Four years of Datomic powered ETL in anger with CANDEL

Speakers

Marshall Thompson, Ph.D.

Benjamin Kamphaus, Ph.D.

Speakers

We were both trained as scientists
- Ben in Earth Science
- Marshall in Genetics and Genomics

We both ended up in the software industry.

PICI & CANDEL were a first return to science for both of us.

CANDEL was an opportunity for us to bring what we learned in software back home.

Mission: to accelerate the development of breakthrough immune therapies to turn all cancers into curable diseases.

The CANDEL Project: History

Federico & Lacey Came to Cognitect (through a Datomic evaluation)

Cancer **D**ata and **E**vidence **L**ibrary

Several Unique Aspects of The CANDEL Project

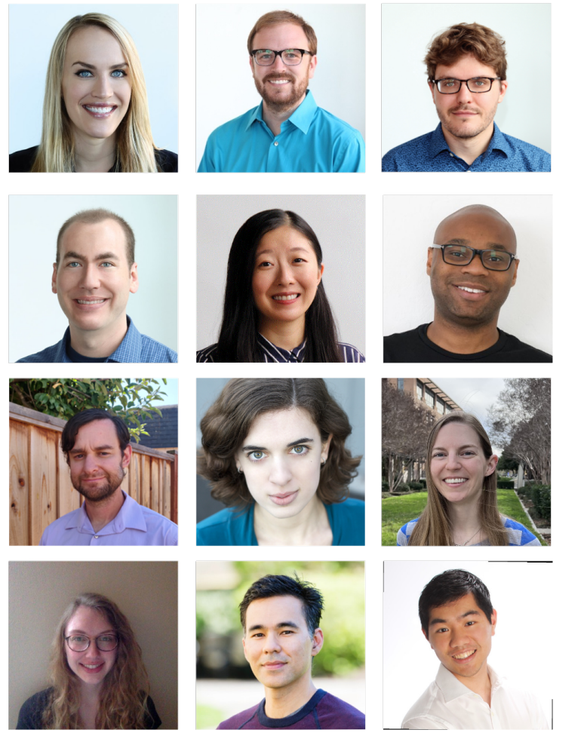
We both spent time on the Datomic product team at Cognitect prior to the project.

We sat at both sides of the table: starting out as the primary consultants from Cognitect, then as both end users and product owners at PICI.

People

Project Kick Off

Dev Team



Bioinformatics Users

CANDEL Project Timeline

- Fall 2018: Project starts between PICI & Cognitect
- Jan 2019: First iteration of CANDEL is up and in use
- Spring 2019: First CANDEL integrations with R
- Summer 2019: Mantis 1.0
- Spring 2019: First versions of RawSugar & Enflame
- Fall 2020: Ben joins PICI full time
- Summer 2021: Marshall joins PICI full time
- Summer 2022: More than 50 Datasets in CANDEL
- Spring 2023: CANDEL Open Sourced

Why CANDEL?

Goal of CANDEL

- increase the pace & quality of research at PICI
- by taking aim at the biggest headaches in biological data:
  - data integration
  - data harmonization

The boring infrastructure problem that bottlenecks biological research

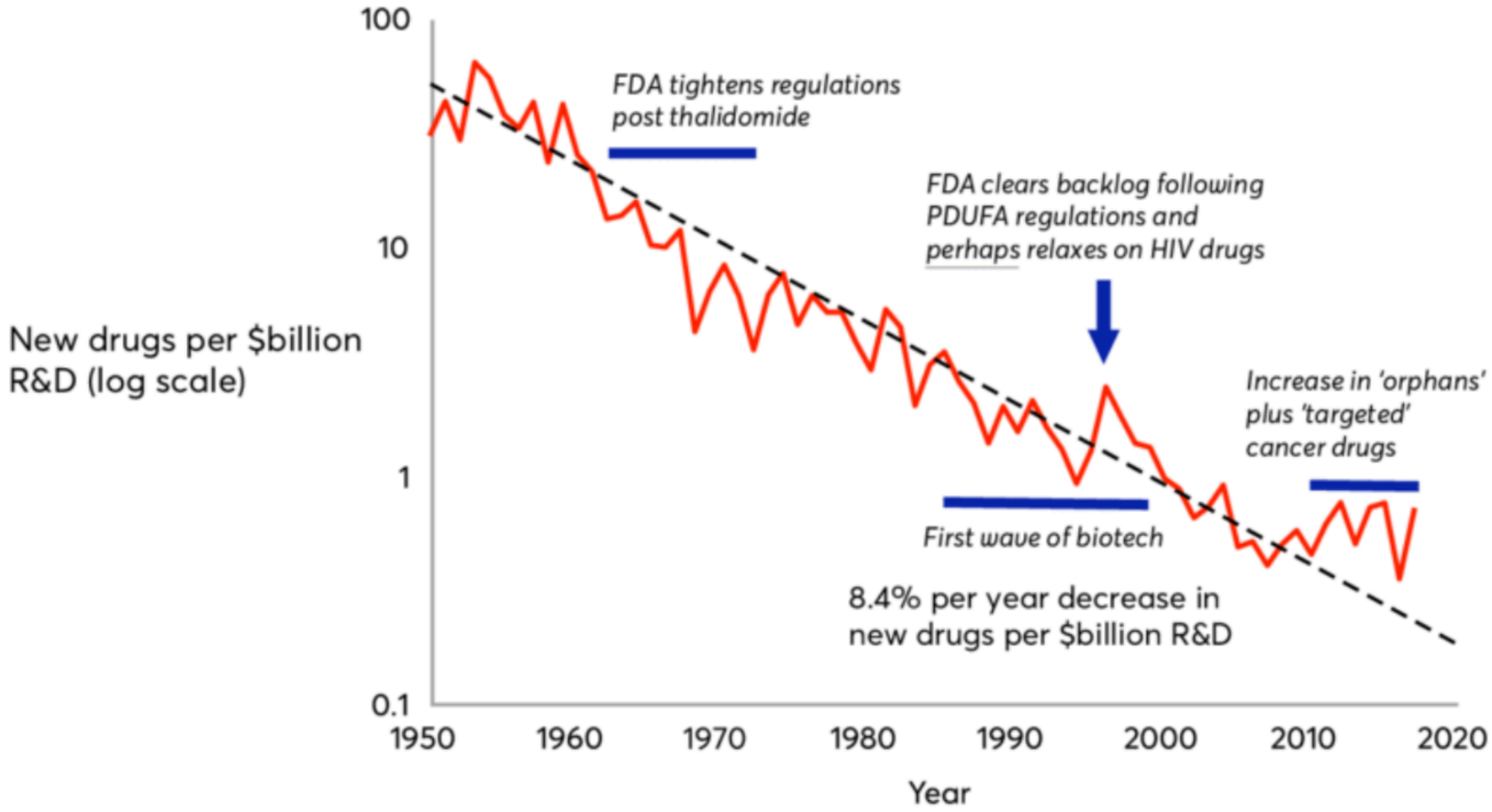
Data Harmonization

- Harmonization is a big problem in biology
- Harmonization is harder than you might expect
- Harmonization is where too much time gets spent instead of science
- Harmonization problems propagate and impact science quality

Data harmonization

- Errors creep in
- People spend too much time hunting them down
- People spend too much time hypothesizing about the wrong things and solving the wrong problems
- Not just data sci or dev time lost but “hammock time” also

Eroom's Law


100
10
1
0.1
New drugs per $billion R&D (log scale)
FDA tightens regulations post thalidomide
FDA clears backlog following PDUFA regulations and perhaps relaxes on HIV drugs
Increase in 'orphans' plus 'targeted' cancer drugs
First wave of biotech
8.4% per year decrease in new drugs per $billion R&D
1950
1960
1970
1980
1990
2000
2010
2020
Year

- 92% of genome sequenced in 2003 ('mission accomplished' point)
- The *cost* of developing a new therapeutic still follows Moore’s law
- cost of sequencing genome has gone from ~$3 billion to $600 in the last twenty years.
- data availability and productions from biological assays have exploded well past exponential growth- greater than Moore’s law

Data's Role

The challenge of handling biological data is not the sole cause of this, but data has either:

- Exacerbated problem
- Had no impact on problem

Neither is acceptable

The Shape of Biological Data

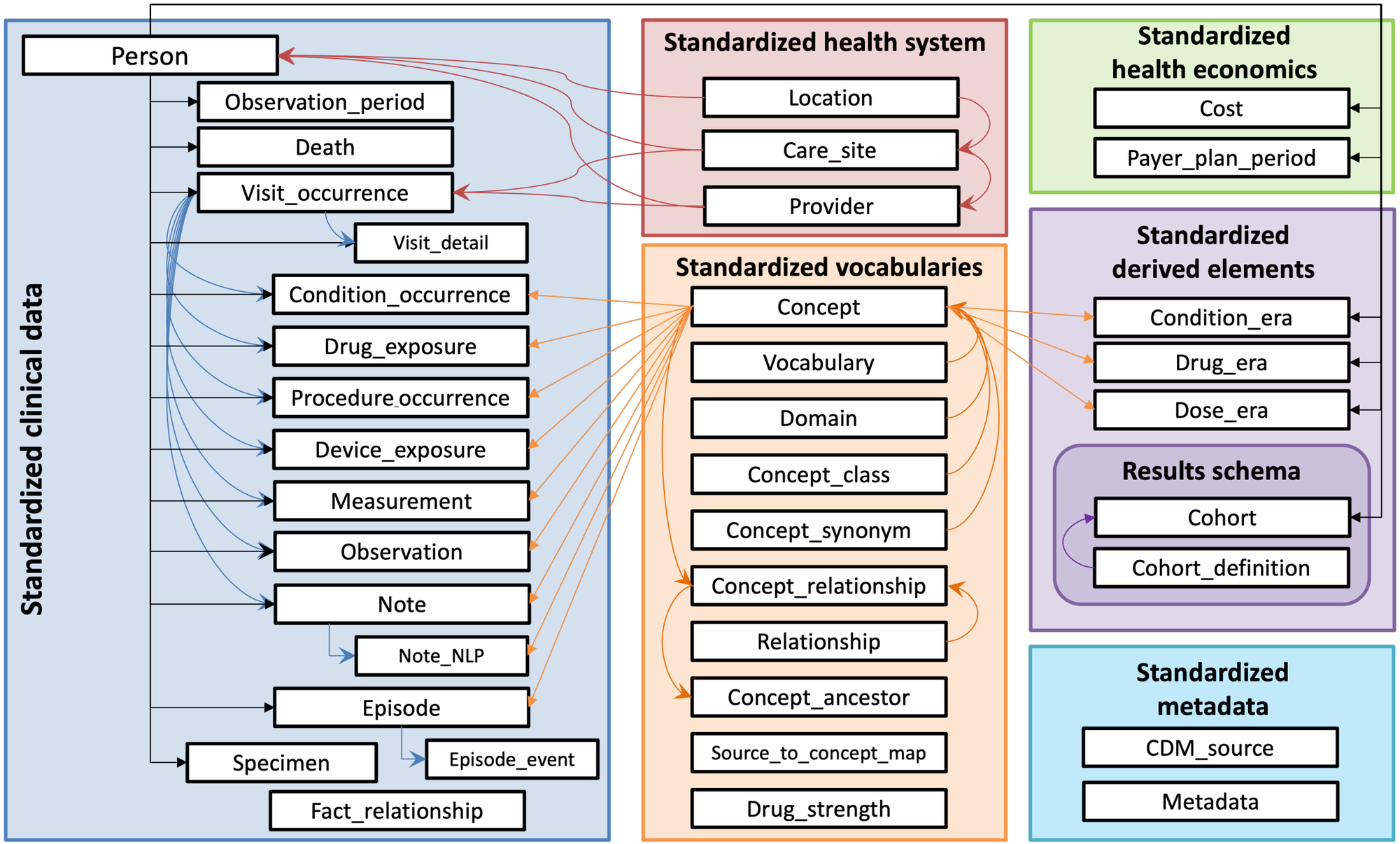
- Bulk Molecular Data
- Patient Outcomes
- Clinical Assays
- Imaging
- Single Cell Data
- Features derived from other assays

PICI's Unique Data Integration & Harmonization Needs

- Partner sites across academia and industry
- New and experimental assays
- Team effort split across multiple projects, trials, etc

The Shape of Biological Data

- Deeply relational, deeply nested
- Sparse, lots of holes
- Large and getting larger

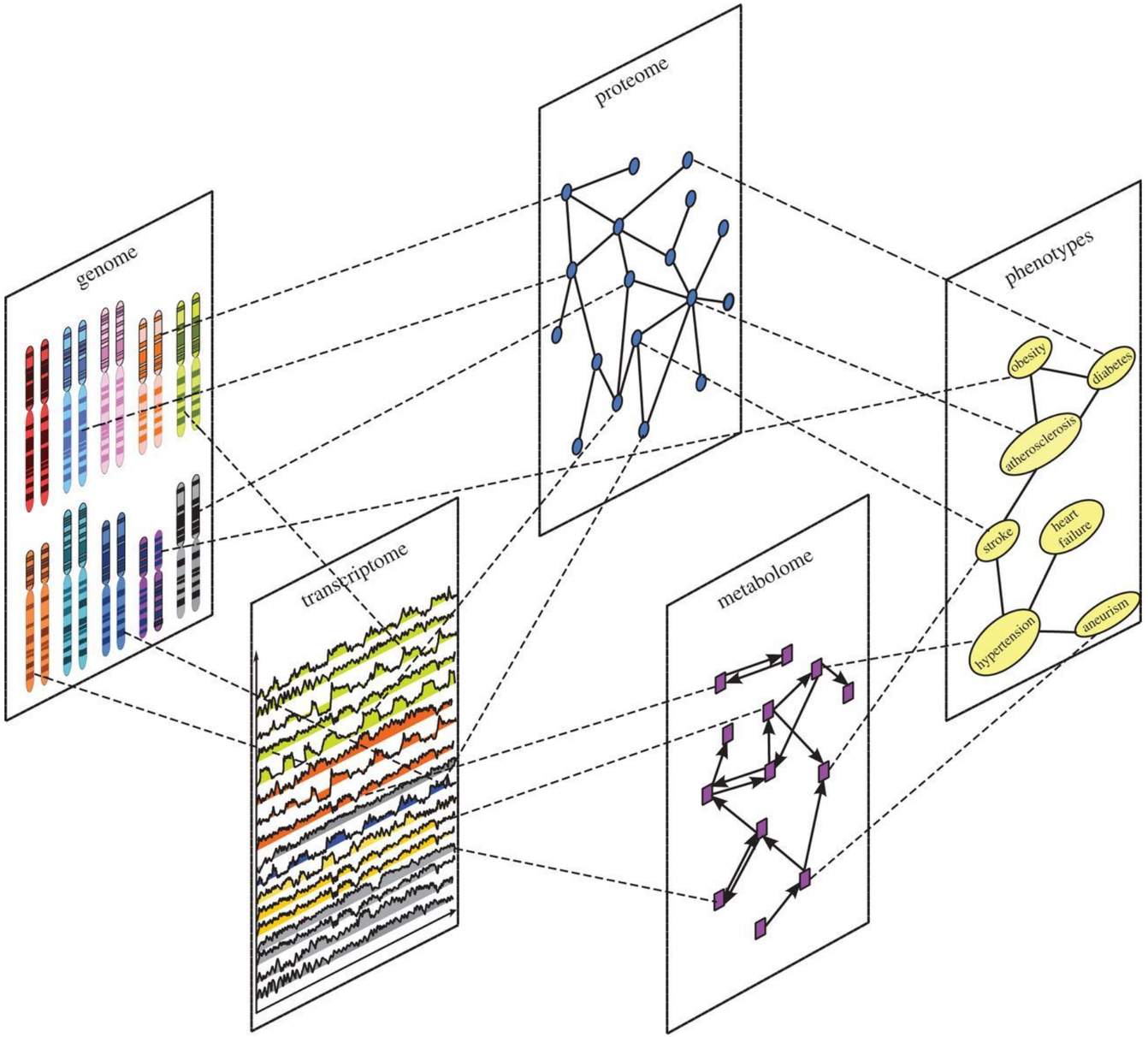

Standardized clinical data
Person
Observation_period
Death
Visit_occurrence
Visit_detail
Condition_occurrence
Drug_exposure
Procedure_occurrence
Device_exposure
Measurement
Observation
Note
Note_NLP
Episode
Episode_event
Specimen
Fact_relationship
Standardized health system
Location
Care_site
Provider
Standardized vocabularies
Concept
Vocabulary
Domain
Concept_class
Concept_synonym
Concept_relationship
Relationship
Concept_ancestor
Source_to_concept_map
Drug_strength
Standardized health economics
Cost
Payer_plan_period
Standardized derived elements
Condition_era
Drug_era
Dose_era
Results schema
Cohort
Cohort_definition
Standardized metadata
CDM_source
Metadata

Omics data at PICI

- Bulk
- Tissue specific
- Image derived
- Cell populations
- Single Cell

- Measured by a particular assay, eg
- RNASeq
- WES
- Metabolic Panel
- Flow

- Measurement targets such as
  - Genes
  - Proteins
  - Mutations
  - Non-trivial relations for all the above


genome
proteome
transcriptome
metabolome
phenotypes
obesity
diabetes
atherosclerosis
stroke
heart failure
hypertension
aneurism

There are Two Other Talks on CANDEL

- Clojure Where it Counts: Tidying Data Science Workflows
- Building a Unified Cancer Immunotherapy Data Library

The first entry in CANDEL: Pret

Pret: automated ETL at the heart of CANDEL:

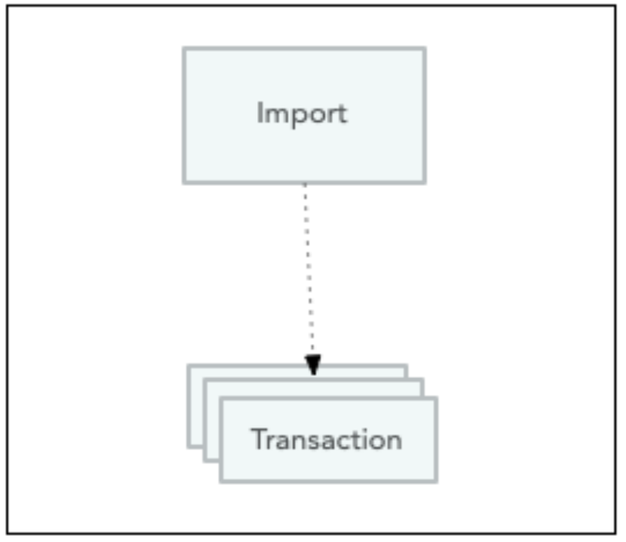
- Turns tables to datoms
- Derives rules for doing so inspecting the schema in Datomic
- Additional layer of schema annotations — the metamodel — assist with this.
- Specs ensure granular data accuracy as well as cross-dataset referential integrity.

```
:samples [{:pret/input-file "processed/samples.txt"
           :id              "Originating ID"
           :subject         "Participant ID"
           :timepoint       "VISIT"
           :specimen        "Source Matcode"
           :container       "BioInventory Group Name"}
```

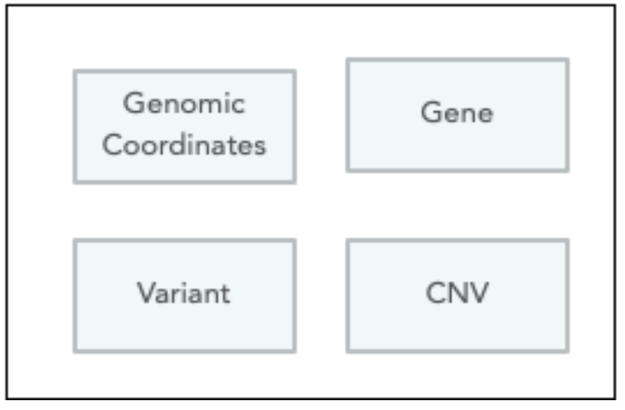
```
:samples [{:pret/input-file "processed/samples.txt"
           :id              "Originating ID"
           :subject         "Participant ID"
           :timepoint       "VISIT"
           :specimen        "Source Matcode"
           :container       "BioInventory Group Name"}
:subjects {:pret/input-file "processed/subjects.txt"
           :id              "USUBJID"
           :race            "RACE"
           :ethnicity       "ETHNIC"
           :therapies
           {:pret/input-file   "processed/therapies.txt"
            :treatment-regimen "TRTACD"
            :order             "order"
            :pret/reverse      {:pret/rev-variable "USUBJID"
                                :pret/rev-attr     :subject/therapies}}}
```

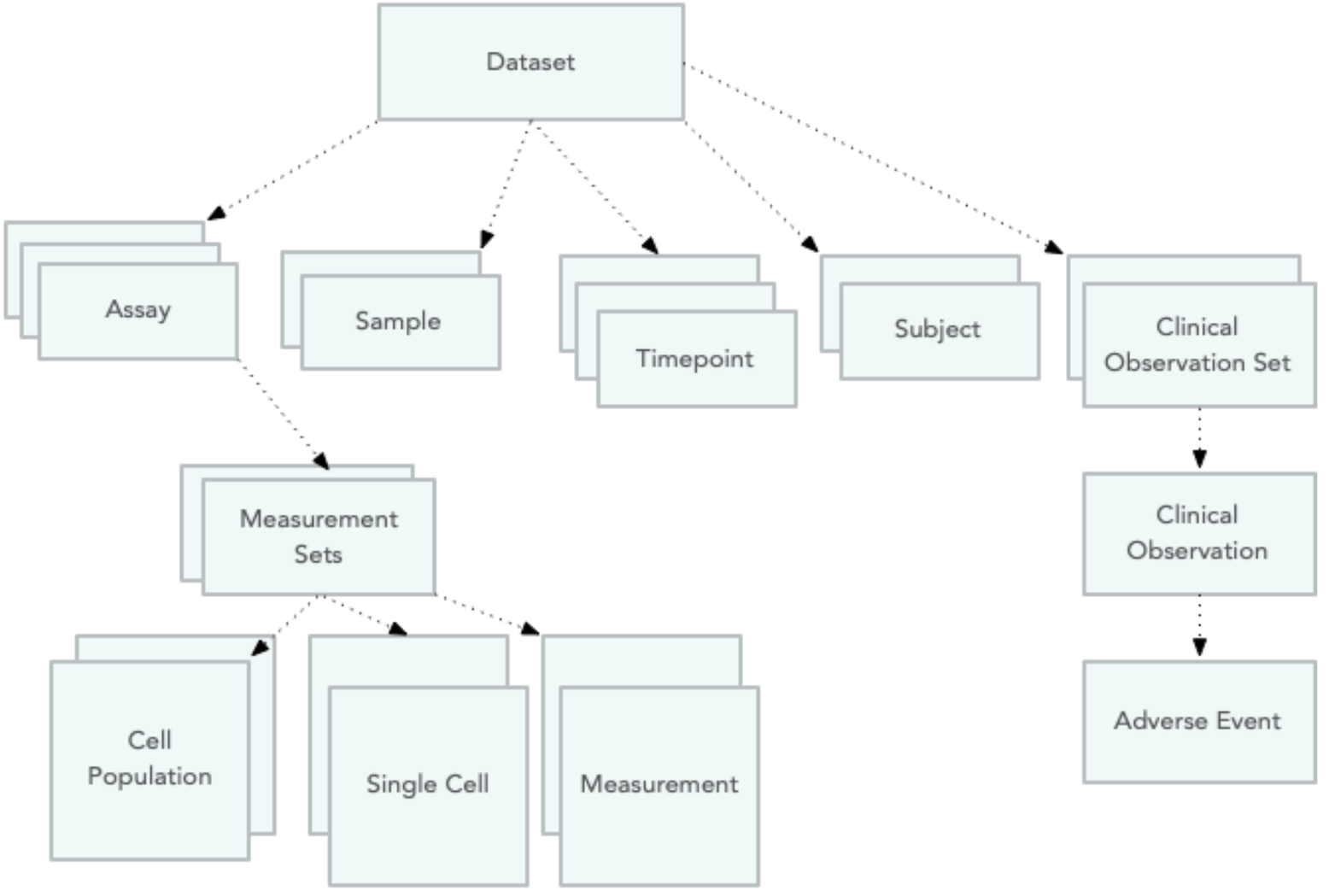
```
{:name "CyTOF"
 :technology          :assay.technology/mass-cytometry
 :description         "CyTOF analysis"
 :measurement-sets
 [{:name                 "Bendall"
   :cell-populations
   [{:pret/input-file  "processed/cell_populations.txt"
     :pret/na          "NA"
     :name             "name"
     :positive-markers "positive.epitopes"
     :cell-type        "cell.type"}
    ...]
```

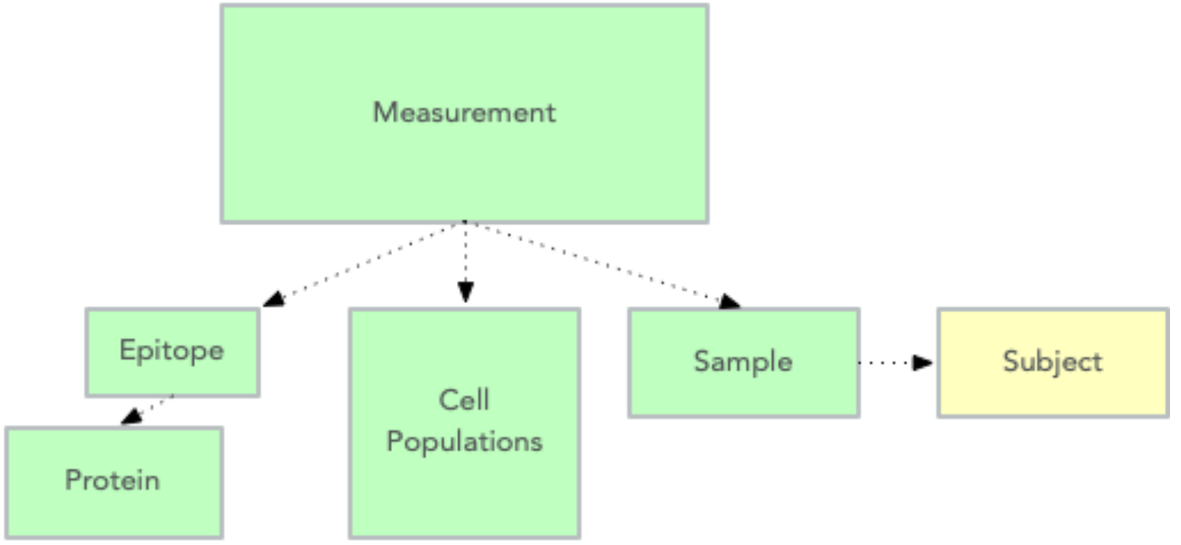
Import Data

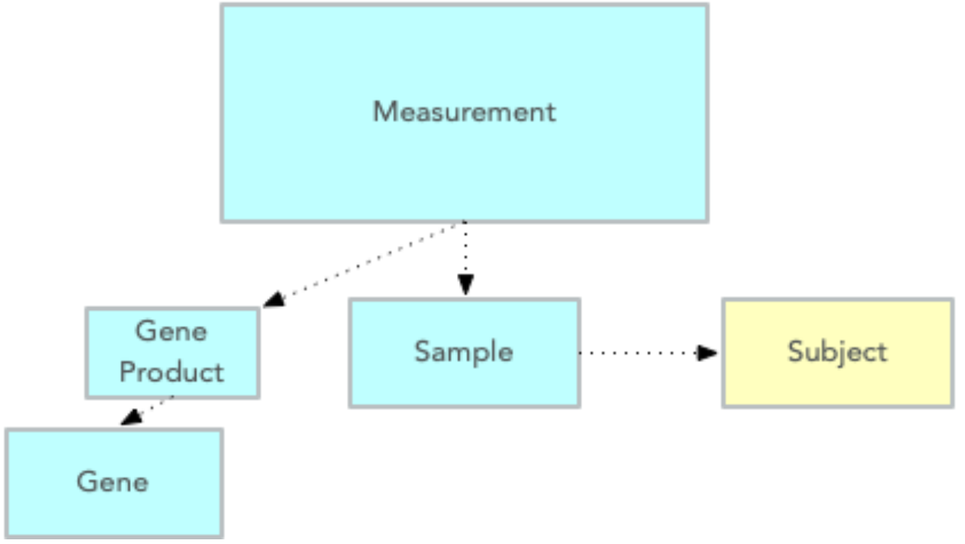


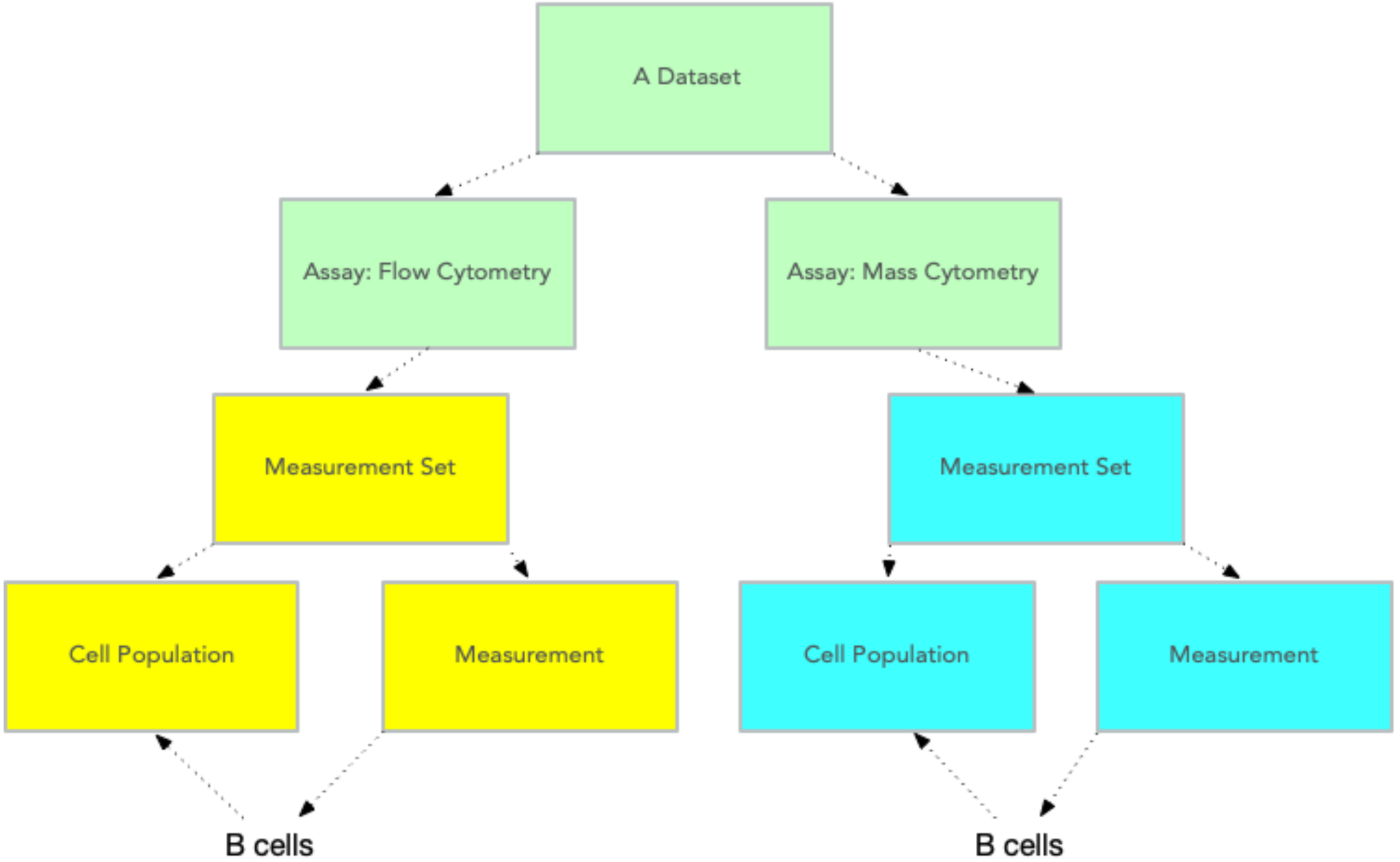
Reference Data






Measurement
Epitope
Protein
Cell Populations
Sample
Subject

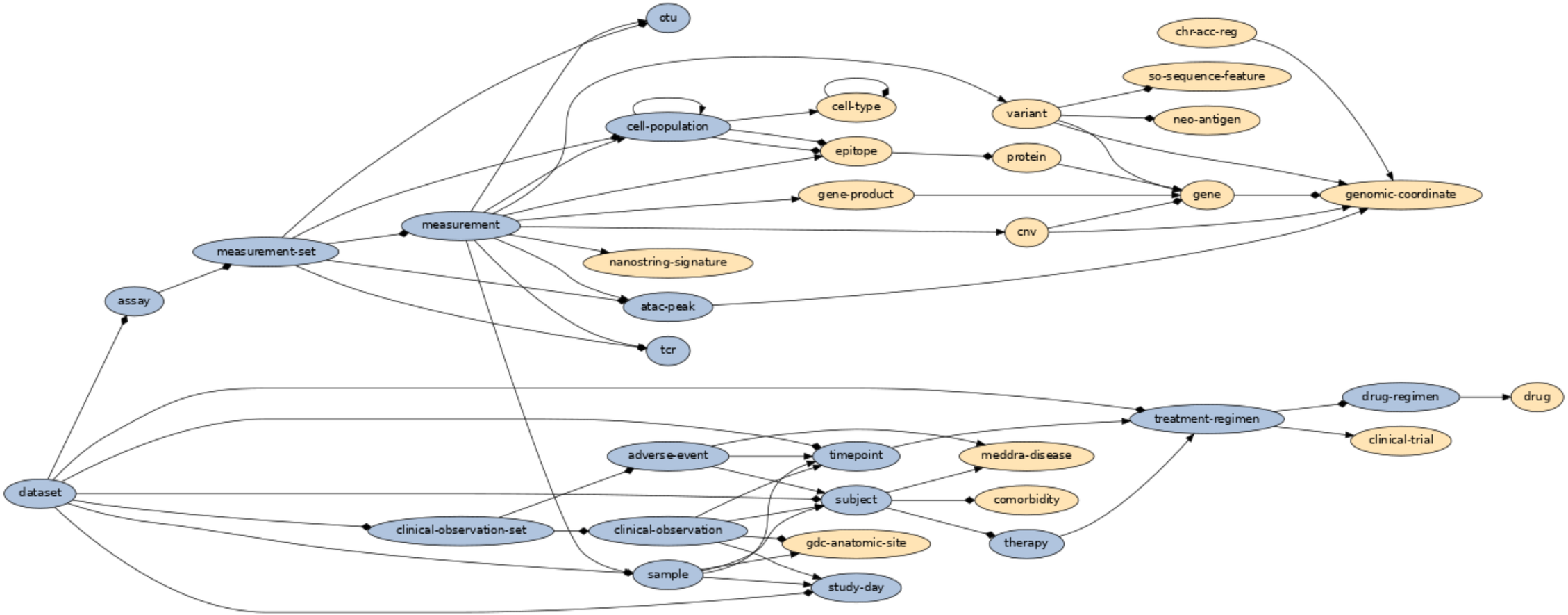

Measurement
Gene Product
Gene
Sample
Subject


A Dataset
Assay: Flow Cytometry
Assay: Mass Cytometry
Measurement Set
Measurement Set
Cell Population
Measurement
Cell Population
Measurement
B cells
B cells

```
{:name             "CyTOF"
 :technology       :assay.technology/mass-cytometry
 :description      "CyTOF analysis performed at Primity Bio"
 :measurement-sets
 [{:name             "Bendall"
   :cell-populations [{:pret/input-file  "processed/cell_populations_Bendall.txt"
                       :pret/na          "NA"
                       :name             "name"
                       :positive-markers "positive.epitopes"
                       :cell-type        "cell.type"}]
   :measurements     [{:pret/input-file "processed/cytof_measurements_Bendall.txt"
                       :pret/na         "NA"
                       :sample          "sample"
                       :cell-population "uniquePopulationName"
                       :pret/variable   "variable"
                       :pret/value      "value"
                       :pret/variables  {"eventCount"               :measurement/cell-count
                                         "normalization.measurement" :measurement/leukocyte-count
                                         "normalized.measurement"    :measurement/percent-of-leukocytes}}]}
  {:name             "Spitzer"
   :cell-populations [{:pret/input-file  "processed/cell_populations_Spitzer.txt"
                       :pret/na          "NA"
                       :name             "name"
                       :positive-markers "positive.epitopes"
                       :cell-type        "cell.type"}]
   :measurements     [{:pret/input-file "processed/cytof_measurements_Spitzer.txt"
                       :pret/na         "NA"
                       :sample          "sample"
                       :cell-population "uniquePopulationName"
                       :pret/variable   "variable"
                       :pret/value      "value"
                       :pret/variables  {"eventCount"               :measurement/cell-count
                                         "normalization.measurement" :measurement/leukocyte-count
                                         "normalized.measurement"    :measurement/percent-of-leukocytes}}]}]}]
```

```
{:kind/name          :assay
 :kind/attr          :assay/name
 :kind/context-id    :assay/name
 :kind/need-uid      :assay/uid
 :kind/parent        :dataset}
{:kind/name          :measurement-set
 :kind/attr          :measurement-set/name
 :kind/context-id    :measurement-set/name
 :kind/parent        :assay
 :kind/need-uid      :measurement-set/uid}
{:kind/name          :cell-population
 :kind/attr          :cell-population/name
 :kind/context-id    :cell-population/name
 :kind/parent        :measurement-set
 :kind/need-uid      :cell-population/uid}
```

Recent CANDEL Data Model



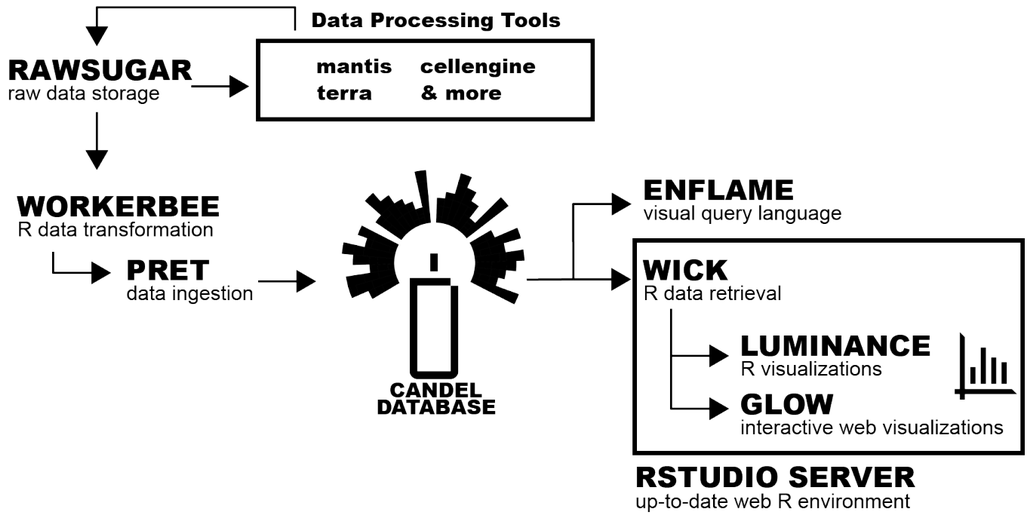
The first tooling for access

- R tools: datalogr, wick, luminance, glow
  - see talk linked on previous slide
- PICI data science used the data in CANDEL primarily from R
- The data model supports query over JSON, so supports:
  - Clojure (both directly from Datomic and as query lib client)
  - Python, or whatever other data science language

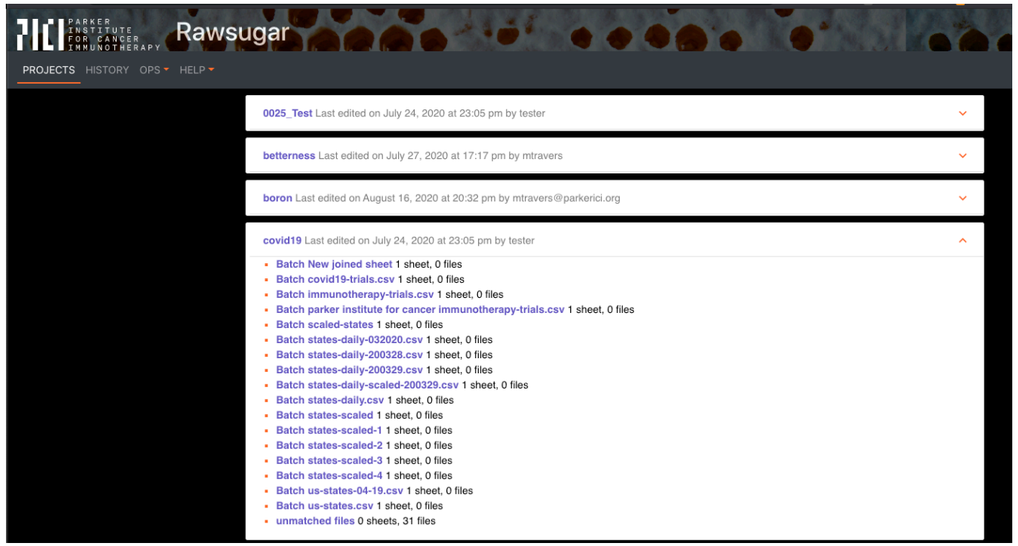
Experience Report: Everything is Always Evolving

CANDEL Ecosystem Grew from Here

- RawSugar, upstream
- WorkerBee, upstream
- Mantis, upstream
- Enflame, analysis enablement, dashboard

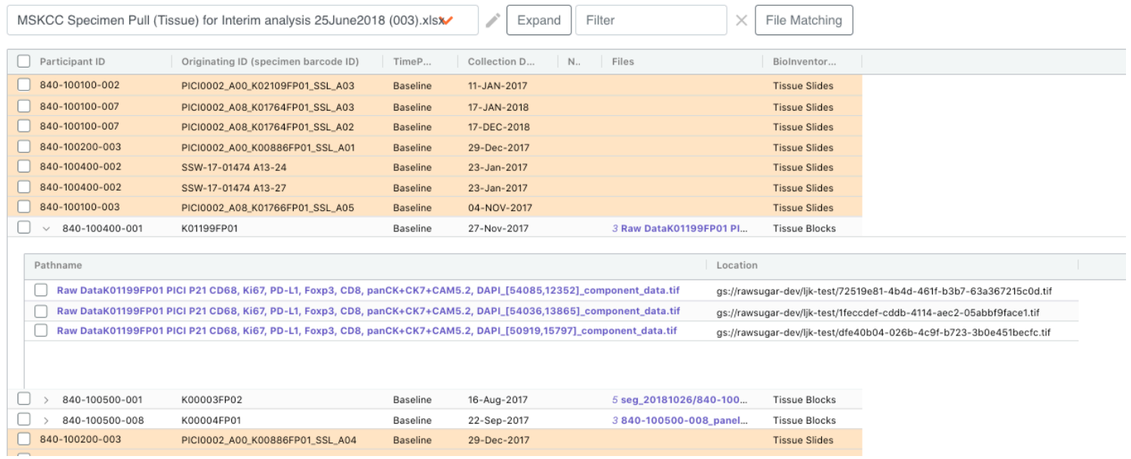


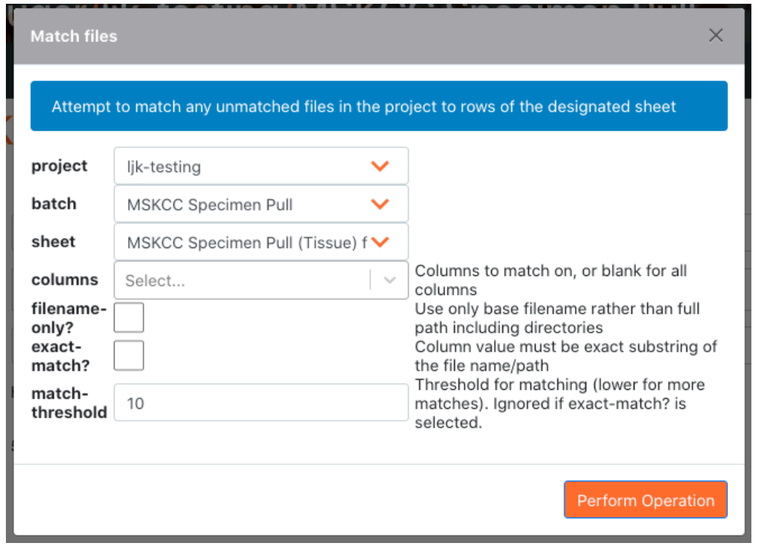
RawSugar



- Structured handling of raw files
- Built on Datomic, includes an immutable history of all file updates and changes

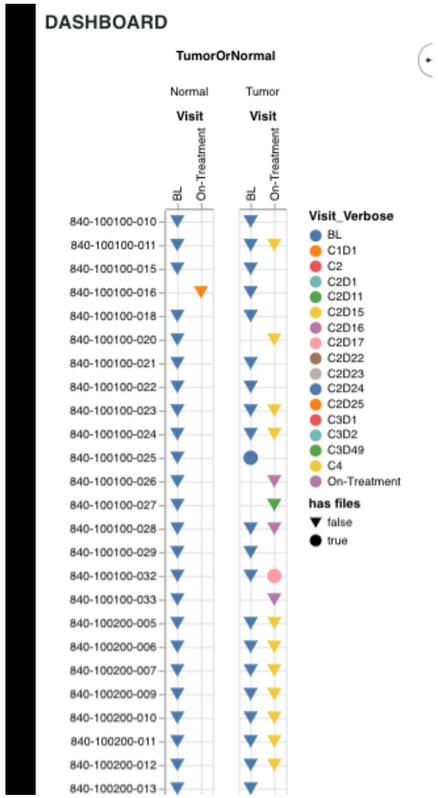
RawSugar





- Relating data as received by vendor with what goes into CANDEL
- Fuzzy file matching & metadata management

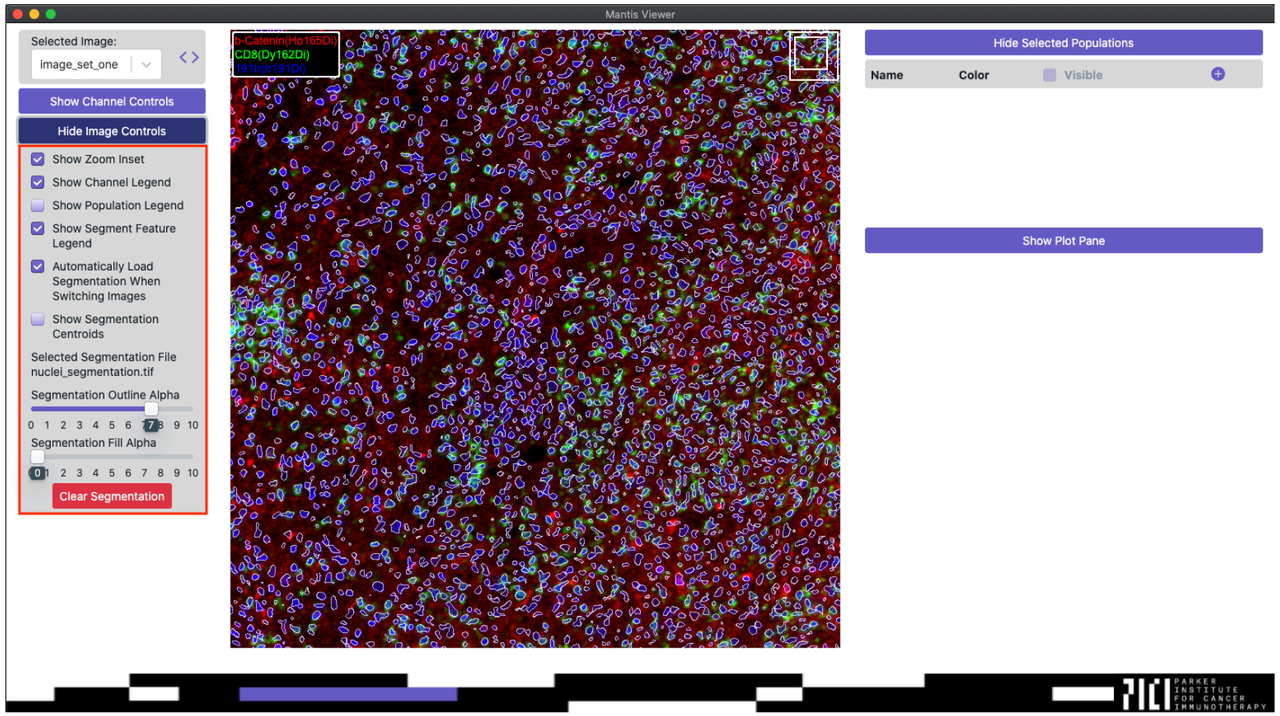
RawSugar



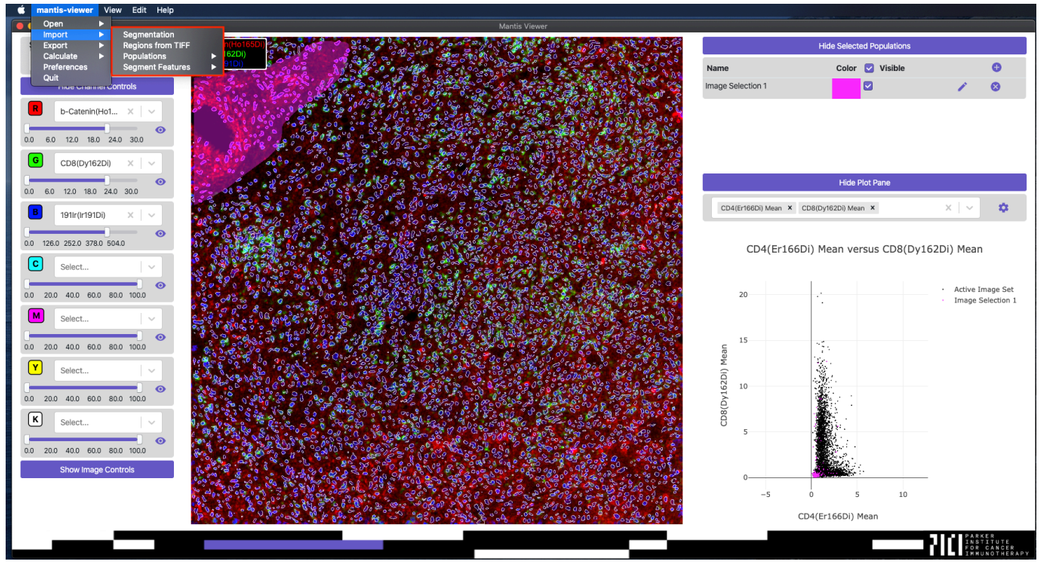
- Organization and management of data to ensure completeness

Mantis

Mantis is a viewing and analysis tool for multi-channel microscopy imaging

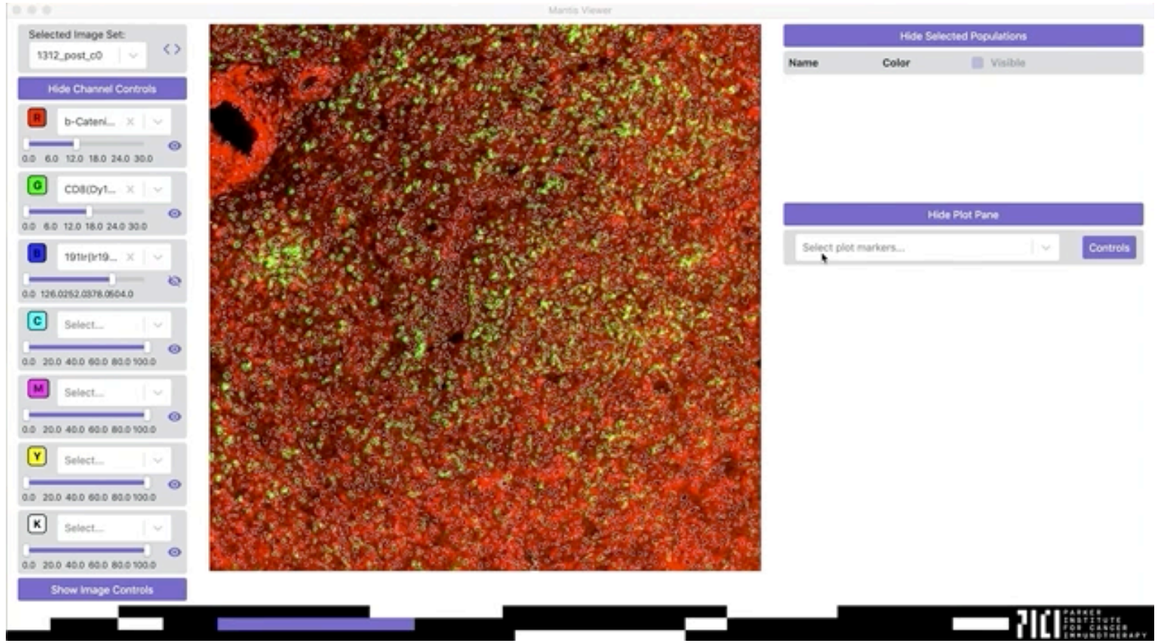


Mantis



- Aggregates features to the level of a cell population or cell
- Cell population or cell level features modeled in CANDEL's data model

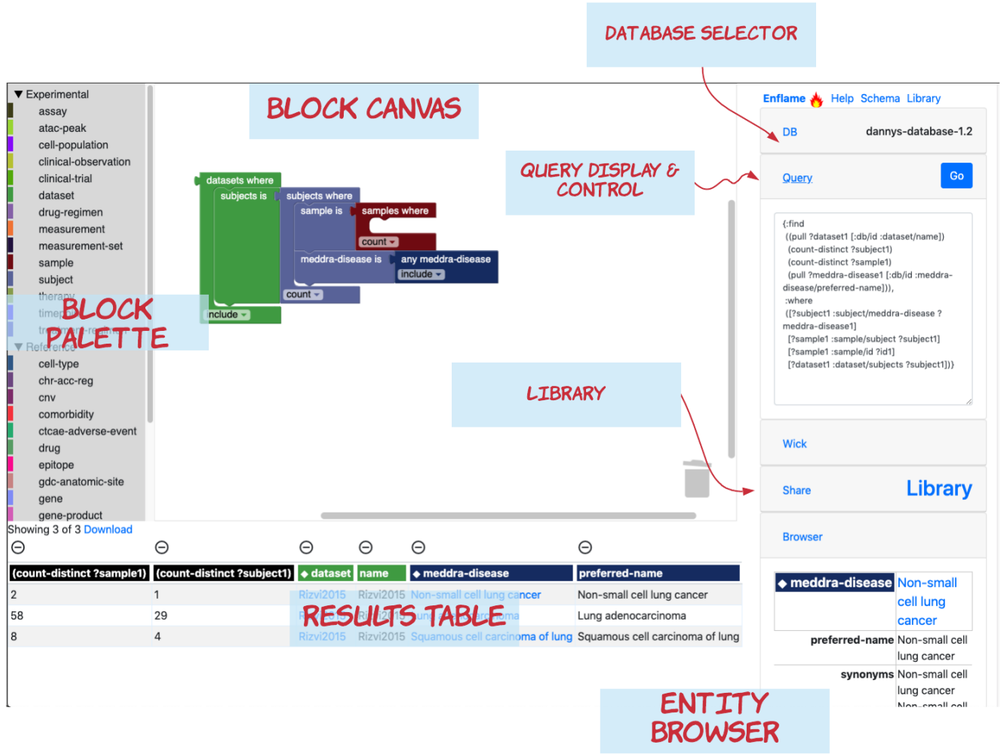
Mantis



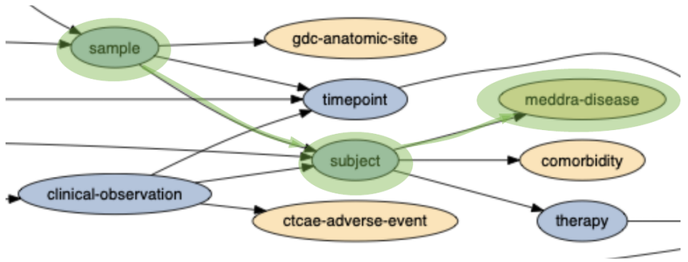
- Scatterplot of marker intensities
- Region and cell population selection and filtering

Enflame

Visual Query



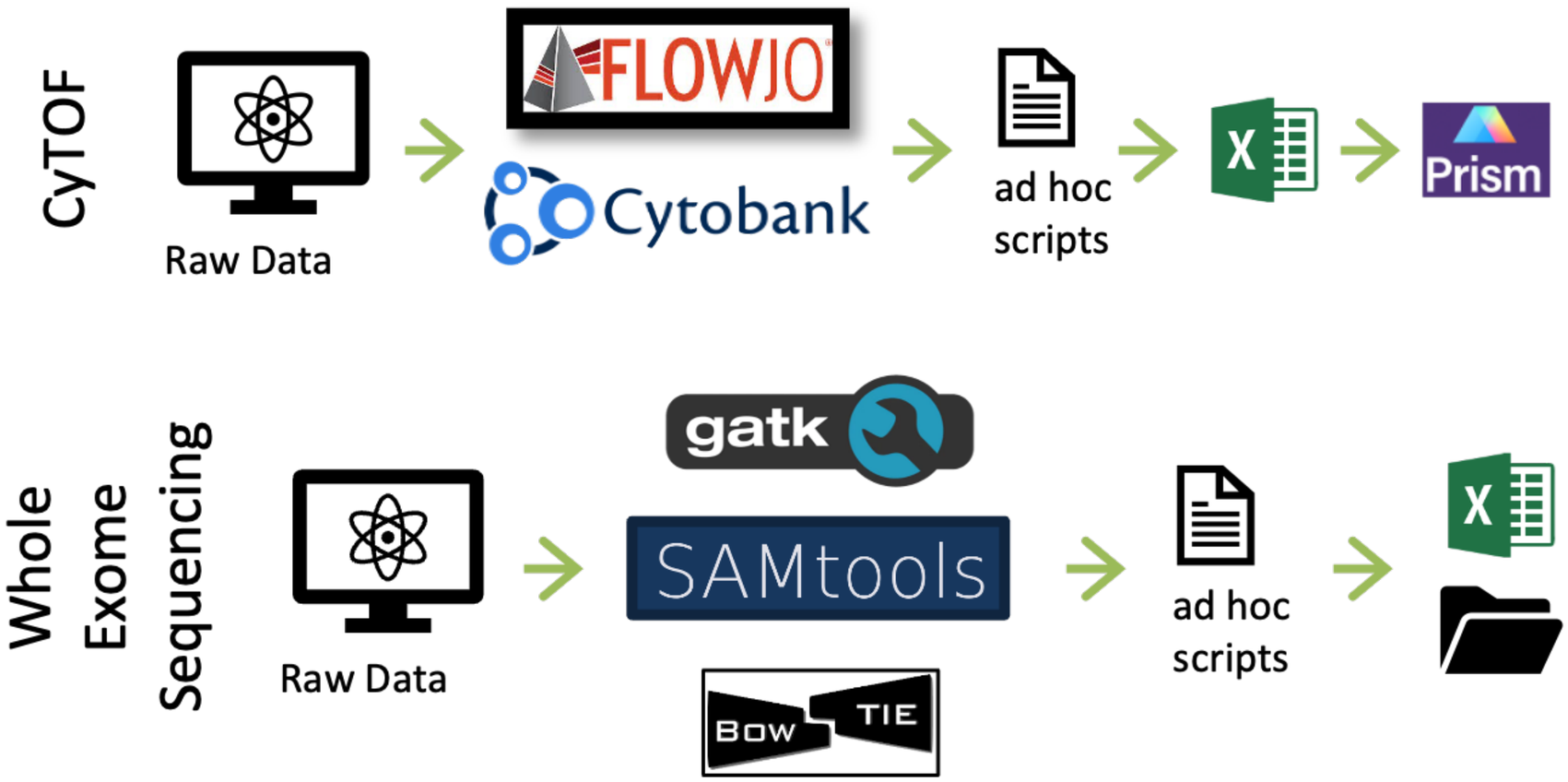
Enflame

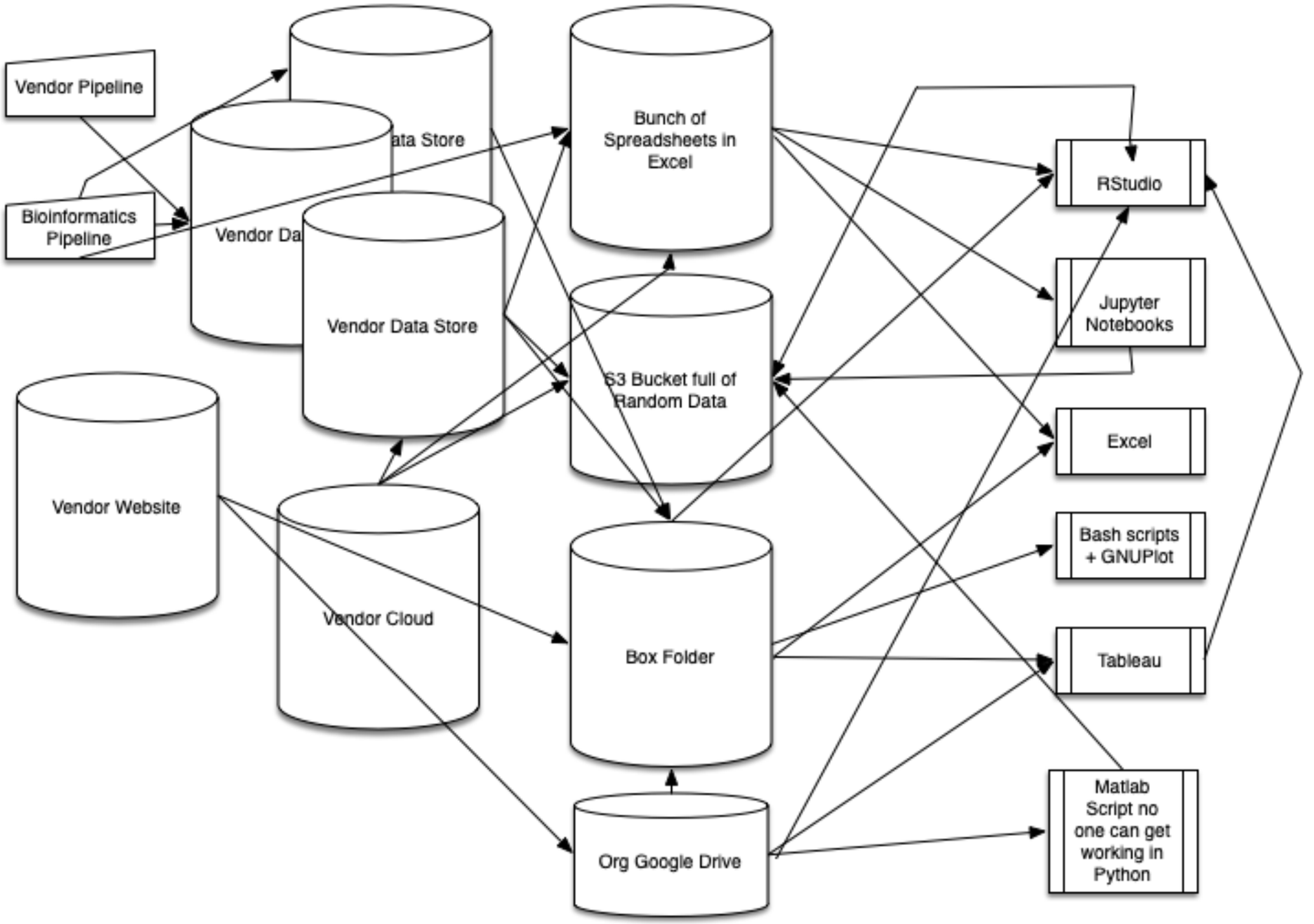


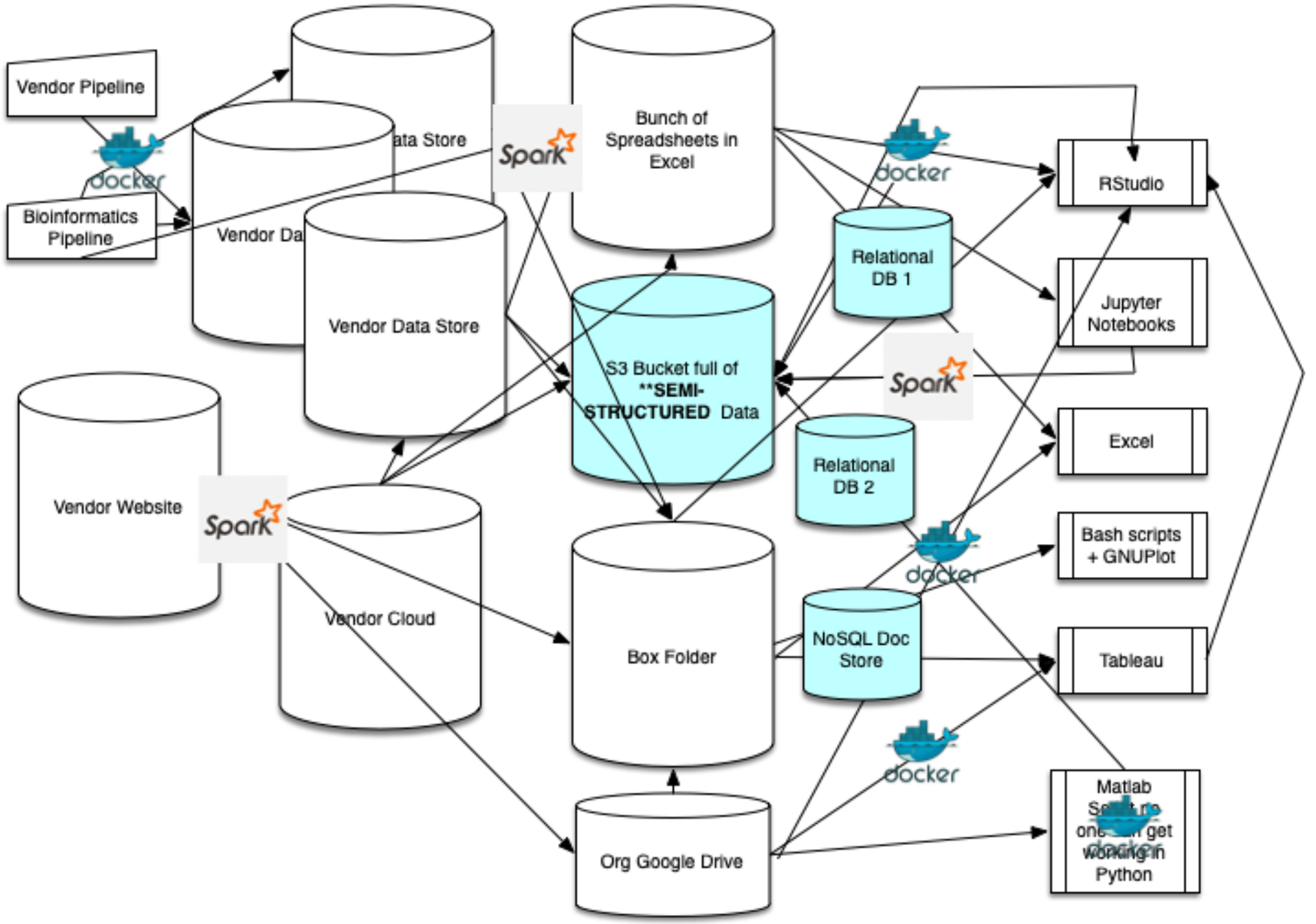
All of these block relationships are dynamically inferred from the DB schema

Using the CANDEL System

Example data history - before CANDEL




Vendor Pipeline
Bioinformatics Pipeline
ata Store
Vendor Da
Vendor Data Store
Vendor Website
Vendor Cloud
Bunch of Spreadsheets in Excel
S3 Bucket full of Random Data
Box Folder
Org Google Drive
RStudio
Jupyter Notebooks
Excel
Bash scripts + GNUPlot
Tableau
Matlab Script no one can get working in Python


Vendor Pipeline
Bioinformatics Pipeline
docker
ata Store
Vendor Da
Vendor Data Store
Spark
Bunch of Spreadsheets in Excel
S3 Bucket full of **SEMI-STRUCTURED Data
Vendor Website
Spark
Vendor Cloud
Box Folder
Org Google Drive
docker
Relational DB 1
Relational DB 2
Spark
NoSQL Doc Store
docker
docker
RStudio
Jupyter Notebooks
Excel
Bash scripts + GNUPlot
Tableau
Matlab
Python

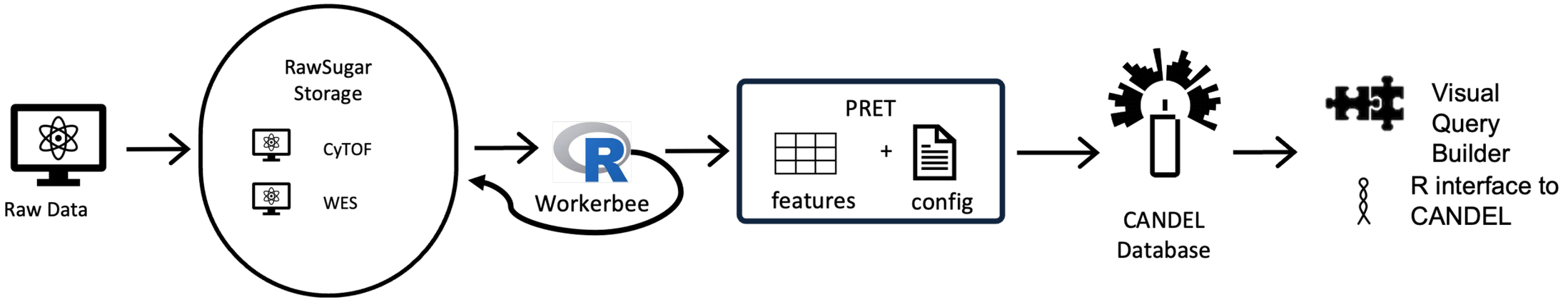
NEWS | 13 August 2021 | Correction 25 August 2021

Autocorrect errors in Excel still creating genomics headache

Despite geneticists being warned about spreadsheet problems, 30% of published papers contain mangled gene names in supplementary data.

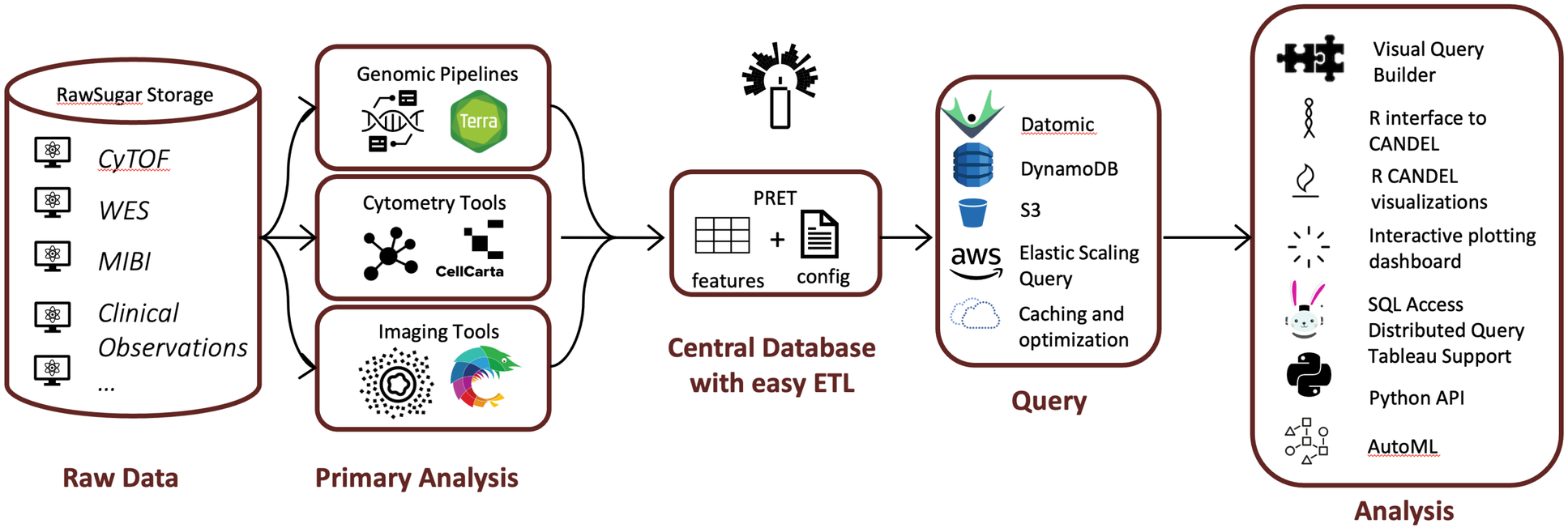

The horror, the horror.

Example data history - with CANDEL



CANDEL provides a unified set of processes, workflows, and validations from data ingestion to analysis

The CANDEL Ecosystem



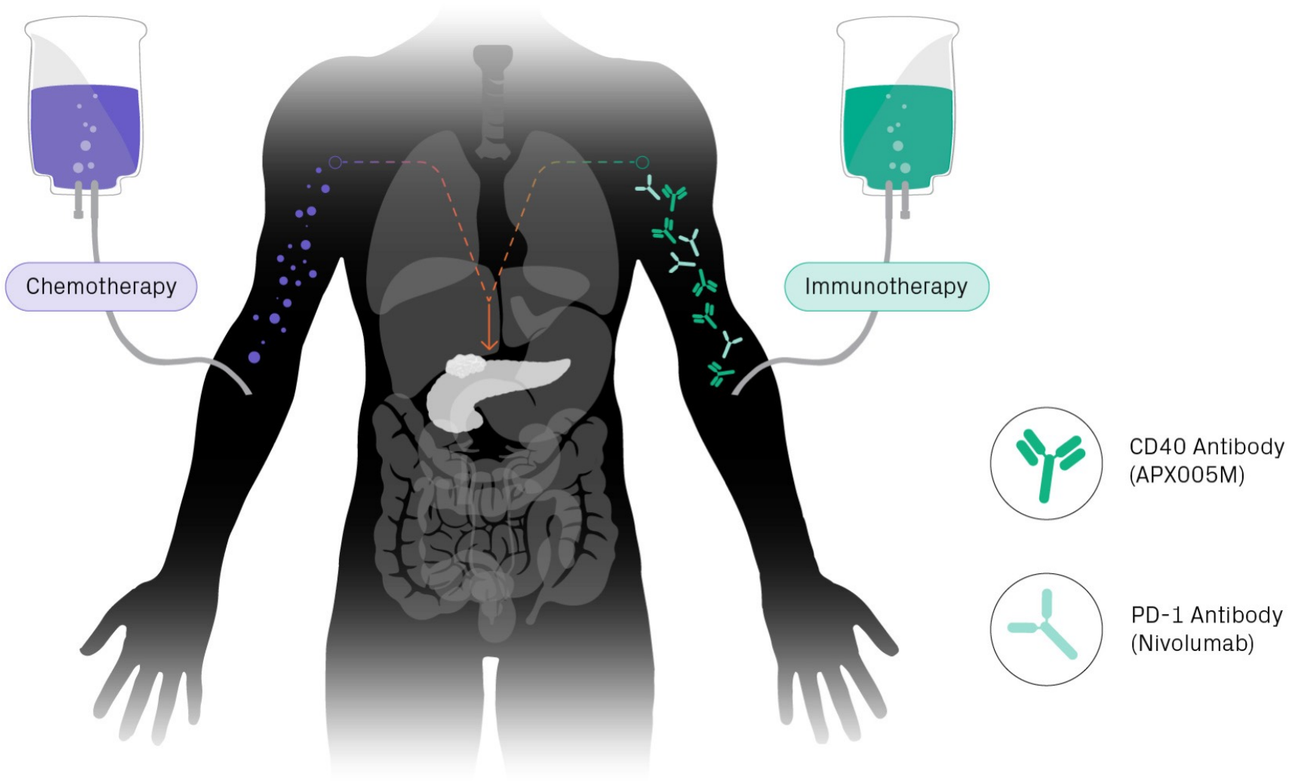
This unified set of processes supports a growing set of assay types and downstream analysis consumers

CANDEL-powered Research: Highlights

- Prince - Phase 2 combination trial in pancreatic cancer
- Morrison 1 - Effect of prior treatment(s) on IO therapy
- Radiohead - Relationship between adverse events and IO therapy
- Revolution - Novel IO combination trial for pancreatic cancer
- Morisson 2 - Multiplex imaging to uncover features of the TME
- Porter - Biomarkers of response in prostate cancer
- Amadeus - Can gut bacteria help destroy tumors? (in preparation)
- McGraw - Biomarkers for checkpoint response in hot vs. cold tumors (in preparation)

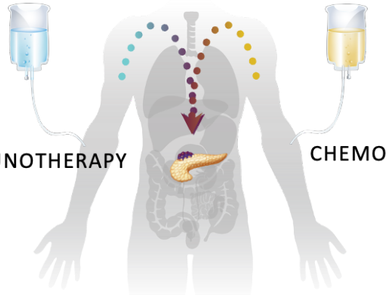
PRINCE Study:

- Phase 2 Trial of chemo + nivolumab and/or sotigalimab
- Multi-omic profiling:
  - Immunophenotyping by CyTOF
  - High-parameter flow cytometry of T lymphocytes
  - Serum proteomics profiling
  - Whole-exome and transcriptome sequencing
  - Multiplex imaging
- Nature Medicine, 2022



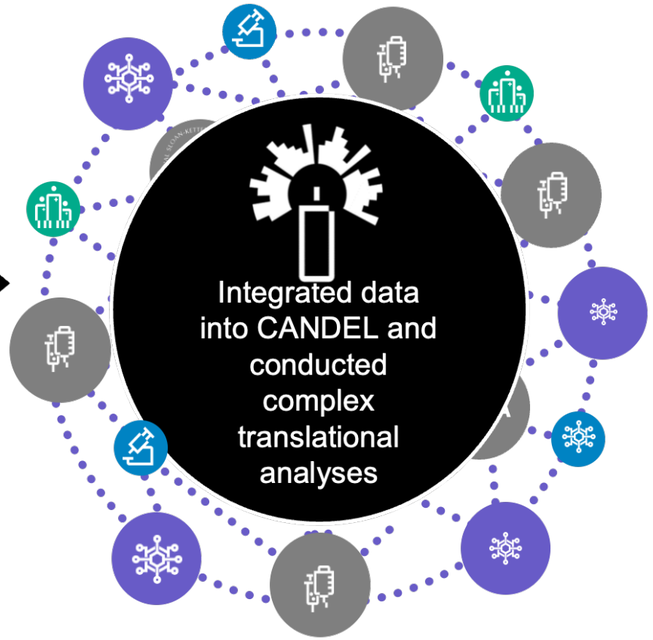
PRINCE Study: Key Findings

Evaluated 105 patients receiving chemo-immunotherapy



- nivolumab (anti-PD-1) + chemo
- APX005M (CD40 agonist) + chemo
- nivolumab + APX005M + chemo

Captured multi-omics data from blood and tumor biopsies collected across multiple timepoints and locations



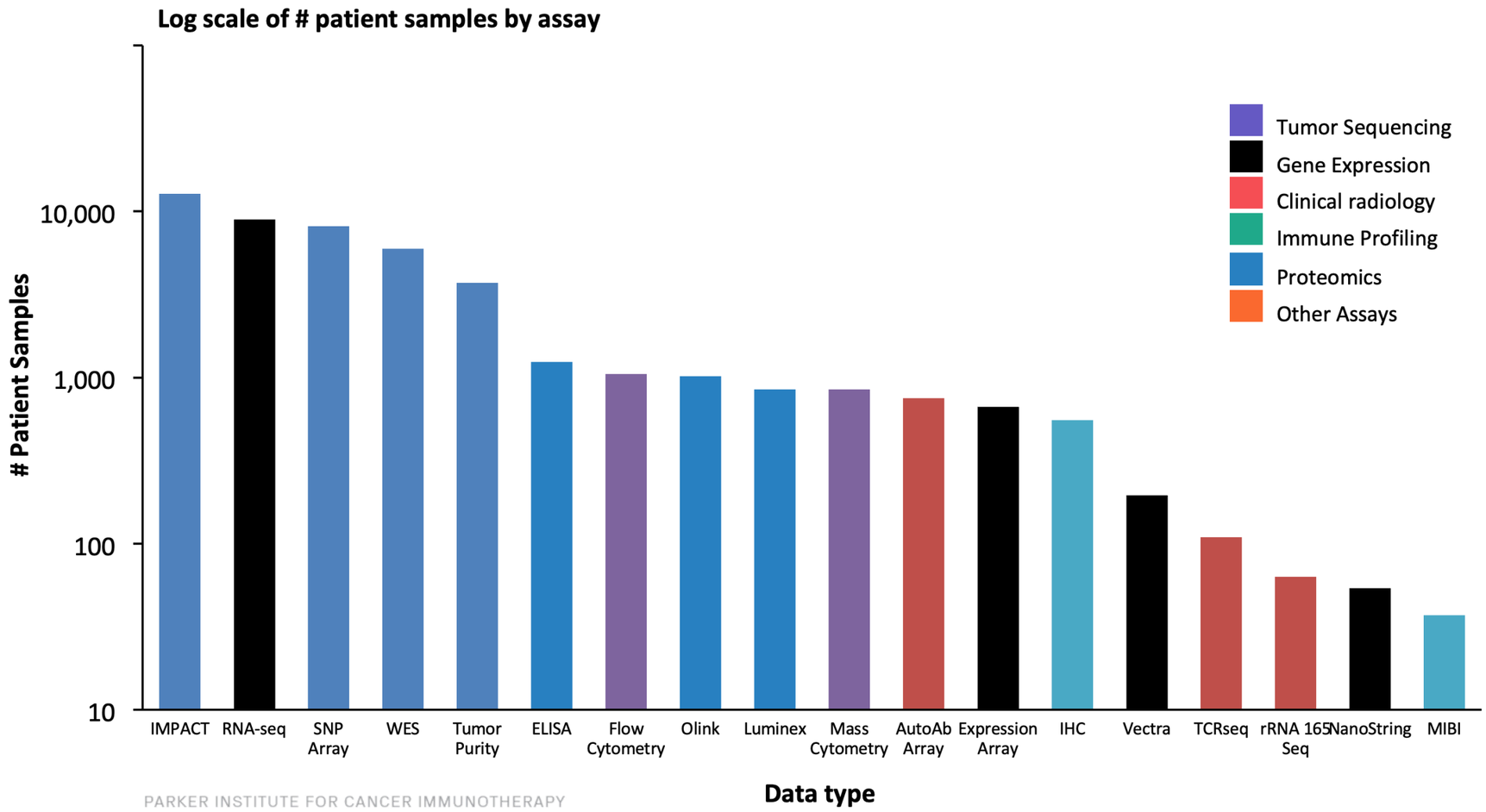
Results

- Improved understanding of each combo regimen's unique MOA and biological response
- Discovered predictive biomarkers for nivolumab and APX005M treatment arms associated with OS benefit
- Demonstrated evidence of antagonism for two immunotherapy combinations

Impact

- Lead investigator has designed a new clinical trial based on the impact of this work.
- Created an opportunity for patients to receive life-extending treatment based on novel biomarkers

Data in CANDEL



Things that went well and things that didn’t

Biggest Win: Designing Around Schema Evolution

If you take nothing else from this talk:

- The schema constantly evolved
- We handled its evolution without code changes by doing things in a data driven fashion
  - Pret infers its data compilation logic from schema and metamodel
  - Enflame, etc. also generates queries, etc. from inspection of schema

Data Harmonization Effort at PICI: Less Work, More Consistent

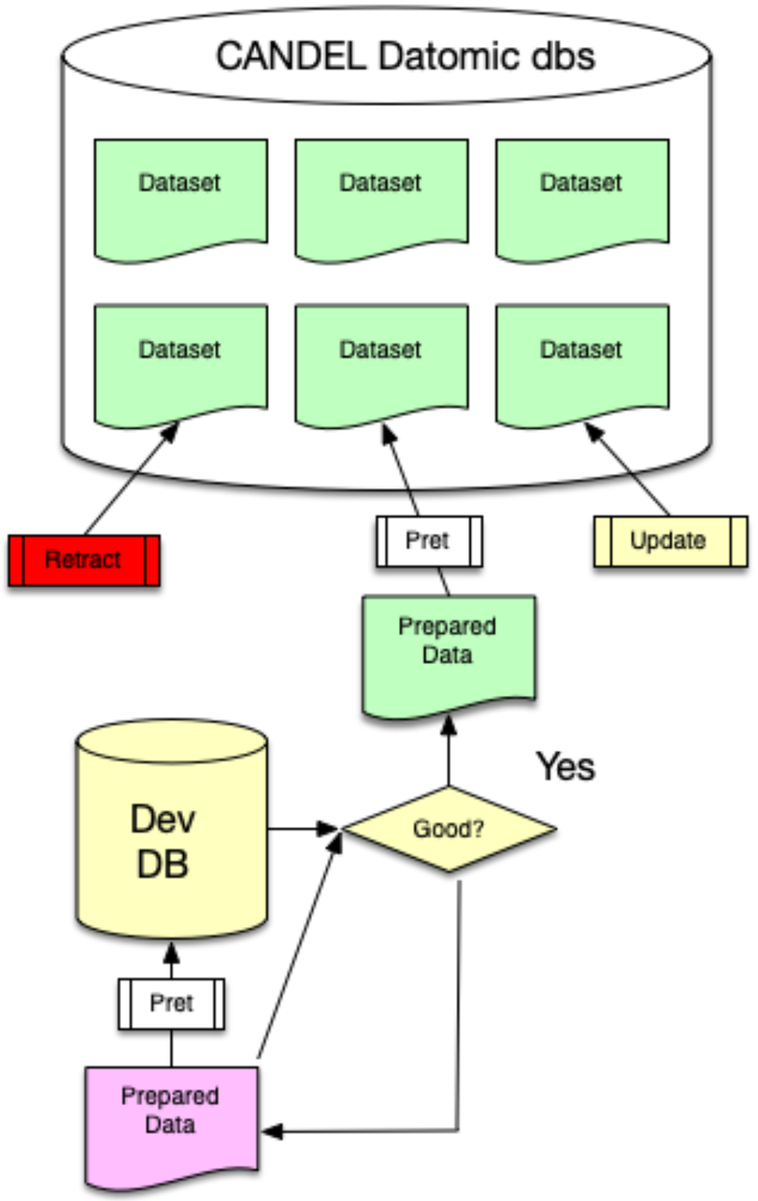
- Data harmonization effort greatly reduced
- Most common pain point was: “I don’t know what this spec error means
- Good problem to have!

These were always real problems with the data! We kept them out of research!

Things We Had to Fix

Handling datasets built against different schema versions

It took a long time to get acceptable solution for dataset evolution, versioning, etc.


CANDEL Datomic dbs
Dataset
Dataset
Dataset
Dataset
Dataset
Dataset
Retract
Pret
Update
Prepared Data
Yes
Dev DB
Good?
Pret
Prepared Data

CANDEL Dashboard

dataset	description	subjects	samples	treatments	mass-cy...
Abida2019	Genomic correlates of clinical outcome in advanced prostate cancer	429	444	ABIRATERONE+VELIPA...	
aerosmith-pici0009-t1...	AEROSMITH patients who developed T1D after ICI & matched controls	30	166	CTLA-4+PD-1, Other, P...	
bailey2016	Genomic analyses identify molecular subtypes of pancreatic cancer	456	456		
eagles	EAGLES Test Dataset	405	2182	NIVO, OTHER, IPI, NIVO...	
gide2019	Distinct Immune Cell Populations Define Response to Anti-PD-1 Mono...	105	91	Nivolumab-3mgkg-2wk...	
hayashi2020	RNA-seq on FFPE samples from 2,928 sections	123	253		
Miao2018	Genomic correlates of response to immune checkpoint blockade in mi...	249	249	anti-CTLA-4 + anti-PD-...	
moffitt-2015	Virtual Microdissection of Pancreatic Ductal Adenocarcinoma Reveals ...	341	357		
Morrison-38	Clinical Data	582	1992	CTLA4-to-PD1, PD1-to...	
morrison2_v1	MORRISON 2	5	37		
msk-impact-2017	10,000 patient in the door data using SNP panels	10336	10945		

Columns Filters

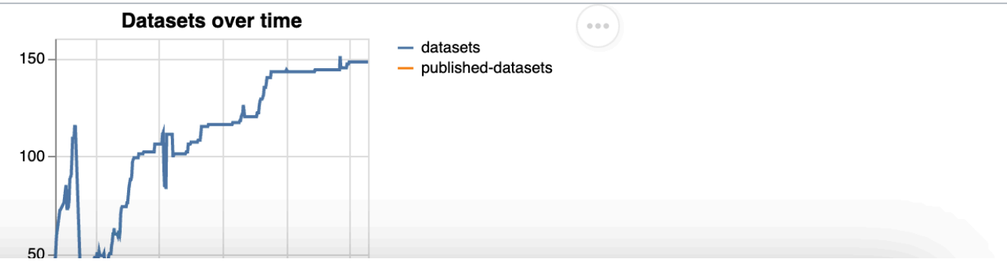
Rows: **148**

Summary

From 30 datasets (published only):

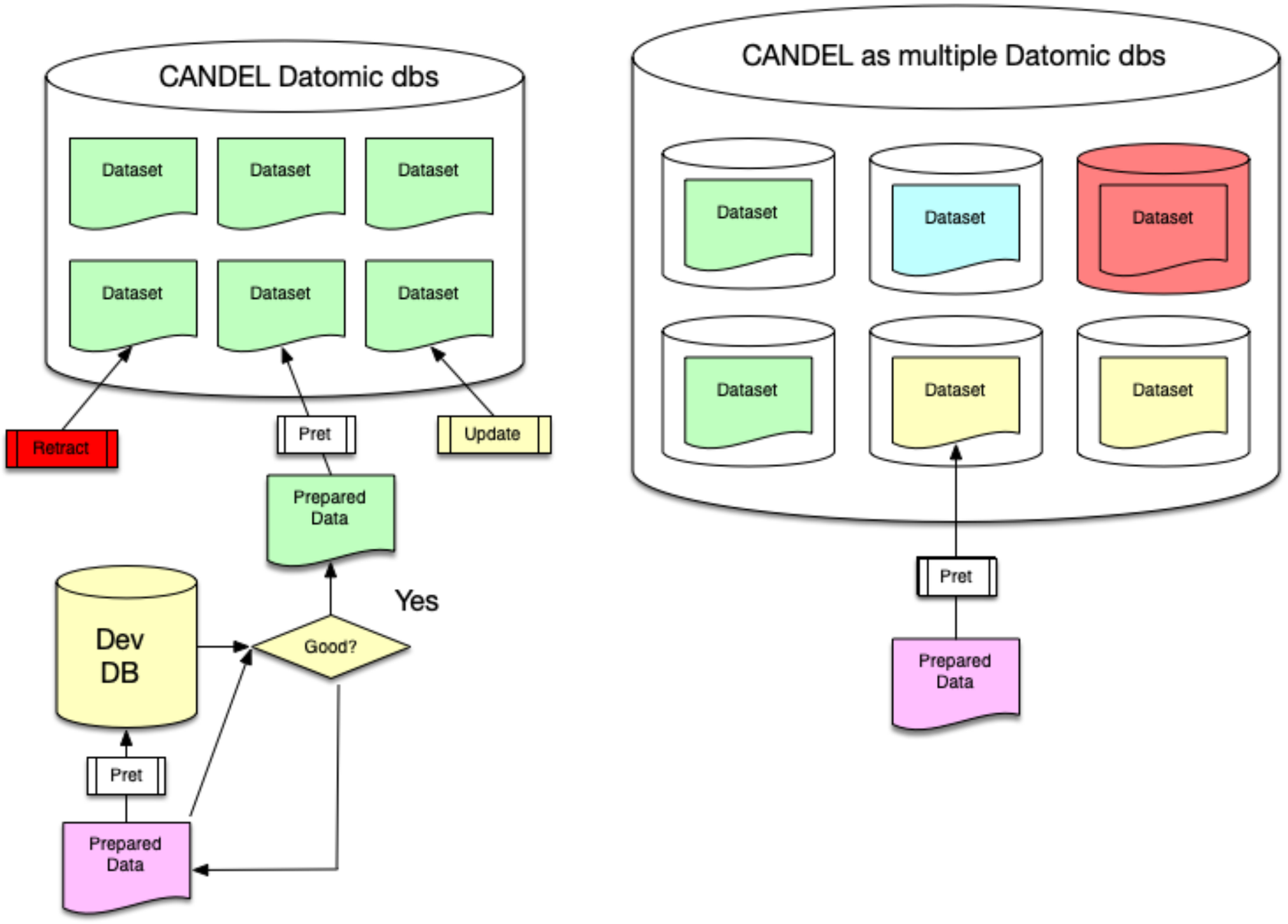
Subjects: 20080

Samples: 34753



Just use multiple Datomic dbs stupid

- transactional boundaries didn’t really matter (entire dataset split across transactions in alrge batches)
- transactions were never 'live'.
- Multiple versions of db could be stood up


CANDEL Datomic dbs
Dataset
Dataset
Dataset
Dataset
Dataset
Dataset
Retract
Pret
Update
Prepared Data
Yes
Dev DB
Good?
Pret
Prepared Data
CANDEL as multiple Datomic dbs
Dataset
Dataset
Dataset
Dataset
Dataset
Dataset
Pret
Prepared Data

Big data and big-ish data

- Big data doesn’t fit well in Datomic, even when it maps to Datoms
- Single cell and image derived features forced data out of Datomic.
  - Solution solved some pain we'd been experiencing with larger measurement sets.

Measurement Matrix

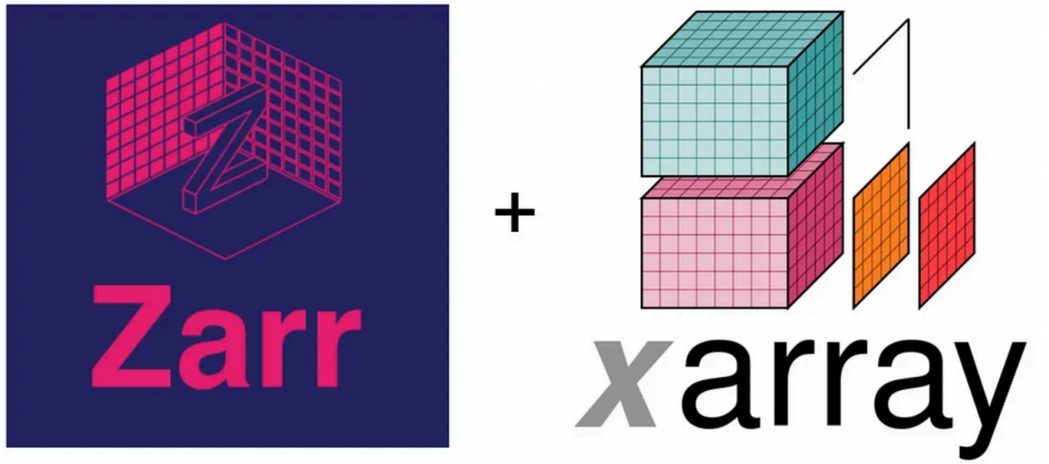
```
:measurement-matrices
[{:name "screening-rna-seq"
  :measurement-type :measurement/read-count
  :pret.matrix/input-file "dense-rnaseq.tsv"
  :pret.matrix/format :pret.matrix.format/dense
  :pret.matrix/column-attribute :measurement-matrix/gene-products
  :pret.matrix/indexed-by {"sample.id" :measurement-matrix/samples}}
 {:name "single cell counts"
  :measurement-type :measurement/read-count
  :pret.matrix/constants {:measurement-matrix/samples "SYNTH-SC-DATA-01"}
  :pret.matrix/input-file "short-processed-counts.tsv"
  :pret.matrix/format :pret.matrix.format/sparse
  :pret.matrix/indexed-by {"barcode" :measurement-matrix/single-cells
                           "hugo" :measurement-matrix/gene-products}}]}]}]
```

References to entities in Datomics (as unordered set), numeric values not.

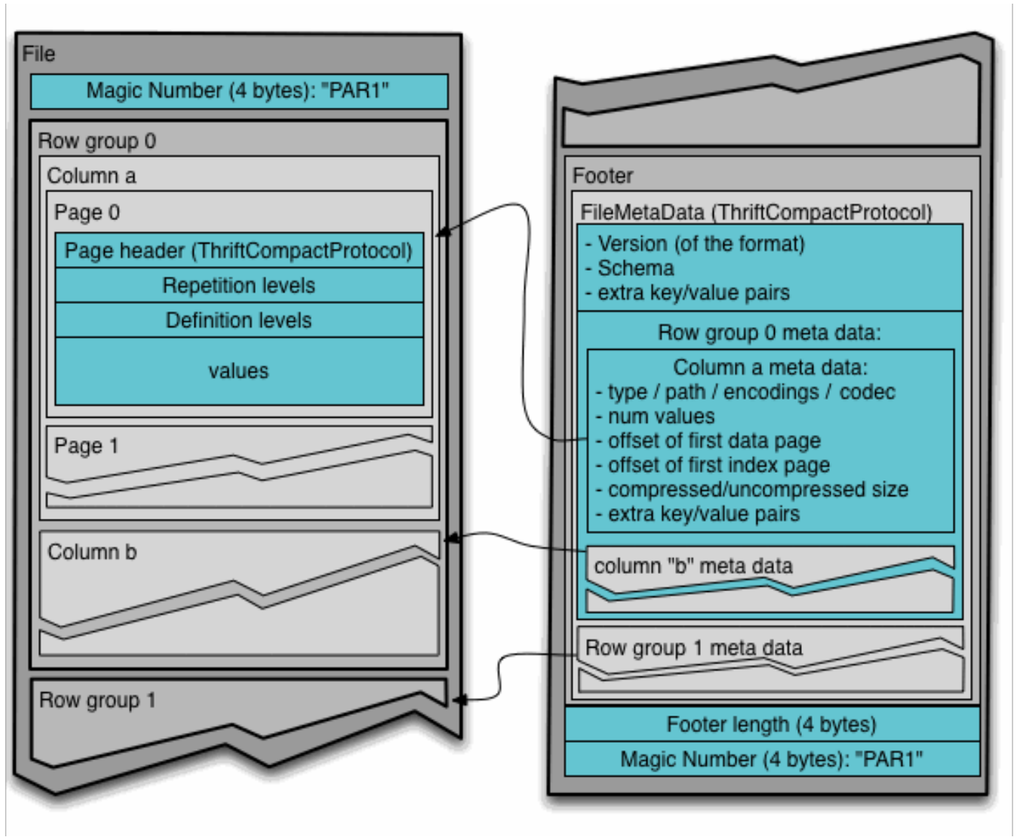
GENE	SAMPLE	VALUE
GENE1	SAMPLE1	1.0
GENE2	SAMPLE1	0
GENE3	SAMPLE1	3.7
GENE3	SAMPLE1	3.7

Measurement Matrix

- We did this in TSV for R and ease of import, arrow/parquet probably more efficient for tables
- For cloud systems, it probably makes more sense to go into zarr or something similar.
- If you want to do that, it would be a minimal code change.



- Other formats could be built in for matrix types, eg. Parquet or Zarr ecosystems.



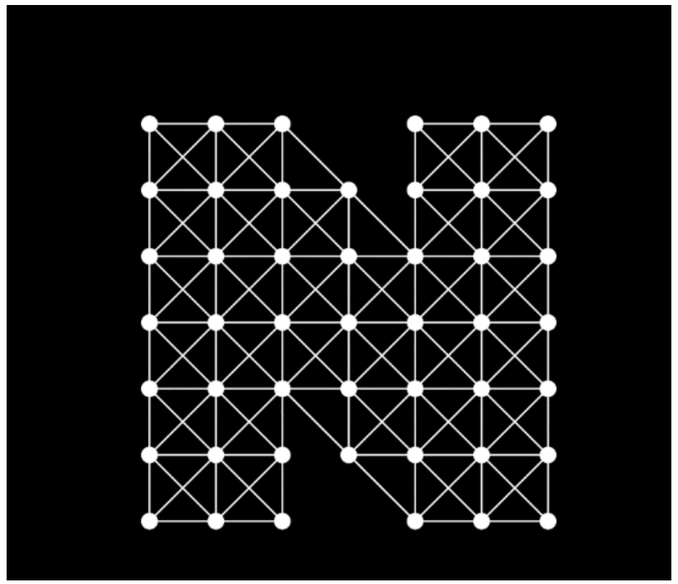
What we're open sourcing

- This is all open sourced now at CANDELBio
- Data harmonization problems aren’t unique to biology
- We expect data unification/harmonization needs across science could benefit from this toolkit
- We expect data science in general or even other ETL workflows can benefit

Where we are now

Getting data at a scale that will support ML, data-driven insights from biology.

Where we are now



RCRF
THE RARE CANCER RESEARCH FOUNDATION

Noetik's mission

- Building large scale datasets from human data that are fit-for-purpose for deep learning.
- Tight feedback loop between comp bio, machine learning, and the lab process.

RCRF is dedicated to curing rare cancers through strategic investments and innovative collaborations that facilitate effective research and accelerate deployment of promising therapies

Thanks

https://github.com/CANDELBio