Converting Data Into Knowledge

Bridging the Gap – Amanda Borens

CRITICAL PATH INSTITUTE

Bridging the Gap

I found DATA!

I know the question I want to ANSWER!

Bridging the Gap

➤ How do we transform data ➜ information?
Especially when we have **minimal data**?
Or when we are **inundated with data**?

➤ Two Case Studies:
Alzheimer's Disease, biomarkers, and longitudinal data
Tuberculosis, ReSeqWHO, and genomics data

Bridging the Gap

I found DATA!

I know the question I want to ANSWER!

How do I go from **Raw Data** to a **Quantitative Solution**?

Image Credit: Getty Images

Alzheimer's Disease

Alzheimer's Disease– *AD Interoperability*

➢ **Research Question:**
"Which longitudinal biomarkers are related to specific clinically-relevant measures of progression and drug effects, at specific disease stages?"

➢ **Data challenge:**
- Combine data from **disparate platforms** to answer question
- No common **data dictionary** or data model
- Create **data marts on demand**

Combining Disparate Data Sets

Combining data from different studies is *more than mapping variables*

Observational Study

ID	GENDER
A1	Male
A2	Male
A3	Female
A4	Female
A5	Male

Clinical Trial #1

SUBJID	SEX
0001	M
0002	F
0003	F
0004	M
0005	F

EHR records

PTID	GENDER
0001	1
0002	1
0003	2
0004	2
0005	1

Pre-clinical Trial #2

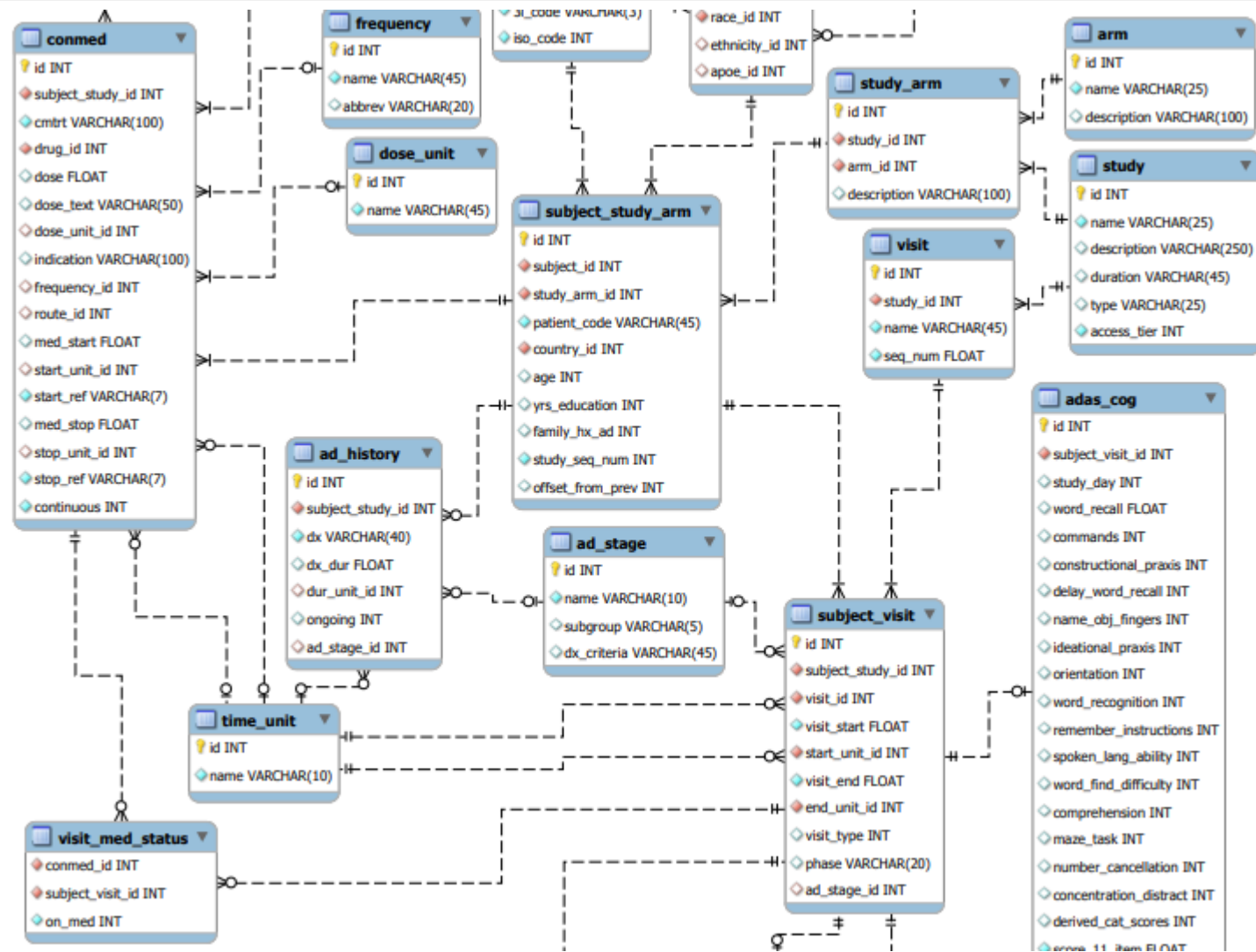| USUBID	SEX
00011	0
00012	1
00013	1
00014	0
00015	1

Standardizing **dates** (eg. medications) **can be a quagmire**
Ranges, start only, month only, year only….

Alzheimer's Disease

- **CPAD clinical trials**
- **Observational studies**
- Future expansion to include **EHR data**

Alzheimer's Disease

| TABLE_NAME	COLUMN_NAME	DATA_TYPE	COLUMN_KEY	COLUMN_COMMENT	VALUES
ad_history	ad_stage_id	INT	MUL	(FK: ad_stage.id) The standardized AD stage term for the diagnosis	
ad_stage	subgroup	VARCHAR		The subgroup within the disease stage. Primarily used for MCI to indicate whether it is "early" or "late" stage MCI.	e.g., "Early MCI", "Late MCI"
ad_stage	name	VARCHAR		The name of the Alzheimer's related disease stage	Normal, MCI, AD
ad_stage	id	INT	PRI	Database assigned numeric ID	
ad_stage	dx_criteria	VARCHAR		The criteria used to make the AD stage diagnosis	
adas_cog	word_recognition	INT		The number of incorrectly recognized words	Score Range: 0 - 12; Not Done Codes: -4 = cognitive impairment reasons, -3 = physical reasons, -2 = subject refused, -1 = other reasons
adas_cog	word_recall	FLOAT		The mean number of words not recalled on the 3 word recall trials	Score Range: 0.0 - 10.0; Not Done Codes: -4 = cognitive impairment reasons, -3 = physical reasons, -2 = subject refused, -1 = other reasons
adas_cog	word_find_difficulty	INT		Rating of the word finding difficulty in spontaneous speech	Score Range: 0 - 5; Not Done Codes: -4 = cognitive impairment reasons, -3 = physical reasons, -2 = subject refused, -1 = other reasons
adas_cog	subject_visit_id	INT	MUL	(FK: subject_visit.id) The visit during which the subject performed the assessment	
adas_cog	study_day	INT		The study day the assessment occurred on	

Tuberculosis – *Critical Path to TB Drug Regimens*

➢**Data challenge:**

- Combine drug sensitivity testing results with genomic variants
- This enables statistical analyses to identify mutations that confer resistance
- Requires **terabytes** of data to elucidate relationships
- Epidemiologists and policy-makers may not have Big Data skills, making information inaccessible

Tuberculosis

➢ Combining statistical resistance reports with global surveillance reports **could guide first-line drug recommendations**

➢**C-Path's role:**

- **Reduce the data burden** by providing tools that give scientists manageable output reports

- **Communicate** the usefulness of the data with 'stories' in reports

Reduce the Data Burden

WGS fastq

**1,500,000 KB**

SNP Report

1,200 KB

Tuberculosis
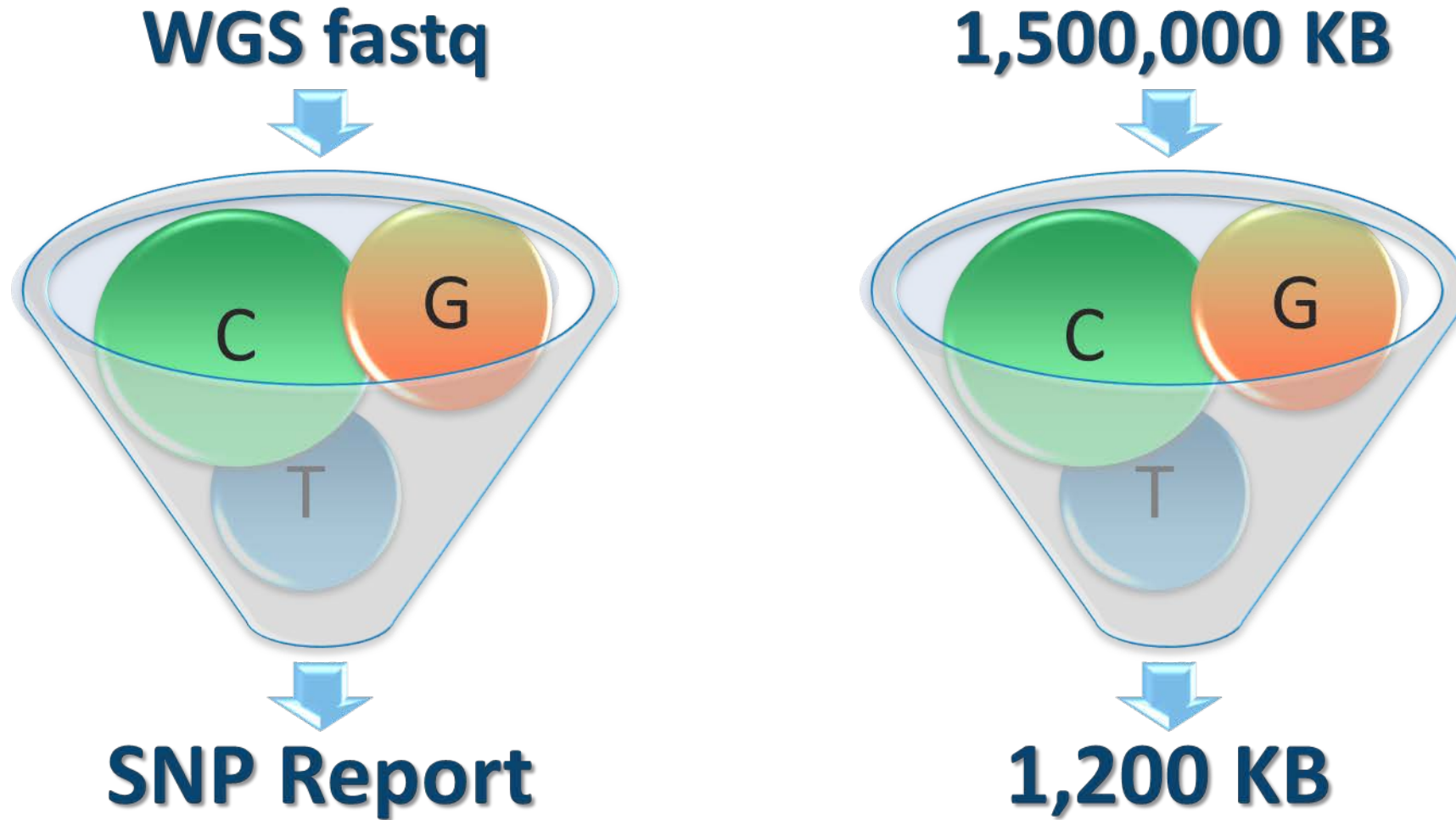
➢ ReSeqTB in CODR platform – not yet optimized for analytics

- SQL algorithm for reporting SNPs that are associated with resistance **ran in 3-8 hours**

➢ ReSeqWHO with redesigned database schema, relations, and indexes

- SQL algorithm for reporting SNPs that are associated with resistance **ran in about *8 seconds***

Bridging the Gap

➢ Combining statistical resistance reports with global surveillance reports **could guide first-line drug recommendations**

➢ **C-Path's role:**

- **Reduce the data burden** by providing tools that give scientists manageable output reports

- **Communicate** the usefulness of the data with 'stories' in reports

Communicate with Reports
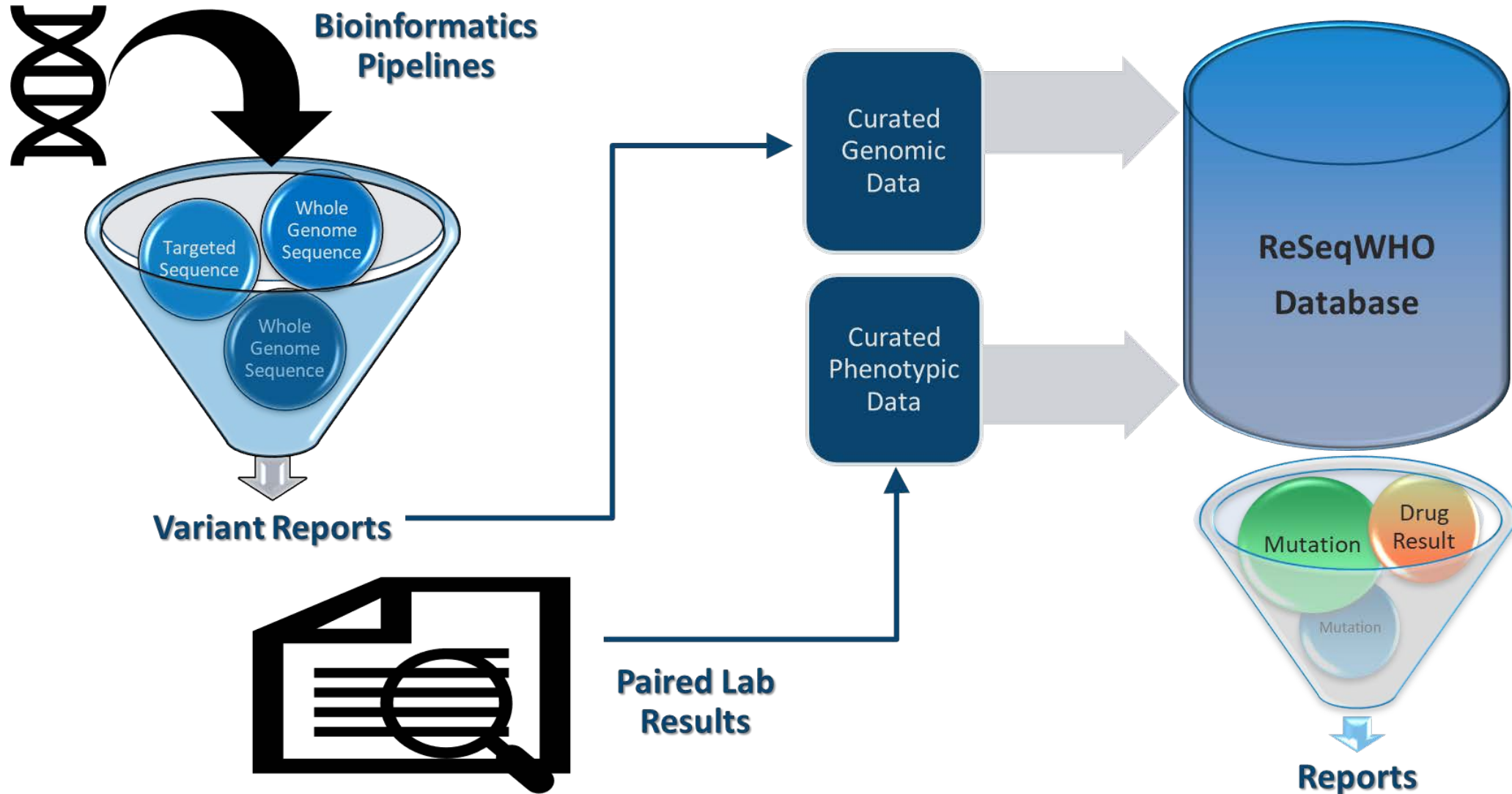
Bridging the Gap

THANK YOU

Amanda J. Borens

Data Platform Architect | Development Manager

aborens@c-path.org

Data Collaboration Center

https://c-path.org/programs/dcc/