 16, 2015

To do:

- Copy all raw MWT folders into the `data` directory inside the `Celegans_wild_isolate_behaviour` directory on WebDav (<https://webdav.sfu.ca/files/leroux-lab/leroux-lab-files/>)
- Create a `.csv` file called `wild_isolate_meta.csv` where each experiment (e.g., plate tested) is a row and which has the following columns: plate (contains date-time stamp, e.g., 20150918\_123423), date, strain, experimenter, and temperature. Save this in the `data` directory on WebDav.
- Create a `.txt` file called `strains.txt` that lists all of the strain names in one column named "strain" and the GPS location of where the strains were isolated from in a second column named "location". Save this in the `data` directory on WebDav.




cloudcks assigned cloudcks on Sep 16, 2015



cloudcks commented on Sep 30, 2015

@tttimbers I have finished the `wild_isolate_meta.csv` file and put it in the `webdav` folder. I took out plates that were obviously contaminated or not synchronized. There is a lot of variability with how many worms are on a plate and the thickness of the bacterial lawn, so I decided to keep in the data from the two days where the OP50 wasn't spread. There are some plates that have very few worms. I noted the extreme cases (1-4 worms) and removed any with no worms, but do we want to eliminate plates with under a certain number of worms? Finally, because there has been a lot of variability with thickness of lawn and number of worms, there are several plates where I can't really tell if there's still bacteria on the plates. I left these plates in but left a note beside them in the `.csv` file (food?). Let me know if you think we should do anything differently.



tttimbers commented on Sep 30, 2015

Author

Thank-you @cloudcks ! Hmmm, censoring based on number of worms, we can but don't necessarily have to... It depends on how we do plotting and statistics. If we need to manually remove the data since we don't have a good plate.

Assignees

cloudcks

Labels

None yet

Projects

None yet

Milestone

No milestone

Linked pull requests


Successfully merging a pull request may close this issue.

Notifications

Unsubscribe

You're receiving notifications because you're watching this repository.

2 participants



Lock conversation

Pin issue

Transfer issue

# Choosing the "best" tool for the job

---



Source: <http://thecontextofthings.com/2015/11/11/work-and-certain-skills-belonging-to-a-few-people/>

Avoid the language wars... silos are worse than not choosing the "perfect" tool.

# Choosing the "best" tool for the job

---

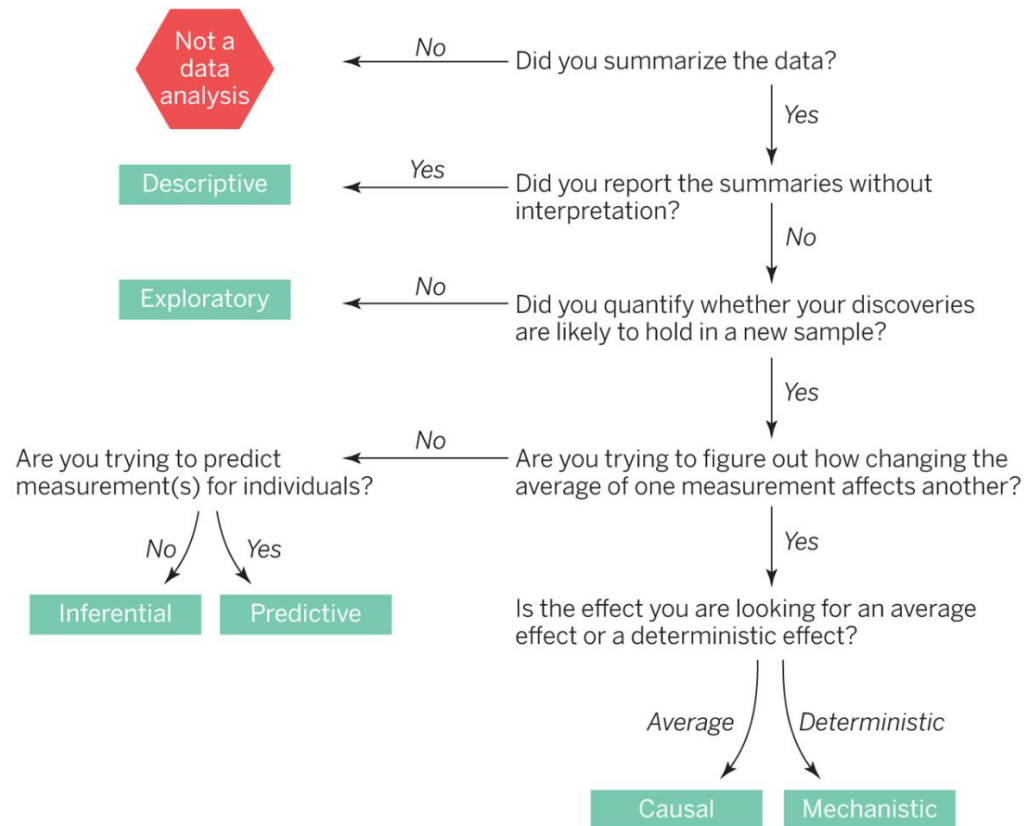


Source: <https://chatsworthconsulting.com/2015/01/29/why-it-is-imperative-to-break-down-silos-now-and-five-ways-to-do-it/>

Sharing of tools and workflows connect silos and leads to collaboration and success

# What is the (statistical) question?

## Data analysis flowchart



Peng, R. & Leek, J. (2015), What is the question?

Science, doi: [10.1126/science.aaa6146](https://doi.org/10.1126/science.aaa6146)

# Descriptive question

---

One that seeks to summarize a characteristic of a set of data. No interpretation of the result itself as the result is a fact, an attribute of the data set you are working with.

1. What is the frequency of viral illnesses in a set of data collected from a group of individuals?
2. How many people live in BC?

# Exploratory question

---

One in which you analyze the data to see if there are patterns, trends, or relationships between variables looking for patterns that would support proposing a hypothesis to test in a future study.

1. What is the frequency of viral illnesses in a set of data collected from a group of individuals?
2. Does air pollution correlate with life expectancy in a set of data collected from groups of individuals from several regions in the United States?

# Inferential question

---

One in which you analyze the data to see if there are patterns, trends, or relationships between variables in a representative sample. We want to quantify how much the patterns, trends, or relationships between variables is applicable to all individuals units in the population.

1. Is eating at least 5 servings a day of fresh fruit and vegetables is associated with fewer viral illnesses per year?
2. Is the gestational length of first born babies different from that of non-first borns?



# Predictive question

---

One where you are trying to predict measurements or labels for individuals (people or things). Less interested in what causes the predicted outcome, just what predicts it.

1. How many viral illnesses will someone have next year?
2. What political party will someone vote for in the next US election?

# Causal question

---

Asks about whether changing one factor will change another factor, on average, in a population. Sometimes the underlying design of the data collection, by default, allows for the question that you ask to be causal (e.g., randomized experiment or trial).

1. Does eating at least 5 servings a day of fresh fruit and vegetables cause fewer viral illnesses per year?
2. Does smoking lead to cancer?

# Mechanistic question

---

One that tries to explain the underlying mechanism of the observed patterns, trends, or relationship (how does it happen?)

1. How do changes in diet lead to a reduction in the number of viral illnesses?
2. How does how airplane wing design change air flow over a wing, leading to decreased drag?

# What is the (statistical) question?

## Common mistakes

REAL QUESTION TYPE	PERCEIVED QUESTION TYPE	PHRASE DESCRIBING ERROR
Inferential	Causal	“Correlation does not imply causation”
Exploratory	Inferential	“Data dredging”
Exploratory	Predictive	“Overfitting”
Descriptive	Inferential	“n of 1 analysis”

Peng, R. & Leek, J. (2015), What is the question?  
Science, doi: [10.1126/science.aaa6146](https://doi.org/10.1126/science.aaa6146)

# A vignette: asking a predictive question

---

Will a patient develop heart disease?

# Where can I access compute resources?

---

Install the data science software stack on your local desktop or laptop:

1. [Mac OSX](#)
2. [Linux \(Ubuntu\)](#)
3. [Window 10](#)

# Where can I access compute resources?

---

Or take advantage of a University-run JupyterHub:

1. For UBC: <https://ubc.syzygy.ca>

Login with your UBC cwl credentials.

(see [here](#) for additional information, and for other institutions)

# Where can I access compute resources?

---

If you need more compute power:

1. UBC has an Advanced Research Computing (ARC) unit: <https://arc.ubc.ca/>

ARC provides support to access Compute Canada's high-performance computing resources, as well as training on using their resources.



# Where can I access compute resources?

---

If you need people and compute resources to help you get started on a project:

1. Submit a proposal to UBC's Amazon Cloud Innovation Centre: <https://cic.ubc.ca/>
2. Submit a proposal to be a partner for UBC's Master of Data Science capstone project: <https://ubc-mds.github.io/capstone/about/>
3. Get support from the Applied Statistics and Data Science Group (ASDa): <https://asda.stat.ubc.ca/>

**THANK-YOU.**

MIKE GELBART, PH.D. & TIFFANY TIMBERS, PH.D.