National Institute of Health ‘All of Us’ Research Program

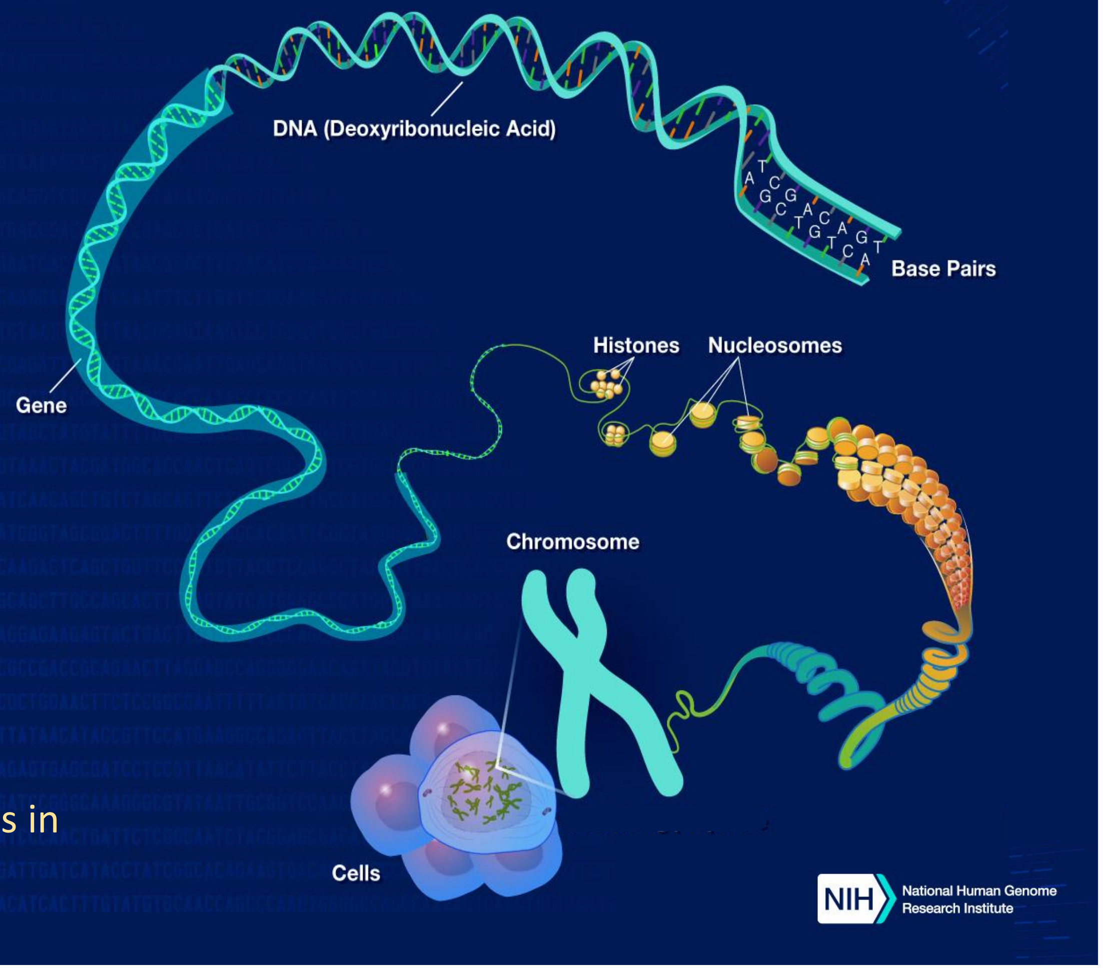
Alexander Simon

DATA607 Data Science in Context Presentation
March 27, 2024

Genomics is the study of the code of life

- DNA is composed of A-T and C-G base pairs similar to 0s and 1s in a computer
- Everyone has a unique DNA sequence

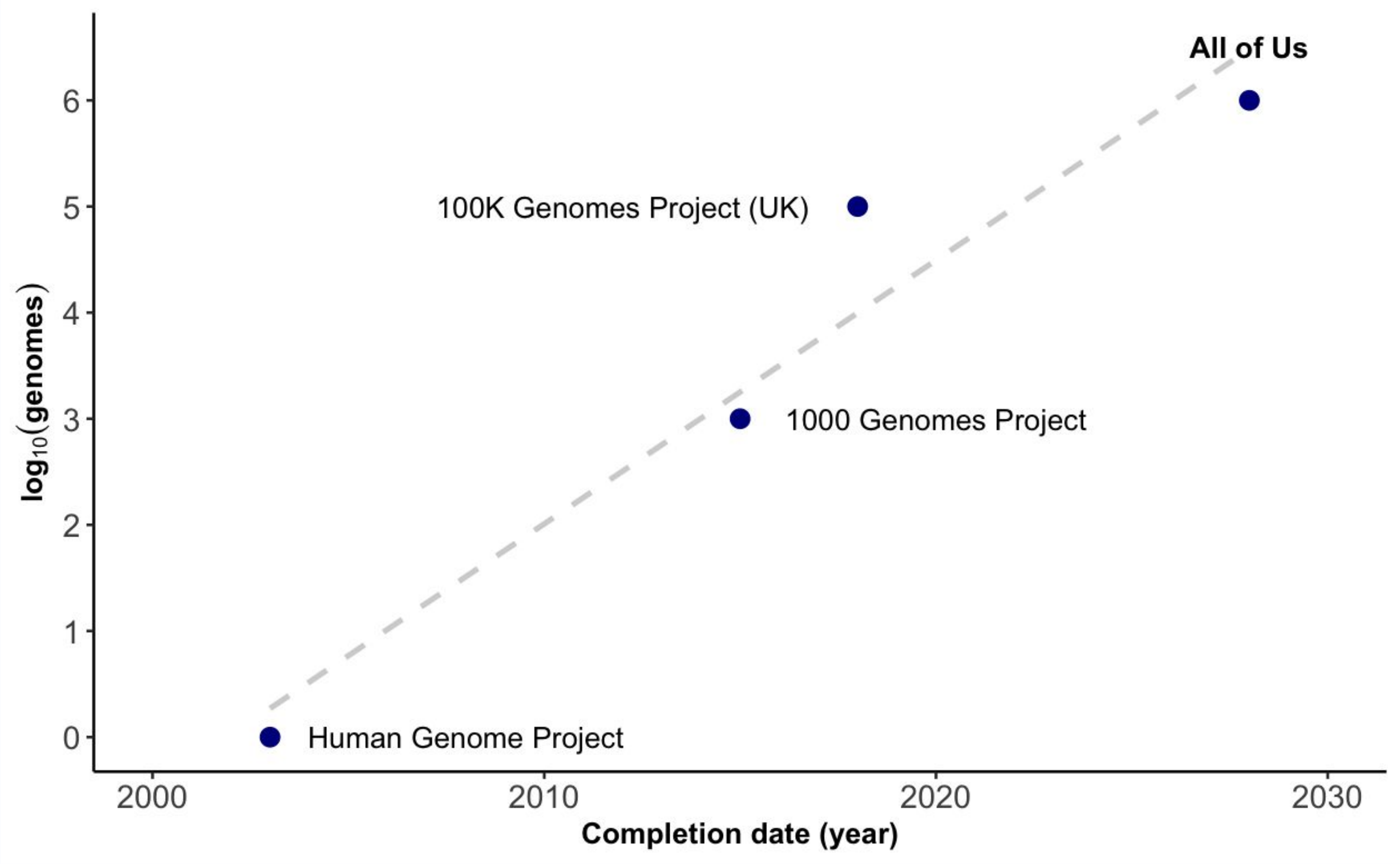
~


DNA (Deoxyribonucleic Acid)
Base Pairs
Gene
Histones
Nucleosomes
Chromosome
Cells
NIH National Human Genome Research Institute

- Errors in the DNA sequence (ie, "bugs") can be fatal or cause disease
- 'All of Us' aims to understand differences in our genomes and improve medicine

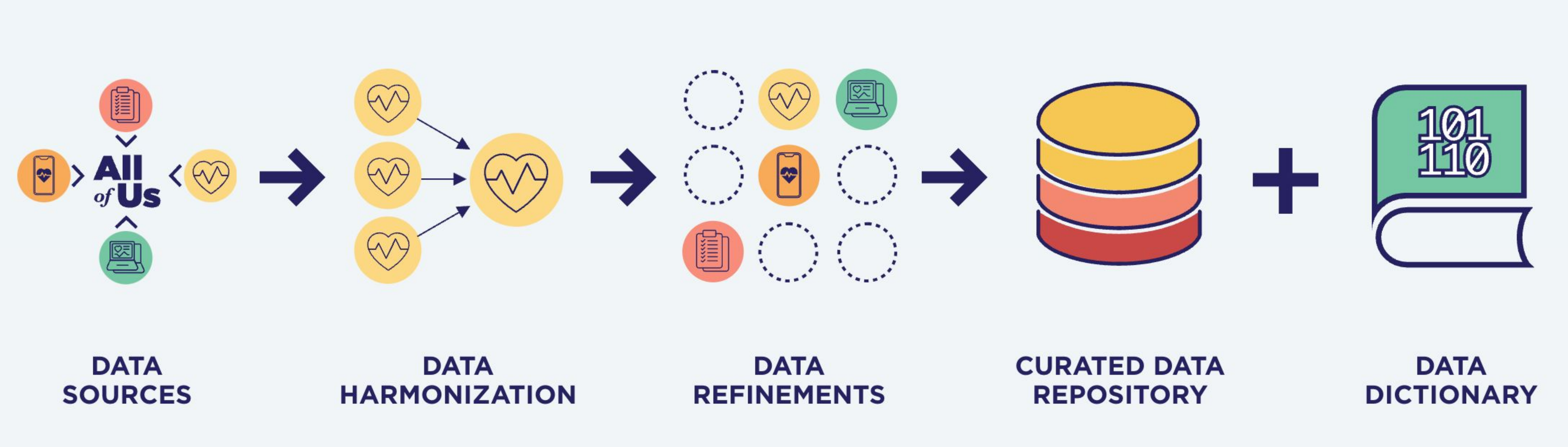
Source: www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics

Increasing scale of genomics projects



Abbreviation: UK, United Kingdom.
Data from www.yourgenome.org/theme/timeline-history-of-genomics/, allofus.nih.gov/about/program-overview/what-makes-all-us-different. Plotted with ggplot2.

Data curation



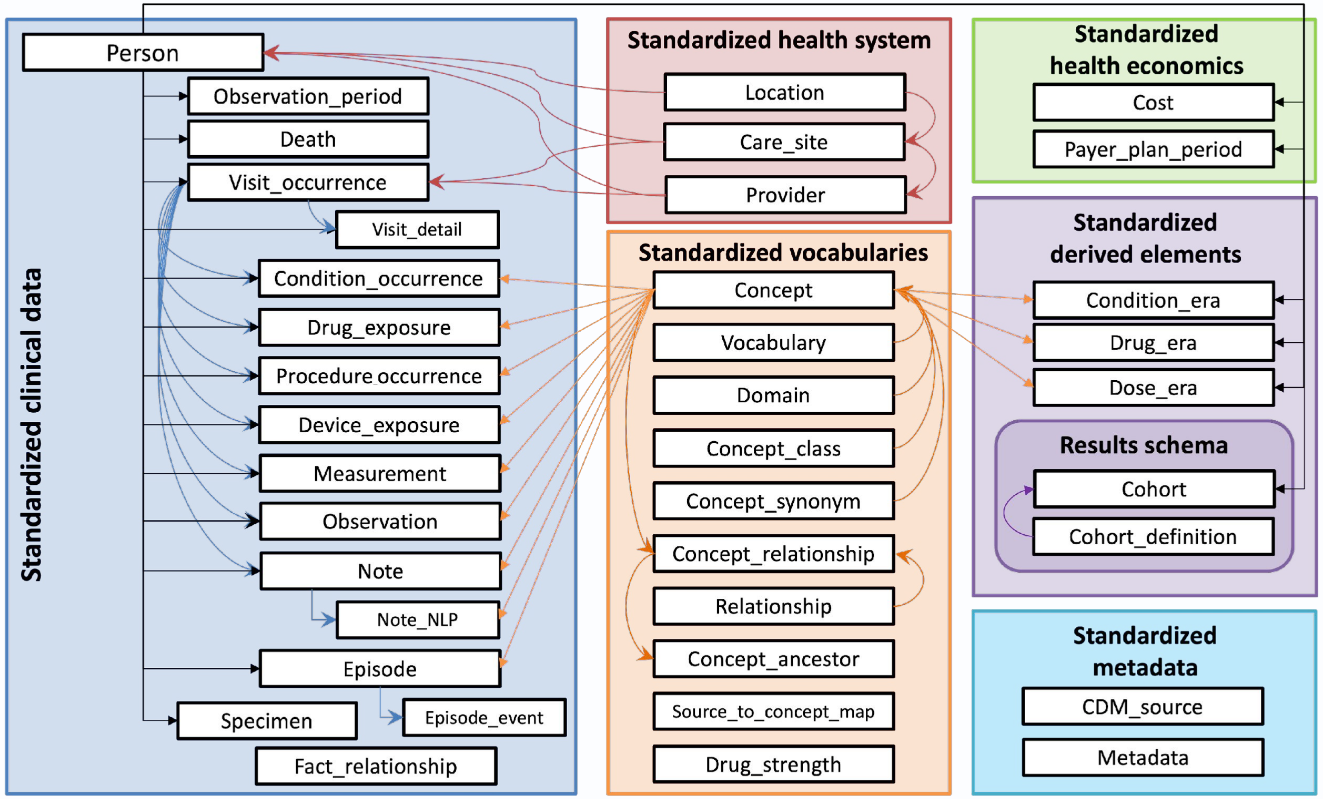
DATA SOURCES → DATA HARMONIZATION → DATA REFINEMENTS → CURATED DATA REPOSITORY + DATA DICTIONARY

- Data sources include patient surveys, electronic health records, genomic sequences
- Data are harmonized using a common data model

Source: www.researchallofus.org/data-tools/methods/

Observational Medical Outcomes Partnership (OMOP) Common Data Model

- Open community standard for observational data from hospitals and insurance providers
  - ○ Common format (data model)
  - ○ Common representation (standardized vocabulary)
  - ○ Library of standard analytic tools
- Implemented in SQL, R, and Python



OMOP entity-relationship diagram (excerpt)

Sources: www.researchallofus.org/data-tools/methods/, ohdsi.github.io/CommonDataModel/

OMOP Data Harmonization Process

Quantitative data quality analysis

Data quality checks

Qualitative data quality analysis

ETL-process

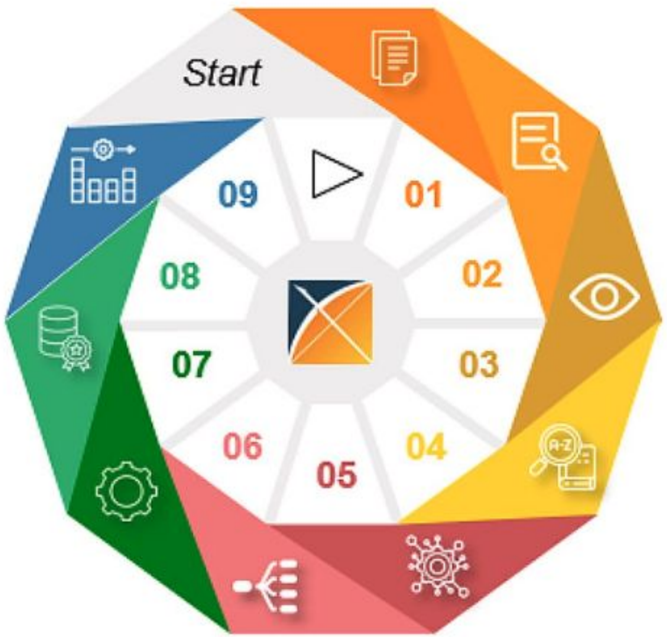
Save OMOP data in target database

Structural mapping

Convert source format to OMOP CDM

Semantic mapping

Map source vocabulary to OMOP CDM



Dataset specification

Define scope of source data for use(s)

Data profiling

Assess source data structure, format

Vocabulary identification

Coverage analysis of vocabularies

Assess extent that source vocabulary already exists in OMOP; identify missing vocabularies

Abbreviations: CDM, common data model; ETL, extract-transform-load; OMOP, Observational Medical Outcomes Partnership.
Image from Henke E., et al. Conceptual design of a generic data harmonization process for OMOP common data model. *BMC Med Inform Decis Mak*. 2024;24:58.
https://pubmed.ncbi.nlm.nih.gov/38408983/

Questions?

For more information about the ‘All of Us’ research program, visit
https://www.joinallofus.org/

Presentation slides available at
https://docs.google.com/presentation/d/1FdJXqmF4aEVTjSECkeu9g4FVEqs-KG5kJnWCVd5ZaWA/edit?usp=sharing