“ELIXIR and Open Data”

View from an ELIXIR Node”

Barend Mons
Prof. Biosemantics, LUMC, Scientific director NBIC,
Head of ELIXIR Node
Netherlands

18 December 2013
ELIXIR Launch, Brussels

European Life Sciences Infrastructure for Biological Information
www.elixir-europe.org

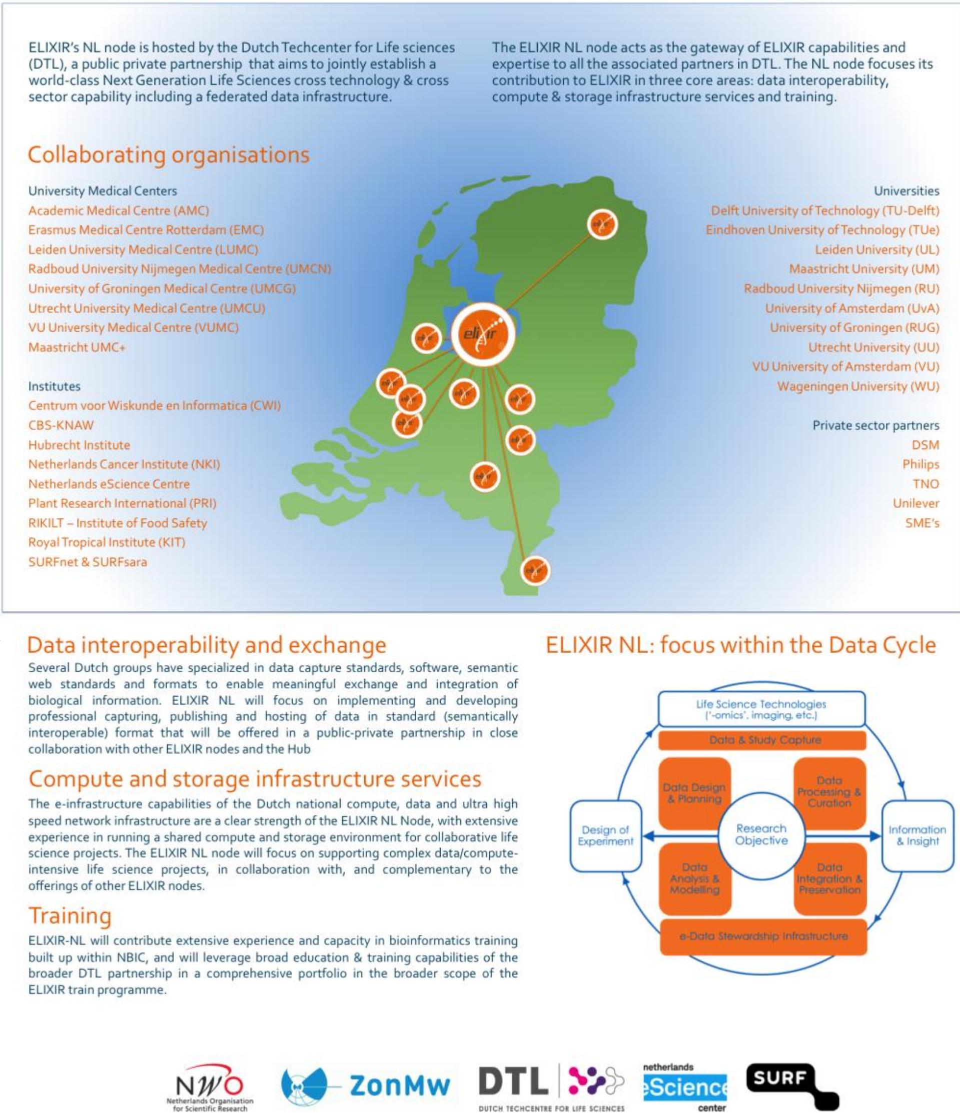
Outline

- The Dutch Node
- Data in the eScience era: Pattern Recognition and Excavation
- Data Collection, Archiving and: Reduction
- Why is ELIXIR important for Open Data?
- What are the needs of clinical institutes and Industry: Open and Managed Data
- Why training as part of ELIXIR?

Data interoperability and exchange

Compute and storage infrastructure services

Training & Education



ELIXIR's NL node is hosted by the Dutch Techcenter for Life sciences (DTL), a public private partnership that aims to jointly establish a world-class Next Generation Life Sciences cross technology & cross sector capability including a federated data infrastructure.

The ELIXIR NL node acts as the gateway of ELIXIR capabilities and expertise to all the associated partners in DTL. The NL node focuses its contribution to ELIXIR in three core areas: data interoperability, compute & storage infrastructure services and training.

Collaborating organisations

University Medical Centers
Academic Medical Centre (AMC)
Erasmus Medical Centre Rotterdam (EMC)
Leiden University Medical Centre (LUMC)
Radboud University Nijmegen Medical Centre (UMCN)
University of Groningen Medical Centre (UMCG)
Utrecht University Medical Centre (UMCU)
VU University Medical Centre (VUMC)
Maastricht UMC+

Institutes
Centrum voor Wiskunde en Informatica (CWI)
CBS-KNAW
Hubrecht Institute
Netherlands Cancer Institute (NKI)
Netherlands eScience Centre
Plant Research International (PRI)
RIKILT – Institute of Food Safety
Royal Tropical Institute (KIT)
SURFnet & SURFsara

Universities
Delft University of Technology (TU-Delft)
Eindhoven University of Technology (TUe)
Leiden University (UL)
Maastricht University (UM)
Radboud University Nijmegen (RU)
University of Amsterdam (UvA)
University of Groningen (RUG)
Utrecht University (UU)
VU University of Amsterdam (VU)
Wageningen University (WU)

Private sector partners
DSM
Philips
TNO
Unilever
SME's

Data interoperability and exchange

Several Dutch groups have specialized in data capture standards, software, semantic web standards and formats to enable meaningful exchange and integration of biological information. ELIXIR NL will focus on implementing and developing professional capturing, publishing and hosting of data in standard (semantically interoperable) format that will be offered in a public-private partnership in close collaboration with other ELIXIR nodes and the Hub

Compute and storage infrastructure services

The e-infrastructure capabilities of the Dutch national compute, data and ultra high speed network infrastructure are a clear strength of the ELIXIR NL Node, with extensive experience in running a shared compute and storage environment for collaborative life science projects. The ELIXIR NL node will focus on supporting complex data/compute-intensive life science projects, in collaboration with, and complementary to the offerings of other ELIXIR nodes.

Training

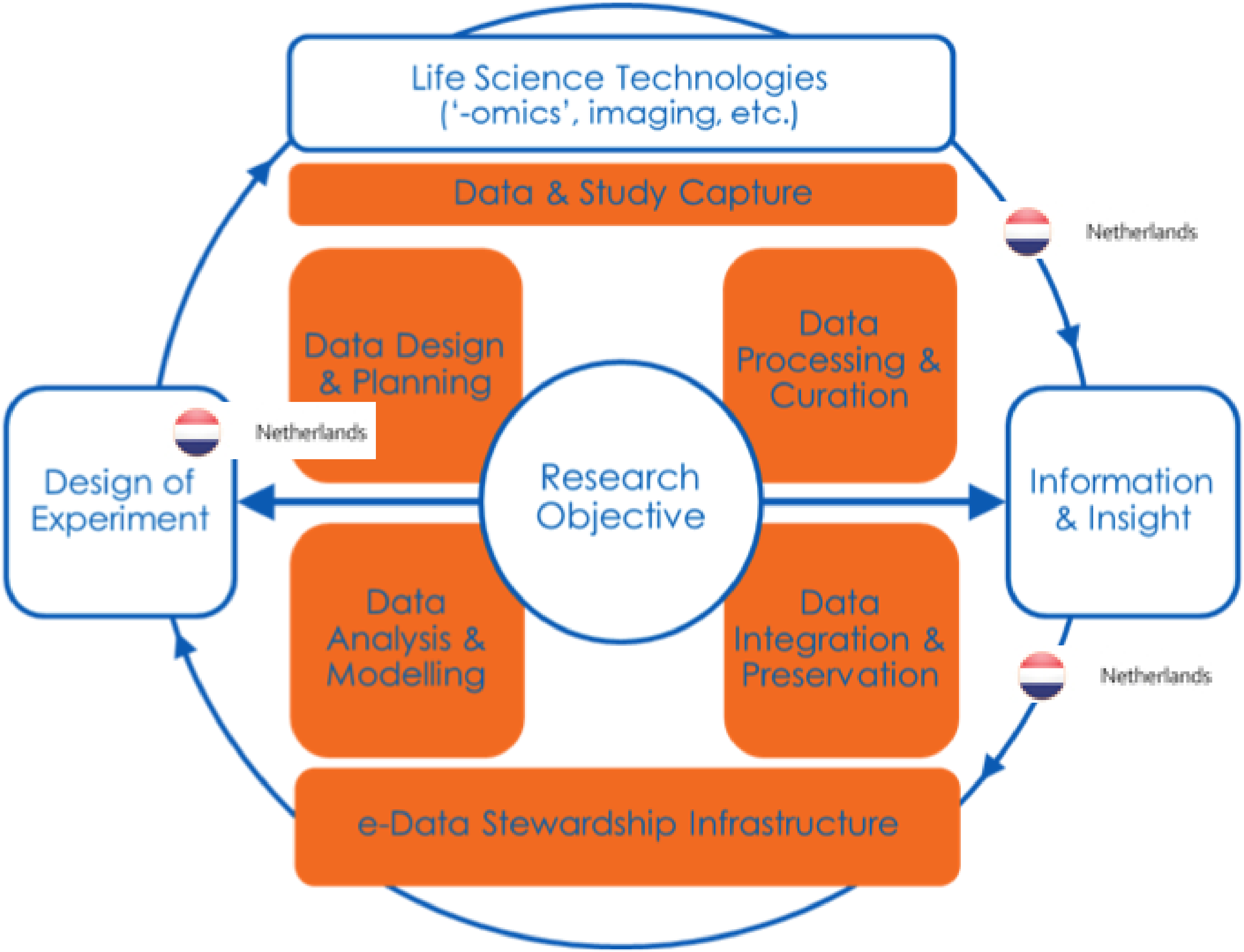
ELIXIR-NL will contribute extensive experience and capacity in bioinformatics training built up within NBIC, and will leverage broad education & training capabilities of the broader DTL partnership in a comprehensive portfolio in the broader scope of the ELIXIR train programme.

ELIXIR NL: focus within the Data Cycle

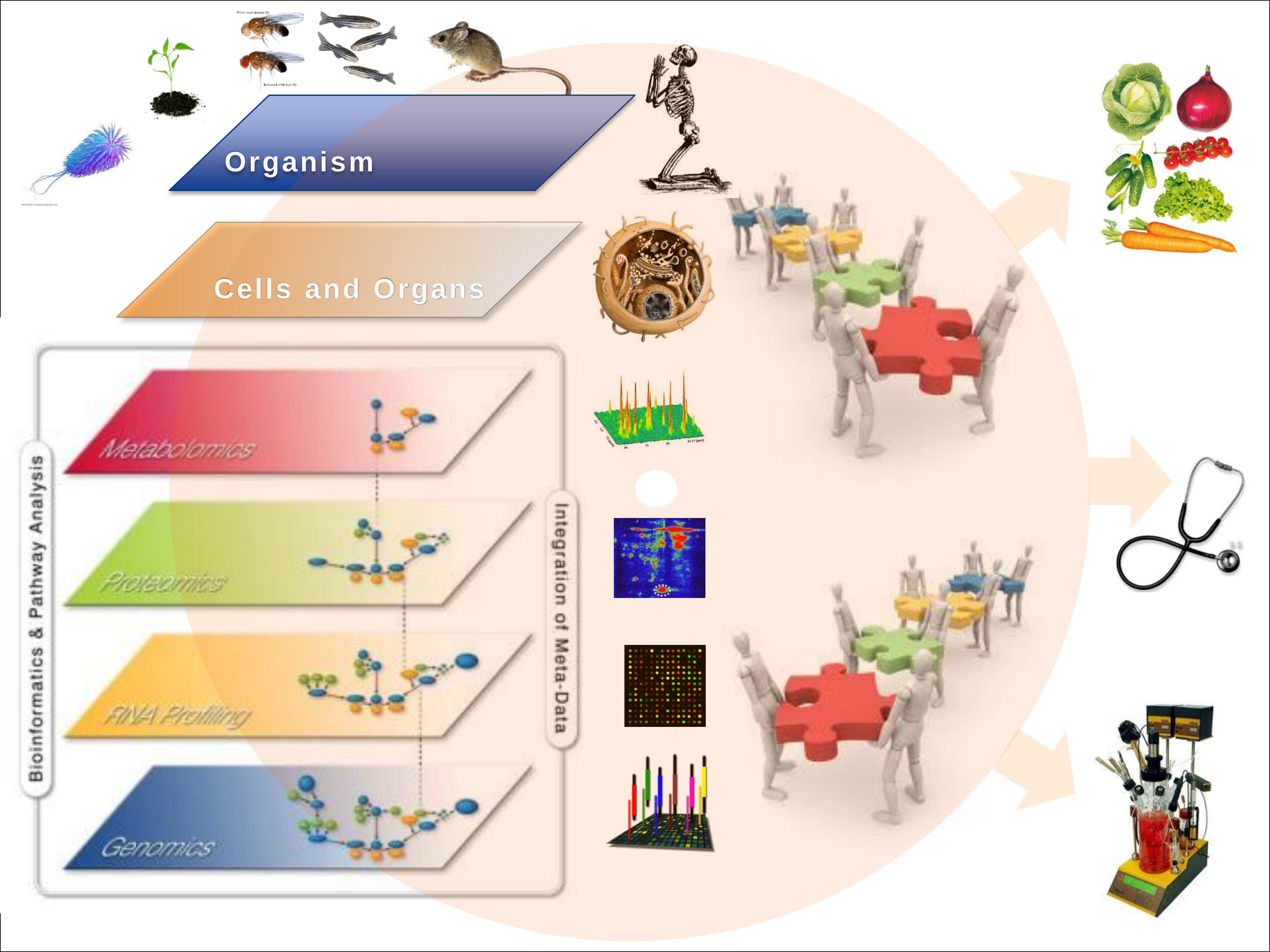
Life Science Technologies ('-omics', imaging, etc.)
Data & Study Capture
Data Design & Planning
Data Processing & Curation
Design of Experiment
Research Objective
Information & Insight
Data Analysis & Modelling
Data Integration & Preservation
e-Data Stewardship Infrastructure

Nodes and a Hub in NL (DTL) > Node in >>>>>>

The Data cycle in eScience


Life Science Technologies
('-omics', imaging, etc.)
Data & Study Capture
Netherlands
Data Design & Planning
Data Processing & Curation
Netherlands
Design of Experiment
Research Objective
Information & Insight
Data Analysis & Modelling
Data Integration & Preservation
Netherlands
e-Data Stewardship Infrastructure

Data Stewardship covers the entire datacycle >>>>>> elixir ?



Organism

Cells and Organs

Bioinformatics & Pathway Analysis

Metabolomics

Proteomics

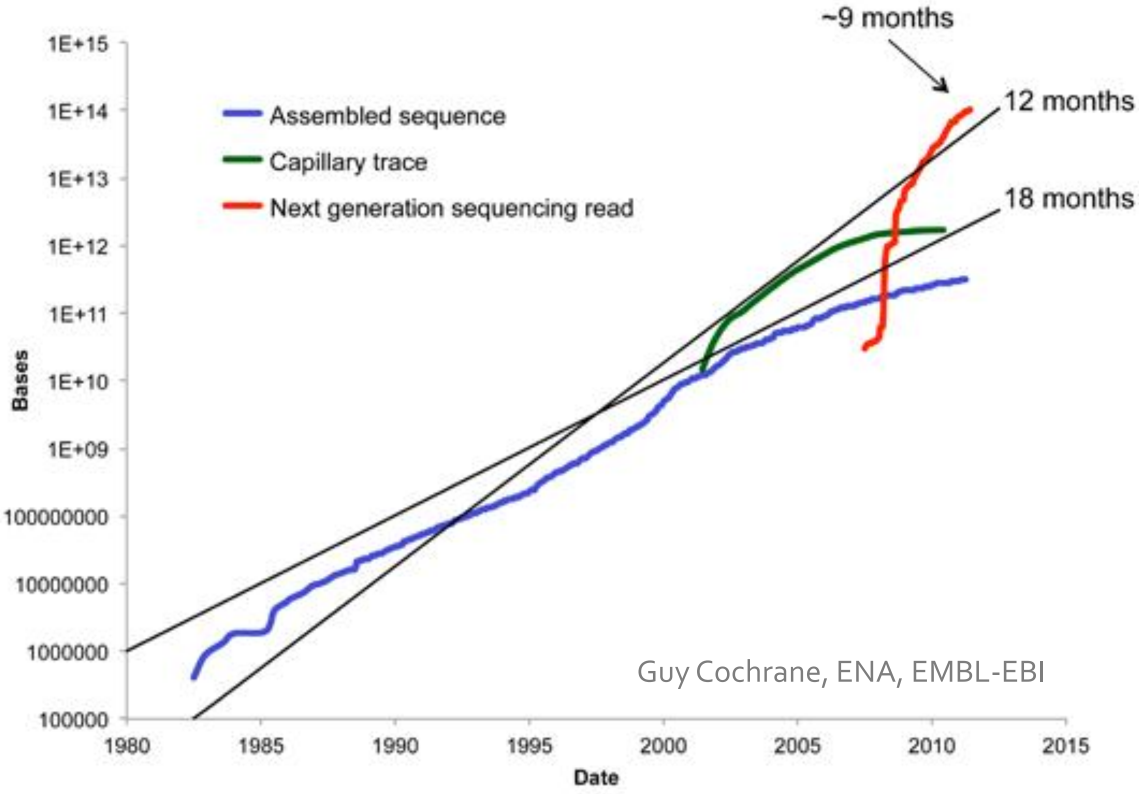
RNA Profiling

Genomics

Integration of Meta-Data

The Data Challenge

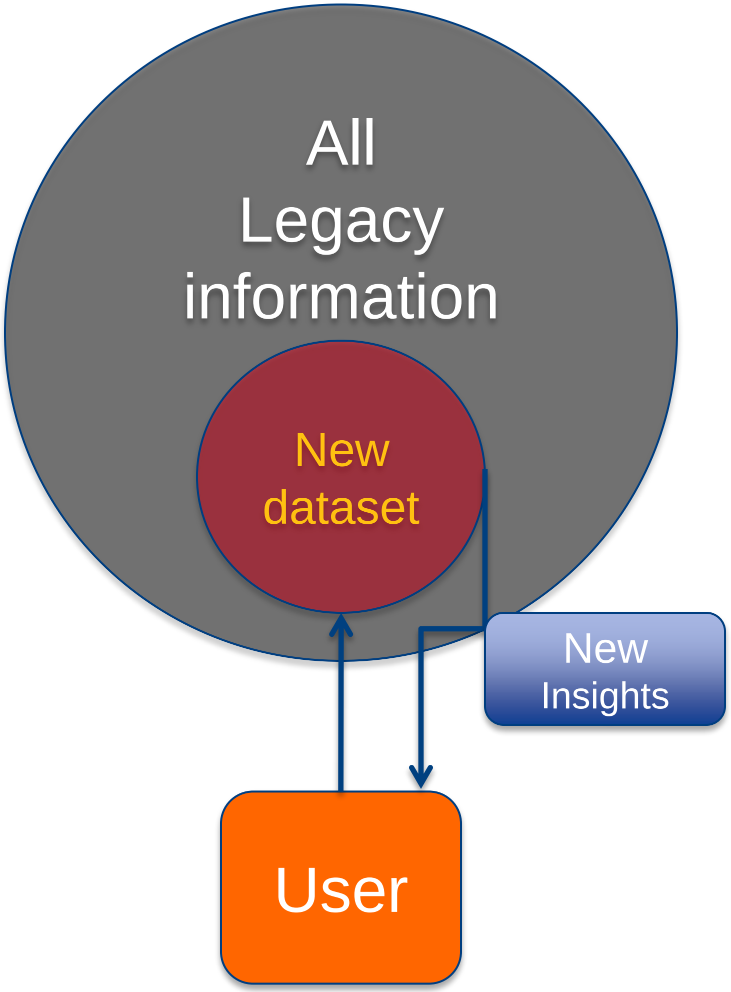
- Computer speed and storage capacity is **doubling every 18 months** and this rate is steady
- DNA sequence data is **doubling every 6-8 months** over the last 3 years and looks to continue for this decade



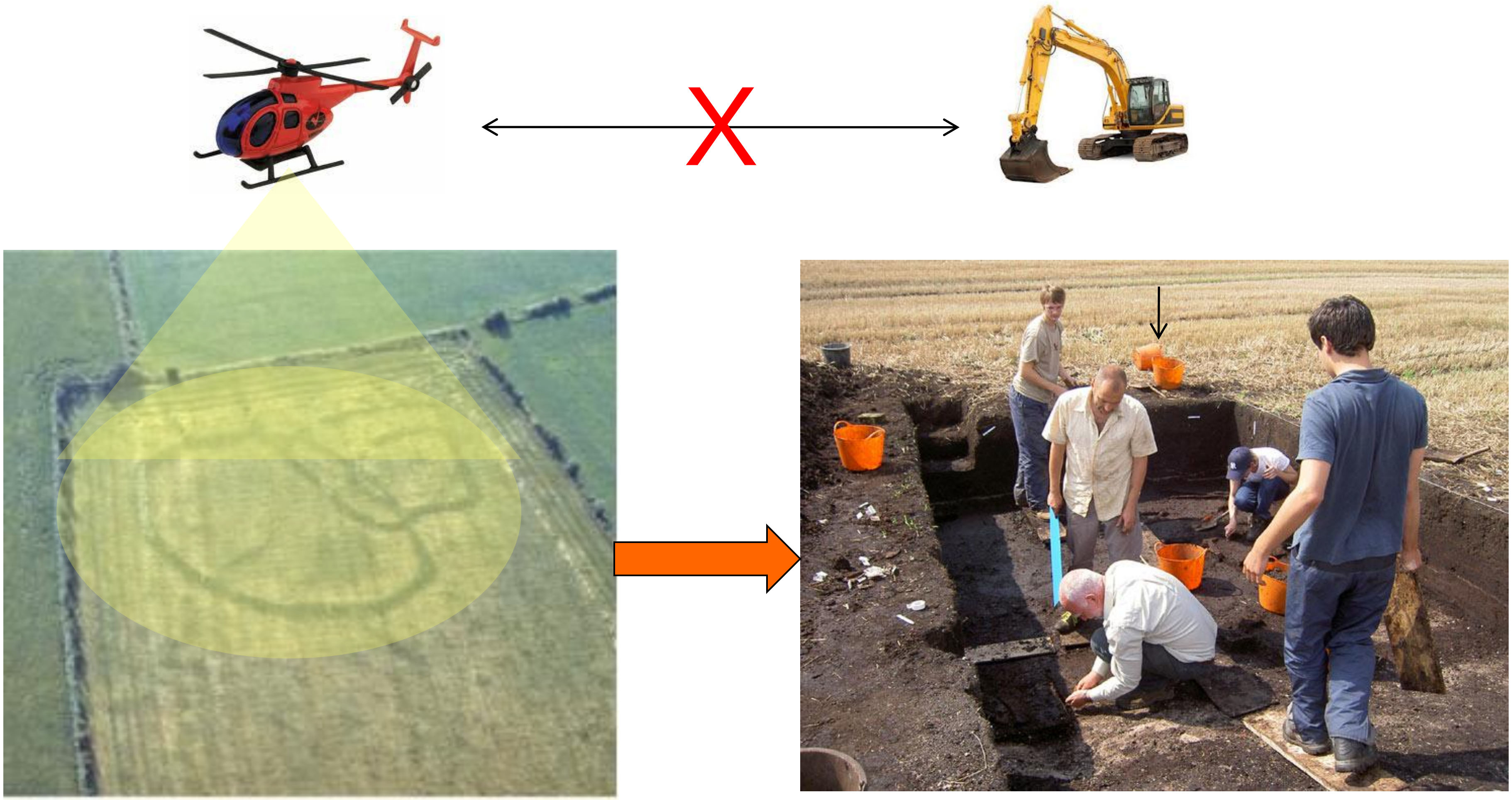
Guy Cochrane, ENA, EMBL-EBI

Proper Data stewardship and analysis may be THE limiting factor in eScience

Simplified eScience



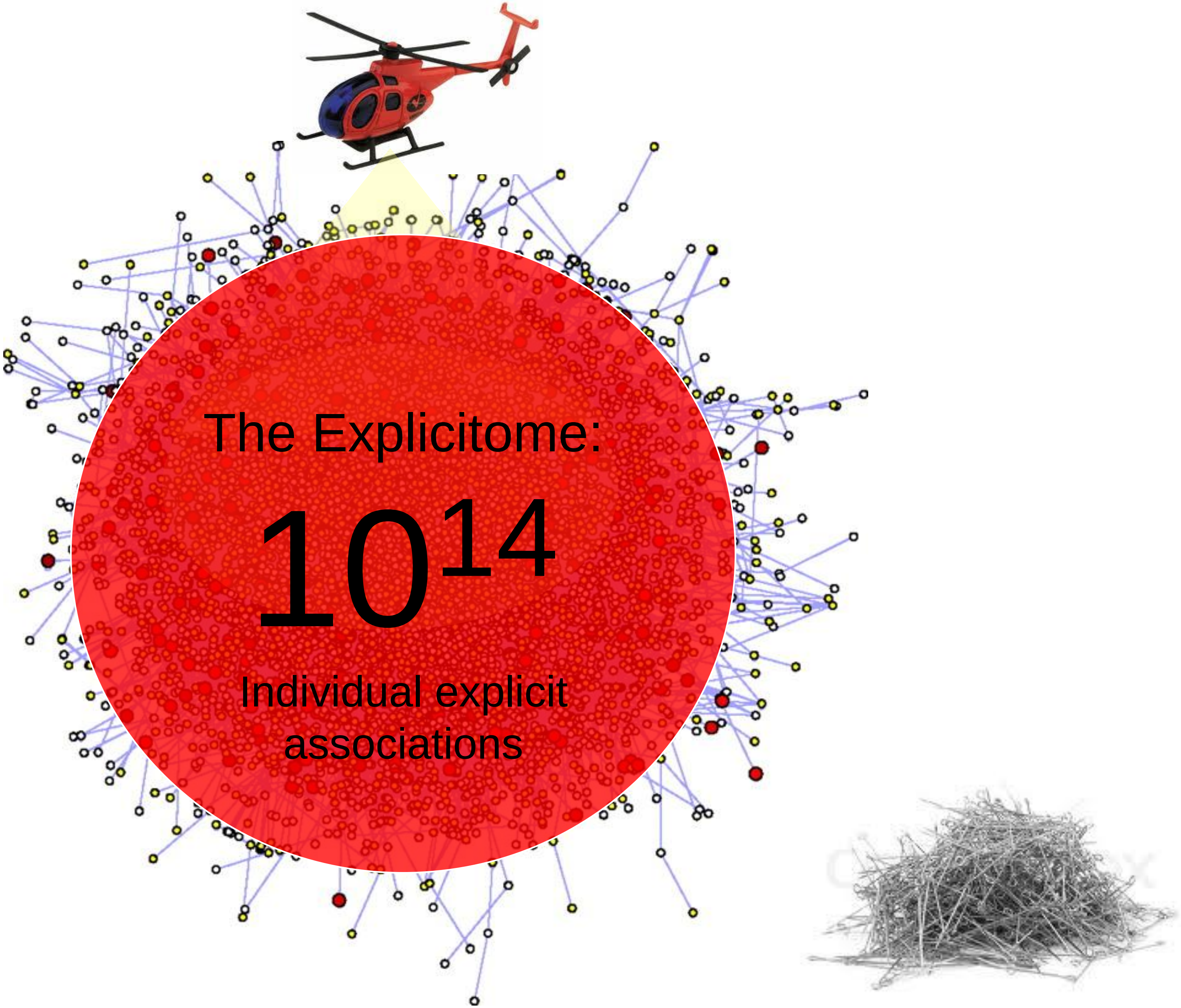
The Goal is Knowledge Discovery, not Data Collection



AREAL SURVEY

DEEP EXCAVATION

Pattern Recognition in Open Data and detailed Excavation should be separated



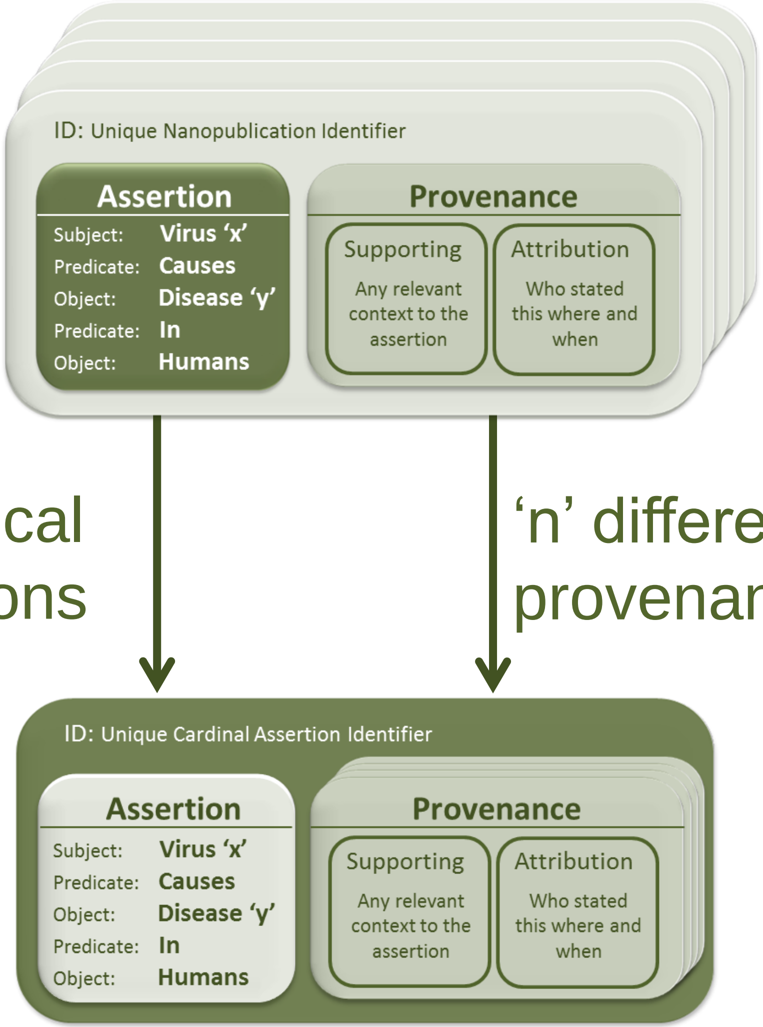
The Explicitome:

10^{14}

Individual explicit associations

How do we discover patterns in ‘Ridiculograms’?

The Semantic Web approach to interoperability



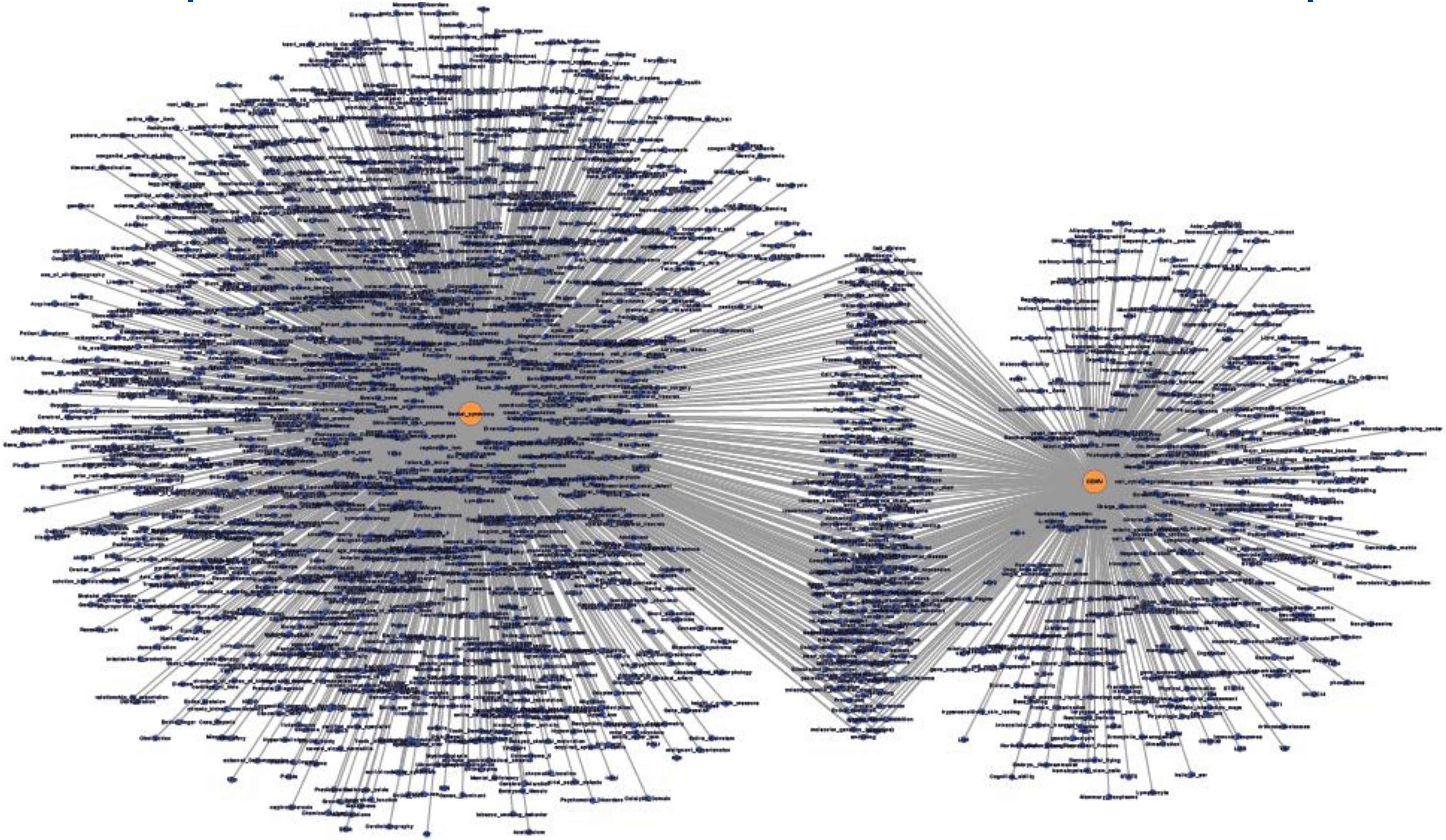
n identical assertions

‘n’ different provenances

Cardinal Assertion

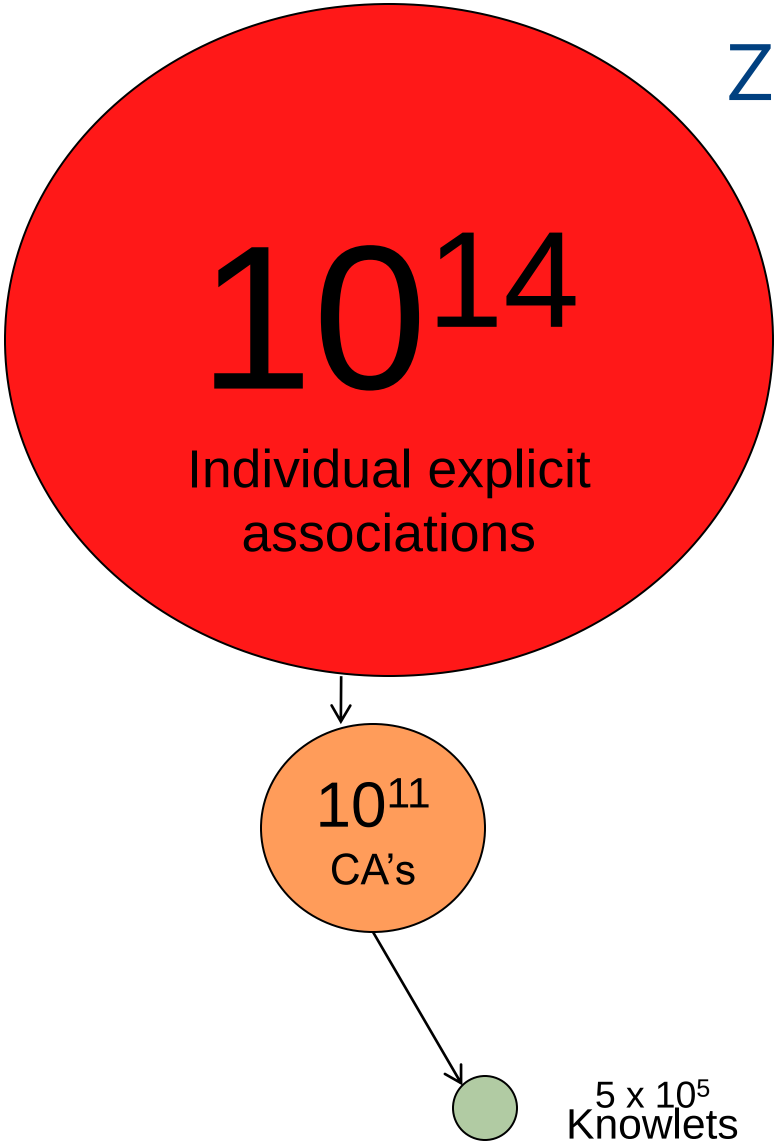
The Unique Explicitome: 10^{11} Cardinal Assertions

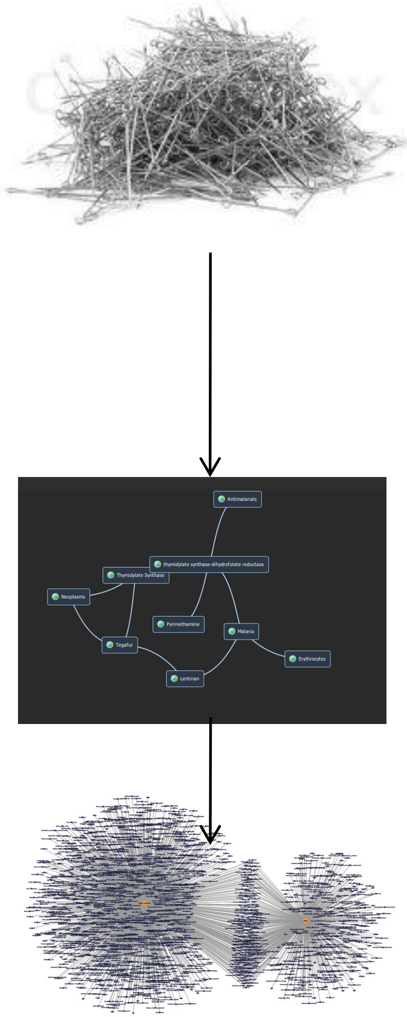
We publish about less than a million LS concepts



10^6 concept clusters (Knowlets)

Zipping the Explicitome



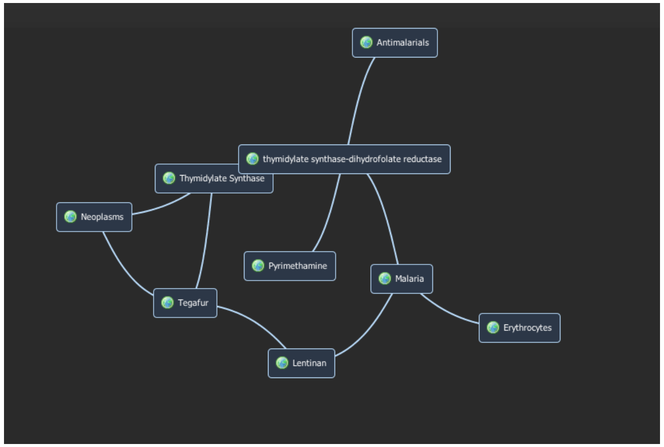


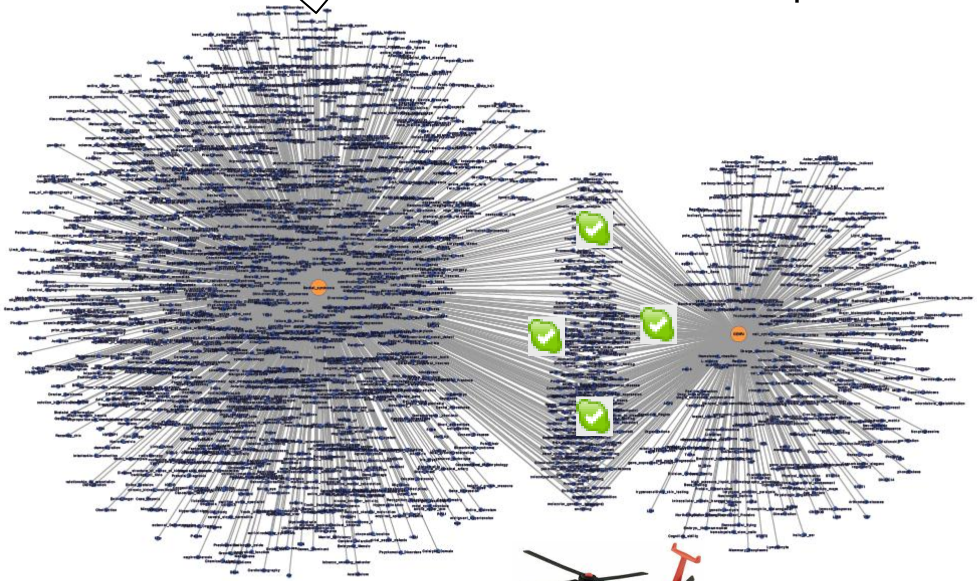
≈99.999996% reduction of infoburden

In silico knowledge discovery for the millions..

experimentation

Enrichment of the explicitome



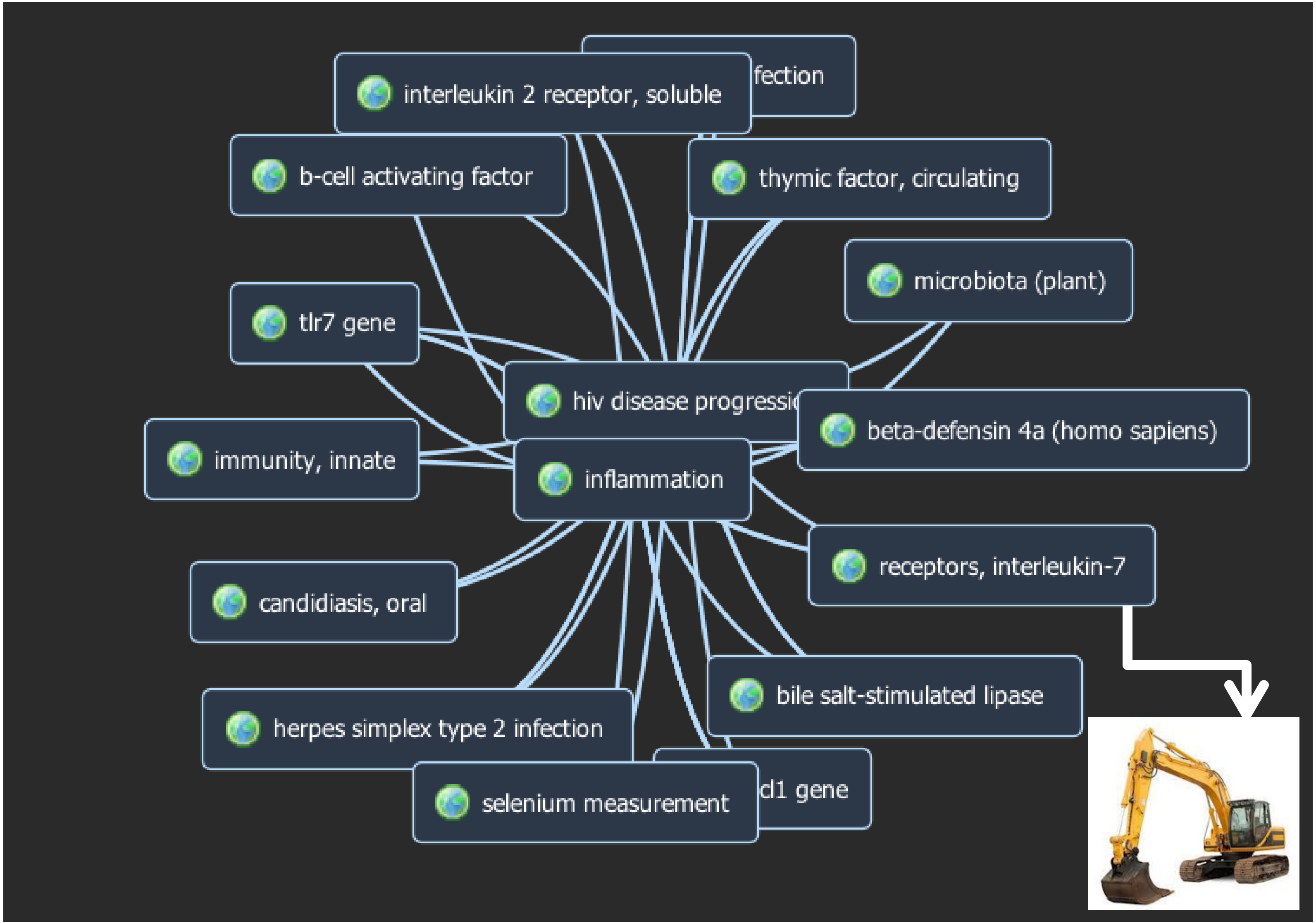


In cerebro rationalisation
And confirmational reading

In silico hypothesis generation

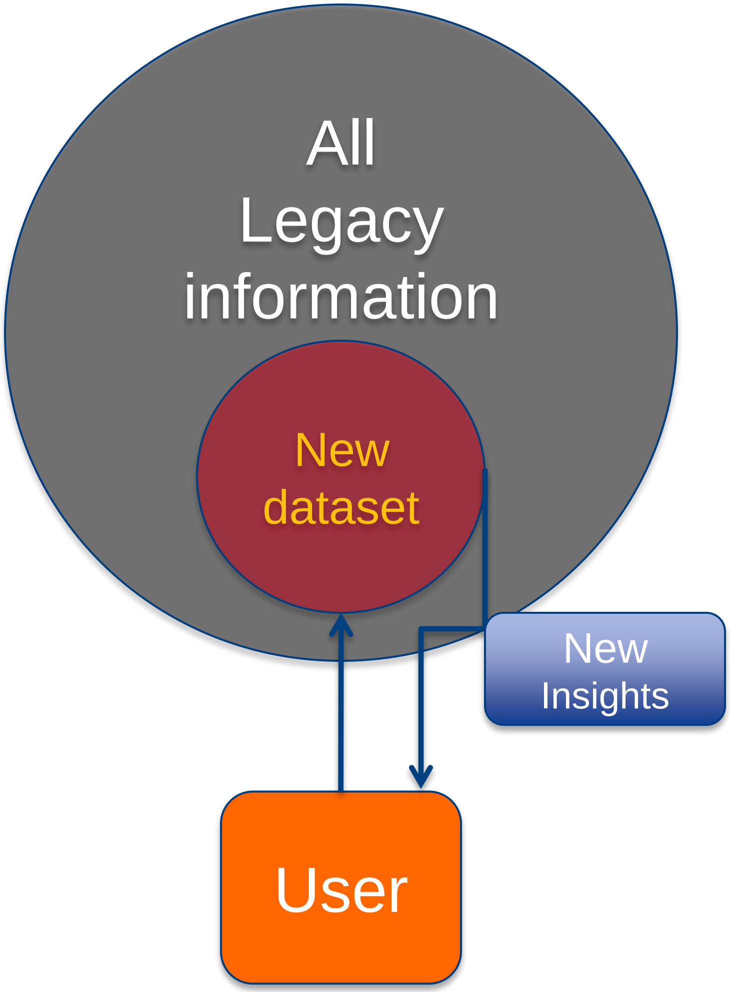
Reasoning takes place on aggregated and zipped data

The Implicitome



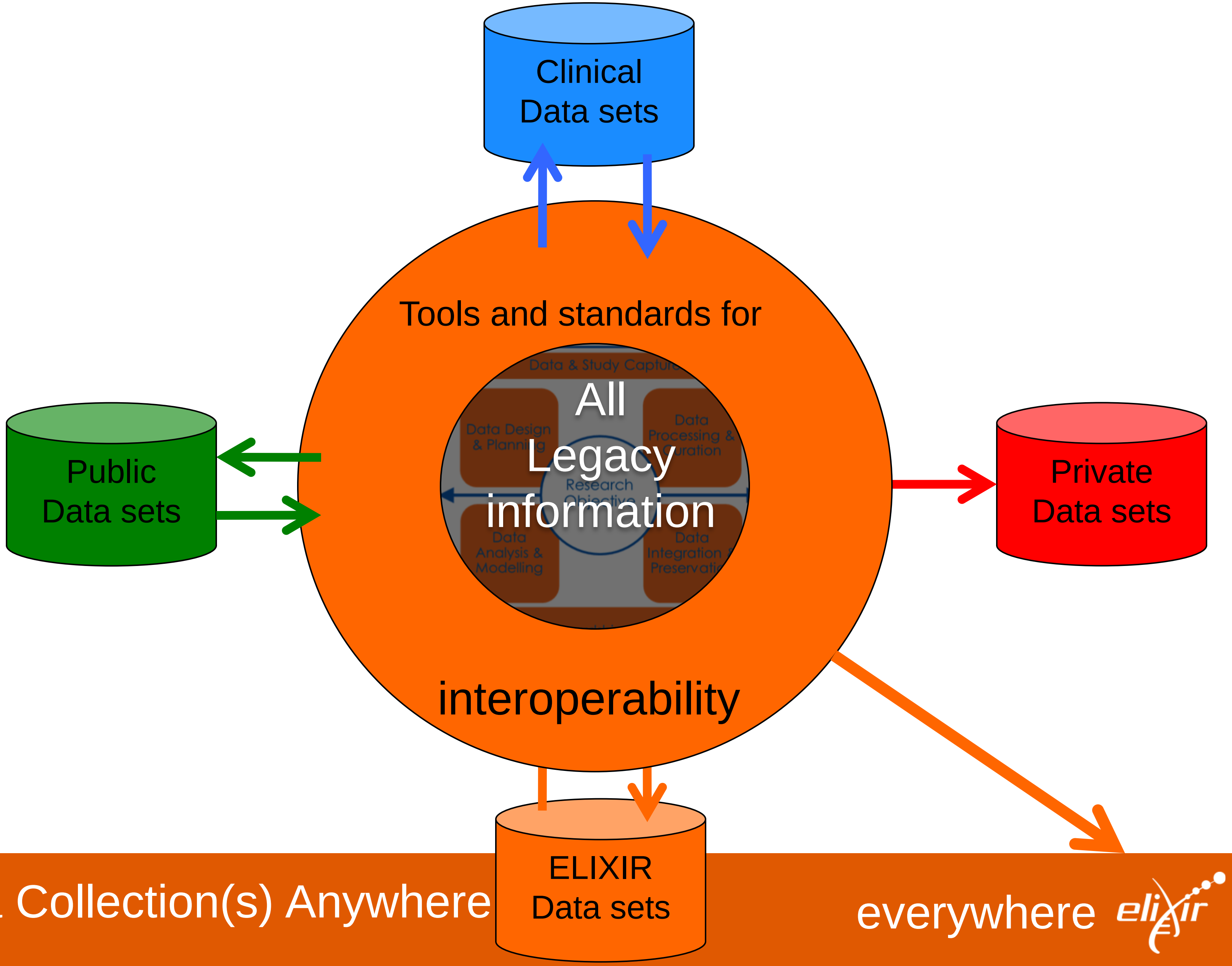
> 1 M hypotheses hidden in the implicitome

eScience & ELIXIR



The Goal is Knowledge Discovery, not Data Collection

The Role of ELIXIR: Open versus Managed



Clinical Data sets

Tools and standards for

All Legacy information

interoperability

Public Data sets

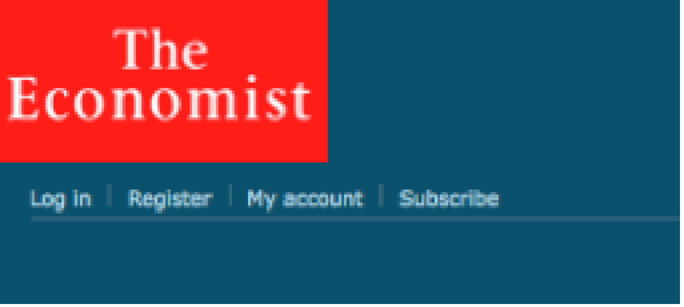
Private Data sets

ELIXIR Data sets

Data Collection(s) Anywhere

everywhere

Data & Study Capture
Data Design & Planning
Data Processing & Curation
Research Objective
Data Analysis & Modelling
Data Integration & Preservation
elixir


The Economist

Log in | Register | My account | Subscribe

Not only 'hardware'

For Big Data to become huge, however, there are still hurdles to leap. For one thing, the tools to analyse data are not yet good enough. And people with the skills to analyse data are scarce and will become scarcer. By 2018 there will be a "talent gap" of between 140,000 and 190,000 people,

Only three things count: Experts Experts Experts

Vision (personal, not necessarily ELIXIR)

- Data Collections are not a goal in themselves
- They ultimately serve **Knowledge Discovery**
- E-datastewardship in a time of plenty is thus also **data zipping**
- E-datastewardship should address the **entire data cycle**
- ELIXIR is more about a trusted partner for the **'Tools&Rules**' for data stewardship and interoperability than about data archiving.
- Interoperability is for **both** humans and computers.
- **Open data needs to be talking to 'closed data'**
- **Data experts** need the place in the egosystem they deserve

SHARED knowledge

is

Double Knowledge