A data sharing model for decentralized research data management

April 13, 2018

Nassib Nassar
Index Data ApS
nassib@indexdata.com

"Wilhelm Ostwald divided scientists into the classical and the romantic . . . . John R. Platt calls them Apollonian and Dionysian . . . .

Support mostly takes the form of grants, and the present methods of distributing grants unduly favor the Apollonian . . . .

A discovery must be, by definition, at variance with existing knowledge."

—Albert Szent-Györgyi (*Science*, June 2, 1972)

Science ≠ Workflows

- Science is methodical and orderly, but also instinctive and chaotic.

- Workflows suggest process. Science is not only about process; it is also about innovating, which can involve an unexpected departure from process.

- In building systems, too much focus on workflows will lead to overly rigid models, a bias in favor of centralization, and monolithic systems.

- A better approach is to insist on simple, independent "software tools", which scientists can either use together in expected ways or arrange in new, unforeseen ways.

Managing data: research and libraries

Researchers . . .

would like to focus on doing research

Libraries . . .

would like to offer research data storage and preservation services

Managing data: research and libraries

Researchers . . .

would like to focus on doing research

use files, databases, spreadsheets, etc.

Libraries . . .

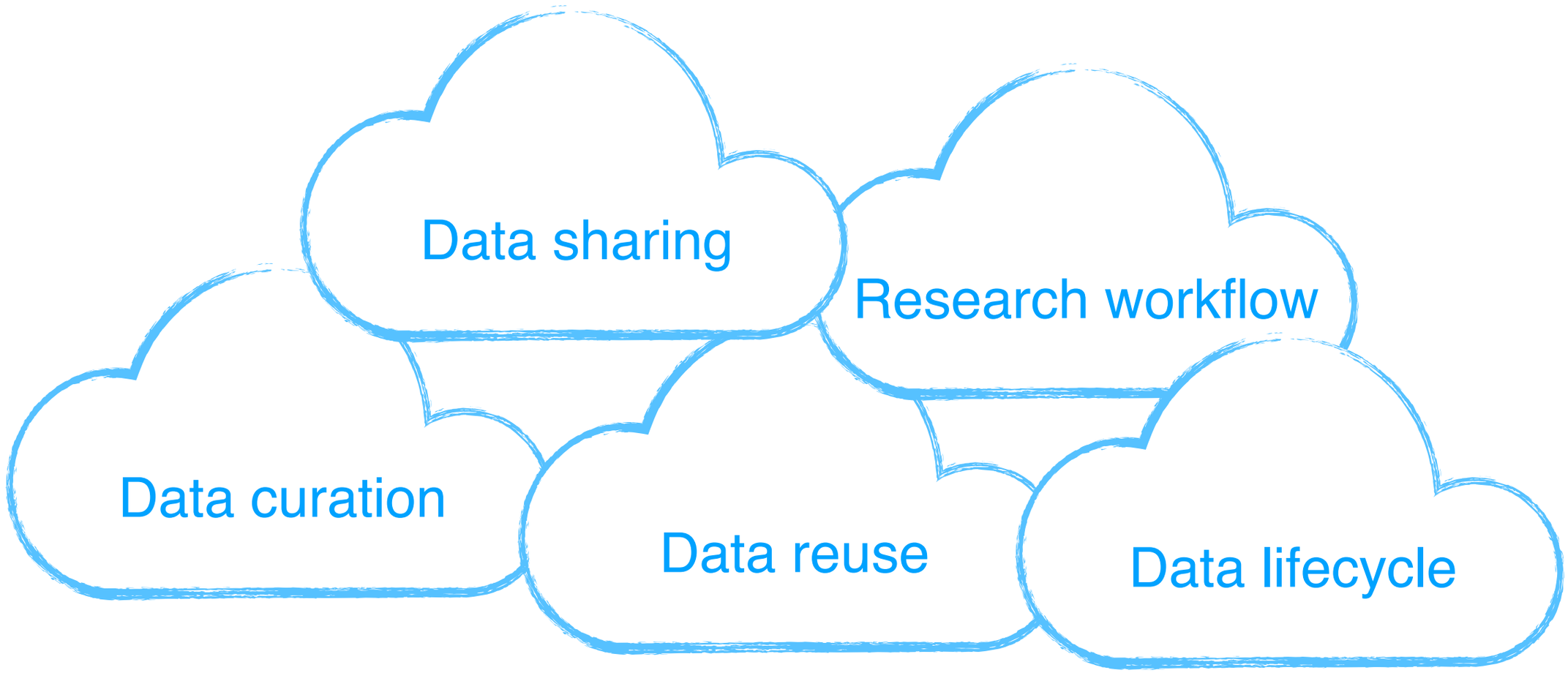
would like to offer research data storage and preservation services

use repositories

Managing data: research and libraries

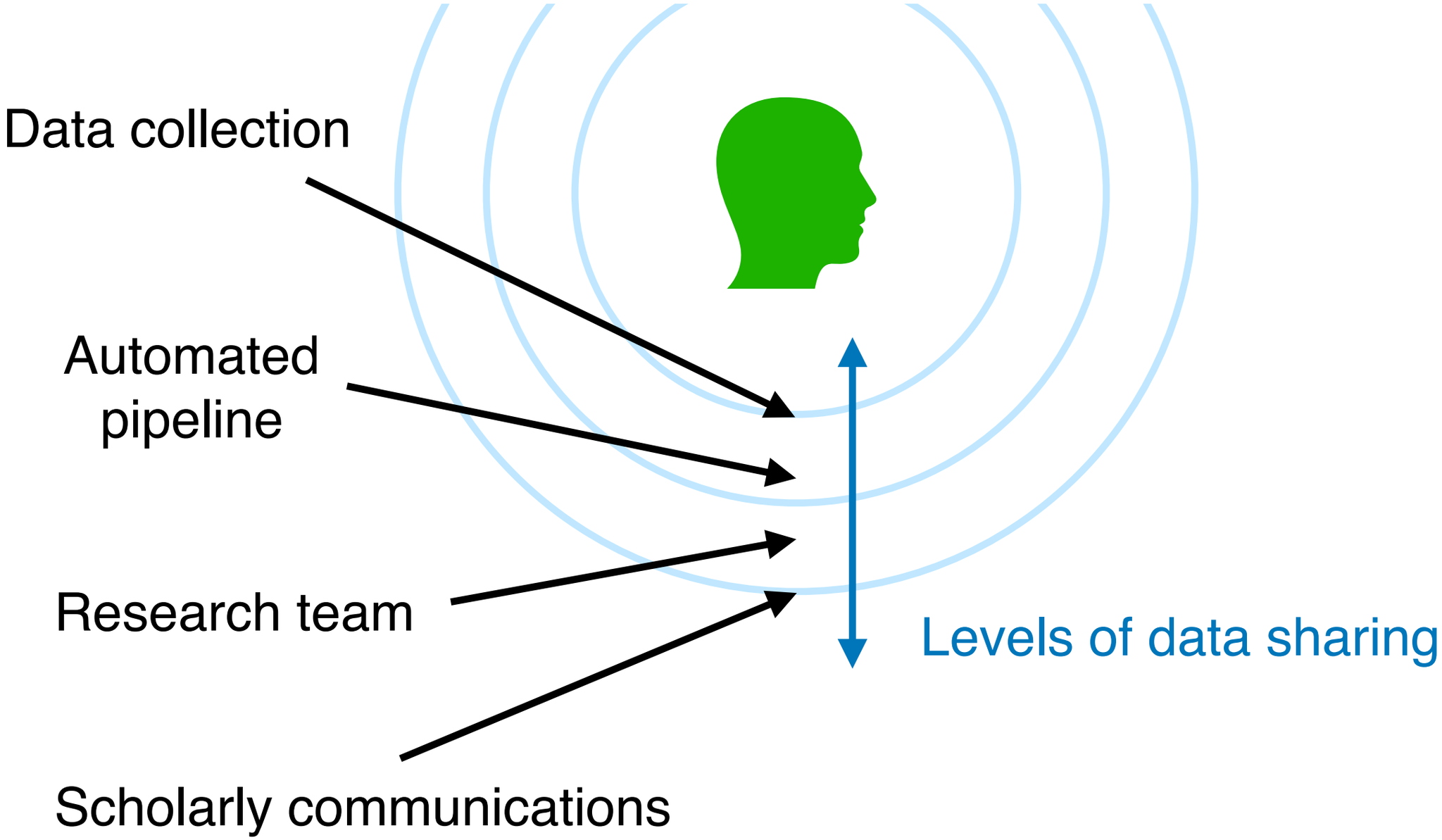
Researchers . . .		Libraries . . .
would like to focus on doing research	**?**	would like to offer research data storage and preservation services
use files, databases, spreadsheets, etc.		use repositories

Research data management concepts

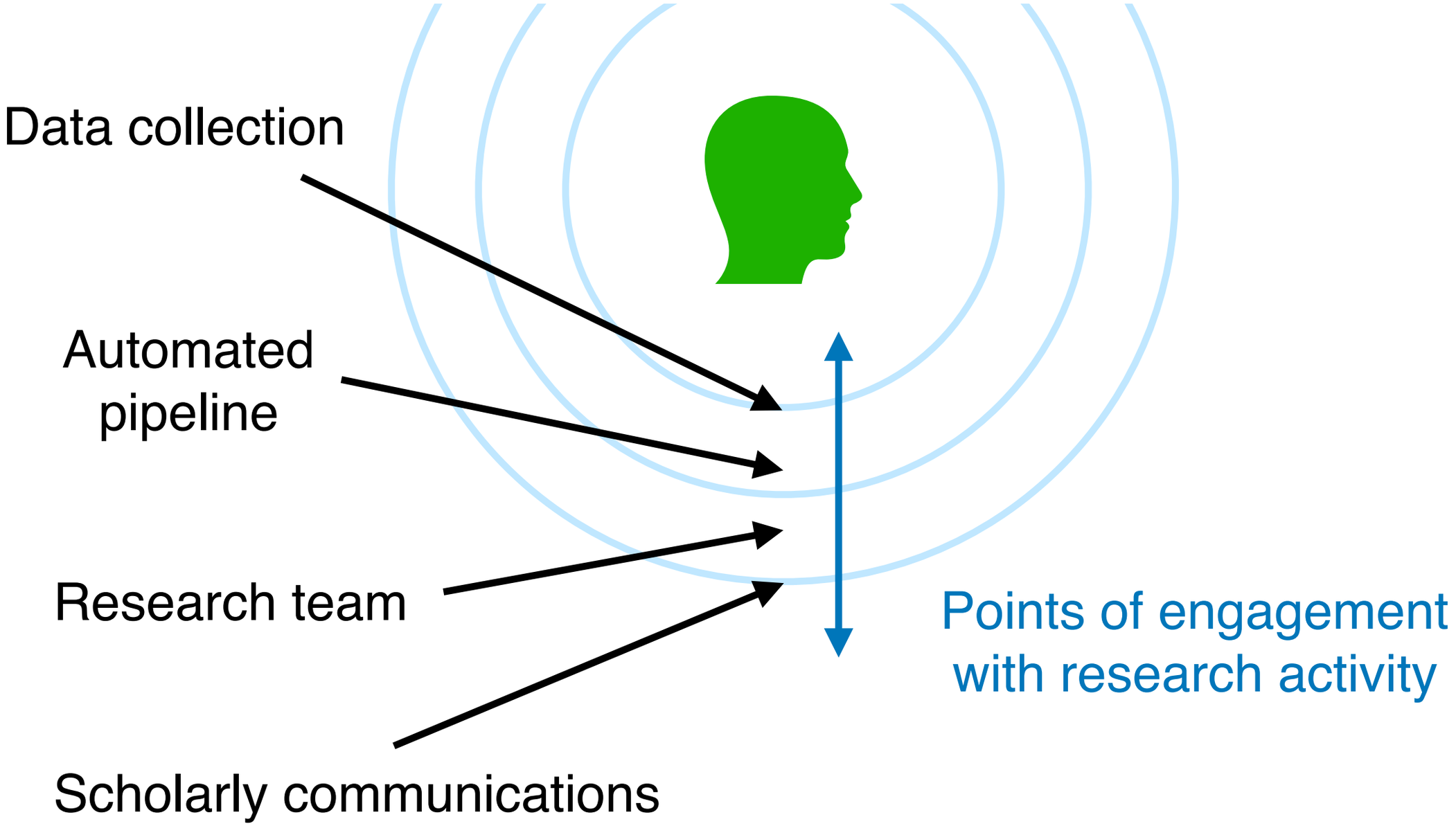


Focus in on data sharing

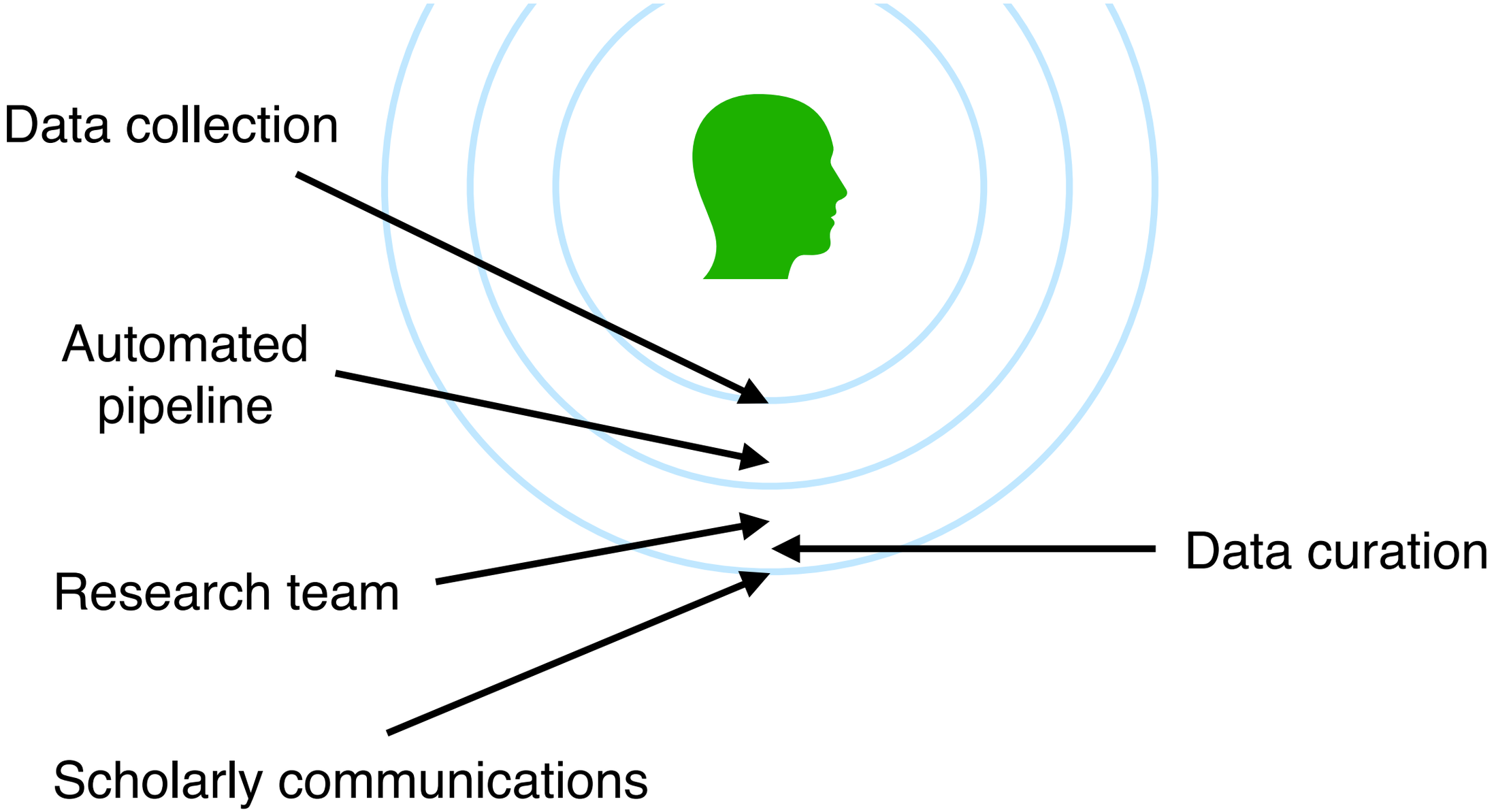
Data sharing: a ubiquitous activity



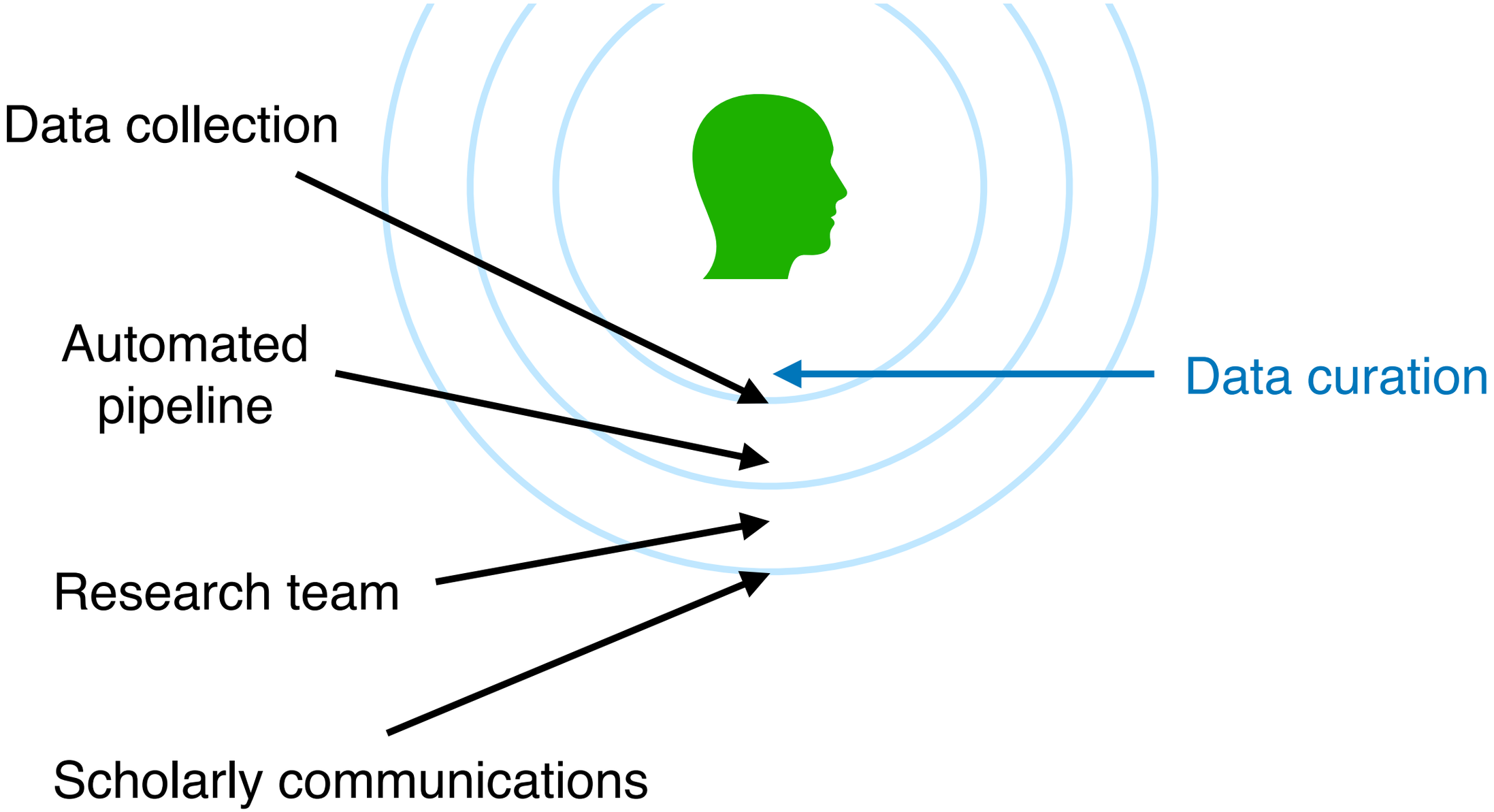
Data sharing: a ubiquitous activity



Data sharing: a ubiquitous activity

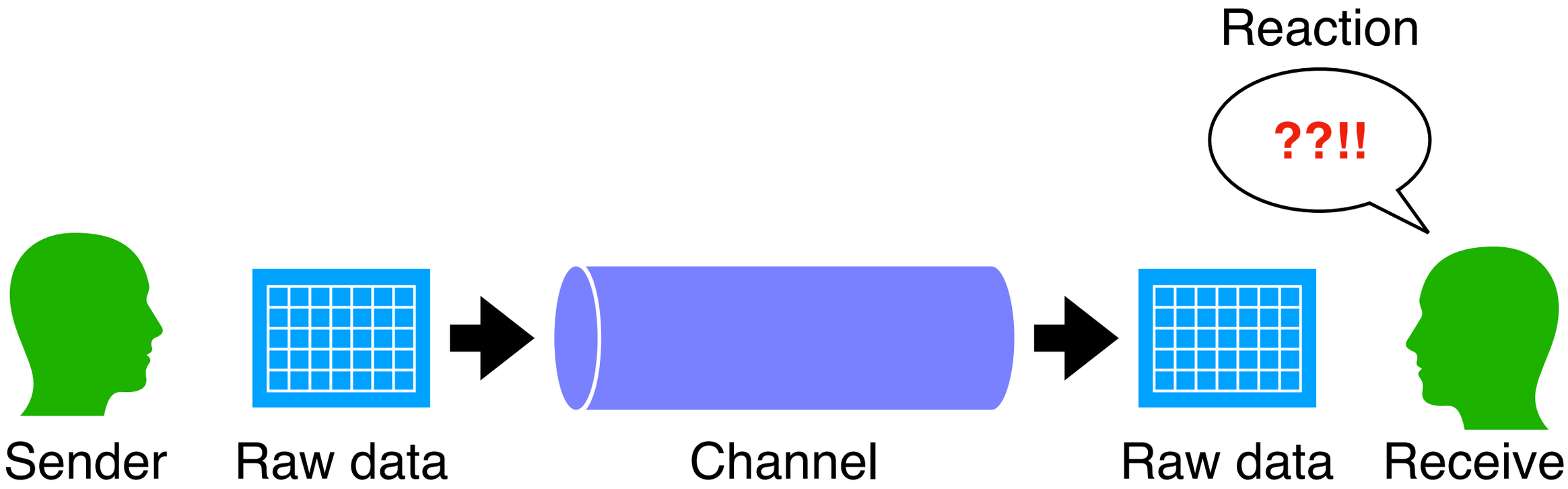


Data sharing: a ubiquitous activity



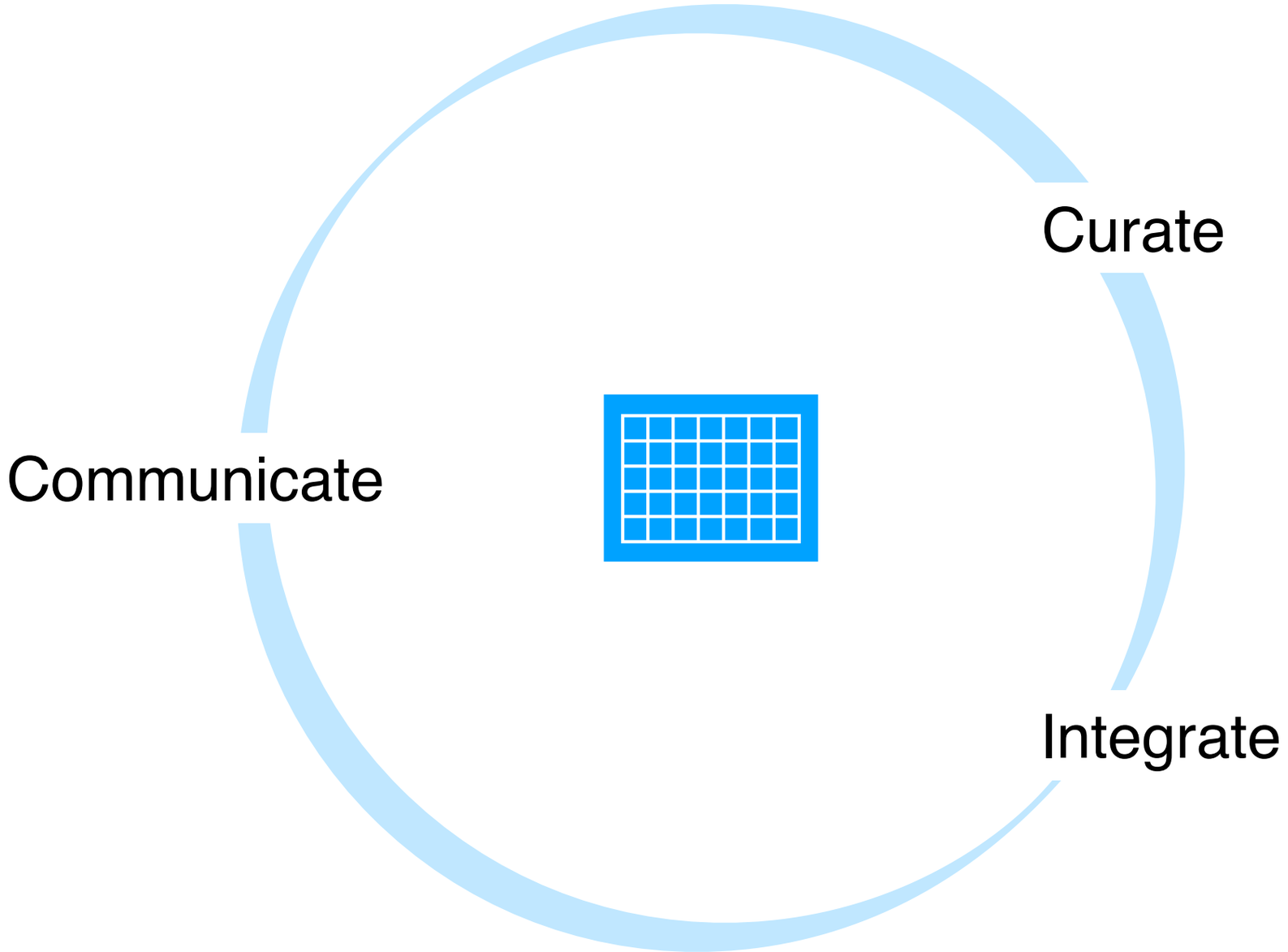
Sharing a data set

Sharing a data set

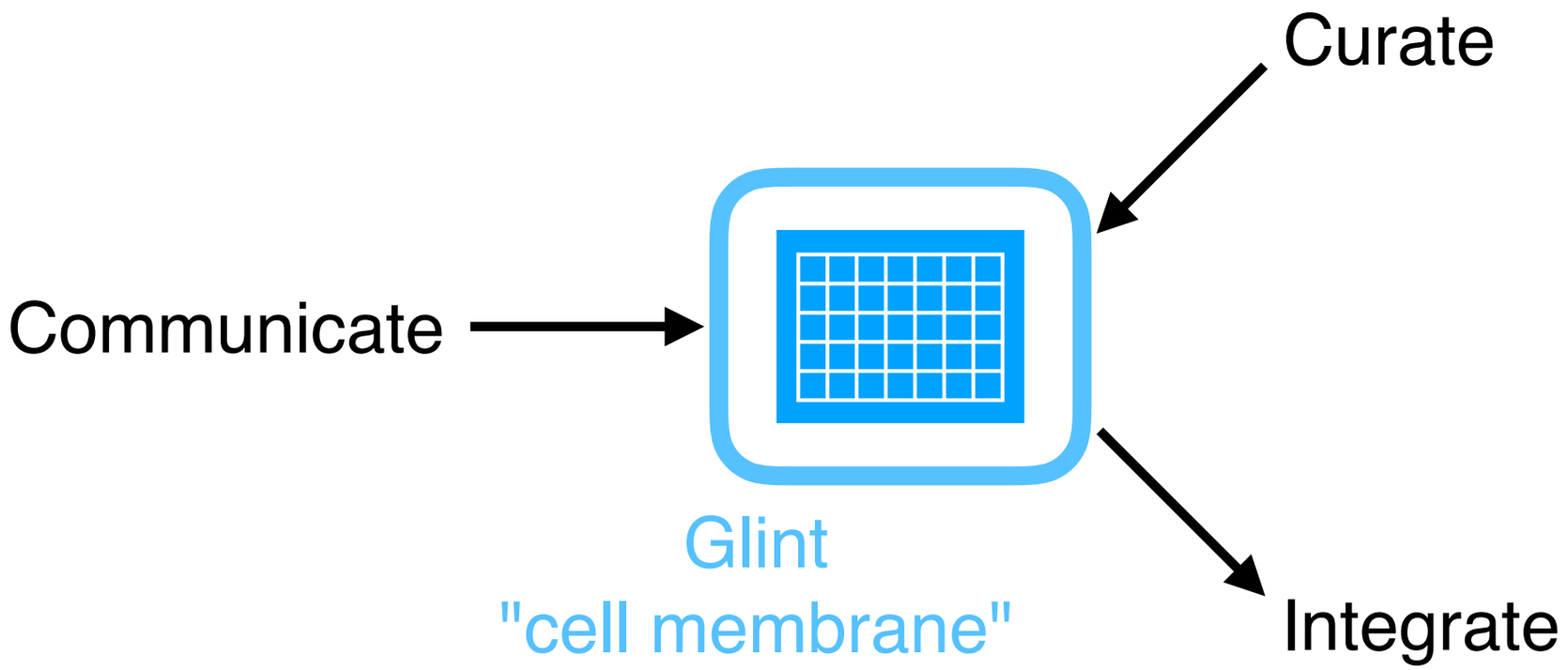


The receiver needs more information about the data

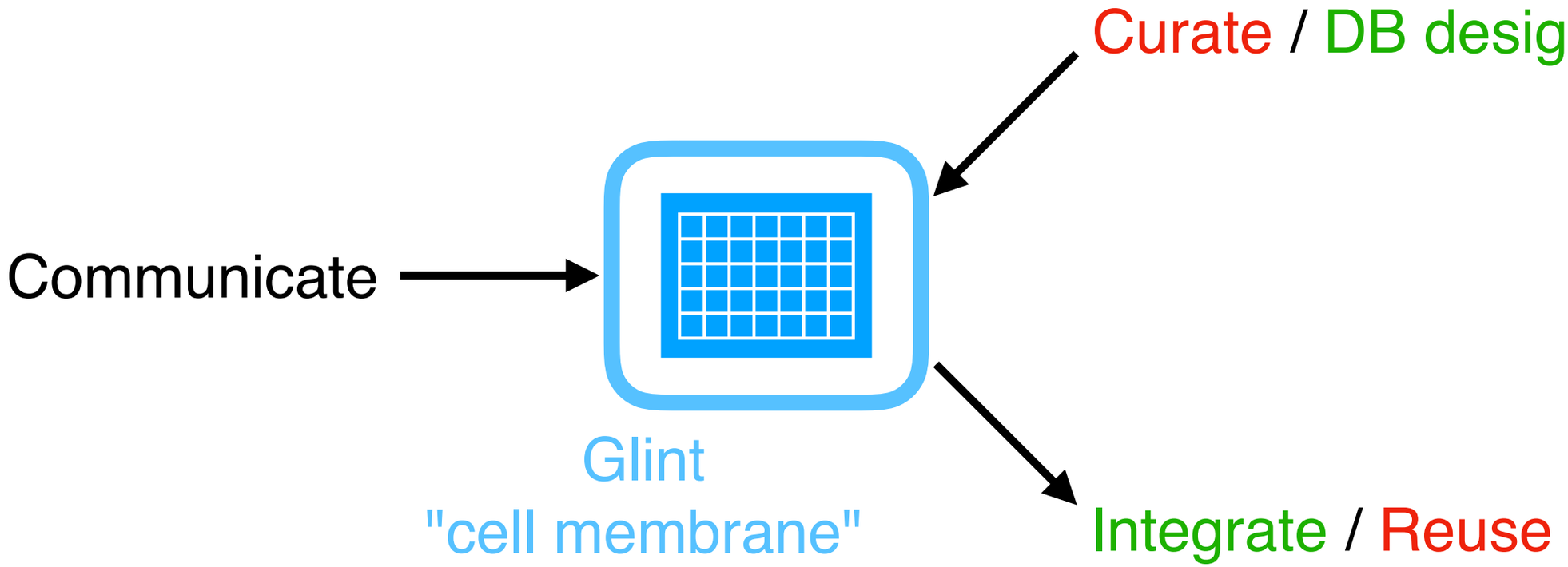
Effective Data Sharing



Glint is software that adds a thin layer of services to data

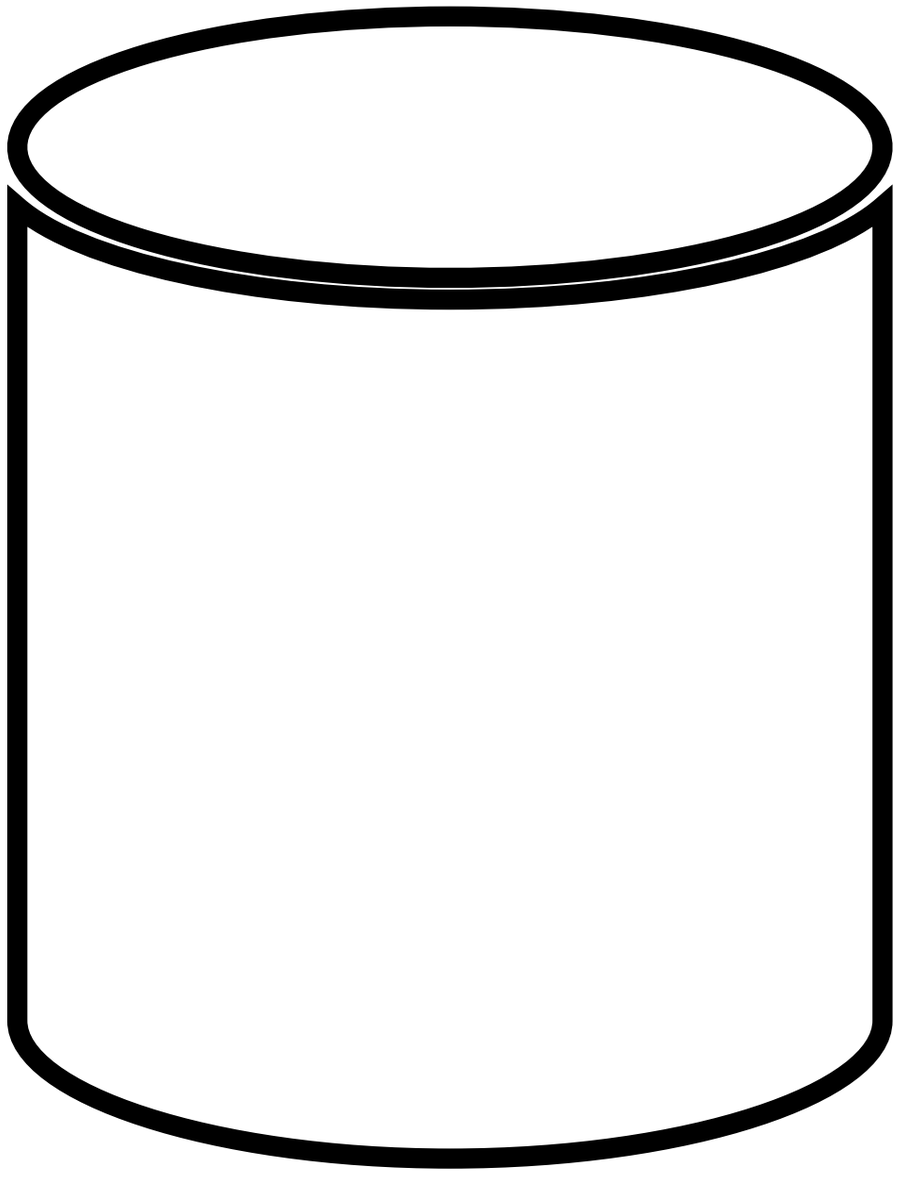


Glint is software that adds a thin layer of services to data



Glint ≠ Repository

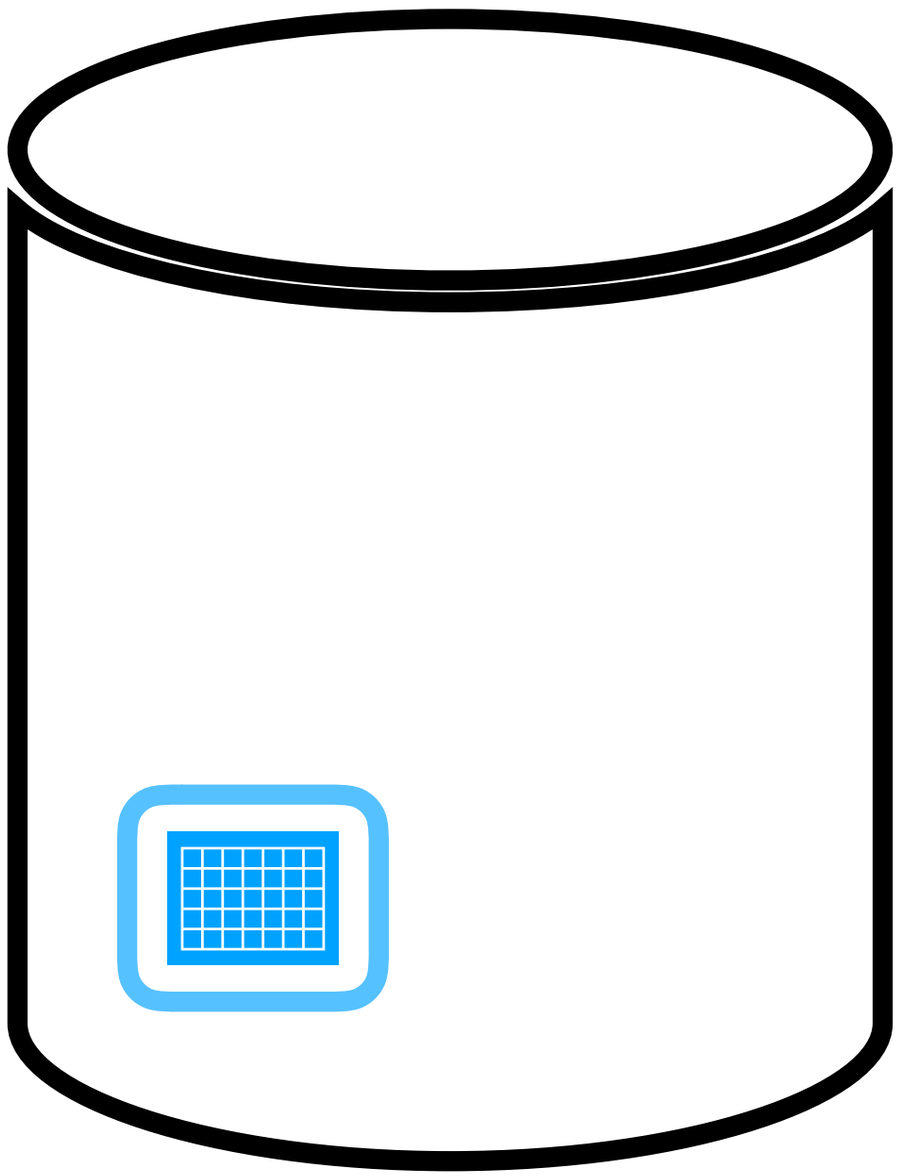
Repositories tend to internalize
and accumulate features

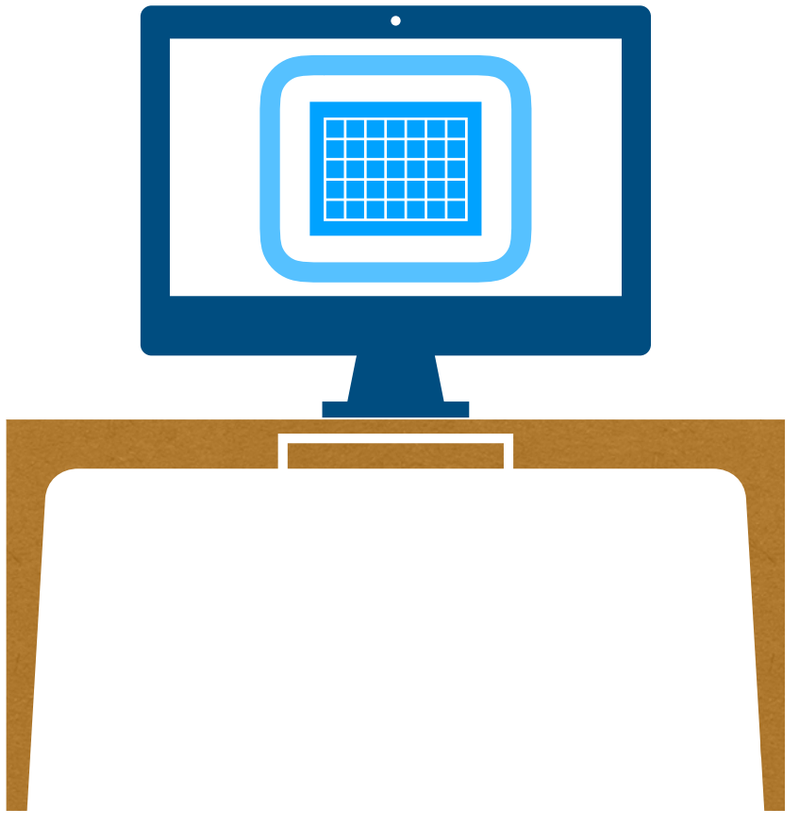


Glint strives to do one thing well,
to be easy to install, and
to integrate easily with other software

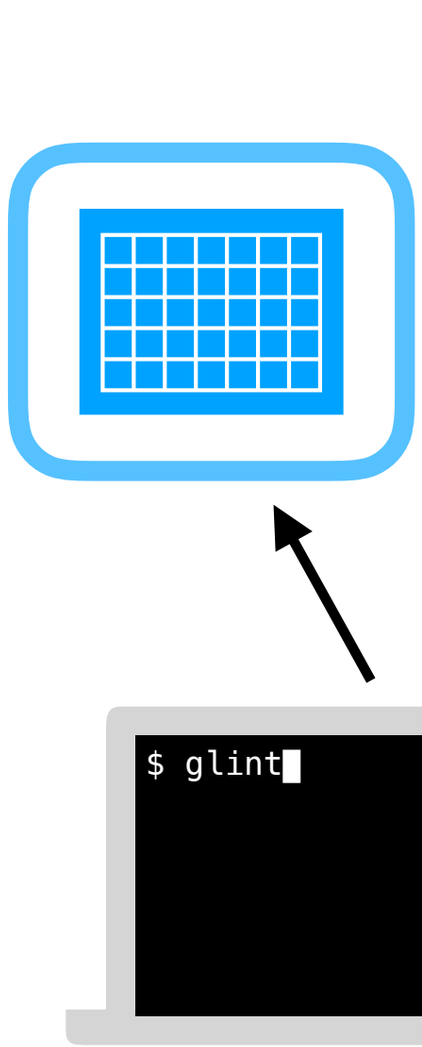
Integrate with database, analysis, and
visualization software, fit into
diverse research workflows, and
curate to the extent possible
when data are created

The data can reside in a repository, on a lab server, etc.





Using Glint

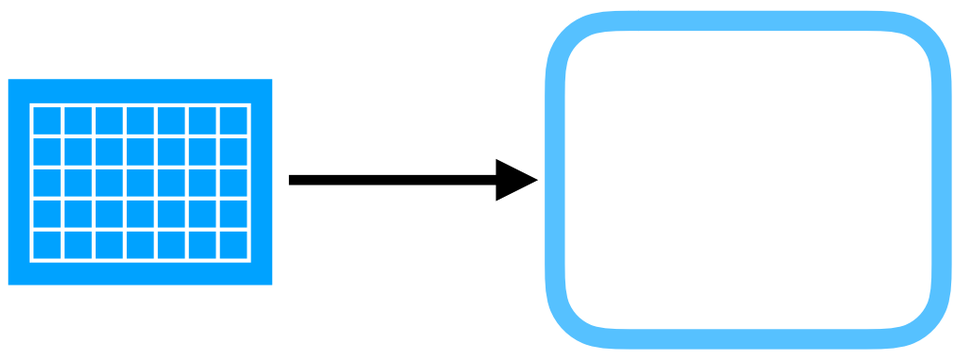


Web-based user interface:
for general users
(work in progress)

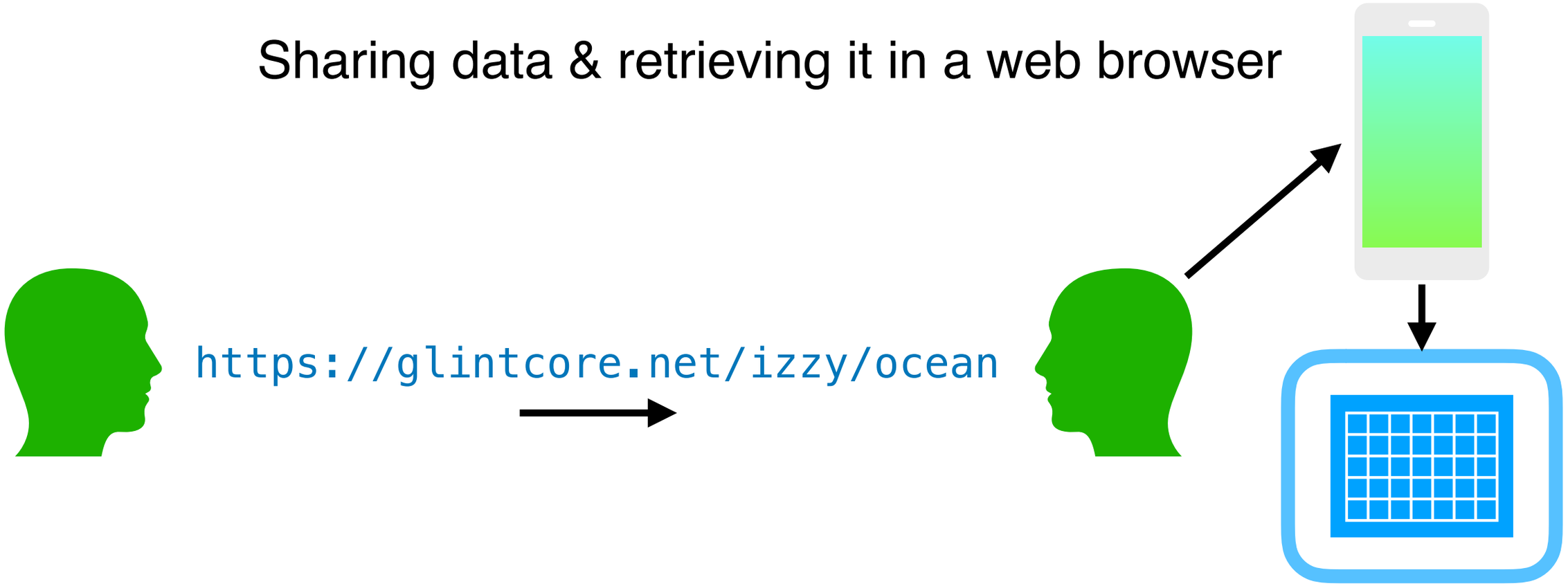
Command line interface: for technical users
& software integrators

Posting data on a Glint server

```
$ glint post ocean.csv
https://glintcore.net/izzy/ocean
```



Sharing data & retrieving it in a web browser



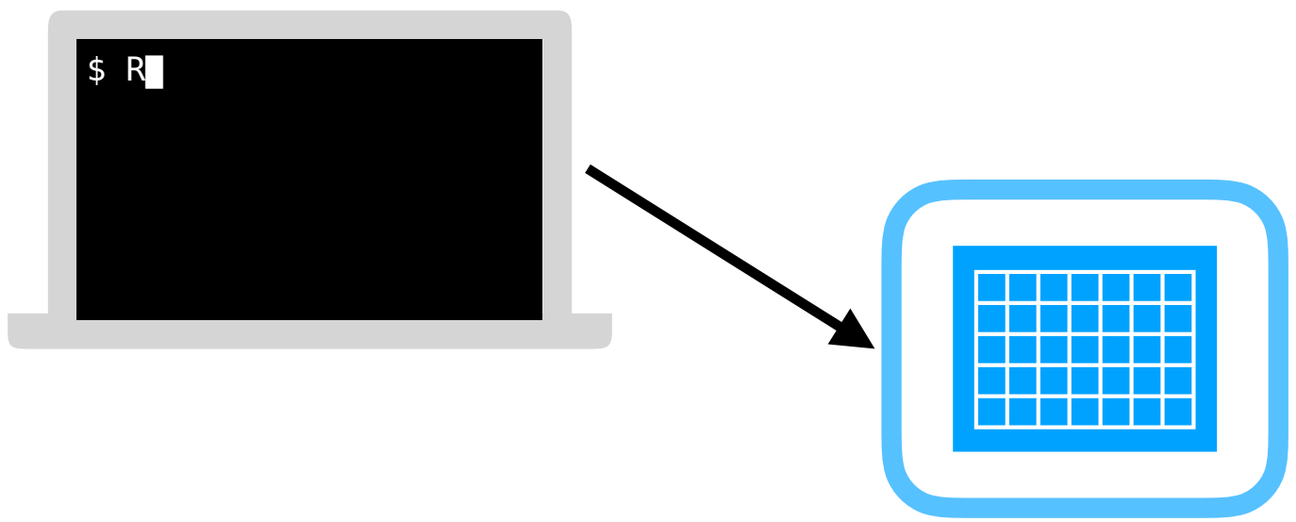
https://glintcore.net/izzy/ocean

Search

izzy / ocean

id	t	record	site_id	air_temp_avg	baro_press_avg	rel_hum_avg	dew_pt_avg	vpr_
1	2016-12-19 17:04:00	8109	1		792.5	171.399993896484		
2	2016-12-19 17:34:00	8110	1		789	163.699996948242		
3	2016-12-19 18:04:00	8111	1		790.400024414062	169.699996948242		
4	2016-12-19 18:34:00	8112	1	12.6400003433228	1012	92.6999969482422	11.5	1.355(
5	2016-12-19 19:04:00	8113	1	13.2600002288818	1011	92.5	12.0799999237061	1.407(

Retrieving data in R

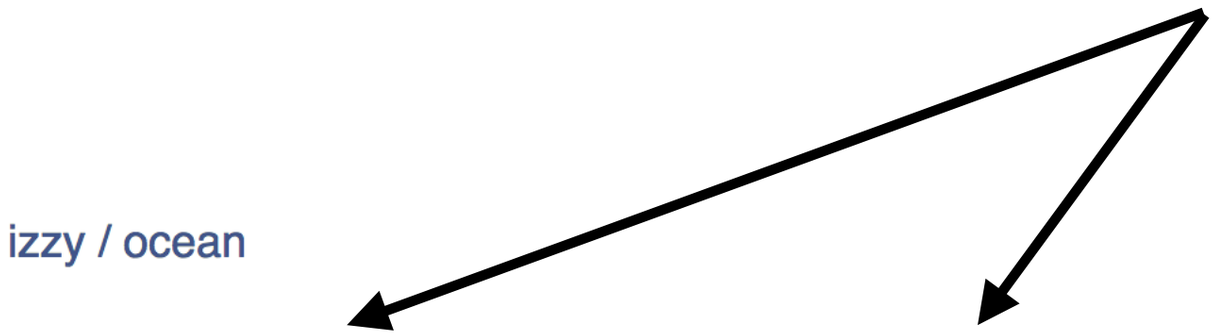



```
> ocean <- read.csv(
                "https://glintcore.net/izzy/ocean" )
```

```
> ocean
  id                   t record site_id air_temp_avg baro_press_avg rel_hum_avg
1  1 2016-12-19 17:04:00   8109       1           NA          792.5       171.4
2  2 2016-12-19 17:34:00   8110       1           NA          789.0       163.7
3  3 2016-12-19 18:04:00   8111       1           NA          790.4       169.7
4  4 2016-12-19 18:34:00   8112       1        12.64         1012.0        92.7
5  5 2016-12-19 19:04:00   8113       1        13.26         1011.0        92.5
  dew_pt_avg vpr_press_avg wind_speed wind_dir stdev wind_gust wtr_lvl_avgreal
1         NA            NA      0.443    26.72 0.048     0.443        1.238093
2         NA            NA      0.443    26.72 0.048     0.443        1.237691
3         NA            NA      0.000     0.00 0.000     0.000        1.238556
4      11.50         1.355      0.000     0.00 0.000     0.000        1.237252
5      12.08         1.408      0.000     0.00 0.000     0.000        1.236872
```

Changing how data are retrieved

`https://glintcore.net/izzy/ocean?show(t,air_temp_avg)as(tsv)`



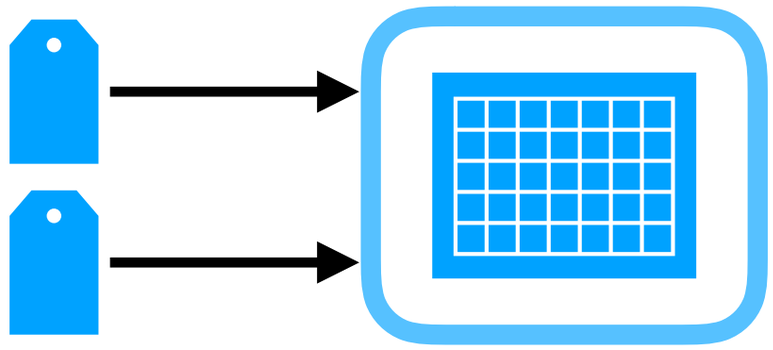
izzy / ocean

id	t	record	site_id	air_temp_avg	baro_press_avg	rel_hum_avg	dew_pt_avg	vpr_
1	2016-12-19 17:04:00	8109	1		792.5	171.399993896484		
2	2016-12-19 17:34:00	8110	1		789	163.699996948242		
3	2016-12-19 18:04:00	8111	1		790.400024414062	169.699996948242		
4	2016-12-19 18:34:00	8112	1	12.6400003433228	1012	92.6999969482422	11.5	1.355(
5	2016-12-19 19:04:00	8113	1	13.2600002288818	1011	92.5	12.0799999237061	1.407(

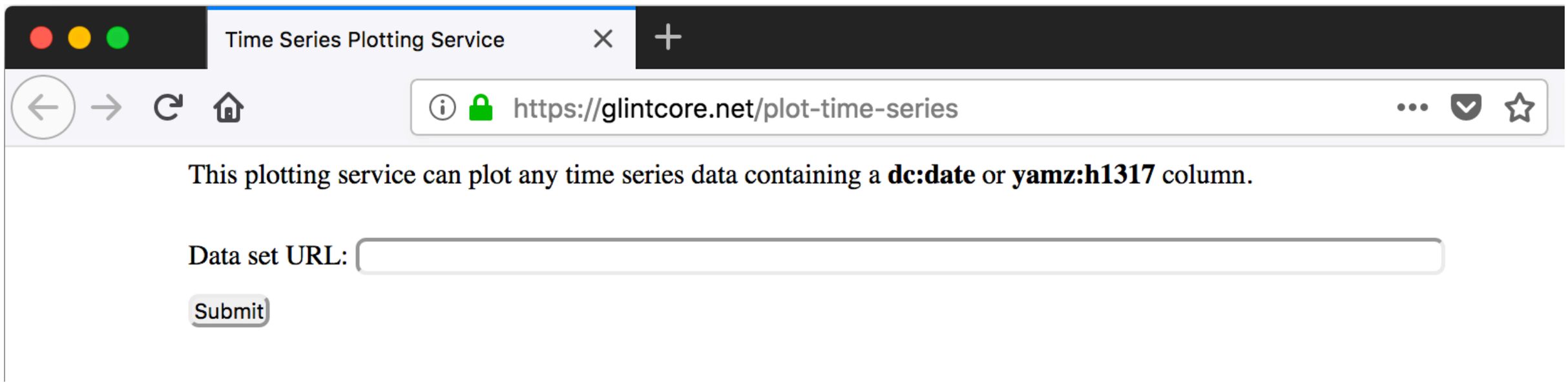
Adding metadata

```
$ glint md ocean.t dc:date

$ glint md ocean.wind_speed yamz:h3846
```

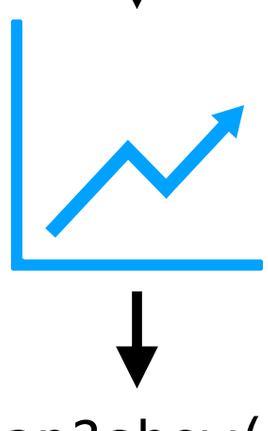


Integrating data with services (1)

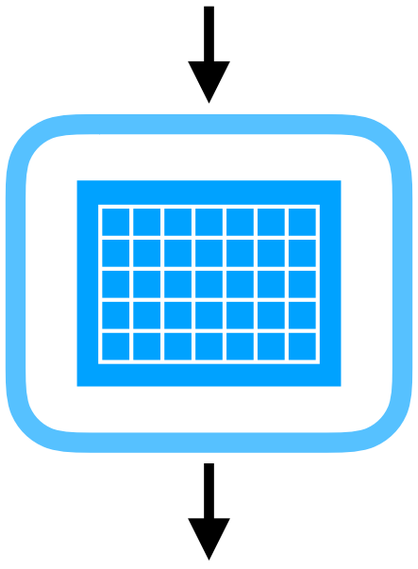


Integrating data with services (2)

```
https://glintcore.net/izzy/ocean?show(t,air_temp_avg,wind_speed)
```



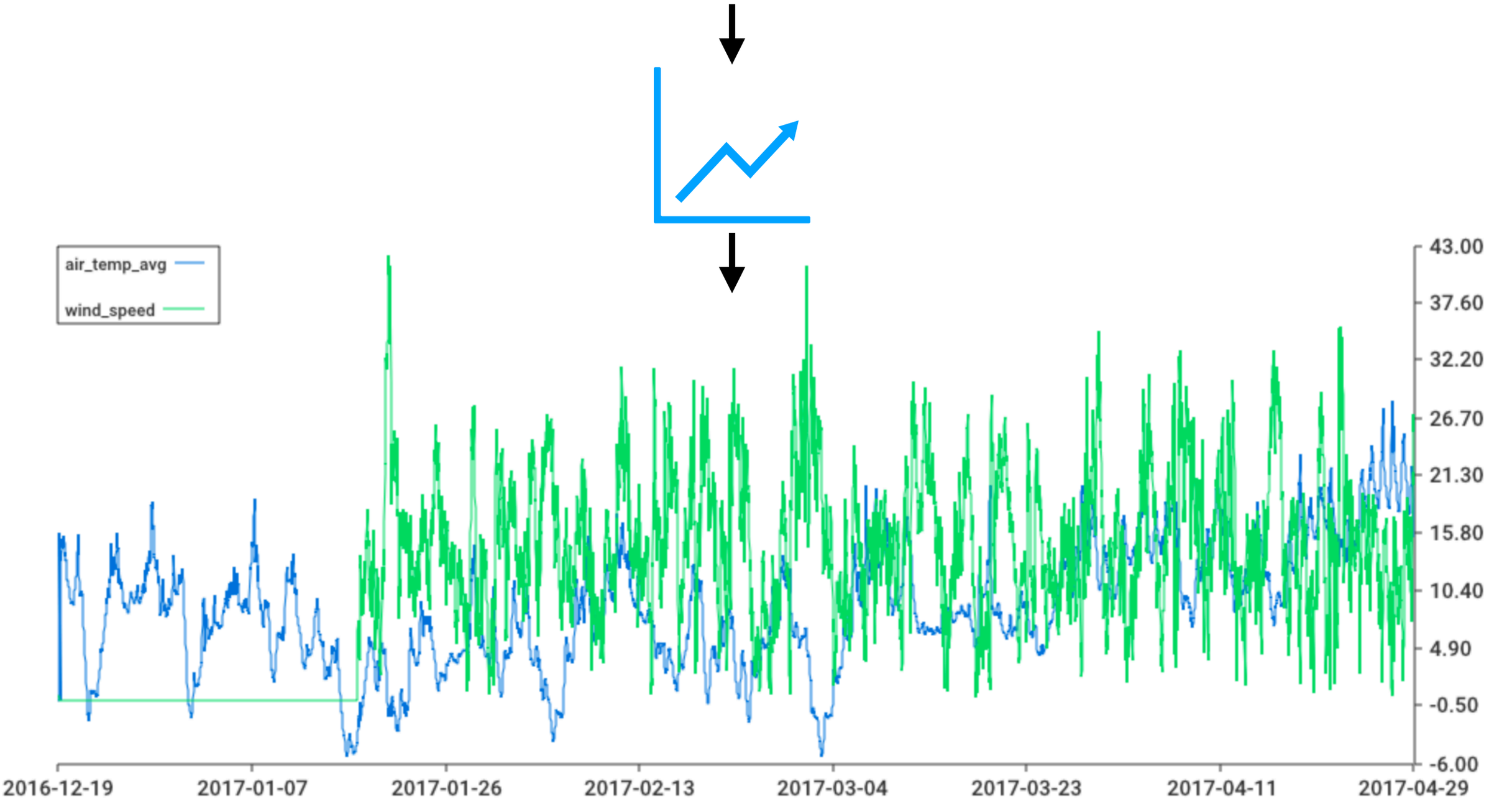
```
https://glintcore.net/izzy/ocean?show(t,air_temp_avg,wind_speed)md()
```



```
t{dc:date},air_temp_avg,wind_speed{yamz:h3846}
```

Integrating data with services (3)

```
t{dc:date},air_temp_avg,wind_speed{yamz:h3846}
```



Integrations

Metadata

URL

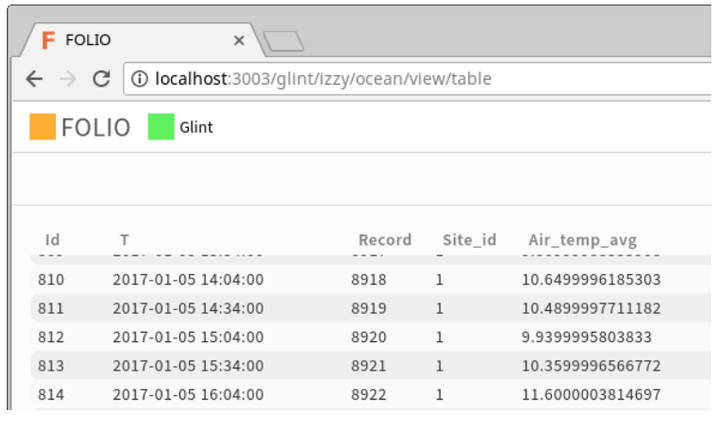
Client

Data formats

Authentication / permissions

Server storage

FOLIO



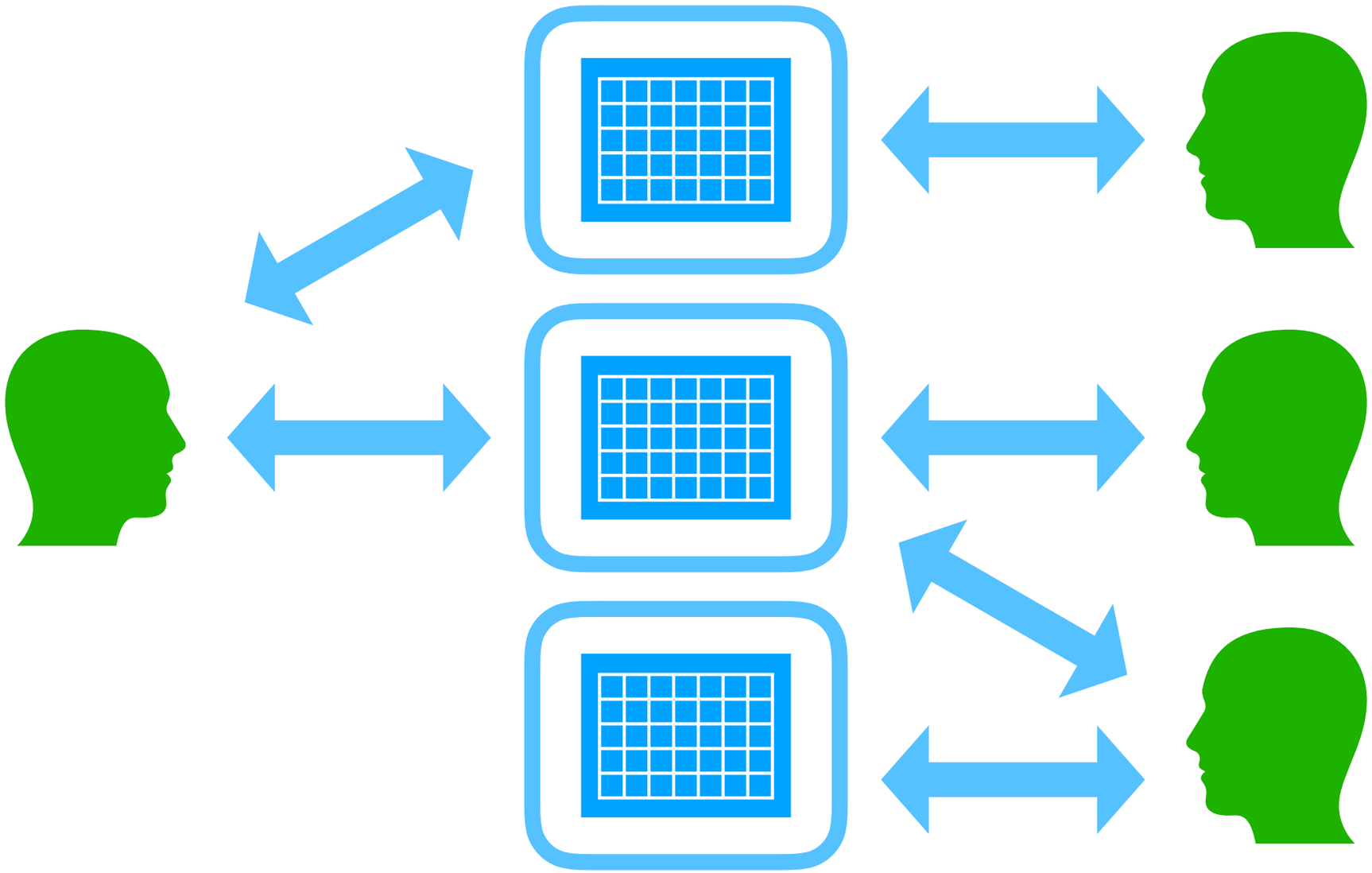
"The generation of most biomedical data is highly distributed and is accomplished mainly by individual scientists or relatively small groups of researchers. Moreover, data also exist in a wide variety of formats, which complicates the ability of researchers to find and use biomedical research data generated by others and creates the need for extensive data 'cleaning.' According to a 2016 survey, data scientists across a wide array of fields said they spend most of their work time (about 80 percent) doing what they least like to do: collecting existing data sets and organizing data. That leaves less than 20 percent of their time for creative tasks like mining data for patterns that lead to new research discoveries."

—Draft NIH Strategic Plan for Data Science (2018)

"The value of research data arises from its use, and the more it is used the greater the social benefits and the higher net welfare."

—Business models for sustainable research data repositories (OECD report, Dec. 6, 2017)

Effective data sharing can accelerate cooperation around data



Suppose that we could share and cooperate around data—forming communities to discuss and understand data better—as easily as we share and discuss interesting articles today

`https://glintcore.net`