Raw Data Reuse: What Does this Mean for CCP4

Eugene Krissinel

CCP4, Research Complex at Harwell, RAL, UK

eugene.krissinel@stfc.ac.uk

iUCR-26 Congress, Melbourne, Australia, 22 August 2023

Data and Software

Data effect on Software ***development***

- data reuse for software improvement
- data reuse for software testing
- data reuse for AI methods

Data and Software ***maintenance***

- data and software are in symbiosis
- data and software do age
- data and software have a cost to maintain

Data and Software ***legacy***

- data and software must be made available in publicly funded research
- data and software must be available for revisiting and revising results
- unlike software, data make scientific evidence

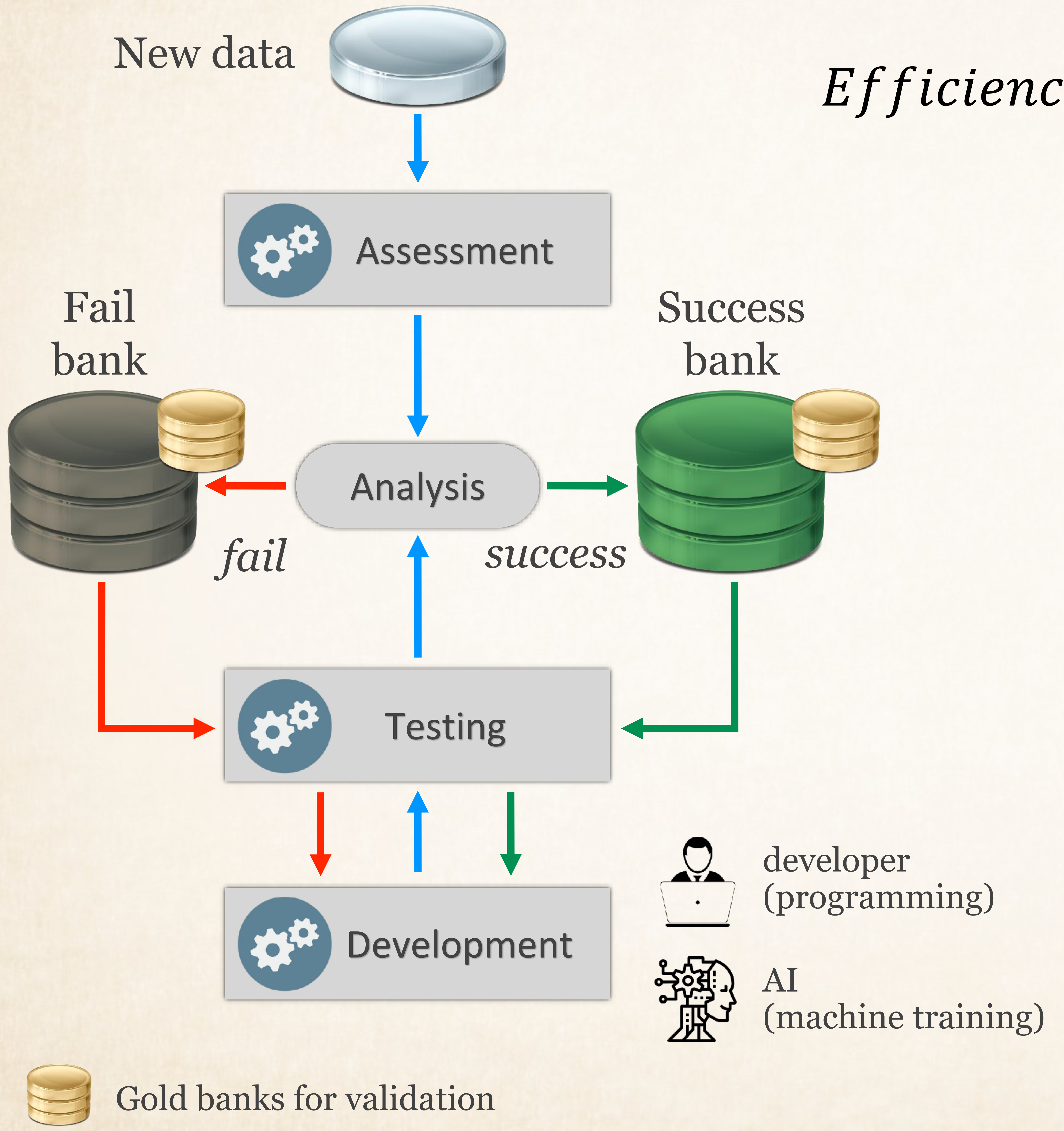
CCP4

Data and Software *Development*

Research Complex at Harwell

UKRI Science and Technology Facilities Council

Scientific Computing



New data

Assessment

Fail bank

Success bank

Analysis

fail

success

Testing

Development

developer (programming)

AI (machine training)

Gold banks for validation

bank mass

$$Efficiency(⚙) = \frac{M_{success}}{M_{fail} + M_{success}} \longrightarrow$$

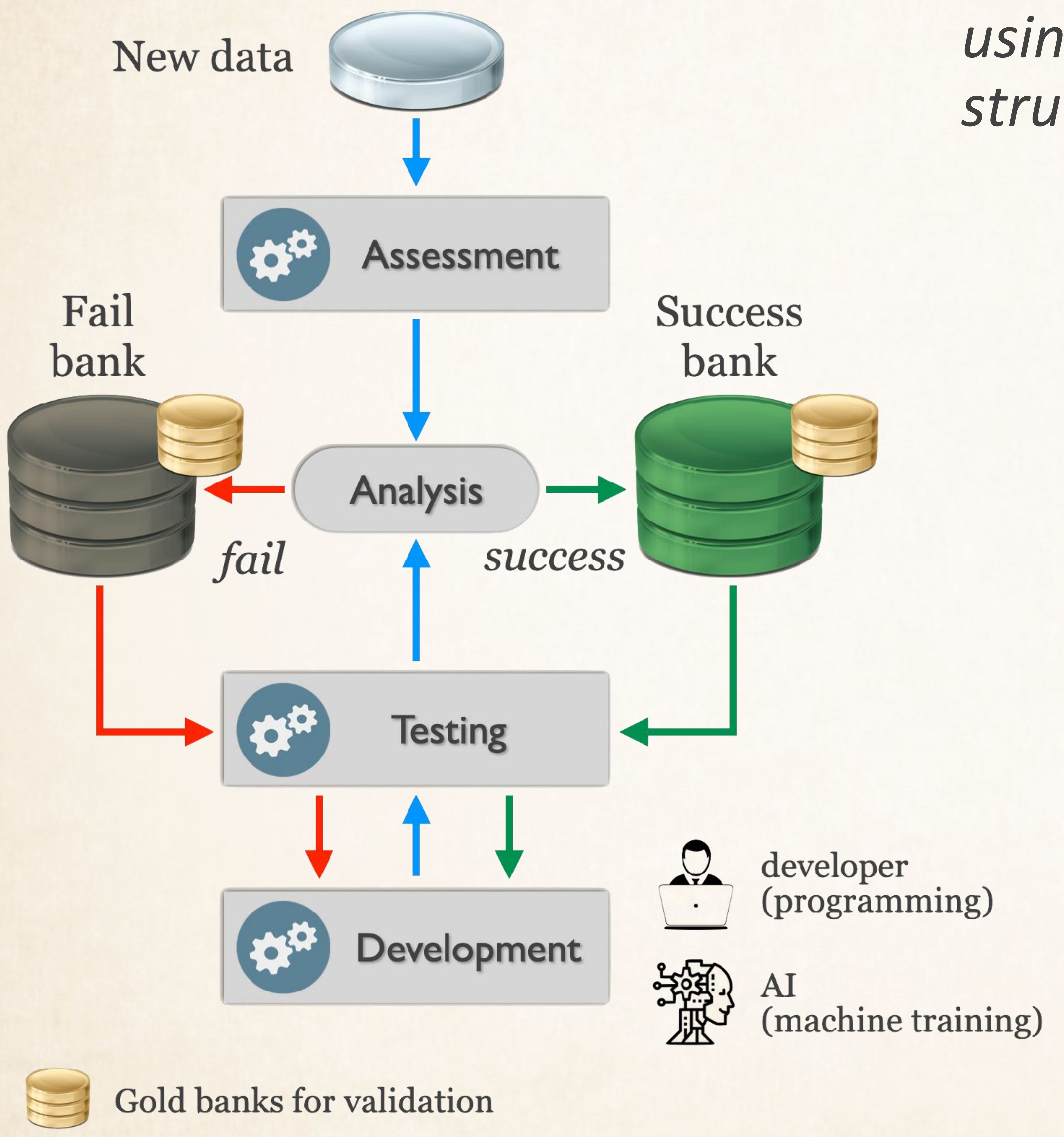
maximise by development

For efficient data (re)use

- maintain **both** Success and Fail banks
- for Analysis, **annotate and classify** data; at least:
  - quality
  - statistical significance
  - redundancy
- assume annotation **changes** with time
- **maintain** annotation software

Reusable data bank >> pile of files

Data Reuse in CCP4



In CCP4, data-driven development is based mostly on using processed (merged and unmerged) data and structure models from the PDB

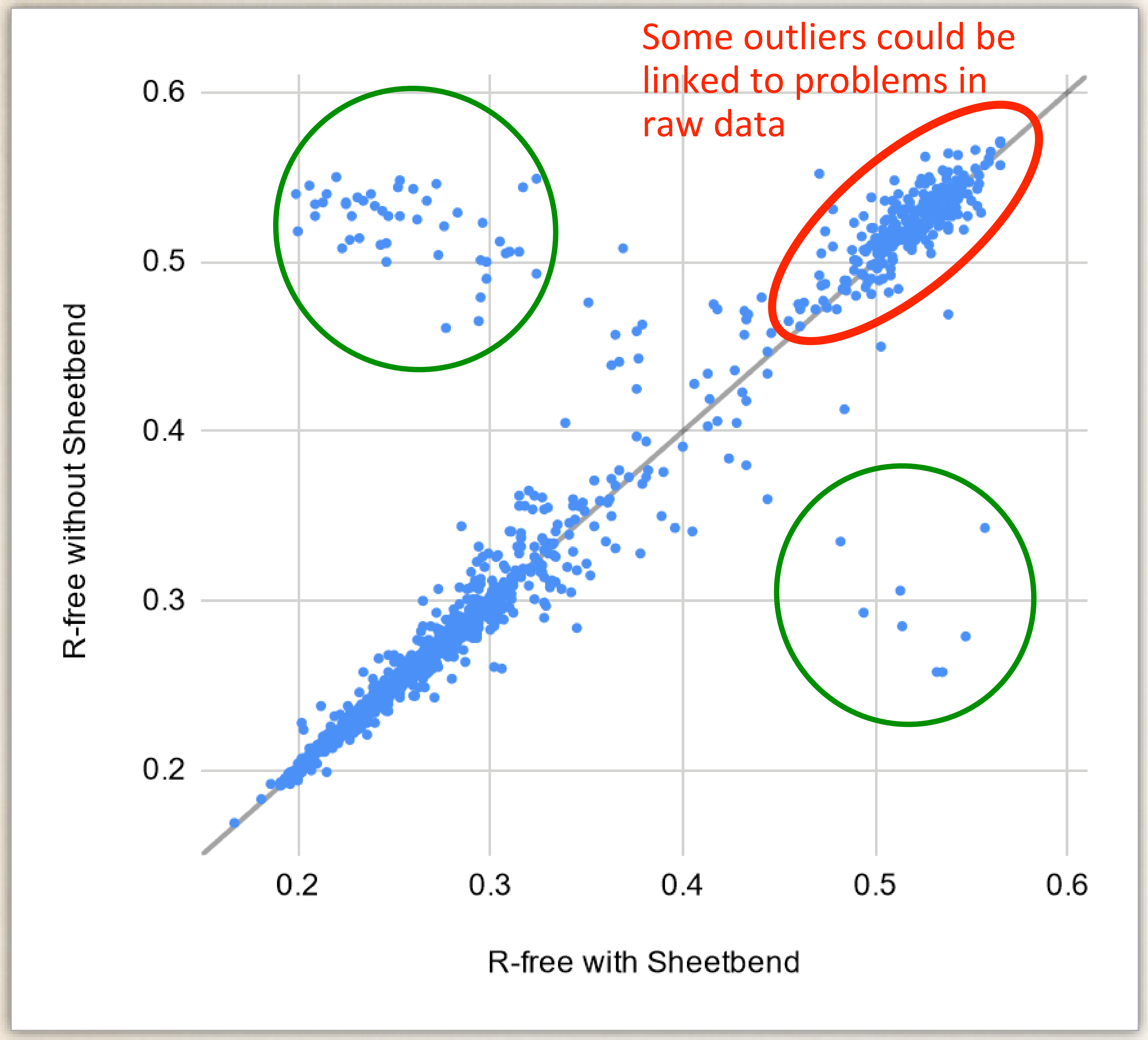
- simplified development loop is often used
- active developments include (but not limited to)

  - model building
  - MR, auto-MR, auto-EP
  - validation

- most of raw data reuse for development occurs in data processing project (DIALS)

  - may be hindered by the absence of comprehensive annotated raw data repository

Data Reuse in CCP4



The effect of shift field refinement on the performance of Modelcraft model building software. Some failures in red area could be due to data processing effects, not verifiable without access to raw data.

Picture provided by Paul Bond and Kevin Cowtan, University of York.

Raw data processing is first suspect for problems in structure solution, model building and refinement

- cannot be verified without access to raw data - benchmarking/training can go wrong
- access to raw data is essential for diagnosing data problems and development of corresponding tools

  - e.g., crystal pathologies; note that it is the **Fail bank** that is needed here

- the access must be programmatic, via **API** rather manual file fetching
- access to raw data is important for crystallography **education**

Data and Software *Maintenance*

Suppose raw data was archived from when CCP4 was founded (1979). Could it be used today? Not unless special care was taken.

- evolution of detectors (effect on processing parameters)
  - photographic film (known from 1885)
  - charge-coupled devices (CCDs, known from 1970s)
  - pixel area detectors (known from 1980s, XRD use from 1990s)

- evolution of image formats (effect on software)

  CBF, MAR, TIFF, MAR345, CBF, SMV, RAXIS, CCP4, STOE, HARVARD, BRUKER, EDF ...

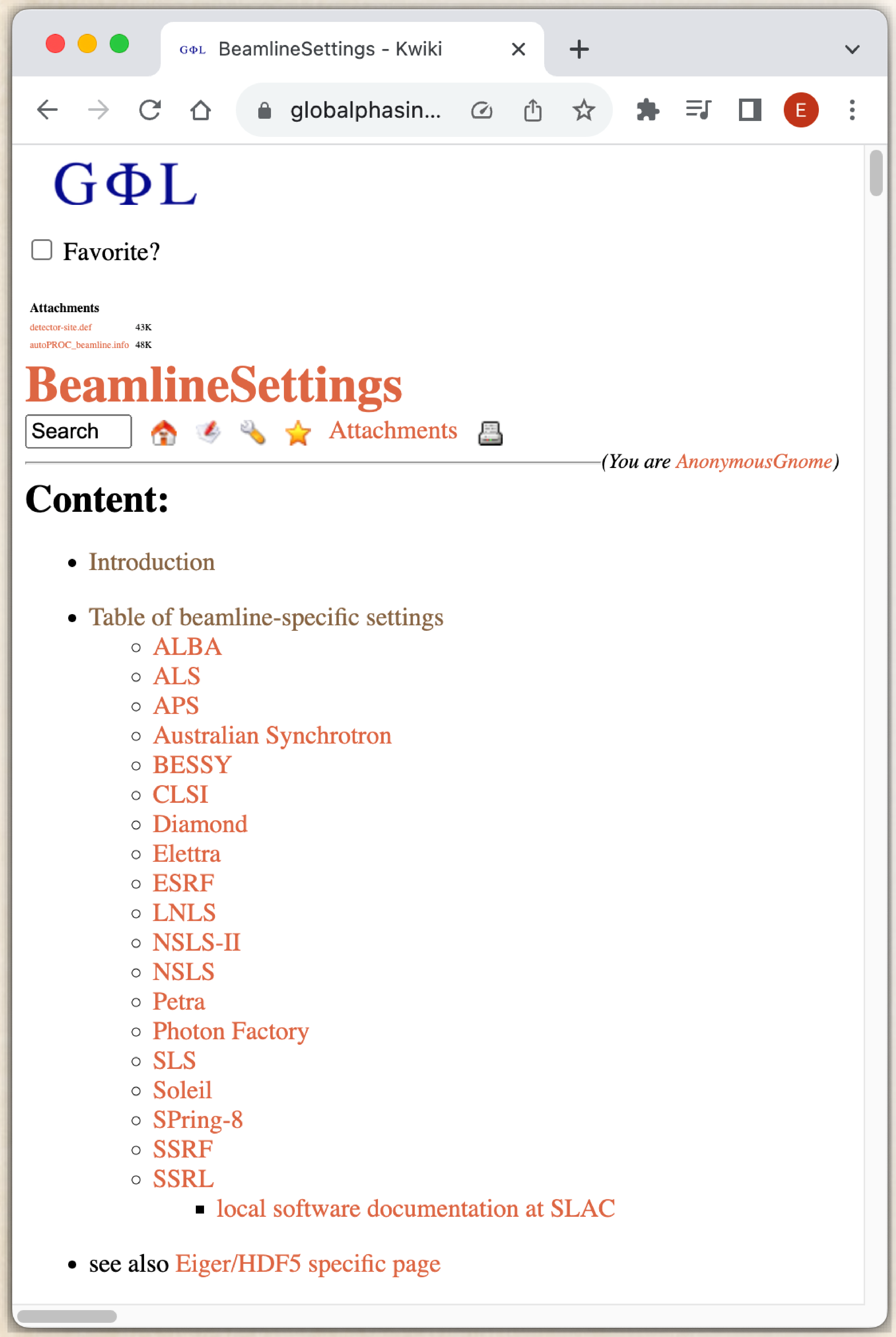
Software should be kept backward-compatible; usually ages much faster than data

- topic is at heart of CCP4 project; software maintenance proves to be a challenge

- programs do get retired for various reasons

Package	*Year*	*Status*	*Vendor/Distributor*
XDS	**1988**	active	Max-Plank
HKL	1997	active	HKL Research Inc.
DIALS	2018	active	Diamond/CCP4
Mosflm	1999	**sunset**	CCP4
d*Trek	1999	**sunset**	Rigaku

Data and Software *Maintenance*



GΦL

□ Favorite?

Attachments

detector-site.def	43K
autoPROC_beamline.info	48K

BeamlineSettings

Search Attachments

(*You are* *AnonymousGnome*)

Content:

- Introduction
- Table of beamline-specific settings
  - ALBA
  - ALS
  - APS
  - Australian Synchrotron
  - BESSY
  - CLSI
  - Diamond
  - Elettra
  - ESRF
  - LNLS
  - NSLS-II
  - NSLS
  - Petra
  - Photon Factory
  - SLS
  - Soleil
  - SPring-8
  - SSRF
  - SSRL
    - local software documentation at SLAC
- see also Eiger/HDF5 specific page

Having just image files is better than nothing but not enough!

- instrument effects
  - damaged pixels and beamline settings vary from place to place, and may change in time.
  - essential for data re-use
  - pixel maps are included in some formats but not in all
  - Global Phasing Ltd. makes excellent job maintaining this information

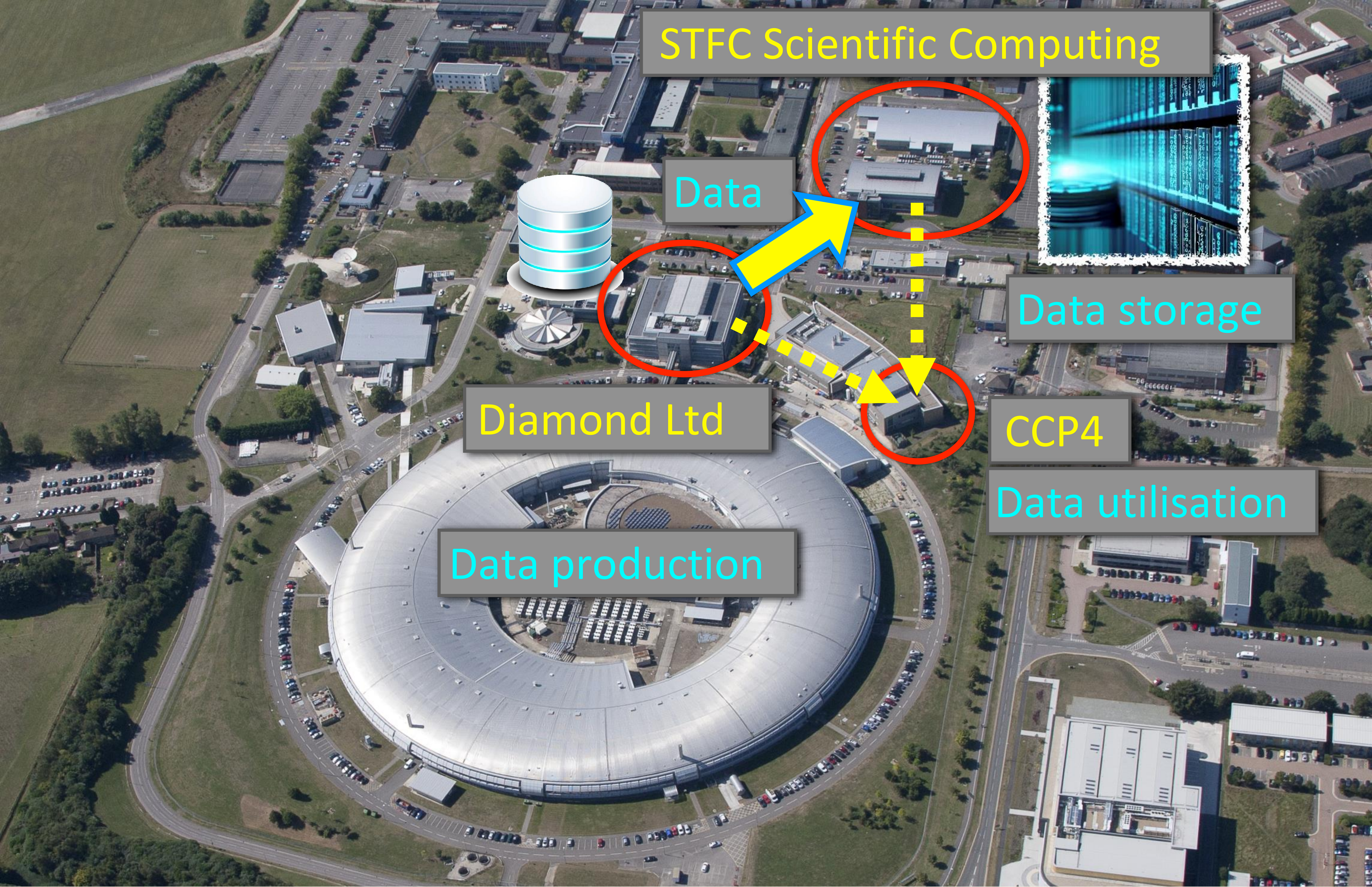
Data and Software *Maintenance*

For efficient, long-term data re-use, consider:

- deposition (proxy) data format
  - must be extendable; mind the future!
- annotation framework
  - extendable for future
  - as automatic as possible (curators may be needed)
- dedicated software support
  - data processing software
  - format adapters
  - annotation software
  - tests and integrity checks
- central gateway
  - repositories may (and probably should) be distributed
  - access API

It is a Big and Costly Project

buy cheap, buy twice



STFC Scientific Computing

Data

Data storage

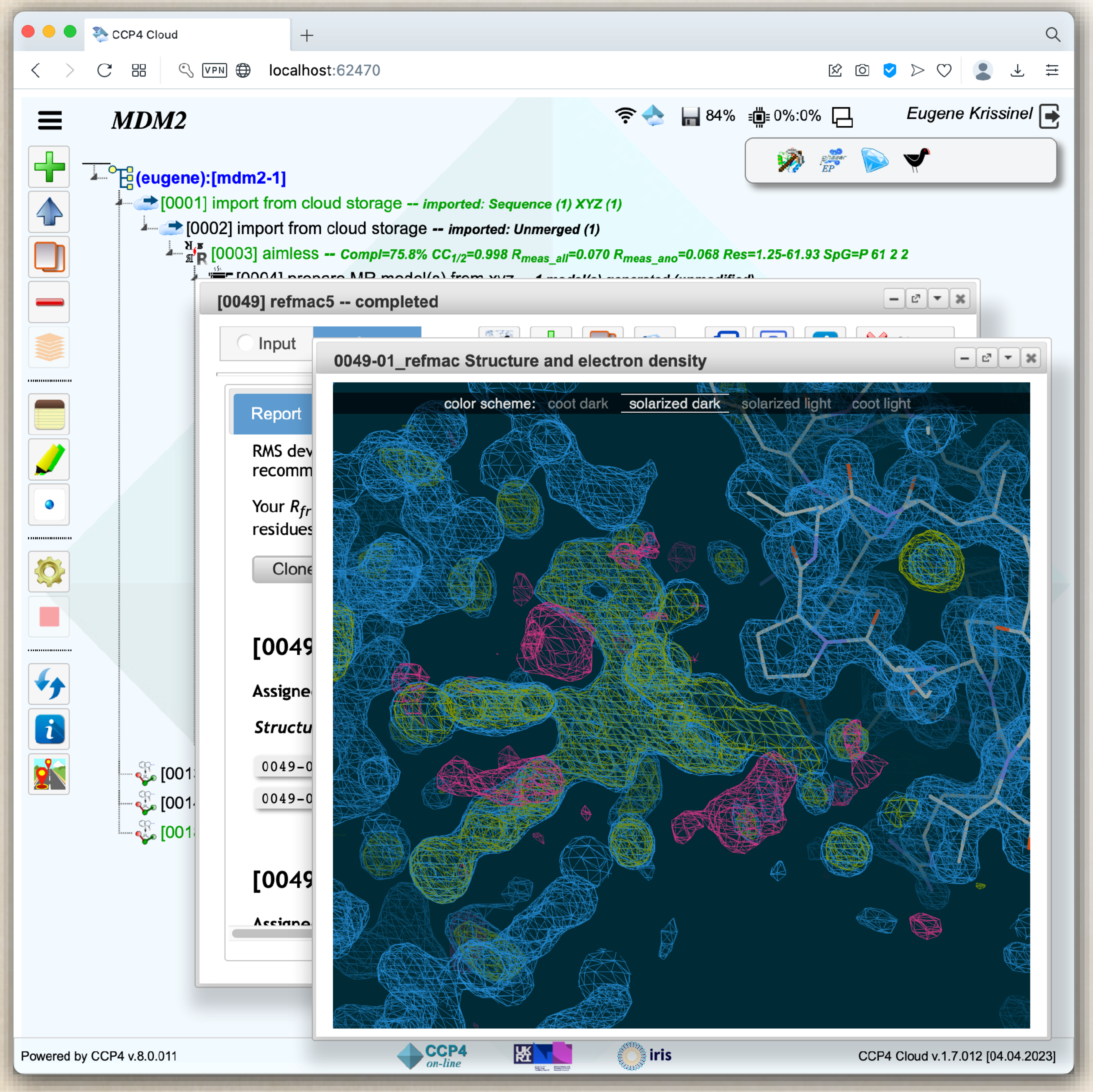
Diamond Ltd

CCP4

Data utilisation

Data production

CCP4 and *Utilisation of Data*



CCP4 Cloud

localhost:62470

MDM2

84% 0%:0%

Eugene Krissinel

(eugene):[mdm2-1]

[0001] import from cloud storage -- *imported: Sequence (1) XYZ (1)*

[0002] import from cloud storage -- *imported: Unmerged (1)*

[0003] aimless -- *Compl=75.8% $CC_{1/2}$=0.998 R_{meas_all}=0.070 R_{meas_ano}=0.068 Res=1.25-61.93 SpG=P 61 2 2*

[0004] prepare MR model(s) from xyz -- 1 model(s) generated (unmodified)

[0049] refmac5 -- completed

Input

Report

RMS dev
recomm

Your R_{fr}
residues

Clone

[0049

Assigne

Structu

0049-0

0049-0

[0049

Assigne

[001

[001

[001

0049-01_refmac Structure and electron density

color scheme: coot dark solarized dark solarized light coot light

Powered by CCP4 v.8.0.011

CCP4 Cloud v.1.7.012 [04.04.2023]

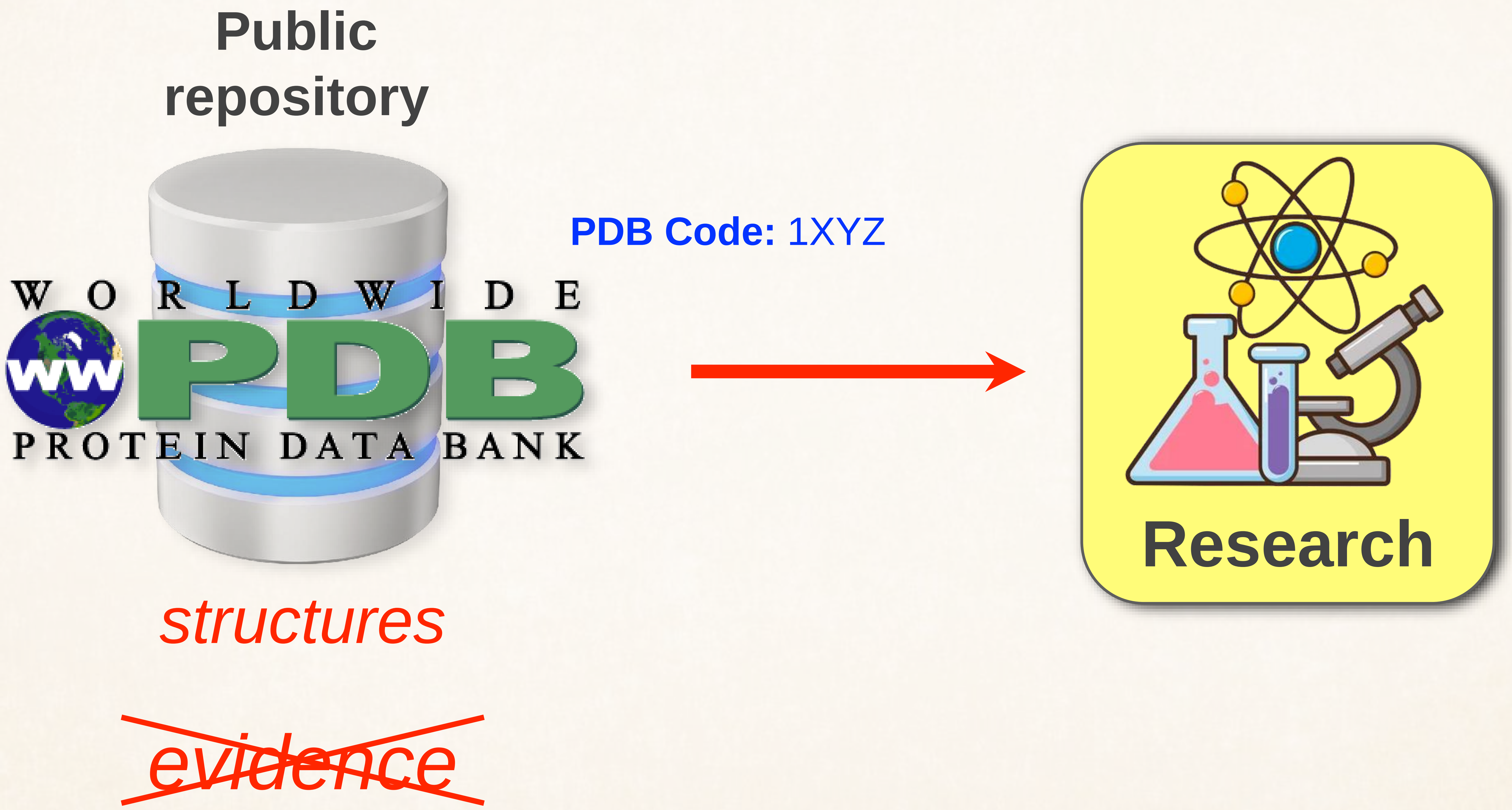
CCP4 Cloud for solving structures online

http://cloud.ccp4.ac.uk

- New way to solve **structures**
- New way to maintain **data**
- New way to maintain **projects**
- New way to use CCP4 **Software**
- Geographically agnostic

Data Trail in Structural Biology

Structural Biology is a data-driven discipline: no data – no research



Public repository

PDB Code: 1XYZ

Research

structures

~~*evidence*~~

CCP4

Data Trail in Structural Biology

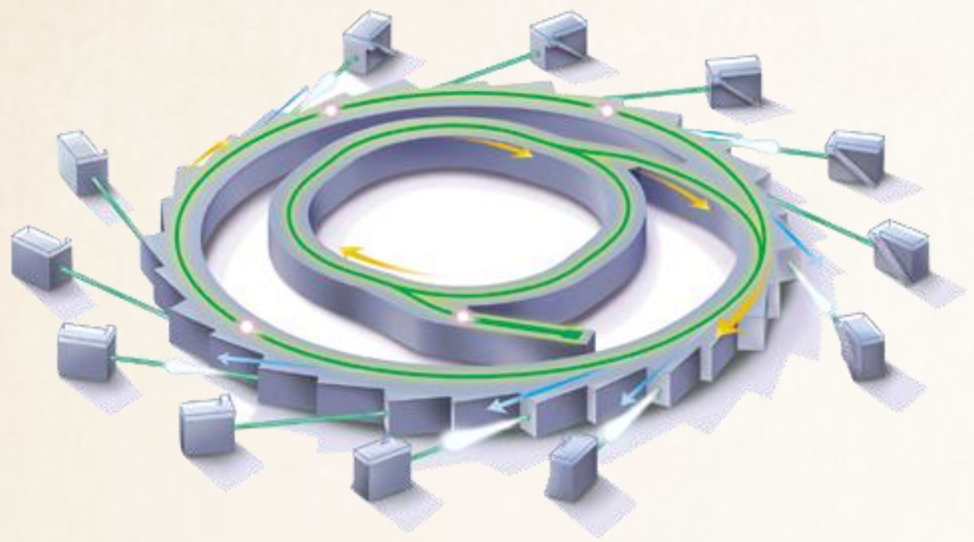
Research Complex at Harwell

UKRI Science and Technology Facilities Council

Scientific Computing

Current state of affairs: technique with limited reproducibility

Experiment



evidence

As a rule, no or limited public access to raw data

Reproducibility is limited

Structure Solution

CCP4, Phenix, Global Phasing ...

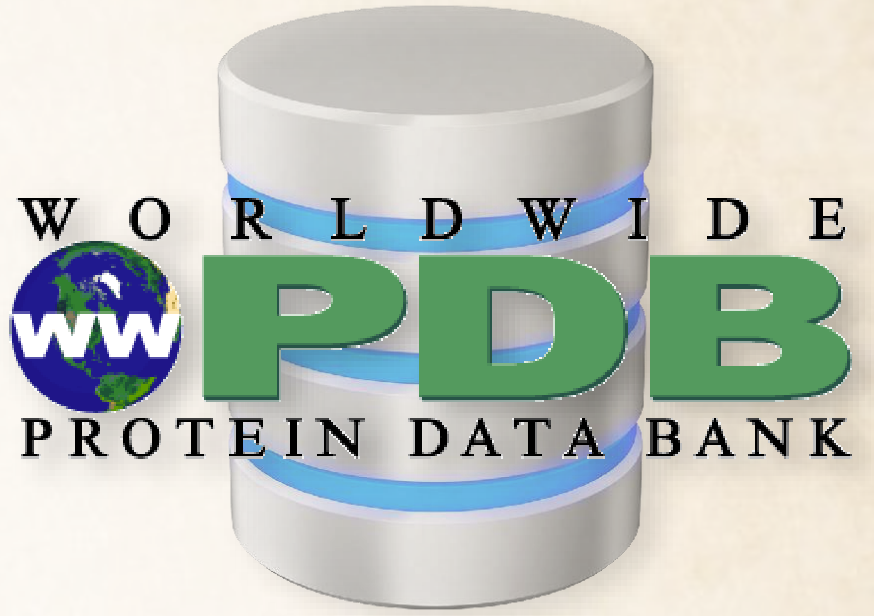
interpretation

- *methods*
- *decisions*
- *assumptions*
- *doubts resolution*
- *validation*
- *alternatives*

As a rule, no public access to all details (publications and local computers)

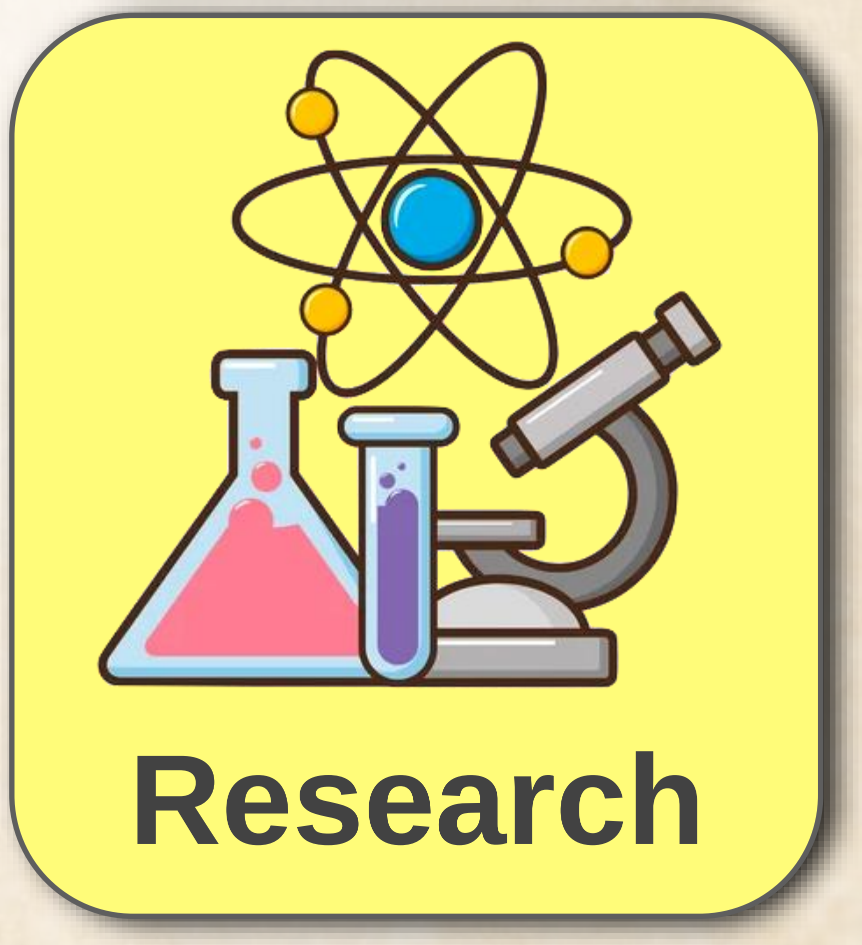
Reproducibility is limited

Public repository



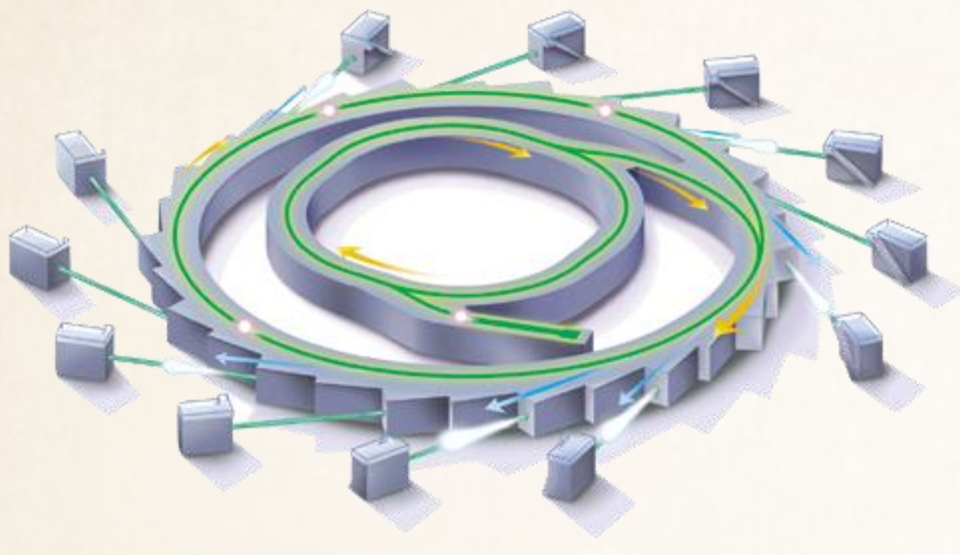
structures

PDB Code: 1XYZ



A better picture: raw data repository cross-linked with the PDB

Experiment



evidence

Public repository

raw data

Image DOI

Structure Solution

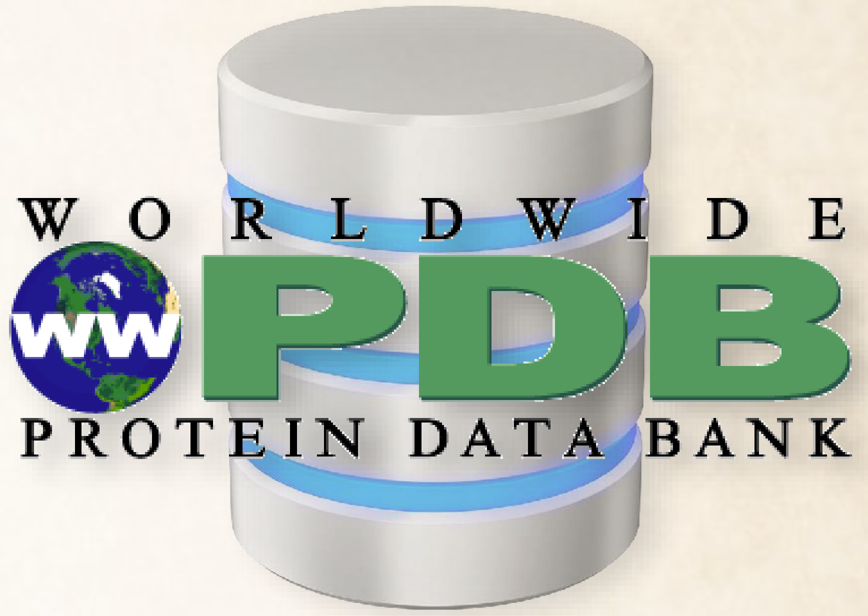
CCP4, Phenix, Global Phasing ...

interpretation

- *methods*
- *decisions*
- *assumptions*
- *doubts resolution*
- *validation*
- *alternatives*

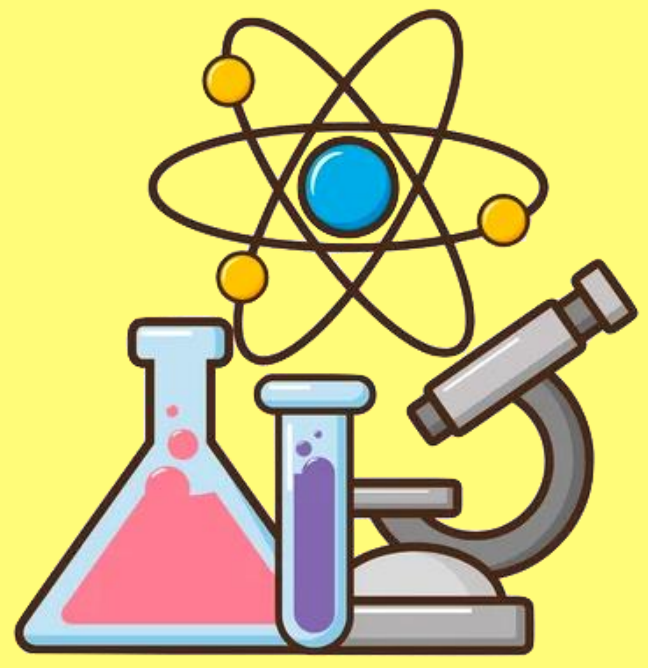
As a rule, no public access to all details

Public repository



structures

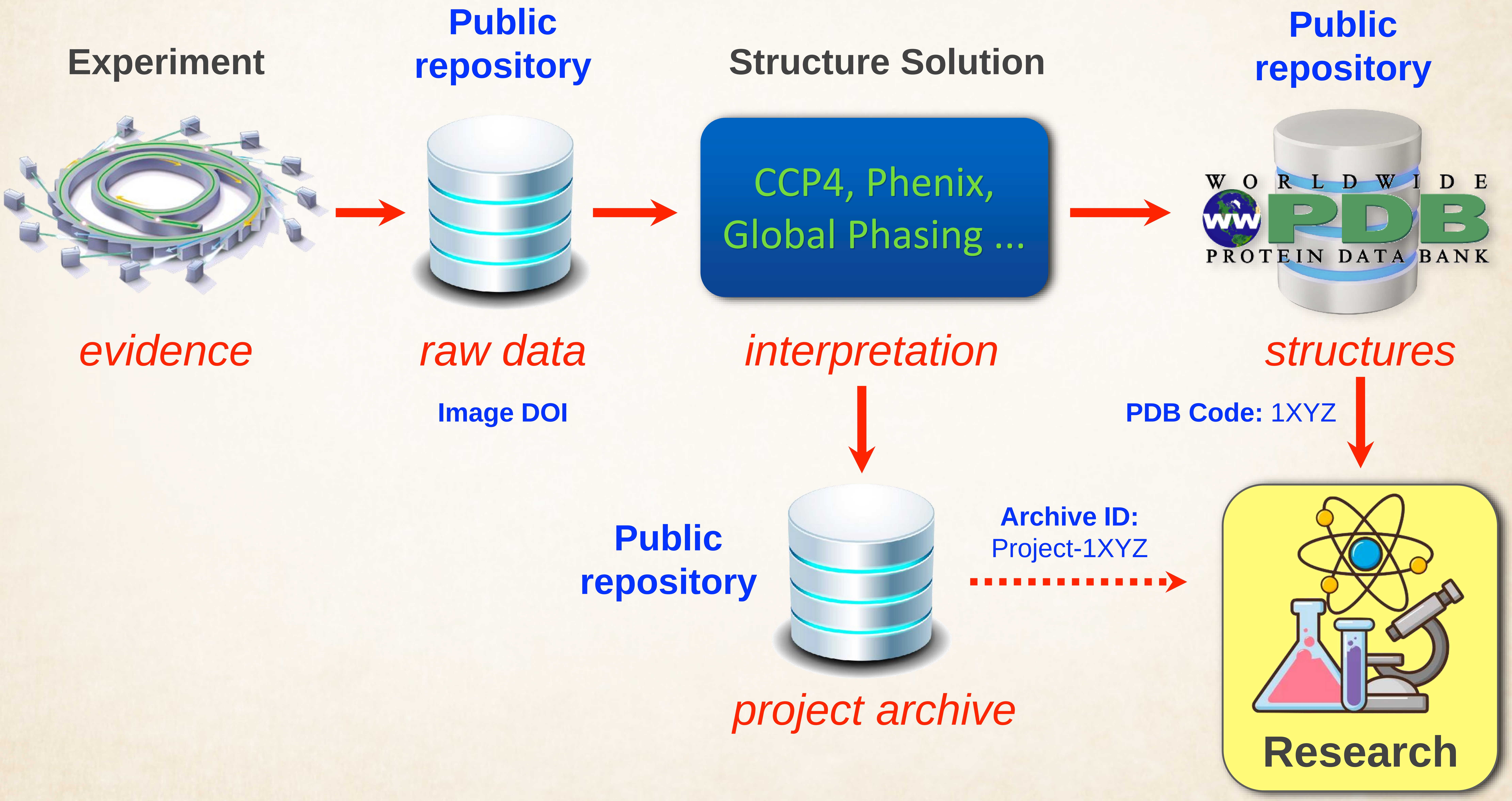
PDB Code: 1XYZ



Research

Data Trail in Structural Biology

Yet better picture



Experiment

evidence

Public repository

raw data

Image DOI

Structure Solution

CCP4, Phenix, Global Phasing ...

interpretation

Public repository

structures

PDB Code: 1XYZ

Public repository

project archive

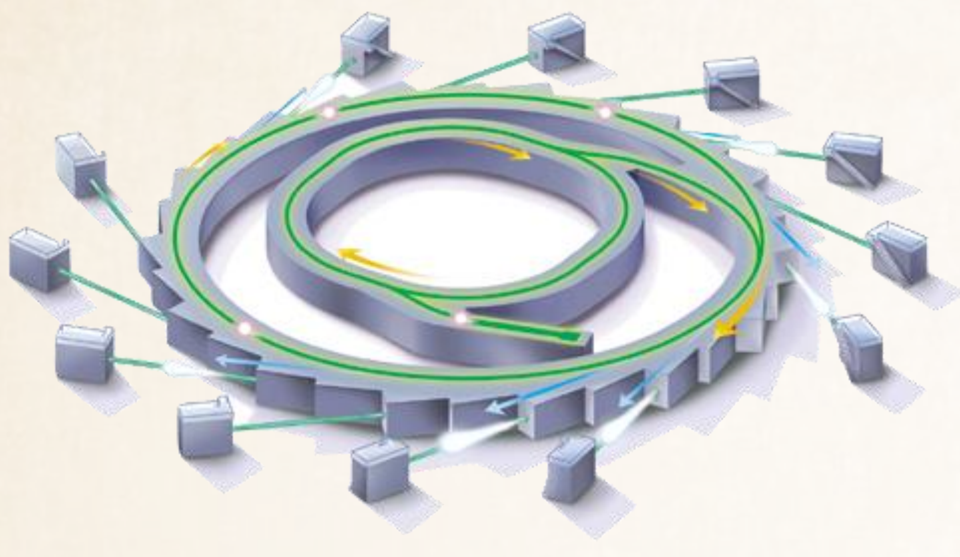
Archive ID: Project-1XYZ

Research

Data Trail in Structural Biology

Possible picture in future

Experiment



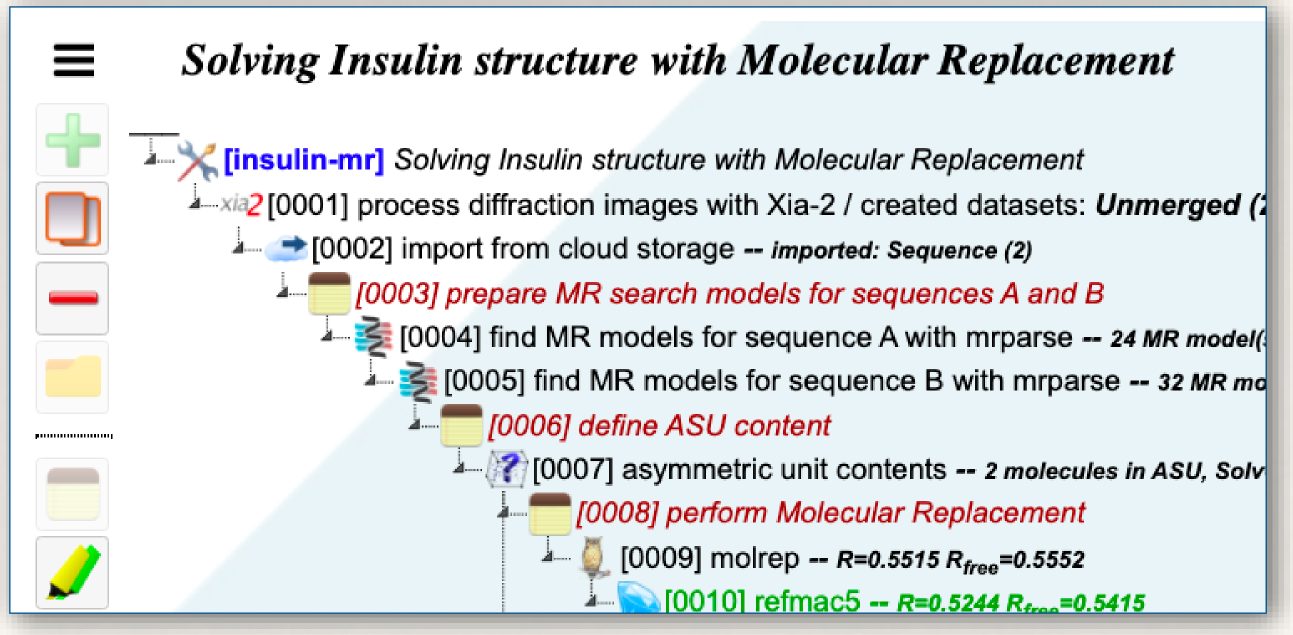
evidence

Public repository

raw data

Image DOI

Structure Solution in CCP4 Cloud



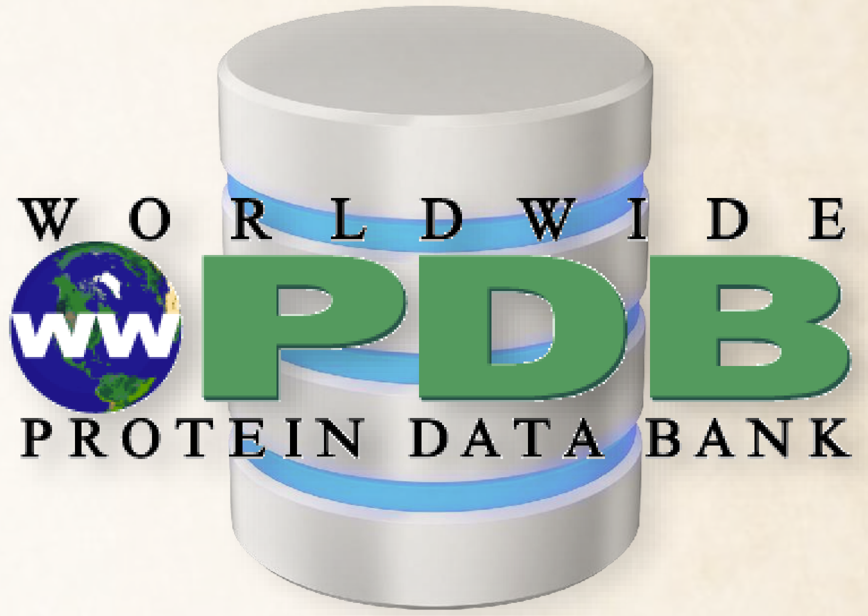
≡ ***Solving Insulin structure with Molecular Replacement***

[insulin-mr] *Solving Insulin structure with Molecular Replacement*

- [0001] process diffraction images with Xia-2 / created datasets: ***Unmerged (***
  - [0002] import from cloud storage -- ***imported: Sequence (2)***
    - *[0003] prepare MR search models for sequences A and B*
      - [0004] find MR models for sequence A with mrparse -- ***24 MR model(***
        - [0005] find MR models for sequence B with mrparse -- ***32 MR mo***
          - *[0006] define ASU content*
            - [0007] asymmetric unit contents -- ***2 molecules in ASU, Solv***
              - *[0008] perform Molecular Replacement*
                - [0009] molrep -- ***R=0.5515 R_{free}=0.5552***
                  - [0010] refmac5 -- ***R=0.5244 R_{free}=0.5415***

interpretation

Public repository



structures

PDB Code: 1XYZ

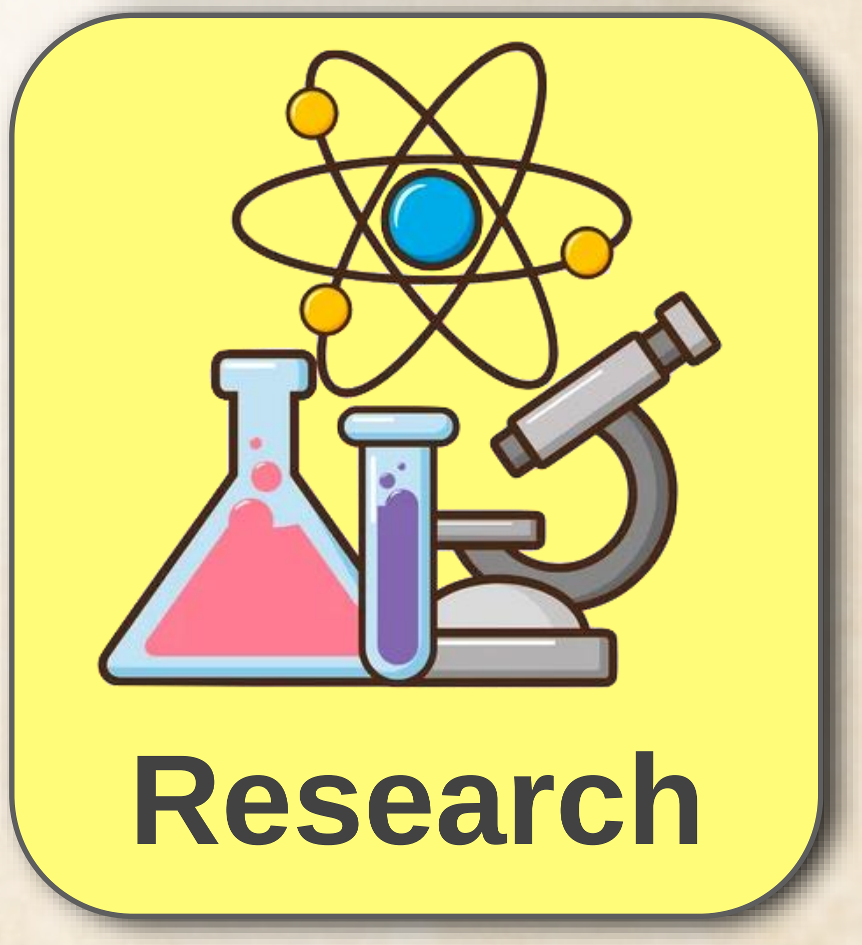
Public repository

(from 01/2023)

Archive ID: CCP4-1XYZ

CCP4 Cloud archive

emerging, partly available



What Raw Data Reuse means for CCP4

If Raw Data were available as routinely as PDB models, CCP4 could:

- do better testing of some components
- further enforce data-driven development practices
- deliver software of higher quality
- help improving data provenance in structural biology

For Raw Data to be available as routinely as PDB models, CCP4 would need to:

- maintain backward compatibility of data processing software
- maintain raw data format converters
- keep in sync with relevant metadata frameworks
- maintain data links with data producing facilities and data repositories
- maintain raw data access facilities in CCP4 front-ends for users

Acknowledgements

CCP4 Core, STFC, Harwell, UK:

Charles Ballard
Ronan Keegan
David Waterman
Andrey Lebedev
Ville Uski
Maria Fando
Tarik Drevon
David McDonagh
Jools Wills
Daniel Garza

MRC/LMB, Cambridge, UK:

Garib Murshudov
Paul Emsley
Robert Nicholls

University of Cambridge, UK:

Randy Read
Airlie McCoy
Robert Oeffner
Tristan Croll

University of York, UK:

Keith Wilson
Kevin Cowtan
Jon Agirre
Paul Bond
Stuart McNicholas
Filomeno Sanchez

EMBL-EBI, Hinxton, UK:

Sameer Velankar
John Berrisford
Deborah Harrus

Leiden University, The Netherlands:

Navraj Pannu
Pavol Skubak

Global Phasing Ltd, Cambridge, UK:

Gerard Bricogne
Clemens Vonrhein
Marcin Wojdyr

University of Liverpool, UK:

Dan Rigden
Adam Simpkin
Jens Thomas

Newcastle University, UK:

Martin Noble
Arnaud Basle

University of Southampton, UK:

Ivo Tews

University of Exeter, UK:

Michael Isupov

EMBL-Hamburg, Germany

Victor Lamzin
Grzegorz Chojnowski

Francis Crick Institute, UK:

Andrew Purkiss

Biotechnology and Biological Sciences Research Council (*BBSRC*) *UK*

Research grants BB/L0070317/1 (2014-2019); BB/Y004019/1 (2019-2024); BB/V015591/1 (2022-2025)