Statistical Engineering: Past, Present and Future

1st International Statistical Engineering Association Summit
West Palm Beach, FL
October 1, 2018

Roger W. Hoerl
Union College
Schenectady, NY

Outline

- The past: some history
- The role of large, complex, unstructured problems
- The present: the current state of statistical engineering
  - The underlying theory
- The future: its ours to claim
- Summary

Some History

- Note that I will focus the history discussion on “statistical engineering” as defined by ISEA (pronounced “I-see-ah”)
- I am not going to discuss the first published use of the this term by Eisenhart in 1950
- Neither will I discuss Shainin’s use of this term, or Steiner and MacKay’s (2005)
- Obviously, these publications predate any use of the term by Ron Snee and me, or by ISEA

This term has been used previously

Some History – RWH & RDS

- RWH & RDS hold private discussions on the limitations of a myopic “tools” orientation to statistics (circa 2006)
  - 2008 Technometrics panel discussion highlights issues within the profession
- Agree on the need for a discipline to formally study how to integrate multiple tools
- Believe they have a “eureka moment”: *statistical engineering*
- Go “public”, beginning in 2009
  - Discuss the need for statistical engineering in 2009 ASMBI article
  - Give talks on statistical engineering at FTC (2009), WQC & JSM (2010)
  - Publish a formal definition in Quality Engineering (2010)
  - Publish in TAS (2017) – first publication in mainstream statistics journal
- Realize that a wider group needs to be involved: approach CAC, GGV ,etc.

Humble origins!

Some History – Key Contributions from Others

- Xiao-Li Meng notes the need to teach statistics students to solve problems that "do not correspond to a recognizable textbook chapter." (TAS 2009)
  - Adds course at Harvard, Stat 399, which "emphasizes deep, broad, and creative statistical thinking"
- MIT President and GE Board member Susan Hockfield gives seminar at GE Global Research (2010)
  - Discusses the relationship between science and engineering
  - RWH is in attendance, taking notes!
- CAC and LL edit a special edition of QE on statistical engineering (2012)
  - Includes numerous case studies
- ASQ Stat Division develops a statistical engineering webpage (2013)

Beginnings of a broad consensus

Some History – ISEA

- PAP and LJF organize NASA & DoD partnership; begin holding conferences oriented towards statistical engineering (2011)
- GGV sends email on 6/26/17 to RWH, RDS, CAC, and PAP, to suggest organizing "a small group to discuss statistical engineering"
- This group begins holding conference calls to discuss path forward
- Group meets in 2017 at FTC in Philadelphia, with SPD, WAJ, & DVN
  - Plans a meeting of interested parties in the DC area in December
- GGV chairs meeting of 14 in Alexandria, VA (12/9-10/2017)
  - Unprecedented gathering, with no affiliation to any professional society or conference
  - Self appointed "steering committee" decides to create a new society
- GGV incorporates ISEA in the Commonwealth of VA (7/9/18)

"Never doubt that a small group of thoughtful, committed citizens can change the world. In fact, it is the only thing that ever has." M. Mead

The Challenge of Large, Complex, Unstructured Problems

- There are many types of applications of statistical engineering; past, present, and future
- However, almost all statisticians agree that there is a gap in the current statistical science body of knowledge relative to **large**, **complex, unstructured problems**
  - These are the problems that get management attention!
- The vast majority of textbook problems have a single, correct answer
- But, many real problems students eventually face are too large, complex, and unstructured to have a "correct" solution
- This was Meng's fundamental point in creating Stat 399
- As explained by Hockfield, creatively integrating individual "parts" from science to solve a real problem is engineering

The statistical science discipline has no answer for these problems

Typical Attributes of Such Problems

- Impact is broad – process performance, financial, customer, social, etc.
- Several departments, groups and functions are involved
- Problem has high degree of complexity involving both technical and non-technical challenges
  - Problem not clearly defined/structured
  - There is no known solution
  - Potential team conflict on how to approach
- Multiple sources of data and information are needed

Non-technical complexity is rarely discussed in statistics texts

Typical Attributes of Such Problems

- More than one technique is required for solution
  - Typically both statistical and non-statistical techniques are required
- Creative use of information technology (IT) is needed
- Long-term successes requires embedding solution into work processes, typically through:
  - Use of custom software
  - Integration with other sciences and disciplines
- Examples of such problems will be presented and discussed at this summit

What literature exists to guide practitioners?

Key Points

- Statistical engineering is **not limited** to large, complex, unstructured problems
- However, statisticians readily admit that these problems cannot be solved with "traditional" statistics, applied or theoretical

Large, complex, unstructured problems are our "ace in the hole"

Statistical Engineering: the Basics

- Definition:
  - The discipline of statistical engineering is the study of the systematic integration of statistical concepts, methods, and tools, often with other relevant disciplines, to solve important problems sustainably.
- In other words, trying to build (engineer) something meaningful from the statistical science "parts list" of tools
  - Focus is on solving problems versus tools, per se
  - Real problems, particularly big problems, require integration of multiple methods
- See special edition of Quality Engineering (2012) on statistical engineering for more background and case studies

Statistical engineering is not a "method" per se

Key Aspects of Definition

- “the study of”
  - Research oriented
  - Statistical engineering has a theory
- “solve important problems sustainably”
  - Results are the “what”, methods and tools are “how’s”
  - Statistical engineering is therefore tool-agnostic
  - Solution must be sustainable over time
- “often with other relevant disciplines”
  - Integration of multiple tools, methods, and even disciplines
  - Information technology usually has a major role to play

Statistical engineering studies how to select and integrate methods in order to solve real problems

A Conjecture

Scientists, engineers, statisticians and other professionals have been building meaningful new things from the statistical science "parts list" of tools for some time, to address large, complex, unstructured problems. However:

- This has typically been done on an ad-hoc basis with little or no underlying theory documented to provide guidance to others
- Applications have generally been "one off's", requiring the "wheel to be reinvented" each time
- This has significantly slowed progress, and led to missed opportunities for impact

Statistical engineering is an old idea, but now is a new discipline

The Underlying Theory

- To be a true discipline, statistical engineering needs a dynamic theory
- But what is theory? "A coherent group of general propositions used to explain a phenomenon"
- Note that the definition of theory does not require mathematics!
- Lindsay et al. (2004): "The issues we raise above have nothing to do with the old distinction between applied statistics and theoretical statistics. The traditional viewpoint equates statistical theory with mathematics and thence with intellectual depth and rigor, but this misrepresents the notion of theory. We agree with the viewpoint that David Cox expressed at the 2002 NSF Work on the Future of Statistics that 'theory is primarily conceptual', rather than mathematical."

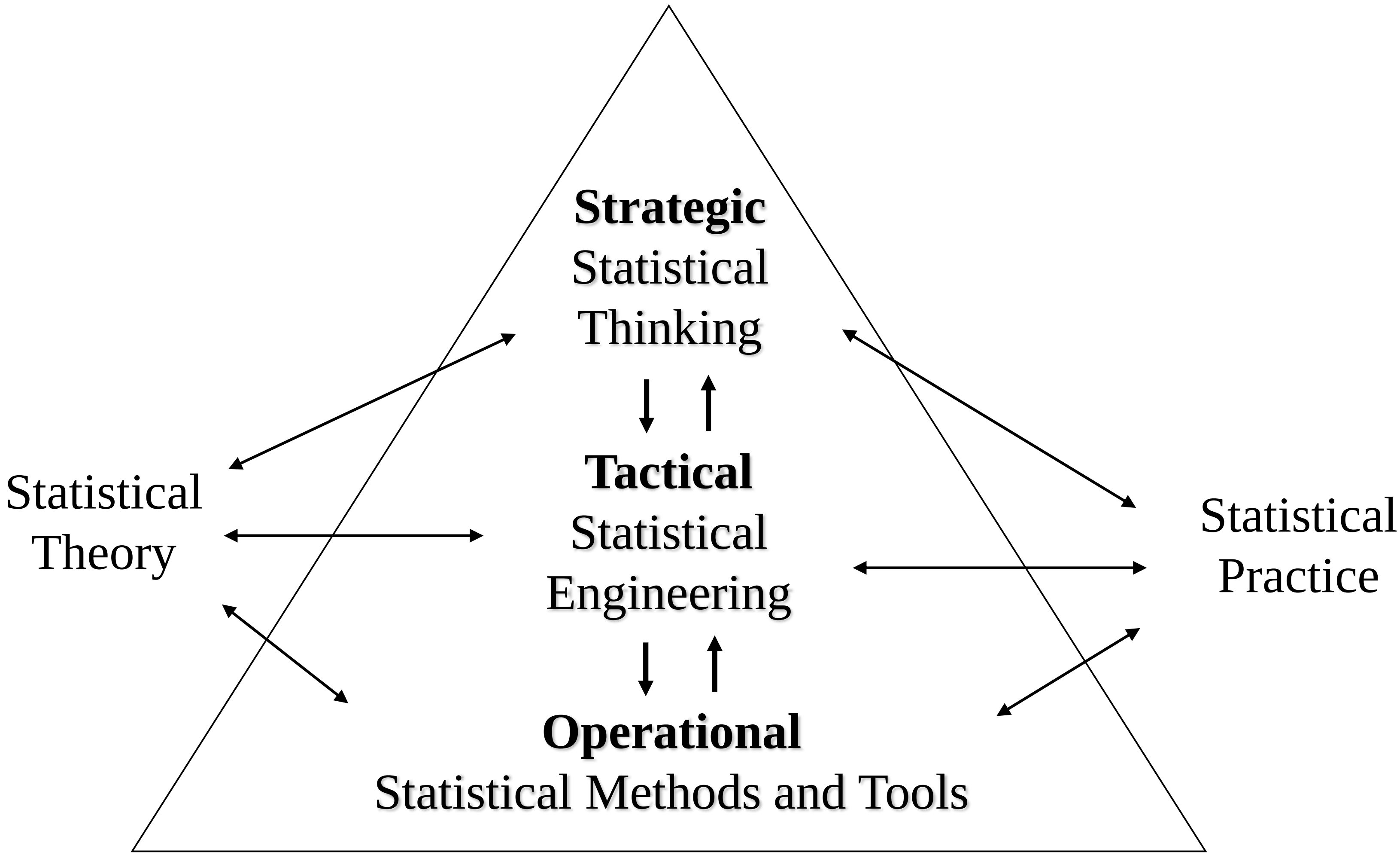
The theory of statistical engineering is not mathematical statistics

The Underlying Theory

- There are four main elements of the current theoretical foundation
  1. The fit with statistical science and statistical thinking
  2. The fundamental principles of statistical engineering
  3. The typical phases of statistical engineering projects
  4. The use of a standard set of “core processes”
- This is clearly a work in progress!
- However, even this rudimentary theory can help people understand and be able to apply statistical engineering

The first key elements of a dynamic theory are in place

1. The Statistics Discipline as a System

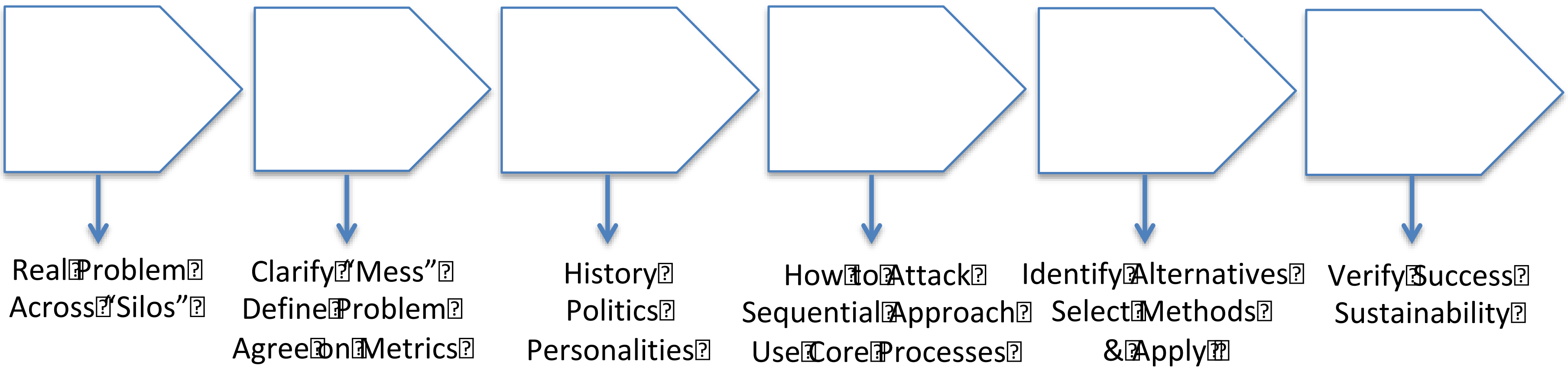


2. Typical Phases of Statistical Engineering Projects

1. **Identify Problems***:* find the high-impact issues critical to achieving the organization's strategic goals - typically crossing "silos"
2. **Provide Structure***:* translate a "mess" into a problem: carefully define the problem, objectives, constraints, metrics for success, and so on
3. **Understand the Context**: identify important stakeholders (e.g., customers, organizations, individuals, management), research the history of the issue, identify unstated complications, such as cultural or political issues, locate relevant data sources
4. **Develop a Strategy**: create an overall, high level approach to attacking the problem, based on phases 2 and 3
5. **Develop and Execute Tactics**: identify diverse methods and disciplines that collectively will accomplish the strategy
6. ***Identify and Deploy Final Solution***: determine and deploy solution, ensure that it actually works, and maintain solution over time

There are no "seven easy steps" to statistical engineering projects

2. Flow of Statistical Engineering Projects



Notes:

- **This is typically NOT a linear process; significant looping back to previous phases is common**
- **Each phase needs to be tailored, depending on the problem structure and context. That is, this is NOT "7 easy steps to solving problems"**
- **Often several projects are required to solve large, complex, unstructured problems**

3. Fundamental Principles of Statistical Engineering

- Understanding of the problem context
- Development of a problem-solving strategy
- Consideration of the “data pedigree”
- Integration of subject matter theory (domain knowledge)
- Use of sequential approaches

These are not always needed for straightforward problems, but they are mandatory for large, complex, unstructured problems.

Understanding of Problem Context

- How will the problem solution be utilized?
  - Our objective is typically an “effective” solution, not the “optimal” solution
  - Best technical or business solution versus best statistical solution
- How did we get where we are today (history)?
- What are the root causes of the problem (beyond symptoms)?
- What solutions have been attempted previously? Why didn’t they work?
- What politics are involved (often unstated)?

There are reasons why the problem remains unsolved

Strategy

- Linking and sequencing tools in novel, logical ways enhances effectiveness, learning, and impact
  - Complex problems can rarely be solved with one method
- Tools-oriented approaches typically produce poor results
  - “Hammer and nail” analogy
  - Debating which tool is “best” is a distraction
- Simple problems only require application of the “correct” tool; complex problems require a strategy
- The strategy for each problem is unique, just as the strategy for each sports opponent is unique

“Failing to plan is planning to fail.” (John Wooden)

Data Pedigree

- All data are ***not*** created equal – sounds obvious, but isn't!
- "Data quality" is subjective and depends on the nature of the problem to be solved
- There is a need for a better-defined and more objective concept: "data pedigree" (Hoerl and Snee 2010)
- The data pedigree is: "documentation of the origins and history of a data set, including its technical meaning, background on the process that produced it, the original collection of samples, measurement processes utilized, and the subsequent handling of the data, including any modifications or deletions made, through the present". (Hoerl and Snee 2018)

Data are "guilty until proven innocent"

Key Elements of Data Pedigree*

- **What the data represent**; that is, a basic explanation of the underlying subject matter knowledge of the phenomenon being measured, including units of measurement.
- Description of the **process that produced the data**, such as a financial process, healthcare process, manufacturing process, and so on.
- Description of **how the "samples" were obtained** from this process that were subsequently measured.
- The specific **measurement process** used to assign numbers or attributes to the "samples".
- The existence (or lack) of **recent analyses of the said measurement system**, such as gage R&R studies, calibration studies, and so on.
- **The history of the data**, documenting the chain of custody - who has had access to the original data, what if any changes or deletions have been made – and access to a "gold standard", i.e., access to a copy of the original data that can be verified.

* Hoerl and Snee (2018)

The problem context determines the degree of rigor required

Subject Matter Knowledge

- Everything we know about the phenomenon of interest, both from the fundamental science or engineering, as well as empirical studies
- Subject matter knowledge is required for actionable statistical analyses
- Such knowledge must guide data collection, and also interpret (make sense of) statistical results
- Statistics becomes ineffective when divorced from subject matter knowledge
  - The only reason for statistics to exist as a discipline is to enhance other disciplines (chemistry, engineering, psychology, economics, etc.)

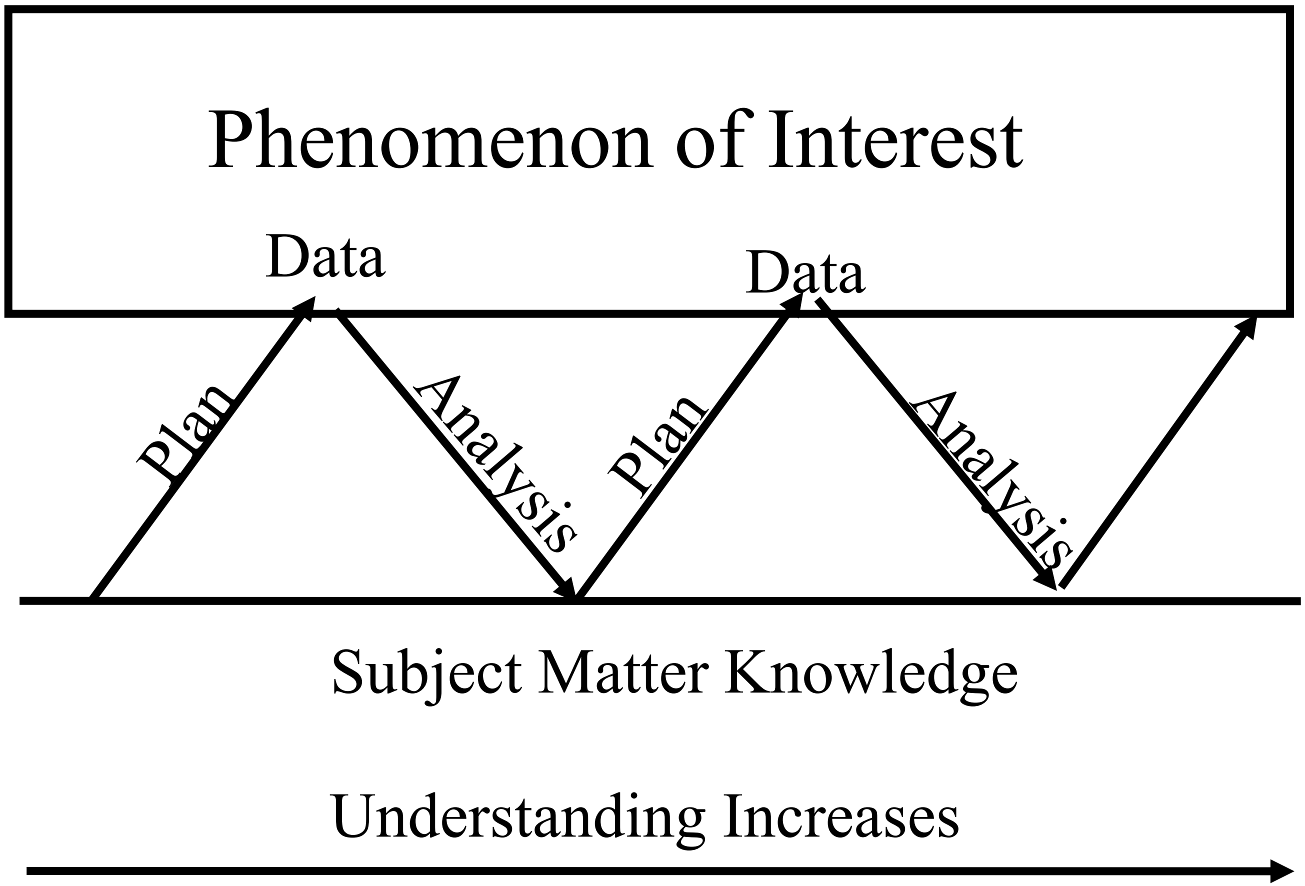
Almost all pioneers in statistics were trained in science or engineering

Sequential Approaches

- Statistical applications should be viewed as part of the ongoing application of the scientific method, not “one shot studies”
- Guiding future studies is often the most beneficial result from analysis of existing data
  - For example, the phases of clinical trials in developing pharmaceuticals
- A sequential approach allows for development of new theory and knowledge, not just testing of existing hypotheses
- A sequential approach fits well with the use of an overall strategy

The scientific method is not built on “one-shot studies”

The Sequential Nature of Statistical Engineering*



* Modified from Box et al. (1978)

4. Core Processes of Statistical Engineering

- The concept of core processes is borrowed from the "unit operations" chemical engineering
- Unit operations involve some specific chemical operation, such as heat transfer, distillation, and so on
- Chemical plants are made up of a sequence of unit operations
  - There are numerous ways to transfer heat – chemical engineers figure out the best one for a given process
- Statistical engineering strategies will involve a sequence of tools selected from the "unit ops" of statistics, i.e., the key statistical activities
  - Non-statistical tools will often be included as well
- These key activities, or "what's", are the core processes

Statistical engineering focuses on the "what" before the "how"

4. Core Processes of Statistical Engineering

- Data acquisition – obtaining high pedigree data
- Data exploration – generating hypotheses
- Model building – of various types
- Drawing inferences – developing broader conclusions
- Solution identification and deployment – solving problems sustainably, based on the analyses
- There are also a set of "overarching competencies", such as project management, teamwork, etc.

The strategy will involve some tailored sequence of tools selected from these processes

Thoughts on the Future

- Statistical engineering has gone from an abstract concept to a defined discipline with its own professional society in just a few years
  - The growth has been truly remarkable
  - See www.isea-change.org
- As with anything new, there has been, and will be, resistance
  - Many journal editors and referees are looking at statistical engineering from a "tools" lens
- In my view, the statistics profession, like the fashion industry, has a bad habit of embracing a sequence of fads
- Statistical engineering could easily become a fad; hot today, gone tomorrow
- Turnover on the ISEA board could result in ISEA gradually becoming just another section or division, with no unique identity

"Historical results do not guarantee future performance"

A Call to Action

- Take a good look around the room (really – do it!)
- If this group doesn't maintain and accelerate the momentum, statistical engineering will have a short half-life
- We can't count on any of the big professional societies to make this happen – its too cross-disciplinary
- Case in point; Germany pioneered the chemical industry, but missed the boat on chemical engineering
  - MIT, and the US in general, pioneered chemical engineering
- Bottom line: its up to us!
- How many other opportunities do you have to make a lasting mark on the profession, and on society?

"Never doubt that a small group of thoughtful, committed citizens can change the world. In fact, it is the only thing that ever has." M. Mead

Summary

- "We predict that in the coming decades, Statistical Engineering will revolutionize the practice of statistics in business, industry and government, and change how statistics is taught and perceived."
  - CAC & PAP: 2012 support letter for Ron Snee for ASA Dixon Award
- Christine and Peter will ultimately turn out to be either prophetic or naïve. Which will it be?

It really is up to the folks in the room!