Multiomics Provider

Institute for Systems Biology

Accelerating *the creation of high-value* **knowledge providers** *from* **raw data**

Gustavo Glusman, PhD

Jennifer Hadlock, MD

Ilya Shmulevich, PhD

Guangrong Qin, PhD

Ryan Roper, MS

Denise Mauldin, MS

Cheng Dai, MEng

Thank you
Theo Knijnenburg
John Earls
Sui Huang
Jared Roach

ISB

NIH National Center for Advancing Translational Sciences

Milestones

- [ ] **M1.1 Ongoing support of Big GIM**
- [ ] **M1.2 Prototype DOCKET Overview**
- [ ] **M1.3 Prototype DOCKET Compare**
- [ ] **M1.4 Prototype DOCKET Integrate**
- [ ] **M1.5 Gather input on DOCKET features**
- [ ] **M1.6 Collaborate on Reasoner Standard API**

Multiomics Data

ISB Wellness

GDSC Cancer

TCGA Cancer

Deliverables

DOCKET code, with Docker container

github.com/PriceLab/DOCKET

Instructions for running prototypes

drive.google.com/drive/u/1/folders/19CT2bu1kzVnXgORhgIQijJd7x8O8Ez6D

Translator Drive > Knowledge Providers

Try-it-yourself document

Sample data

Big GIM: **G**ene **I**nteraction **M**iner, and

Big CLAM: **C**ell **L**ine **A**ssociation **M**iner:

biggim.ncats.io/api

github.com/PriceLab/translator-bigquery-api

RAW DATA → KNOWLEDGE

wellness
cancer
EHRs
many more!

handcrafted analytics → expert curation, validation → scientific communication / software engineering

papers!

APIs!

Big GIM

Big CLAM

DOCKET:

automated analytics → facilitated curation → standardized deployment

- "Data Overview, Comparison, and Knowledge Extraction Tool"
- Utilities for extracting knowledge, evaluating, visualizing, serving...
- The resulting dossier, a resource readily available for ARAs

(1) Contribute multiomic Knowledge Providers

(2) Automate the creation of KPs via DOCKET

Big GIM: Gene Interaction Miner

Developed during Translator Phase 1 (Theo & John, Blue Team)
In use by workflows, collaborators, ARAs; M1.1: ongoing support

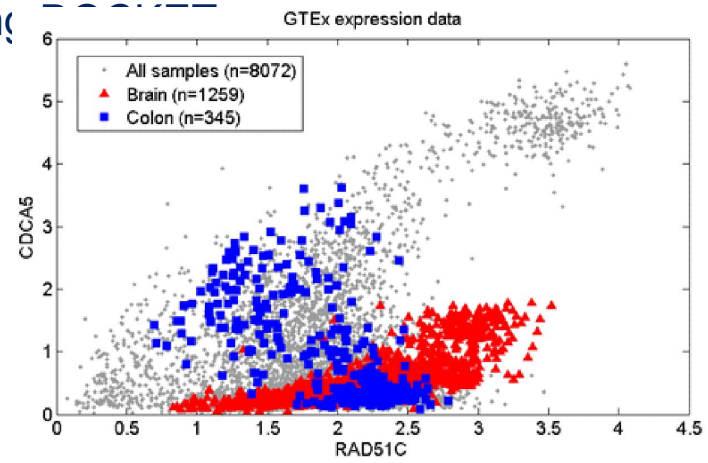
- **Precompute** all pairwise gene expression correlations, in tissue context
- **Curate** to ensure correctness
- **Store** results in Google Cloud Bigtable
- **Serve** via API as proto-KP on gene similarity

Gene 1	Gene 2	GTEX expression correlations		TCGA expression correlations		GIANT functional interaction		BioGRID direct interaction
tissue →		Skin	Colon	SKCM	COAD	Skin	Colon	
RAD51C	CDC6	0.65	-0.68	0.47	0.51	0.62	0.52	0
KPNA3	BRCA2	0.70	-0.60	0.65	0.57	0.08	0.06	0
PDS5B	BRCA2	-0.27	-0.57	0.66	0.69	0.12	0.10	1
...	...							

410M rows

400 columns

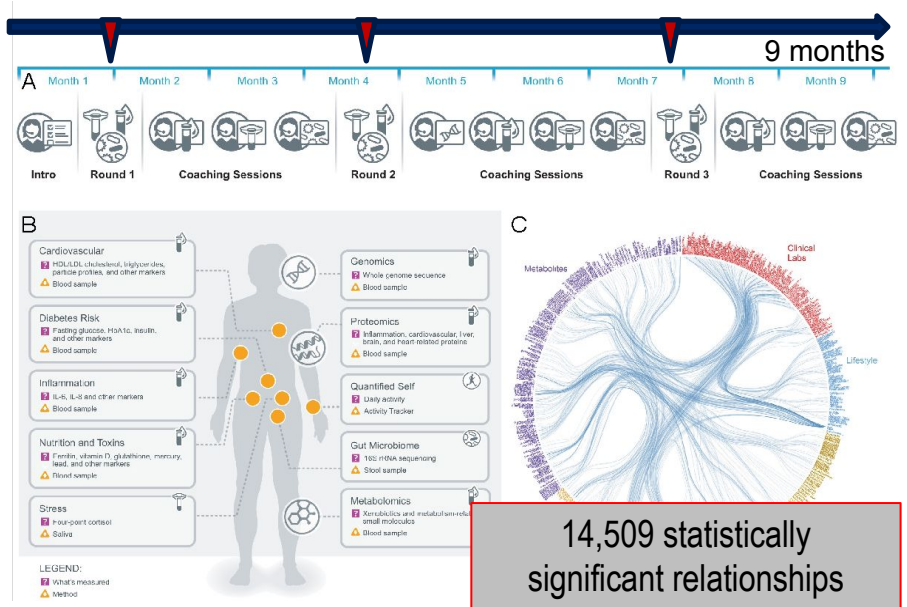
sing



Multiomics Data to Knowledge: Wellness

2016, Pioneer 100 Project: N=108

Whole genomes, 213 clinical lab tests, 643 metabolites, 392 proteins, gut microbiome



14,509 statistically significant relationships between individual features

nature biotechnology

A wellness study of 108 individuals using personal, dense, dynamic data clouds

Nathan D Price[1,2,6,7], Andrew T Magis[2,6], John C Earls[2,6], Gustavo Glusman[1], Roie Levy[1], Christopher Lausted[1], Daniel T McDonald[1,5], Ulrike Kusebauch[1], Christopher L Moss[1], Yong Zhou[1], Shizhen Qin[1], Robert L Moritz[1], Kristin Brogaard[2], Gilbert S Omenn[1,3], Jennifer C Lovejoy[1,2] & Leroy Hood[1,4,7]

Today: N~5000, 96 data snapshots

Genetics, clinical labs, metabolites, proteins, microbiome, medications, activity, sleep, weight, hormones, assessments of vitals, diet, digestive health, lifestyle, etc.

nature biotechnology

Article | Published: 02 September 2019

Blood metabolome predicts gut microbiome α-diversity in humans

Journals of Gerontology: Biological Sciences
cite as: J Gerontol A Biol Sci Med Sci, 2019, Vol. 74, No. S1, S52–S60
doi:10.1093/gerona/glz220

THE GERONTOLOGICAL SOCIETY OF AMERICA

OXFORD

Healthy Longevity 2019: Supplement Article

Multi-Omic Biological Age Estimation and Its Correlation With Wellness and Disease Phenotypes: A Longitudinal Study of 3,558 Individuals

SCIENTIFIC REPORTS

OPEN

Genetic Predisposition Impacts Clinical Changes in a Lifestyle Coaching Program

Cell Reports
Article

A Multi-omic Association Study of Trimethylamine N-Oxide

Ohad Manor,[1,3,*] Niha Zubair,[1] Matthew P. Conomos,[1] Xiaojing Xu,[1] Jesse E. Rohwer,[1] Cynthia E. Krafft,[1] Jennifer C. Lovejoy,[1,2] and Andrew T. Magis[1]
[1]Arivale, Inc., Seattle, WA 98104, USA
[2]Institute for Systems Biology, Seattle, WA 98109, USA

International Journal of Obesity

Original Article | Published: 12 September 2017

Habitual sleep duration and sleep duration variation are independently associated with body mass index

X Xu, M P Conomos, O Manor, J E Rohwer, A T Magis & J C Lovejoy

New Results

Gut Microbiome Pattern Reflects Healthy Aging and Predicts Extended Survival in Humans

Tomasz Wilmanski, Christian Diener, Noa Rappaport, Sushmita Patwardhan, Jack Wiedrick, Jodi Lapidus, John C. Earls, Anat Zimmer, Gustavo Glusman, Max Robinson, James T. Yurkovich, Deborah M. Kado, Jane A. Cauley, Joseph Zmuda, Nancy E. Lane, Andrew T. Magis, Jennifer C. Lovejoy, Sean M. Gibbons, Leroy Hood, Eric S. Orwoll, Nathan D. Price

doi: https://doi.org/10.1101/2020.02.26.966747

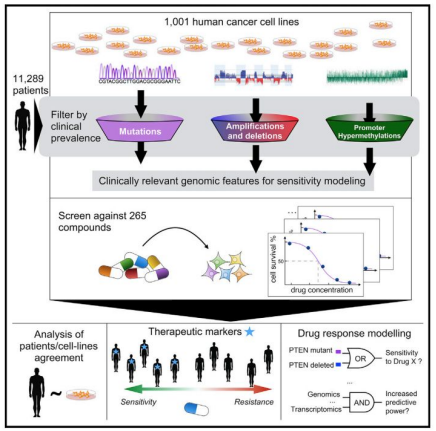
Multiomics Data to Knowledge: Cancer

Which **gene mutations** in any of **these pathways** are associated with sensitivity or resistance to any of **these drugs** in the cell lines from **these tumor types**?

Cell

A Landscape of Pharmacogenomic Interactions in Cancer

Graphical Abstract



Authors

Francesco Iorio, Theo A. Knijnenburg, Daniel J. Vis, ..., Julio Saez-Rodriguez, Ultan McDermott, Mathew J. Garnett

Correspondence

um1@sanger.ac.uk (U.M.), mg12@sanger.ac.uk (M.J.G.)

In Brief

A look at the pharmacogenomic landscape of 1,001 human cancer cell lines points to new treatment applications for hundreds of known anti-cancer drugs.

Highlights

- We integrate heterogeneous molecular data of 11,289 tumors and 1,001 cell lines
- We measure the response of 1,001 cancer cell lines to 265 anti-cancer drugs
- We uncover numerous oncogenic aberrations that sensitize to an anti-cancer drug
- Our study forms a resource to identify therapeutic options for cancer sub-populations

Accession Numbers

GSE68379
E-MTAB-3610

Mutation + **Drug response**

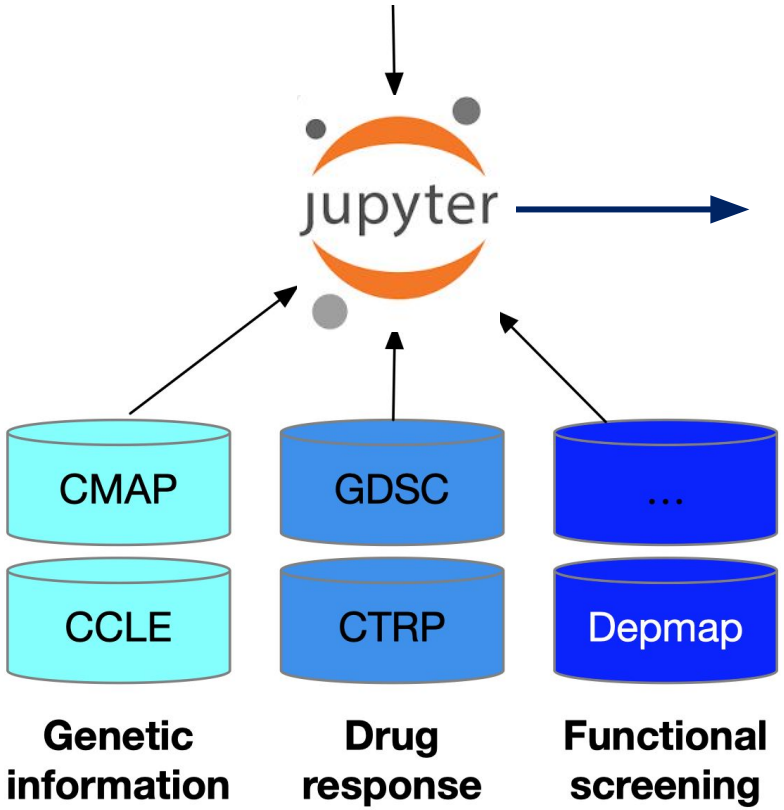
Expression

Methylation

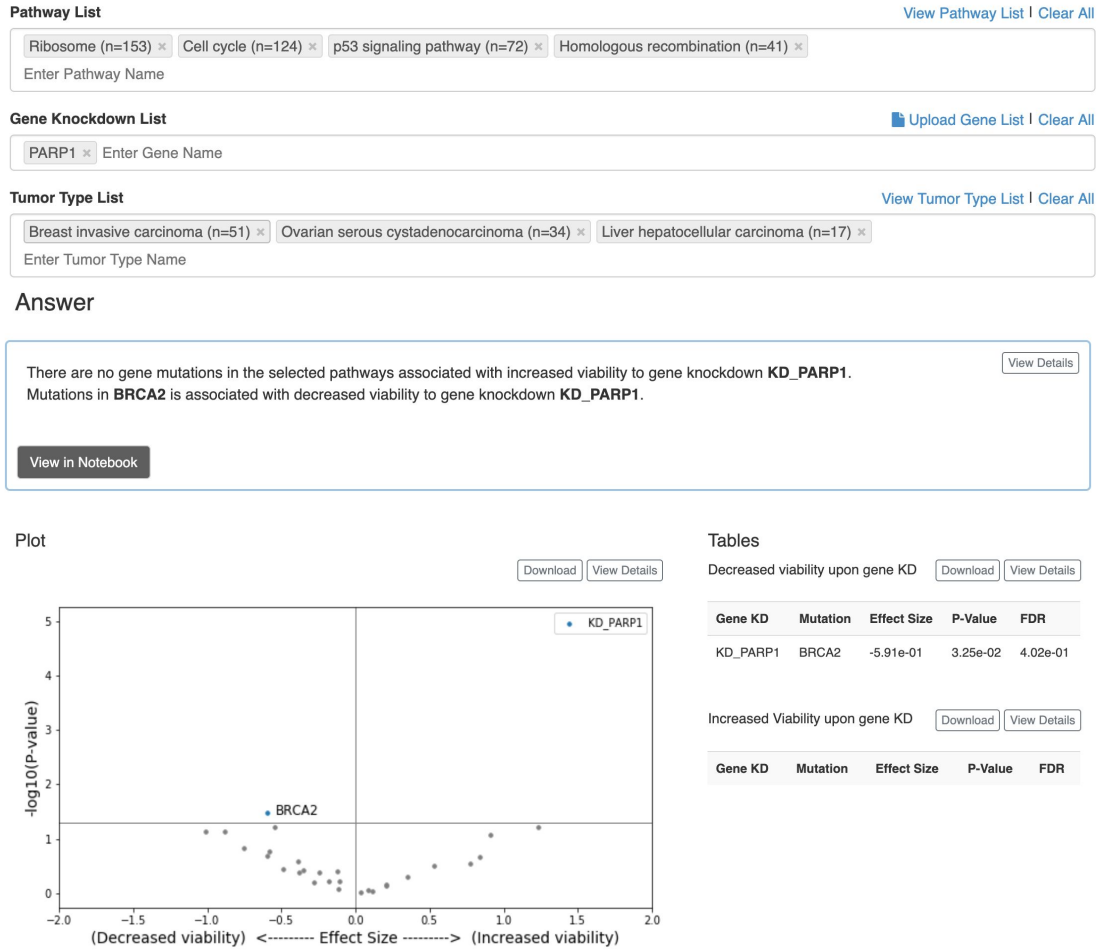
1001 cancer cell lines, 265 anti-cancer drugs
Over 30 tumor types

Multiomics Data to Knowledge: Cancer

Which **gene mutations** in any of **these pathways** are associated with relative increased or decreased viability of cells that have been exposed to any of **these gene knockdowns** in the cell lines from **these tumor types**?



Statistical modeling



EHR Phenotype Data to Knowledge

Providence St. Joseph Health

Clin Transl Sci. 2019 Jul;12(4):329-333.

Clinical Data: Sources and Types, Regulatory Constraints, Applications.

Ahalt SC[1], Chute CG[2], Fecho K[1], Glusman G[3], Hadlock J[3], Taylor CO[2], Pfaff ER[4], Robinson PN[5], Solbrig H[2], Ta C[6], Tatonetti N[6], Weng C[6]; Biomedical Data Translator Consortium.

Opportunity

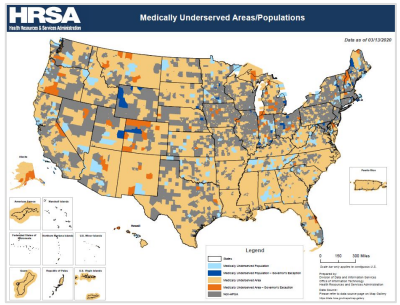
*Private data to public knowledge**

4.3 million patients

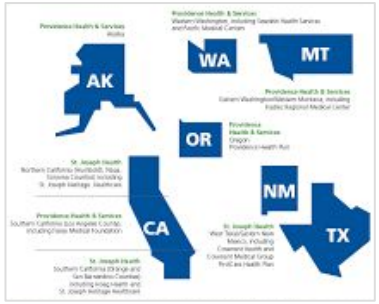
- 7 states from Alaska to Texas
- 51 hospitals, 829 clinics
- Medically underserved areas
- Longitudinal records

30 billion data elements

- 330 M encounters
- 27 M diagnoses
- 7 M vital signs
- 260 M drug details
- 947 M lab results
- 11 M surgical logs
- Full text notes



HRSA 2020



PSJH 2020

Challenge

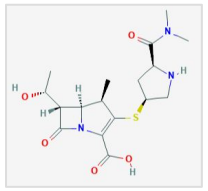
Systemic biases are refractory to scale

Accurate phenotype requires higher fidelity than ICD-10 code
Biomedical logic supports substratification and classification

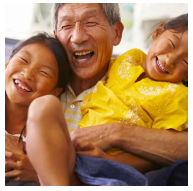
Examples:

- HbA1C correlations, going beyond billing label for diabetes
- Empirical treatment for suspected sepsis vs. clinical sepsis
- Care spread across large catchment areas
- Maternal care, links between maternal and fetal records

Code Black 2020



PubChem 2020



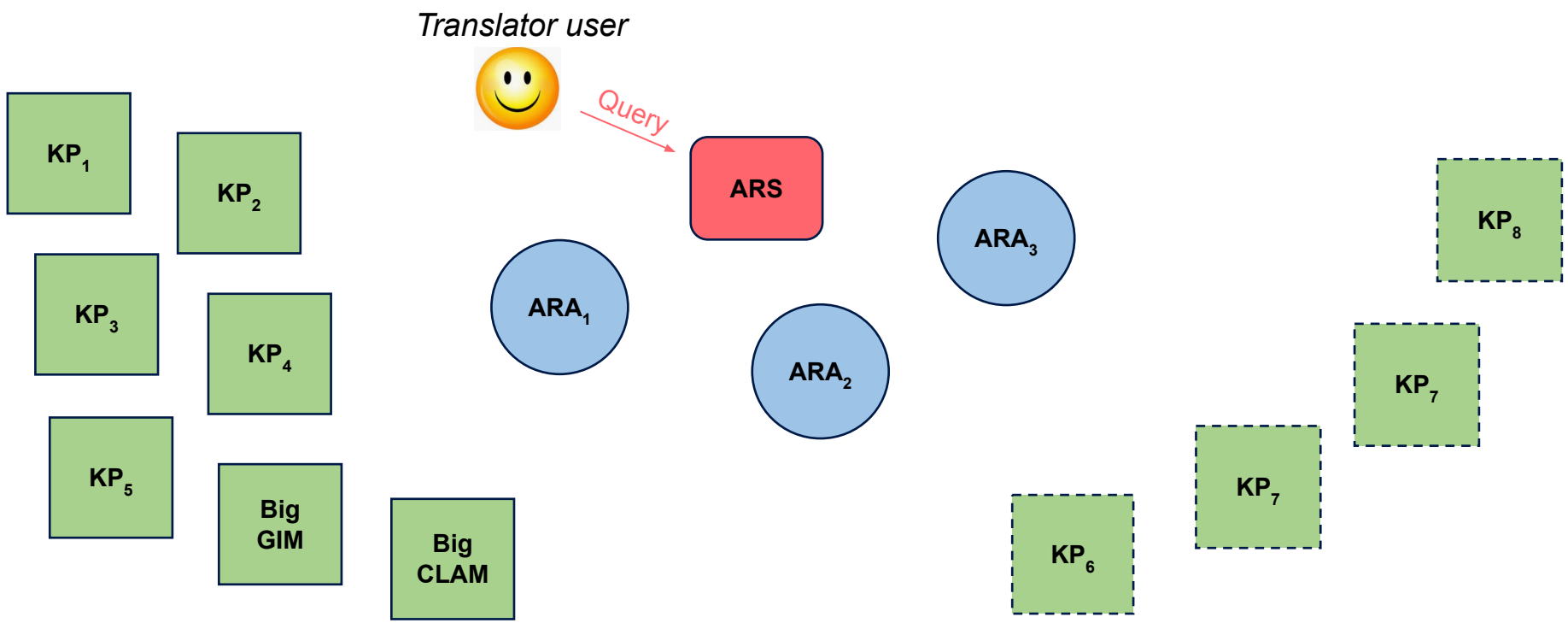
PSJH 2020

CDC 2015

The following list provides a sense of the breadth of data types to be considered for the Biomedical Data Translator, but it is not exhaustive:

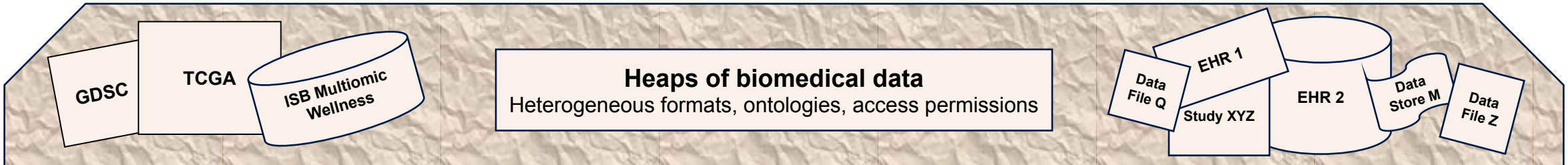
- Diseases
- Signs of disease
- Symptoms (from patient registries and natural history studies)
- Patient-reported outcomes
- Electronic health records
- Clinical encounters
- Prescription data
- Health insurance claims data
- Diagnostic labs
- Biomedical imaging data
- Adverse event reports
- Biomarkers
- Organ systems
- Subanatomy
- Tissue types
- Cell types
- Cell lineages
- Cell processes
- Organelles
- Orthologs/Animal models
- Microbiome data
- Molecular mechanisms
- Signaling pathways
- Molecular and cellular networks
- Environmental factors
- Disease etiologies
- Proteomes
- Proteins
- Post translational modifications (PTMs)
- Co-factors
- Transcriptomes
- Epigenomes
- Genes
- Genetic mutations
- Functional polymorphisms
- Therapeutic interventions
- Intervention exposure
- Pharmacokinetics/Pharmacodynamics
- Clinical trial data

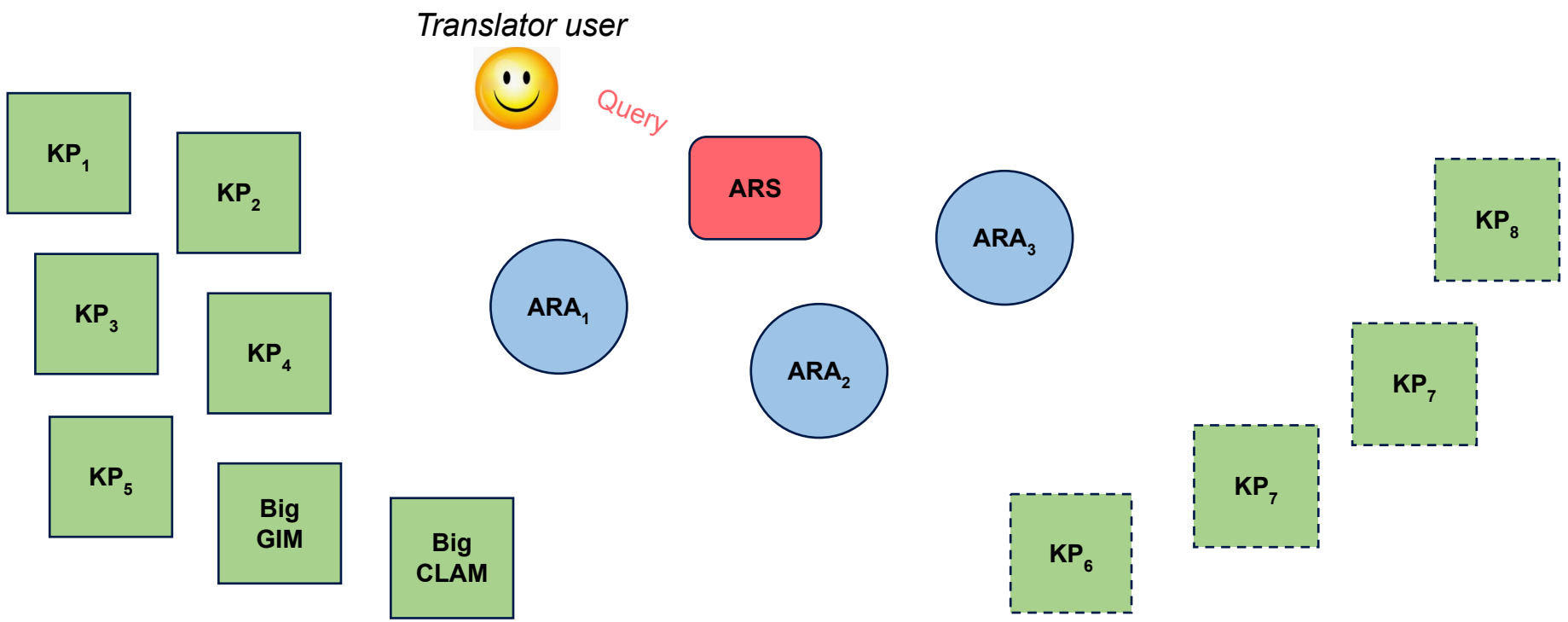
https://ncats.nih.gov/translator/about/data

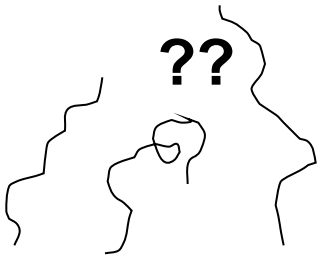


The important question, motivating DOCKET:

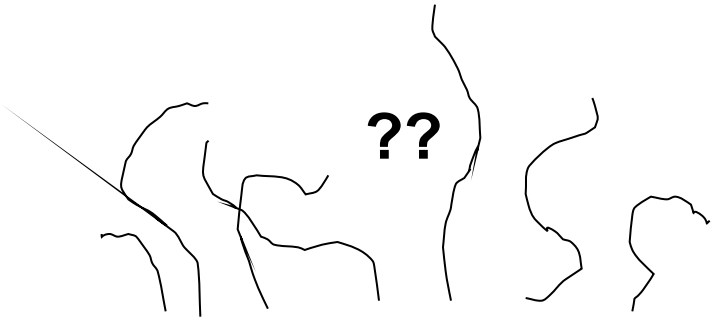
How are we going to generate KPs to cover all desired data types? (And how much will it cost?)



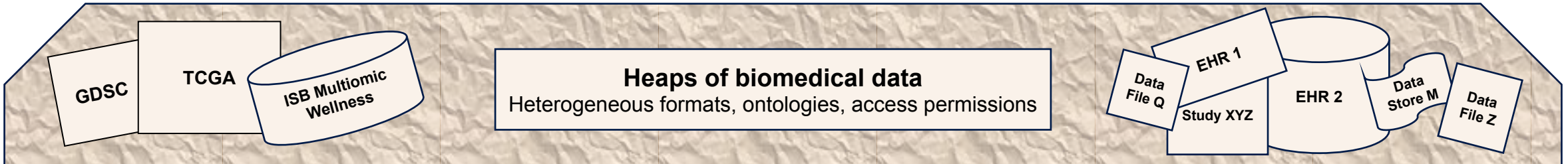


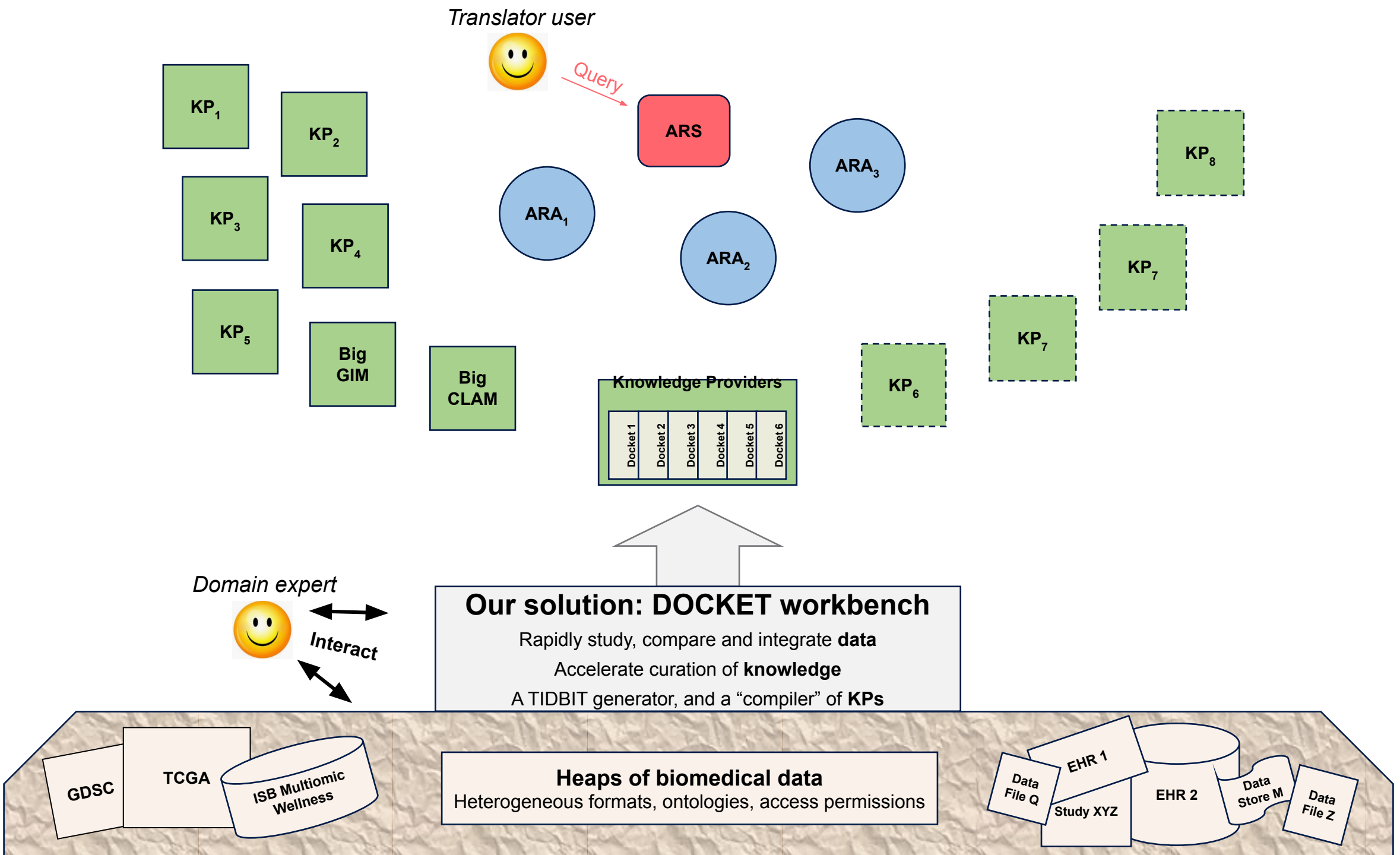


Domain expert



The first mile problem




Translator user
Query
ARS
KP1
KP2
KP3
KP4
KP5
Big GIM
Big CLAM
ARA1
ARA2
ARA3
KP8
KP7
KP7
KP6
Knowledge Providers
Docket 1
Docket 2
Docket 3
Docket 4
Docket 5
Docket 6
Domain expert
Interact
Our solution: DOCKET workbench
Rapidly study, compare and integrate data
Accelerate curation of knowledge
A TIDBIT generator, and a “compiler” of KPs
GDSC
TCGA
ISB Multiomic Wellness
Heaps of biomedical data
Heterogeneous formats, ontologies, access permissions
EHR 1
Data File Q
Study XYZ
EHR 2
Data Store M
Data File Z

Overview of DOCKET (prototype)

M1.2

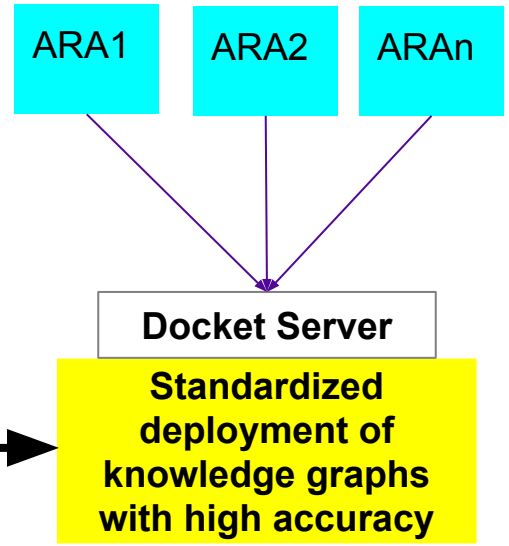
DOCKET Overview:
a system for better understanding the data

M1.3

DOCKET Compare:
compare different datasets
compare different snapshots

M1.4

DOCKET Integrate:
statistical models to rationalize the connections among different entities or features



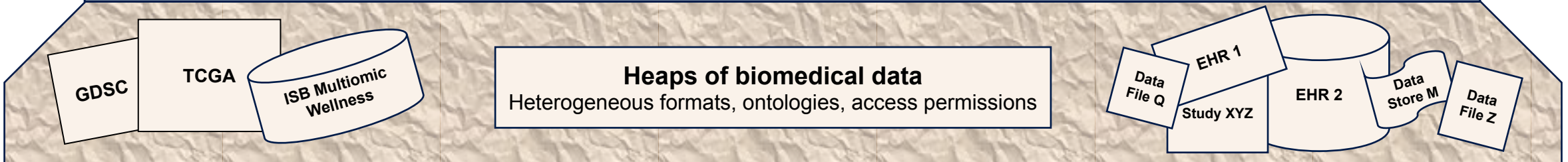
Automated analytics, facilitated curation → **Comparison among different resources** → **Integration of data from different levels or different resources**

Our solution: DOCKET workbench

Rapidly study, compare and integrate **data**

Accelerate curation of **knowledge**

A TIDBIT generator, and a “compiler” of **KPs**



Features of the DOCKET workbench

What we mean by...
DOCKET: The utilities for extracting knowledge, evaluating, visualizing, serving...
A docket: The resulting dossier, a resource readily available for ARAs

M1.2

DOCKET Overview:
a system for better understanding the data

- DOCKET Study — Automated pipeline from data to docket
- DOCKET Overview — Visualize contents of docket, interact to refine

M1.3

DOCKET Compare:
compare different datasets
compare different snapshots

- DOCKET Match — Compare dockets to find common data types
- DOCKET Compare — Compare versions of the data to assess knowledge robustness

M1.4

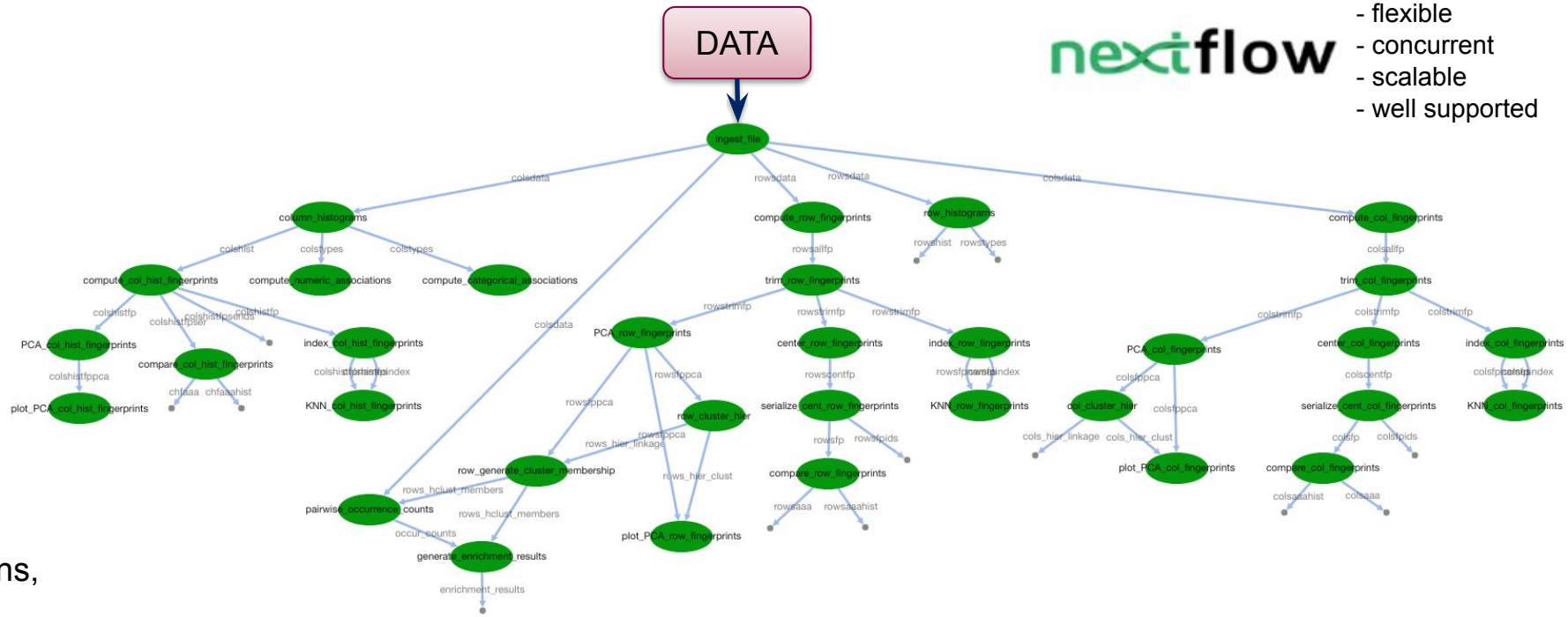
DOCKET Integrate:
statistical models to rationalize the connections among different entities or features

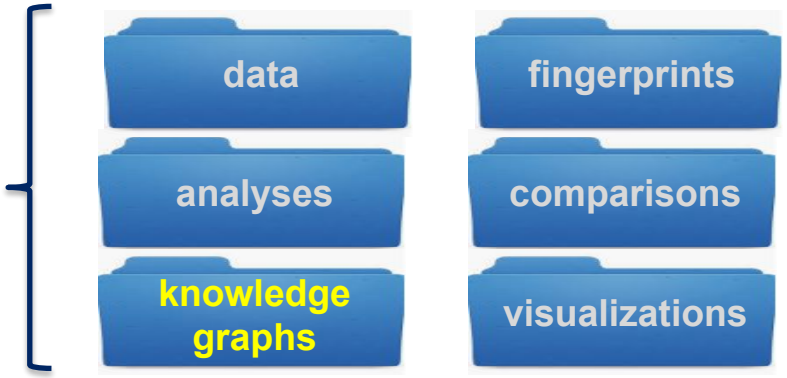
- DOCKET Merge — Merge compatible tables
- DOCKET Integrate — Statistical modeling to derive knowledge

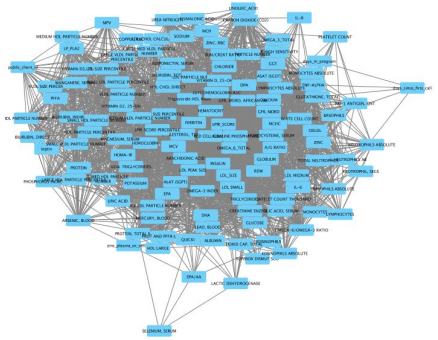
DOCKET Study

An automated, concurrent analysis pipeline

- Ingest data
- Statistics, distributions
- Dimensionality reduction (fingerprinting, PCA) – both row-wise and column-wise
- Hierarchical clustering
- Enrichment analysis
- PCA visualizations
- Data indexing
- All-against-all comparisons
- k-nearest neighbors
- Pairwise column associations, numerical and categorical (bivariate statistics)



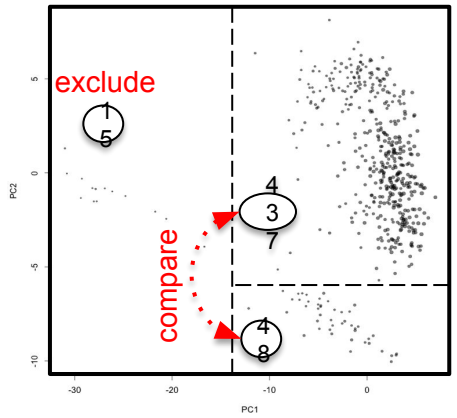




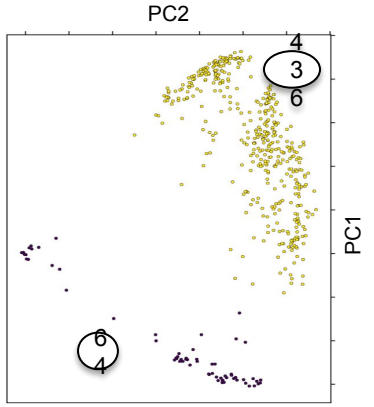
DOCKET Study

Automated analytics, curation support

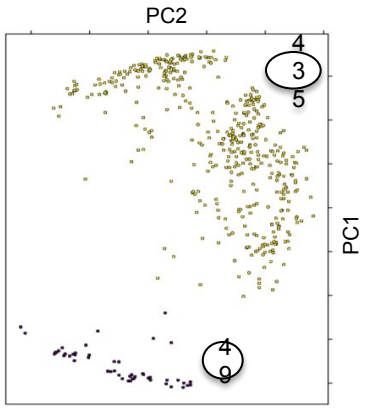
St. Jude Life cohort N=500



automation



curation, refinement



Data types: demographics, body metrics, genetic tests, diagnoses (chronic), secondary primary cancers, chemotherapy + dosage, radiation + dosage

5.0x Adrenal insufficiency
4.9x Cerebellar dysfunction
4.6x Growth hormone defficiency
3.0x Vascular disease
2.6x Obstructive sleep apnea
2.5x Hypothyroidism
2.5x Cerebrovascular accident
1.8x Radiation to thyroid
1.7x GERD

Term	p-value
Hypothyroidism	1.7E-08
Adrenal insufficiency	9.9E-08
Childhood GHD	4.3E-06
Adult GHD	1.6E-05
Hearing loss	1.7E-05
Cerebellar dysfunction	3.7E-05

Enrichment analysis, unclear statistics

Knowledge on pediatric oncology multi-morbidities: association between disease entities (biolink)

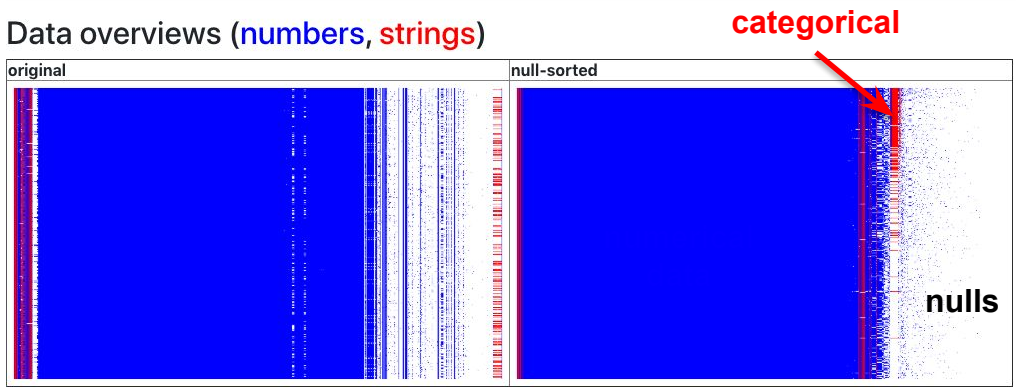
DOCKET Overview

Interactive report □ discovery and curation

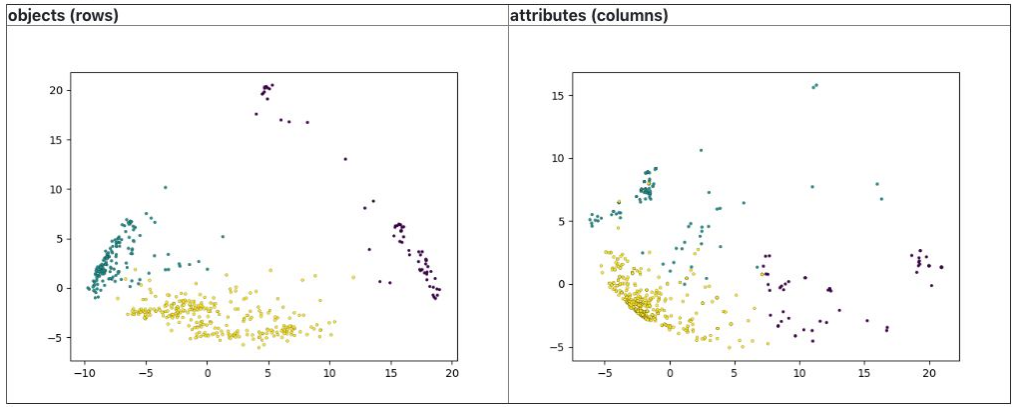
DOCKET Overview for SJL500 nonLabVars

Docket path: demo_dockets/SJL500_nlv-200
Table size: 500 rows x 838 columns (41 of which are empty, 765 are numerical)

Data overviews (numbers, strings)



PCA (PC1 vs. PC2)



Numerical associations

Show 10 entries — search data

variableA	variableB	N	rho	pval
chemodose.vinca_dose_prim	chemodose.vinca_dose_10	347	1.000	0.00e+00
chemodose.vinca_dose_prim	chemodose.vinca_dose_any	346	1.000	0.00e+00
chemodose.vincrist_dose_5	chemodose.vcr_dose_5	331	1.000	0.00e+00
chemodose.mtx_iv_dose_10	chemodose.mtx_iv_dose_prim	91	1.000	0.00e+00
chemodose.mtx_iv_dose_10	chemodose.mtx_iv_dose_any	91	1.000	0.00e+00
chemodose.mtx_dose_10	chemodose.mtx_dose_any	156	1.000	0.00e+00
chemodose.cortico_dose_10	chemodose.cortico_dose_any	15	1.000	0.00e+00
chemodose.cortico_dose_10	chemodose.cortico_dose_prim	15	1.000	0.00e+00
chemodose.cortico_dose_10	chemodose.cortico_dose_5	15	1.000	0.00e+00
chemodose.mercapto_po_dose_5	chemodose.mercapto_po_dose_10	9	1.000	0.00e+00

Showing 1 to 10 of 528 entries — Previous 1 2 3 4 5 … 53 Next

Categorical associations

Show 10 entries — search data

variableA	variableB	N	Theil_B_A	Theil_A_B
chemodose.mercapto_po_dose_10	chemodose.mercapto_po_dose_5	409	1.000	1.000
chemodose.mercapto_po_dose_10	chemodose.mercapto_po_dose_any	409	1.000	1.000
chemodose.mercapto_po_dose_10	chemodose.mercapto_po_dose_prim	409	1.000	1.000
chemodose.mercapto_po_dose_10	chronic.Cardiovascular_Thrombus	10	1.000	1.000
chemodose.mercapto_po_dose_5	chemodose.mercapto_po_dose_any	409	1.000	1.000
chemodose.mercapto_po_dose_5	chemodose.mercapto_po_dose_prim	409	1.000	1.000
chemodose.mercapto_po_dose_5	chronic.Cardiovascular_Thrombus	10	1.000	1.000
chemodose.mercapto_po_dose_any	chemodose.mercapto_po_dose_prim	409	1.000	1.000
chemodose.mercapto_po_dose_any	chronic.Cardiovascular_Thrombus	10	1.000	1.000
chemodose.mercapto_po_dose_prim	chronic.Cardiovascular_Thrombus	10	1.000	1.000

Showing 1 to 10 of 1,104 entries — Previous 1 2 3 4 5 … 111 Next

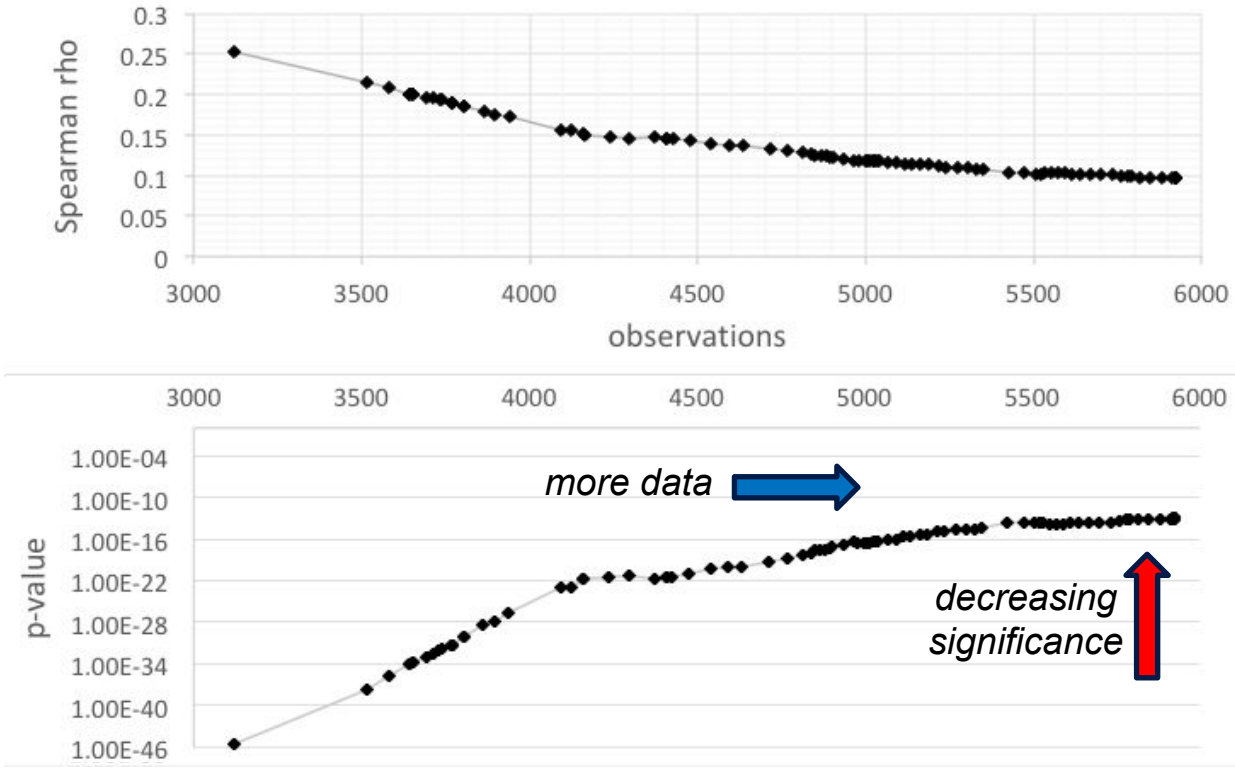
DOCKET Compare

Derive *knowledge robustness* statistics

ISB wellness data set, September 2017 correlations

- Glycohemoglobin A1C (LOINC 4548-4) and neutrophils (LOINC 26507-4): N=4970, rho=0.091, p-value=1.60E-10

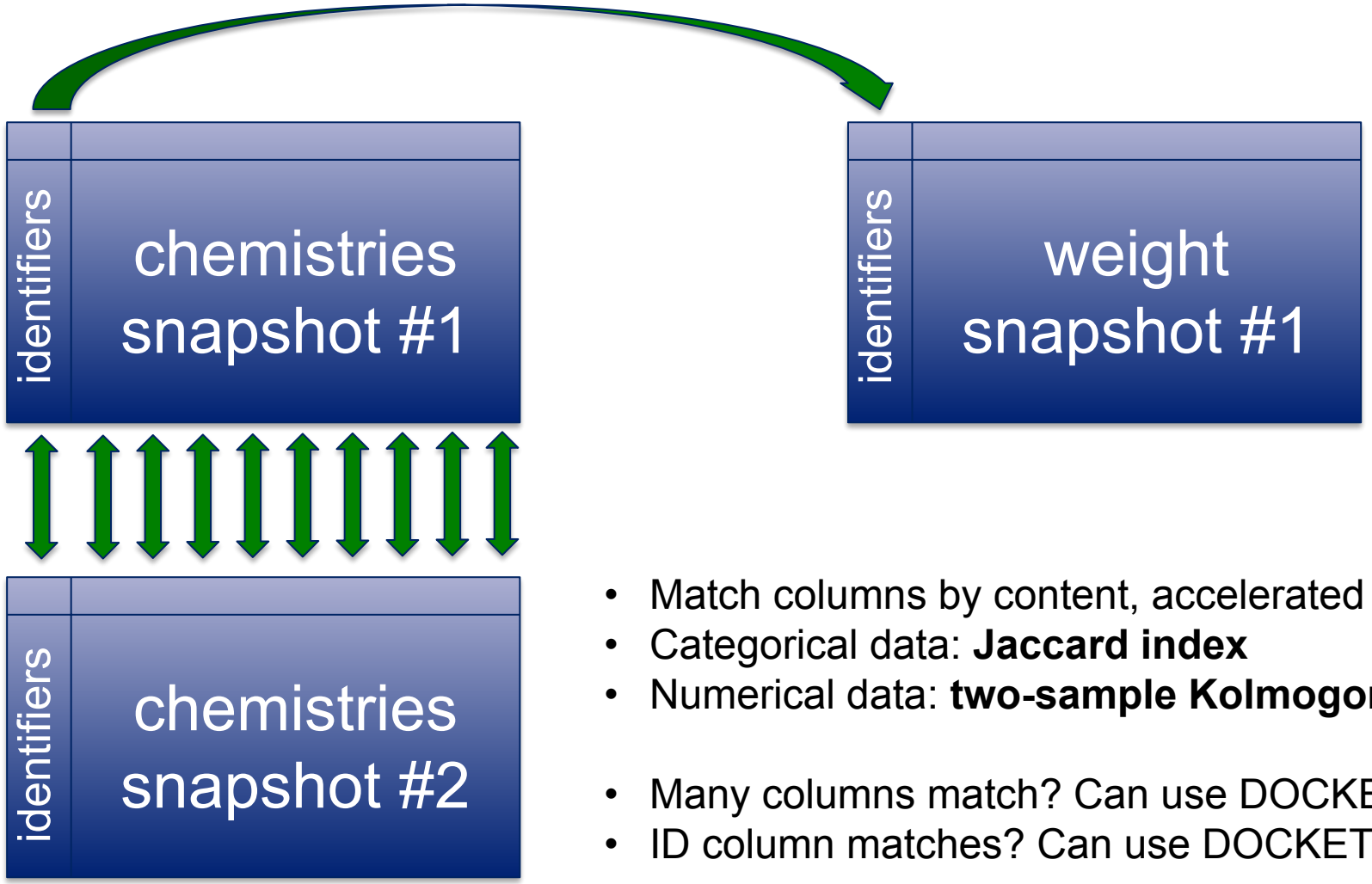
- Glycohemoglobin A1C (LOINC 4548-4) and basophils (LOINC 706-2): N=3121, rho=0.252, p-value=2.48E-46



Confidence metric of knowledge robustness
- support for expert curation and validation
- metric of quality and validity of results returned to ARAs

DOCKET Match

Compare dockets to find what they share in common

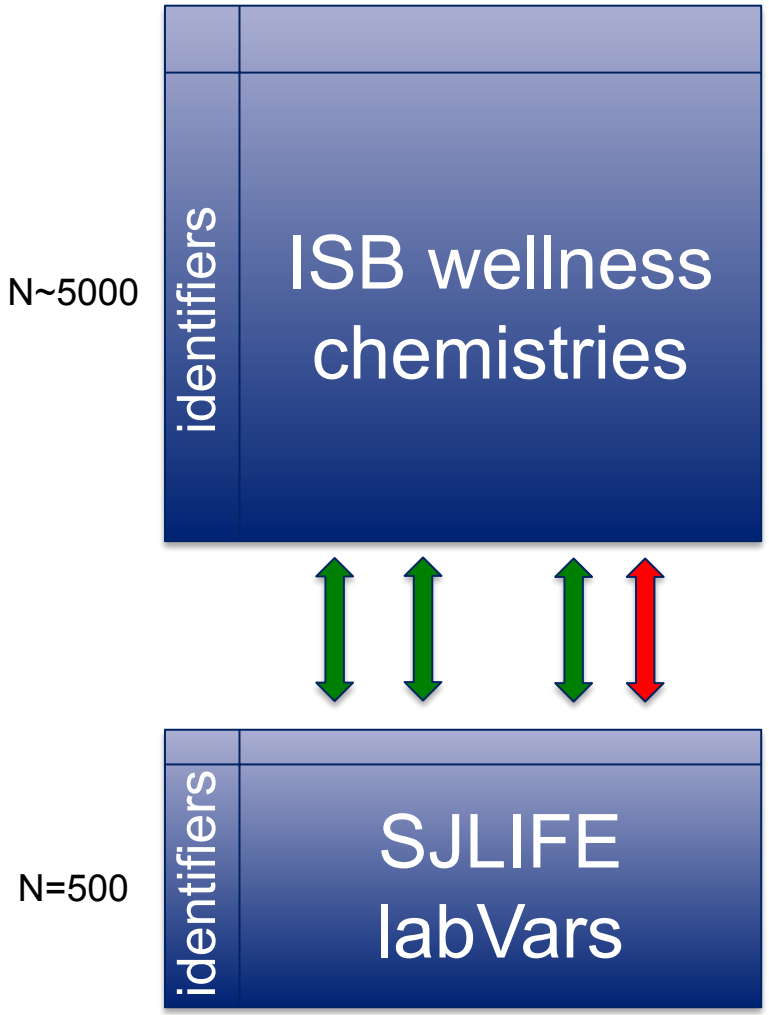


- Match columns by content, accelerated via fingerprints
- Categorical data: **Jaccard index**
- Numerical data: **two-sample Kolmogorov-Smirnov test**

- Many columns match? Can use DOCKET Compare
- ID column matches? Can use DOCKET Merge, Integrate

DOCKET Match

Can also match content from different data sets



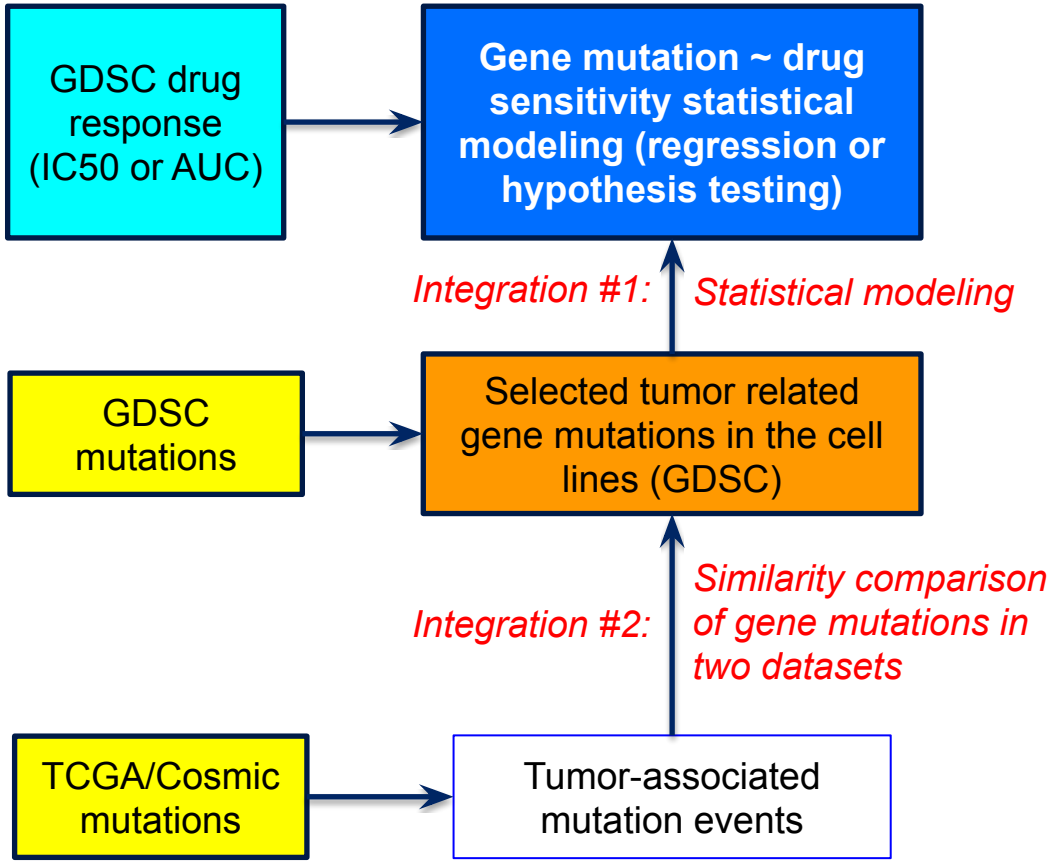
Wellness “chemistries” vs. SJLIFE “labVars”:

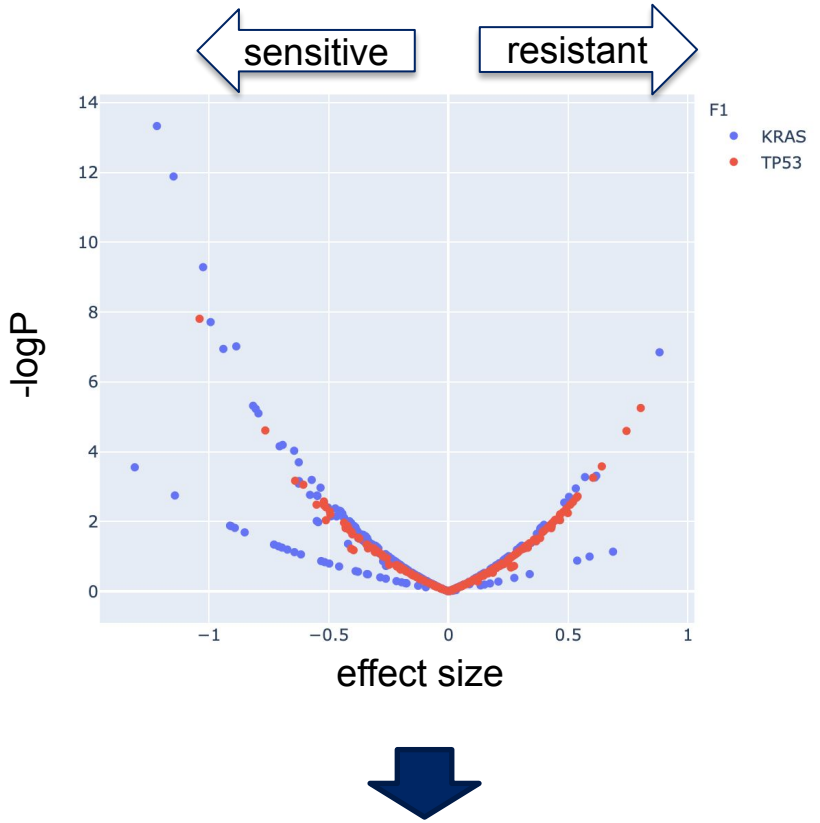
- OK “MCH” (785-6) □□ “labs_numeric.MCH”
- OK “HEMOGLOBIN” (718-7) □□ “labs_numeric.Hgb”
- info gain “RED CELL COUNT” (789-8) □□ “labs_numeric.RBC_mm3”
- QC ? “HOMA-IR” □□ “labs_numeric. Thyroid_Stimulating_Hormone_Analyzer”

Assistance to expert curation!

DOCKET Integrate

Statistical modeling (drug response vs. mutations, from multiple resources)



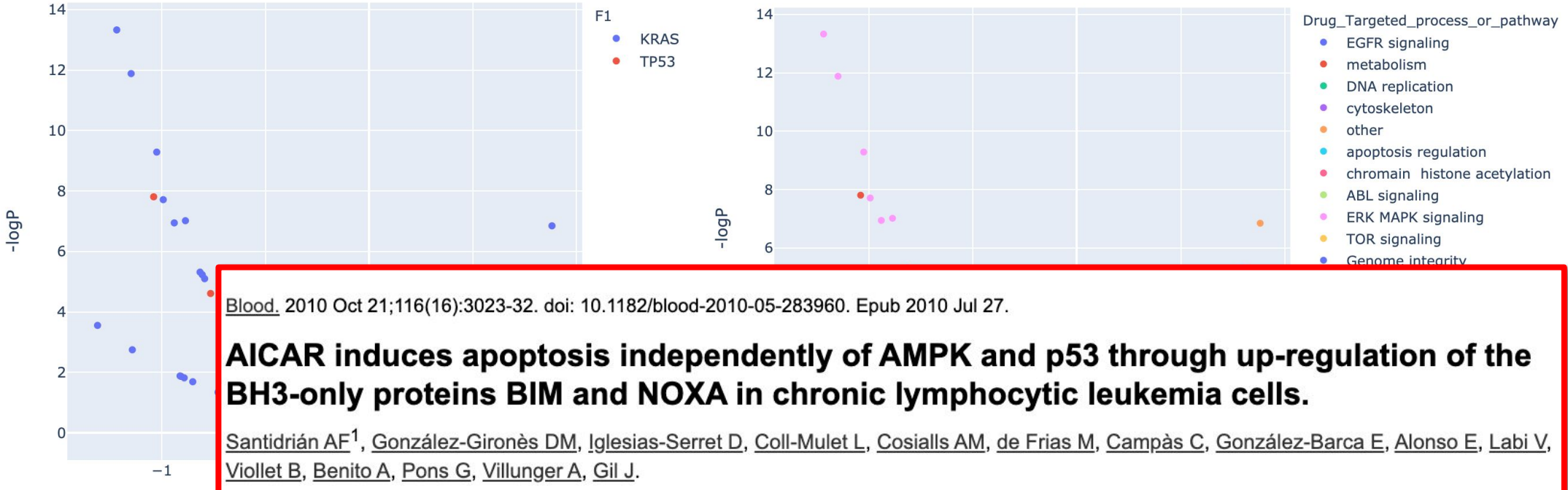


KP: Graphs and tables for users and ARAs

Biolink knowledge graph:
Disease --- Gene --- Drug (association: drug sensitivity)

DOCKET Integrate

Annotate knowledge graphs with external resources



Blood. 2010 Oct 21;116(16):3023-32. doi: 10.1182/blood-2010-05-283960. Epub 2010 Jul 27.

AICAR induces apoptosis independently of AMPK and p53 through up-regulation of the BH3-only proteins BIM and NOXA in chronic lymphocytic leukemia cells.

Santidrián AF[1], González-Gironès DM, Iglesias-Serret D, Coll-Mulet L, Cosialls AM, de Frias M, Campàs C, González-Barca E, Alonso E, Labi V, Viollet B, Benito A, Pons G, Villunger A, Gil J.

Author information

F1	Drug_Name	SE	FDR	Drug_Putative_Target	Drug_Targeted_process_or_pathway
KRAS	RDEA119 (rescreen)	-1.217725	0.000795	MAP2K1 (MEK1), MAP2K2 (MEK2)	ERK MAPK signaling
KRAS	AZD6244	-1.148194	0.001684	MAP2K1 (MEK1), MAP2K2 (MEK2)	ERK MAPK signaling
KRAS	Trametinib	-1.024762	0.015079	MAP2K1 (MEK1), MAP2K2 (MEK2)	ERK MAPK signaling
KRAS	PD-0325901	-0.993351	0.043625	MAP2K1 (MEK1), MAP2K2 (MEK2)	ERK MAPK signaling
TP53	AICAR	-1.039567	0.043625	AAPK1 (AMPK) agonist	metabolism

literature validation

Enrichment of drug sensitivity or resistance:

Sensitivity pairs:

	Gene	mol_action	p
0	KRAS	ERK MAPK signaling	0.000150
1	TP53	metabolism	0.047059

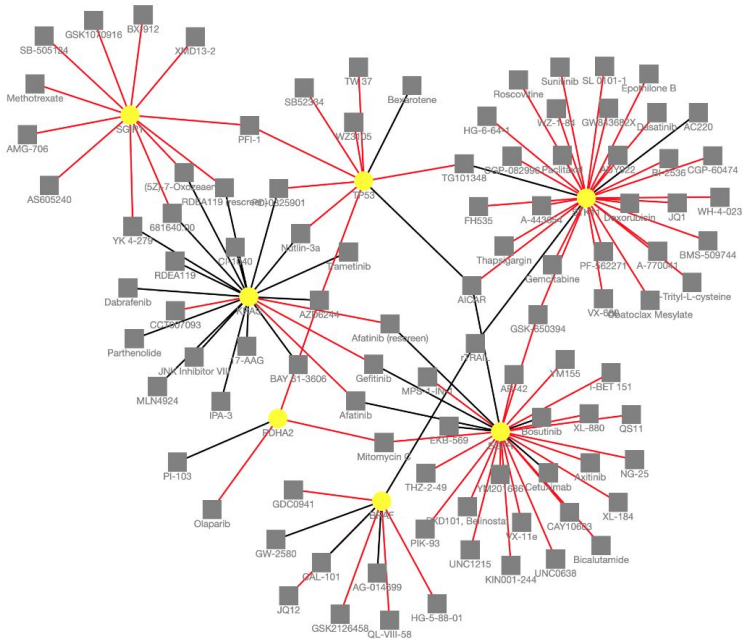
Resistance pairs:

	Gene	mol_action	p
0	KRAS	EGFR signaling	0.003023

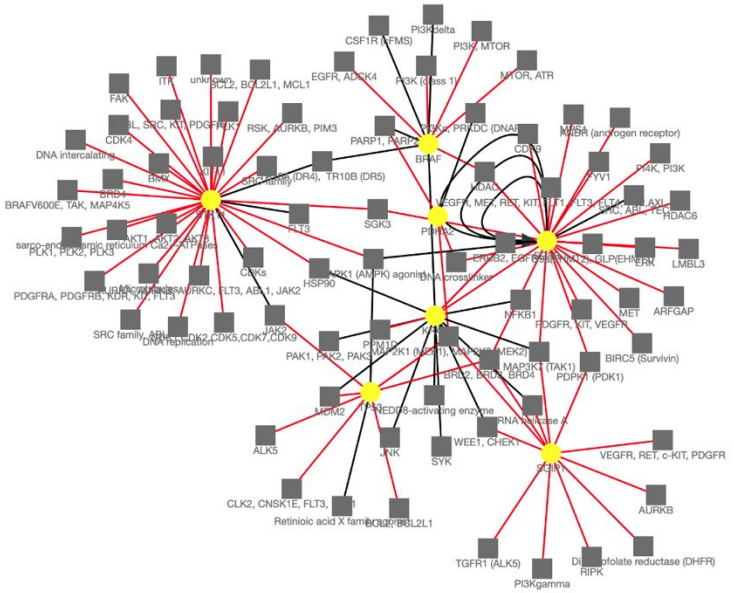
DOCKET Integrate

Visualization for knowledge graphs

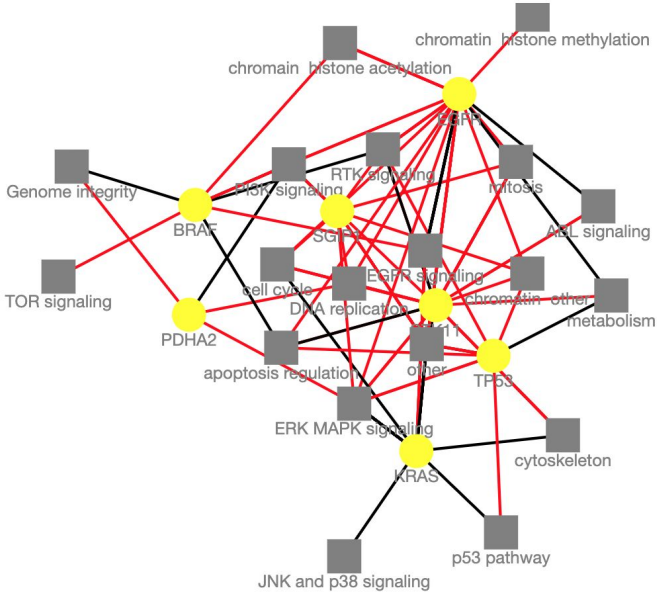
Association between drug and gene mutations



Association between Drug targets and gene mutations



Association between Targeted_process_or_pathway and gene mutations



Evidence level: GDSC data, LUAD, sample size: 64, P < 0.05, t-test

Provide visualizations of graphs

Milestones

M1.1 Ongoing support of Big GIM

- Migration to new Google Cloud Platform, connected with Translator users
- Deployed on NCATS server, BDD tests 91% code overage

M1.2 Prototype DOCKET Overview

- Implemented **Study** (nextflow pipeline), **Overview** (interactive reports)
- Automatically studied wellness and cancer multiomics data
- Yielded visualizations and knowledge graphs

M1.3 Prototype DOCKET Compare

- Implemented **Match** (data types) and **Compare** (versions)
- Demonstrated ability to quantify knowledge stability

M1.4 Prototype DOCKET Integrate

- Implemented **Merge** (join tables with human expert *validation*)
- Implemented **Integrate** (statistical analysis, Jupyter notebooks)
- Demonstrated integrative analysis of cancer data

M1.5 Gather input on DOCKET features

- Designed survey, followed up with respondents
- CI, [C]D: Containerized DOCKET, established BDD test framework

M1.6 Collaborate on Reasoner Standard API

Deliverables

DOCKET code, with Docker container

github.com/PriceLab/DOCKET

Instructions for running prototypes

drive.google.com/drive/u/1/folders/19CT2bu1kzVnXgORhgIQijJd7x8O8Ez6D

Translator Drive > Knowledge Providers

Try-it-yourself document

Sample data

Big GIM: **G**ene **I**nteraction **M**iner, and

Big CLAM: **C**ell **L**ine **A**ssociation **M**iner:

biggim.ncats.io/api

github.com/PriceLab/translator-bigquery-api

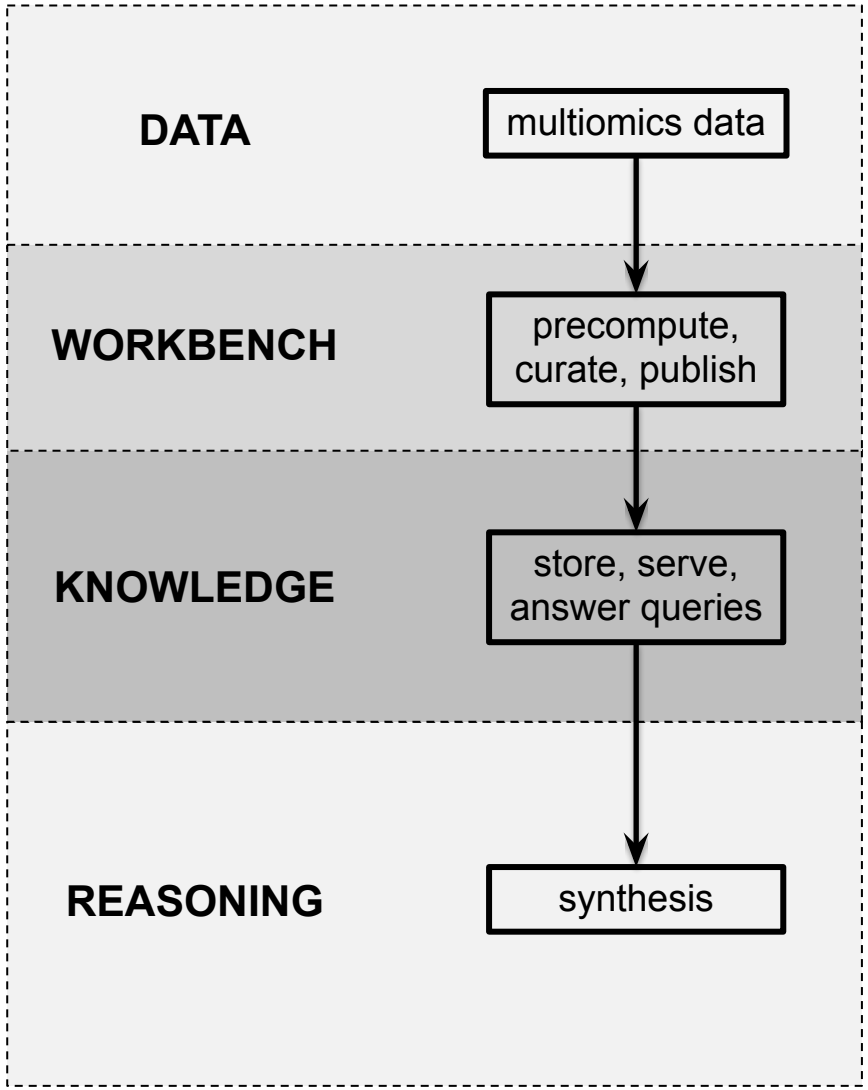
DEMO TIME

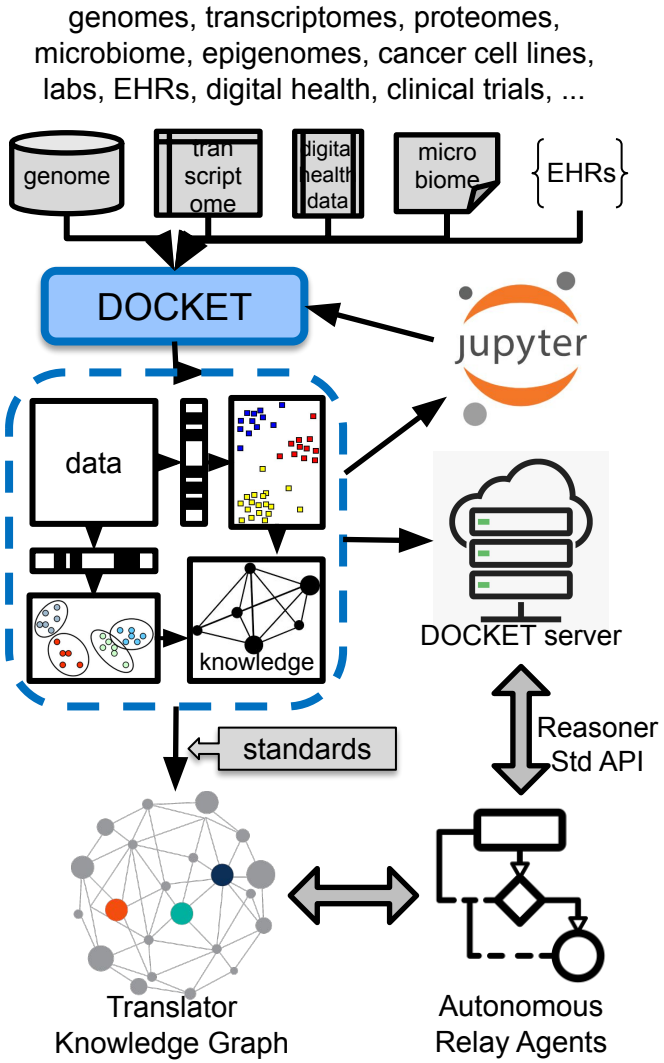
EXTRA SLIDES

Survey Highlights

High priority scenarios and features

Scenarios	
Allowing subject matter experts to curate and validate knowledge without needing to run code.	100%
Coordinating parallel workflow across groups of people curating and validating knowledge	83%
Features	
Support for rapid exploration, analysis, visualization and curation of raw data from heterogeneous locations, storage formats and versions.	100%
Automatically generated descriptive statistics and visualization to support understanding of data structure, content, missing data, and outliers.	100%
Automatically detected suggestions for observations in different data sources that are likely to be referring to the same concept.	100%
Automatically generated machine learning analyses, such as dimensionality reduction, clustering and enrichment.	100%
Automatically detected suggestions for mapping to Biolink entities and associations.	83%
A user interface for rapid human validation and curation of which correlations should be added as associations into your knowledge graph.	83%
Ability to incorporate existing provenance data from one or more data sets into your KP.	83%
Automatic generation of provenance information when you merge data from different sets, transform data, map to biolink entities and associations.	83%
Ability to ingest your graph and metadata and automatically deploy a knowledge graph.	83%


DATA
multiomics data
WORKBENCH
precompute, curate, publish
KNOWLEDGE
store, serve, answer queries
REASONING
synthesis


genomes, transcriptomes, proteomes, microbiome, epigenomes, cancer cell lines, labs, EHRs, digital health, clinical trials, ...
genome
tran script ome
digita health data
micro biome
EHRs
DOCKET
jupyter
data
knowledge
DOCKET server
standards
Reasoner Std API
Translator Knowledge Graph
Autonomous Relay Agents

DOCKET features and information flow