DATA

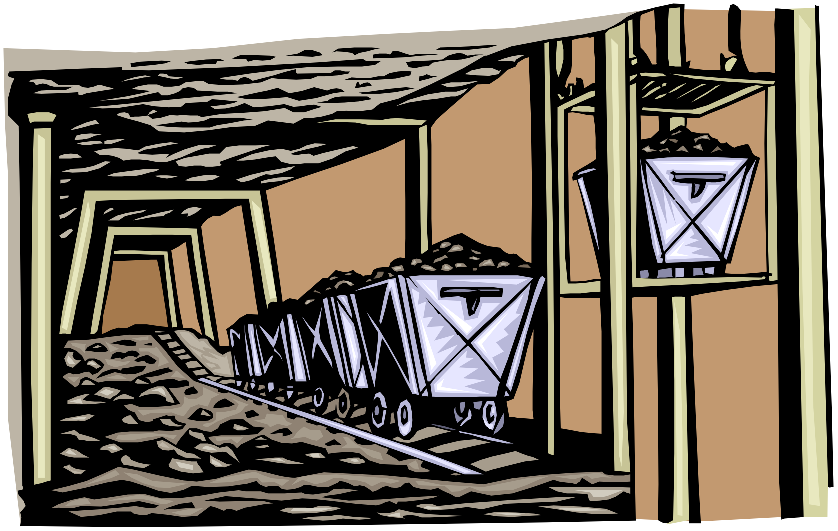
"BIG DATA" COMPUTING

- Big data means different things to different people
  - Data mining from TB → PB of data.
    - Web search, image analysis, climate data analysis
  - Continuous stream analysis
    - Wireless sensor fusion, dynamic planning and control, inline scientific analysis
  - Dense/complex data handling
    - (some) bioinformatics, business analysis

- It may be that these are all different… but there does seem to be some commonality

QUESTIONS TO THINK ABOUT

- How do you think about information coming out of big data?



Questions to think about

- Does big data kill "science"? Does statistical inference replace model building?

QUESTIONS TO THINK ABOUT

- What are the sources of data that are available and relevant for university research that would support industrial concerns?

DISCUSSION LEADERS

- Joel Saltz – Emory
- Doug Blough – GT ECE
- Alex Gray – GT CSE
- Ron Oldfield – Sandia
- Calton Pu – GT CS
- Scott Klasky – ORNL

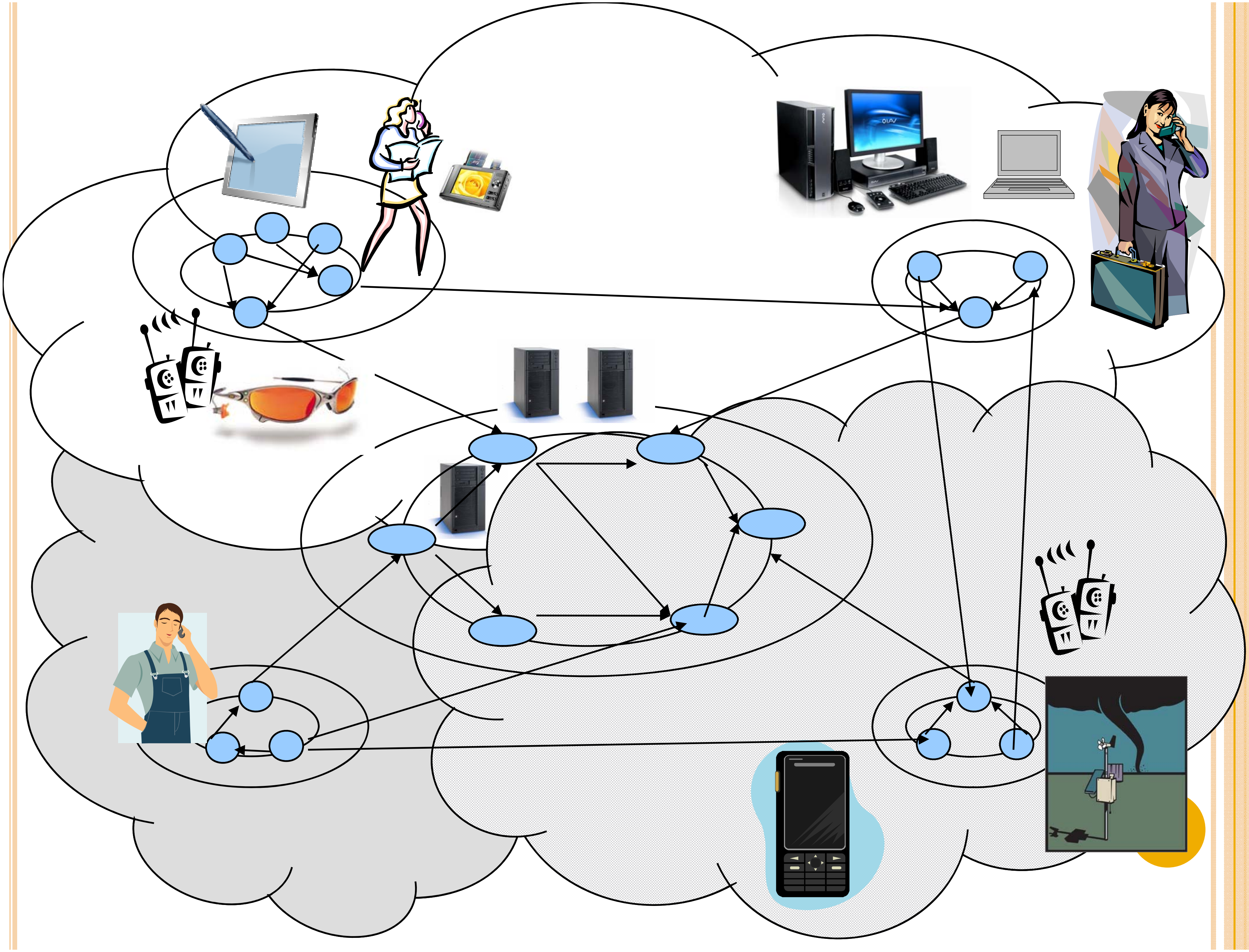
SavvyData

Self-handling data for the data explosion

Brought to you by a cast of thousands…

DATA EXPLOSION

- The future is swarming with data
  - Not that that surprises anyone….
- The present is also swarming with data
  - But, on the whole, we don't know what to do with it
  - Not that long ago, NASA was just ditching some of the feeds of satellite data, simply because it didn't know where to put it.
- Data is, frequently, well formatted
  - But may be poorly formatted for what you **want** to extract from it.
  - Heterogeneous format handling will be the rule, not the exception.

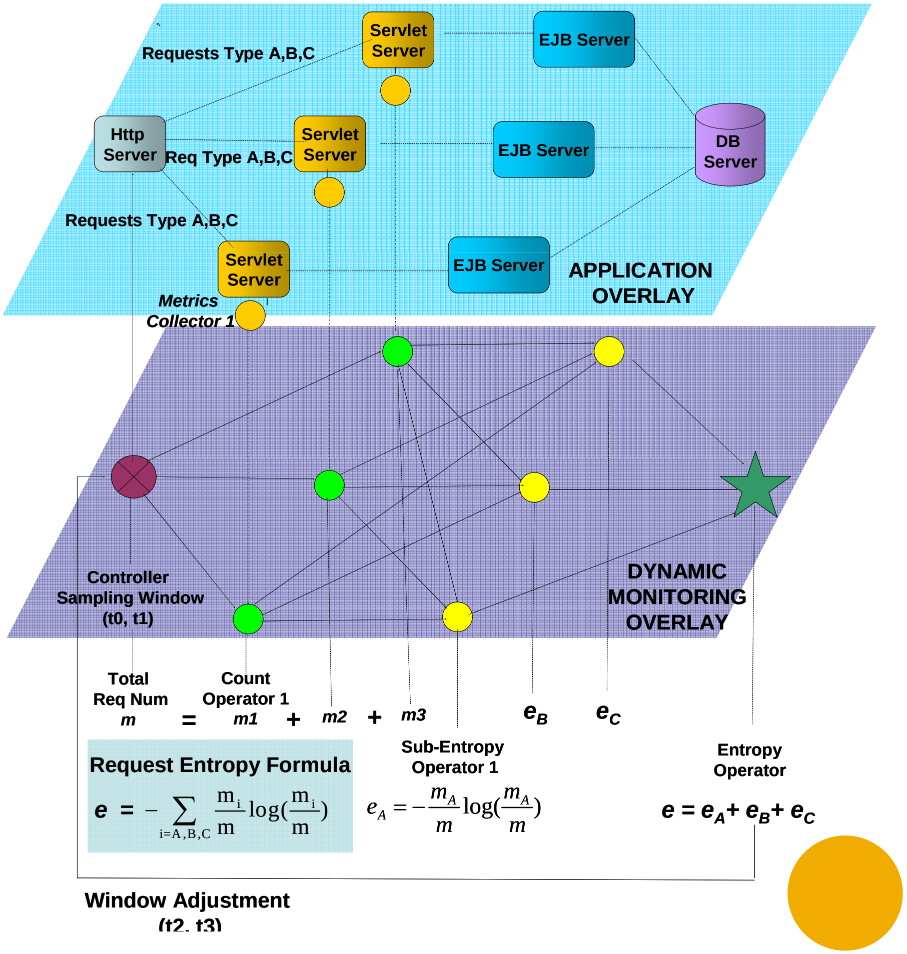


SavvyData

- Two observations:
  - Flops are free. So now, the key is the same as in real estate. Location, Location, Location.
  - Line between Metadata and Data is now blurry. (Ore vs needle)
- Data, data expression, and the data handling process needs to be integrated.
  - Self-organization will be key.
  - Abstraction should lift application awareness of specific locality (ie specific file names) while enabling the platform to localize
- SavvyData is a middleware abstraction allowing a self-* data access

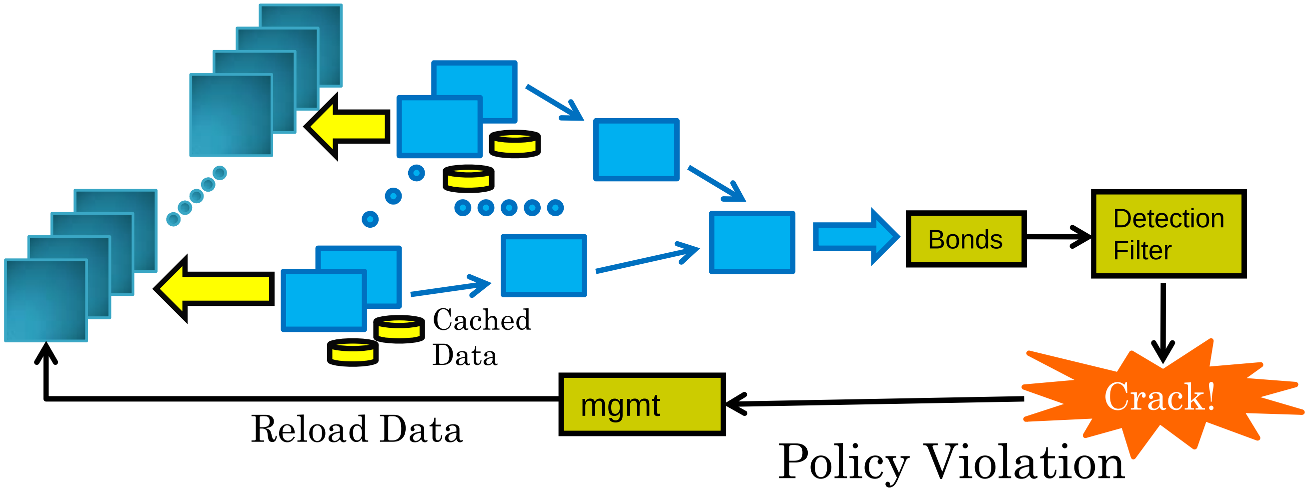
Monitoring Example: Statistical Overlay Analysis

- Monitoring streams push data through overlay
  - Entropy measures are dynamically inserted into streams
  - Data classification can take place dynamically



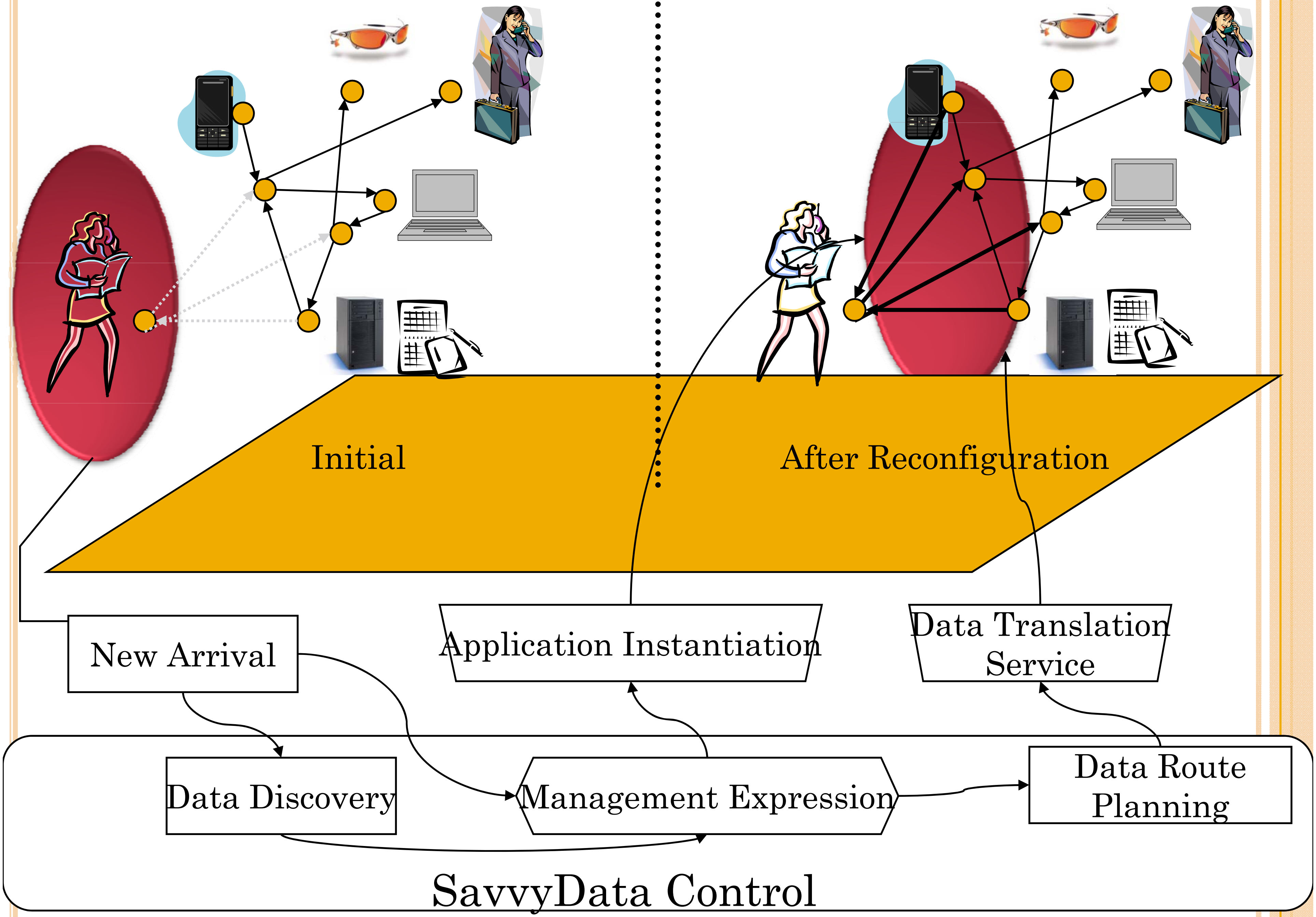
Science Example: Scientific Workspace

- Motivating application is based on a multi-scale material physics model
  - Exploiting locality of data (caching)
  - Improvements in time to discovery for relevant material properties
  - Automatic policy actuation



Technical Innovation Details

- Self-describing data
  - Data correlations should also be extendable and self-discoverable
    - "I am data 7 of 9, and the 'most useful' of 12"
  - Leverage existing work by G. Eisenhauer (& many others) over the last 15 years on self-describing data packets
- Self-routing data
  - Control plane for metadata-based routing
  - Efficient discovery of introduced metadata tags
- Self-annotating data
  - Dynamic morphing/extension of data in a context and location
    - Code specialization, Dynamic re-typing, etc.



Initial

After Reconfiguration

New Arrival

Application Instantiation

Data Translation Service

Data Discovery

Management Expression

Data Route Planning

SavvyData Control