The SDM Center Data Integration Effort and Beyond

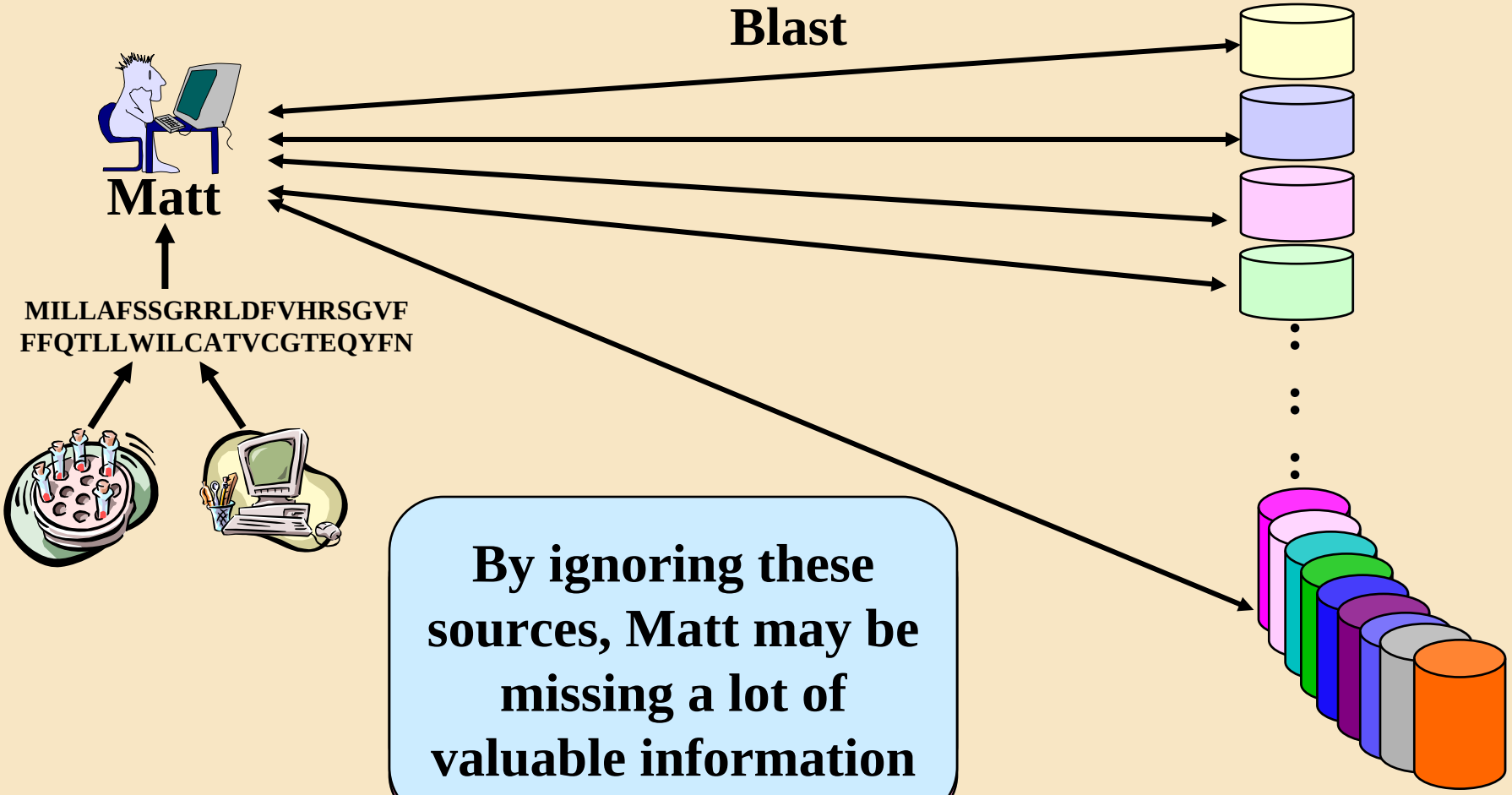
Terence Critchlow

Center for Applied Scientific Computing

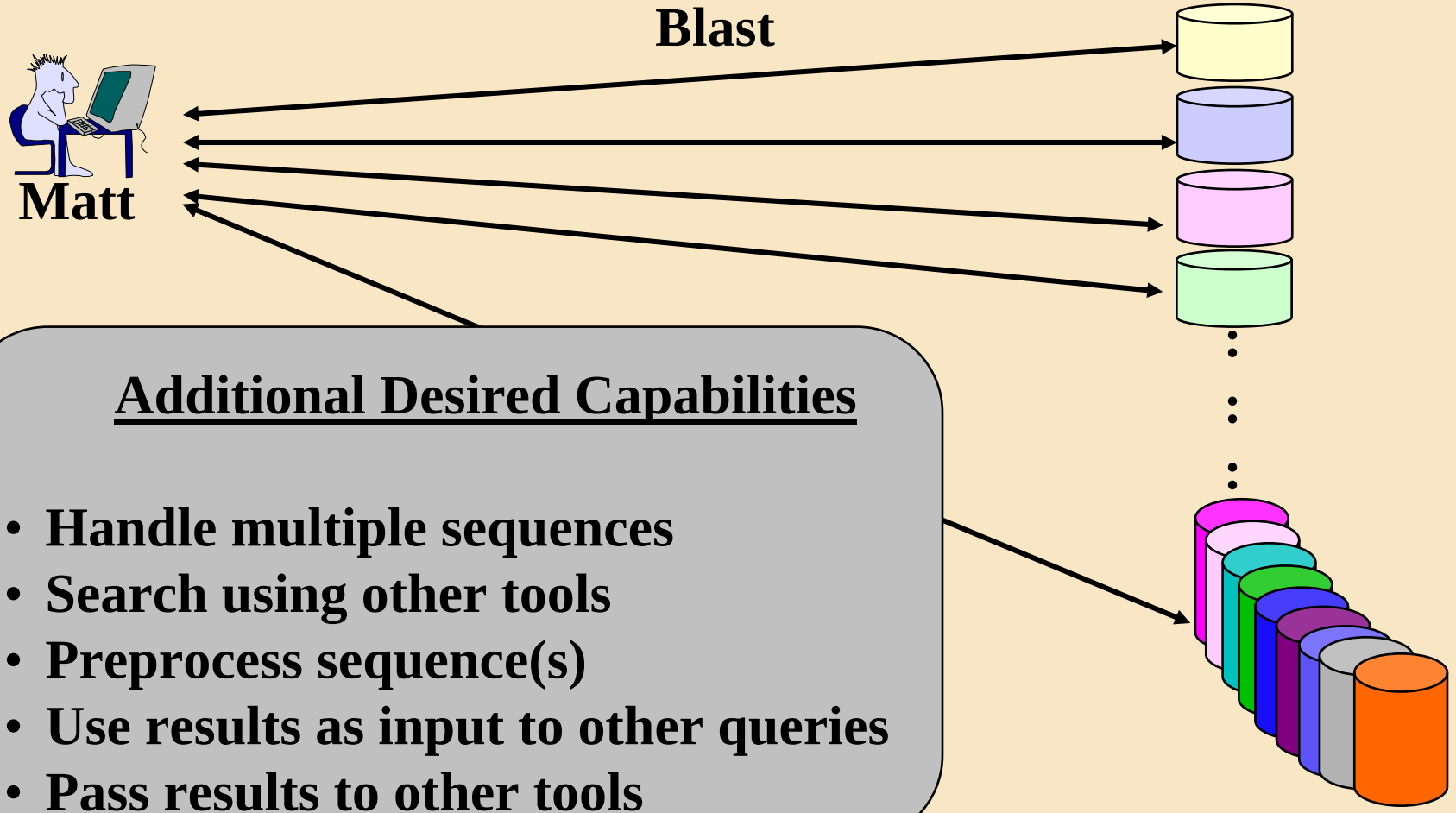
Lawrence Livermore National Laboratory

January 2002

Example:
Find everything related to a sequence

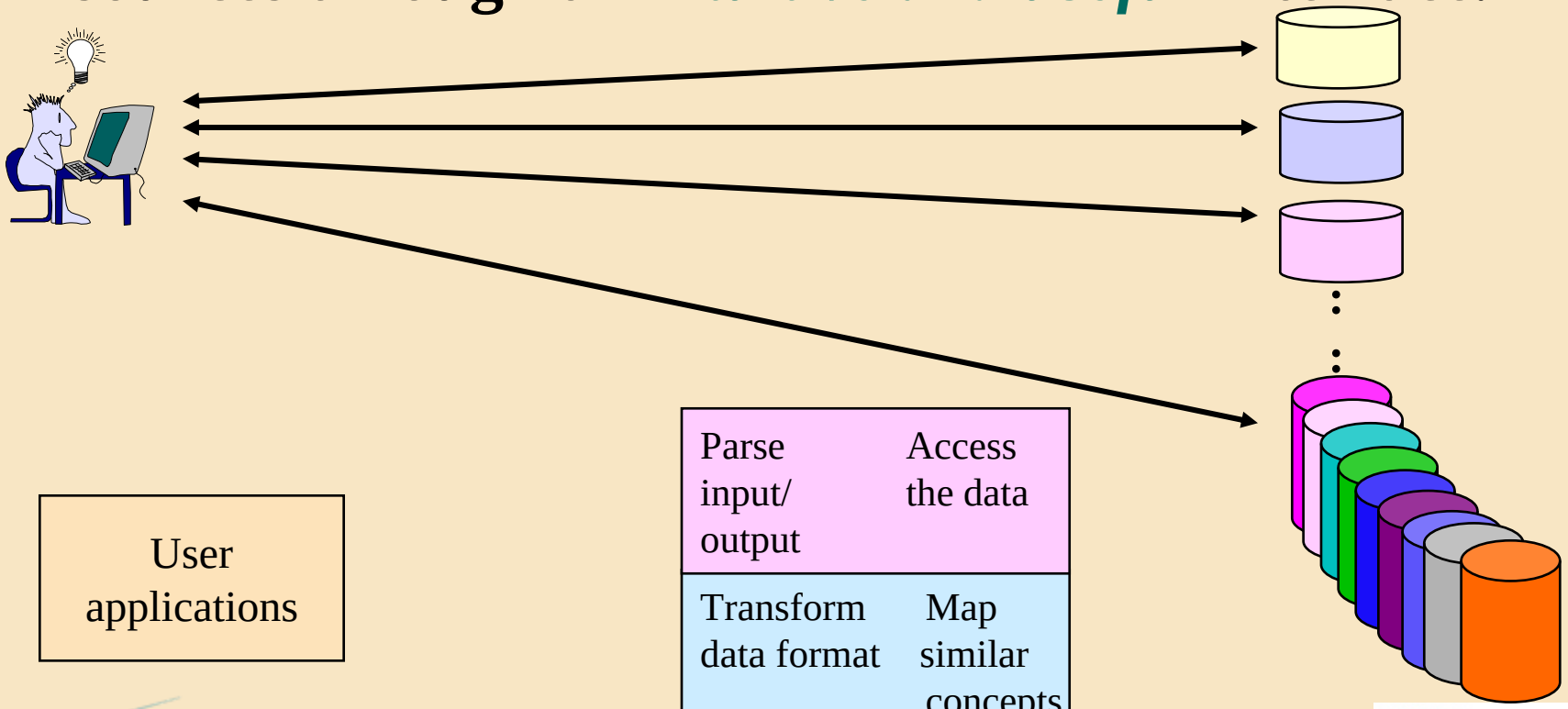


Example:
Find everything related to a sequence



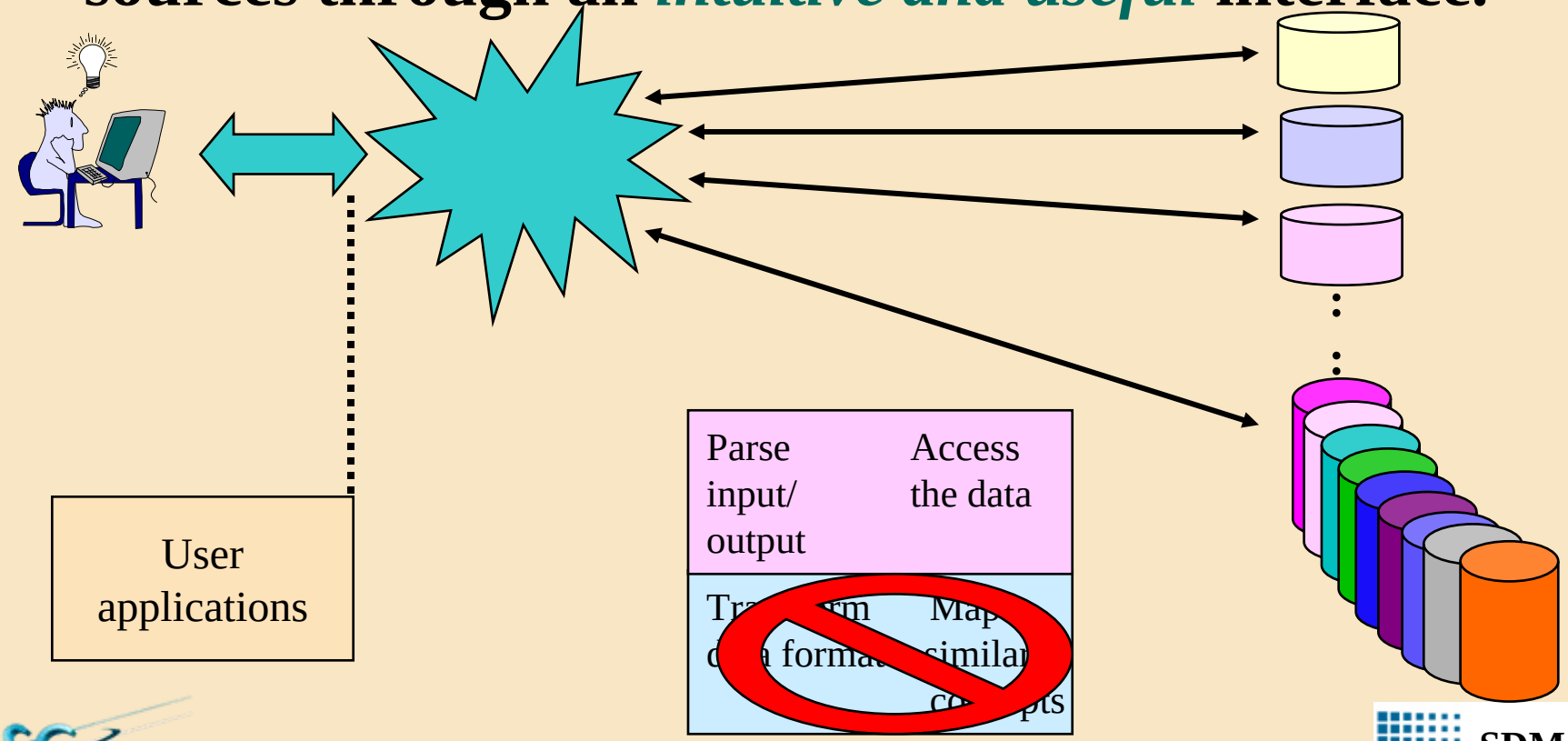
What is the ideal environment?

A *single* location that provides *effective* access to a *consistent* view of data and tools from *many* sources through an *intuitive and useful* interface.

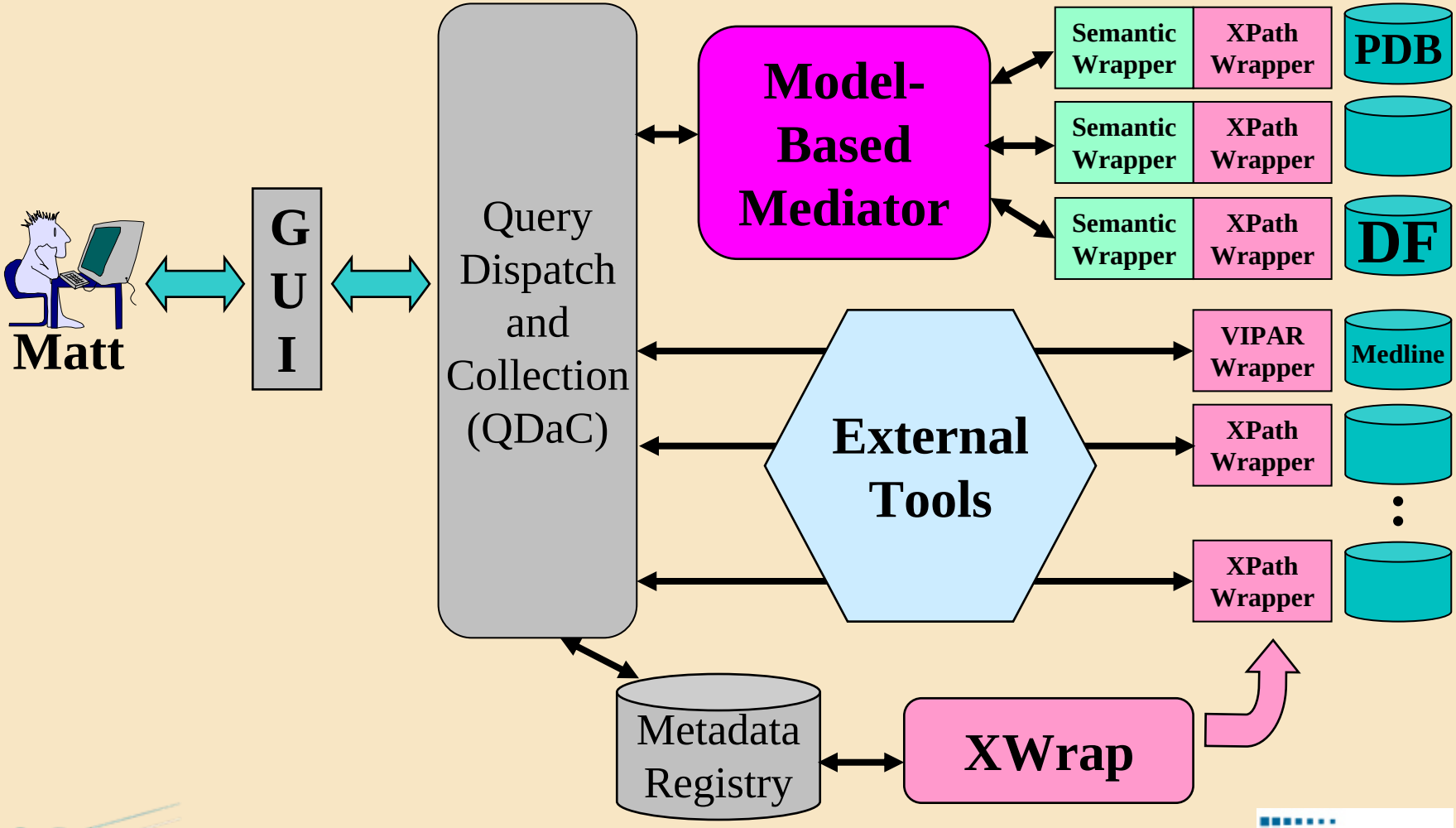


What is ~~the ideal~~ a realistic environment?

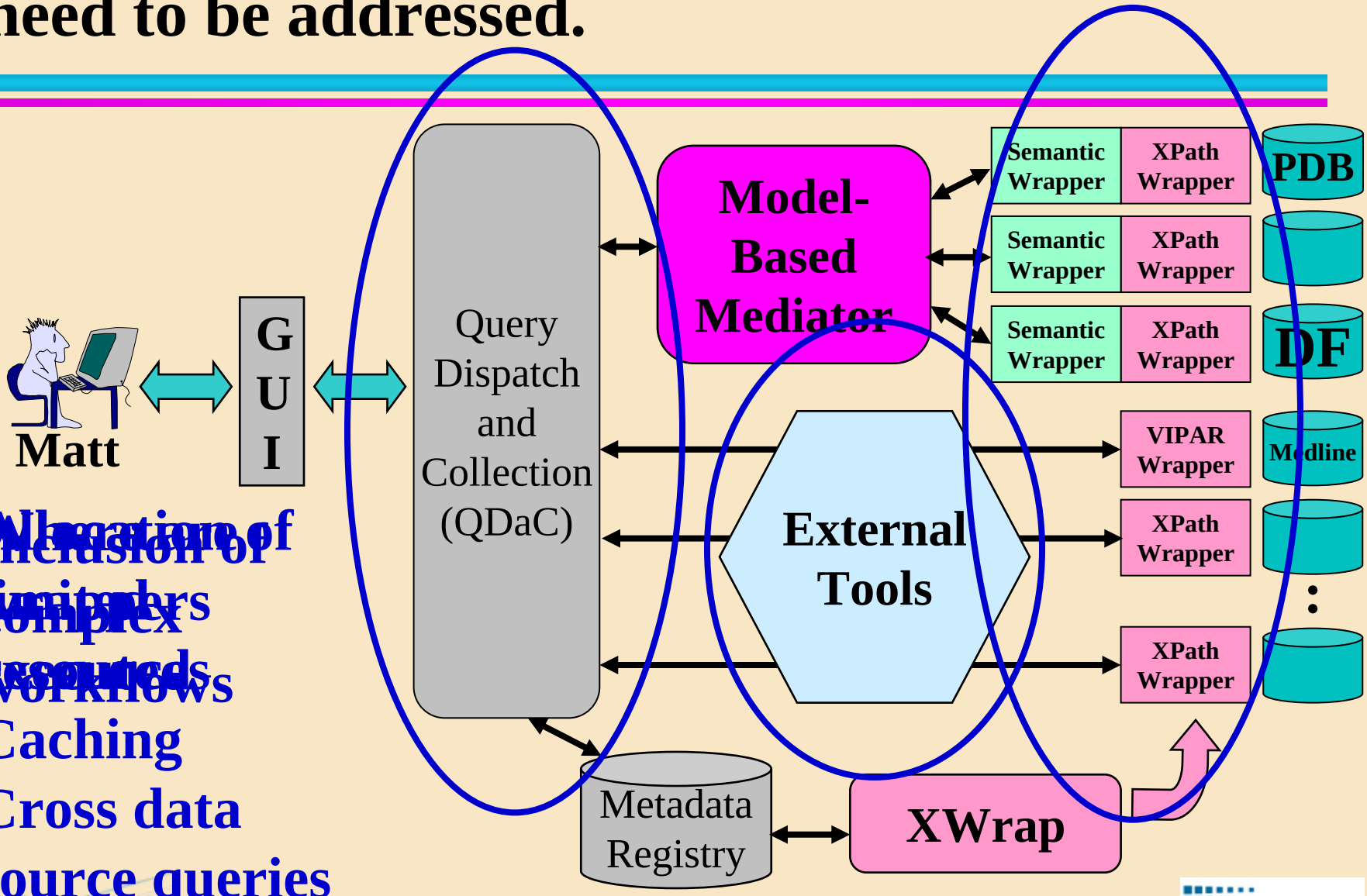
A *single* location that provides *effective* access to ~~a *consistent* view of~~ data and tools from *many* sources through an *intuitive and useful* interface.



SDM Center Data Integration Infrastructure

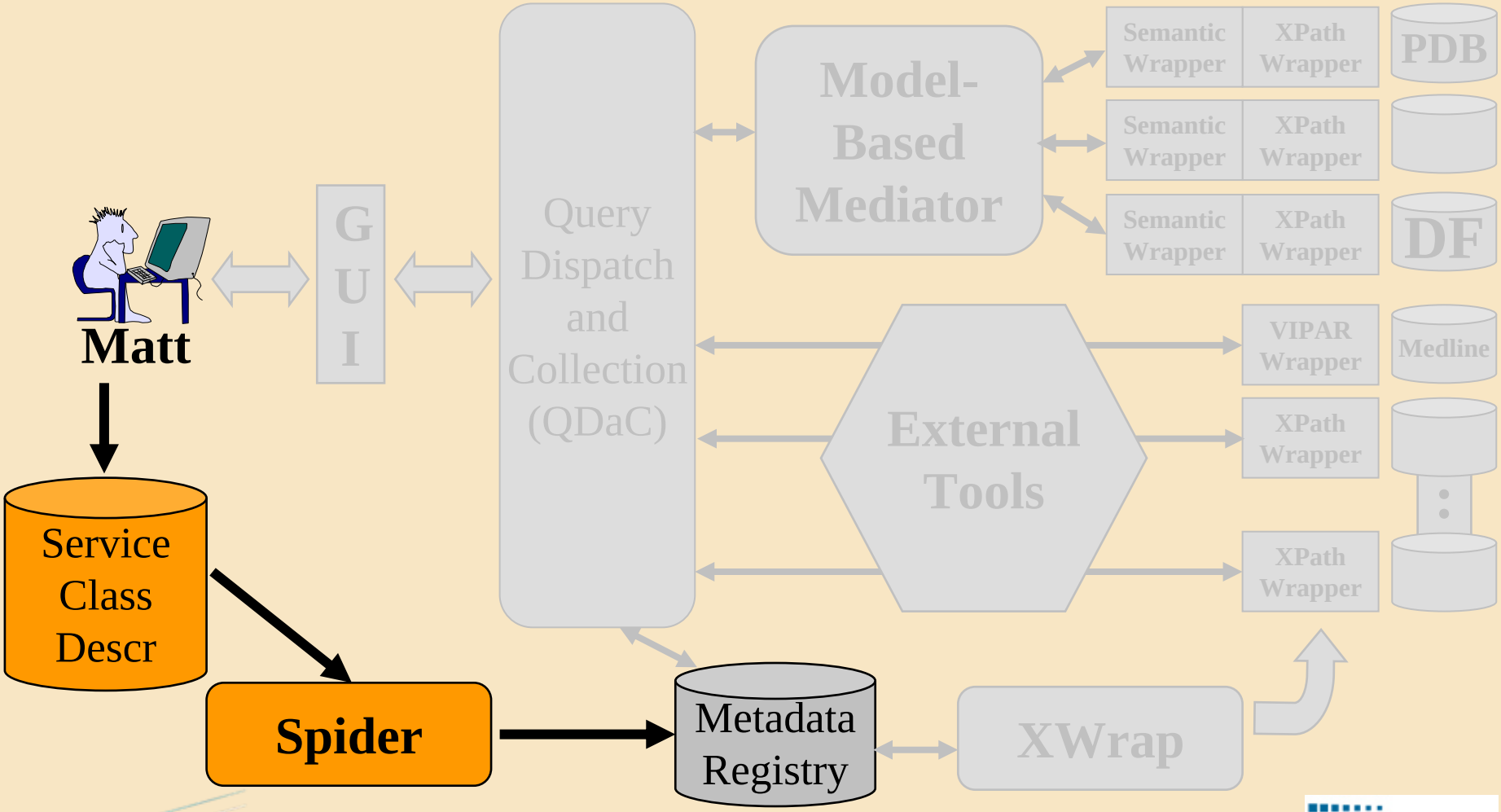


There are a lot of CS research issues that still need to be addressed.



- Integration of complex workflows
- Caching
- Cross data source queries

How does this contribute to a scalable infrastructure?



Standards – why don’t we have them yet?

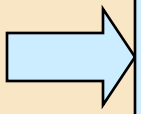
Standards imply semantics

Semantics are HARD!

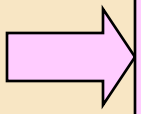
Standards – why don't we have them yet?

Challenges

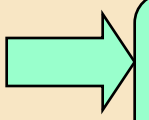
- Genomics is a complex field where there are more exceptions to the rules than rules themselves
- Technology is constantly evolving and the terminology has to keep up
- Different genomics communities use the same terms in different ways



Able to express the complex concepts found in this domain



Extensible while retaining backward compatibility



Interaction between multiple standards

What is the answer?

What is the answer?

- **Forced standards**
  - **Won't work in a evolving scientific environment**
- **Ontologies are becoming popular**
  - **DAML OIL**
    - **XML based representation for ontology exchange**
    - **Is being promoted as an approach to dealing with this problem**
    - **Unclear whether it will be sufficiently robust for this environment**

Scientists need to decide semantics are important enough to focus time and energy on

Conclusions

- **Efforts are beginning to address data accessibility issues**
  - **SciDAC SDM Center - data integration infrastructure**
  - **DataFoundry - scalable data access**

- **Providing consistent semantics is one of the biggest challenges remaining**
  - **Need support from scientists if current efforts are to be successful**

People

LLNL

- Terence Critchlow (lead)

Georgia Tech

- Calton Pu
- Ling Liu
- David Buttler
- Dan Rocco
- Henrique Paques
- Wei Han

SDSC

- Bertram Ludaescher
- Amarnath Gupta
- Ilkay Altintas

Agent Technology

- Tom Potok (ORNL)
- Mladen Vouk (NCSU)

Target Users

- Matt Coleman (LLNL)
- Allen Christian (LLNL)
- Phil Bourne (PDB)

Questions?

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-ENG-48.