Ana2

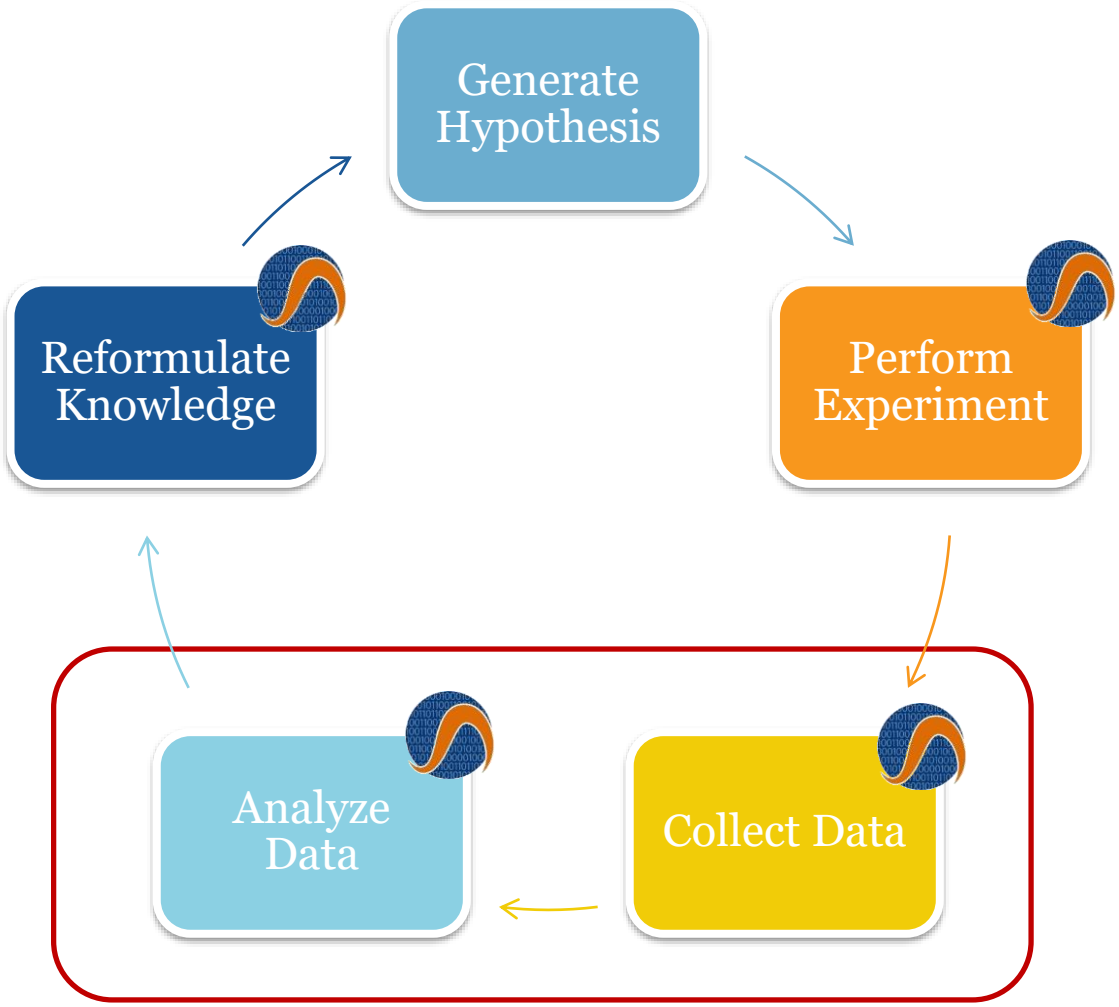
Ana2:

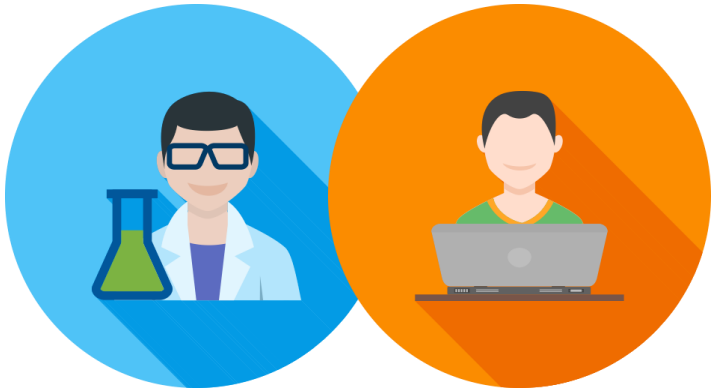
Systematic, centralized archiving and automation of raw analytical instrument data

Corey E. Bakalarski, PhD
Scientist
Depts. Of Microchemistry, Proteomics & Lipidomics
and Bioinformatics & Computational Biology
Genentech Research and Early Development (gRED)

Genentech
A Member of the Roche Group

Managing the Scientific Knowledge Lifecycle





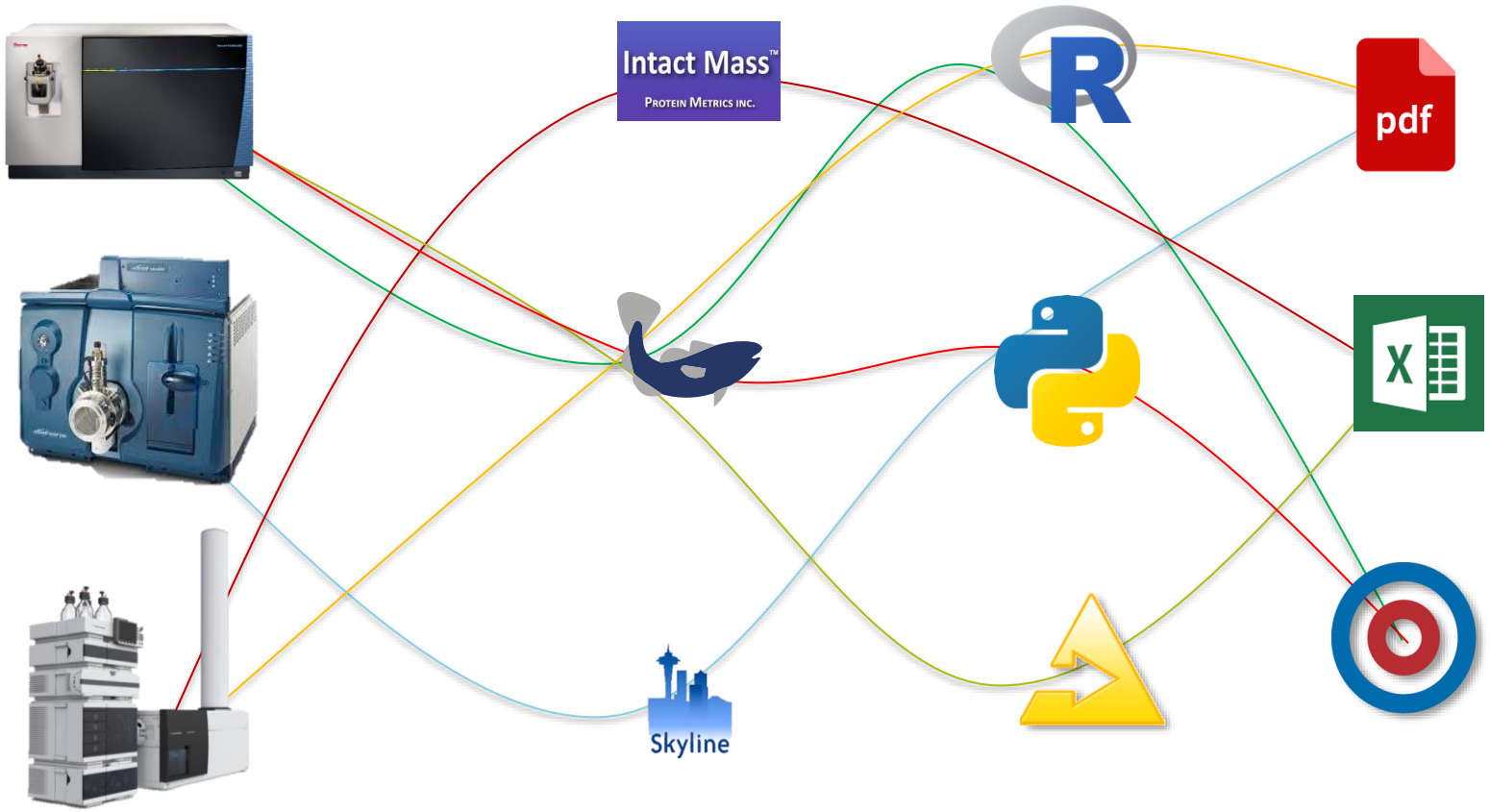
The FAIR Guiding Principles:

- To be Findable
- To be Accessible
- To be Interoperable
- To be Reusable

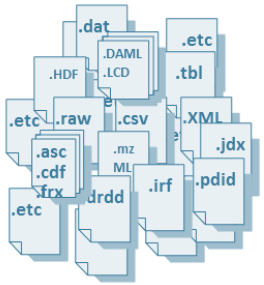
A Diversity of Data and Analysis Tools: The Research Perspective

Many
Workflows
× Many
Instruments
× Many
Formats
× Many
Analysis Types
× Many
Result Types

A Challenging
Landscape to
Connect
Information



Process
Material
Equipment
Result



.adf

But where do we start?

Our raw instrument data is the starting point for all of our biological knowledge…

…and the foundation upon which everything is built.

A Broad Approach, Driving Broad Adoption

Ana2

- Easily Accessible, Anywhere, Anytime
- Single Source of Truth
- Common Interface
- A Sense of Ownership
- A Bridge to Automation
- Starting Point for Standardization

Key Deliverables of Ana, Our Organization Platform

Standardization

Consistent Formats for Consistent Handling

Flexibility

Capable of Capturing Varied Data Types

Scalability

Capable of Simultaneous Performance on 100+ instruments

Evolution

Able to Grow with the Data as it Evolves

- Single, Centralized Data Resource (+CRO data)
- Potential for Cross-platform Integration and Data Mining
- Enables Automation,
- Decreases Costs
- Increases Throughput

Continuous Integration for Laboratory Data

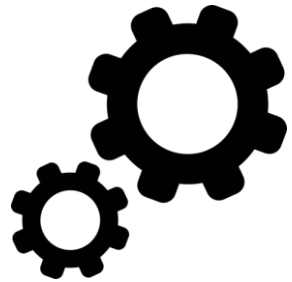
Push vs. Pull: Performance

Lightweight Monitoring Agent

Ana submission servers

Scale to 100s of Instruments

Automate Data Processing



Analysis Tools

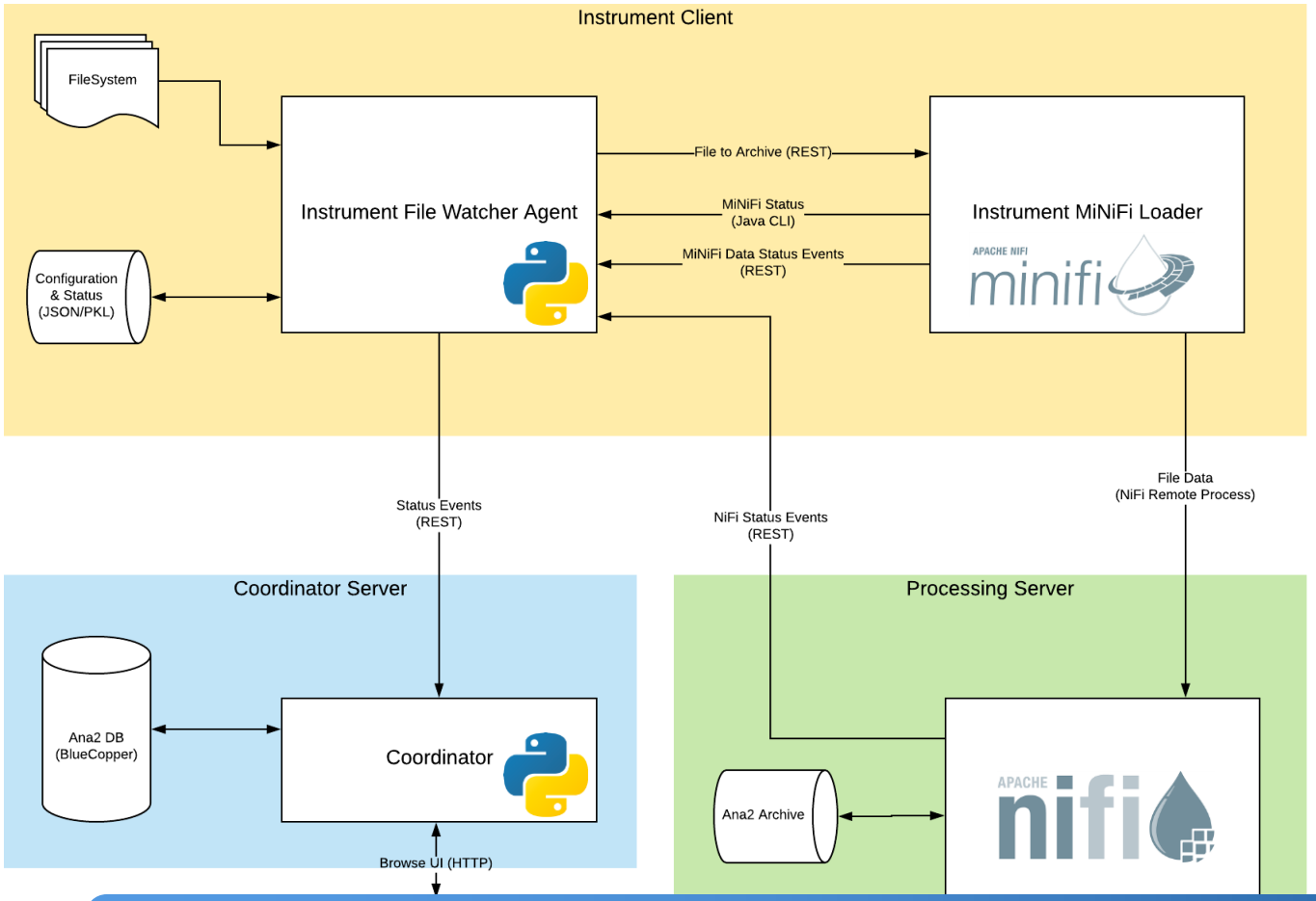
RESTful Web Services

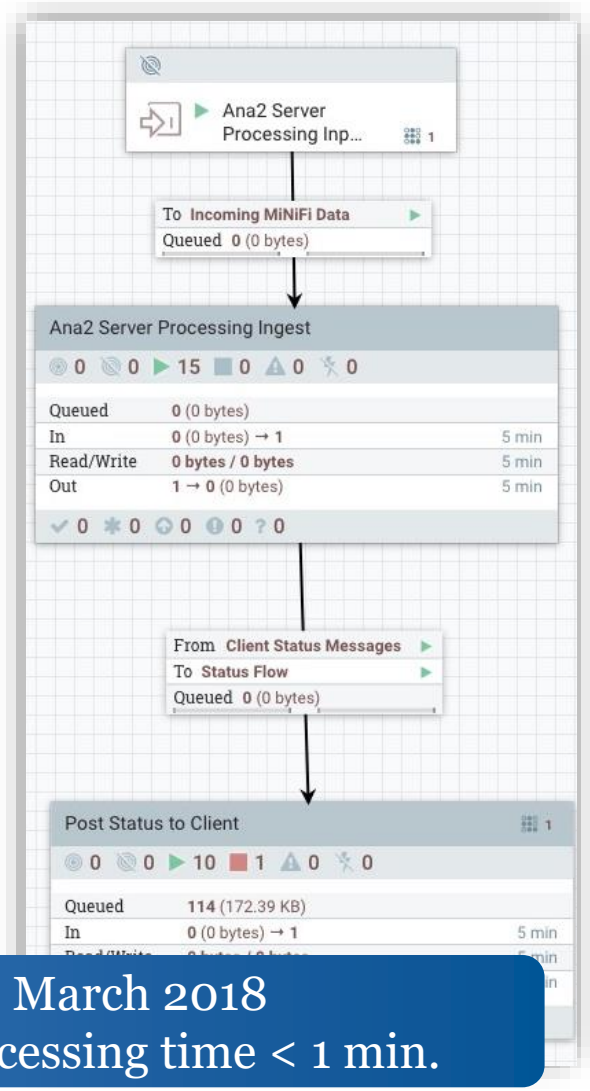
Ana Data Lake

Ana web-based user and administration interface

Whole Dataset Delivery

A Scalable, Rapidly Deployable Architecture





Started development in January 2018; Deployed Beta in March 2018
Processed 50,000 files from 12 instruments to date: average processing time < 1 min.

Focusing on a Lightweight, Agile Approach

Agent

- Minimize Footprint (Python)
- Plugin Architecture

Ingest

- Apache NiFi
- Centralized Processing

Lifecycle

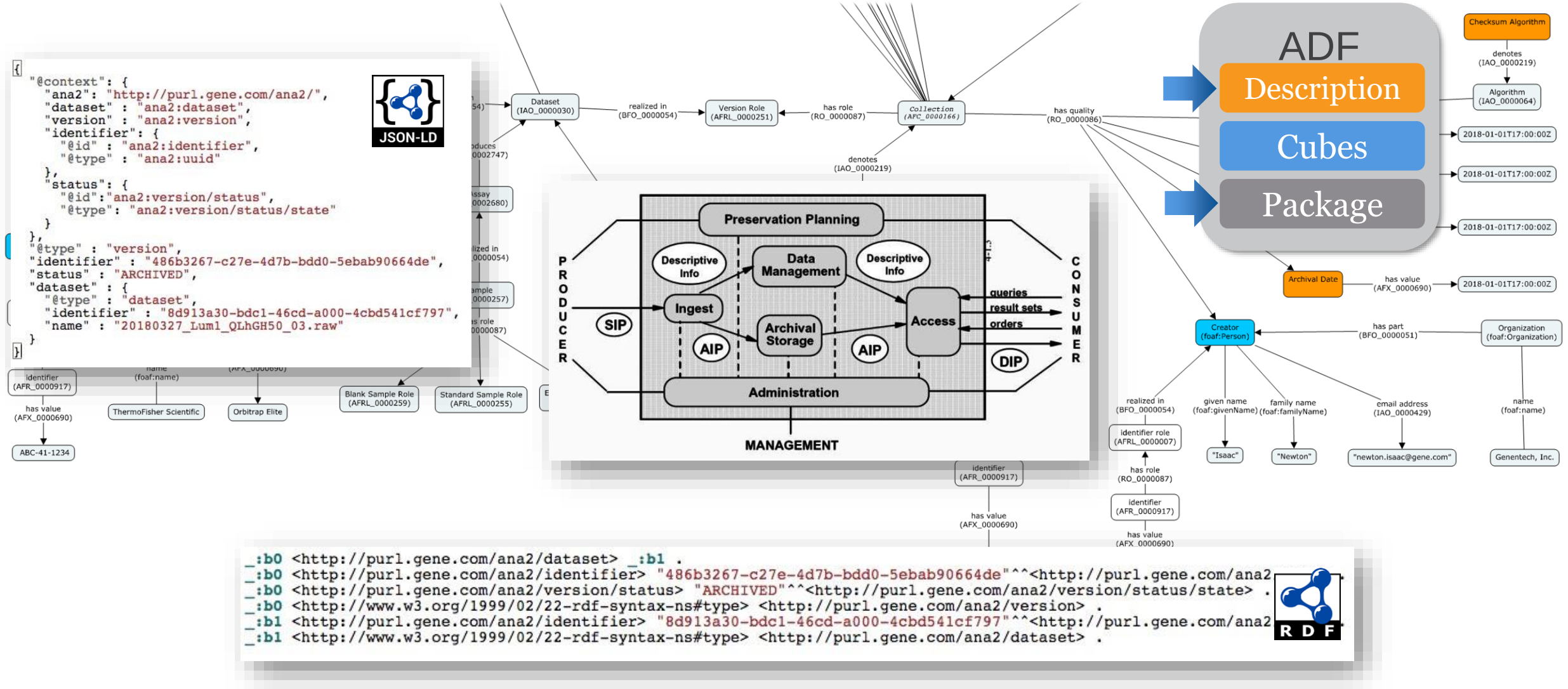
- Modular Approach
- Simplified Installation, Maintenance

Interface

- Streamlined Experience
- Simple REST Interface

Allotrope Formats Solve the Data Packaging Challenge

Drafting a Flexible Common Archival Data Model



Partnering with APN Member Companies To Advance Allotrope Technologies

When are analyses complete?

How can metadata be extracted?

ThermoFisher SCIENTIFIC

Agilent

SHIMADZU
Excellence in Science

Retrieving Data from Ana2

Ana2 Client Status

Name	Hostname	Status	Last Update	Location	IP Address	Identifier
Ana Instrument 02	ANAFILES02.roche.com	ACTIVE	2018-04-17 10:19:46	None	10.36.244.113	051cb574-232f-11e8-a4db-005056ae47c9
Ana Instrument 03	ANAFILES03.roche.com	OFFLINE	2018-03-22 10:49:30	None	10.36.244.114	ae7d1a06-2235-11e8-b620-005056ae5850
MPL Elite1	Lab413042-Eli1.gene.com	ACTIVE	2018-03-29 16:26:03	None	10.29.104.6	d2c23722-235f-11e8-aab3-d067e5e7f504
MPL Lumos1	LAB413040-LUM1.gene.com	ACTIVE	2018-04-17 10:19:45	None	1	
Ana Instrument 04	ANAFILES04.roche.com	ACTIVE	2018-04-17 10:19:37	None	1	
MPL Sciex Qtrap 6500	lab413042-Q6500.gene.com	ACTIVE	2018-04-17 10:19:35	None	1	
MPL Lumos2	LAB413040-LUM2.gene.com	ACTIVE	2018-04-17 10:19:46	None	1	


```
{recordsTotal: 49367, recordsFiltered: 49367, draw: 8,…}
  ana2_data: [{first_seen_date: "2018-04-16 23:07:46", version_id: 60963,…},…]
    0: {first_seen_date: "2018-04-16 23:07:46", version_id: 60963,…}
    1: {first_seen_date: "2018-04-16 22:11:37", version_id: 60962,…}
    2: {first_seen_date: "2018-04-16 22:38:05", version_id: 60961,…}
    3: {first_seen_date: "2018-04-16 22:00:16", version_id: 60960,…}
    4: {first_seen_date: "2018-04-16 22:08:23", version_id: 60959,…}
    5: {first_seen_date: "2018-04-16 21:39:18", version_id: 60958,…}
    6: {first_seen_date: "2018-04-16 21:38:43", version_id: 60957,…}
    7: {first_seen_date: "2018-04-16 21:28:42", version_id: 60956,…}
    8: {first_seen_date: "2018-04-16 20:52:02", version_id: 60955,…}
      client: "MPL Elite2"
      current_status: "ARCHIVED"
      current_status_comment: null
      current_status_date: "2018-04-16 21:44:56"
      dataset_id: 68885
      filename: {,…}
      first_seen_date: "2018-04-16 20:52:02"
      original_filename: "20180416_Eli2_isoB_hGH_midJT_02.raw"
      original_path: "C:/Xcalibur/data/standards/20180416_Eli2_isoB_hGH_midJT_02.raw"
      size: "69.9 MB"
      version_id: 60955
      version_uuid: "bbad1148-2fef-471e-9b71-e53b442bed90"
    9: {first_seen_date: "2018-04-16 21:09:00", version_id:
  draw: 8
  recordsFiltered: 49367
  recordsTotal: 49367
```


API

Ana2 Processed datasets (49,365)

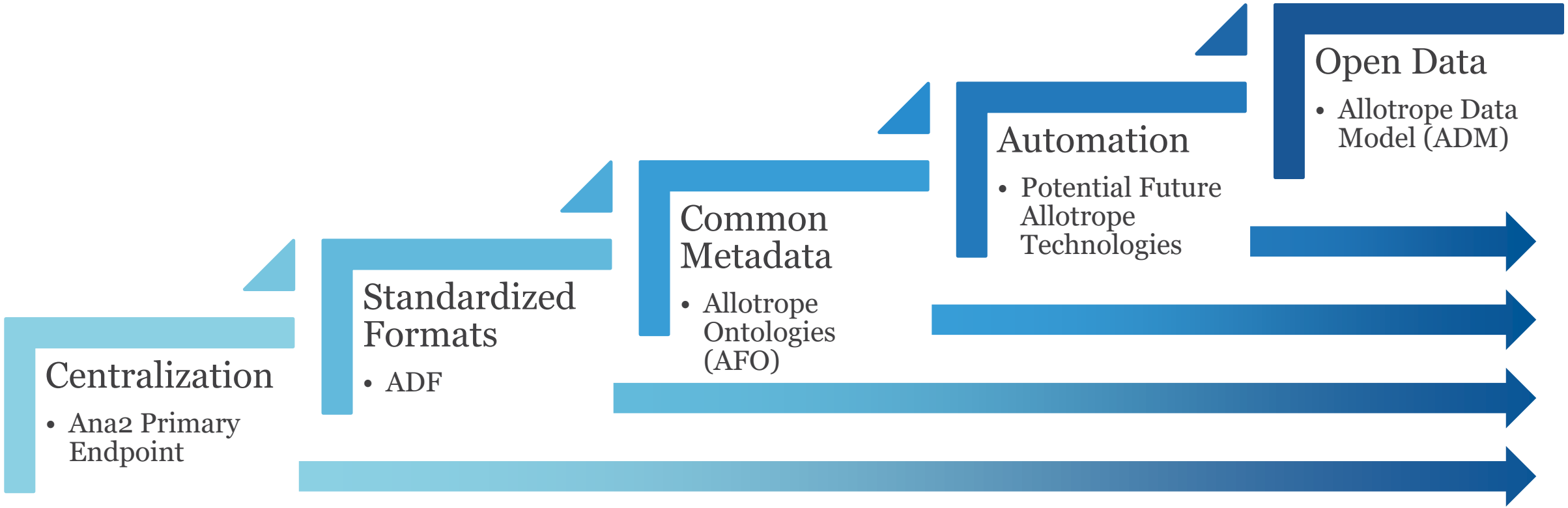
Client Status

Show 10 entries Filter Filenames:

Date	Uuid	Status	Instrument	Filename	Size	Comment
2018-04-17 01:44:11	f6c4a07…	ARCHIVED	MPL Fusion	20180416_Fus1_QLhGH50_CAL_newcolumn_08.raw	41.0 MB	
2018-04-17 01:26:24	4fdf867…	PROCESSING	MPL Elite2	20180416_Eli2_QLhGH50_02.raw	30.5 MB	
2018-04-17 01:12:51	f6d7fd7…	ARCHIVED	MPL Fusion	20180416_Fus1_QLhGH50_CAL_newcolumn_07.raw	55.3 MB	
2018-04-17 00:53:53	2730c36…	PROCESSING	MPL Elite2	20180416_Eli2_QLhGH50_01.raw	33.5 MB	
2018-04-17 00:40:57	f329c37…	ARCHIVED	MPL Fusion	20180416_Fus1_isoB_hGH_05.raw	35.6 MB	
2018-04-17		ARCHIVED	MPL Elite2	20180416_Eli2_QLcometK562Std_01.raw	132.7 MB	
		READY_TO_ARCHIVE	OMNI-BD San Francisco	6500bm041318_Abc_LysC_IAQUA_SC_SC2-09.wiff.scan	195.8 kB	
		ARCHIVED	MPL Fusion	20180416_Fus1_isoB_hGH_04.raw	34.2 MB	
		READY_TO_ARCHIVE	OMNI-BD SF	6500bm041318_Abc_LysC_IAQUA_SC_SC2-08.wiff.scan	350.6 kB	
		ARCHIVED	MPL Elite2	20180416_Eli2_OBJ0012345_Def_QLabcd.raw	88.2 MB	

Previous 1 2 3 4 5 … 4937 Next

A Layered, Evolutionary Approach to Data Organization



Our Approach: Deploy ADF Widely to Corral Data and Layer on Open Data as ADM Evolves

Allotrope: Paying Dividends for Ana2

Process does not Define Product

- Using standards like AFO, ADM, ADF allow innovation and modularization
- Open standards ensure data remains accessible (now and in the future)

Open Standards Ensure Accessibility

- Now: allow automation and centralization
- Future: provide compatibility for legacy data
- Renewed focus on the science

A Beneficial Community - with a Purpose

- Opportunities to shape the future to be beneficial for all: data models and software
- Connections and a common goal

Genentech

A Member of the Roche Group

bakalarski.corey@gene.com

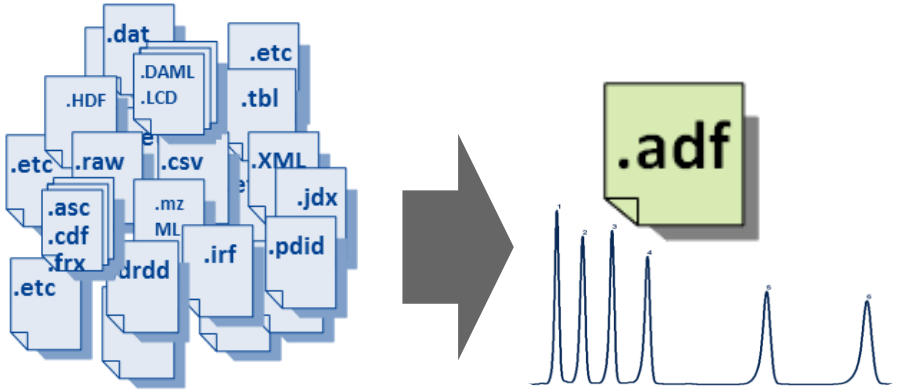
APPENDIX

Questions? Contact bakalarski.corey@gene.com

Genentech
A Member of the Roche Group

Vendor-Specific Formats

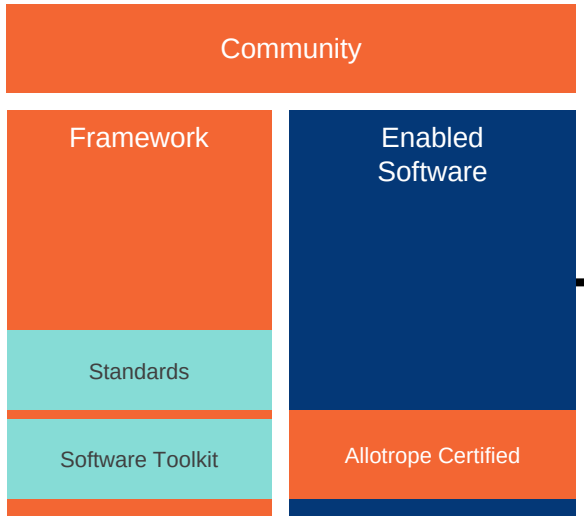
Data in Standard Format

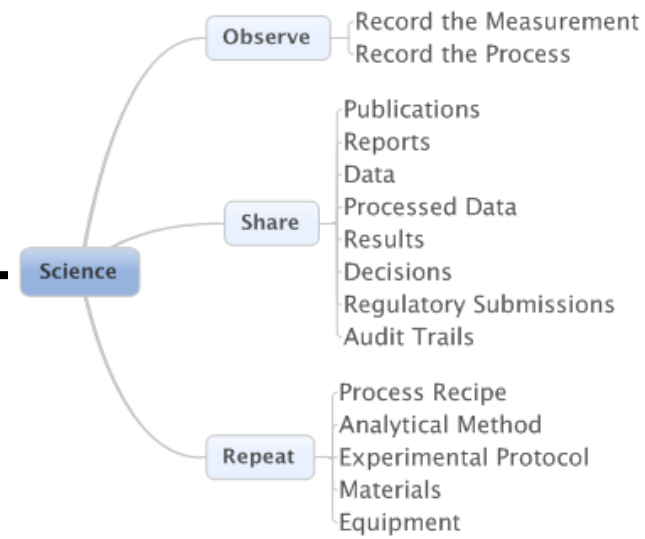


Paper, unstructured text for methods, regs, recipes

A Standard vocabulary & metadata structure

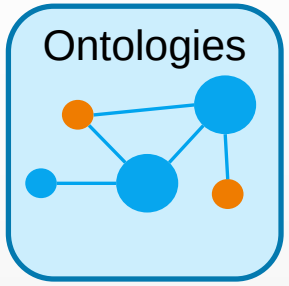
How do we fix the root cause of inefficiencies & data integrity challenges?



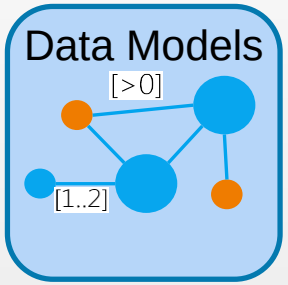


Allotrope Foundation Mission:
Revolutionize the way we acquire, share and gain insights from scientific data, through a community and the framework for standardization and linked data.

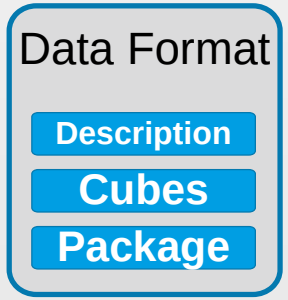
Allotrope's Three Foci



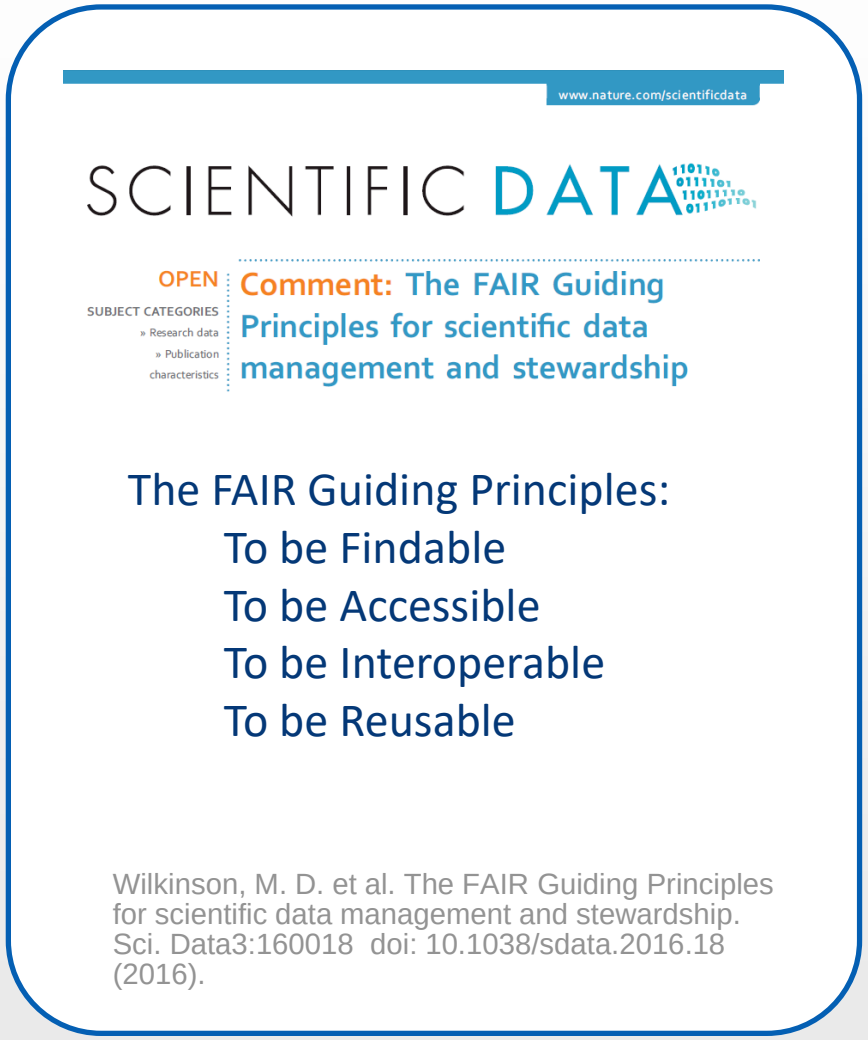
- **Ontologies** describe and relate the parts of a scientific analysis



- **Data Models** write the rules about what parts must be there



- Standardized **Data Format** to record data and metadata


www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

The FAIR Guiding Principles:
To be Findable
To be Accessible
To be Interoperable
To be Reusable

Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data3:160018 doi: 10.1038/sdata.2016.18 (2016).