DATA FARMING:
METHODS FOR THE PRESENT,
OPPORTUNITIES FOR THE FUTURE

Susan M. Sanchez

SEED Center for Data Farming
Operations Research Department
Naval Postgraduate School

ISIM 2017 Research Workshop
Durham, U.K.
Data Mining vs. Data Farming
Data Mining vs. Data Farming

- Miners seek valuable buried nuggets
Data Mining vs. Data Farming

- Miners seek valuable buried nuggets
  - Miners have no control over what’s there or how hard it is to separate it out
Data Mining vs. Data Farming

- Miners seek valuable buried nuggets
  - Miners have no control over what’s there or how hard it is to separate it out
  - Data Mining seeks valuable information buried within massive amounts of data
Data Mining vs. Data Farming

• Miners seek valuable buried nuggets
  - Miners have no control over what’s there or how hard it is to separate it out
  - Data Mining seeks valuable information buried within massive amounts of data
• Farmers cultivate to maximize yield
Data Mining vs. Data Farming

- Miners seek valuable buried nuggets
  - Miners have no control over what’s there or how hard it is to separate it out
  - Data Mining seeks valuable information buried within massive amounts of data

- Farmers cultivate to maximize yield
  - Farmers manipulate the environment to their advantage: pest control, irrigation, fertilizer, etc.
Miners seek valuable buried nuggets

- Miners have no control over what’s there or how hard it is to separate it out
- Data Mining seeks valuable information buried within massive amounts of data

Farmers cultivate to maximize yield

- Farmers manipulate the environment to their advantage: pest control, irrigation, fertilizer, etc.
- Data Farming manipulates simulation models to advantage with designed experimentation
Data Mining vs. Data Farming

- Miners seek valuable buried nuggets
  - Miners have no control over what’s there or how hard it is to separate it out
  - Data Mining seeks valuable information buried within massive amounts of data

- Farmers cultivate to maximize yield
  - Farmers manipulate the environment to their advantage: pest control, irrigation, fertilizer, etc.
  - Data Farming manipulates simulation models to advantage with designed experimentation

One way of thinking of big data…any data set that pushes against the limits of currently available analysis technology
“Harnessing vast quantities of data rather than a small portion, and privileging more data of less exactitude, opens the door to new ways of understanding. It leads society to abandon its time-honored preference for causality, and in many instances tap the benefits of correlation.”

 Walls Streeters have the fastest computers, most sophisticated software and biggest databases money can buy, and yet many failed to see the 2008 crash coming. The hope that Big Data will make economics and other social sciences truly scientific— that is, precise and predictive— remains, for now, a fantasy.”

1 Hogan, J., 2014. “So Far, Big Data is Small Potatoes”
“Harnessing vast quantities of data rather than a small portion, and privileging more data of less exactitude, opens the door to new ways of understanding. It leads society to abandon its time-honored preference for causality, and in many instances tap the benefits of correlation.”

“Wall Streeters have the fastest computers, most sophisticated software and biggest databases money can buy, and yet many failed to see the 2008 crash coming. The hope that Big Data will make economics and other social sciences truly scientific— that is, precise and predictive— remains, for now, a fantasy.”

Simulators don’t have to choose!

Correlation = 0.947

1Hogan, J., 2001. “So Far, Big Data is Small Potatoes”
Large-scale computational experiments are transformative

“Petaflop machines like Roadrunner have the potential to fundamentally alter science and engineering…[allowing scientists to] perform experiments that would previously have been impractical.”

Large-scale computational experiments are transformative

“Petaflop machines like Roadrunner have the potential to fundamentally alter science and engineering...[allowing scientists to] perform experiments that would previously have been impractical.”


Experimentation is hard:
“$2^{100}$ is forever”
—Maj Gen Jasper Welch

Even with today’s most powerful computers, brute force exploration of 100 variables at 2 levels for a simulation that runs in one second would take many times the age of the universe... so we need to be smart!
Moore’s Law is not enough!

The “curse of dimensionality” cannot be solved by hardware alone.

Experimentation is hard:

“$2^{100}$ is forever”
—Maj Gen Jasper Welch

Even with today’s most powerful computers, brute force exploration of 100 variables at 2 levels for a simulation that runs in one second would take many times the age of the universe… so we need to be smart!

Large-scale computational experiments are transformative

“Petaflop machines like Roadrunner have the potential to fundamentally alter science and engineering…[allowing scientists to] perform experiments that would previously have been impractical.”


Petaflop = 1 quadrillion ops/second  
Cost of “Roadrunner”= $133 million
Large-scale computational experiments are transformative

“Petaflop machines like Roadrunner have the potential to fundamentally alter science and engineering...[allowing scientists to] perform experiments that would previously have been impractical.”


Moore’s Law is not enough!

The “curse of dimensionality” cannot be solved by hardware alone.

Experimentation is hard:

“$2^{100}$ is forever”

—Maj Gen Jasper Welch

Even with today’s most powerful computers, brute force exploration of 100 variables at 2 levels for a simulation that runs in one second would take many times the age of the universe... so we need to be smart!

Data farming helps overcome the curse of dimensionality...

With large-scale efficient experimental designs, we generate “better big data” and regularly study hundreds of factors for longer-running simulations in hours, days, or weeks on high-performance computing clusters...

Petaflop = 1 quadrillion ops/second
Cost of “Roadrunner”= $133 million
Simulation is different

Response Surface Complexity

maximal screening
main effects
iid errors

1st order

2nd order

smooth

minimal assumptions
non-smooth
complex errors

Simulation Experiments

Physical Experiments

Number of Factors

Many

Few

sequential bifurcation (SB)
Efficient R5 FF and CCD
coarse grids ($2^k$ factorial)

R4

R5

central composite (CCD)

combined designs

Latin hypercube (LH)

frequency designs
differential grids

fine grids ($m^k$ factorial)
Simulation is different

What is?
What if?
What matters?
What could be?
What should be?

How might we get there?
Simulation is different

How might we get there?

“think big!” — factors, features, flexibility
Simulation is different

Analyst may be more “expensive” than the data

What is?
What if?
What matters?
What could be?
What should be?

“think big!” — factors, features, flexibility
Simulation is different

Analyst may be more “expensive” than the data

• Don’t focus on keeping the #design points or #reps small, if you can make runs in a reasonable time

What is?
What if?
What matters?

What could be?

What should be?

How might we get there?

“think big!” — factors, features, flexibility
Simulation is different

Analyst may be more “expensive” than the data

• Don’t focus on keeping the #design points or #reps small, if you can make runs in a reasonable time

Make few assumptions

What is?
What if?
What matters?
What could be?
What should be?

How might we get there?

“think big!” — factors, features, flexibility
Simulation is different

Analyst may be more “expensive” than the data
  • Don’t focus on keeping the #design points or #reps small, if you can make runs in a reasonable time

Make few assumptions
  • Retain flexibility in the design and analysis, “explore” the output to gain understanding: we advocate space-filling designs

What is?
What if?
What matters?
What could be?
What should be?

How might we get there?

“think big!” — factors, features, flexibility
Simulation is different

Analyst may be more “expensive” than the data

- Don’t focus on keeping the #design points or #reps small, if you can make runs in a reasonable time

Make few assumptions

- Retain flexibility in the design and analysis, “explore” the output to gain understanding: we advocate space-filling designs

You can always double check!

“think big!” — factors, features, flexibility
Simulation is different

Analyst may be more “expensive” than the data
- Don’t focus on keeping the #design points or #reps small, if you can make runs in a reasonable time

Make few assumptions
- Retain flexibility in the design and analysis, “explore” the output to gain understanding: we advocate space-filling designs

You can always double check!
- Confirmation runs of your simulation can be made to see how well your metamodels perform at previously untested design points

“What is?”
- What is?

“What if?”
- What if?

“What matters?”
- What matters?

“What could be?”
- What could be?

“What should be?”
- What should be?

How might we get there?

“think big!” — factors, features, flexibility
Simulation is different

Analyst may be more “expensive” than the data

• Don’t focus on keeping the #design points or #reps small, if you can make runs in a reasonable time

Make few assumptions

• Retain flexibility in the design and analysis, “explore” the output to gain understanding: we advocate space-filling designs

You can always double check!

• Confirmation runs of your simulation can be made to see how well your metamodels perform at previously untested design points

Statistical significance is NOT practical importance

“think big!” — factors, features, flexibility

What is?
What if?
What matters?
What could be?
What should be?
How might we get there?
Simulation is different

Analyst may be more “expensive” than the data

• Don’t focus on keeping the #design points or #reps small, if you can make runs in a reasonable time

Make few assumptions

• Retain flexibility in the design and analysis, “explore” the output to gain understanding: we advocate space-filling designs

You can always double check!

• Confirmation runs of your simulation can be made to see how well your metamodels perform at previously untested design points

Statistical significance is NOT practical importance

• You may throw out many metamodlel terms with low p-values without sacrificing explanatory power or affecting the decision

“think big!” — factors, features, flexibility
SEED student impact...

**Saving money: Major Chris Nannini, USA**
“The Analysis Of The Assignment Scheduling Capability For Unmanned Aerial Vehicles (ASC-U) Simulation Tool.”
M.S. in Operations Research

**New methods: LTC Tom Cioppa, USA**
“Efficient Nearly Orthogonal and Space-Filling Experimental Designs for High-Dimensional Complex Models”
Ph.D. in Operations Research

**Helping the Fleet: LT Chad Kaiser, USN**
“Air Defense Against UAS Kamikaze Saturation Attack”
M.S. in Operations Research

“If the Division Commanders want a UAV at their level and have nothing else, we ought to give it to them.”

“The UAV modeling...harvested $6 billion in savings and 6,000 to 10,000 billets, that’s a brigade’s worth of soldiers. Over 20 years that allowed us to avoid a cost of $20 billion.”
– Michael F. Bauman (2007), Director of the United States Army Training and Doctrine Command Analysis Center.

U.S. Navy TACBUL AD 09-01, “USN Surface Weapon System Capabilities and Limitations against low slow flyers (Helicopters/Small Aircraft/UAVs).”
Graphs from a few examples

• **STORM**
  - U.S. Navy campaign analysis model
  - ~40MB of input data spread over 150 input files
  - A single replication takes hours to complete, yields tens or hundreds of GBs of output data (mix of database fields, large flat files)
  - Graphs shown come from a notional training scenario
  - See “Improving U.S. Navy campaign analysis with big data” by Morgan, Schramm, Smith, Lucas, McDonald, Sanchez, Sanchez, & Upton (2017), forthcoming in Interfaces

• **Fleet management model (matlab)**
  - Australia using for its naval helicopter fleet (30 year lifetime)
  - Exploring how results depend on different policies
  - Graphs shown are based on notional data
21 responses indicate whether or not the naval campaign has gone well for the notional Carthage empire (“Blue” side)
Delving deeper

We can look at correlations or clusters of the responses

![Heatmap and scatter plot](image)

### Table: Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th>SubLosses</th>
<th>SubSurfaceLosses</th>
<th>AmphibLosses</th>
<th>CarrierLosses</th>
<th>C2_RedAdvSAMsitesDead_count</th>
<th>C2_isRedCarrierDead_count</th>
<th>C2_BlueAirSupremacy_count</th>
</tr>
</thead>
<tbody>
<tr>
<td>SubLosses</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SubSurfaceLosses</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AmphibLosses</td>
<td>0.08</td>
<td>0.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CarrierLosses</td>
<td>0.04</td>
<td>0.23</td>
<td>0.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2_RedAdvSAMsitesDead_count</td>
<td>0.14</td>
<td>-0.27</td>
<td>0.1</td>
<td>0.27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2_isRedCarrierDead_count</td>
<td>0.15</td>
<td>-0.21</td>
<td>0.06</td>
<td>0.21</td>
<td>0.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2_BlueAirSupremacy_count</td>
<td>0.12</td>
<td>-0.04</td>
<td>-0.11</td>
<td>0.07</td>
<td>0.39</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.08</td>
<td>-0.15</td>
<td>-0.35</td>
<td>0.06</td>
<td>0.19</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

![Graph showing replications by cluster](image)
Heat maps can show conditions, events, or resources over time, rather than just end-of-run results.
How often are different rule conditions triggered? These side-by-side graphs show median times and how often different phases of the plans fired, by cluster.

delving deeper
Code line #s for phases A, B, C, D, E triggering

Pink boxes show code line #s for next event, Gray boxes show median time (top) and number of replications where condition triggered (bottom): the latter also corresponds to shading
Identify design points that satisfy multiple criteria
then use partition trees / other metamodels to identify important factors and thresholds
Genes represent factors of interest, such as UAV speed or search pattern.

Initial parents are chosen to give diversity in input space.

Offspring are mutants of their parents.

Successive generations of ARTeMIS reveal the Pareto frontier of non-dominated solutions. The absence of points in the lower left indicates that a moderate cost must be incurred in order to reduce loss.

“loss” in the robust sense: minimizing average (squared deviation from target) across the noise space.
Future simulation **clients**:

- **Complex problems**
  
  Avoid Type III errors, embrace computational tractability

- **Complex questions**
  
  Big simulation data leads to new and interesting insights

- **Comfort with computerized and computer-based decisions**
  
  more data science professionals, new tools for simulation community, greater range of applications
Future simulation methods:

• Continual processing
  Simulation study as a process, not an end state;
  Leverage unused computing cycles

• Changing areas of research emphasis
  Multi-objective, adaptive procedures,
  parallel computing, large-scale simulation experiments, exploration/optimization on
  adaptively updated metamodels

• Causal computerized decision making
  Simulation: the gold standard for model-driven big data and inferential decision
  making within big data analytics
Future simulation models

- Living models
  Automated links between real-world data capture and simulation modeling environments
  Models / confederations of models that evolve over time
  Listener event graph objects (LEGOs)
Future research / application by data scientists

• Anticipate changes, suggest “what if?”
  Massive sensitivity analysis of metamodels might indicate what to watch for, suggest interventions for direct experimentation

• Methods that leverage structure
  Data from designed experiments has some different characteristics than observational data

• Smarter computational agents
  Intelligent agents search through model-driven data sets, identifying important factors and interesting features
find resources, or add yourself to our mailing list

http://harvest.nps.edu
bonus
Thus, of course, the correlation between the input and output is very close, and the number of acquired observations is 40.

The fastest packages were DiceKriging, laGP, sklearn, and DACE, which only took a minute to fit the data.

The next slowest is mlegp, taking about eight minutes, with GPy only slightly faster.

The run times are enormous differences between the fastest and the slowest packages performing the same task.

Following Sacks et al. (1989) the surface is modeled as a mean, a factor of over 1000 between the fastest and the slowest packages performing the same task.

Figure 2: Borehole 4-D and 8-D comparison. All four plots are on the same scale.

Figure 3: Run times (seconds) for Borehole 8-D with $d=8$, $n=160$.

The output is one dimensional, the input is $x_1, \ldots, x_d$. The surface is modeled as a mean, variance, and 1. These are stored in the rows of the design matrix $X$.

The OTL circuit function is used. Ben-Ari and Steinberg (2007) use a test function that describes an output transformer.

The OTL push-pull circuit. There are six input parameters, five for resistors (b, c, d, d, d). The remaining parameter is the gain of the output transformer.

X has a multivariate normal distribution. If we have a factor of over 1000 between the fastest and the slowest packages performing the same task.

The fastest packages were DiceKriging, laGP, sklearn, and DACE, which only took a minute to fit the data.

The next slowest is mlegp, taking about eight minutes, with GPy only slightly faster.

The run times are enormous differences between the fastest and the slowest packages performing the same task.

Following Sacks et al. (1989) the surface is modeled as a mean, a factor of over 1000 between the fastest and the slowest packages performing the same task.

Figure 2: Borehole 4-D and 8-D comparison. All four plots are on the same scale.

Figure 3: Run times (seconds) for Borehole 8-D with $d=8$, $n=160$.

The output is one dimensional, the input is $x_1, \ldots, x_d$. The surface is modeled as a mean, variance, and 1. These are stored in the rows of the design matrix $X$.

The OTL circuit function is used. Ben-Ari and Steinberg (2007) use a test function that describes an output transformer.

The OTL push-pull circuit. There are six input parameters, five for resistors (b, c, d, d, d). The remaining parameter is the gain of the output transformer.

X has a multivariate normal distribution. If we have a factor of over 1000 between the fastest and the slowest packages performing the same task.

The fastest packages were DiceKriging, laGP, sklearn, and DACE, which only took a minute to fit the data.

The next slowest is mlegp, taking about eight minutes, with GPy only slightly faster.

The run times are enormous differences between the fastest and the slowest packages performing the same task.