ICFA Data Lifecycle Panel: Introduction

ICFA Data Lifecycle Panel meeting - April 15, 2024

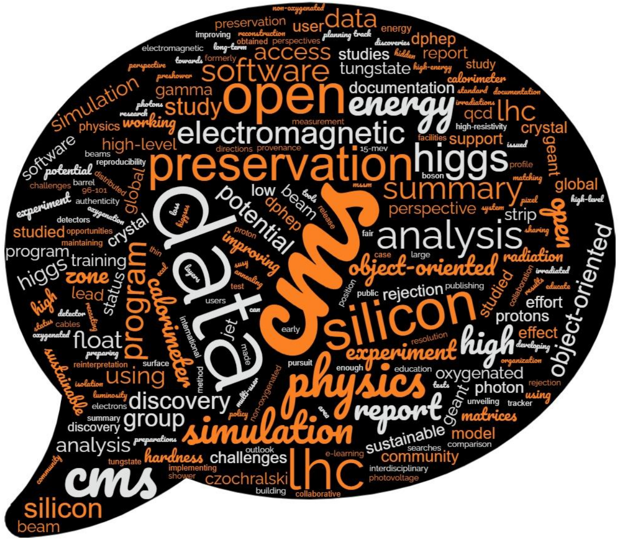
Kati Lassila-Perini
Helsinki Institute of Physics - Finland
CMS Data preservation and open access coordinator

1

Who am I?

My background and reasons for agreeing to chair the panel.

Kati Lassila-Perini



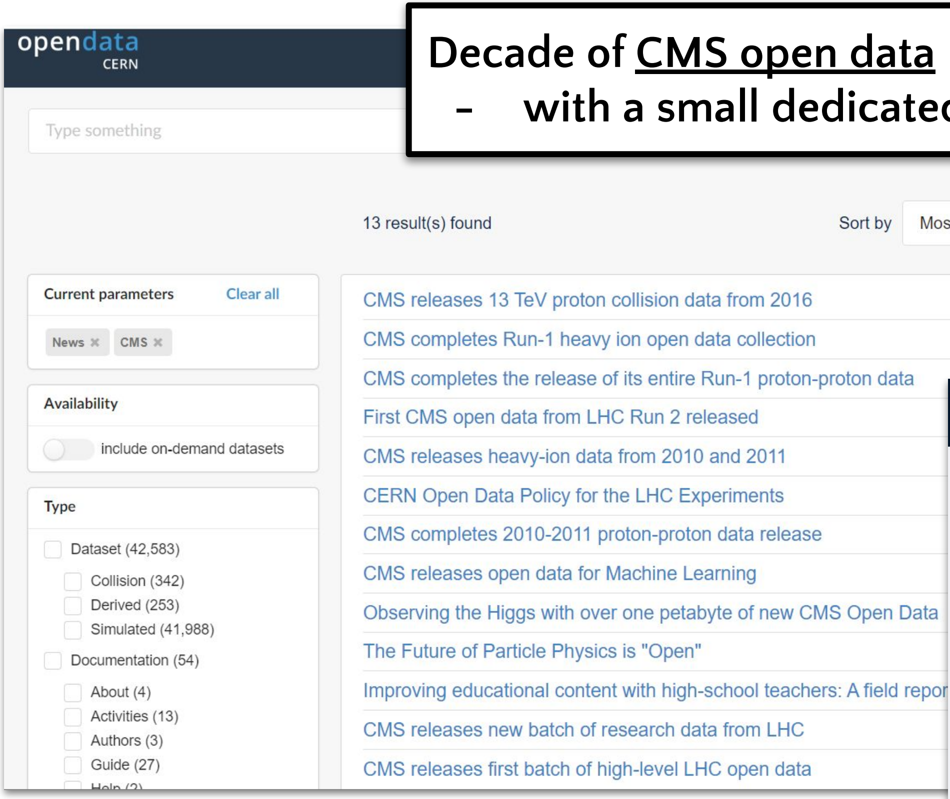
Experimental particle physicist, Ph.D

Research coordinator at Helsinki Institute of Physics (HIP)

CMS data preservation and open access (DPOA) coordinator

kati.lassila-perini @ cern.ch

Decade of CMS open data – with a small dedicated team



13 result(s) found

Sort by Most recent

Current parameters Clear all

News ✖ CMS ✖

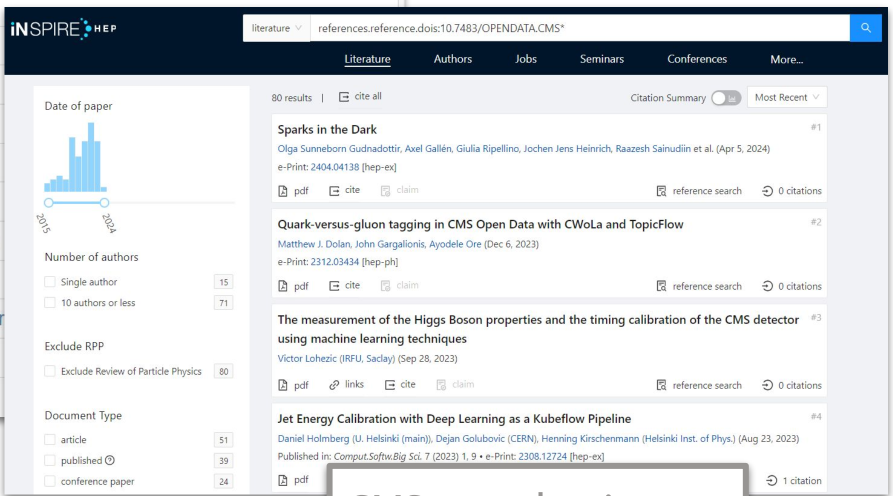
Availability

include on-demand datasets

Type

- □ Dataset (42,583)
  - □ Collision (342)
  - □ Derived (253)
  - □ Simulated (41,988)
- □ Documentation (54)
  - □ About (4)
  - □ Activities (13)
  - □ Authors (3)
  - □ Guide (27)
  - □ Help (2)

- CMS releases 13 TeV proton collision data from 2016
- CMS completes Run-1 heavy ion open data collection
- CMS completes the release of its entire Run-1 proton-proton data
- First CMS open data from LHC Run 2 released
- CMS releases heavy-ion data from 2010 and 2011
- CERN Open Data Policy for the LHC Experiments
- CMS completes 2010-2011 proton-proton data release
- CMS releases open data for Machine Learning
- Observing the Higgs with over one petabyte of new CMS Open Data
- The Future of Particle Physics is "Open"
- Improving educational content with high-school teachers: A field repor
- CMS releases new batch of research data from LHC
- CMS releases first batch of high-level LHC open data



literature references.reference.dois:10.7483/OPENDATA.CMS*

80 results | cite all

Citation Summary Most Recent

Date of paper

Number of authors

□ Single author	15
□ 10 authors or less	71

Exclude RPP

□ Exclude Review of Particle Physics	80

Document Type

□ article	51
□ published	39
□ conference paper	24

Sparks in the Dark #1
Olga Sunneborn Gudnadottir, Axel Gallén, Giulia Ripellino, Jochen Jens Heinrich, Raazesh Sainudiin et al. (Apr 5, 2024)
e-Print: 2404.04138 [hep-ex]
pdf cite claim reference search 0 citations

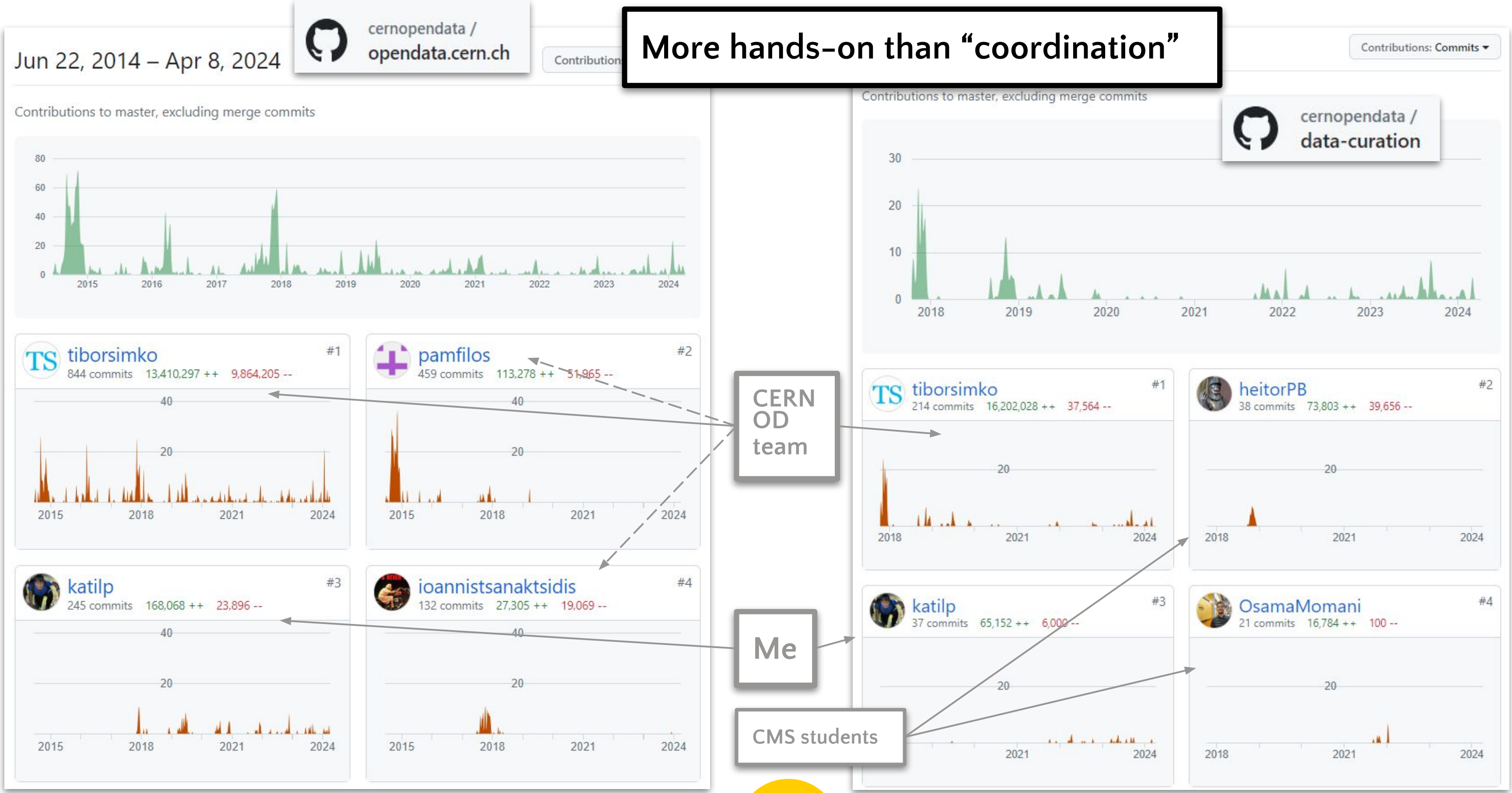
Quark-versus-gluon tagging in CMS Open Data with CWoLa and TopicFlow #2
Matthew J. Dolan, John Gargalionis, Ayodele Ore (Dec 6, 2023)
e-Print: 2312.03434 [hep-ph]
pdf cite claim reference search 0 citations

The measurement of the Higgs Boson properties and the timing calibration of the CMS detector using machine learning techniques #3
Victor Lohezic (IRFU, Saclay) (Sep 28, 2023)
pdf links cite claim reference search 0 citations

Jet Energy Calibration with Deep Learning as a Kubeflow Pipeline #4
Daniel Holmberg (U. Helsinki (main)), Dejan Golubovic (CERN), Henning Kirschenmann (Helsinki Inst. of Phys.) (Aug 23, 2023)
Published in: *Comput.Softw.Big Sci.* 7 (2023) 1, 9 • e-Print: 2308.12724 [hep-ex]
pdf 1 citation

CMS open data in use

More hands-on than “coordination”



cernopendata / opendata.cern.ch

Jun 22, 2014 – Apr 8, 2024

Contribution

Contributions to master, excluding merge commits

tiborsimko #1
844 commits 13,410,297 ++ 9,864,205 --

pamfilos #2
459 commits 113,278 ++ 51,965 --

katilp #3
245 commits 168,068 ++ 23,896 --

ioannistsanaktsidis #4
132 commits 27,305 ++ 19,069 --

cernopendata / data-curation

Contributions: Commits

Contributions to master, excluding merge commits

tiborsimko #1
214 commits 16,202,028 ++ 37,564 --

heitorPB #2
38 commits 73,803 ++ 39,656 --

katilp #3
37 commits 65,152 ++ 6,000 --

OsamaMomani #4
21 commits 16,784 ++ 100 --

CERN OD team

Me

CMS students

“CV”

Are CMS open data FAIR?

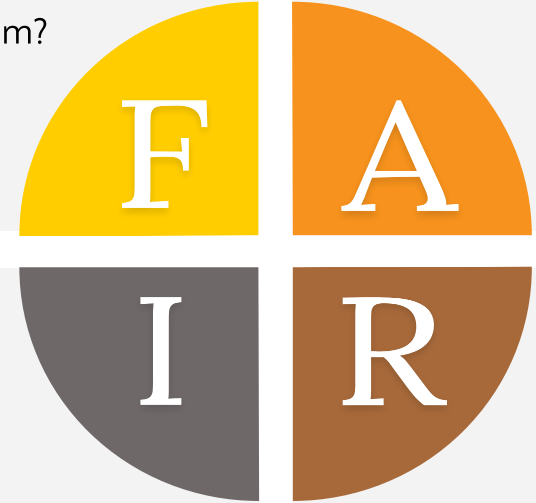
FINDABLE

Do you know where to look for them?

Can you find what you need?

ACCESSIBLE

Can you download them?



Are they in some common format?

Do you have the tools to open the data files?

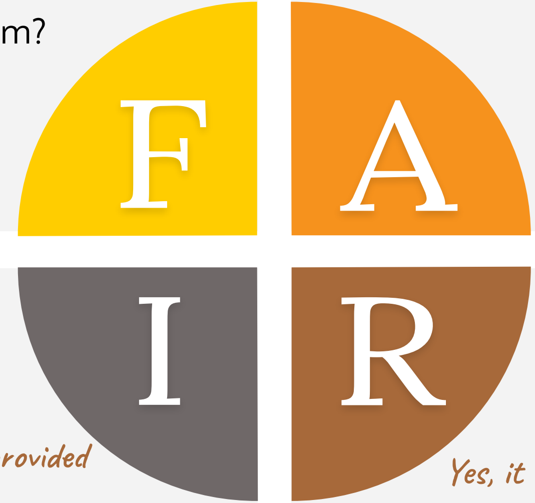
INTEROPERABLE

Do you know how to use?

Can you make new research with them?

REUSABLE

FAIR? Yes...



FINDABLE

Do you know where to look for them?

Yes, CERN open data portal

Can you find what you need?

Yes, there are search functions

ACCESSIBLE

Can you download them?

Yes, or they can be streamed with XRootD

Are they in some common format?

Partly yes, partly no...

Do you have the tools to open the data files? *Yes or no, if no, they are provided*

INTEROPERABLE

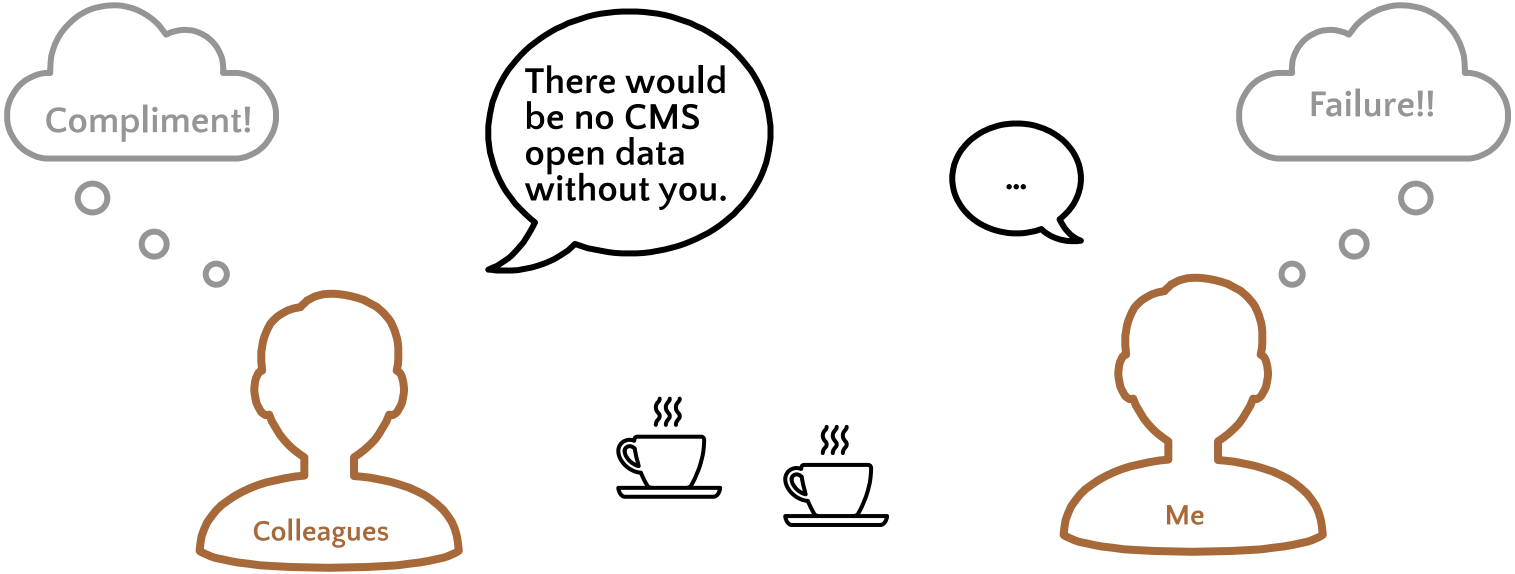
Do you know how to use?

There are instructions to get started...

Can you make new research with them?

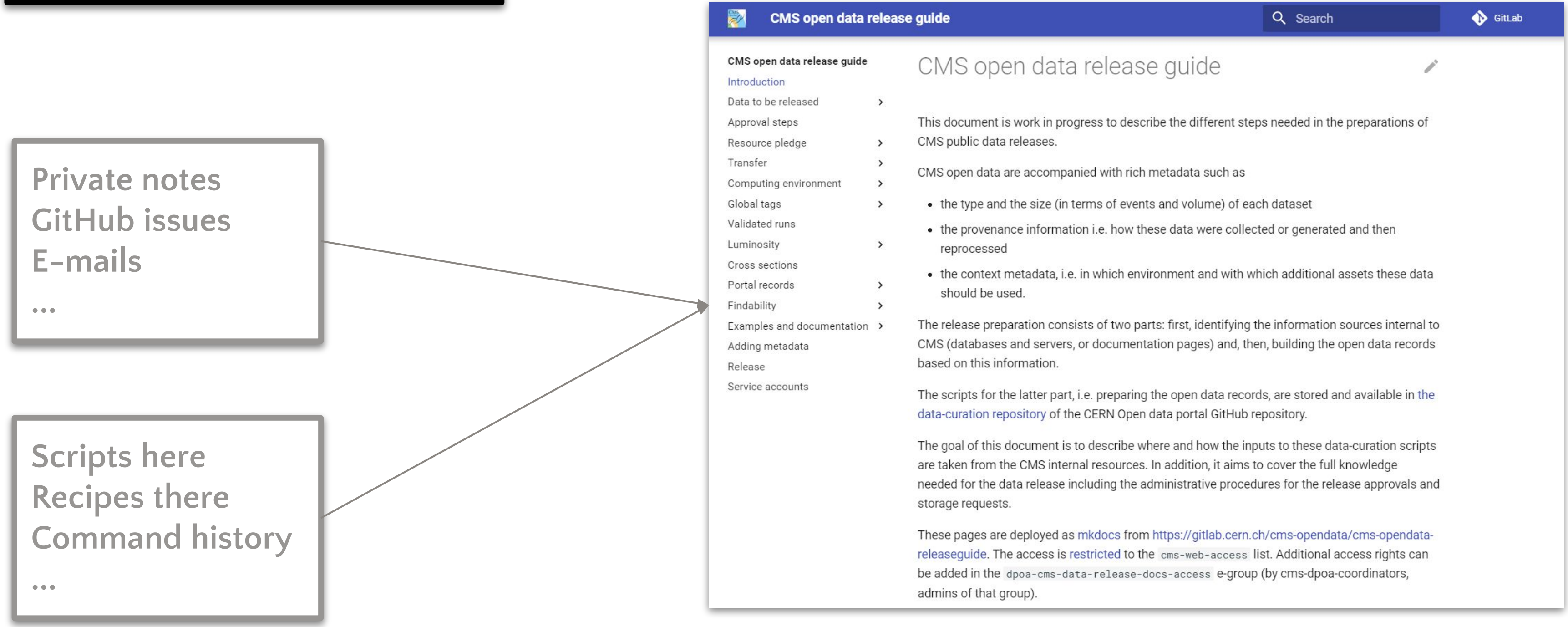
Yes, it takes at least as much as for the CMS people

REUSABLE



...but, some years ago...

To make the process FAIR…



Private notes
GitHub issues
E-mails
…

Scripts here
Recipes there
Command history
…

CMS open data release guide

This document is work in progress to describe the different steps needed in the preparations of CMS public data releases.

CMS open data are accompanied with rich metadata such as

- the type and the size (in terms of events and volume) of each dataset
- the provenance information i.e. how these data were collected or generated and then reprocessed
- the context metadata, i.e. in which environment and with which additional assets these data should be used.

The release preparation consists of two parts: first, identifying the information sources internal to CMS (databases and servers, or documentation pages) and, then, building the open data records based on this information.

The scripts for the latter part, i.e. preparing the open data records, are stored and available in the data-curation repository of the CERN Open data portal GitHub repository.

The goal of this document is to describe where and how the inputs to these data-curation scripts are taken from the CMS internal resources. In addition, it aims to cover the full knowledge needed for the data release including the administrative procedures for the release approvals and storage requests.

These pages are deployed as mkdocs from https://gitlab.cern.ch/cms-opendata/cms-opendata-releaseguide. The access is restricted to the `cms-web-access` list. Additional access rights can be added in the `dpoa-cms-data-release-docs-access` e-group (by cms-dpoa-coordinators, admins of that group).

… I needed to change how I work.

2

Data Lifecycle panel

The mandate is broad, even overwhelming…

My expertise is CMS/CERN/Open data/FAIR-centric

Diverse range of expertise within the panel – looking forward!

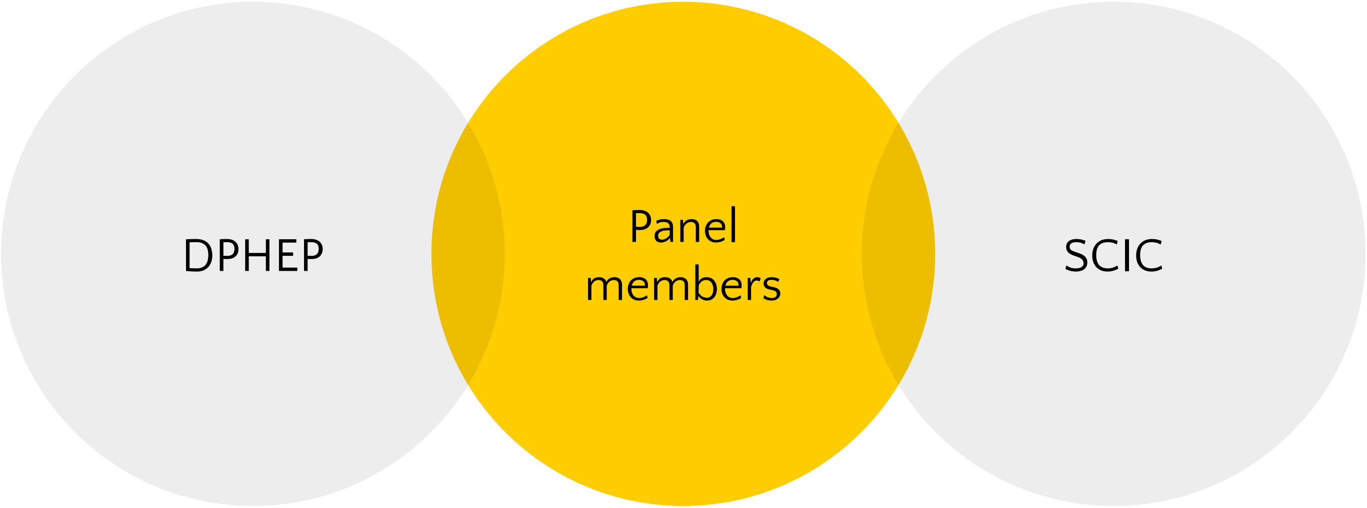
Practicalities

ICFA statement on the Data Lifecycle Panel

Mandate of the Data Lifecycle Panel

“

Rely on broad expertise



Practicalities

Frequency of the meetings?

- Monthly? Set a regular day and time?

Privacy of the agendas?

- Agendas public, but attendance and recordings members only?

Communications:

- Mailing list: ICFA-Data-Lifecycle-panel at cern.ch
- Collect input from the members through “surveys” for each meeting.
- Other channels?

Web site

- https://icfa.hep.net/icfa-panel-on-the-data-lifecycle/

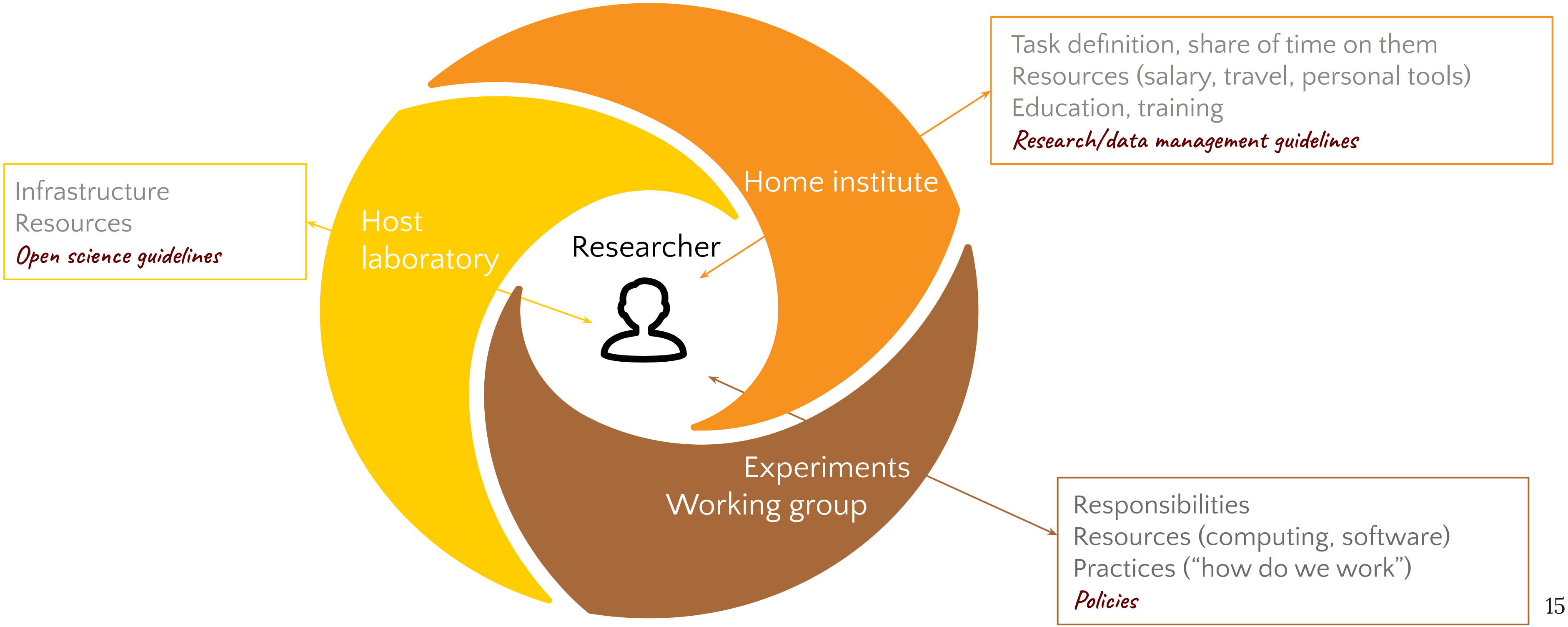
3

Stakeholders

Individuals researchers, within the collaborations, carry out the work.

The surrounding stakeholders may either empower or restrict the researchers' ability to adopt best practices for the full “data lifecycle".

Stakeholders



Infrastructure
Resources
Open science guidelines

Host laboratory

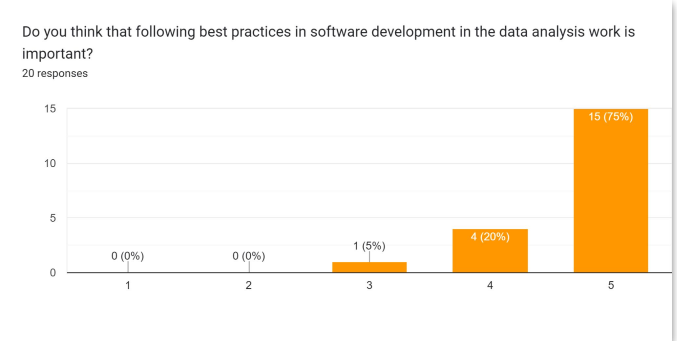
Researcher

Home institute

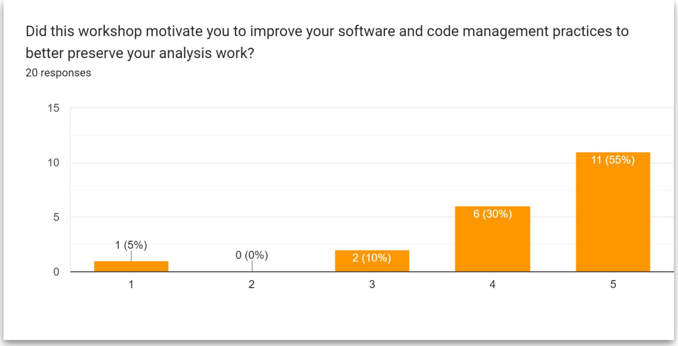
Task definition, share of time on them
Resources (salary, travel, personal tools)
Education, training
Research/data management guidelines

Experiments
Working group

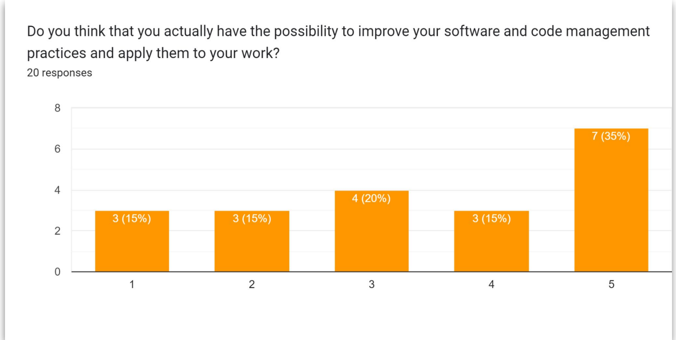
Responsibilities
Resources (computing, software)
Practices (“how do we work”)
Policies


Do you think that following best practices in software development in the data analysis work is important?

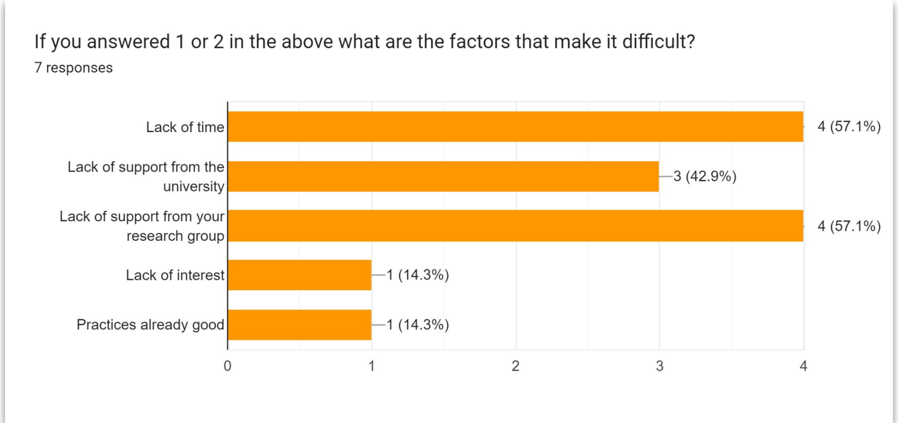
20 responses


Did this workshop motivate you to improve your software and code management practices to better preserve your analysis work?

20 responses


Do you think that you actually have the possibility to improve your software and code management practices and apply them to your work?

20 responses


If you answered 1 or 2 in the above what are the factors that make it difficult?

7 responses

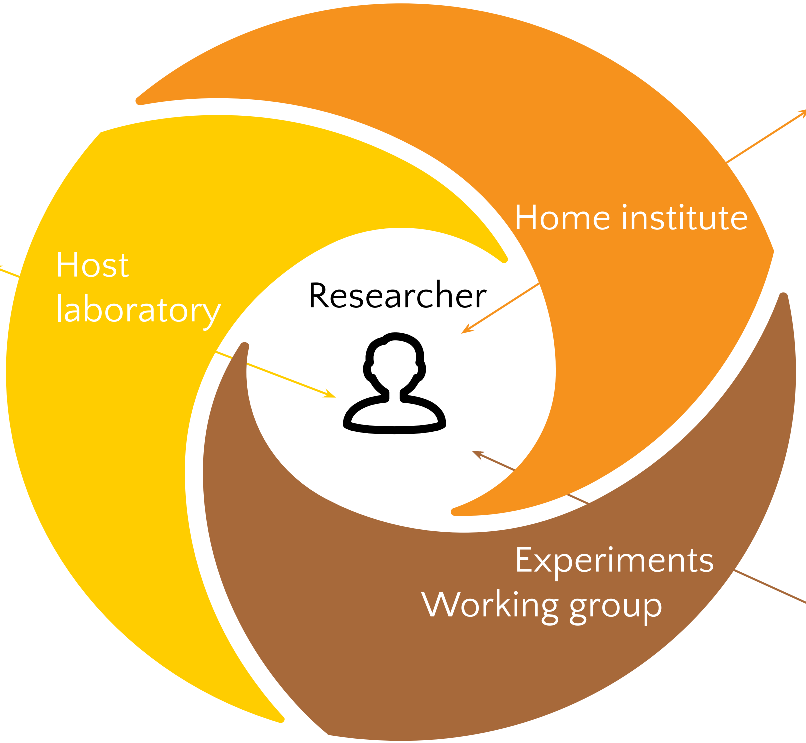
Anecdotic feedback from an Open data -workshop for PhD students in physics (outside HEP)

4

Guiding principle: Bridge the gap

between words and actions
between prototyping and implementation

4.1 Bridge the gap - between words and action

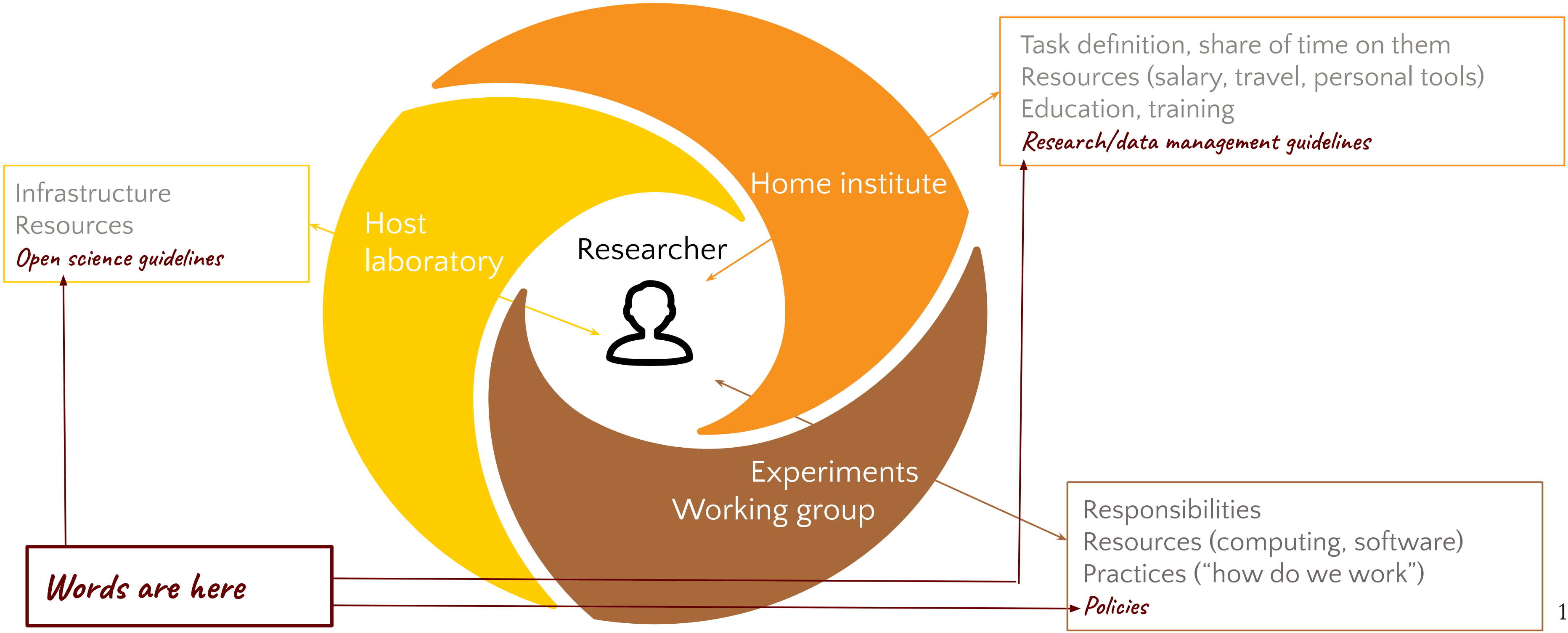


Infrastructure
Resources
Open science guidelines

Task definition, share of time on them
Resources (salary, travel, personal tools)
Education, training
Research/data management guidelines

Responsibilities
Resources (computing, software)
Practices (“how do we work”)
Policies

4.1 Bridge the gap - between words and action



Infrastructure
Resources
Open science guidelines

Host laboratory

Researcher

Home institute

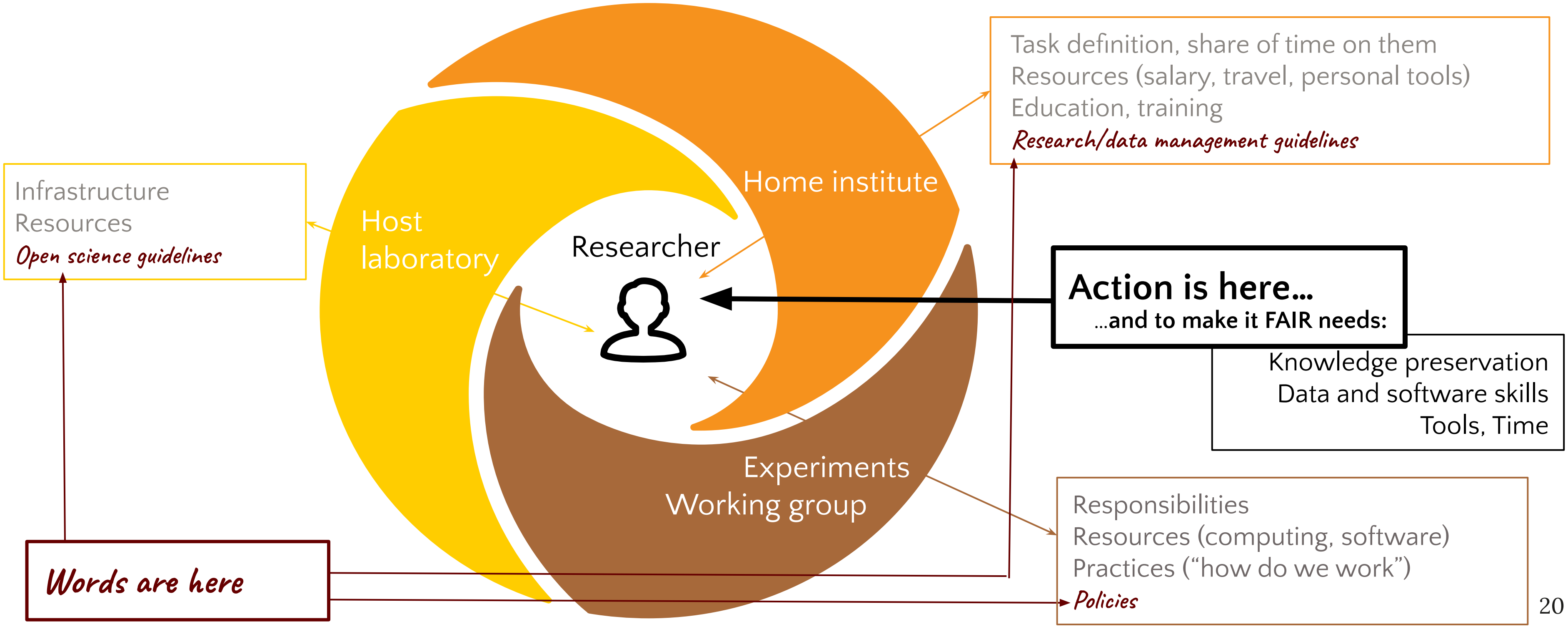
Task definition, share of time on them
Resources (salary, travel, personal tools)
Education, training
Research/data management guidelines

Experiments
Working group

Responsibilities
Resources (computing, software)
Practices (“how do we work”)
Policies

Words are here

4.1 Bridge the gap - between words and action



Infrastructure
Resources
Open science guidelines

Task definition, share of time on them
Resources (salary, travel, personal tools)
Education, training
Research/data management guidelines

Action is here…
…and to make it FAIR needs:

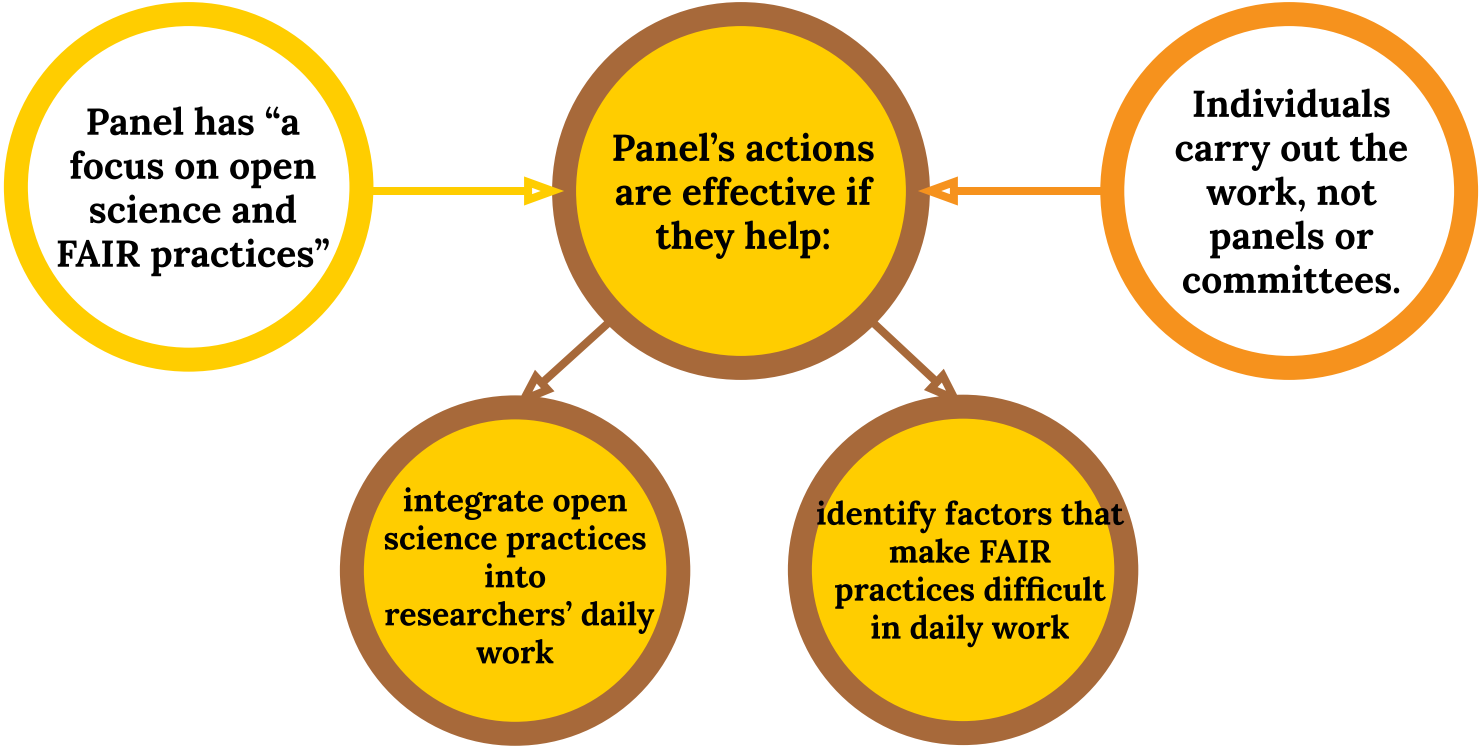
Knowledge preservation
Data and software skills
Tools, Time

Responsibilities
Resources (computing, software)
Practices ("how do we work")
Policies

Words are here

○: from panel's mandate

How to assess panel's actions on open science and FAIR practices?



Panel has "a focus on open science and FAIR practices"

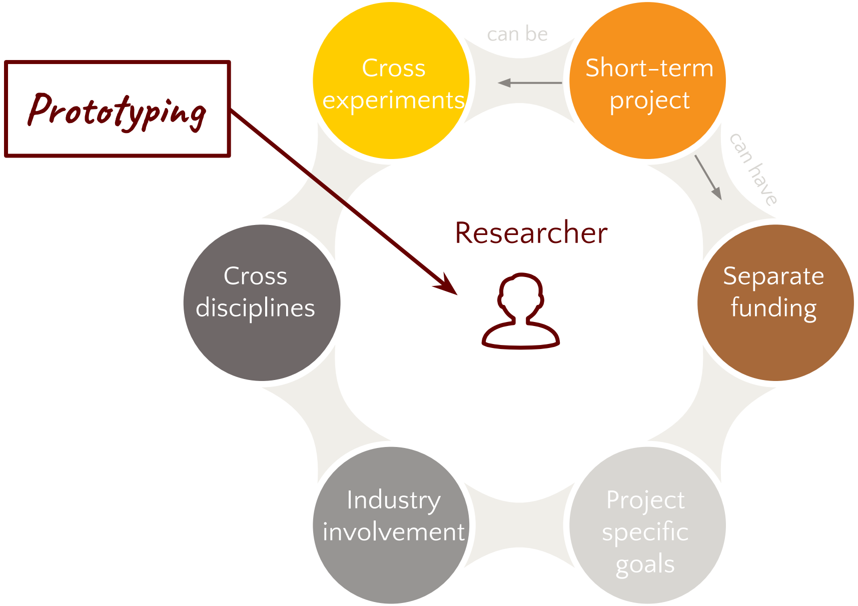
Panel's actions are effective if they help:

Individuals carry out the work, not panels or committees.

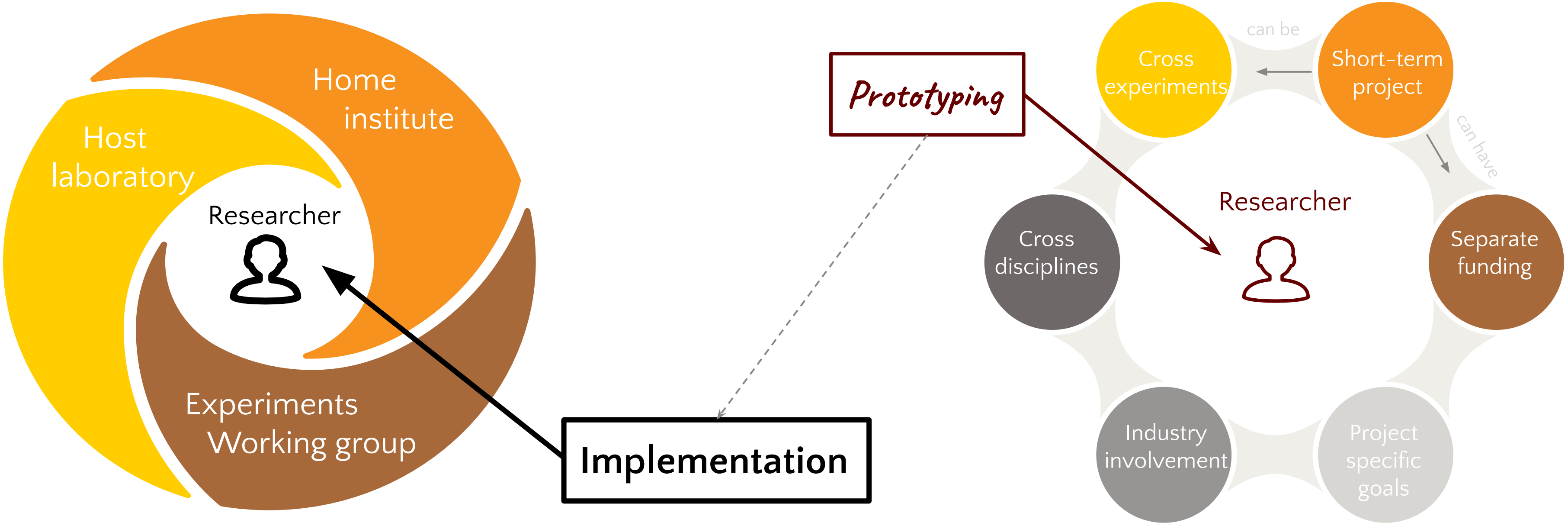
integrate open science practices into researchers' daily work

identify factors that make FAIR practices difficult in daily work

4.2 Bridge the gap - between prototyping and implementation



4.2 Bridge the gap - between prototyping and implementation



Home institute

Host laboratory

Researcher

Experiments Working group

Implementation

Prototyping

Cross experiments

can be

Short-term project

can have

Researcher

Cross disciplines

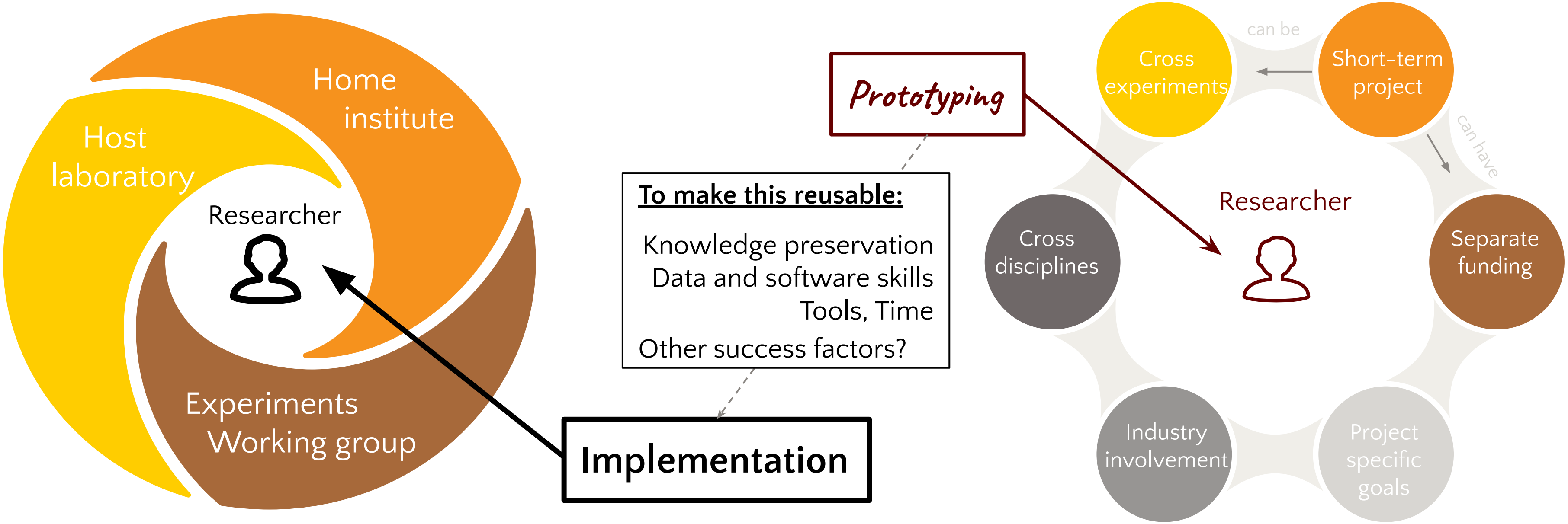
Separate funding

Industry involvement

Project specific goals

4.2

Bridge the gap - between prototyping and implementation

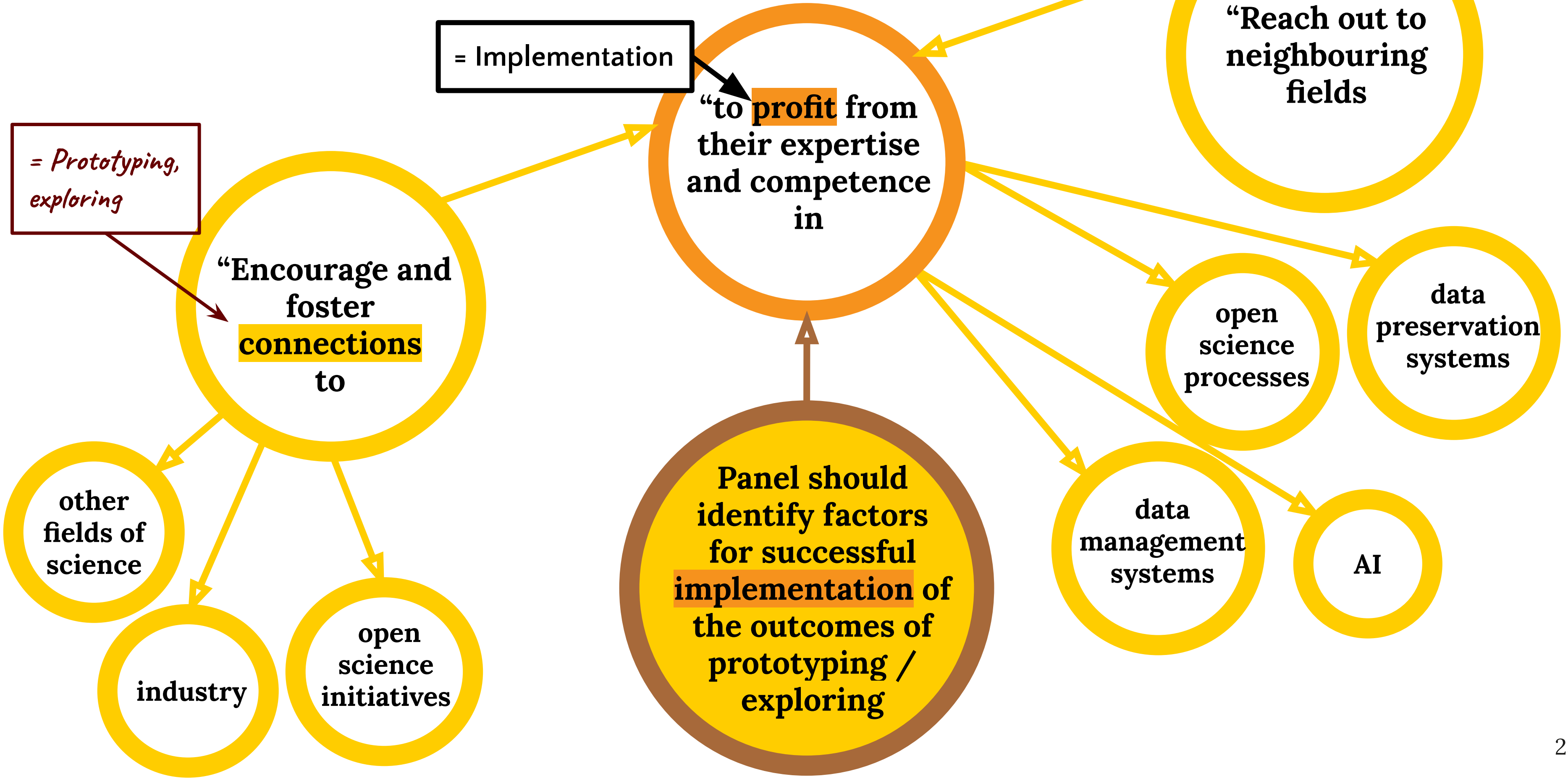


To make this reusable:

Knowledge preservation
Data and software skills
Tools, Time

Other success factors?

○ ○: from panel's mandate



= Implementation

"to profit from their expertise and competence in

"Reach out to neighbouring fields

= Prototyping, exploring

"Encourage and foster connections to

other fields of science

industry

open science initiatives

Panel should identify factors for successful implementation of the outcomes of prototyping / exploring

open science processes

data preservation systems

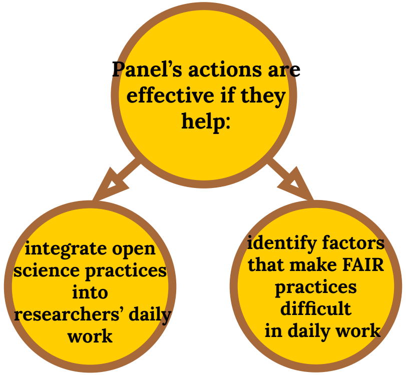
data management systems

AI

5

Your input

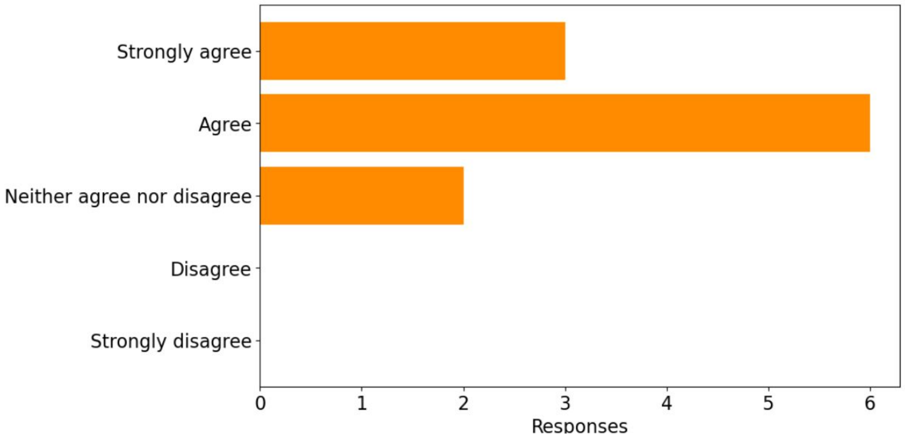
Thanks to everyone having responded the survey!



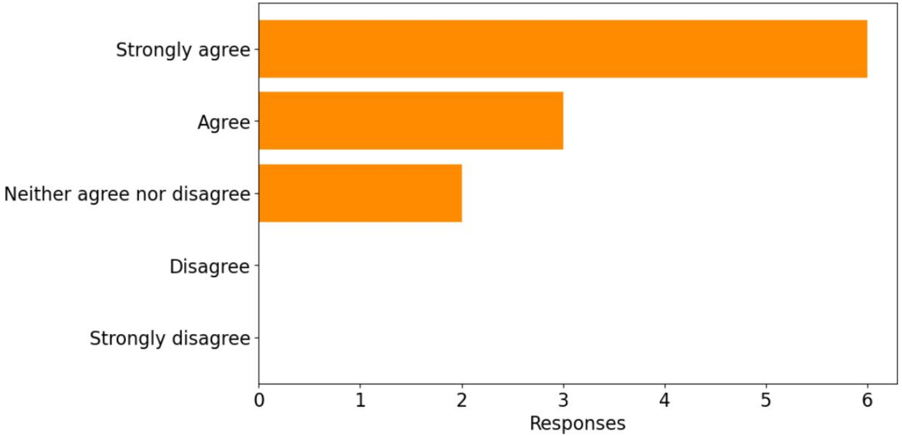
Panel’s actions are effective if they help:

integrate open science practices into researchers’ daily work

identify factors that make FAIR practices difficult in daily work



Panel should identify factors for successful implementation of the outcomes of prototyping / exploring



I think developing a way to assess the panel's positive impacts is very important, though it is challenging. I think we can do a lot of good work in helping to spread best practices from where they are developed to the wider community.

This panel covers so much ground that I don't know exactly where to start. I think it would be better to start with the definitions of the words so that all panel members have the same understanding in the discussion (e.g. What is the data lifecycle? the data lifecycle may also vary depending on the scale of the experiment...)

I suggest we do analyze the mandate and establish a 2-3 year plan of action in some of the directions (we will not be able to follow everything at the same level of involvement). I believe that the guiding principle for the panel's action is to have a strategic approach for the longer term (as opposed to already well covered aspects in the community).

I hope that this body can be an advocate for common tools and services, because the experiments themselves can not do that. They have a hard enough time advocating for funding of their own operations. In the end both sides are needed.

We need to define a multi-branched structure that combines advanced technology initiatives, pathways from prototyping into production and making clear to the community technology trends and how they are related to the coming "data intensive" and analysis challenges, and possible solutions following in-depth prototyping and integration with some of the major data management and analysis toolsets of the collaborations. We also need to propagate knowledge and trends in advanced networking and how these can be exploited to solve some of the challenges of the HL LHC era.

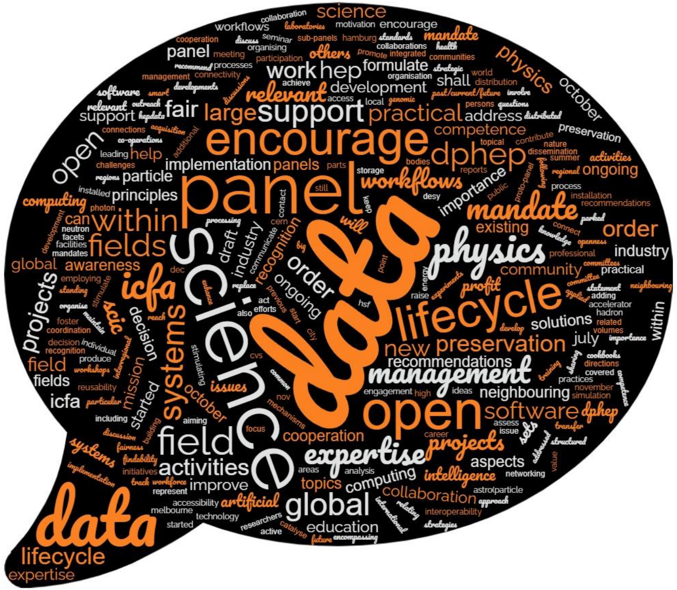
Outlook

Looking forward to the next panel meeting in May (tbc).

Discuss the plan of action

Panel substructure?

Report at the ICFA meeting in July



Thank you!

Questions?

And thanks to SlidesCarnival for this free presentation template