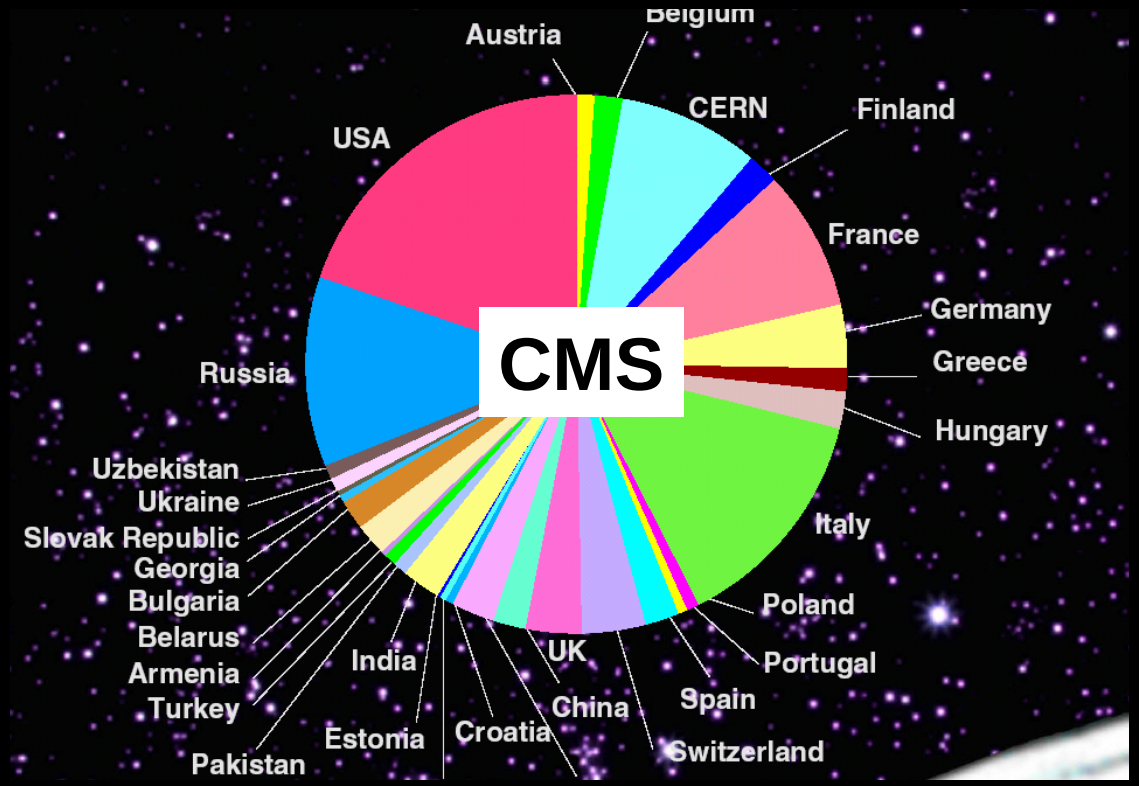
# Facilitating Science Collaborations for the LHC: Grid Technologies

**Richard Cavanaugh**
**University *of* Florida**

**Shaping the Future of Collaboration in Global Science Projects**
**11-13 December, 2006, CICG, Geneve**

# Big Science ⇒ Global Collaborations

- No single country can build the LHC
  - Too much money
  - Worldwide distributed technical expertise
- E.g. HEP collaborations each consist of:
  - ~1000s scientists
  - ~100s institutes
  - ~10s countries




- Massive data collections belong to the participating world community
  - All participating collaborators get equal access to data
  - Democratisation of science

# Problem Solving at the LHC

## Technical Challenges

- One of the most complex instruments ever built by humankind
  - The LHC Accelerator
  - The four LHC Experiments
- Network intensive:
  - From ~200 Gbps (2008)
  - To ~1 Tbps (2013)
  - Across & among world regions
- Data and computationally intensive
  - From Petabyts (2008) to Exabytes of Shared Data
  - $10^5$ processors evolving with technology; $10^5$ jobs

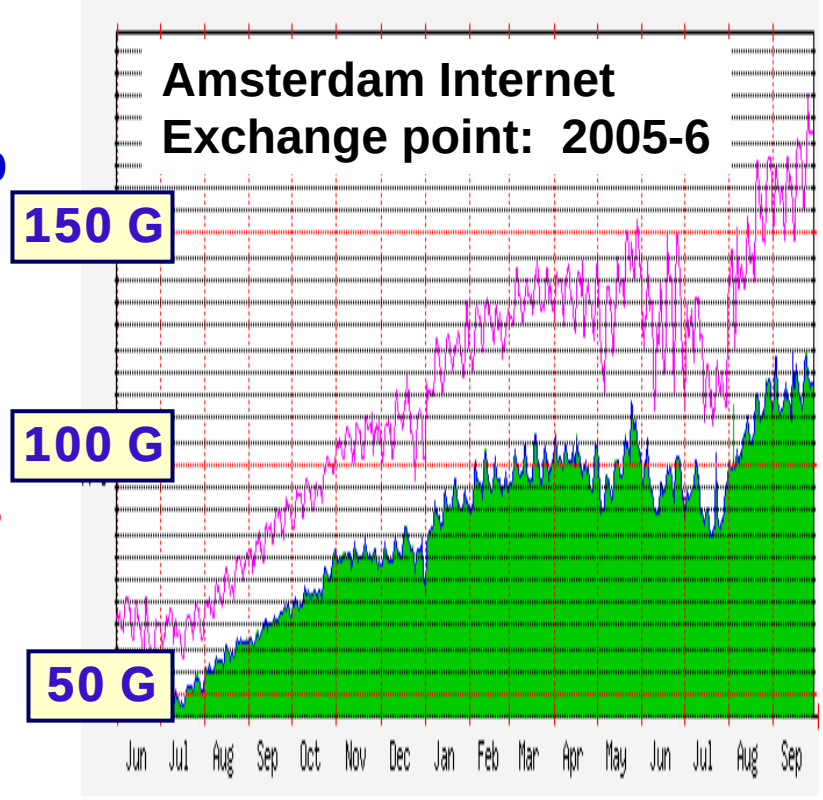
## Social Challenges

- Teams organized around common goals
  - Communities: "Virtual organizations"
- Diverse membership & capabilities
  - Heterogeneity is a strength not a weakness
- Geographic and political distribution
  - No location/organization possesses all required skills and resources
- Must adapt as a function of the situation
  - Adjust membership, reallocate responsibilities, renegotiate resources

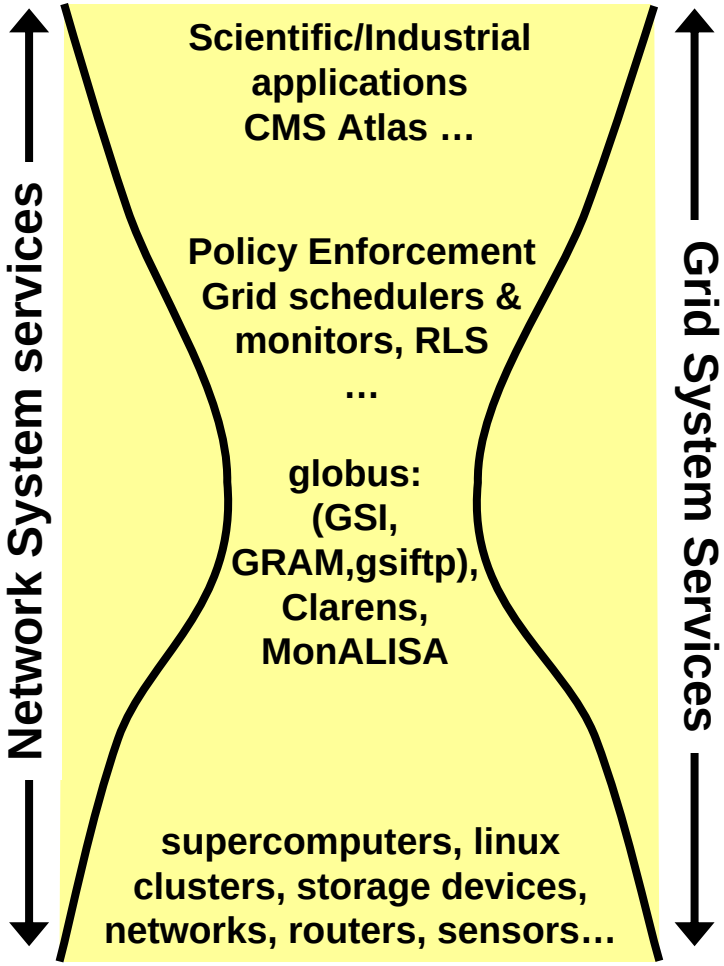
# Collaboration size also drives need for new technology

- **In the 80's-90's HEP transitioned to 100's of collaborators per experiment**
  - **How best to manage?**
  - **How best to share information, internationally?**
  - **Development of the World Wide Web**
- **LHC represents "phase transition" in collaboration size → 1000's of collaborators**
  - **how best to govern?**
  - **how best to work together, globally?**
  - **Development of the Grid**

**Networking is key to facilitating collaboration**
**→ increasing number of people who need to share increasing amount of information!**



# Early Architecture of the "Grid"




- User applications
  - Scientific applications, industrial tools, games…
- Grid service applications
  - Tools necessary to control and monitor access to the Grid
  - Directory brokering, Grid monitoring and diagnostics, Grid schedulers…
- Grid middleware layer
  - Software that ties it all together
  - Presents consistent, secure and dependable interface for grid apps
- Fabric layer:
  - Computers, Storage, sensors and networks
  - Provide the physical resources and raw capabilities to the infrastructure

**Evolving Architecture: As complexity scale increases (data, users, algos)**
**→ increasingly automated (hide & manage complexity)**
**→ fully distributed, intelligent-agent based arch. (improve robustness)**

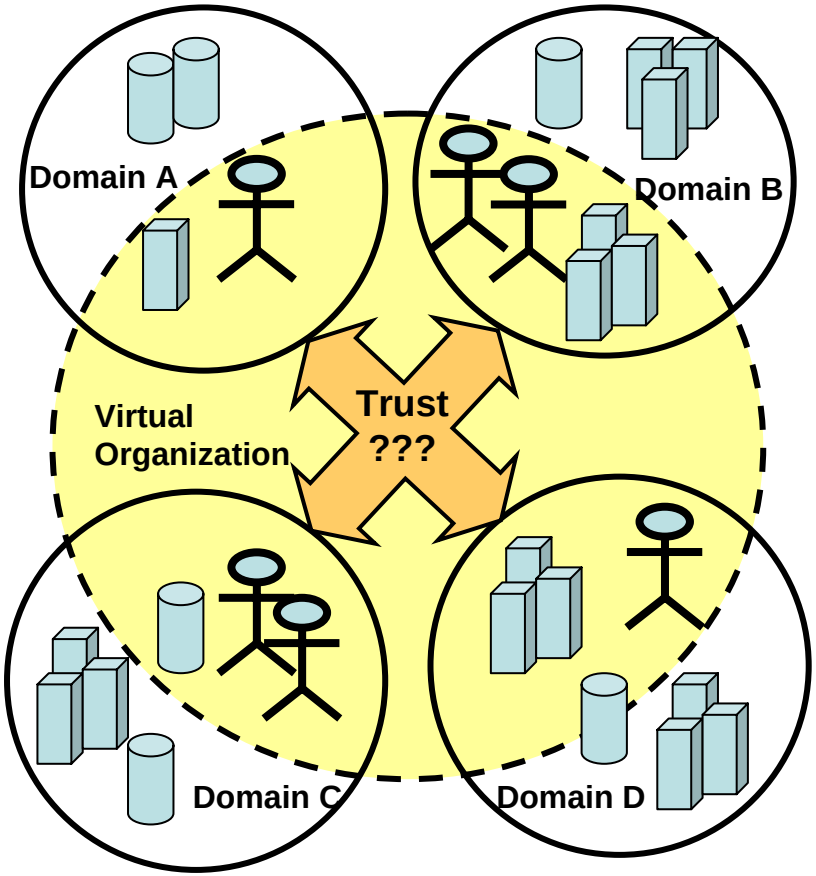
# LHC Data Analysis – Essential Components

- **Data Processing:** All data needs to be reconstructed, first into fundamental components like tracks and energy deposition and then into "physics" objects like electrons, muons, hadrons, neutrinos, etc.
  - Raw → Reconstructed → Summarized
  - Simulation, same path. Critical to understanding detectors and underlying physics.
- **Data Discovery:** We must be able to locate events of interest (Databases)
- **Data Movement:** We must be able to move discovered data as needed for analysis or reprocessing (Networks)
- **Data Analysis:** We must be able to apply our analysis to the reconstructed data
- **Collaborative Tools:** Vital to sustaining global collaborations
- **Policy and Resource Management:** We must be able to share, manage and prioritise in a resource scarce environment

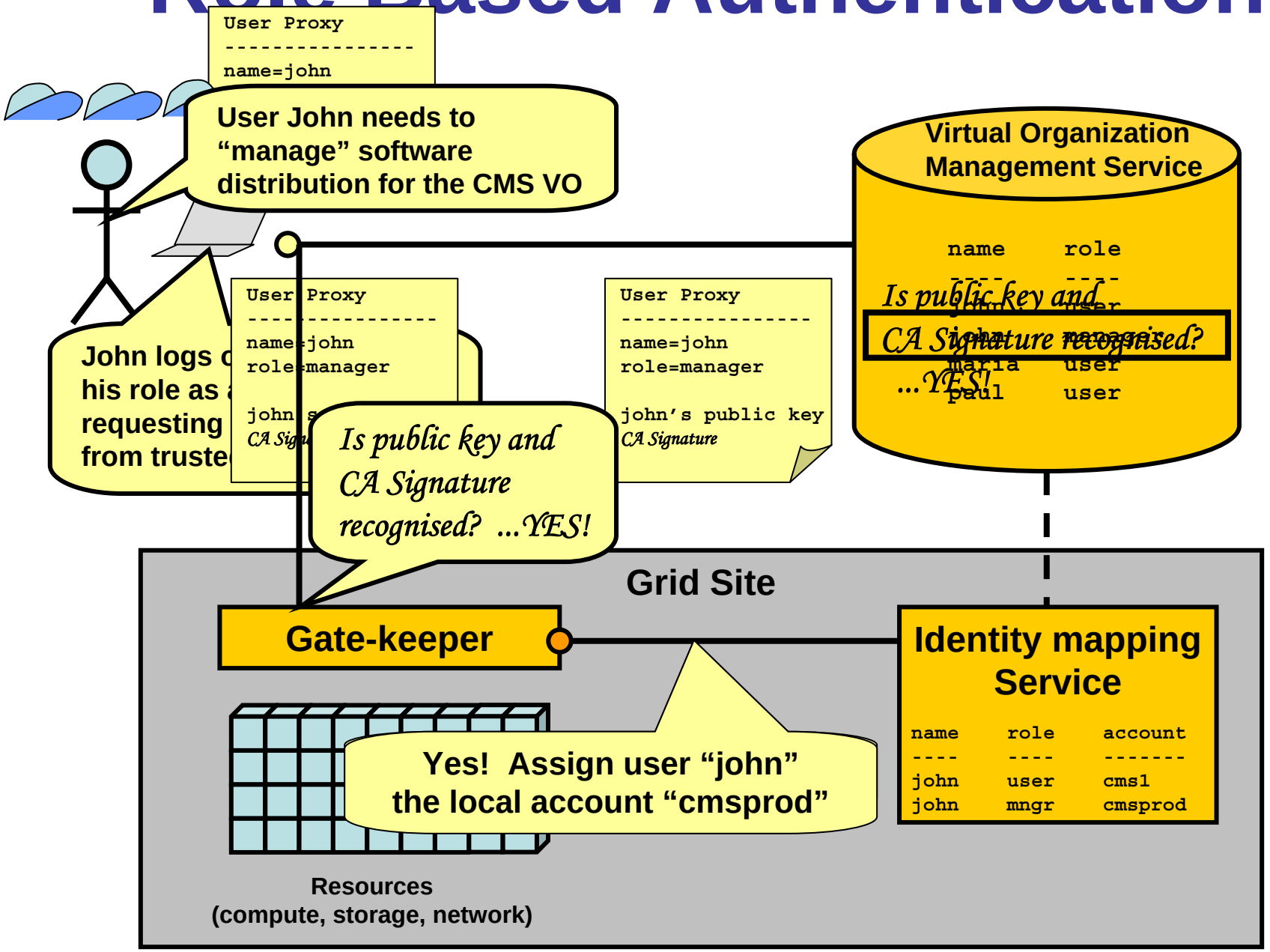
**All components integrated in end-to-end system!!**
**System able to automatically and intelligently adapt!**

# Virtual Organization

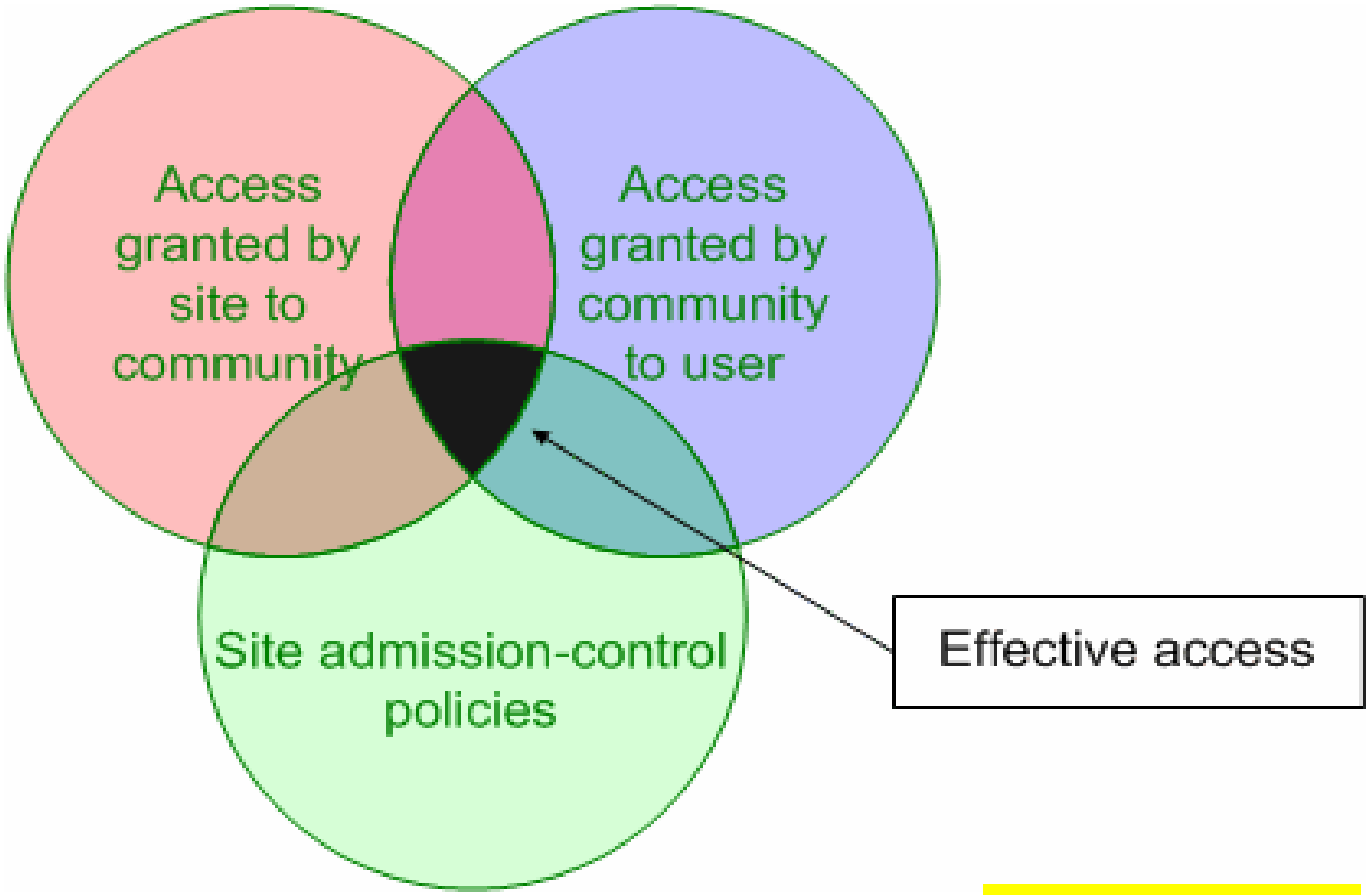
- **Group of people who want to collaborate**
- **Involves multiple organization and security domains**
  - **Users from multiple domains**
  - **Resources from multiple domains**
- **Desire to share resources and information**
- **Vital to establish Trust!**
  - **Authentication & Authorization**



# Role Based Authentication

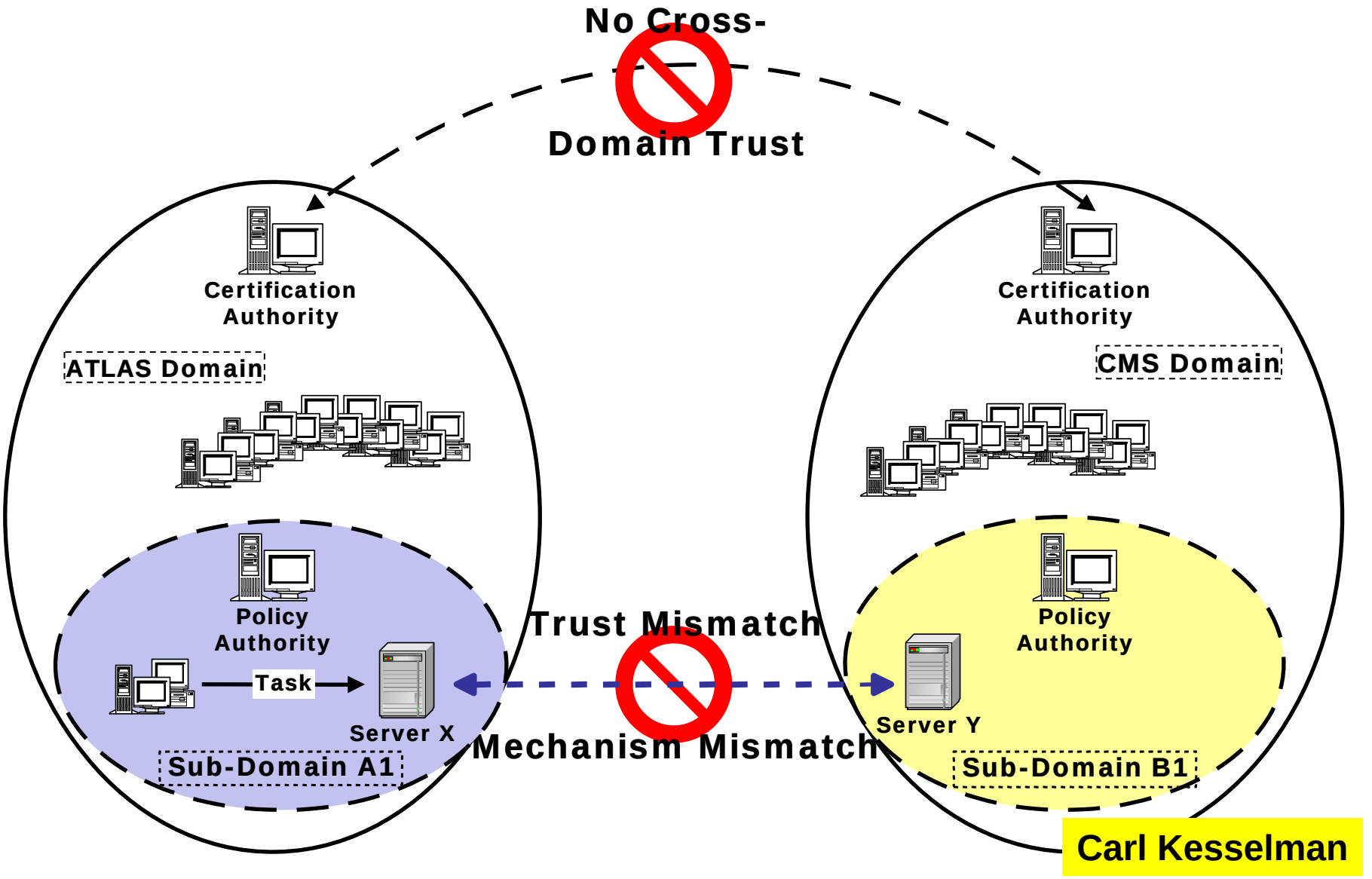


# Effective Policy, Governing Access within a Collaboration



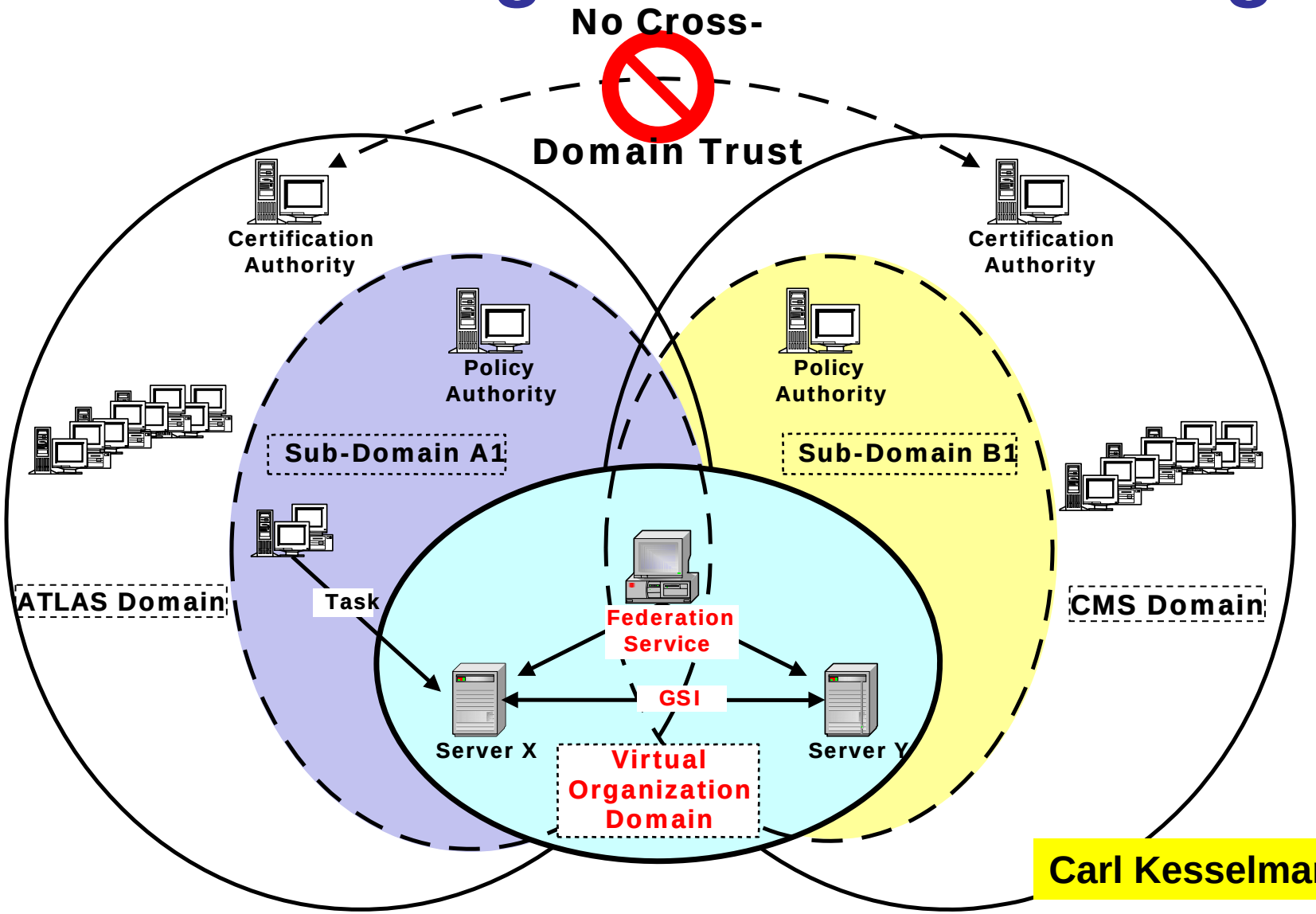

Carl Kesselman

# Multi-Institution Issues



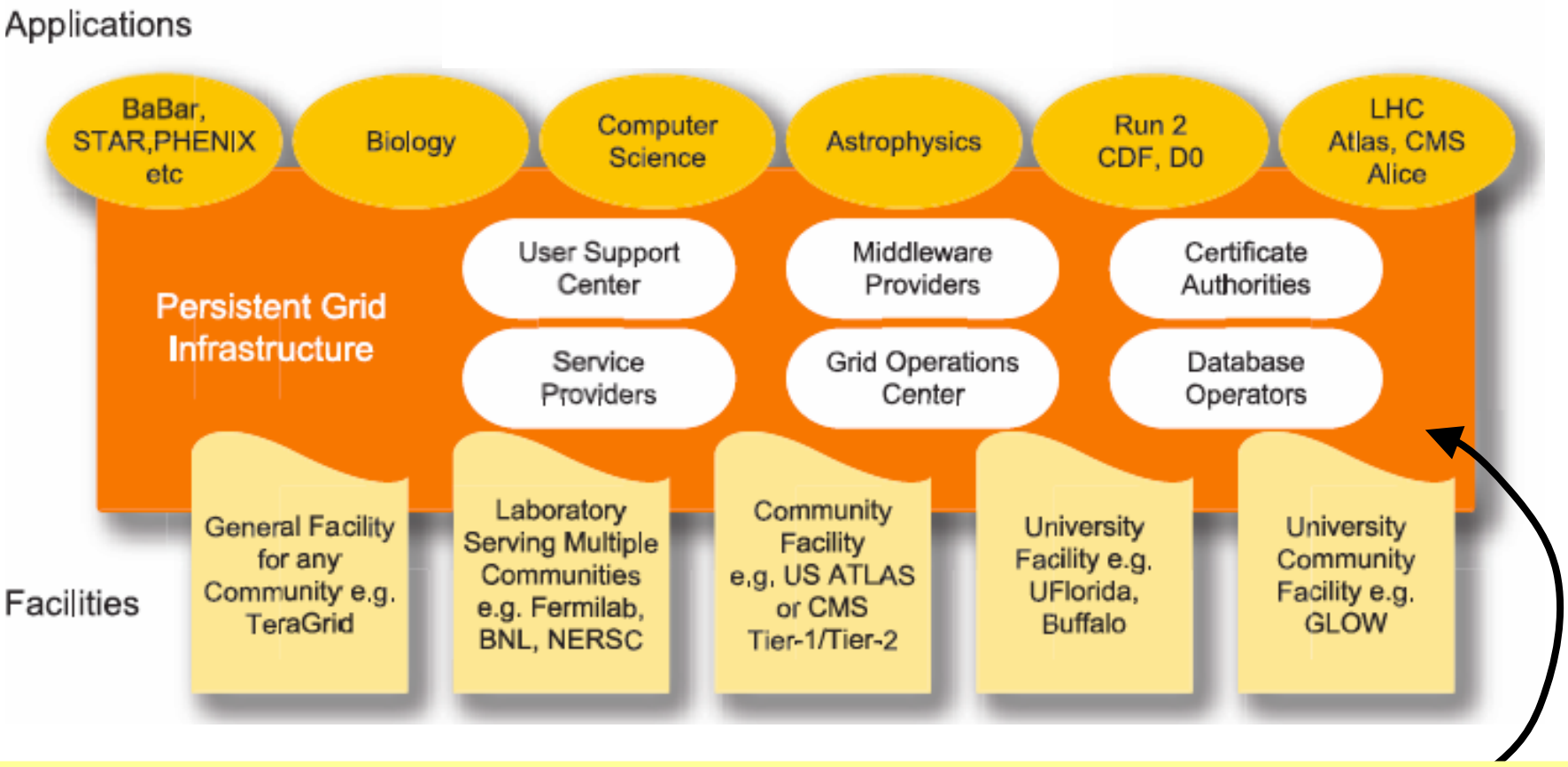

Carl Kesselman

# Grid Solution:
# Use Virtual Organization as Bridge



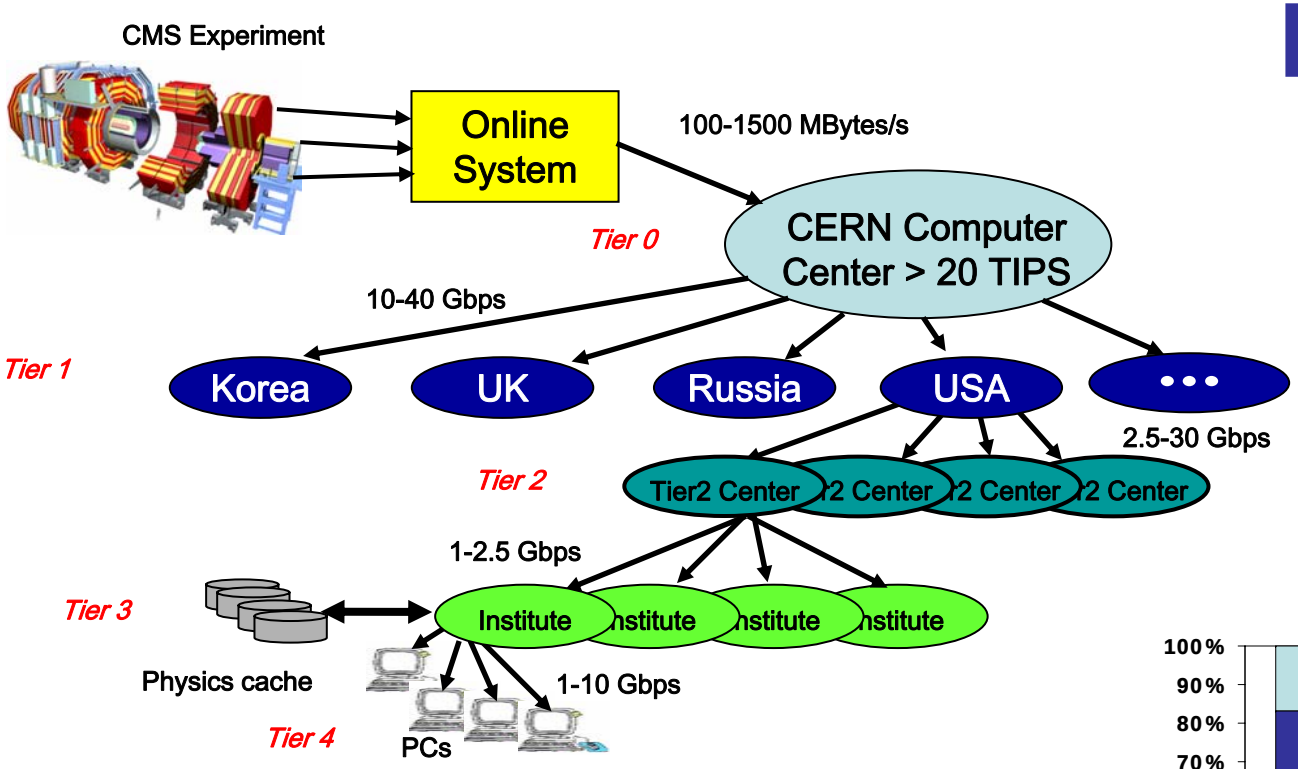

Carl Kesselman

# Enabling Environment for Collaboration

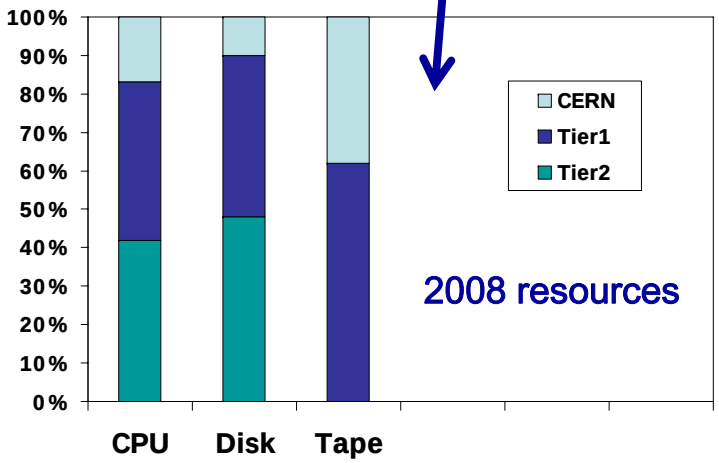



- **Contains Pervasive "Controls" (Grid & Network System Services) → Capable System**
- **Able to Invest Intelligence into System →**
  - **Shield Users from Complexity**
  - **Enable System to Scale with Tolerable Level of Manpower**

# World-wide LHC Computing Grid Hierarchy

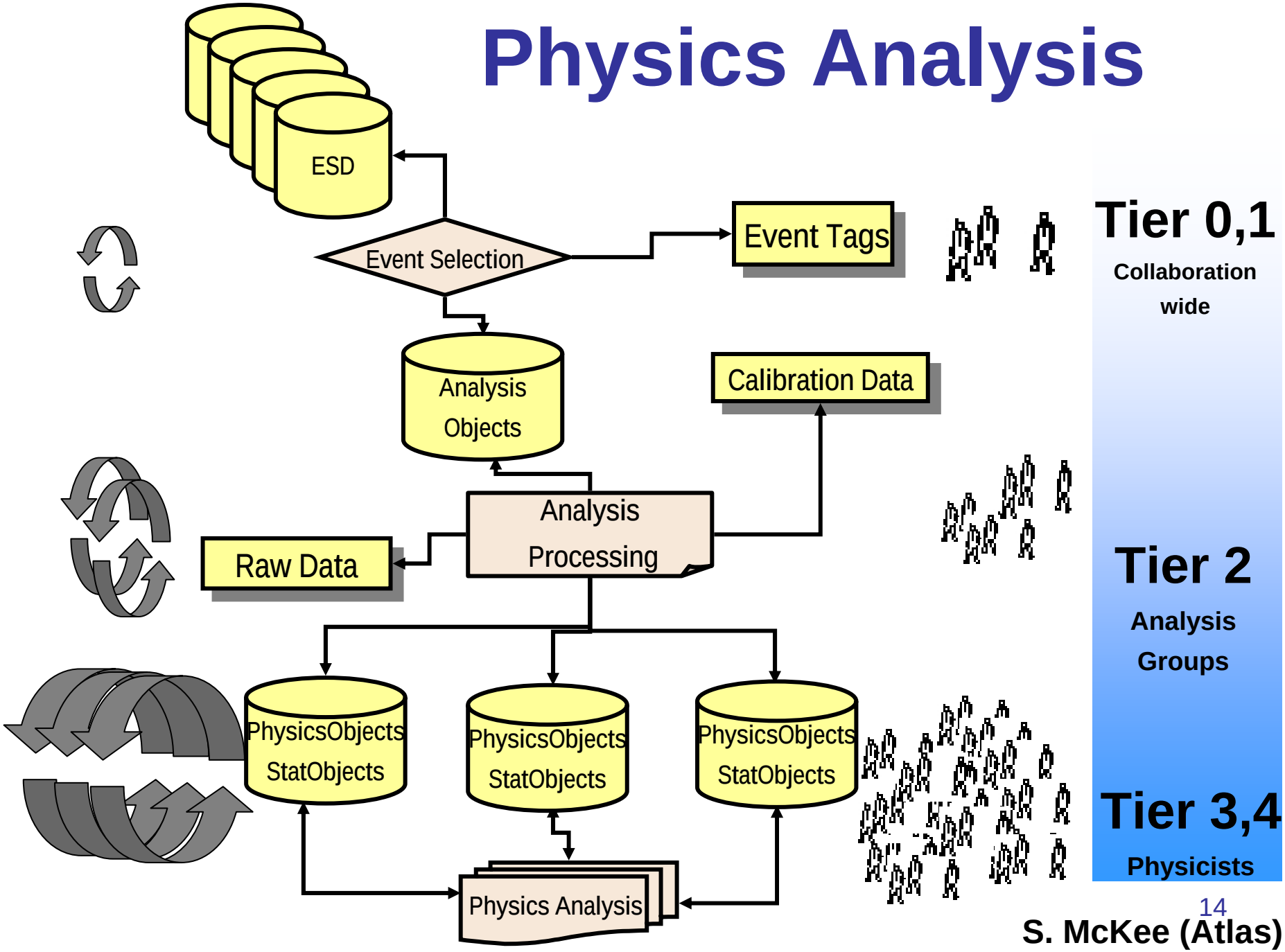



Most IT resources outside of CERN




- ~10s of Petabytes/yr by 2007-8
- ~1000 Petabytes in < 10 yrs

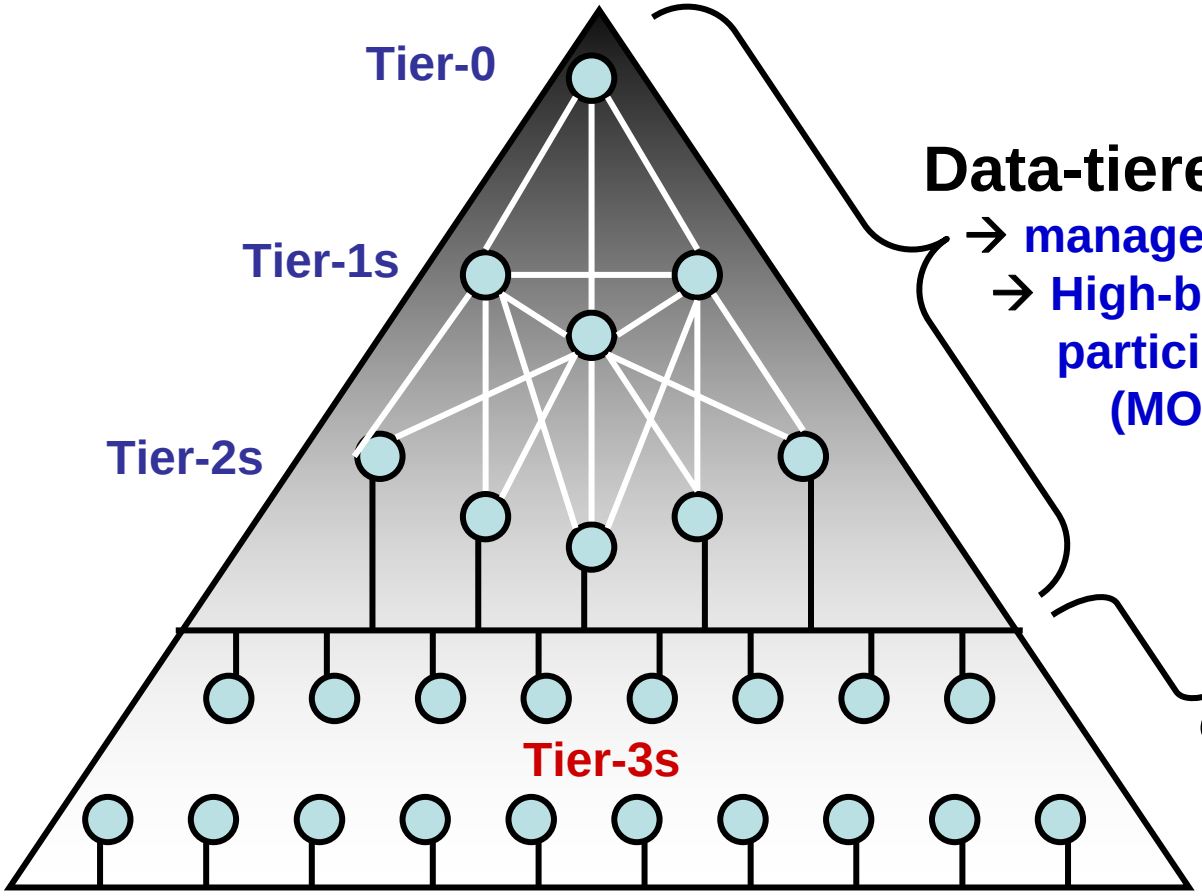
# Physics Analysis






S. McKee (Atlas)

# Lowering the Barrier for Collaboration




**Data-tiered system**

→ managed tasks

→ High-barrier for participation (MOUs, etc)

Examples: WLCG, EGEE, OSG

**Data-bus system**

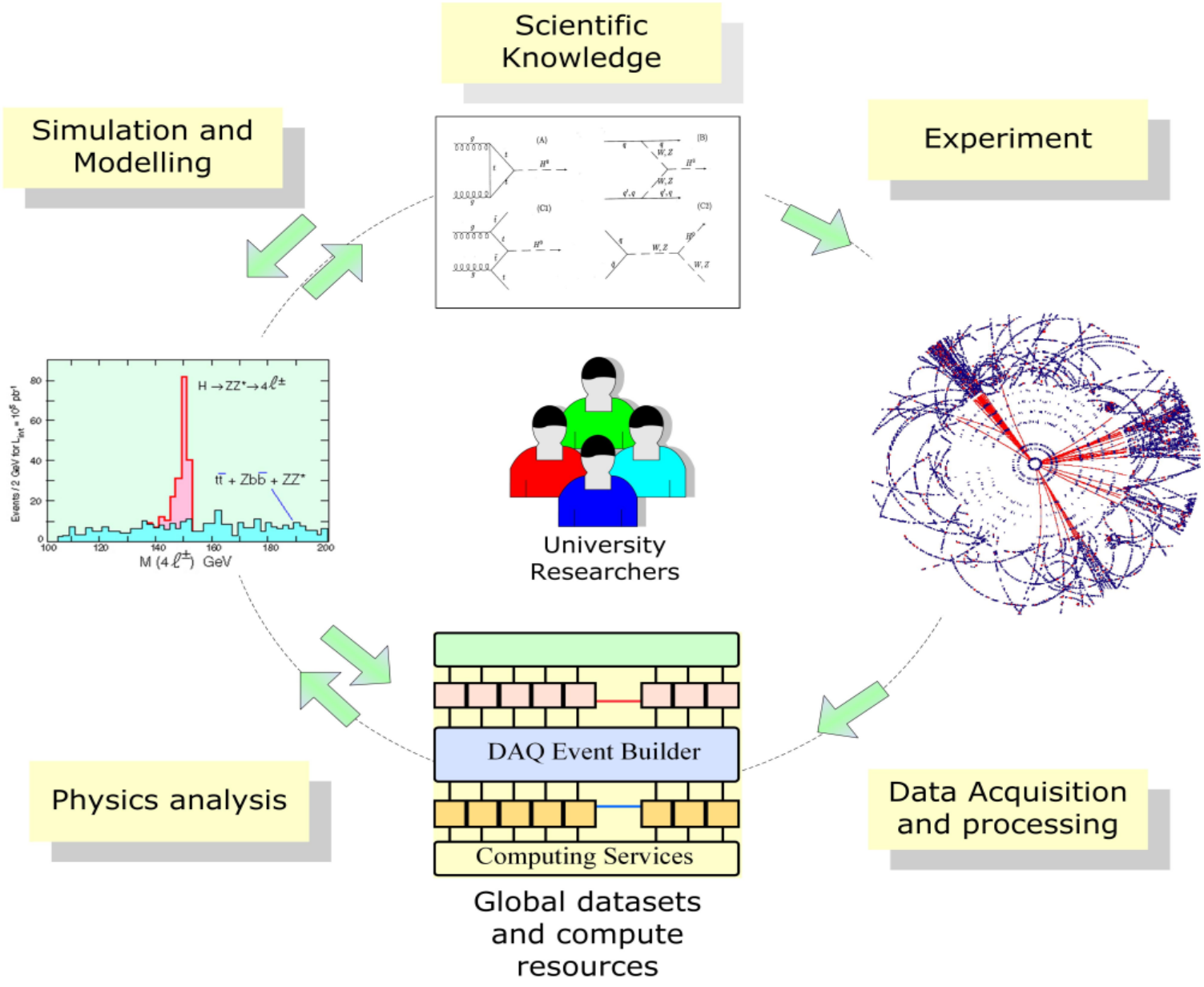
→ chaotic activity

→ Low-barrier for participation (just plug-in)

Examples:
Grid-enabled Analysis Environment, UltraLight

**See H. Newman's talk**

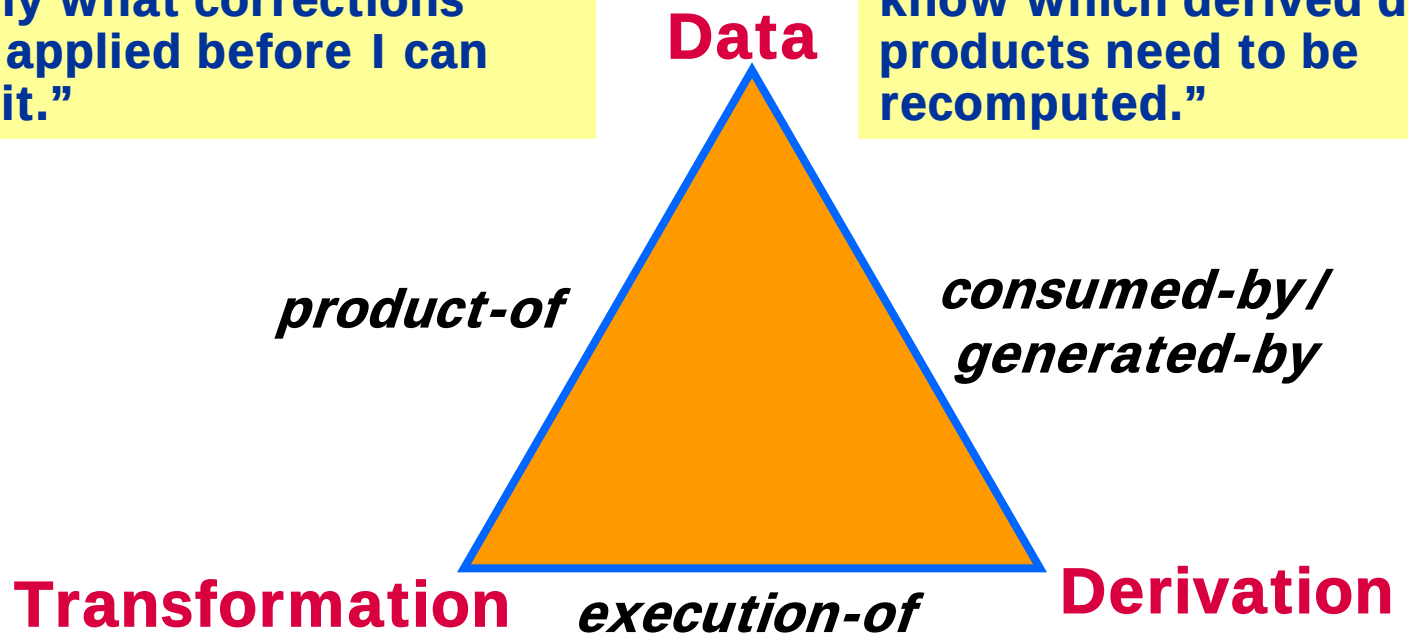
# Scientific Search and Discovery



# Virtual Data Motivations

**"I've found some interesting data, but I need to know exactly what corrections were applied before I can trust it."**

**"I've detected a muon calibration error and want to know which derived data products need to be recomputed."**

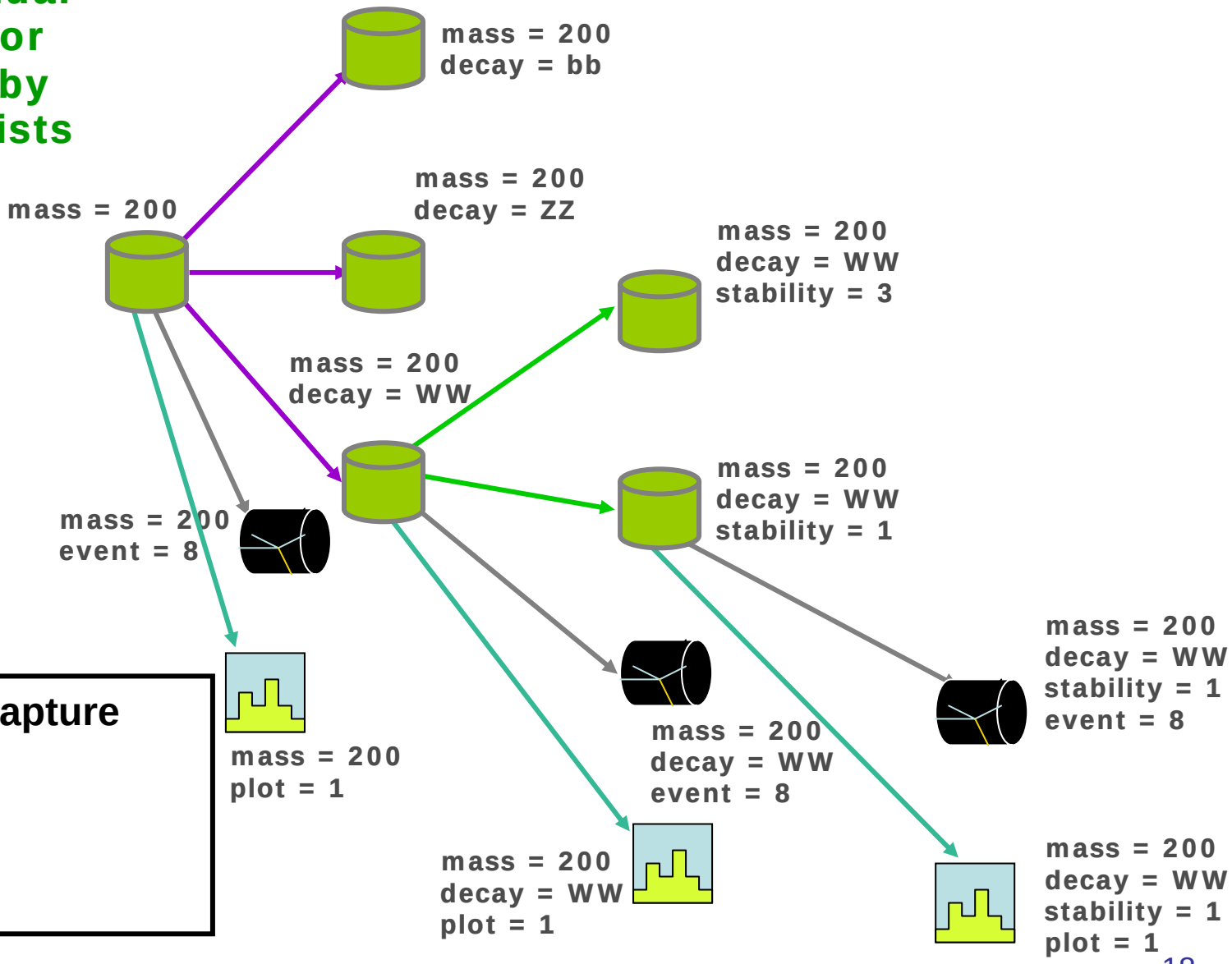



**"I want to search a database for 3 muon SUSY events. If a program that does this analysis exists, I won't have to write one from scratch."**

**"I want to apply a forward jet analysis to 100M events. If the results already exist, I'll save weeks of computation."**

# Virtual Data in HEP Analysis

**Define a virtual data space for exploration by other scientists**



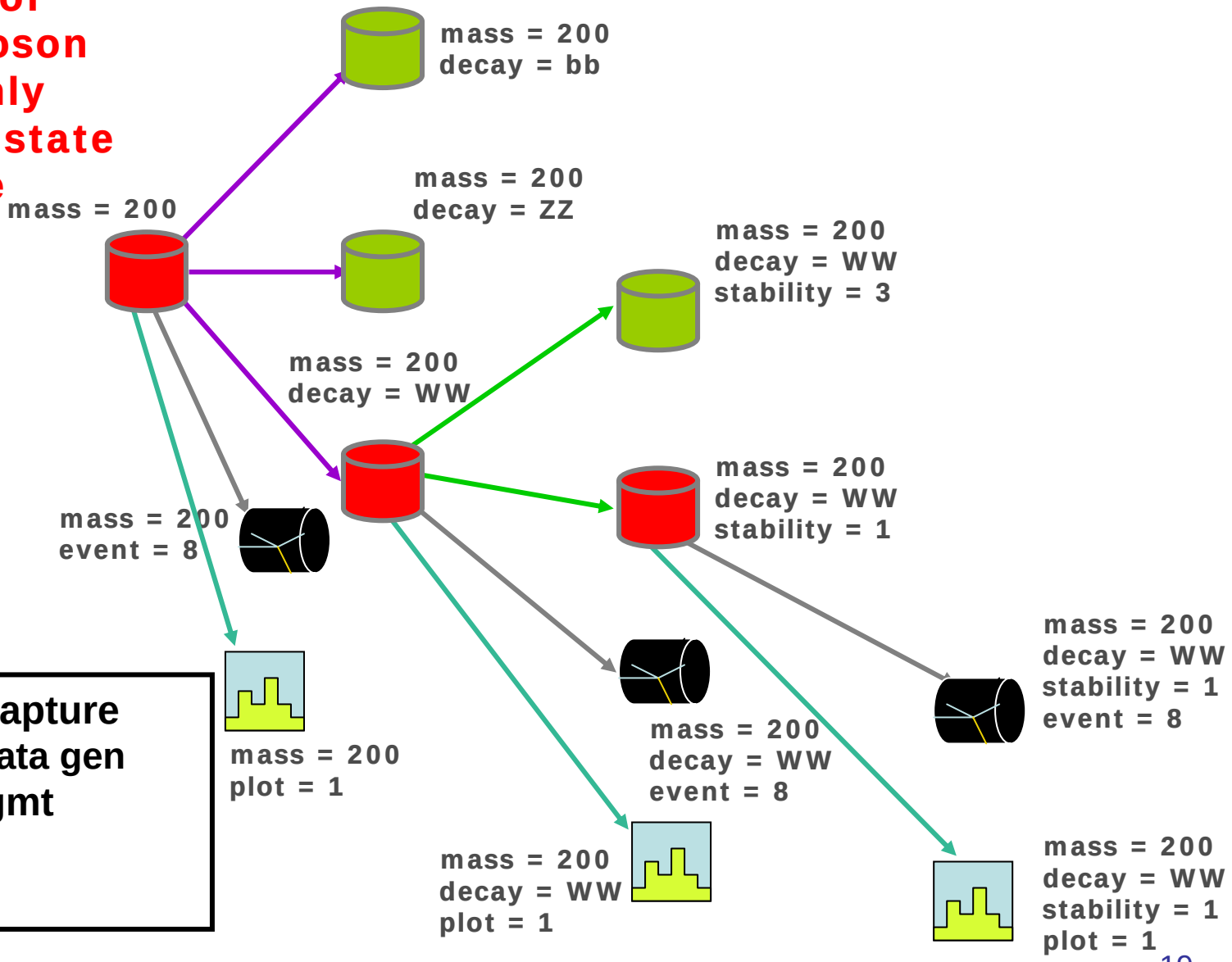

- **Knowledge capture**



**See http://www.griphyn.org/chimera and http://bourilko.web.cern.ch/bourilko/caves.html**

# Virtual Data in HEP Analysis

**Search for WW decays of the Higgs Boson for which only stable, final state particles are recorded?**




- Knowledge capture
- On-demand data gen
- Workload mgmt

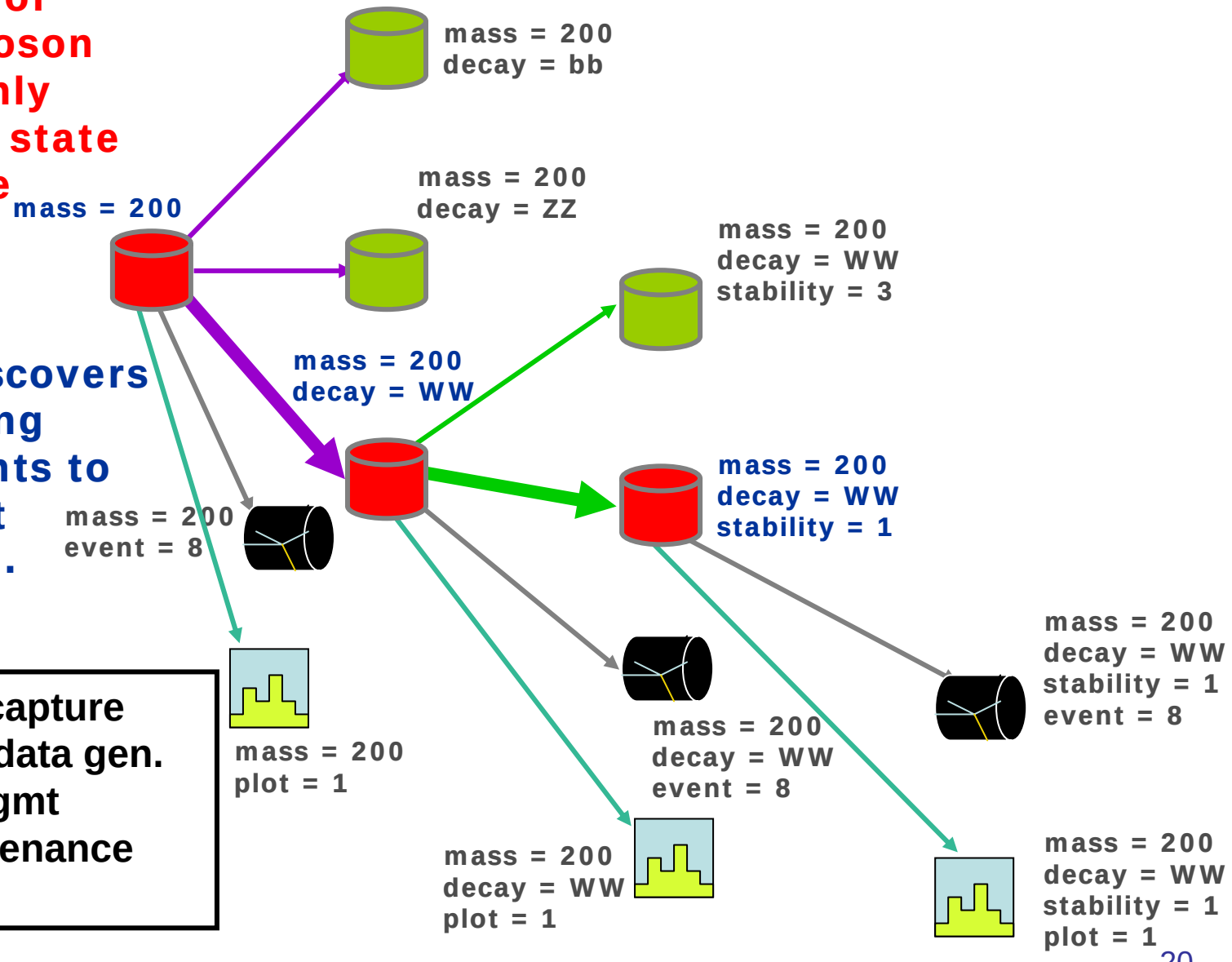


See http://www.griphyn.org/chimera and http://bourilko.web.cern.ch/bourilko/caves.html

# Virtual Data in HEP Analysis

**Search for WW decays of the Higgs Boson for which only stable, final state particles are recorded?**

**Scientist discovers an interesting result – wants to know how it was derived.**



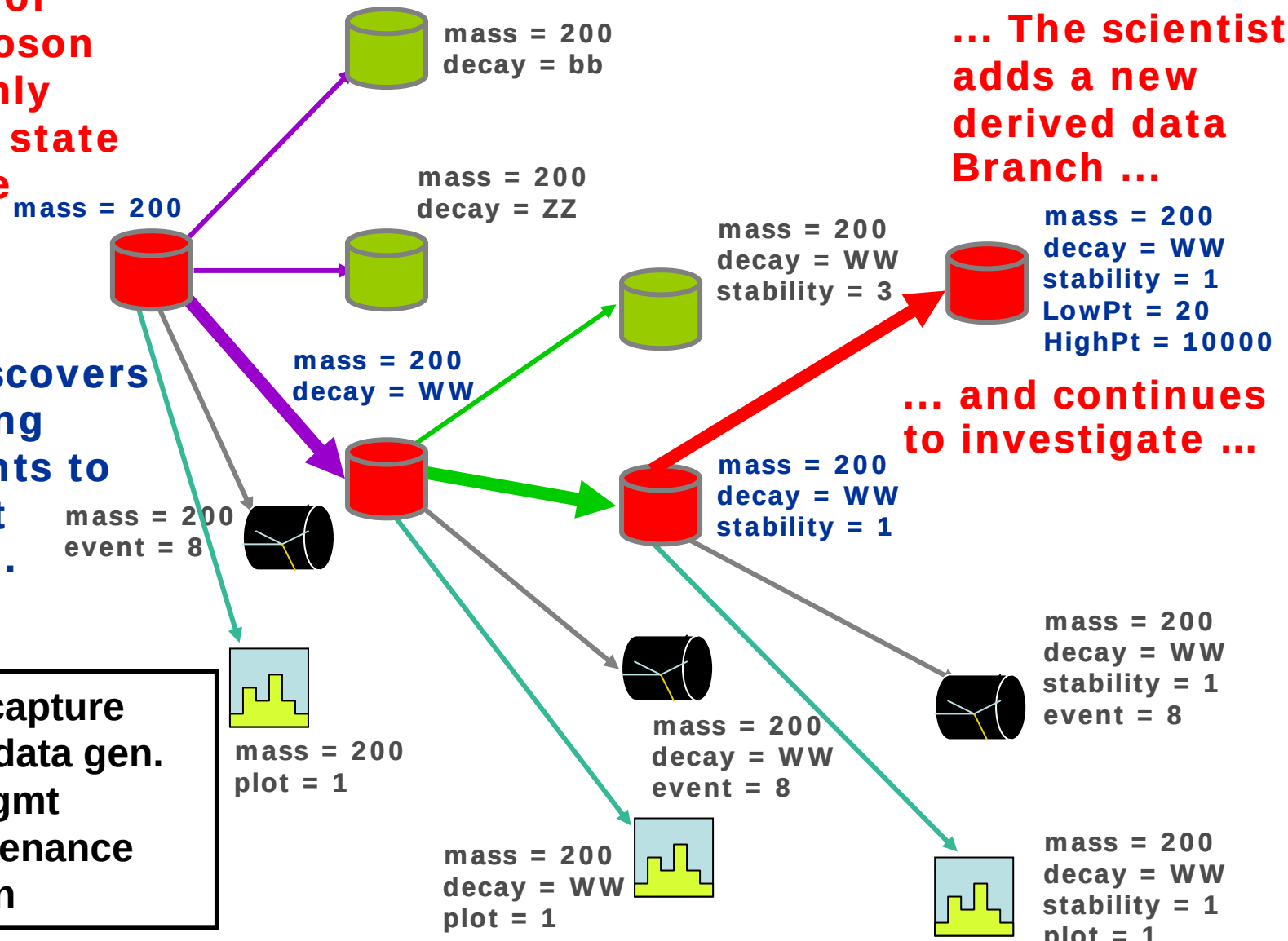

- Knowledge capture
- On-demand data gen.
- Workload mgmt
- Explain provenance



See http://www.griphyn.org/chimera and http://bourilko.web.cern.ch/bourilko/caves.html

# Virtual Data in HEP Analysis

**Search for WW decays of the Higgs Boson for which only stable, final state particles are recorded?**




**Scientist discovers an interesting result – wants to know how it was derived.**

**… The scientist adds a new derived data Branch …**

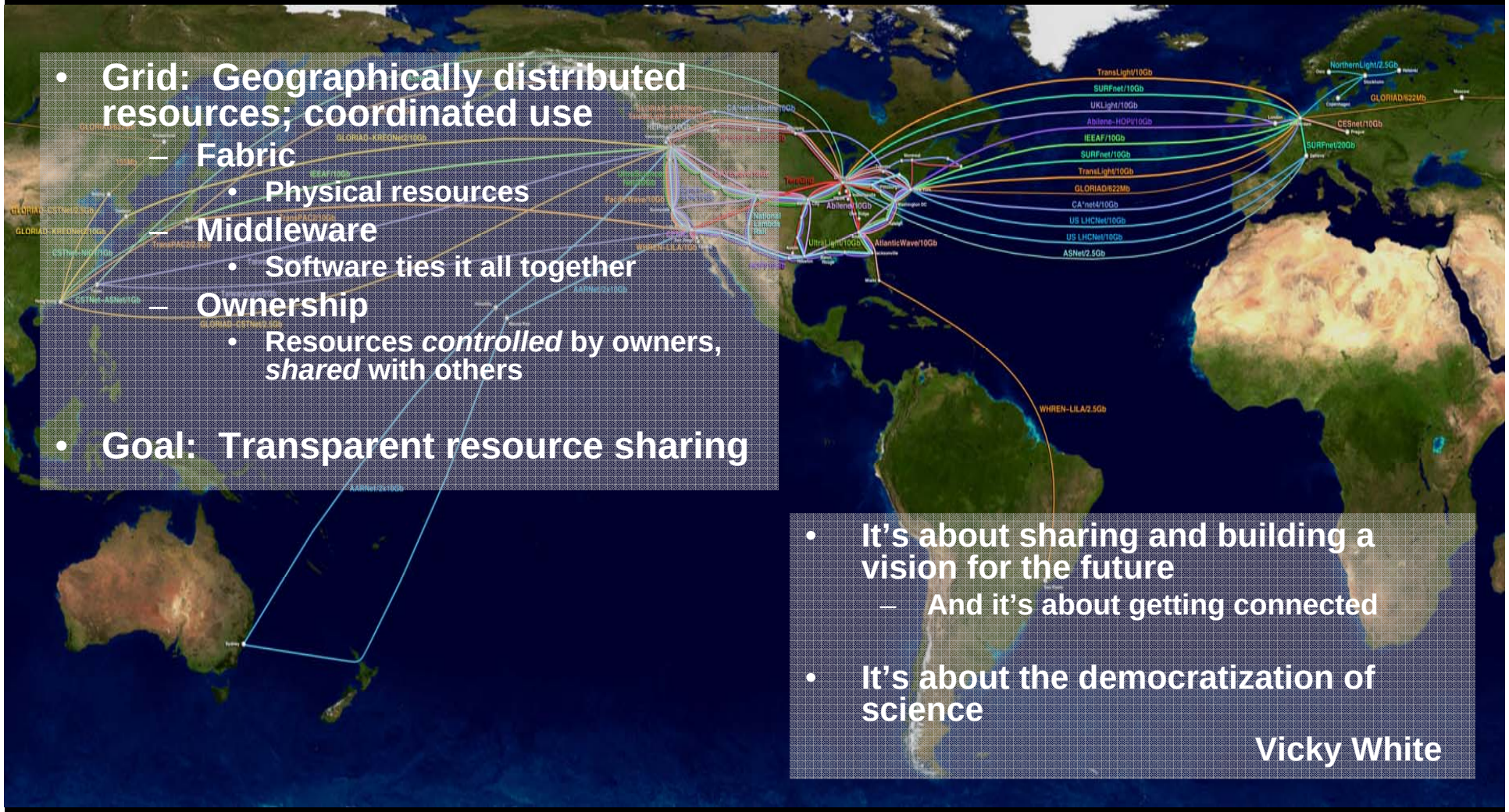
**… and continues to investigate …**

- Knowledge capture
- On-demand data gen.
- Workload mgmt
- Explain provenance
- Collaboration



See http://www.griphyn.org/chimera and http://bourilko.web.cern.ch/bourilko/caves.html

# The Grid - its really about collaboration!




- **Grid: Geographically distributed resources; coordinated use**
  - Fabric
    - Physical resources
  - Middleware
    - Software ties it all together
  - Ownership
    - Resources *controlled* by owners, *shared* with others
- **Goal: Transparent resource sharing**

- **It's about sharing and building a vision for the future**
  - And it's about getting connected
- **It's about the democratization of science**

**Vicky White**