Harmonization of data syntax and semantics for large-scale translational research

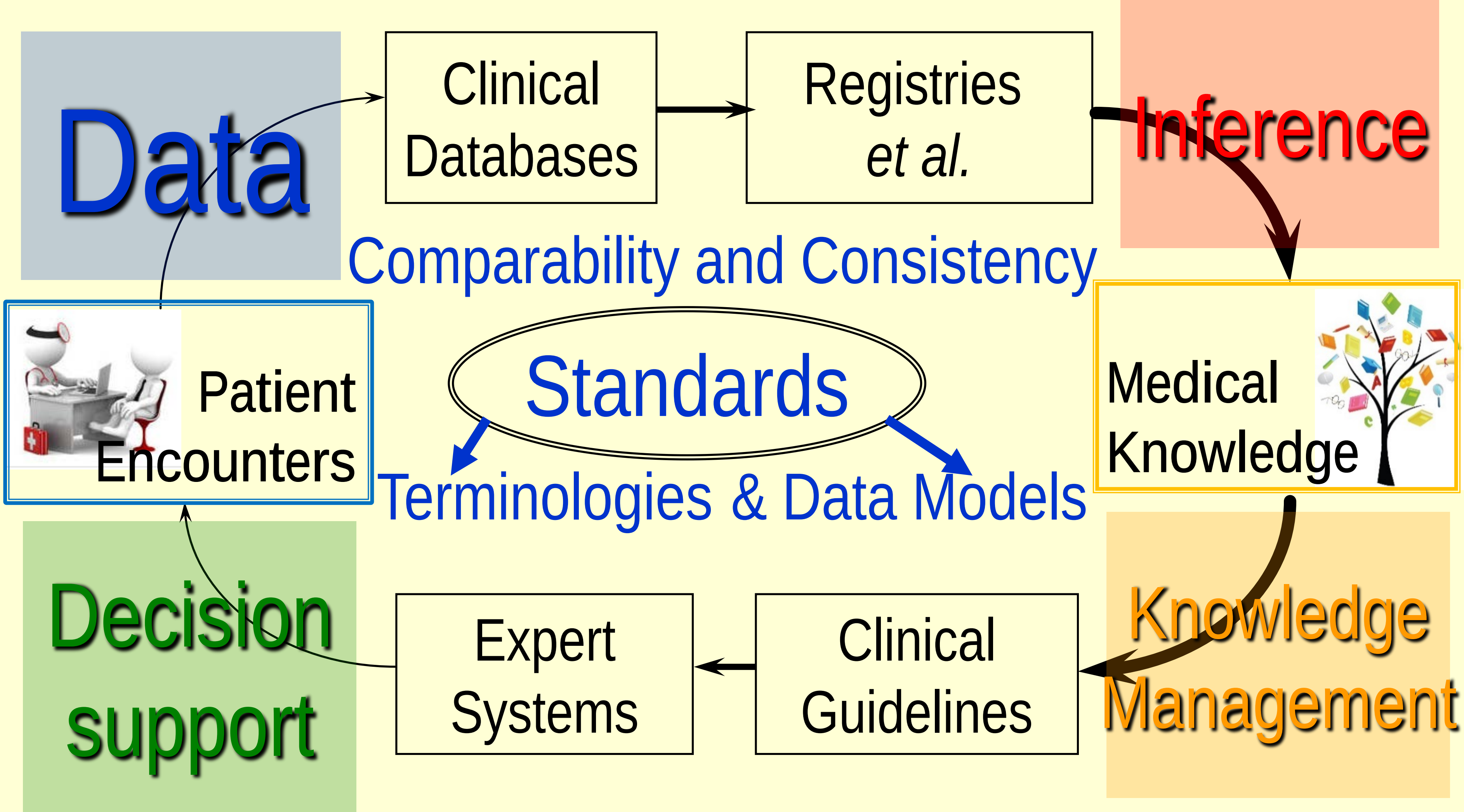
Why worry about clinical data comparability and consistency, and how to fix it

Christopher G. Chute, MD DrPH
Bloomberg Distinguished Professor of Health Informatics
Professor of Medicine, Public Health, and Nursing
Chief Health Research Information Officer
Deputy Director, Institute for Clinical and Translational Research
Johns Hopkins University, Baltimore, MD, USA

Genomic Medicine XI: Implementation

La Jolla, 6 Sept 2018

From Practice-based Evidence to Evidence-based Practice



Data

Clinical Databases

Registries *et al.*

Inference

Comparability and Consistency

Patient Encounters

Standards

Terminologies & Data Models

Medical Knowledge

Decision support

Expert Systems

Clinical Guidelines

Knowledge Management

Foundations for Learning Health System

Precision Medicine
The same, but more so.

- PM requires data and knowledge
- The questions one may need to ask are unknown
- The sources of data are heterogeneous
- The patients are individuals, though can be considered as “small homogeneous groups”
- How to assemble data into *comparable and consistent* format *is the challenge*
- Analytics is, relatively, the easy part

Genotype to Phenotype

- Genomic data quality and reproducibility
  - Well recognized principle
  - Subject of resources and effort
- Clinical data quality and reproducibility
  - More challenging, non-protocol, opportunistic
  - Data quality efforts established for Quality Metrics
- Rational focus for research secondary use of Clinical Data
  - *Comparability and consistency*

Comparable and Consistent Clinical Data

Two options:

- Map what you have to what you need
  - Hopelessly tangled spaghetti
  - Redundant and non-scalable work
- Embrace a “common data model” (CDM)
  - Map what you have to the CDM
  - Define canonical form
  - Preferentially conduct research analyses using mutually agreed upon CDM format

CDM Nirvana
(once chosen and adopted)

- Clear hub and spoke harmonization
  - Canonical hub
- Map once, use many
  - Obviates redundant work
- Data creation is CDM semantics where practical
- Defines practical data interoperability

CDM Hades

- Happy to use CDM, as long as it is mine
- Lets agree to map among CDMs
  - Oxymoron of CDM plurality
- I am going to “extend” the CDM for my use case
  - Everybody making non-comparable extensions
- I am going to make a new CDM for my use case
- I am going to change the CDM for my use case

- Recipe for non-interoperability

Which CDM?
High Profile *Research* CDMs

- Sentinel – FDA surveillance for adverse events
  - Derived from health services research tact
  - Emphasized administrative data
- PCORNet CDM
  - Adaption of Sentinal; clinically brittle
- ACT – CTSA shared model (i2b2 adaptation)
- OMOP/OHDSI – Pharma initiated
  - Focus on large population questions
- TriNetX – Commercial, interoperable nodes
  - Has the advantage of *working*, industry sponsored

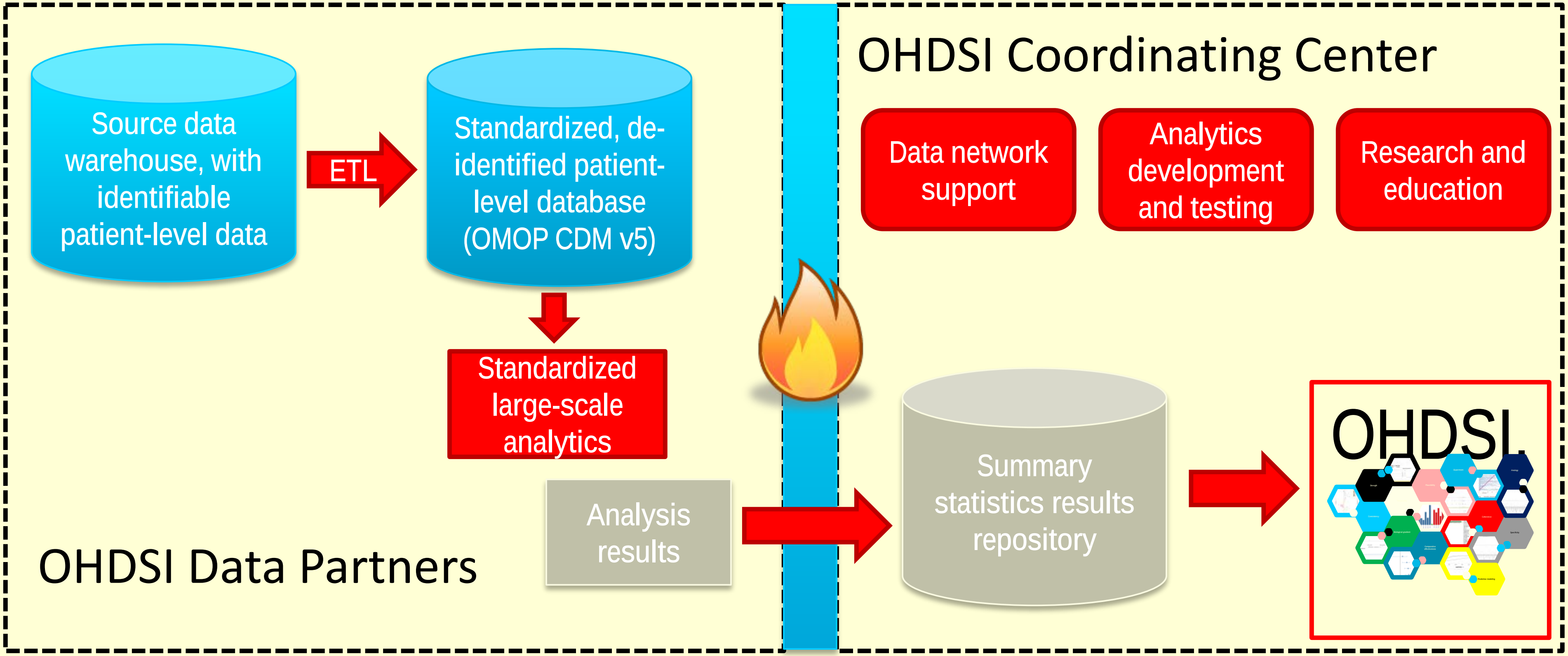
Evidence OHDSI seeks to generate from observational data

George
Hripcsak
OHDSI

- Clinical characterization = tallying
  - Natural history: Who has diabetes, and who takes metformin?
  - Quality improvement: What proportion of patients with diabetes experience complications?
- Population-level estimation = causality
  - Safety surveillance: Does metformin cause lactic acidosis?
  - Comparative effectiveness: Does metformin cause lactic acidosis more than glyburide?
- Patient-level prediction = prediction
  - Precision medicine: Given everything you know about me, if I take metformin, what is the chance I will get lactic acidosis?
  - Disease interception: Given everything you know about me, what is the chance I will develop diabetes?

How OHDSI Works

George
Hripcsak
OHDSI



OHDSI Data Partners

Source data warehouse, with identifiable patient-level data

ETL

Standardized, de-identified patient-level database (OMOP CDM v5)

Standardized large-scale analytics

Analysis results

OHDSI Coordinating Center

Data network support

Analytics development and testing

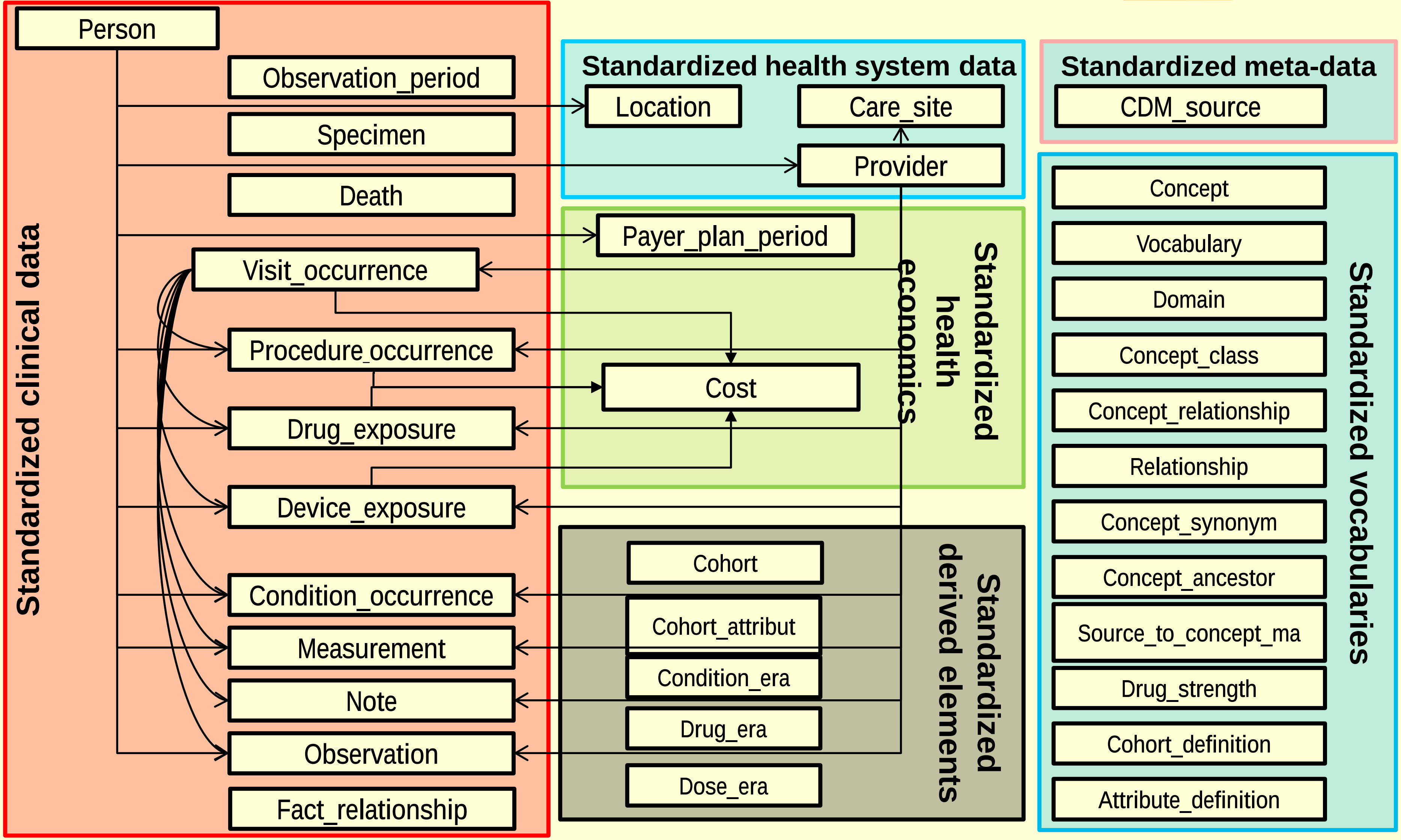
Research and education

Summary statistics results repository

JOHNS HOPKINS
MEDICINE

Deep information model
OMOP CDM v5

George
Hripcsak
OHDSI



Standardized clinical data

Person

Observation_period

Specimen

Death

Visit_occurrence

Procedure_occurrence

Drug_exposure

Device_exposure

Condition_occurrence

Measurement

Note

Observation

Fact_relationship

Standardized health system data

Location

Care_site

Provider

Standardized health economics

Payer_plan_period

Cost

Standardized derived elements

Cohort

Cohort_attribut

Condition_era

Drug_era

Dose_era

Standardized meta-data

CDM_source

Standardized vocabularies

Concept

Vocabulary

Domain

Concept_class

Concept_relationship

Relationship

Concept_synonym

Concept_ancestor

Source_to_concept_ma

Drug_strength

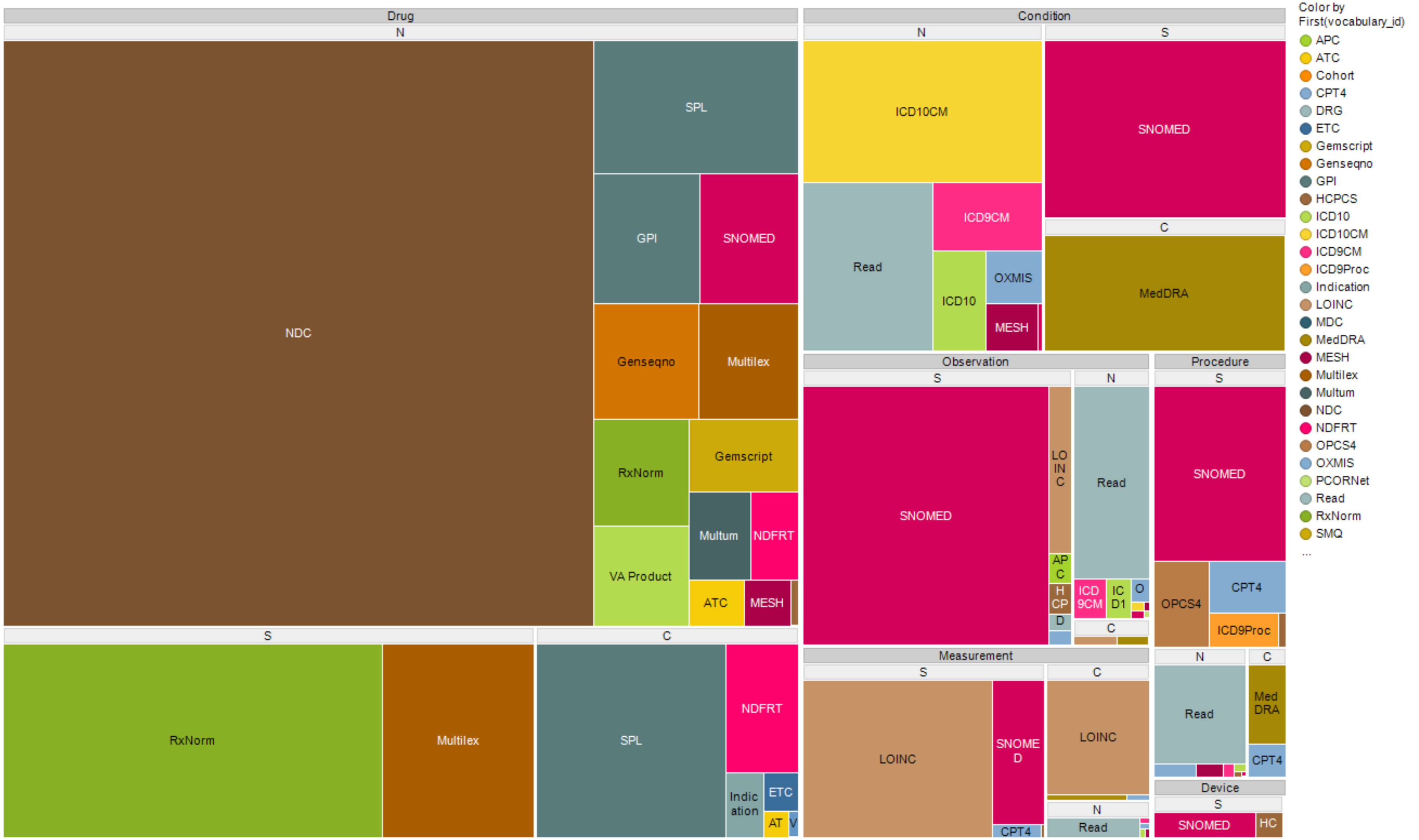
Cohort_definition

Attribute_definition

Extensive vocabularies (80)

George
Hripcsak
OHDSI

Breakdown of OHDSI concepts by domain, standard class, and vocabulary



George
Hripcsak
OHDSI

emerge network
ELECTRONIC MEDICAL RECORDS AND GENOMICS

FDA U.S. FOOD & DRUG
ADMINISTRATION

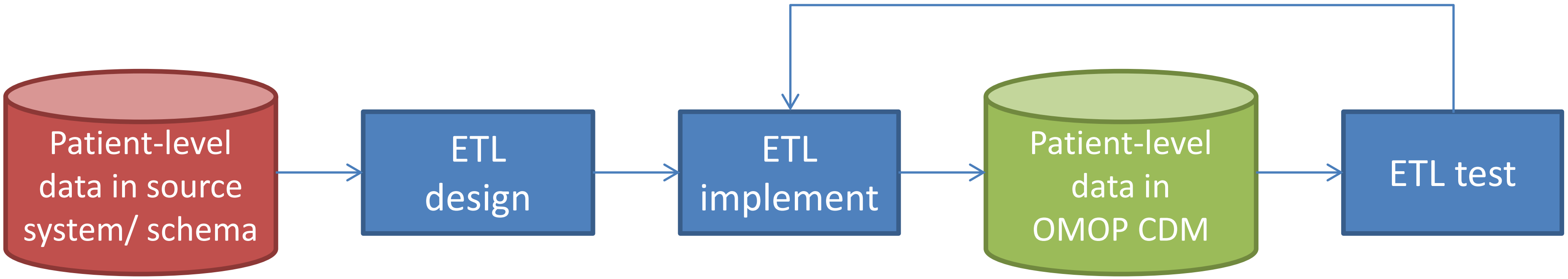
NIH National Institutes of Health
All of Us Research Program

NIH NATIONAL CANCER INSTITUTE

imi innovative
medicines
initiative

George
Hripcsak
OHDSI

Tools to convert your data



Patient-level data in source system/ schema

ETL design

ETL implement

Patient-level data in OMOP CDM

ETL test

OHDSI tools built to help

WhiteRabbit: profile your source data

RabbitInAHat: map your source structure to CDM tables and fields

ATHENA: standardized vocabularies for all CDM domains

Usagi: map your source codes to CDM vocabulary

CDM: DDL, index, constraints for Oracle, SQL Server, PostgresQL; Vocabulary tables with loading scripts

ACHILLES: profile your CDM data; review data quality assessment; explore population-level summaries

OHDSI Forums: Public discussions for OMOP CDM Implementers/developers

http://github.com/OHDSI

Large-Scale Research CDMs
Intrinsic Limitations

- Large-scale data models are inevitably optimized for specific use-cases
- Prematurely binding a model to a large-scale presumes a use-case, presumes the questions
- Orthogonal questions require serial outer-joins
    - SQL servers slow to a crawl
- The larger the model, the more brittle its reuse
- Thus, the question is: what is the *optimal size* of a canonical data model

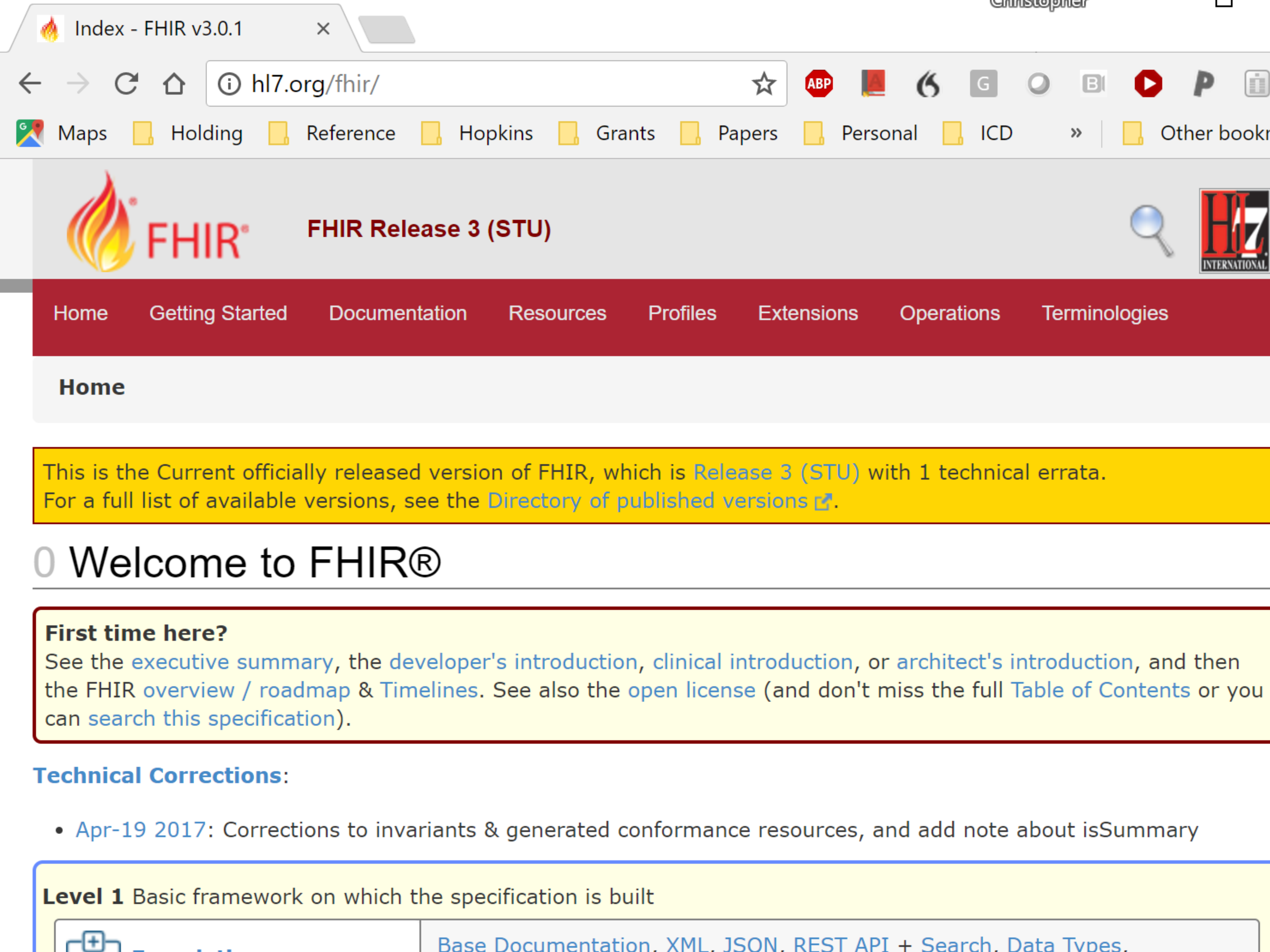
Goldilocks and the Three Data Scales

- Models that are *too small* lead to incoherency
  - At the limit is inchoate data

- Models that are *too big* lead to brittle structures that cannot efficiently address unanticipated questions
- Our previous work (SHARPn.org) suggests that the data element level is "*just right*"
  - e.g. laboratory observation, medication order, diagnostic assertion

Clinical Standards

- The clinical health information technology community has made enormous progress in the past decade
- International agreement
- Pragmatic adoption
- RESTful resources (modern IT architecture)
- *Obviates need* for research specific CDM



FHIR Release 3 (STU)

Home

This is the Current officially released version of FHIR, which is Release 3 (STU) with 1 technical errata.
For a full list of available versions, see the Directory of published versions.

0 Welcome to FHIR®

First time here?
See the executive summary, the developer's introduction, clinical introduction, or architect's introduction, and then the FHIR overview / roadmap & Timelines. See also the open license (and don't miss the full Table of Contents or you can search this specification).

Technical Corrections:

- Apr-19 2017: Corrections to invariants & generated conformance resources, and add note about isSummary

Level 1 Basic framework on which the specification is built

	Base Documentation, XML, JSON, REST API + Search, Data Types,

FHIR Resources define a Goldilocks level of Clinical Data Organization

- “Resources” are:
  - Small logically discrete units of exchang
  - Defined behaviour and meaning
  - Known identity / location
  - Smallest unit of transaction
  - “of interest” to healthcare
  - Like v2 Segments/v3 CMETs
  - 3 parts: discrete, narrative & extensions
  - 100-150 ever

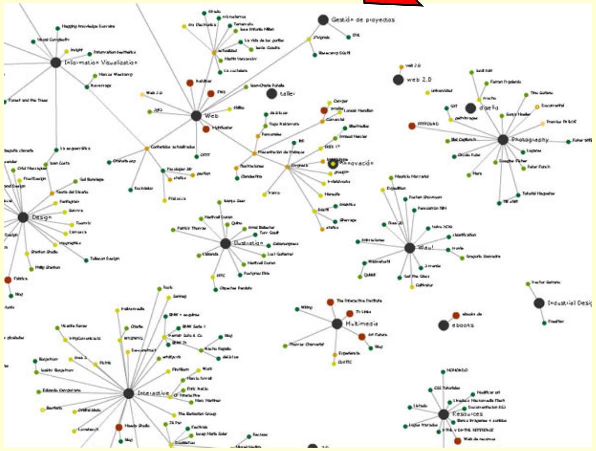
FHIR as the ultimate CDM
Right-sized Specification

FHIR Resources & CIMI Archetypes

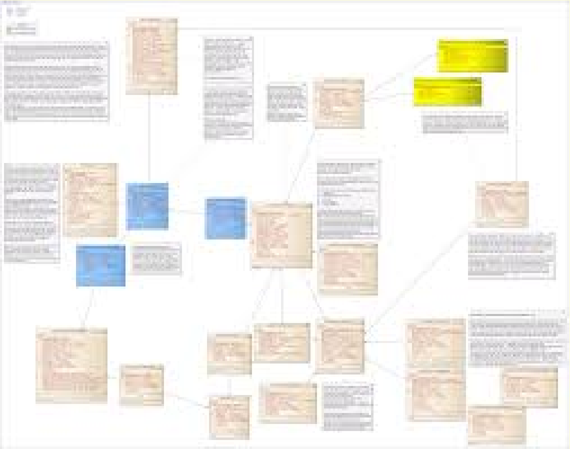
- Demographics
- Observations
- Medications
- Procedures
- …

Data Marts

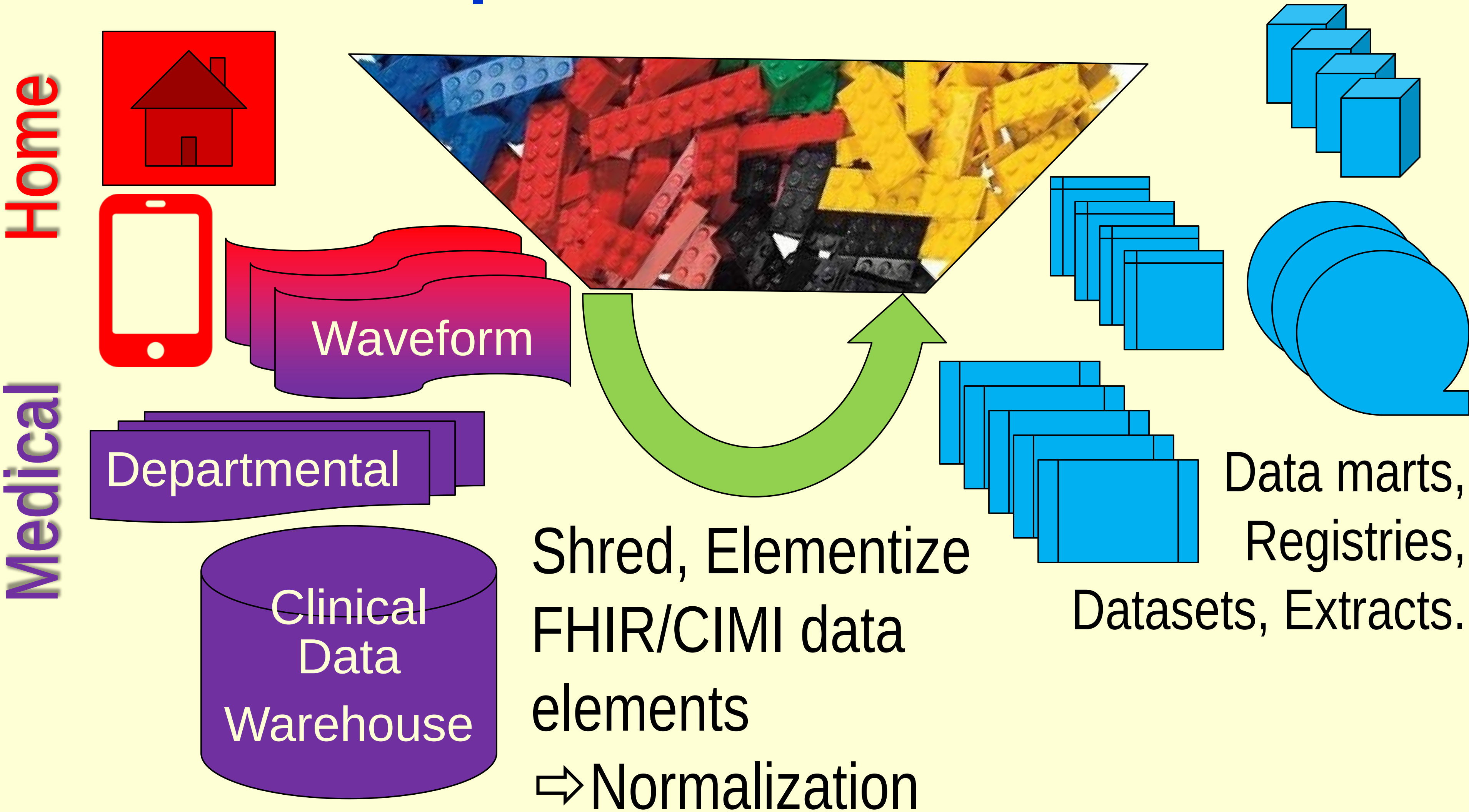
- Registries
- Protocols
- Studies
- Cohorts
- …



VS.



Pluripotent Data Model



Shred, Elementize FHIR/CIMI data elements
⇨Normalization

Data marts,
Registries,
Datasets, Extracts.

Research Adoption of FHIR

- All of Us: Synch for Science
- NCATS FDA data interoperability
- Genomic Results resource specification
- CTSA Next Generation Repository project
  - Under Center for Disease to Health (CD2H)

Where is This Going?

- Biomedical practice and research are data, information, and knowledge intensive
- Comparable and consistent data representation are pre-requisite for efficient clinical analytics
- Canonical data rendering is a prerequisite for analytics, particularly in Precision Medicine
- Data element scale models are optimal for Precision Medicine
- FHIR Resources are the obvious candidate