

# The Angle Project: Some Lessons for Data Mining



October 12, 2007

The Angle Project Team

Talk by Robert L Grossman

# Angle Project Team

Robert L. Grossman, Anushka Anand, Shirley Connelly,  
Yunhong Gu, Matt Handley, Michal Sabala, Rajmonda Sulo,  
Dave Turkington and Lee Wilkinson  
National Center for Data Mining  
**University of Illinois at Chicago**

Ian Foster, Ti Leggett, Mike Papka, Mike Wilde  
**University of Chicago and Argonne National Laboratory**

Joe Mambretti  
**Northwestern University**

Bob Lucas and John Tran  
Information Sciences Institute  
**University of Southern California**

# Thanks to:

- Vipin Kumar
- Sanjay Ranka

# Part 1

# Problem



- Discover in near real time **suspicious** behavior in IP traffic collected from geographically distributed sites.
- Important for a variety of problems, including protecting cyberinfrastructure.

# What Does Suspicious Mean?



- Even if a data mining application costs \$100M, there are still only a finite number of analysts who look at alerts.
- Suspicious = high score = analyst looks at it.

# Angle

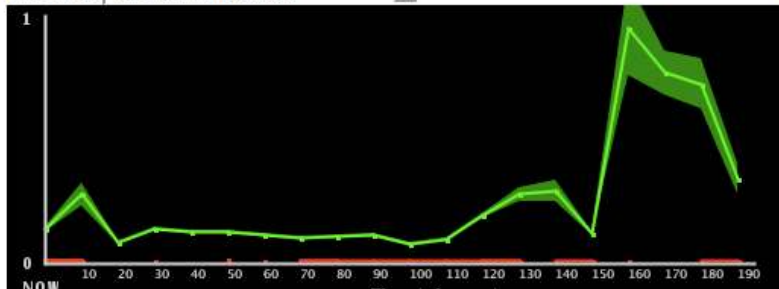
Select data to score: Location:  Time:

Show IPs with score above:



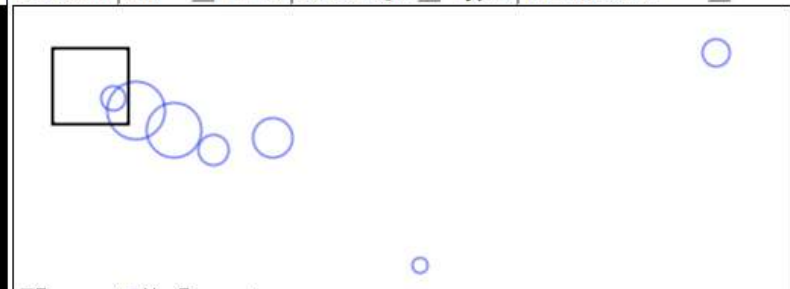
Select feature to examine:

Feature:

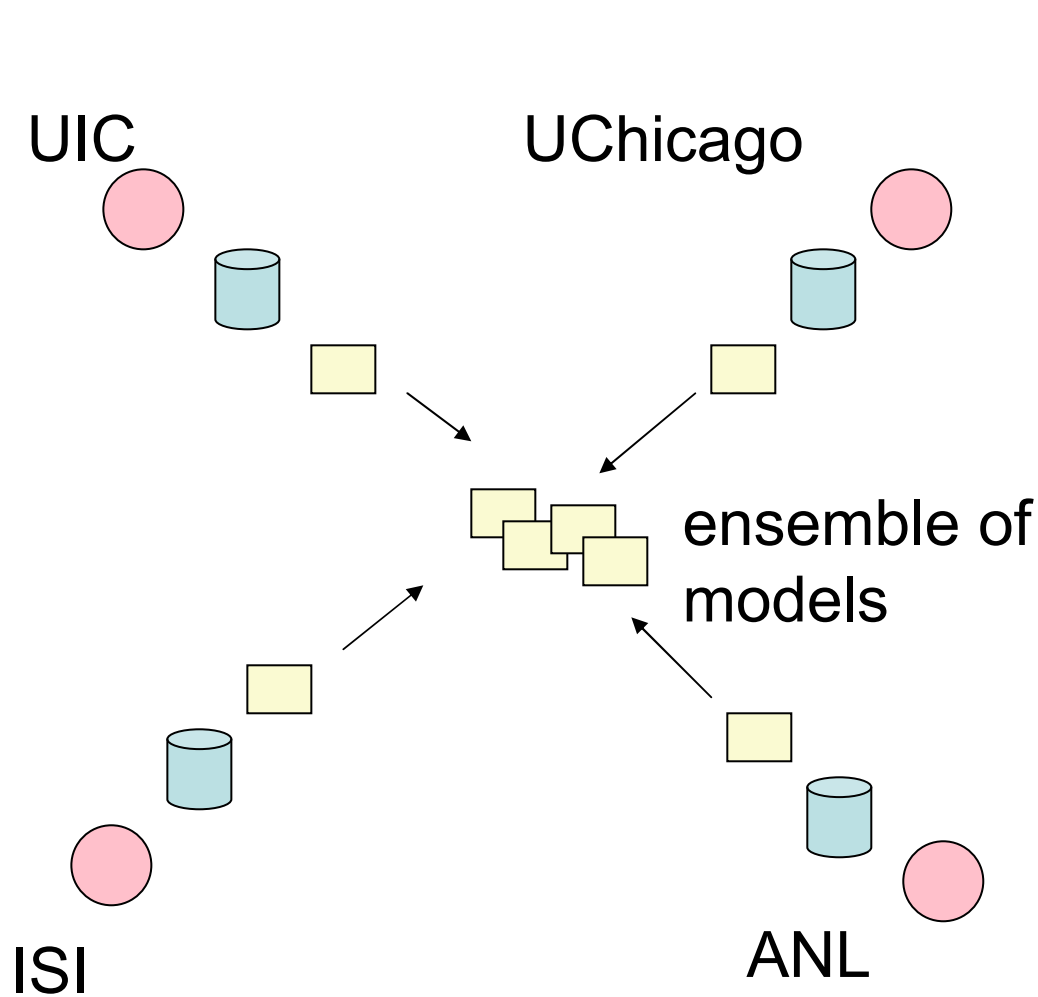






Select model:

Location:  Time:  Type:



# Success Story in Distributed Data Mining: Version 1



- Sensors  capture packets on commodity internet
- Use local data  to build local models  to classify known exemplars.
- Gather local models  to produce global model.

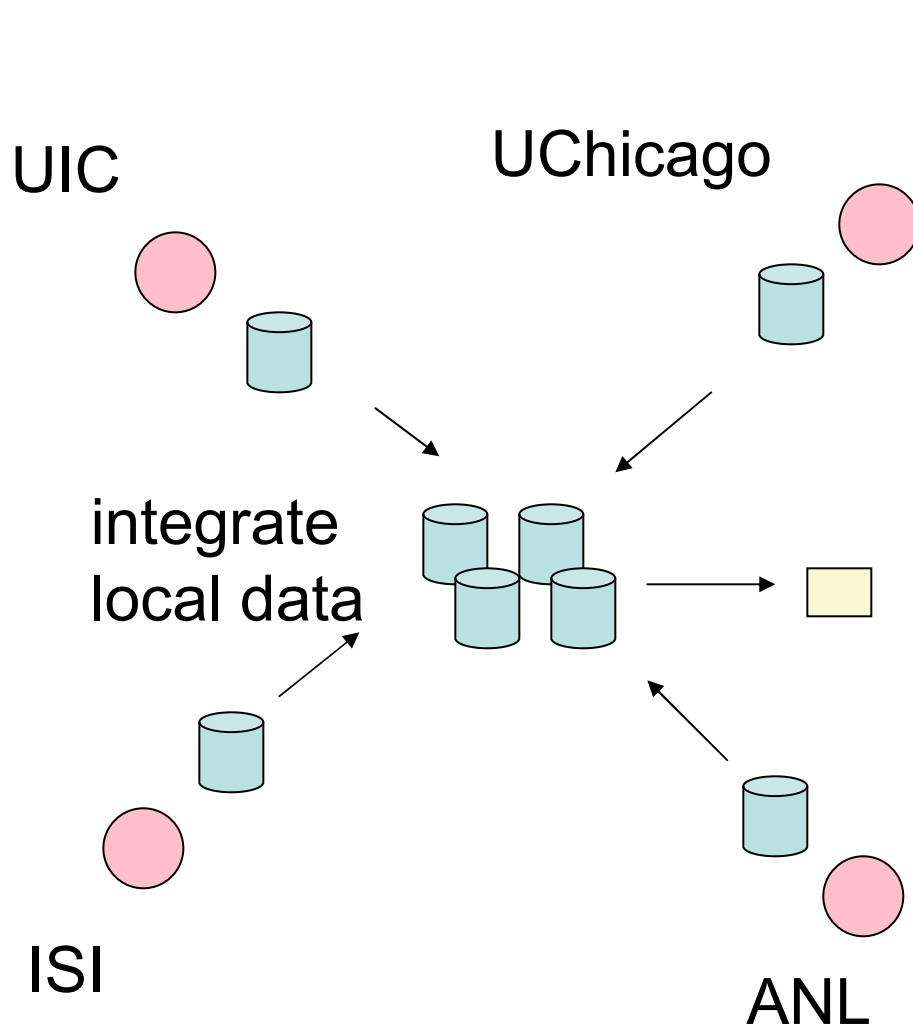






# Why does this work?

Probably because, of the  
unreasonable effectiveness of  
ensembles in data mining.

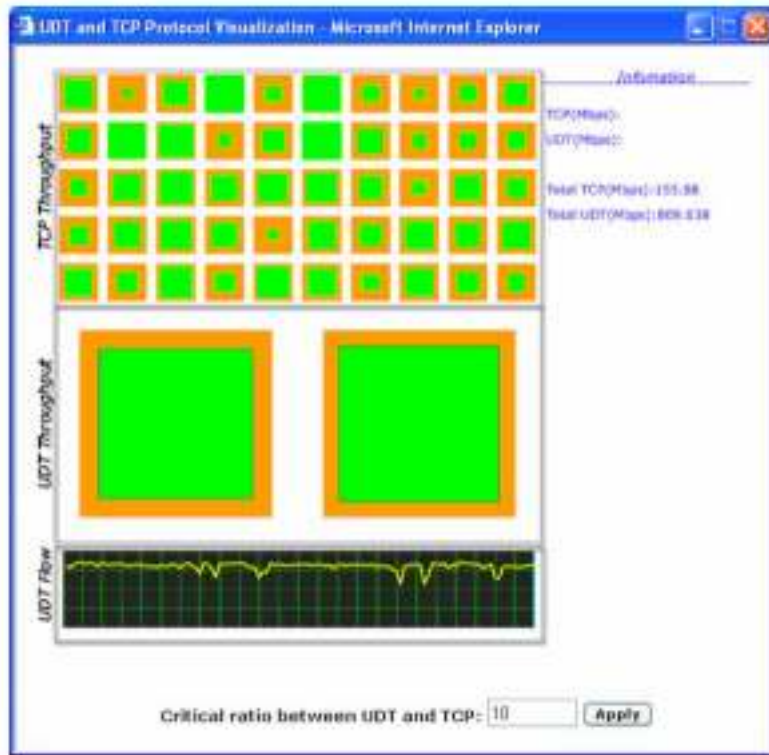
Cf. Eugene Wigner, The Unreasonable Effectiveness of  
Mathematics in the Natural Sciences, Comm. Pure  
Applied Math., 1960.

# Success Story in Distributed Data Mining: Version 2



- Sensors  capturing packets on commodity internet
- Transport local data  using high perf. networks
- Integrate all local data  to produce global model  that classifies known exemplars





































# Why does this work?



Because after ten years of research and development, we can finally transport large data sets over high performance networks.

## Nodes Status

### Nodes Performance Statistics

ID	Name	Location	CPU	Memory (GB)	NIC (Gb/s)	Disk Space (GB)	Status
1	SL-1	Chicago	Opteron	2	10	1500	
2	SL-2	Chicago	Opteron	2	10	1500	
3	Starlight-1	Chicago	Xeon	2	1	360	
4	Starlight-3	Chicago	Xeon	2	1	360	
5	Starlight-4	Chicago	Xeon	2	1	360	
6	QueensU-1	Kingston	Xeon	1	1	360	
7	QueensU-2	Kingston	Xeon	1	1	360	
8	QueensU-3	Kingston	Xeon	1	1	360	
9	QueensU-4	Kingston	Xeon	1	1	360	
10	JGN2-1	Tokyo	Opteron	2	10	1500	
11	APAN	Tokyo	Opteron	2	10	1500	
12	JGN2-2	Tokyo	Opteron	2	1	1500	
13	Amsterdam-1	Amsterdam	Xeon	1	1	360	
14	Amsterdam-2	Amsterdam	Xeon	1	1	360	
15	Amsterdam-3	Amsterdam	Xeon	1	1	360	
16	Amsterdam-4	Amsterdam	Xeon	1	1	360	
17	Amsterdam-5	Amsterdam	Xeon	1	1	360	
18	Amsterdam-6	Amsterdam	Xeon	1	1	360	
19	Amsterdam-7	Amsterdam	Xeon	1	1	360	
20	Amsterdam-8	Amsterdam	Xeon	1	1	360	
21	Amsterdam-9	Amsterdam	Xeon	1	1	360	
22	Amsterdam-10	Amsterdam	Xeon	1	1	360	
23	Caltech-1	Pasadena	Opteron	2	10	2000	
24	Caltech-2	Pasadena	Opteron	2	10	2000	
25	Caltech-3	Pasadena	Opteron	2	10	2000	
26	Datatag-1	Geneva	Opteron	2	10	1500	
27	Datatag-2	Geneva	Opteron	2	10	1500	
28	Datatag-3	Geneva	Opteron	2	10	1500	
29	NASA-1	Greenbelt	Opteron	2	10	2000	
30	NASA-2	Greenbelt	Opteron	2	10	2000	
31	NASA-3	Greenbelt	Opteron	2	10	2000	
32	NASA-4	Greenbelt	Opteron	2	10	2000	
33	KISTI-1	Daejeon	Opteron	2	10	2000	
34	KISTI-2	Daejeon	Xeon	2	10	1500	
36	LAC-1	Chicago	Opteron	4	10	5000	
37	LAC-2	Chicago	Opteron	4	10	18000	

**Distributed data mining requires  
the proper infrastructure.**

# Four Questions

- Why is this the wrong problem?
- Why is this the wrong algorithm?
- Why is this the wrong architecture?
- Why aren't more systems like this deployed in the field?

# Part 2 - What is a Better Problem?

For many problems, we have  
very few exemplars.



# TJ Max Compromise



For many problems it critical to identify something new, interesting and relevant. We call this **emergent** behavior.

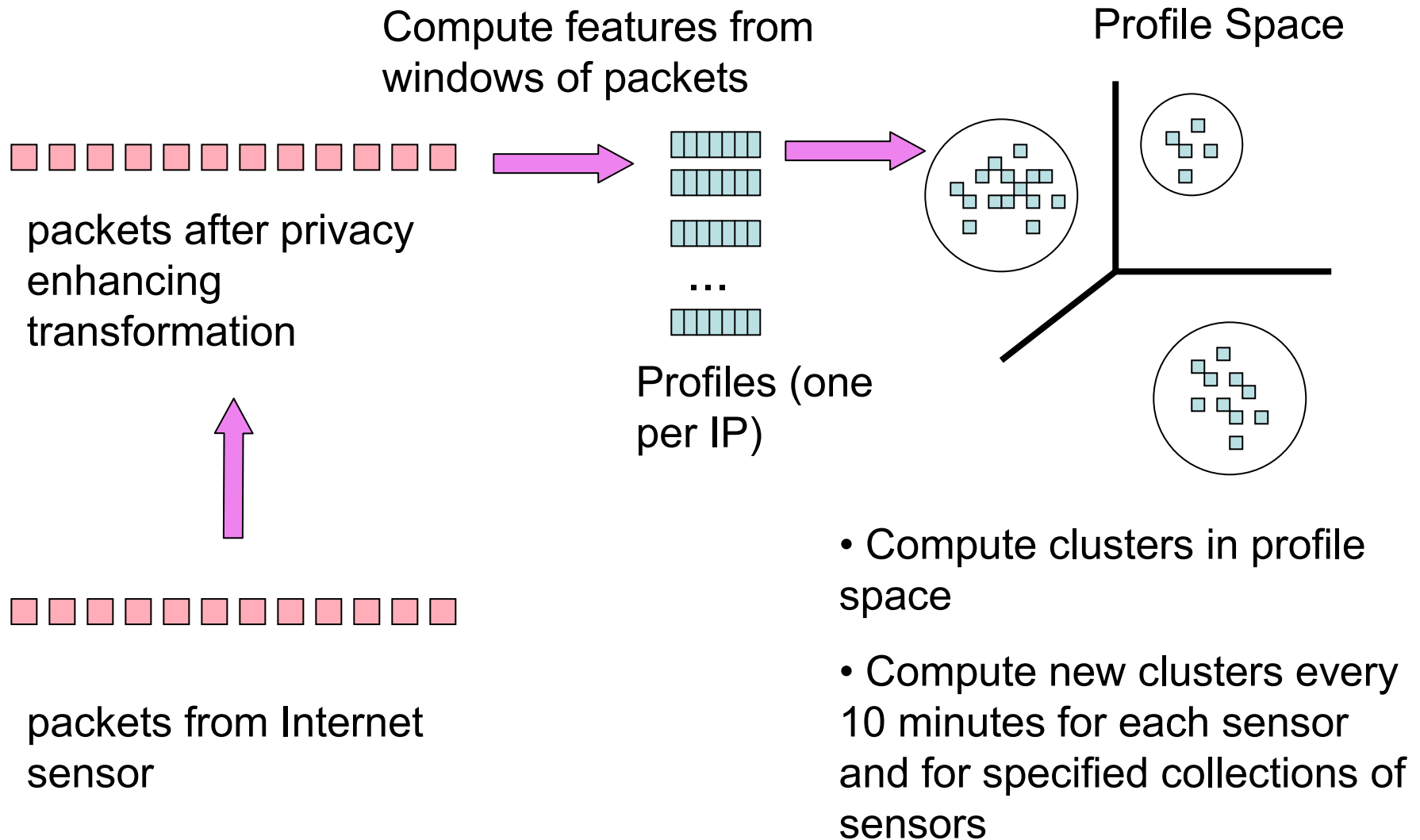
# Better Problem

Find algorithms that discover new types of emergent behavior from unlabeled data.

# Part 3. What is a Better Algorithm?

Any algorithm that doesn't require very much labeled data (and isn't biased toward the existing exemplars).

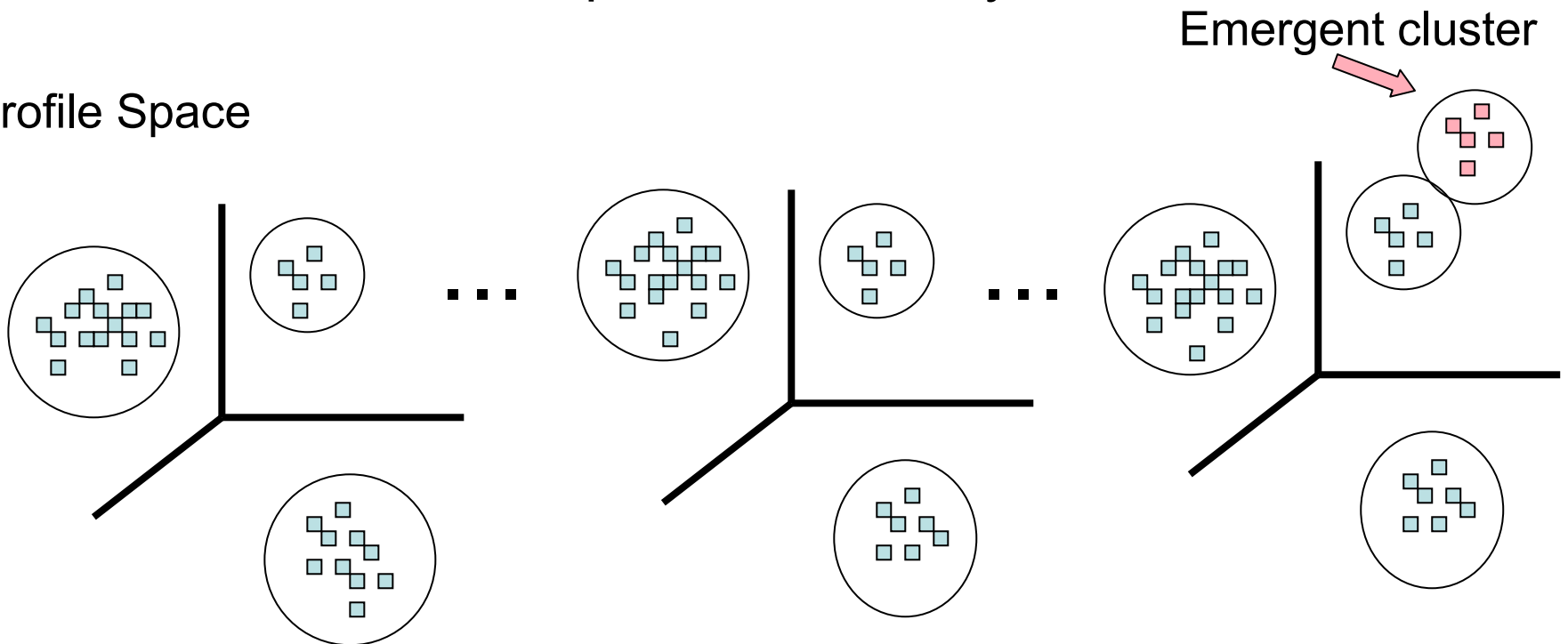
# From Packets to Features to Clusters



# What is Emergent Behavior?

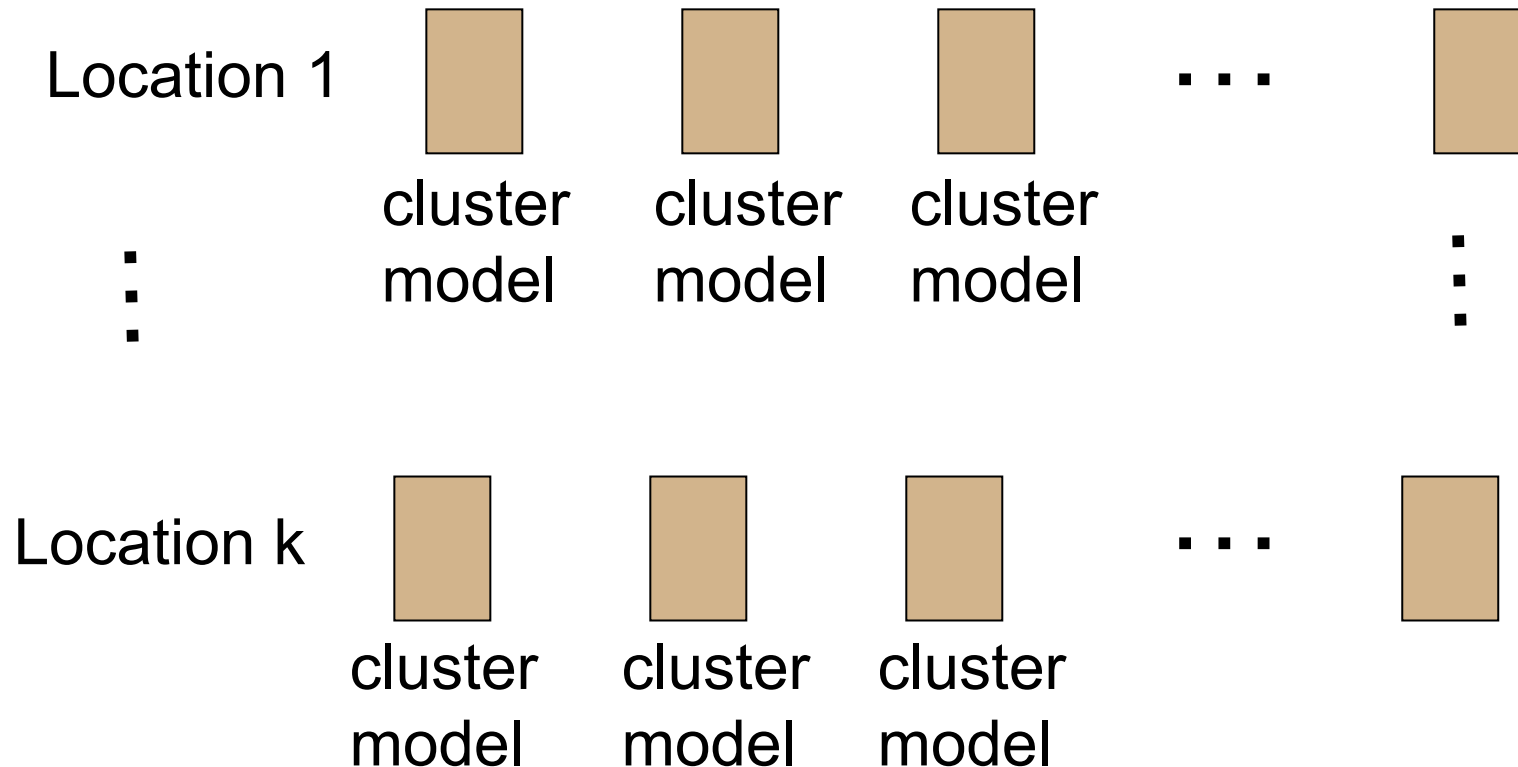
- No agreed to definition
- Our working definition - new unusual behavior we haven't seen before as indicated by emergence of new clusters after a period of stability.

Profile Space



Clusters are **stable** for a period ... Then new cluster emerges

# Identifying Emergent Clusters



- Think of this as an algorithm to do “meta-analysis” of 50,000+ cluster models to identify emergent clusters.



# Part 4: What is a Better Architecture?

# Key Question - What Resource is Scarce?

- Scarce processors wait for data
  - assume that cycles precious
  - wait for an opening in the queue
  - scatter the data to the processors
  - and gather the results

**Supercomputer Center Model (local)**

**Grid/Data Grid (distributed)**

---
- Persistent data wait for queries
  - assume data is precious
  - persistent data waits for queries
  - computation done locally
  - results returned

**Data Center Model (local)**

**Data Cloud (distributed)**

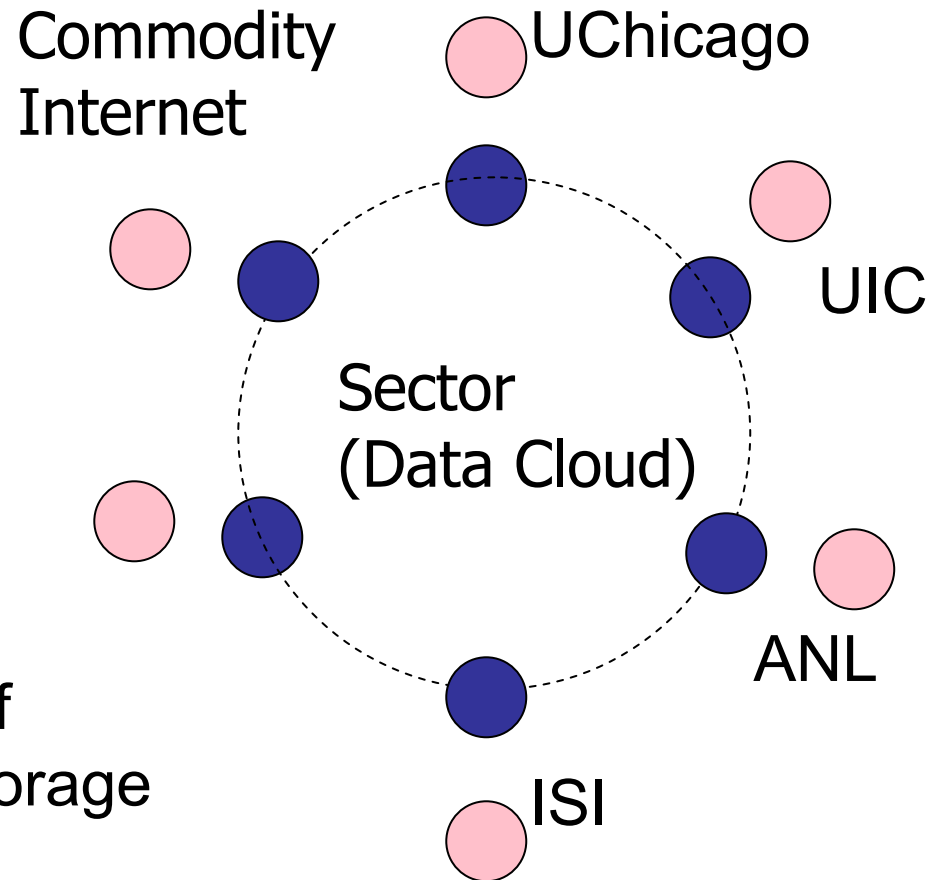
Supercomputer Center Model

or

Data Center Model?

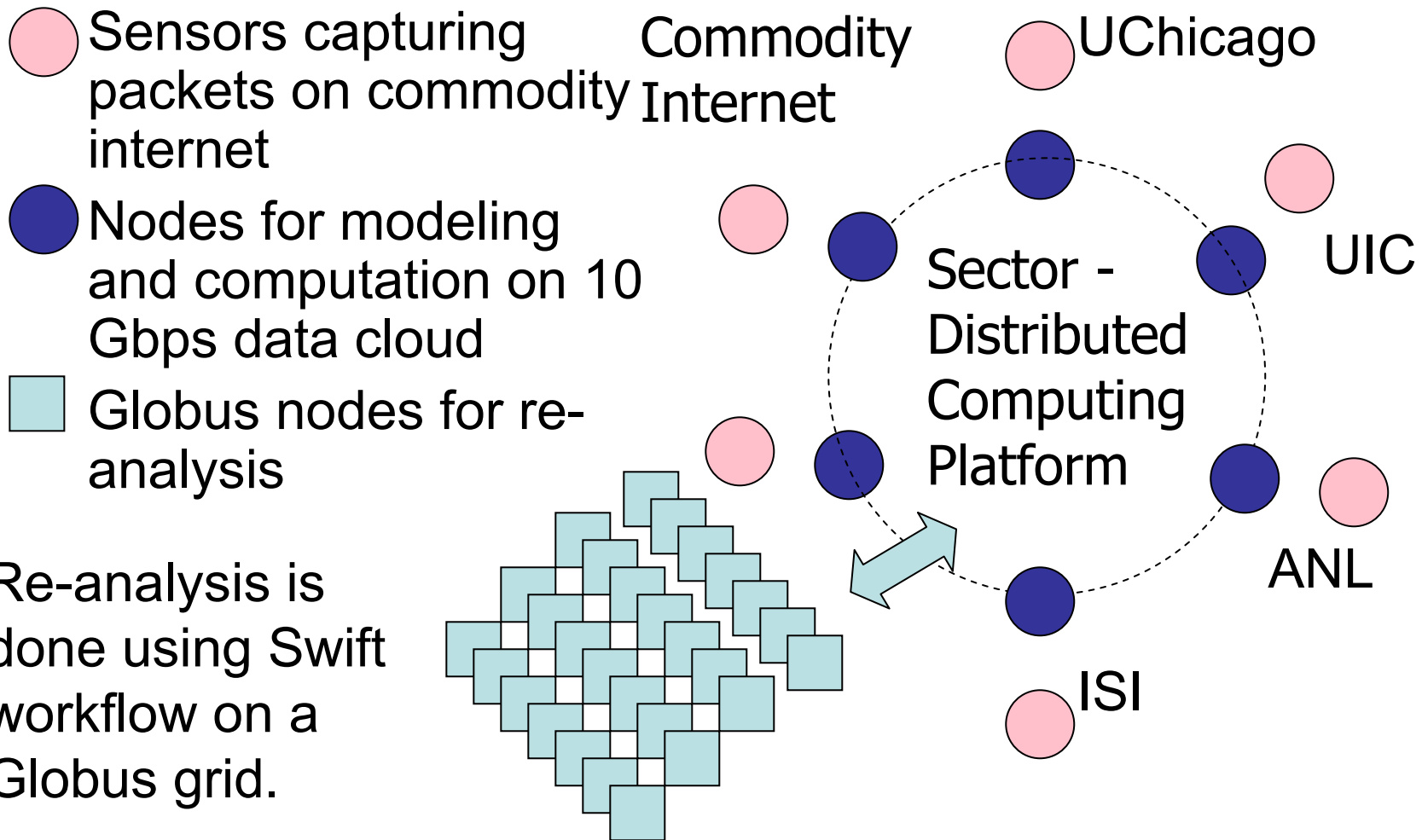
# Angle Architecture: Part 1 - Data Cloud

- Sensors capturing packets on commodity internet
- Nodes for modeling and computation on 10 Gbps data cloud



Data cloud = A collection of persistent and managed storage and computing resources connected by the Internet.

# Angle Architecture: Part 2 - Grid



# Part 5: What is Required for Deployment?

# Deployment Requires

- A data cloud infrastructure with distributed clusters and high performance networks.
- Software (Sector and UDT to manage the data)
- Angle data sets properly anonymized
- New concepts (emergent)
- New algorithms (algorithms to identify stable clusters)
- New visualization techniques for visualizing large collections of cluster models.
- Workflow algorithms for alerting and collaboration.

Without all this, the system  
cannot be used.



# Part 6

## What Are We Leaving to the Community?

- A terabyte (and growing) of data for the community.
- Tools for building data clouds to support data mining and distributed data mining.
- Some algorithms for identifying emergent clusters.
- Some papers.
- An invitation to join the project.

# Please Join Us at SC07



- Angle is one of two semi-finalists in the HPC Analytics Challenge at SC 07.

Thank you.

Some papers at  
[www.rgrossman.com](http://www.rgrossman.com)