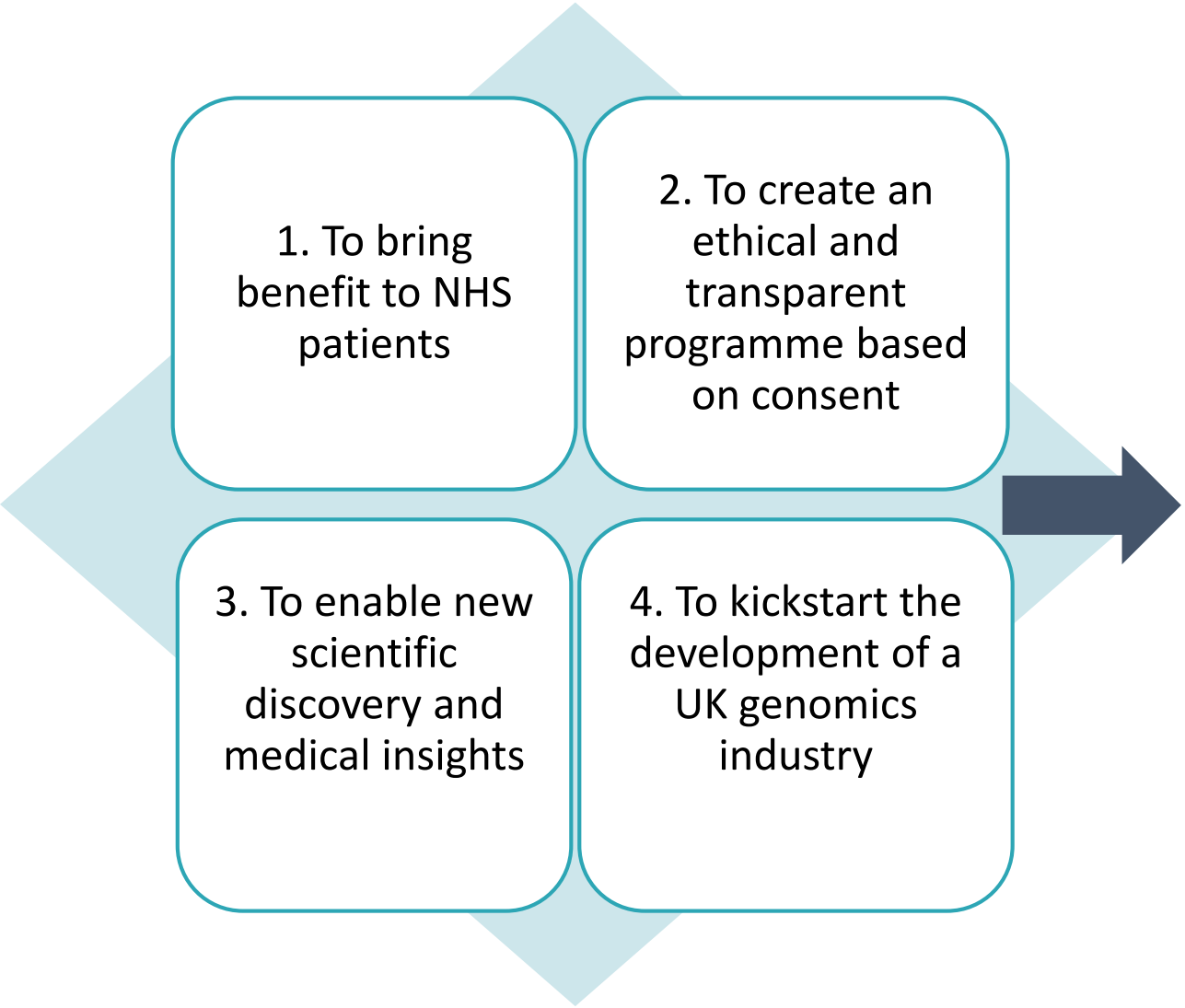
Genomics england

Accelerating Discovery

James Holman

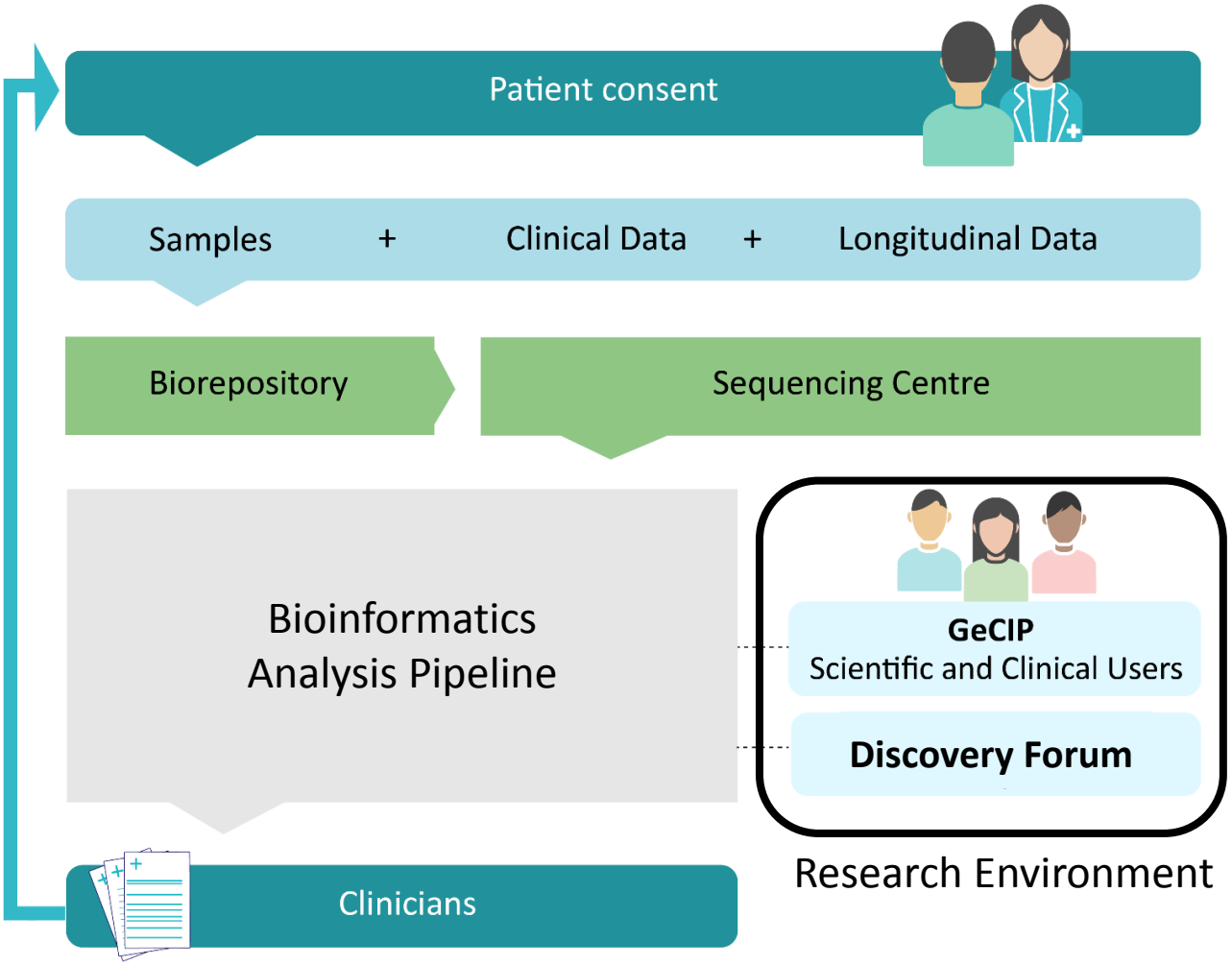
June 2019

Genomics England



- Rare diseases
- Certain cancers

How it works


Patient consent
Samples + Clinical Data + Longitudinal Data
Biorepository
Sequencing Centre
Bioinformatics Analysis Pipeline
GeCIP
Scientific and Clinical Users
Discovery Forum
Research Environment
Clinicians

Progress to date

Samples

120,500
Samples collected from NHS GMCs and sent to biorepository

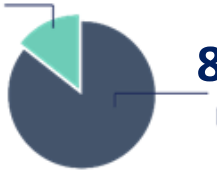
36,859 cancer

83,641
rare disease

Genomes

109,072
Genomes sequenced

24,782 cancer



84,290
rare disease

Analysis and Results

Results for **76,527** genomes sent to NHS GMCs

Research

Data release of genomes in RE **91,271**

2,424 GeCIP members with access to RE

Over **100** Discovery Forum members (13 full members)

A safe environment for research

Genomics england

Virtual Desktop — Inuvika

Managed File Transfer — GO ANYWHERE® Managed File Transfer

Data and documentation

Genomes stored in by date folders on Isilon share

Clinical data stored in LabKey

Release notes, data dictionaries, user guides.

Confluence

Tools and analysis

Terminal allows command line querying of the data

R and Rstudio allows statistical analysis of the data

Firefox browser allows access to whitelisted sites

Access to cluster and the submission node to run large scale analysis

Collaboration and workflow

Research registry – to promote collaboration and enforce publication moratorium

Rocket Chat social media, IM platform for communication

Docker to run custom software and workflows

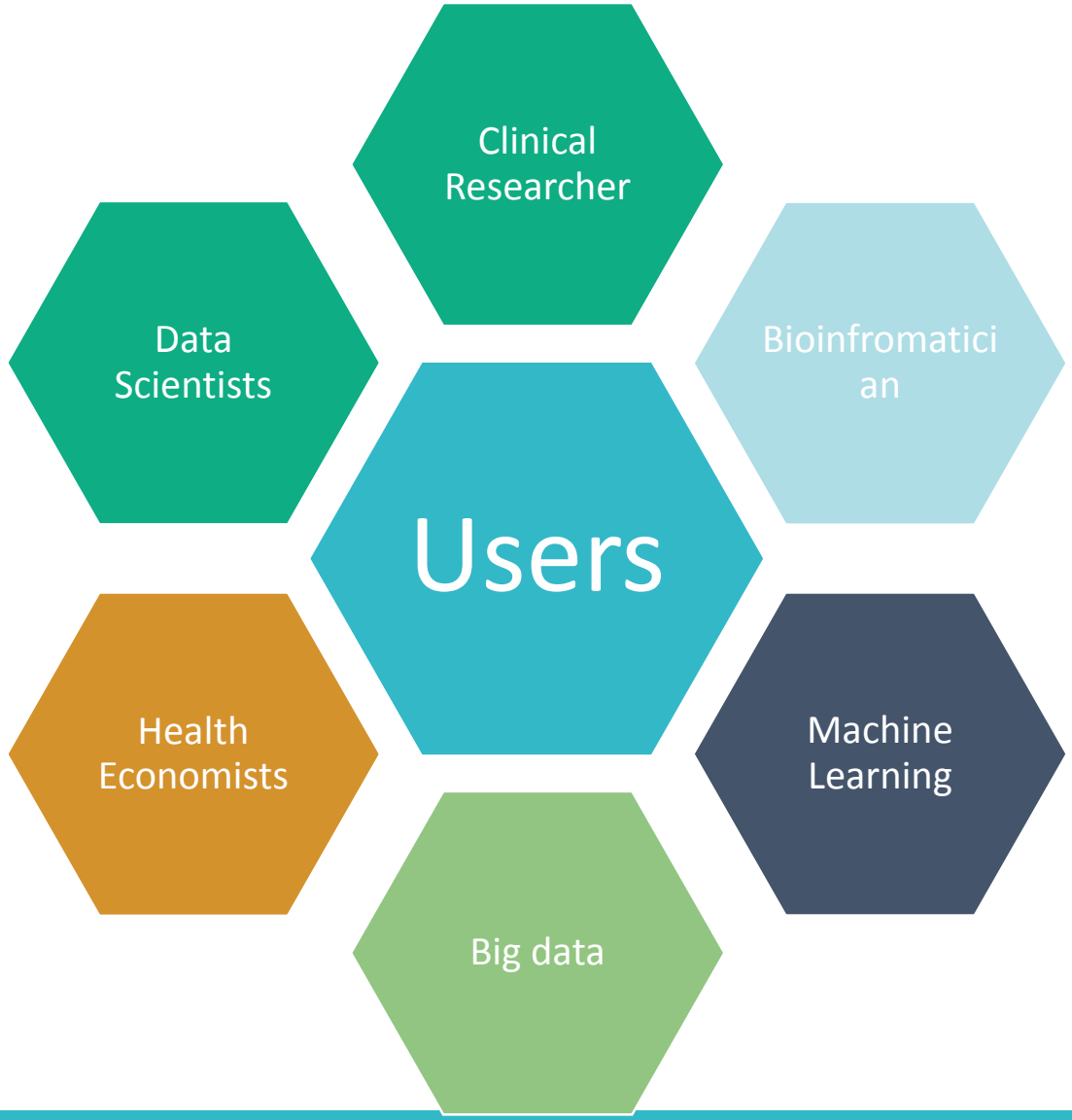
Data that's available today

Genomes	**91,271 genomes** • **22,091 Cancer** • **69,172 Rare Disease**	**Primary clinical data**	**94,285 participants** • **20,475 Cancer** • **73,810 Rare Disease**

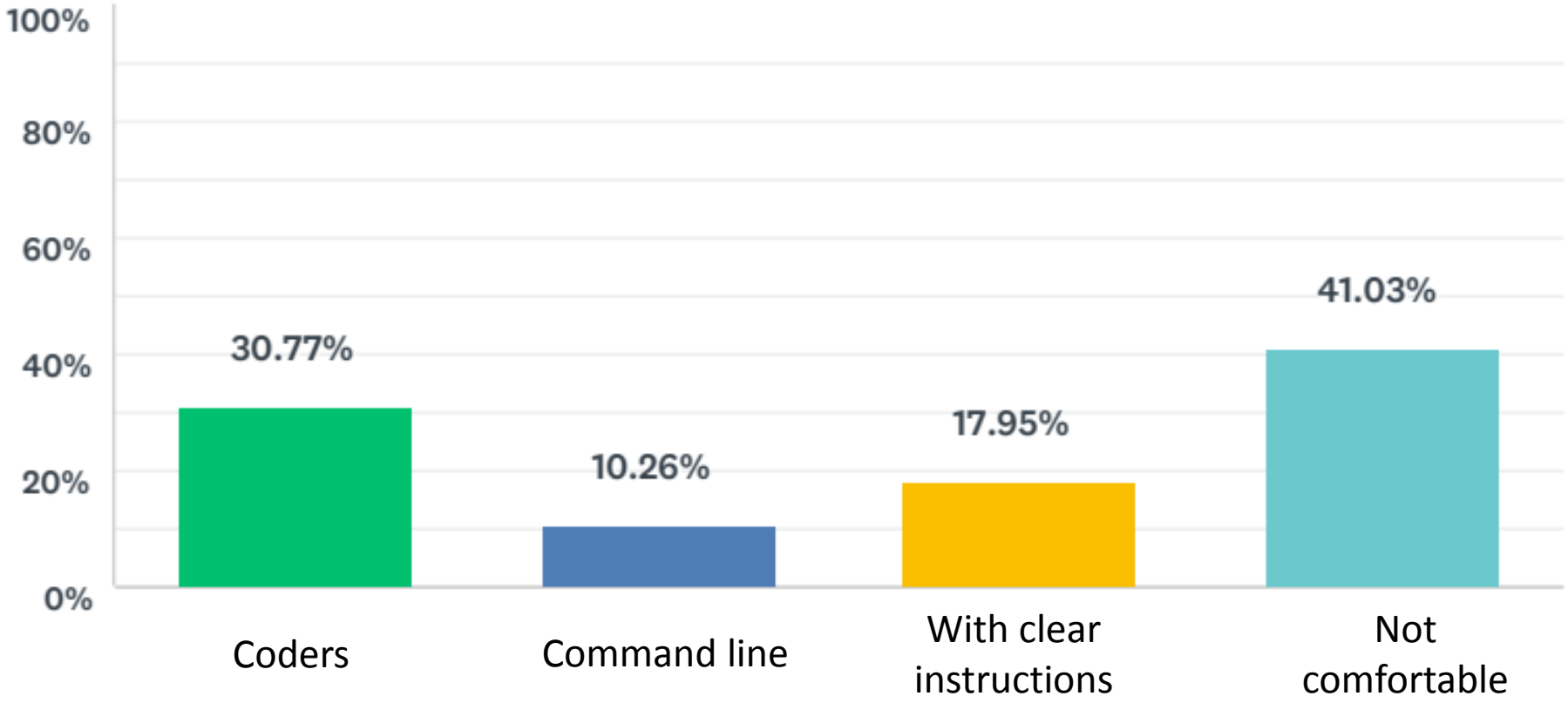
Secondary data	• Hospital Episode Statistics (HES) • Diagnostic Imaging Dataset (DID) • Patient Reported Outcome Measures (PROMs)	• Mental Health Minimum Data Set • Office for National Statistics (ONS) • Systemic Anti-Cancer Therapy Data Set (SACT)

Clinically interpreted data & QC	• **21,873 families** with Tier 1, 2 and 3 variants from interpretation pipeline • **4,763 families** with GMC exit questionnaires • **45,743** tiered and quality checked rare disease genomes; **19,098** quality checked cancer genomes	**Quick view tables**	• Key information from different tables, merged and filterable • Merged with QC data • Allow cohort-building and project feasibility assessment

Diverse needs



Programming Competency



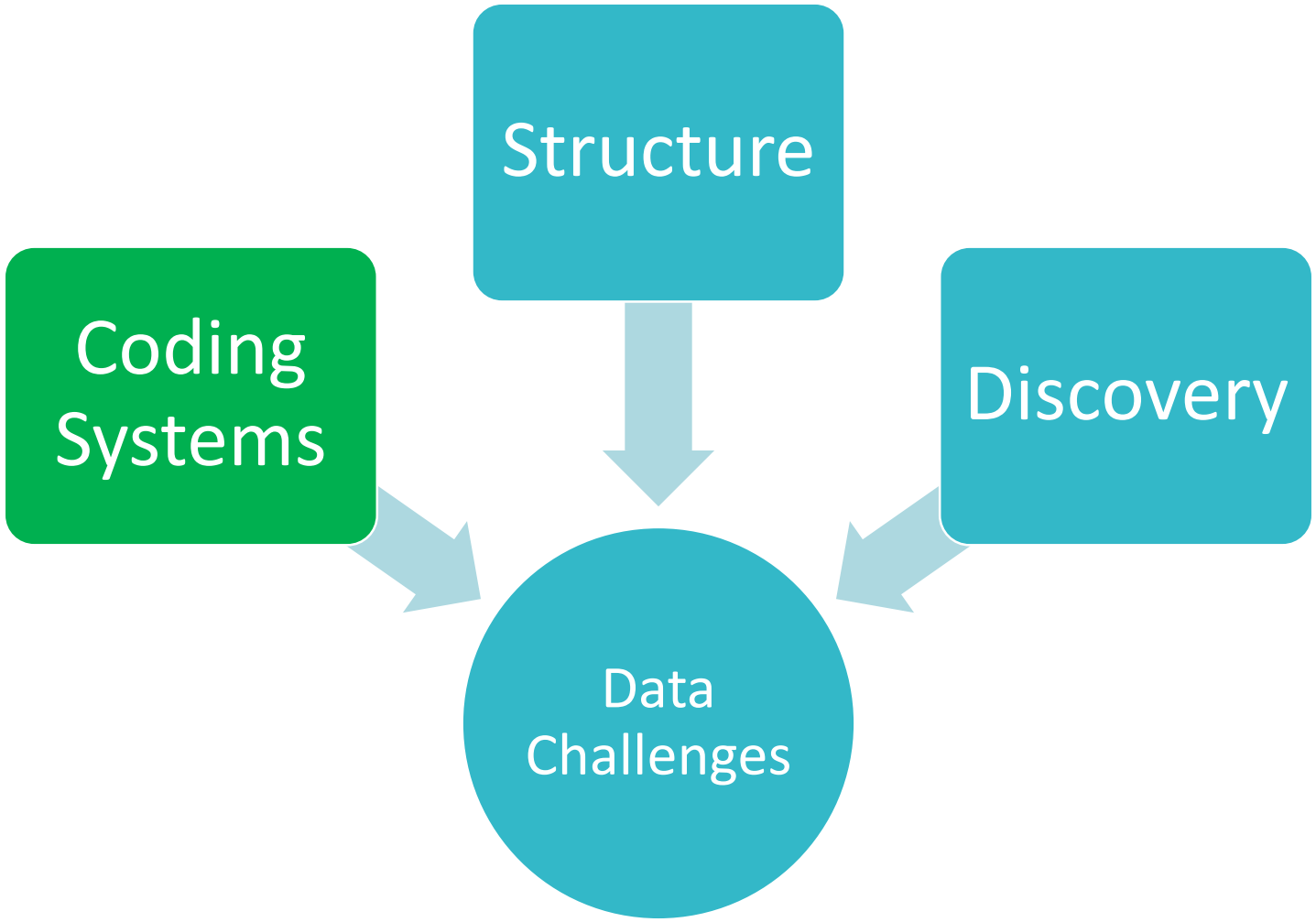
User insights

“I can’t easily see what data is available as I need to query many tables”

“its not user friendly and difficult at times to navigate the different datasets.”

“It is difficult to find the data you need in a joined up way…”

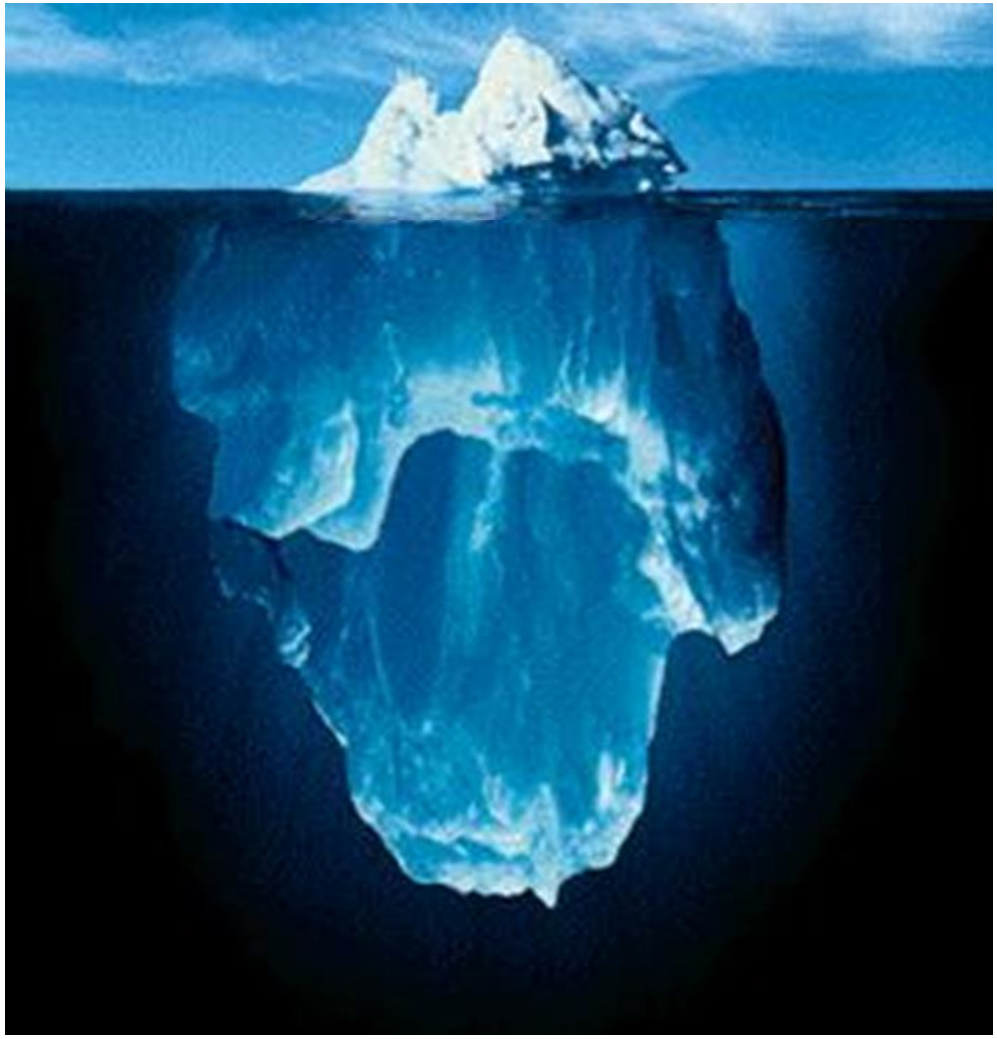
Democratise Research



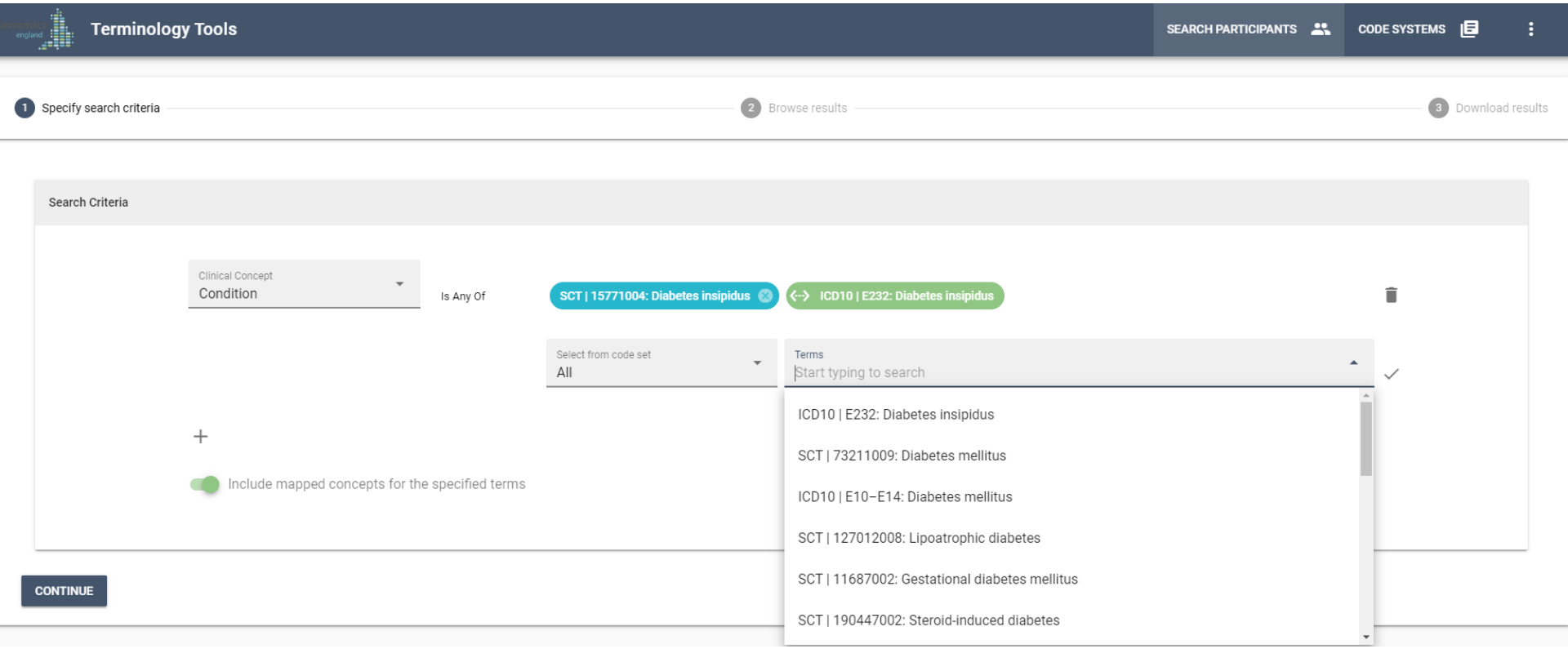
Not more codes!

- SNOMED CT
- HPO
- ICD10
- ICD-0
- OPCs
- Read
- A&E codes
- GEL Disease Name

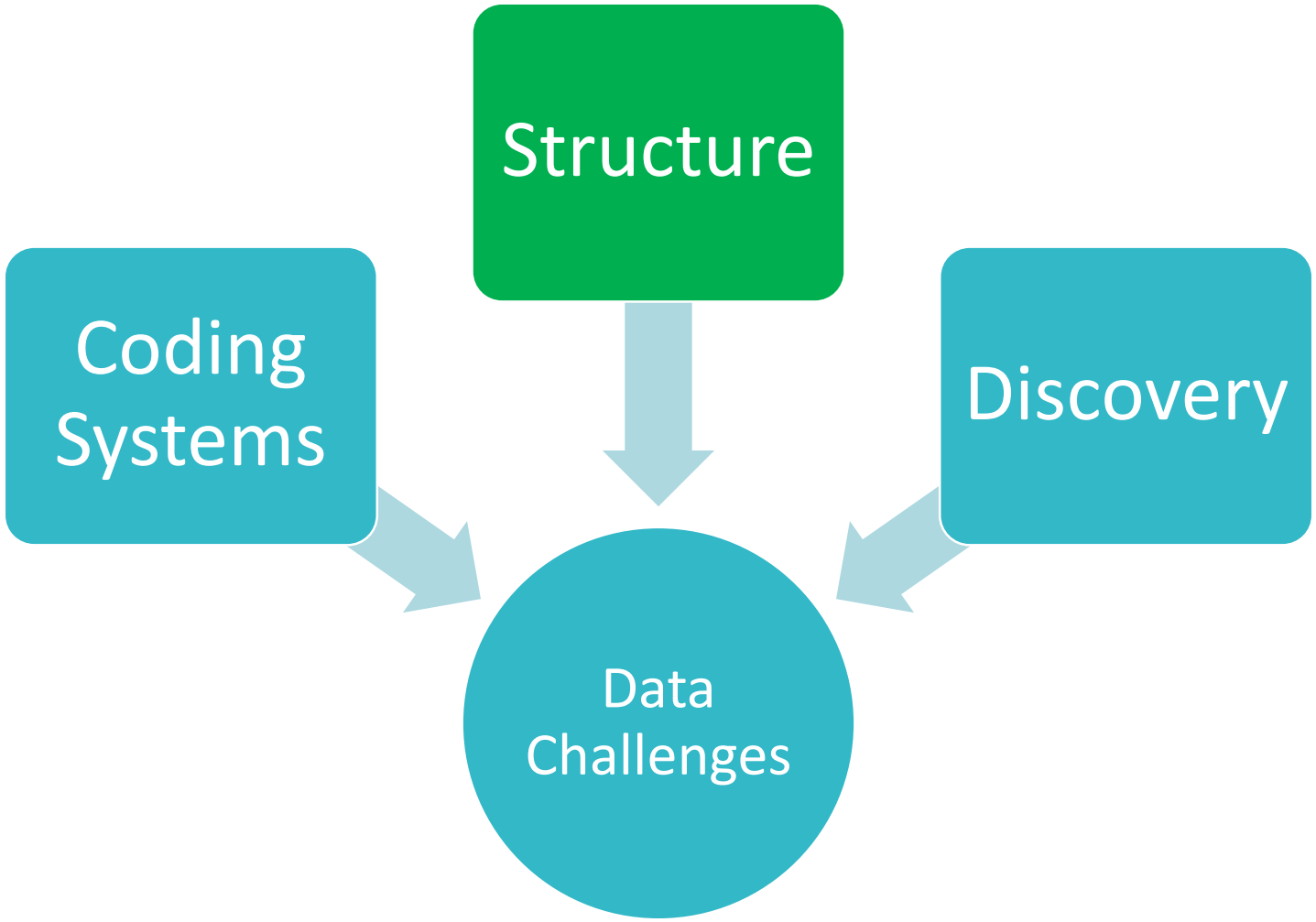
Leverage Ontologies



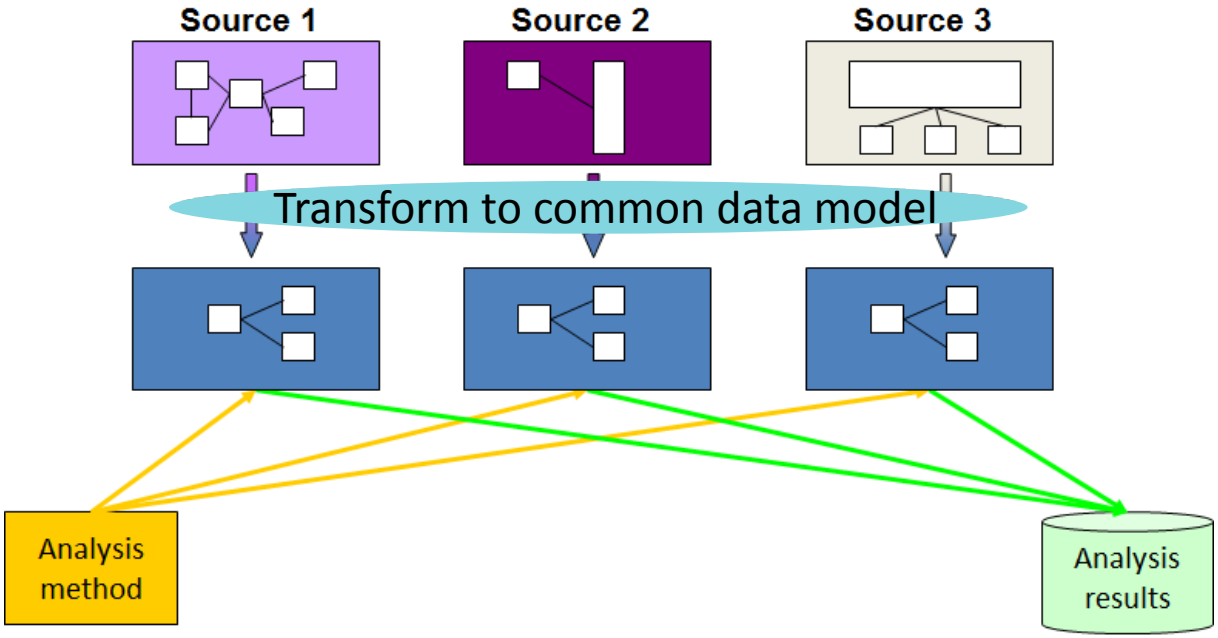
Terminology Service



Democratise Research



Common Data Model



- Application agnostic data model
- Describes the structure and data semantics in relation to the organisations business process.

CDMs - for clinical research data

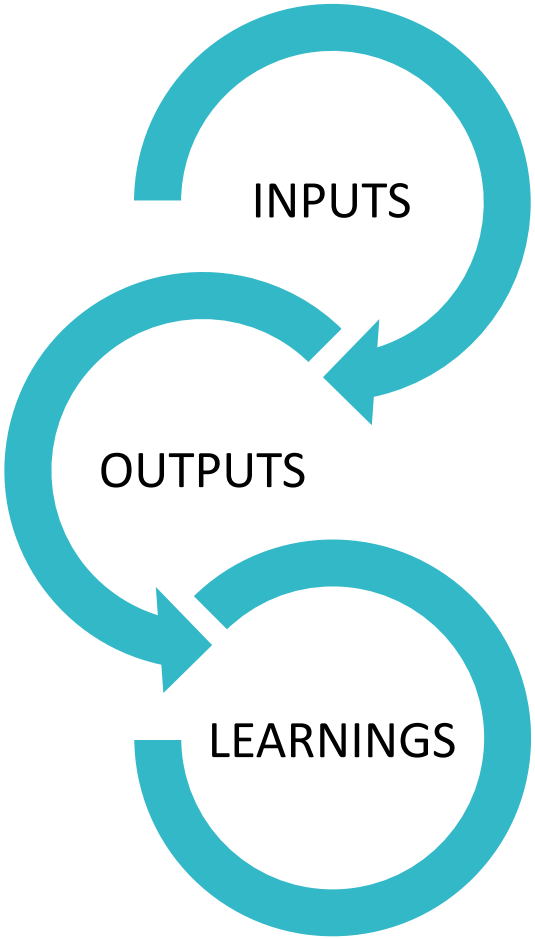
- **O**bservational **M**edical **O**utcomes **P**artnership (OMOP) Common Data Model
- Sentinel
- National Patient-Centered Clinical Research Network (PCORnet)
- Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM)
- And others.

Why a CDM?

Ensure shared, consistent understanding
Facilitate information integration
Decrease barriers to reuse

Minimize the effort to develop cohorts and analyze data

An evaluation

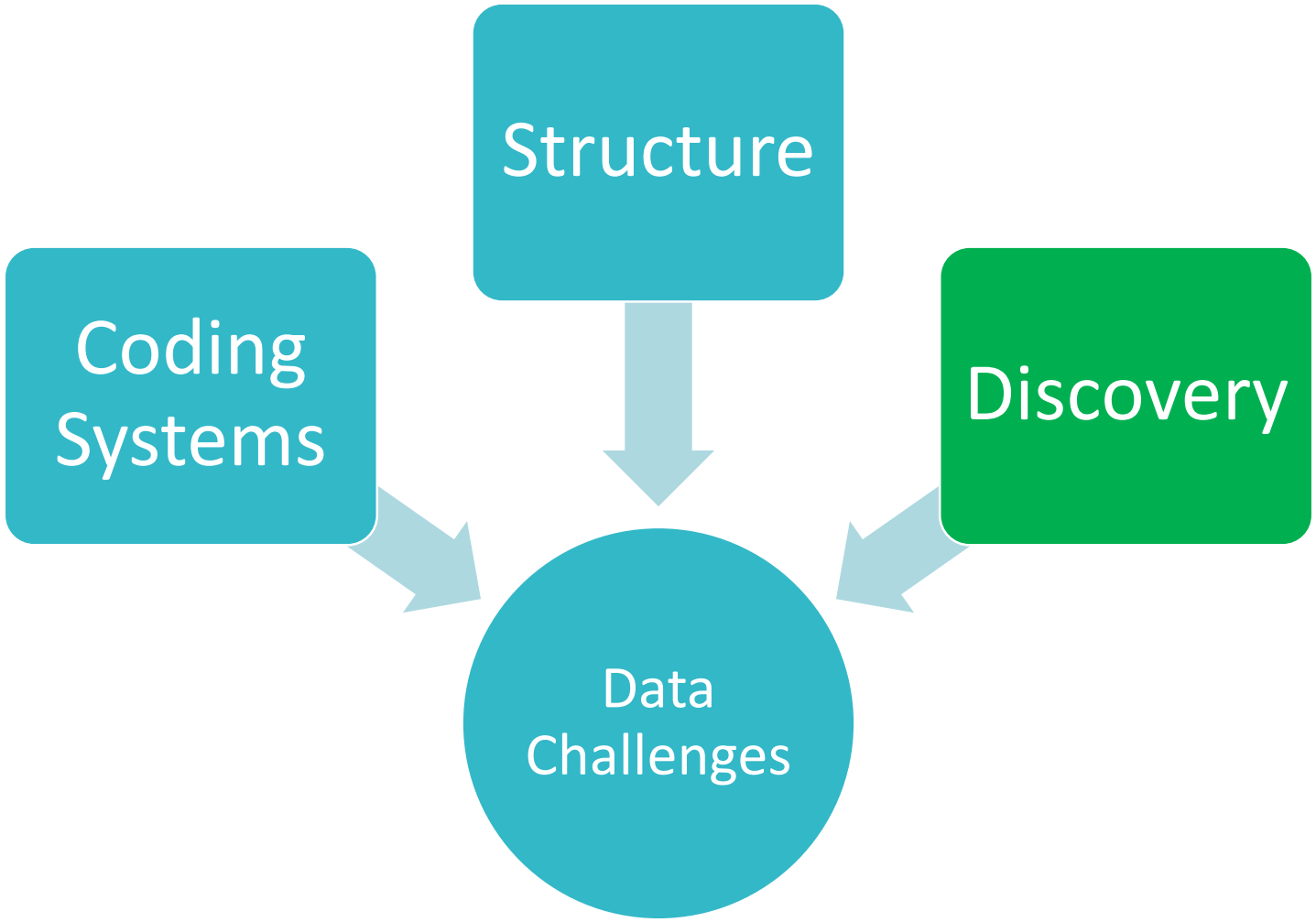


- Workshops
- Use cases
- Existing mappings
- Profiling

- Gap analysis
- New mappings
- ETL design
- Evaluation

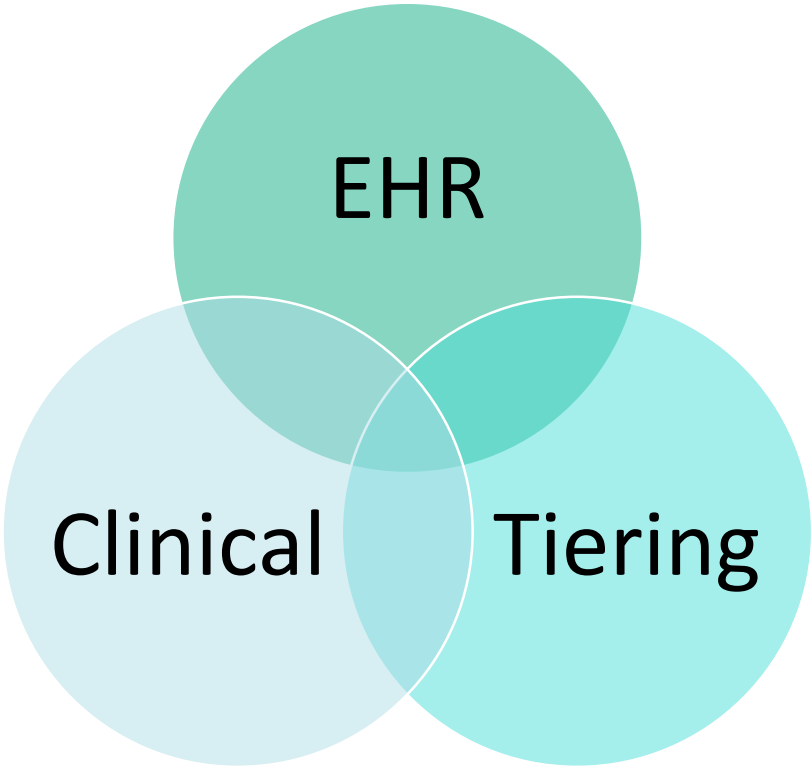
- OMOP genomic extension
- Hybrid approach for pedigree
- Hybrid approach for samples
- Reworked internal models
- Improved metadata and docs

Democratise Research

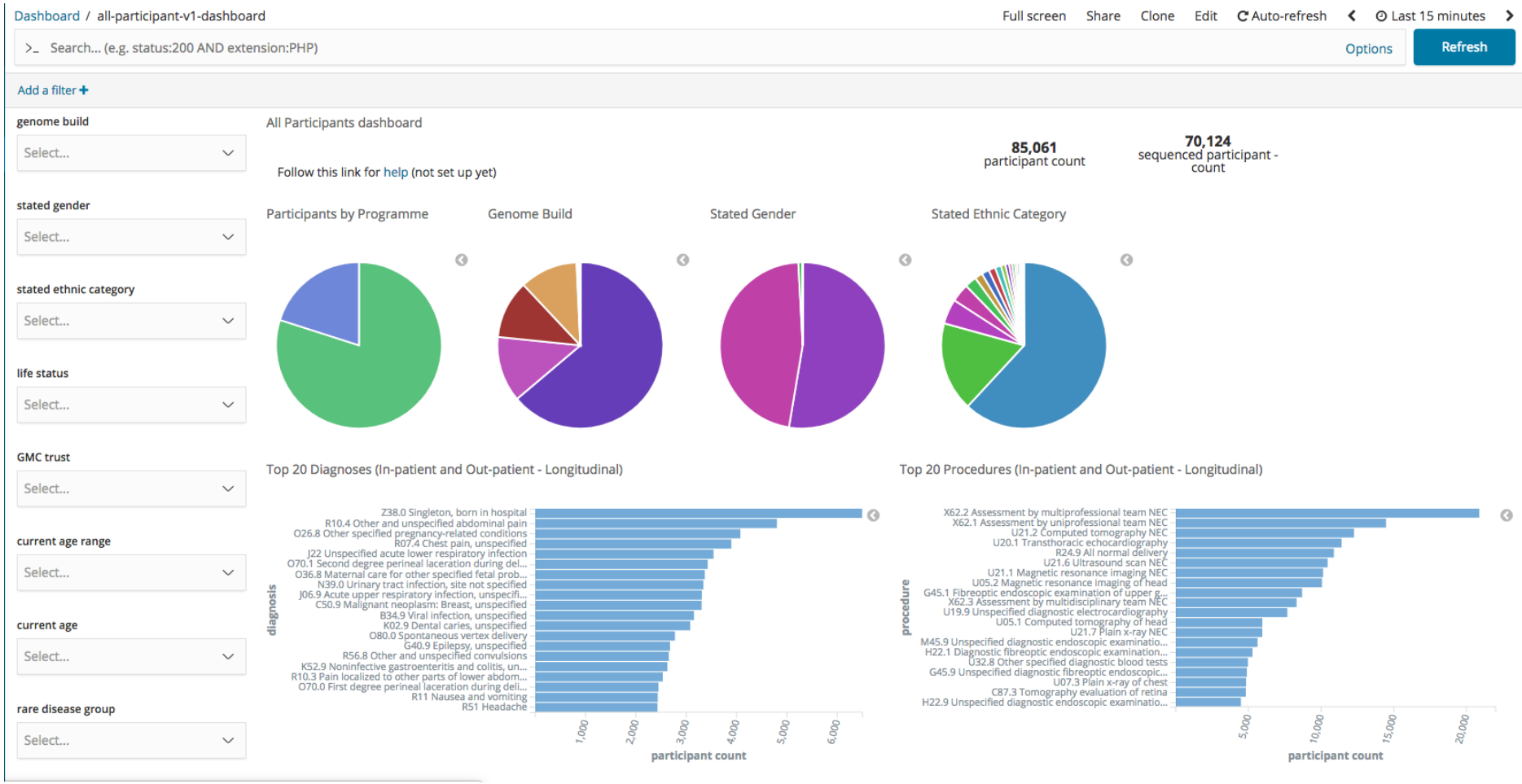


Discovery Portal

- Simple Data Access
- Quick answers to ad hoc queries
- “Visualisations” for easy comprehension



Discovery Portal



4 tables and 85 columns


BIG
DATA
INFORMATION
STORAGE
LARGE
SETS
ANALYSIS
SEARCH
PETABYTES
TECHNOLOGIES
MANAGEMENT
SOFTWARE
MILLION
DEVELOPMENT
TECHNOLOGY
SCIENCE
GOVERNMENT
VOLUME
PROCESS
RESEARCH
INTERNET
LEARNING
WORLD
EVERY
BASED
NEW
BILLION
PARALLEL
FUTURE
SAN
NAS
SOCIAL
YEARS
RESULTS
DRIVE
REPLICATION
HANDLE
SOURCES
HAND
CAPACITY
CHALLENGES
LIMITS
COMPLEX
ANNOUNCED
PAST
INCLUDE
NETWORKS
HYPOTHESIS
END
MOST
NEARLY
MAPREDUCE
CRITIQUES
STATE
CRITIQUE
PEOPLE
TRAFFIC
USE
CURRENTLY
APPROACH
MANY
OUT
ONE
FACTOR
SYSTEMS
MANAGE
ONLINE
DAY
WORLDWIDE
MAKING
INTEGRATION
MASSIVE
DEFINITION
USING
FRAMEWORK
BETWEEN
STATISTICS
BUSINESS
EFFECTIVE
VISUALIZATION
DATABASES
SIZE
ANALYTICS
PROCESSING
SIMULATIONS
EXABYTES
DETERMINE
WELL
SET
DEPARTMENT
QUERIES
HIGH
USERS
HIGHER
ALGORITHMS
COMPANIES
AMOUNT
RATE
REAL
STRUCTURE
WORK
APPLICATIONS
INFRASTRUCTURE
TERMS
YEAR
PRIVATE
CAPABILITIES
SUCCESSFUL
NEEDED
USES
LESS
HUGE
COMPARED
SHARED
VARIOUS
TIMES
ARCHITECTURE
PRIVACY
INTERNATIONAL
IMPLEMENTATION
BECOMES
SECOND
NATIONAL
INTELLIGENCE
FLOW
APPROACHES
PROCESSES
PARADIGM
FUNDING
RECORDS
COST
INSIGHT
TOOLS
CENTER
INITIATIVE
ECONOMIC
PROJECT
DISTRIBUTED
PER
RELEVANT
TERABYTES
MARKET
VARIETY
ANNUAL
USED
RELATED
TIME
SENSOR
WOULD
NOW

Stay in touch

@genomicsengland #genomes100k

Like the ‘Genomics England’ page

Follow ‘Genomics England’

Subscribe to our newsletter:
www.genomicsengland.co.uk/sign-up

www.genomicsengland.co.uk