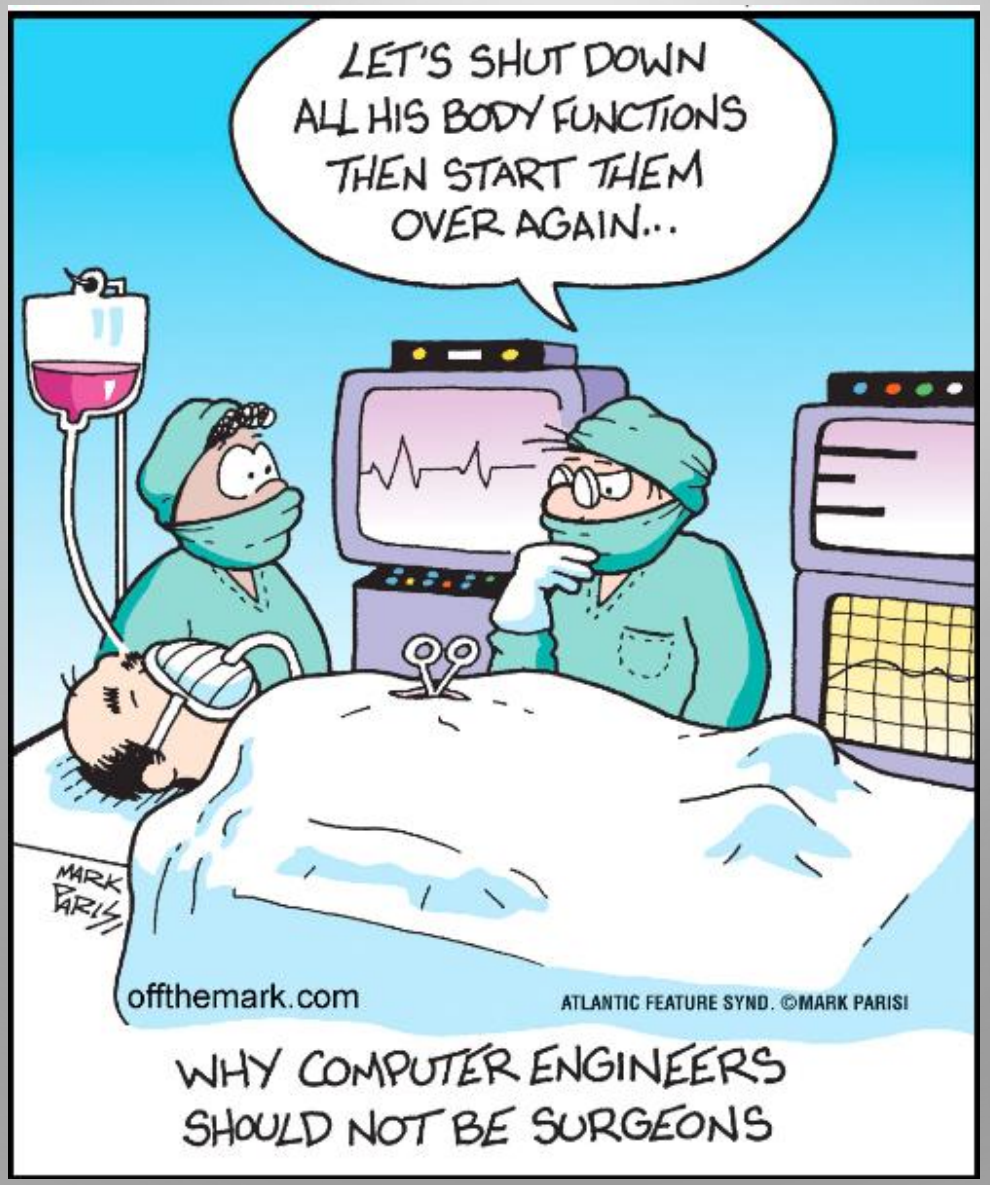
DATA AND PRACTICES

DR ANJALI KULKARNI

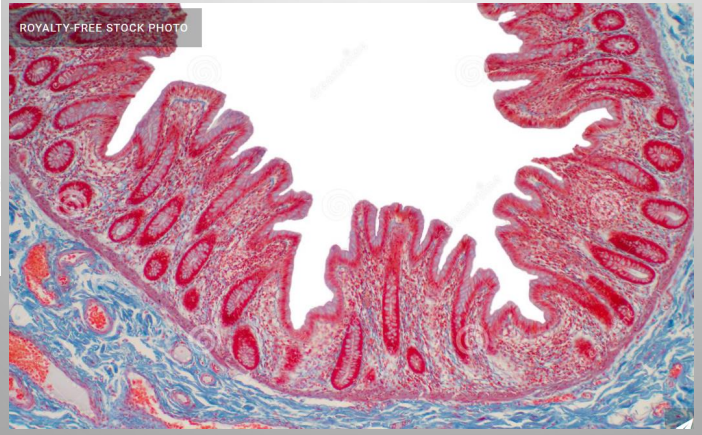
VP-4BASECARE RWD STRATEGY AND ANALYTICS


LET'S SHUT DOWN ALL HIS BODY FUNCTIONS THEN START THEM OVER AGAIN...
MARK PARISI
offthemark.com
ATLANTIC FEATURE SYND. ©MARK PARISI
WHY COMPUTER ENGINEERS SHOULD NOT BE SURGEONS

CLINICAL DATA FORMATS

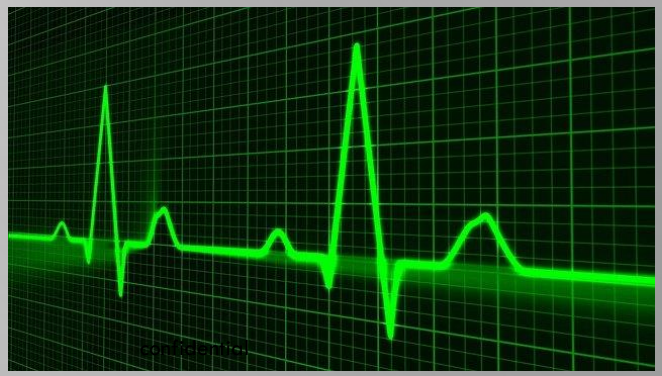
Text , Numerical data

HW-ST911	the unique instrument name
111	the run id
C0N4WACXX	the flowcell id
5	flowcell lane
1101	tile number within the flowcell lane
2249	'x'-coordinate of the cluster within the tile
2216	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 (paired-end or mate-pair reads only)
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on
TTAGGC, CGATC	index sequence



Imaging data

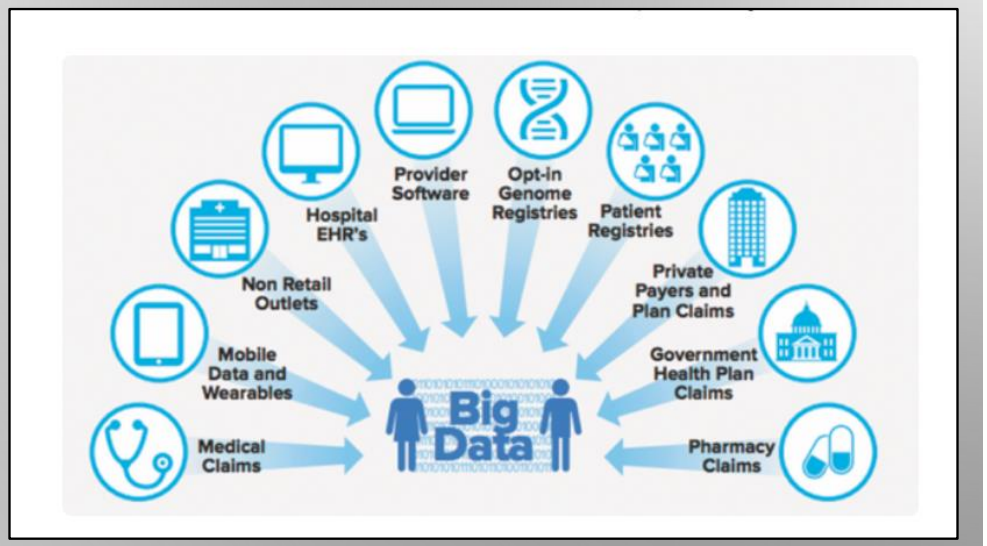
ECG Signals



Complex Genomic VCF , FASQ

```
@HWI-ST911:111:C0N4WACXX:5:1101:2249:2216 1:N:0:TTAGGC CGATC:@@@FF
NATGGCACCATTAAAAAGAATGTTTTATATGGTGTTGAGAAGGACAAAGCTGAAGAAGAAATTTAGTCTGCACTTGATGTTGCAAATGCAAAGAAA
+
#2A2<CCFHIIIIIIIIIIIIIGCCHIIIGIIIFFHIIDGHIGIIIIIIICHGIIIGGCECEGICFHCECDEFFFFDEEEEEDDDDDCDDCDDDDBC
@HWI-ST911:111:C0N4WACXX:5:1101:2509:2197 1:N:0:TTAGGC CGATC:1+4=B
NATGAGATAAATCAATTGTCTTTAATGAAGTACAGTCTTTGAATAATGAGTTTTGAACTCTTCTGCAACTTTTTGGAAACTTTAAAGTTTGTAATG
+
#4A2<AADHIIIIIIIIIHHIIIIIIIIIIFGIII@GIIFIIIGIIIIEIDHEHIIIHIIIIIIIIIIICHIIIHHEEDFFFFFEEECEEEADDFC
@HWI-ST911:111:C0N4WACXX:5:1101:3746:2179 1:N:0:TTAGGC CCATC:+11+A
NATGTCATCCATCTTTTCCTATCTAAAAAAGAATCAAAAAAGGGATAGTACAGAGGGAAAGTTCAATCCAGAGGACGATGAAACACTGATTGATGG
+
```

Today, approximately 30% of the world's data volume is being generated by the healthcare industry.



Real World data market is set to increase to 1.66billion $ by 2029

https://www.meticulousresearch.com/product/real-world-data-market-5297

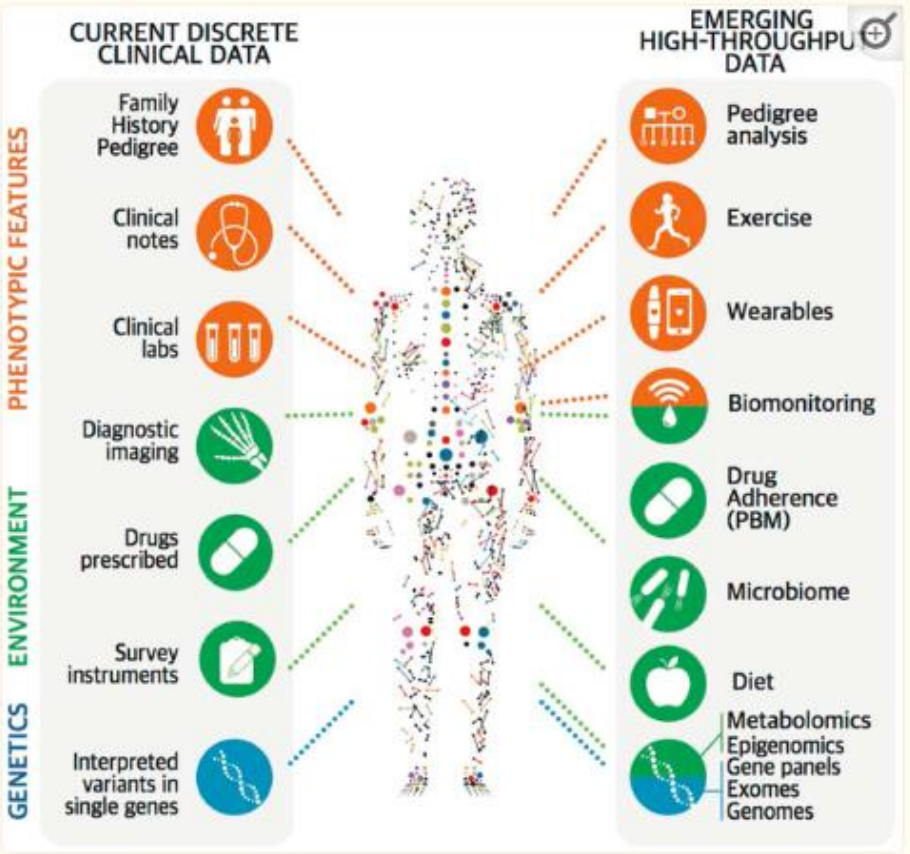


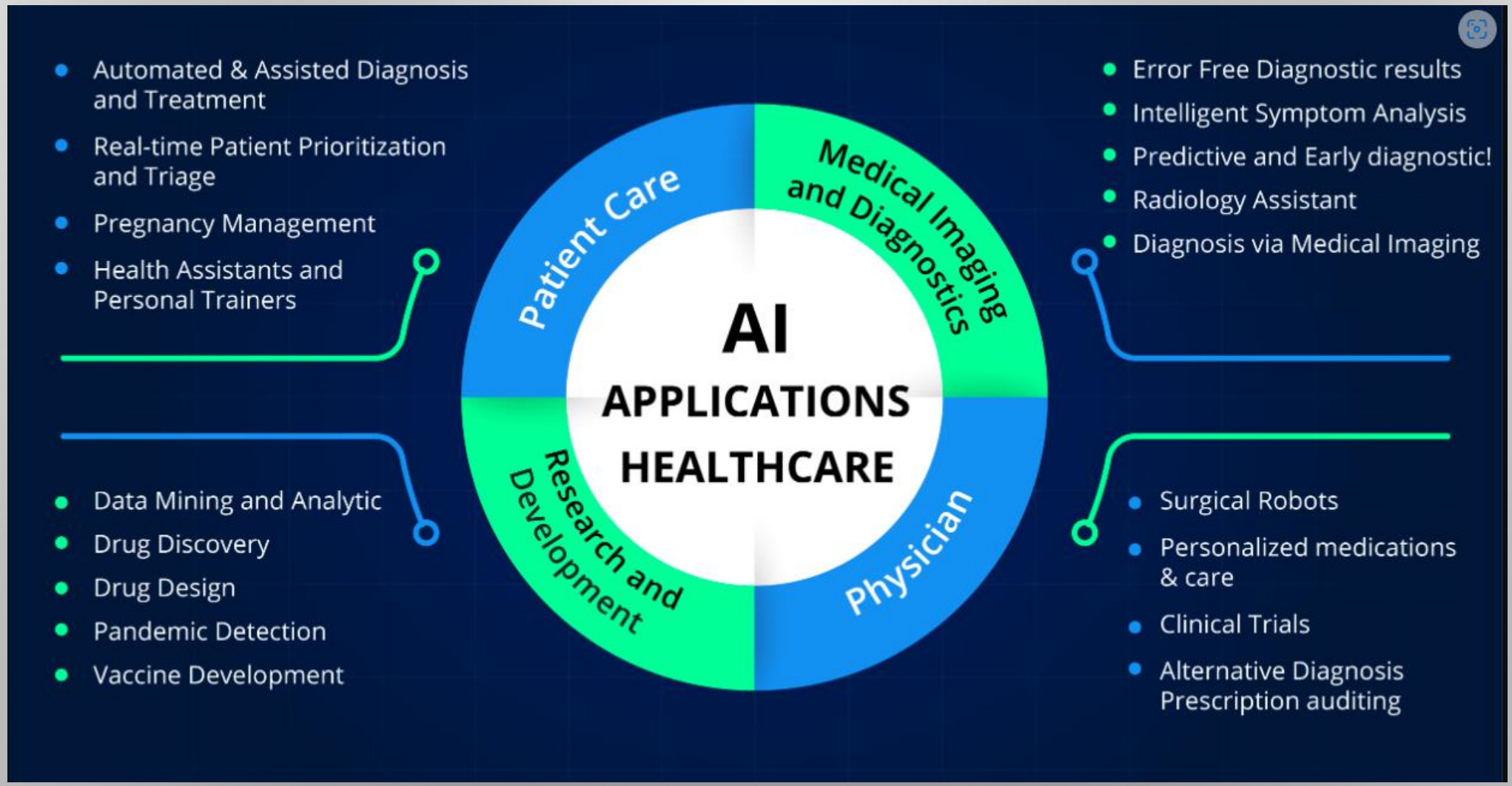
Figure 1.

Multimodal Clinical and High-Throughput Data, Captured in Diverse Ways.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6503847/

COMPLEXITY OF HEALTH CARE DATA

Sources of training data	Data structure types	Data locations
■ Images ■ Videos ■ Emails ■ Documents, PDFs ■ Quantitative data from databases ■ Spreadsheets ■ Text content ■ Voice and audio	■ Structured data □ Queryable with a well-defined schema ■ Unstructured data □ No well-defined schema, not queryable ■ Semi-structured data □ Limited structure and queryability	■ On-premises databases ■ Data warehouses ■ Data Lakes ■ Cloud sources ■ Edge devices ■ Hybrid ■ Sensors/IoT



2021-04-01.png (5001×2626) (techment.com)

SYNTHETIC DATA

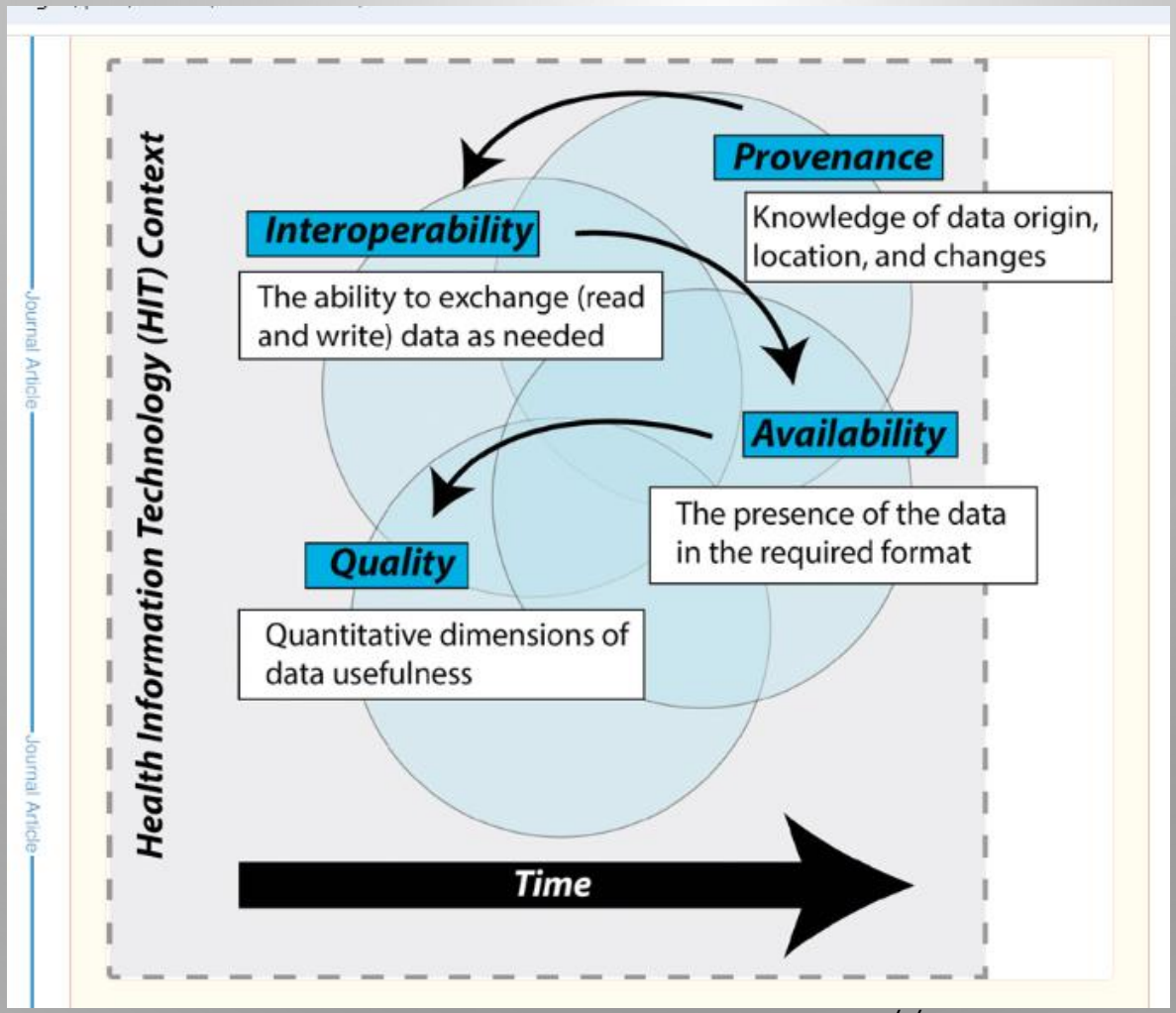
- **Synthetic data is artificially generated data that mimics the structure and properties of real-world data.**

- It is created using mathematical models and algorithms to generate realistic datasets without compromising privacy.

- Used in clinical trials , researchers can use synthetic data to create a so called "synthetic control arm"

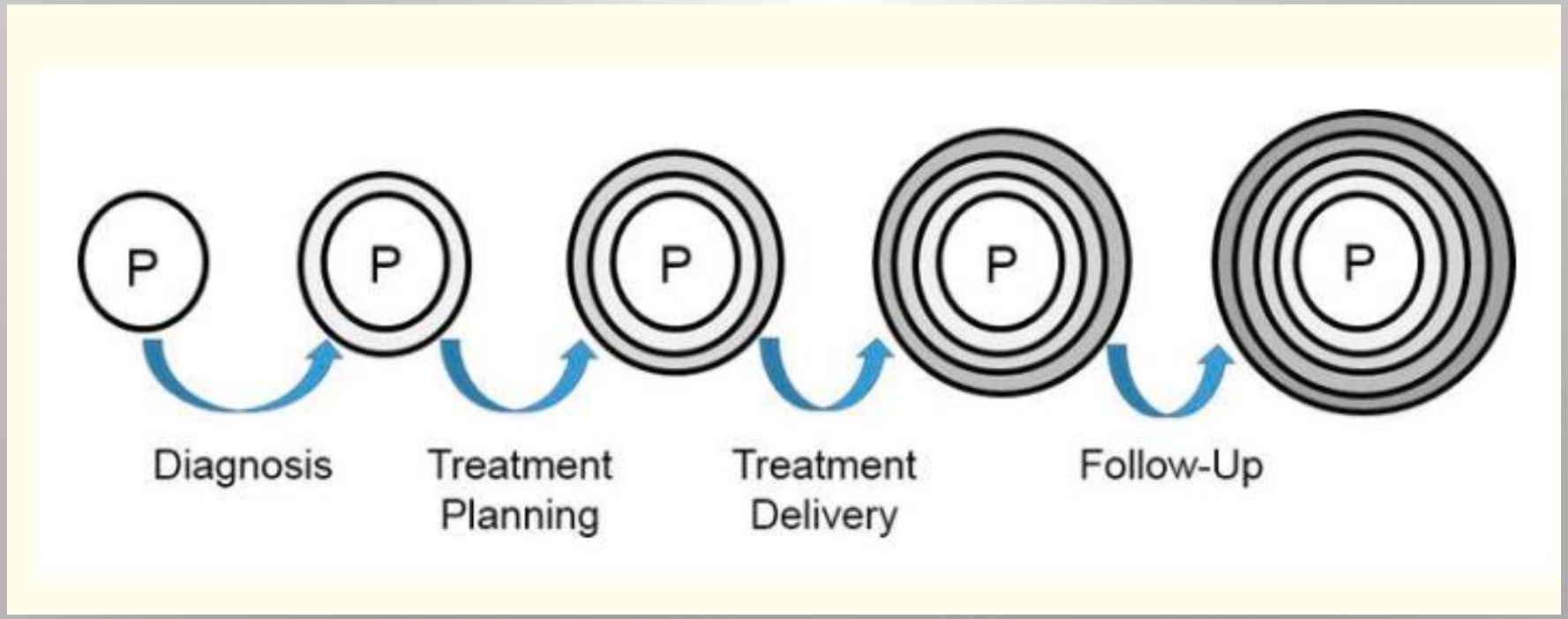
- Using synthetic data for healthcare research brings many benefits, including privacy protection, data availability, scalability, research collaboration, reproducibility, and addressing issues related to data bias and representativeness.

Data Readiness conceptual framework



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8294946/

Radiation Oncology data burden !!



Standardization in healthcare data collection and management is very important

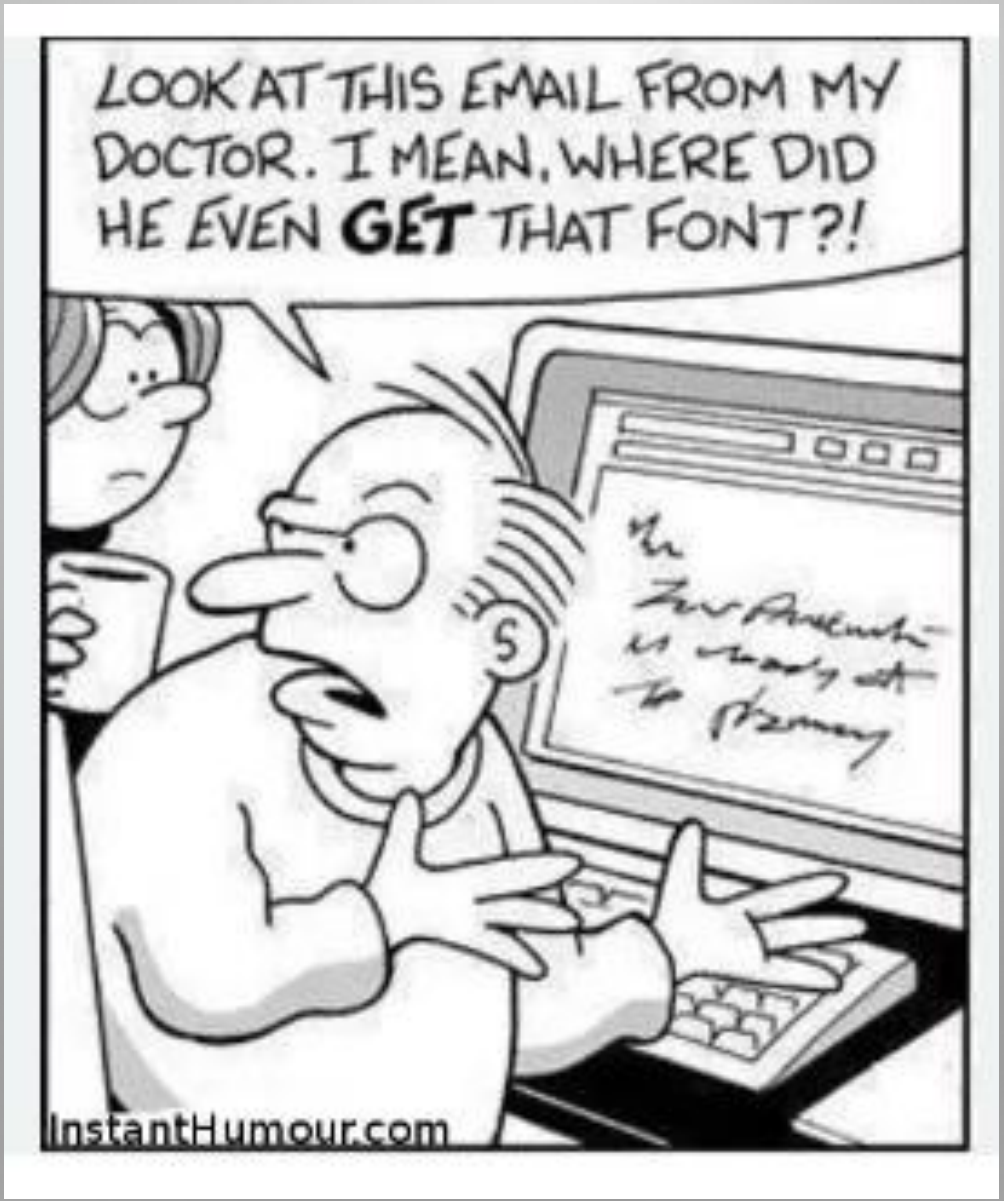
Making Data Computable

Standards, Terminologies, and Ontologies Widely Used in Clinical Medicine.*

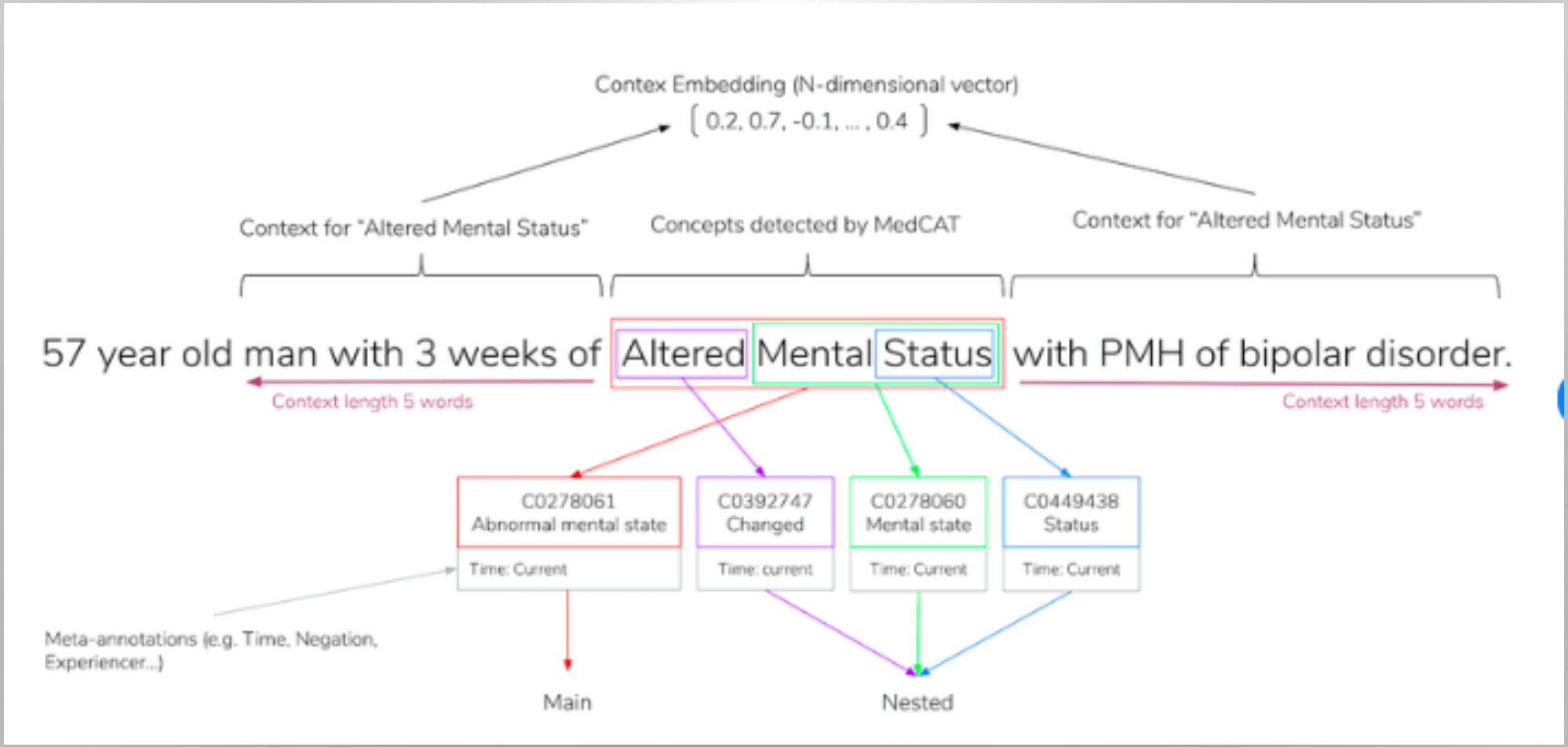
Type of Data	Ontology	Example of Term
Diagnoses	Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) ICD Orphanet Rare Disease Ontology (ORDO) National Cancer Institute Thesaurus (NCIT)	Triple-negative breast carcinoma (NCIT:C71732)
Phenotypic abnormalities	Human Phenotype Ontology (HPO)	Bronchopulmonary sequestration (HP:0010960)
Medications	RxNorm DrugBank ChEMBL	Panobinostat (CHEMBL483254)
Adverse reactions	Ontology of Adverse Events (OAE)	Injection-site induration (OAE:0000323)
Procedures	Medical Dictionary for Regulatory Activities (MedDRA)	Cardiac aneurysm repair (MEDDRA/10007514)
Laboratory examinations	Logical Observation Identifiers Names and Codes (LOINC)	Creatinine in serum or plasma (LOINC:2160-0)
Imaging data	Digital Imaging and Communications in Medicine (DICOM) RadLex	Periosteal cortical thinning (RID45761)

Ontologies differ from terminologies in that ontologies define relationships between concepts in a way that allows computational logical reasoning, enabling the drawing of conclusions from related assertions

Major challenge in the unstructured clinical notes


LOOK AT THIS EMAIL FROM MY DOCTOR. I MEAN, WHERE DID HE EVEN GET THAT FONT?!
InstantHumour.com

Natural Language processing –NLP



Need for Natural language processing (NLP) to derive insights

1. **Information Extraction:** Extraction of drugs, events, diseases through named entity recognition

2. **Specific clinical insight or question answering**: Provide answers to clinical questions after extracting entities from unstructured text

3. **Clinical Decision Support Systems**: Build clinical decision support systems by extracting various useful information for making diagnosis and therapeutic decisions,

4. **Medical Resource Allocation**: building patient triage systems, so that medical resources can attend to critical cases with priority

5. **Patient Clinical Data Summarization:** Extract and summarize longitudinal patient clinical journey

6. **Clinical Research:** NLP may enable efficient clinical trial design. Patient recruitment, clinical trial analytics

7. **Mental /Occupational Health Monitoring**: NLP-driven smart healthcare has great value in predicting/diagnosing and treating mental health conditions.

8. **Drug Review and Safety Monitoring** Adverse drug events discovery and drug safety monitoring

https://arxiv.org/pdf/2110.15803.pdf

Digital transformation and AI knocking--

- ✓ Adopt the digital applications
- ✓ Periodic data quality checks
- ✓ Data audits
- ✓ Training Good Data practices

Thank you