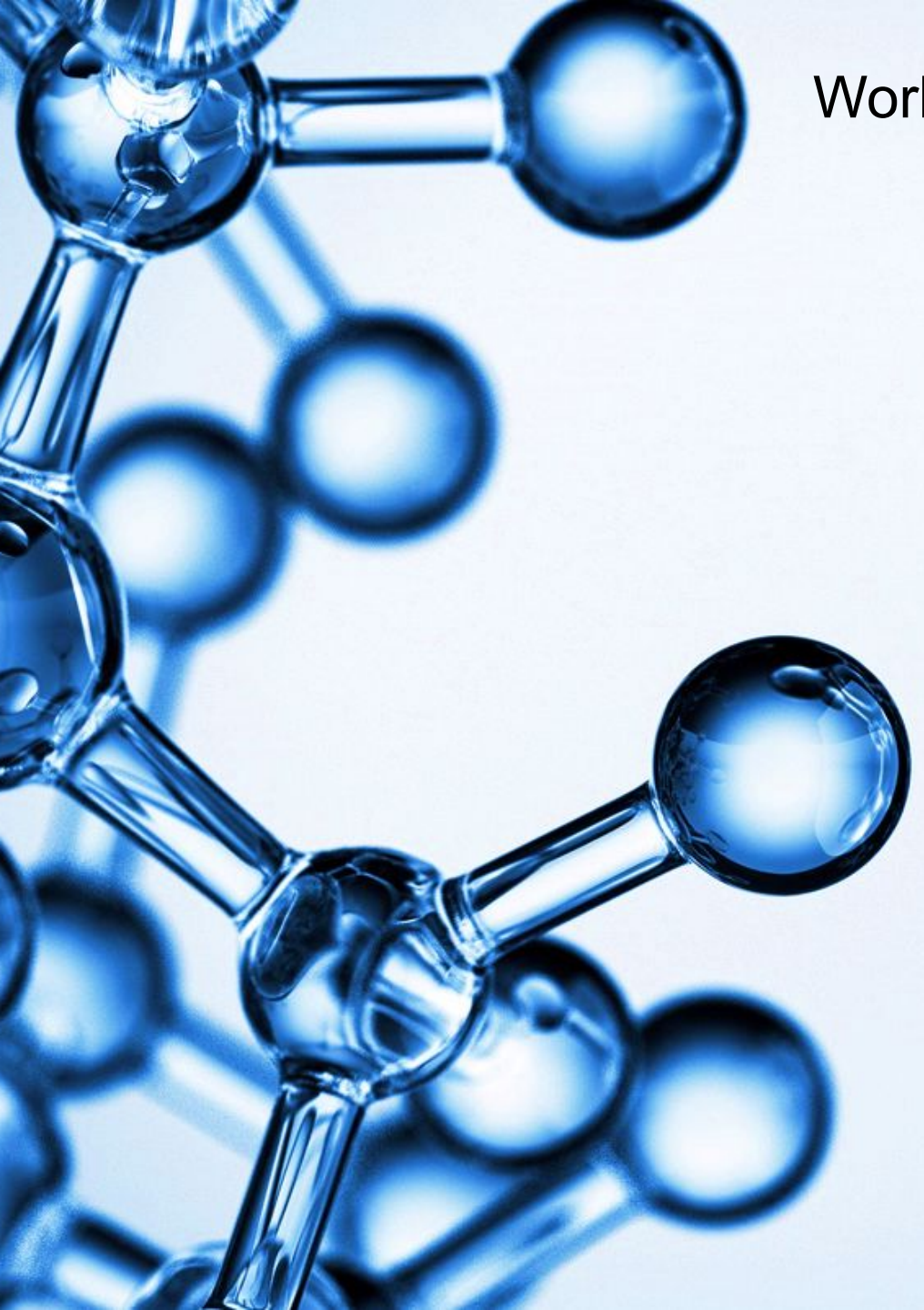


Workshop on Synergizing Biomedical Ontologies 2021

CDD Annotator and Perspectives from the Data FAIRy Initiative

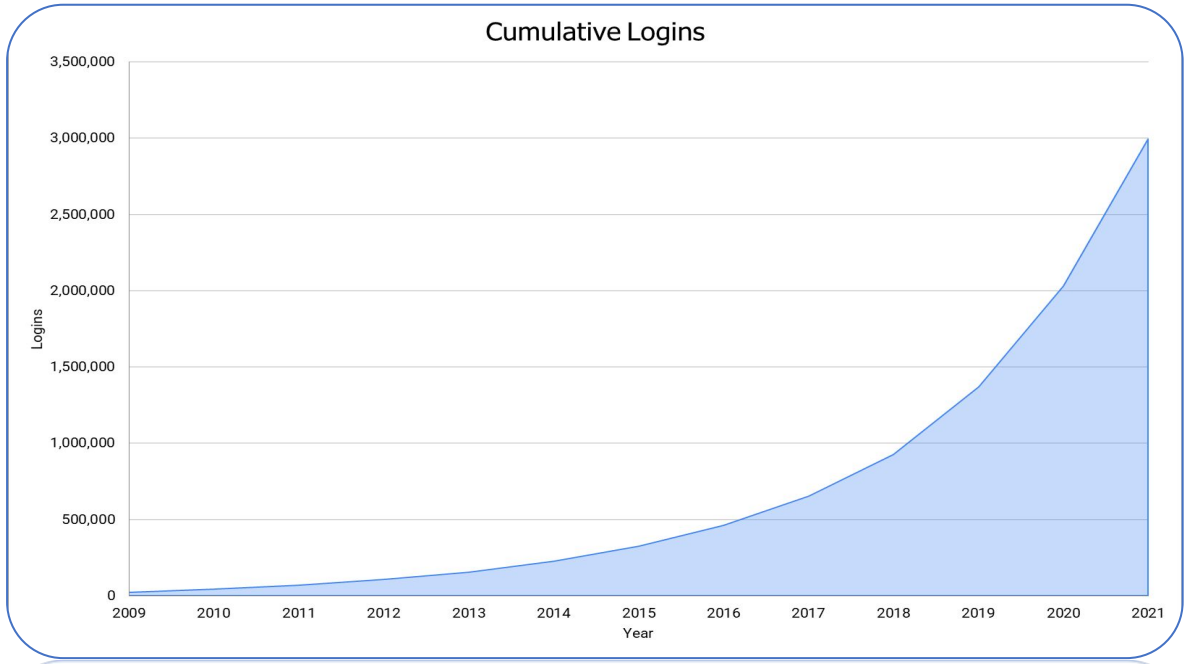
Samantha Jeschonek, Ph.D.

Collaborative Drug Discovery

15 July 2021

Collaborative Drug Discovery

Started in San Francisco in 2004, CDD provides cloud-based data management and discovery platforms to the scientific community worldwide. Our software engineers and application scientists have a rich variety of backgrounds and expertise in many fields, which they leverage to support your modern drug research informatics needs.



COLLABORATIVE DRUG DISCOVERY

CDD's platforms for (meta)data management

CDD VAULT®
Complexity Simplified™

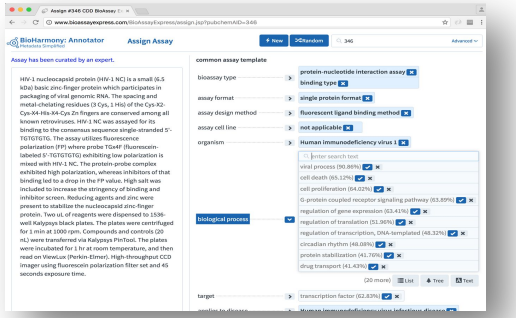
The industry's most trusted cloud-based drug discovery data management platform

Manage

BioHarmony
DATA SIMPLIFIED

Your centralized source for semantic drug data

Discover

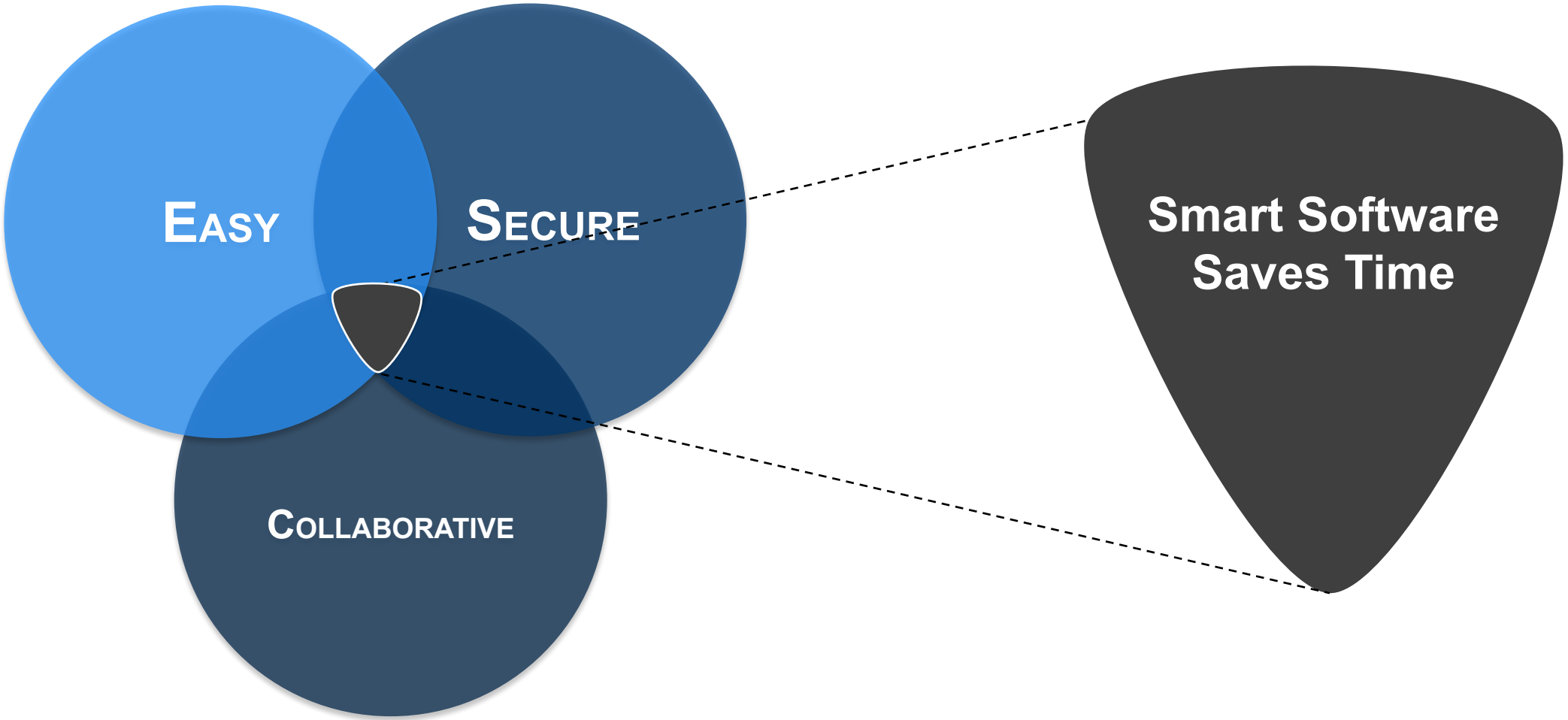


Annotator
METADATA SIMPLIFIED

Automatically converts your assay metadata into semantic content

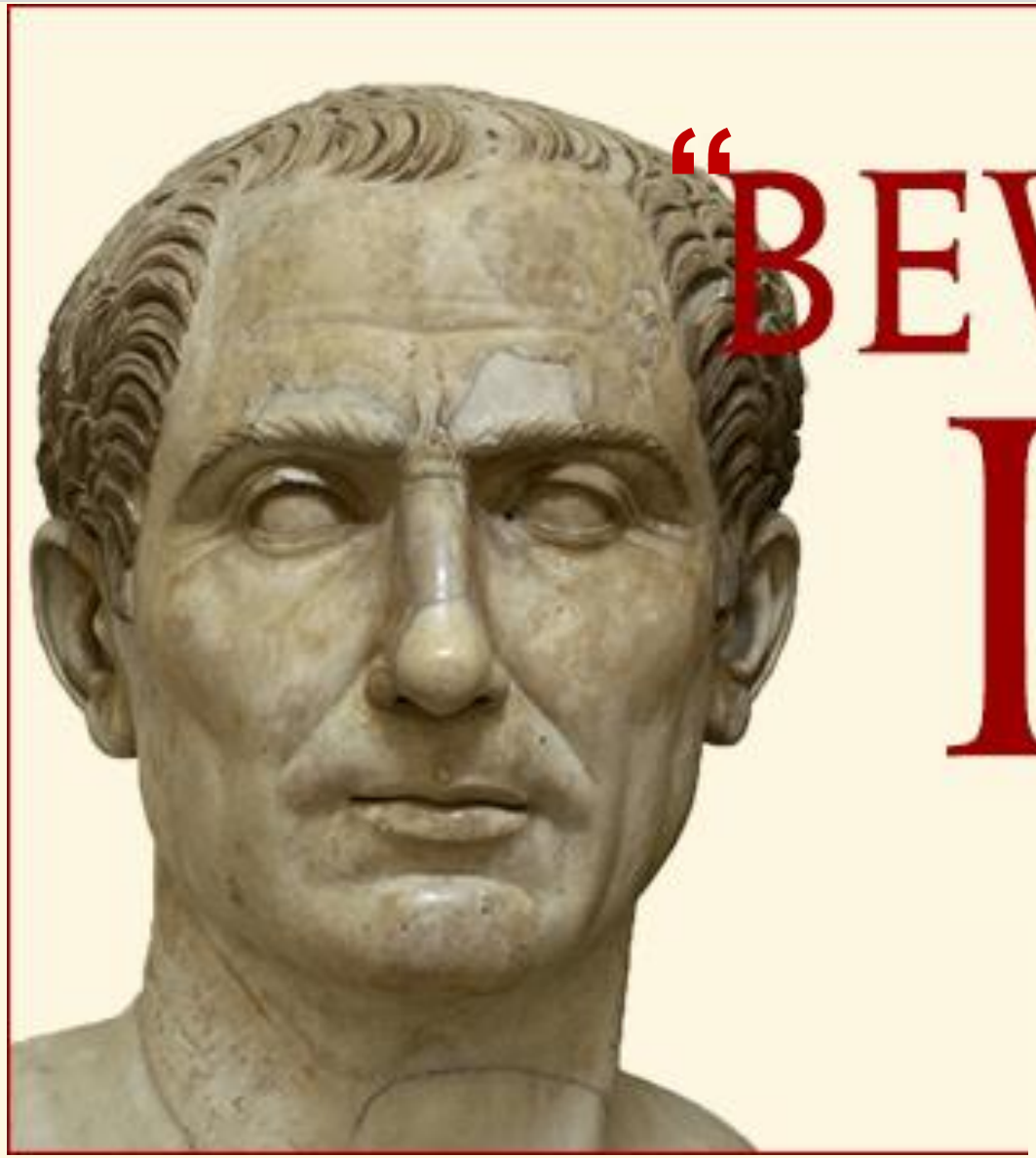
Annotate

CDD's principles to save you *TIME*!



CDD COLLABORATIVE DRUG DISCOVERY

Are you capturing enough data today to answer *tomorrow's* questions?

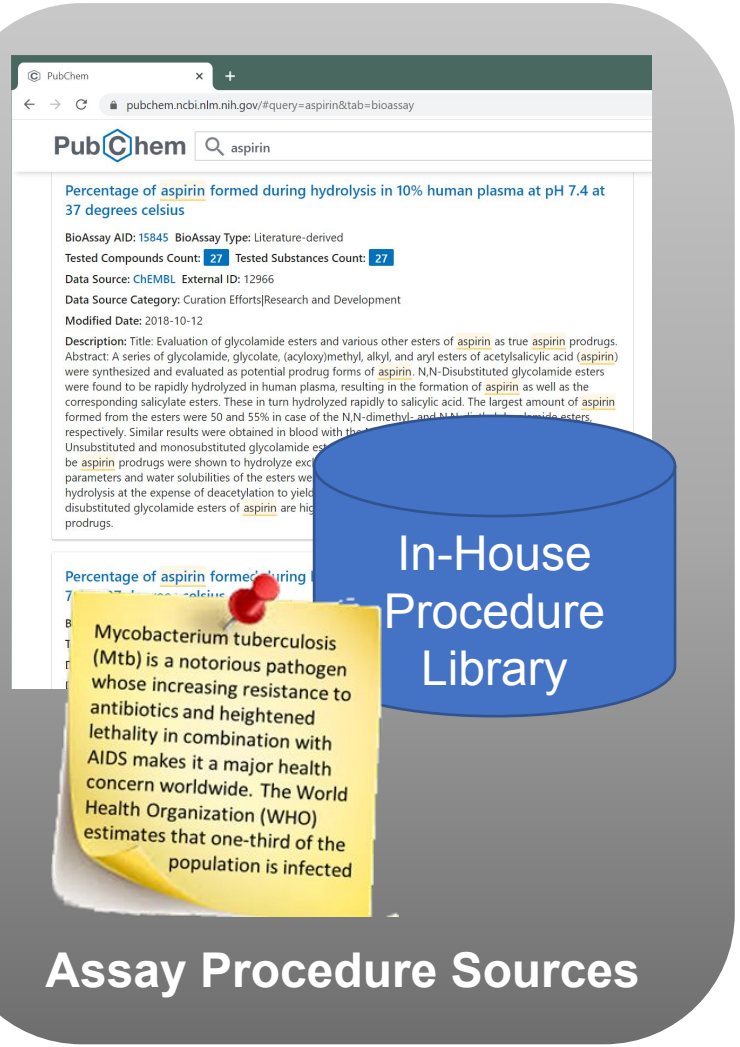


"BEWARE THE IDES…

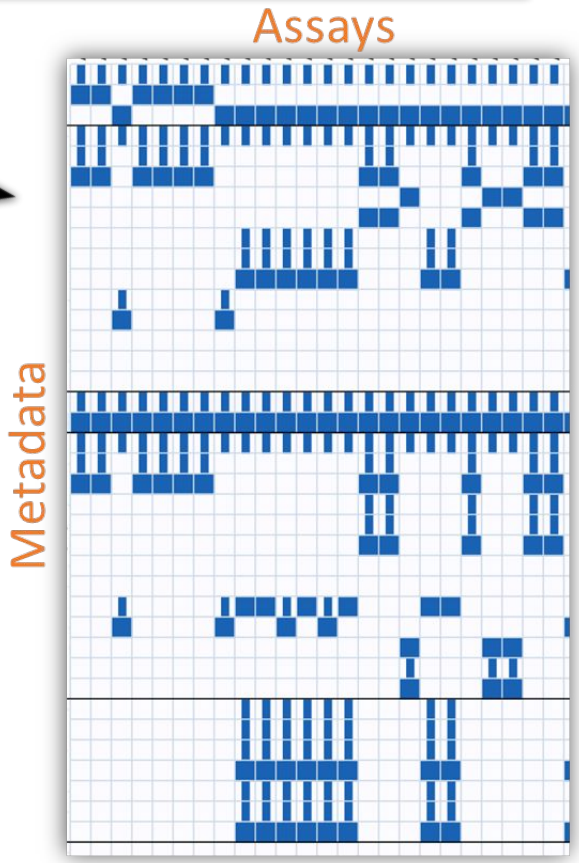
…cribed my assays poorly"

And now I don't have the data I need,
The time I once had and
The answers I wanted

Annotator (previously *BioAssay Express*) enables true assay informatics



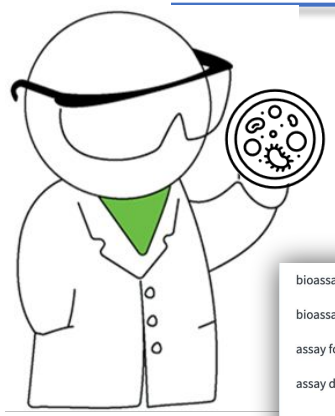
NLP & ML & Human Guidance



Assay Informatics

COLLABORATIVE DRUG DISCOVERY

Annotator also enables new assay registration



bioassay type	functional
bioassay	beta lactamase reporter gene assay
assay format	cell based format
assay design method	reporter gene method beta lactamase induction
assay supporting method	stable transfection
assay cell line	CHO cell
organism	Homo sapiens
biological process	serotonin receptor signaling pathway Gi/o-coupled serotonin receptor activity
target	5-hydroxytryptamine receptor 1a
applies to disease	cognitive disorder mental depression
assay mode of action	activation agonism
result	EC50
screening campaign stage	confirmatory assay
assay footprint	1536 well plate

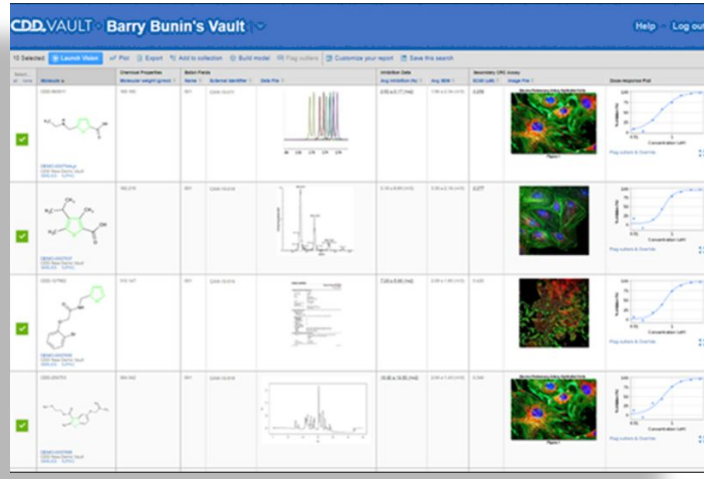
Assay Registration

Assay Annotations

Auto-Generate Text Description

This is *a confirmatory assay* to identify potential treatments for **human immunodeficiency virus infectious disease**, by investigating the biological process of **viral RNA genome packaging**, ...ally targeting **see Gene** ...**mmunodeficiency virus** ... **type/protein-nucleotide** ...**y** in *a single protein* ...*uorescent ligand binding* ...line *not applicable* was ...ucted in *1536 well plates*

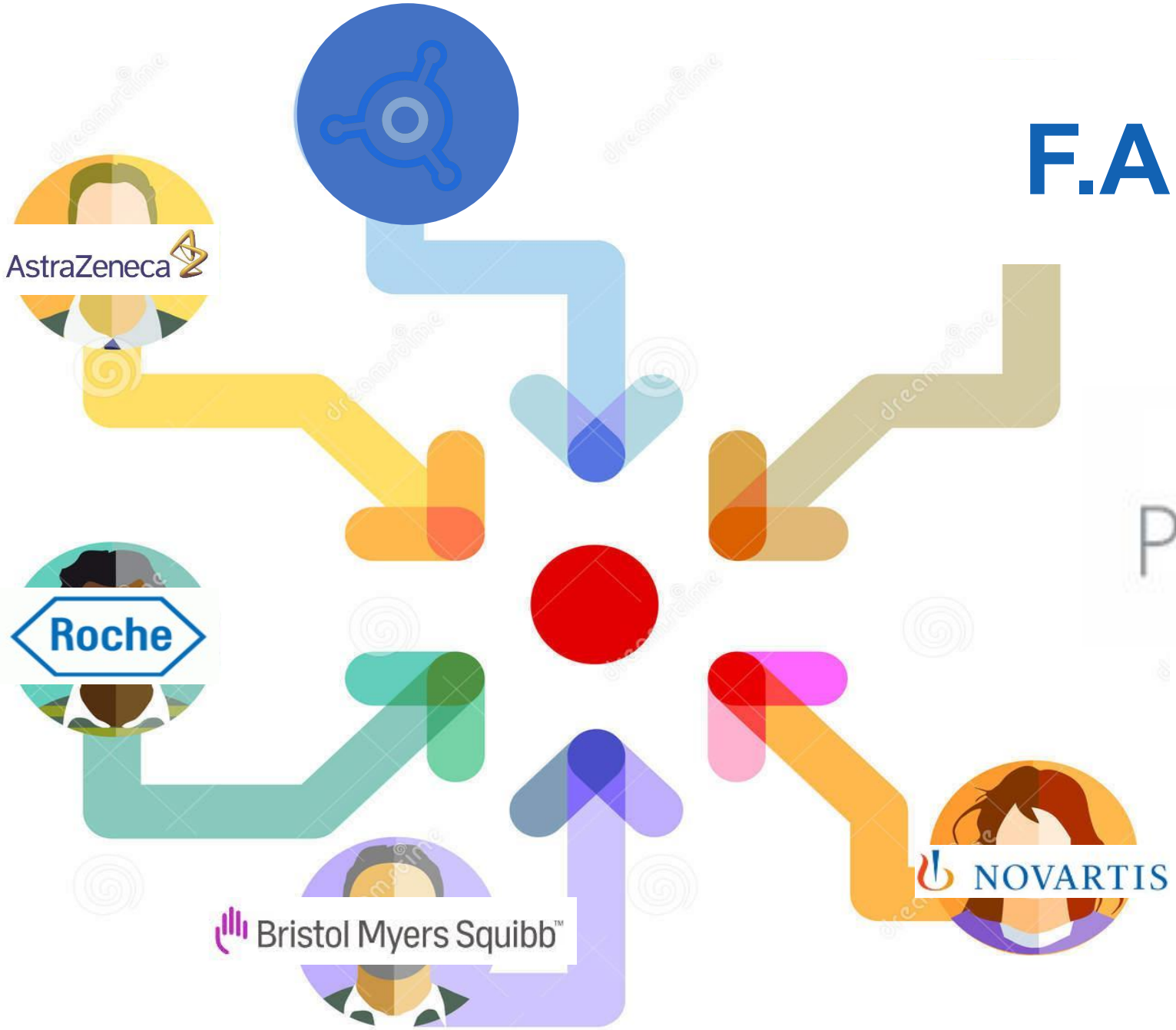
Human & Machine-Readable Content



Project Data Analysis

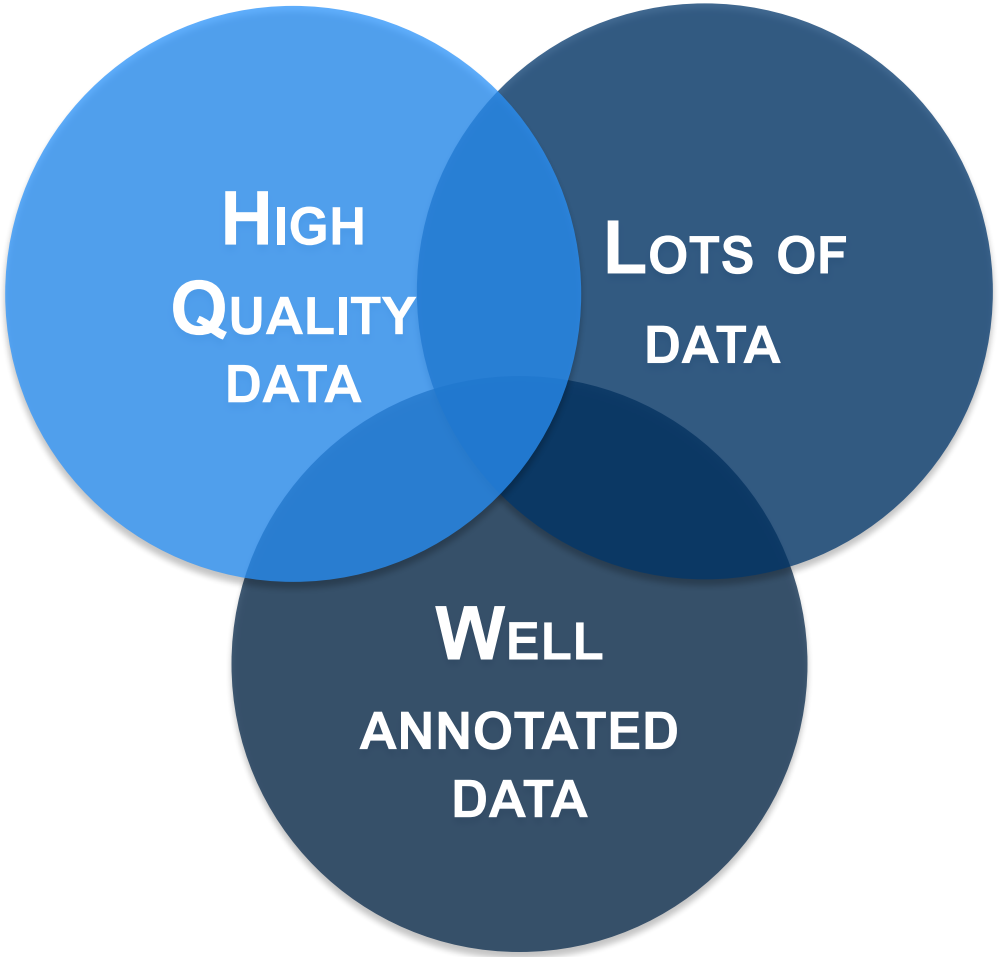
CDD COLLABORATIVE DRUG DISCOVERY

F.A.I.R. Mindsets



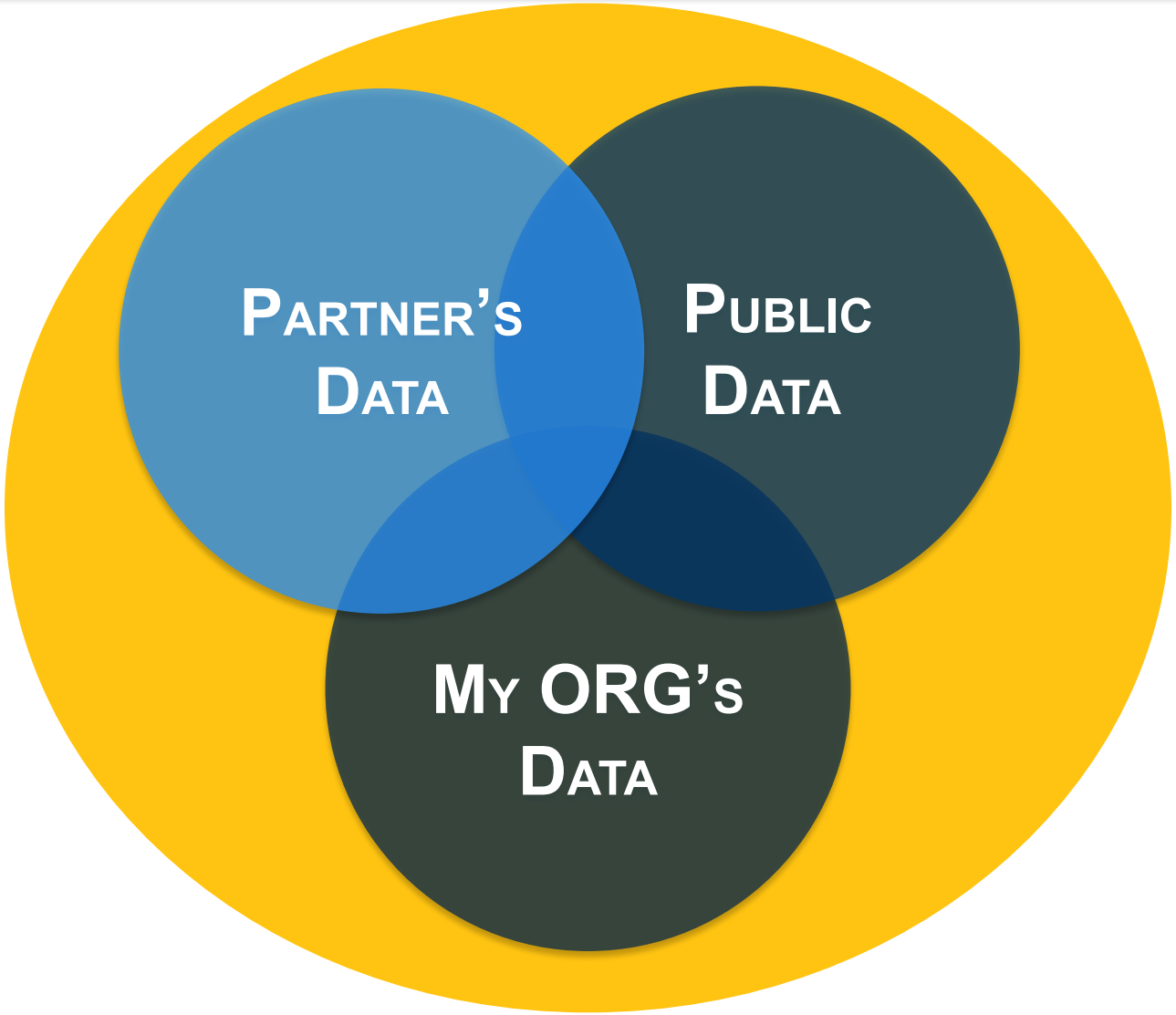
Pistoia Alliance

Community Needs



COLLABORATIVE DRUG DISCOVERY

Siloed Data is Not Helpful



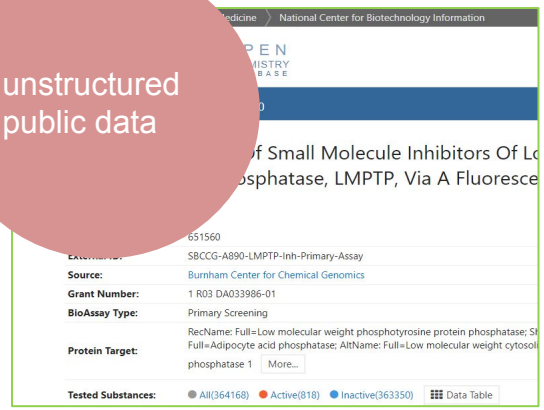
CDD COLLABORATIVE DRUG DISCOVERY

The proposed DataFAIRy operational model (2018)

A vision by Isabella Feierberg…

Cost-shared annotation of public domain bioassay descriptions with high quality, using an agreed data model, making data FAIR

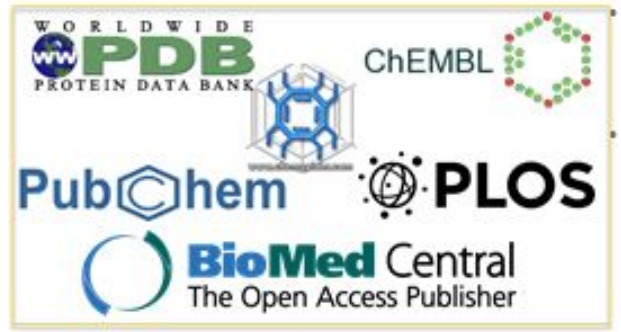
unstructured public data

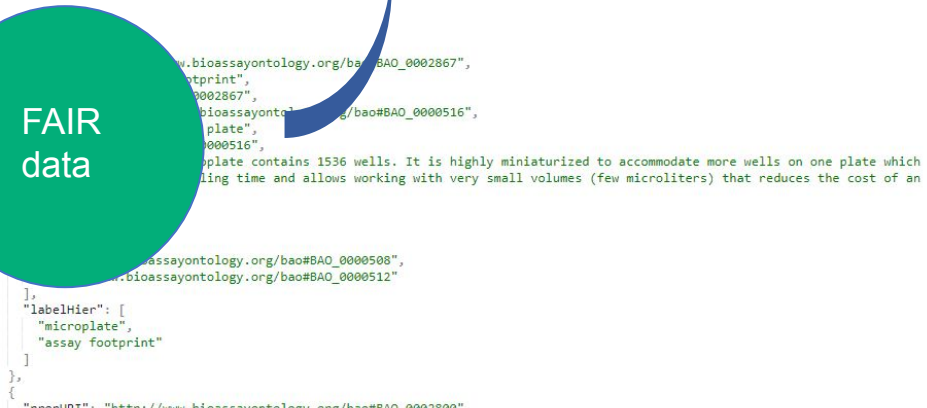


DataFAIRy Partners

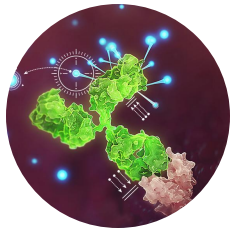
Curation and QC by independent domain experts

FAIR data



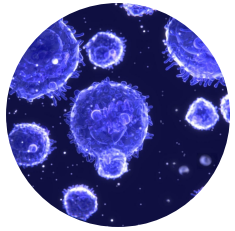


Project team



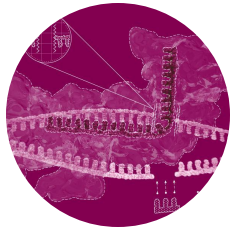
Novartis

Anosha Siripala
Gabriel Backiananthan



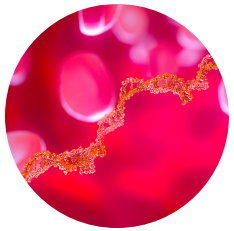
BMS

Dana Vanderwall*



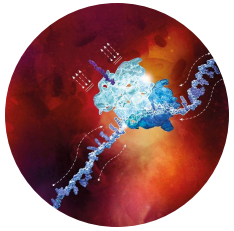
Roche

Rama Balakrishnan
Martin Romacker



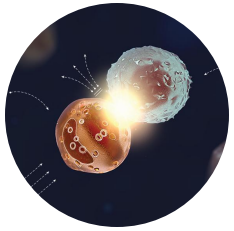
AstraZeneca

Tim Ikeda
Isabella Feierberg*



Collaborative Drug Discovery

Samantha Jeschonek
Jason Harris
Whitney Smith



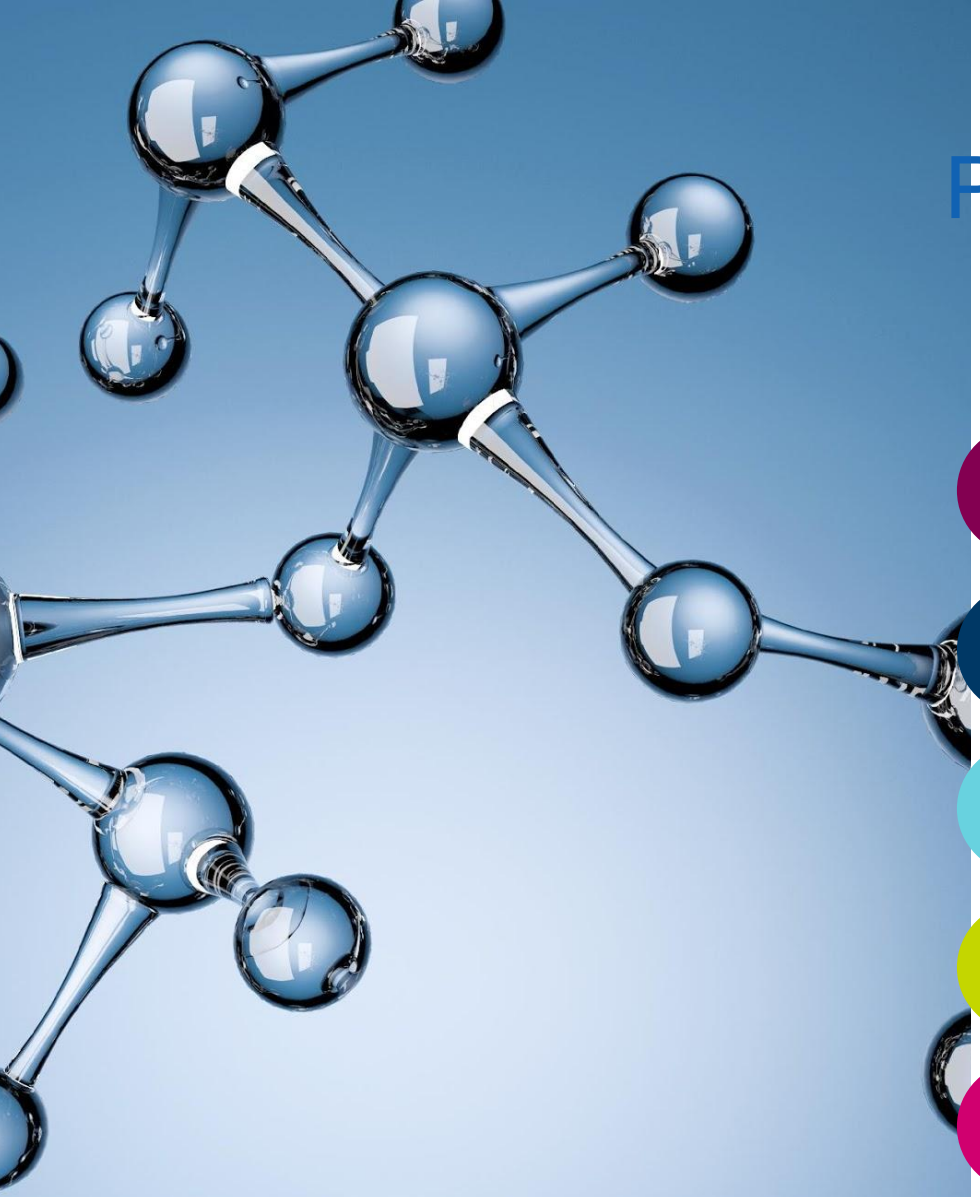
Pistoia Alliance

Vladimir Makarov
Thomas Liener

Bristol Myers Squibb™ AstraZeneca NOVARTIS Roche CDD COLLABORATIVE DRUG DISCOVERY

Pistoia Alliance

Create FAIR data
with intention

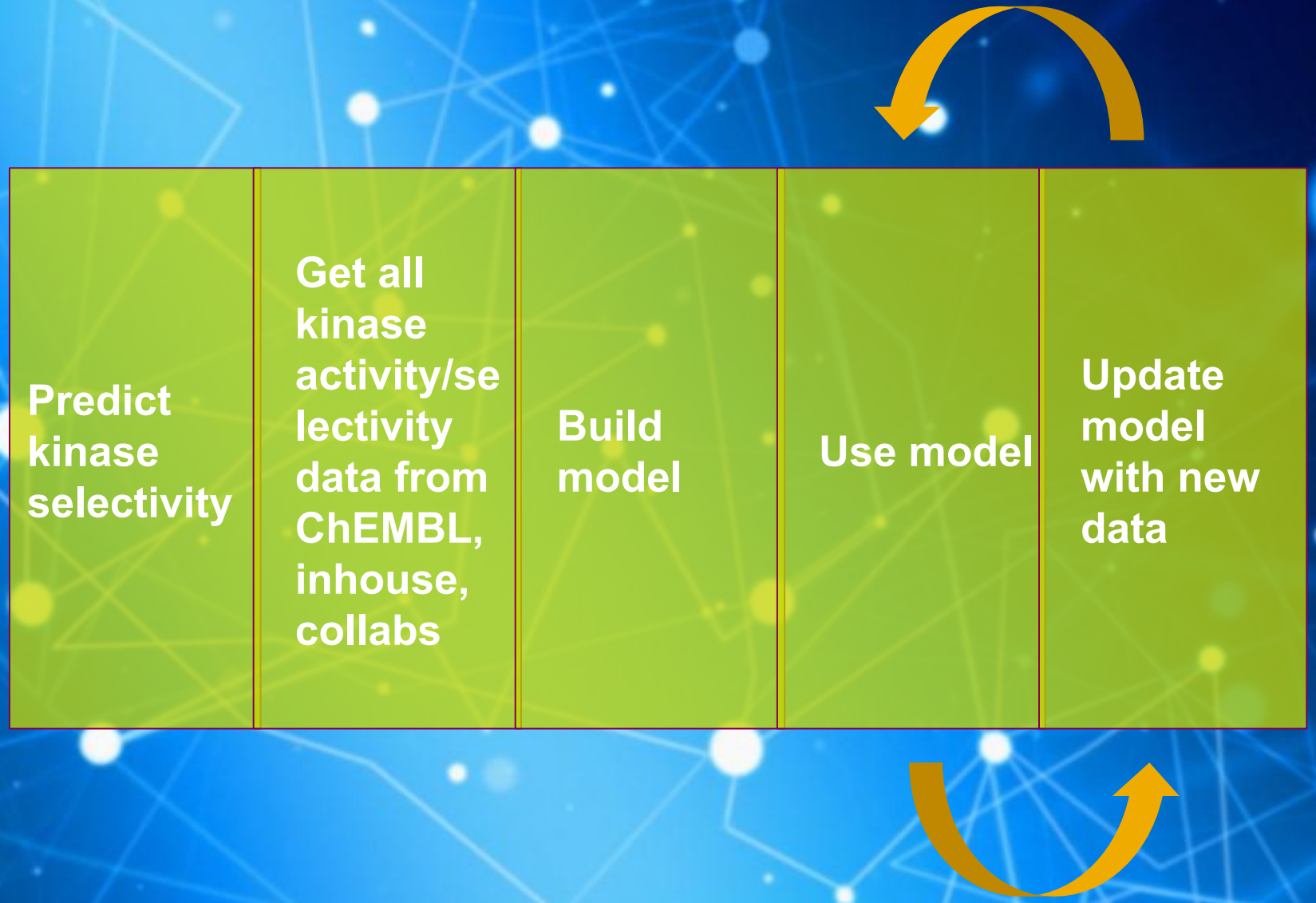


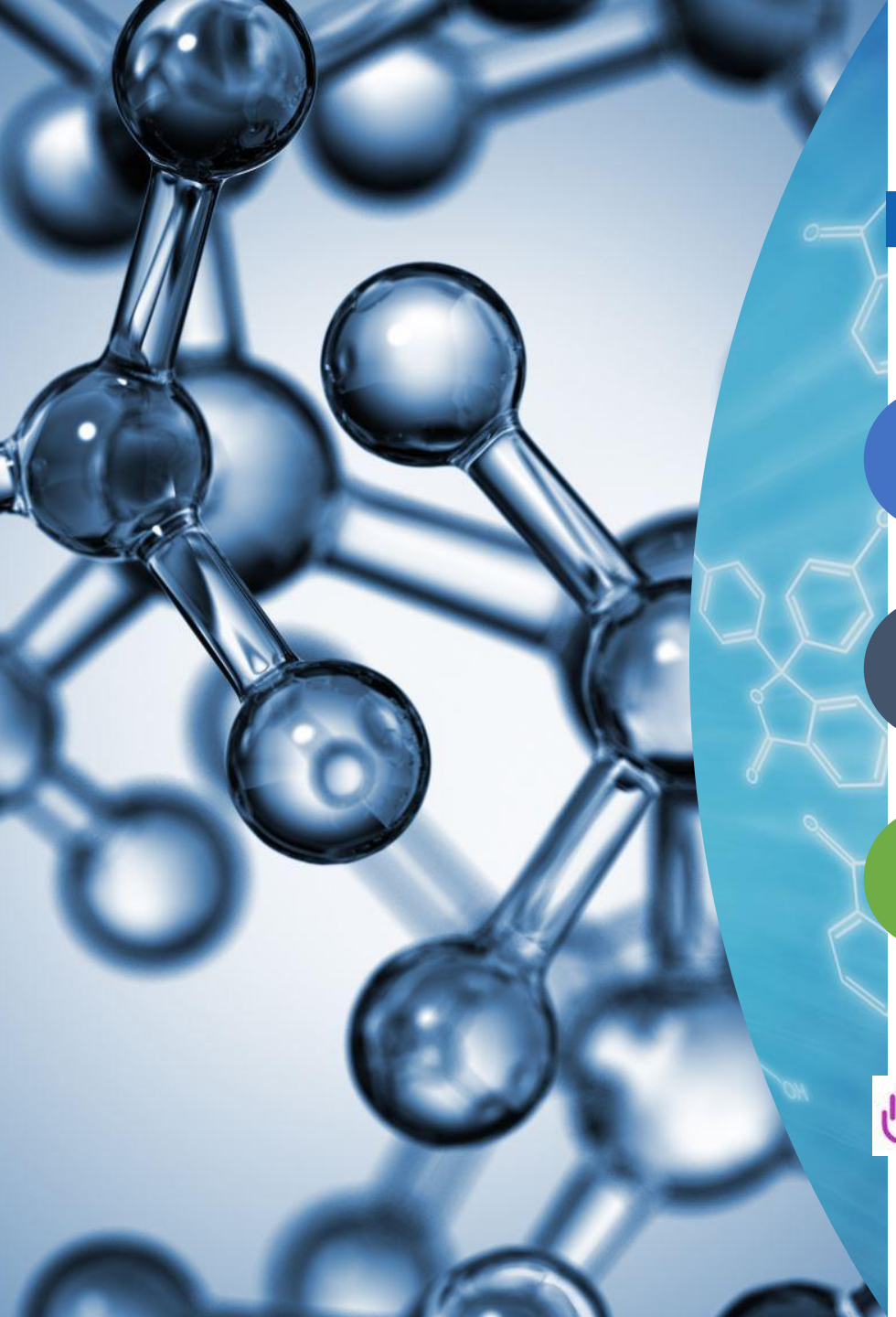
Pilot Project - Business questions

26 initial questions, pruned down to 15, across 5 main categories

1. Biology oriented literature mining for discovery project planning
2. Assay technology oriented
3. Chemistry/tool compound oriented
4. Specific assay conditions
5. Computational chemogenomic modelling (e.g., target activity, ”PAINS”)

Example


Predict kinase selectivity
Get all kinase activity/selectivity data from ChEMBL, inhouse, collabs
Build model
Use model
Update model with new data

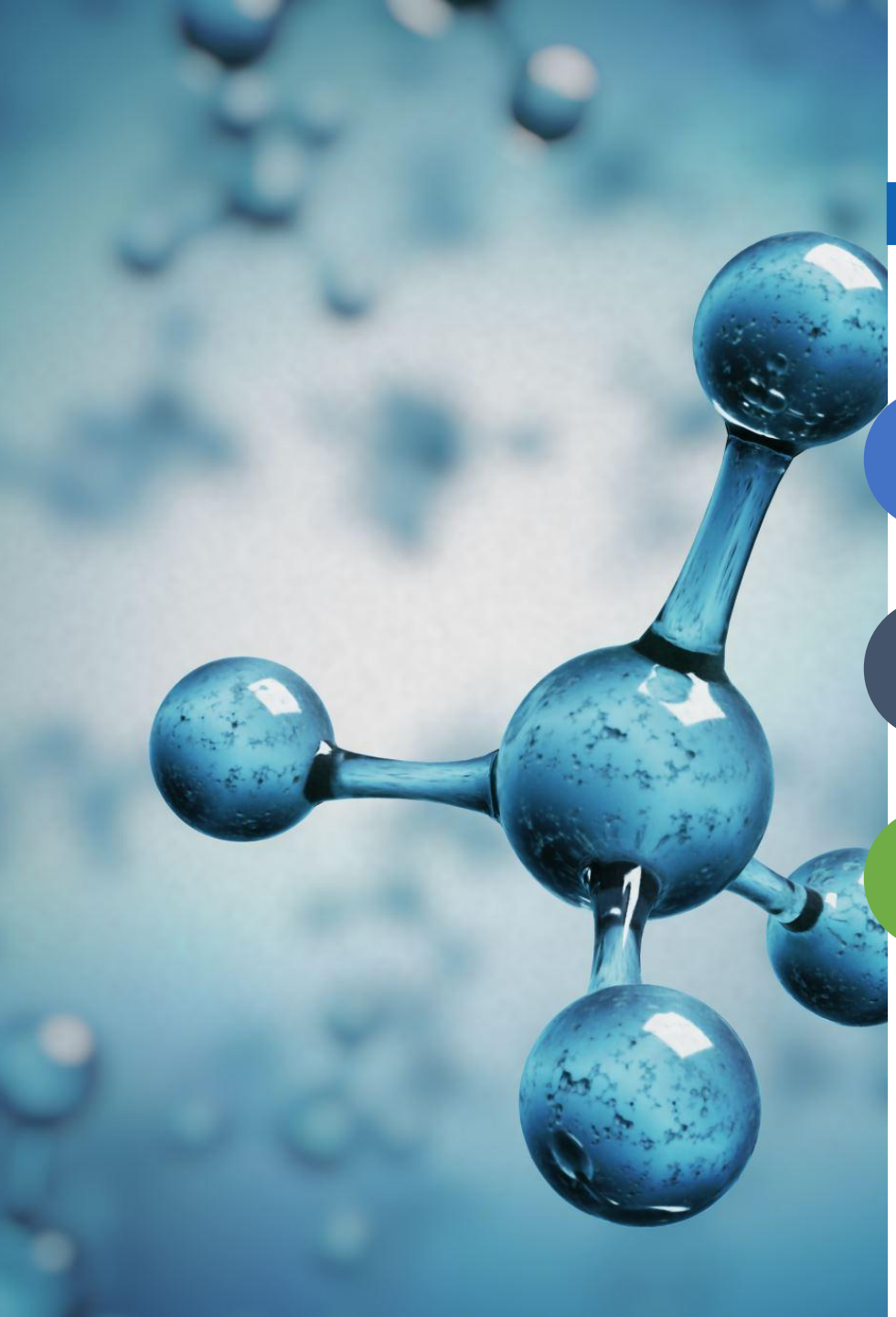


Pilot project (2020) – Summary

- Feasibility study, guidance for a larger initiative, example creation
- Curation of 496 public domain assay descriptions were converted into FAIR information objects **using an agreed data model**, which was guided by jointly defined business questions. Upload of the metadata to PubChem.
- Learning points were captured along with recommendations for future endeavors

Bristol Myers Squibb™ AstraZeneca NOVARTIS Roche

Pistoia Alliance



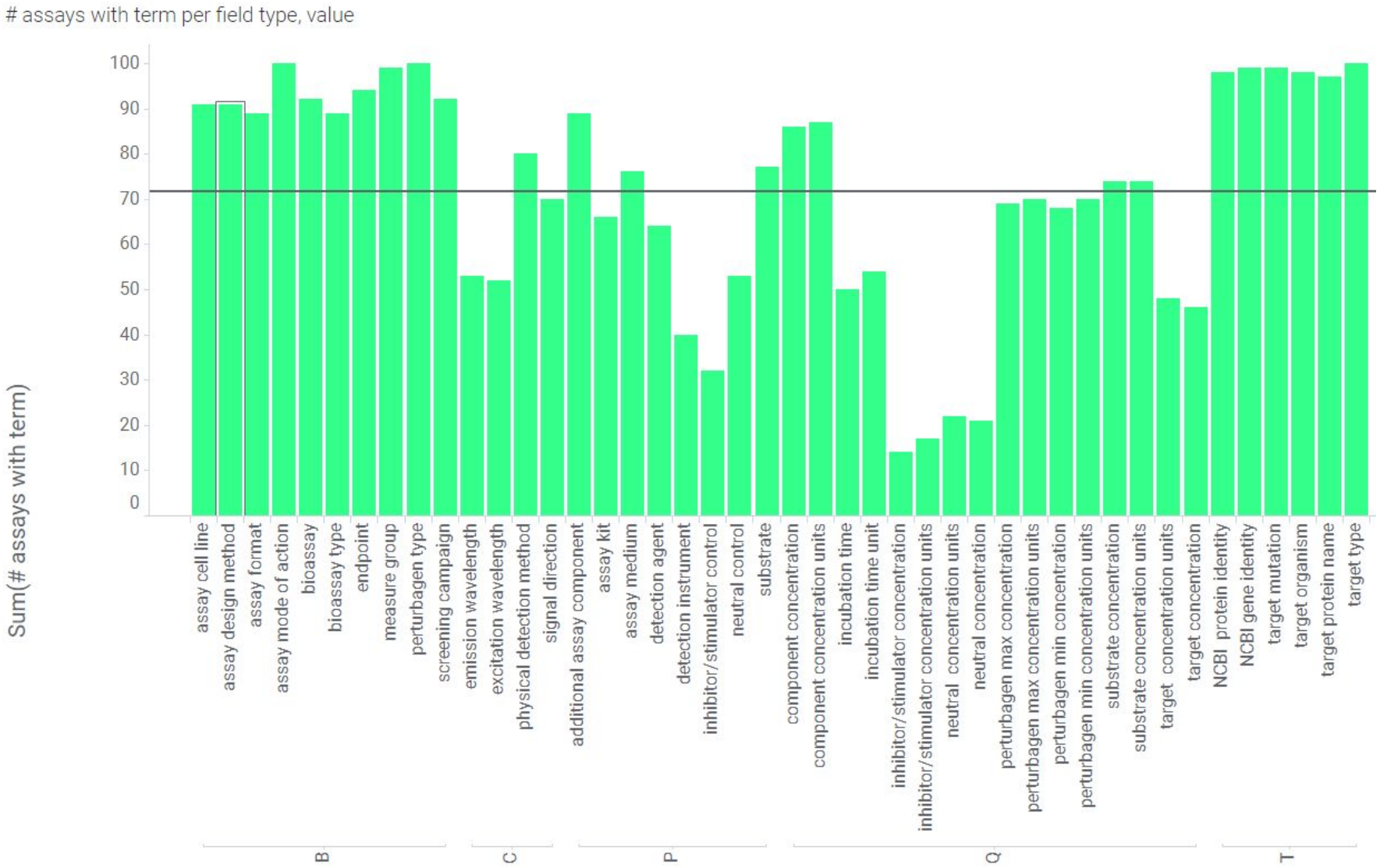
Pilot Project – Assay Selection

1. 245 Commercial panel assays: ThermoFisher's kinase selectivity Z'-lyte panel
2. 42 PubChem NCATS assays – qHTS, large datasets
3. 210 publication assays: ChEMBL assays where the target is EGFR, and the reference is Open Access

100 were subjected to manual QC by project team members

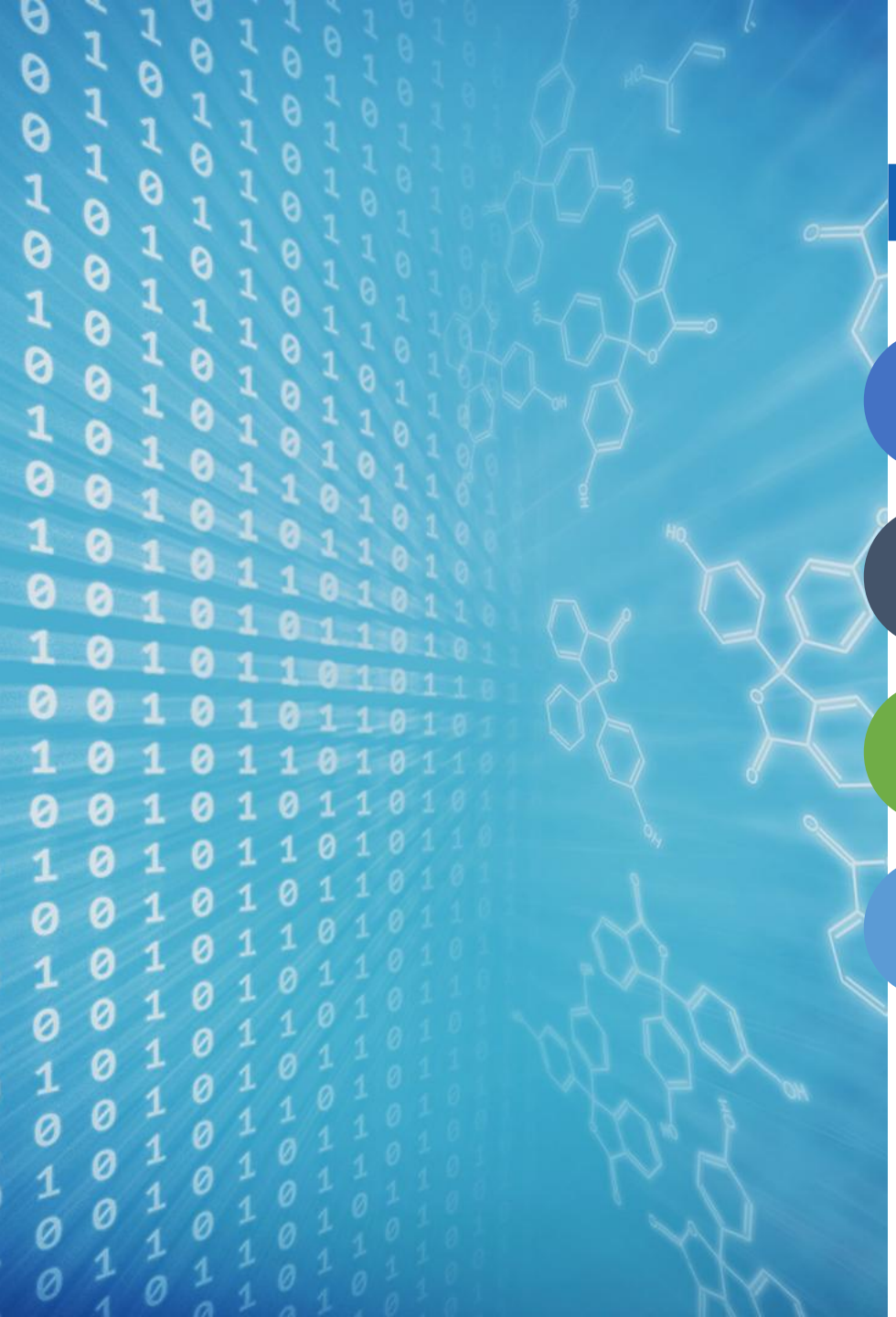
Annotator
METADATA SIMPLIFIED

How well did the pilot assays get annotated?



Pilot Project – Learnings

1. Review of supplements & citations = High cost. Choose assays wisely.
2. Commercial assay panels were the easiest to annotate (low-hanging fruit)
3. No persistent links exist for commercial assay panel protocols
4. Errors propagate between papers
5. Fully automated is not fully accurate: Benefit from good work practices: audit trail, versioning, iterative QC
6. Need for published assay protocols to be well-annotated in public databanks and linked to the publication
7. Need for a common community data standard for future assay publications.
8. Need for flexible & synergized Ontologies (e.g.: OntoloBridge)



Next steps:

1. **Scale up** (x 10-100) in next steps. Having more partners lowers cost per partner per assay and overhead cost
2. **Optimize** process, data sources, tools, QC within quality constraints. Define quality metrics.
3. Define and promote a **community standard** for assay reporting and publishing -- **align** with vendors, publishers, government agencies.
4. Attract new project members and sufficient funding to start the next phase

Thanks to

AstraZeneca
Nigel Green
David Hayes
Tom Plasterer

BMS
Rick Bishop

Janssen
Herman van Vlijmen

Novartis
Fabien Pernot

MMV
Jeremy Burrows

PubChem
Evan Bolton

ChEMBL
Anna Gaulton
Andrew Leach

Roche
Olivier Roche

Medicines Discovery Catapult
John Overington
Mark Davies

Pangeadata.ai
Vibhor Gupta

University of Miami
Stephan Schürer

BioSci Consulting
Scott Wagers

Collaborative Drug Discovery
Barry Bunin
Frank Cole
Alex Clark
Hande Kücük McGinty
(now Univ. Of Ohio)

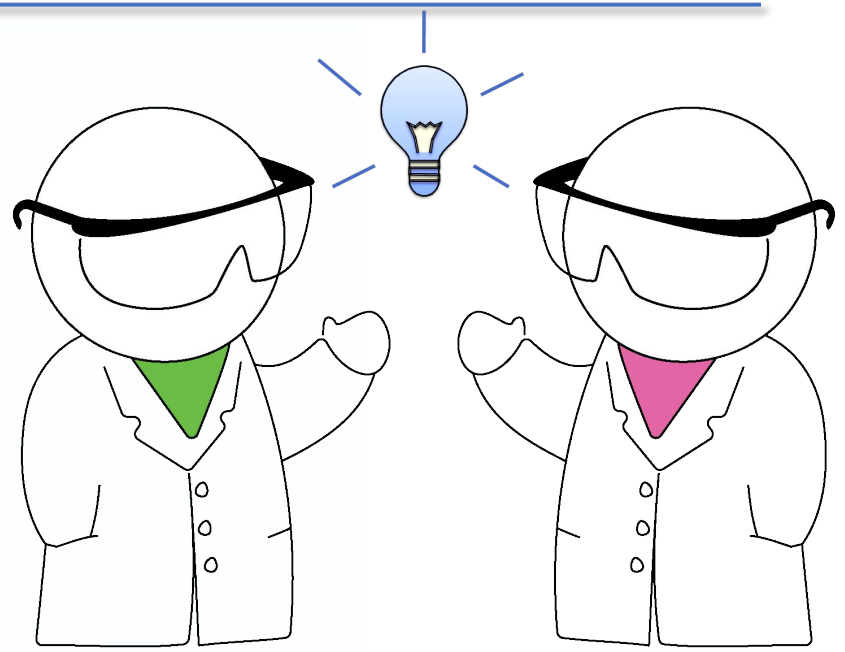
Pistoia Alliance
Carmen Nitsche *(now CCDC)*
Nick Lynch *(now Curlew Research)*

Metadata Managed! – Annotator In Action

Unique ID | assay ID | common assay template | Template | Sections | new content

Protocol Full Text

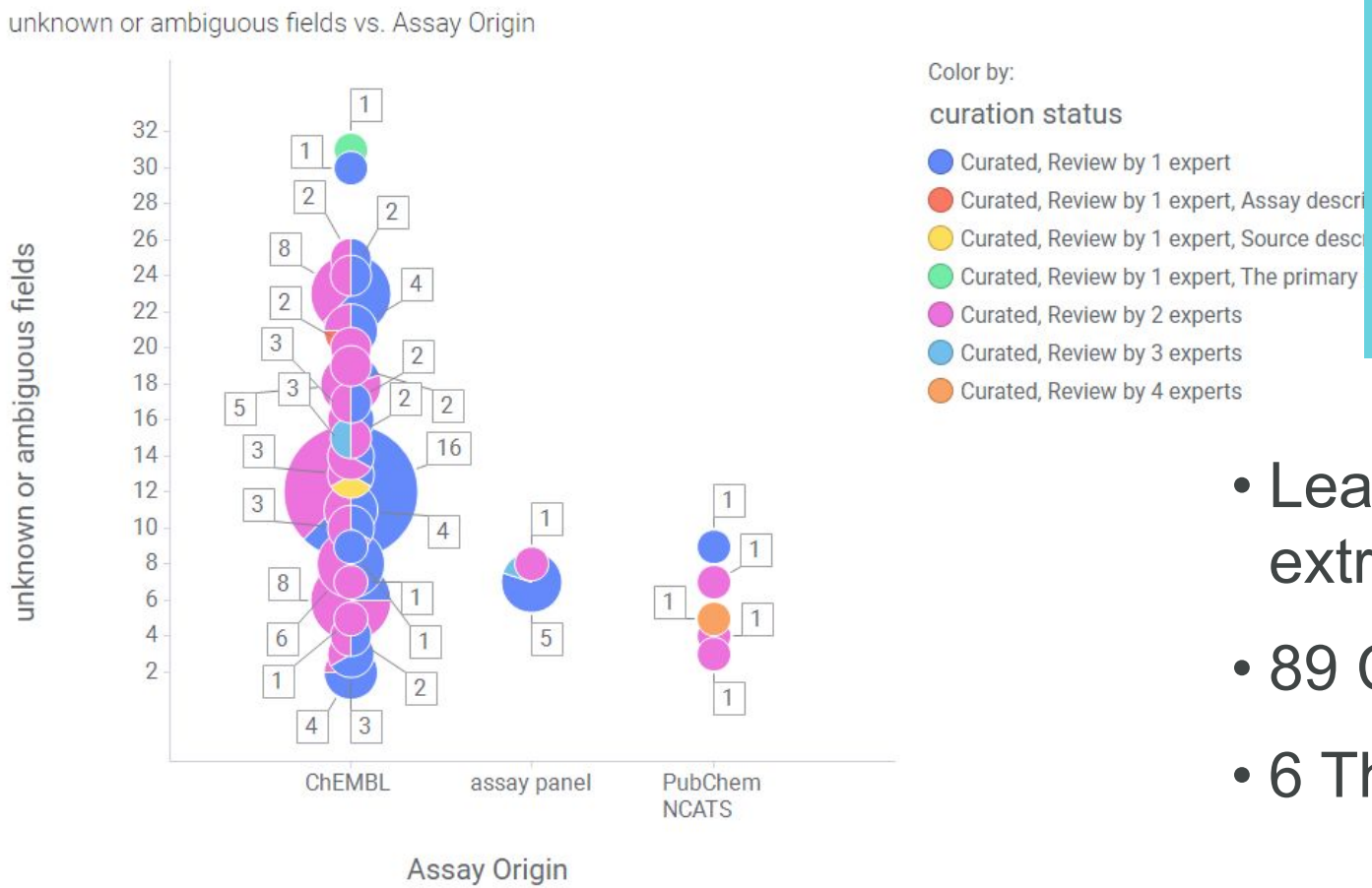
assay title
bioassay type
bioassay
assay format
assay design method
assay supporting method
assay cell line
organism
biological process
target
applies to disease
assay mode of action
result
screening campaign stage
assay footprint



Findable **A**ccessible **I**nteroperable **R**eusable

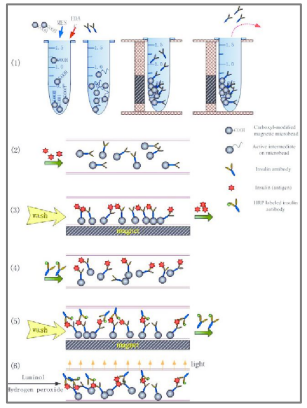
COLLABORATIVE DRUG DISCOVERY

Pilot: 100 QC:d assays (~20%)



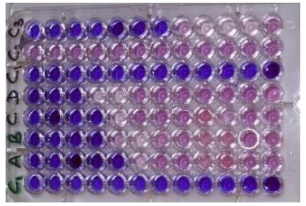
- Learning points are largely extrapolating on the 100 QC:d assays
- 89 ChEMBL assays, 5 NCATS assays
- 6 ThermoFisher panel assays QC:d

Are you capturing enough data today to answer *tomorrow's* questions?



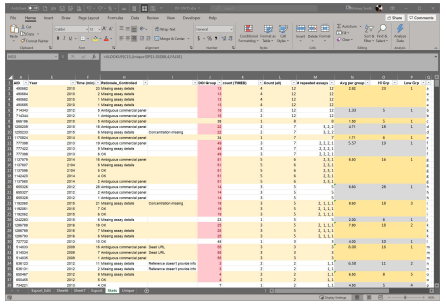
Procedures

Instrumentation

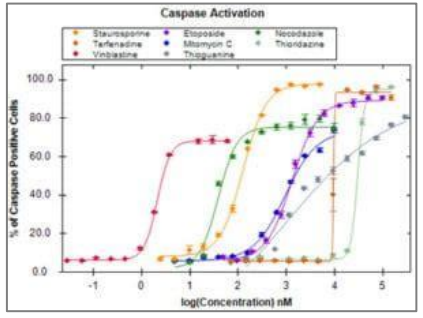


Samples & Results

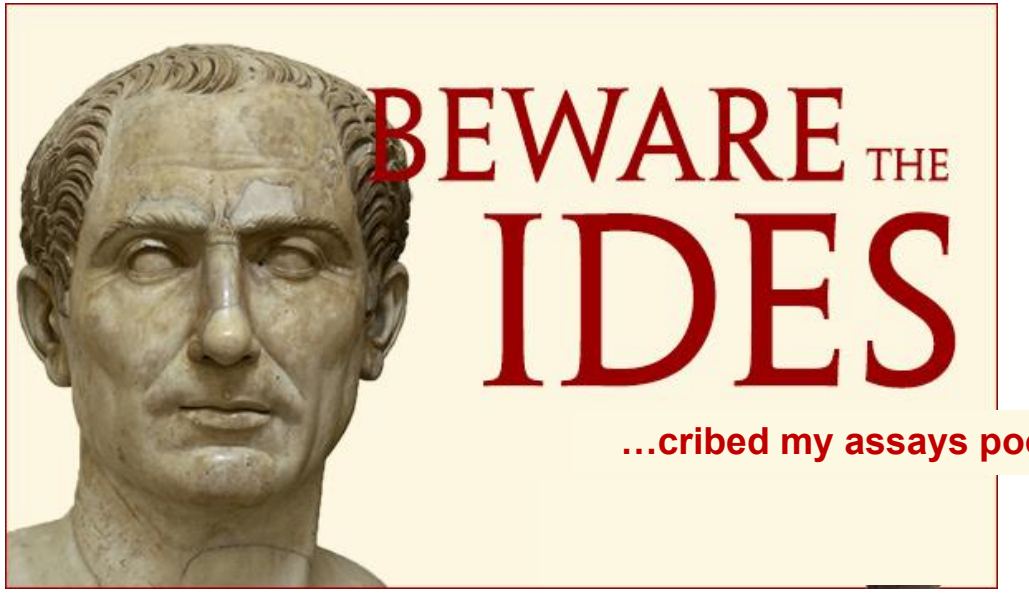
Compound	Result (mM)	Note
XYZ-0001234	12	Added too much Buffer 1
XYZ-0001235	105	
XYZ-0001236	>200	Didn't look gud
XYZ-0001237	<5	
XYZ-0001238	<5	
XYZ-0001239	>200	Looked bad
XYZ-0001240	55	
XYZ-0001241	56	
XYZ-0001242	20	
XYZ-0001243	25	Sent to Fred for follow-up (New Fred, not Old Fred)



Raw Data & Observations



Data Analysis & Assumptions



Why DataFAIRy?

There is a need for FAIR public domain data with high quality annotations using public ontologies and a common data model

Available metadata in (public domain) data repositories is often insufficient for answering current and future business questions

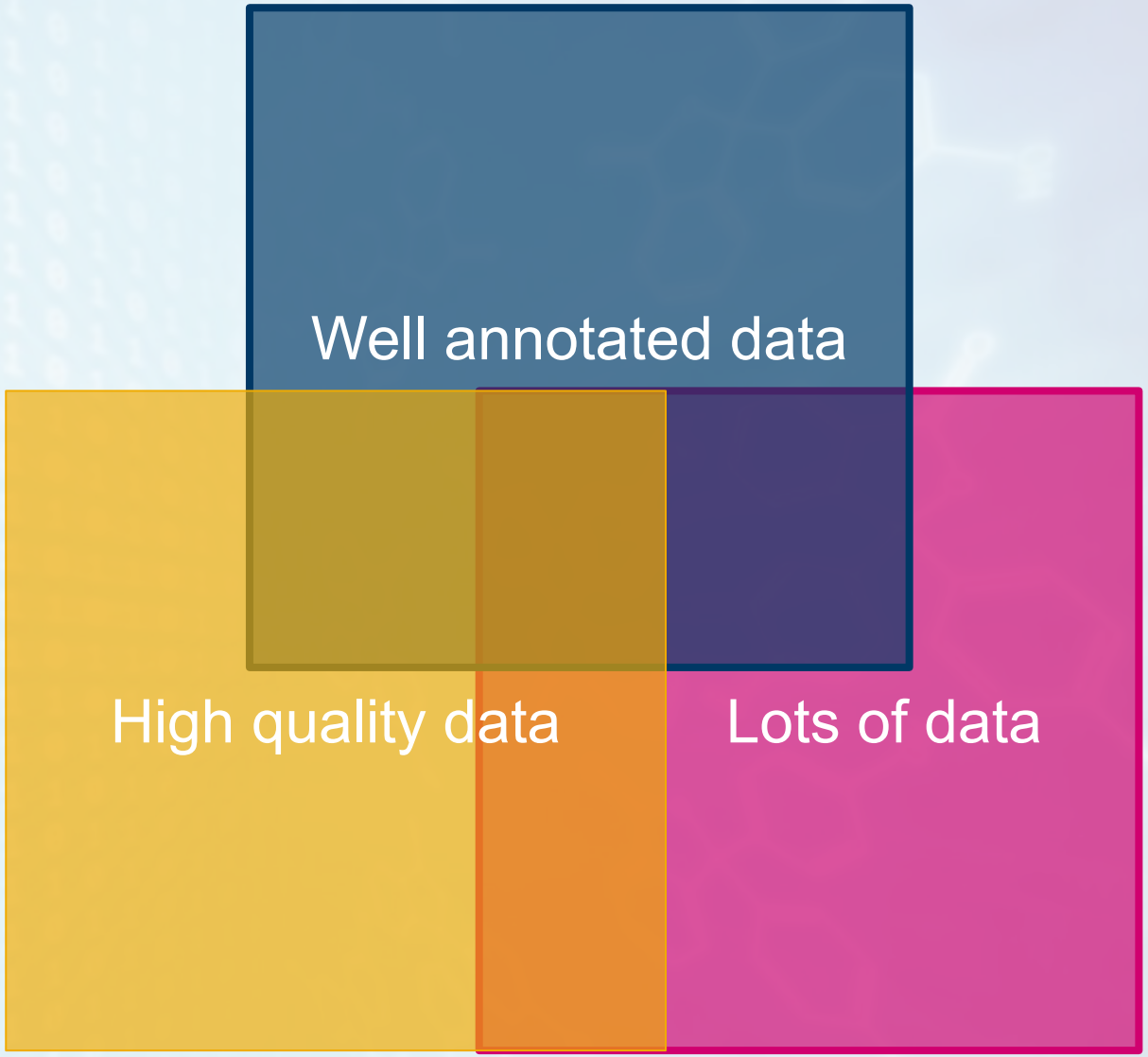
Substantial investments are being made in AI, ML and FAIR data across life science industry and academia

Pharma companies already pay for curation of partially overlapping public domain data (e.g., ChEMBL, papers, chemistry patents)

Small molecule bioassays make up a good pilot case

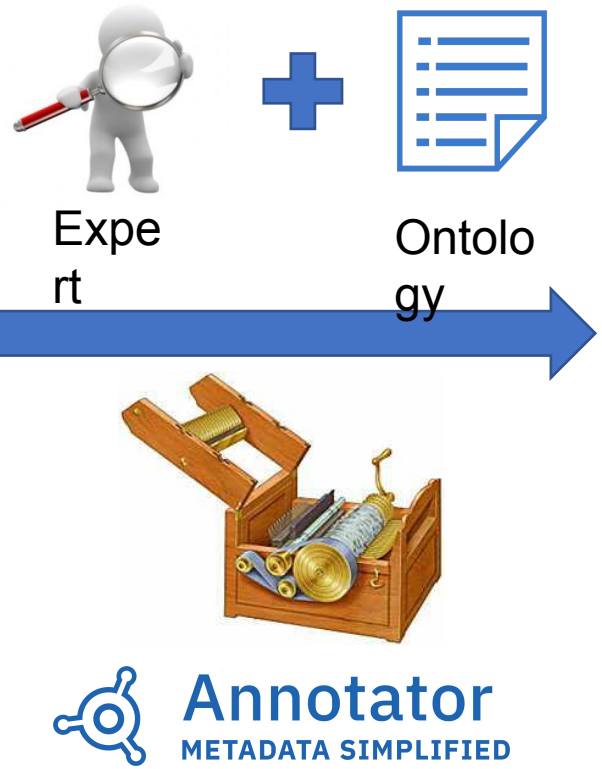
- Project planning – what is available in the public domain?
- Assay development, e.g., assay conditions and tool compounds
- Chemogenomic model building
- Enriching public chemogenomics data with FAIR metadata will show impact across the cheminformatics domain

What we want



A Generic Ontology-Driven Annotation Tool

Ambiguous, Unstructured Data



Unambiguous, Structured Data

Example annotation use cases: Biochemical, cell, phenotypic assays, pharmacology assays, clinical assay data, bicycles…

CDD COLLABORATIVE DRUG DISCOVERY