Fundamentals of Data Infrastructures

Dublin, March 2014

Welcome & Introduction
Adam Carter
EPCC, The University of Edinburgh
Training Coordinator, EUDAT

EUDAT

Timetable

09:00	Registration & Coffee
09:15	Welcome & Introduction - Adam Carter
09:30	Data Discovery - Introduction: Why (benefits of reusing data), How EUDAT's services help with this (in general) - Adam Carter
09:45	PIDs - Adam Carter
10:05	Metadata and Ontologies - Adam Carter
10:30	B2SHARE Nordic - An example of a service that facilitates Data Discovery and uses PIDs and Metadata - Teemu Kemppainen
11:00	Coffee
11:30	Legal Issues in Research Data Collection &Sharing - Pawel Kamocki
12:50	Wrap Up - Adam Carter
13:00	Close & Lunch courtesy of EUDAT

EUDAT

What do we mean by Data Infrastructures?

- The services, and underlying hardware that allow researchers and data managers to:
  - Store Data
  - Move Data
  - Organise Data
  - Find Data
  - Share Data

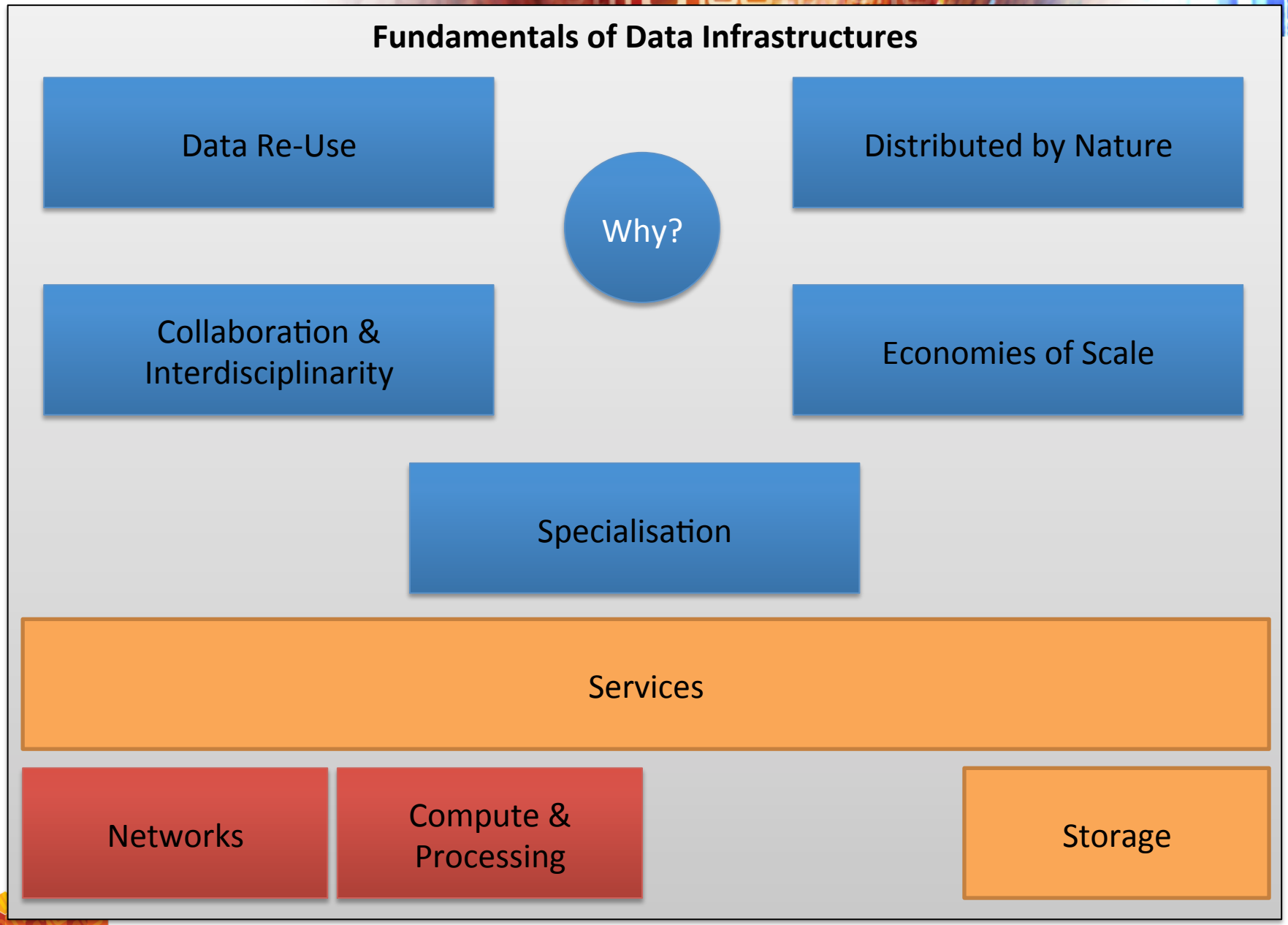
EUDAT

Why Data Infrastructures?

- You're all working with data
  - Yours, other people's
- Data has value
- Mechanisms which support data re-use add value to that data
- They can help with big data, and small…
- They can support new ways of doing science
  - Data Science / Fourth Paradigm / Datascopes

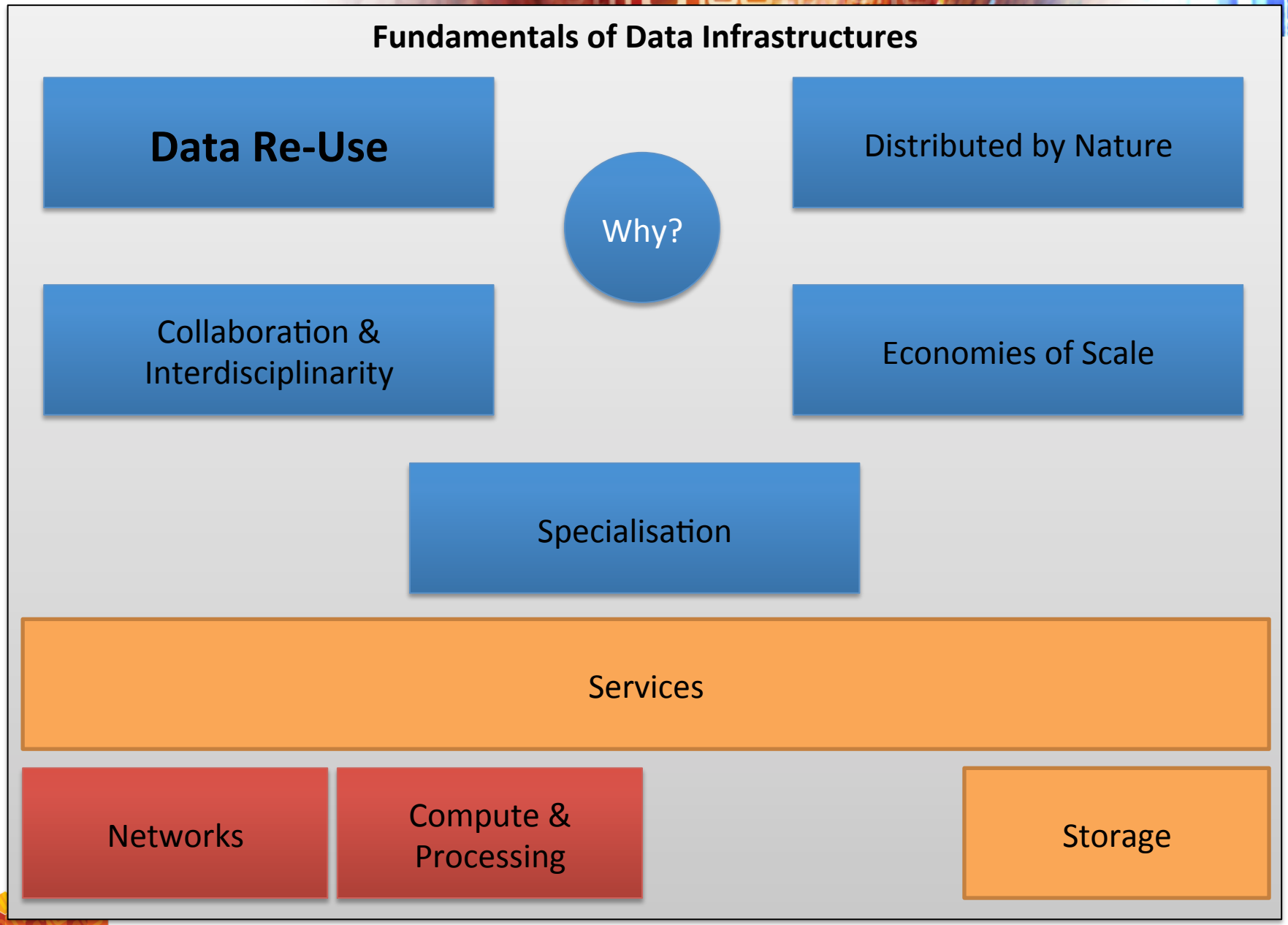
EUDAT

Fundamentals of Data Infrastructures



EUDAT

Fundamentals of Data Infrastructures



EUDAT

Fundamentals of Data Infrastructures

PIDs

Metadata

Moving Data

Processing Data

Authentication & Authorisation

Licensing & Data Ownership

Data Management Planning

Data Curation & Preservation

EUDAT

Fundamentals of Data Infrastructures

PIDs

Metadata

Moving Data

Processing Data

Authentication & Authorisation

Licensing & Data Ownership

Data Management Planning

Data Curation & Preservation

EUDAT

PIDs

- The idea is that it's useful to **identify** your data with some *unique*, *persistent* label (which, in general, is *not* its location, but which can be used to *discover* its location)
  - e.g.: **10.1045/march99-bunker**
- This allows the data to be referred to unambiguously
  - in publications, by machines & services, …

EUDAT

Metadata

- Data about data
  - e.g. Description of what it relates to, where and when it was collected, how it is formatted, …
- Helps with:
  - Finding data (search on the metadata)
  - Understanding data
- There are standard formats for metadata and standard mechanisms for sharing it

EUDAT

Moving Data

- Data Staging
  - Getting data to where you need to use it (often a compute resource, but also, e.g., a data infrastructure)
- Replication of Data
  - Automated mechanisms for copying data, usually between geographically remote sites for redundancy and/or data locality (faster access)
- Dataflows
  - Automated workflows based on a flow of data

EUDAT

Processing Data

- Such a big area, that it's a subject in its own right
  - We don't discuss computing, HPC, HTC, Grid, Cloud, etc. here
- But there are overlaps with data infrastructures, e.g.,
  - Bringing compute to the data
  - Workflows
  - Data-intensive compute architectures

Authentication & Authorisation

- Authentication: Who you are
- Authorisation: What you can do
- Necessary prerequisites for ***trust***
- Possibly contrary to first impressions:
  - They *help* with sharing of data
- Mechanisms to **describe** who should have access to what, and what they can do with it, and to **implement** these access restrictions

EUDAT

Licensing & Ownership

- Rights *and* responsibilities
- Copyright
- Ownership of data
  - Who decides what can be done with the data, and by whom?
  - Who has responsibility for looking after the data (and ensuring that it can be re-used in the future)

EUDAT

Data Curation & Preservation

- Also a huge field in its own right
- Overlaps with data infrastructures:
  - Providing services and hardware to those with responsibility for looking after others' data and allowing the *data curator* or *data owner* to do what they need to do to look after the data
- There's often a shared responsibility, e.g., a data centre will look after the bits on disk, and a library will look after making sure they can be interpreted by future generations

EUDAT

Data Management & Planning

- This is about good research practice
- It's the process you go through to ensure that your data can be re-used in the future
- This supports:
  - Reproducible science
  - Open, verifiable results
  - Re-use of data by you, your colleagues, and the wider research community
- It's about defining things like ownership

EUDAT

Timetable

09:00	Registration & Coffee
09:15	Welcome & Introduction - Adam Carter
09:30	Data Discovery - Introduction: Why (benefits of reusing data), How EUDAT's services help with this (in general) - Adam Carter
09:45	PIDs - Adam Carter
10:05	Metadata and Ontologies - Adam Carter
10:30	B2SHARE Nordic - An example of a service that facilitates Data Discovery and uses PIDs and Metadata - Teemu Kemppainen
11:00	Coffee
11:30	Data access rights and mechanisms, open, sensitive and confidential data, and copyrights and licensing issues - Pawel% Kamocki
12:50	Wrap Up - Adam Carter
13:00	Close & Lunch courtesy of EUDAT

EUDAT

EUDAT