ETL & Data Quality

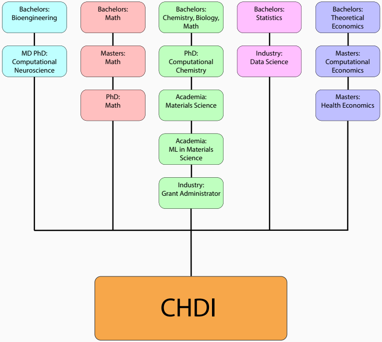
Digital Transformation of Healthcare

Michoel Snow, MD PhD

Center for Health Data Innovations

Who are we?

- Center for Health Data Innovations (CHDI)
  - formerly, the Clinical Research Informatics (CRI) core
- Part of both Einstein and Montefiore
- Develop infrastructure based on informatics technologies
- Links Einstein's translation science engine to Montefiore's learning healthcare system



Bachelors: Bioengineering

MD PhD: Computational Neuroscience

Bachelors: Math

Masters: Math

PhD: Math

Bachelors: Chemistry, Biology, Math

PhD: Computational Chemistry

Academia: Materials Science

Academia: ML in Materials Science

Industry: Grant Administrator

Bachelors: Statistics

Industry: Data Science

Bachelors: Theoretical Economics

Masters: Computational Economics

Masters: Health Economics

CHDI

What do we do?

- PROOFcheck
  - Department - Critical Care
  - Respiratory failure prediction
  - EMR based alerts
- Metastatic Epidural Spinal Cord Compression
  - Department - Radiation Oncology
  - Early identification and remediation of spinal met progression
- Outpatient Appointment Attendance
  - Department - Medicine
  - Determine the probability of a patient not showing up to their appointment
  - Optimize patient appointments

Digital Transformation of Healthcare

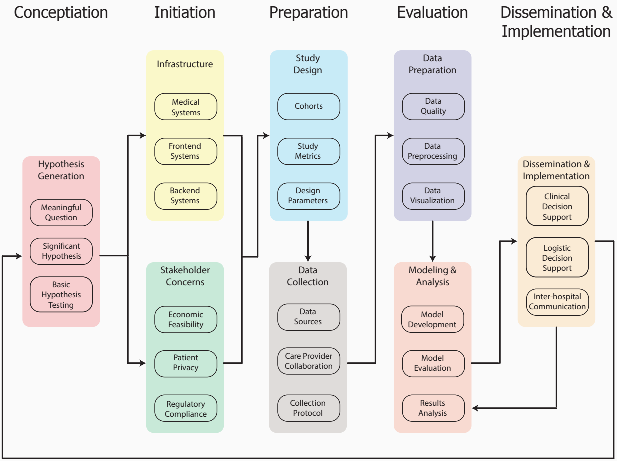
- What kind of questions can I answer using automatically collected data?
- What kind of data is collected by the hospital and how can I access the data?
- How much will it cost/save the hospital to implement the study as well as act on its results?
- What do I need to consider when designing a study using patient data?
- How can I integrate automatic systems with collaborators to collect the desired data?
- How do I transform the data from its collected format to a format useful for analysis?
- How can I integrate the results of my study within the hospital system?

Case Study

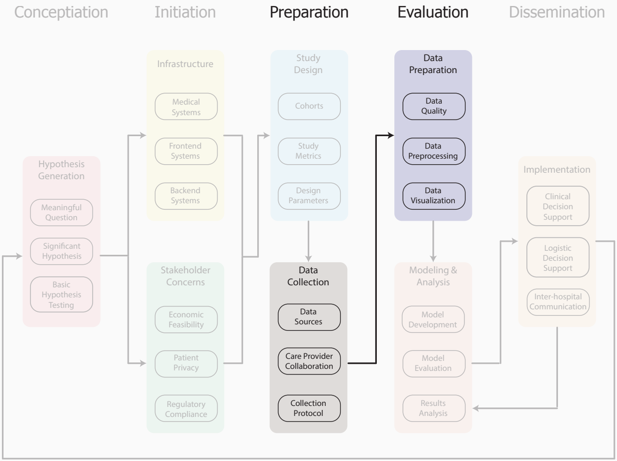
You have developed a new method of detecting sepsis in patients, which you think is better than the current sepsis criteria.

- How can you determine if your claim is true, retrospectively?

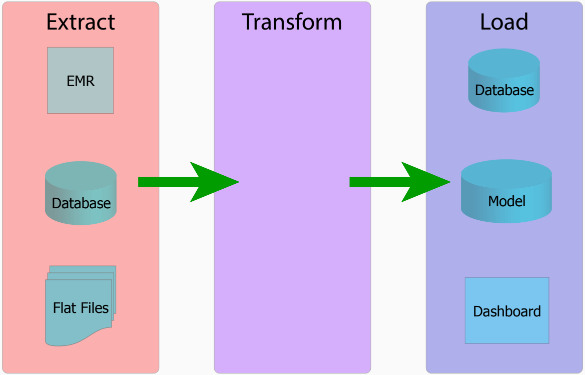
Bioinformatics Pipeline



ETL & Data Quality

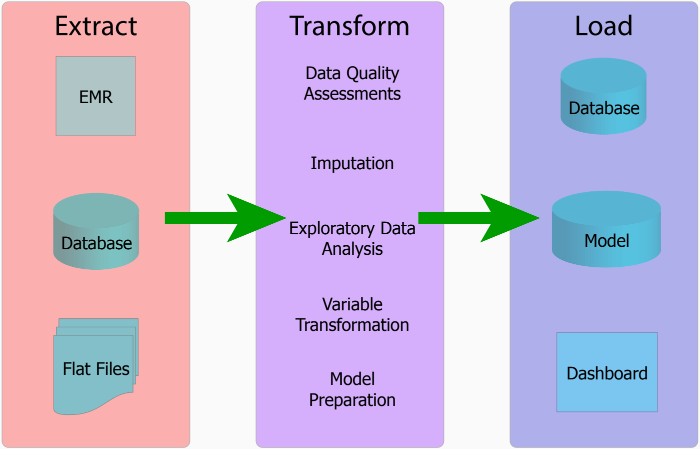


Extract, Transform and Load (ETL)



- What transformations would you want to do to your extracted data?

Extract, Transform and Load (ETL)



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 - 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?

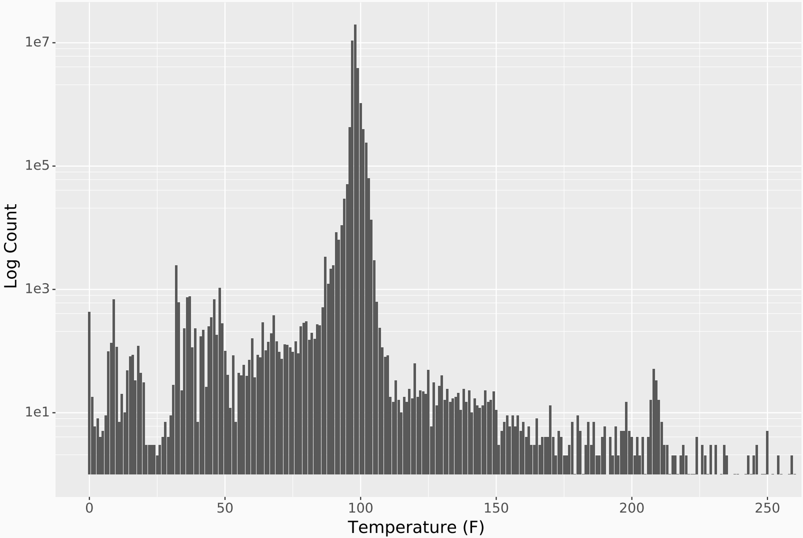
Quality Control - Sepsis Case Study

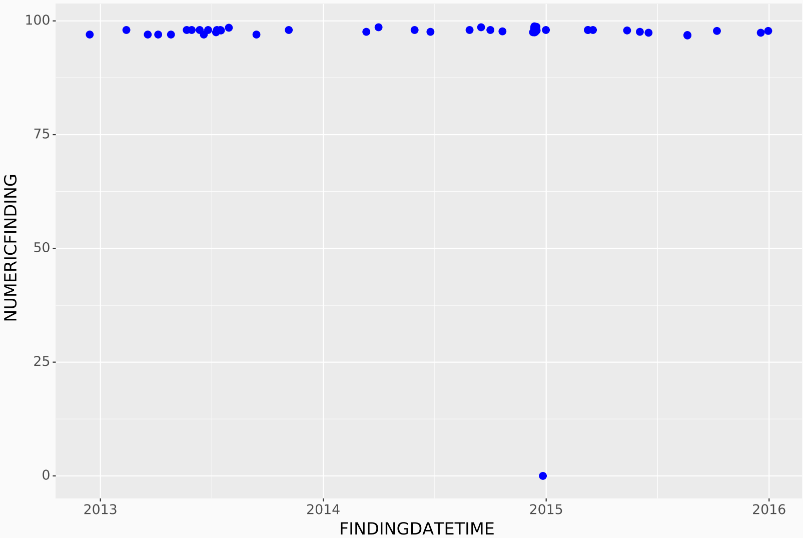
My Sepsis metric depend on the following parameters

- Temperature
- Respiratory Rate
- BP
- HR

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)


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

Sources

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

Thank You

https://github.com/MichoelSnow/crtp

msnow1@montefiore.org