Introduction to Data Science

CS3300 Data Science

RJ Nowling

Introductions

- What is your name?
- What is your major?
- What did you do this summer?
- What excites you about Data Science?
- Do you know what you want to do after graduation?

Who am I?

- Dr. RJ Nowling
- Ph.D. in Computer Science & Engineering
  - University of Notre Dame
  - Simulating dynamics of molecules (Molecular Dynamics)
  - Genomes of insect vectors (Bioinformatics)
- Industry Experience
  - 2 years at Red Hat – working on open-source big data platforms
  - 2 years at AdRoll – Data Science Engineering (everything for a real-time recommendation system)
- 1 year at MSOE ☺
- In my spare time, I like to ride my bike and take my dogs to the dog park

Reading

- Chapter 1 of *The Data Science Design Manual*

What is Data Science?

- Extract actionable knowledge from data – Dr. Jay Urbain
- “The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it.”
- Application of the scientific method to data
  - Exploring data to generate hypotheses
  - Evaluating hypotheses with visualizations, predictive models, and statistics
  - Communicating those hypotheses and evidence to others

Modes of Inquiry: Hypothesis Driven

- Traditional scientific method
- We form a hypothesis
- We design an experiment (including collecting data) to test the hypothesis
- If the experiment is able to reject or disprove the hypothesis, we generate a new hypothesis
- Otherwise, we design another experiment to test the hypothesis
- Eventually, if we are unable to disprove the hypothesis, it becomes a law

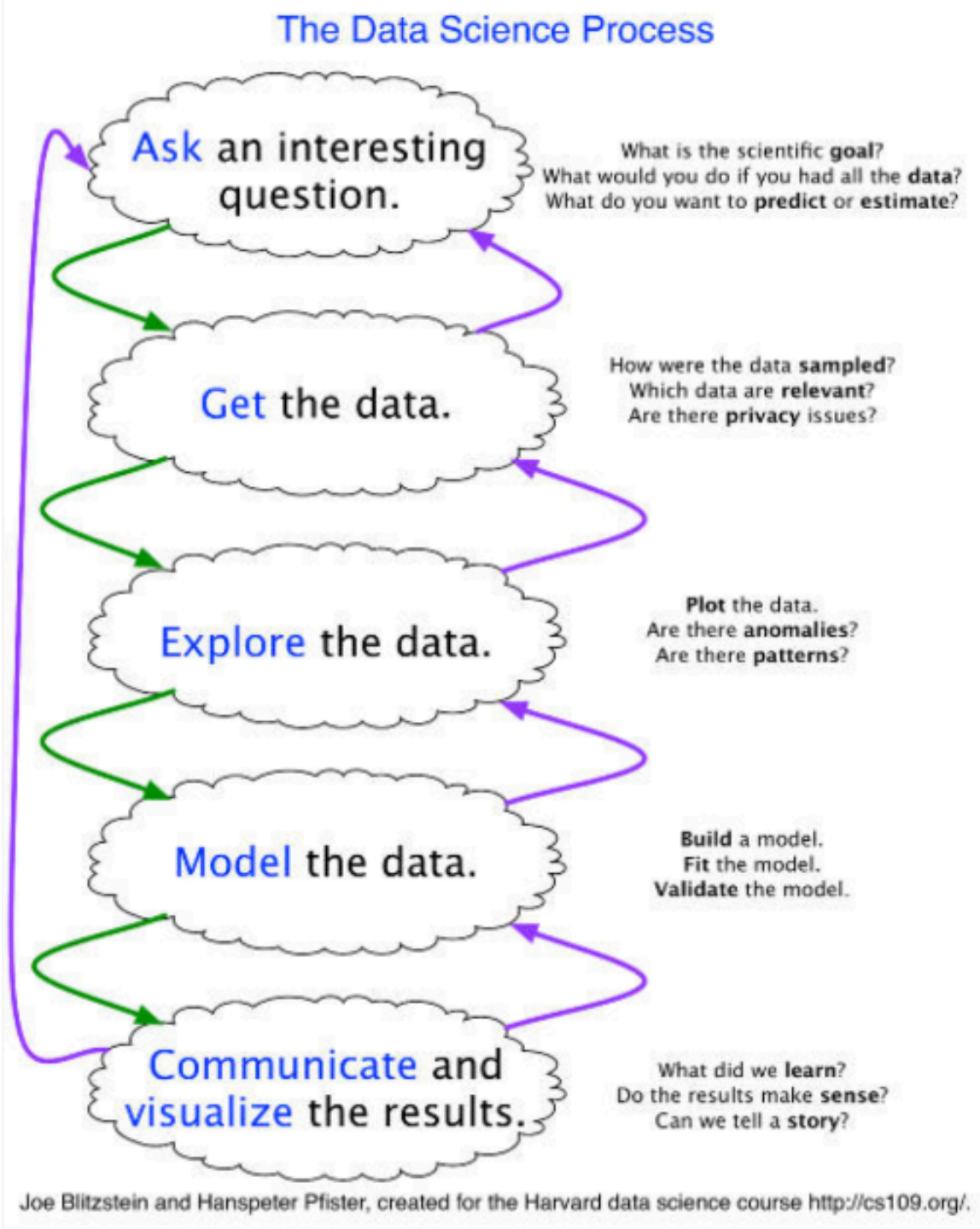
Modes of Inquiry: Method Driven

- Traditional engineering approach
- We focus on understanding a method or technique
- We identify when and how to apply that technique
- We becomes experts at applying the technique

Modes of Inquiry: Data Driven

- We start with a data set
- We explore the data set to identify patterns
- From these patterns, we ask questions and form hypotheses
- We may be able to use the data to answer the hypothesis or may need to design a new experiment
- This is a new mode of inquiry and what makes Data Science different from traditional science and engineering.

Data Science Process



Online Advertising Infrastructure Case Study

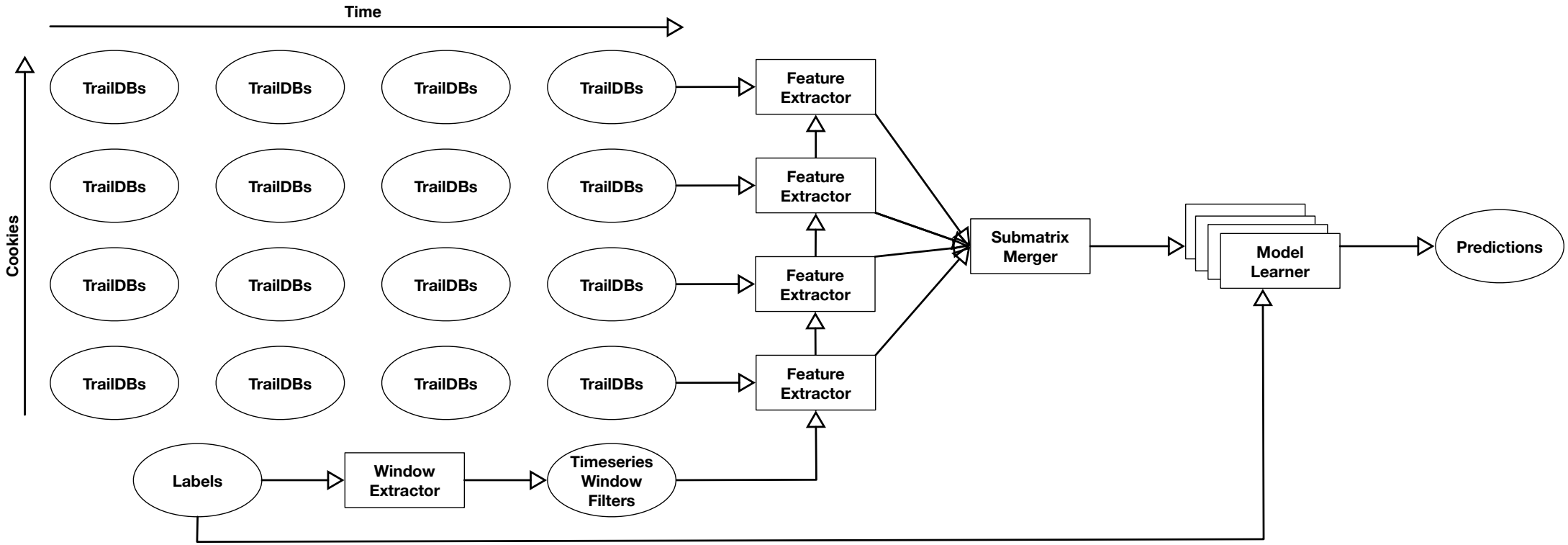
- A is an online advertising platform
- A uses user data (e.g., web pages viewed) from your web site to determine which ads to show users and how much to pay for those ads
- One part of A's system is a recommendation system – used to generate candidate ads for a user which are refined by another system further down the pipeline
- Recommendations generated for 1 billion cookies per day by a batch job

Online Advertising Infrastructure Case Study

Or "why am I getting a paged in the middle of the night saying that our machine learning pipeline hasn't finished running?"

(More profanities in real life…)

Machine Learning Pipeline



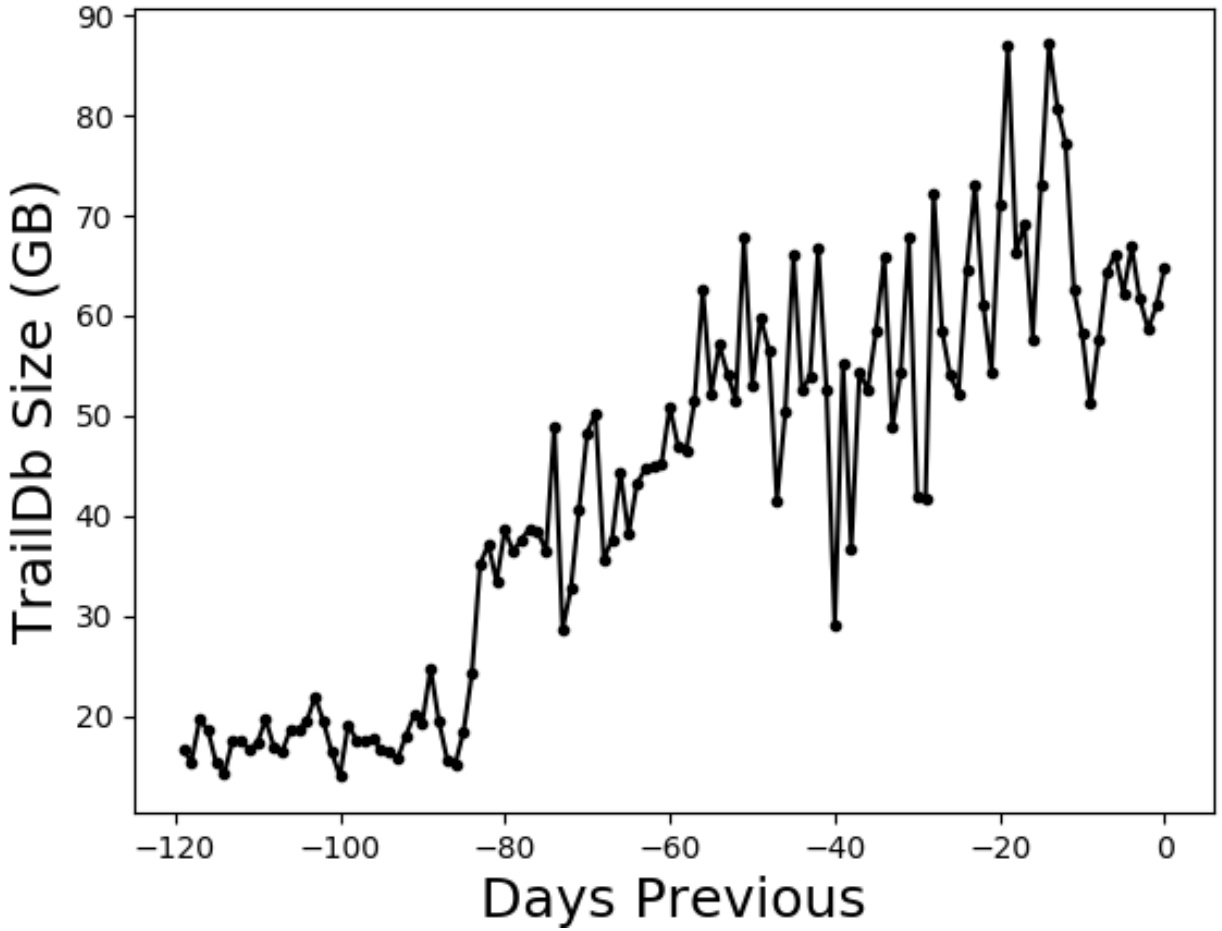
Problem

- We monitor pipeline completion times
- Pipeline run times started going up
- Potential they won't complete in desired period
- More data -> longer run time
- More data -> higher infrastructure costs (e.g., more EC2 instances and higher S3 storage costs)

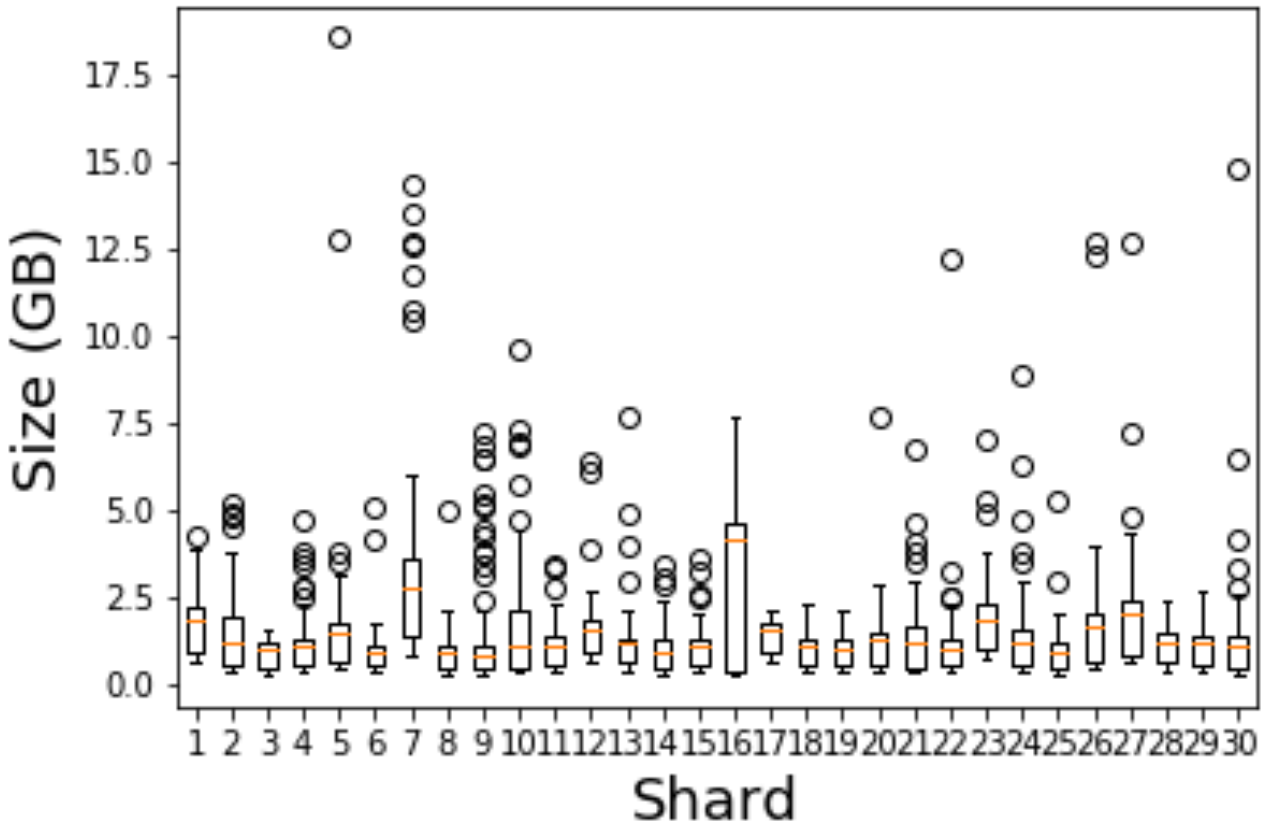
Data Collection

- Customers inject code into their web site to allow A to track visitors' page views
- Record every ad impression and click

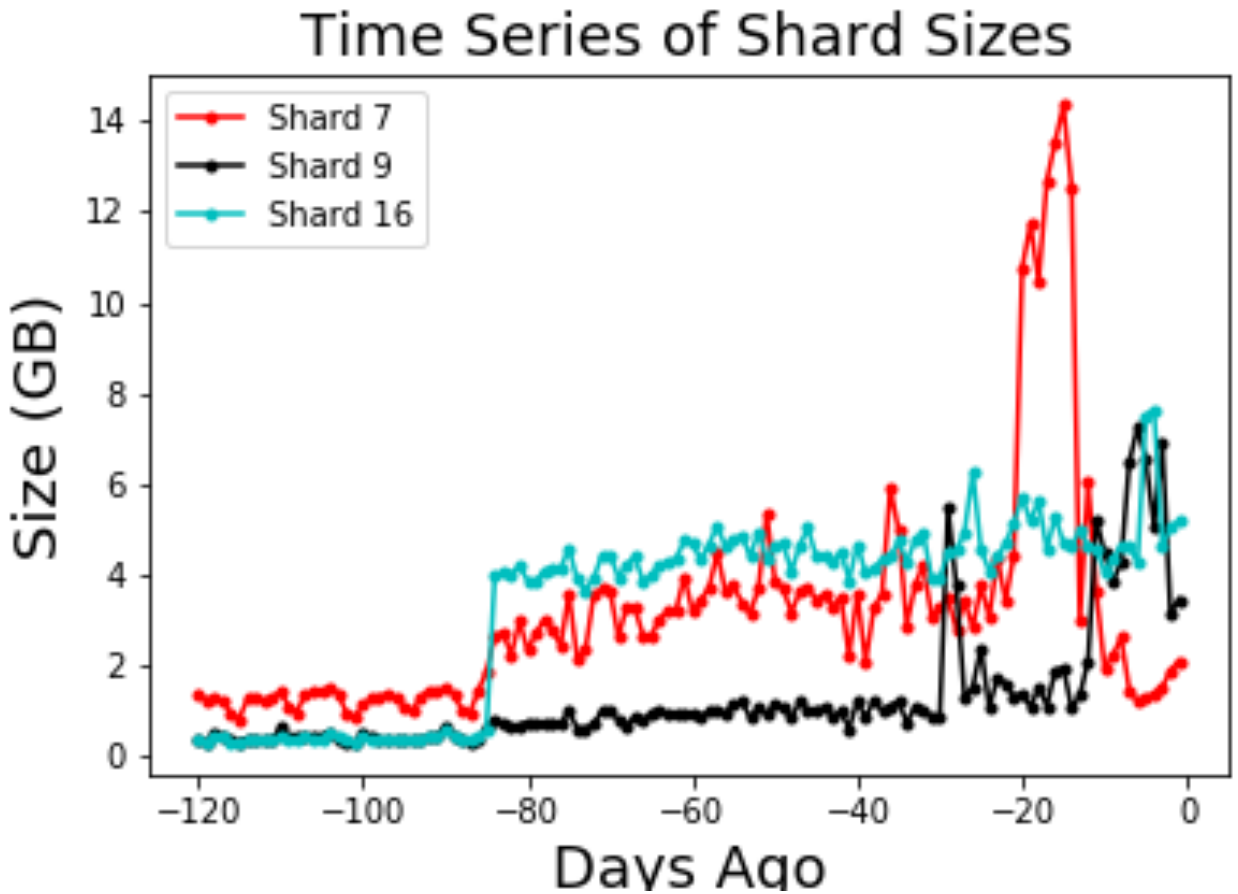
Daily Data Size Over Last 120 Days



Daily Size Distributions by Shard



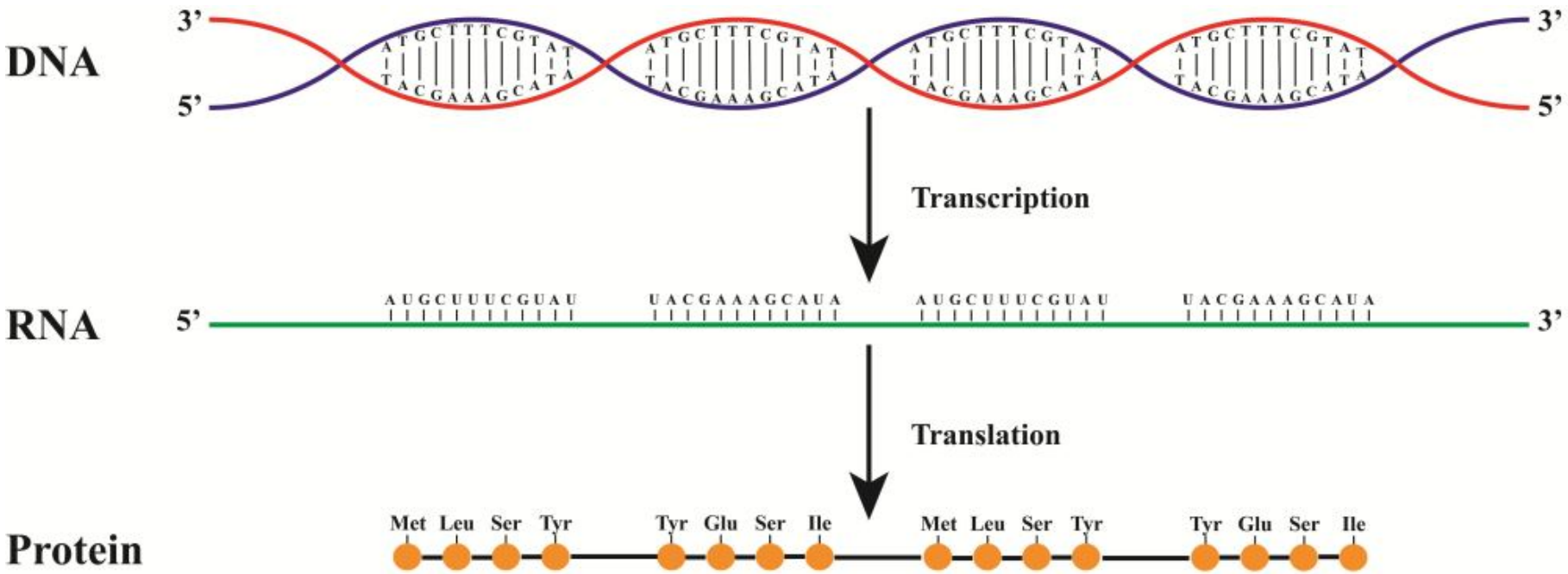
Time Series of Chosen Shards



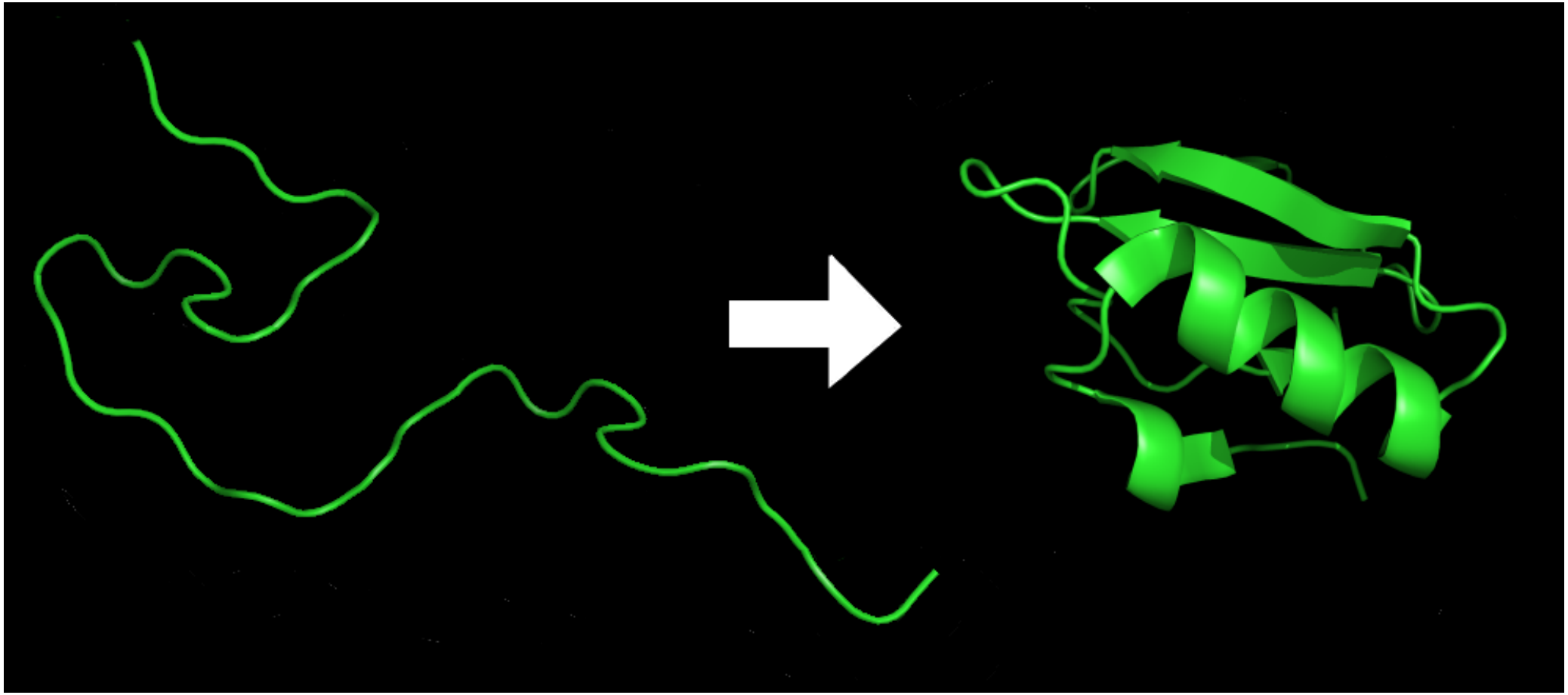
Action Items

- Duplicate data during data model transition
  - Wait until transition is complete
- Mobile advertiser data
  - Filter out data
- Highly-redundant user attribute data from third party vendor
  - Short term: Filter out third-party user attribute data
  - Medium term: Implement lightweight monitoring and alerting system that checks properties of the data over time
  - Longer term: De-duplicate data and make available as source for feature engineering

Biochemistry



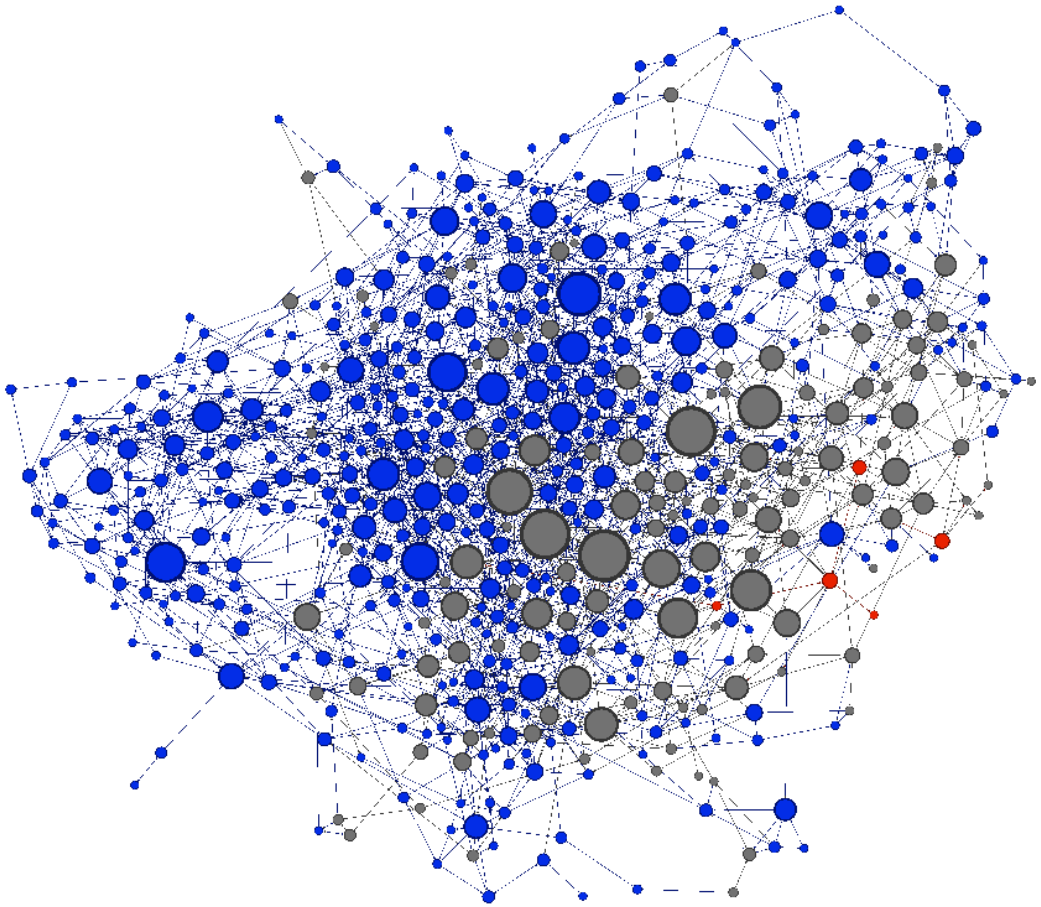
Protein Folding

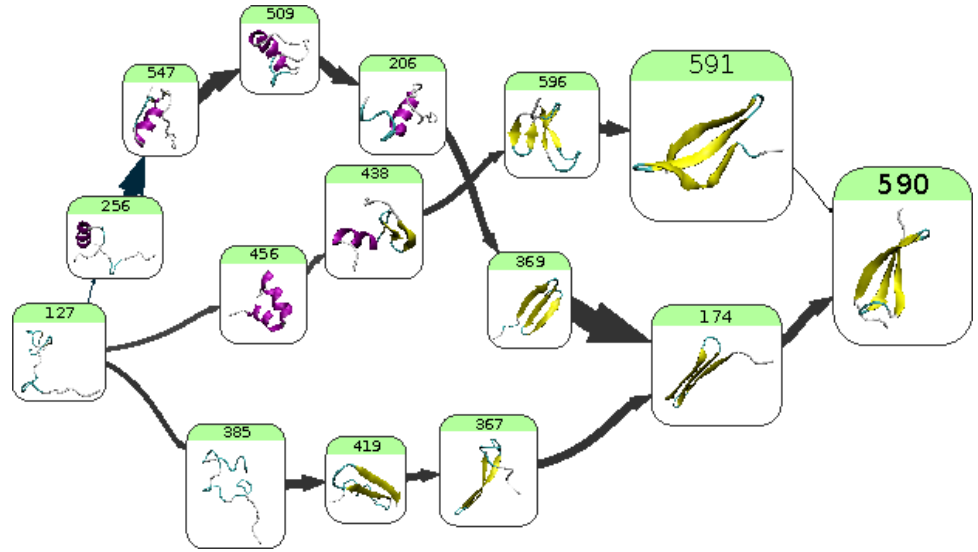


Molecular Dynamics

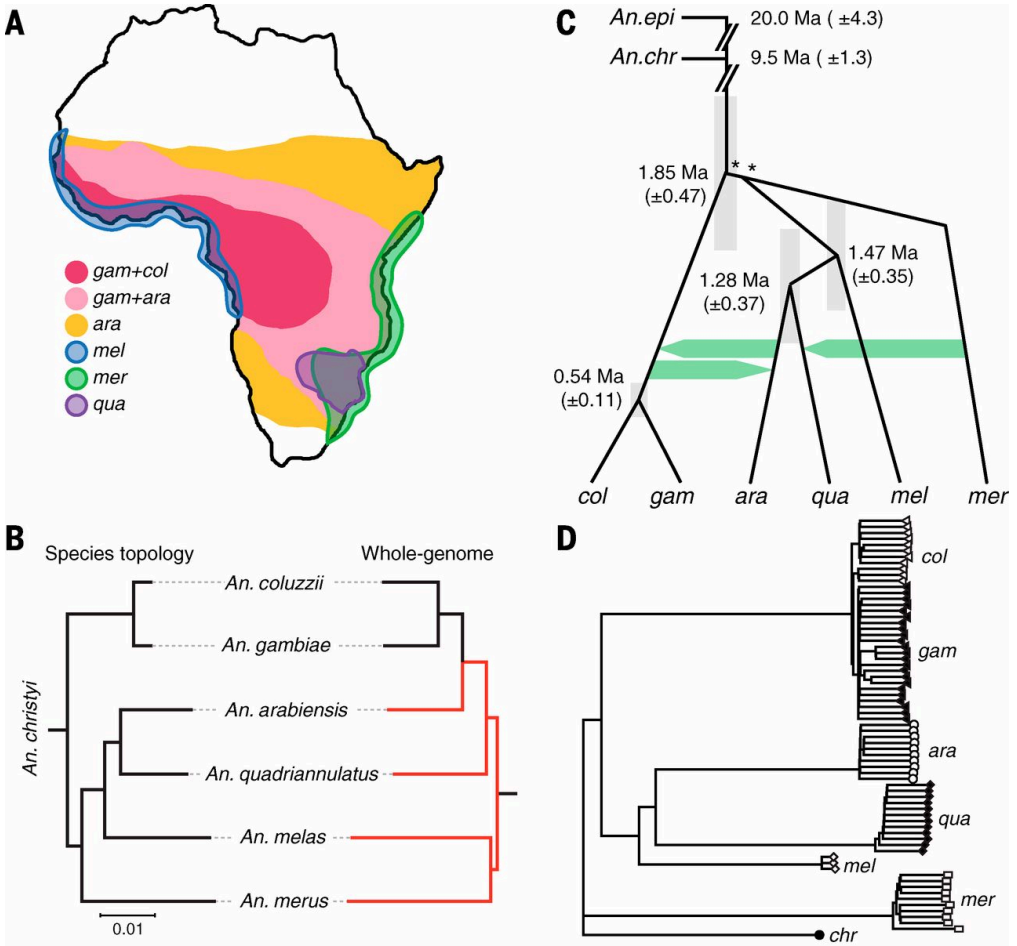
- Simulate dynamics of proteins, liquids, etc. at the atomic level
- NTL9
- B1 domain of Protein G
- Very resource intensive
  - GPUs
  - Thousands of machines
  - Folding@Home
- Generates terabytes of very high dimensional data

Folding Pathways with Clustering and Markov Models

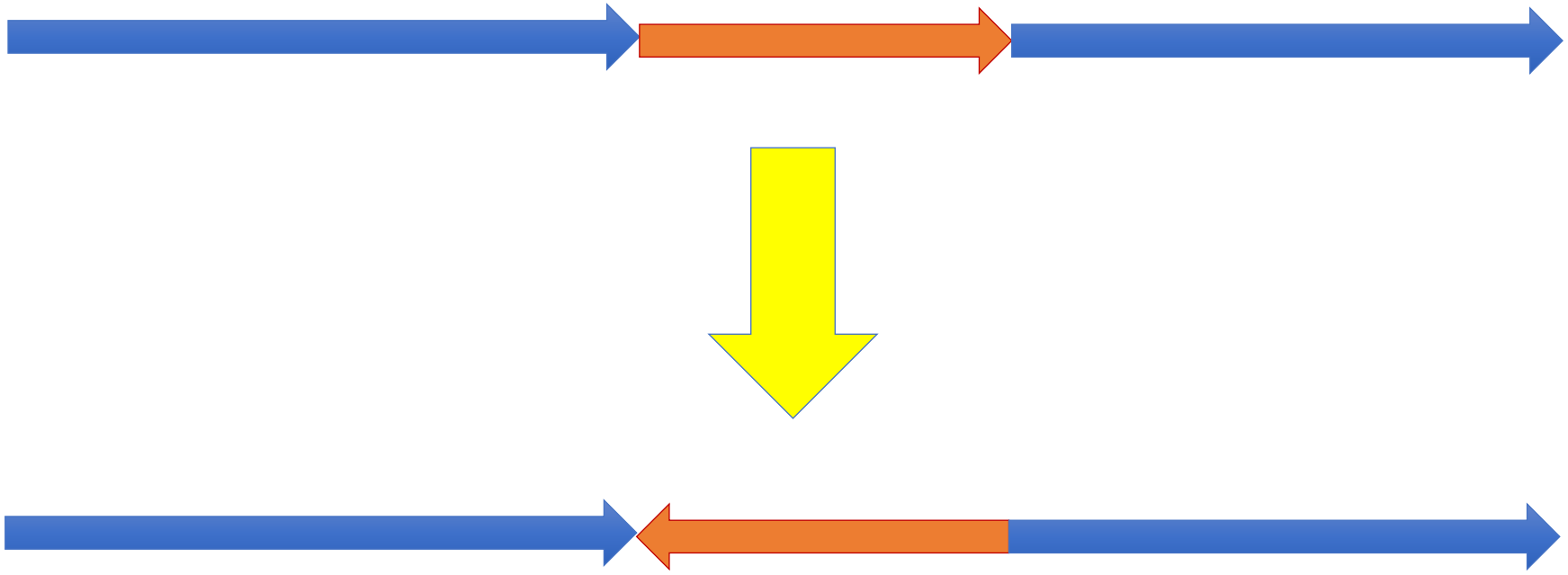




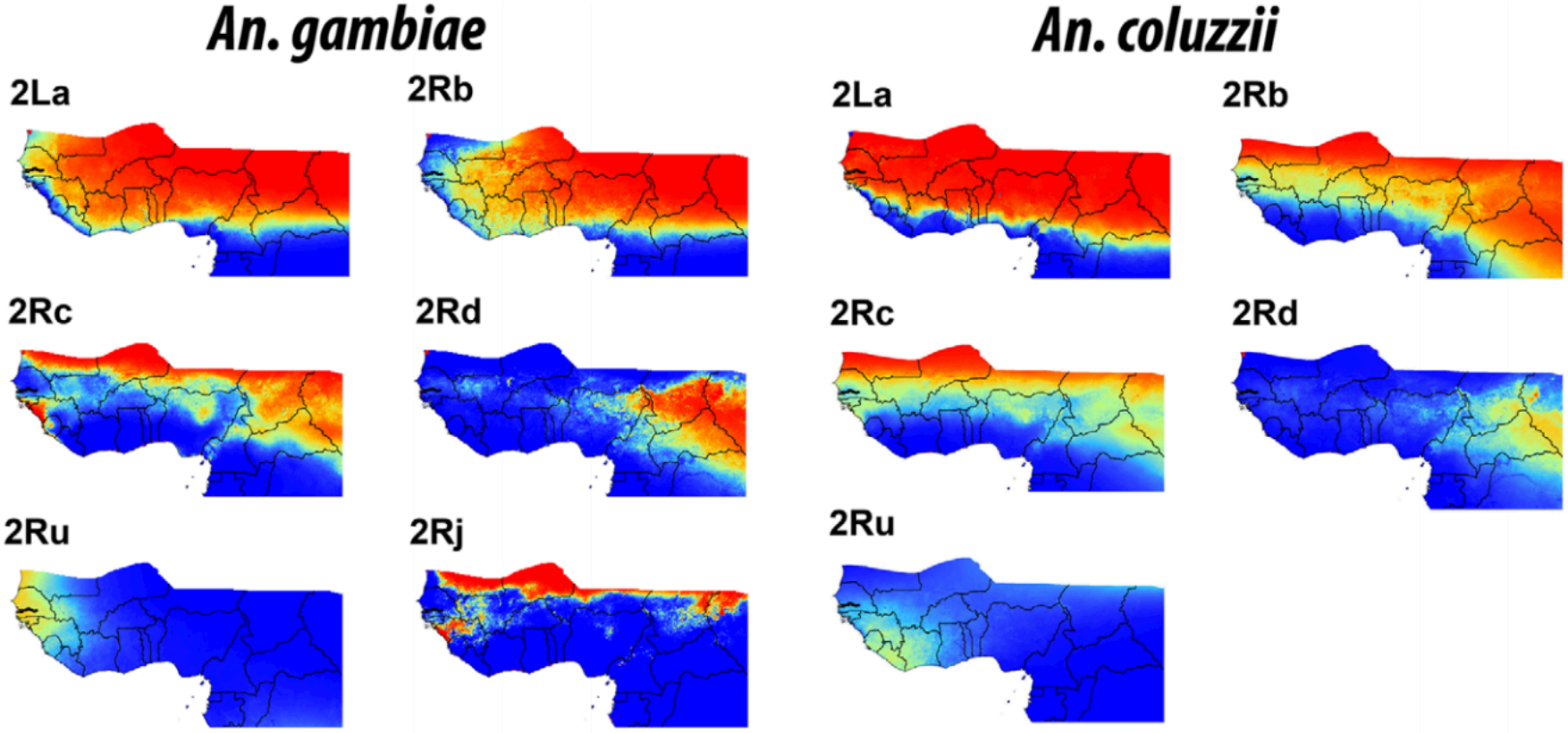
Biology (Genomics)

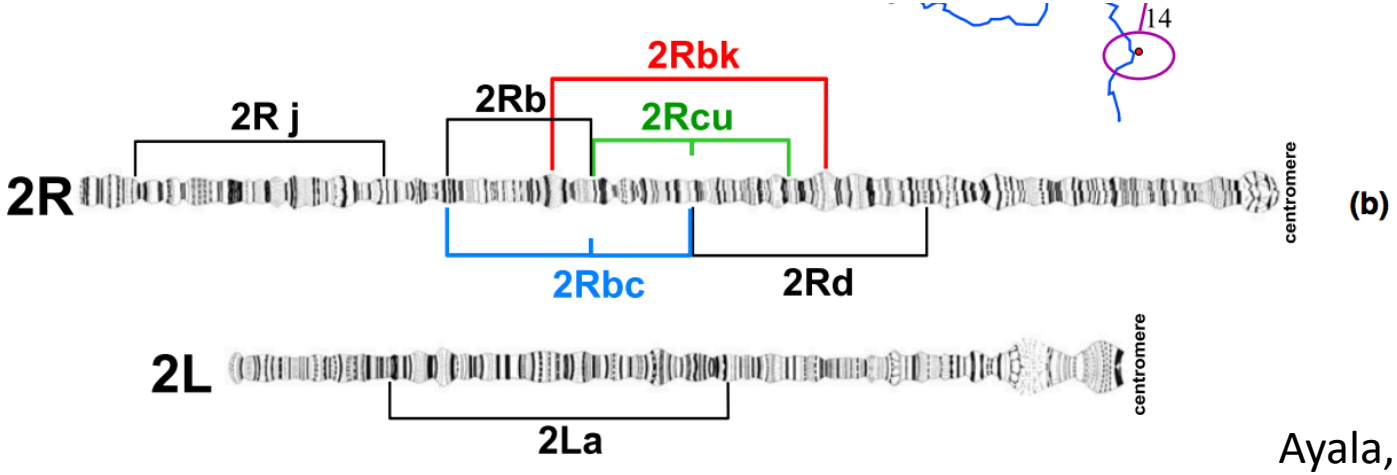


What is an Inversion?



Inversions





Ayala, et al. *Evolution*. 2017; Caputo, et al. *Mol. Ecology*. 2014.

Importance of Inversions

- Thought to play an important role in ecological adaptation by enabling the accumulation of beneficial alleles

  Fuller, et al. Bioarxiv. 2017.; Love, et al. *Mol. Ecol.* 2016.

- 2La in *Anopheles gambiae*
  - Thermal tolerance of larvae
    Rocca, et al. *Malaria Journal*. 2009.
  - Enhanced desiccation resistance
    Gray, et al. *Malaria Journal*. 2009.
  - Susceptibility to malaria parasite species
    Riehle, *Elife*. 2016.

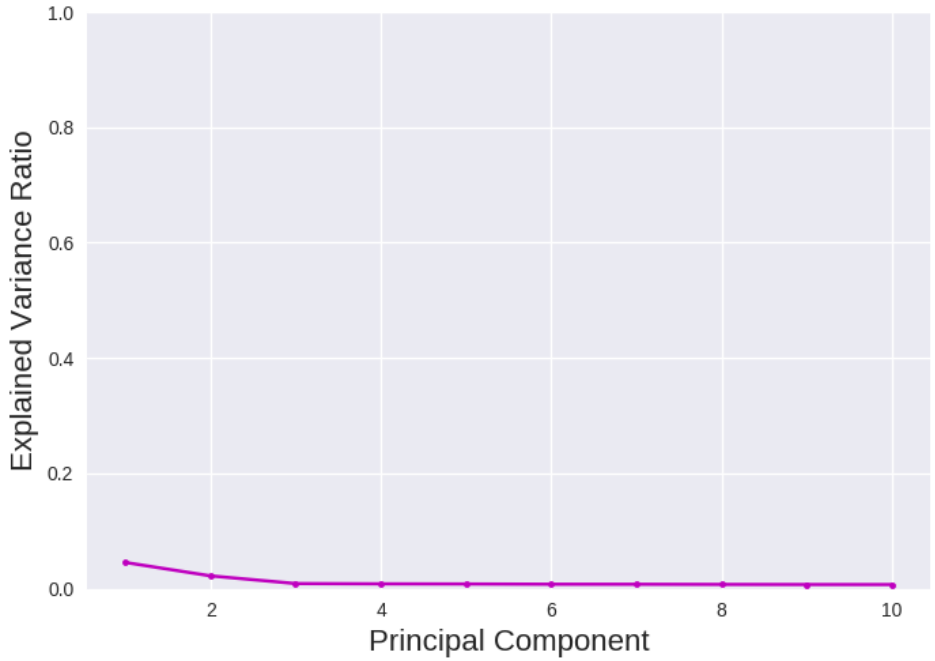
Single Nucleotide Polymorphisms (SNPs)

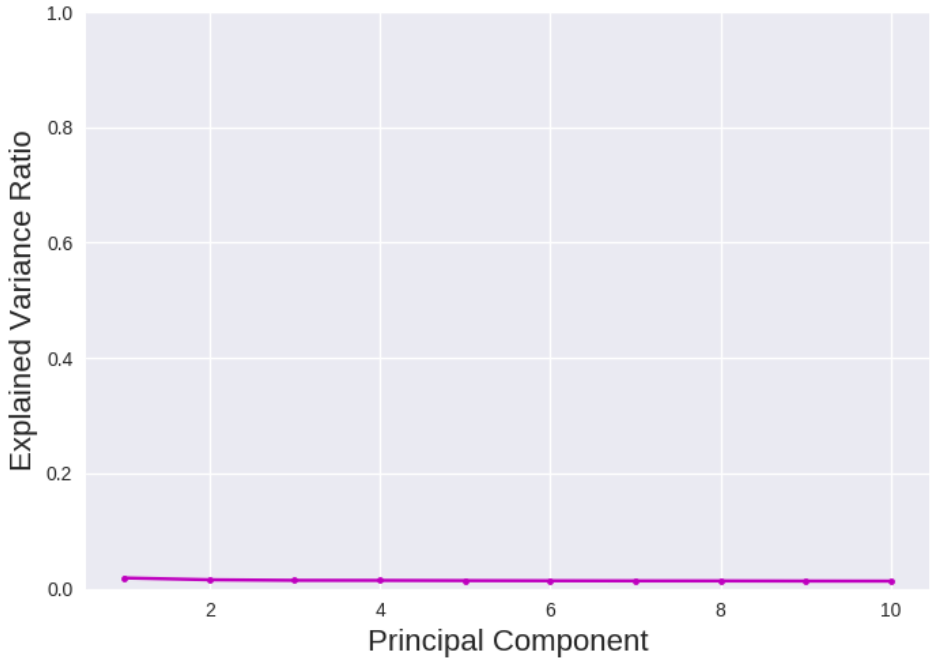
An. gambiae	ATGCATGCATTCATGC
An. gambiae	ATGCATCCATACATGC
An. gambiae	ATGCATGCATTCATGC
An. gambiae	ATGCATCCATACATGC
An. coluzzii	AAGCATGCATTCATGC
An. coluzzii	AAGCATGCATACATGC
An. coluzzii	AAGCATGCATTCATGC
An. coluzzii	AAGCATGCATACATGC

Single Nucleotide Polymorphisms (SNPs)

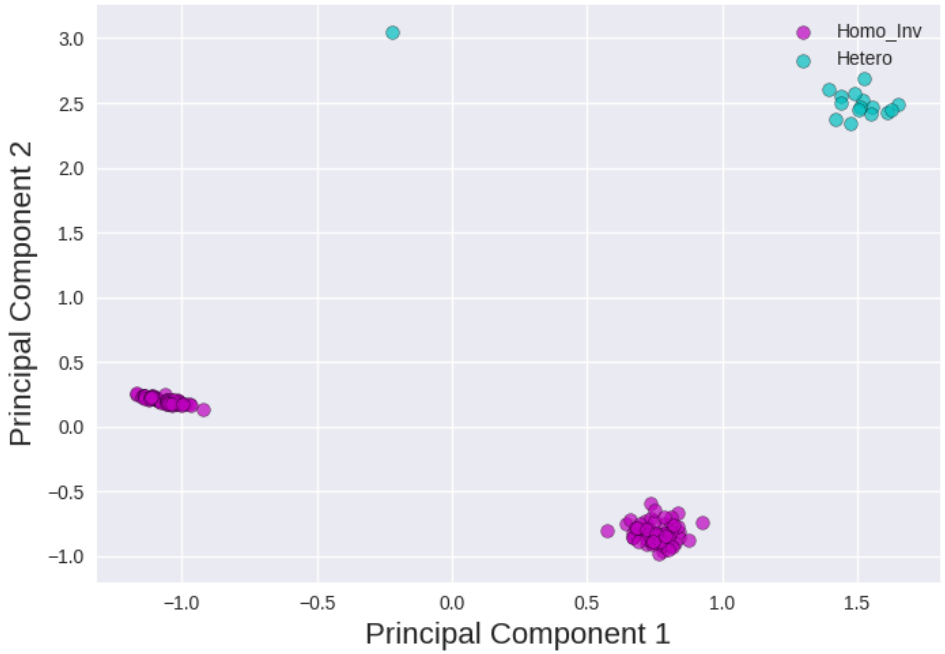
An. gambiae	T	G	T
An. gambiae	T	C	A
An. gambiae	T	G	T
An. gambiae	T	C	A
An. coluzzii	A	G	T
An. coluzzii	A	G	A
An. coluzzii	A	G	T
An. coluzzii	A	G	A

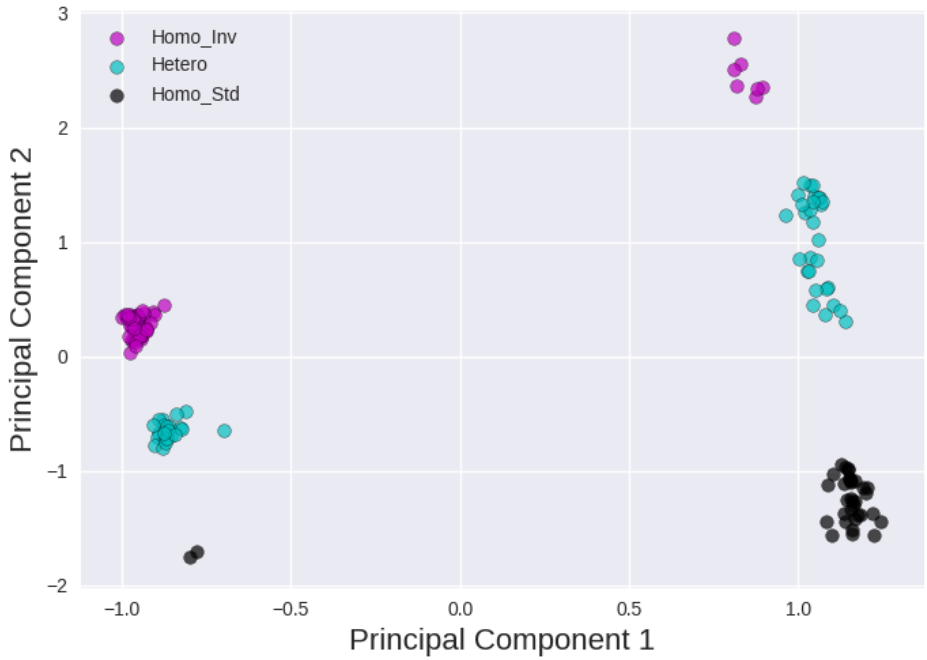
Picking Number of Components



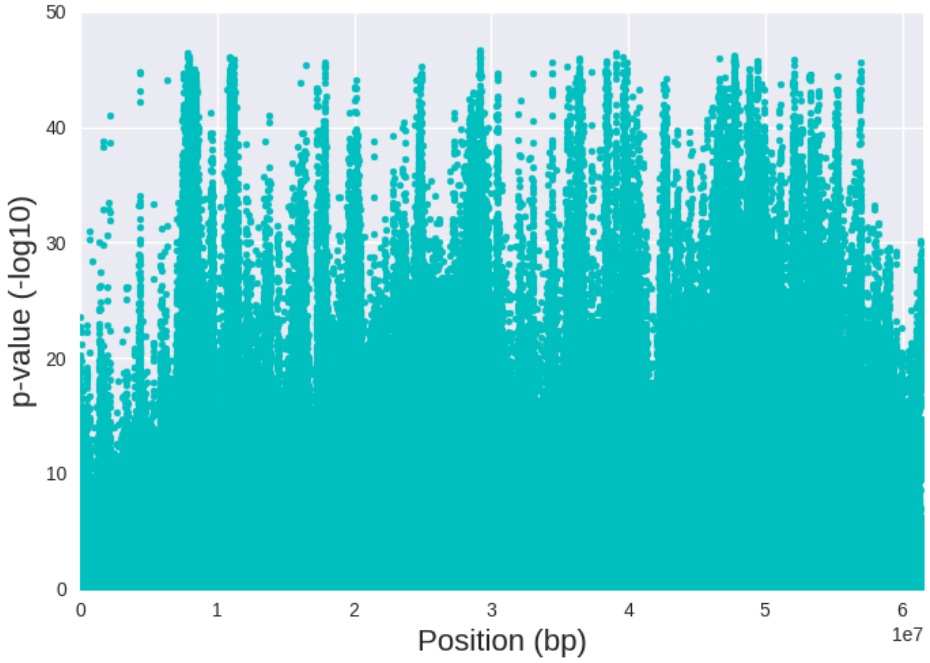


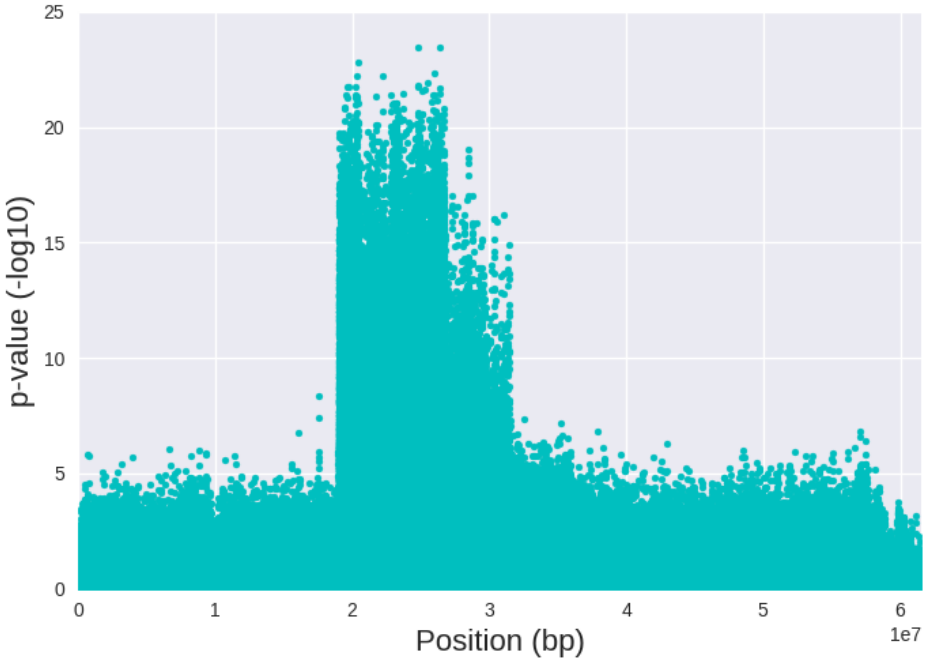
PCA of *Anopheles* Mosquitoes





PC-SNP Associations on *Anopheles* 2R





Data Science Skills / What You'll Learn

- Data munging – parsing, scraping, formatting, cleaning data
- Scientific process – exploring data to observe patterns, stating a hypothesis, and proving or disproving the hypothesis (e.g., using models, statistics, or visualizations)
- Communication and Visualization – reports, tables, graphs, interactive data applications, summary statistics
- Statistics – traditional analysis
- Machine learning – modeling relationships, prediction
- Domain knowledge – business, science, etc.