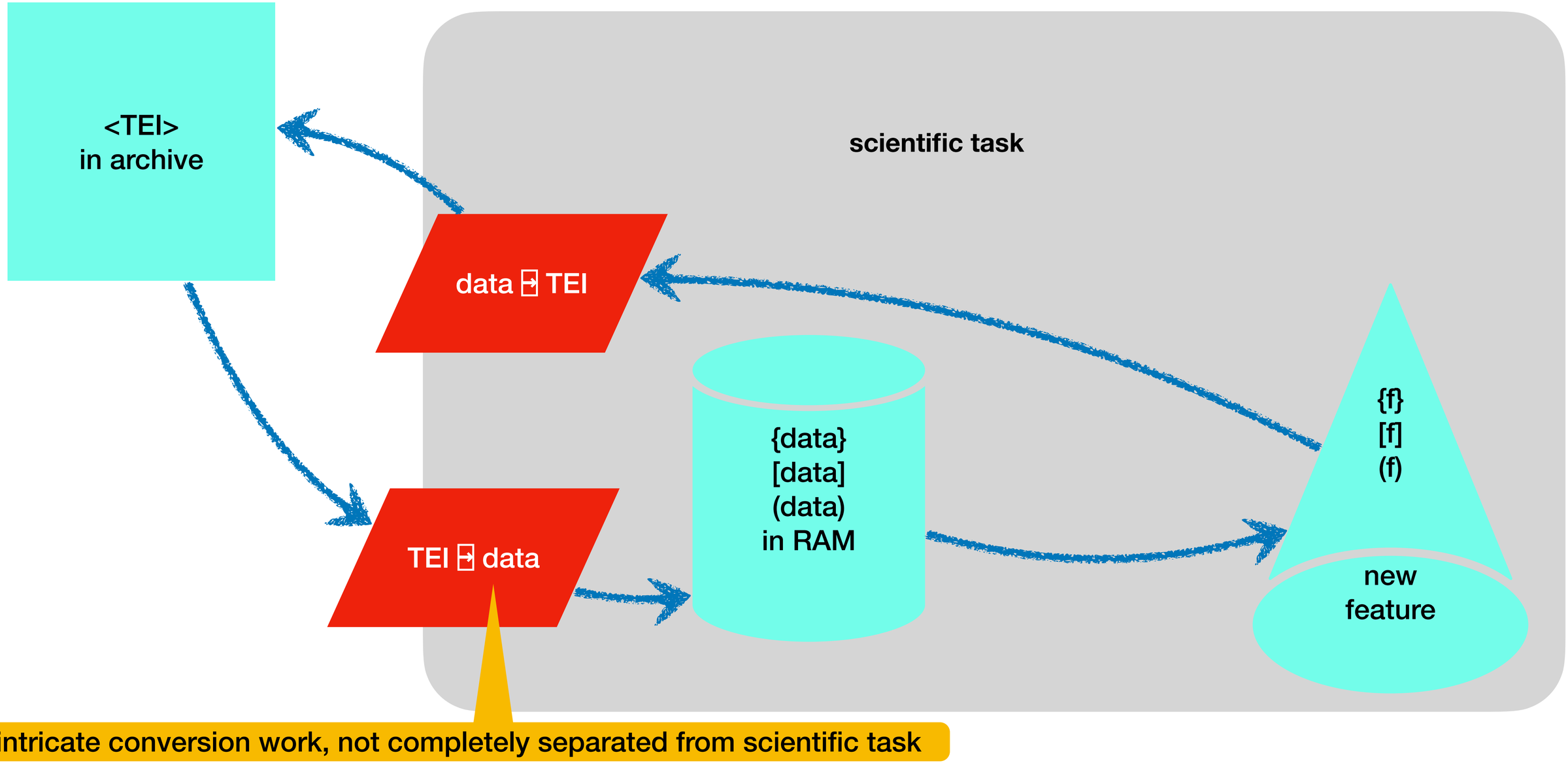
corpus enrichment workflow

Dirk Roorda

2021-05-19

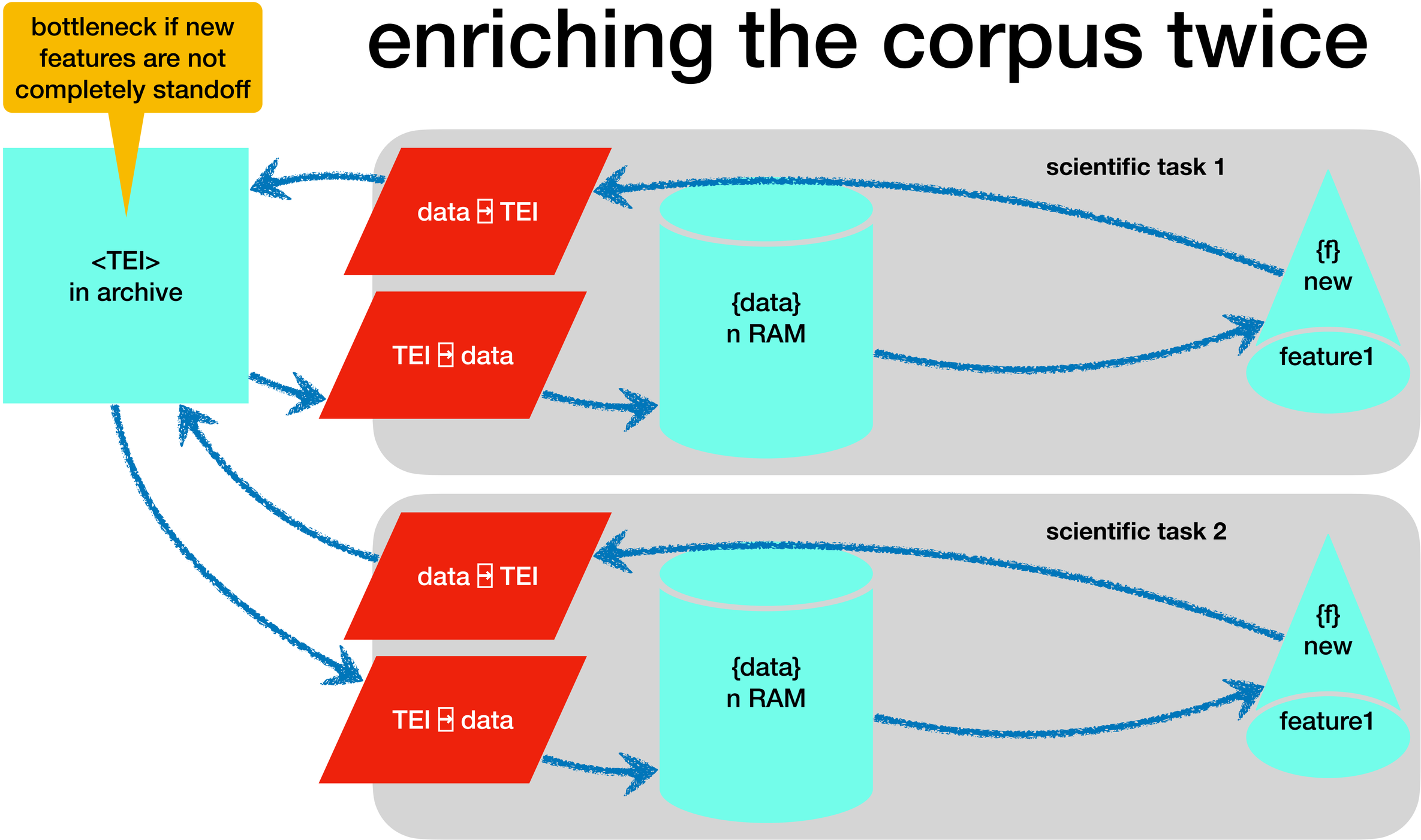
enriching a TEI corpus



problem: multiple scientific tasks

- several enrichments (linguistic, knowledge, interpretation)
- done by independent teams
- who do not mutually align their activities

enriching the corpus twice



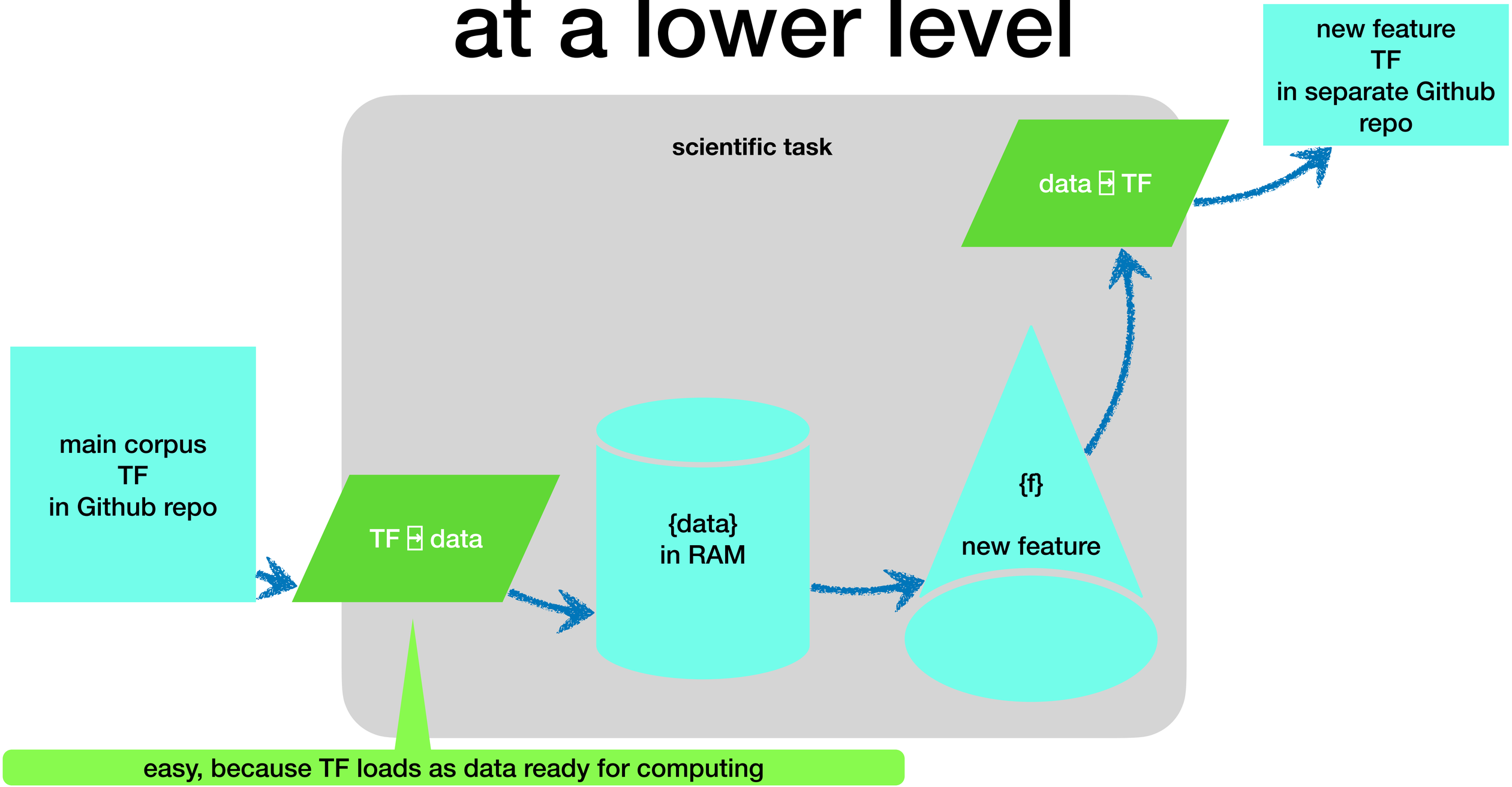
issues

- conversion interferes with scientific work
- each group adjusts conversion to own purposes
- duplication of work between groups, difficult to disentangle
- the logistics of bringing back results to the archive is difficult

at a more technical level

- scientific workflows need data in generic structures
- we can turn a corpus in such a structure once and for all
- enrichments are new data in the same structures but physically separate
- enrichments are stored where they are created
- there is support for auto-downloading data from GitHub
- not from one repo, but from multiple repos

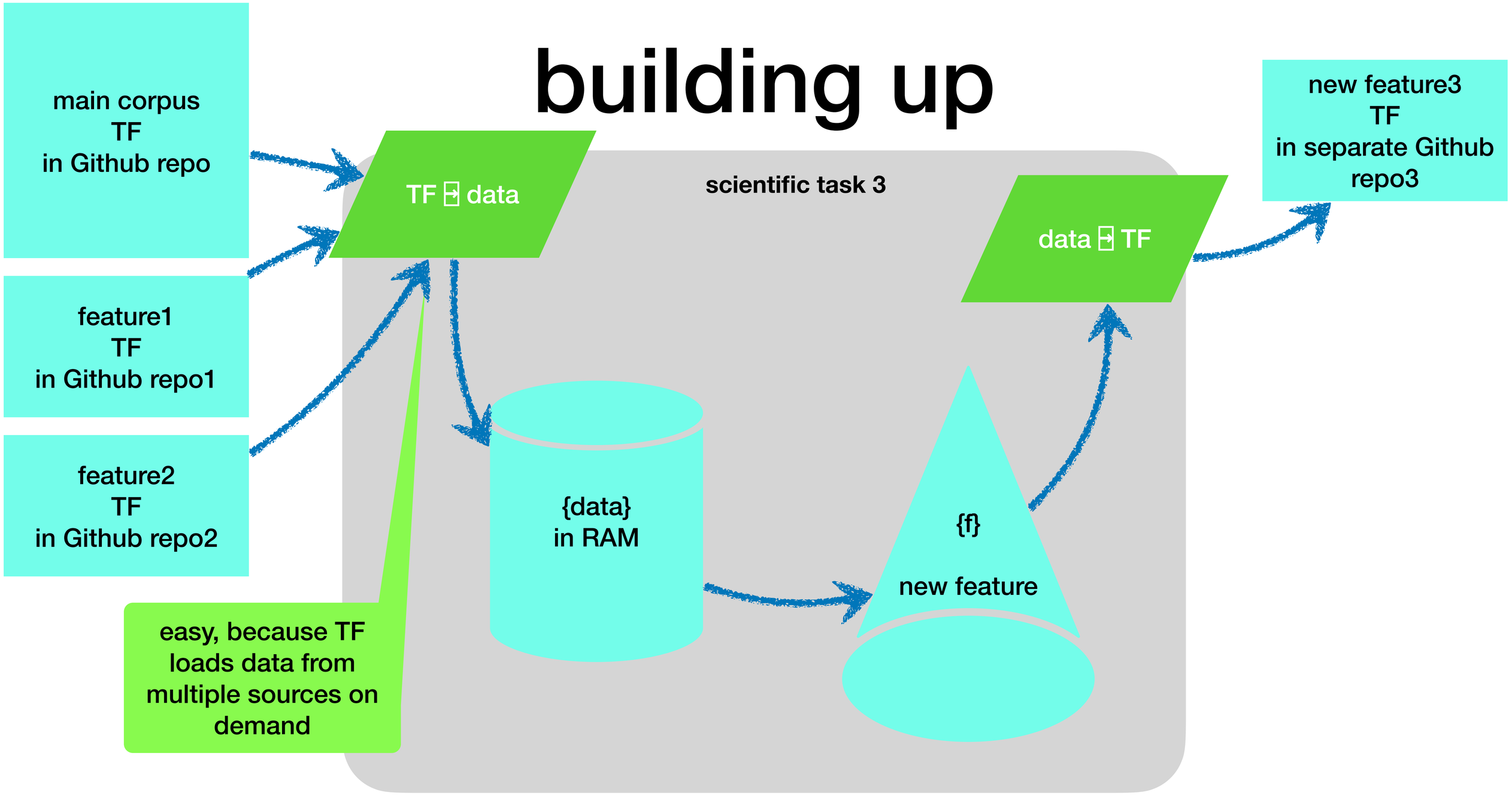
at a lower level



build on top of each other

- because of radical separation of concerns and because of radical standoff of enrichments:
- we can stack up the scientific work of multiple groups
- and make it available to yet other groups

building up


main corpus
TF
in Github repo
feature1
TF
in Github repo1
feature2
TF
in Github repo2
TF ⍈ data
scientific task 3
{data}
in RAM
{f}
new feature
data ⍈ TF
new feature3
TF
in separate Github
repo3
easy, because TF
loads data from
multiple sources on
demand

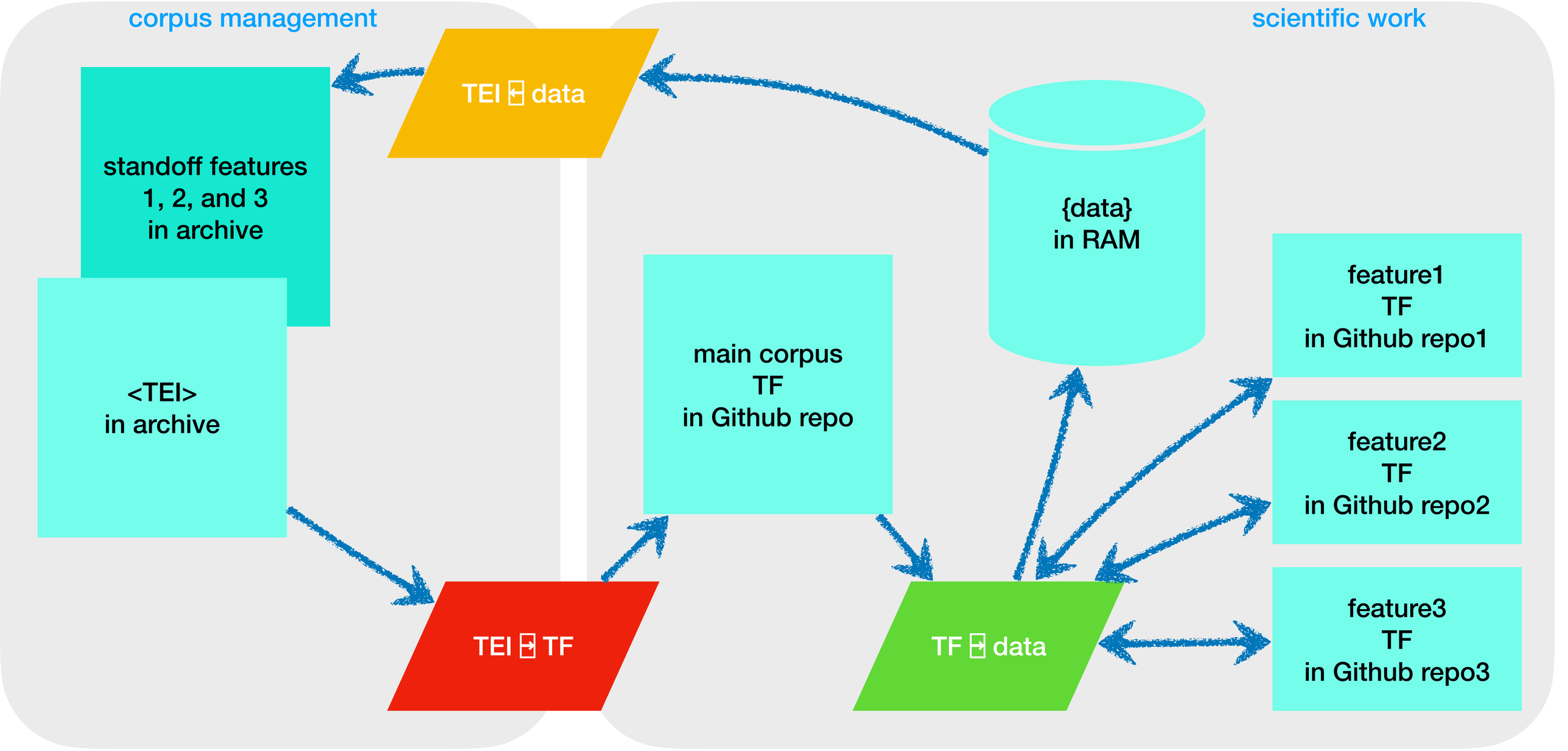
traits of the lower level

- distributed authorship of scientific features
- agile merging of corpus and existing features
- separation of concerns when building scientific features

At this level we can support complex, distributed, asynchronous workflows.

And we can bring the results back to the higher level.

bringing it back to TEI



separation

- the scientific work and the corpus maintenance are now separated
- with a clear interface between them
- Text-Fabric is just an example of a toolkit that support this state of affairs
- The crucial principles are
  - stand-off annotation (physical)
  - separation of concerns