BIOLOGY & DATA SCIENCE

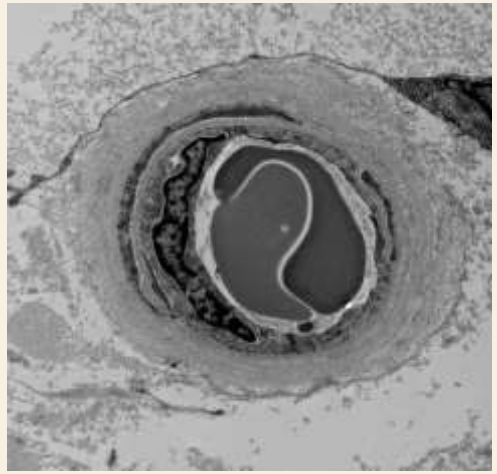
Justin Waterfield, MMCi

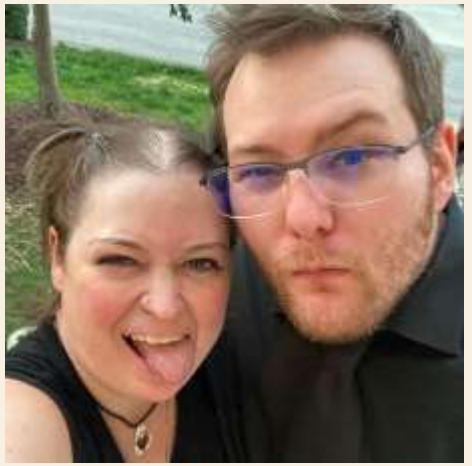
Bioinformaticist
RTI, International
Winthrop, Class of 2009

OUTLINE

- Personal Background & Introduction
- The Data Science Process
- What is Bioinformatics?
- Use Case #1 – Biology
- Use Case #2 – Healthcare
- Informatics TLDR
- Thinking Ahead, Tips & Tricks!
- Closing Thoughts – Q&A

MY STORY…





DATA SCIENCE PROCESS

DATA COLLECTION + STORAGE

Data Engineer

DATA PREPARATION

Data Scientist

EXPLORATION & VISUALIZATION

Data Analyst

EXPERIMENTATION & PREDICATION

Machine Learning Scientist

WHAT IS A BIOINFORMATICIST?

Unofficial Definition:
A scientist or analyst who uses data, often biological in origin, to interpret results and derive insights.

AKA – Someone who knows something about science, biology, or healthcare. Who can use that expertise to interpret data and do something useful with it.

Official Definition:
A discipline that develops software and tools for understanding biological data.

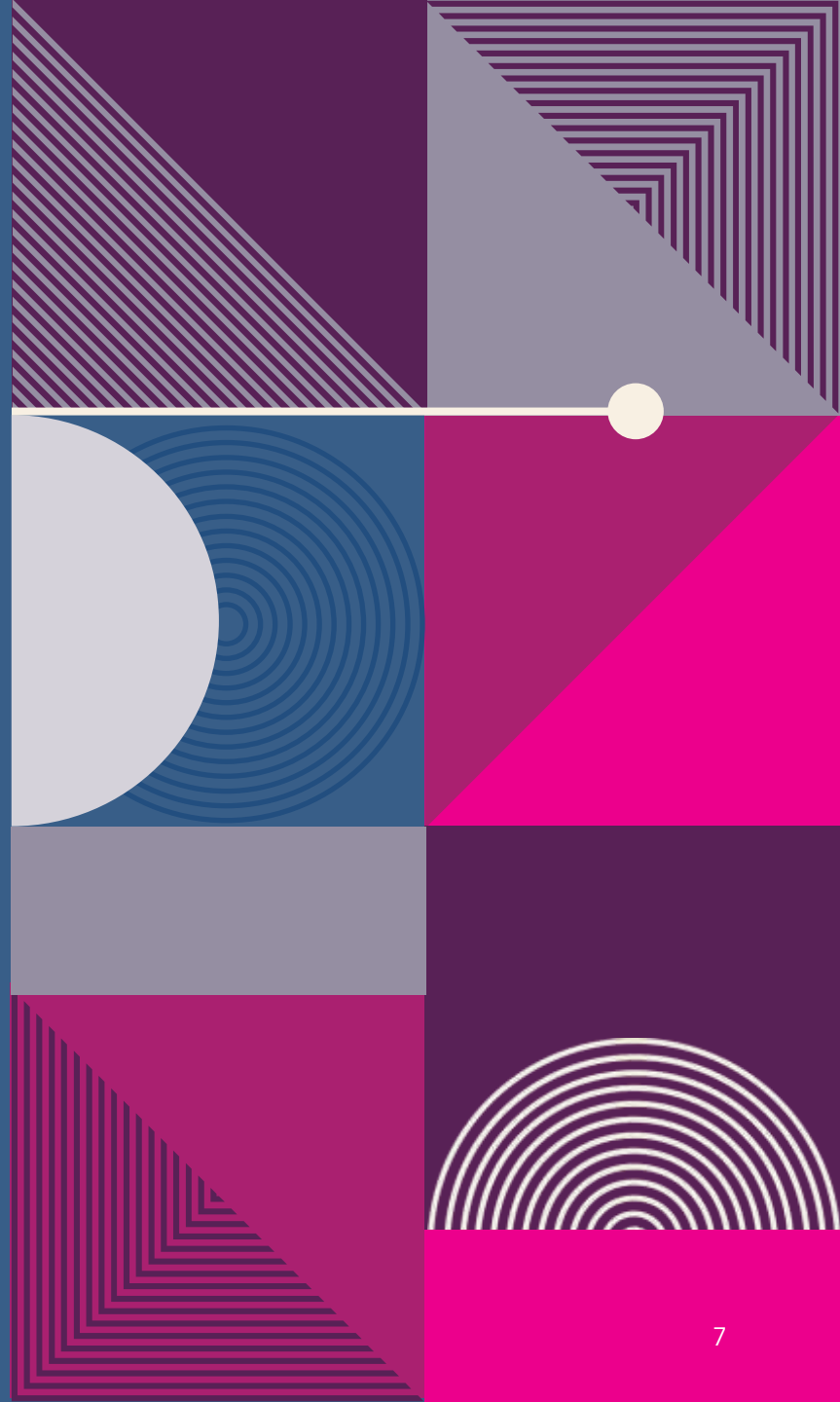
Example:
Genomics, Proteomics, Metabolomics, Pharmacogenomics, etc. etc.

These all fuse the base study (Genetics) + technology (-omics).

SO WHAT?

AKA – Why should you care?

- The future of science, healthcare, and business is all data focused.
- **The average scientist, clinician, and researcher will need to have critical data skills to advance their career.**
- Our generational problems such as cancer, climate change, precision medicine, and artificial intelligence all need data science skills.



USE CASE #1 - BIOLOGY

GENOMIC SEQUENCING

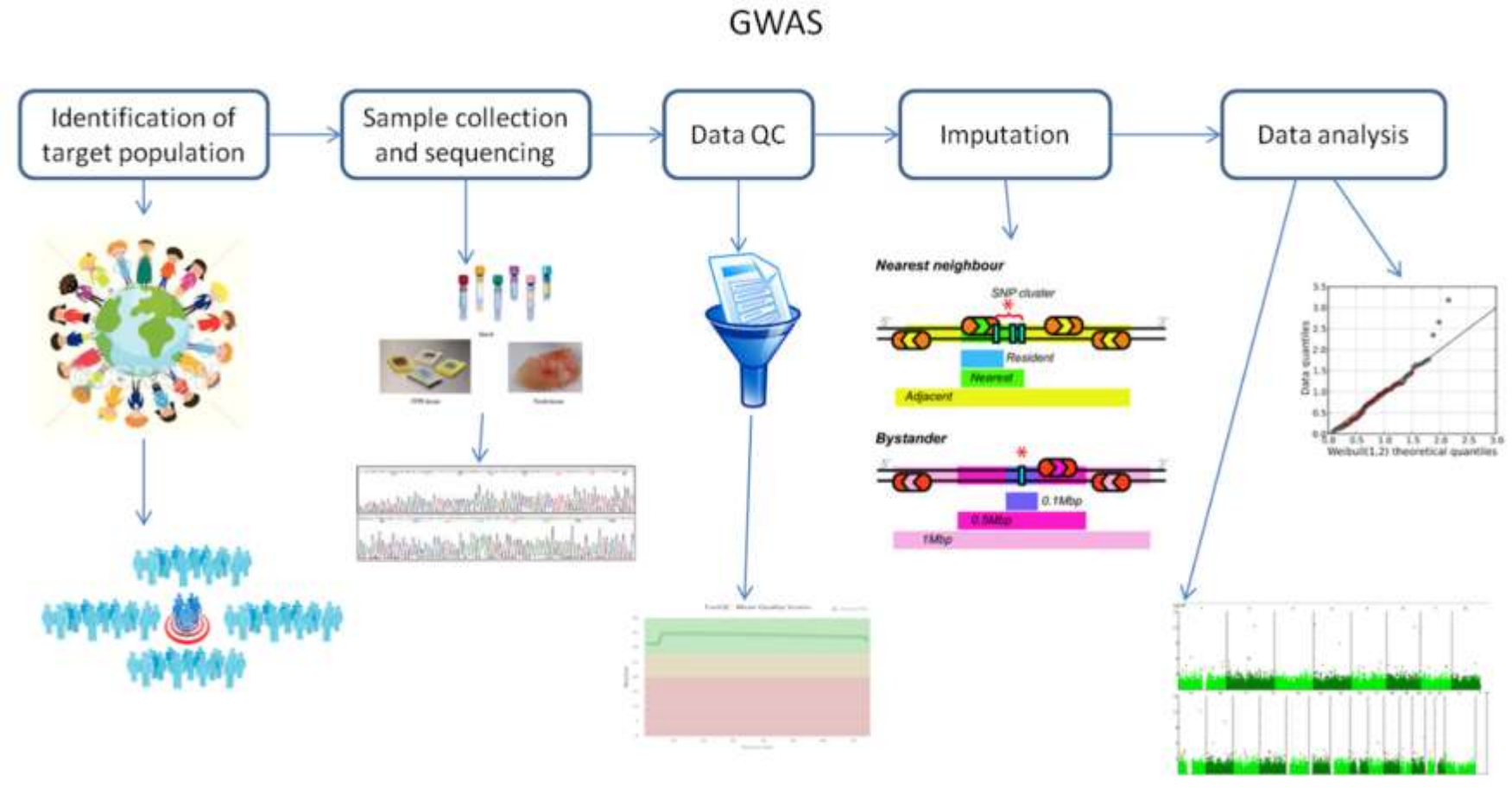
- Genome-Wide Assocation Studies (GWAS)
- Standard approach for querying genetic data for disease risk
- Technological advances have lowered cost (think 23&me)
- National Human Genome Research Institute (NHGRI) publishes large amount of data
- All of Us! → Precision Medicine/Health

DATA CATALOGUING

- GWAS data must be encoded, variables set, phenotypes determined
- What is your population?
- Biases, concerns, variables?
- This is the field of population health, which has a major undercurrent of data literacy
- Can be extrapolated to ecology, evolution, botany, entomology, virology, etc. etc.

Source: Bioinformatics Challenges in Genome-Wide Association Studies (GWAS) | SpringerLink (duke.edu)

GWAS VISUAL



Source: IJMS | Free Full-Text | Critical Analysis of Genome-Wide Association Studies: Triple Negative Breast Cancer Quae Exempli Causa | HTML (mdpi.com)

USE CASE #2 - HEALTHCARE

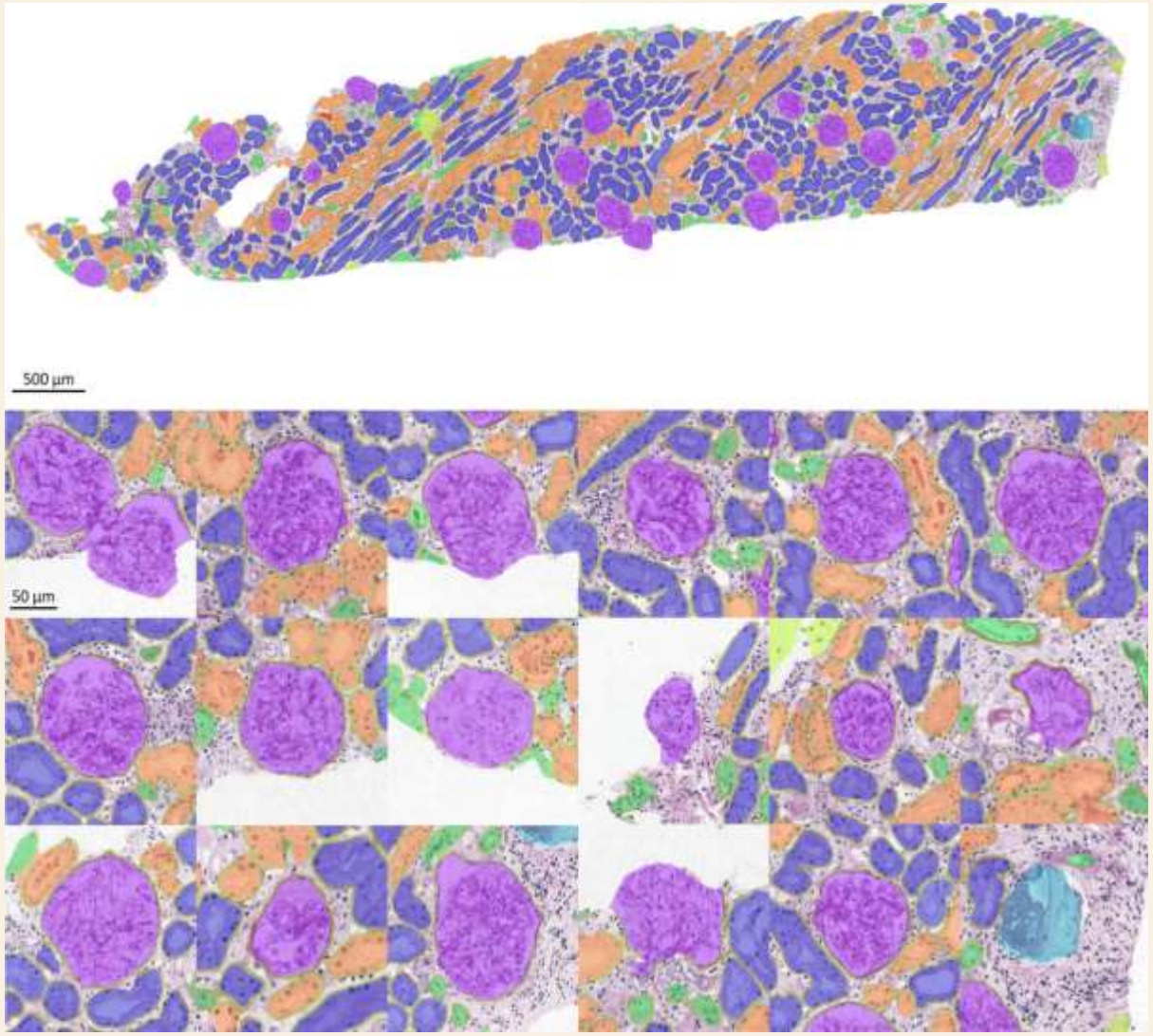
DIGITAL PATHOLOGY

- Pathology is the study of disease processes
- A biopsy (muscle/kidney/brain/etc) is removed and examined for cancer/disease/virus/etc
- Traditionally, that biopsy was viewed microscopically by technologists and pathologists via normal light microscopy
- Now, slides are scanned via slide scanner and enables work from home (WFH) and remote diagnosis. As well as artificial inelligence

LABORATORY DATA + EHR

- The Electronic Health Record (EHR) / Electronic Medical Record (EMR) is the core of all healthcare informatics
- The patient's entire data history/record is stored here
- Your data is stored in terms of variables/resources: Patient / Age/ Name / Medication
- Artificial Intelligence/Machine Learning is huge right now in this area

PATHOLOGY VISUAL



Source: Deep Learning–Based Histopathologic Assessment of Kidney Tissue | American Society of Nephrology (asnjournals.org)

INFORMATICS IN A NUTSHELL

DATA GOVERNANCE

- A good data model goes a long way
- Understanding what, where, and why your data exists
- Data is cheap (usually), time is not
- Privacy and security are crucial, more than ever
- In biology, linking your physical samples to your databases – as well as well defined metadata

DATA QUALITY

- Bad data costs money
- Costs USA ~$3 **trillion** annually!
- Biological data is often uncurated, especially not by experts in specific field
- As databases grow, the problem becomes exponential, but the talent/staff is not growing at same rate
- Scientists rarely agree…

TIPS AND TRICKS!

LEARN 2 CODE

- Start to learn some basic form of programming and coding.
  - R / Python / SQL
  - DataCamp

T-SHAPE

- Science rewards having deep expertise in one area, but a breadth of domain knowledge.

GET INVOLVED!

- Communities, organizations, memberships. Say yes to things…but also know your limits!

FAILURE IS OKAY!

DATA SCIENCE IS STILL SCIENCE!


QUESTIONS?

THANK YOU

Justin Waterfield

www.justinwaterfield.com
www.linkedin.com/in/waterfieldj