Statistical Engineering: An Idea Whose Time Has Come?

ENBIS
Nancy, France
September 3rd, 2018

Roger W. Hoerl, Union College
Schenectady, NY
USA

Outline

- The challenge of large, complex, unstructured problems
- A financial case study
- The statistical engineering approach
- The fundamentals of statistical engineering
- Summary

The Challenge of Large, Complex, Unstructured Problems

- The vast majority of textbook problems have a single, correct answer
- However, many real problems students eventually face are too large, complex, and unstructured to have a "correct" solution.
- Others have noted this oversight previously
- For example, Meng (2009) introduced Stat 399, Problem Solving in Statistics, at Harvard
  - Stat 399, "…emphasizes deep, broad, and creative statistical thinking instead of technical problems that correspond to a recognizable textbook chapter."

Are students prepared for the "real world"?

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

How should practitioners attack such problems?

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

What literature exists to guide practitioners?

A Personal Case Study

- Problem: GE Capital announced losses of over $125 million on WorldCom bonds that went into default
- Their question to GE Global Research: "Is this just the cost of doing business in the financial sector, or could we have predicted these losses with enough lead time to lower our exposures?"

Classic problem in finance - unsolved

A Personal Case Study

- Challenges included
  - Financial theory ("efficient market hypothesis") says you can't predict defaults ahead of the market
  - GE Capital needed to trade in large quantities
  - No commonly accepted definition of "default"
  - Limited internal data – no set of "universal data"
  - No defined measure of success

Does this sound like a typical textbook problem?

A Personal Case Study

Approach Taken:

- Cross-functional team organized
  - Statistics, operations research, machine learning, quantitative finance, business expertise
  - Spread between upstate New York, Bangalore (India), and Stamford (Connecticut)
- Developed definition of default, and metrics to document success and failure (step zero)
  - Typical for unstructured problems

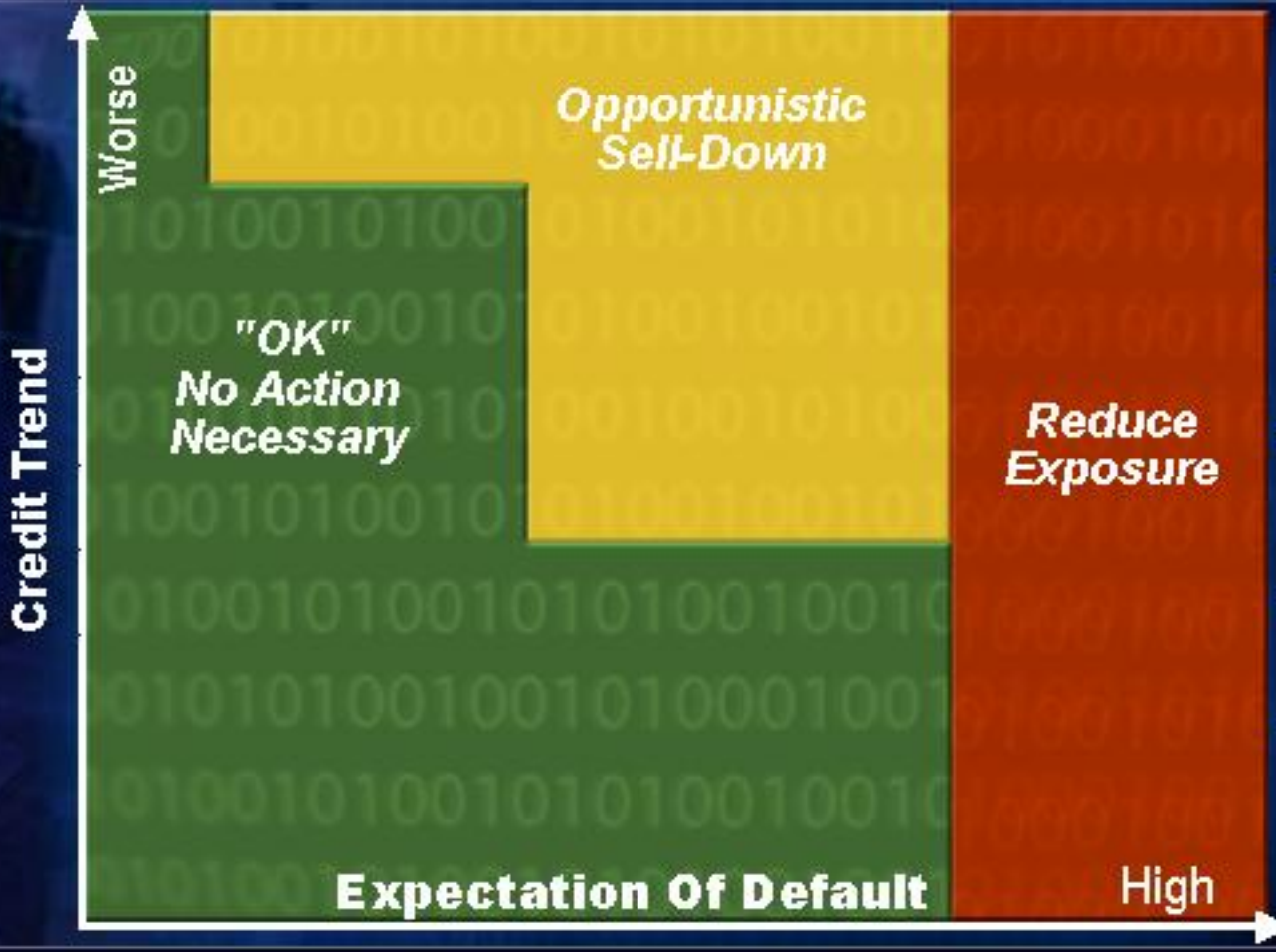
No “template” to follow

A Personal Case Study

- Data obtained externally – needed to merge different data sources
  - Eventually set up direct data feed from Wall Street
- Final prediction methodology utilized:
  - Publicly available default predictor as an input
    - Engineering vs. pure science approach
  - Smoothing algorithms, classification and regression trees (CART), simulation, and Markov Chains
- Developed "control plan" to detect need for retuning the algorithm: used censored data methods from reliability

Does this look like a typical textbook solution?

"CREDIT ALERT"
PROPRIETARY PREDICTIVE TOOL



New Tool!

- Evaluated 5 Yrs Credit Trends – 3,500+ Companies
- Statistically Derived Indicators Of Credit Deterioration
- Validated Against Our Own Portfolio History
- Applicable Across All Commercial Portfolios

Up To 12 Months "Early Read" – Very Promising!

A Personal Case Study

Results:

- GE Capital performed a simulation study of the final prediction system – without our involvement
  - Evaluated their potential financial results, had they used this system in the past year for all trading
  - Results were positive in the hundreds of millions of dollars
- This system was subsequently incorporated into underwriting procedures for large financial deals
  - "Embedding" statistical methods into business processes
- The team received a patent for the *system* – not for the *algorithm* (US20030229556A1)

Solving large, complex, unstructured problems produces impact

The Statistical Engineering Approach

- How did the team attack this problem? Answer: using what we now call statistical engineering
- 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 "hows"
  - 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

SE 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 offs", 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 perhaps a new discipline

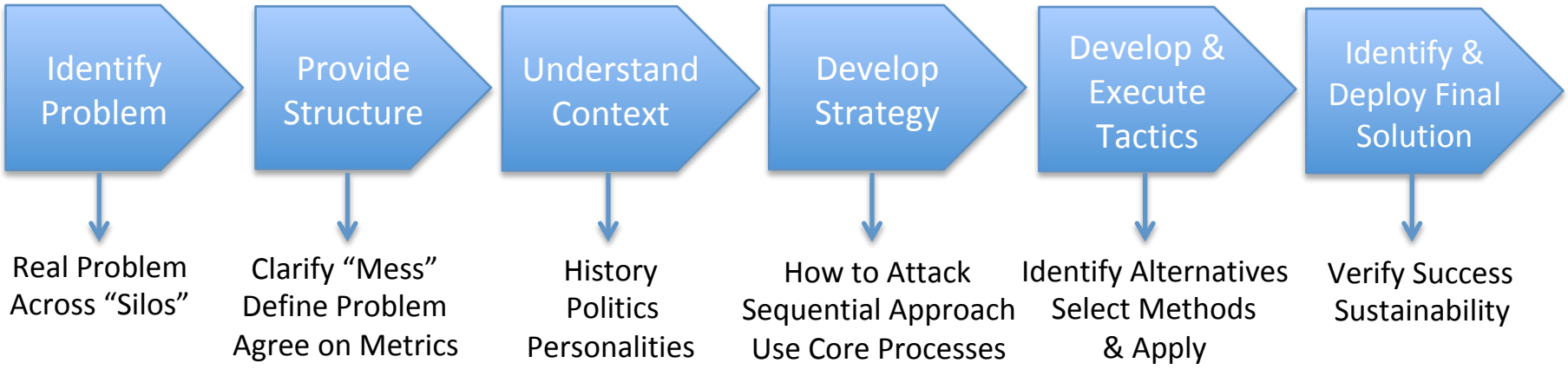
Fundamentals of Statistical Engineering

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. *Create Structure:* convert 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 it solution over time

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

Flow of Statistical Engineering Projects



Note:

- 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

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?
- How did we get where we are today? What are the root causes?
- What solutions have been attempted previously? Why didn't they work?
- What politics are involved (often unstated)?
- Our objective is typically a "useful model", not "optimal" model
  - Best technical or business solution versus best statistical solution

There are reasons 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
    - Consider France in the World Cup

Large, complex problems require a strategy

Data Pedigree

- All data are ***not*** created equal – sounds obvious, but isn't!
- Have the data been modified, filtered, "cleaned", or altered in any way, since collected?
- Does a "gold standard" copy of the original data exist?
- Models should never be more complex than can be supported by the available data
- To properly analyze data we must understand the process that produced them
- Documentation of statistical analyses should include limitations, including restrictions on application of the analyses

Data are "guilty until proven innocent"

Subject Matter Knowledge

- 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 aspect of 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 existing hypotheses
- A sequential approach fits well with the use of an overall strategy

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

Summary

- Textbook problems do not prepare people for the large, complex, unstructured problems they will face in the real world
  - A different approach is needed
- An engineering paradigm seems to work for these complex problems
- Statistical engineering provides a framework that can accelerate the learning curve in attacking such problems
- Statistical engineering is currently being developed and documented as a discipline, i.e., it is a “work in progress”
  - Including development of an underlying theory
- It does, indeed, appear to be an idea whose time has come!

Significant implications for statistical practice, education, and research