Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Categories of Data in Digital Health

Borchert, Dr. Schapranow

Data Management for Digital Health

Winter 2023

Agenda
Pillars of the Lecture

Hasso Plattner Institut

Medical Use Cases

Biology Recap

Oncology

Nephrology

Infectious Diseases

Technology Foundation

Data Sources

Data Formats

Processing and Analysis

Software Architectures

Machine Learning

Data

ML

Refine

Evaluate

Prediction + Probability

Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

Agenda
Pillars of the Lecture

Hasso Plattner Institut

Medical Use Cases

Biology Recap

Oncology

Nephrology

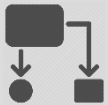
Infectious Diseases

Technology Foundation

Data Sources

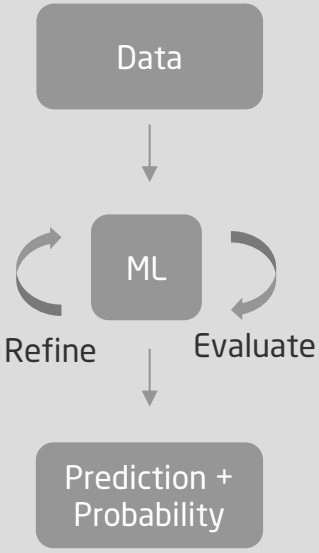
Data Formats

Processing and Analysis



Software Architectures

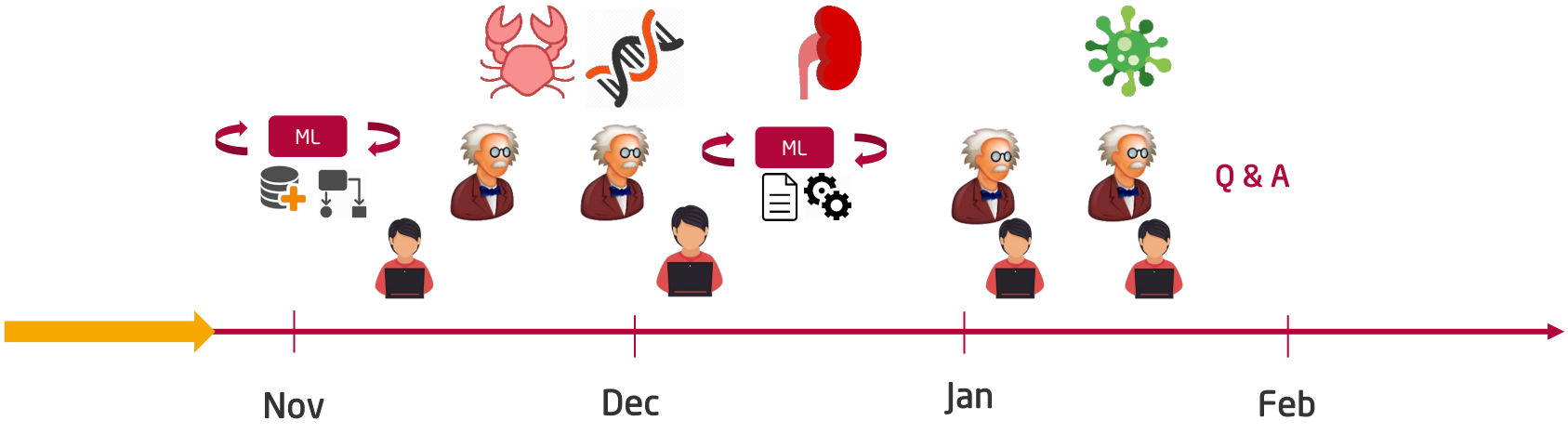
Machine Learning



Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

Lecture Schedule



- Lecture Kickoff
- Actors in Healthcare
- Digital Health Data

- Machine Learning (ML) Foundations
- Use Case Oncology
- Biology Recap

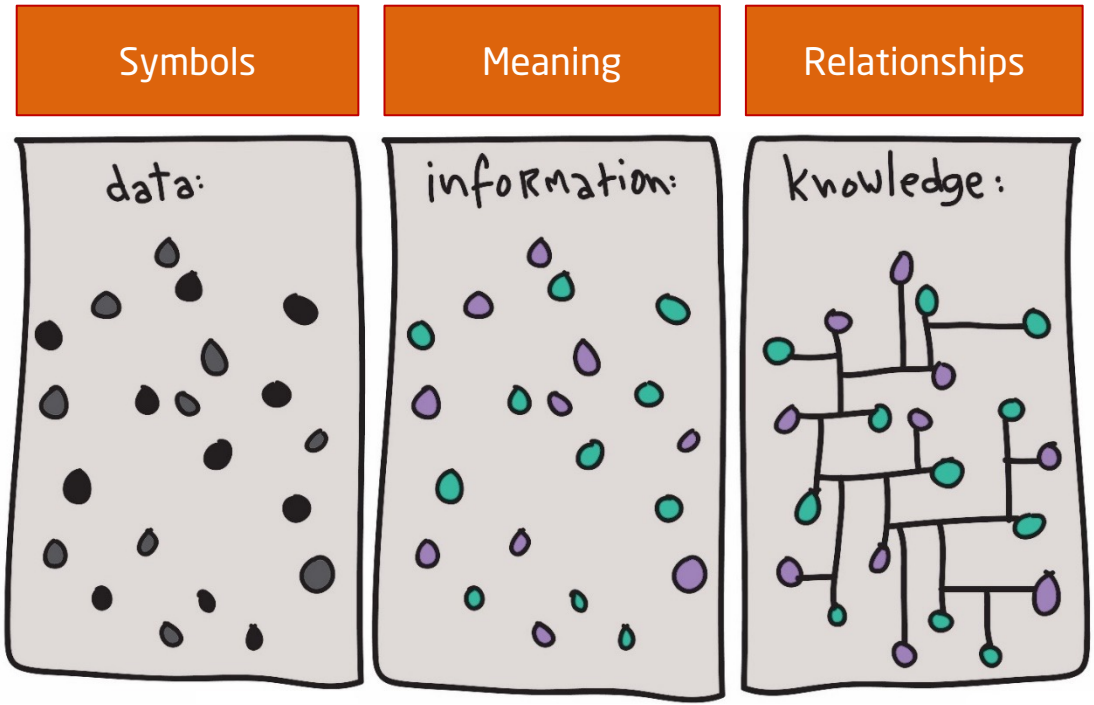
- Natural Language Processing
- Use Case Nephrology & Intensive Care
- Supervised ML & Deep Learning

- Use Case Infectious Diseases
- Unsupervised ML

Agenda

- Terms you should know
- Overview about Digital Health Data
  - Sources of Data
  - Data Characteristics
  - Data management
  - Challenges

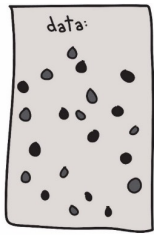
Terms You Should Know

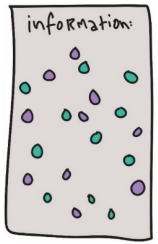


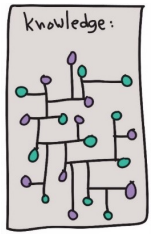
Source: David Somerville based on the original by Hugh McLeod

Zins, Chaim. "Conceptual approaches for defining data, information, and knowledge." Journal of the American society for information science and technology 58.4 (2007): 479-493.

Terms You Should Know







- 01000100 01001001 01000001 01000010 01001100 01001111
- DIABLO := Interpret(01000...01111)
- DIABLO is a gene on Chr. 12 associated with cancer progression known from a publication

Terms You Should Know

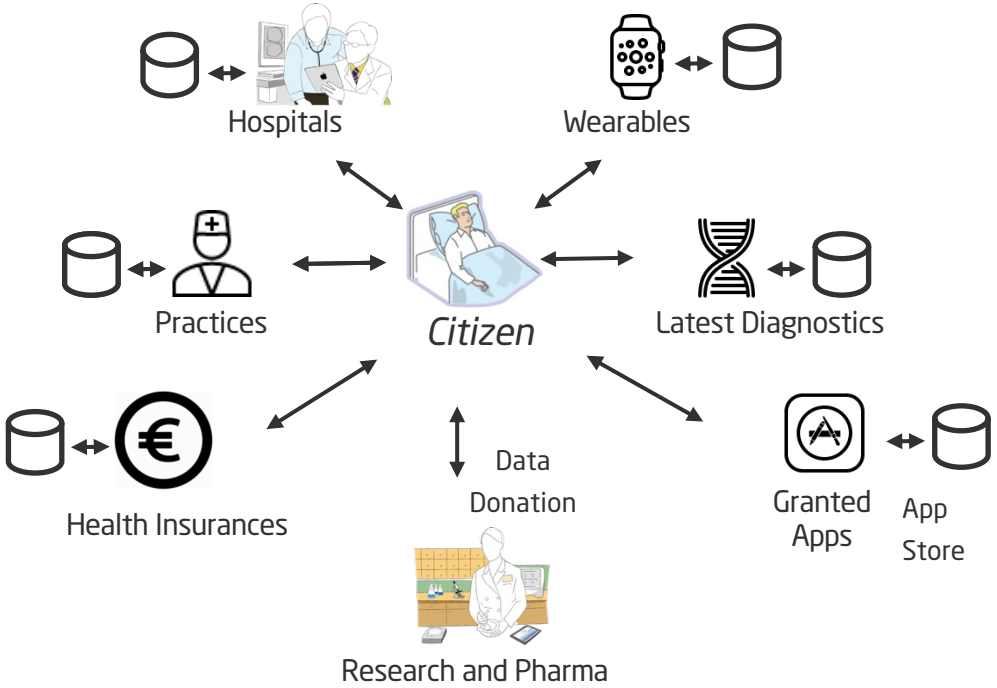
- **Data** := Information symbolized as material-energetic pattern, a.k.a. message

- **Information** := Something knowable, i.e. something we could know
  - **Perceiving** := Perceived information from someone else
  - **Interpretation** := Received information
  - **Processing** := Derived information
  - **Forgottenness** := Lost information

- **Knowledge** := Stored experience

Wendt, Siegfried. Nichtphysikalische Grundlagen der Informationstechnik: Interpretierte Formalismen. Springer-Verlag, 1991.

Sources of Digital Health Data



Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

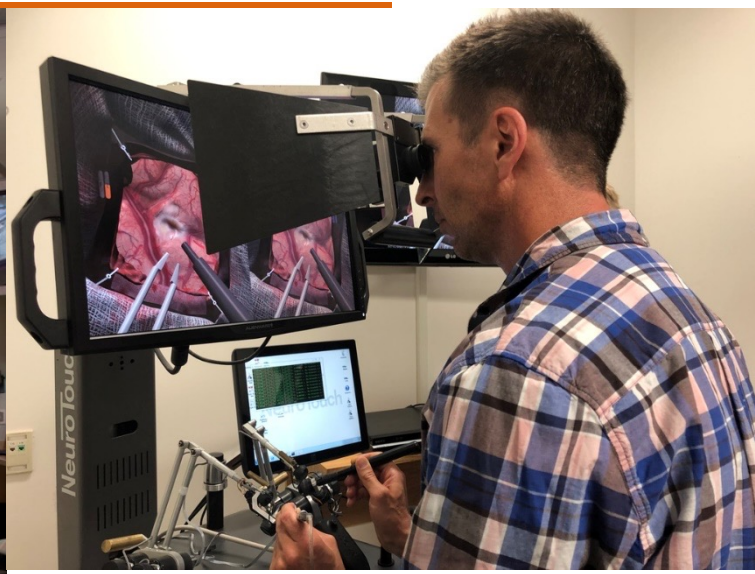
Sources of Digital Health Data: Laboratories as Data Generators

https://medizingeschichte.charite.de/forschung/arzneimittelforschung_in_der_ddr/zeitzeugeninterviews/
https://www.thumbay.com/news/thumbay-laboratories-and-medical-research-facility-expand-their-services/

Sources of Digital Health Data: Surgery Room 1.0: No Data

https://www.mwv-berlin.de/meldung/!/id/59

Sources of Digital Health Data: Surgery Room 2.0: Built upon Data



M. Schapranow / HPI

http://innovativevirtualconsulting.com/digital-or-planning/

https://vimeo.com/143245835

Bear in mind: FAIR principles

- Builds on a 2016 initiative
- Agreement on how to improve use of (research) data
- Current focus on findability an accessibility, e.g. data obtained in research consortia

FAIR Principles

Compliance

Findability

Resource and its metadata are easy to find by both, humans and computer systems. Basic machine readable descriptive metadata allows the discovery of interesting data sets and services.

- **F1**. Resource is uploaded to a public repository.
- **F2**. Metadata are assigned a globally unique and persistent identifier.

Accessibility

Resource and metadata are stored for the long term such that they can be easily accessed and downloaded or locally used by humans and ideally also machines using standard communication protocols.

- **A1**. Resource is accessible for download or manipulation by humans and is ideally also machine readable.
- **A2**. Publications and data repositories have contingency plans to assure that metadata remain accessible, even when the resource or the repository are no longer available.

Interoperability

Metadata should be ready to be exchanged, interpreted and combined in a (semi)automated way with other data sets by humans as well as computer systems.

- **I1**. Resource is uploaded to a repository that is interoperable with other platforms.
- **I2**. Repository meta- data schema maps to or implements the CG Core metadata schema.
- **I3**. Metadata use standard vocabularies and/or ontologies.

Reusability

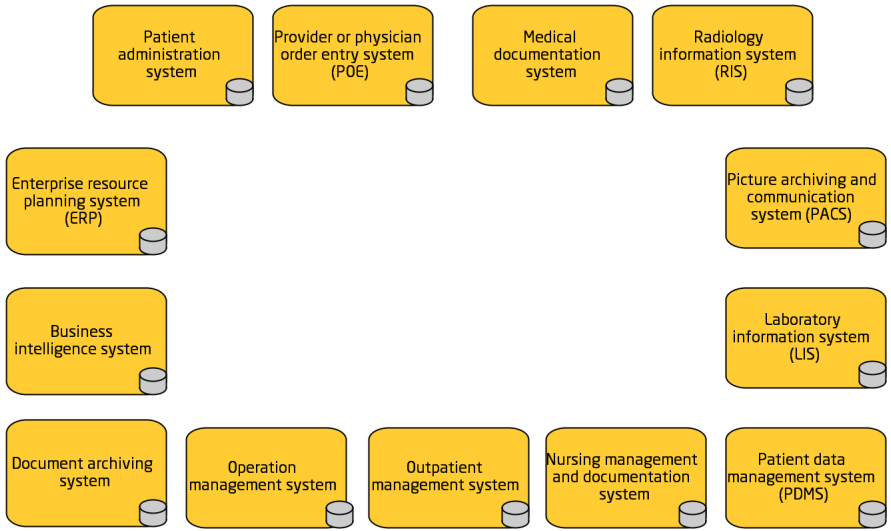
Data and metadata are sufficiently well-described to allow data to be reused in future research, allowing for integration with other compatible data sources. Proper citation must be facilitated, and the conditions under which the data can be used should be clear to machines and humans.

- **R1.** Metadata are released with a clear and accessible usage license.
- **R2.** Metadata about data and datasets are richly described with a plurality of accurate and relevant attributes.

http://www.go-fair.org/fair-principles/https://ccafs.cgiar.org/open-access-and-fair-principles#.X7JTYi1Q1TZ

Sources of Digital Health Data: Hospital Information Systems

- Historic, decentralized development
- Challenge: specific systems per data source and use case
- Data exchange: unidirectional or via message broker



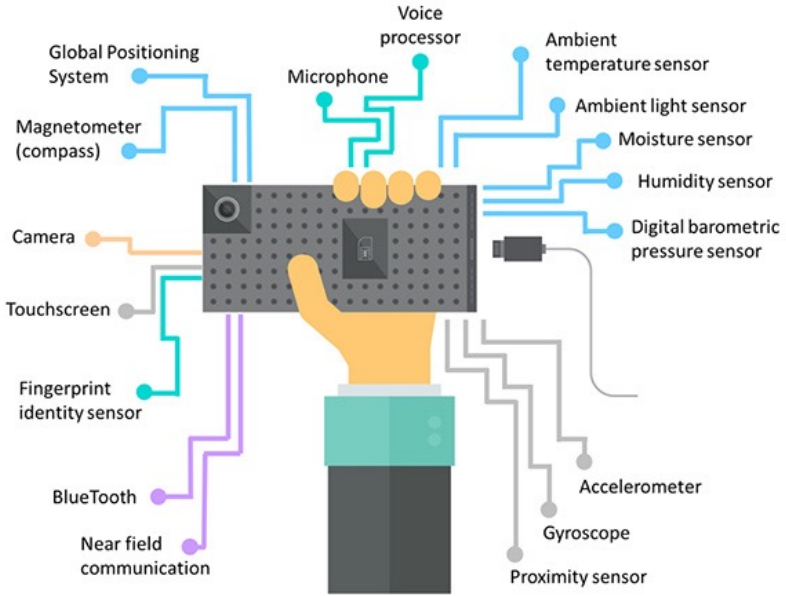
A. Winter, R. Haux, E. Ammenwerth, B. Brigl, N. Hellrung, F. Jahn. Health Information Systems: Architectures and Strategies (2010)

Sources of Digital Health Data: Patients: Self-reported Outcomes

- New dimension: Patient-Reported Outcomes (PROs)
- Pros:
  - Enables continues monitoring of high-risk patients
  - Documentation of personal events, feelings, situations
  - Patient diary to recall special situations
  - Integrates personal feedback of individual patients
- Cons:
  - Quality and source of data
  - Need to react
  - How to ensure reaction in time?

http://www.rethinkingclinicaltrials.org/cores-and-working-groups/patient-reported-outcomes-2/

Sources of Digital Health Data: Phone := Sensors + Computational Unit



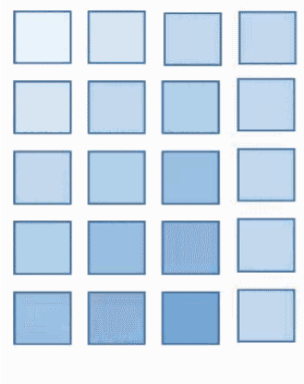
https://study-volunteers.iconplc.com/icon-views/blog/2017/06/16/leveraging-smartphone-potential/

- Accelerometer / motion sensor + three-axis gyroscope
- GPS + compass
- Proximity + ambient light
- Barometer + ambient temperature
- Touch sensor + Touch ID
- Cameras + Face ID
- Microphone
- Radio / WiFi + NFC
- Moisture (analogue inside device)
- **+ On-site computing infrastructure**

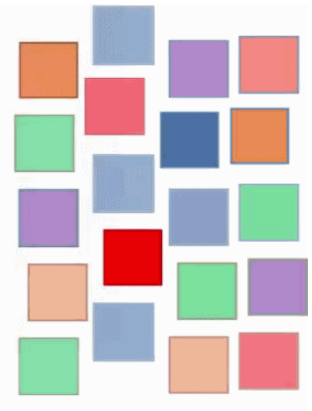
Interactive Sessions

- Task: Add examples for digital health data of the corresponding category

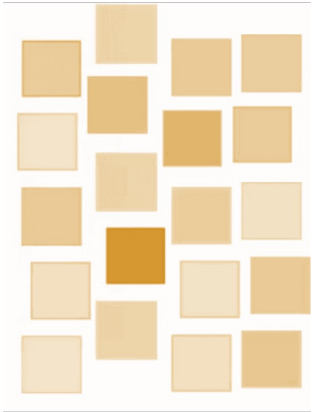
Structured



Unstructured



Semi-structured



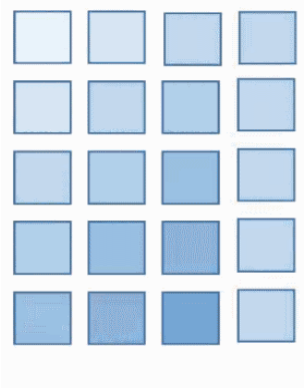
Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

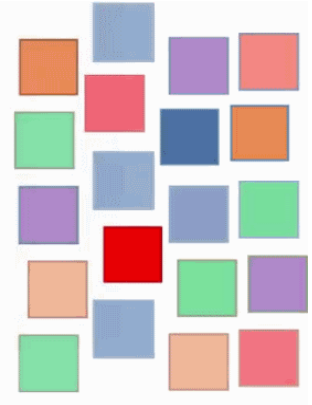
Categories of Digital Health Data

- **Structured:** well-organized, stored in relational databases, allows effective analysis
- **Unstructured:** no common data model, not directly machine-readable
- **Semi-structured:** unstructured, some tools may support analysis

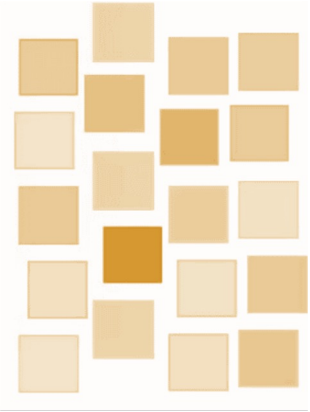
Structured



Unstructured

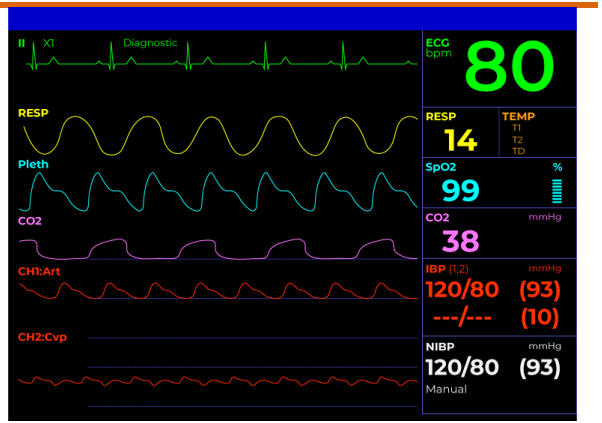


Semi-structured

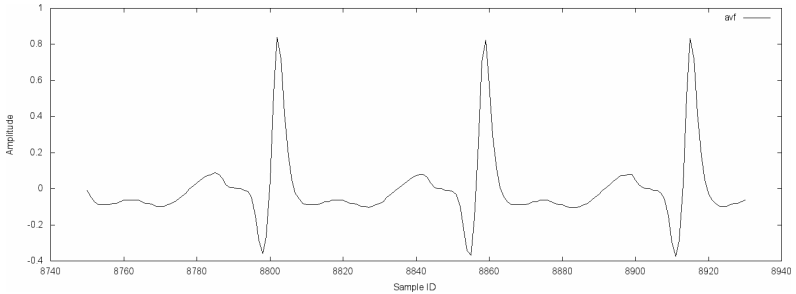


Categories of Digital Health Data: Structured Data

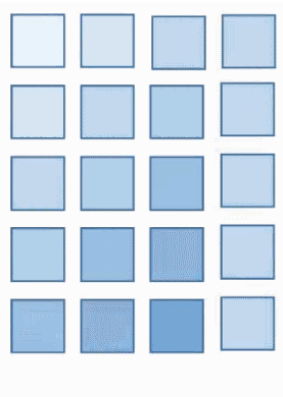
- Standardized, consistent input
- Machine processable
- Examples of structured data:
  - Data points
  - Longitudinal data
  - Laboratory values
  - Sensor data
  - Genomic data



https://www.cardiacdirect.com/how-to-read-a-patient-monitor/



Structured



Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

Categories of Digital Health Data: Sensor Data

- Data acquired by medical equipment in equidistant time
- Examples
  - Patient bedside monitoring
  - Electrocardiogram (ECG) monitors, pulse oximetry, blood pressure
  - Wearables, e.g. blood pressure, accelerator

Structured

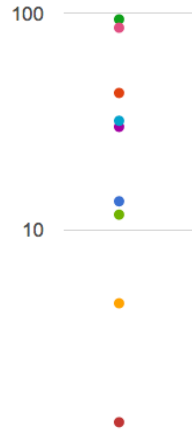
SantaMedical's Finger Pulse Oximeter

Heal Force Portable ECG monitor

https://www.cnet.com/health/apple-watch-ecg-app-what-cardiologists-want-you-to-know/

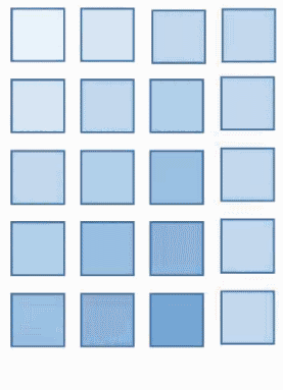
Categories of Digital Health Data: Data Points

- **Data points** := Acquired once or multiple times in (non-)equidistant times
- Provides a single point in time impression
- Examples: Lab results



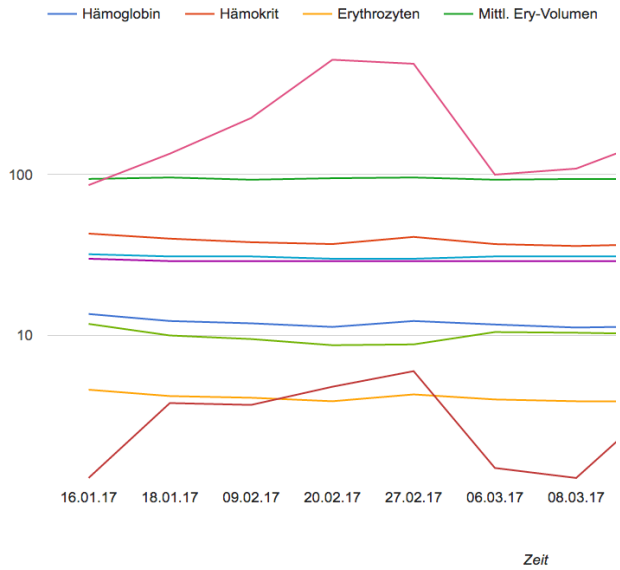
- Pro: Can provide just-in-time insights
- Con: Does not provide holistic view

Structured

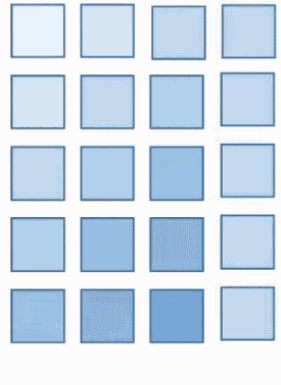


Categories of Digital Health Data: Longitudinal Data

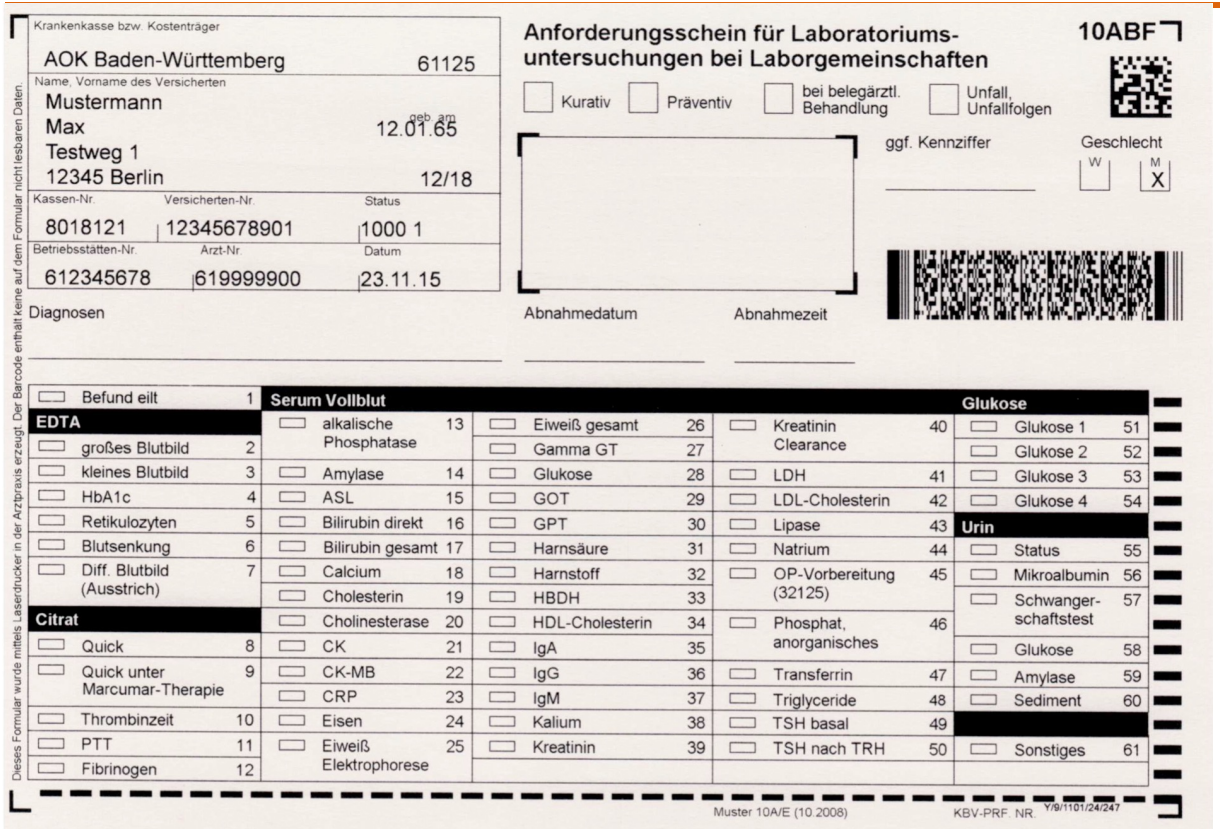
- **Longitudinal data** := Multiple measurements over (equidistant) time spans
- Examples:
  - Lab values
  - Clinical studies
  - Observational studies
- Pro: Can provide a more holistic view on changes of data over time
- Con: More time and effort required



Structured



Categories of Digital Health Data: Order for Laboratory Tests (German Muster 10A)



Krankenkasse bzw. Kostenträger
AOK Baden-Württemberg 61125
Name, Vorname des Versicherten
Mustermann
Max geb. am 12.01.65
Testweg 1
12345 Berlin 12/18
Kassen-Nr. 8018121 | Versicherten-Nr. 12345678901 | Status 1000 1
Betriebsstätten-Nr. 612345678 | Arzt-Nr. 619999900 | Datum 23.11.15
Diagnosen

Anforderungsschein für Laboratoriums-untersuchungen bei Laborgemeinschaften

10ABF

☐ Kurativ ☐ Präventiv ☐ bei belegärztl. Behandlung ☐ Unfall, Unfallfolgen

ggf. Kennziffer

Geschlecht W ☐ M X

Abnahmedatum

Abnahmezeit

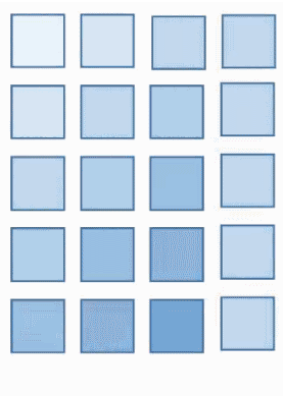
Dieses Formular wurde mittels Laserdrucker in der Arztpraxis erzeugt. Der Barcode enthält keine auf dem Formular nicht lesbaren Daten.

☐ Befund eilt	1	**Serum Vollblut**								**Glukose**	
EDTA		☐ alkalische Phosphatase	13	☐ Eiweiß gesamt	26	☐ Kreatinin Clearance	40	☐ Glukose 1	51		
☐ großes Blutbild	2			☐ Gamma GT	27			☐ Glukose 2	52		
☐ kleines Blutbild	3	☐ Amylase	14	☐ Glukose	28	☐ LDH	41	☐ Glukose 3	53		
☐ HbA1c	4	☐ ASL	15	☐ GOT	29	☐ LDL-Cholesterin	42	☐ Glukose 4	54		
☐ Retikulozyten	5	☐ Bilirubin direkt	16	☐ GPT	30	☐ Lipase	43	**Urin**			
☐ Blutsenkung	6	☐ Bilirubin gesamt	17	☐ Harnsäure	31	☐ Natrium	44	☐ Status	55		
☐ Diff. Blutbild (Ausstrich)	7	☐ Calcium	18	☐ Harnstoff	32	☐ OP-Vorbereitung (32125)	45	☐ Mikroalbumin	56		
		☐ Cholesterin	19	☐ HBDH	33			☐ Schwanger-schaftstest	57		
Citrat		☐ Cholinesterase	20	☐ HDL-Cholesterin	34	☐ Phosphat, anorganisches	46				
☐ Quick	8	☐ CK	21	☐ IgA	35			☐ Glukose	58		
☐ Quick unter Marcumar-Therapie	9	☐ CK-MB	22	☐ IgG	36	☐ Transferrin	47	☐ Amylase	59		
		☐ CRP	23	☐ IgM	37	☐ Triglyceride	48	☐ Sediment	60		
☐ Thrombinzeit	10	☐ Eisen	24	☐ Kalium	38	☐ TSH basal	49				
☐ PTT	11	☐ Eiweiß Elektrophorese	25	☐ Kreatinin	39	☐ TSH nach TRH	50	☐ Sonstiges	61		
☐ Fibrinogen	12										

Muster 10A/E (10.2008) KBV-PRF. NR. Y/9/1101/24/247

http://www.wenger.de/de/gesundheitswesen/blankoformularbedruckung/blankoformularbedruckung.php

Structured

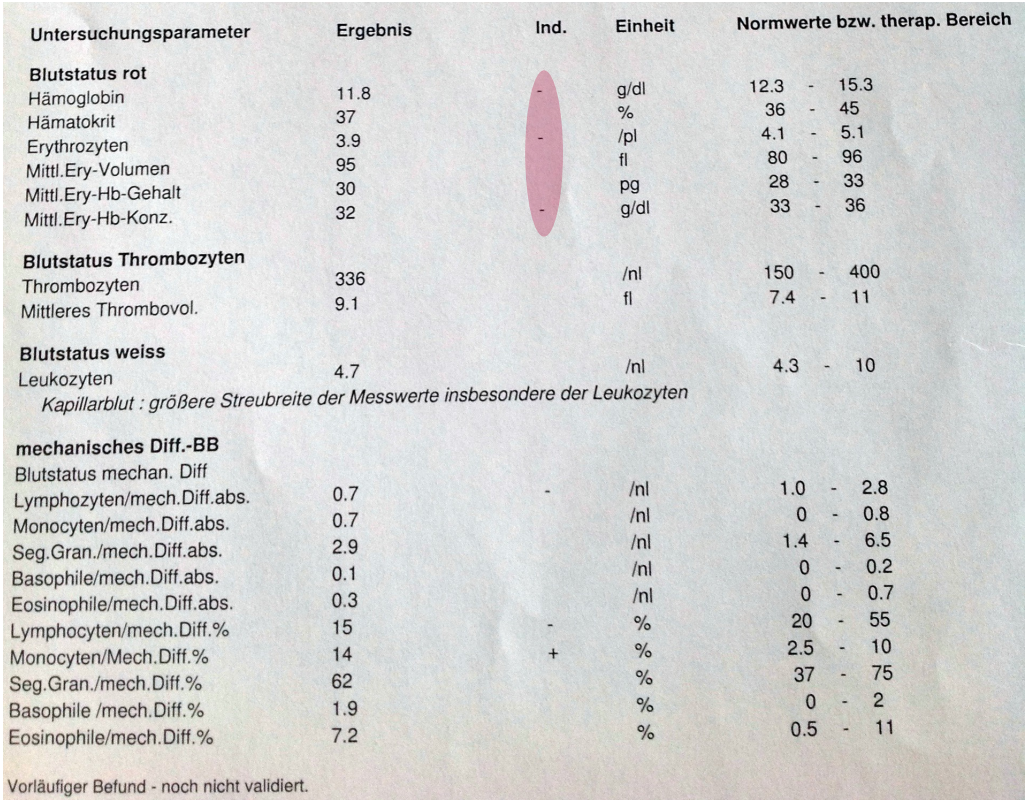


Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

Categories of Digital Health Data: Laboratory Results

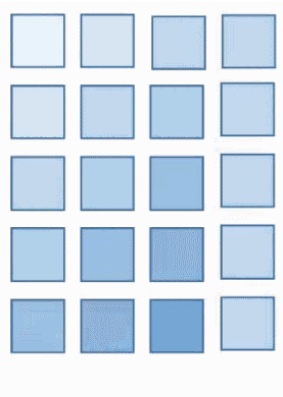
- List of medical attributes and their “normal” thresholds
- Exceeded values are highlighted
- Standardized encoding using Logical Observation Identifiers Names and Codes (LOINC)
- LOINC was initiated 1994 in the U.S.



Untersuchungsparameter	Ergebnis	Ind.	Einheit	Normwerte bzw. therap. Bereich
Blutstatus rot				
Hämoglobin	11.8	-	g/dl	12.3 - 15.3
Hämatokrit	37		%	36 - 45
Erythrozyten	3.9	-	/pl	4.1 - 5.1
Mittl.Ery-Volumen	95		fl	80 - 96
Mittl.Ery-Hb-Gehalt	30		pg	28 - 33
Mittl.Ery-Hb-Konz.	32	-	g/dl	33 - 36
Blutstatus Thrombozyten				
Thrombozyten	336		/nl	150 - 400
Mittleres Thrombovol.	9.1		fl	7.4 - 11
Blutstatus weiss				
Leukozyten	4.7		/nl	4.3 - 10
Kapillarblut : größere Streubreite der Messwerte insbesondere der Leukozyten				
mechanisches Diff.-BB				
Blutstatus mechan. Diff				
Lymphozyten/mech.Diff.abs.	0.7	-	/nl	1.0 - 2.8
Monocyten/mech.Diff.abs.	0.7		/nl	0 - 0.8
Seg.Gran./mech.Diff.abs.	2.9		/nl	1.4 - 6.5
Basophile/mech.Diff.abs.	0.1		/nl	0 - 0.2
Eosinophile/mech.Diff.abs.	0.3		/nl	0 - 0.7
Lymphocyten/mech.Diff.%	15	-	%	20 - 55
Monocyten/Mech.Diff.%	14	+	%	2.5 - 10
Seg.Gran./mech.Diff.%	62		%	37 - 75
Basophile /mech.Diff.%	1.9		%	0 - 2
Eosinophile/mech.Diff.%	7.2		%	0.5 - 11

Vorläufiger Befund - noch nicht validiert.

Structured



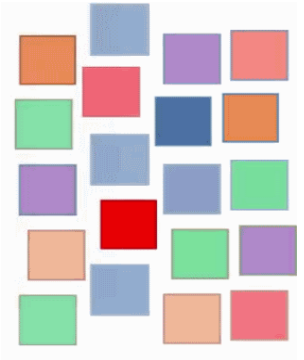
Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

Categories of Digital Health Data: Unstructured Data

- Approx. 50 documents per in-patient stay
- 60-80% is textual, generated from humans for humans
- Example from literature:
  - 350-bed institution with an avg. seven day Length of Stay (LOS)
  - Approx. 1,600 of text documents daily
  - Textual output comparable to a daily newspaper

Unstructured



Categories of Data in Digital Health

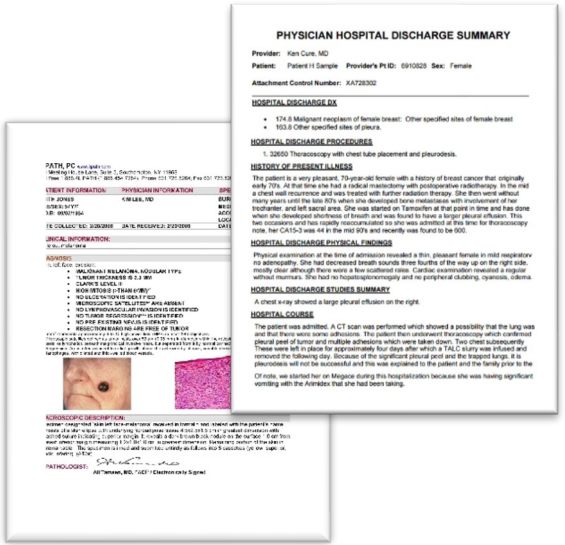
Data Management for Digital Health, Winter 2023

Schmücker P, 2012: Dokumentenaufkommen und eArchivierung in Krankenhäusern - Entwicklung und Stand heute.

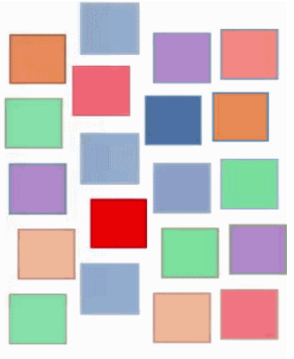
Categories of Digital Health Data: Text Documents

Hasso Plattner Institut

- Unstructured data, i.e. not directly machine-readable / -processable
- Examples:
  - Discharge letters
  - Doctor letters
  - Clinical reports, e.g. pathology or radiology
  - Medical literature



Unstructured



Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

Categories of Digital Health Data: Discharge Letters

Hasso Plattner Institut

PHYSICIAN HOSPITAL DISCHARGE SUMMARY

Provider: Ken Cure, MD

Patient: Patient H Sample **Provider's Pt ID:** 6910828 **Sex:** Female

Attachment Control Number: XA728302

HOSPITAL DISCHARGE DX

- 174.8 Malignant neoplasm of female breast: Other specified sites of female breast
- 163.8 Other specified sites of pleura.

HOSPITAL DISCHARGE PROCEDURES

1. 32650 Thoracoscopy with chest tube placement and pleurodesis.

HISTORY OF PRESENT ILLNESS

The patient is a very pleasant, 70-year-old female with a history of breast cancer that originally early 70's. At that time she had a radical mastectomy with postoperative radiotherapy. In the mid a chest wall recurrence and was treated with further radiation therapy. She then went without many years until the late 80's when she developed bone metastases with involvement of her trochanter, and left sacral area. She was started on Tamoxifen at that point in time and has done when she developed shortness of breath and was found to have a larger pleural effusion. This two occasions and has rapidly reaccumulated so she was admitted at this time for thoracoscopy note, her CA15-3 was 44 in the mid 90's and recently was found to be 600.

HOSPITAL DISCHARGE PHYSICAL FINDINGS

Physical examination at the time of admission revealed a thin, pleasant female in mild respiratory no adenopathy. She had decreased breath sounds three fourths of the way up on the right side. mostly clear although there were a few scattered rales. Cardiac examination revealed a regular without murmurs. She had no hepatosplenomegaly and no peripheral clubbing, cyanosis, edema.

HOSPITAL DISCHARGE STUDIES SUMMARY

A chest x-ray showed a large pleural effusion on the right.

HOSPITAL COURSE

The patient was admitted. A CT scan was performed which showed a possibility that the lung was and that there were some adhesions. The patient then underwent thoracoscopy which confirmed pleural peel of tumor and multiple adhesions which were taken down. Two chest subsequently These were left in place for approximately four days after which a TALC slurry was infused and removed the following day. Because of the significant pleural peel and the trapped lungs, it is pleurodesis will not be successful and this was explained to the patient and the family prior to the

Of note, we started her on Megace during this hospitalization because she was having significant vomiting with the Arimidex that she had been taking.

SAMPLE DISCHARGE SUMMARY

Primary Diagnosis: 40 week IUP with delivery of a liveborn infant
Secondary Diagnosis: Advanced Maternal Age; Prolonged second stage of labor with maternal exhaustion
Procedure Performed:

1. Spontaneous Vaginal Delivery with delivery of live male infant weighing 7# 5oz at 1542 on January 3, 2012 with APGARS of 8 at one minute and 9 at five minutes.
2. Placement of Intrauterine Pressure Catheter.

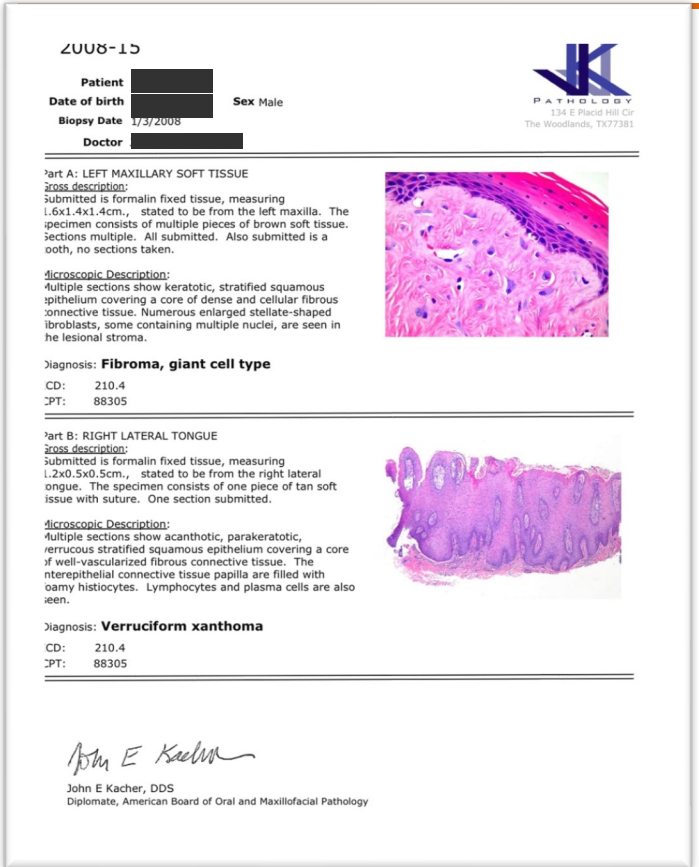
Reason for Hospitalization: This 36yo G2P1001 presented at 40 weeks gestation by an LMP of 3/12/11 with an EDC of 1/3/12 in spontaneous labor. This pregnancy has been complicated by advanced maternal age. QS performed at 17 weeks was within normal limits and a genetic amniocentesis was offered and declined. Prenatal laboratory data showed blood type B+ with a negative antibody screen, Rubella Immune, VDRL nonreactive, HepBsAg negative, Diabetic Screen 120, HIV nonreactive. She remained normotensive throughout her pregnancy. At the time of admission she reported positive fetal movement and denied loss of fluid.
Physical Exam on Admission: Temperature 98.4. Pulse 94. Respirations 16. Blood pressure 128/78. Fetal Heart Rate 150's and reactive. Uterine contractions q 4 minutes. HEENT within normal limits. Heart regular. Lungs clear. Abdomen gravid with a fundal height appropriate for gestational age. Extremities 2+ DTR's and trace edema. Cervical exam 4 cm/80%/-1.
Lab and X-Ray Data: Predelivery H&H of 12.4 and 36.2 respectively. Platelets 221.

Hospital Course: The patient was admitted in spontaneous labor in the morning of January 3rd. was reactive and reassuring throughout the course of her stay in labor and delivery. Her labor progressed well and at 0900 hours, she had spontaneous rupture of membranes with a return of fluid. At that time, her cervix was dilated to 6 cm/90%/0. Epidural anesthesia was requested and obtained. Her labor then quickly progressed and the patient was noted to be completely dilated at +1 station at 1100 hours. She was then allowed to push. After pushing for 2 hours, the patient brought the vertex to the perineum, but was unable to continue her expulsive efforts. The infant delivered by outlet forceps over a midline episiotomy. ***Please see operative report for full details.*** The patient and infant did well. She is breast-feeding the infant well, and has remained afebrile with minimal lochia since delivery. The patient was voiding and ambulating without difficulty by the evening of PPD #0. She declined any contraception at the time of discharge, and was deemed stable for discharge on PPD 2.

Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

Categories of Digital Health Data: Pathology Reports



2008-15

Patient [redacted]
Date of birth [redacted] Sex Male
Biopsy Date 1/3/2008
Doctor [redacted]

PATHOLOGY
134 E Placid Hill Cir
The Woodlands, TX77381

Part A: LEFT MAXILLARY SOFT TISSUE
Gross description:
Submitted is formalin fixed tissue, measuring 1.6x1.4x1.4cm., stated to be from the left maxilla. The specimen consists of multiple pieces of brown soft tissue. Sections multiple. All submitted. Also submitted is a tooth, no sections taken.

Microscopic Description:
Multiple sections show keratotic, stratified squamous epithelium covering a core of dense and cellular fibrous connective tissue. Numerous enlarged stellate-shaped fibroblasts, some containing multiple nuclei, are seen in the lesional stroma.

Diagnosis: **Fibroma, giant cell type**

ICD: 210.4
CPT: 88305

Part B: RIGHT LATERAL TONGUE
Gross description:
Submitted is formalin fixed tissue, measuring 1.2x0.5x0.5cm., stated to be from the right lateral tongue. The specimen consists of one piece of tan soft tissue with suture. One section submitted.

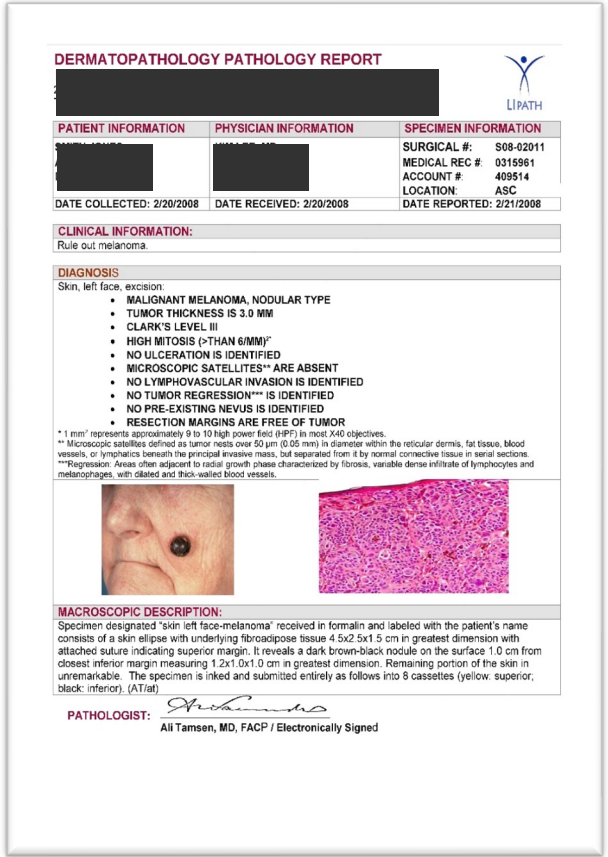
Microscopic Description:
Multiple sections show acanthotic, parakeratotic, verrucous stratified squamous epithelium covering a core of well-vascularized fibrous connective tissue. The interepithelial connective tissue papilla are filled with foamy histiocytes. Lymphocytes and plasma cells are also seen.

Diagnosis: **Verruciform xanthoma**

ICD: 210.4
CPT: 88305

John E Kacher

John E Kacher, DDS
Diplomate, American Board of Oral and Maxillofacial Pathology



DERMATOPATHOLOGY PATHOLOGY REPORT

[redacted]

LIPATH

PATIENT INFORMATION	PHYSICIAN INFORMATION	SPECIMEN INFORMATION
[redacted]	[redacted]	SURGICAL #: S08-02011 MEDICAL REC #: 0315961 ACCOUNT #: 409514 LOCATION: ASC
DATE COLLECTED: 2/20/2008	DATE RECEIVED: 2/20/2008	DATE REPORTED: 2/21/2008

CLINICAL INFORMATION:
Rule out melanoma.

DIAGNOSIS
Skin, left face, excision:
- MALIGNANT MELANOMA, NODULAR TYPE
- TUMOR THICKNESS IS 3.0 MM
- CLARK'S LEVEL III
- HIGH MITOSIS (>THAN 6/MM)*
- NO ULCERATION IS IDENTIFIED
- MICROSCOPIC SATELLITES** ARE ABSENT
- NO LYMPHOVASCULAR INVASION IS IDENTIFIED
- NO TUMOR REGRESSION*** IS IDENTIFIED
- NO PRE-EXISTING NEVUS IS IDENTIFIED
- RESECTION MARGINS ARE FREE OF TUMOR

* 1 mm² represents approximately 9 to 10 high power field (HPF) in most X40 objectives.
** Microscopic satellites defined as tumor nests over 50 µm (0.05 mm) in diameter within the reticular dermis, fat tissue, blood vessels, or lymphatics beneath the principal invasive mass, but separated from it by normal connective tissue in serial sections.
***Regression: Areas often adjacent to radial growth phase characterized by fibrosis, variable dense infiltrate of lymphocytes and melanophages, with dilated and thick-walled blood vessels.

MACROSCOPIC DESCRIPTION:
Specimen designated "skin left face-melanoma" received in formalin and labeled with the patient's name consists of a skin ellipse with underlying fibroadipose tissue 4.5x2.5x1.5 cm in greatest dimension with attached suture indicating superior margin. It reveals a dark brown-black nodule on the surface 1.0 cm from closest inferior margin measuring 1.2x1.0x1.0 cm in greatest dimension. Remaining portion of the skin in unremarkable. The specimen is inked and submitted entirely as follows into 8 cassettes (yellow: superior; black: inferior). (AT/at)

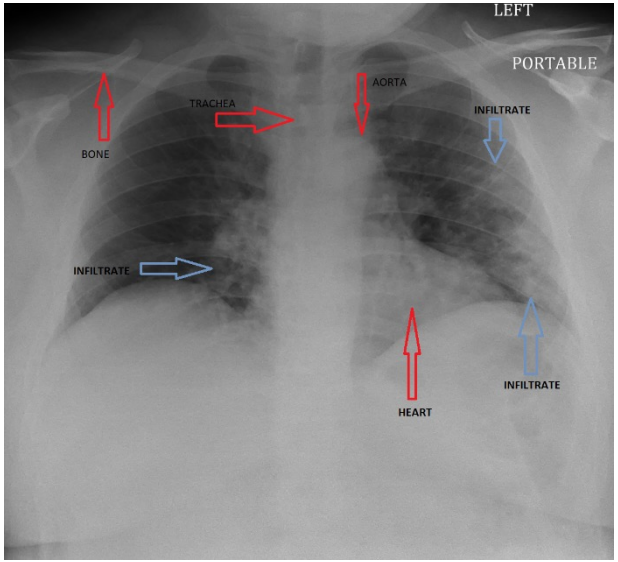
PATHOLOGIST:
Ali Tamsen, MD, FACP / Electronically Signed

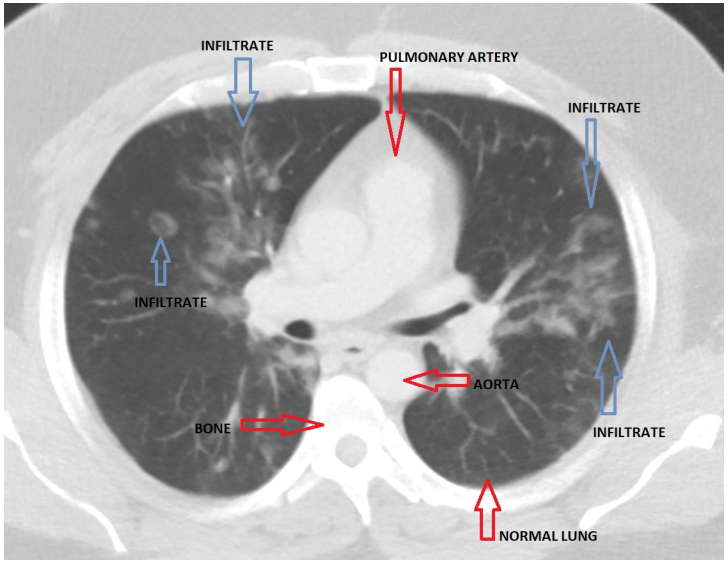
Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

Categories of Digital Health Data: Imaging Data

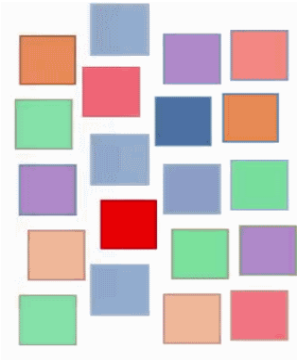
- Creates huge amounts of unstructured data in short period of time
- Equipped with structured meta data
- Requires post-processing and interpretation





https://imgur.com/gallery/PpyzwvH#oaLhIhl

Unstructured

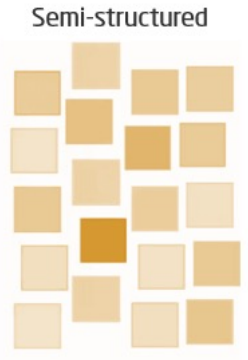


Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

Categories of Digital Health Data: Semi-structured

- Mixture of unstructured data, but partially structured
- Examples:
  - Structured form, but with free-text fields
  - Documents with a fixed outline, but free-text paragraphs
  - Imaging data combined with detailed meta data, e.g. DICOM



Categories of Digital Health Data: Paper-based Prescriptions

Krankenkasse bzw. Kostenträger
AOK Rheinland-Pfalz

Name, Vorname des Versicherten
Mustermann
Erika
Heidestraße 17
51147 Köln

geb. am
12.08.1964
10/14

Kassen-Nr. 106415300
Versicherten-Nr. A123456789
Status 1000 1
Betriebsstätten-Nr. 271111100
Arzt-Nr. 654321161
Datum 10.07.2012

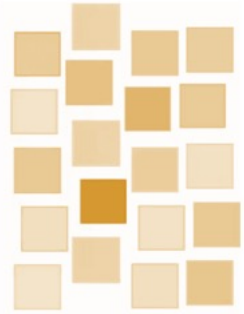
Rp. (Bitte Leerräume durchstreichen)
Antistressin Impfstoff Amp. 10 x 0.5 ml
Muster Pharma GmbH

27/1111100
Psychologische Gemeinschaftspraxis
Dr. med. Markus Mustermann
Dr. rer. nat. Erik Mustermann
Dorfheidestraße 1
51069 Köln

Bei Arbeitsunfall auszufüllen
Abgabedatum in der Apotheke
Unfalltag
Unfallbetrieb oder Arbeitgebernummer

2711111004

Semi-structured

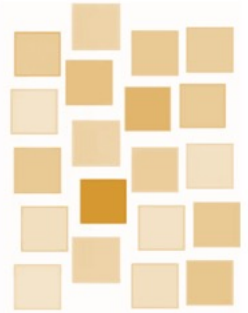


Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

Categories of Digital Health Data: Encyclopedias, Forums, Social Media

Semi-structured



Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

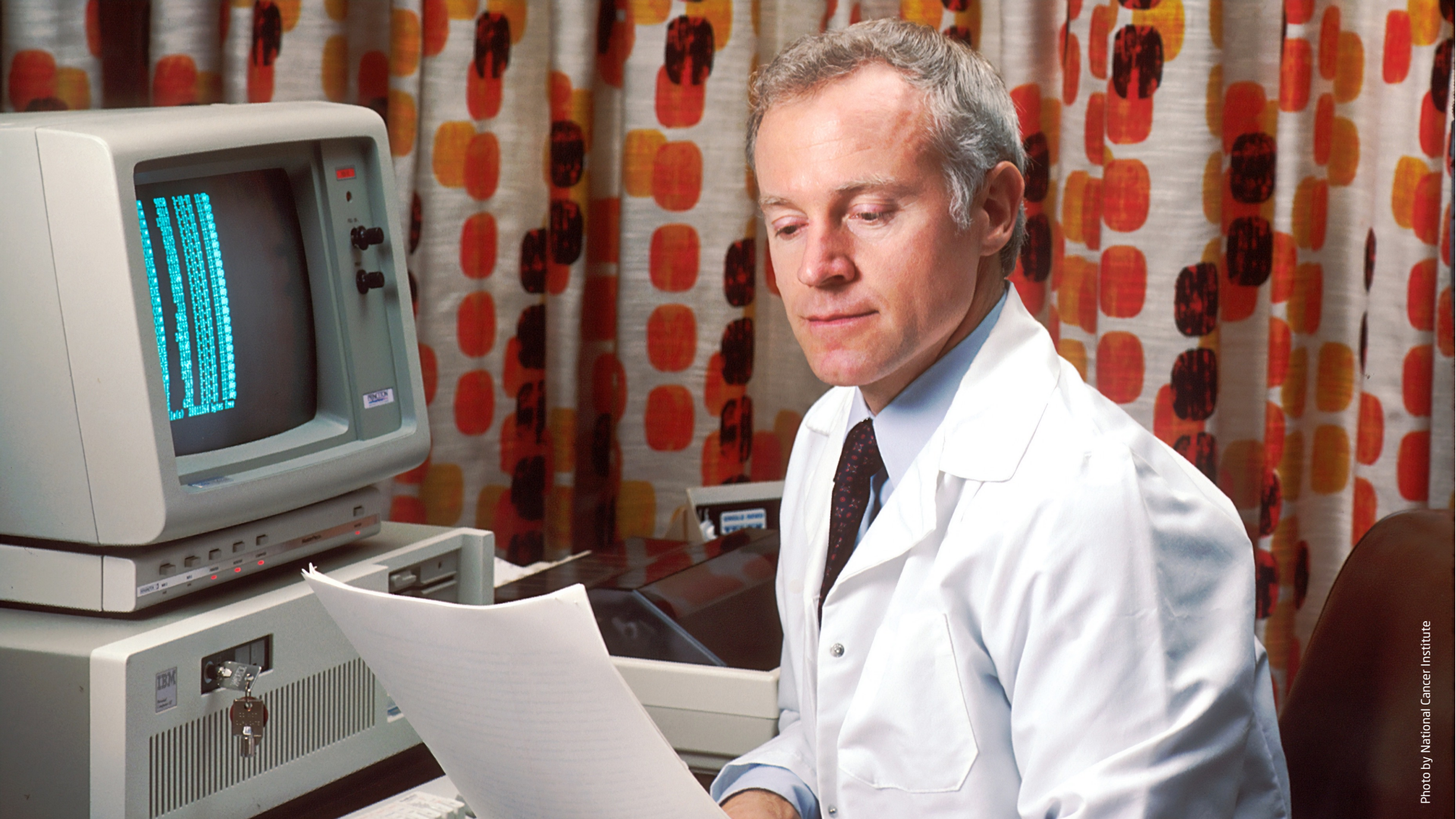
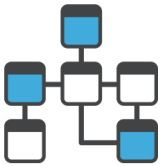


Photo by National Cancer Institute

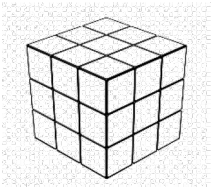
Data Management
Transactional vs. Analytical Workload

- **Data Management** := Methods, concepts, organizational and technical procedures to handle data in such a way that it supports daily business processes

Traditionally individual systems for



- **Online Transaction Processing (OLTP)**
  - Regular business operations, requires interactive user behavior
  - Workload: write and read
- **Online Analytical Processing (OLAP)**
  - Dedicated data models for real-time data analysis to support business decisions
  - Workload: mainly read

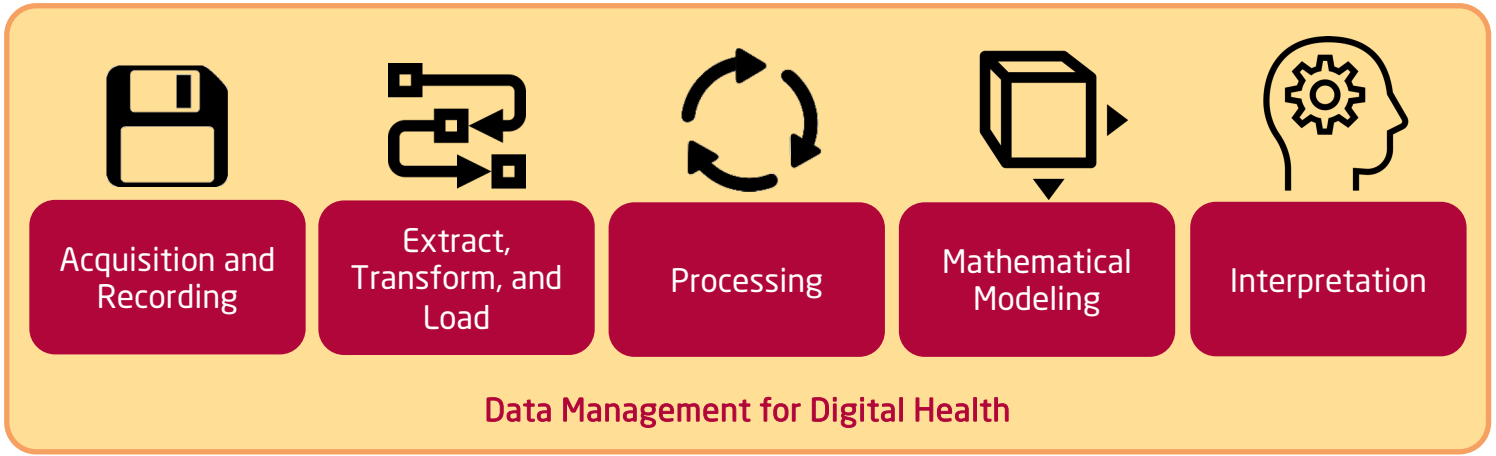


Data Analytics and Data Mining

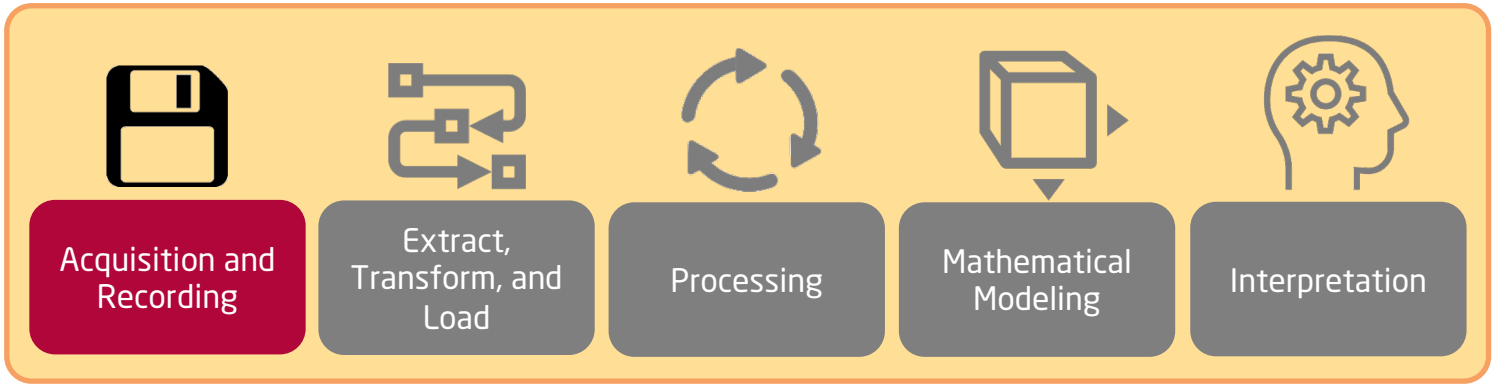
- **Data Analytics:** "…techniques used to analyze and acquire intelligence from big data [1]", e.g.
  - Use of MS Excel to perform descriptive analytics
  - Use Business Intelligence tools to generate graphical reports
- **Data Mining:**
  - "…discovery of interesting, unexpected or valuable structures in large datasets [2]"
  - "…turns a large collection of data into knowledge [3]"
  - Explorative analysis of available data sets, e.g.
    - Apply deep learning to discover hidden patterns in clinical notes
    - Use of support vector machines to forecast clinical outcomes

[1] Gandomi, A., & Haider, M. "Beyond the hype: Big Data Concepts, Methods, and Analytics. International Journal of Information Management, (2015).
[2] Hand, D. J. "Principles of Data Mining. Drug Safety", 30(7), (2007), pp. 621-622.
[3] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

Data Management
Process Perspective



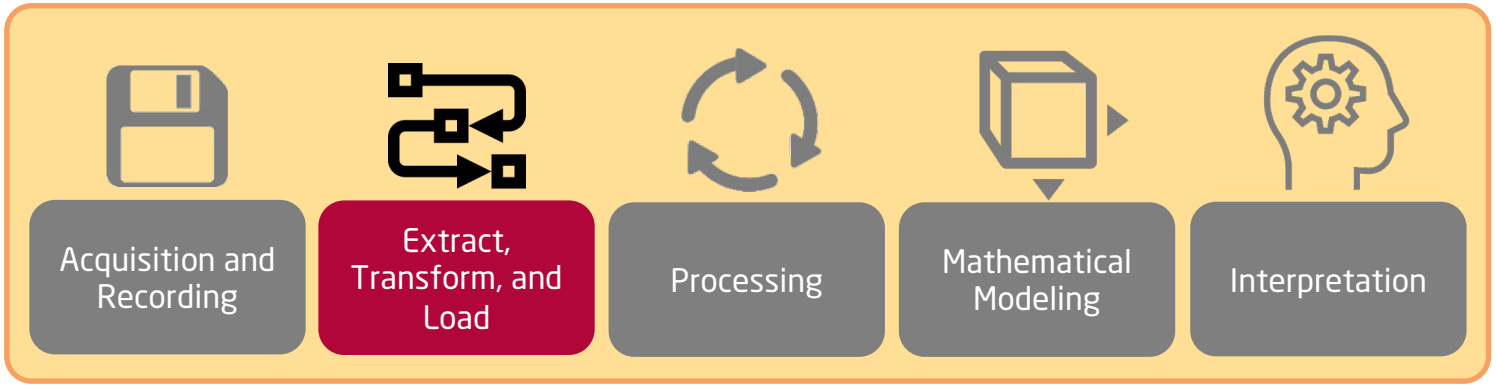
Data Management
Acquisition



- Identify necessary data sources
- Set-up measurement devices

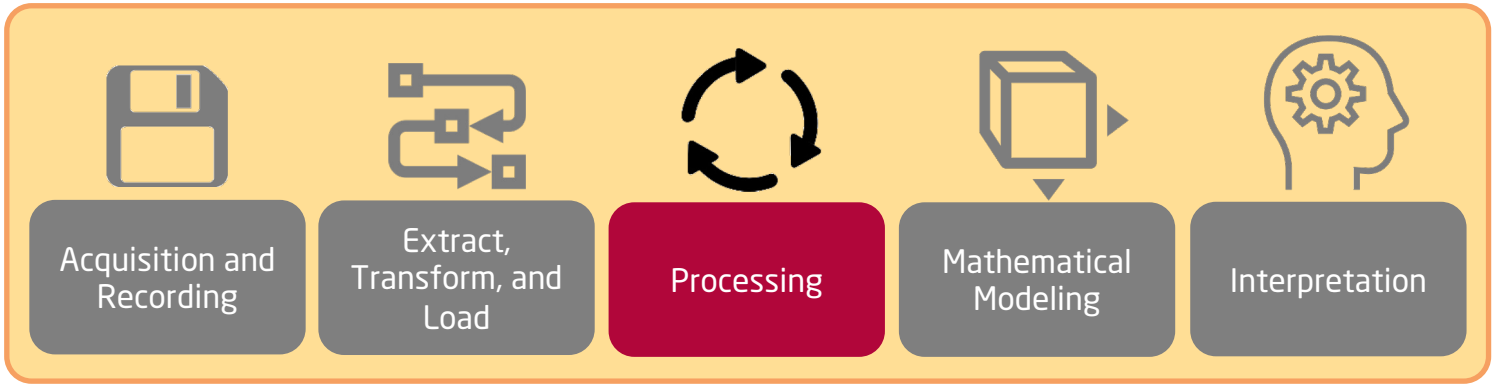
- Examples: Select data sources, setup sensors/wearables, record data, etc.

Data Management
Extract, Transform, Load



- Extract data from identify data sources
- Transform data format if required
- Load data into data processing system
- Examples: Extract from clinical system, harmonize formats, insert into research db

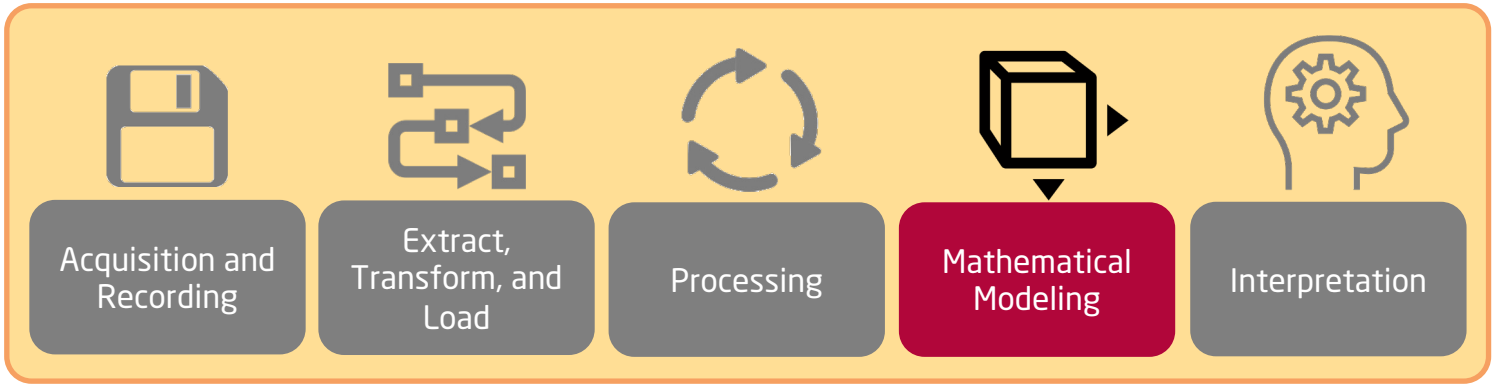
Data Management
Processing



- Perform data (pre-)processing prior to modeling

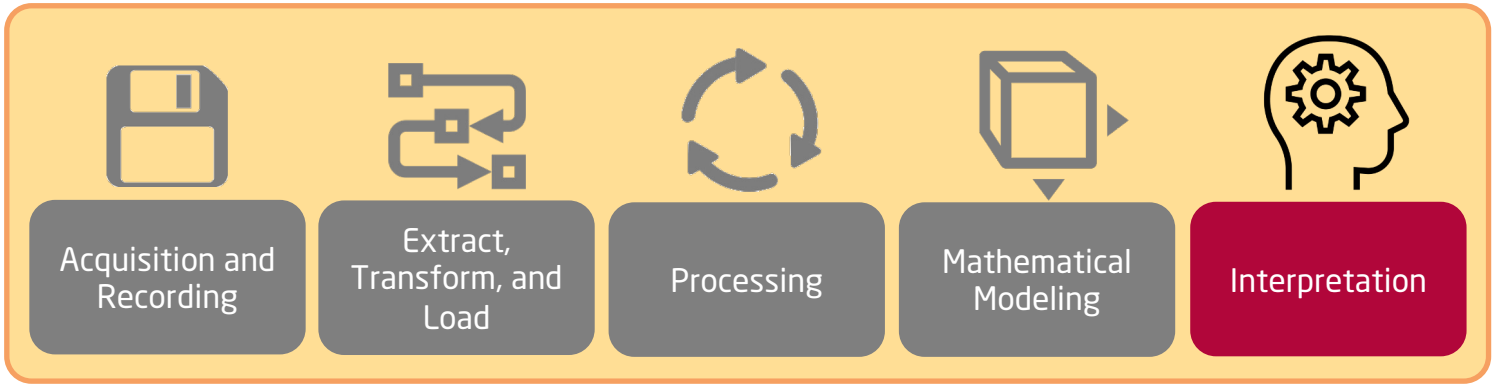
- Examples: Variant calling on genome data, extract features from a biological signal

Data Management Modeling



- Employ mathematical tools to make predictions and detect patterns
- Examples: Predict patient outcomes on historic data, detecting anomalies in images

Data Management
Interpretation



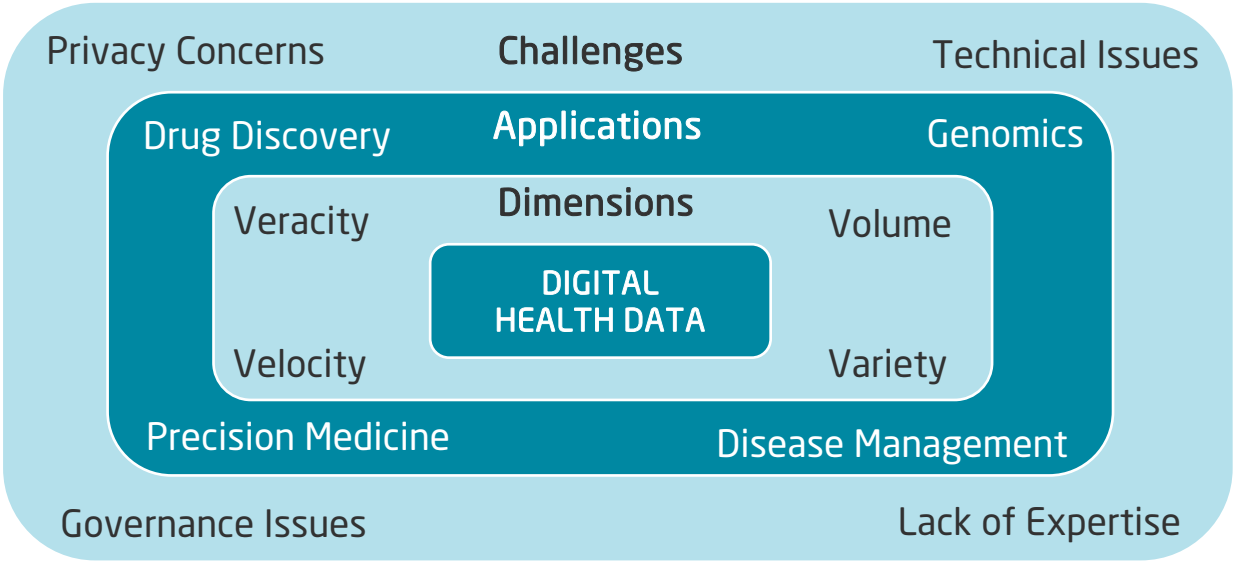
- Make sense of processing results, interpret results in the medical context

- Examples: Interpret results with subject-matter expert, try to identify causalities

Challenges?

Photo by Jukan Tateisi

Challenges of Digital Health Data

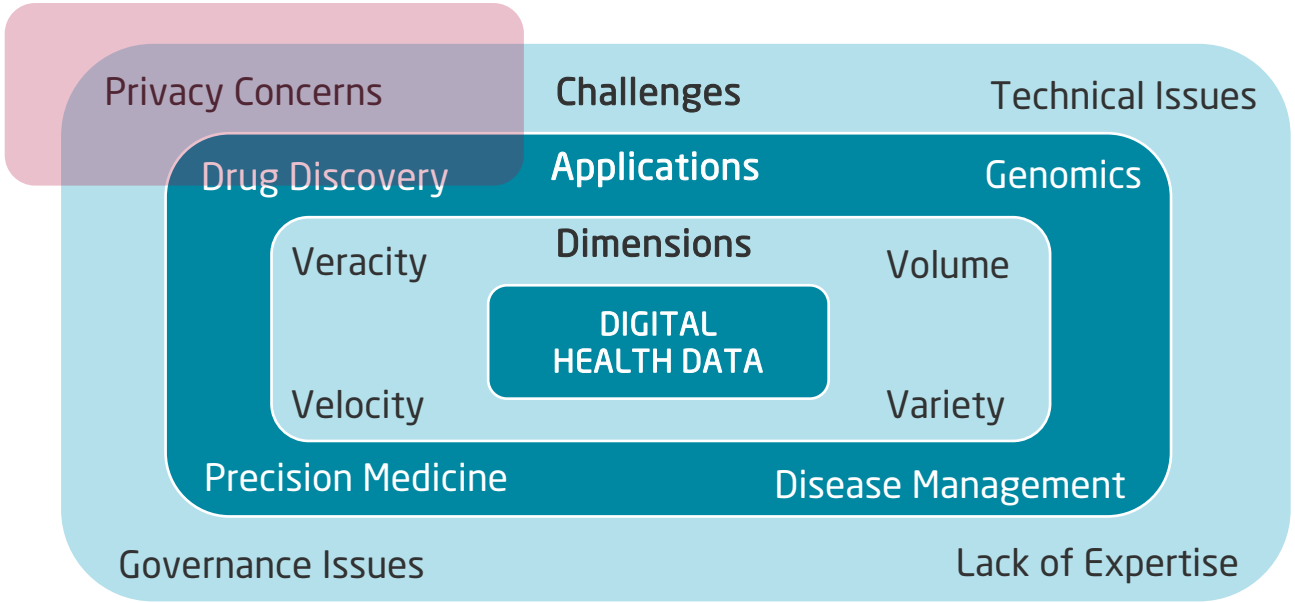


Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

Mehta, Nishita, and Anil Pandit. "Concurrence of big data analytics and healthcare: A systematic review." International journal of medical informatics 114 (2018): 57-65.

Challenges of Digital Health Data
Privacy Concerns



Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

Mehta, Nishita, and Anil Pandit. "Concurrence of big data analytics and healthcare: A systematic review." International journal of medical informatics 114 (2018): 57-65.

Hasso Plattner Institut

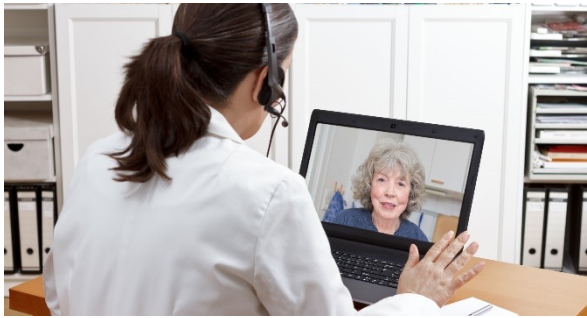
Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

Challenges of Digital Health Data Privacy Concerns

Hasso Plattner Institut

- Data privacy and trust are the foundation for patient-doctor relationships
- Disclosure of sensitive data may lead to social stigma and discrimination
- Data used for research requires explicit patient consent



https://www.aerzteblatt.de/nachrichten/110997/Telemedizin-Kraeftiger-Schub-fuer-Videosprechstunden

- Legal regulations:
  - General Data Protection Regulation (GDPR)
  - Health Insurance Portability and Accountability Act (HIPAA)
  - The Personal Information Protection and Electronic Documents Act (PIPEDA)

GDPR

HIPAA Health Insurance Portability & Accountability Act

PIPEDA

Appari, Ajit, and M. Eric Johnson. "Information security and privacy in healthcare: current state of research." International journal of Internet and enterprise management 6.4 (2010): 279-314.

EU General Data Protection Regulation Structure (effective since 2018)

Hasso Plattner Institut

GDPR

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

- Chap. I (Art. 01 – 04) General provisions
- Chap. II (Art. 05 – 11) Principles
- Chap. III (Art. 12 – 23) Rights of the data subject
- Chap. IV (Art. 24 – 43) Controller and processor
- Chap. V (Art. 44 – 50) Transfer of personal data

https://gdpr-info.eu/

EU General Data Protection Regulation Structure

- Chap. VI (Art. 51 – 59) Independent supervisory authorities
- Chap. VII (Art. 60 – 76) Cooperation and consistency
- Chap. VIII (Art. 77 – 84) Remedies, liability and penalties
- Chap. IX (Art. 85 – 91) Provisions relating to specific processing situations
- Chap. X (Art. 92 – 93) Delegated acts and implementing acts
- Chap. XI (Art. 94 – 99) Final provisions

EU General Data Protection Regulation Examples

- Art. 05: Processing of identifiable patient data
- Art. 30: Responsibilities
- Art. 32: Protection regulations
- Art. 33: Reporting of security leaks to official authorities (typically 72hrs)
- Art. 34: Reporting of security leaks to individuals
- Art. 35: Risk assessment

Data Protection Tips for your Digital Health projects

Hasso Plattner Institut

- Raise awareness for data protection measures
- Always assess the legal basis for accessing data
- Check specifically for personal data of children
- Implement privacy-by-design and privacy-by-default settings
- Evaluate existing contracts
- Implement and document data protection measures if not already present
- Organize reporting structures and test them
- Implement information duties and rights of data subjects
- Start and continue to document

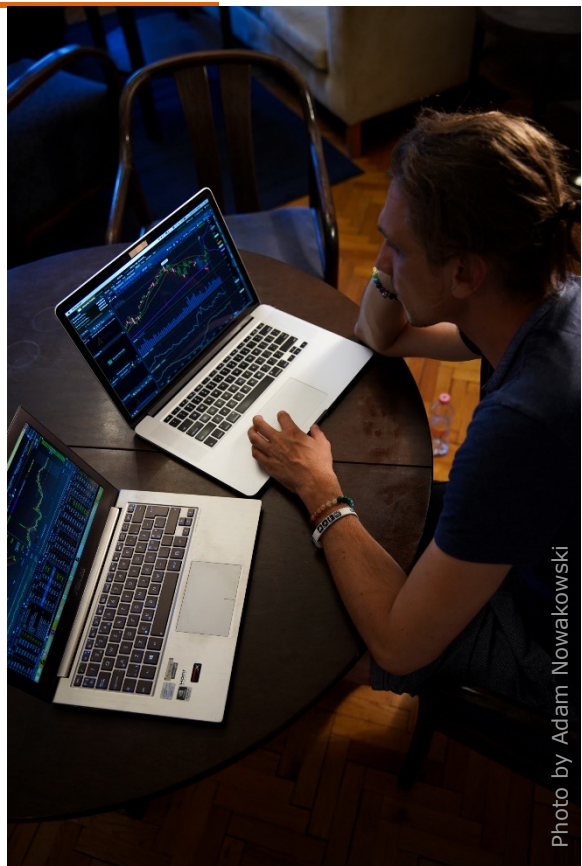
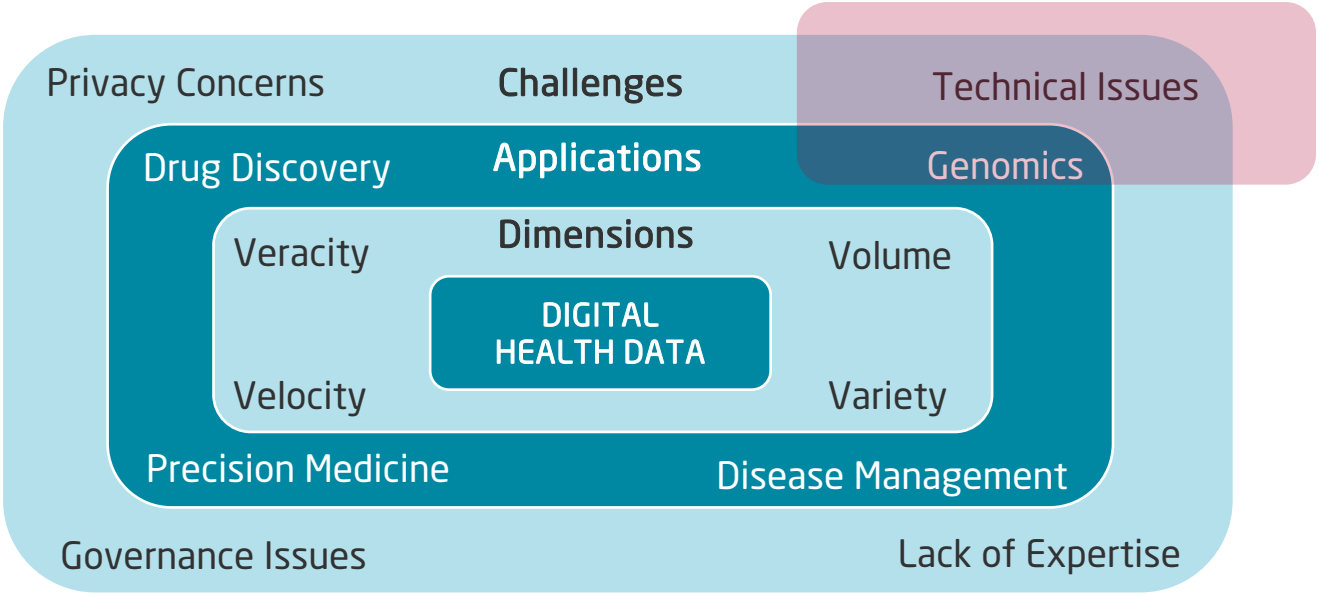


Photo by Adam Nowakowski

Challenges of Digital Health Data
Technical Issues



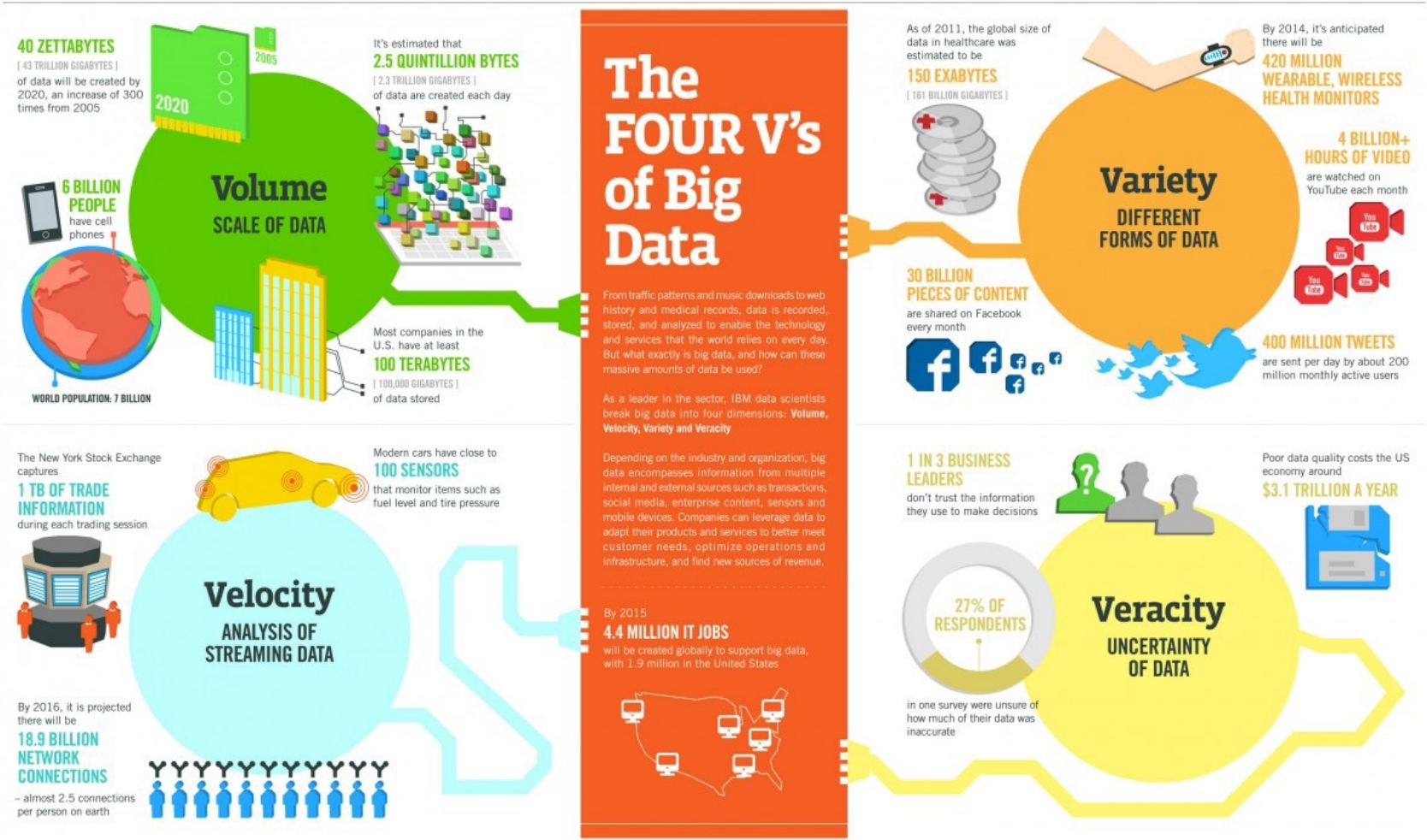
Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

Mehta, Nishita, and Anil Pandit. "Concurrence of big data analytics and healthcare: A systematic review." International journal of medical informatics 114 (2018): 57-65.

Challenges of Digital Health Data
Technical Issues

- Sheer size of data, e.g. *omics and sensor data
- Processing of large datasets requires high-performance computing
- Interoperability issues, e.g. technical and semantic
- Specific data schemas and required transformations
- Lack of complete data sets
- Reproducibility



asso
lattner
nstitut

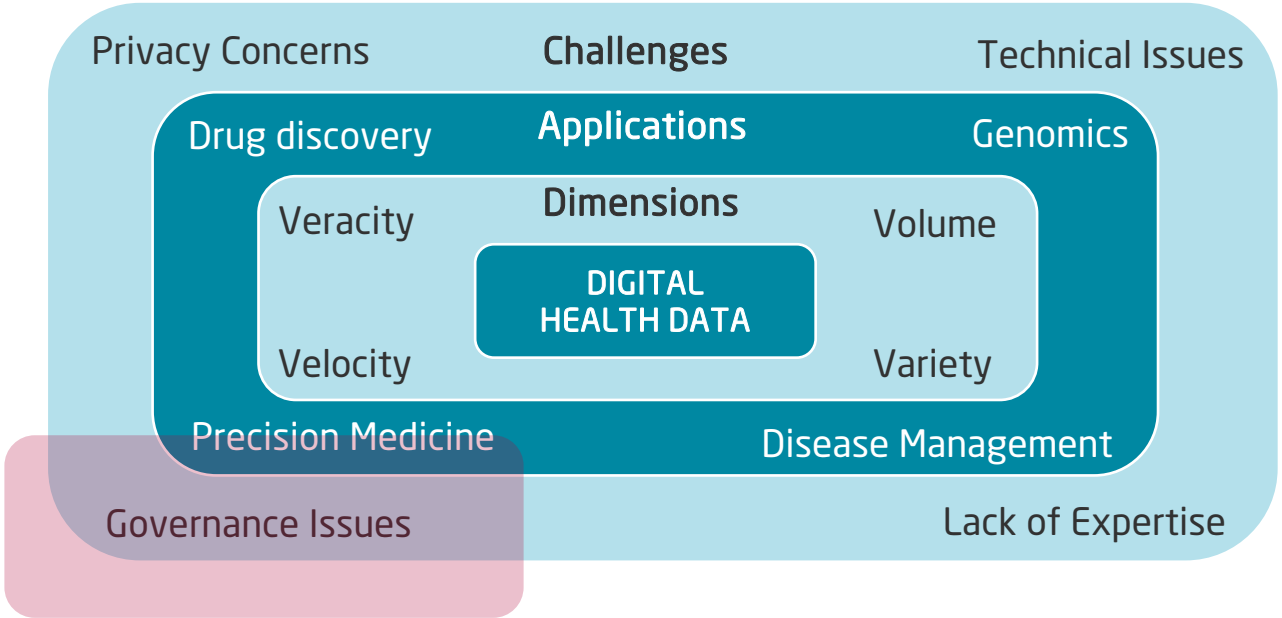
Data in

ent for
Winter

[1] The
[2] Baro

https://www.ibmbigdatahub.com/infographic/four-vs-big-data

Challenges of Digital Health Data Governance Issues



Categories of Data in Digital Health

Data Management for Digital Health, Winter 2023

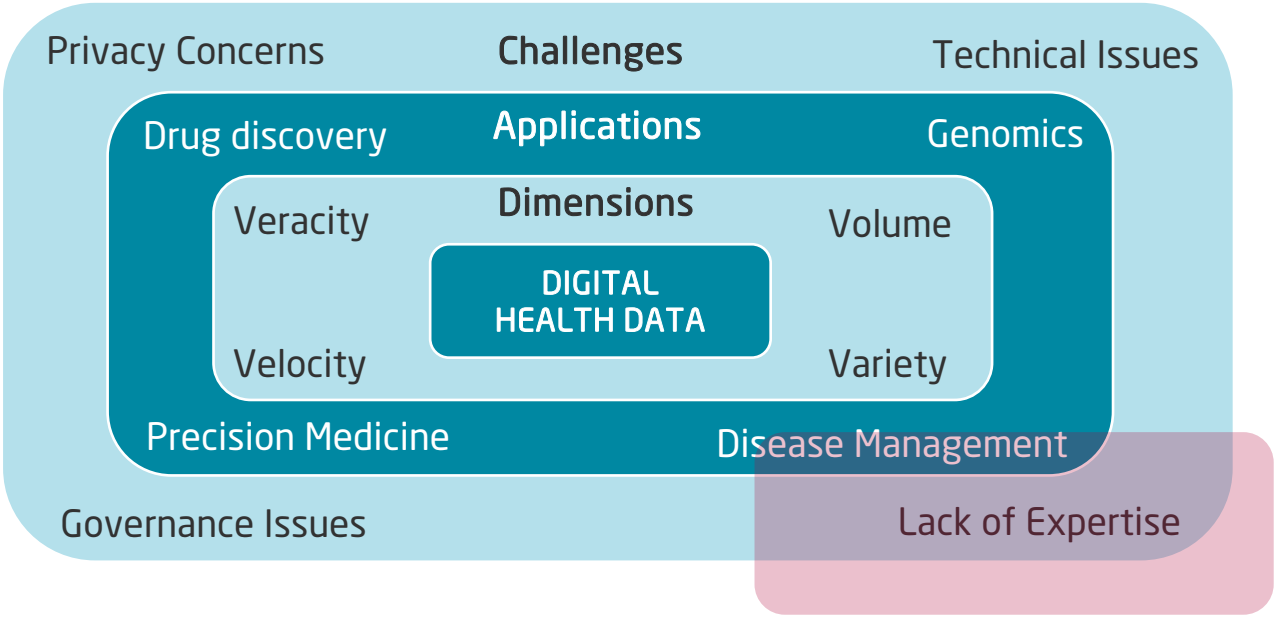
Mehta, Nishita, and Anil Pandit. "Concurrence of big data analytics and healthcare: A systematic review." International journal of medical informatics 114 (2018): 57-65.

Challenges of Digital Health Data Governance Issues

- How to find relevant data? → Meta data catalogues (FAIR)
- Responsibility, e.g. data stored in inter- or intra-institutional silos (FAIR)
- Integration between health care actors / adoption of data standards and ontologies in practice → National data governance initiatives (FAIR)
- Healthcare data is mainly of administrative nature, but retrospective analysis with existing data can support research → Re-use (FAIR)

Categories of Data in Digital Health

Challenges of Digital Health Data
Lack of Expertise



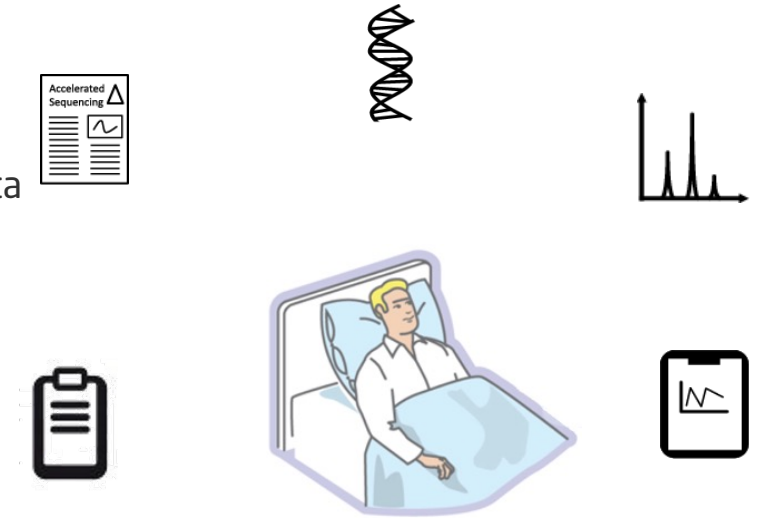
Categories of Data in Digital Health

Mehta, Nishita, and Anil Pandit. "Concurrence of big data analytics and healthcare: A systematic review." International journal of medical informatics 114 (2018): 57-65.

Consider it done!

What to Take Home?

- Data, information, knowledge
- Data characteristics
- Structured, semi-structured, unstructured data
- Challenges in data handling
- Privacy and sensitivity
- Big data



Categories of Data in Digital Health