YarcData
Getting to **Eureka!** faster™

Scaling up Graph Search Performance in Hardware …

to unlock the true potential of Big Data

Amar Shan

YarcData Inc.

WHO Is YarcData?

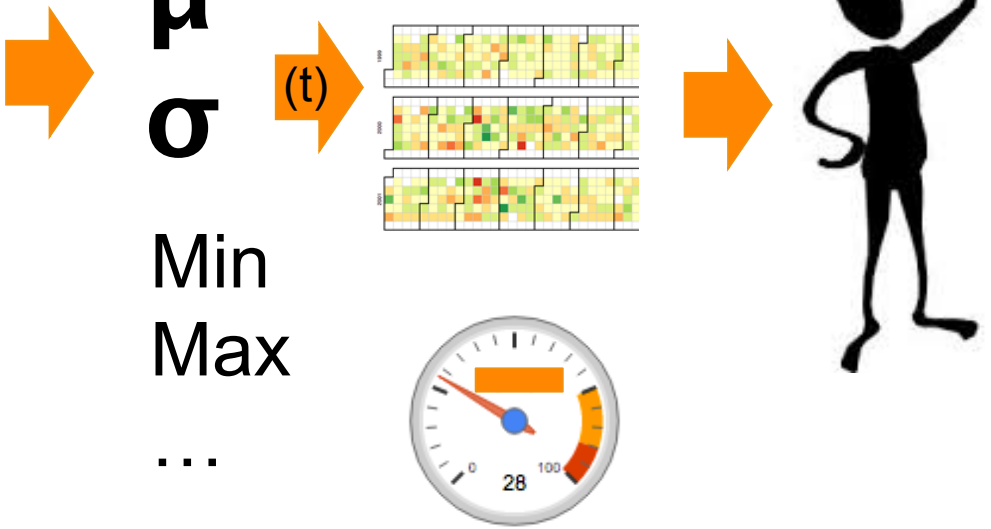
YarcData
Getting to Eureka! faster

Is Cray still around???

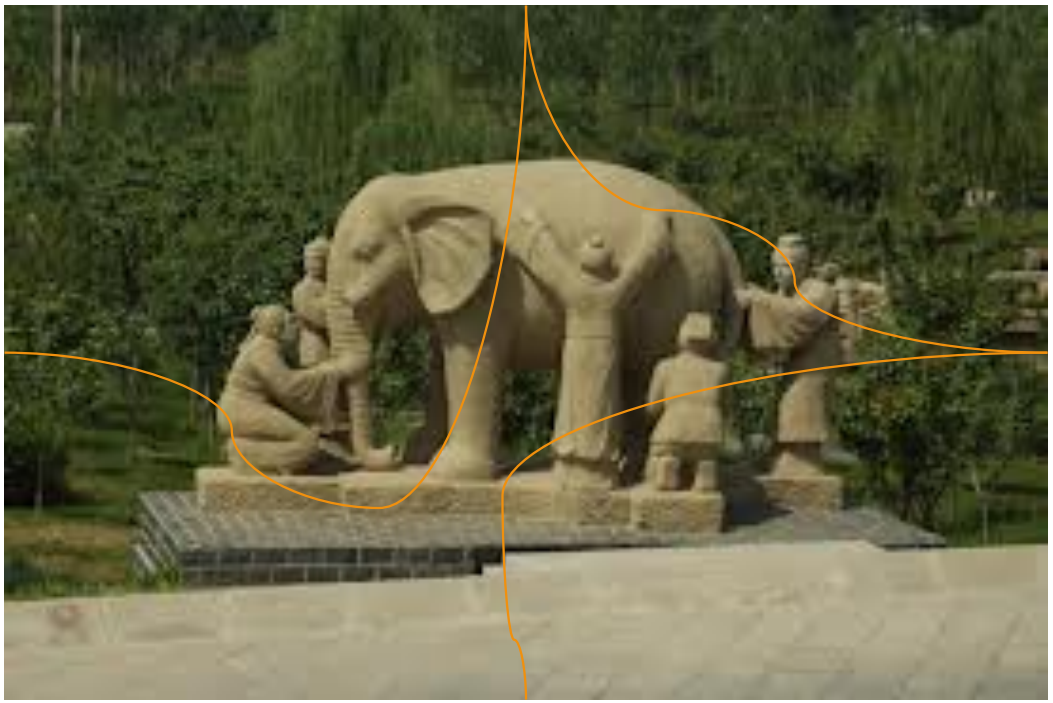
YarcData
Getting to Eureka! faster

Gain insights by “summarizing” from big data…

employing an analysis of a complex subject into a simplified, less detailed form; of, pertaining to, or employing reductionism; reductionistic.

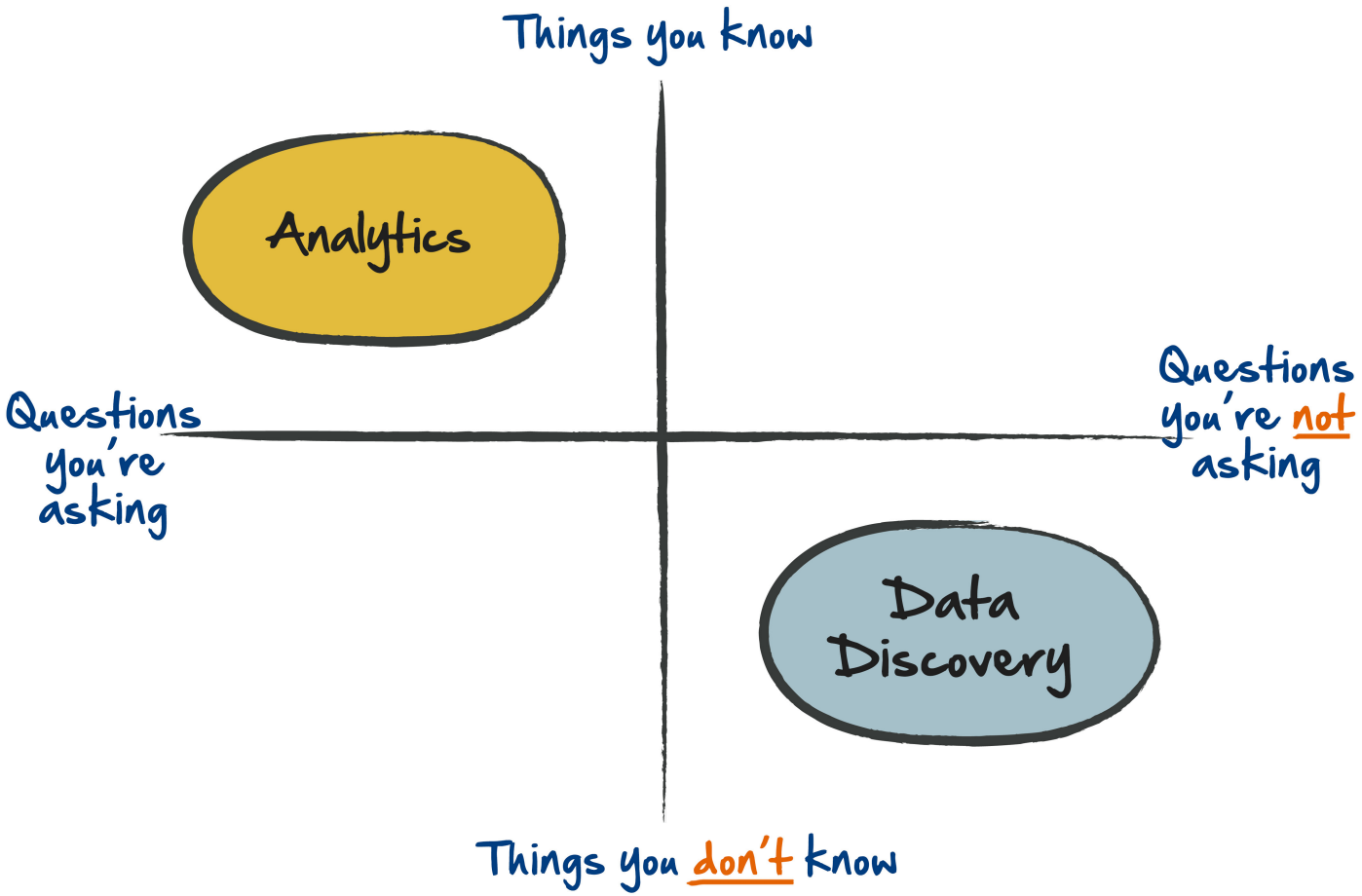


Except…



The big (data) question

"Take all these different data sources and put them together and then help me find something about the data that I don't already know…"



YarcData
Getting to Eureka! faster

The cycle of discovery in cancer research…

TCGA

FDA

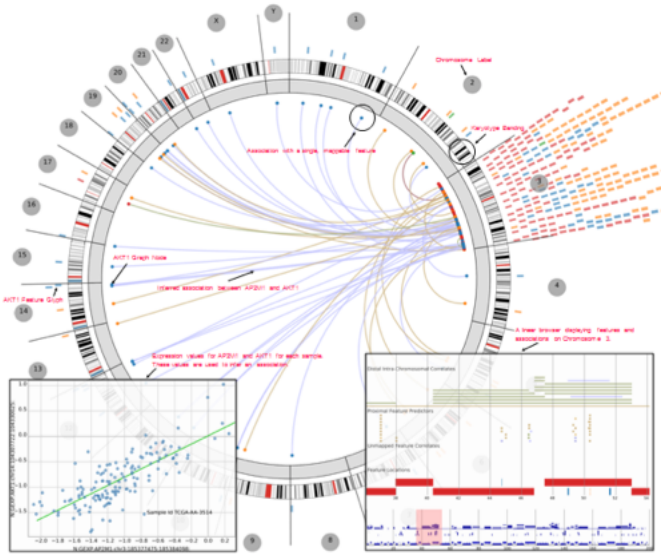
DrugBank
Open Data Drug & Drug Target Database

Pathway Commons

Human Protein Reference Database

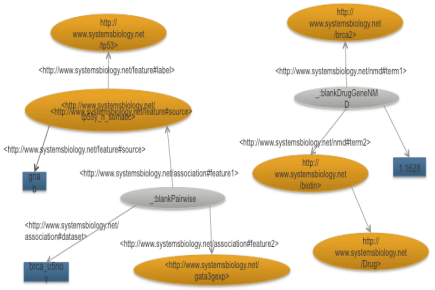
DOMINE
Database of Protein Domain Interactions

Others …



Formulate Partial Hypothesis – Statistical analysis used to assist

Load or Augment Required Data from many diverse data sources

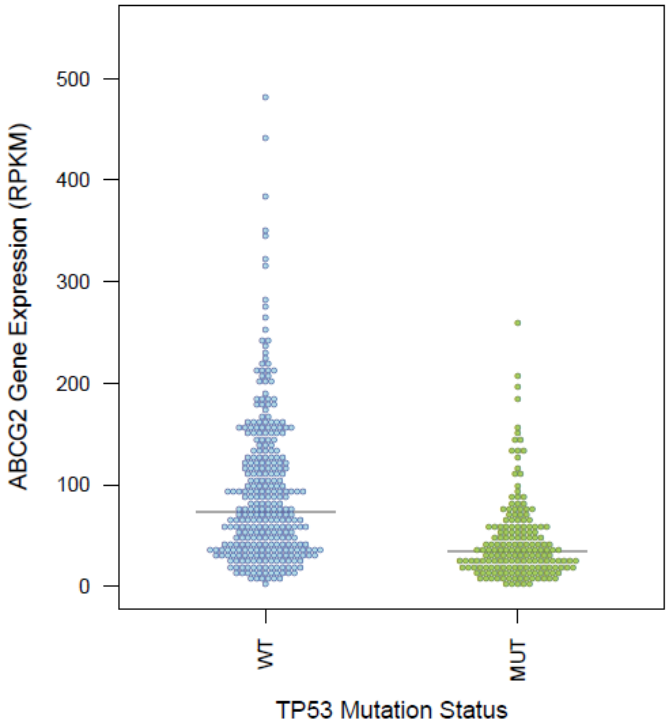


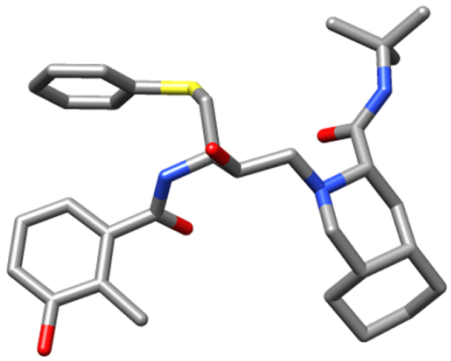
Validate Hypothesis against entire dataset – 5.3 billion relationships

YarcData
Getting to **Eureka!** faster

Recent Discovery: Repurpose HIV drug for Cancer

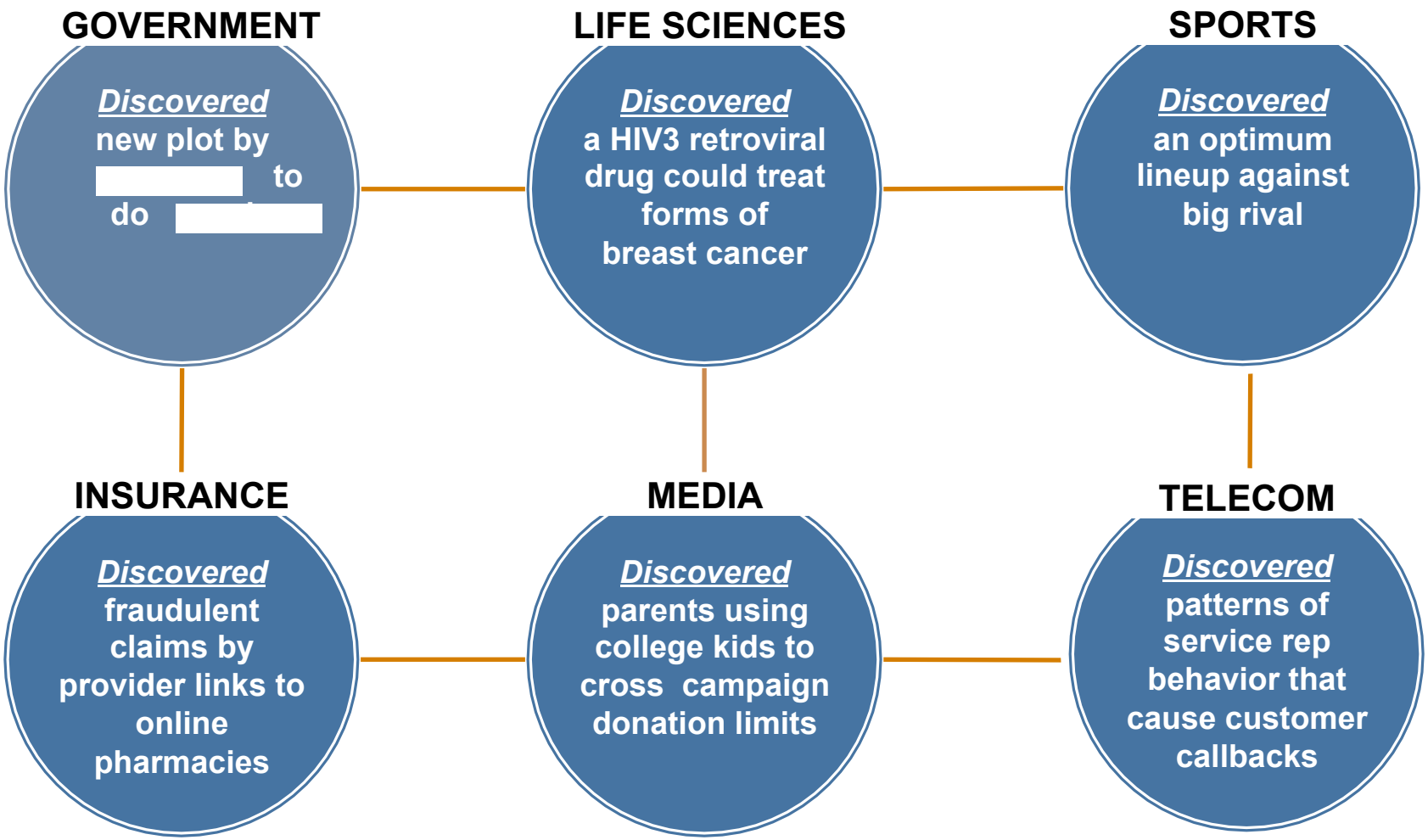
- **TP53** is frequently mutated in most tumor types
- **ABCG2**, also known as Breast Cancer Resistance Protein (BCRP), is associated with TP53 mutation in TCGA breast cancer data
- **Nelfinavir**, an HIV protease inhibitor, also binds ABCG2 and many other proteins
- High-throughput cell line screening of breast cancer cells recently identified Nelfinavir as a selective inhibitor. “It can be brought to HER2-breast cancer treatment trials with the same dosage regimen as that used among HIV patients. “ [Shim *et al.*





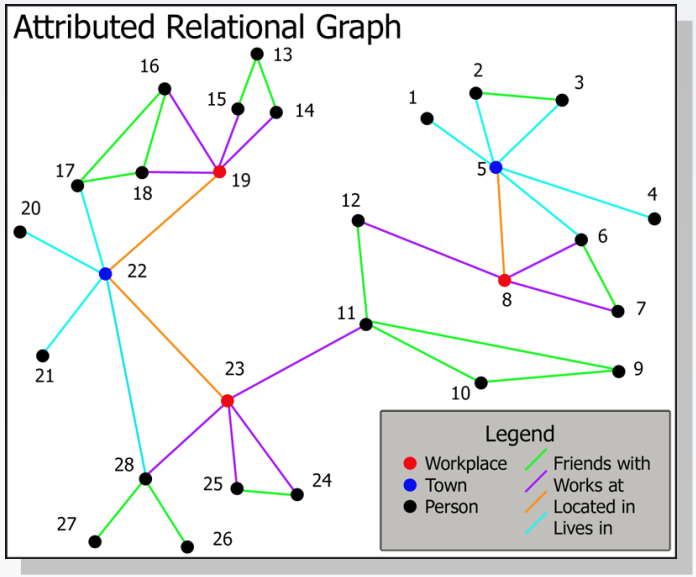
They discovered this in just 6 weeks!

YarcData: Eureka!



YarcData
Getting to Eureka! faster

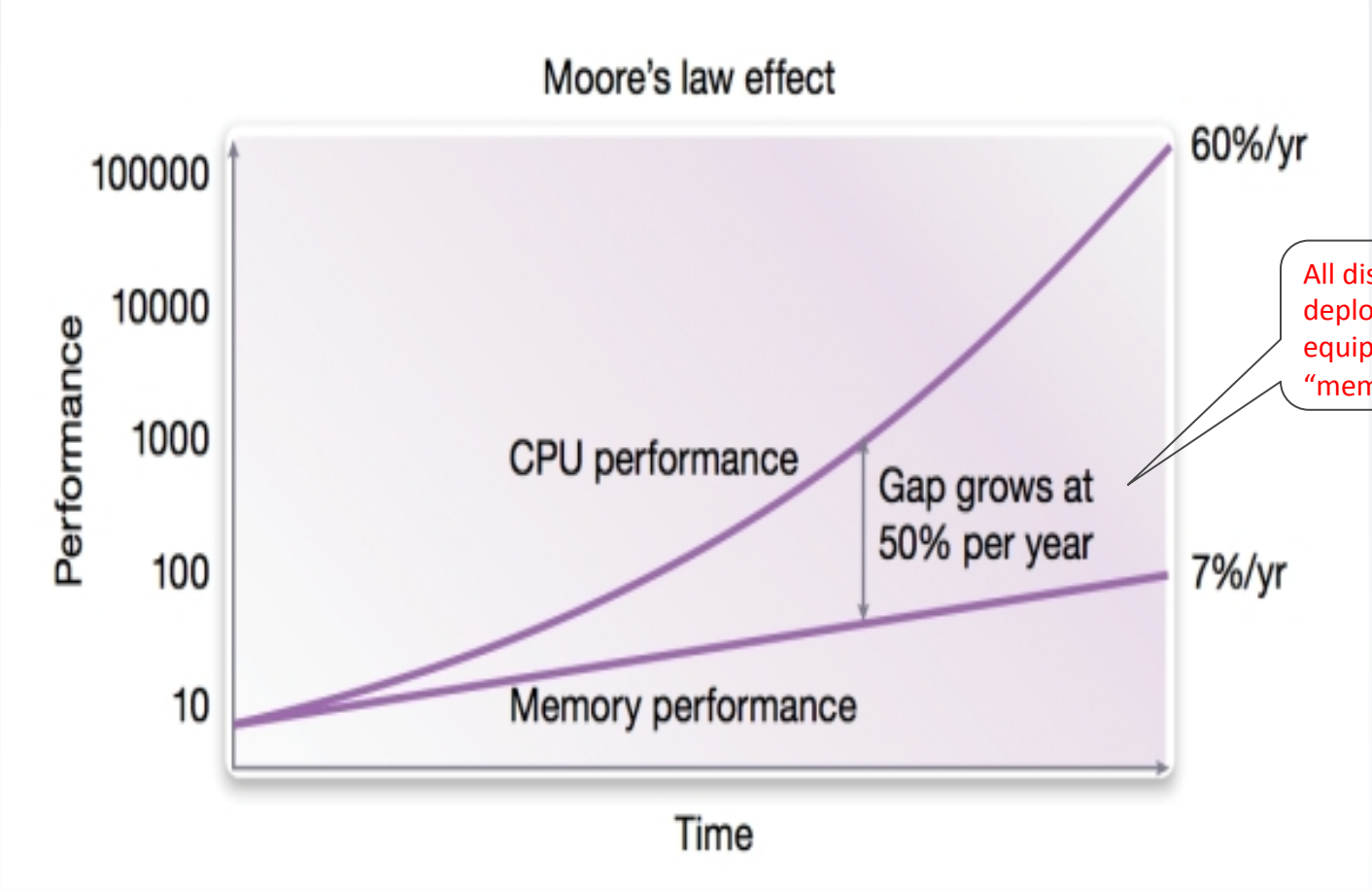
THE THING IS… DISCOVERY IS CACHE BUSTING…



- Discovery involves finding previously unknown patterns of relationships in the data.
- The process of discovery is never linear, and the questions to be answered are not known in advance.
- How then can you lay out memory for contiguous access???? *You can't.*
- **Discovery analytics involve random patterns of memory access, and are always cache busting!**

SO WHAT IF DISCOVERY ALGORITHMS ARE CACHE BUSTING?

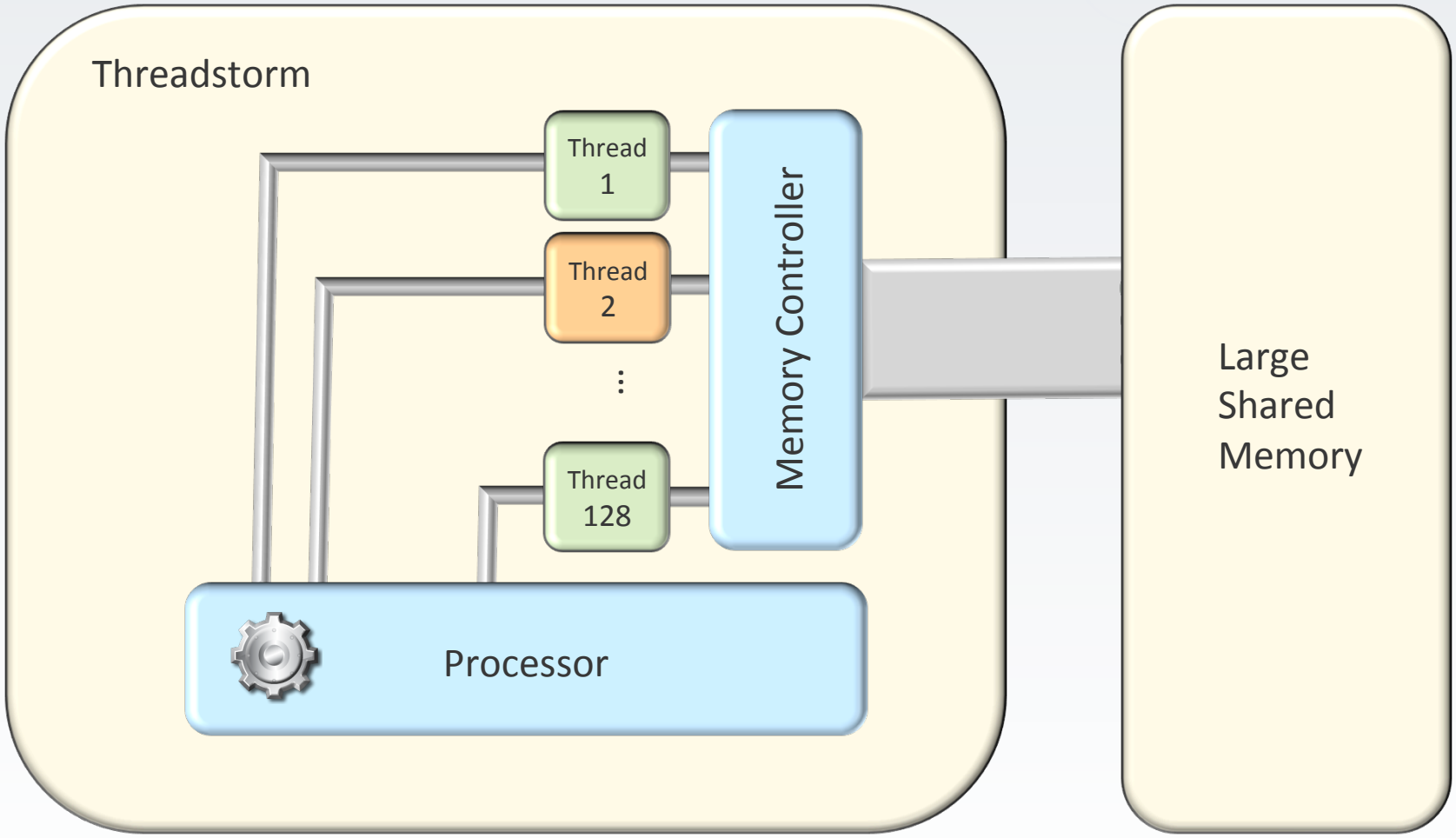
YarcData
Getting to Eureka! faster



Source: Scientific Programming, IEEE Computing Society

THREADSTORM: TOLERATES MEMORY LATENCY

YarcData
Getting to Eureka! faster

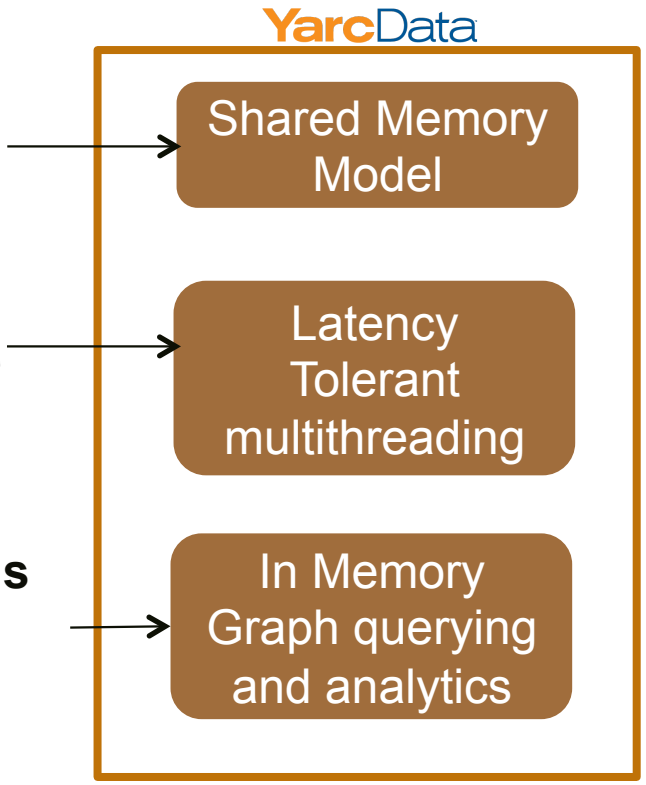


YarcData: Purpose-built for Data Discovery

Do not know the relationships In the data

Do not know the desired insight or the right question to ask

Do not know the paths linkages to explore diverse data sets



Sandia National Laboratories

	# PROCESSORS	TIME
Traditional Approaches	48	10.8 Hours
YarcData	32	30 sec

1,944 Times Faster !

Institute for Systems Biology

"In the amount of time it takes to validate one hypothesis, we can now validate 1000 hypotheses – increasing our success rate significantly." *– Dr. Ilya Shmulevich*

YarcData Getting to Eureka! faster

Requirements for Storage

- Bandwidth
  - Large scale graph loading
  - Checkpoints
  - Large result sets communicated between applications
  - *Parallel File System is ideal*
- Latency
  - Hadoop Giraph & other graph engines go to disk for graph nodes
  - Caching is ineffective – low latency reads are key…
  - Moving “hot data” into fast storage (SSD) is very useful
- Posix compliance

"You've got to be very careful if you don't know where you're going, because you might not get there."

Thank You!

Amar Shan
e. amar@cray.com
c. 778-881-4600