

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

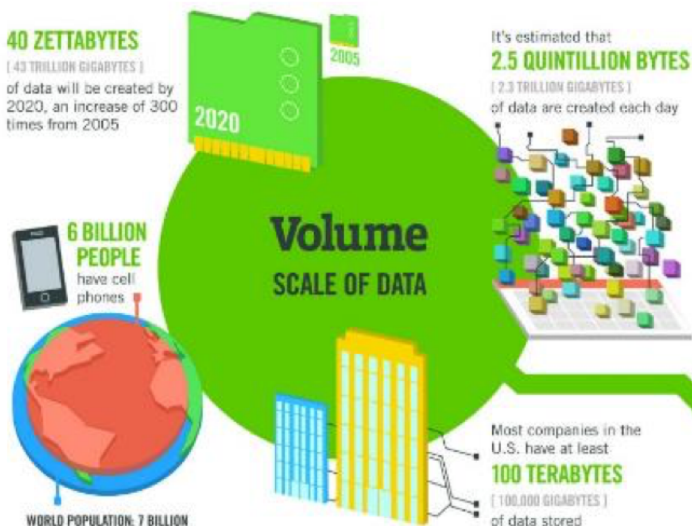
Why Data Management?



- 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?

40 ZETTABYTES

[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES
[161 BILLION GIGABYTES]



30 BILLION PIECES OF CONTENT
are shared on Facebook every month



Variety

DIFFERENT FORMS OF DATA

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month



400 MILLION TWEETS
are sent per day by about 200 million monthly active users



The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session



Velocity

ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
— almost 2.5 connections per person on earth



Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure



1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions



Poor data quality costs the US economy around
\$3.1 TRILLION A YEAR



27% OF RESPONDENTS

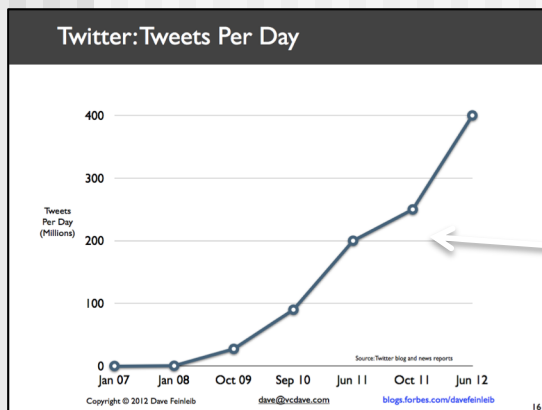
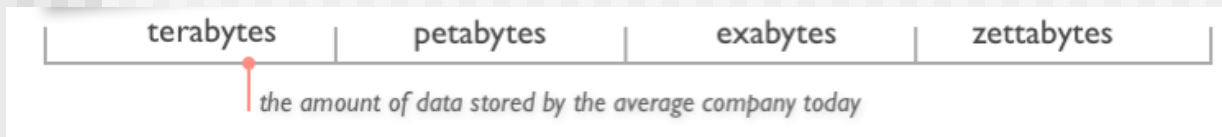
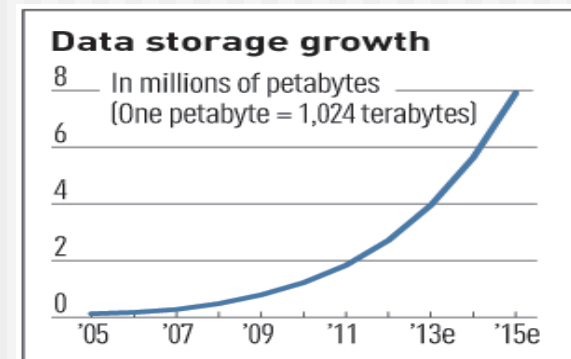
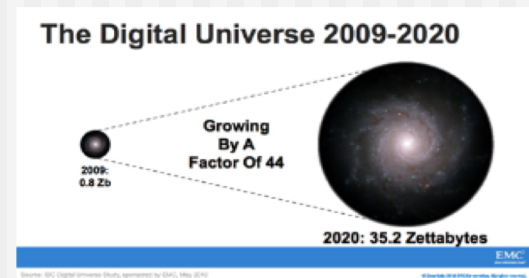
in one survey were unsure of how much of their data was inaccurate

Veracity

UNCERTAINTY OF DATA

Characteristics of Big Data: #1-Scale (Volume)

- **Data Volume**
 - 44x increase from 2009 to 2020
 - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially

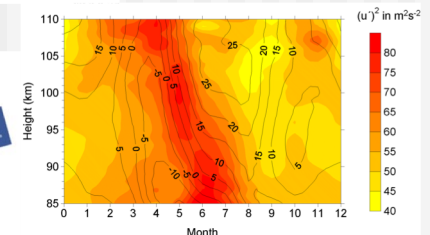
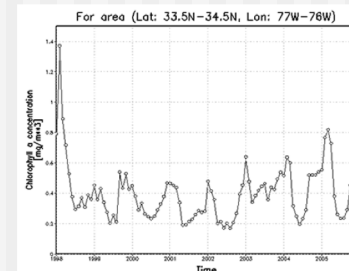
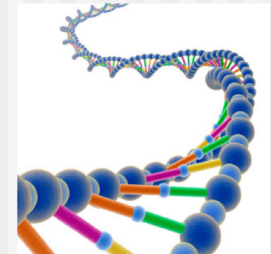
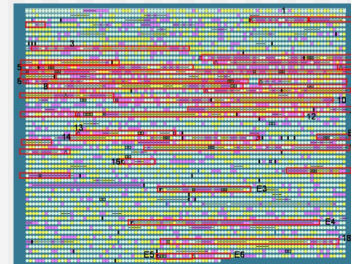


*Exponential increase in
collected/generated data*

Characteristics of Big Data:

#2-Complexity (Variety)

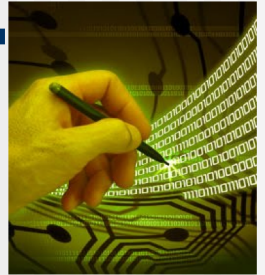
- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc...
- Static data vs. streaming data
- A single application can be generating/collecting many types of data



To extract knowledge → all these types of data need to be linked together

Characteristics of Big Data:

#3-Speed (Velocity)

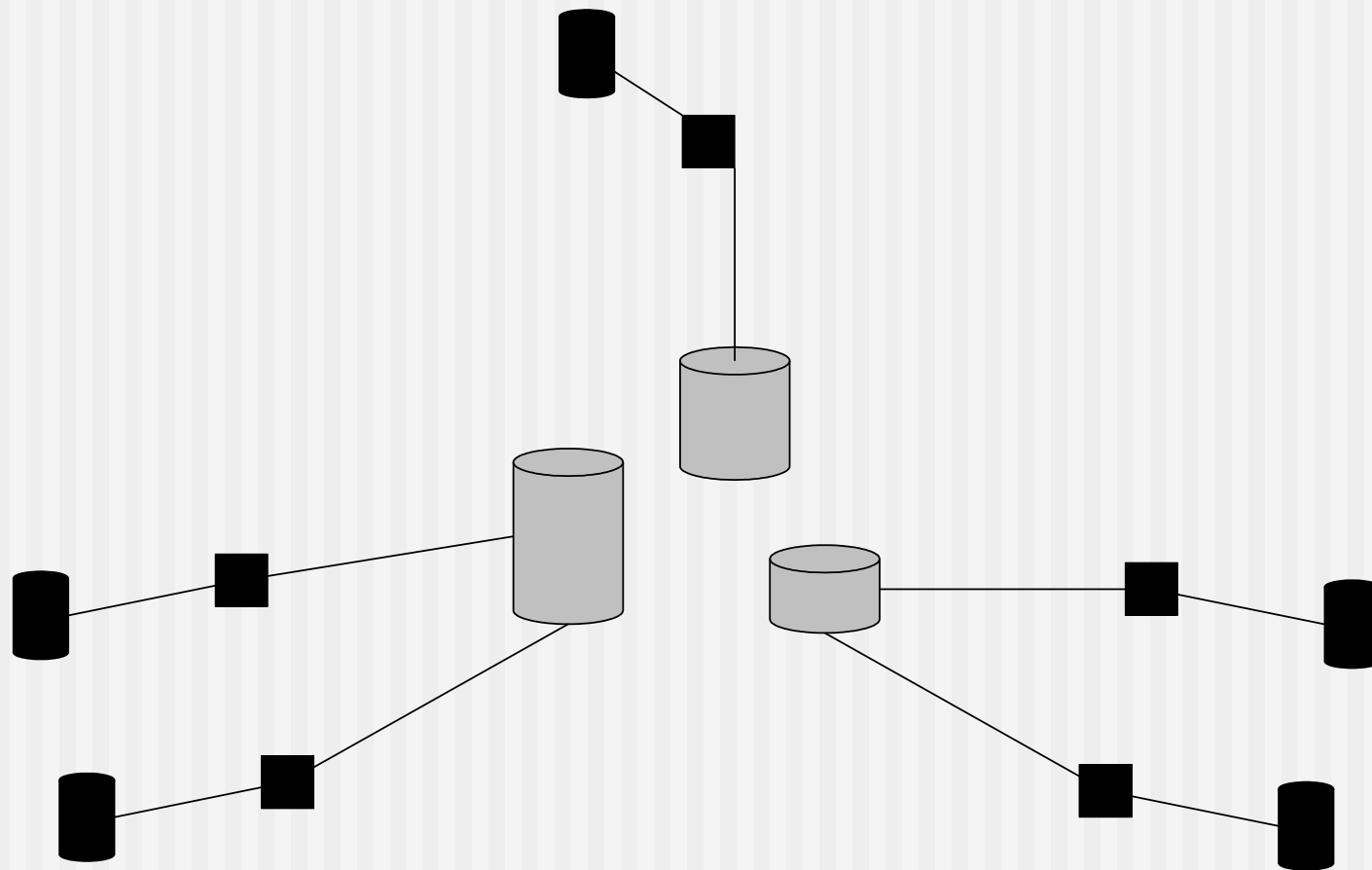


- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions → missing opportunities
- The bottlenecks are technology & skillsets
- **Examples**
 - **E-Promotions:** Based on your current location, your purchase history, what you like → send promotions right now for store next to you
 - **Healthcare monitoring:** sensors monitoring your activities and body → any abnormal measurements require immediate reaction

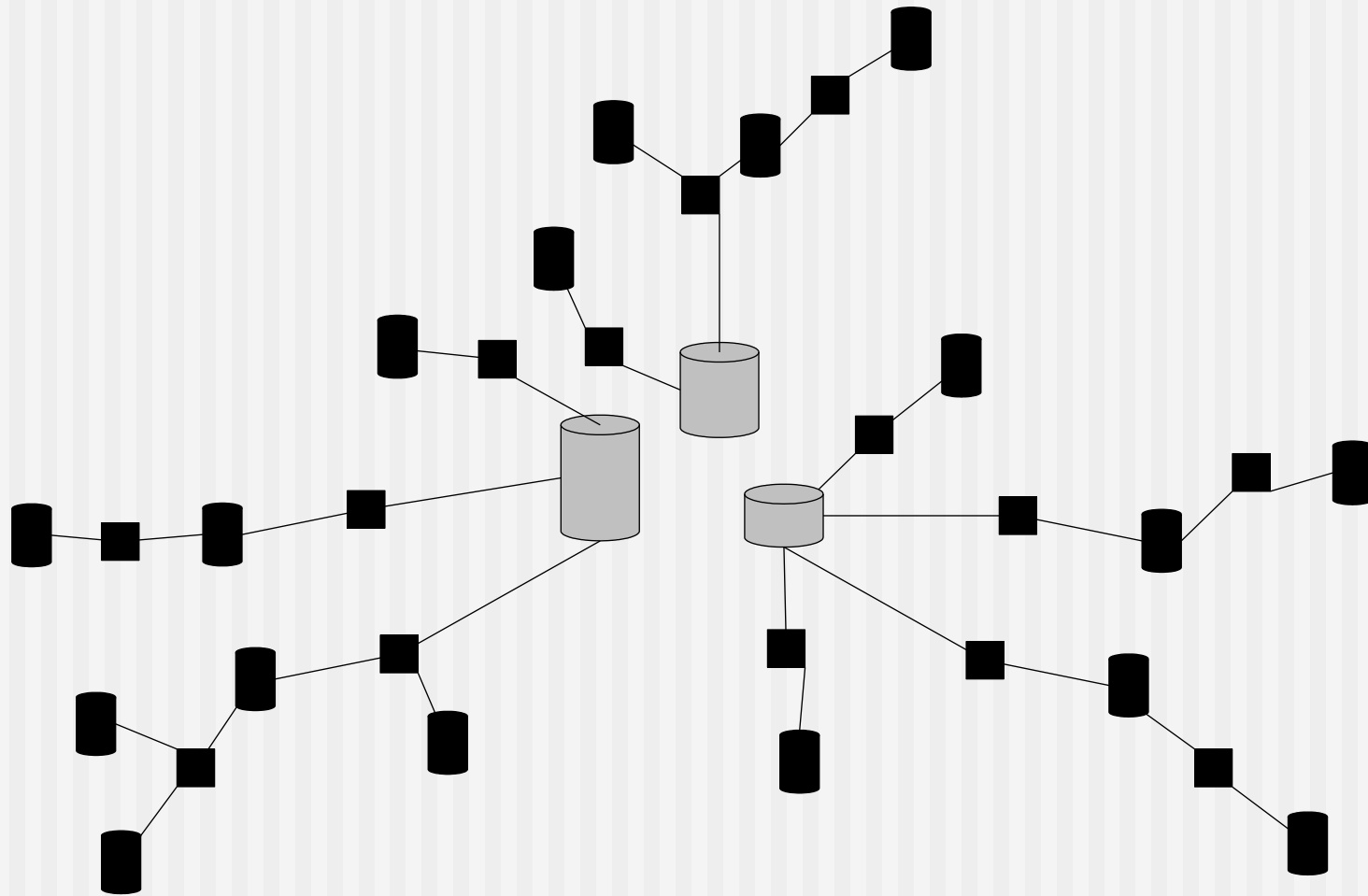
In the Beginning, life was simple...



But...

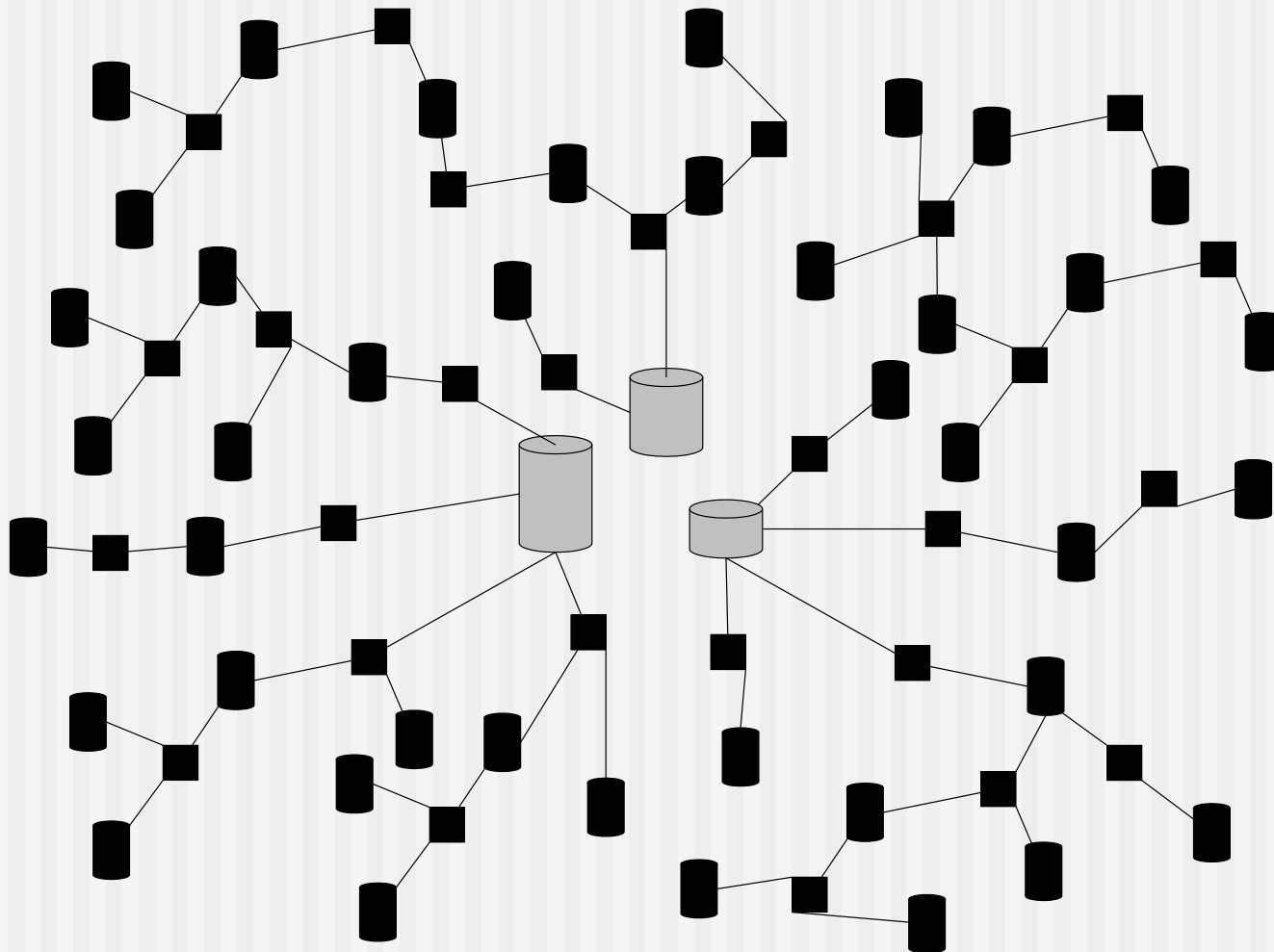


Our information needs...



Kept growing. (The Spider web)

A type of database schema



SOURCE: William H. Inmon

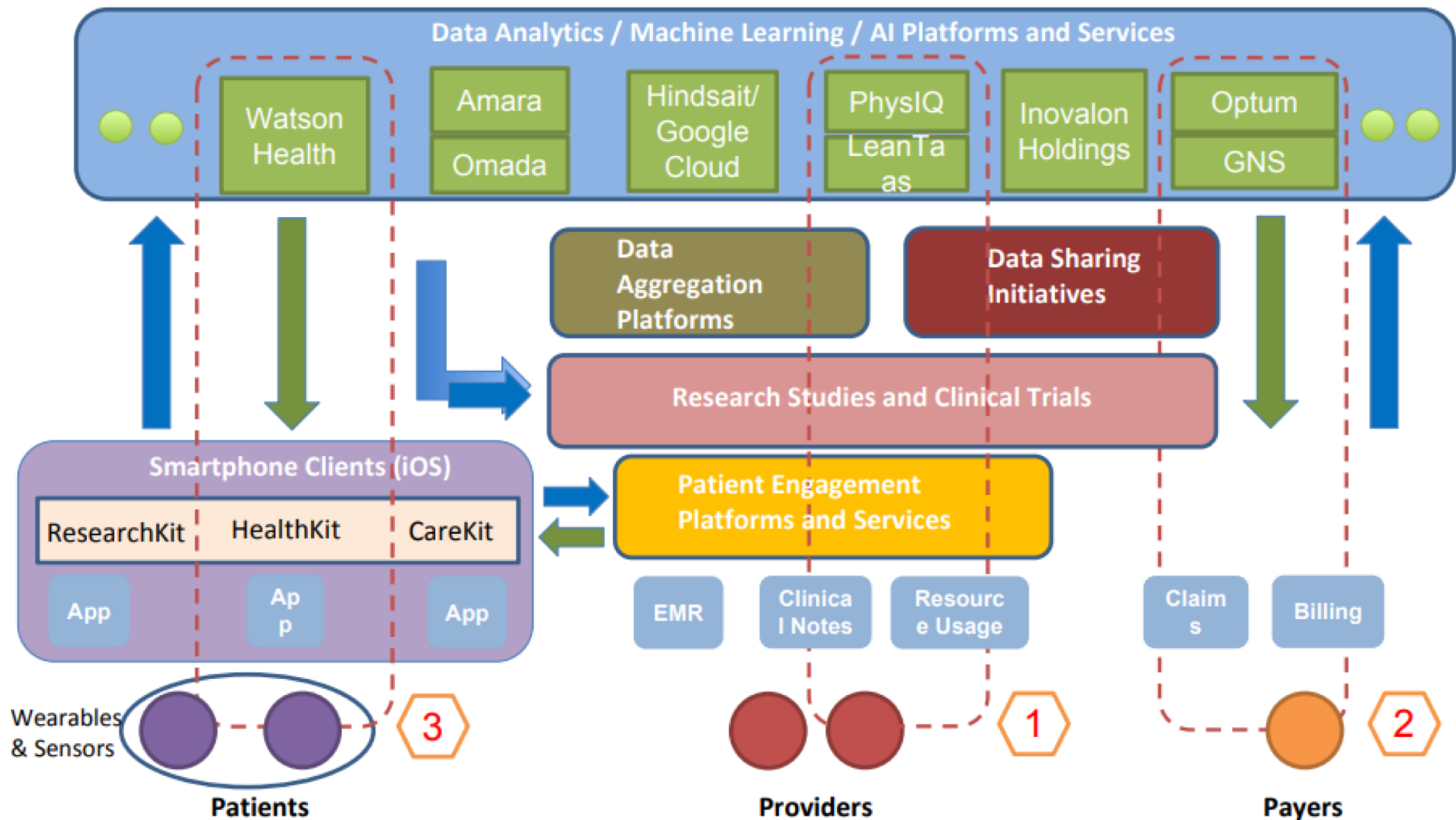
Mapping Value to Technology



Examples

- 1 Risk stratification/patient identification for integrated-care programs
- 2 Risk-adjusted benchmark/simulation of hospital productivity
- 3 Identification of patients with negative drug-drug interactions
- 4 Identification of patients with potential diseases ("patient finder")
- 5 Evaluation of clinical pathways
- 6 Evaluation of drug efficacy based on real-world data
- 7 Performance evaluation of integrated-care programs and contracts
- 8 Identification of inappropriate medication
- 9 Systematic reporting of misuse of drugs
- 10 Systematic identification of obsolete-drug usage
- 11 Personal health records

Emerging Landscape...Digital Health Gold Rush



Database Schema

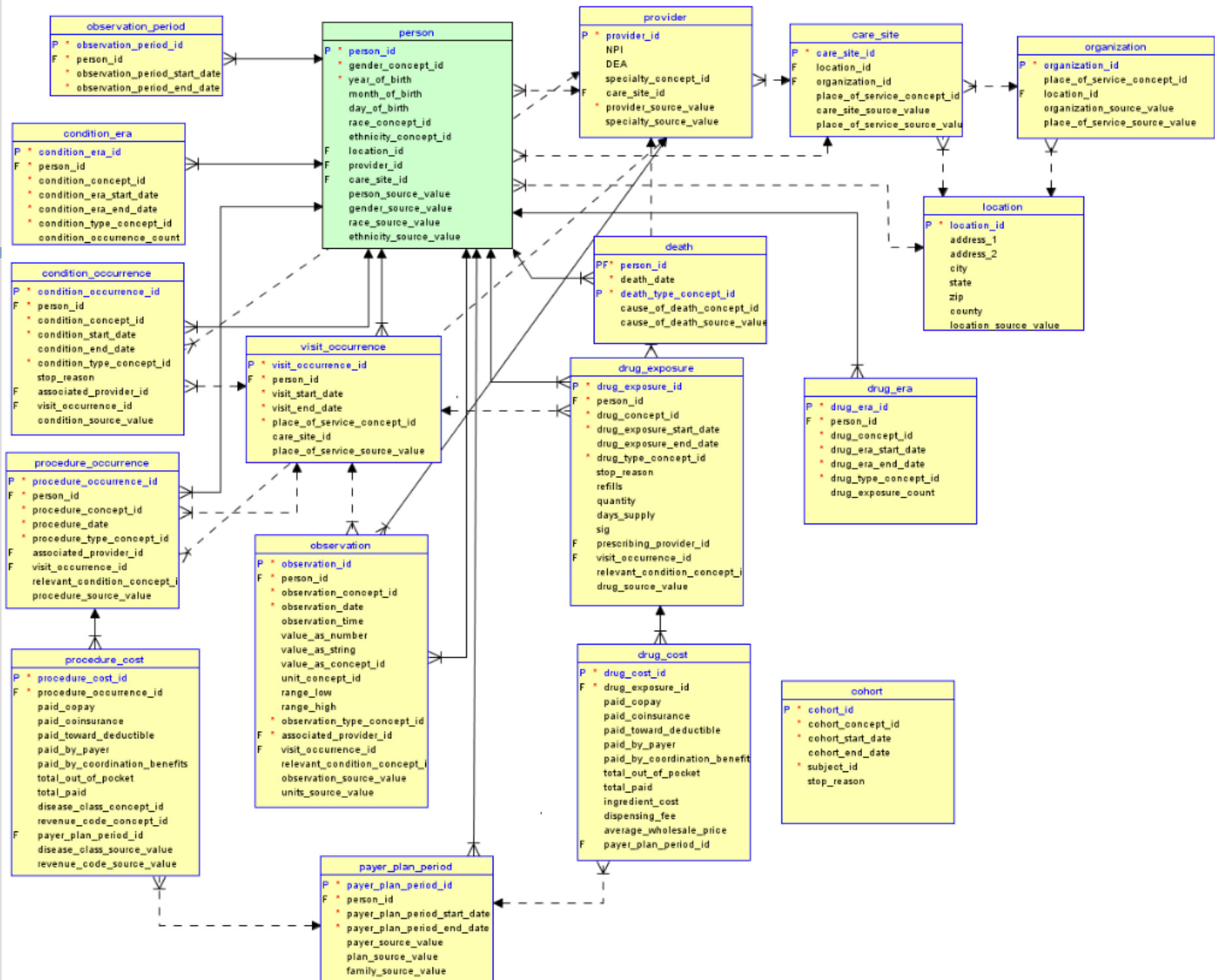
Medical Records

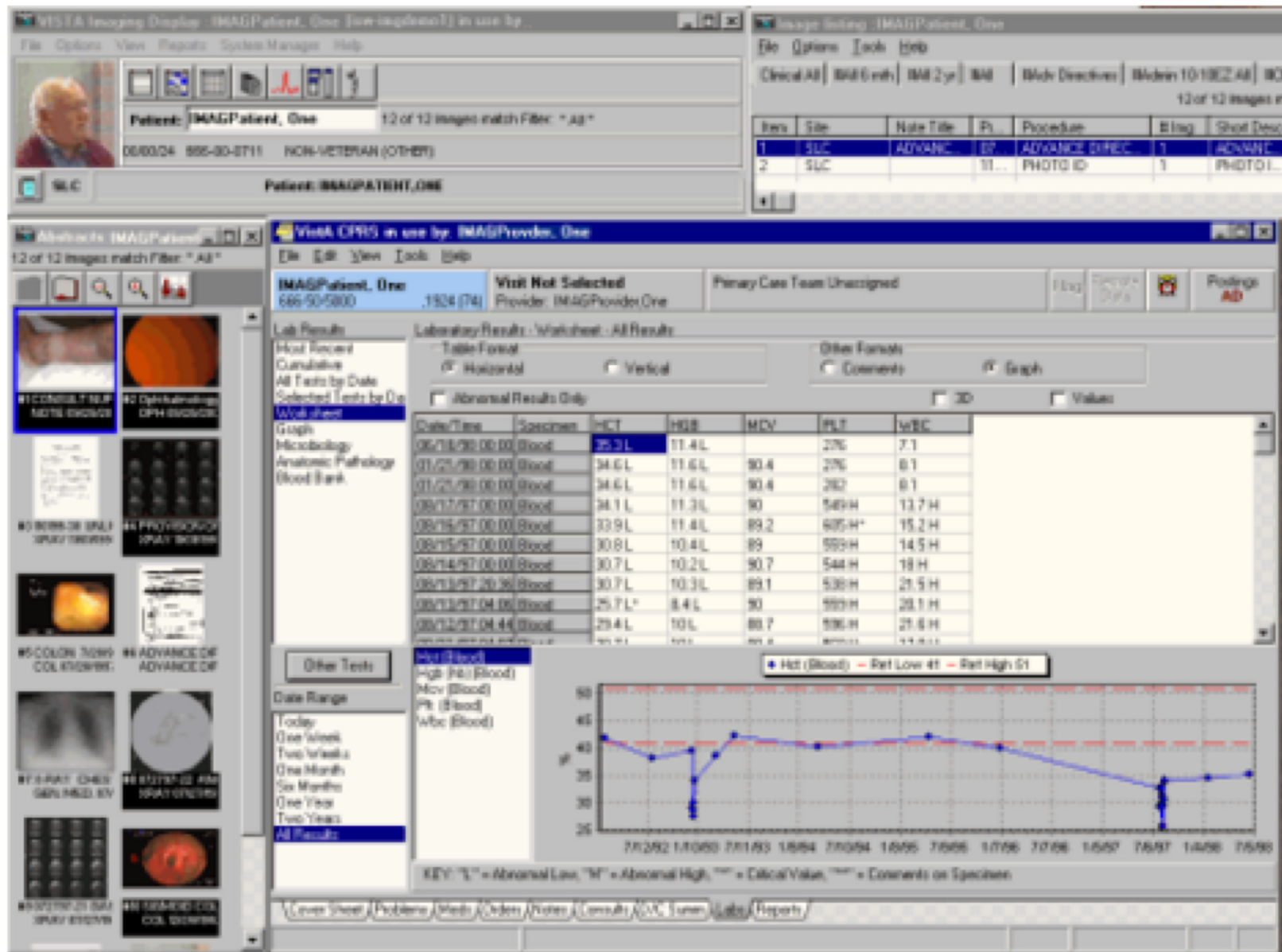
- Collects all patient information in a central repository
- Provide central point of access for
 - Patients
 - Care providers: physicians, nurses, etc.
 - Billing
 - Insurance reconciliation
- Hospitals, physicians, patients, procedures, records



MONGODB
WORLD'16 FOR DATA SCIENCE

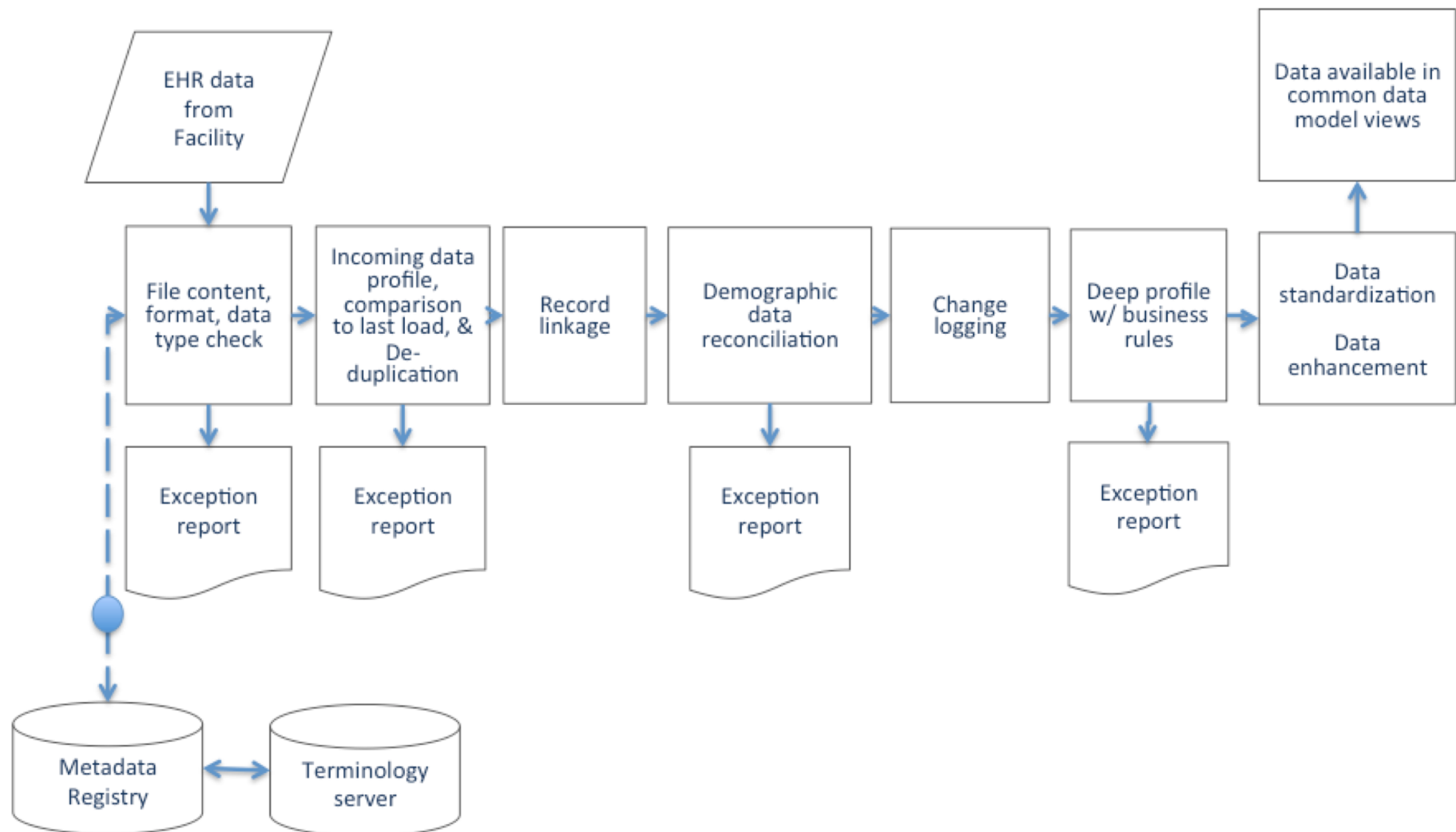
...time & effort to digitize this?





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

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.

Data powers UCSF work
in healthcare, research, education and administration



Request Clinical Data

For Research

For Clinical & Quality Initiatives

[Learn about the new simplified process!](#)

See Presentations & Handouts from
the 2017 Clinical Data Colloquium

New! Available for Research Use:

- [Data De-Identification Validation Service](#)
- [STOR Data](#)

Coming Soon:

- **Training opportunities** for machine learning and Natural Language Processing (NLP)
- Access to additional data assets and tools

[Learn about the vision for data.ucsf.edu](#)

Request Clinical Data for Research

UCSF, ZSFG and more!

- Electronic health record data from
 - STOR – now available! Records go back to 1988. [Read details about STOR](#)
 - APeX – Records from 2012 onward
 - ZSFG and other DPH facilities
- **Cost:** hourly recharge may apply

New! Simplified process for streamlining UCSF regulatory and compliance requirements. [Learn more!](#)

Request Clinical
Data for
Research

Data Assets

- ▶ Request Clinical Data for Research

[Request Clinical Data for Clinical & Quality Initiatives](#)

[De-identification Validation Service](#)

Data De-identification Validation

UCSF now has a Data De-identification Validation Service which enables you to collaborate with experts to verify that your datasets have been de-identified in accordance with HIPAA regulations and obtain the required certification. The data can then be shared with non-UCSF partners. [Learn more!](#)

More clinical data available for research:

Multi-institution

- **UCReX** – explore and obtain counts from 15 million+ patient records representing the combined patients of the five UC medical centers. [Read more](#)
- **PCORNet (PSCANNER) & OMOP** – patient-centered SCAlable National Network for Effectiveness Research (pSCANNER) is a network that will integrate data from existing networks and surrounding health systems covering over 37 million patients. [Read more](#)

Coming Soon

- **Training opportunities** for machine learning and Natural Language Processing (NLP)
- Access to additional data assets

[Learn about the vision for data.ucsf.edu](#)

Clinical Data Request Process: for Research

The Clinical Data Request Process allows you to request clinical data sets for use in research. You can also request to share clinical data sets with external third parties to facilitate work collaborations and vendor engagements.

We've set up a system that will automatically guide your request through the completion of the required steps. For detailed process steps, see below.

- [For Internal UCSF Faculty, Staff, Students](#)
- [For External Third Parties](#)

For Internal UCSF Faculty, Staff, Students

1. [Submit a Request for CTSI Data Extraction Consultation](#)

A consultant will work with you, provide guidance on whether you need de-identified, limited, or identified data sets, as applicable, will work with you to define a data specification that will be used by the Reporting Technical Team to fulfill your request.

2. The CTSI Consultant will help you to submit a clinical data request form to fulfill your request.

3. The ServiceNow system inform Principal Investigator (PI) of request

The ServiceNow system emails the Principal Investigator specified in the Clinical Data Request Form to notify them of your research data request. No action is required by the Principal Investigator or you for this step.

4. Verify current attestation form

The ServiceNow system checks whether you have a current Statement of Responsibility attestation form on file. An attestation form is considered current if it was signed by you during the past 2 years. If you do not have a current attestation form on file, the attestation form is emailed to you via DocuSign to read and sign.

Clinical Data Request Process

► For Research

[For Clinical and Quality Initiatives](#)

[IRB Checklist](#)

[Research Compliance](#)

[Instructions for Staff Providing Clinical Data](#)

Next stop: data.ucsf.edu/lexicon



COLLABORATE AND INNOVATE
through data-driven
insights

LEARN MORE >

Discovery. Innovation. Collaboration.

We partner to extract fresh insights from health care data.

Prevention after delivery

Find new ways to monitor type 2 diabetes in mothers at risk.

An alternative: [optumlabs.com](https://sp.ucop.edu/sites/its/immediate/ciosupport/internal/optumlabs/default.aspx) partnering with UCSF
<https://sp.ucop.edu/sites/its/immediate/ciosupport/internal/optumlabs/default.aspx>

Major Tasks in Data Management (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 Data Management (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
 - eg, 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 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 99999999”

*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