

Biostat 202: Data Management & Storage

“IT’S SINK OR SWIM AS THE TIDAL WAVE OF DATA APPROACHES”

Nature 399:517 10 June 1999

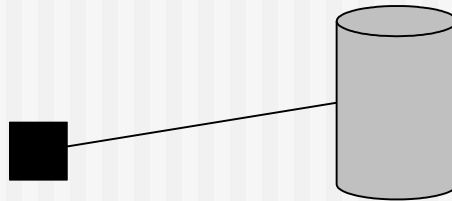


Isabel Elaine Allen
Professor, Biostatistics & Epidemiology
isabel.allen@ucsf.edu

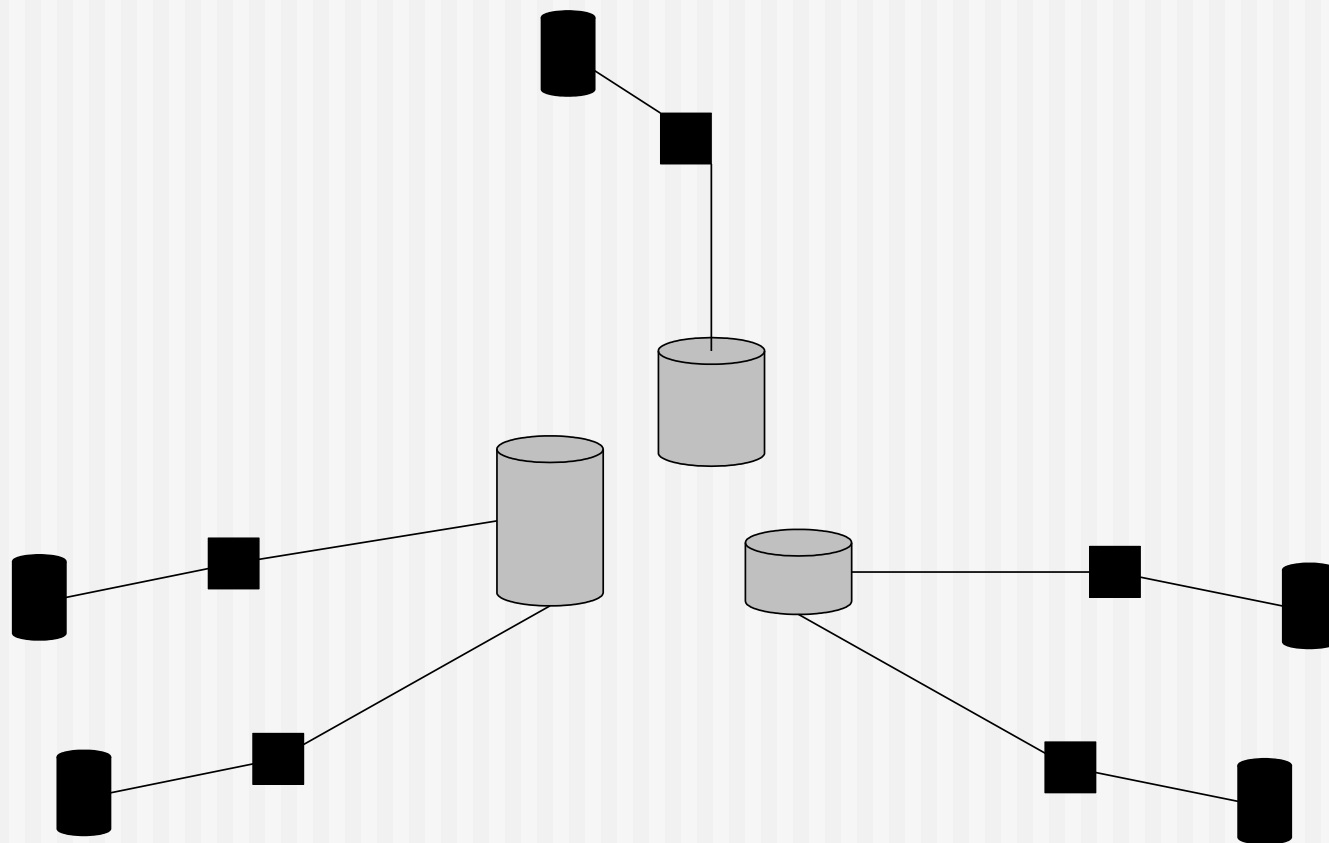
Why Data Management?

- Data in the real world is dirty
 - **Incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - **Noisy**: containing errors or outliers
 - **Inconsistent**: containing discrepancies in codes or names
 - **Not well controlled**: convenience samples
 - **Security**: how will you protect the data?

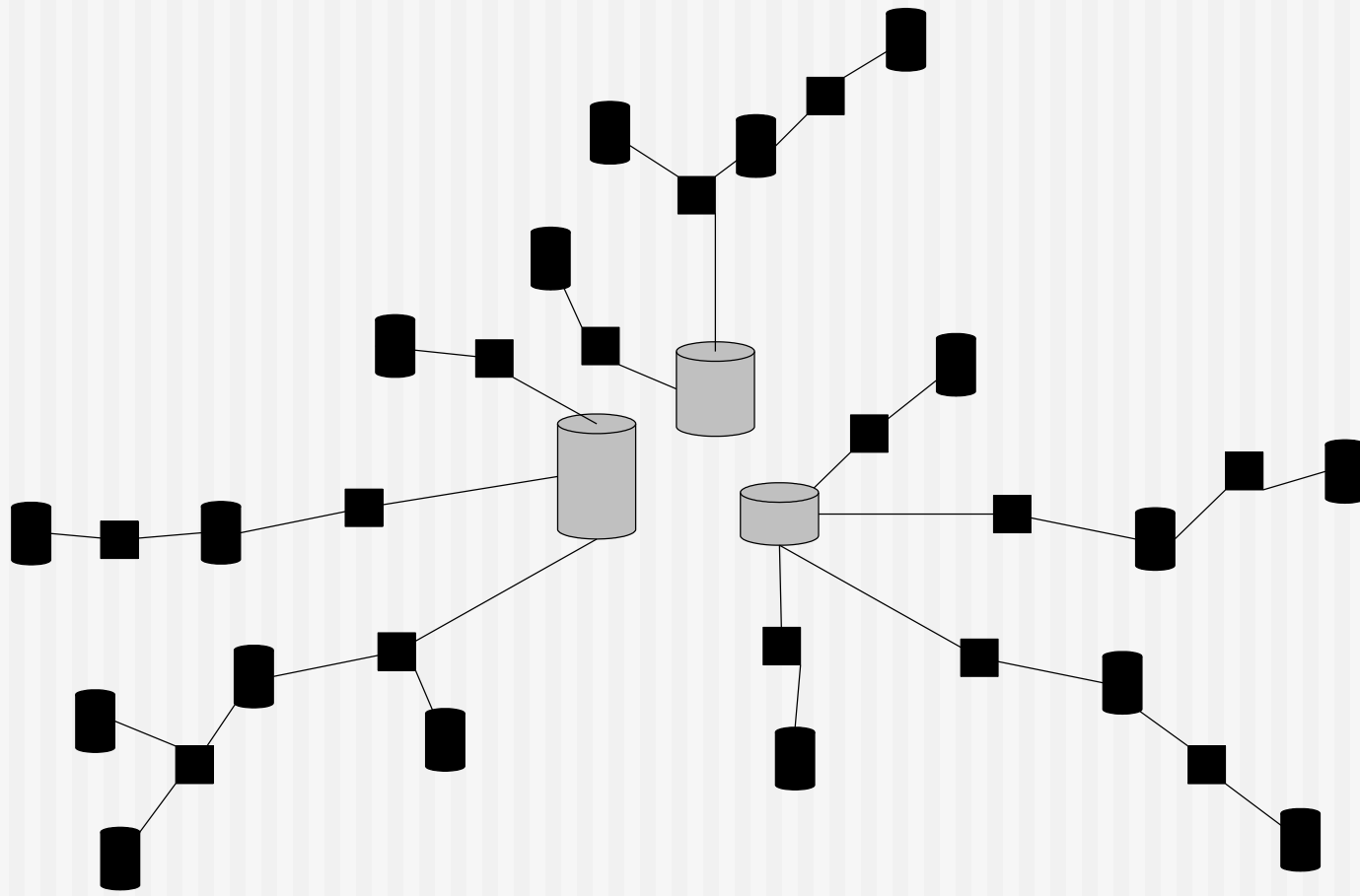
In the Beginning, life was simple...



But...

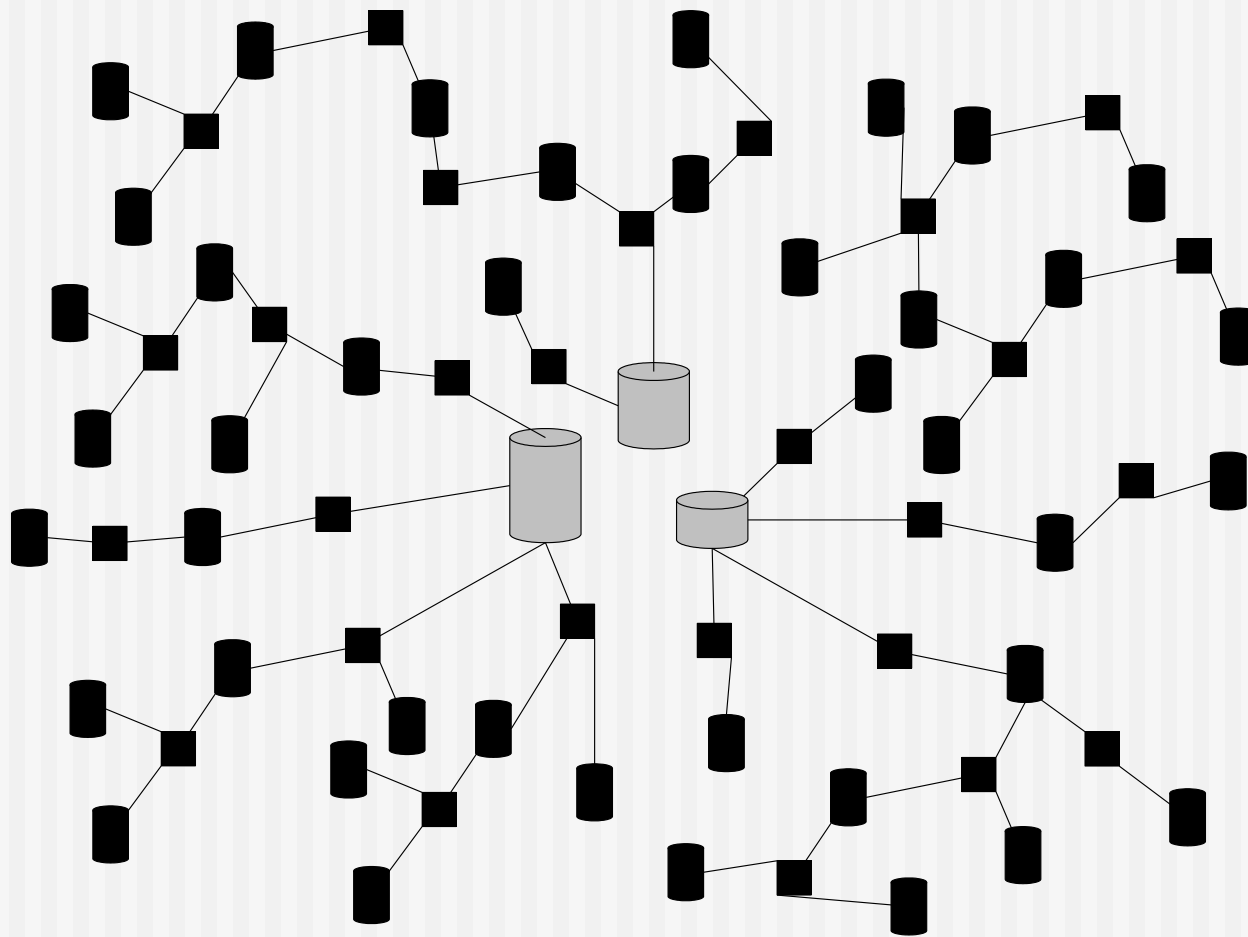


Our information needs...



Kept growing. (The Spider web)

A type of database schema



SOURCE: William H. Inmon

Database Schema

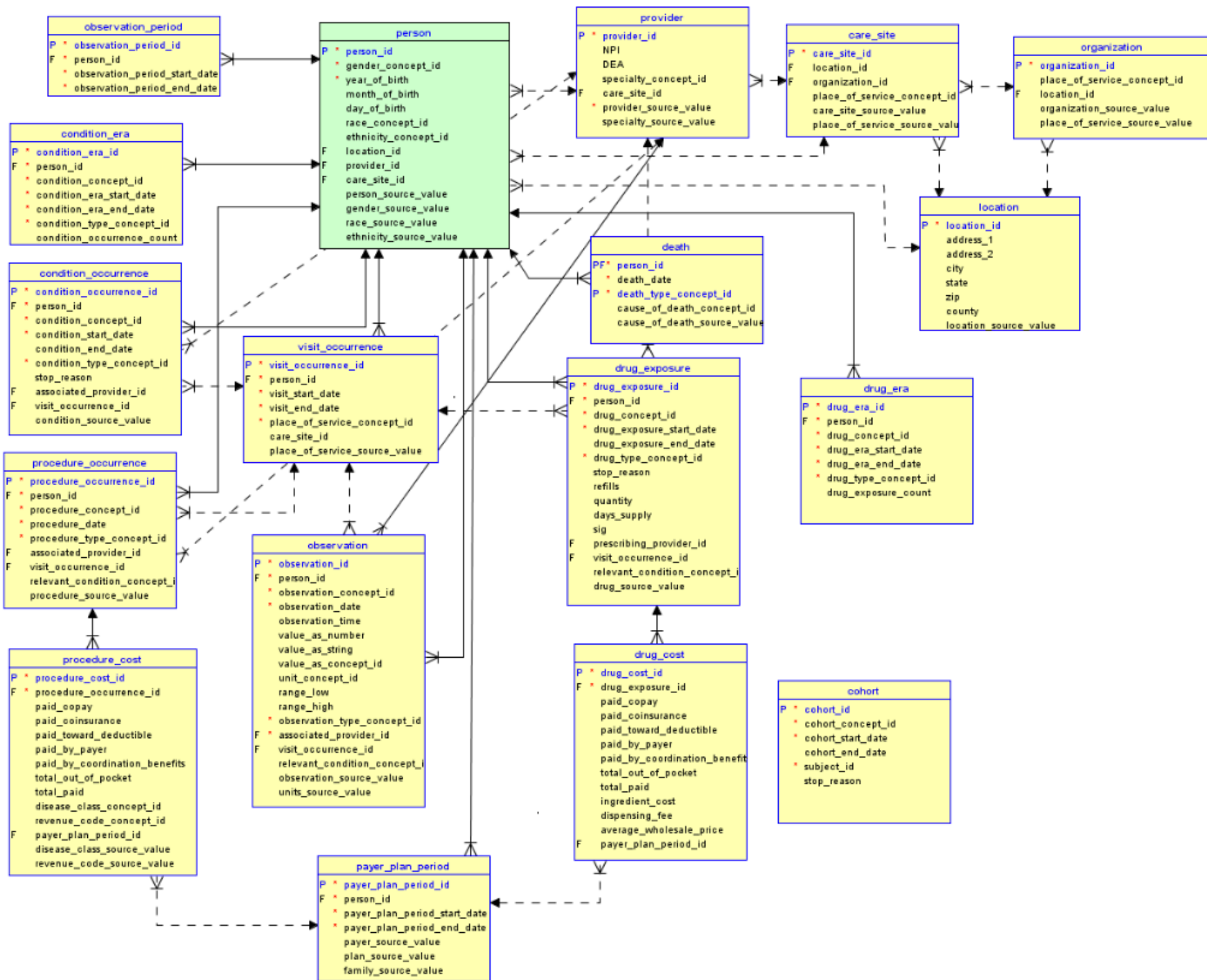
Medical Records

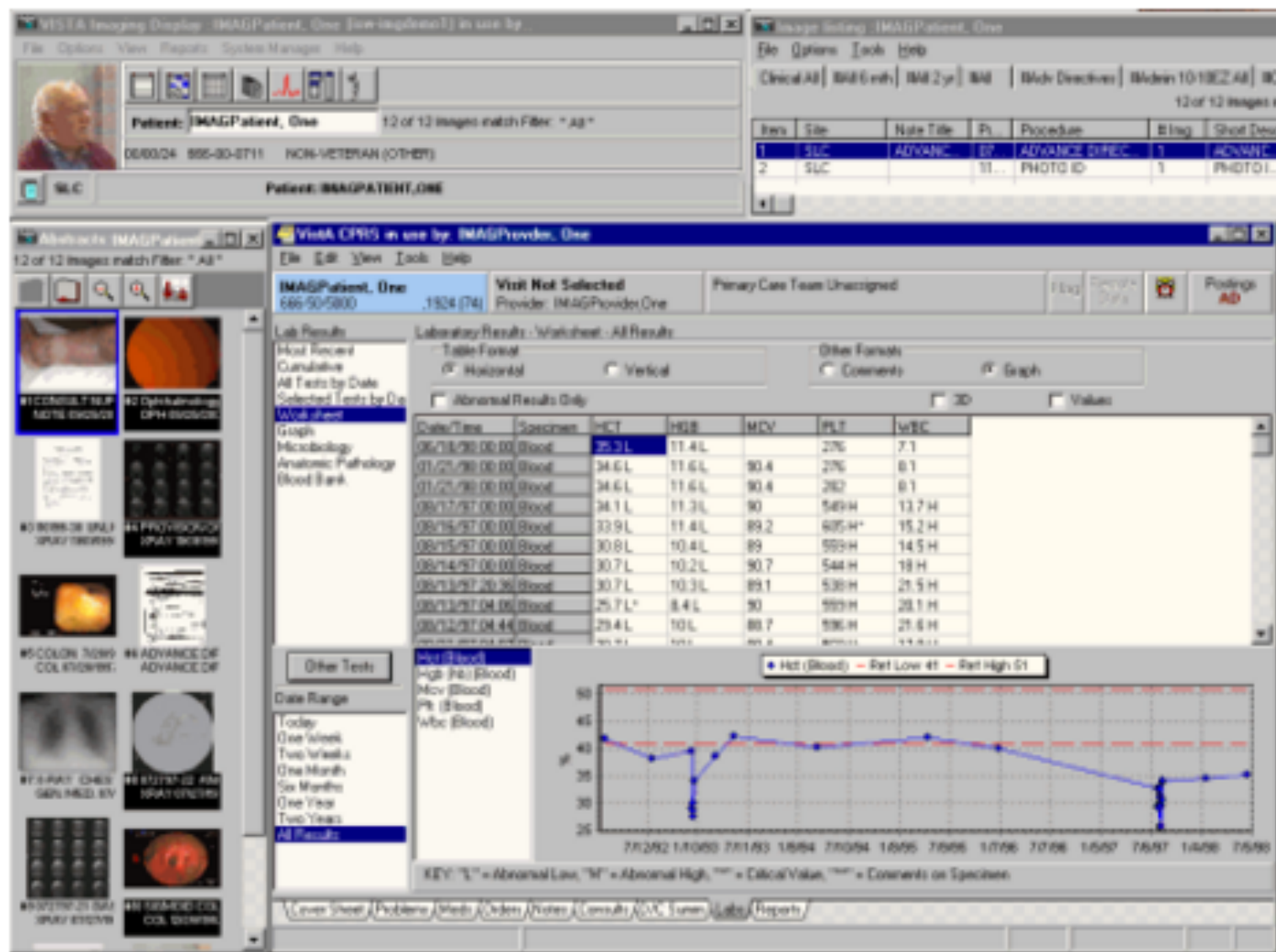
- Collects all patient information in a central repository
- Provide central point of access for
 - Patients
 - Care providers: physicians, nurses, etc.
 - Billing
 - Insurance reconciliation
- Hospitals, physicians, patients, procedures, records



MONGODDB
WORLD'16 FOR EARTHY IDEAS

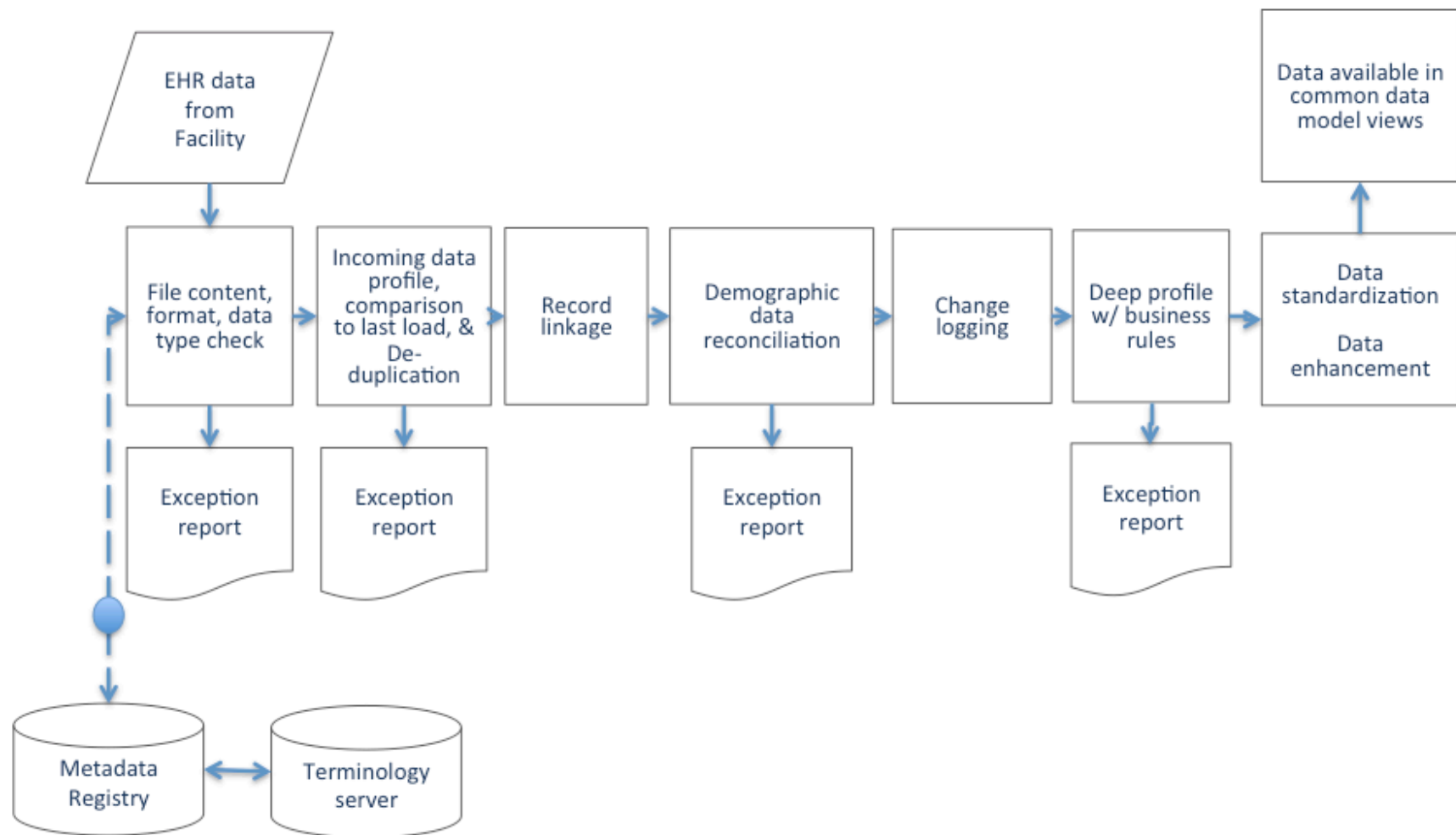
...time & effort to digitize this?





patient record view from an image-based electronic health record

How does it get there?



...So now you want to put the data on your laptop

Do UCSF Databases Differ From Data On Your Laptop?

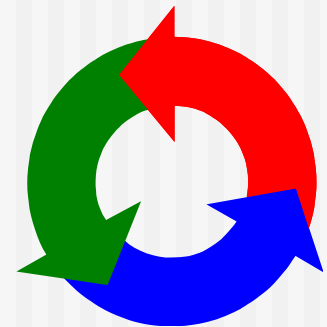
- Size?
- Goals?
- Privacy?
 - Anonymizing
 - HIPPA
 - Stored only on secure server
- Structure & Merging & Analyzing?
 - MyResearch & RedCap
 - EMR's (Epic vs. Apex vs. ??)
 - Social media: Twitter & Google etc

Some Database Terminology: Data Warehouses, Data Marts (Data Lakes), & Data Stores

- Data Warehouse – The queryable source of data in the enterprise. It is comprised of the union of all of its constituent data marts. Could be APEX or EPIC or ?
- Data Mart or Data Lake – A logical subset of the complete data warehouse. Often viewed as a restriction of the data warehouse to a single group of related data. Could be Benioff Hospital or ER only.
- Data Store – Could be a place to query, create an online dataset of tweets, and purchase from Twitter or a software company pre-processing the data.

Building your database: Importing Data

- Identify data sources
- Extract the needed data from existing systems to a temporary *data staging* area
- Merge, Append, Transform and Clean the data
 - Resolve data type conflicts
 - Resolve naming and key conflicts
 - Remove, replace, correct, or flag bad data
 - Conform Dimensions
- Load the data
- Pre-processing the data



An Example using APEX & Ambulance Transportation Company Records*

- Identify all patients seen in SFGH ER between 1/2014 and 12/2014
- Identify all transports to the ER in ambulances during the same period
- Extract primary visit reason
- Extract medical history
- Extract demographics
- Cross-code patient ID to link databases
- Merge datasets
- Anonymize data
- Export data = 43,560 records

*2014 Hall, Rodriguez, Allen et al. EMS-STARs:Emergency Medical Services "Superuser" Transport Activations, a retrospective study *Prehospital Emergency Care* 19(1)L 61-67.

Major Tasks in Preprocessing (1)

- Data cleaning

- Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies (i.e. different units)

- Data integration

- Integration of multiple databases, data cubes, or files
- Inner or outer merge, append



Major Tasks in Preprocessing (2)

- Data transformation
 - Normalization and aggregation
- Data reduction
 - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
 - Part of data reduction but with particular importance, especially for numerical data

Data Cleaning

- Data cleaning tasks
 - Fill in missing values
 - Identify outliers and smooth out noisy data
 - Correct inconsistent data*

*2005 Leff, Allen et al, Inside Employment Interventions: Effects of Job Development and Job Support on Competitive Employment of Persons with Severe Mental Illness, *Psychiatric Services*, 56(10): 1237-1245.

Missing Data (1)

- Data is not always available
 - E.G., Many observations have no recorded value for several attributes (which may be linked), such as previous visit information or demographics



Missing Data (2)



- Missing data may be due to
 - Refuse/unable to answer/inconsistent answer
 - Equipment malfunction
 - Inconsistent with other recorded data and thus deleted
 - Data not entered due to misunderstanding
 - Certain data may not be considered important at the time of entry
 - No register history or changes of the data
- Missing data may need to be inferred and imputed

How to Handle Missing Data? (1)

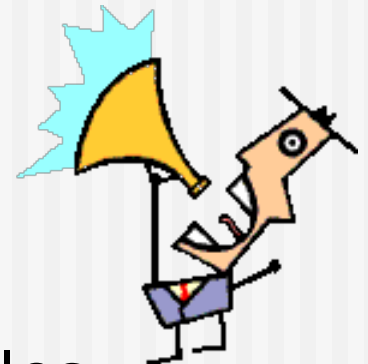
- Ignore the observation: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably)
- Fill in the missing value manually: tedious + not feasible for any medium/large dataset
- Use a global constant to fill in the missing value: e.g., “Unknown”, a new class, ‘999’?

How to Handle Missing Data? (2)

- Use the mean to fill in the missing value (okay)
- Use the mean for all samples belonging to the same class to fill in the missing value: (smart)
- Use the most probable value to fill in the missing value: inference-based such as bayesian formula or decision tree or regression (smarter)
- Multiple imputation techniques examining patterns of missingness (smartest?)
- What can Modeler do for you? (Upcoming!)

Noisy Data (1)

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
 - Faulty data collection instruments
 - Data entry problems
 - Data transmission problems
 - Technology limitation
 - Inconsistency in naming convention
 - Identifiable bias?



...let's measure the tables

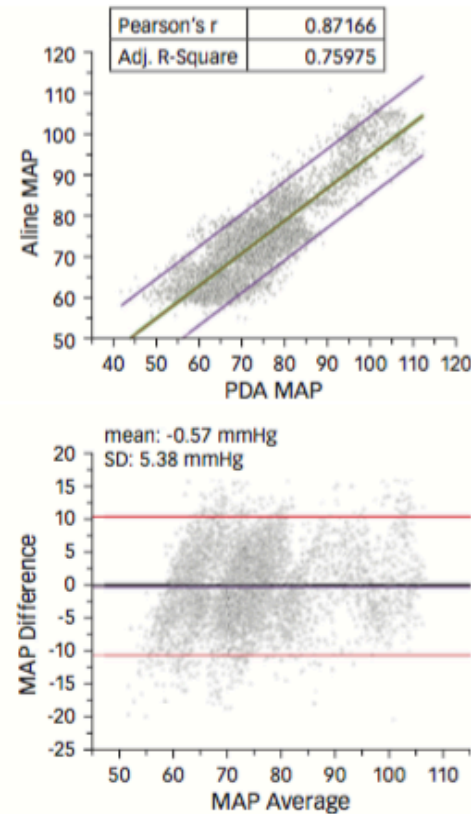
Noisy Data* (2)

- Other data problems which require data cleaning
 - Duplicate records
 - Incomplete data
 - Inconsistent data



Data Cleaning Example*

Correlation and Bland-Altman graphs / FIGURE 1



Note: The correlation (top, linear fit with 95% confidence bounds) and (bottom) Bland Altman graphs are used for mean arterial pressure comparison of both blood pressure measurement systems.

*2016 Allen & Seaman Plotting for Quality *Journal of Quality Progress*, August

*2016 Gratz, Allen, Seaman et al Continuous Non-Invasive Finger Cuff Caretaker® comparable to Invasive Intra-Arterial Pressure in Patients Undergoing Major Intra-Abdominal *Canadian Anesthesia Journal* (to appear).

Data Integration (1)

- Data integration:
 - Combines data from multiple sources into a coherent store
 - Merging vs. appending data
- Database integration
 - Integrate **metadata** from different sources
 - Entity identification problem: identify real world entities from multiple data sources, e.g., A.Patient-id $\equiv$ B.cust-#

Data Integration (2)

- Detecting and resolving data value conflicts
 - For the same real world entity, attribute values from different sources are different
 - Possible reasons: different representations, different scales, e.g., Metric vs. British units

Redundant Data



- Redundant data occur often when integrating of multiple databases
 - The same attribute may have different names in different databases
 - One attribute may be a “derived” attribute in another table, e.g., GFR or Creatinine Clearance

Data Integrity Problems: Examples

- Same person, different spellings
 - Agarwal, Agrawal, Aggarwal etc...
- Multiple ways to denote company name
 - Persistent Systems, PSPL, Persistent Pvt. LTD.
- Use of different names
 - Mumbai, Bombay
- Different account numbers generated by different applications for the same customer
- Required fields left blank
- Invalid product codes collected*
 - manual entry leads to mistakes
 - “in case of a problem use 9999999”

*2016 Chu, Chi, Allen et al Durability of Flexible Ureteroscopy & Predictors of Repair: A prospective multicenter study (WIP)

SPSS Modeler for Data Management

- Data audit
- Variable derivation
- Anonymizing
- Merging & Appending
- Missing Data
- Examining small subgroups

...to the lab for today