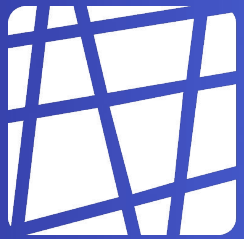
The Build Fellowship

BUILDFELLOWSHIP.COM



Open Avenues

Weekly Updates

- Please provide a quick update on either:
  - Something you did/saw this week that you thought was interesting
  - What you're looking forward to about this week's workshop

(Reminder - please have your cameras on if possible)

The Build Fellowship

OpenVenues

The Build Fellowship

Workshop 3
Data Analysis & Visualization

Recap

Sessions Overview

- Workshop 1 – Project Introduction & Setup
- Workshop 2 – Genomic Data (A2 Assignment)
- **Workshop 3 – Data Analysis & Visualization (A3 Assignment)**
- Workshop 4 – Featurization & Baseline Modeling (A4 Assignment)
- Workshop 5 – Model Training Approaches (Final Assignment Set)
- Workshop 6 – Model Tuning
- Workshop 7 – Performance Evaluation (Final Assignment Code/Testing Due)
- Workshop 8 – Results Presentation & Wrap up (Final Presentation Due)

Data Collection

1. Reviewed publicly available genomes & AMR data on BV-BRC
2. Downloaded metadata on AMR tests
3. Downloaded genome metadata
4. Subset to Escherichia coli + cefepime only
5. Download genomic sequences
6. Download CARD data

Assignment:

7. Aligned the CARD gene sequences against our genome sequences

Data used this Workshop

AMR Data (subset to E coli + cefepime) - downloaded and reviewed in class last week

Genome Summary Data (subset to E coli) - downloaded and reviewed in class last week

Aligned Genes to Genomes (from assignment):

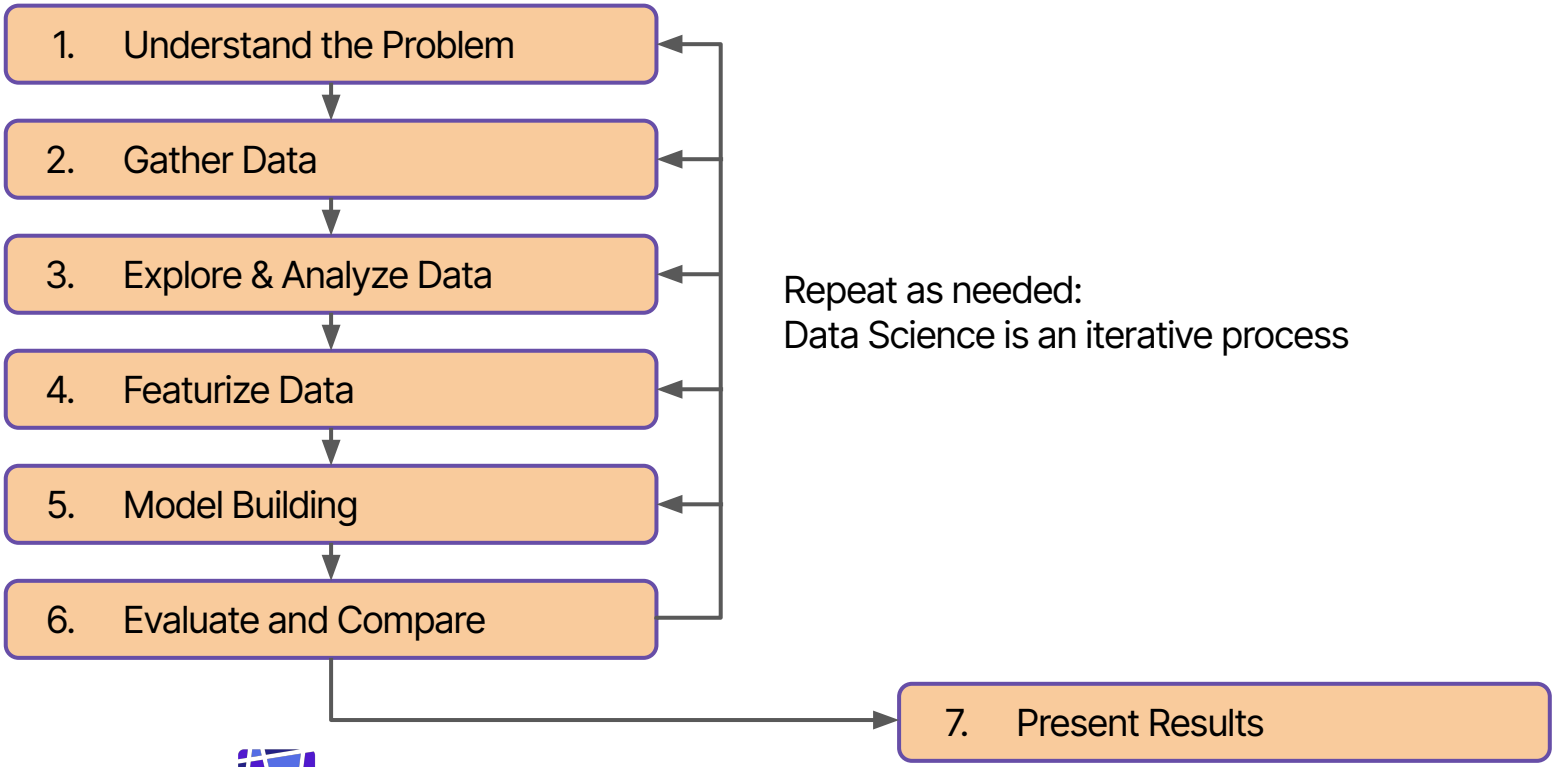
	ref_name	contig	res_gene	match_start	match_end	match_qual	query_str
0	562.42782	562.42782.con.0004	gb|U00096.3|-|3324062-3324911|ARO:3003386|Ecol...	96839	97688	849M	ATGAAACTCTTTGCCCAGGGTACTTCACTGGACCTTAGCCATCCTC...
1	562.42782	562.42782.con.0009	gb|AP009048.1|+|3760295-3762710|ARO:3003303|Ec...	61667	64082	2415M	ATGTCGAATTCTTATGACTCCTCCAGTATCAAAGTCCTGAAAGGGC...
2	562.42782	562.42782.con.0030	gb|BA000007.3|+|4990267-4994296|ARO:3003288|Ec...	21749	25778	4029M	TTACTCGTCTTCCAGTTCGATGTTGATACCCAGCGAACGAATCTCT...
3	562.42782	562.42782.con.0001	gb|U00096.3|-|2336792-2339420|ARO:3003294|Ecol...	153373	156001	2628M	TTATTCTTCTTCTGGCTCGTCGTCAACGTCCACTTCCGGAGCGATT...
4	562.42782	562.42782.con.0004	gb|AP009048.1|-|3172159-3174052|ARO:3003316|Ec...	242171	244064	1893M	ATGACGCAAACTTATAACGCTGATGCCATTGAGGTACTCACCGGGC...

Data Science Process

The Build Fellowship

OpenAvenues FOUNDATION

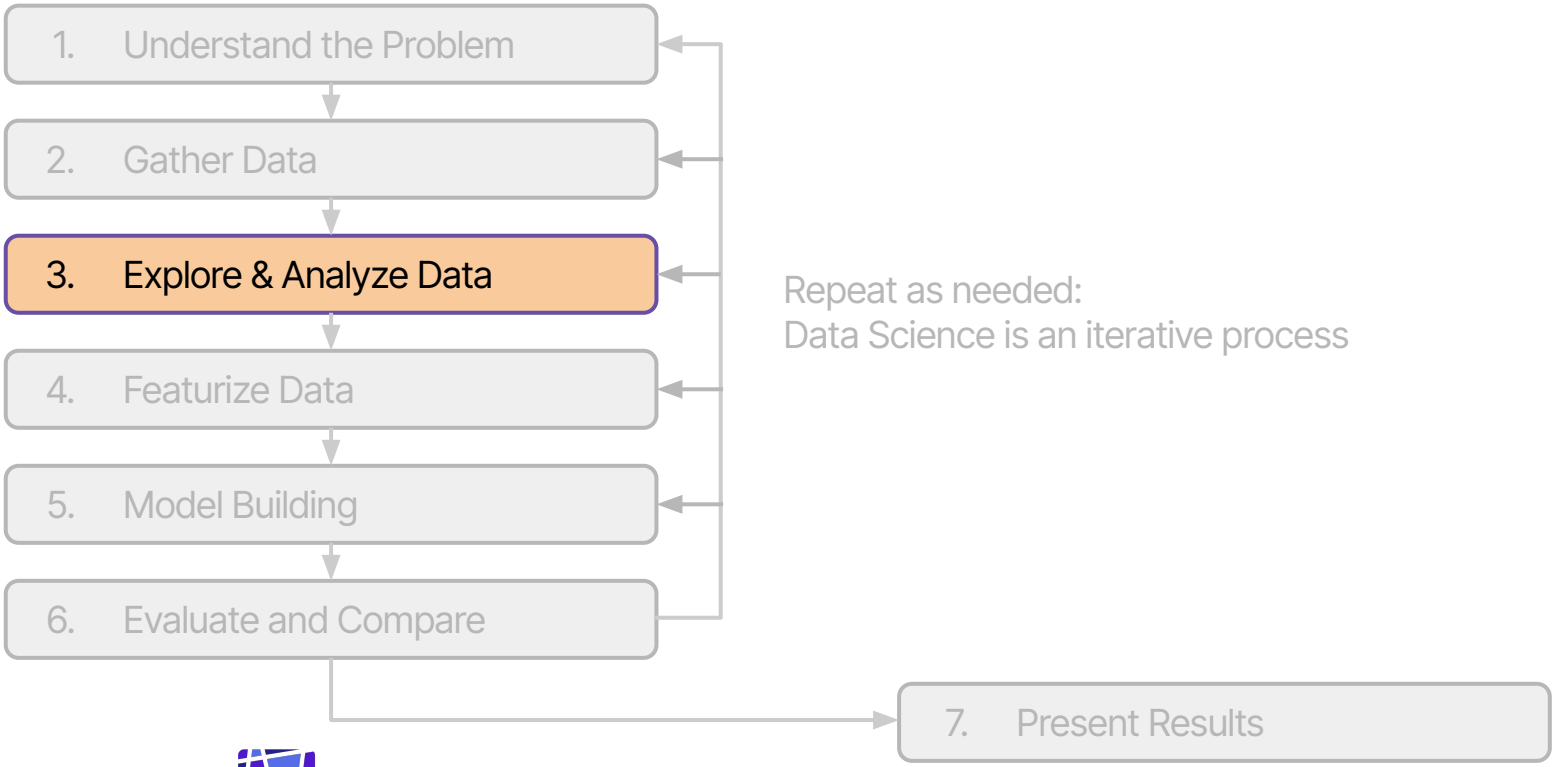
What is the Data Science Process?



The Build Fellowship

OpenAvenues FOUNDATION

What is the Data Science Process?



The Build Fellowship

OpenAvenues FOUNDATION

QUIZ TIME !?

Why should we visualize data?

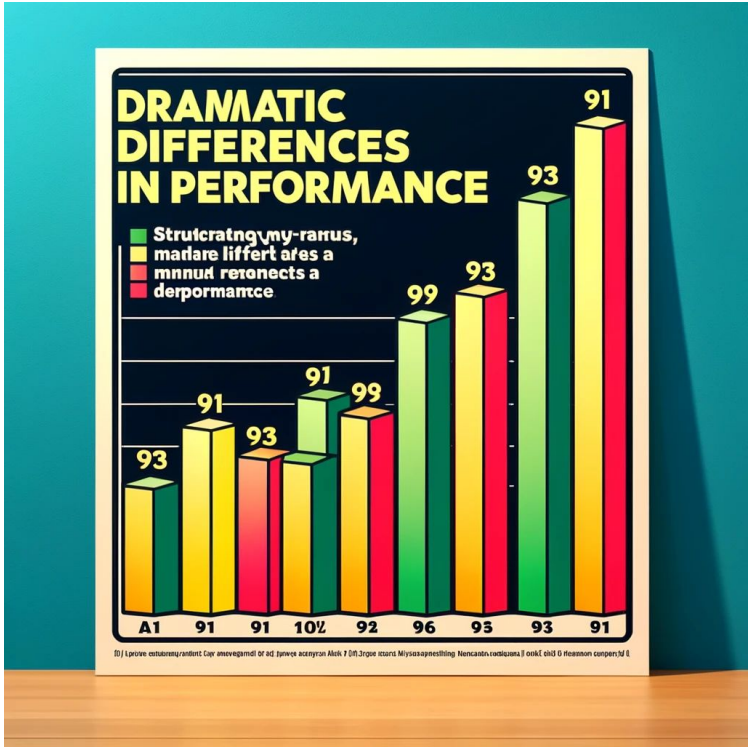
a) To make complex patterns easier to understand
b) To help us form hypotheses
c) To hide or manipulate information
d) To summarize and present information to others

The Build Fellowship

OpenAvenues

Why is EDA Important?

1. Data Quality
2. Understanding Patterns
3. Generating Hypotheses
4. Informing Model Selection
5. Communicating Findings



ChatGPT can make some really misleading graphs

The Build Fellowship

OpenAVenues FOUNDATION

Best Practices

EDA approaches will inherently vary depending on the type of problem and data.

A few common best practices:

- Take a manual look through your data
- Summarize your descriptive statistics (missingness, counts, data types)
- Use a combination of tables & visualization
- Don’t be afraid to iterate

When building a visualization also:

- Think about your target audience (scientific vs consumer)
- Formulate your question ahead of time
- Decide which specific facet you’re trying to visualize (comparison, distribution, time series)
- Make it understandable (always add titles, label your axes)
- Summarize your interpretations

Python Libraries

The Build Fellowship

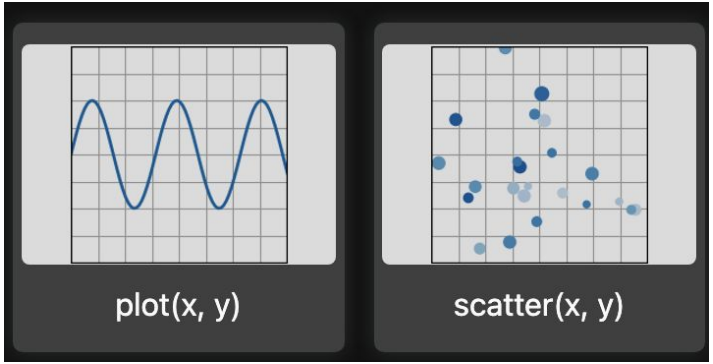
OpenAvenues FOUNDATION

Matplotlib & Seaborn

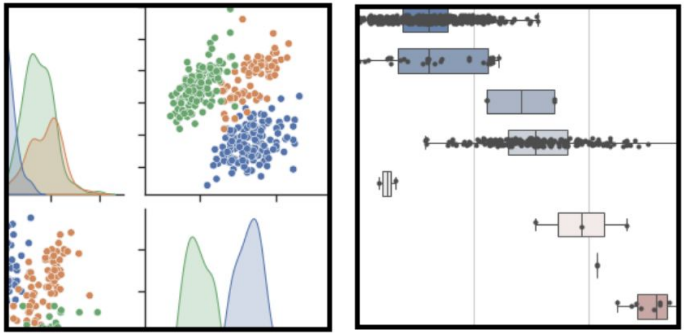
Matplotlib is the most ubiquitous plotting library for Python
- Highly flexible
- Low level

Seaborn is a fantastic extension to Matplotlib that provides a higher level interface for common statistical visualizations
- Easy interfaces
- Attractive visuals
- Good summarization

We'll be leveraging both together in this workshop (see documentation for more details)



[1]



[2]

Alternatives

Whilst we'll be using Matplotlib + Seaborn there are a huge array of tools & libraries available for exploring and visualizing data, many of which serve a specific niche or purpose:

Plotly - Interactive visualizations

Altair - Declarative python library

Tableau/Power BI - Interactive dashboards

R + RShiny - Statistical analysis & dashboarding

D3 - Complex & highly interactive visualizations (JavaScript)

Examples for Genomic Data

The Build Fellowship

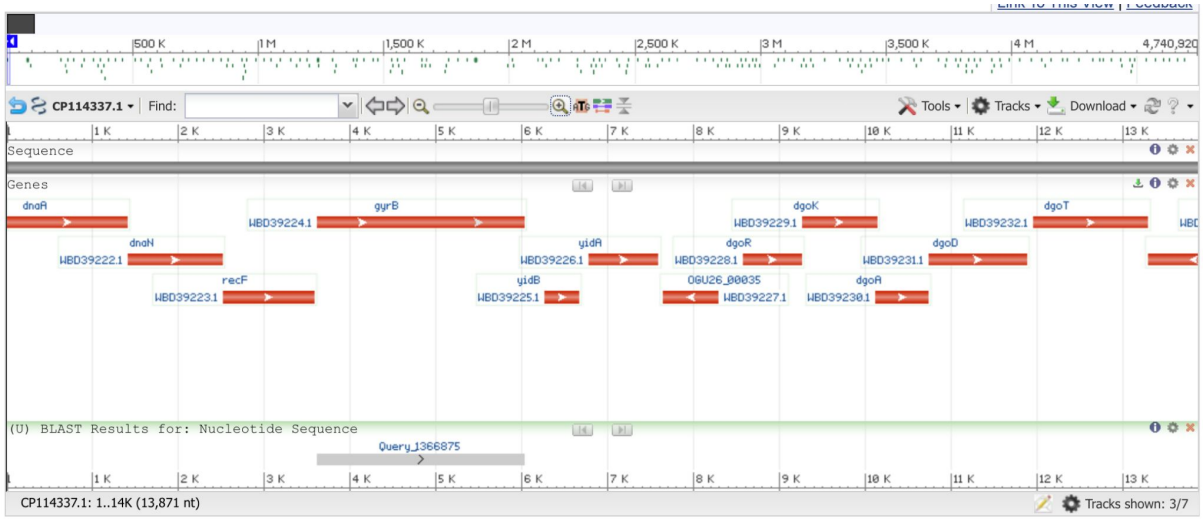
OpenAvenues FOUNDATION

NCBI Blast

In the assignment last week, you aligned resistance genes to the Escherichia coli genomes

NCBI Blast is a tool for doing the same but with a webportal and a set of great genomic visualizations:

- As an example I took a random genes from our CARD data and ran it through Blast
- Check it out at [this link]



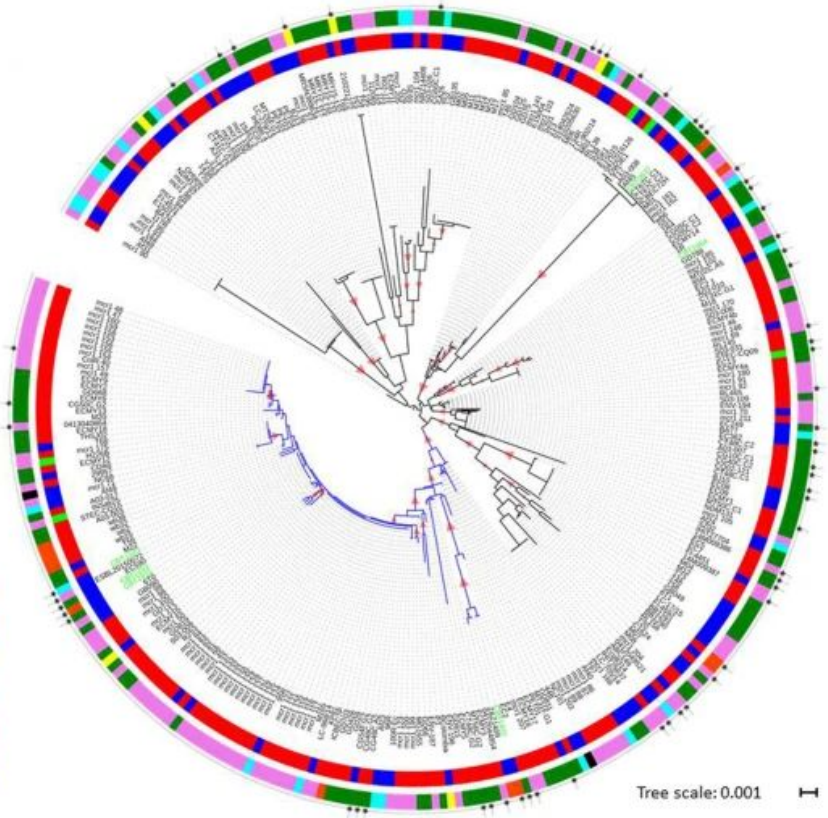
[3]

Escherichia coli Phylogeny

In this build project we're working with Escherichia coli

It's common to want to understand the grouping and relationships between strains

This circle of doom on the right represents a subset of the E coli "Phylogeny"



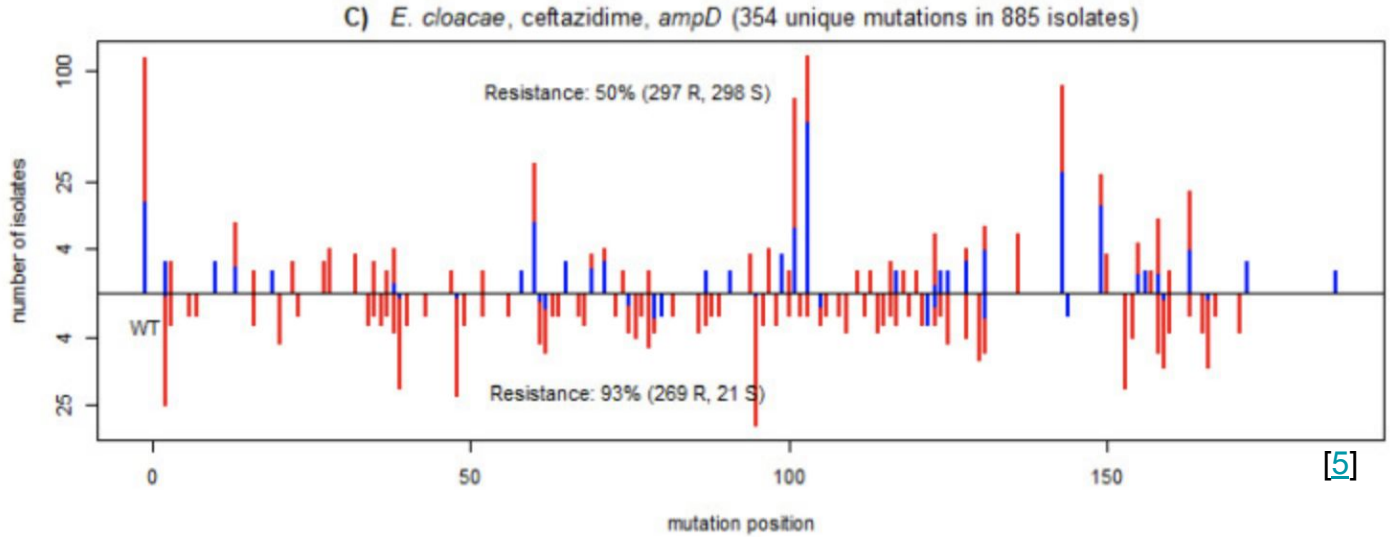
The Build Fellowship

OpenAV

Gene Mutations

In a 2021 publication Majek et al. showed a really interesting relationship between mutations in a gene and AMR phenotype

- Great use of positional information in X (gene position) and Y (counts)
- Color to show S vs R phenotype



[5]

Workshop 3
EDA

The Build Fellowship

OpenAvenues FOUNDATION

References

[1]:https://matplotlib.org/stable/plot_types/index.html

[2]:https://seaborn.pydata.org/

[3]:https://blast.ncbi.nlm.nih.gov/Blast.cgi

[4]:https://www.nature.com/articles/s41598-017-15539-7

[5]:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8657983/