Digital Transformation of Healthcare

Data Cleaning

Michoel Snow, M.D. Ph.D., Glen Ferguson, Ph.D.

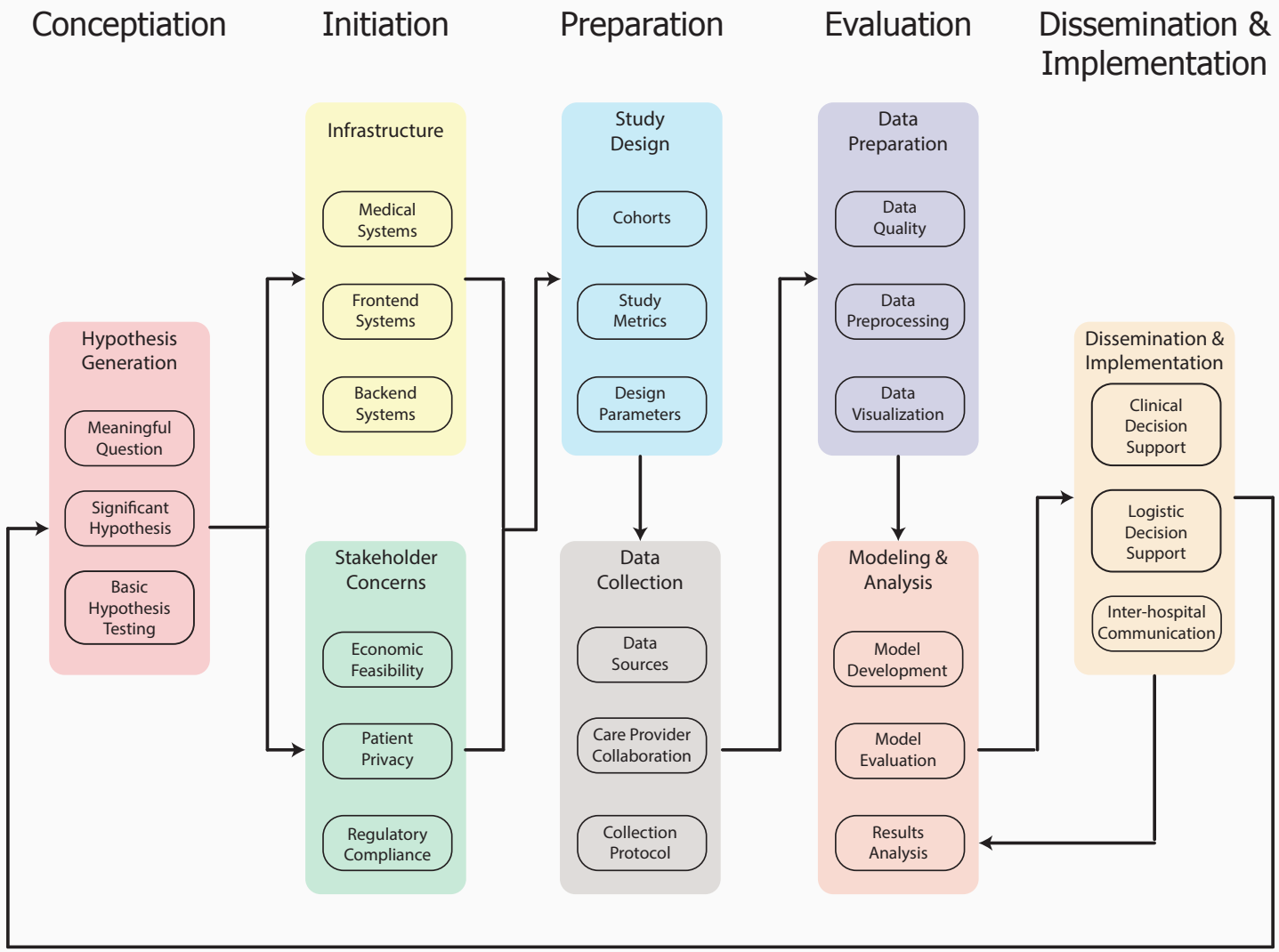
Center for Health Data Innovations

Data Cleaning

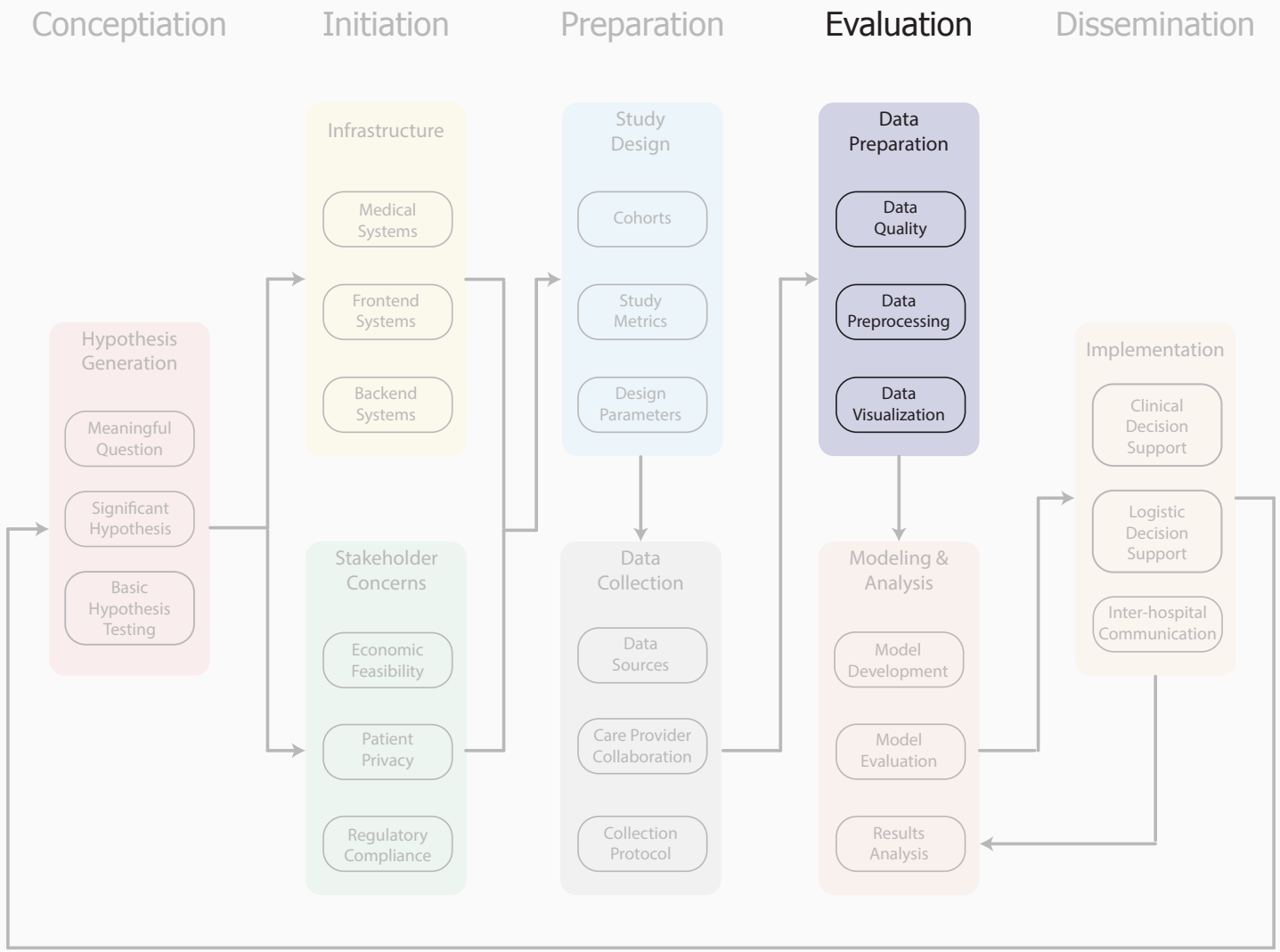
After this lecture students will be able to

- Discuss and apply the steps involved in cleaning data for modeling
- Design a process for the imputation of missing data
- Build a bioinformatics pipeline starting from given data

Bioinformatics Pipeline

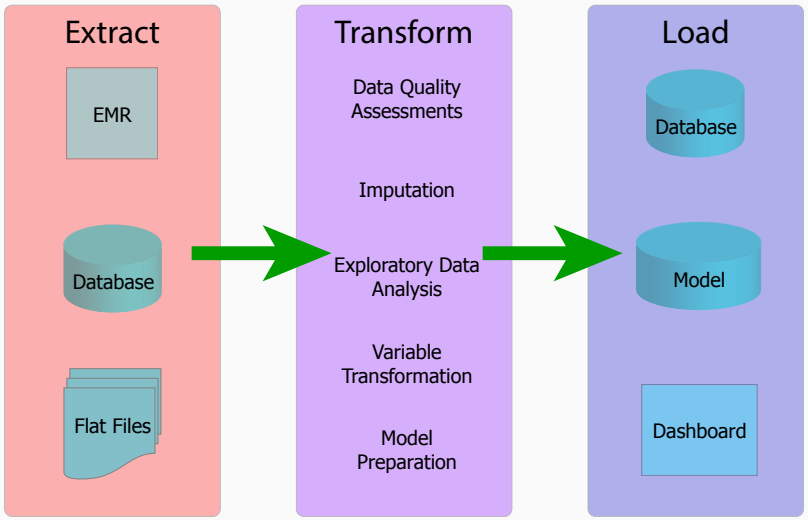


Data Cleaning



Scenario

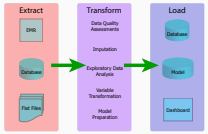
- You are trying to optimize the use of pain medication regimens for pediatric sickle cell patients
- You have just collected all the data on all peds hem-onc patients on the floor for the past month.
- How can you systematically assess, organize and prepare the data for modeling and analysis?
    - What are the steps to transform raw data into a usable format



Scenario

Scenario

- You are trying to optimize the use of pain medication regimens for pediatric sickle cell patients
- You have just collected all the data on all peds hem-onc patients on the floor for the past month.
- How can you systematically assess, organize and prepare the data for modeling and analysis?
  - What are the steps to transform raw data into a usable format



- Data encoding
- variable correctness, i.e., what percentage of the data is missing (and does that percentage make sense (birth date vs death date), are all the values the same
- variable quality
- Data organization/ time frequencies - Ensure that time blocks used in time series data are appropriate to the task (Do you need to group your data or can they all be left as individual data points, e.g., do I want to know all of my patient's lab values every 6 hours, or can I take them as they come in. Loaf of bread model vs streaming model.)
- drop corrupt data, variables and time blocks
- impute and/or mask missing variables

Imputation and Extrapolation

- what are the different reasons why data might be missing
- What are the different ways that data could be missing
- Can we develop a systematic way to deal with missing data
  - pain score
  - pain medication usages
  - retic count
  - infection status
  - imaging results
- How do you evaluate imputation

Imputation and Extrapolation

Imputation and Extrapolation

- what are the different reasons why data might be missing
- What are the different ways that data could be missing
- Can we develop a systematic way to deal with missing data
  - pain score
  - pain medication usages
  - retic count
  - infection status
  - imaging results
- How do you evaluate imputation

- data could be mcar, mar (dependent on an observed variable, which can be controlled for), mnar or because we are slicing the data into chunks smaller than the sampling rate or there is no value, e.g., death date for a living patient
  - How is the data missing, e.g., first values, random values, ...
  - when is there too much missing data
  - correlation/heatmap of missingness

  - categorical fill, cases where missingness tells you information e.g., missing date of x-ray could be another class in that column, as empty but not missing
  - ffill, mean fill, back fill
  - drop blocks if data is bad or empty
  - Drop beginning blocks if empty, drop end blocks if dead or discharged
  - **Imputation** - MICE, NN, KNN
  - Anything which is not imputed is masked (-9999, not 0)

- check through hand created missingness accuracy as well as downstream calculations

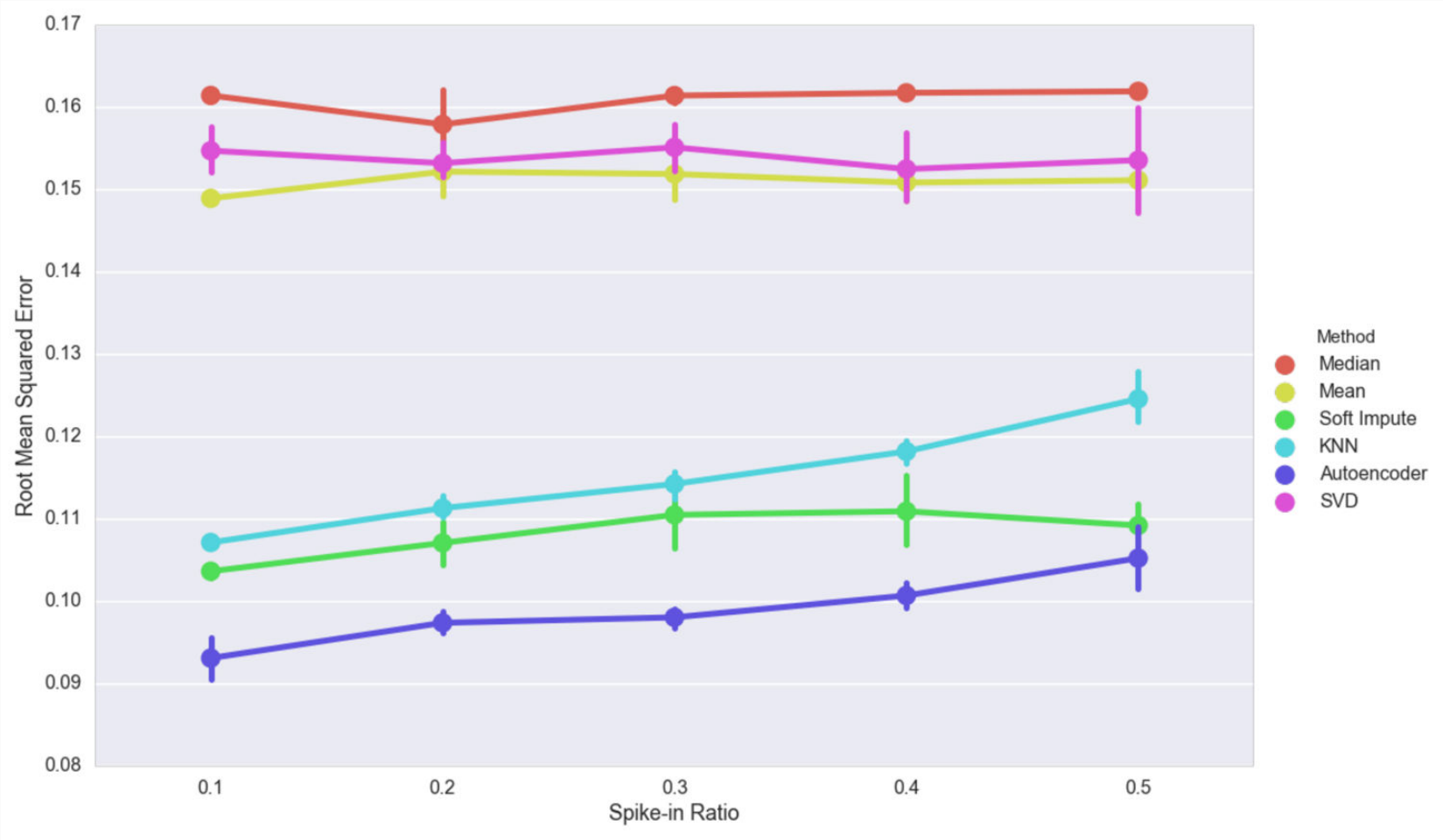
Imputation Example

Missing Data Imputation in the Electronic Health Record Using Deeply Learned Autoencoders

- Researchers started from an ALS clinical trials database of 10,723 patients
- The dataset includes patient demographic data, family history, concomitant medications, vital sign measurements, laboratory results, and patient clinical history
- They removed data using either an MCAR or MNAR approach
- For end metrics they considered both accuracy in imputation and ALS functional rating scale

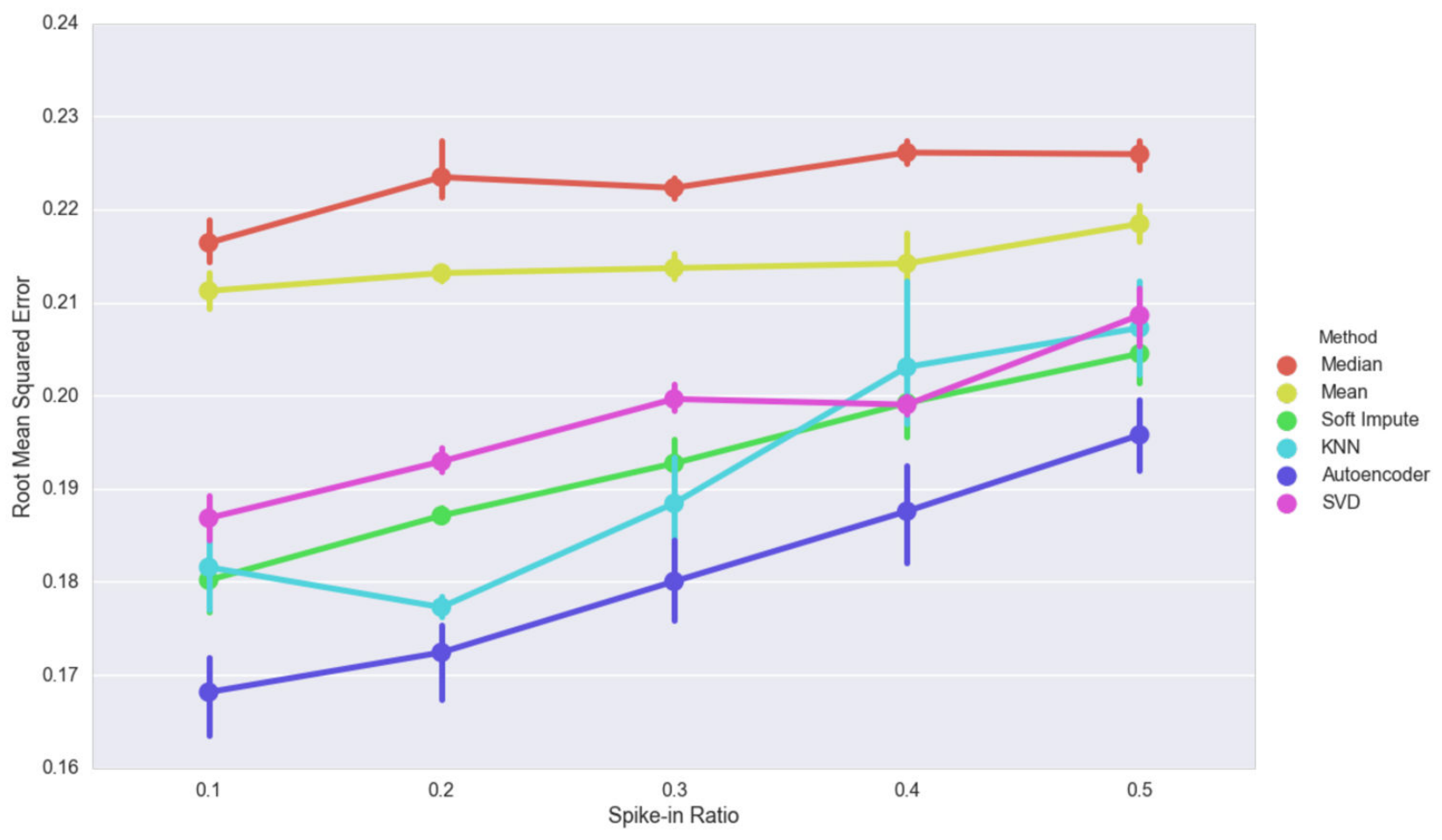
Imputation Example

Missing Completely at Random

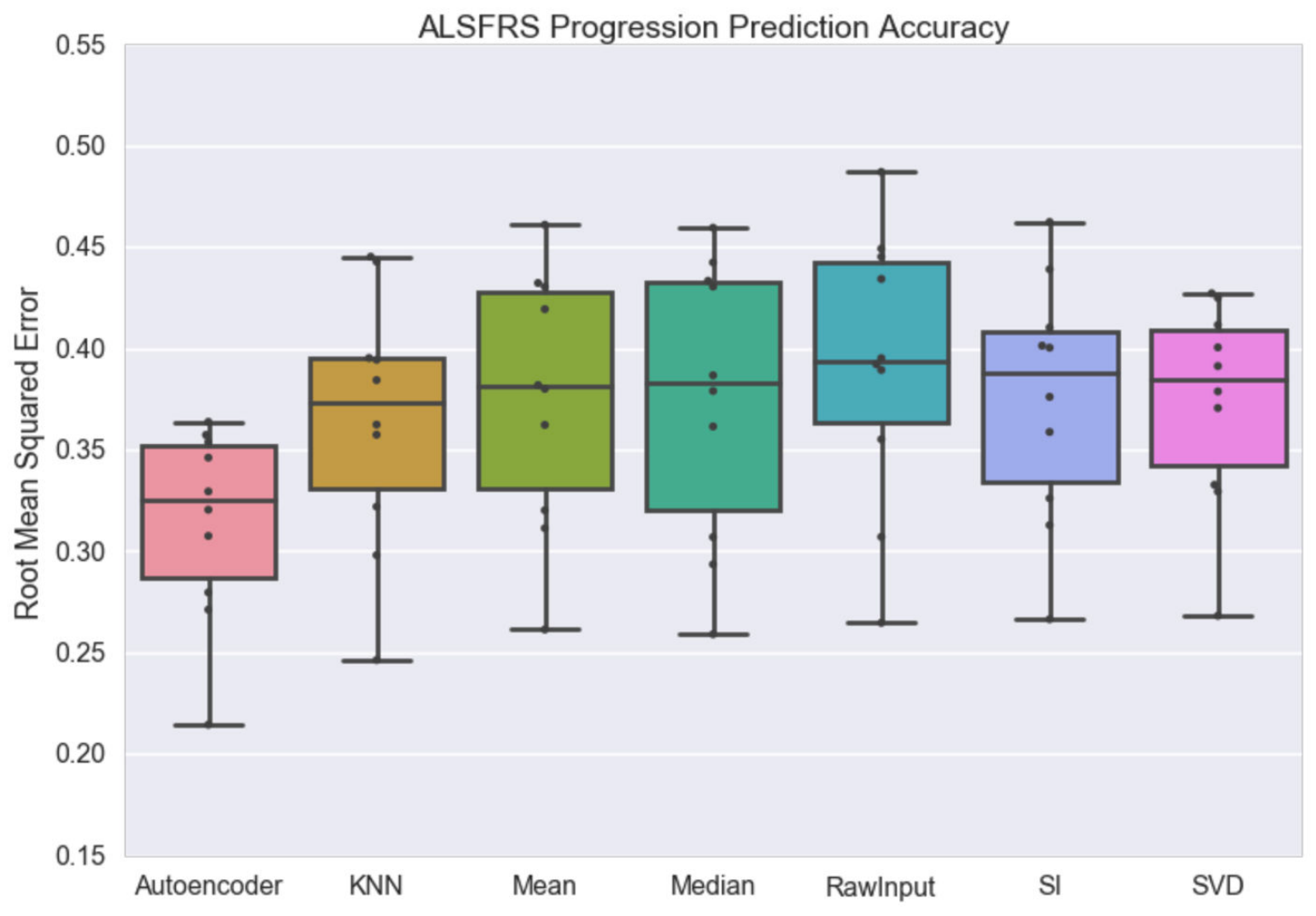


Imputation Example

Missing Not at Random



Imputation Example



Pipeline

- Using the pediatric sickle cell example let's walk through building the second half of the bioinformatics pipeline