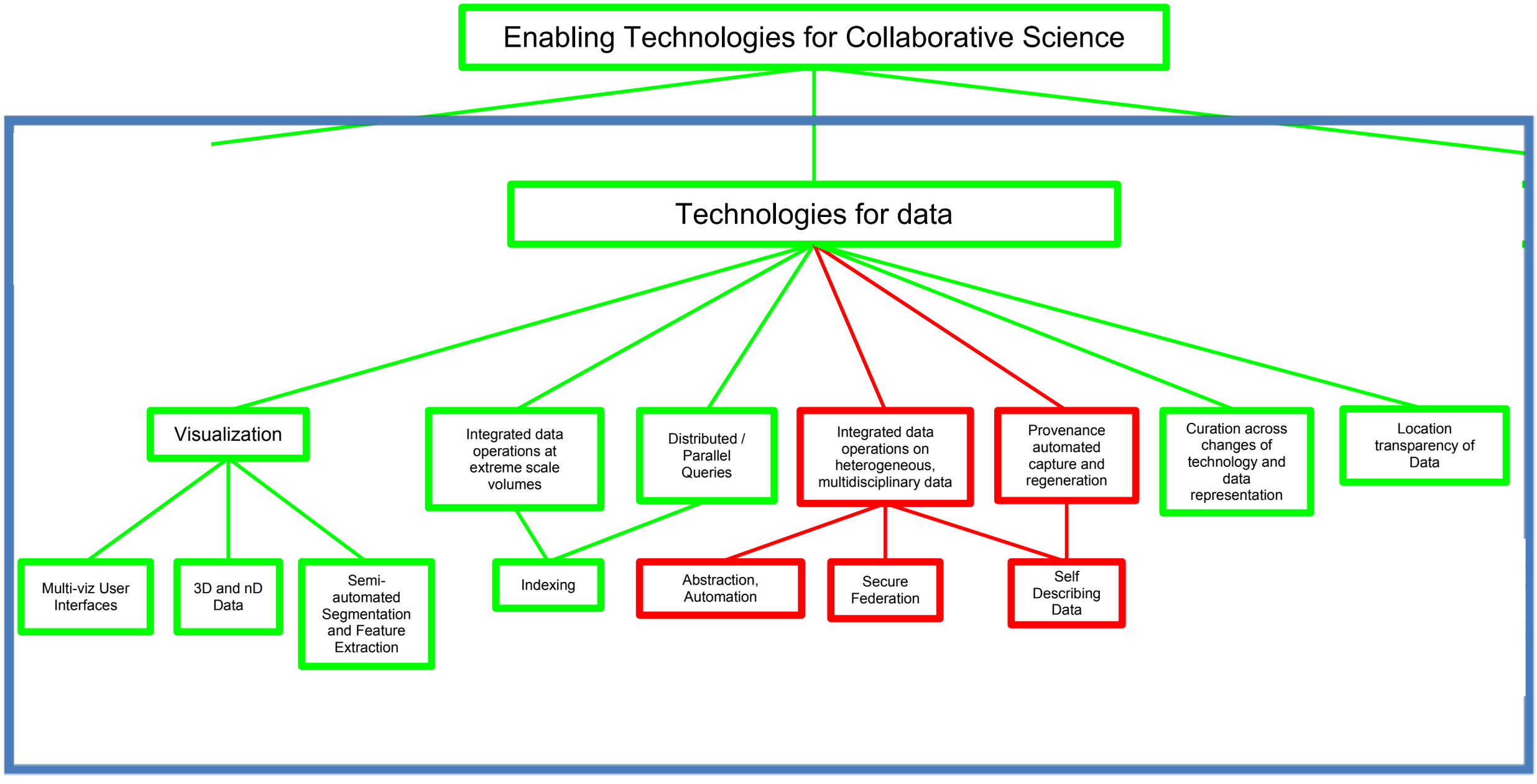
Technologies for Data Breakout

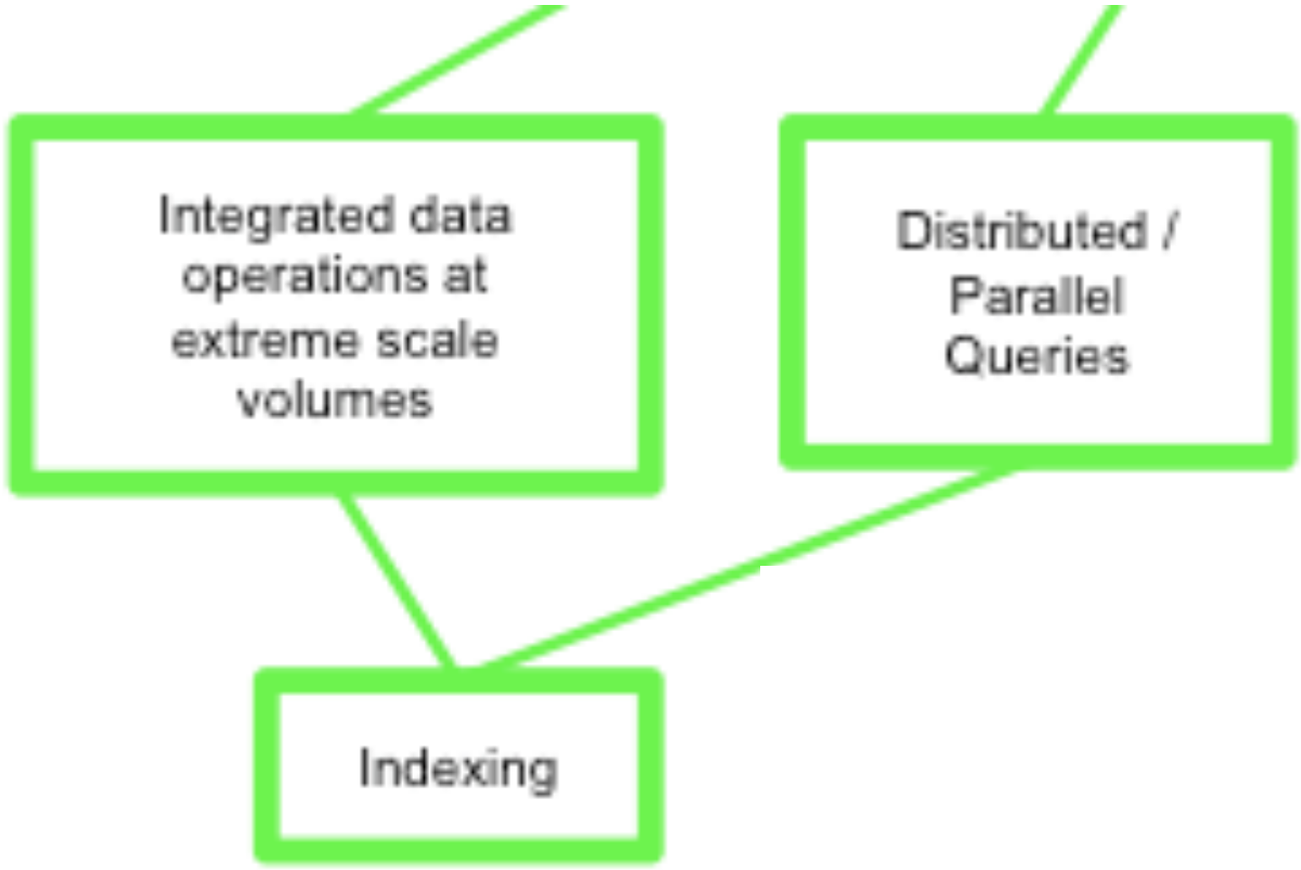
The needs for R

(under the implied assumption that D is also very important)

Summary Slides


Enabling Technologies for Collaborative Science
Technologies for data
Visualization
Multi-viz User Interfaces
3D and nD Data
Semi-automated Segmentation and Feature Extraction
Integrated data operations at extreme scale volumes
Indexing
Distributed / Parallel Queries
Abstraction, Automation
Integrated data operations on heterogeneous, multidisciplinary data
Secure Federation
Provenance automated capture and regeneration
Self Describing Data
Curation across changes of technology and data representation
Location transparency of Data

- Revisit boxes for scalable data stores. Can three be collapsed into one?



Individual Research Areas we did not prioritize as a group:

- Research to ensure scaling of access to petabyes of data stored as blocks/datasets on disk/in databases.
  - Applicable to databases and files.
- Ensure data I/O keeps pace with capabilities of the underlying hardware systems.

- Research to support access/querying on data across heterogeneous data sources – within a or across multiple science domains.
- Research into integration/embedding of statistical methods into querying layers/services.
  - Being done elsewhere but certainly needed for collaborative use of data.

- Research into signal processing algorithms, techniques and methods are applicable in collaborative (as well as non-collaborative) environments.
  - Frameworks for collaboration that allow dynamic inclusion of these

Research into Repeatable, Regeneratable, Trustable data

- Semantics
  - Is XML a usable standard here? Is there research into how to make it more easily accessible to scientists?
  - Discover and involve experts at the right time and place in the scientific collaboration lifetime.
  - Continuous – never ends.

Research into Repeatable, Regeneratable, Trustable data

- Provenance
  - Full, accurate, evolvable, validated.
  - Plus that needed, extensibility, for multi-decade lifetime datasets where the format/storage/semantics need to be migrated (multiple times) to new technologies.

Research into standard Provenance and instrument/convert existing data, apply Google tools and test data discovery, utility of the reuse, method etc.

- Is this Experimental Computer Science or Computer Science?

Research into discoverable/auto-generated provenance information for complex and long-lived data.

- value proposition was under discussion.

Research into the architectures, definitions, common interfaces for

- General frameworks that support access to dynamically changing, arbitrary and evolving algorithm codes.
- Research in ability to effectively share the storage of data and dynamically readjust the storage available, with the access rights, and the clean up, in a collaborative environment

Research into frameworks to Provide user/scientist access and analysis across multi-disciplinary data sources.

- Prototype this across 3 disciplines (x2 teams?)
- How to prove this is general “enough” is “generalizable”
- Remember instrumentation and monitoring when don’t function as expected.
- Is this covered sufficiently elsewhere?
- Most useful in Bio. (cf Kbase)

Research into algorithms, statistical and analytic methods

- Useful for collaborative as well as non-collaborative science, but still very important.
- Some differences when delivering to a team working together

? Research program to

Build a prototypical " framework" /collaborative tools suite/ interoperable toolset* in the context of a specific workflow (chosen by someone else) that incorporates all/most of the capabilities of "Technologies of Data" except for Visualization and adds "pluggable data transformation" and "discover/search" and

demonstrate this framework works at small, medium and large scales, and that it can be extended/reduced to apply to a second/third workflow.

? Independent, interoperable, integratable tools ?

a matter of degree?