Just how open are we, really?

Mark Gahegan and Ben Adams

Centre for eResearch & Department of Computer Science,
The University of Auckland, New Zealand

Making data & code 'available' is simply not good enough

Desirable state: metadata and data semantics are used to support data discovery, reuse and integration.

Producers of data generate these descriptions based on their own context and understanding of what the data are good for.

Does this help a potential consumer? The consumer needs to know if the data are fit for their purpose, not for the producer's purpose.

'Openness' does not begin and end with data and methods, we need also to be open about what we do with them!

Turning the question around

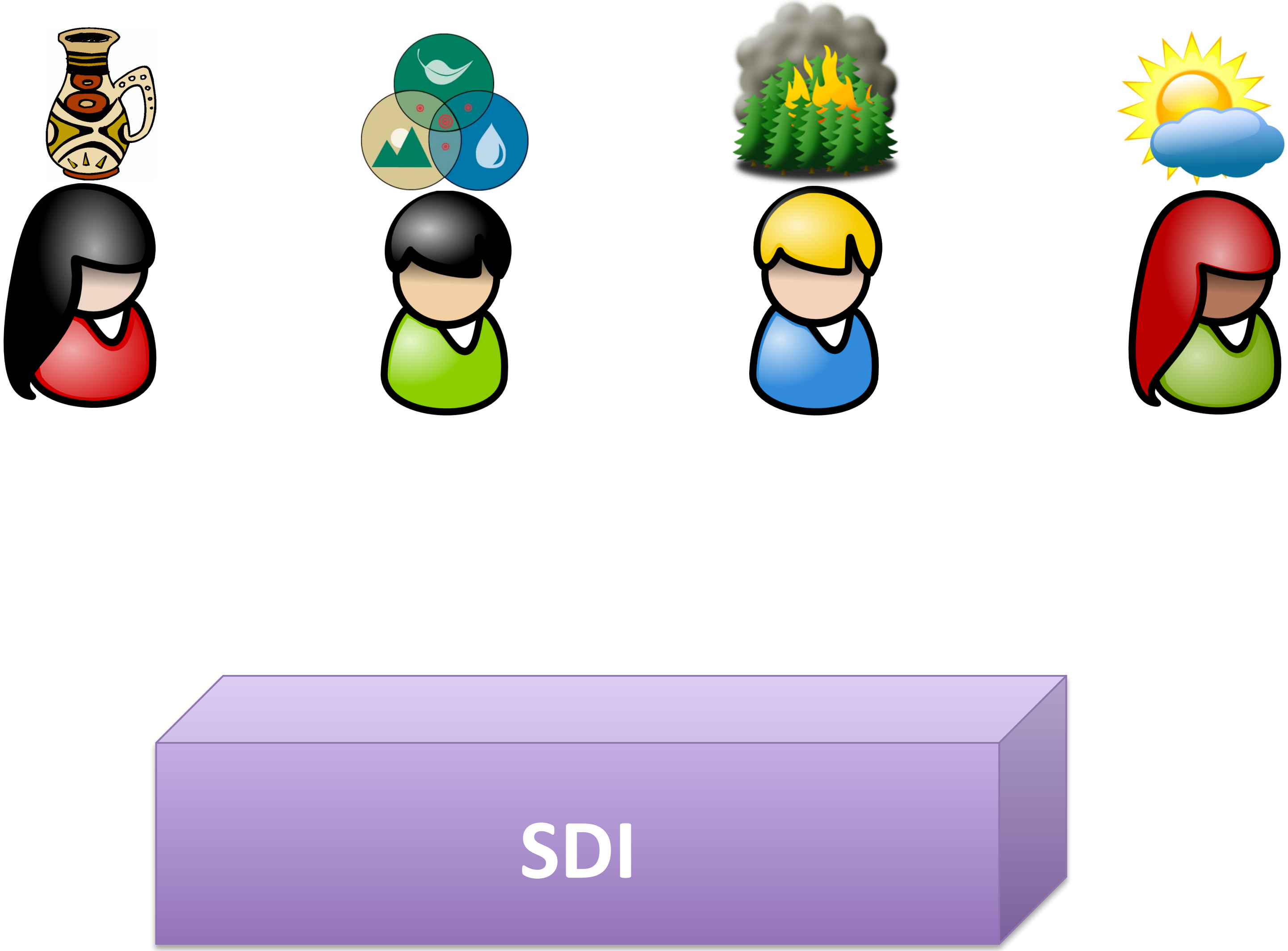
Not: *How does a data producer understand the world?*
But rather: *What does a consumer wish to know?*

So, as well as asking:

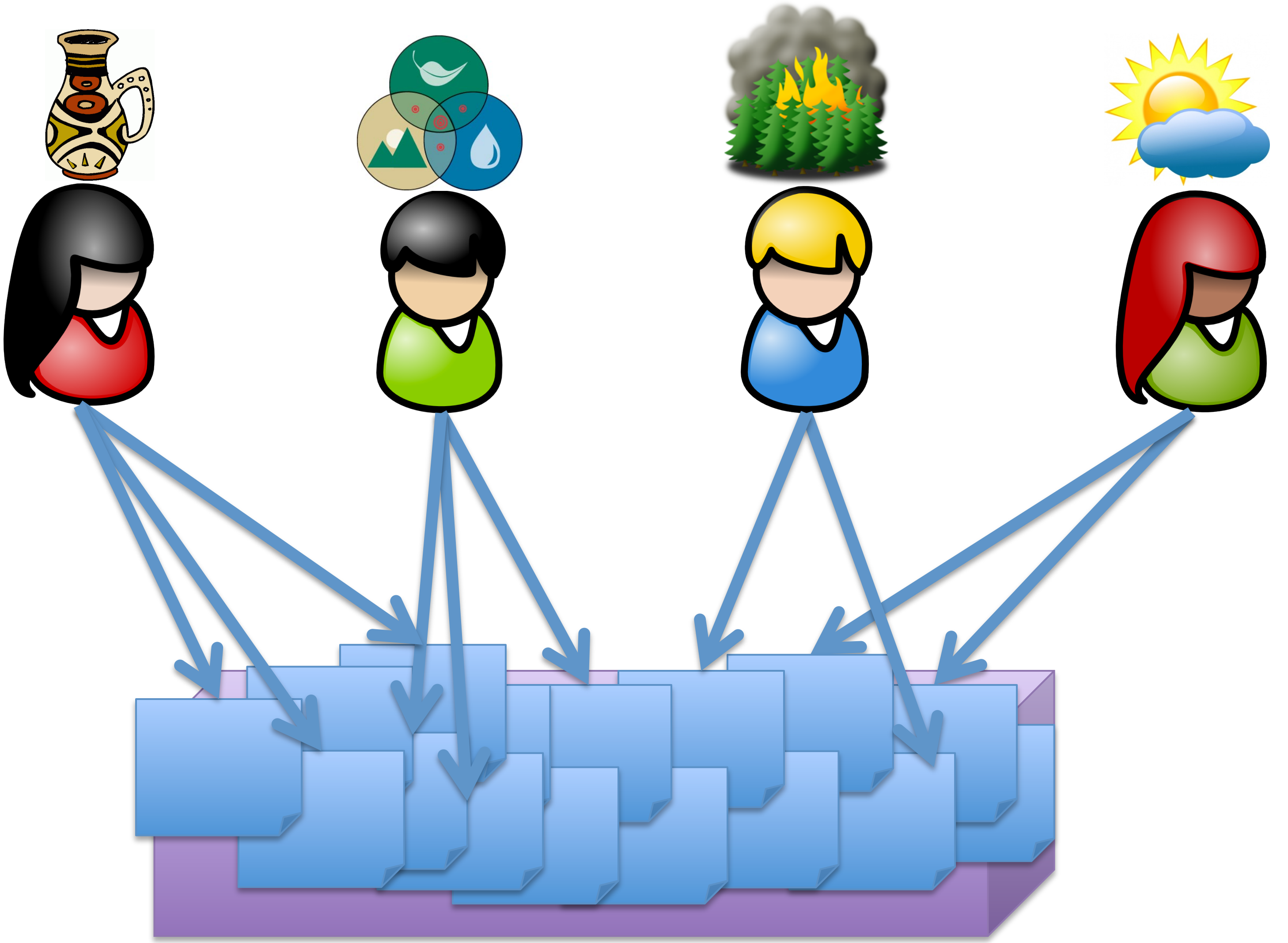
> "How do we share our data?"

…we should also be asking:

> "What kinds of properties are shown to be useful in experience to facilitate data reuse?"



SDI



Meta-model

GOAL: evaluate the utility of the various descriptive facets that *could* be captured

- We have constructed a generative model to explore the options
- The model has a set of description spaces, that represent themes that we believe (initially) may be useful
- Within these spaces we measure an ordinal distance to some kind of desired `optimal' state, as simply as we can

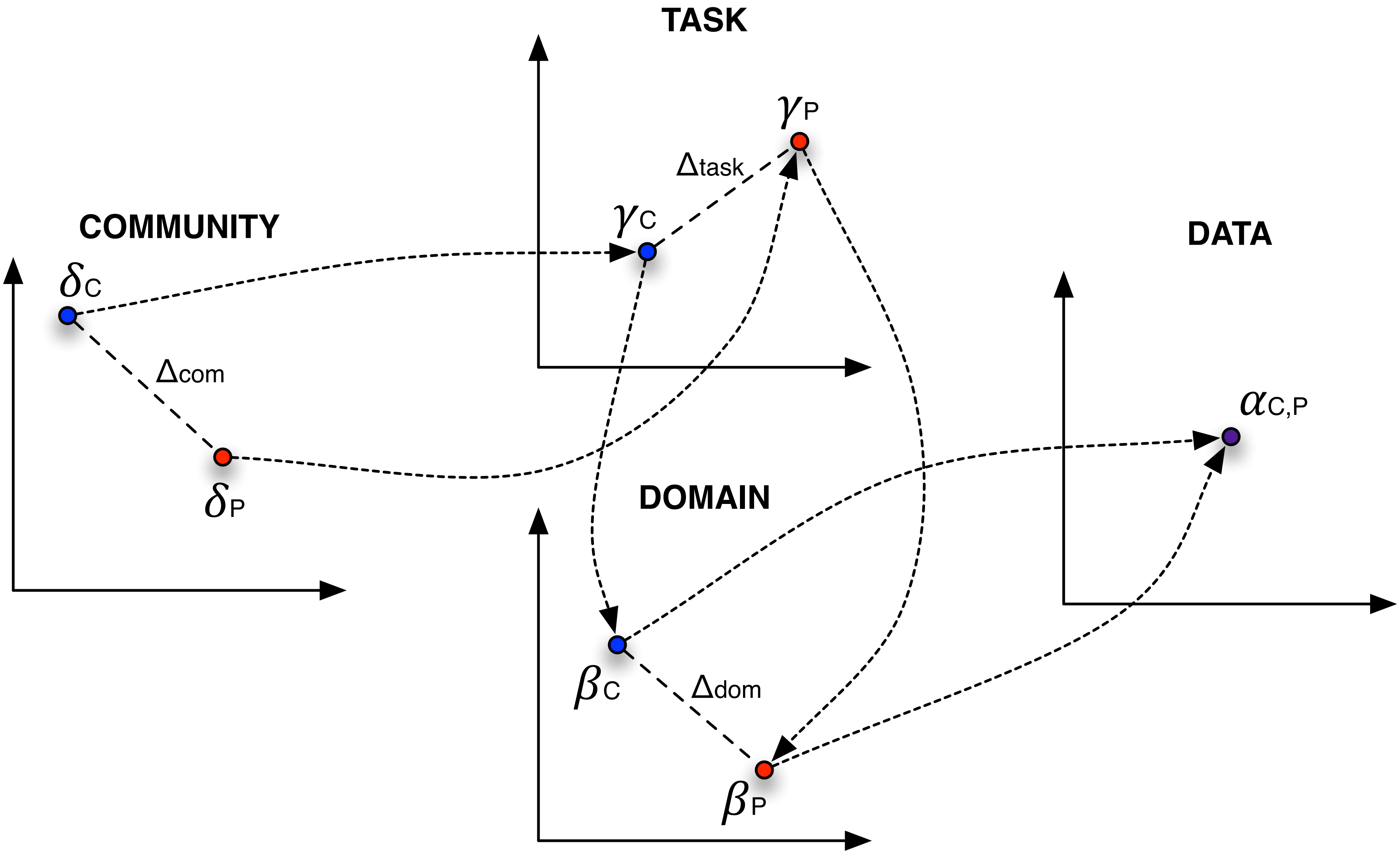
Facets

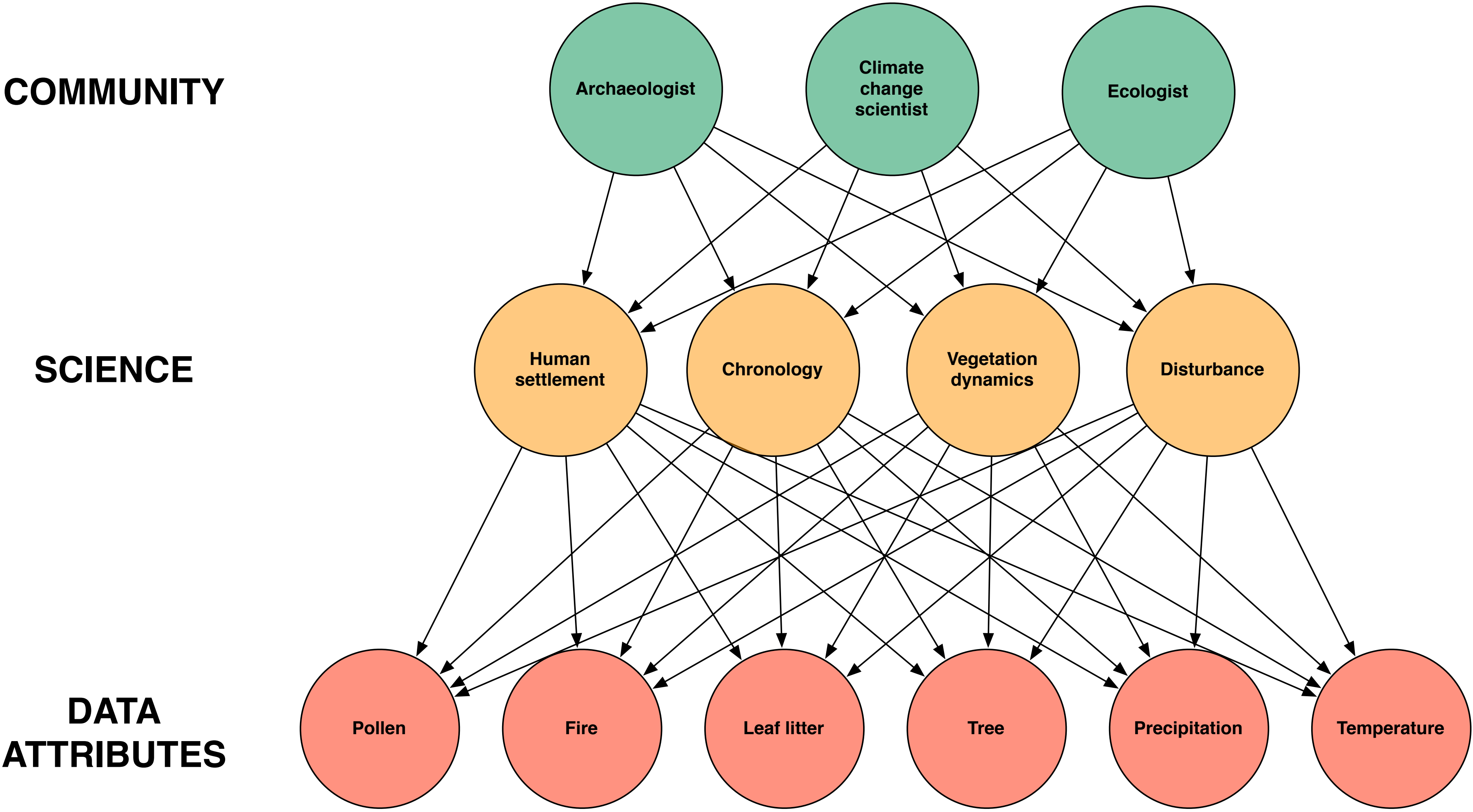
- **Spatio-temporal**: When & where is it?
  - Spatio-temporal frameworks
- **Thematic**: What is it?
  - Attribute schema & domain semantics
- **Process**: How was it made and thus how confident are we in it?
  - Quality (accuracy & uncertainty), Provenance (lineage)
- **Community**: Who can use it? Why was it made? What is it used for?
  - Motivation, access and licensing
  - Authority (governance & trustworthiness)

Knowledge of Community

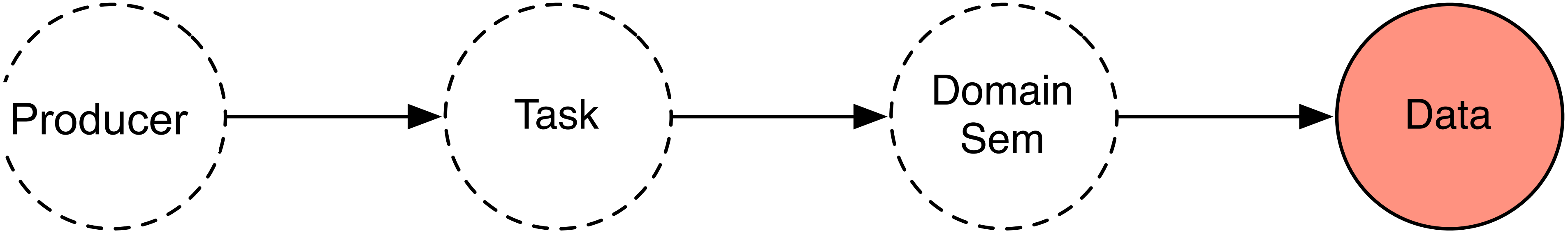
Knowledge of Science

Schema + Syntax of Data + Metadata





An example of descriptive terms from DataONE mapped to our generative model

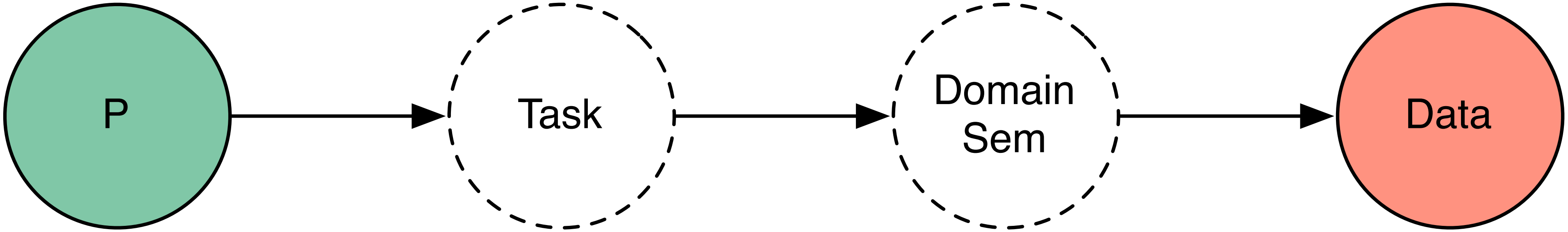


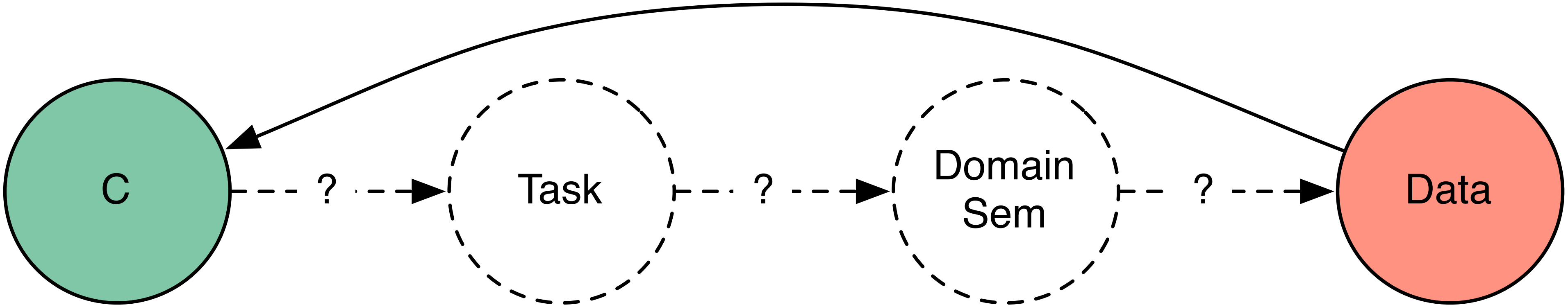
Producer

Task

Domain Sem

Data


P
Task
Domain
Sem
Data



C

?

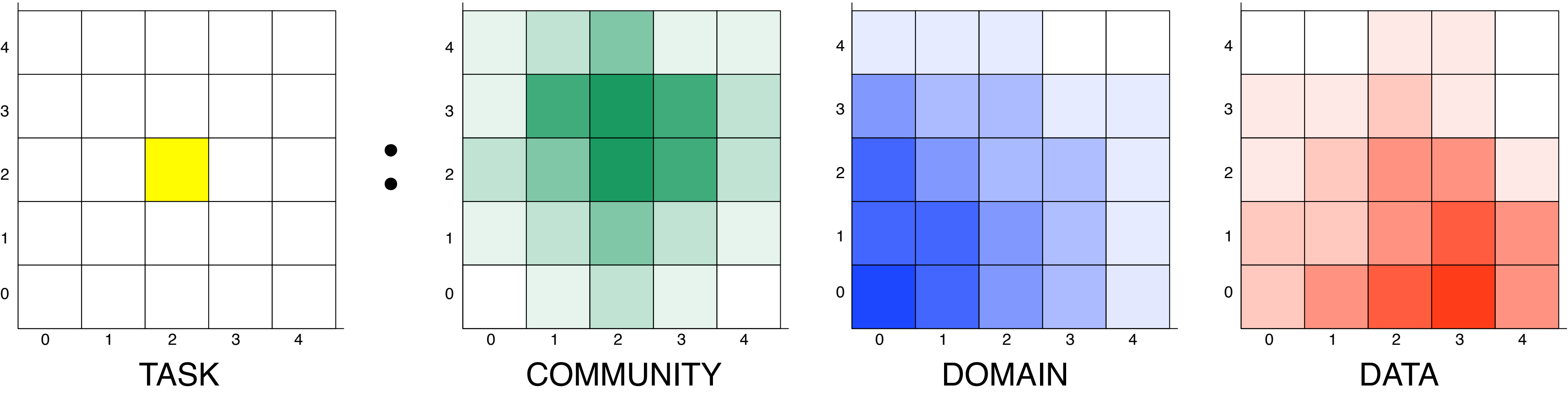
Task

?

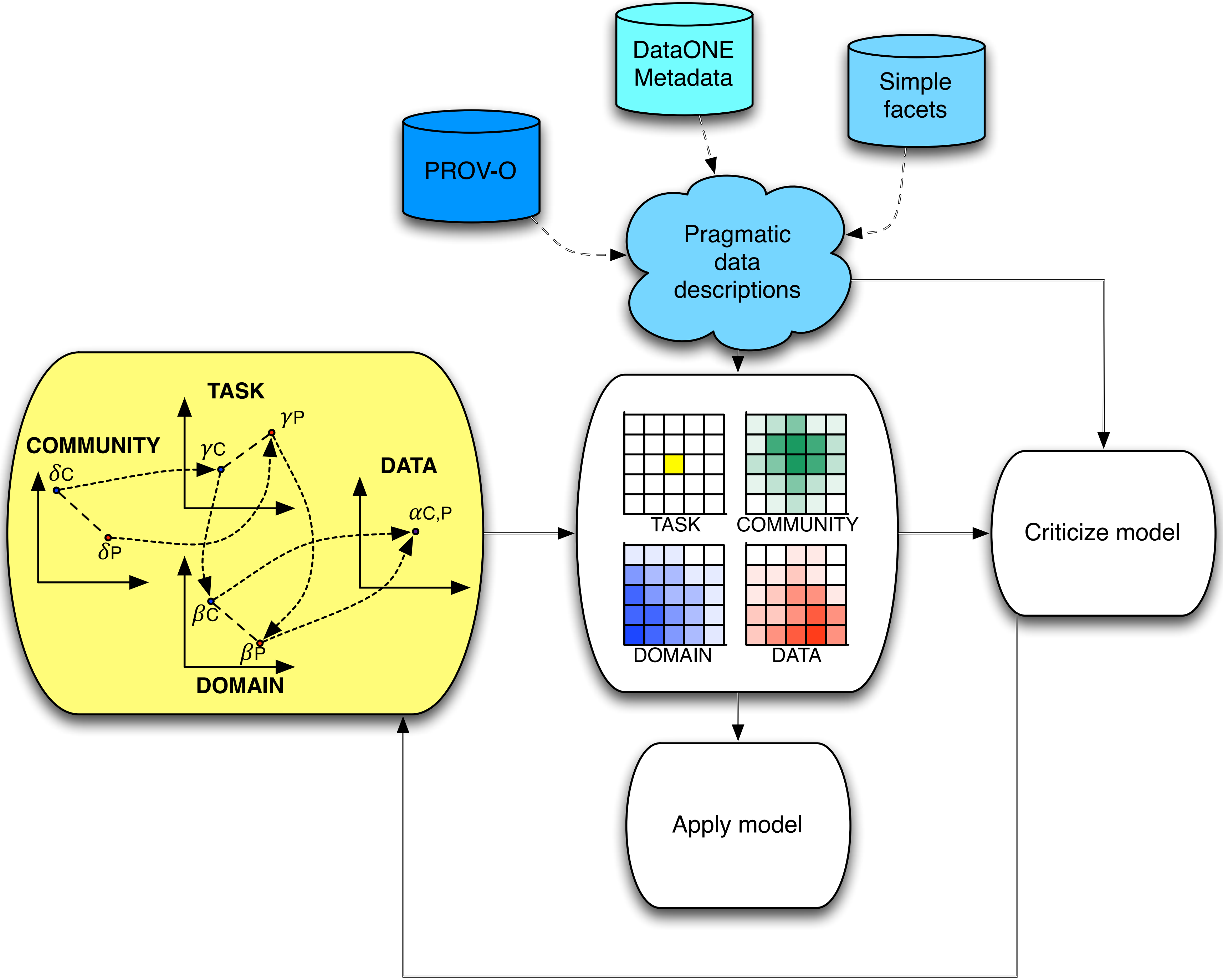
Domain
Sem

?

Data



Given a partial descriptive vector, we can estimate missing values



PROV-O

DataONE Metadata

Simple facets

Pragmatic data descriptions

Criticize model

Apply model

TASK

COMMUNITY

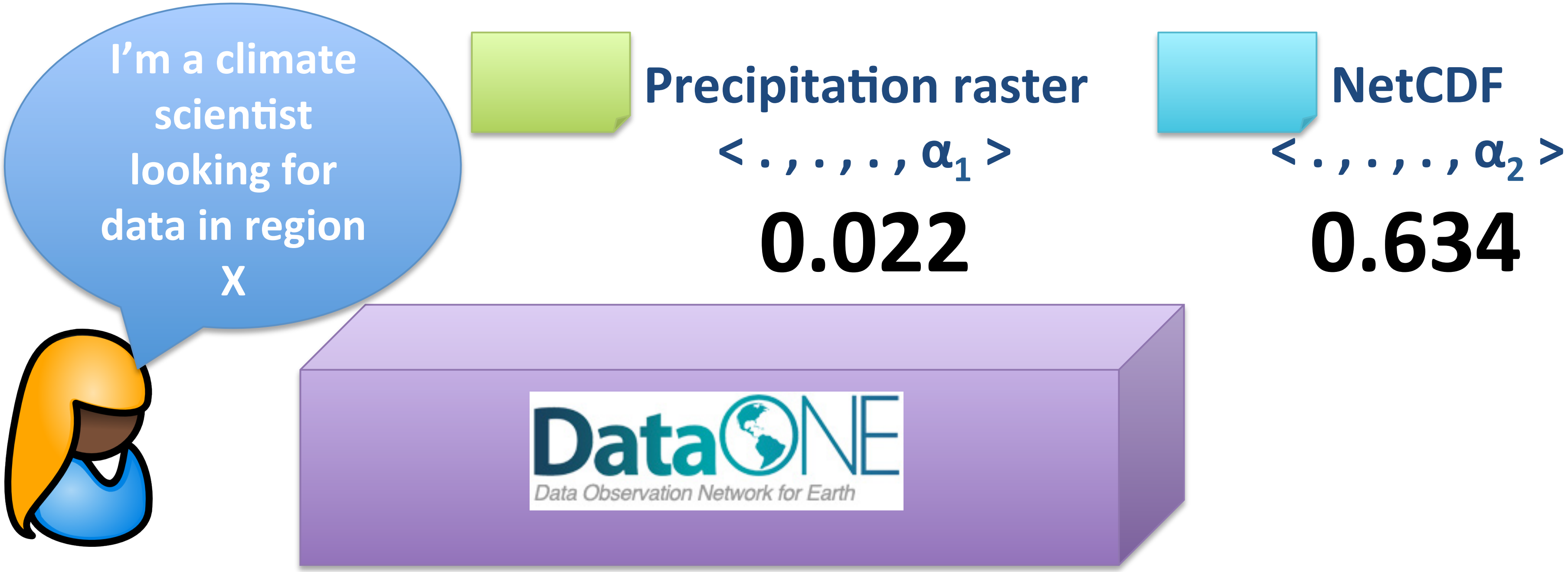
DOMAIN

DATA

We can associate relevance scores between a user's needs and data descriptions we have captured.

- climate scientist (δ_0)

Scores for two real datasets in DataONE


I'm a climate scientist looking for data in region X
DataONE
Data Observation Network for Earth

🟩 Precipitation raster

$< . , . , . , \alpha_1 >$

0.022

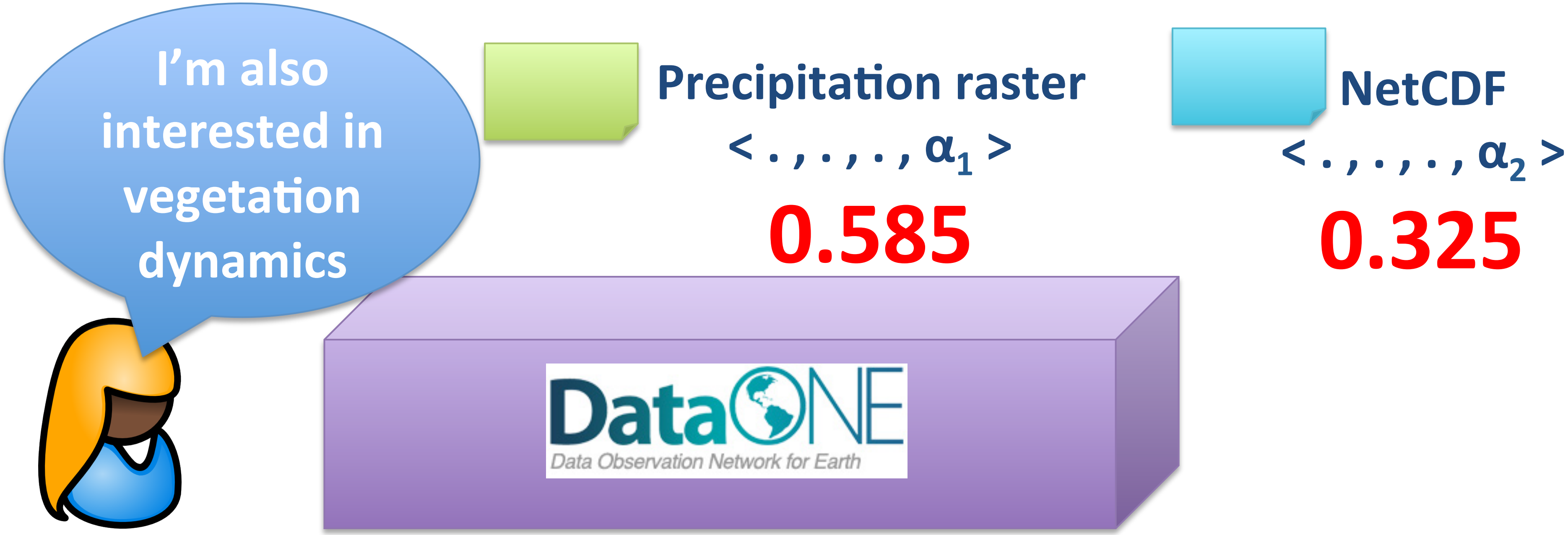
🟦 NetCDF

$< . , . , . , \alpha_2 >$

0.634

Learning new information about the user will change the scores

- climate scientist (δ_0)
- vegetation dynamics (β_0)



I'm also interested in vegetation dynamics

Precipitation raster

$< . , . , . , \alpha_1 >$

0.585

NetCDF

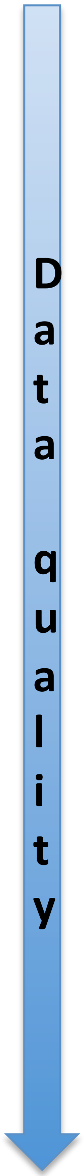
$< . , . , . , \alpha_2 >$

0.325

Manifesto for Open GeoSpatial

1. Share data, methods, code, workflows, protocols.
2. Data and metadata should be persistent, *identified*, federated and linked
3. Build or learn strong descriptions of data creation and data use
4. Expose this provenance and the use-cases
5. ***Learn*** which kinds of data descriptions are most effective at communicating fitness-for-use…

End



1. Online submission of data set for publication with basic metadata

* 2. Editor verifies that the data set is within the scope of the journal

3. Automated tools check data set for obvious omissions and errors.

4. Online tools ingest and integrate data & generate tables of statistics

** 5. Potential errors and omissions reported to data set author and/or editors

6. Data set acts on this feedback

7. Automated data checks verify that data set is complete and standardised.

*** 8. Data editor confirms that resubmitted data and metadata are correct

9. Independent peer review of data

10. Publish data to a wider scientific audience for comment

11. Author responds to referees' comments

12. Editor makes a publishing decision based on quality standard achieved by data set, (including reject and revise and resubmit).

**** 13. Data and metadata are published online. The data has its own webpage that tracks its use, or is integrated into the authoritative subject databases,

***** 14. Papers are published that consumed the data and any errors found have been corrected