Types of Healthcare Data
What Do You Think Makes Up Health Data?
Health Data is Many Things
Big Data in Healthcare
How Big is “Big”

GROWTH OF HEALTHCARE DATA

48% increase annually

153 exabytes

2,314 exabytes

2013

2020

(one exabyte = one billion gigabytes)

Source: http://ow.ly/2okF30ftiSO
How Big is “Big”
“Big Data” in Healthcare

Volume

Data at Rest

Terabytes to Exabytes of existing data to process
“Big Data” in Healthcare

<table>
<thead>
<tr>
<th>Volume</th>
<th>Velocity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data at Rest</td>
<td>Data in Motion</td>
</tr>
<tr>
<td>Terabytes to Exabytes of existing data to process</td>
<td>Streaming data, requiring milliseconds to seconds to respond</td>
</tr>
</tbody>
</table>
### “Big Data” in Healthcare

<table>
<thead>
<tr>
<th>Volume</th>
<th>Velocity</th>
<th>Variety</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data at Rest</td>
<td>Data in Motion</td>
<td>Data in Many Forms</td>
</tr>
<tr>
<td>Terabytes to Exabytes of existing data to process</td>
<td>Streaming data, requiring milliseconds to seconds to respond</td>
<td>Structured, unstructured, text, multimedia,...</td>
</tr>
</tbody>
</table>
## “Big Data” in Healthcare

<table>
<thead>
<tr>
<th>Volume</th>
<th>Velocity</th>
<th>Variety</th>
<th>Veracity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data at Rest</strong></td>
<td><strong>Data in Motion</strong></td>
<td><strong>Data in Many Forms</strong></td>
<td><strong>Data in Doubt</strong></td>
</tr>
<tr>
<td>Terabytes to Exabytes of existing data to process</td>
<td>Streaming data, requiring milliseconds to seconds to respond</td>
<td>Structured, unstructured, text, multimedia,…</td>
<td>Uncertainty due to data inconsistency &amp; incompleteness, ambiguities, latency, deception, model approximations</td>
</tr>
</tbody>
</table>
Health Data Hits All 4 V’s

- **Genomic Data**
  - ~3GB
  - It is estimated the average hospital will generate 665 Tb of data.

- **Monitoring Equipment**
  - Can Generate 1000 Readings per second

- **Medical image archives**
  - Increasing by 20-40% annually.

- **X-Ray**
  - ~30MB
  - The number of hospitals using health information technology has more than doubled in the last two years.

- **4.9 Million remote monitoring devices**

- **80% coding variability among diagnosis and lab tests**
Types of Health Data
Omics Data

Broadly defined, omics represents the study of information contained within an individual’s genome and the biological derivatives of these genes.

- Omics thus includes fields concerned with the study of:
  - genes (genomics),
  - gene expression and RNA (transcriptomics),
  - proteins (proteomics),
  - metabolites (metabolomics)
  - lipids (lipidomics)

<table>
<thead>
<tr>
<th>Omics Data</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome</td>
<td>The full complement of genetic information both coding and non coding in the organism</td>
</tr>
<tr>
<td>Proteome</td>
<td>The protein-coding regions of the genome</td>
</tr>
<tr>
<td>Transcriptome</td>
<td>The population of mRNA transcripts in the cell, weighted by their expression levels</td>
</tr>
<tr>
<td>Physiome</td>
<td>Quantitative description of the physiological dynamics or functions of the whole organism</td>
</tr>
<tr>
<td>Metabolome</td>
<td>The quantitative complement of all the small molecules present in a cell in a specific physiological state</td>
</tr>
<tr>
<td>Phenome</td>
<td>Qualitative identification of the form and function derived from genes, but lacking a quantitative, integrative definition</td>
</tr>
<tr>
<td>Morphome</td>
<td>The quantitative description of anatomical structure, biochemical and chemical composition of an intact organism, including its genome, proteome, cell, tissue and organ structures</td>
</tr>
<tr>
<td>Interactome</td>
<td>List of interactions between all macromolecules in a cell</td>
</tr>
<tr>
<td>Glycome</td>
<td>The population of carbohydrate molecules in the cell</td>
</tr>
<tr>
<td>Secretome</td>
<td>The population of gene products that are secreted from the cell</td>
</tr>
<tr>
<td>Ribonome</td>
<td>The population of RNA-coding regions of the genome</td>
</tr>
<tr>
<td>Orfeome</td>
<td>The sum total of open reading frames in the genome, without regard to whether or not they code; a subset of this is the proteome</td>
</tr>
<tr>
<td>Regulome</td>
<td>Genome-wide regulatory network of the cell</td>
</tr>
<tr>
<td>Cellome</td>
<td>The entire complement of molecules and their interactions within a cell</td>
</tr>
<tr>
<td>Operome</td>
<td>The characterization of proteins with unknown biological function</td>
</tr>
<tr>
<td>Transportome</td>
<td>The population of the gene products that are transported; this includes the secretome</td>
</tr>
<tr>
<td>Pseudome</td>
<td>The complement of pseudogenes in the proteome</td>
</tr>
<tr>
<td>Functome</td>
<td>The population of gene products classified by their functions</td>
</tr>
<tr>
<td>Translatome</td>
<td>The population of proteins in the cell, weighted by their expression levels</td>
</tr>
<tr>
<td>Foldome</td>
<td>The population of gene products classified through their tertiary structure</td>
</tr>
</tbody>
</table>
The field of genomics is often broken down into additional subfields, such as:

- Genes
- Single-nucleotide polymorphism (SNP)
- Short tandem repeat (STR)
- Genome-wide association studies (GWAS)

GWAS has been widely applied to omics data as a popular approach for assessing the association of SNPs with various phenotypic traits, as well as for assessing the genetic etiology of diseases.
What Do We Use it For? – Precision Medicine

• Disease Links
  – Currently over 2,000 genetic tests available to aid in the diagnosis and therapy for more than 1,000 different diseases.

• Nutrigenomics, has investigated genome-wide influences on nutrition and the role of genetic polymorphisms in dietary-influenced disease.
Pharmacogenomics has explored how omics data can be utilized to identify the treatment efficacy of various medications and medication dosages for a particular individual.
Demographic Data

• Intrinsic
  • Physiological characteristics:
    – Age
    – Sex
    – Height,
    – Weight (to some extent)
    – Ethnicity
  • They may also include less obvious characteristics, such as allergies:
    – foods or medications.

• Extrinsic
  • Often characteristics derived from an individual’s environment and lifestyle, which may include:
    – Address
    – Martial status
    – Religion
    – Employment
      • type, location, and salary
    – Insurance plan
What Do We Use it For?

Controlling Confounding Factors

No Control
Simple correlation
$r = 0.49, p < .001$

Control
Controlling for recorded temperature
Partial $r = -0.02, p = .81$

Stratify Data for Analysis
Ie. Age and sex
Demographics as a Research Tool
Wellness Data
Commonly Associated with Fitness Trackers

- Headbands
- Sociometric badges
- Camera clips
- Smartwatches
- Sensors embedded in clothing

- Accelerometer
- Altimeter
- Digital camera
- Electrocardiogram
- Electromyograph
- Electroencephalogram
- Electrodensitometer
- Location GPS
- Microphone
- Oximeter
- Bluetooth proximity
- Pressure
- Thermometer
Wellness Data

- Most prominently associated with wearable fitness tracking
  - Fitbits, apply watches, Jawbones, etc.

- Wireless scales

- Digital pill boxes
  - Medication adherence

- Personal medical devices:
  - Digital glucometers
  - Personal blood pressure cuffs
  - Pulse oximeters.

- Pervasive Monitoring Tools (Home Monitoring)
  - Beds, chairs, and fall detection within flooring [22].
Hettler proposed a hexagon model, which included physical, emotional, social, intellectual, occupational, and spiritual wellness.

Today wellness data extends far beyond physical sensors to aspects of an individual’s social activity and dietary habits, and daily life.
What Do We Use it For?

• To maintain connections with individuals between formal clinical visits

• Pervasive monitoring
  – Real-time, multi-stream integration

• To provide individuals a means to monitor and quantify their health.

• Research Studies
  – Determining clinical trials efficiency
  – Evaluating novel therapeutics
  – Measuring functional recovery in patients
More Recently - mHealth
Administrative Data

• Financial data
  – Primarily public and private insurance claims.
    • Managed care plans, hospital discharge datasets, and revenue cycle management organizations.
    • Released by Department of Health and Human Services (HHS)

• Logistical data pertaining to several aspects of patient care,
  – Care-team composition and staffing metrics,
  – Resource utilization
    • Service metrics (e.g., inpatient, outpatient, and emergency department visits),
    • Medication usage
    • performed diagnostic tests and procedure

• Quality Assessments
  – Performance analyses of the service providers (i.e., reviews of clinician performance)
  – Patient satisfaction surveys
  – Hospital Quality Measures
What Do We Use it For?

Average US Health Spending Per Person in 2014: **$9,700**

- **$800** Unnecessary Services
  - e.g. Duplicate tests
- **$720** Excess Administrative Costs
  - e.g. Billing/coding errors
- **$495** Inefficient Care Delivery
  - e.g. Test results not shared
- **$400** Inflated Prices
  - e.g. Excessive profits
- **$285** Fraud
  - e.g. False Claims
- **$210** Prevention Failures
  - e.g. Missed flu shot

**$2,910** - Total Wasted Spending Per Person
What Do We Use it For?

- Quality and efficiency of care provided.

- Resulting burden of fraud on the healthcare system. A large body of research has been dedicated to the application of patient claims data and statistical methods to detect healthcare fraud.

- Incorrectly coded and billed patient charges can be identified and recovered without the need for a manual review of all service claims.
What Do We Use it For?

Staffing data not only to develop models to ensure adequate staffing, but to better understand the dynamics of the care environment from the provider’s perspective.

Targeting individuals, with specific resources used by population subgroups and how those resources are best utilized has the potential to greatly improve the quality of care.
Aggregate Data

- Data typically curated by large national and government entities capturing community, facility, district, province, country, and global levels
  - Among the most prominent are the United States Census Bureau
    - Decennial Census of Population and Housing, Economic Census, Census of Governments, American Community Survey (ACS) and Economic Indicators.
  - Bureau of Labor Statistics,
  - Agency for Healthcare Research and Quality (AHRQ).
Community Health Rankings

- Data can be drawn from many less-conventional channels. In fact, the WHO maintains an extensive set of such indicators, which are often used for a variety of purposes including program management, allocation of resources, monitoring country progress, performance-based disbursement, global reporting, and so on.

Electronic Medical Records (EMR)
Lots of Interfaces.. One Goal
Reality is a bit more complex..
Reality is a bit more complex..

Preliminaries
A wealth of Data – EMR Usage
Medical Information Mart for Intensive Care (MIMIC III v 1.4)
Comprises over 58,000 hospital admissions for 38,645 adults and 7,875 neonates.

Clinical Codes: Diagnoses, medications, procedures, and lab tests.

Demographics: Age, Sex, Patient Chart Data

620-bed tertiary academic medical center
77-bed Level I trauma center

Clinical records matched with physiologic data

Collected 2001-2012
Diagnoses

• MIMIC provides two forms of diagnosis information: ICD Diagnosis, Diagnosis Related Groups (DRG)
  – Diagnosis codes are billing codes!

• ICD Diagnosis Codes:
  – All Codes are in the ICD-9 Standard
  – Ordered by “priority”

• DRG:
  – Classification based on clinical factors and utilization of resources
    • 2 Types: All Payers / Health Care Financing Administration (now known as CMS DRGs)
ICD – 9 Codes

The International Statistical Classification of Diseases and Related Health Problems (commonly known as the ICD) provides alpha-numeric codes to classify diseases and a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances and external causes of injury or disease.

Infectious and Parasitic Diseases (001-009)
Neoplasms (140-239)
Endocrine, Nutritional and Metabolic Diseases, and Immunity Disorders (240-279)
Diseases of the Blood and Blood-Forming Organs (280-289)
Mental Disorders (290-319)
Diseases of the Nervous System and Sense Organs (320-389)
Diseases of the Circulatory System (390-459)
Diseases of the Respiratory System (460-519)
Diseases of the Digestive System (520-579)
Diseases of the Genitourinary System (580-629)
Complications of Pregnancy, Childbirth and the Puerperium (630-676)
Diseases of the Skin and Subcutaneous Tissue (680-709)
Diseases of the Musculoskeletal System and Connective Tissue (710-739)
Congenital Anomalies (740-759)
Certain Conditions Originating in the Perinatal Period (760-779)
Symptoms, Signs, and Ill-defined Conditions (780-799)
Injury and Poisoning (800-999)
DRG

- Patient classification scheme which provides a means of relating the type of patients a hospital treats to the costs the hospital incurs.
  - DRGs consist of classes of patients that are similar clinically and in terms of their consumption of hospital resources.

- The All Patient DRGs (AP-DRGs) are an expansion of the basic DRGs to be more representative of non-Medicare populations such as pediatric patients.

- Each Base APR-DRG is divided into 4 subclasses
  - Two types of Subclasses:
    - Severity of Illness (SOI)
    - Risk of Mortality (ROM)
Procedures

- MIMIC Provides procedural data in two formats:
  - ICD
    - Typically Less Detailed
    - Ordered, but not timestamped
  - Current procedural terminology (CPT)
    - Maintained, and copyrighted by the American Medical Association (AMA)
• 5 Digits Codes - Fall into three categories:
  – Category I – These five-digit codes have descriptors which correspond to a procedure or service. Codes range from 00100 - 99499.
  – Category II – These alphanumeric tracking codes are used for execution measurement. Using them is often optional.
  – Category III – These are provisional codes for new and developing technology, procedures, and services. The codes were created for data collection and assessment of new services and procedures.

• Two-character modifiers are appended to CPT codes to report special circumstances and to clarify or modify the description of the procedure.
Medications

- Extensive information is provided for the medications offered to patients:
  - Dosage and unit
    - E.g. 1500 mg
  - Form and Form Unit
    - E.g. 3 TAB
  - Route
    - IV (Intravenous) / PO (taken orally)
Same Drugs, Different Names

Since there are many forms of the same drug, MIMIC provides: Generic Sequence Number (GSN) and the National Drug Code (NDC).

This is a platform called RxNorm which helps to map these ingredients.
MIMIC provides detailed clinical notes across a range of categories:
- ECG
- Respiratory
- Radiology
- Discharge summary
- Rehab Services
- Nursing/other
- Nutrition
- Pharmacy
- Social Work
- Case Management
- Physician
- Consult

MS is a 77 y/o woman who presents with a rash. The rash began one week ago and first appeared on her back and extended to her abdomen on the left side. She first noticed the rash after working in her garden. For the first 3 days she had no associated symptoms, however on the 4th day, she began to experience severe pain “like being stung by a bee”. The pain is worse in the evenings and she is having difficulty sleeping. She has been taking Tylenol every 4-6 hours and using Vitamin E lotion with minimal relief. She is worried she may have poison ivy.

She has a history of hypertension. Her medications include: Nifedipine XL 30mg daily and Metoprolol 50mg BID.

MS is an elderly woman who appears energetic and in no distress. VS: BP 130/80, P 60, RR 12, T 37. Skin exam reveals an erythematous rash of grouped vesicles with clear fluid, approximately half of which are crusted over. The rash extends from the midline of the back anterior to the left side of the T12/L1 dermatome region.

A new onset painful rash in a dermatomal distribution. This is most likely secondary to herpes zoster because of the location, distribution and associated pain. A less likely possibility is contact dermatitis, given her recent gardening. She may also have a cellulitis, but the lack of temperature doesn’t support this.

#1 MS will continue to use Tylenol for pain relief.
#2 MS was informed that she may continue to experience pain after the rash resolves.
#3 MS should return to clinic if the pain becomes more severe or if she develops warm, redness or a fever.
Microbiology and Lab Results

• The lab results contains both in-hospital laboratory measurements \textit{and} out of hospital laboratory measurements from clinics which the patient has visited. This data contains:
  – Value (can be non-numeric)
  – Unit (if appropriate)
  – Flag
    • Indicates whether the value is considered abnormal, using pre-defined thresholds.

• Microbiology results capture data from various cultures
  – Specimen which is tested for bacterial growth
  – Organism that grew (none, i.e. negative culture)
  – Antibiotic Testing
Chart Events

• This is by far the largest table in MIMIC and contains an immense array of information, reflective of anything put into a patient's medical chart.

• This can include:
  – Patients' routine vital signs
  – Ventilator settings,
  – Code status,
  – Mental status
  – Even small items like bed adjustments
  – Much more...
Waveforms

Waveform Database contains thousands of recordings of multiple physiologic signals ("waveforms") and time series of vital signs ("numerics") collected from bedside patient monitors in adult and neonatal intensive care units (ICUs).

- Contains 22,317 waveform records and 22,247 numerics records, which have been matched and time-aligned with 10,282 MIMIC-III Clinical Database records.

- Waveforms almost always include:
  - ECG signals, continuous arterial blood pressure (ABP), Fingertip photoplethysmogram (PPG) signals, and respiration

- Numerics typically include:
  - Heart and respiration rates, SpO2, and systolic, mean, and diastolic blood pressure,
Navigating MIMIC
Tracking a Patient

• **PATIENTS**: Every unique patient in the database

• **ADMISSIONS**: Every unique hospitalization for each patient

• **ICUSTAYS**: Every unique ICU stay in the database

• **SERVICES**: Clinical service under which a patient is registered

• **TRANSFERS**: Patient movement from bed to bed within the hospital, including ICU admission and discharge

• **CALLOUT**: Information regarding when a patient was cleared for ICU discharge and when the patient was actually discharged
Tracking a Patient

- Each individual is given a unique SUBJECT_ID
  - This will only correspond to one person across all of MIMIC
- Each time that individual comes to a hospital they receive a hospital ID HADM_ID
  - One individual can have multiple HADM_ID
- During their admission to the hospital the individual can utilize services in multiple intensive care units. Each time they are admitted to an ICU they receive an ICUSTAY_ID
  - Multiple ICU stays either within the same hospitalization, or across multiple hospitalizations, or both,
Some Data is Collected at the Hospital-Level

- DIAGNOSES_ICD: Hospital assigned diagnoses, coded using the International Statistical Classification of Diseases and Related Health Problems (ICD) system
- DRGCODES: Diagnosis Related Groups (DRG), which are used by the hospital for billing purposes.
- LABEVENTS: Laboratory measurements for patients both within the hospital and in outpatient clinics
- MICROBIOLOGYEVENTS: Microbiology measurements and sensitivities from the hospital database
- PRESCRIPTIONS: Medications ordered, and not necessarily administered, for a given patient
- PROCEDURES_ICD: Patient procedures, coded using the International Statistical Classification of Diseases and Related Health Problems (ICD) system
Some Data is Collected at the ICU-Level

- CAREGIVERS: Every caregiver who has recorded data in the database (defines CGID)
- CHARTEVENTS: All charted observations for patients
- DATETIMEEVENTS: All recorded observations which are dates, for example time of dialysis or insertion of lines.
- INPUTEVENTS_CV: Intake for patients monitored using the Philips CareVue system while in the ICU
- INPUTEVENTS_MV: Intake for patients monitored using the iMDSof MetaVision system while in the ICU
- NOTEEVENTS: Deidentified notes, including nursing and physician notes, ECG reports, imaging reports, and discharge summaries.
- OUTPUTEVENTS: Output information for patients while in the ICU
- PROCEDUREEVENTS_MV: Patient procedures for the subset of patients who were monitored in the ICU using the iMDSoft MetaVision system.
Code Lookups

- D_ICD_DIAGNOSES: Dictionary of International Statistical Classification of Diseases and Related Health Problems (ICD) codes relating to diagnoses
- D_ICD_PROCEDURES: Dictionary of International Statistical Classification of Diseases and Related Health Problems (ICD) codes relating to procedures
- D_ITEMS: Dictionary of ITEMIDs appearing in the MIMIC database, except those that relate to laboratory tests
- D_LABITEMS: Dictionary of ITEMIDs in the laboratory database that relate to laboratory tests
Announcements

• Office Hours:
  – Next Week 8/27-8/31:
    • Monday and Wednesday from 4pm-7pm in iCeNSA 384 Nieuland
    – All subsequent weeks it will return to normal times
    • Wednesday 1 PM to 3 PM / Thursday 10 AM to 12 PM in 308 Cushing

• Assignment 1 out today
  – Due 8/31
  – Requesting Data and Setting up Python Environment for Labs, Assignments, and Projects
Next Class – Data Understanding