Career Development Series 2020

Storing and Managing Data in 21st Century

Presentation will begin at 12:00 PM (PT)

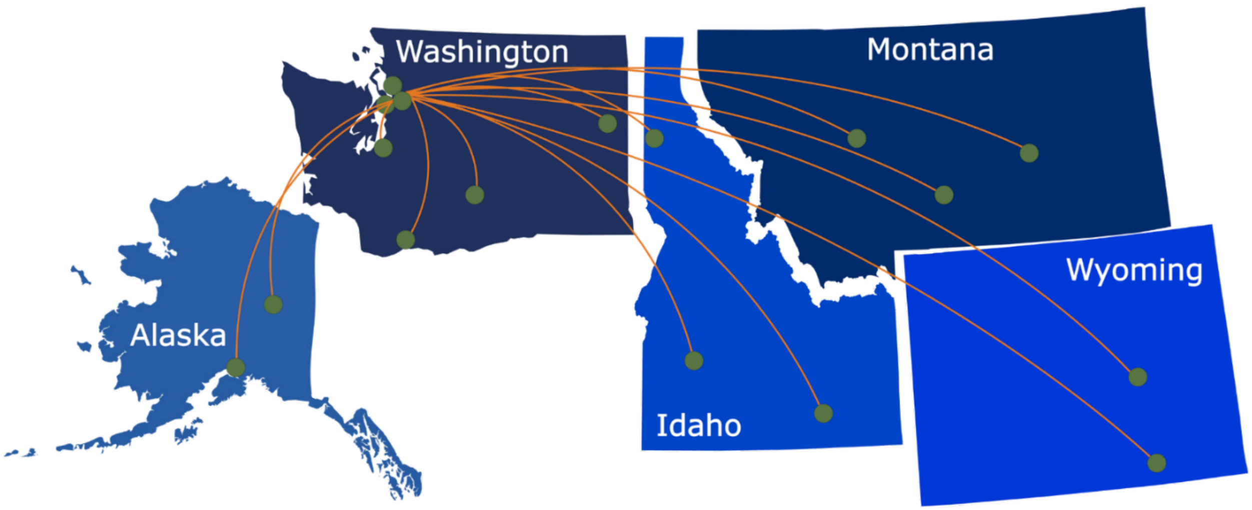
ITHS | Institute of Translational Health Sciences

ACCELERATING RESEARCH. IMPROVING HEALTH.

UNIVERSITY of WASHINGTON

FRED HUTCH
CURES START HERE

Seattle Children's
HOSPITAL · RESEARCH · FOUNDATION



What We Offer:

1. **Research Support Services:** Members gain access the different research services, resources, and tools offered by ITHS, including the ITHS Research Navigator.

2. **Community Engagement:** Members can connect with regional and community based practice networks

3. **Education & Training:** Members can access a variety of workforce development and mentoring programs and apply for formal training programs.

4. **Funding:** Members can apply for local and national pilot grants and other funding opportunities. ITHS also offers letters of support for grant submissions.

Contact our **Director of Research Development**

- **Project Consultation**
- **Strategic Direction**
- **Resources and Networking**

Melissa D. Vaught, Ph.D.
ithsnav@uw.edu
206.616.3875

Upcoming Career Development Series 2020

July 22 – Evidence Synthesis Primer: A Step by Step Guide

No ITHS CDS held in the month of August

Sept. TBD – Teaching How to Give Constructive Feedback

ITHS | Institute of Translational Health Sciences
ACCELERATING RESEARCH. IMPROVING HEALTH.

Feedback

At the end of the seminar, a link to the feedback survey will be sent to the email address you used to register.

Career Development Series 2020

Storing and Managing Data in 21st Century

Presented by Sean Mooney, PhD

Chief Research Information Officer (CRIO) and professor in the Department
of Biomedical Informatics and Medical Education at the University of Washington

ITHS | Institute of Translational Health Sciences
ACCELERATING RESEARCH. IMPROVING HEALTH.

Learning Objectives

1. Attendees will be aware of the challenges for modern research data management.

2. Attendees will understand the basic concepts of making data more FAIR (Findable, Accessible, Interoperable and Reuseable).

3. Attendees will acquire ideas on where research data technologies are headed.

Poll Question

Please answer the question in the poll

What this talk is not going to do

In this talk, I am *not* going to:

- Lessons on how to program
  - Learn R or Python (http://python.org/doc/ - click on tutorial)
- Lessons on how to use SQL
  - Start with MySQL (http://mysql.com/ and take tutorial)
- Lessons on how to build REDCap forms
  - Start with our REDCap Trainings
- Teach you how to administer computer environments in the 'Cloud'
  - Start with AWS – S3, EC2, etc
  - https://docs.aws.amazon.com/AmazonS3/latest/gsg/GetStartedWithS3.html

What I am going to tell you

In this talk, I am going to:

- Focus on why data management is hard
- Focus on current trends in computer resources on the internet, ‘cyberinfrastructure’
- Focus on principles of creating useable scientific data
- And discuss why and how data management is becoming more collaborative

UW Medicine Research Information Technology

UW Medicine

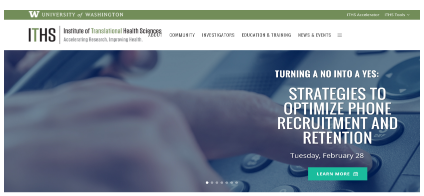
UW SCHOOL OF MEDICINE

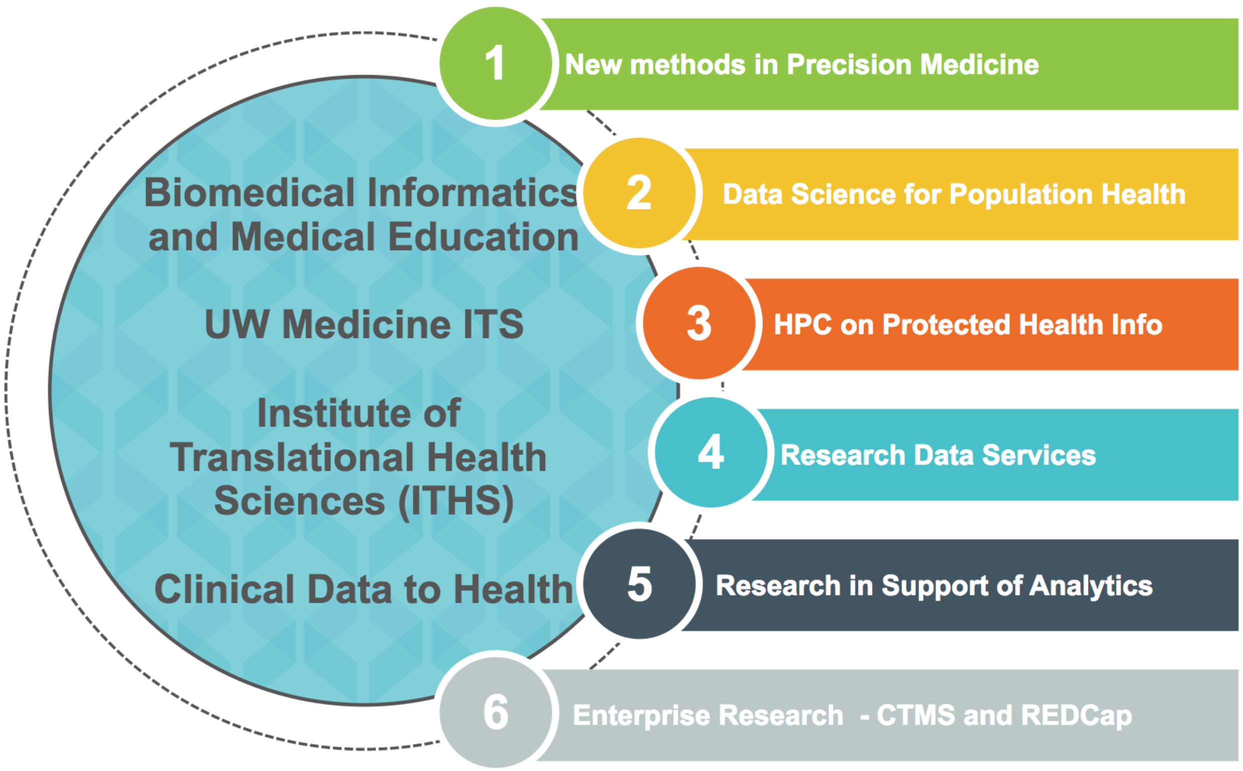
BIOMEDICAL INFORMATICS AND MEDICAL EDUCATION

Melissa Habrat, Director of Research IT, UW Medicine

UW Medicine

INFORMATION TECHNOLOGY SERVICES





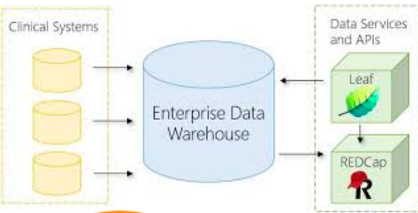
ITHS | Institute of Translational Health Sciences
ACCELERATING RESEARCH. IMPROVING HEALTH.

What is cyberinfrastructure? A headache!

REDCap
Research Electronic Data Capture

OHDSI

NCBI



How do we connect tools and data together when they are developed on different platforms with different philosophies?

i2b2 tranSMART FOUNDATION

amazon web services

Home grown electronic systems

Prepackaged Software:
- R, Perl, Python
- MatLab
- Commercial data management

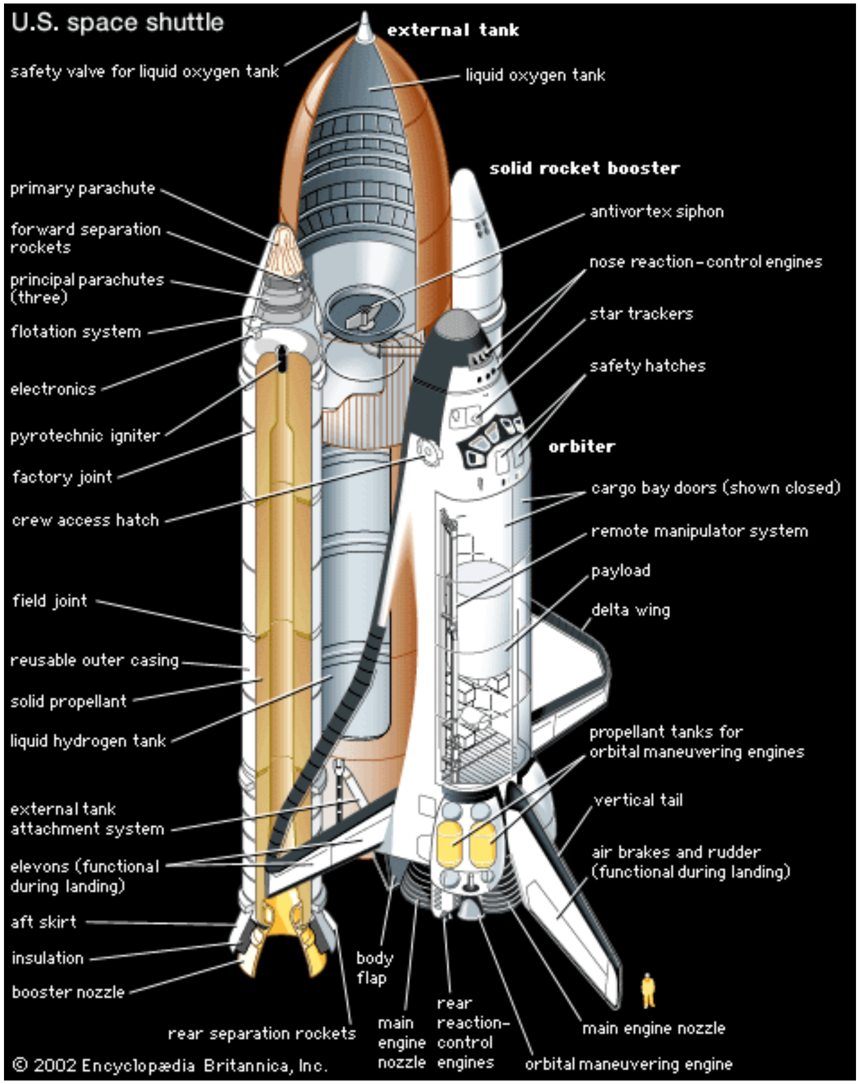
NIH Funded Networks and Data Coordinating Centers

ITHS Institute of Translational Health Sciences
ACCELERATING RESEARCH. IMPROVING HEALTH.

pcornet®

Microsoft Azure

How do I see it? A space program as an analogy

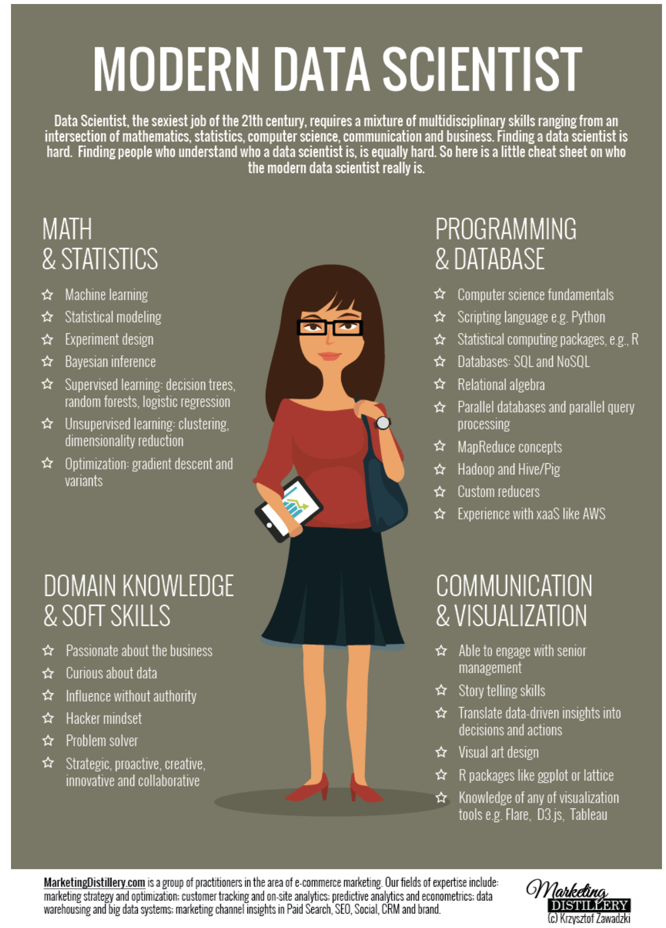


- Each space shuttle, before it was retired, cost about $2.1 *billion*
- Thousands of parts must be designed and assembled to create a working space craft
- *Today biomedical informatics is much like a space craft, lots of parts we must put together to operate!*

We have to put the ship together!

- Collectively we are constructing the components to connect research and clinical data electronically.
- These components do not necessarily work together!
- However, it is up the service providing informatician at the institution to build the '*space ship*' from these components
- *Scientists are the astronauts!*

Data is transforming both research about healthcare and its delivery



http://www.marketingdistillery.com/2014/11/29/is-data-science-a-buzzword-modern-data-scientist-defined/

Institutions are investing in data in the health sciences…

Stanford University recently created a Department of Data Science in their School of Medicine

Harvard Medical School has created a Department of Biomedical Informatics

UW has recruited faculty heavily in informatics over past five years (>10 faculty)

This is creating a research demand that touches many areas of computing

There is lots of different types of data

EHR Data

Clinical Text

Clinical Image Data

Social and environmental determinants of health

Sensor data and smart device data

Online data and social media data

Inpatient sensor and data streams data

Genome data

Molecular data (genomics/proteomics/metabolomic/etc)

Clinical case report form data

Claims/administrative data

What are some others?

How we manage data has evolved considerably over past two decades

- For example, clinical form data collection

- Rough history (from oldest to modern):
  - Paper case report forms
  - Excel spreadsheets
  - Access databases
  - Other database systems
  - Home grown websites
  - REDCap, cloud file storages, cloud based databases, etc

- The change is that for many challenges in data, we now have off-the-shelf tool solutions

The rise of ’FAIR’ data

One of the most discussed concepts is the idea that data should open and useable by researchers as easily as possible. One model is the ‘FAIR’ framework.

FAIR is the concept that research data should be:

1. findable,
2. accessible,
3. interoperable and
4. reuseable

What does that mean in practice?

The rise of 'FAIR' data

FAIR was first described in this publication from 2016

Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons -Show fewer authors

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

’FAIR’ data: Findable

Findable data: From https://www.go-fair.org/fair-principles/

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the **FAIRification process**.

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

’FAIR’ data: Findable

Findable data: From https://www.go-fair.org/fair-principles/

There’s a lot going on there. Let’s break it down a bit.

Findable means that an investigator somewhere else should be able to find your data, usually on the internet.

Two important concepts here:

Persistent identifiers, and

Metadata

Metadata registered in a findable resource

’FAIR’ data: Findable

Persistent identifiers:

- Universal identifier for data (like a web address)
- Generally, you won’t need to worry too much about this
- Examples include Pubmed IDs, DOIs, database IDs

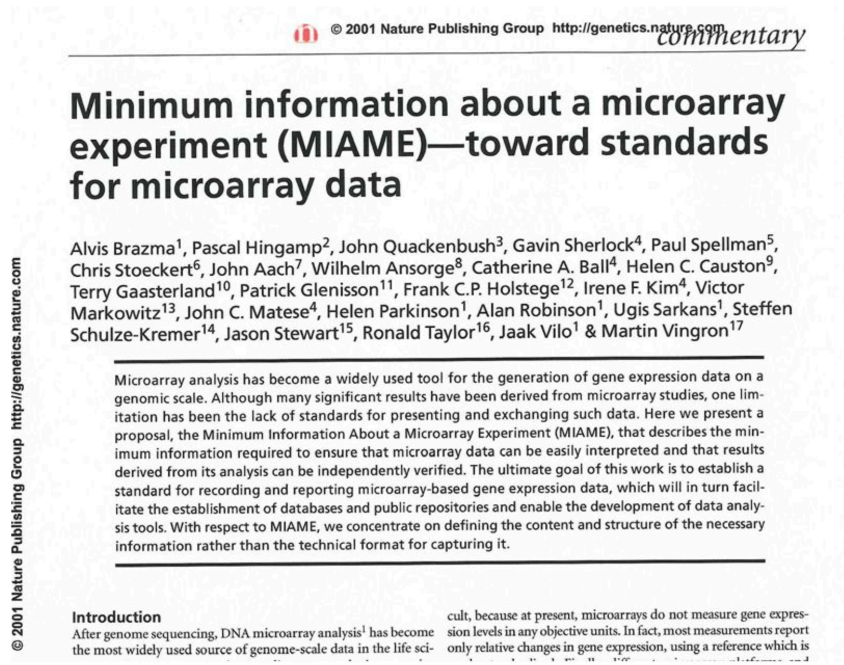
Metadata:

- Very important and something you should think about
- Metadata is data that describes data
- Often overlooked
- Should be annotated richly by the investigator and you should spend sometime with it when asked
- Biomedical ontologies are great for annotating metadata
- Metadata can be findable in resources such as Google, but we need more

Metadata

There are some standards about how metadata should be collected an early version was MIAME, now http://fairsharing.org/

MIAME element	GEO Entity	GEO attribute
Experiment	Series	Title
		Type
		Summary
		Overall design
		PubMed ID
		Web link
Biological samples, preparation extraction and labeling	Sample	Organism
		Label
		Label protocol
		Extracted protocol
		Extracted molecule
		Growth protocol
		Treatment protocol
		Source
		Biomaterial provider
		Description
		Characteristic
Array	Platform	Title
		Distribution
		Technology type
		Manufacturer
		Manufacturer Protocol
		Catalog number
		Coating
		Support
		Description
Hybridization	Sample	Hybridization Protocol
		Description
		Sample type
Measurement	Sample	Scan protocol
		Data processing


© 2001 Nature Publishing Group http://genetics.nature.com

commentary

Minimum information about a microarray experiment (MIAME)—toward standards for microarray data

Alvis Brazma[1], Pascal Hingamp[2], John Quackenbush[3], Gavin Sherlock[4], Paul Spellman[5], Chris Stoeckert[6], John Aach[7], Wilhelm Ansorge[8], Catherine A. Ball[4], Helen C. Causton[9], Terry Gaasterland[10], Patrick Glenisson[11], Frank C.P. Holstege[12], Irene F. Kim[4], Victor Markowitz[13], John C. Matese[4], Helen Parkinson[1], Alan Robinson[1], Ugis Sarkans[1], Steffen Schulze-Kremer[14], Jason Stewart[15], Ronald Taylor[16], Jaak Vilo[1] & Martin Vingron[17]

Microarray analysis has become a widely used tool for the generation of gene expression data on a genomic scale. Although many significant results have been derived from microarray studies, one limitation has been the lack of standards for presenting and exchanging such data. Here we present a proposal, the Minimum Information About a Microarray Experiment (MIAME), that describes the minimum information required to ensure that microarray data can be easily interpreted and that results derived from its analysis can be independently verified. The ultimate goal of this work is to establish a standard for recording and reporting microarray-based gene expression data, which will in turn facilitate the establishment of databases and public repositories and enable the development of data analysis tools. With respect to MIAME, we concentrate on defining the content and structure of the necessary information rather than the technical format for capturing it.

Introduction

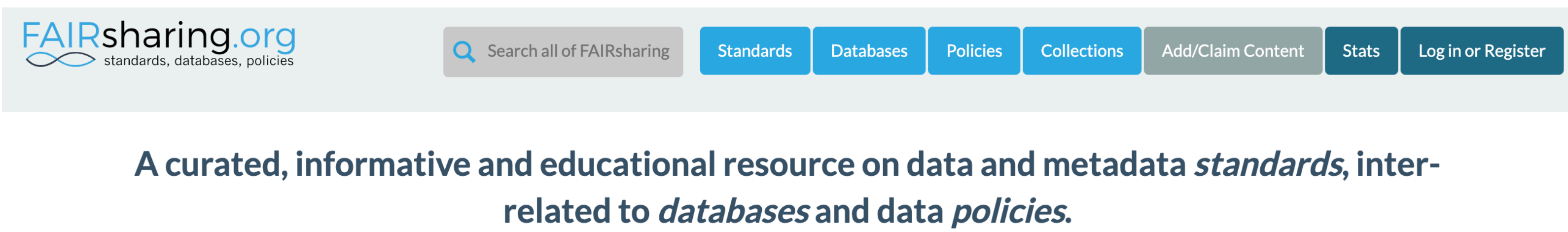
After genome sequencing, DNA microarray analysis[1] has become the most widely used source of genome-scale data in the life sci- … cult, because at present, microarrays do not measure gene expression levels in any objective units. In fact, most measurements report only relative changes in gene expression, using a reference which is

© 2001 Nature Publishing Group http://genetics.nature.com

ITHS | Institute of Translational Health Sciences
ACCELERATING RESEARCH. IMPROVING HEALTH.

Metadata

FAIR Sharing provides a list of standards for different types of data
http://fairsharing.org/



An Important Aside: Biomedical Ontologies

Ontologies are more than nomenclature, more than a controlled vocabulary and more than a taxonomy but capture all of those in their use

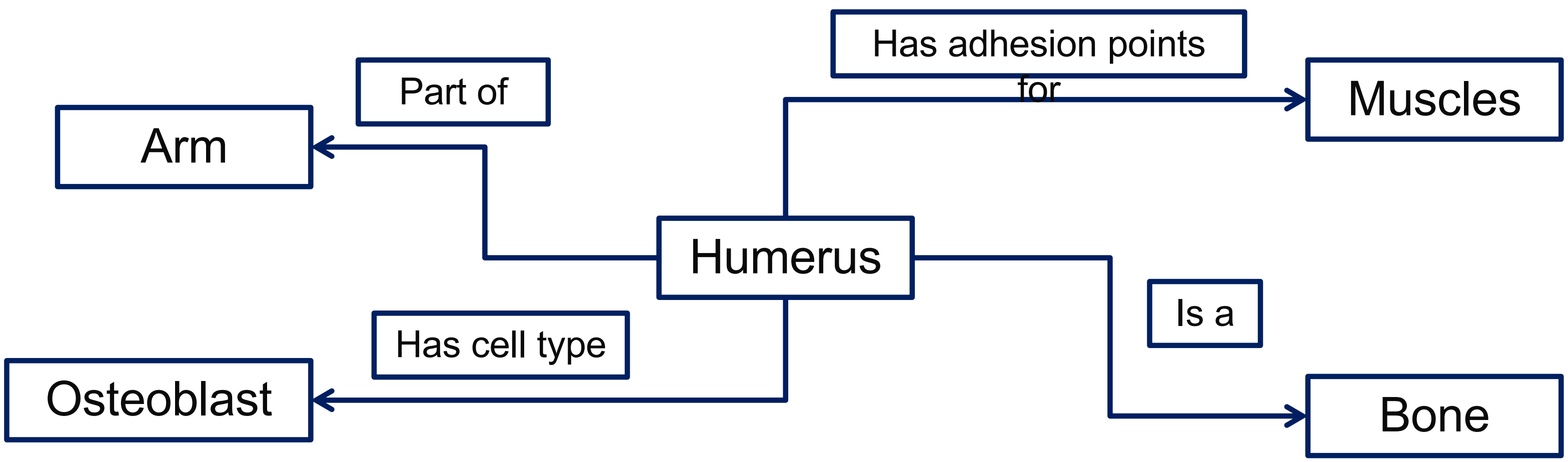
What is an ontology?

Formal way of representing knowledge in which terms are described both by their meaning and their relationship to each other.

When this framework is used to represent biological knowledge the result is a bio-ontology.

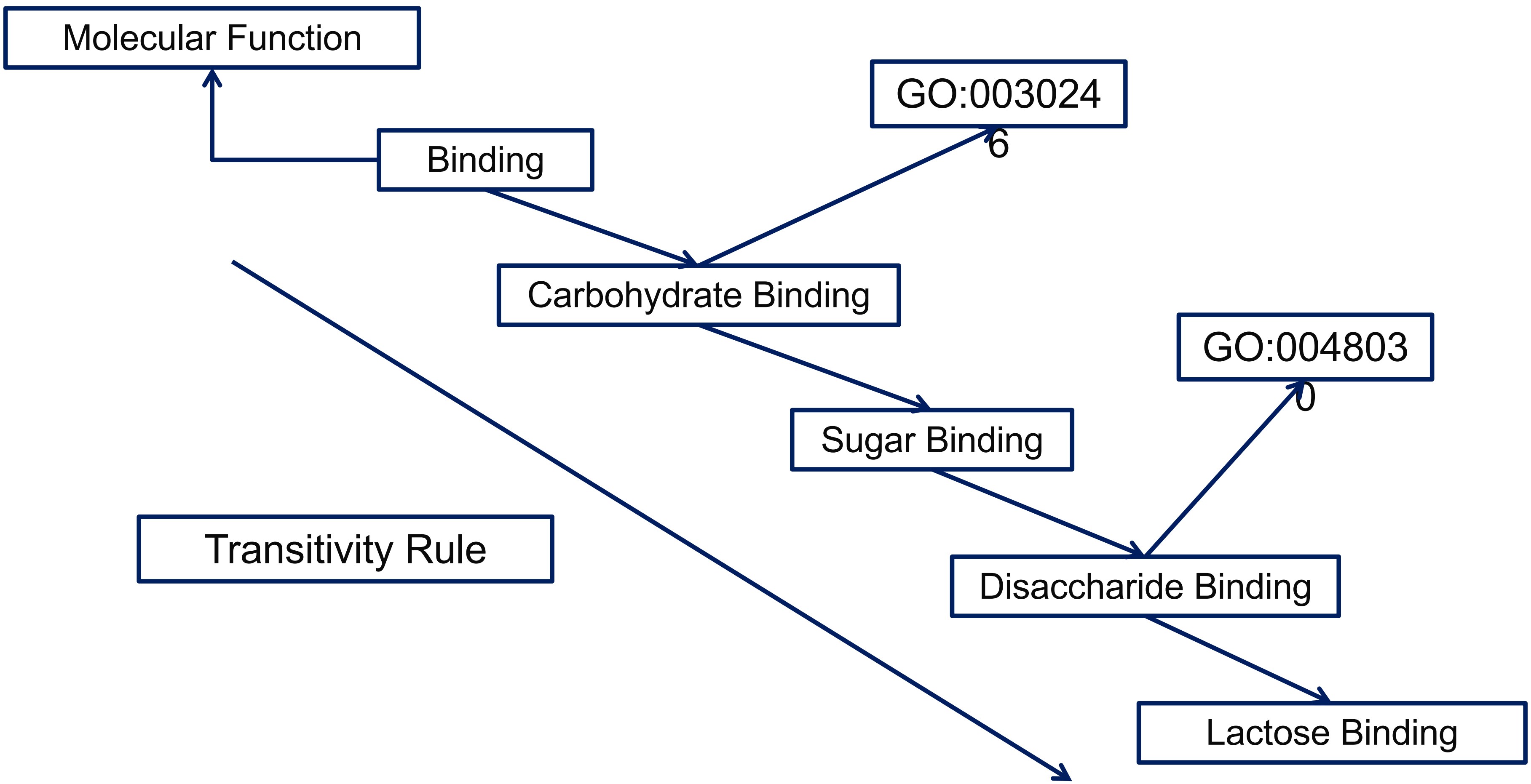
Representation of data within an ontology

- Ontologies can be represented as graphs
- The nodes in the graph represent terms.
- The edge represents the relationship between the nodes.
- Example



Representation of data within an ontology

Example from Gene Ontology



ITHS | Institute of Translational Health Sciences
ACCELERATING RESEARCH. IMPROVING HEALTH.

Open Biological And Biomedical Ontologies

Communities of scientists build ontologies in the biomedical domain.

Goals:

- ► Interoperability between ontologies
  - Common design
  - Implementation
- ► Smoothen data and information integration, retrieval, annotation
- ► Natural language processing and decision support

National Center for Biomedical Ontology (NCBO)

- ► Repository
- ► Single point to access all ontologies
- ► http://bioportal.bioontology.org/

Open Biomedical Ontologies Consortium (OBO)

Ontologies

NCBO integrates over 200 ontologies.

High level categories

- Anatomy
- Phenotype
- Experimental conditions
- Genomic and Proteomic
- Chemistry
- Health

Example of Ontologies

- Couple of examples for each category
- High level categories
  - Anatomy
    - Foundational model of anatomy
    - Drosophila gross anatomy
  - Phenotype
    - C. elegans phenotype
    - Human Disease
  - Experimental conditions
    - Tissue Microarray ontology
    - Ontology of clinical research
  - Genomic and Proteomic
    - Gene Ontology
    - Protein Ontology
  - Chemistry
    - Chemical entities of biological interest
    - Lipid Ontology
  - Health
    - Medical subject headings (MSH)
    - SNOMED clinical terms

Gene Ontology

Initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases.

Three domains

- ► Cellular component
  - parts of a cell or its extracellular environment
  - Mitochondria, Ribosome
- ► Molecular function
  - the elemental activities of a gene product at the molecular level
  - Binding, catalysis
- ► Biological process
  - operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms
  - Oxidative phosphorylation, cell death

Gene ontology is structured as a directed acrylic graph.

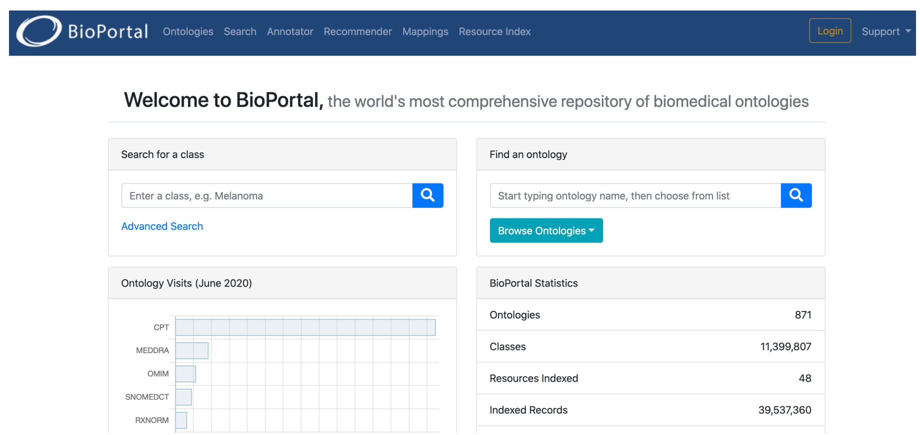
- ► Each term has defined relationship with one or more terms.

Biomedical ontologies: Summary

Biomedical ontologies provide nomenclature for research disciplines but so much more, including:

- Relationships between concepts ‘femur’ *is a* ‘bone’
- Universal identifiers for a concept
- The text label of a concept (e.g. ‘tumor’) and it’s synonyms (e.g. ‘neoplasm’, ‘cancer’, etc)
- The precise definition of a term
- For more info go to bioportal

https://bioportal.bioontology.org/



Ontologies make metadata and data universal

’FAIR’ data: Accessible

Data should be available, like on the internet. Metadata should be available.

Ways to make data available:

- Attach to a paper or manuscript
- Include in a publicly curated database
- Archive in a library archive
- Other ways

’FAIR’ data: Interoperable

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

’FAIR’ data: Interoperable

Interoperable data means that data should work with other data – data should follow a standard and be useable with other data that would be theoretically useful to use

Some concepts:

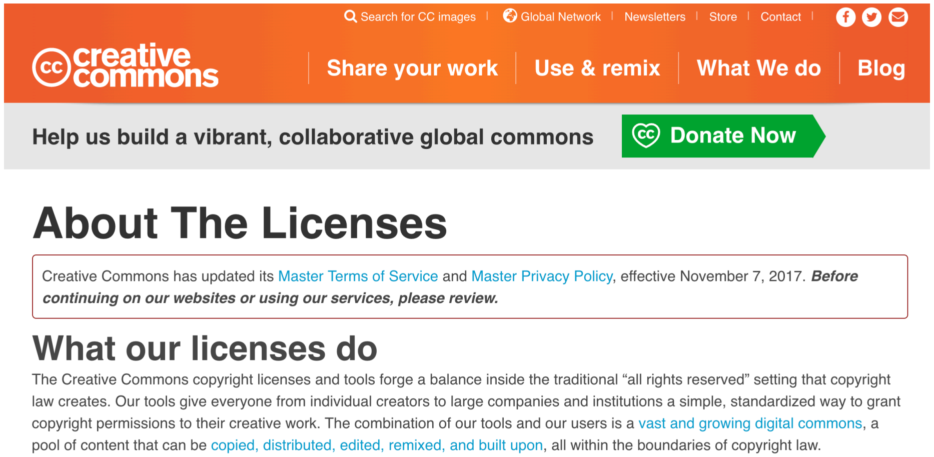
- Data should be collected in a standardized way
- Data should use standard nomenclature (ontologies)
- Data should use standard forms (e.g. PROMIS instruments for Patient Reported Outcomes)
- Data should use standard data models for data when available (such as OMOP for patient EHR data)

’FAIR’ data: Reuseable

Reuseable data should be licensable for reuse

Some concepts:

- Data use agreements should be least restrictive as possible with other considerations (e.g. patient privacy)
- Use of open access copyrights on data such as Creative Commons licensing
- Not just papers! Data too



What FAIR doesn't cover

FAIR doesn't solve all problems:

- Data quality
- Sustainability
- Utility
- Impact

What else?

Some examples in real world

Some standard examples:

- Genomic data
- Phenotypic metadata
- EHR Data
- Case Report Form data

Raw Genomic/Bioinformatics Data

Genomic datasets have standard file formats, data models and nomenclature :

- Genetic – VCF files
- Next generation sequencing – FASTQ
- Sequence databases – FASTA and variants
- Protein structures – Protein Databank Format (PDB)
- Many others …

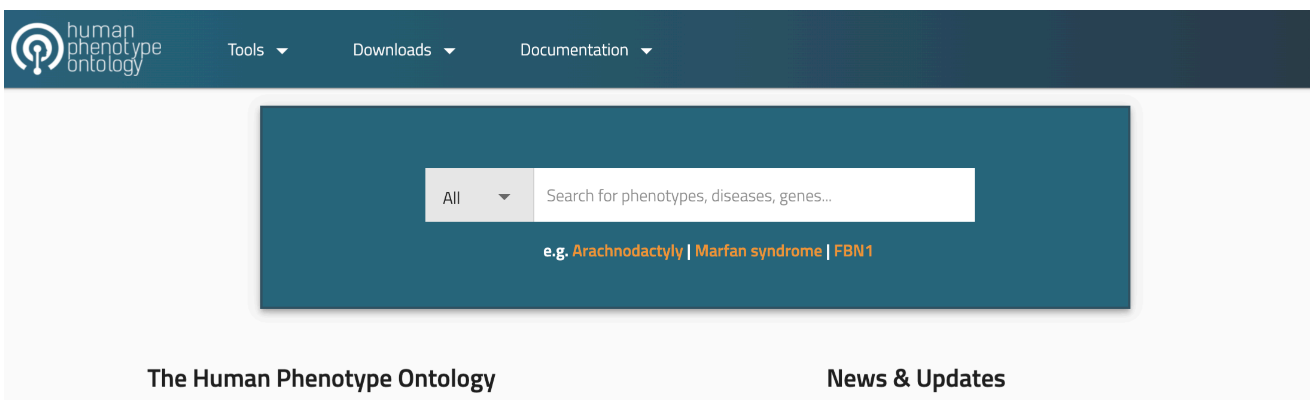
VCF File Example

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo
sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS      ID         REF   ALT    QUAL  FILTER   INFO                               FORMAT       NA00001
NA00002          NA00003
20     14370    rs6054257  G     A      29    PASS    NS=3;DP=14;AF=0.5;DB;H2            GT:GQ:DP:HQ
0|0:48:1:51,51  1|0:48:8:51,51   1/1:43:5:.,.
20     17330    .          T     A      3     q10     NS=3;DP=11;AF=0.017                GT:GQ:DP:HQ
0|0:49:3:58,50  0|1:3:5:65,3     0/0:41:3
20     1110696  rs6040355  A     G,T    67    PASS    NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ
1|2:21:6:23,27  2|1:2:0:18,2     2/2:35:4
20     1230237  .          T     .      47    PASS    NS=3;DP=13;AA=T                    GT:GQ:DP:HQ
0|0:54:7:56,60  0|0:48:4:51,51   0/0:61:2
20     1234567  microsat1  GTC   G,GTCT 50    PASS    NS=3;DP=9;AA=G                     GT:GQ:DP     0/1:35:4
0/2:17:2         1/1:40:3
```

Phenotypic Data

Example: RNA-Seq dataset from human cells that compared normal vs abnormal:

- Metadata will likely include phenotypic information
- Human Phenotype Ontology (HPO) includes information about abnormality



Phenotypic Data

Short stature HP:0004322

A height below that which is expected according to age and gender norms. Although there is no universally accepted definition of short stature, many refer to "short stature" as height more than 2 standard deviations below the mean for age and gender (or below the 3rd percentile for age and gender dependent norms).

Synonyms: *Decreased body height, Small stature, Stature below 3rd percentile, Height less than 3rd percentile, Short stature*

Cross References: *UMLS:C0349588, SNOMEDCT_US:237836003*

Export Associations

Disease Associations | Gene Associations | LOINC Associations

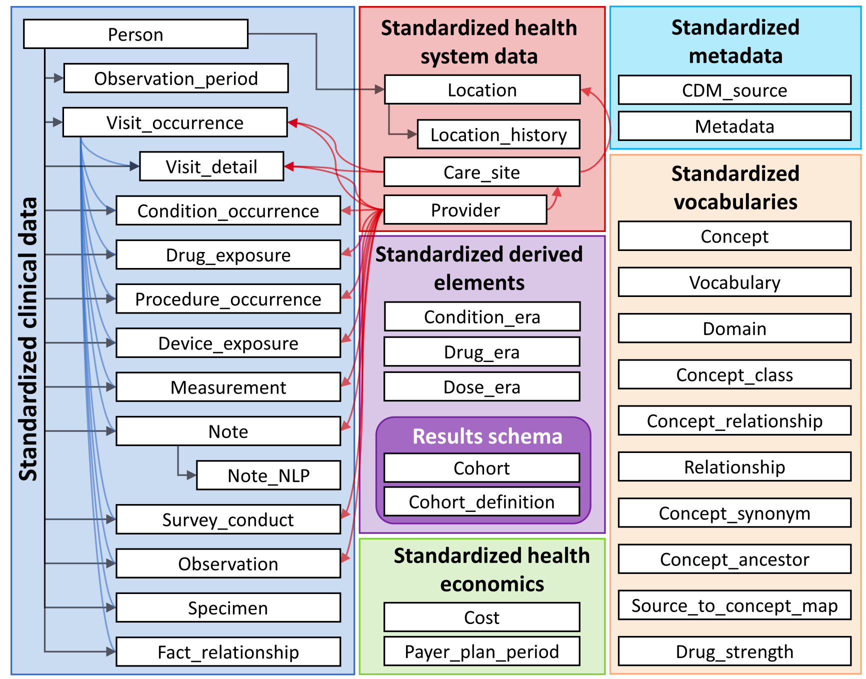
Disease Id	Disease Name	Associated Genes
DECIPHER:76	12q14 Microdeletion Syndrome	
ORPHA:94063	12q14 Microdeletion Syndrome	LEMD3 [23592] HMGA2 [8091]
ORPHA:412035	13q12.3 Microdeletion Syndrome	

Electronic Health Record - EHR Data

EHR data is mapped to 'common data models' or CDMs

Some concepts:

- Models enable comparative effectiveness research, registry construction and data science
- Rely on ontologies and more basic controlled vocabularies to code specific fields, such as diagnosis or medications
- OMOP (figure), PCORI PopMedNet, are examples



EHR Data Networks

Because of EHR Data Models, data is more shareable than ever before

This has created networks of EHR data for research (and other purposes!). These include:

1. OMOP/OHDSI – upwards of 1 Billion Patients
2. PCORNet – pediatric and adult datasets
3. I2b2
4. Accrual to Clinical Trials (ACT)

Example: UW EDW

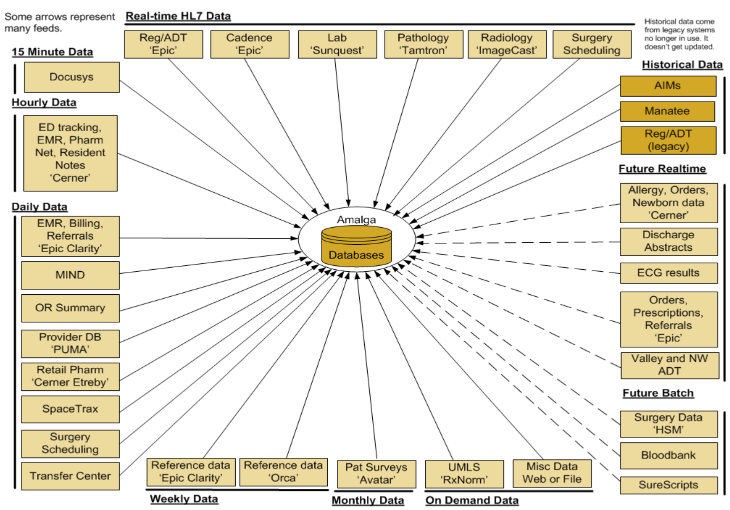
UW Medicine
UNIVERSITY OF WASHINGTON MEDICAL CENTER
HARBORVIEW MEDICAL CENTER
NORTHWEST HOSPITAL & MEDICAL CENTER

Adam Wilcox, PhD
Chief Analytics Officer

Beth Britt, PhD
Director of Analytics

UW EDW Example Continued

4.3 million patients
>50 TB of data
>150 data streams



UW Medicine Enterprise Data Warehouse®

We are working tirelessly to make data useable

We (UW) are making the EDW better

UW Medicine
EHR Data

Adam Wilcox, BIME, and Beth Britt, ITS, has built a draft **OMOP** dataset

Steve Mooney, Epidemiology, is leading effort to **Geocode** all UW Medicine Patients

Meliha Yetisgen, BIME, is annotating **clinical narrative text** with **computable phenotypes and biomedical concepts** and **social determinants of health**

Christina Banderagoda, Civil Engineering, has been working on annotating **environmental determinants of health**

Nic Dobbins built **Leaf** as a self-service tool for access

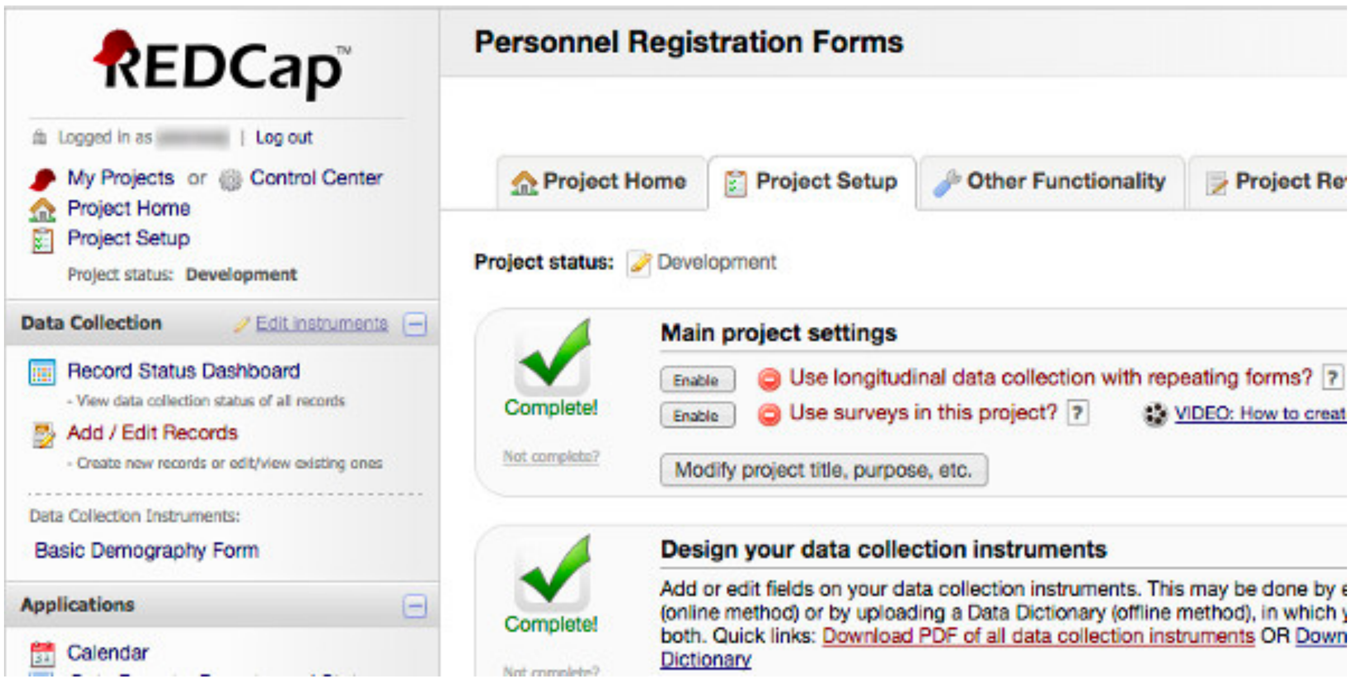
Andrea Hartzler is beginning conversations to better capture **outcomes** through patient reported outcome measures

Several are working on making the cloud more clinical research friendly

Form Data in Clinical Research

REDCap has revolutionized data collection for clinical research

Self service creation of forms for data collection which supports automated QA, ontologies and has an API for computer access



Some Messages

Many tools exist for data management, you should use them and not try and re-invent the wheel

- REDCap – forms
- CDMs - EHR data
- Genomics/Genetics – standard file formats exist
- Standard Nomenclatures – Biomedical Ontologies

The Cloud for Data Management

The Cloud has become ubiquitous:

- Your email is in the cloud (Office265/Gmail/etc)
- DropBox/Box/OneDrive is in the cloud
- Websites are almost all in the cloud now
- Data is stored in the cloud

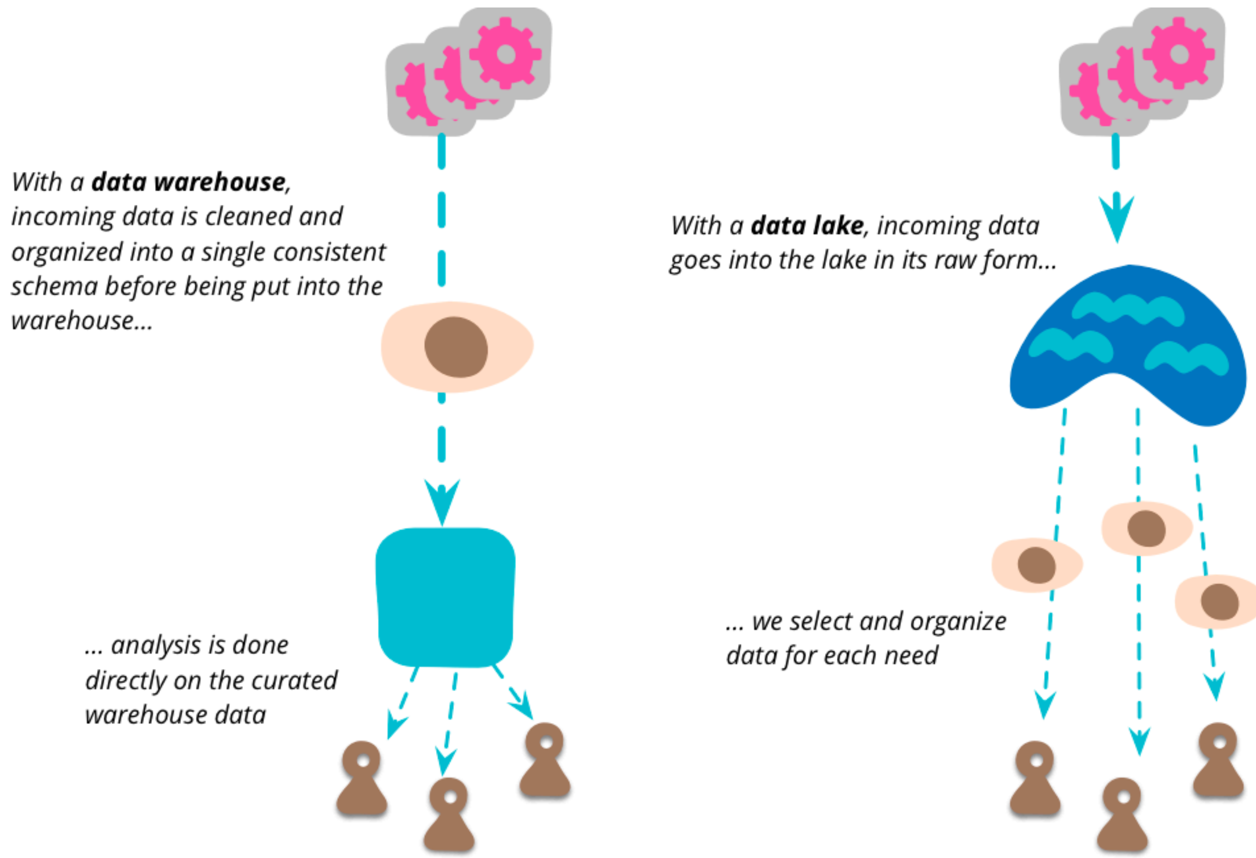
Databases

Databases have grown over time:

- Relational Database – tables (like excel spreadsheets) with fields that are 'relational' with other tables
- Data Warehouse – a collection of multiple sources of data, generally a single location of all of an organization's data and generally curated/integrated/harmonized
- Data Lake – a large, cloud based repository of data, could be databases, documents, images, videos, etc. Usually is not 'relational' instead relying on key – value approaches.

Data Warehouse vs Data Lake

Data warehouses and Data lakes have different uses

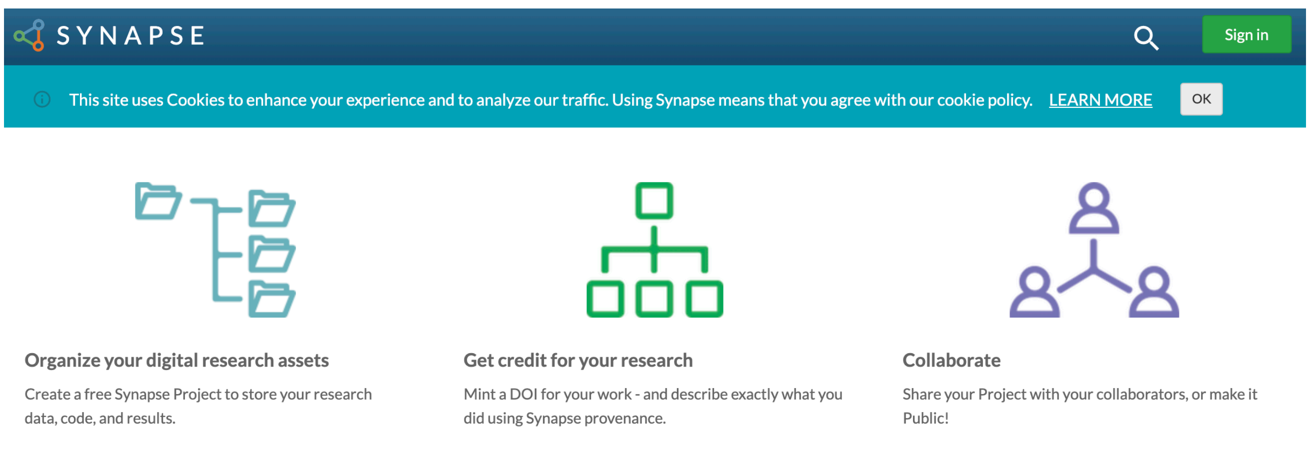


https://martinfowler.com/bliki/DataLake.html

Tools That Make Data Management Easier

Tools that make data management easier:

- REDCap – form data
  - Secure, HIPAA Compliant
- Sage Synapse - https://www.synapse.org/
  - Team based data management
  - Free and easy to use



Data Security

Ensuring the security of data is arguably the most important concern when dealing with datasets

The first step is knowing your data and what you are required to do:

- *Was the data licensed?* What are the terms of that license?
- *Is the data under a Data Use or Transfer Agreement?* What are your obligations under that agreement?
- *Does the data fall under protection under law, such as HIPAA, FERPA or local laws?*

HIPAA and FERPA

		Who must comply?	Protected information	Permitted disclosures[1]
FERPA	The **Family Educational Rights and Privacy Act (FERPA)** is a federal law enacted in 1974 that protects the privacy of **student education records.** The Act serves two primary purposes: 1. Gives parents or eligible students more control of their educational records 2. Prohibits educational institutions from disclosing "personally identifiable information in education records" without written consent 	• **Any public or private school:** – Elementary – Secondary – Post-secondary • **Any state or local education agency** Any of the above must receive funds under an applicable program of the US Department of Education	**Student Education Record:** Records that contain information directly related to a student and which are maintained by an educational agency or institution or by a party acting for the agency or institution	• School officials • Schools to which a student is transferring • Specified officials for audit or evaluation purposes • Appropriate parties in connection with financial aid to a student • Organizations conducting certain studies for or on behalf of the school • Accrediting organizations • Appropriate officials in cases of health and safety emergencies • State and local authorities, within a juvenile justice system, pursuant to specific state law • To comply with a judicial order or lawfully issued subpoena
HIPAA	The **Health Insurance Portability and Accountability Act (HIPAA)** is a national standard that protects sensitive **patient health information** from being disclosed without the patient's consent or knowledge. Via the Privacy Rule, the main goal is to • Ensure that individuals' health information is properly protected while allowing the flow of health information needed to provide and promote high quality health care and to protect the public's health and well-being. 	• Every healthcare provider who electronically transmits health information in connection with certain transactions • Health plans • Healthcare clearinghouses • Business associates that act on behalf of a covered entity, including claims processing, data analysis, utilization review, and billing	**Protected Health Information[2]:** Individually identifiable health information that is transmitted or maintained in any form or medium (electronic, oral, or paper) by a covered entity or its business associates, excluding certain educational and employment records	• To the individual • Treatment, payment, and healthcare operations • Uses and disclosures with opportunity to agree or object by asking the individual or giving opportunity to agree or object • Incident to an otherwise permitted use and disclosure • Public interest and benefit activities (e.g., public health activities, victims of abuse or neglect, decedents, research, law enforcement purposes, serious threat to health and safety) • Limited dataset for the purposes of research, public health, or healthcare operations

1. **Permitted disclosures** mean the information can be, but is not required to be, shared without individual authorization.
2. **Protected health information** or **individually identifiable health information** includes demographic information collected from an individual and 1) is created or received by a healthcare provider, health plan, employer, or healthcare clearinghouse and 2) relates to the past, present, or future physical or mental health or condition of an individual; the provision of healthcare to an individual; or the past, present, or future payment for the provision of healthcare to an individual; and
   (i) That identifies the individual, or
   (ii) With respect to which there is a reasonable basis to believe the information can be used to identify the individual.

For more information, please visit the Department of Health and Human Services' HIPAA website and the Department of Education's FERPA website.

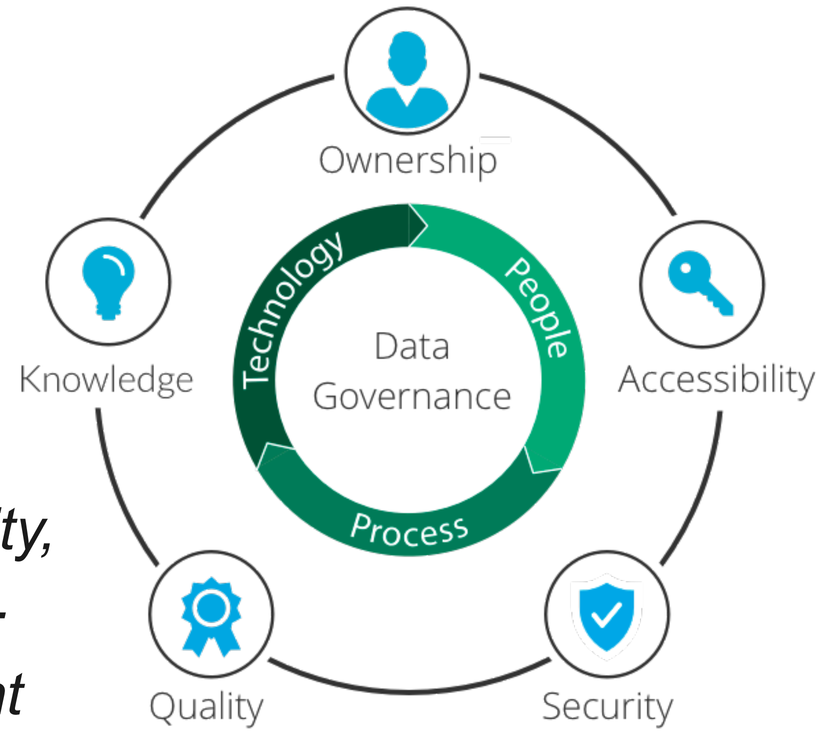
A word on data governance

Some data requires governance, such as sensitive, organizational, etc.

"Data governance is the practice of identifying important data across an organization, ensuring it is of high quality, and improving its value to the business.

A data governance policy is a document that formally outlines how organizational data will be managed and controlled."

https://www.imperva.com/learn/data-security/data-governance/



Data Sharing

Data governance policies inform data sharing work for research

If you share data with another investigator outside of your institution a data use agreement or data transfer agreement is generally required

Your institution can advise on how to draft and execute those agreements.

Generally your signing official will sign off on them

Next Generation Tools

Some next generation tools to be aware of:

- **Jupyter Notebooks** - The Jupyter Notebook is a powerful tool for interactively developing and presenting data science projects. - https://www.dataquest.io/blog/jupyter-notebook-tutorial/
- **GitHub** – GitHub is a powerful tool for collaboratively created and maintained code or data projects
- **Slack** – Collaborative chat and sharing for teams
- **Google Drive/MS Teams/etc** – Collaborative everything

The Future

Data is becoming more collaborative and shared. In order to have an impact it has to be more FAIR

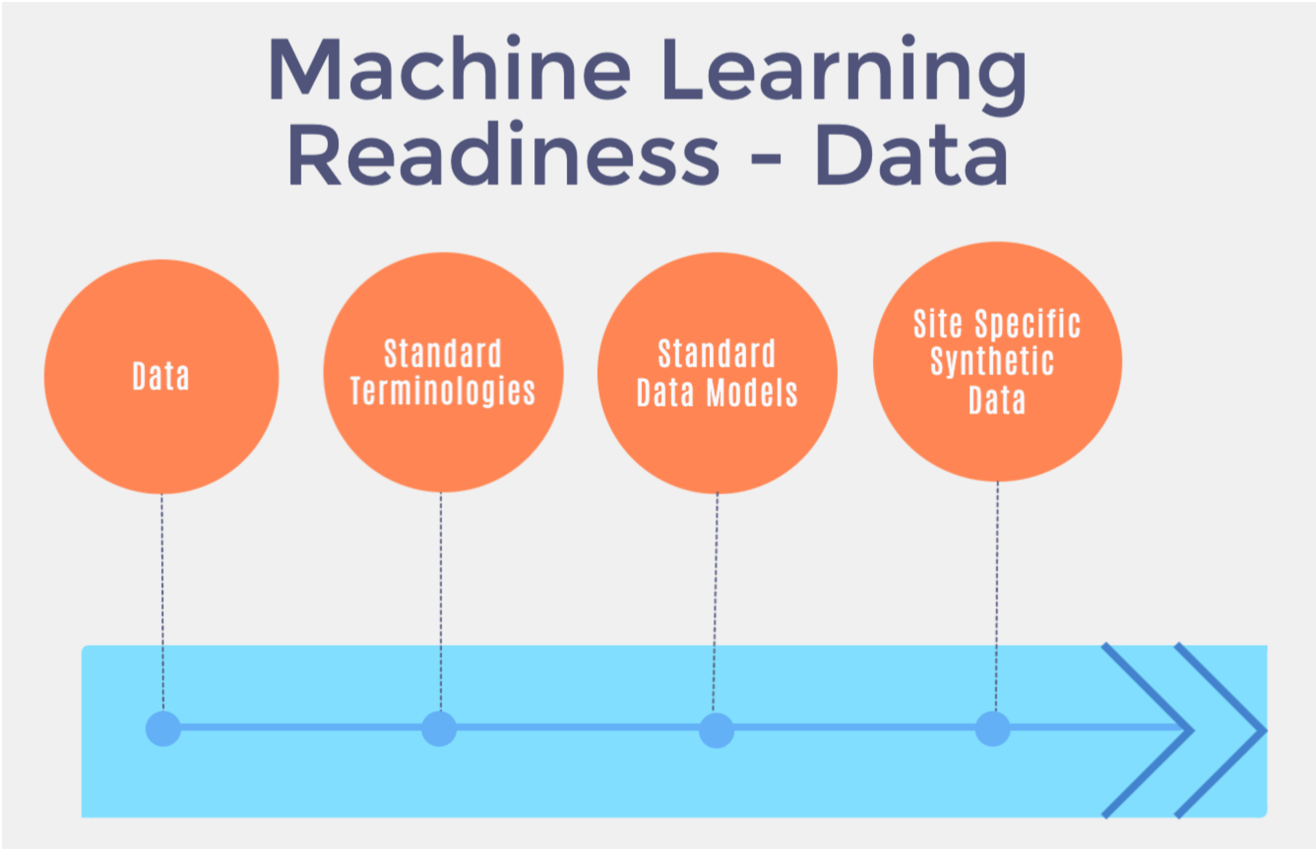
Use of standards, such as ontologies or common data models, is becoming ever important in data collection and use

Every field is different so you have to learn the standard approaches for you

Sandboxes (or enclaves) are being more widely used and won't share data in a traditional way

Using Data from Electronic Health Records

Over time, real world data from EHRs has become more ready to use for research

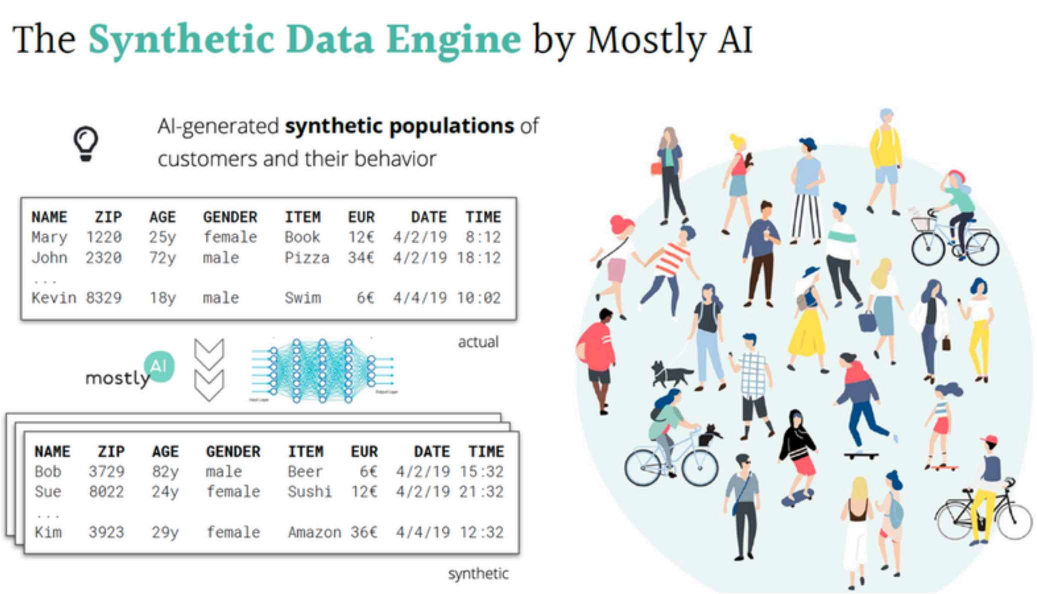


A word on synthetic data

EHR data is increasingly being distributed as a synthetic extract.

Synthetic data:

- Shares the properties of real data
- Does not contain any real patients
- Synthetic patients will 'look' as much like real patients without any aspects being mappable to any real patient
- Protects privacy

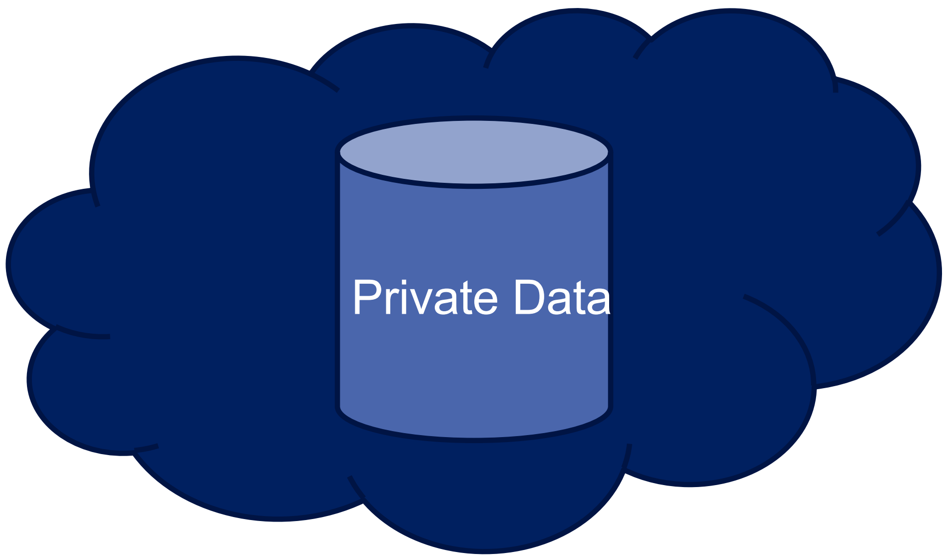


https://www.plugandplaytechcenter.com/resources/synthetic-data-solution-bypassing-headache-data-privacy/

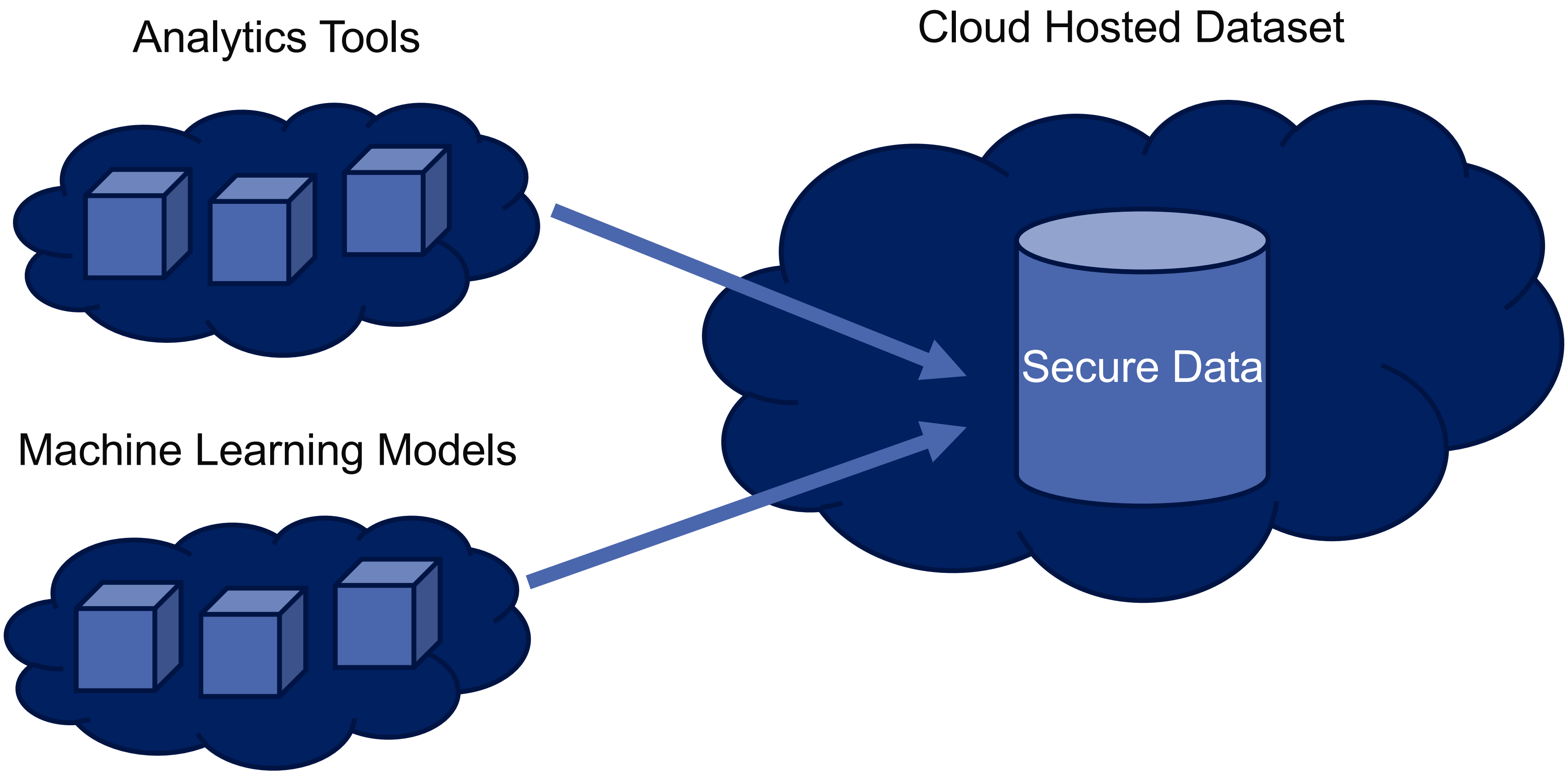
In the future we are going to see more approaches that involved ‘sandboxes’

The cloud is enabling the creation of analytics sandboxes or enclaves, where data is accessed and analyzed but not shared

Cloud Hosted Dataset

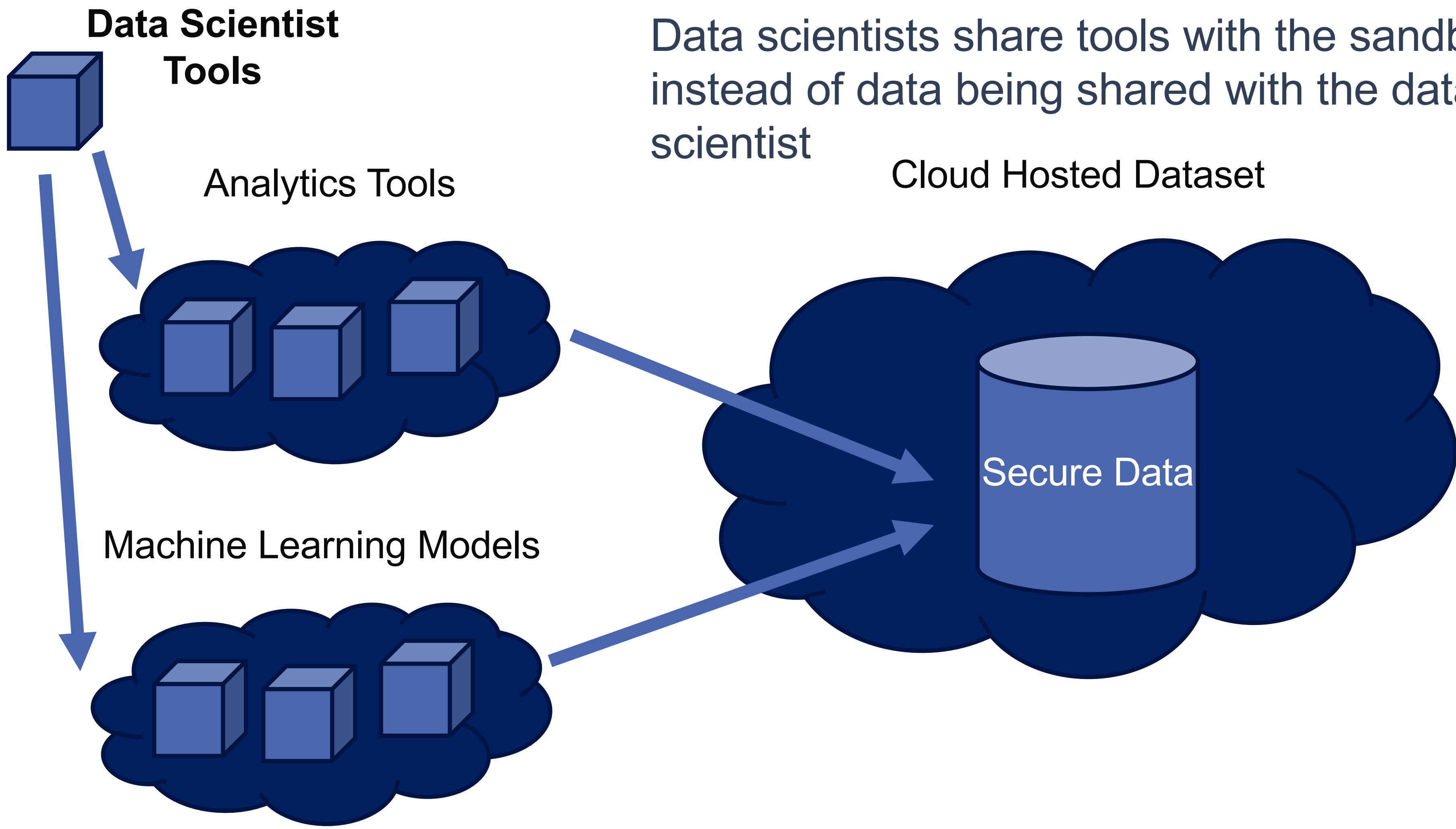


Using the cloud, analytics can be arbitrarily applied to that data



Tools are shared with data instead of data being shared

Data scientists share tools with the sandbox instead of data being shared with the data scientist



Examples of use of Sandboxes in Research

There are a growing number of use of sandboxes or secure enclaves for data analysis:

- All of Us research program
- N3C Covid-19 research collaborative
- Model To Data 'DREAM' Data challenges

Global Unique Identifiers

Another future innovation being rolled out in clinical research is GUIDs – Global Unique Patient/Research Participant Identifiers

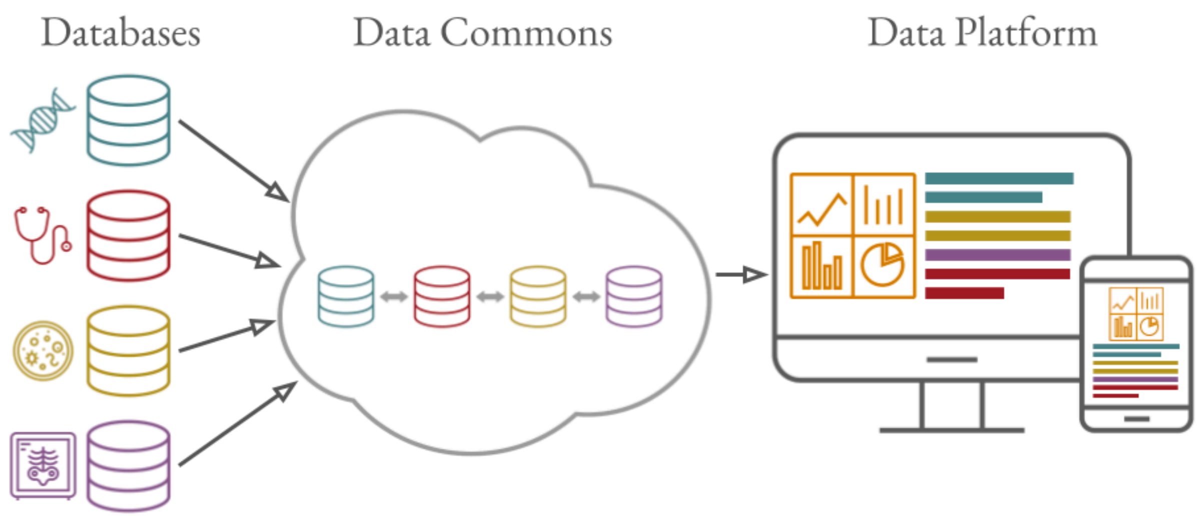
Usually a unique identifier (number, string of letters/numbers) that maps uniquely back to an individual. Enables linking datasets by knowing whether a participant participated in either project

Useful for:

- Removing duplication
- Building aggregate datasets
- Doesn't require PHI for sharing

The Rise of ‘Data Commons’

Data Commons are places where data of a particular interest can be contributed and accessed – leverages many of the technologies we just described



https://commons.cri.uchicago.edu/pcdc/

Advice

Some advice

- Spend time with data curation and think about how data will be used in the future
- Do not under resource data efforts
- Collaborate with an biomedical informaticist
- Understand that high value research datasets may not be ‘downloadable’ but are still able to be used

Questions?

Thank You!

ITHS | Institute of Translational Health Sciences
ACCELERATING RESEARCH. IMPROVING HEALTH.

Feedback Survey

A link to the feedback survey has been sent to the email address you used to register.

Please get out your device, find that email, and spend a few moments completing that survey before you leave today.

Tip: If on a mobile device, shift view to landscape view (sideways) for better user experience.