Digital Transformation of Healthcare

ETL & Assessing Data Quality

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

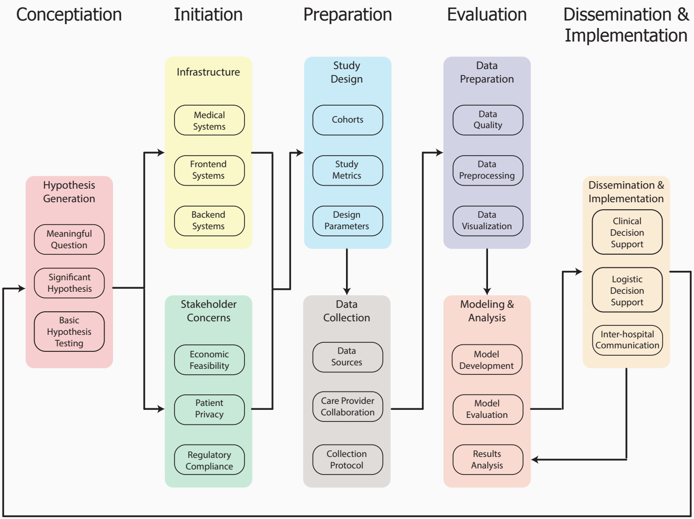
Center for Health Data Innovations

Assessing Data Quality

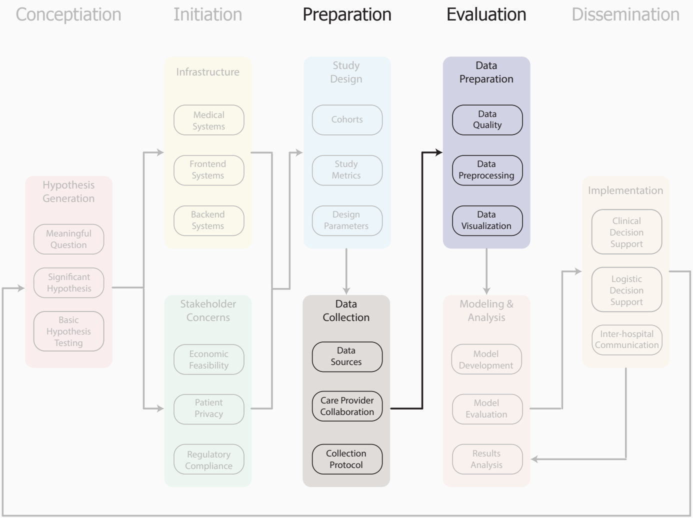
After this lecture students will be able to

- Describe the components of an ETL pipeline
- Extract variables from data and incorporate additional information
- Assess the quality of data
- Trace the steps where data quality can be affected
- Examine data for problems and discuss possible causes
- Design a process for the imputation of missing data

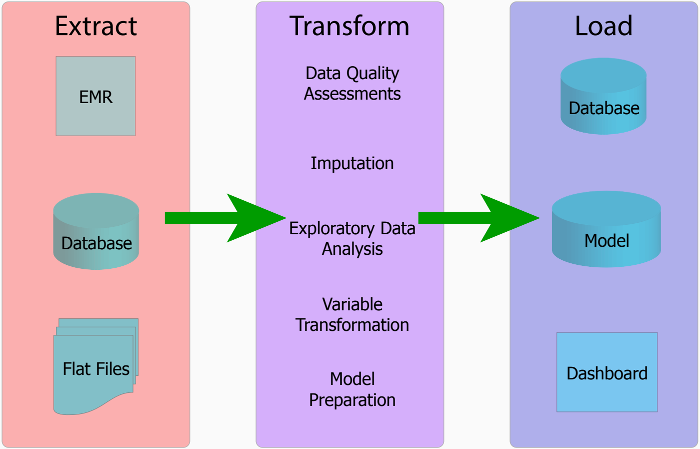
Bioinformatics Pipeline



ETL & Data Quality



Extract, Transform and Load (ETL)



Extract

EMR

Database

Flat Files

Transform

Data Quality Assessments

Imputation

Exploratory Data Analysis

Variable Transformation

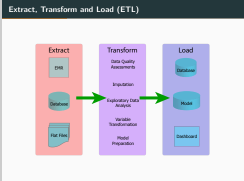
Model Preparation

Load

Database

Model

Dashboard



Extract, Transform and Load (ETL)

Extract

EMR

Database

Flat Files

Transform

Data Quality Assessments

Imputation

Exploratory Data Analysis

Variable Transformation

Model Preparation

Load

Database

Model

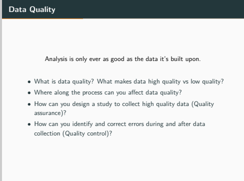
Dashboard

- Variable transformations
  - Non-numerical to numerical data
  - Extracting implied or hidden variables
  - Connecting extraneous variables to current variables, e.g., weather to date
- How can you convert non-numerical data to numerical data
  - what kinds of non-numerical data are there - categorical (blood group, degree of Ab susceptibility/resistance), semi/unstructured (text, audio, video), boolean (test results), dates
- What makes a transformation useful?
  - What are all the different possible things you can extract from a date (season, days since equinox, hours since noon, is it happy hour, weekend)
  - What other information can you connect to a date (weather, precipitation, barometric pressure, traffic accidents, sunrise, sunset, federal holiday, famous birthdays)

Data Quality

Analysis is only ever as good as the data it's built upon.

- What is data quality? What makes data high quality vs low quality?
- Where along the process can you affect data quality?
- How can you design a study to collect high quality data (Quality assurance)?
- How can you identify and correct errors during and after data collection (Quality control)?

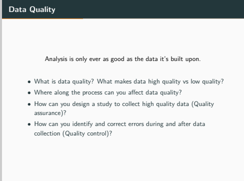


Data Quality

Analysis is only ever as good as the data it's built upon.

- What is data quality? What makes data high quality vs low quality?
- Where along the process can you affect data quality?
- How can you design a study to collect high quality data (Quality assurance)?
- How can you identify and correct errors during and after data collection (Quality control)?

- Data quality consists of the objective [accuracy, validity (not outside range of possibilities, all data is for the same pt, formatting requirements, DICOM dates), reliability (dx matches problem list matches coding), legibility (units, shorthand)] and subjective [completeness]
- Steps
  - Definition/Design - lack of clear definitions for data items/collection, incompatible units, precision, scope, depth
  - Collection - not enough documentation (drug given/dosage altered but no start and end date), non-adherence to data definitions (collecting data outside of protocol time), human variance/error (bp cuff, RR, incorrect units), Orders are placed (procedures, medications) which are not connected to a rationale or sufficient reason
  - Processing - interpretation ('initial' lab, diagnosis date), coding error (mis-entering information such as order of birthdate, or height as 9 cm instead of 90 cm), random (mistyping, illegible handwriting), software errors, Assigning codes to problems treated vs problems tested and ruled out, which complaints do you code/document (doctors as coders)



Data Quality

Analysis is only ever as good as the data it's built upon.

- What is data quality? What makes data high quality vs low quality?
- Where along the process can you affect data quality?
- How can you design a study to collect high quality data (Quality assurance)?
- How can you identify and correct errors during and after data collection (Quality control)?

- quality assurance - training of personnel (mock exams and reporting), site visits, reduce open-ended questions
- quality control - data monitoring (compare to independent source), hand verification, entering data in twice (by different sources), consistency checks

Quality Assurance - DICOM

- DICOM - Digital Imaging and Communications in Medicine - is the international standard for medical images and related information. It defines the formats for medical images that can be exchanged with the data and quality necessary for clinical use
- DICOM groups information into data sets, e.g., an x-ray would contain the patient ID within the file, so that the image can never be separated from this information by mistake.
- DICOM Value Representations

https://www.dicomstandard.org/about/

Quality Assurance - DICOM

name	VR	value
Group Length	UL	532
Image Type	CS	DERIVED
SOP Class UID	UI	1.2.840.10008.5.1.4.1.1.2
SOP Instance UID	UI	1.2.840.114356.2008.11.30.12.34.2.329.999
Study Date	DA	20081230
Content Date	DA	20081230
Study Time	TM	122731
Content Time	TM	12299.0000
Modality	CS	CT
Institution Name	LO	Manhasset Diagnostic Imaging
Station Name	SH	
Study Description	LO	MOSES CT Outside Reference Images
Procedure Code Sequence	SQ	[{(0008, 0100): (0008, 0100) Code Value ...
Code Value	SH	MOSESOUTREFCT
Coding Scheme Designator	SH	GEIIS
Coding Scheme Version	SH	0
Code Meaning	LO	MOSES CT Outside Reference Images
Series Description	LO	Reformatted
Referenced SOP Class UID	UI	1.2.840.113619.2.51762891606.1649.1005918257.250
Referenced SOP Instance UID	UI	1.2.840.114356.2008.11.30.12.34.2.329.1301

Quality Assurance - DICOM

name	VR	value
Study Date	DA	20081230
Content Date	DA	20081230
Study Time	TM	122731
Content Time	TM	12299.0000

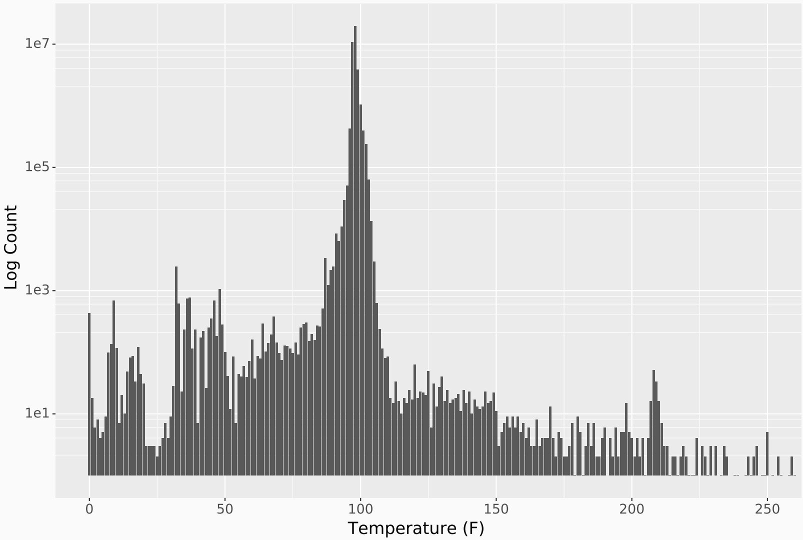
- **DA** - A string of characters of the format YYYYMMDD
- **TM** - A string of characters of the format HHMMSS.FFFFFF.
  - One or more of the components MM, SS, or FFFFFF may be unspecified as long as every component to the right of an unspecified component is also unspecified

Whose fault is this?

Quality Control - Temperature

How can I find the temperatures recorded from every patient in the hospital

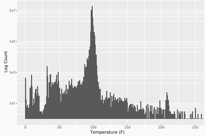
To the SQL



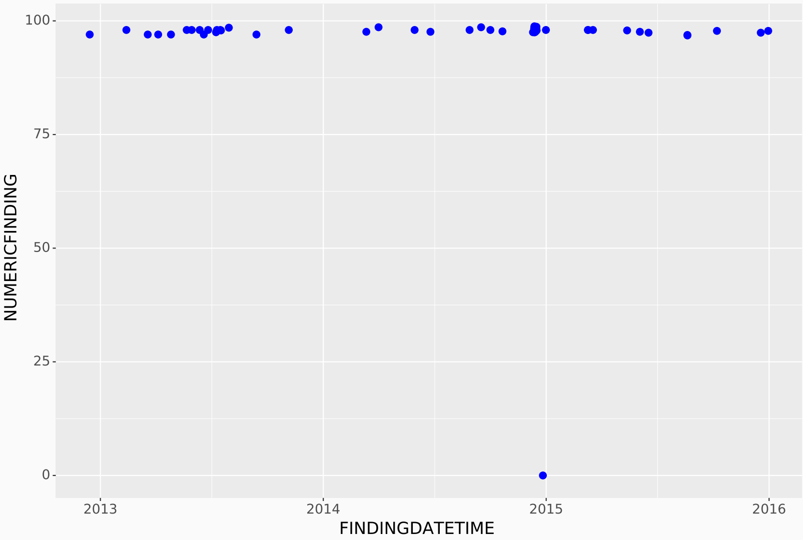
1e7
1e5
1e3
1e1
Log Count
0
50
100
150
200
250
Temperature (F)

2018-11-01

Digital Transformation of Healthcare



- How do you interpret the temps around 0 (probably had to enter something), how about around 37F (centigrade), how about 212, how about 95 (MICE)
- Let's take a look at an individual patient's data, who had a temp of 0
- maybe the data is being pulled from 5 different hospitals and it's the ETL which is causing the errors, because it doesn't know Celsius from F
- let's pick a patient whose temperature is zero and see what the rest of their temperature values look like (and then look at her results for that date)


100
75
50
25
0
NUMERICFINDING
2013
2014
2015
2016
FINDINGDATETIME

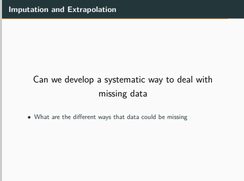
Associated Values

FINDINGDATETIME	FINDINGDESC	NUMERICFINDING
2014-12-26	PULSE OXIMETRY	97.00
2014-12-26	WEIGHT/SCALE (ounces)	2800.16
2014-12-26	HEIGHT (inches)	62.00
2014-12-26	Diastolic Blood Pressure	82.00
2014-12-26	Systolic Blood Pressure	139.00
2014-12-26	HEIGHT (CM)	157.48
2014-12-26	PULSE	75.00
2014-12-26	BODY MASS INDEX	32.13
2014-12-26	O2 SAT%	97.00
2014-12-26	TEMPERATURE (F)	0.00
2014-12-26	Systolic Blood Pressure	139.00
2014-12-26	WEIGHT (KG)	79.38
2014-12-26	Diastolic Blood Pressure	82.00

Imputation and Extrapolation

Can we develop a systematic way to deal with missing data

- What are the different ways that data could be missing



Imputation and Extrapolation

Can we develop a systematic way to deal with missing data

- What are the different ways that data could be missing

- data could be MCAR, MAR, MNAR or because we are slicing the data into chunks smaller than the sampling rate
- Missing Data Procedure
  - **Variable correctness** - var correctly derived/appropriate to include, e.g., complete or near-complete missingness or same value in all rows.
  - **Time freq** - Ensure that time blocks used in time series data are appropriate to the task
  - Determine how frequently every variable is measured
  - Use the frequency range from the previous step for each variable to do ffill
  - Encounters without data or good data cannot add value and should be dropped.
  - Drop beginning blocks if empty, drop end blocks if dead or discharged
  - **Imputation** - MICE, NN. Cases where imputation should not be done are when the missingness itself is significant or if the imputation cannot be done by adding another class. An example of the latter is would be if an x-ray is performed. X-rays not being performed are another class that can be added to the column.
  - Anything which is not imputed is masked (-9999, not 0)

Sources

- WHO data quality
- Healthcare Data Warehousing and Quality Assurance
- (2002). Defining and improving data quality in medical registries JAMIA, 9(6), 600-611.