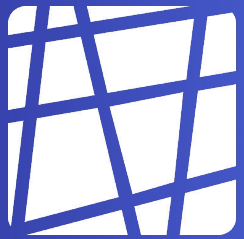
The Build Fellowship

BUILDFELLOWSHIP.COM



Open Avenues

The Build Fellowship

Exploring the Bacterial Genome using Data Science

Introductions

The Build Fellowship

OpenAvenues FOUNDATION

Who am I?

Hayden Sansum

Lead Data Scientist at Day Zero Diagnostics

Data Science professional for 9+ years

Interests

- Coffee (drinking and making)
- Hiking & Skiing
- Miniature painting
- Baldur's Gate 3
- Sewing

Who are you?

- Want to encourage open discussion during these workshops
  - We will start every session with everyone giving an update on their week!
  - For each session come prepared with anything you did/saw that you thought was interesting or what you're looking to learn from the upcoming session

- For today, please tell us a little about yourself!:
  - Name + School & Course of Study
  - What do you want to learn from this workshop?
  - One hobby or an event you enjoyed recently?




The Build Fellowship

OpenAvenues FOUNDATION

Why are we here?

The Build Fellowship

OpenAvenues FOUNDATION

QUIZ TIME !?

What is “Antibiotic Resistance”?

a) When the human body becomes immune to antibiotics
b) When antibiotics become more effective over time
c) When bacteria evolve to survive exposure to antibiotics
d) When antibiotics start a protest

The Build Fellowship

OpenAvenues

Antimicrobial Resistance

"Antimicrobial resistance (AMR) is one of the top global public health and development threats. It is estimated that bacterial AMR was directly responsible for 1.27 million global deaths in 2019 and contributed to 4.95 million deaths" [1]

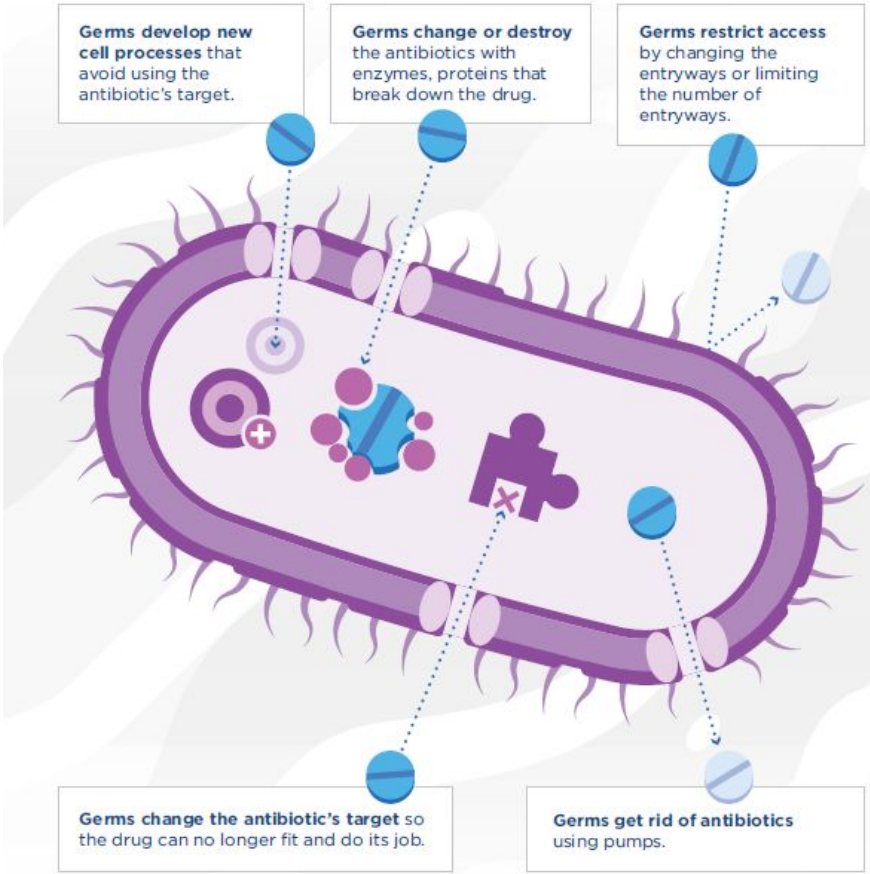
- Antibiotics are a key line of defense
- Even simple cuts or wounds can lead to deadly outcomes
- Use of antibiotics promotes further spread of resistance
- Resistance = harder to treat infections and to predict which antibiotics will be effective

The Build Fellowship

OpenAvenues FOUNDATION

Resistance Overview

- Antibiotics are varied in their mechanistic effect, only certain drugs will be effective against certain species
- Binary Fission - bacteria divide to reproduce, these rapid division cycles can make bacteria susceptive to selective pressure
- Mutations, gene transfer and bacteriophages all contribute to changes to the bacterial genome
- A single base pair change in DNA could be completely inconsequential or result in a modified protein and phenotypic changes



The Build Fellowship

OpenAvenues FOUNDATION

U.S. Department of Health and Human Services
Centers for Disease Control and Prevention

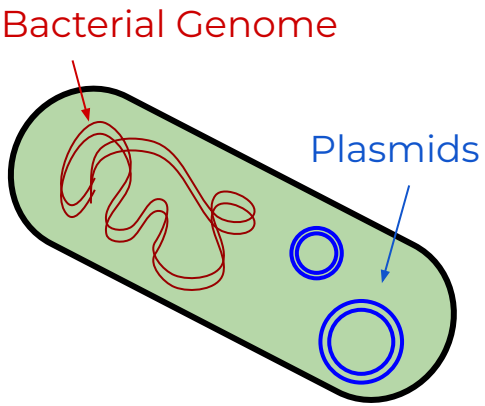
[2]

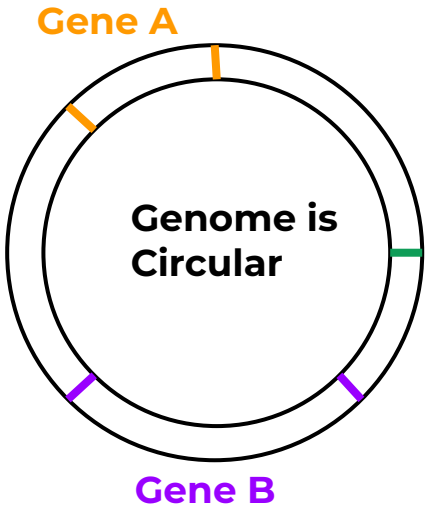
Bacterial Genomes

- AGCT nucleotides make up the fundamental building blocks of DNA
- Strings of nucleotides encode for proteins - functional elements on the bacterial cell
- Bacteria can also gain new pieces of DNA through conjugative transfer - these plasmids can carry resistance genes
- AT and GC form pairs and DNA can be written in forward or reverse complements:

→ AACCCGA

TCGGGTT ←





A : Adenine

C : Cytosyne

G : Guanine

T : Thymine

The Build Fellowship

OpenAvenues FOUNDATION

How does Data Science Help?

- Data Science is uniquely positioned to drive positive change when it comes to AMR and bacterial genomes
- It is very difficult to experimentally determine and prove a causal relationship between a gene and a phenotype (see gene knockout studies [3])
- Data science however is fantastic at picking out correlations amongst large, complex data sets
- Data for supervised machine learning:
  - Bacterial antibiotic resistance can be measured
  - Bacterial genomes can be sequenced into nucleotide strings
- Can we build models which leverage bacterial DNA data to predict AMR phenotype?

The Build Fellowship

OpenAvenues FOUNDATION

Course Overview

The Build Fellowship

OpenAvenues FOUNDATION

Learning Objectives

This is fundamentally a Data Science project and the focus is on breaking down the Challenge into a reusable “data science process”

In data science, understanding the subject of application is always important, so learning about genomics and bacteria is crucial to being able to build a good solution

Objectives:

- Tackle a data science (predictive modeling) problem from start to finish (e.g. a take home interview exercise)
- Set up a reproducible Python environment and understand how to structure code for analysis
- Process tabular data and sequencing data (bacterial genomes) into machine learning features
- Train a well-fit machine learning model and comparison baseline model
- Analyze machine learning models and understand how to use cross-validation and statistical tests to compare between models

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**)

Accessing Files

One single place for accessing all course materials:

- Through the main course folder (link to access here)

Notes:

- Large Data files are only stored in the main course folder (under `data/`)
- Final projects should be submitted in the `submissions/` folder - make a new folder with your name
- Git and Github are not being taught as part of this course. Any existing familiarity?
    - Encourage you to use Git & Github for packaging and presenting your work (especially the final project)

Expectations

The Build Fellowship

OpenAvenues FOUNDATION

Office Hours and Getting Help

Three main methods for support during this course:

1. Office hours:
   - I'm available **3pm-4pm Fridays** and **6pm-7pm Tuesdays**
   - Reach out on slack (see below) or via email to book a slot in these times

2. Slack
   - There is a course slack channel, join using this link
   - Please join the "march-2025-workshop" channel
   - Use this channel for all questions, you are also encouraged and welcome to answer each others questions

3. Recordings of each workshop are available to rewatch in the course folder

The Build Fellowship

OpenAvenues FOUNDATION

Requirements for Course Completion

To formally complete the course you must:

1. Attend at least **5/8 workshops** throughout the course (attending all 8 is highly encouraged!)
2. You must **submit your final project assignment** (this will be set on Workshop 5 and will be due for Workshop 8)

Each workshop is split into:

1. Career discussion & questions
2. Recap / Teaching Component (slides)
3. Interactive Tutorial (jupyter notebook)
4. Assignment (jupyter notebook)

Assignments are designed to be relatively short and example code is provided

The Build Fellowship

OpenAvenues FOUNDATION

Workshop 1 Setup

The Build Fellowship

OpenAvenues FOUNDATION

Set up File Structure

To make everything work smoothly, use the same directory structure as present in the course folder.

To get set up please:

1. Make a top level folder with the course name (all course files will live in here)
2. Make an empty ‘data’ folder within this top level folder (don’t download the data folder!)
3. Download Workshop 1 from the shared folder and place in the top level
4. Download the environment yaml file (first try the one called “etbg_env.yml” - if using windows you may need to use “etbg_env_nobio.yml”)

Resulting Structure:

- ExploringTheBacterialGenome
  - data
  - Workshop 1 - Introduction & Setup
  - etbg_env.yml

The Build Fellowship

OpenAvenues FOUNDATION

Assignment - 0

The Build Fellowship

OpenAvenues FOUNDATION

What do you aim to achieve?

It's important to put yourself in the right mindset and think about what you're looking to learn!

I'm here to support you and help you get the most out of this project and reach your goals.

Prior to next week:

- Write up a short paragraph on what you think you will learn from this course
- Imagine you will be adding this to your resume/portfolio
- You'll come back to this paragraph at the end of the project as a reflection

Send this to me by next Friday (via email)

GitHub Usage

Quick Poll - What is everyone's experience level with version control?

- I won't be formally teaching Git or using GitHub but it is a great habit to get into
- Used extensively in industry

Option:

- Push submissions and weekly progress to GitHub
- Pair up to perform code reviews
- Tag me in pull requests to get feedback

Will still need to upload submissions to the project folders

Final note on Workshops 3 & 5

Due to other commitments I'll be unavailable on:

- **Friday 28th March (Workshop 3)**
- **Friday 11th April (Workshop 5)**

I plan to shift those workshops to another nearby day/time.

I appreciate everyone might not be able to make a new time but I will send out a poll via Slack shortly and try to optimize the adjust scheduling.

Records will, as usual, be available from the project folders.

References

[1] - https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)02724-0/fulltext

[2] - https://www.cdc.gov/drugresistance/about/how-resistance-happens.html

[3] - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10440060/