

MAI4CAREU

Master programmes in Artificial
Intelligence 4 Careers in Europe

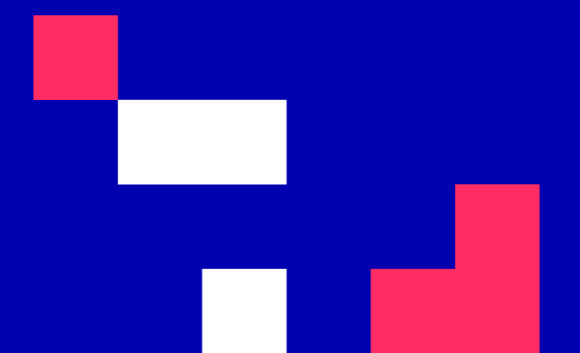


University of Cyprus

MAI643: Artificial Intelligence in Medicine

Kalia Orphanou

January – May 2023



Clinical Data : Their Acquisition, Storage and Use

(Some material drawn from Hersh, WR, 2022. Health Informatics: Practical Guide, 8th Edition, slides)

Clinical Data: Their Acquisition, Storage and Use

CONTENTS

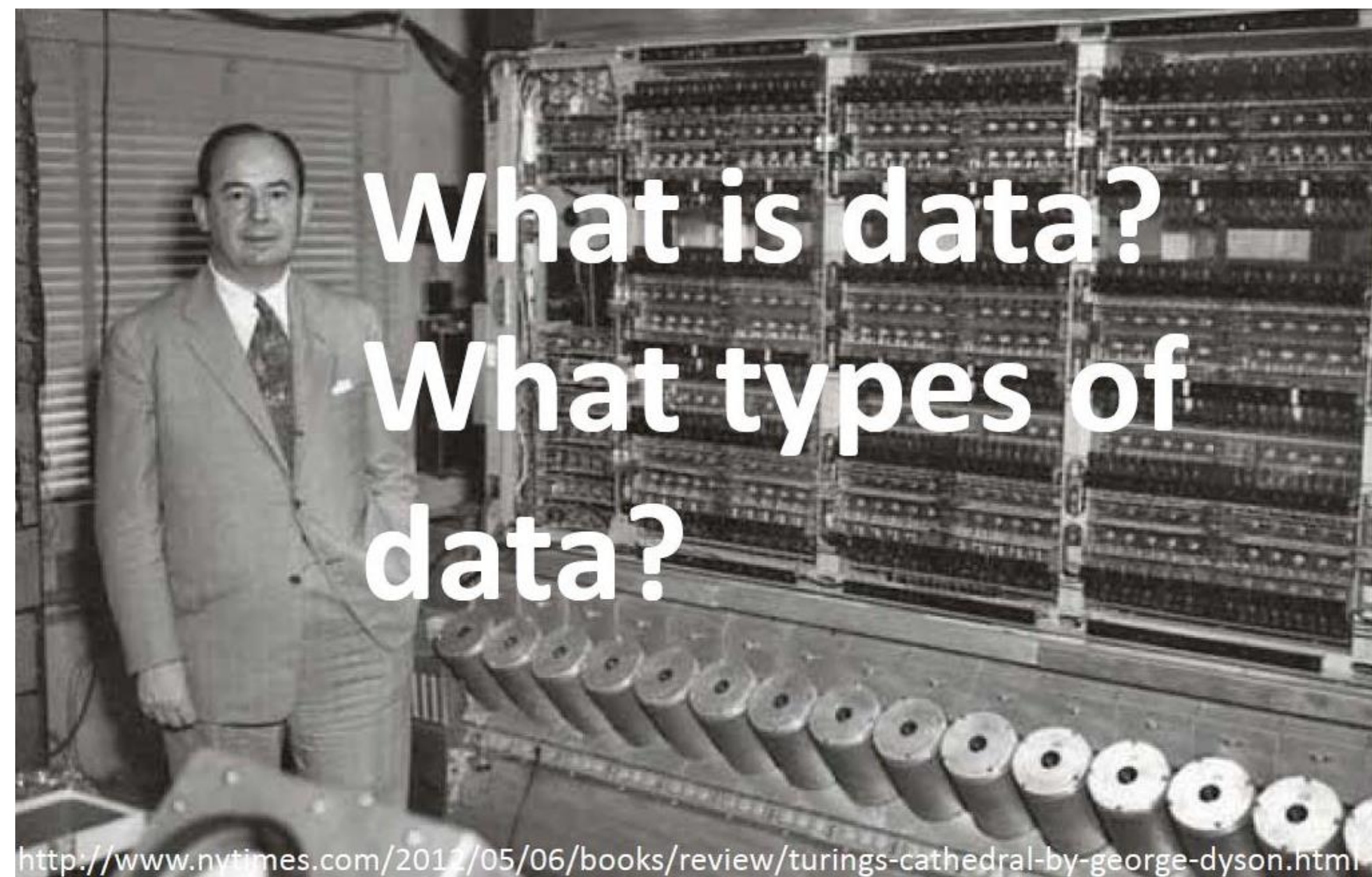
1. Types of clinical data
2. Sources of clinical data
3. Uses and coding of clinical data
4. History of health records
 - a. Personal health records
 - b. Electronic health records
5. Processing of clinical data
 - a. Machine learning methods
 - b. Feature engineering
6. Narrative clinical data and Natural Language Processing (NLP)
7. Storage of clinical data
 - a. Federated databases
 - b. Cloud servers
 - c. Data integration
8. Future of clinical data

INTENDED LEARNING OUTCOMES

Upon completion of this unit on Clinical Data, students will be able to:

1. Point out the different types of clinical data
2. Overview the main sources, users and uses of clinical data
3. Appreciate the benefits and challenges of electronic health records
4. Overview the different methods of storing clinical data
5. Differentiate between the methods used for processing either structured or unstructured clinical data
6. Grasp the importance of feature engineering before applying ML in clinical data
7. Appreciate the challenges of applying NLP techniques for textual clinical data
8. Discuss the future of collecting, storing, using health data

Unit 4



What clinical data are?

- A datum is a single observation
- Clinical data are the collection of observations about a patient
 - Example from John Halamka of Geek Doctor blog fame (<https://dmice.ohsu.edu/hersh/halamka-record.pdf>), part of Personal Genomes Project (<https://www.personalgenomes.org/us>)
- Each datum about a patient has a minimum of four elements:
 - the patient (Bill Hersh)
 - the attribute (heart rate)
 - the value of the attribute (50 beats per minute)
 - the time of the observation (1:00 pm on 7/1/1990 – many ways to record dates!)

Hersh, 2022

Where do clinical data come from?

- Disease – pain, altered body function, etc.
- Monitoring – follow-up of ongoing problems
- Preventive measures (primary, secondary) – screening
- Pre-work/school examination

From Hersh, 2022

Types of Clinical Data

- Narrative – recording by clinician
- Numerical measurements – blood pressure, temperature, lab values
- Coded data – selection from a controlled terminology system
- Textual data – other results reported as text
- Recorded signals – EKG, EEG
- Pictures – radiographs, photographs, and other images

Data Sources

- Birth and death records
- Medical records at physician offices, hospitals, nursing homes, etc.
- Medical databases within various agencies, universities, and institutions.
- Physical exams and laboratory testing
- Disease registries
- Self-report measures: interviews and questionnaires
- Social media posts

Who is collecting the data?

- Health care practitioners
- Nurses: Making observations for future references, assessment of patients to physicians
- Physicians: Collection and interpretation
- Office staff and admissions personnel – demographic data
- Radiologists technicians perform the Xray examinations and radiologists interpret the Xray results.
- Public health researchers & organizations
 - Population medical data – surveys and social media posts
- Others (insurance companies, pharmacists, HIT vendors)

Assessment of a stable patient

- Chief complaint
- History of the present illness
- Past medical history
- Social history
- Family history
- Review of systems
- Physical examination
- Testing – lab, x-ray, other
- Assessment and plan

From Hersh, 2022

Uses of clinical data

Form basis of historical record:

- Support communication among providers
- Anticipate future health problems
- Record standard preventive measures
- Coding and billing
- Provide a legal record
- Support clinical research

From Hersh, 2022

Some complications of data

- Circumstances of observation – e.g., how was heart rate taken? pulse? EKG?
- Uncertainty – how accurate is patient reporting, measurement, device?
- Time – what level of specificity do we need?
- Imprecision vs. inaccuracy

From Hersh, 2022

Coding of Data

- Historically performed by a Clinical Coding Specialist (CCS)
- Major purpose has historically been for reimbursement
- A core issue in biomedical informatics has been how to generate and use coded data for other purposes.
- Trade-offs
- Standardization of language vs. freedom of expression

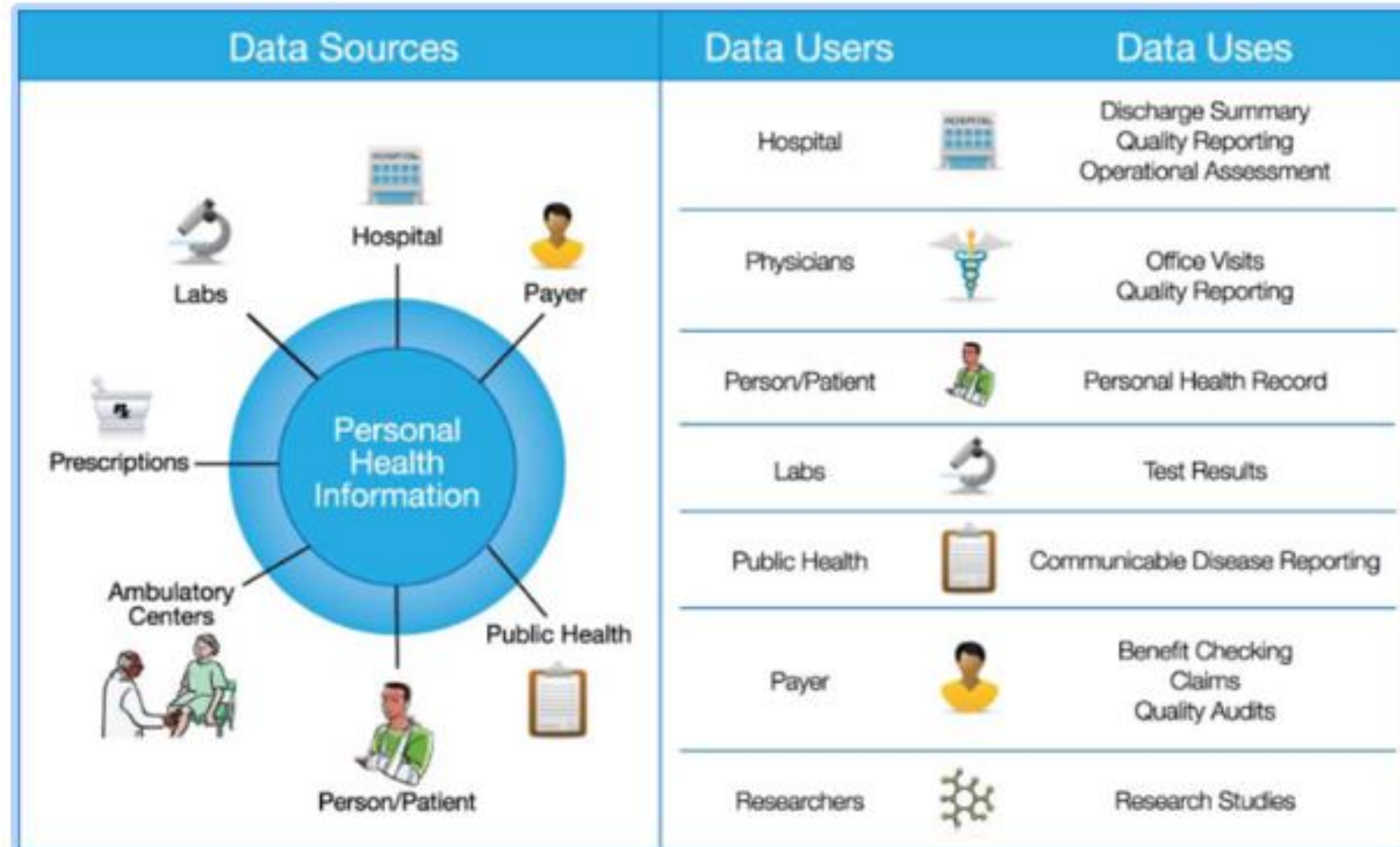
From Hersh, 2022

Coding of Data

- Time to narrate vs. code
- Other difficulties
- Creating and maintaining coding systems
- Structuring coding systems to capture meaning

From Hersh, 2022

Many sources, users, and uses of clinical data



From Hersh, 2022

Another important part of data: Social determinants of health (SDOH)

(Artiga, KFF, 2018)

Economic Stability	Neighborhood and Physical Environment	Education	Food	Community and Social Context	Health Care System
Employment	Housing	Literacy	Hunger	Social integration	Health coverage
Income	Transportation	Language	Access to healthy options	Support systems	Provider availability
Expenses	Safety	Early childhood education		Community engagement	Provider linguistic and cultural competency
Debt	Parks	Vocational training		Discrimination	Quality of care
Medical bills	Playgrounds	Higher education		Stress	
Support	Walkability				
	Zip code / geography				

Health Outcomes
Mortality, Morbidity, Life Expectancy, Health Care Expenditures, Health Status, Functional Limitations

From Hersh, 2022



Electronic Health Records (EHR)

History and perspective of the medical record

- Data can be organized as:
 - Practitioner (physician)-centered
 - Patient-centered
- Orientations (not mutually exclusive) include
 - Time-oriented – organized chronologically
 - Department-oriented – organized by department
 - Problem-oriented – organized by focus on problems

From Hersh, 2022

History and perspective of the medical record (ctd)

- Earliest medical records were physician-oriented
- Hippocrates said over 2,500 years ago that the medical record should:
 - Accurately reflect course of disease
 - Indicate possible causes of disease
- Before era of widespread medical diagnostic testing, record consisted mostly of observations
- Overview video – <https://www.youtube.com/watch?v=WxQhtPmmwNY>

From Hersh, 2022

Modern-day medical record

- Mixture of patient- and problem-oriented approaches
- Usually has problem list, which may or may not be well-kept
- In general, each provider or institution maintains its own record
- Creator of the medical record is assumed to be “owner,” although patients can request access

From Hersh, 2022

Additional challenges in the modern era

- Coordinating care requires better communication among providers
- Increasing cost of care requires justification and documentation of expenditures
- Patients change plans, so their records should be portable
- Informed consumers desire more participation in care decisions, which includes access to their records
- They also want security and other protections of their information

From Hersh, 2022

Additional challenges in the modern era

- No single vendor has complete solution
- Many take “best of breed” approach to matching components
- Now must align goals for using **electronic health records (EHR)** following the **HITECH Act**.
- HITECH Act encourages healthcare providers to adopt EHR and improve privacy and security protections of the clinical data.

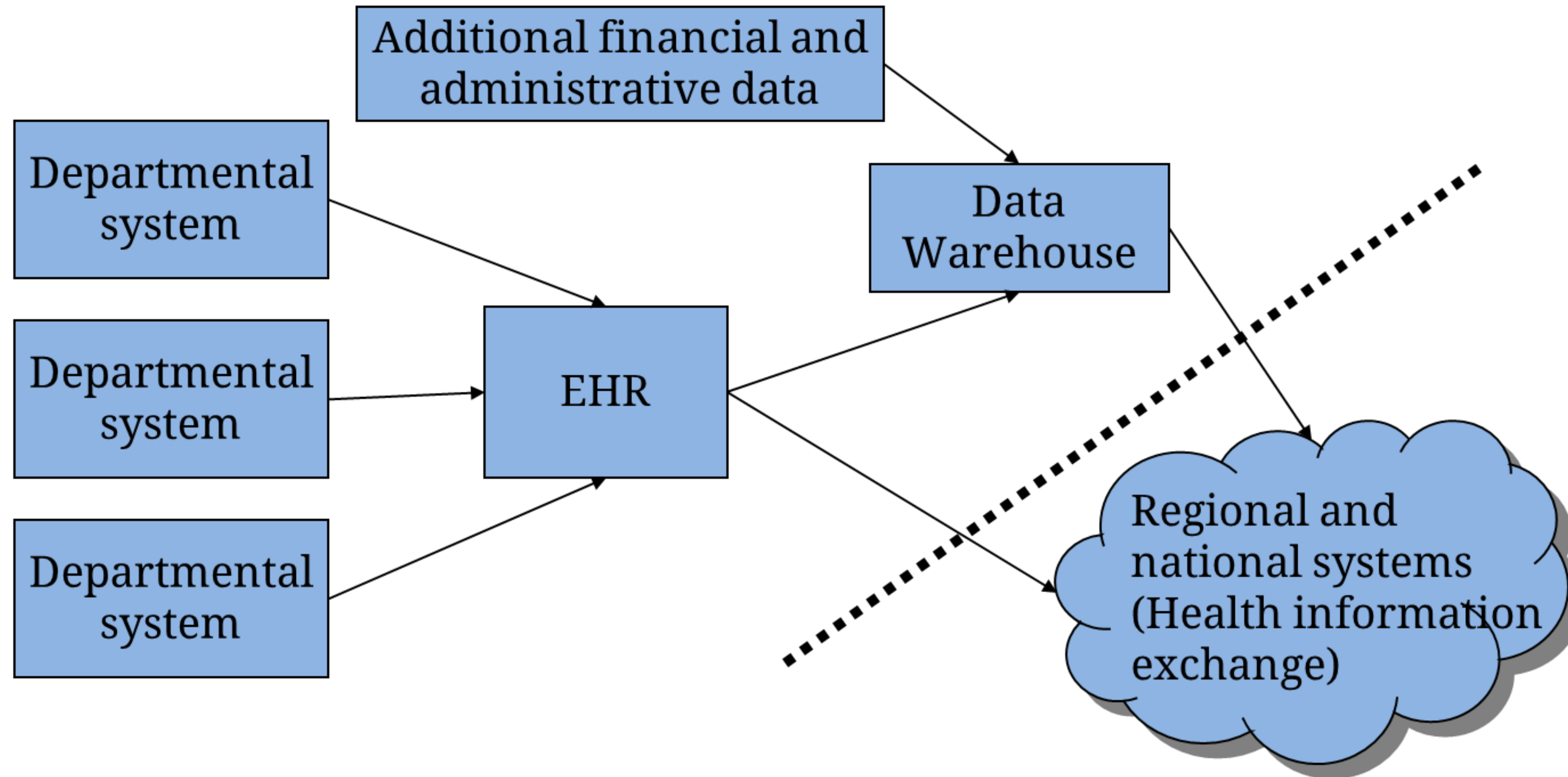
From Hersh, 2022

All functions should address five healthcare quality criteria

- Improve patient safety
- Support delivery of effective patient care
- Facilitate management of chronic conditions
- Improve efficiency
- Have feasibility of implementation

From Hersh, 2022

EHR data flow (typically in hospitals or large clinics)



From Hersh, 2022

Benefits and challenges of EHR

Benefits

- Improved patient care
- Personal health records
- Data science and artificial intelligence
- Clinical decision support
- Quality measurement and improvement
- Multi-user ubiquitous access to patient data
- Clinical research (re-uses of clinical data)
- Better communication with other providers and with patients
- Public health

Challenges

- Data quality (i.e. incomplete/inaccurate data, without documentation)
- Inadequate adherence to standards, resulting in lack of interoperability
- Understanding clinical narrative text
- Free-form entry by historical methods (non-structured data)
- Privacy, confidentiality, and security
- Implementation (time-consuming, cost)

From Hersh, 2022

Personal health record (PHR) definition

- A collection of information about your health i.e. a folder of medical papers or your health details kept in your smartphone.
- Information stored:
 - Your doctor's names and phone numbers
 - Allergies, including drug allergies
 - Your medications, including dosages
 - List and dates of illnesses and surgeries
 - Chronic health problems, such as high blood pressure
 - Living will or advance directives
 - Family history
 - Immunization history
 - Exercise, dietary habits

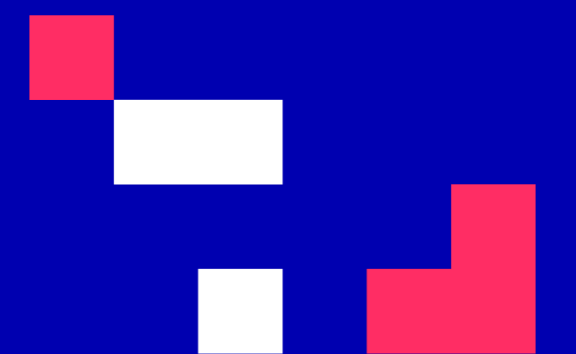
From Hersh, 2022

Types of Personal Health Records

- Tethered ([Epic MyChart](#))
 - Stored in healthcare provider's EHR
 - Often allows communication with provider
 - May allow patient to add information
- Standalone ([WebMD Health Manager](#))
 - Isolated application on individual computer – may use mobile device or Web site
- Interconnected or integrated ([AppleHealth](#))
 - Separate application but integrated with provider EHR(s) for, e.g., test results, scheduling, data collection, etc.

From Hersh, 2022

Processing Clinical Data



Role of AI Scientist

- Aggregate collected data from different data sources
- Process multiple types of data using AI techniques
- Develop medical decision-support systems
- Communicate the results with clinicians/patients

AI Scientist: Information to be Processed

- What is the patient's history (development of a current illness; other diseases that coexist or have resolved; pertinent family, social, and demographic information)?
- What symptoms has the patient reported?
- When did they begin, what has seemed to aggravate them, and what has provided relief ?
- What physical signs have been noted on examination?
- How have signs and symptoms changed over time?

AI Scientist: Information to be Processed (ctd.)

- What laboratory results have been, or are now, available?
- What radiologic and other special studies have been performed?
- What medications are being taken and are there any allergies?
- What other interventions have been undertaken?
- What is the reasoning behind the management decisions?

Categories of Clinical Data

- Structured data (or tabular data)
 - Represented in a form of a matrix where each row corresponds to a patient record (sample)
 - Columns represent the features, and
 - One or more columns include the value to be predicted.

- Unstructured data
 - Image data
 - Waves representing sound
 - Text

Feature extraction is necessary for unstructured data

Feature Engineering for Clinical Data

- Medical Images
 - Application of **Deep Learning** techniques to automatically extract and select features
 - **Feature extraction:** Computation of image properties such as color, texture, or shape
 - **Feature selection:** to detect only the most relevant features

- Biomedical Signal Processing
 - Automatic processing applying for patient monitoring, medical diagnosis and rehabilitation purposes – Deep learning techniques
 - **Feature extraction:** Representation of biosignals by means of Fourier and wavelet basis functions and auto-regressive parameters in a feature vector.
 - Feature selection to detect only the most relevant features

Feature Engineering for Clinical Data (Ctd.)

▪ Narrative/Text Data

- Linguistic features extracted using natural-language processing (NLP) techniques
- Representation of the extracted features as feature vector/matrix or structured data
- Feature selection methods to find the most relevant ones

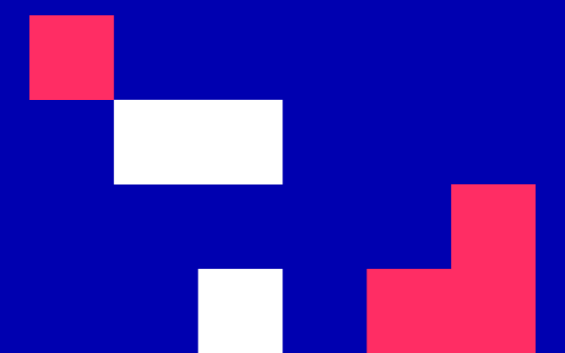
▪ DNA Microarray Data

- Structured data with few samples and a large amount of features
- Feature selection to eliminate redundant and irrelevant features and to help experts detect underlying relationships between gene expression and a given disease.

Benefits of Feature Selection in Clinical Data

- Dimensionality reduction
- Cost reduction by avoiding non-necessary exams/tests
- Extraction of information from images
- Understanding the reasons behind disagreements regarding disease diagnosis among image-analysis experts

Narrative (Textual) Data



Sources of Clinical Textual Data

- in medical databases i.e PubMed
- in the scientific medical literature
- social media posts
- clinical notes - In health care facilities where patient information mainly occurs in narrative notes and reports

Natural Language Processing Approaches in Medicine

- Information extraction – process narrative to structured data
- Question answering – finding answers in text
- Text summarization – summarizing from a large amount of content
- Text generation – from speech
- Text classification – categorizing text into known types
- Machine translation
- Sentiment analysis and emotion detection
- Conversational agents – medical chatbots

Information Extraction

▪ Looking for patterns in text

- NLP can support health consumers and patients looking for information about a particular disease or treatment.
- Extract social risk factors from clinical notes or/and social media to understand the prevalence and progression of a disease in a specific community.

Question Answering

- user submits a natural language question, which is then automatically answered by a QA system.

- health care consumers, health care professionals, and biomedical researchers frequently search the Web to obtain information about a disease, a medication, or a medical procedure.

Text Summarization/Generation

- From medical literature and web resources
 - One or several documents given as input and produces a single, coherent text, which synthesizes the main points of the input documents.
- **Text generation**
 - A doctor can read his/her diagnostic of a patient or an X-ray, and the NLP system generates a Word document to be sent to the patient or other doctors.

Machine Learning and Clinical Text Data

- **First era (early era of AI):** Manually developed lexicons, grammars and algorithms.
- **Second era:** Machine learning applied to lexicons and grammars
- **Third era:** Deep learning applied to all

(Deng, 2018)

Use Cases for NLP in Cancer Care

- Identify potential clinical trial matches
 - Advances information extraction from complex patient documents
 - Precise information retrieval for clinical case histories and outcome studies
 - Apply predictive models and care coordination rules to unstructured patient narratives*
 - Analyze patient narratives for insights into treatment outcomes
-
- ***Patient narratives:** A written or spoken account of a person's health history and experiences with diseases, treatments, and healthcare providers.
 - Individual's health status includes physical, emotional, and social factors that may impact their well-being
- From Hersh, 2022

Open AI chatbot – Medical Diagnosis

K

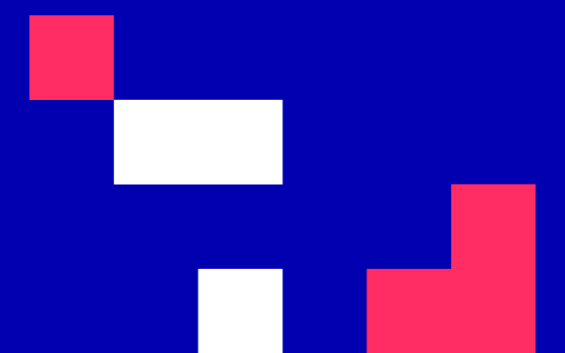
I have cough, fever, running nose, chest pain what do I have?



It's possible that you have a respiratory infection such as the flu or a cold. Chest pain with these symptoms can also be a sign of pneumonia. It is recommended to see a doctor for a proper diagnosis and treatment plan.

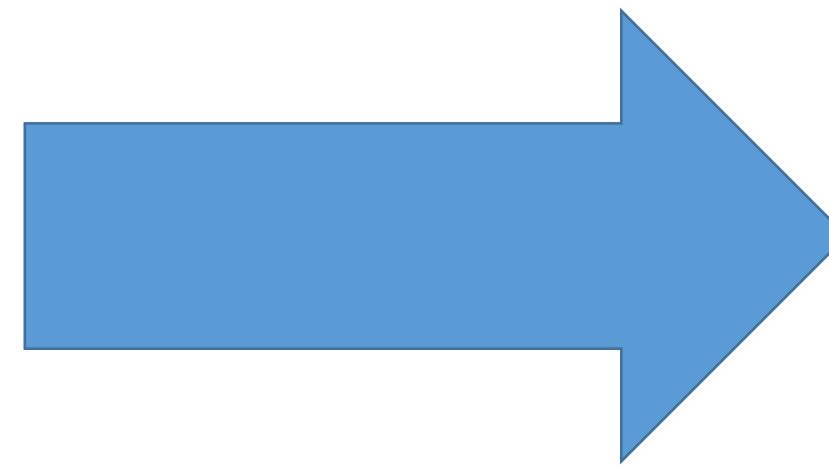


Data Storage



Data Integration

- Mobile/social media data
 - Text data
 - Signal data
 - Measurements of pressure/heart rates
- Telehealth data
 - Text data
- EHR data
 - Laboratory test results
 - Patient history (Temporal data)
- Socio/economic data
- Molecular and genetic data



**Aggregation of different
data modalities**

Interoperability

From Hersh, 2022

Data Integration (ctd.)

Meta-dimensional analysis can be divided into three categories:

- **Concatenation-based integration** involves combining data sets from different data types at the raw or processed data level before modelling and analysis;
- **Transformation-based integration** involves performing mapping or data transformation of the underlying data sets before analysis, and the modelling approach is applied at the level of transformed matrices;
- **Model-based integration** is the process of performing analysis on each data type independently, followed by integration of the resultant models to generate knowledge about the trait of interest.

From Shortliffe, 2022

Data Federation Models

- **Local Data Store model:** Every data holder hosts its own data on its own server. Others can view and analyze the data after creating an account to have access to these data but they cannot download them.
- **One Single Centralized Datastore model:** Data from multiple sources are aggregated into one “**data warehouse**” for secure data access and analysis.
- **Federated query model:** Using **federated databases**; when independent geographically dispersed databases are networked in such a way that they can respond to queries as if all the data were in a single virtual database.
 - Data requesters can submit a query to a federated query service and have that query be routed to all databases participating in that federation.
 - Data holders have full access to their data

From Herch, 2022

Cloud Computing for Storing Clinical Data

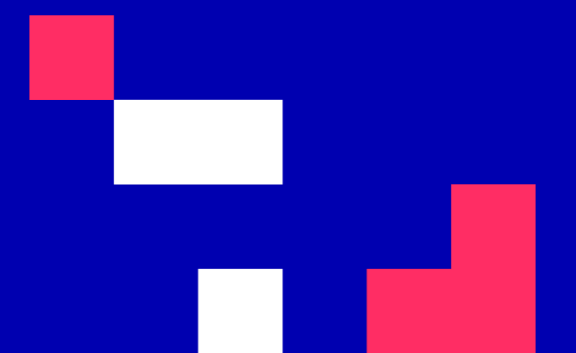
- Healthcare's digital transformation leads to the greater need for data sharing between systems and physicians.
- Enable **collaboration** between physicians and other medical data providers
- Cloud storage of big data
 - Imaging or diagnostic data, data from patient monitors within the hospital, or from remote monitoring solutions
- **Data privacy, security, monitoring platforms**
 - Backup data
- Cloud-based solutions enable the integration of informatics applications
 - Performance support for ML

Data Resharing and Reuse

- **FAIR (Findable, Accessible, Interoperable, Reusable) principle**
- **Findability** requires indexing and shared metadata and persistent Digital Object Identifiers (DOIs) such as from DataCite or other services that span disciplines.
- **Accessibility** brings up data rights, ownership, access policies, and fair (as in just) credit for data sharing—all of which are wide open issues.
- **Reusability** include both reusability by humans and computers

From Hersh, 2022

Future of Clinical Data and Computation



Motivational Example

Andre is a 47-year-old man with mild Type 2 diabetes. He was returning from a business trip overseas when he felt short of breath, out of sorts, and had occasional sharp chest pains. He signed onto a **telehealth service** offered through his employer. The **telehealth service's chatbot** interviewed Andre, using an avatar that was Hispanic, as Andre is. After an initial set of questions, the chatbot handed over the case **to a human physician, who conducted a video consultation** with Andre while reviewing **his electronic health record data** along with **his respiratory rate, body temperature, oxygen saturation and other data from his smartwatch**. The physician recommended that Andre get evaluated in person at the nearest Emergency Room (ER). Andre is getting worried. On his way to the ER, Andre **asks Siri** what he might have. Siri tells him scary diagnoses like pneumonia, and something called pulmonary embolism. **Siri explains** that pulmonary embolism is when a blood clot forms in a leg after prolonged sedentariness (like a long flight) and breaks off to the lungs causing chest pain and shortness of breath.


From: Shortliffe et al., 2022

Motivational Example

At the ER, Andre was first seen by a **resident physician-in-training** who ordered **multiple tests including labwork, a chest x-ray, and a chest CT**. Based on those data, a decision support system ran **predictive models that resulted in a ranked list of differential diagnoses**, with an intermediate probability for pulmonary embolism. The resident presented the case to Dr. Jackson, **the attending ER physician**. After reviewing the data and output, Dr. Jackson went to talk with Andre and examine him. She noticed crackles in the lungs, an S3, a prominent right-sided cardiac lift and elevated jugular venous pressure. On further questioning, Andre mentioned that he had had a **“bad cold” about 1 month before** and had been feeling unwell even before the business trip. Suspecting biventricular failure from viral myopericarditis, Dr. Jackson ordered an **echocardiogram** and admitted Andre.

From: Shortliffe et al., 2022

Real-time Clinical Data

- The ability to access such data in real time requires:
 - health data interoperability encompassing network computing,
 - data standards, and
 - sociotechnical data sharing mechanisms
- Siri and the decision support system in the ER  **automated reasoning**

Open Challenges for Processing Clinical Data

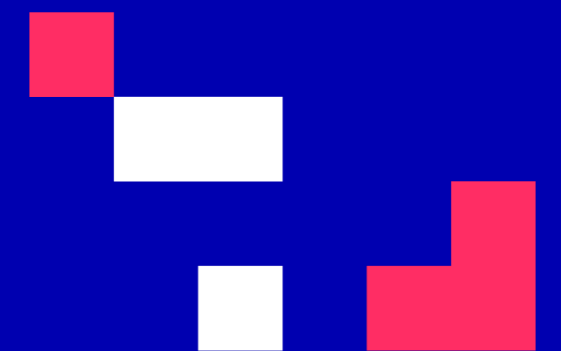
- **Data annotation:** time-consuming and expensive process
- Big data – many data that are not annotated or they are inaccurately annotated
- Difficult to aggregate and use them for machine learning.
- New **semi-supervised approaches** are emerging that rely on small amounts of labelled data to predict missing labels for larger datasets but they may encompass bias.

Open Public Data

- General Repositories
 - GenBank, EMBL, HMCA, ...
- Specialized by data types
 - UniProt/SwissProt, MMMP, KEGG, PDB, ...
- Specialized by organism
 - WormBase, FlyBase, NeuroMorpho, ...
- Details: <http://hci-kdd.org/open-data-sets>
- The Research Resource for Complex Physiologic Signals: physionet.org
- A curated list of medical data for machine learning
<https://github.com/beamandrew/medical-data>

SUMMARY

- There are multiple types of clinical data and multiple sources for collecting these data
- Stakeholders who are involved in collecting/processing clinical data include the clinicians, nurses, specialized doctors, hospitals' receptionists, patients, AI/data scientists...
- The main steps for processing clinical data using machine learning methods include feature engineering (feature construction/feature selection)
- Natural language processing techniques are used to process textual clinical data for information extraction, classification and summarization.
- Data storage for clinical data varies from federated databases, electronic health records and cloud servers.
- Sharing clinical data among multiple medical providers is very important. The data should be aggregated and follow the FAIR principle.



Discussion

- As AI scientists, which do you believe will be the major challenges that you have to deal when processing clinical data and how you would try to handle them?
- Which type of clinical data do you feel that is the most challenging to process and why (your personal opinion)?

References

- Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet.* 2015 Feb;16 (2):85–97. doi: 10.1038/nrg3868. Epub 2015 Jan 13. PMID: 25582081
- Hersh, WR, 2022. *Health Informatics: Practical Guide*, 8th Edition.
- Shortliffe, Edward H., and James J. Cimino, eds. *Biomedical informatics: Computer applications in health care and biomedicine.* Springer Science & Business Media, 2021.
- Sim I. Mobile Devices and Health. *N Engl J Med* 2019; 381:956–968.
- V.L. Patel and T.A. Cohen's chapter in T.A. Cohen, V.L. Patel and E.H. Shortliffe (editors), *Intelligent Systems in Medicine and Health: The Role of AI*, Springer, 2022