DISCOVER THE OCEAN. UNDERSTAND THE PLANET.

OCEAN NETWORKS CANADA INNOVATION

ACQUISITION, PRESERVATION AND DISTRIBUTION OF COMPLEX SCIENTIFIC DATA

Research Data Canada Webinar

Benoît Pirenne

November 26, 2013

AN INITIATIVE OF University of Victoria

DISCOVER THE OCEAN. UNDERSTAND THE PLANET.

[OR: SCIENCE DATA MANAGEMENT]

OCEAN NETWORKS CANADA INNOVATION

Research Data Canada Webinar

Benoît Pirenne

November 26, 2013

AN INITIATIVE OF University of Victoria

DISCOVER THE OCEAN. UNDERSTAND THE PLANET.

OCEAN NETWORKS CANADA INNOVATION

[OR: TOWARDS DATA STEWARDSHIP FACILITIES]

Research Data Canada Webinar

Benoît Pirenne

November 26, 2013

AN INITIATIVE OF University of Victoria

WHAT IS DATA MANAGEMENT?

- It’s an integrated *process*
  - takes data from acquisition to distribution
  - applies optional transforms (e.g., calibration) along the way
  - associates data with its complete description (**metadata**), and possibly later with related **publications**
  - is auditable, repeatable, quality-controlled
  - includes **hardware**, **software**, **processes** and **people**
  - **evolves with practice and technology**

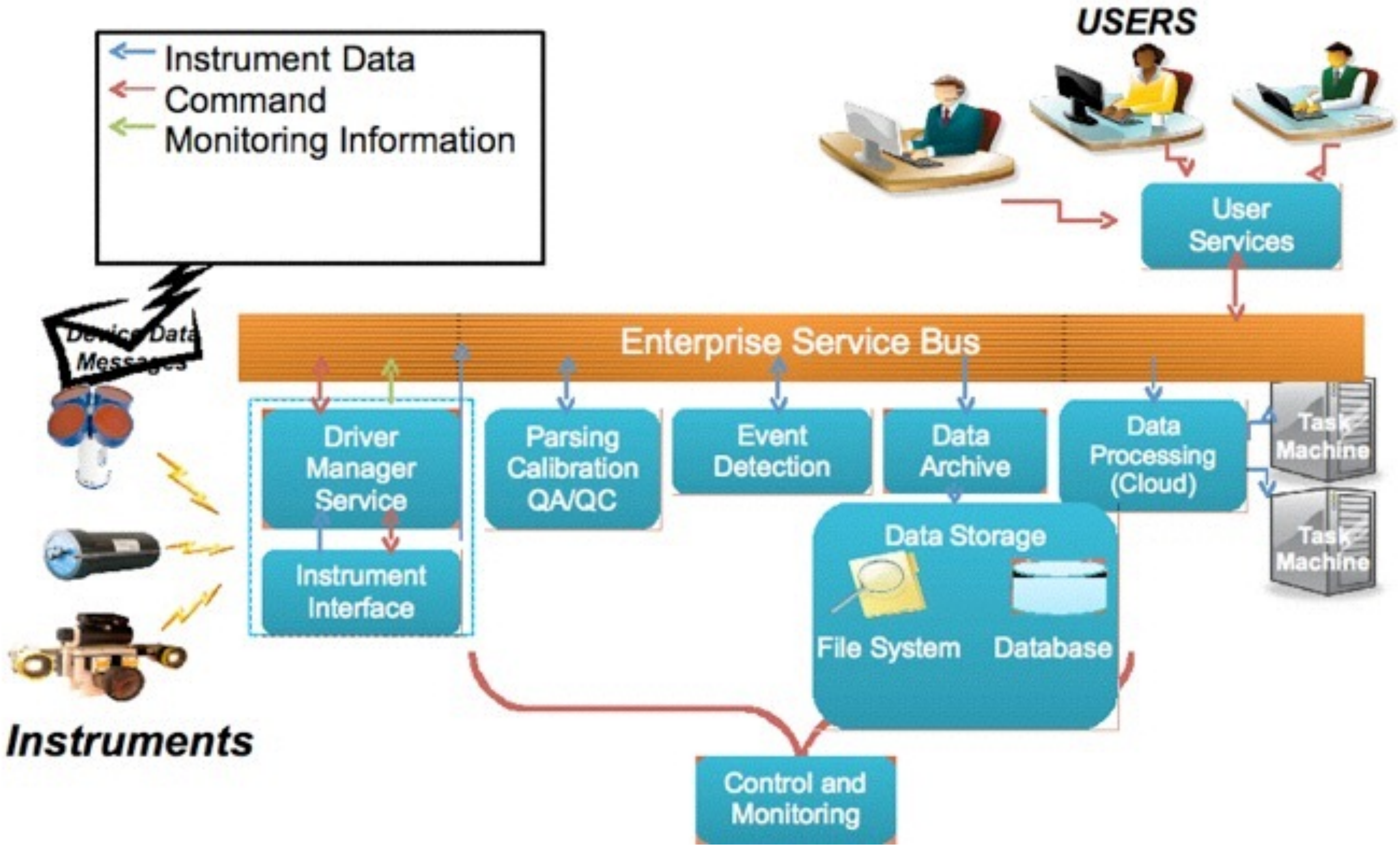
WHAT IS DATA MANAGEMENT?

- The process applies to any type of data/discipline
  - images from astronomy detector
  - current velocity vectors in the ocean
  - audio recordings from first nation elders
  - scans from old manuscript
  - lab test results
  - …

DISCOVER THE OCEAN. UNDERSTAND THE PLANET.

WHAT IS DATA MANAGEMENT?

OCEAN NETWORKS CANADA INNOVATION



WHY DATA MANAGEMENT?

- Science research equipment and programmes are costly to setup and/or operate and therefore data must be **re-used** and **shared** with many other users
- There is a potential for **new insight to emerge from a re-use** of the data
- The project build time is significant and its operational lifetime unclear (e.g., space experiment, Arctic exploration, …)
- Observations/findings are **unique** and cannot easily be reproduced (e.g., observation of poorly known, possibly transient phenomena)

WHY DATA MANAGEMENT?

- The science team behind the data is large and internationally distributed
- The science requires long time series
- The opportunity to recalibrate later and improve data quality as sources are better understood
- Need to optimize and audit the use of resources
- Need to support outreach and education efforts

HOW CAN WE AFFORD DM?

- Data Management *is* affordable
  - Experience shows that across disciplines, the average cost to set up a DM is ~10% of the costs of the projects it supports
  - Experience shows that the burden of operating a DM is about 10% of the overall projects operating costs
  - DM costs fall down further when projects are no longer operational

CHALLENGES OF DM

- Hardware:
  - not much of an issue. Really. (thanks, Gordon Moore!)
  - Instruments and experiments producing highest data rates usually not ready before relevant information technology catches up
  - *As long as we maintain the funding formula!*

CHALLENGES OF DM

- Hardware:

[LHC's 25PB/yr]: *"Storing the data is not a problem: hard drives are cheap and getting cheaper. The challenge is preserving knowledge that is less commonly stored — the software, algorithms and reference plots specific to each experiment. These often degrade or disappear with time", says Cristinel Diaconu* (nature.com Nov. 26, 2013)!

- *As long as we maintain the funding formula!*

CHALLENGES OF DM

- ❖ Data description (metadata)
  - Requires having dedicated staff with the memory of assets and holdings, which causes a significant cost increase --- or you do it right and make it part of the design!
  - Essential for, and part of, data quality assessment
  - Includes calibration, annotations, space-time info, ..., ownership, access authorizations, ...

CHALLENGES OF DM

- Data access
  - Search through data (not always possible), search through metadata
  - Metadata encoding and transport standards needed
  - Data formats are discipline-specific
  - Uniform, interoperable access is a huge challenge (e.g., VO)

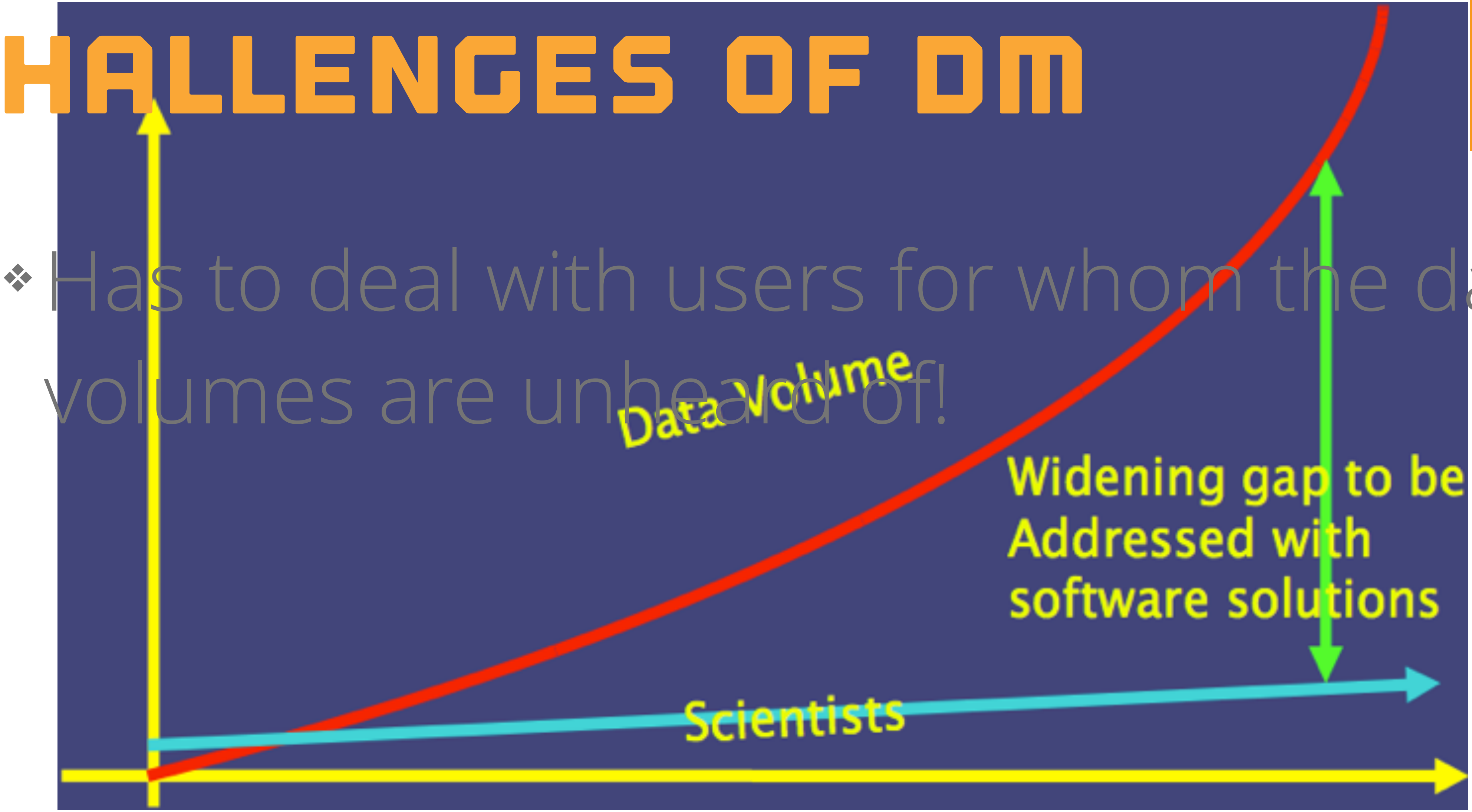
CHALLENGES OF DM

- Convince PIs and funding agencies that good Data Management is important.
  - But this battle is by now almost won. (NSF, TC3+, ...)

OCEAN NETWORKS CANADA INNOVATION

CHALLENGES OF DM

- ❖ Has to deal with users for whom the data volumes are unheard of!



TOWARDS DATA STEWARDSHIP FACILITIES

- At the service of many projects in a related disciplines
- Provides long-term data storage, access and stewardship, well beyond the lifetime of individual projects
- Need is particularly acute for small projects
- Avoid the creation of many ad-hoc systems that can’t be maintained long-term
- Economies of scale, discipline-specific expertise

DSF FOR USERS

- ❖ Are a one-stop-shop for data in a given discipline, and a portal to international resources
- ❖ Allow scientists to focus on science, not on data management
- ❖ Ensure stewardship of data beyond project funding
- ❖ Ensure data will remain citable

DSF FOR FUNDING AGENCIES

- Ability make economies of scale
- DSF gather expertise in data management *and* in the science disciplines
- DSF have the wherewithal to remain at the leading edge of technology
- Adoption easier with users used to entrust their most precious data to "the Cloud", and work using remote compute resources
- With similar international peers, have a voice at the interoperability and standards table

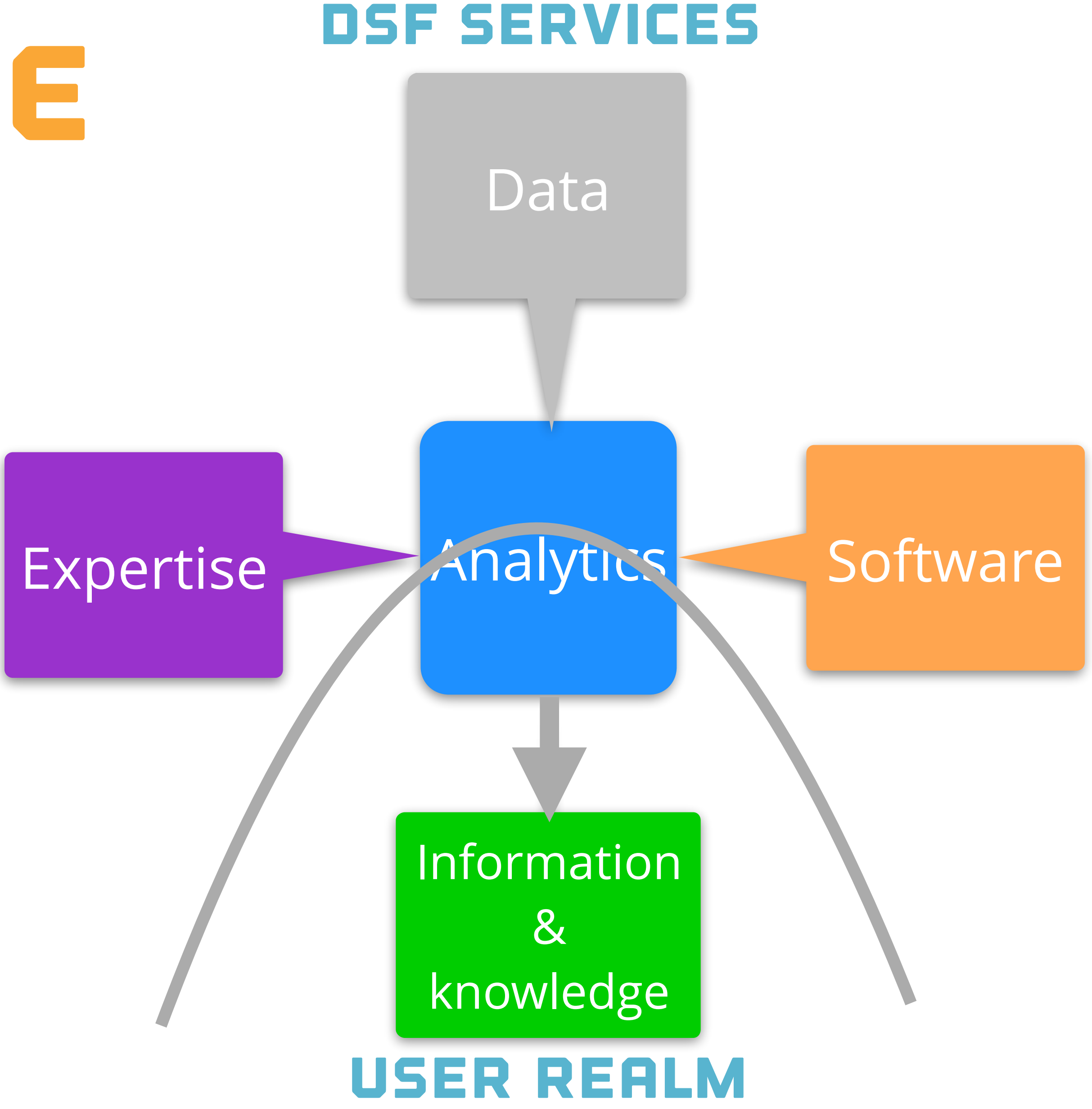
CHALLENGES FOR DSF

In progress: use of clouds increasing

In progress: more and more open data policies around

- Need buy-in from PIs regardi
  - Development of trust w ntities managing their data
  - The definition of a(n open) data policy, sharing of data
  - Being thorough with data/experimentation *description* (Metadata)
  - Realizing that data management is not achieved with a bit of hardware and software

DSF ROLE



DSF SERVICES

Data

Expertise

Analytics

Software

Information & knowledge

USER REALM

DSF STRUCTURE EXAMPLE

SOFTWARE DEV & QC

- Polyvalent IT staff

SYSTEMS & OPS

- On call
- Network, systems management
- Redundancy management
- Support software management

DATA STEWARDSHIP

- Data QA/QC
- Data annotation
- Overall metadata quality
- User support

CANADIAN DSF EXAMPLES

- Canadian Astronomy Data Centre (CADC) is a great example of discipline specific Data Stewardship Facility
- Canadian Polar Data Network (CPDN) — includes multi-disciplinary data
- Canadian Research Data Centre Network (CRDCN) (social and population health statistics)
- …

THANK YOU