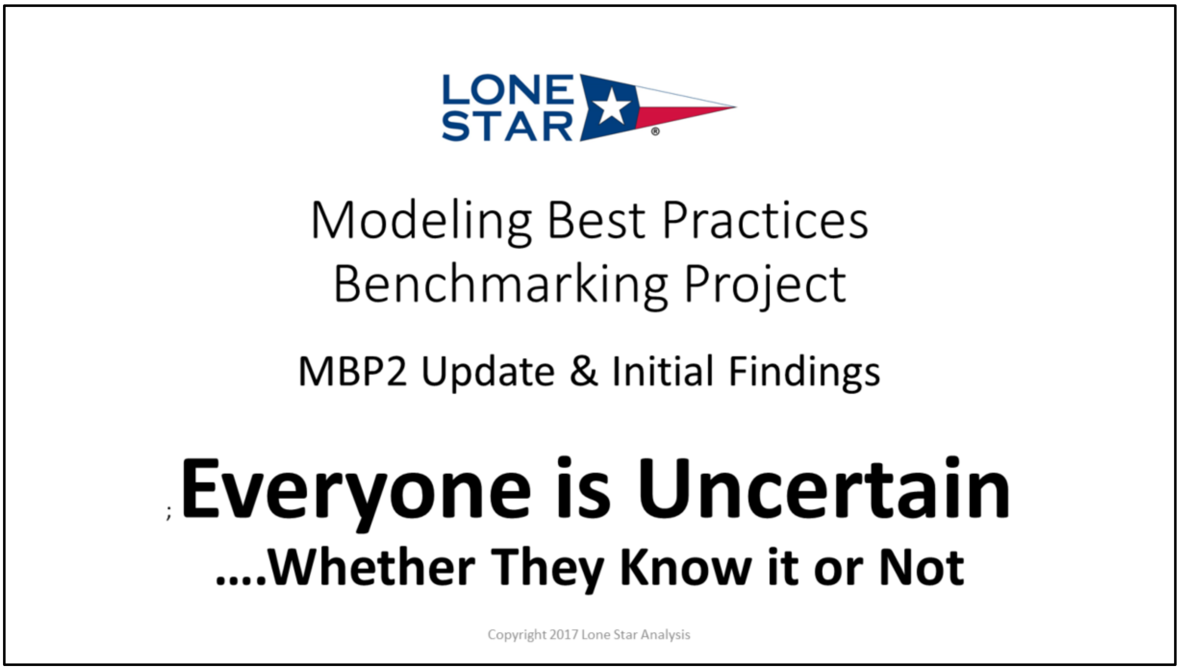


Modeling Best Practices Benchmarking Project

MBP2 Update & Initial Findings

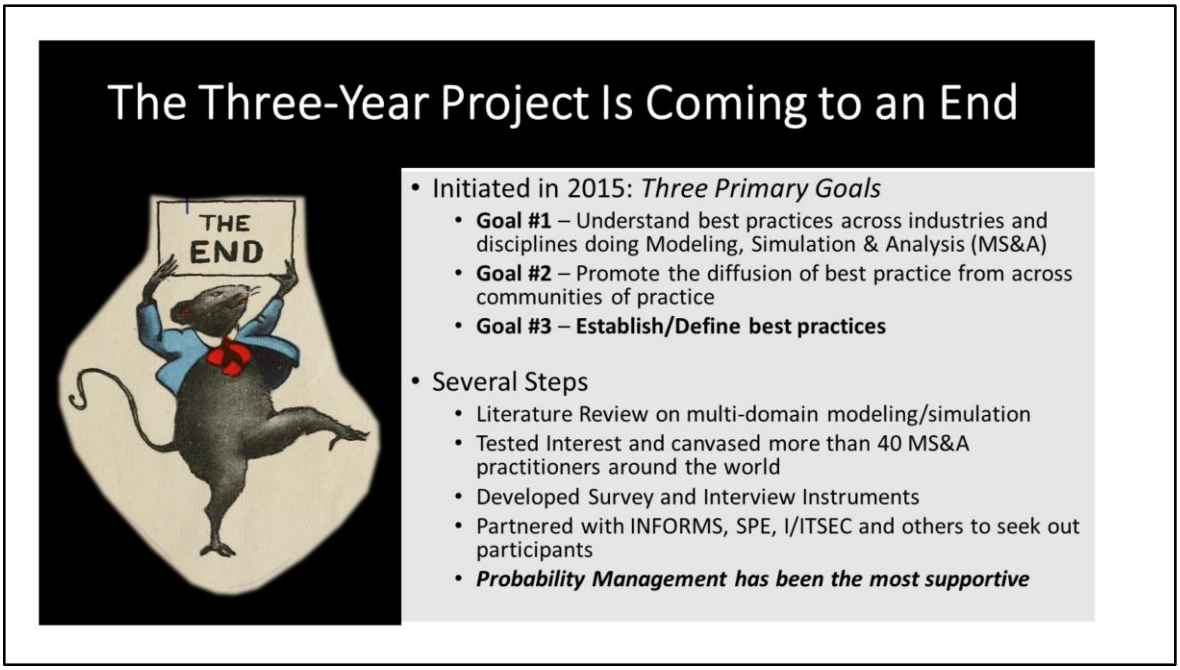
Everyone is Uncertain

….Whether They Know it or Not

Thank you for the chance to talk about MBP2. We have been on a long journey. Today we are far enough along to know some of the results, so we'll share some initial findings.

Some of our focus today may not prove to be the most important findings, or the most valuable for the Modeling, Simulation, and Analysis community. Rather, we will talk today about some interesting things, or at least things which interest me; five uncertainties.

Since this presentation was created for Probability Management, the leading group in the "Arithmetic of Uncertainty" this seems like a great place to begin our series to report the findings of MBP2.



The Three-Year Project Is Coming to an End

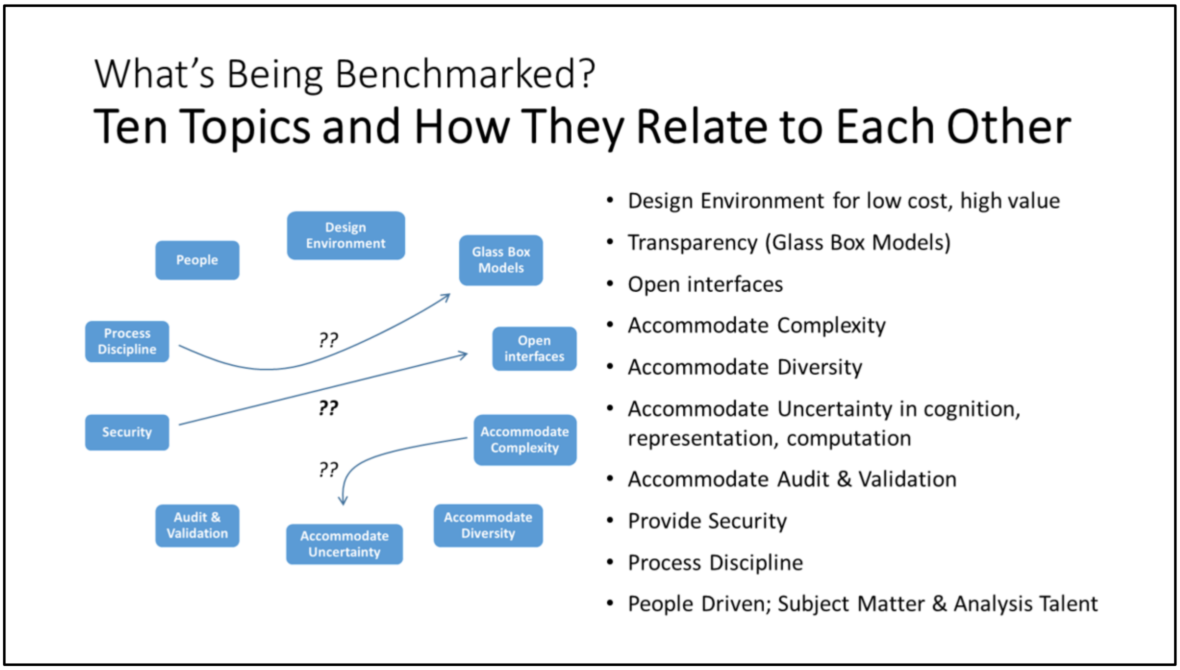
- Initiated in 2015: *Three Primary Goals*
  - **Goal #1** – Understand best practices across industries and disciplines doing Modeling, Simulation & Analysis (MS&A)
  - **Goal #2** – Promote the diffusion of best practice from across communities of practice
  - **Goal #3 – Establish/Define best practices**
- Several Steps
  - Literature Review on multi-domain modeling/simulation
  - Tested Interest and canvased more than 40 MS&A practitioners around the world
  - Developed Survey and Interview Instruments
  - Partnered with INFORMS, SPE, I/ITSEC and others to seek out participants
  - ***Probability Management has been the most supportive***

The Modeling Best Practices Benchmarking Project (MBP2) seeks to survey a broad range of modeling practice and share insights among professionals who might not otherwise have the opportunity to gain insight from outside their own communities.

Modeling, simulation and analysis (MS&A) supports a wide range of economic, academic and governmental efforts. Different practitioner communities have agreed on MS&A practices. But there is little interaction among communities. As a result, best practices in one MS&A community may not be familiar to others.

MBP2 goals are to understand best practices across industries and disciplines, and to promote diffusion of best practice, and to define best practice as a set of standards which apply broadly.

Acknowledgments attached to this presentation provides a partial list of organizations supporting MBP2; no organization has been more supportive than Probability Management.



What's Being Benchmarked?

Ten Topics and How They Relate to Each Other

People
Design Environment
Glass Box Models
Process Discipline
??
Open interfaces
Security
??
Accommodate Complexity
??
Audit & Validation
Accommodate Uncertainty
Accommodate Diversity

- Design Environment for low cost, high value
- Transparency (Glass Box Models)
- Open interfaces
- Accommodate Complexity
- Accommodate Diversity
- Accommodate Uncertainty in cognition, representation, computation
- Accommodate Audit & Validation
- Provide Security
- Process Discipline
- People Driven; Subject Matter & Analysis Talent

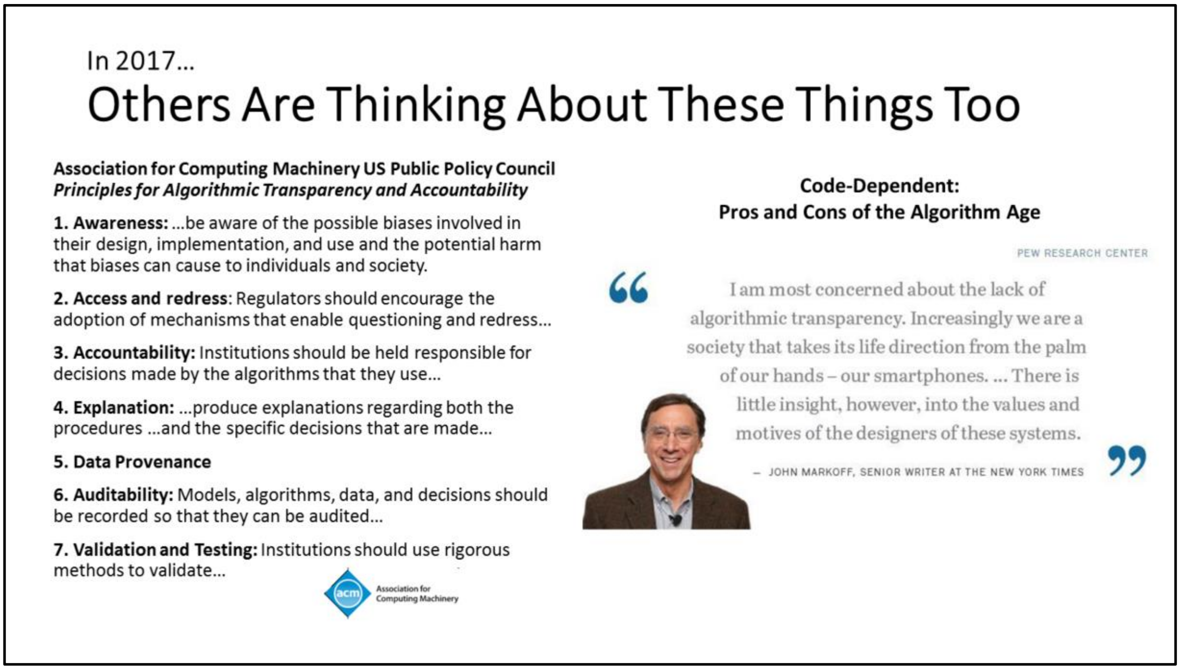
When we began our benchmarking we faced the challenge of choosing specific topics to explore. Because there was very little literature covering multi-domain MS&A, we developed, and revised a list of topics until we settled on these ten topics.

We tested these topics in early interviews and found high performing organizations could easily identify with them. And, these organizations had clear relationships existing among these areas.

In contrast, low performing organizations struggled to discuss these topics, and had no clear understanding of their relationships. This contrast led us to believe the list was useful for MS&A practitioners, across a range of performance levels.

Based on the early interviews, we designed a survey instrument using this framework.

In the meantime, some machine learning advocates have made claims about generic application of those methods. So, the interest in general MS&A outside one discipline has increased.



In 2017…

Others Are Thinking About These Things Too

Association for Computing Machinery US Public Policy Council
Principles for Algorithmic Transparency and Accountability

1. Awareness: …be aware of the possible biases involved in their design, implementation, and use and the potential harm that biases can cause to individuals and society.

2. Access and redress: Regulators should encourage the adoption of mechanisms that enable questioning and redress…

3. Accountability: Institutions should be held responsible for decisions made by the algorithms that they use…

4. Explanation: …produce explanations regarding both the procedures …and the specific decisions that are made…

5. Data Provenance

6. Auditability: Models, algorithms, data, and decisions should be recorded so that they can be audited…

7. Validation and Testing: Institutions should use rigorous methods to validate…

Code-Dependent:
Pros and Cons of the Algorithm Age

PEW RESEARCH CENTER

> “I am most concerned about the lack of algorithmic transparency. Increasingly we are a society that takes its life direction from the palm of our hands – our smartphones. … There is little insight, however, into the values and motives of the designers of these systems.”
>
> – JOHN MARKOFF, SENIOR WRITER AT THE NEW YORK TIMES

It is gratifying to see that today, others have identified many of these same issues as topics for study and concern.

In January 2017, the Association for Computing Machinery US Public Policy Council (USACM) published the statement shown on the left. It can be found on the web here: http://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

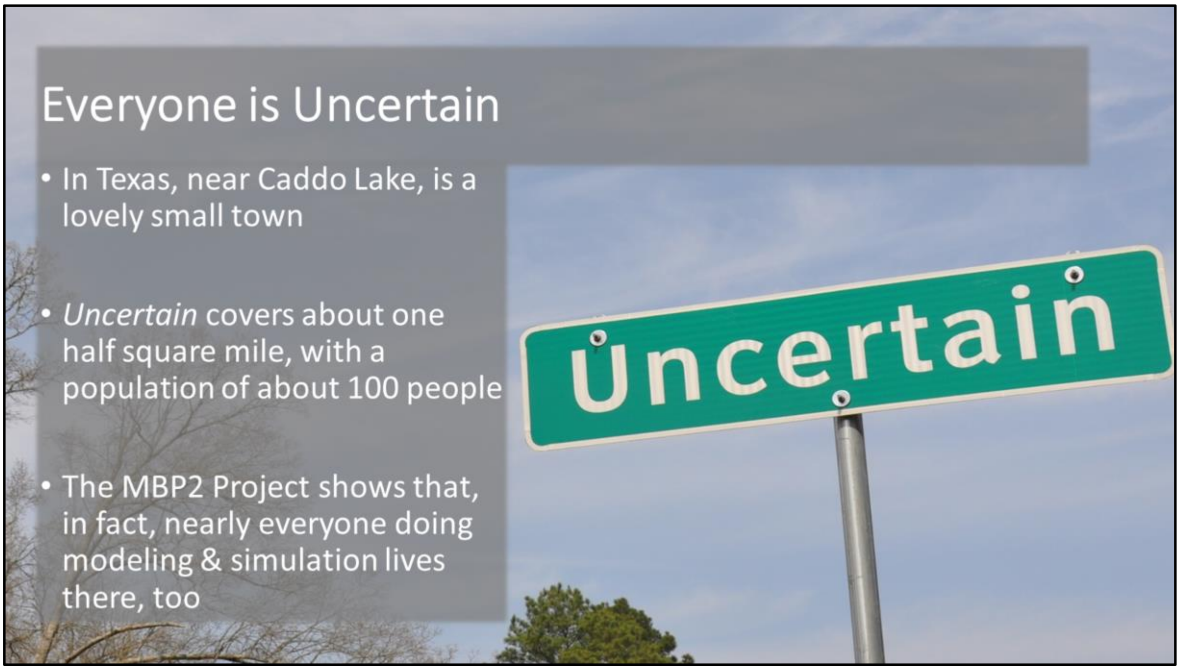
In February, the Pew Research Center for Internet, Science and Tech released a lengthy study report, *Code-Dependent: Pros and Cons of the Algorithm Age,* found here; http://pewrsr.ch/2kslvuK. Many of the Pew concerns are echoed by a group of authors who published in Scientific American https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/

Meanwhile, The EU is about to enforce the General Data Protection Regulation (GDPR), “the most important change in data privacy regulation in 20 years.” It imposes penalties for breaches of data privacy. Organizations running afoul of GDPR can be fined 4% of annual global turnover or €20 Million (whichever is greater). A summary of the regulations can be found here; http://www.eugdpr.org/article-summaries.html

These are not just regulations on privacy protection; they extend to transparent use. Under the new regulations, personal data of all types “must be processed lawfully, fairly, and in a manner transparent to the data subject.” The regulations control how data is collected and greatly restricts data use. Holding data for analysis alone will be a potential source of legal risk under the new rules. *Black box algorithms could create nearly unlimited liability.*

Big Data advocates who contend they should be able to discover new relationships and patterns in consumer data will have to be careful how they explore the unknown while regulators demand use of data for “specified, explicit purposes and only those purposes.“

Other nations, including the UK and Australia, are debating, or are in the process of adopting similar laws and regulations



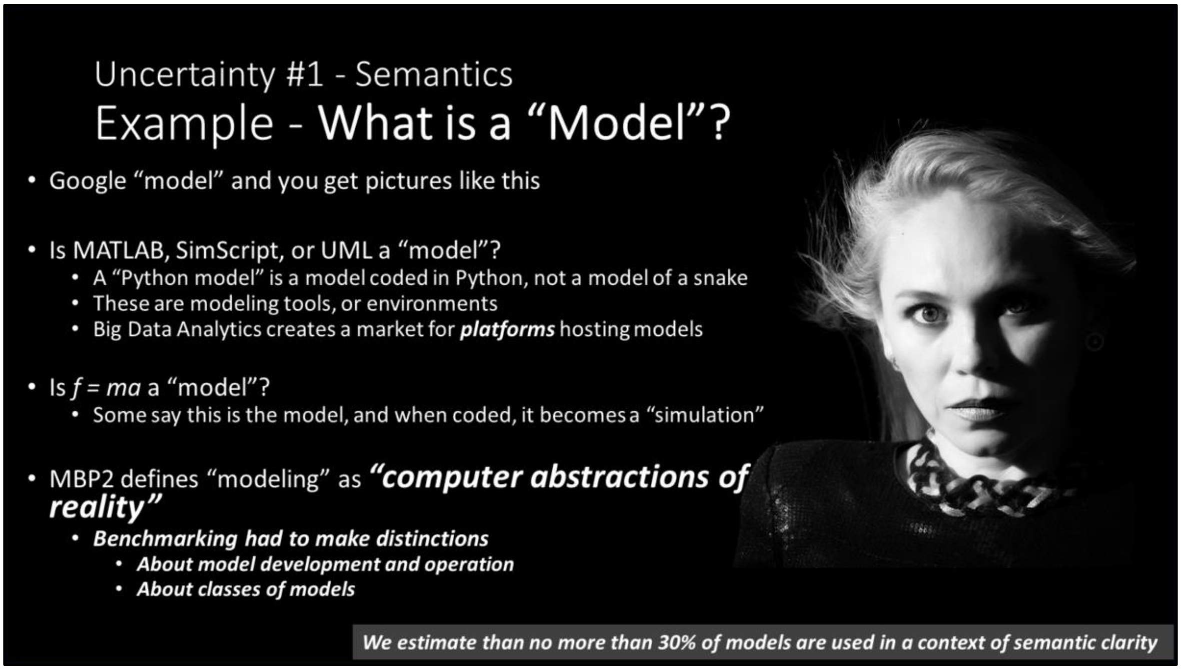
Everyone is Uncertain

- In Texas, near Caddo Lake, is a lovely small town
- *Uncertain* covers about one half square mile, with a population of about 100 people
- The MBP2 Project shows that, in fact, nearly everyone doing modeling & simulation lives there, too

So, clearly there is a great deal we could cover in our early results. A place to start is where this presentation focuses; one group of results – ***uncertainty is pervasive.***

Of course, this is not an exhaustive list of findings, and perhaps not even the most important findings.

The MBP2 project shows us the population of Uncertain is a lot bigger than the Census Bureau reports. Only a tiny fraction of the MS&A community is dealing effectively with five uncertainties we will talk about today.



Uncertainty #1 - Semantics

Example - What is a "Model"?

- Google "model" and you get pictures like this
- Is MATLAB, SimScript, or UML a "model"?
  - A "Python model" is a model coded in Python, not a model of a snake
  - These are modeling tools, or environments
  - Big Data Analytics creates a market for ***platforms*** hosting models
- Is *f = ma* a "model"?
  - Some say this is the model, and when coded, it becomes a "simulation"
- MBP2 defines "modeling" as ***"computer abstractions of reality"***
  - ***Benchmarking had to make distinctions***
    - ***About model development and operation***
    - ***About classes of models***

We estimate than no more than 30% of models are used in a context of semantic clarity

If we had to define "model" in a survey of those doing modeling for a living, that should suggest how difficult some semantic issues proved to be. What is a "model" is just one example of challenging semantic issues. This was due, in part to the international and multi-domain span of MBP2. It makes sense that different communities have developed different syntax.

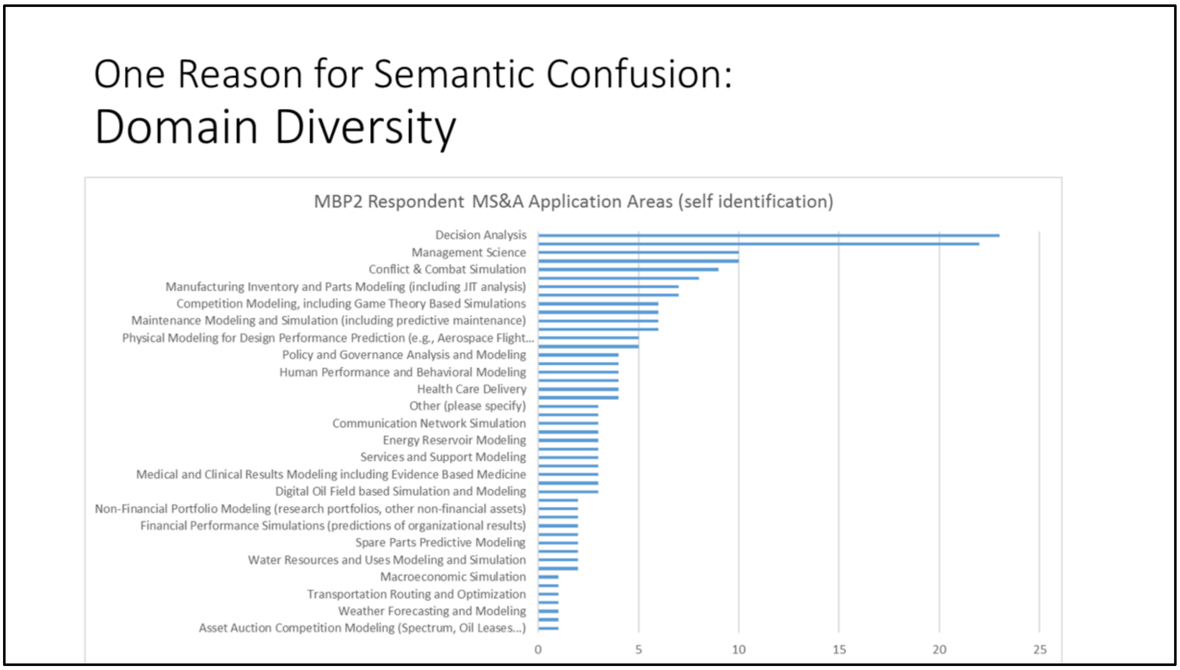
MBP2 defines "modeling" as *"computer abstractions of reality."* Usually this means math representations, which may be called "simulation" or "computer forecasting." In some cases, it is called "big data" if it involves computer enabled analysis. "Big data" algorithms, neural nets, deep learning models are all "models" for the purposes of MBP2. Models may be based on "wisdom of crowds" and involve some form of computer-enable or web-enabled group elicitation, or forecasting competition.

But, The MBP2 definition excludes individual humans guessing, no matter how gifted they are, because this is not amenable to benchmarking and not generally computer enabled.

Some of the semantic confusion is caused by informal communication by those who deliver MS&A. A "Python Model" might actually be about the behavior of snakes, but more likely, it was coded in the Python Programming Language.

While we did not directly measure semantic clarity, but some survey questions, and nearly all our interviews showed significant semantic risk. For example, do decision makers know which visualizations are actually measured data and which are a simulation? Is there clear understanding about whether machine learning is ongoing, or frozen?

We estimate than no more than 30% of models are used in a context of semantic clarity.

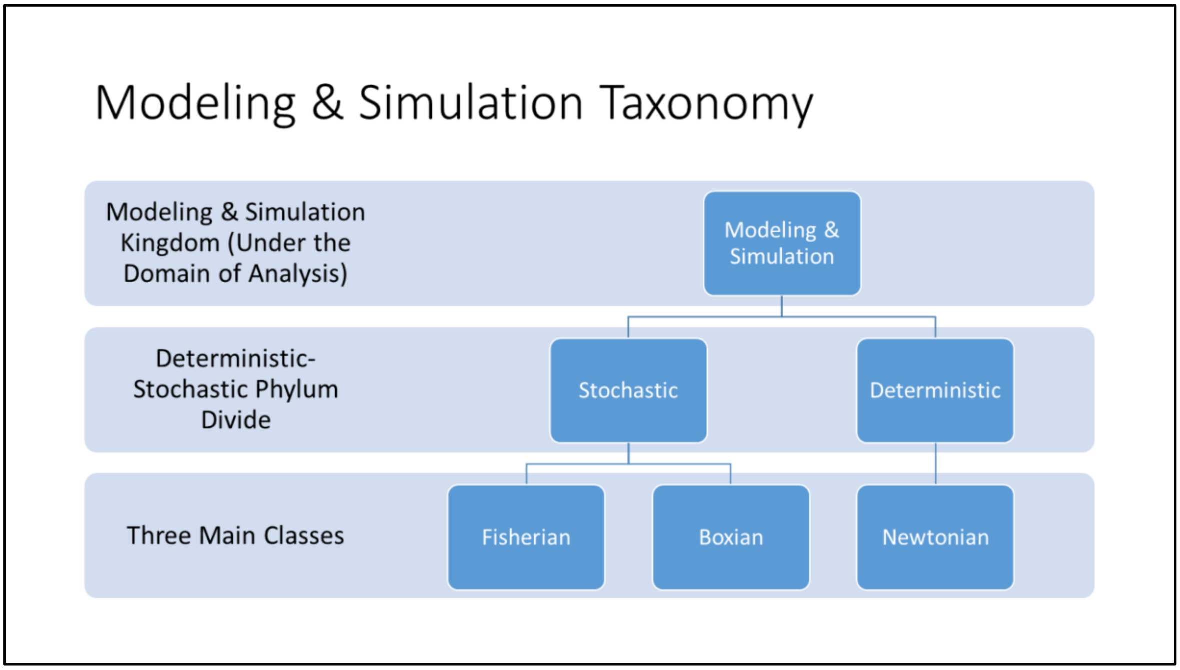


One Reason for Semantic Confusion: Domain Diversity

MBP2 respondents were diverse, representing more than 40 different disciplines and applications of MS&A, and respondents came from several nations.

It’s not surprising that a practitioner who does petroleum reservoir simulations uses different semantics than an aerospace structures engineer.

A majority of respondents said they did “Decision Analysis.” This is another example of semantic confusion. Some of these respondents are INFORMS members, and practice Decision Analysis as defined by Ron A, Howard of Stanford. But many simply seem to be reflecting the idea that MS&A is usually conducted to help someone make a decision.



Modeling & Simulation Taxonomy

Modeling & Simulation Kingdom (Under the Domain of Analysis)

Modeling & Simulation

Deterministic-Stochastic Phylum Divide

Stochastic

Deterministic

Three Main Classes

Fisherian

Boxian

Newtonian

To deal with semantic confusion, we used this taxonomy. For our interviews, we classify the subject MS&A into one of four types, Fisherian, Boxian, Newtonian, or hybrid.

By “Newtonian” we mean any purely deterministic model based on rigid laws, generally accepted as correct. While this includes physics, it is more than just Newton per se.

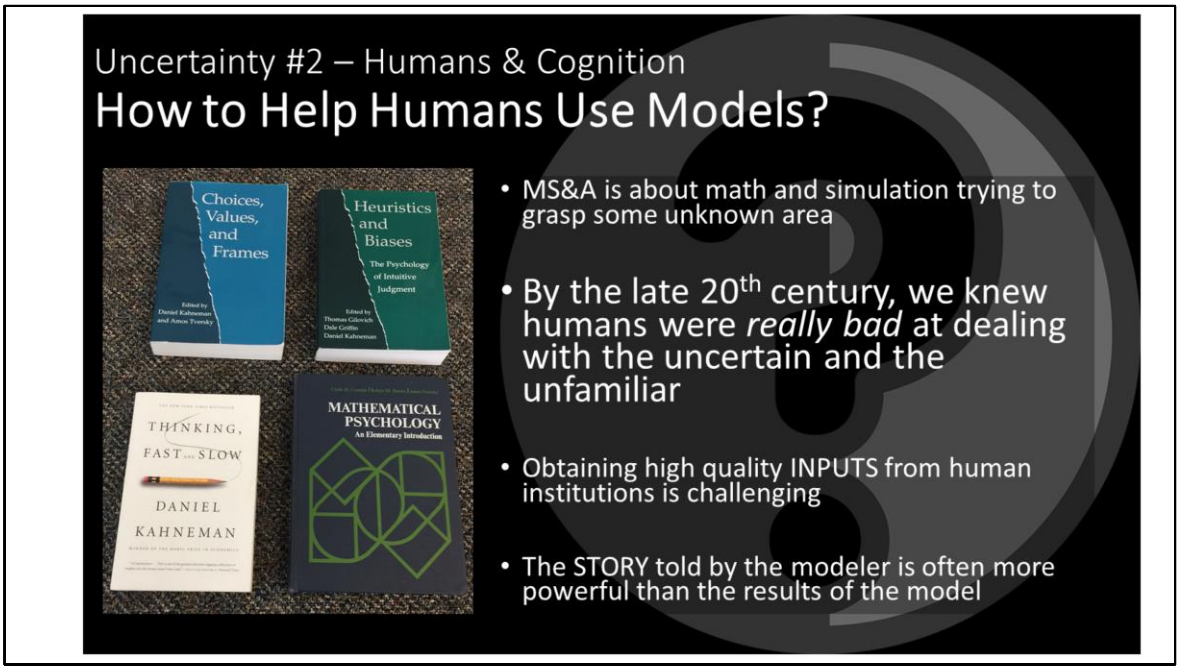
By “ Fisherian” we mean MS&A consistent with the general philosophy of Sir Ronald Aylmer Fisher FRS(R.A. Fisher). This class is characterized by

Emphasis on empirical data
Finding correlations, neighbor groupings (often termed “Analytics”)
Making inferences and conclusions from data, including “Big Data”
Examples include neural network based machine learning, Bayesian Spam filters, several statistical modeling suites

By “Boxian” we mean MS&A consistent with the general philosophy of George Edward Pelham Box FRS (G.E.P. Box). This class is characterized by

Cause and effect relationships (what Box called “mechanistic” models)
Blending empirical data and estimates when measured data is not possible
Generating insight, even when precision is impractical
Examples include Monte Carlo methods (implicit use of Bayes, vs. explicit conditional probabilities)

We don’t claim this taxonomy is perfect, but it was useful for our benchmarking purposes, and proved to help interview respondents describe their work.


Choices, Values, and Frames
Edited by Daniel Kahneman and Amos Tversky
Heuristics and Biases
The Psychology of Intuitive Judgment
Edited by Thomas Gilovich, Dale Griffin, Daniel Kahneman
Thinking, Fast and Slow
Daniel Kahneman
Mathematical Psychology
An Elementary Introduction

Uncertainty #2 – Humans & Cognition

How to Help Humans Use Models?

- MS&A is about math and simulation trying to grasp some unknown area
- By the late 20th century, we knew humans were *really bad* at dealing with the uncertain and the unfamiliar
- Obtaining high quality INPUTS from human institutions is challenging
- The STORY told by the modeler is often more powerful than the results of the model

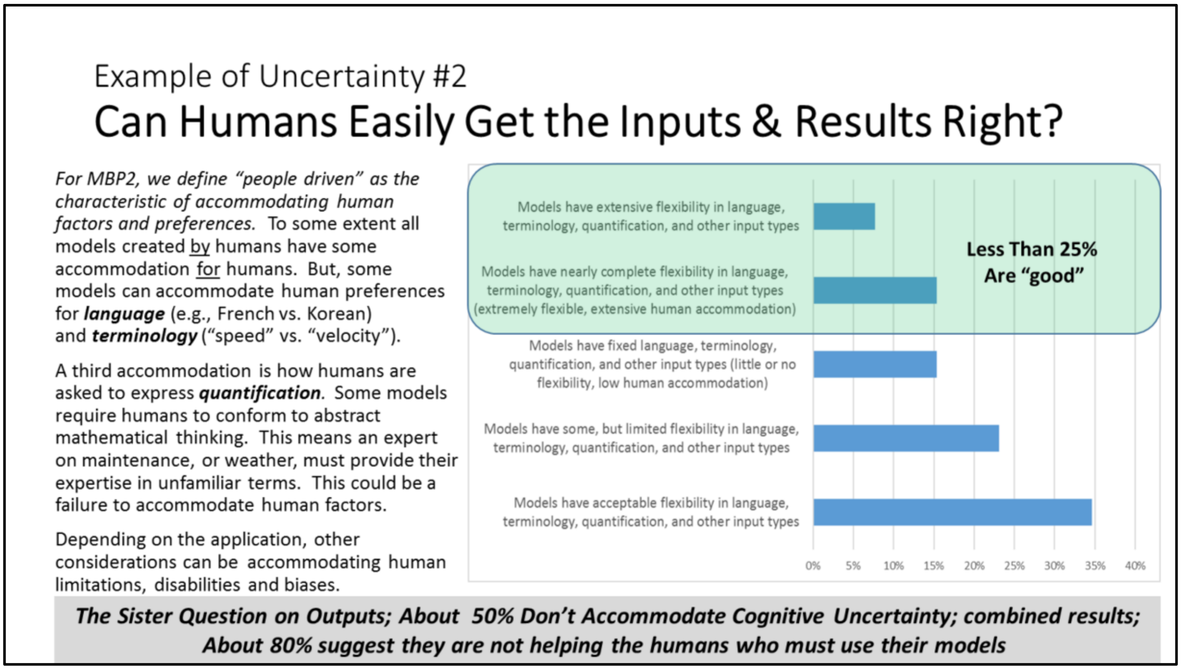
Beginning with Miller's "Magical Number Seven" paper, in 1956, we have seen a 60 year assault on assumptions about human cognitive power.   Brutal reality (and repeatable experiments) show our very limited capacity for information processing.

In particular we know humans are limited information processing channels in at least three ways.

- Humans have limited perception of dynamic range; we don't comfortably understand large variations. For most people a million dollars feels "more different" than a dollar, than a trillion dollars to a million, but the ratios are identical.
- A second problem is our lack of resolving power. We quantify most perceptions to something like 2 or three bits of information.
- A third problem, related to both of these is our poor intuitive grasp of uncertainty; we don't do odds very well.

There are of course other problems we could cite, but nearly every MS&A problem faces these three human limitations. So, for the MS&A community human limitations have important implications.

Broadly in MBP2 we called this "accommodating humans" and we explored accommodating human inputs and providing humans with MS&A outputs.



Example of Uncertainty #2

Can Humans Easily Get the Inputs & Results Right?

For MBP2, we define "people driven" as the characteristic of accommodating human factors and preferences. To some extent all models created by humans have some accommodation for humans. But, some models can accommodate human preferences for ***language*** (e.g., French vs. Korean) and ***terminology*** ("speed" vs. "velocity").

A third accommodation is how humans are asked to express ***quantification***. Some models require humans to conform to abstract mathematical thinking. This means an expert on maintenance, or weather, must provide their expertise in unfamiliar terms. This could be a failure to accommodate human factors.

Depending on the application, other considerations can be accommodating human limitations, disabilities and biases.

Models have extensive flexibility in language, terminology, quantification, and other input types
Models have nearly complete flexibility in language, terminology, quantification, and other input types (extremely flexible, extensive human accommodation)
Models have fixed language, terminology, quantification, and other input types (little or no flexibility, low human accommodation)
Models have some, but limited flexibility in language, terminology, quantification, and other input types
Models have acceptable flexibility in language, terminology, quantification, and other input types
Less Than 25% Are "good"
0%
5%
10%
15%
20%
25%
30%
35%
40%

The Sister Question on Outputs; About 50% Don't Accommodate Cognitive Uncertainty; combined results; About 80% suggest they are not helping the humans who must use their models

We asked a pair of similar questions on inputs and outputs. The input question is shown here.

On the left is an example of the kind of definitions needed before asking questions for MBP2. Feedback from respondents was positive – they felt these long, specific definitions and questions were worthwhile.

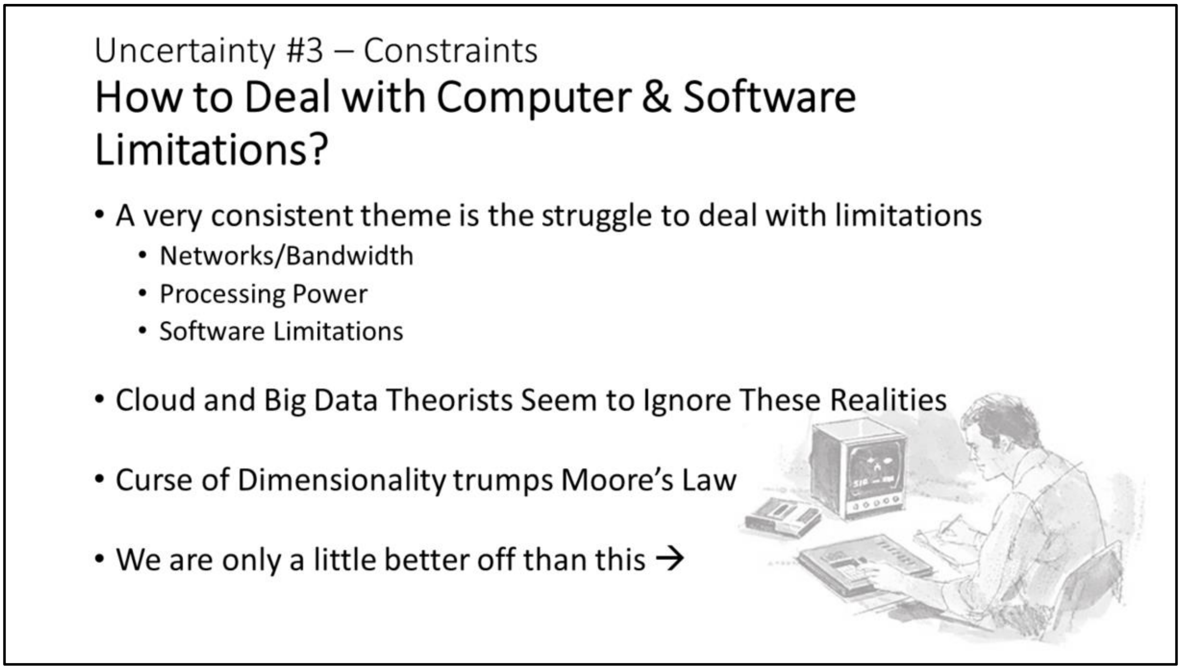
Sadly, the results show that only about 20% of respondents indicated they were able to accommodate the humans who had to use their models.

Outputs were a little better approaching about 50%. But if you believe that bad inputs threaten the quality of outputs, this is a problem.

Some practitioners suggest that "smart" users can adapt themselves to MS&A limitations. But Stanovich and West (1998 -2000) demonstrated weak correlation between raw IQ and things like statistical reasoning, correlation detection, and hypothesis testing. Lone Star research suggests that formal training in statistics can actually make people worse at the intuitive understanding of these matters.

We can find no evidence to suggest that "smart users don't need this kind of help."

So, even though we know humans are bad at uncertainty, about 80% of modelers suggest they aren't helping the humans who must use their models.



Uncertainty #3 – Constraints

How to Deal with Computer & Software Limitations?

- A very consistent theme is the struggle to deal with limitations
  - Networks/Bandwidth
  - Processing Power
  - Software Limitations
- Cloud and Big Data Theorists Seem to Ignore These Realities
- Curse of Dimensionality trumps Moore's Law
- We are only a little better off than this →

Most interview subjects mentioned the limitations of their host computer hardware, their network bandwidth, or their host software.

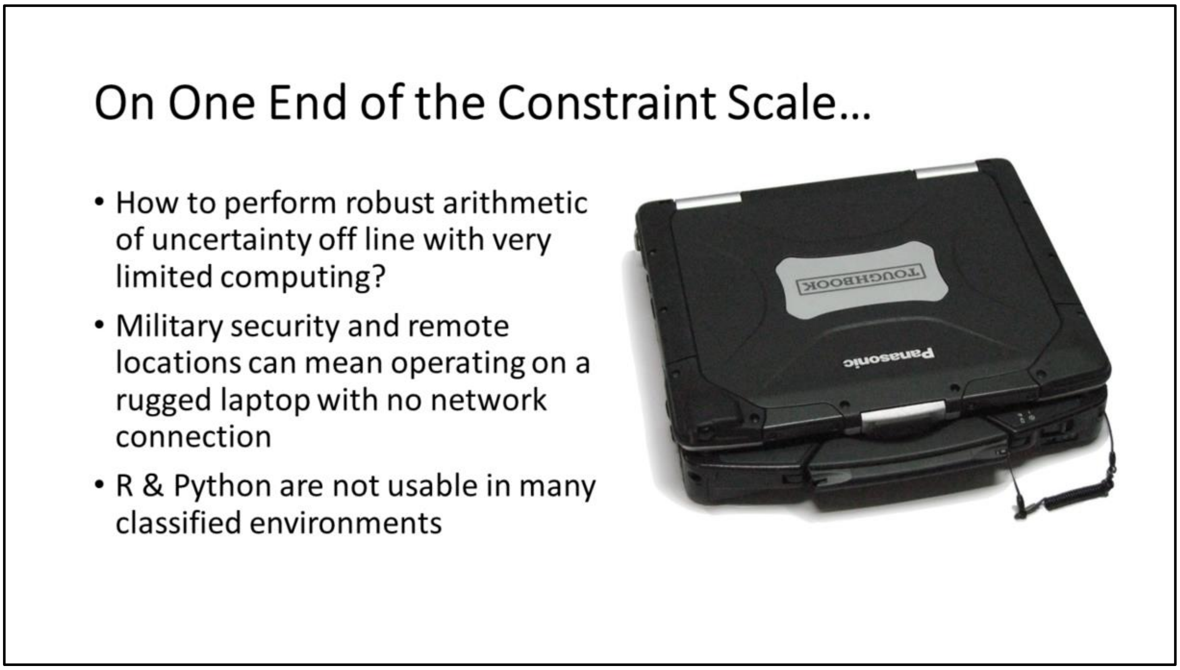
A great deal of data science deals with static data sets. Processing may take weeks or even years. This has led some Big Data theorists to underplay how difficult it can be to create MS&A results soon enough to matter. Some Cloud computing advocates, in particular seem vulnerable to this trap.

The fellow on the right is using a 1977 RCA desktop computer. Depending on the measure of power, the computer in which this presentation was created is somewhere between 8,000 to 8,000,000 times more powerful. That is a huge improvement in the 40 years since 1977. But it is not an infinite improvement.

Some curse of dimensionality problems require more processing steps than the number of hydrogen atoms in the solar system. Compared to that scale, the old RCA, a pretty good laptop and the world's best supercomputer are all roughly the same – not big enough.

We did not survey this factor, but it came up in nearly all our benchmark interviews, and in many of the pre-interviews. What we observed was that hardware and software limitations were a serious concern for every organization we deemed to be potentially best in class. The only people who didn't seem to worry about these matters were practitioners who seemed to be short (or far short) of being best in class contenders.

It seems quite striking to see how little these matters are discussed in big data papers. Google authors are an exception. Perhaps practitioners who do a great deal of big data work know how hard this is, while those who just study and write about it don't realize "There be monsters here?"



On One End of the Constraint Scale…

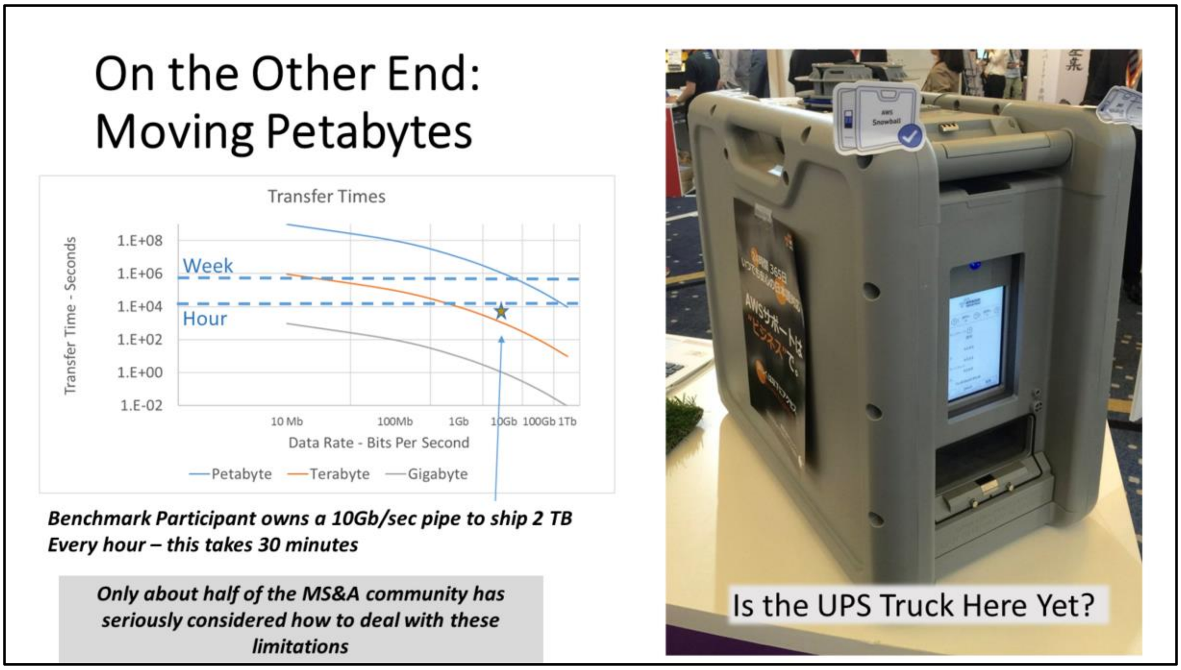
- How to perform robust arithmetic of uncertainty off line with very limited computing?
- Military security and remote locations can mean operating on a rugged laptop with no network connection
- R & Python are not usable in many classified environments

TOUGHBOOK
Panasonic

One kind of constraint is dealing with known limitations.

- Being off line
- Lacking the time to obtain good measured data
- Using fairly small processing platforms
- Being restricted in terms of the software titles available

SIPMath was cited in one interview as an example of a powerful tool for MS&A with these limitations



On the Other End: Moving Petabytes

Benchmark Participant owns a 10Gb/sec pipe to ship 2 TB Every hour – this takes 30 minutes

Only about half of the MS&A community has seriously considered how to deal with these limitations

Is the UPS Truck Here Yet?

If you had a 1 gigabit connection, you would be very fast. The U.S. Federal Communications Commission defines “broadband” as a MINIMUM of 25 megabits per second. Globally there are less than 10% of wireless connections with this much AVERAGE bandwidth. In many places gigabit speed is just a fantasy.

But to really do terabyte and petabyte big data, *really big data pipes* are needed.

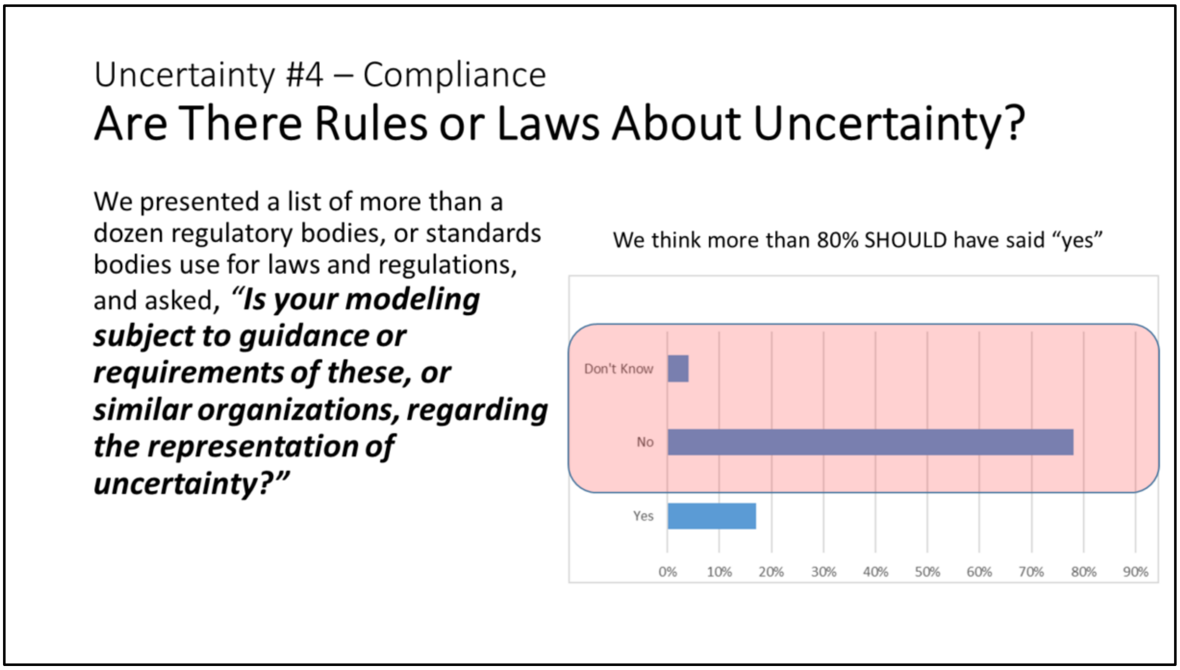
Walmart is a good example. They have a large number of medium to large connections. These correspond to each store or distribution facility. Together, they constitute about 2.5 Terabytes per hour. They can do this because they don’t EVER have all that data in one place. And, they quickly discard it. Walmart can’t keep all their data; their cloud can store less than one day of data at this rate.

Evernote is another example. To transfer “just” 3 petabytes to Google Cloud Storage took about 70 days.

One benchmark participant generates 2 Terabytes of data per hour. It takes 30 minutes to transmit over a 10 Gb channel. They do this 24 hours a day. For modelers who work on a massive scale, none of the cloud based approaches is likely to work. This is why there is still a super computer market even though you can rent as many processors as you need in the cloud.

The picture on the right is the Amazon Web Service Snowball. Data is loaded into it, and physically shipped.

We did not explicitly measure the topic of hardware and software limitations. However, based on the interviews and pre-interviews, ***we estimate that only about half of the MS&A community has seriously considered how to deal with these limitations, even if their own limits are modest compared to these large problems.***



Uncertainty #4 – Compliance

Are There Rules or Laws About Uncertainty?

We presented a list of more than a dozen regulatory bodies, or standards bodies use for laws and regulations, and asked, ***"Is your modeling subject to guidance or requirements of these, or similar organizations, regarding the representation of uncertainty?"***

We think more than 80% SHOULD have said "yes"

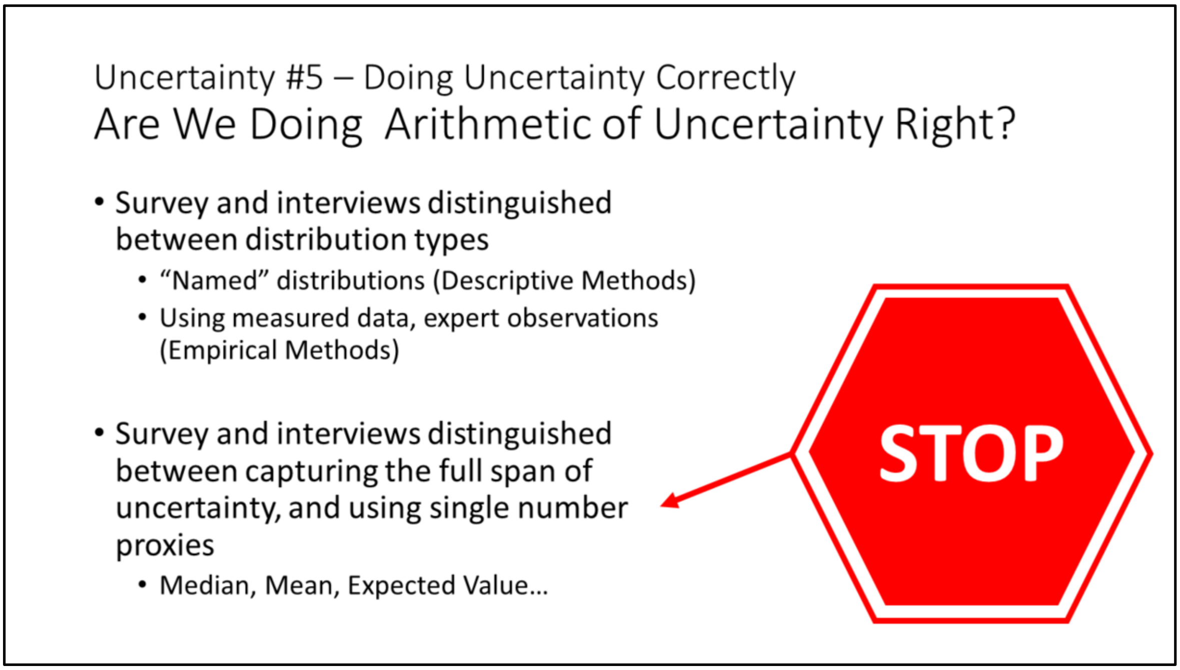
The survey shows, in fairly clear terms that a sizable number of modelers don't know they are breaking the law or ignoring some regulation. Note that this question is very narrow. *It does NOT deal with the emerging regulations on transparency and privacy like the ones cited from the ACM, the EU, and Pew research.*

This question is simply asking about doing the arithmetic of uncertainty in a way which complies with laws and regulations.

An example is the estimation of petroleum reserves. While some firms attempt to estimate reserves to the percentile of probability as recommended by SPE norms, our detailed questioning showed very few knew what the SPE methods were, or why they were required to provide investor relations and the finance department with those numbers.

Based on these results, we estimate that more than 80% of modelers don't know whether they need to meet specific standards in the MS&A.

They are uncertain about their level of compliance – or they should be uncertain if they knew a little more.



Uncertainty #5 – Doing Uncertainty Correctly

Are We Doing Arithmetic of Uncertainty Right?

- Survey and interviews distinguished between distribution types
  - “Named” distributions (Descriptive Methods)
  - Using measured data, expert observations (Empirical Methods)
- Survey and interviews distinguished between capturing the full span of uncertainty, and using single number proxies
  - Median, Mean, Expected Value…

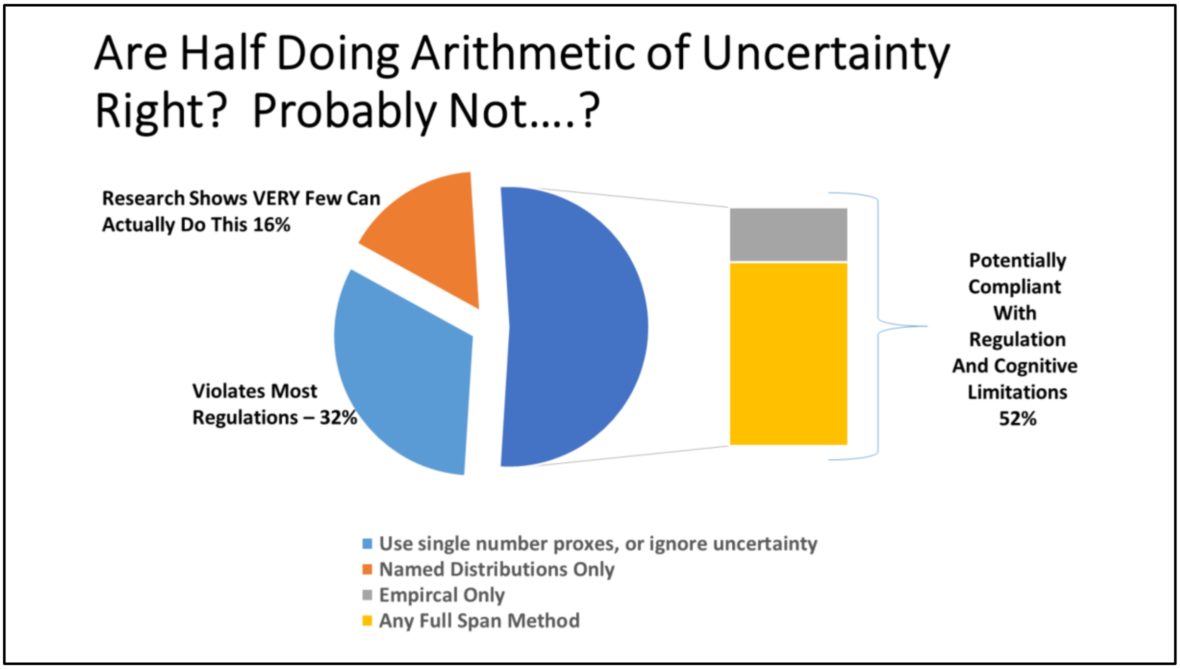
The survey and interviews explored uncertainty in several ways. One line of inquiry had to do with math methods.

While there is a great deal of diversity, we found most respondents reporting one or more of three methods.

- Using spans of uncertainty (distribution) by applying named distributions (Gaussian, Poisson…)
- Using spans of uncertainty by applying empirical methods (SIPs, parametrically specified…)
- Using single number proxies (median, expected value…)

Of course use of single number proxies is a bad thing. At best it is mathematical malpractice if it claims to represent uncertainty. At worst it violates the law.

We wanted to see how widespread the different methods were.



Are Half Doing Arithmetic of Uncertainty Right? Probably Not….?

Looking at how people deal with uncertainty math can be compared with regulatory issues. Doing the arithmetic of uncertainty "right" has both mathematical and legal ramifications.

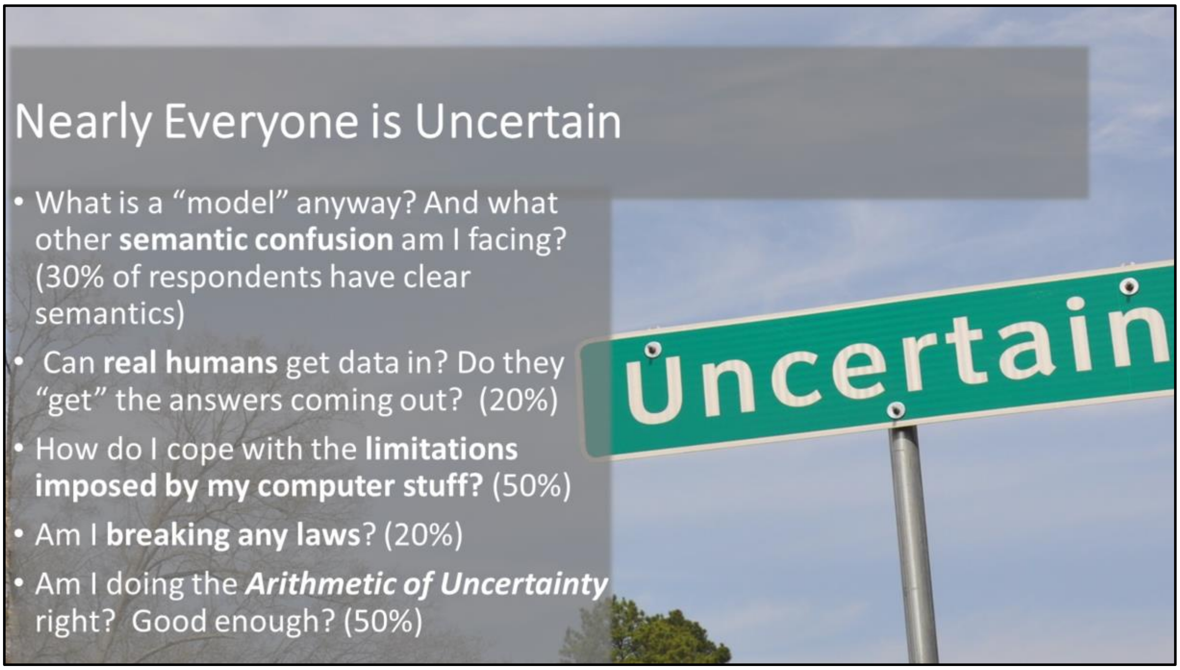
Using single number proxies for a span of uncertainty is a violation of most of the standards and regulations which apply to the arithmetic of uncertainty. The 32% of respondents who report doing this are at risk of such infractions.

About 16% report relying on named distributions only. Prior research shows how few practitioners are able to successfully employ these methods.

That leaves about 52% of respondents using full spread uncertainty modeling which has a chance of being complaint with regulations, and, compatible with human cognition (limitations). Are all of those respondents doing the arithmetic of uncertainty correctly? Probably not. The 52% are only *potentially* compliant.

If we guess that half of those using "named distributions" and about one fourth of the others using full span methods ***then we'd have to estimate less than half our practitioners are doing the Arithmetic of Uncertainty correctly***, and, remember – these are people who are interested in best practices. Presumably they are better than the typical MS&A practitioners.

To summarize – there are at least five common uncertainties we found in our benchmarking.



Nearly Everyone is Uncertain

- What is a "model" anyway? And what other **semantic confusion** am I facing? (30% of respondents have clear semantics)
- Can **real humans** get data in? Do they "get" the answers coming out? (20%)
- How do I cope with the **limitations imposed by my computer stuff?** (50%)
- Am I **breaking any laws**? (20%)
- Am I doing the ***Arithmetic of Uncertainty*** right? Good enough? (50%)

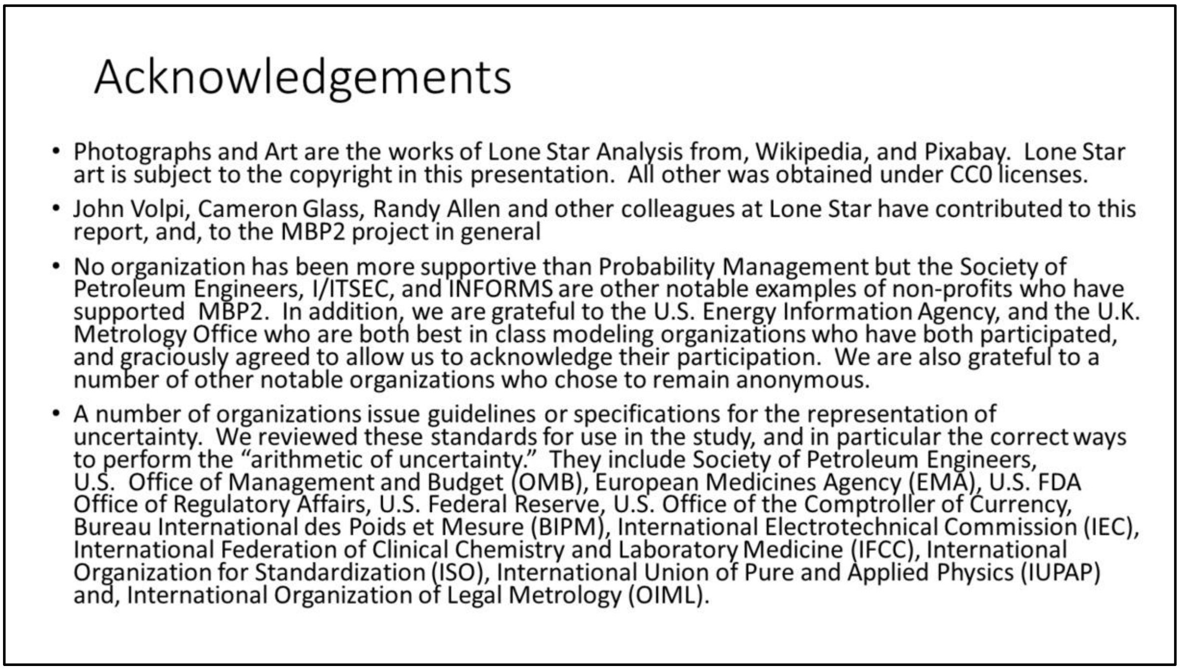
This presentation focuses on one group of results – ***uncertainty is pervasive among MS&A practitioners (or should be).***

The percentages shown are an estimate of our respondents who are effectively dealing with each one of the uncertainties. It's easy to see that very few practitioners are performing best in class practices in all of our benchmarking areas. The percentages shown by each topic are our estimate of the MS&A community we surveyed and interviewed who are dealing effectively with these questions.

On any given topic, no more than half the MS&A has been able to address these issues.

So.. Only a tiny fraction of the MS&A community is dealing effectively with all five uncertainties we talked about today.

And… Nearly everyone is facing these uncertainties, whether they know it our not.



Acknowledgements

- Photographs and Art are the works of Lone Star Analysis from, Wikipedia, and Pixabay. Lone Star art is subject to the copyright in this presentation. All other was obtained under CC0 licenses.
- John Volpi, Cameron Glass, Randy Allen and other colleagues at Lone Star have contributed to this report, and, to the MBP2 project in general
- No organization has been more supportive than Probability Management but the Society of Petroleum Engineers, I/ITSEC, and INFORMS are other notable examples of non-profits who have supported MBP2. In addition, we are grateful to the U.S. Energy Information Agency, and the U.K. Metrology Office who are both best in class modeling organizations who have both participated, and graciously agreed to allow us to acknowledge their participation. We are also grateful to a number of other notable organizations who chose to remain anonymous.
- A number of organizations issue guidelines or specifications for the representation of uncertainty. We reviewed these standards for use in the study, and in particular the correct ways to perform the "arithmetic of uncertainty." They include Society of Petroleum Engineers, U.S. Office of Management and Budget (OMB), European Medicines Agency (EMA), U.S. FDA Office of Regulatory Affairs, U.S. Federal Reserve, U.S. Office of the Comptroller of Currency, Bureau International des Poids et Mesure (BIPM), International Electrotechnical Commission (IEC), International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), International Organization for Standardization (ISO), International Union of Pure and Applied Physics (IUPAP) and, International Organization of Legal Metrology (OIML).