SIMULATION INFORMATICS

SAND2011-3203C

A data-driven methodology for synthesis from simulation results

David F. Gleich

John von Neumann Postdoctoral Fellow

Sandia National Laboratories

Livermore, CA

A bit more background on me

Before

Undergrad Harvey Mudd College (Joint CS/Math)

Internships Yahoo! and Microsoft Internships

Graduate Stanford (Computation and Mathematical Engineering)
Thesis: Models and Algorithm's for PageRank Sensitivity

Internships Intel, Joint Library of Congress/Stanford, Microsoft

Post-doc 1 Univ. of British Columbia

Now

Post-doc 2 2010 John von Neumann fellow @ Sandia

In August

Tenure-Track position at Purdue's CS department

It's time to ask:
What can science learn from Google?

-- Wired Magazine (2008)

http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

A tale of two computers …

Cielo (#10 in top 500)

Separate storage system

*This platform is **awesome** for simulating physical systems*

Cost ~$50 million

SNL/CA Nebula Cluster

2TB/core storage

*This platform is **awesome** for working with data*

Cost $150k

Simulation: The Third Pillar of Science

21st Century Science in a nutshell

Experiments are not practical / feasible.

Simulate things instead.

But do we trust the simulations?

We're trying!

Model Fidelity
Verification & Validation (V&V)
Uncertainty Quantification (UQ)

Insight and confidence requires multiple runs.

The problem Simulation ain't cheap!

We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.
--Wired (again)

21.1st Century Science
in a nutshell?

Simulations are too expensive.

Let data provide a surrogate.

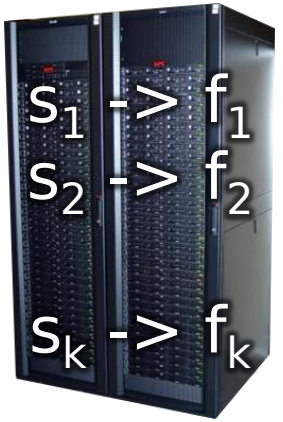
Input Parameters

s

Time history of simulation

f

The Database



Our CSRF project Simulation Informatics

To enable analysts and engineers to hypothesize from inexpensive data computations instead of expensive HPC computations.

The People

Myself

Paul G. Constantine
(2009 von Neumann)

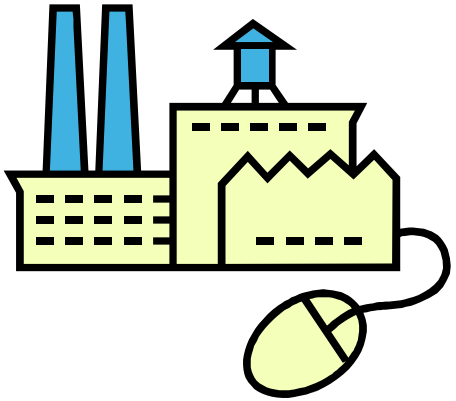
Stefan Domino

Jeremy Templeton

Joe Ruthruff

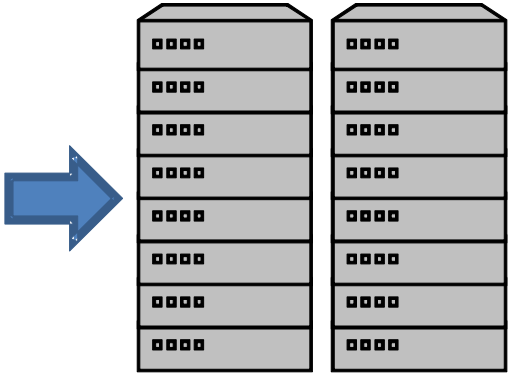
The idea and vision: Store the runs!

Supercomputer



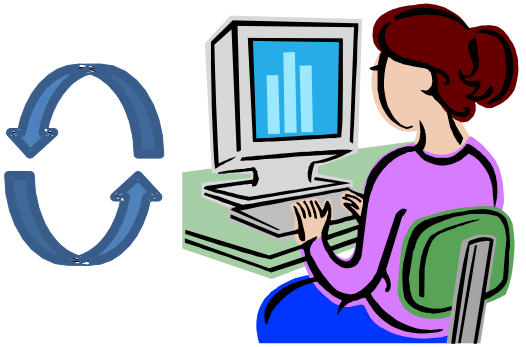
Each multi-day HPC simulation generates gigabytes of data.

MapReduce cluster



A MapReduce cluster can hold hundreds or thousands of old simulations …

Engineer



… enabling engineers to query and analyze months of simulation data for statistical studies and uncertainty quantification.

MapReduce

Originated at Google for indexing web pages and computing PageRank

The idea

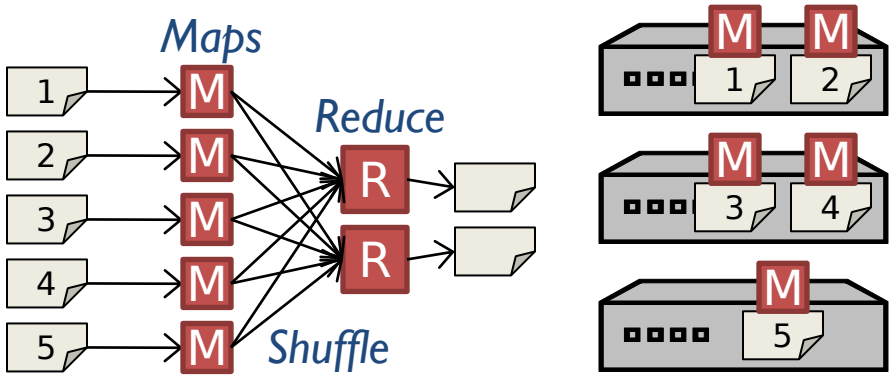
Bring the computations to the data.

Express algorithms in data-local operations.

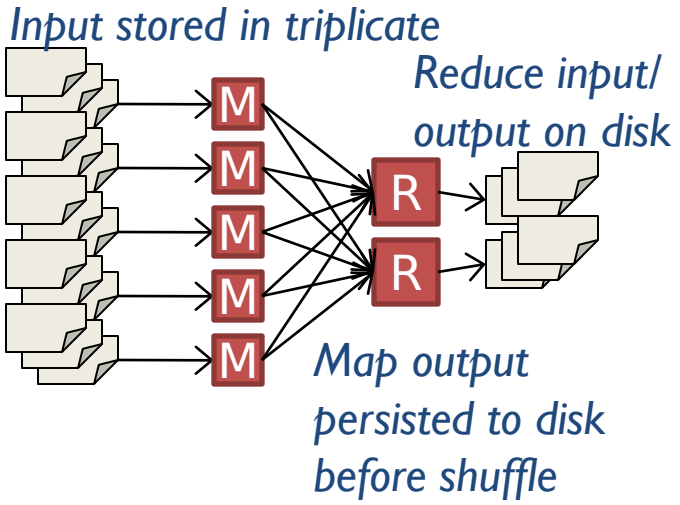
Implement one type of communication: shuffle.

Shuffle moves all data with the same key to the same reducer

Data scalable



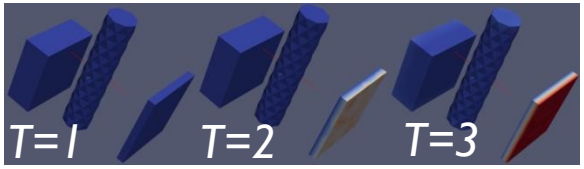
Fault tolerance by design



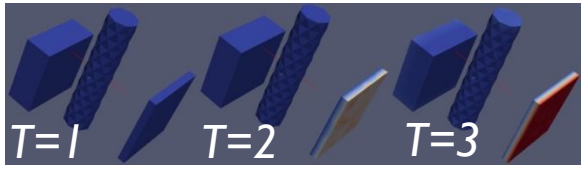
Mesh point variance in MapReduce

Three simulation runs for three time steps each.

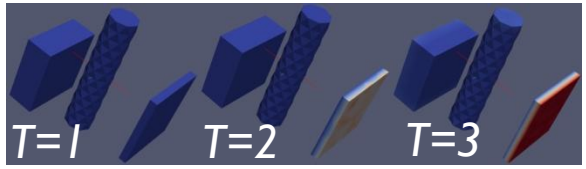
Run 1



Run 2

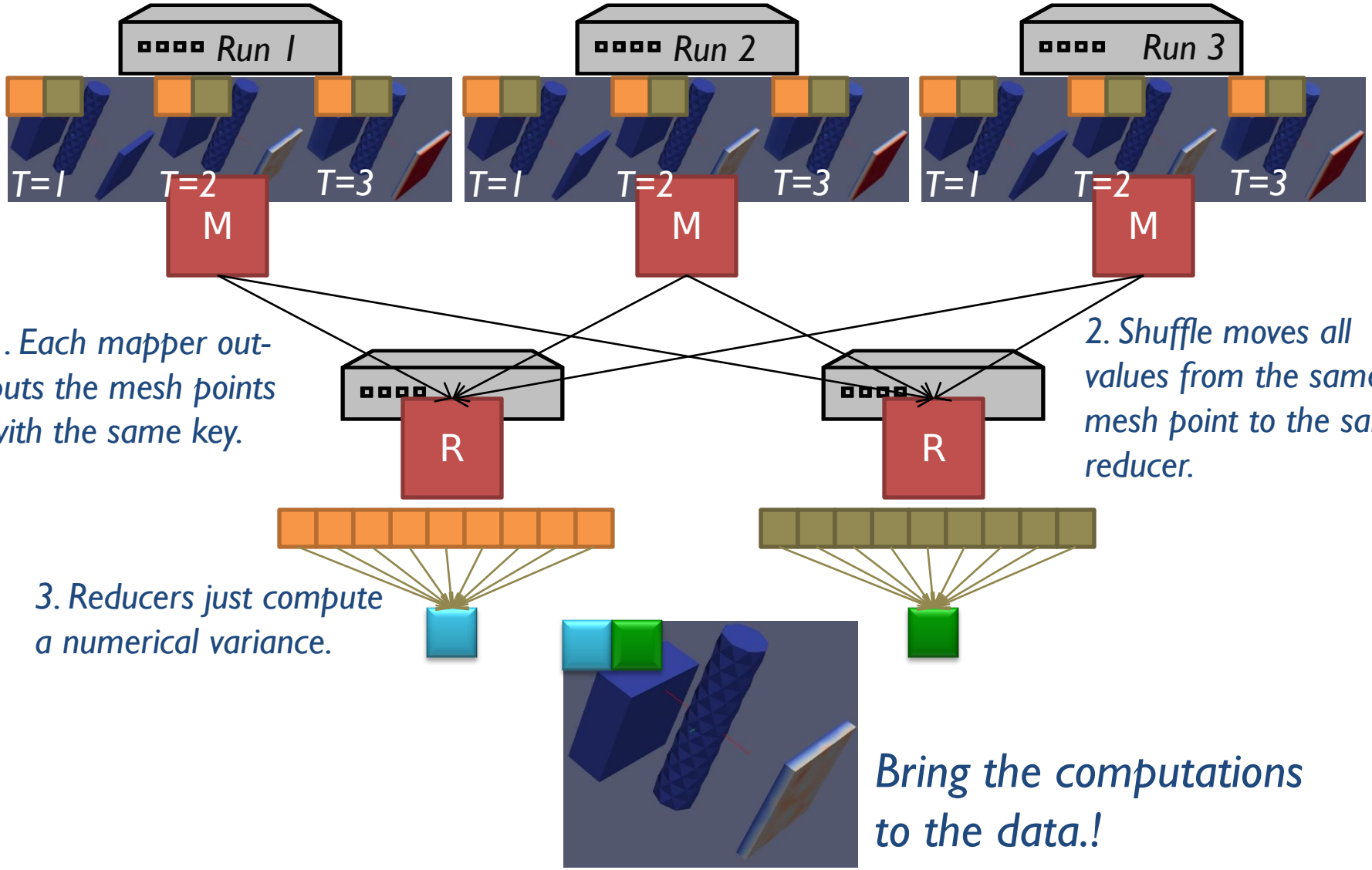


Run 3



Mesh point variance in MapReduce

Three simulation runs for three time steps each.



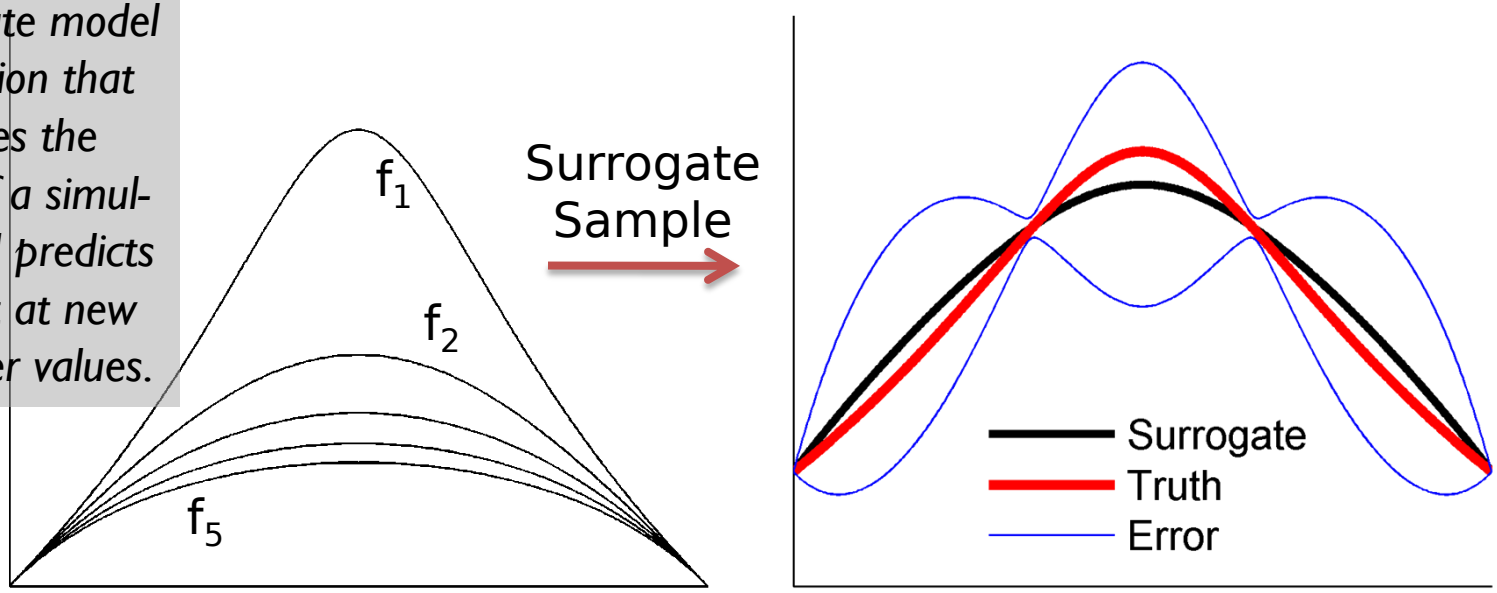
Hadoop???

Hadoop: an open-source MapReduce system supported by Yahoo!

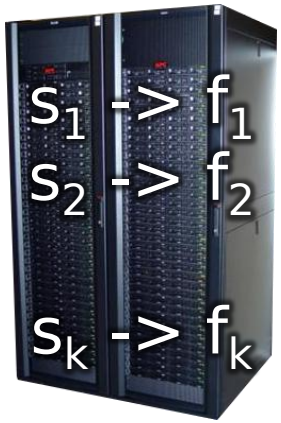
Photo credit, New York Times (2009)
http://www.nytimes.com/2009/03/17/technology/business-computing/17cloud.html

MapReduce and Surrogate Models

A surrogate model is a function that reproduces the output of a simulation and predicts its output at new parameter values.



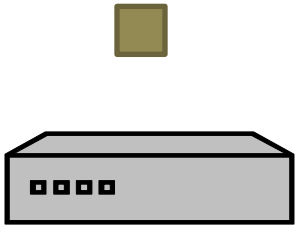
The Database



On the MapReduce cluster

Extraction

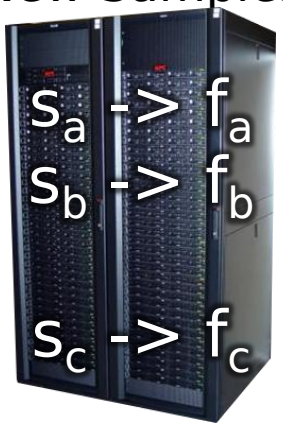
The Surrogate



Just one machine

Interpolation

New Samples



On the MapReduce cluster

MapReduce + Simulations

Thermal Race problem

Joe Ruthruff and Jeremy Templeton

Heating an unclassified NW model to simulate a fire

Goal make sure the weak link fails before the strong link

Objective determine where parameters are interesting

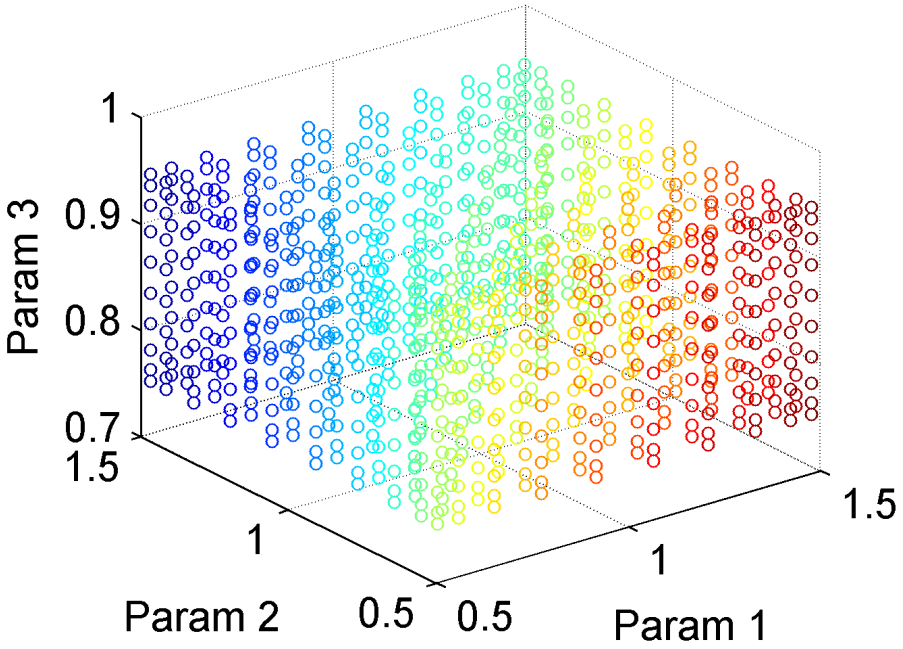
Parameters material properties

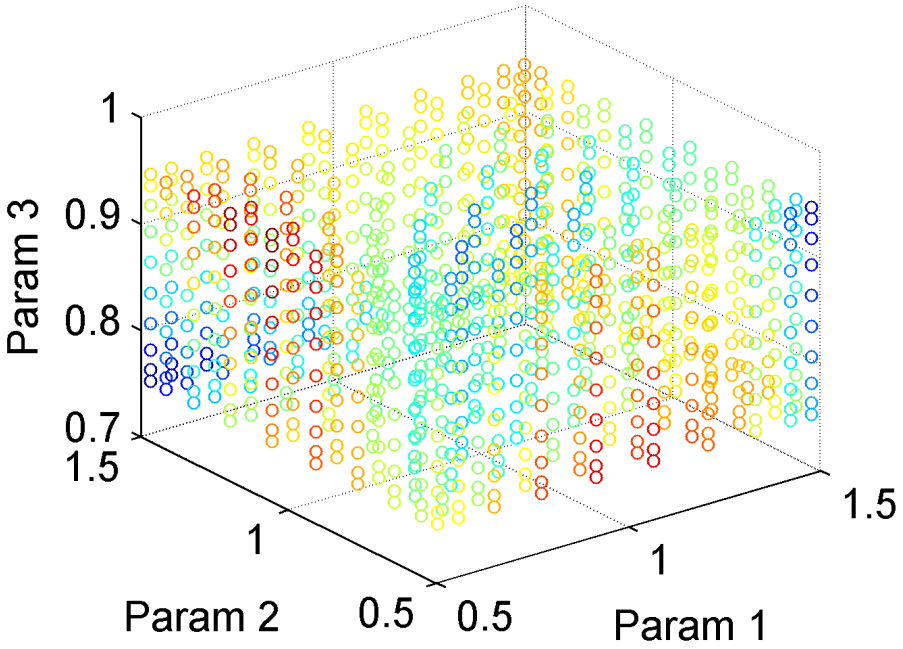
1000 sample parameter study using

30 min per run
700GB of Raw Data
~64GB of data for MapReduce

Our results
Extraction **30 min**
Interpolation **5 min**

Data driven Green's functions





The simulation varies most in the corners of the parameter space.

These analyses will help understand which parameters matter, and where to continue running the simulations.

What can science learn from Google?

The importance of data.

For more about this, see a talk and paper by Peter Norvig

"Internet-Scale Data Analysis"
http://www.stanford.edu/group/mmds/slides2010/Norvig.pdf

Halevy et al. "The unreasonable effectiveness of data." IEEE Intelligent systems, 2009.

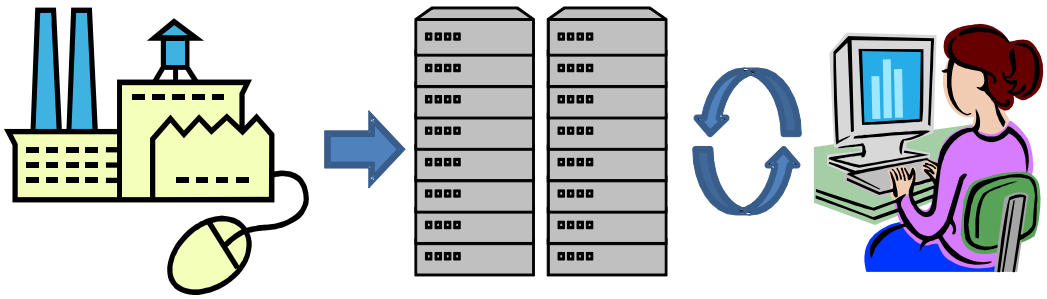
The future?

MapReduce computers embedded within super computers as the parallel file system and data analysis platform.

The best of both worlds.

Our work

Laying the foundation for useful surrogate models that will make data based simulation a compelling addition to the simulation revolution.



Any questions?