



Data-Intensive Workflows

A journey to a Holistic Framework for
Data-Intensive Workflows

Ian Corner – Design and Implementation Lead – May 2016

INFORMATION MANAGEMENT AND TECHNOLOGY (IMT)



CSIRO – Who we are

Commonwealth Scientific and Industrial Research Organisation

5319

talented staff

\$1billion+
budget

Working
with over
2800+
industry
partners

55

sites across
Australia

Top 1%
of global
research
agencies

Each year
6 CSIRO
technologies
contribute
\$5 billion to
the economy

CSIRO – Our Mission

Strategy 2020 – Australia’s Innovation Catalyst

**Create value
for customers
through
innovation
that delivers
positive
impact for
Australia**



Projects and teams – creative, entrepreneurial, collaborative teams tackling big challenges through science, technology and innovation



Customer value – delivering value through innovative solution for customers in industry, government and community



Impact delivery – creating new economic, environmental and social impact for Australia

CSIRO – What we do

1854

patents
Biggest
patent holder
in Australia
30% involve
collaboration

150+

spin-out
companies
worth \$1bn in
market
capitalisation

300

licenses
Most with
Australian
companies



WiFi

Extended-wear
contact lenses

UltraBattery

**Building
IQ**

WASP

Zebedee

Globally our
publications are

**Top
1%**

in 15 of 22
research fields

1,200+

schools
benefit from
our scientists
in schools
program

**200,
000+**

people visit
our public
facilities and
visitor centres

CSIRO – Our Collections

Commonwealth Scientific and Industrial Research Organisation

Australian National Insect Collection

12,000,000 specimens (+100,000 per year)

Australian National Fish collection

5,000 species

Australian National Algae Culture Collection

1,000 strains of more than 300 micro-algae species

Australian National Herbarium

1,000,000 herbarium (Captain Cook's 1770 expedition to Australia)

Australian National Wildlife Collection

200,000 irreplaceable specimens of wildlife

<http://www.csiro.au/en/Research/Collections>

CSIRO – Yesterdays Collections

Physical collections, Captured and Preserved



<http://www.csiro.au/en/Research/Collections/ANIC>

CSIRO – Today's Collections

We need collections digitised, discoverable, consumable

The screenshot displays the CSIRO Data Access Portal interface. At the top left is the CSIRO logo. A blue header bar contains the text "Data Access Portal" and links for "Contact Us" and "Help". On the right, a "Registered Users" section shows a login form with the text "Login Using Partners" and a dropdown menu. The username field contains "cor22d" and the password field is masked with dots. A "LOGIN" button is positioned below the password field. Below the header, there are navigation tabs for "SEARCH", "BROWSE", and "DOMAIN SEARCH" with a magnifying glass icon. A breadcrumb trail reads "Home > Domain Search > ATNF Pulsar Observation Search". The main content area is titled "ATNF PULSAR OBSERVATION SEARCH" and includes a "New search" link. Under the heading "INSTRUCTIONS", there are three paragraphs of text. To the right, a search form is titled "Source Name / Position" and includes fields for "Source Name", "CONE SEARCH", "Right Ascension" (with a placeholder "hh:mm:ss.ss (J2000)"), "Declination" (with a placeholder "dd:mm:ss.ss (J2000)"), and "Search Window" (with a placeholder "arcmin"). There are "hide all" and "hide" links to the right of the search form.

<http://data.csiro.au/>

CSIRO – Today's Collections

Commonwealth Scientific and Industrial Research Organisation



RV Investigator is our state-of-the-art marine research vessel, supporting Australia's atmospheric, oceanographic, biological and geosciences research from the tropical north to the Antarctic ice-edge.

<http://www.csiro.au/en/Research/Facilities/Marine-National-Facility/RV-Investigator>

Data-Intensive Workflows

Where we started



Data-Intensive Workflows

As data growth and proliferation continued to outpace research grade infrastructure, we considered a new approach?

CSIRO started by asking what good is our data if it:

is unable to be found?

can not speak?

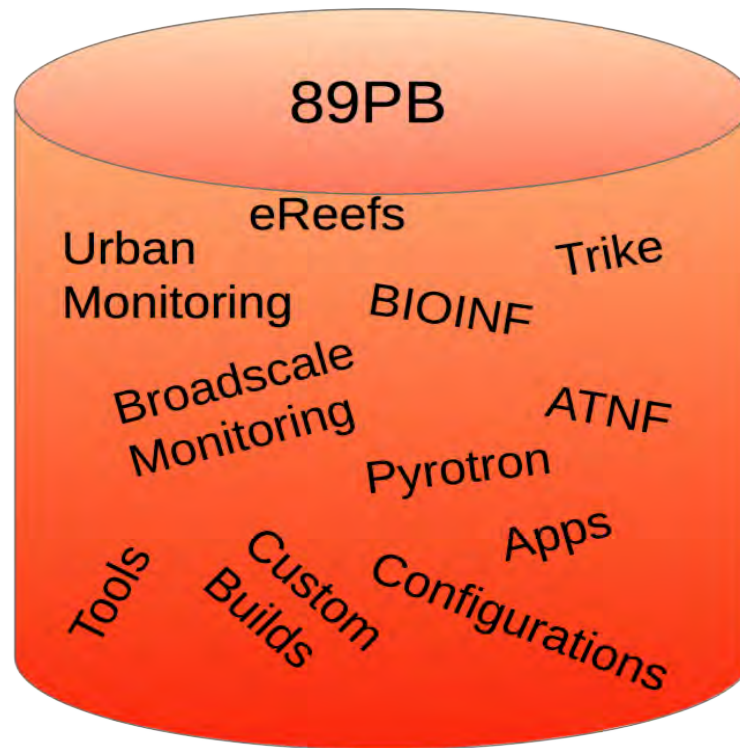
only ever repeats the same story?

can not repeat the same story twice?

speaks so slowly the message is lost?

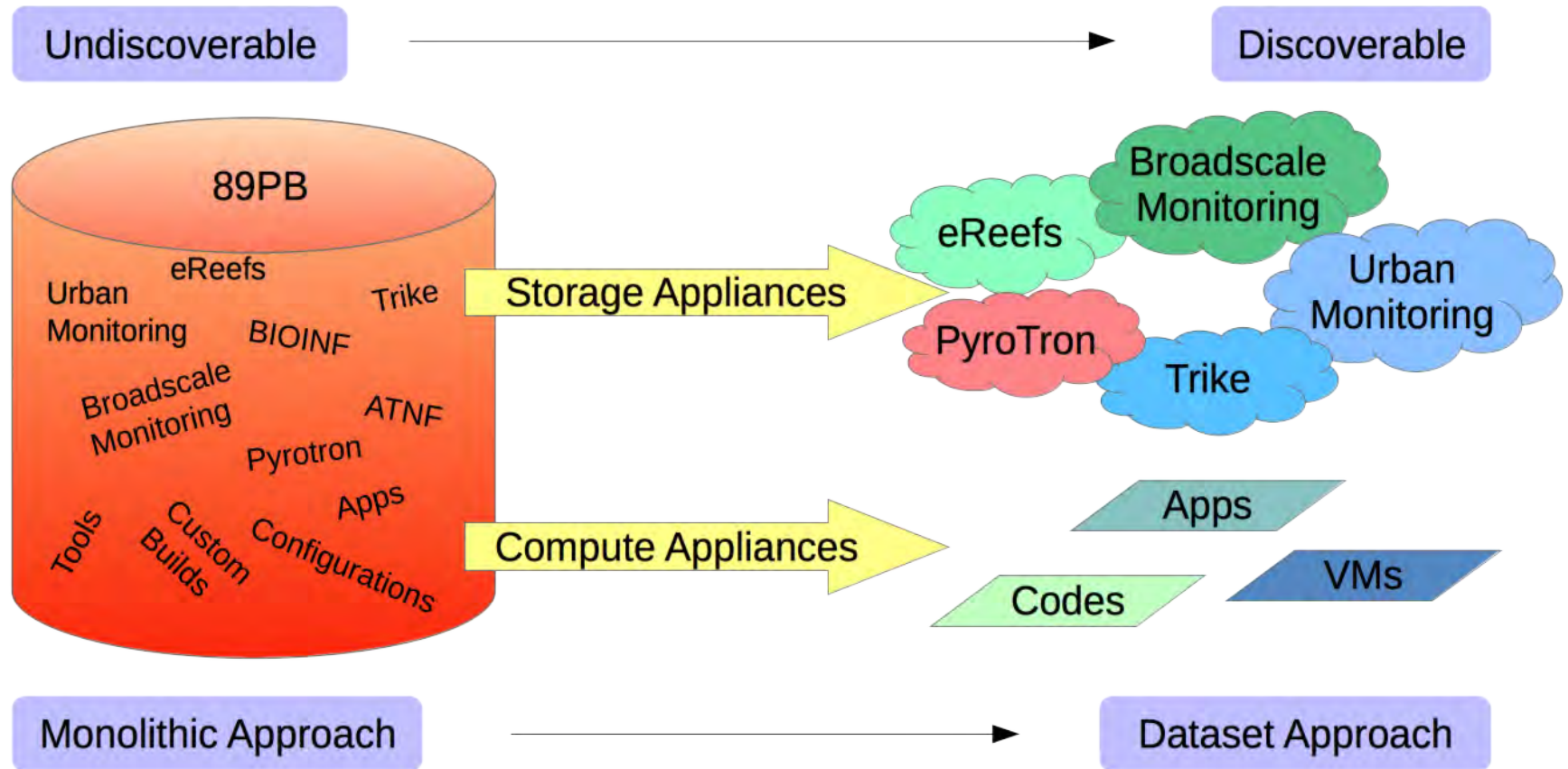
Data-Intensive Workflows

Lets revisit the “monolithic approach”



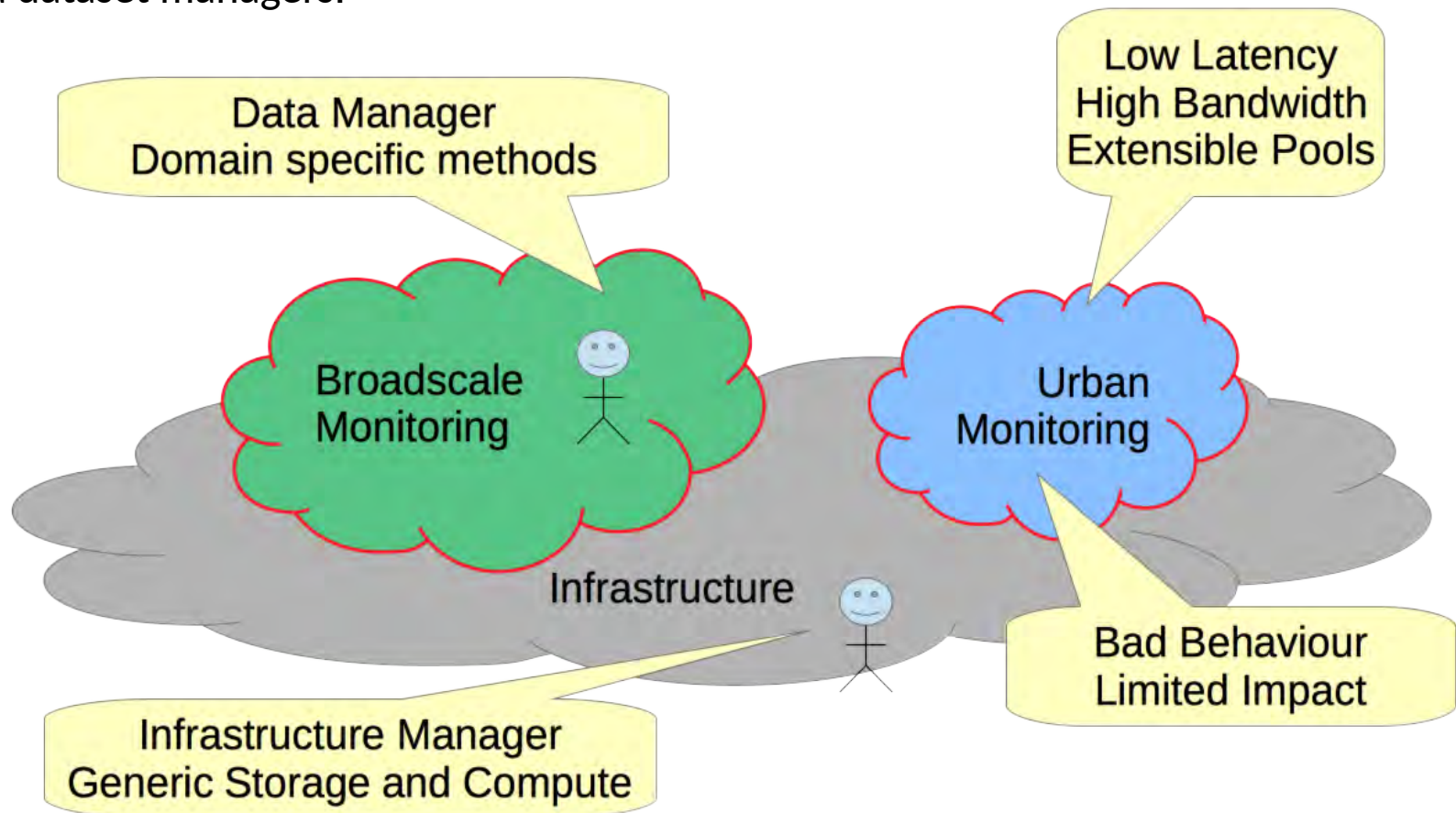
Data-Intensive Workflows

We split the monolithic file systems into named and discoverable 'datasets.'



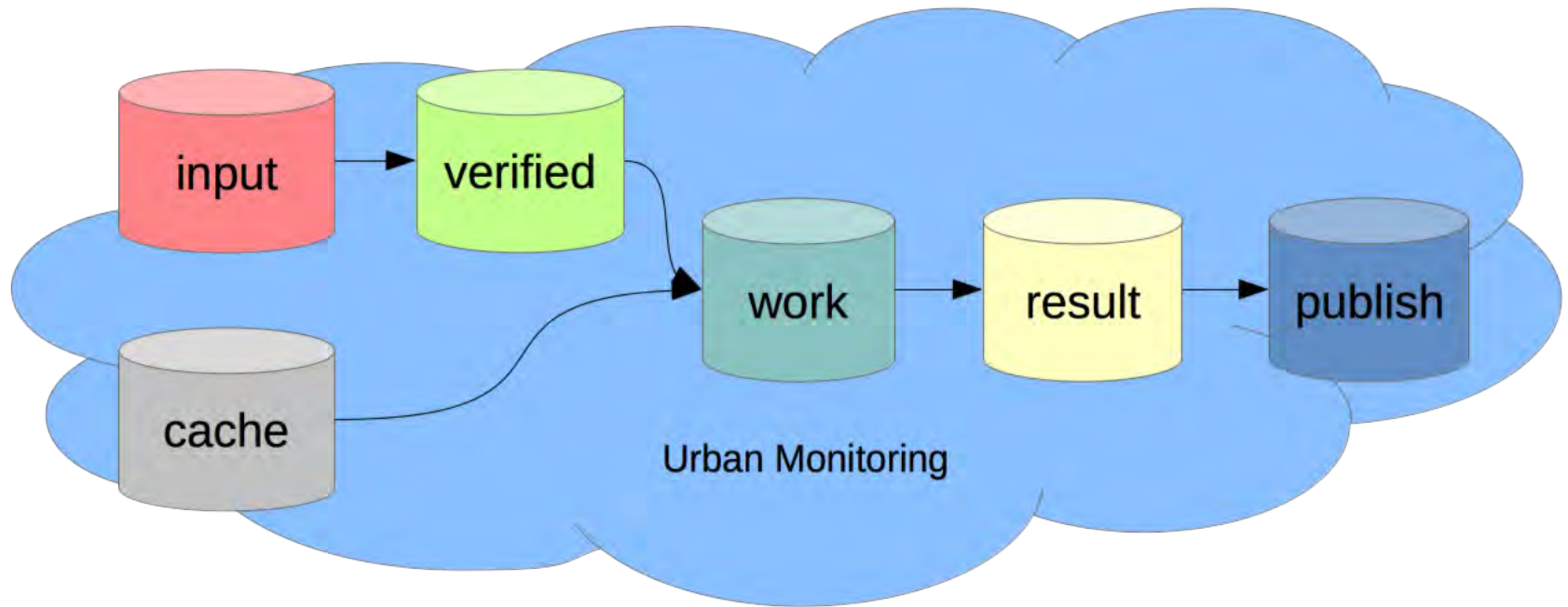
Data-Intensive Workflows

The 'dataset' approach delineated the 'responsibility' between infrastructure owners and dataset managers.



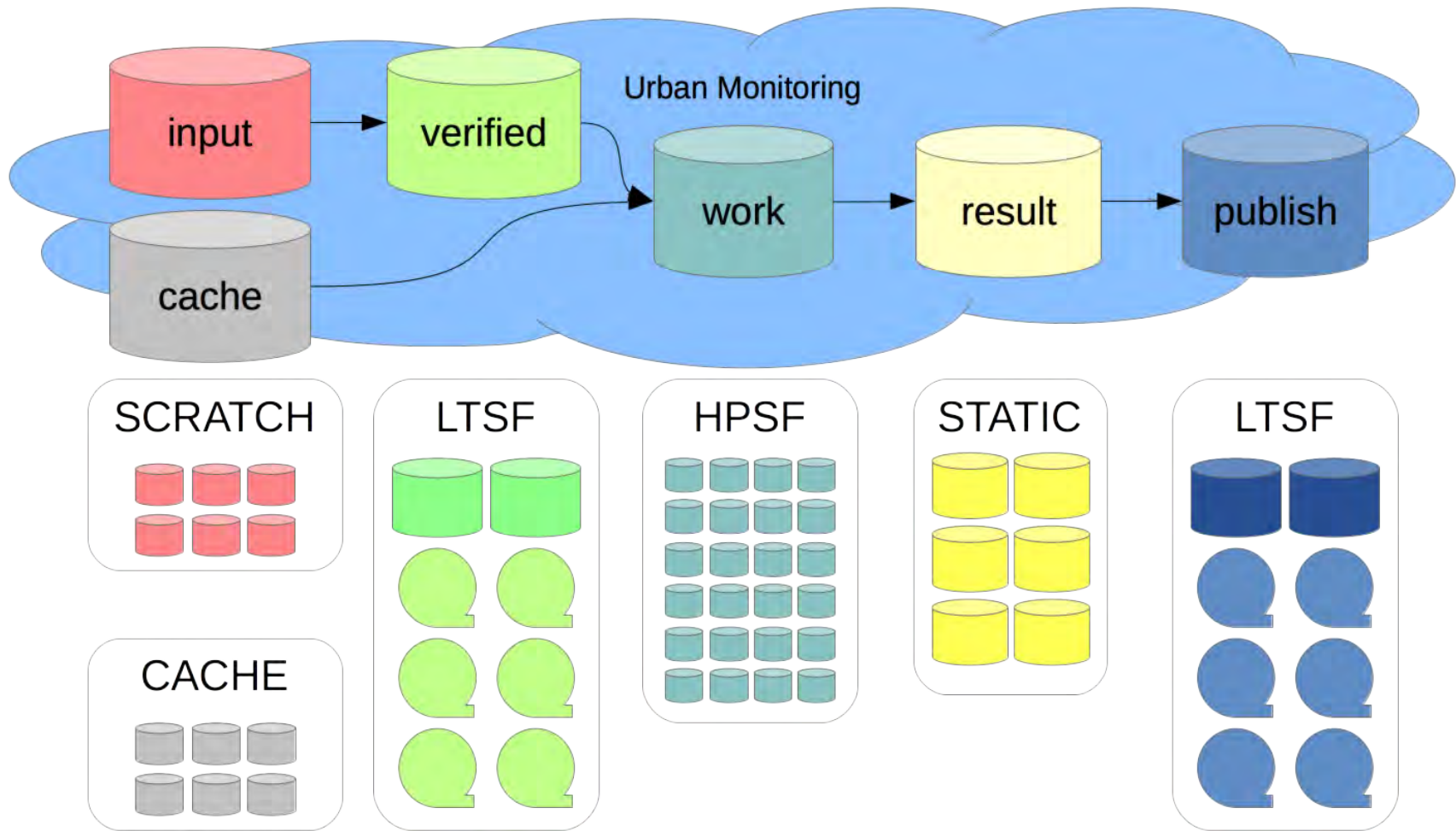
Data-Intensive Workflows

Within the dataset we developed 'categories' as a tool for data management.



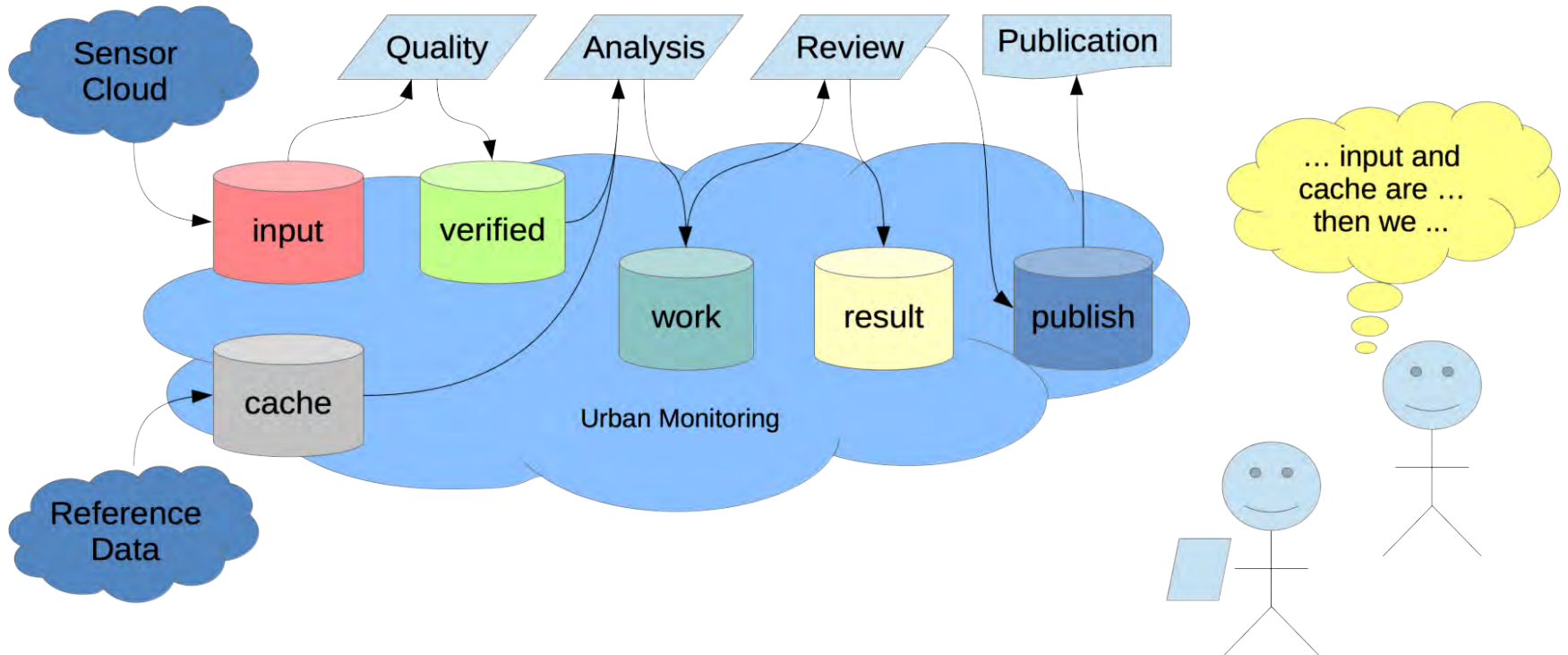
Data-Intensive Workflows

Categories enabled mapping of the workflow to technology of best fit.



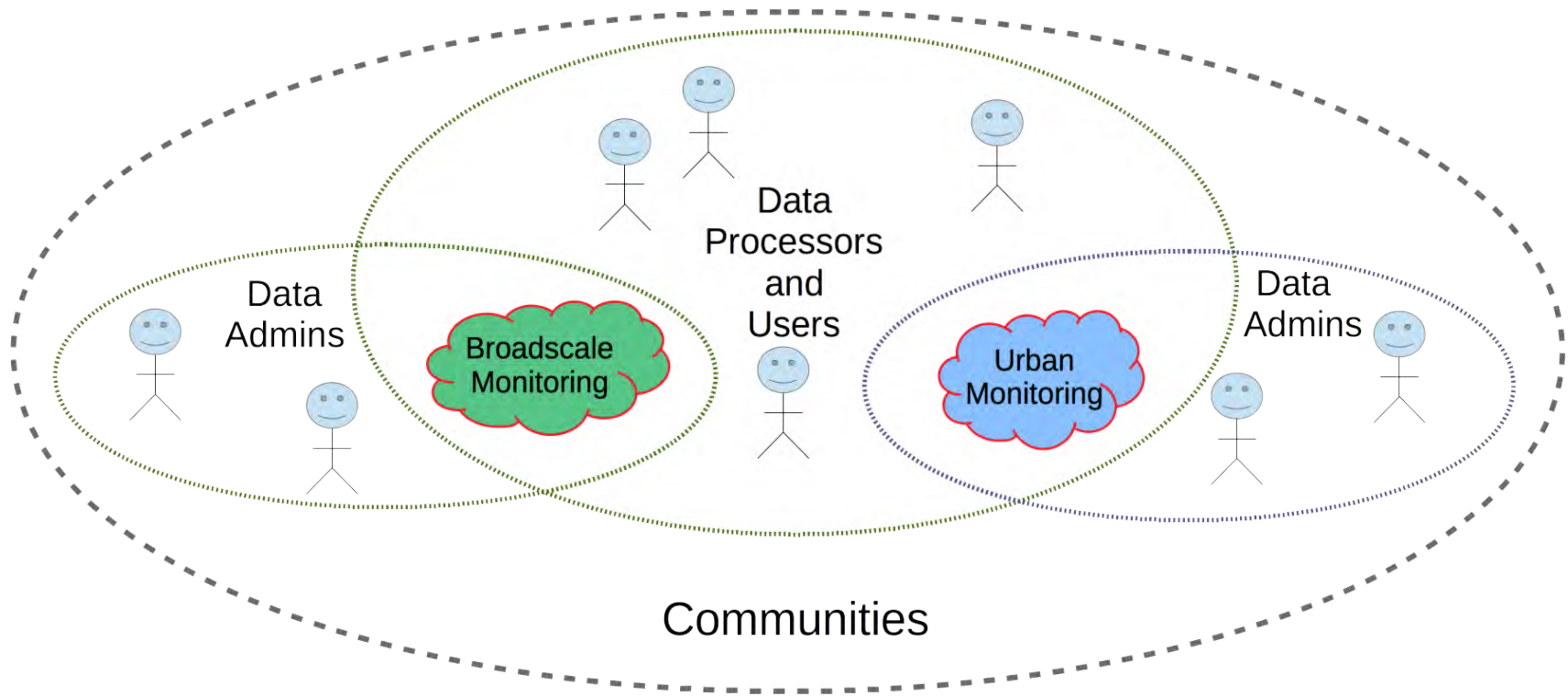
Data-Intensive Workflows

Categories “kick” started the discussion about workflows.



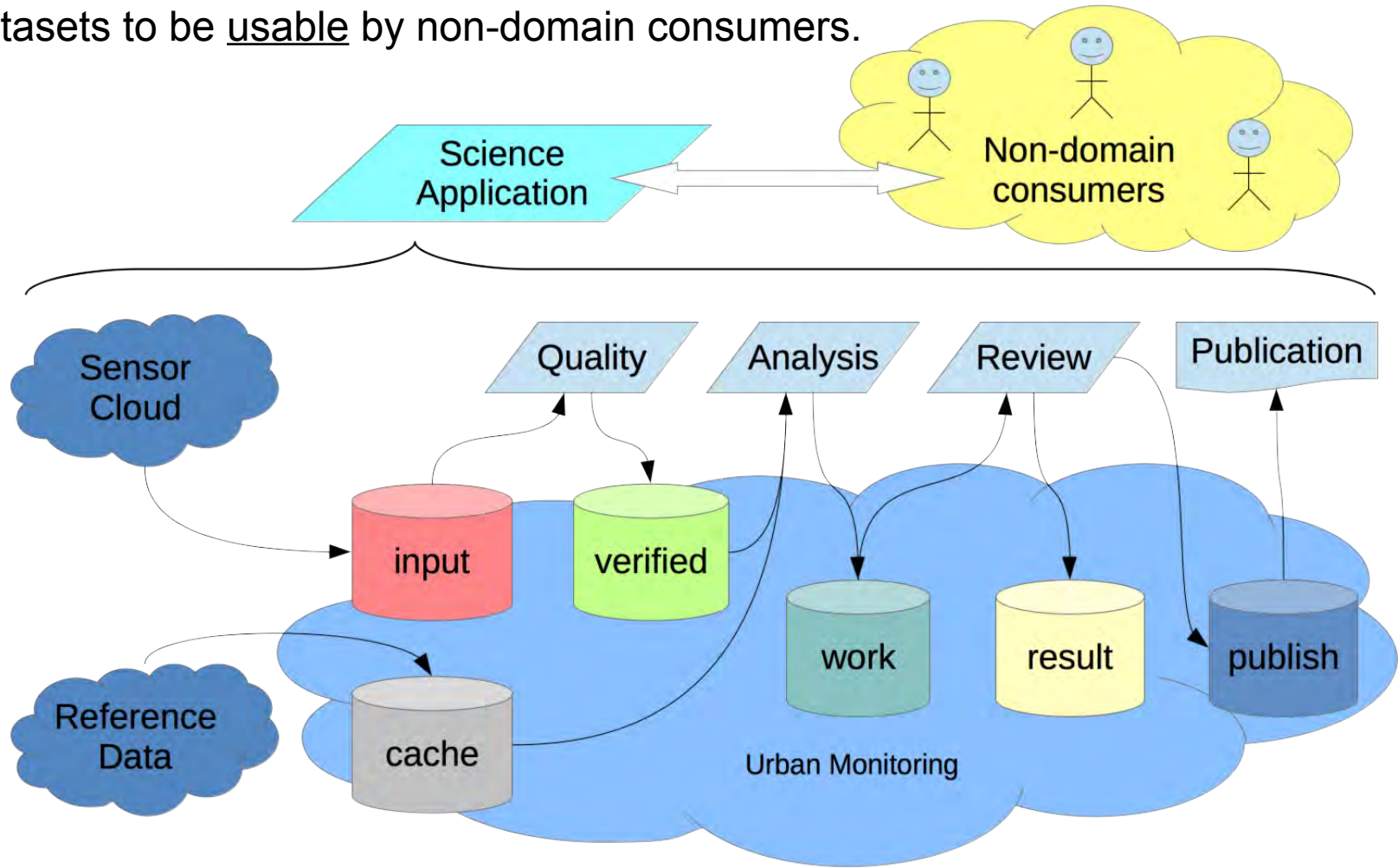
Data-Intensive Workflows

We established the 'relationships' between owners, domain specialists, users, consumers, and infrastructure.



Data-Intensive Workflows

As workflows matured, “science apps” evolved enabling domain specific datasets to be usable by non-domain consumers.



Data-Intensive Workflows – Science Applications

The Pyrotron - CSIRO National Bushfire Research Facility



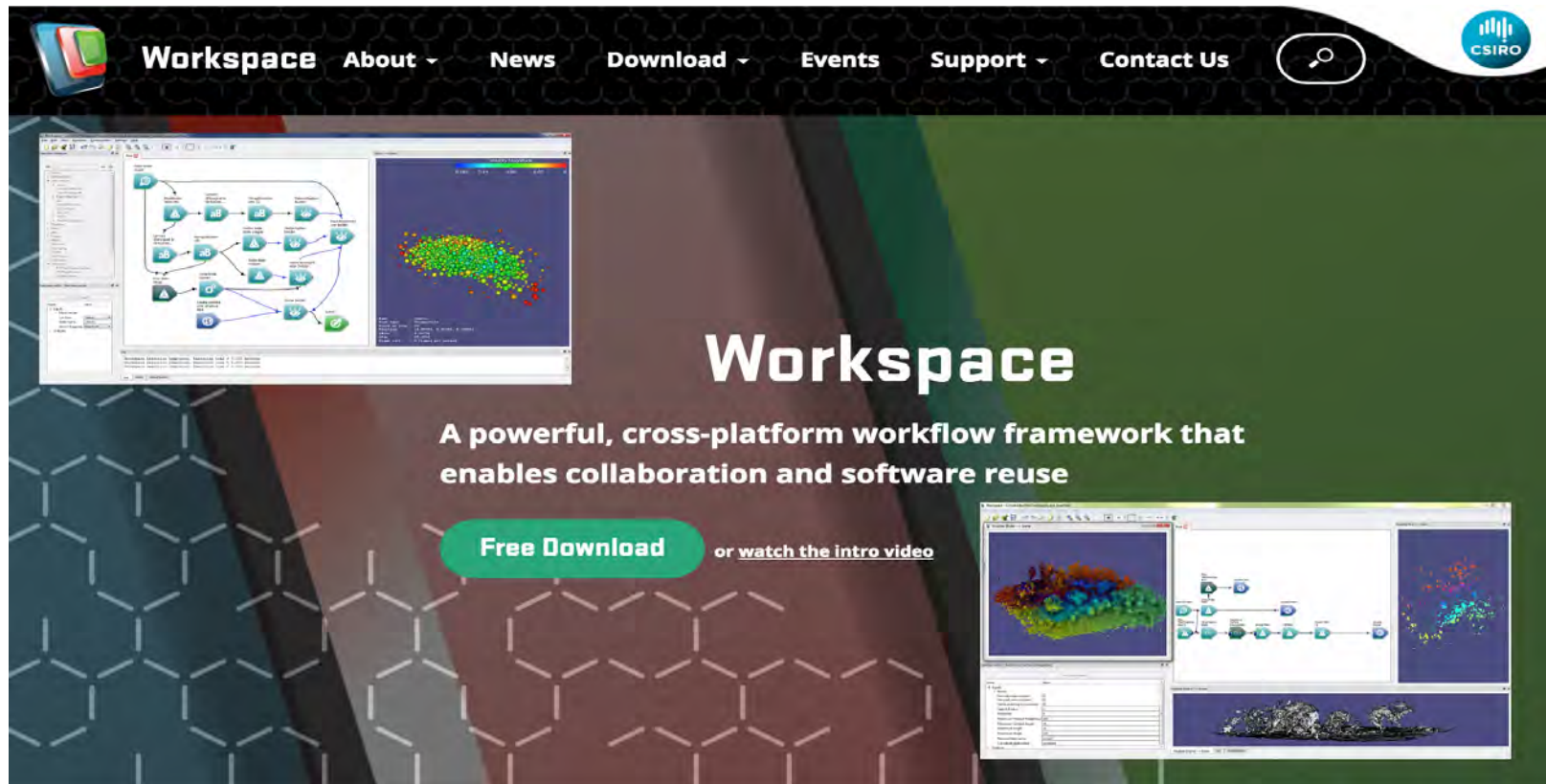
Pyrotron: A national bushfire research facility

The Pyrotron is used to study the combustion and spread of fires in bushfire fuel under controlled conditions. It aims to improve fire safety and fire-fighting for Australian communities.

<http://www.csiro.au/en/Do-business/Services/Testing-and-technical-services/Enviro/Pyrotron>

Data-Intensive Workflows – Science Applications

CSIRO – Workspace - Intuitive Workflow Development Tool



Workspace

About ▾ News Download ▾ Events Support ▾ Contact Us

CSIRO

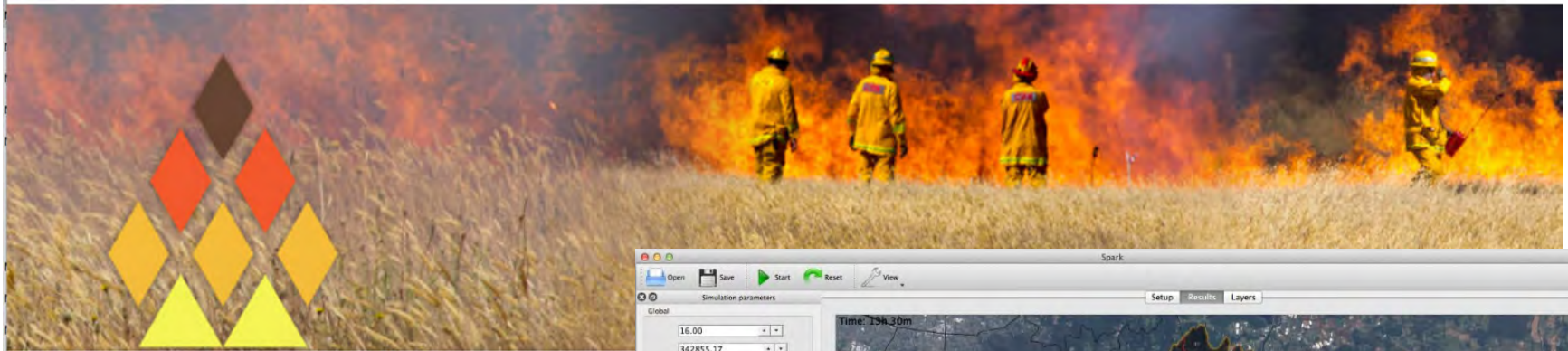
A powerful, cross-platform workflow framework that enables collaboration and software reuse

[Free Download](#) [or watch the intro video](#)

<https://research.csiro.au/workspace/>

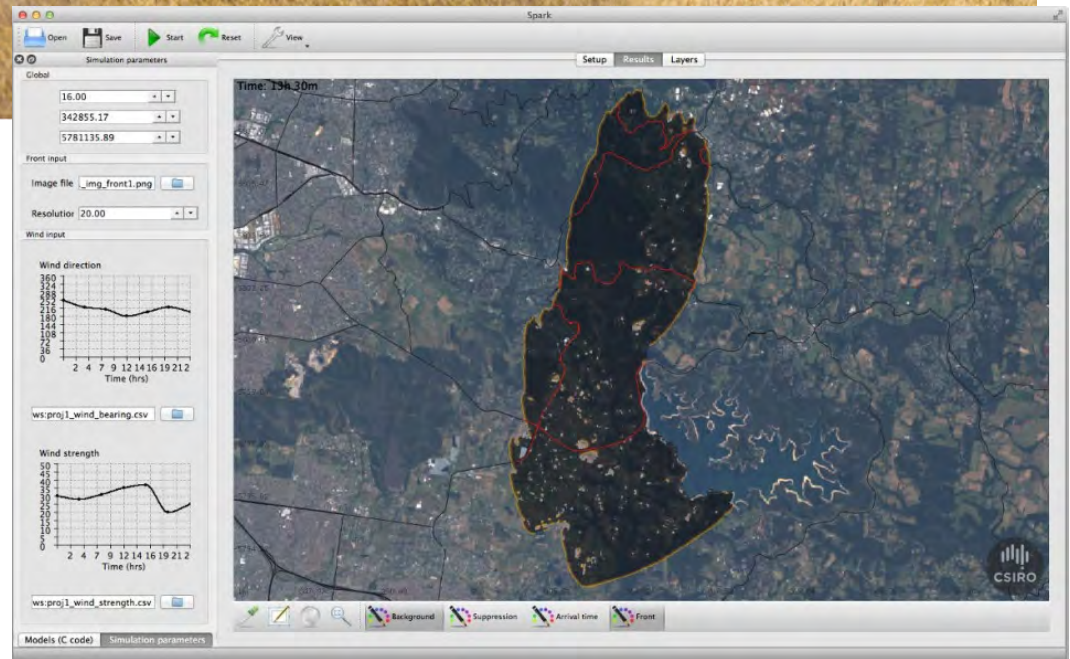
Data-Intensive Workflows – Science Applications

CSIRO – SPARK – A wild fire simulation tool



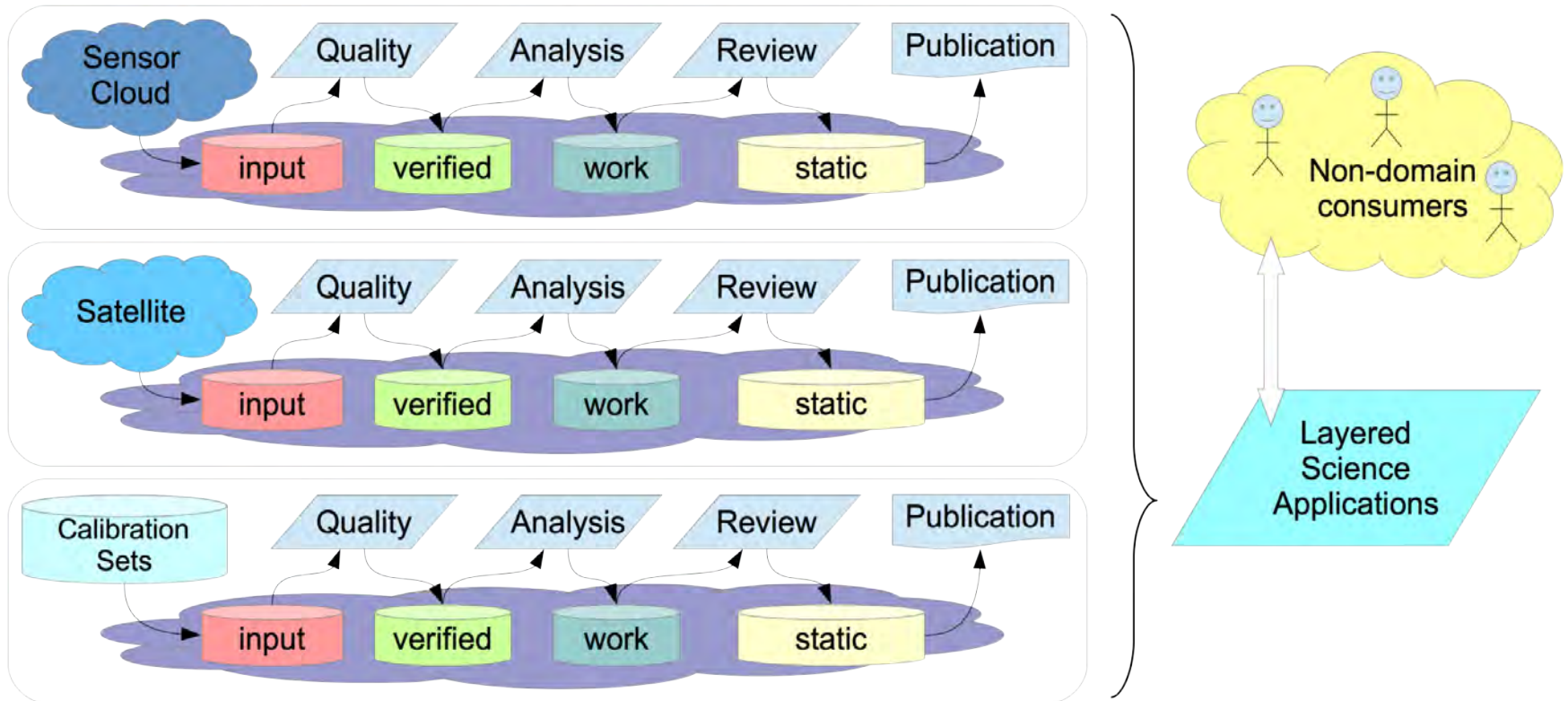
SPARK – A wildfire simulation framework for researchers and experts in the disaster resilience field.

<https://research.csiro.au/spark/>



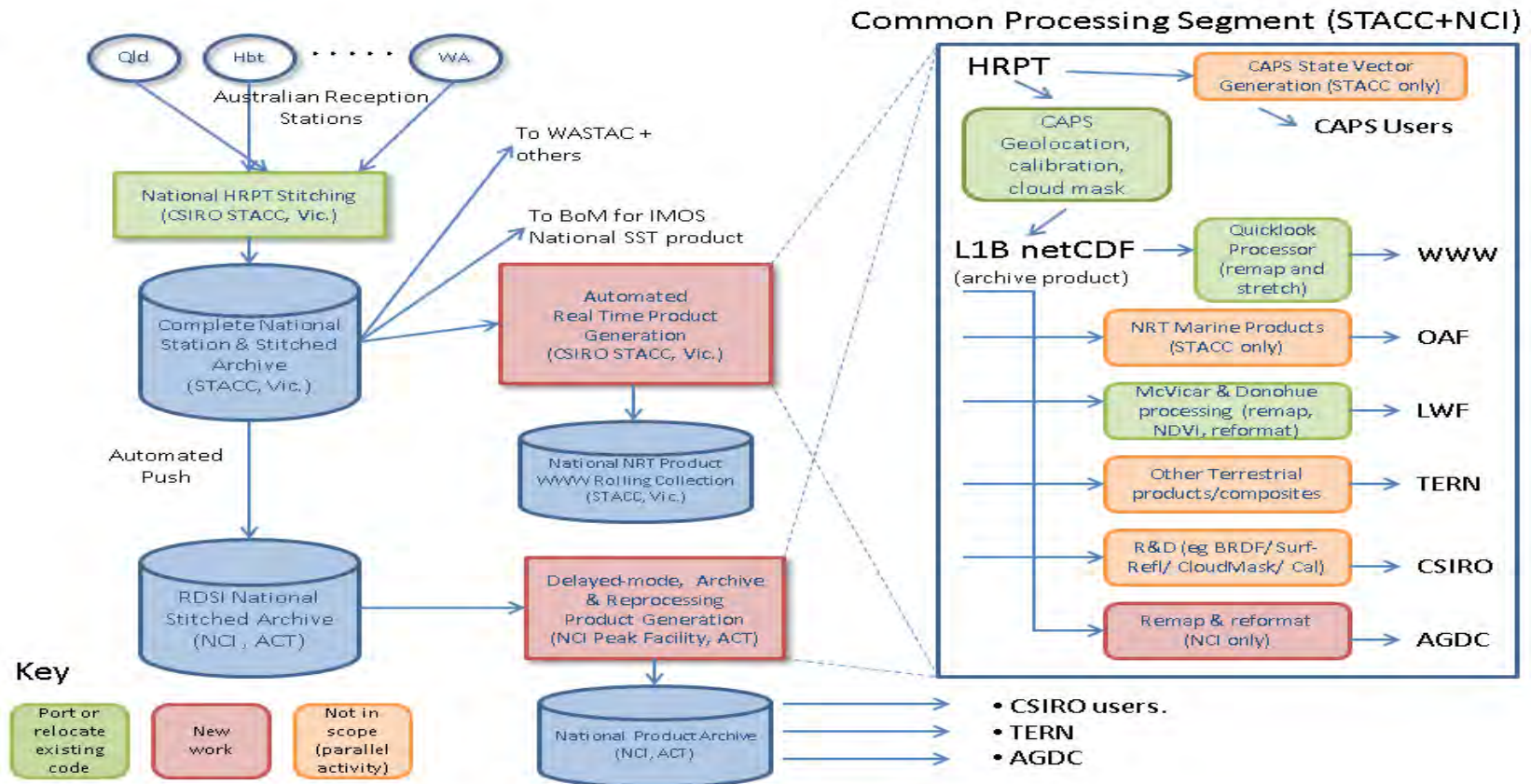
Data-Intensive Workflows

Our leading edge researchers combined domain specific workflows to produce higher value layered products.



Data-Intensive Workflows

Our leading edge researchers combined domain specific workflows to produce higher value layered products.



Data-Intensive Workflows

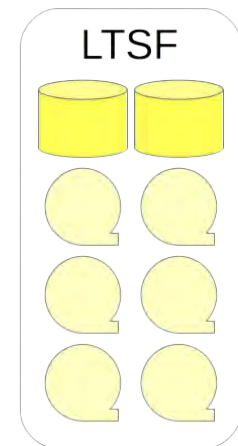
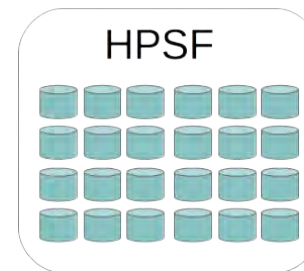
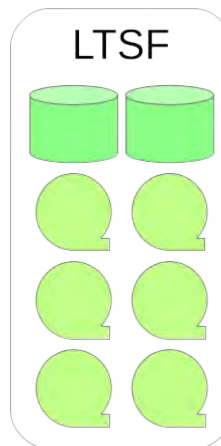
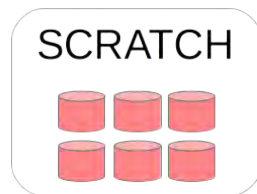
How we matured



Data-Intensive Workflows

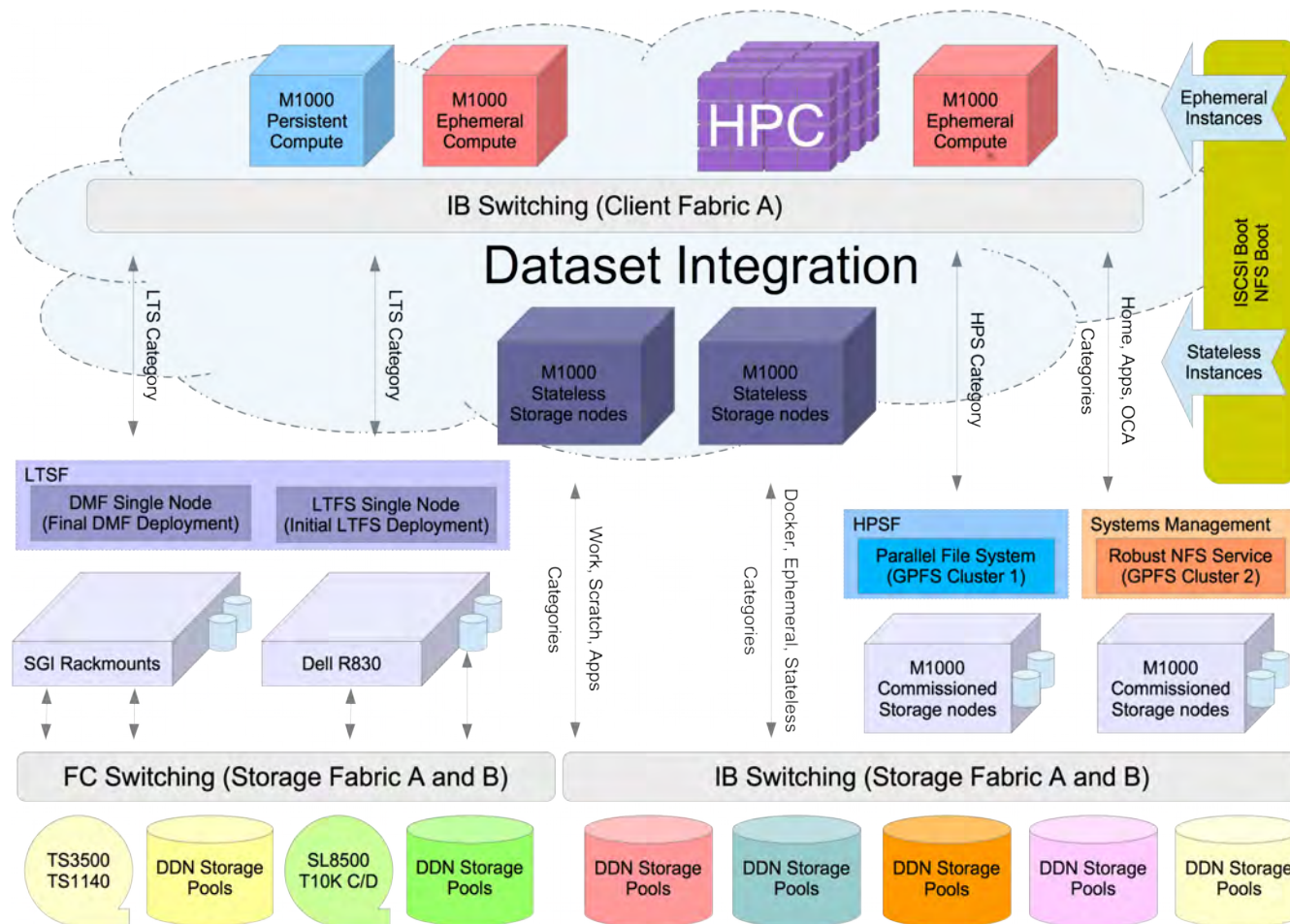
Below the line 'technology' is a consumable, replaceable, discardable commodity.

BELOW THE LINE



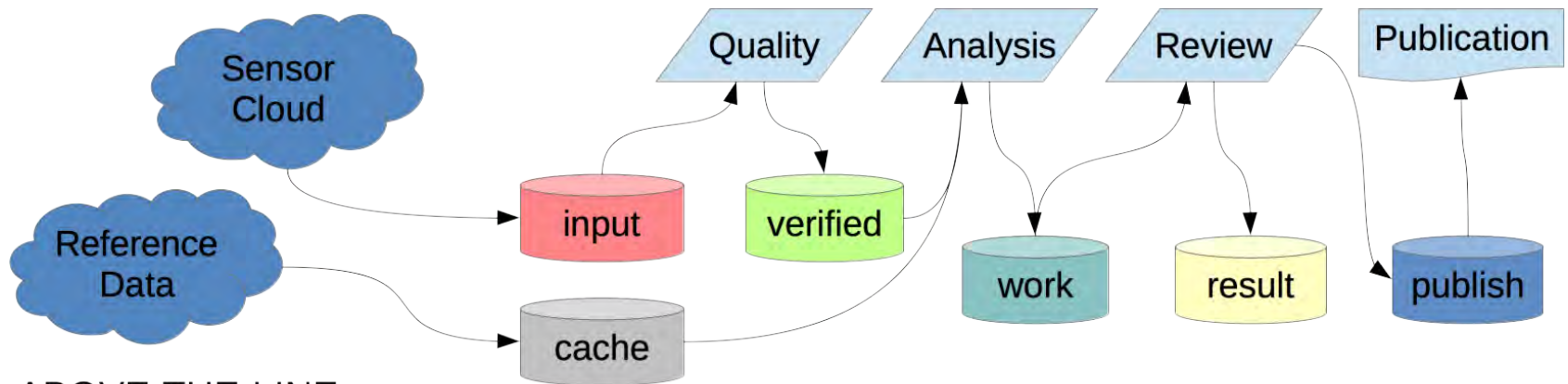
Data-Intensive Workflows

Below the line - the “fit for purpose” pool of generic infrastructure



Data-Intensive Workflows

CSIRO's value proposition is the "Workflow."



ABOVE THE LINE

BELOW THE LINE

Data-Intensive Workflows

Crossing the line we deliver to the 'current' profile of the researchers workflow.

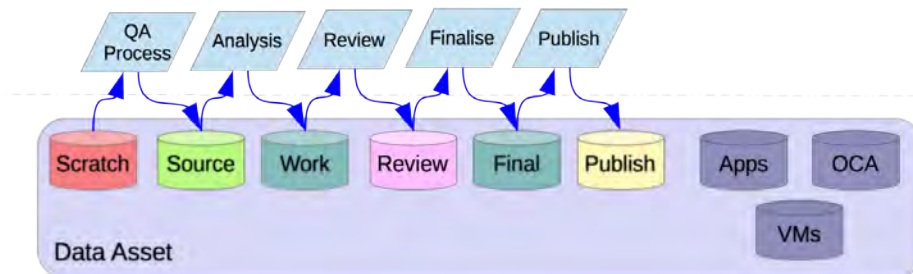
ABOVE THE LINE

BELOW THE LINE

CROSSING THE LINE
We “MAP” to the
Technology of Best Fit

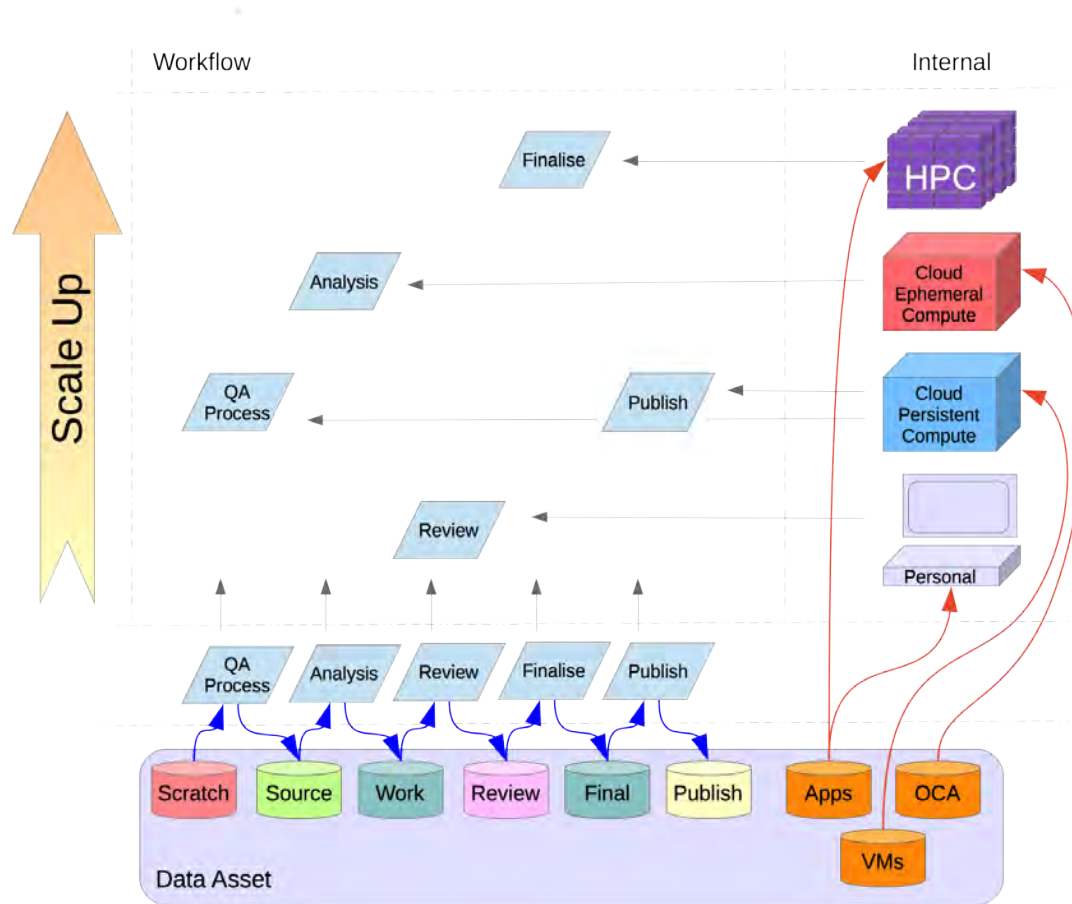
Data-Intensive Workflows

Layers of abstraction enabled us to “scale up.”



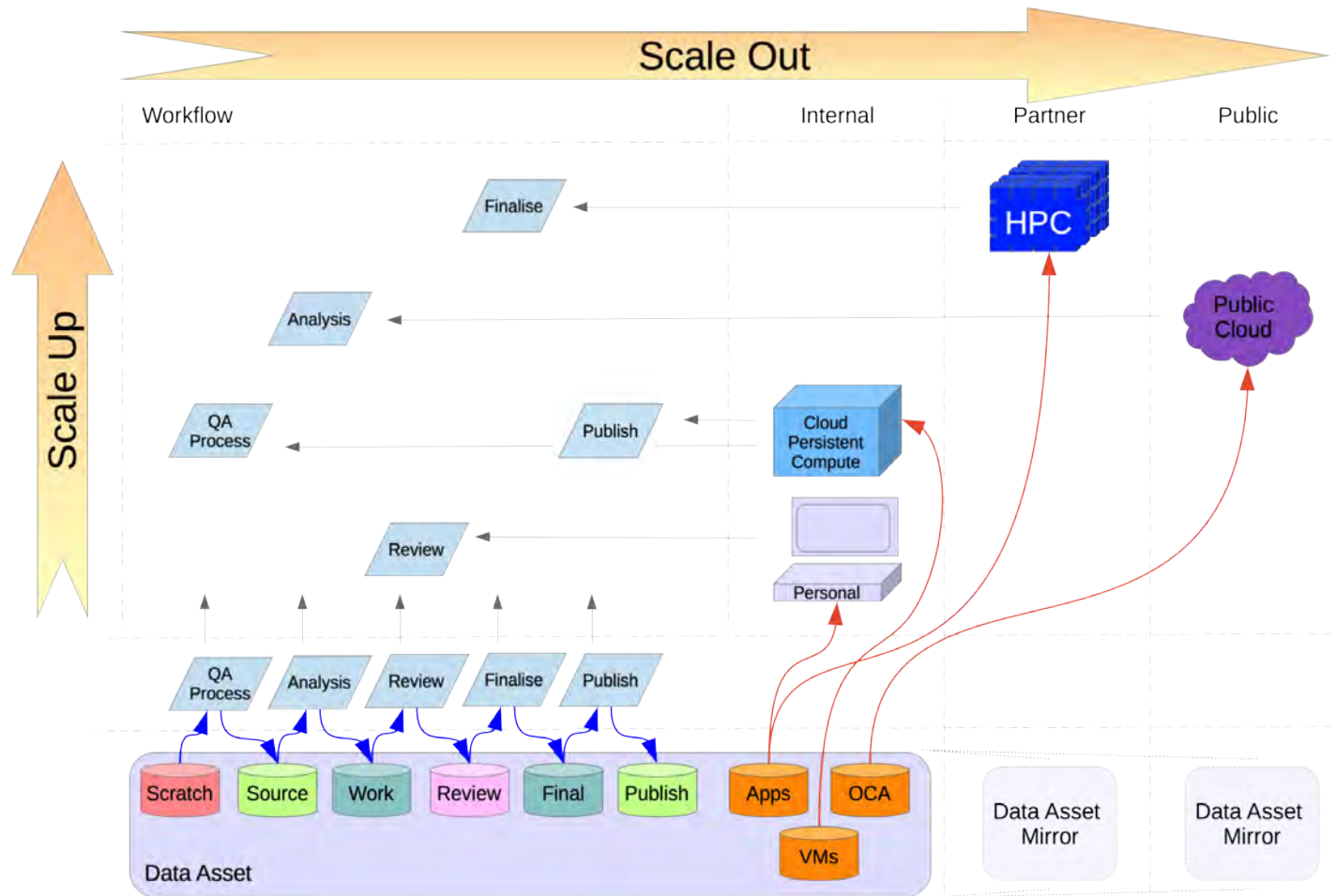
Data-Intensive Workflows

Layers of abstraction enabled us to “scale up.”



Data-Intensive Workflows

Layers of abstraction enabled us to “scale out.”



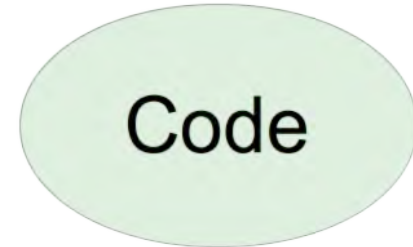
Data-Intensive Workflows

Summary

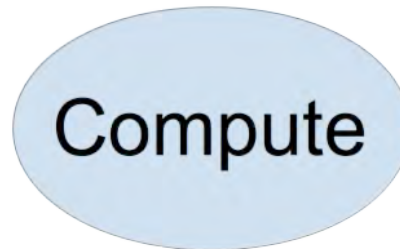


Where we started

We came from a position where data, code and compute were isolated by the approach to HPC infrastructure.

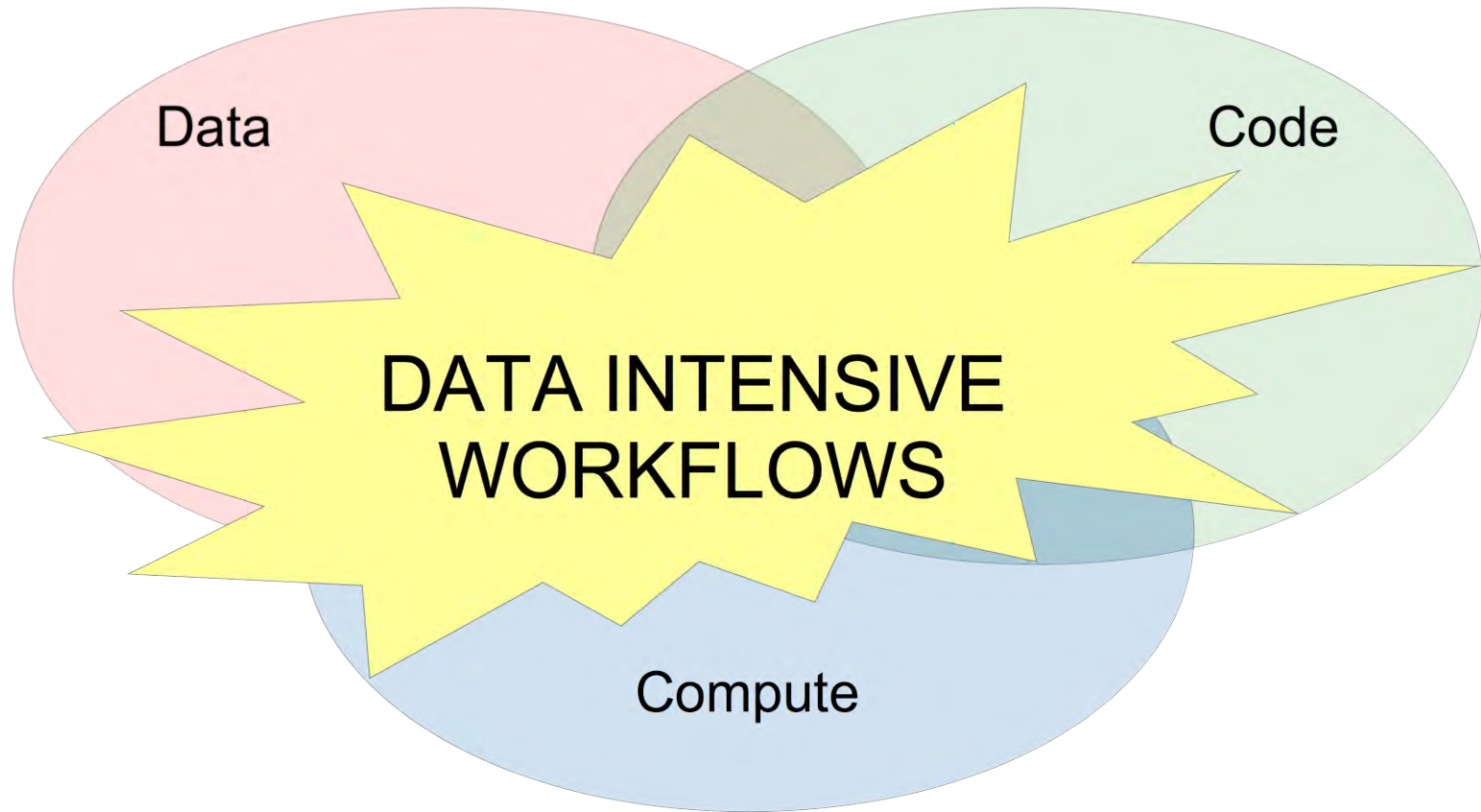


Isolated



What we did – Brought Data to Life

We engineered a solution where data, code and compute are all now directly connected.



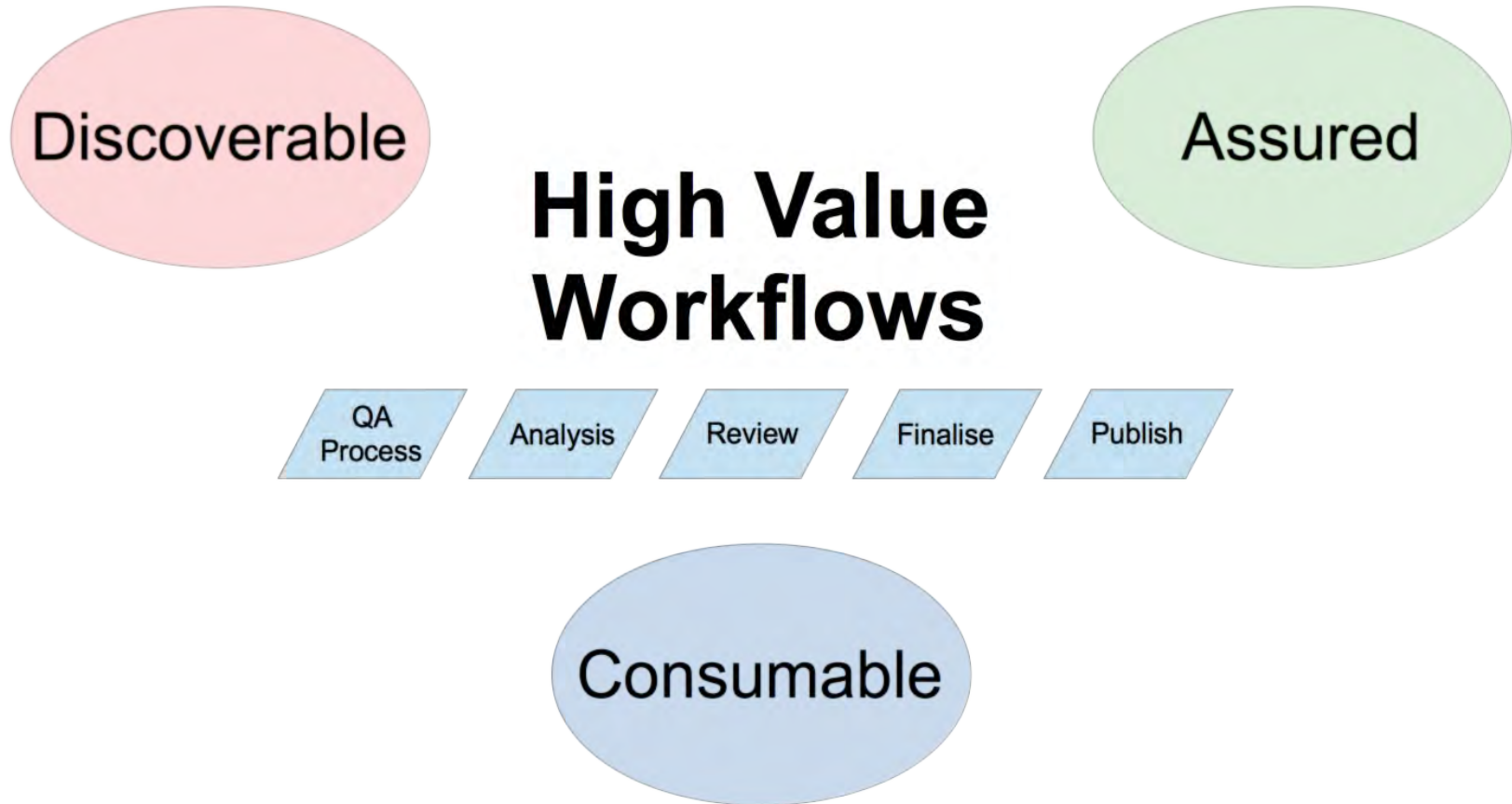
High Value Information:

**Discoverable,
Assured, and
Consumable.**



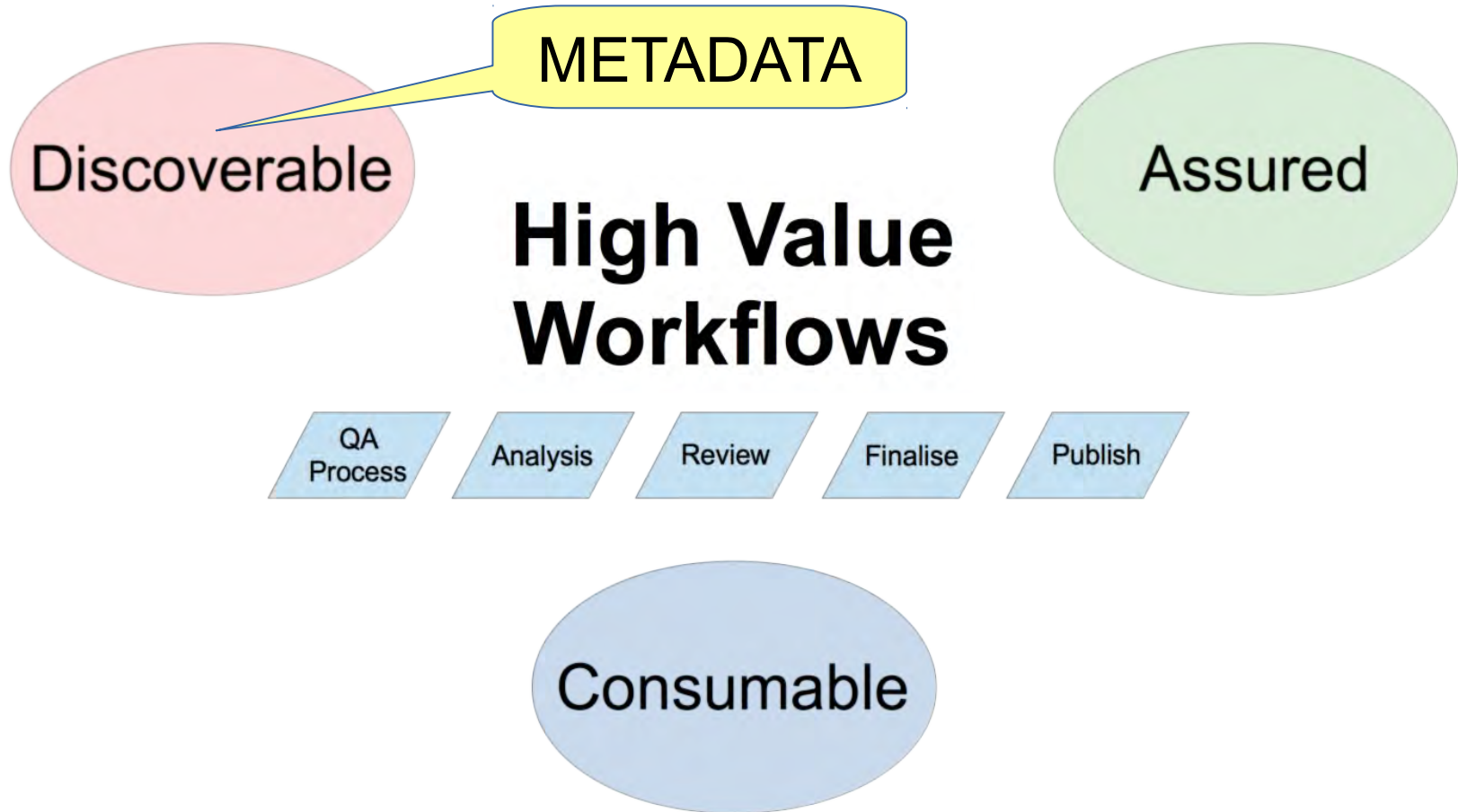
Data-Intensive Workflows

CSIRO's data-intensive workflows are a valuable source of information.
How do we discover them, trust them and consume them?



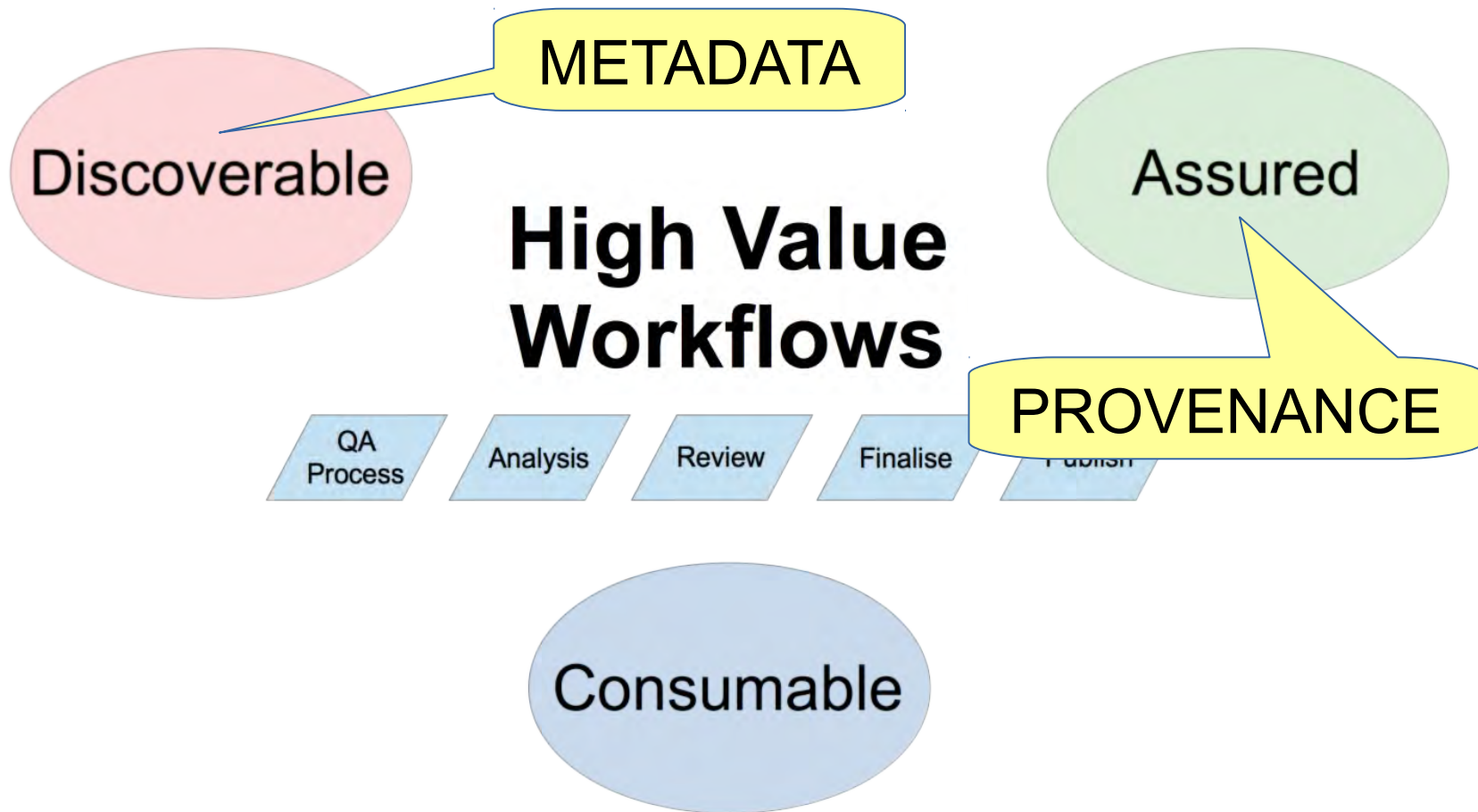
Data-Intensive Workflows

CSIRO's data-intensive workflows are a valuable source of information.
How do we **discover** them, trust them and consume them?



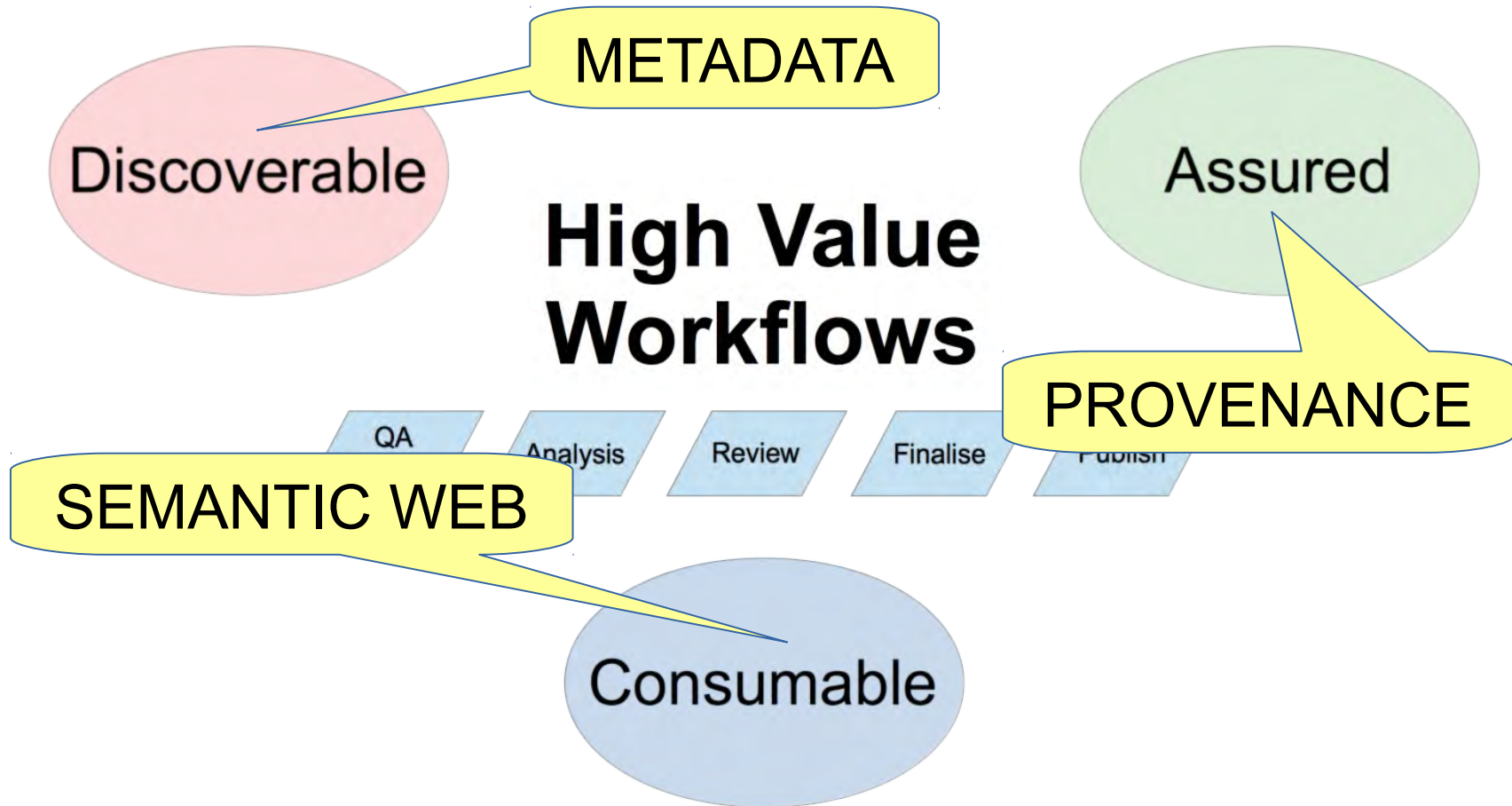
Data-Intensive Workflows

CSIRO's data-intensive workflows are a valuable source of information.
How do we discover them, **trust** them and consume them?



Data-Intensive Workflows

CSIRO's data-intensive workflows are a valuable source of information.
How do we discover them, trust them and **consume** them?



Discoverable – Metadata

Metadata is a pathway to making data and workflows discoverable.

Lets look at Wikipedia:

<https://en.wikipedia.org/wiki/Metadata>

Metadata is "data that provides information about other data". Two types of metadata exist: structural metadata and descriptive metadata. Structural metadata is data about the containers of data. Descriptive metadata uses individual instances of application data or the data content.

Consumable – Semantic Web

Pragmatic use of the web.

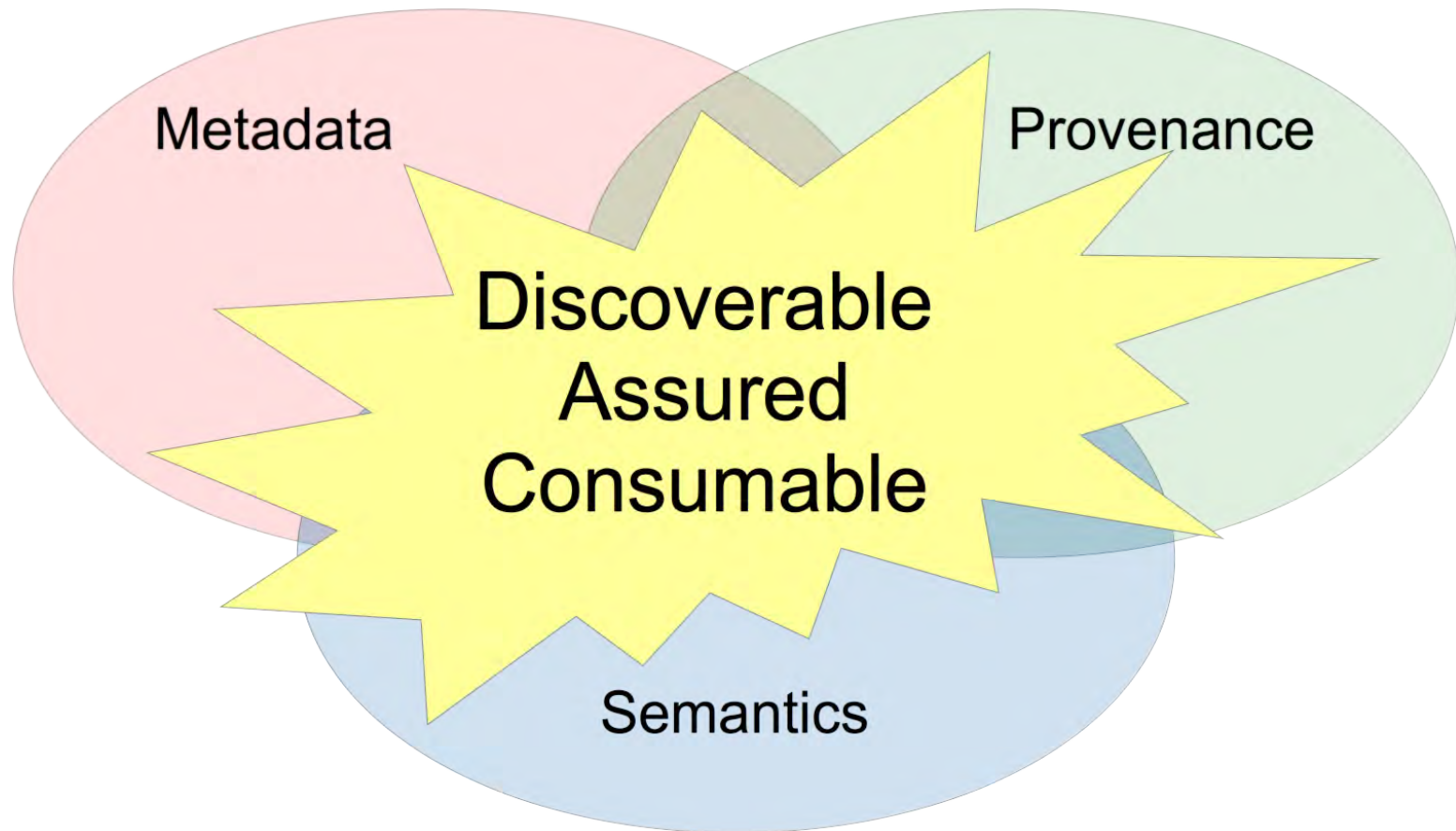
Lets look at Wikipedia:

https://en.wikipedia.org/wiki/Semantic_Web

The Semantic Web is an extension of the Web through standards by the World Wide Web Consortium (W3C). The standards promote common data formats and exchange protocols on the Web, most fundamentally the Resource Description Framework (RDF).

Linking Metadata, Provenance and Semantics

We need to link *metadata*, *provenance* and *semantics* in an automatic and extensible manner to increase our value.



Context Capture - The Future

Preserving Metadata

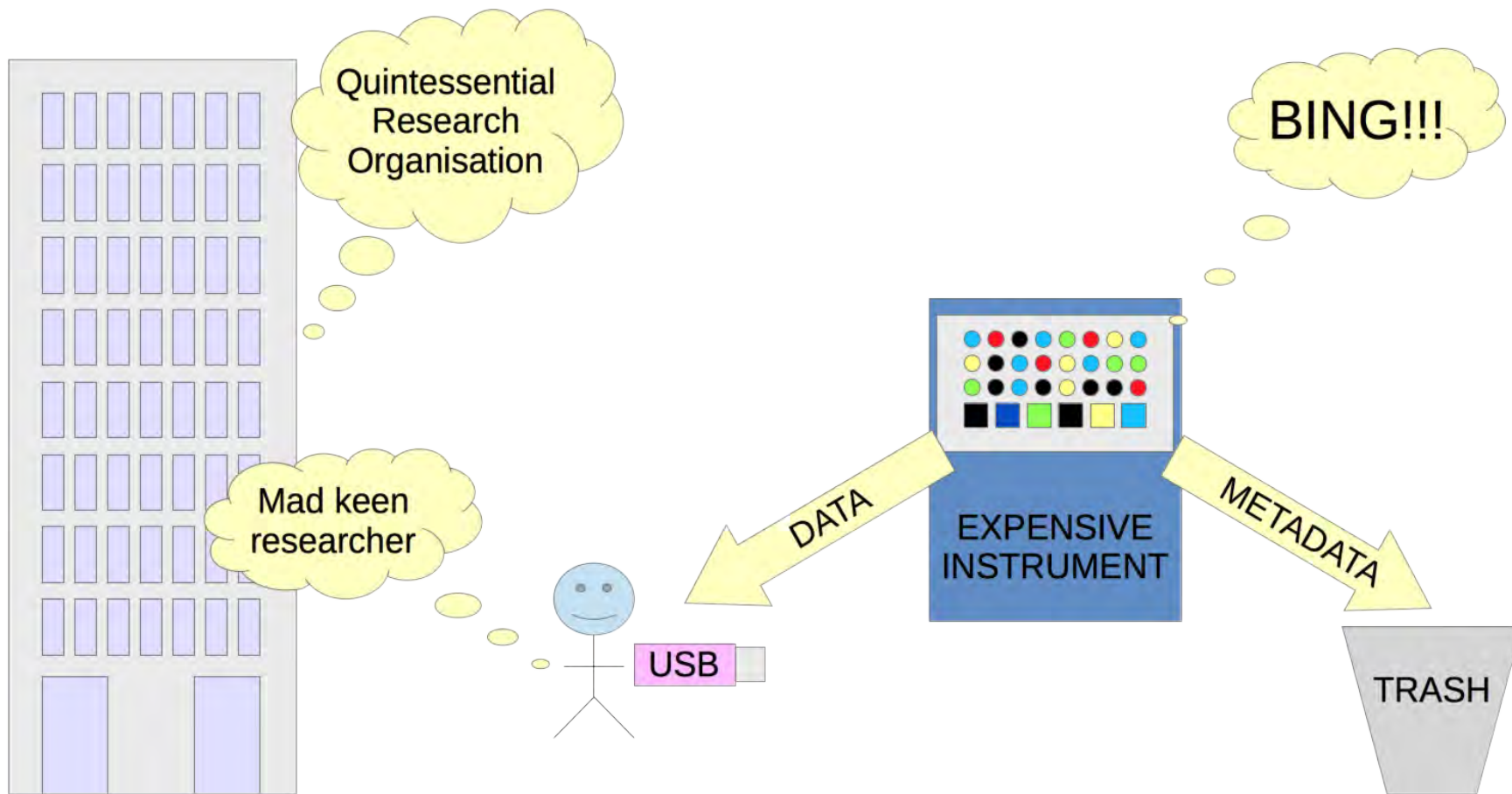
Establishing Provenance

Presenting via Semantic Web



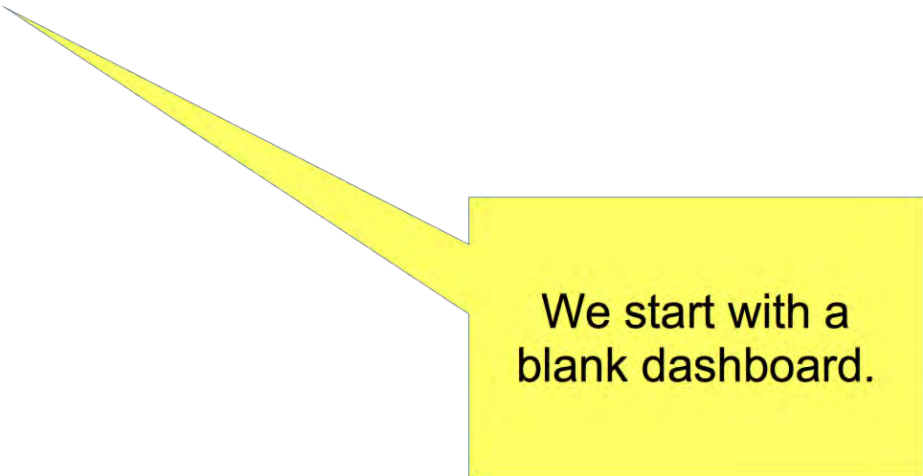
Stripping Context - The Past

What are we currently losing



Context Capture – The Future

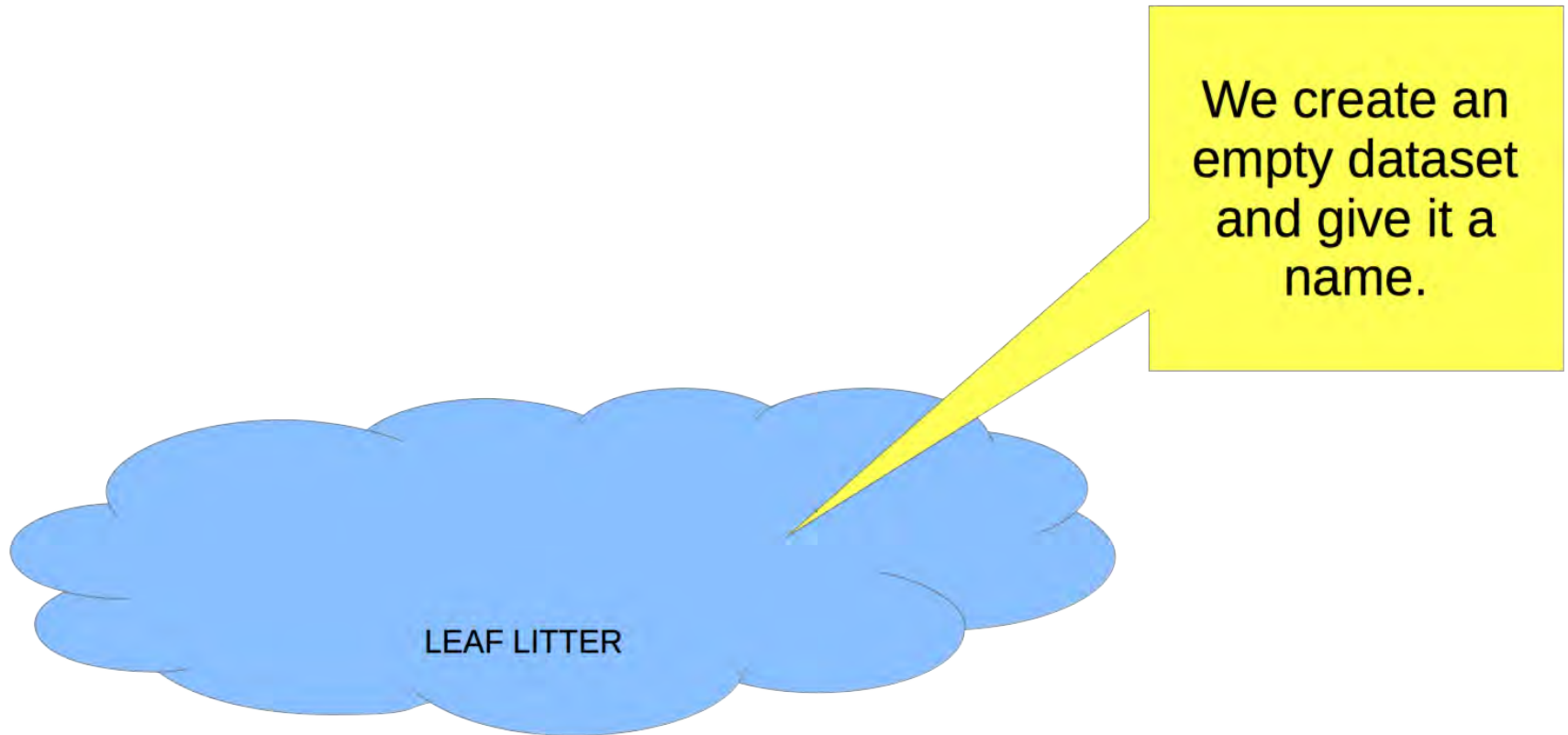
Preserving the context of discrete events



We start with a blank dashboard.

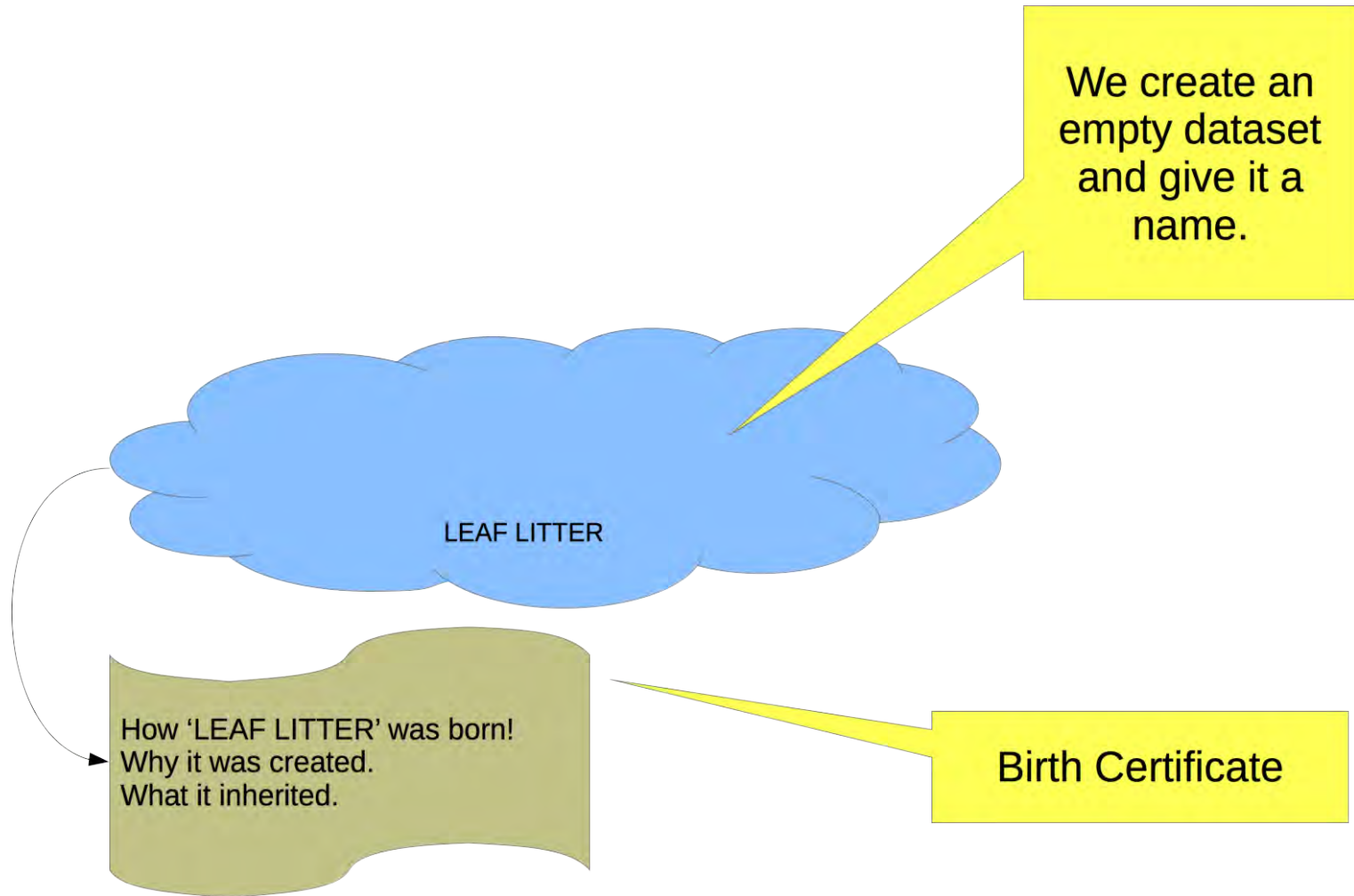
Context Capture - The Future

Create a blank dataset



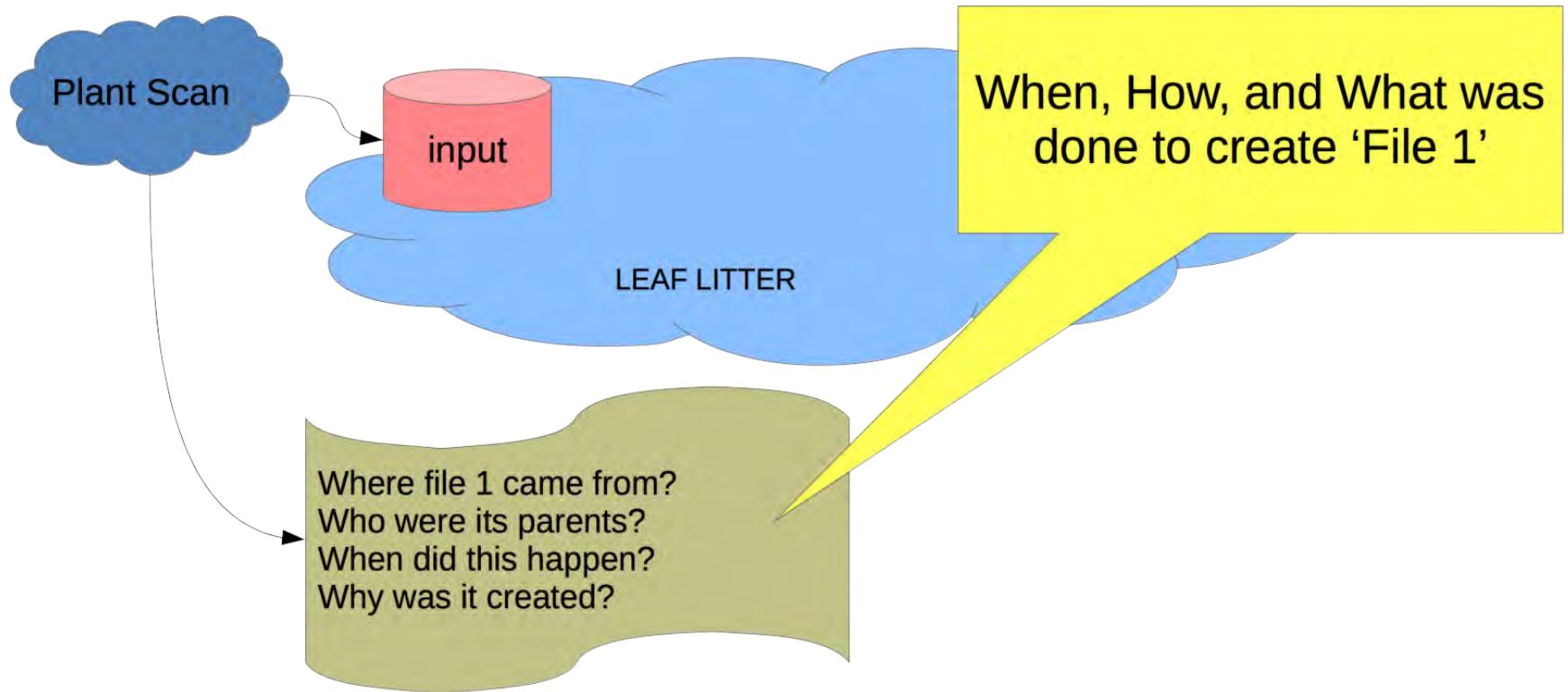
Context Capture - The Future

Create a blank dataset - Preserving metadata, establishing provenance, ...



Context Capture - The Future

Ingesting sensor data - Preserving metadata, establishing provenance, ...



Context Capture – The Future

Lets consider a real world application - PlantScan

PlantScan provides non-invasive analyses of plant structure (topology, surface orientation, number of leaves), morphology (leaf size, shape, colour, area, volume) and function by utilising cutting edge information technology including high resolution cameras and three-dimensional (3D) reconstruction software.

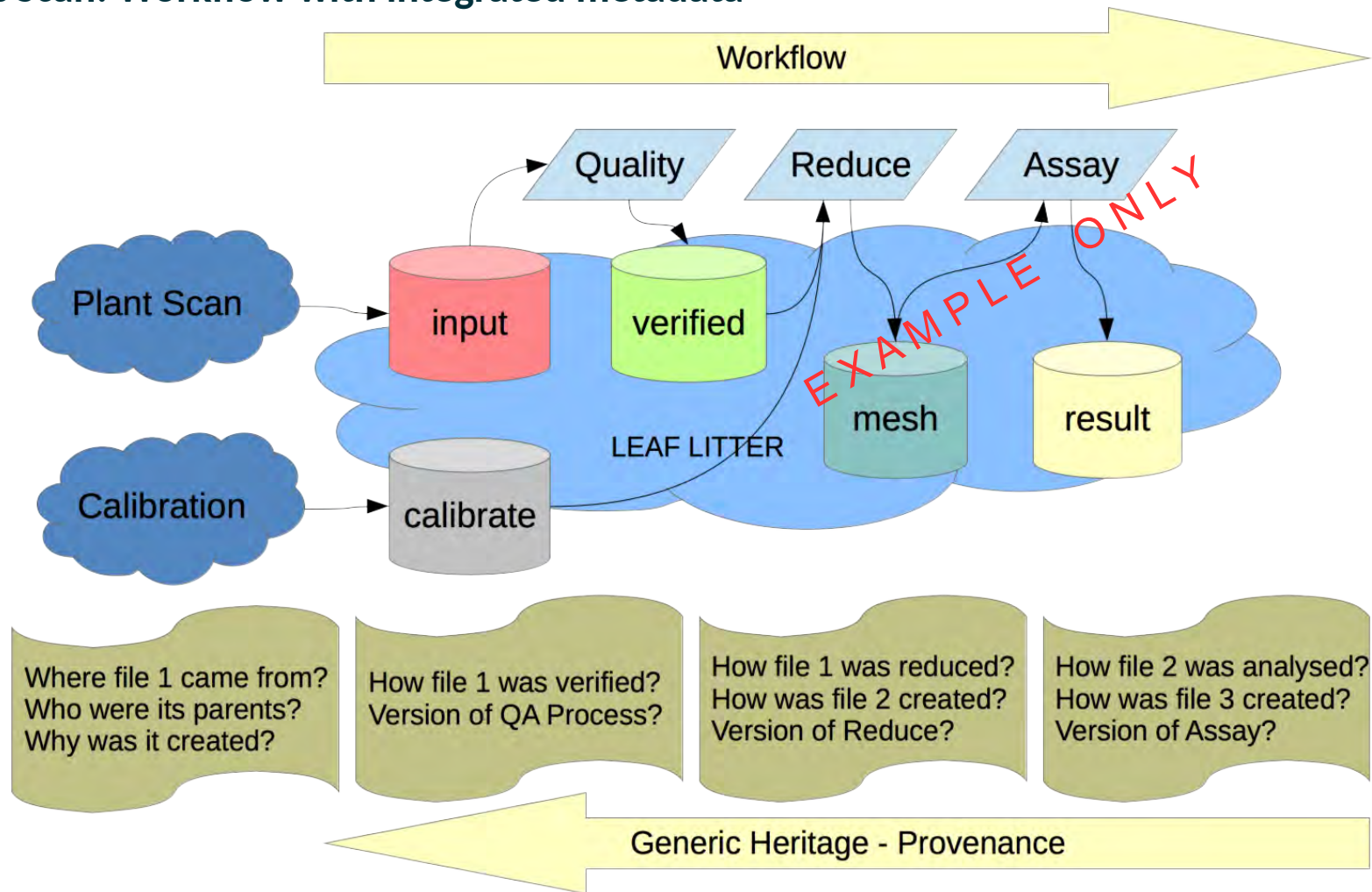


- ✓ Plant surface mesh reconstruction
- ✓ Morphological mesh segmentation
- ✓ Accurate phenotypic data extraction
- ✓ Longitudinal matching

<http://www.plantphenomics.org.au/services/plantscan/>

Context Capture - The Future

Plant Scan: Workflow with integrated metadata



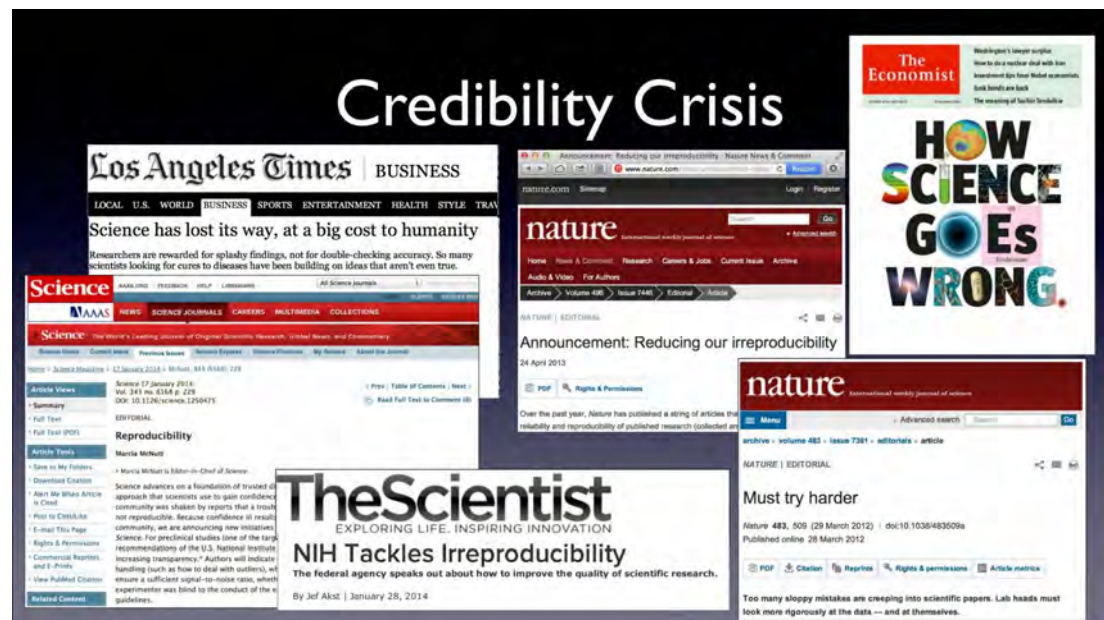
Context Capture - The Future

Benefit 1 - Repeatable Analysis

Now there is no 'simple' fix to this issue.

But if we issue the 'birth certificate' before the baby leaves the hospital then we have at least improved our position.

We reduce the size of the problem.



Context Capture – The Future

Benefit 2 – Quality Assurance

File_1 became File_2 using version 56 of Fred_1.

Now if File_1 had a calibration issue.

Or Fred_1 had an analysis bug.

Guess what? We reduce our problem more.



Context Capture – The Future

Other benefits

Benefit 3: Immediate Consumption

Context Capture – The Future

Other benefits

Benefit 3: Immediate Consumption

Benefit 4: Benchmarks

Context Capture – The Future

Other benefits

Benefit 3: Immediate Consumption

Benefit 4: Benchmarks

Benefit 5: Infrastructure Management

Context Capture – The Future

Other benefits

Benefit 3: Immediate Consumption

Benefit 4: Benchmarks

Benefit 5: Infrastructure Management

Benefit 6: Failures

Context Capture – The Future

Summary

An example from the past

In Unix everything is a file (pretty much) there are a set of well written simple tools which you can tie together in an ad-hoc way to produce high value outcomes in a dynamic yet robust manner.

Moving to the future

Everything is a dataset, there are a set of published, well proven and tested set of ‘research’ workflows which you can tie together in an ad-hoc way to produce high value outcomes in a dynamic yet robust manner.

Summary



Context Capture – The Future

Summary

We made it to a place where data, code and compute are now tightly coupled – Researchers focus on the workflow.

As workflows proliferate we want to make sure they exist in an ecosystem where they can be discovered, assessed and consumed.

Thank You

INFORMATION MANAGEMENT AND TECHNOLOGY (IMT)

