Debunking common myths and misconceptions of Data

Anthony J. Scrriffignano, PhD
SVP/Chief Data Scientist
Dun & Bradstreet
1-April, 2021
The opportunities to use data in new and amazing ways are changing before our eyes...
The fabric of data is represented in the connectedness of everything
Matching supply and demand, vetting relationships

Basic understanding: supply chain disruption

Integrated value chains: Impact
AGENDA

CHANGING TIMES
CHANGES IN THE ENVIRONMENT OF ALL THINGS DATA

COMMON MYTHS / INCONVENIENT TRUTHS
THE IMPLICATIONS OF TRIVIAL INFERENCE FROM DATA

THINKING ABOUT THE FUTURE
ADVICE, PERSPECTIVE
CHANGING TIMES

CHANGES IN THE ENVIRONMENT OF ALL THINGS DATA
Data, data, data…
The changing nature of data and “digital everything”

- Localization
- Autonomous
- Adversarial
- Unstructured
- Data
- Intentionally manipulated
- Ethical
- Personal
- Permissible Use
- Meaning
- Veracity
- Velocity
- Volume
- Variety
- Synthesis
- Curation
- Edge
- Governance
- Privacy
- Big Data
- Bias
- Secure
- Redundancy
- Incomparable
- Latent
- Incomplete
- Explainable
- Benefit
Data is new ... Data

**Water**
- Finite supply
- Poorly distributed
- Quality varies greatly
- Infinitely recyclable

**Oil**
- Finite supply
- Poorly distributed
- Controlled by sovereign states
- Requires refinement
- Destroyed to reap benefit when used

**Currency**
- Finite supply
- Poorly distributed
- Different value in different sovereign states
- Consumed to benefit when used
- Value varies over the time

**Data**
- Infinitely growing hypergeometrically
- Data begets data
- Ubiquitous
- Controlled by the informed
- Requires refinement
- Contains bias
Part of the challenge is that we keep using the same words to describe different things.
When is enough enough?

**DATA IN HAND**
- Opportunity Formulation
  - Relative size
  - Key question

**DISCOVERABLE DATA**
- Dispositive Threshold
  - Estimate
  - Triangulate

**EXISTING BUT INACCESSIBLE DATA**
- Decision Elasticity
  - Bias
  - Opportunity cost

**DATA YET TO BE AVAILABLE**
What does it mean?
Understanding the stability of an environment

Points to consider:

**Character**
- Rate of change (Data at rest, in motion, stream)
- Ontological variation (ability to connect, synthesize)
- “Highest and best” source
- Latency
- Granularity

**Quality**
- Stability (variation)
- Metadata (likely usefulness)
- Context (source and use)
- Required transformations (data standardization, normalization)
- Known or identifiable bias
- Heuristic evaluation
THE TRUTH - When possible, especially with positivistic questions, compare to an authoritative source
  • Beware data in motion (e.g. exchange rates, population data)
  • Retain provenance and permissible use
  • Retain metadata
  • Consider standardization, but beware the impact of multiple methods

THE WHOLE TRUTH – Use empirical methods to assess missingness
  • Look for patterns in missing data that may indicate systemic remediation opportunities
  • Consider metadata violations or misinterpretations that may contribute to the situation
  • Consider standardization, but beware the impact of multiple methods

NOTHING BUT THE TRUTH – Assess carefully, and be ready to “unlearn”
  • Build in techniques for redacting information that fails subsequent adjudication
  • Assess the degree to which the data contributes to the solution vs. adding “noise”
  • Instantiate methods to control for data inclusion based on epistemology and usefulness
End-to-End Model for Data Stewardship

- **Discovery**
  - identifying new sources of information
  - Estimating quality and veracity
  - understanding changes in the data environment and customer demands
  - managing cost efficiencies
  - developing new technologies/processes for identification of data assets

- **Curation**
  - creating ways to transform data to meet current and evolving needs
  - Reposing data in a way that it can be best used considering data in motion, data at rest, etc.
  - Creating derived data assets (e.g., signals, graphs)
  - Detecting anomalies, cliques, and other useful relationships

- **Synthesis**
  - creating ways to derive new insight from data
  - developing new/improved capabilities
  - Inferring metadata or other ontology
  - Discovering quixotic or useful methods of inference

- **Governance and Quality Assurance** must exist into/out of the model as well as within and between steps

- **Fabrication and Delivery** will depend on implementation specifics and can range from manual processes to highly automated services
Curation example: Language transformation

We must also consider regional nuance:

- Writing with implied orthography (e.g. Arabic)
- Declension (e.g. Greek) παράδειγμα
- Definite / indefinite articles missing (e.g. Russian) местоимение
- Measure words (e.g. Chinese) 一个杯子
Synthesis Example: Malfeasance

**BUSINESS-TO-HIGH RISK & FRAUD**
- Trade –(AR & AP)
- Bust out
- Business Id Theft
- Financial Statement
- Payment / Trade-Ring
- Business Credit Coach

**EMERGING**
- Cyber
- Cryptoeconometrics
- Convergence (e.g. IoT and Crypto)
- Unexplainable AI
- Fake news
- Dark Web
- Silent Data Breach

**THE FRAUD TREE**

- Corruption
  - Bribery
  - Illegal Gratification
  - Economic Expropriation

- Asset Misappropriation
  - Embezzlement
  - Theft
  - Robbery

- Financial Statement Fraud
  - Inflation
  - Deflation
  - Graft
  - Kickback

**The Fraud Tree Overview**

EMERGING Examples:
- Synthesis Example: Malfeasance
- Trade –(AR & AP)
- Bust out
- Business Id Theft
- Financial Statement
- Payment / Trade-Ring
- Business Credit Coach

- Cyber
- Cryptoeconometrics
- Convergence (e.g. IoT and Crypto)
- Unexplainable AI
- Fake news
- Dark Web
- Silent Data Breach
COMMON MYTHS AND INCONVENIENT TRUTHS

The implications of trivial inference from data...
We live in an age of promise. Big data and advanced methods are making things possible that were science fiction only a few short years ago.
How do we know?
Jumping to the tools and technology before considering the question…
Thinking about data as we receive it

- Expectations
- Sampling
- Bias
- Missingness
- Imperative
- Regression
- Learning

You are the only person in the meeting.
It is important to challenge what we think is true... and what we fear is true.

<table>
<thead>
<tr>
<th><strong>MYTHS</strong></th>
<th><strong>TRUTH</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>More Data is better</td>
<td>Data vs. noise</td>
</tr>
<tr>
<td>Lots of data in 1 place is sufficient to learn</td>
<td>Data at rest vs. data in motion</td>
</tr>
<tr>
<td>AI can find answers</td>
<td>AI methods have preconditions</td>
</tr>
<tr>
<td>Machine learning will find hidden truth</td>
<td>Regression vs. unprecedented change</td>
</tr>
<tr>
<td>Natural language processing removes all language barriers</td>
<td>Language is constantly changing</td>
</tr>
<tr>
<td>Machine Translation is good enough</td>
<td>Many unmet challenges remain in linguistics</td>
</tr>
</tbody>
</table>
A careful look at data and analytic insight...

<table>
<thead>
<tr>
<th>Myths</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magic wand</td>
<td>Collection of models and algorithms, often applied at the same time</td>
</tr>
<tr>
<td>Computational power is its own convocation!</td>
<td>Not yet – assumptions and data yield “built in” results and output</td>
</tr>
<tr>
<td>AI can replace human intelligence</td>
<td>Not yet – Maybe never. AI constrained by data, data topography</td>
</tr>
<tr>
<td>AI always produces the best possible answer</td>
<td>Even sophisticated models can fail at tasks a human finds easy; model assumptions can propagate bias!</td>
</tr>
<tr>
<td>Computers in lieu of people</td>
<td>AI will NOT produce results without intelligent people behind it</td>
</tr>
<tr>
<td>AI transcends traditional programing (e.g. Skynet)</td>
<td>Still code: make one typo in the code and the program won’t run</td>
</tr>
</tbody>
</table>
Implications…
The dialogue about data is complex – and personal
Why would we assume it’s all for the good?
‘Black Cat’ problems

Dealing with “Black Cat” problems
• Signals – internal / external / derived
• Nodes / edges, full and partial graph measures
• Systemic measures – quality, character
• Anomaly detection / Graph inspection
  – Anisotropism
  – Betweenness centrality
  – Hierarchical simplification / Sparse graph analysis
  – Cliques
  – Clustering coefficients
• Data sensing – new sources / uses
• Triggers – events, observations

Scare the cat
Turn on the lights
Find damage to the furniture
Remove the furniture
Bring in some cat food

Examples
• Fraud
• New business instantiation
• Black Swan events
THINKING ABOUT THE FUTURE

ADVICE, PERSPECTIVE
Best Practices: Key strategies to extracting insight

- Multisourcing of information and mindful partnering
- Perpetual adjudication and segmentation of veracity
- People in the process in the right places
- No people in the process in the right places
- Innovating on the Source vs. Innovating on the process
- Meaningful learning from fraud and volatility in data
- Consider all V’s simultaneously
- Empirical process (repeatable)
- New technologies: perspective on myth vs. reality
Thoughts on innovation in the current context

<table>
<thead>
<tr>
<th>Innovation Focus</th>
<th>Explanation</th>
<th>Key Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understand the Universe</td>
<td>Assess data curation vs. rate of change in the environment. What is no longer true? What assumptions have changed?</td>
<td>Rate of disruption exceeds rate of curation. Lack of ground truth. Actions taken to react may not be visible with old discovery methods.</td>
</tr>
<tr>
<td>Understand Patterns of Disruption</td>
<td>Establish methods to assess disruption based without relying on “learning” from prior trends.</td>
<td>Many traditional modeling and machine learning methods are not appropriate.</td>
</tr>
<tr>
<td>Support Decisions</td>
<td>Provide information to enterprise and agencies in a timely manner to reduce the impact of the crisis.</td>
<td>Recognize, react, adapt, recover, repeat.</td>
</tr>
<tr>
<td>Analytics for the “New Normal”</td>
<td>Establish AI and analytic methods that survive ongoing disruption.</td>
<td>Hybrid methods, confirmation bias, other sources of variation.</td>
</tr>
<tr>
<td>Innovation in the context of Disruption</td>
<td>Deliver products and services in time to be relevant.</td>
<td>Regulatory compliance, permissible use, veracity, regression.</td>
</tr>
<tr>
<td>Opportunity Cost</td>
<td>Understanding the cost and impact of a failure to act.</td>
<td>Market pressure, increasing customer demands, timing/relevance.</td>
</tr>
</tbody>
</table>
The key to progress: Making new mistakes.
Making things “Computable”

**Formal Theory of Computation:**
Solving problems efficiently and effectively using an accepted mode of algorithmic treatment

3 major branches:
1. Automata theory – study of abstract “machines”
2. Computability theory – tractability, reducibility
3. Computational complexity theory – assessing the effort to solve

**Scientific Methods**
- Formulating a research question
- Understanding what research has already been done / what is assumed to be true
- Picking methods and understanding preconditions
- Analyzing the data with a careful eye to bias and empirical rigor

**Cross-functional teams**
- Include as broad a cross-section of thinking as possible
- Assign peer review and opposition roles to make sure that the work is defendable
- Account for dynamic changes in the system as you are implementing a solution

**Think about the future**
- Understand how you would know when conditions have changed sufficiently to warrant reevaluation
- Ensure a process for captured learnings
- Challenge the value proposition from the perspective of the customer or user
The risks are real as well

Risks to be considered

- Focusing on the wrong problem
- Ignoring the trends: AI, IoT and other emerging technologies are here to stay; unprecedented risk to be disrupted
- Betting on the wrong trend/technology can be VERY costly, both in terms of investment but also from market share standpoint
- Investing in new capabilities in light of regulatory uncertainty
- Complying with emerging and future regulations massively non-trivial
- Bad guys get further ahead, and faster!

Emerging dialogue

- Inequality
- AI Bias
- Adversarial Manipulation
- Federation of Technology
- Open Source
- Data Rights
- Intellectual Property
- Agency
- Explainability
Future Thinking: What is Driving the Change?

Source: frankdiana.net

Disruptive Evolution
A. Scalfignano, D&B
New innovation questions to ask…

- What do we have to believe?
- What sort of bias is in our questions?
- Who are the disruptors?
- How do we define innovation?
- What can we learn from what we don’t know?
- How can we best use the smart humans?
- What do we learn today that will be relevant in 3 years?
New context questions to ask…

- What do we have to believe? (Bias, preconditions)
- What happens if we do nothing? (New malfeasance, eroding relevance)
- What is the problem or challenge we are attempting to address?
- What NEW skills do we need?
- Is the environment changing faster than the data that describes it?
“You must look within for value, but beyond for perspective.”
Denis Waitley, author
Thank you

Anthony J. Scriffignano, PhD
@scriffignano1
Also LinkedIn
Abstract:

Business decisions have always been informed by data: data about the suppliers, customers, and other counterparties in a relationship, the nature of a transaction and historical risk are but a few examples. With the advent of AI and other advanced analytical capabilities at the forefront of decision-making, it is crucial to challenge our longstanding beliefs and practices regarding, the discovery, curation, and synthesis of data. Our data is changing in curious and sometimes alarming ways.

In this session, Dr. Anthony Scriffignano, SVP/Chief Data Scientist at Dun and Bradstreet, will explore how the massive availability and changing nature of data is altering the way modern business decisions are made. He will discuss some of the common myths and inconvenient truths about the data we use to make decisions. The session, both relevant and at times irreverent, will challenge how we think about the data that is underpinning so much of what we do in the modern world.