Data Science and Prediction*

Vasant Dhar

Professor
Editor-in-Chief, Big Data
Co-Director, Center for Business Analytics, NYU Stern
Faculty, Center for Data Science, NYU

*Article in Communications of the ACM, Vol. 56 No. 12, December 2013
http://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltext
Themes: What Makes Prediction Hard?

• Noise (physical versus social systems)
  – “Physics has 3 theories that explain 99% of observed phenomena, whereas Psychology has 99 theories that explain 3% of observed phenomena.”

• Not knowing the right question to ask (formulation)
  – “If only you knew what to ask, I’d show you something really interesting!”

• Not having the right/Enough data: observational versus experimental
  – “Am I looking under the lamppost for the key because...?”

• Combining machine and human intelligence
  – “Surely human and machine intelligence can augment each other?”

• Believing in the analysis!
  – “I don’t believe the result. Go find the mistake!”

@vasantdhar @digitalarun
The Data Landscape and Applications

- **Financial Markets**
  - What will the market do tomorrow?
  - Will the retail sector pull back within a month?
- **Healthcare**
  - Who will become sick in the near future?
  - How will some respond to a medication?
- **Marketing**
  - Who will respond to what offer?
  - Is a customer likely to attrit shortly?
- **Social/Product Networks**
  - Will demand for XXX go up next week given the activity of its neighbors?
  - How should I craft my message so that it “spreads” through the network? i.e. where should I “seed” it?
- **What is the “sentiment” in a collection of textual data?**
  - Does the sentiment have any predictive power?
Data Science and Prediction

“Data Science is the study of the generalizable extraction of knowledge from data”*

A key epistemic requirement for new knowledge (and its “actionability”) is its ability to **predict** and not just **explain**

1. Noise

Physical Systems: theory is expected to be "complete"

Social/Health Systems: incomplete models intended to be partial approximations of reality, often based on assumptions of human behavior known to be simplistic.
What is Noise Anyway?

How would you order these on the continuum?

Source: LMCM analysis.
Skill and Luck in Baseball: Batting Avg, Singles, and Strikeout YoY Correlation

Baseball Metrics By Skill and Luck

SO Rate: Strikeout Rate
IP HR Rate: In-play home run rate
BB Rate: Walk rate
OBP: On-base percentage
IP 2+3 Rate: In-play doubles and triples rate
IP AVG: In-play batting average
IP S Rate: In-play singles rate
AVG: Batting average
Is This Ordering Credible?

Reversion to the mean exists in activities that combine skill and luck. It is useful to know where the problem lies on the continuum above. Knowing where we lie in the continuum allows us to anticipate outcomes. Illusion of control is a factor in luck situations!
Disentangling Skill and Luck: The Formula

\[
\text{Variance(observed)} = \text{Variance (skill)} + \text{Variance (luck)}
\]

\[
\text{Variance(skill)} = \text{Variance (observed)} - \text{Variance (luck)}
\]

For win/loss outcomes, stdev = \( \frac{p(1-p)}{\sqrt{n}} \)
where \( p \) = prob of outcome (i.e. win), and \( n \) = number of cases (i.e. games)

This is observable from the data

Depends on sample size
Prediction in Noisy Domains (Markets)*

*From: Dhar, V., Prediction in financial markets: The case for small disjuncts, ACM transactions on Intelligent Systems and Technology, volume 2, No 3, April 2011
2. Asking the Right Question

“Patterns Emerge Before Reasons for Them Become Apparent”

Asking the right question is therefore critical: “If only you knew what question to ask me, I’d give you very interesting answers from the data.”

Keep moving on? Dig for causality?
What is the Right Question Here?*

Are complications associated with the yellow meds?
Or with the gray meds?
Or the yellows in the absence of the blues?
Or is it more than three yellows or three blues?
Or is it the greens in “quick succession?”
Or does it have to do with “lifestyle choices?!” (i.e. Bias? Gather mo data?)

http://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltex
High Level View of Model Discovery

Decision Rule or Trading Strategy (i.e. the “question”)

Better part → Breeds → Better part → Breeds → Better part → Breeds → Better part

Worse part → Drops Out → Worse part → Drops Out → Worse part → Drops Out → Worse part

Solution Quality

Best

Average

Worst

Iterations
Solutions Can Represent Arbitrary Data Structures

Arrays and sequences

Trees

Boolean Expressions

I.e. X1 between 0 and 0.8
An Interesting Pattern?

<table>
<thead>
<tr>
<th>Patient</th>
<th>Age</th>
<th>#Medications</th>
<th>Complication</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>52</td>
<td>7</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>57</td>
<td>9</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>43</td>
<td>6</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>33</td>
<td>6</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>8</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>49</td>
<td>8</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>58</td>
<td>4</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>62</td>
<td>3</td>
<td>No</td>
</tr>
<tr>
<td>9</td>
<td>48</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>37</td>
<td>6</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Age $\geq 37$
AND
#Medications $\geq 6$
\Rightarrow Complication = Yes (100% confidence)
3. Observational Data and Acquiring New Data

- Observational data may answer questions without explicitly asking anyone anything!
- It may also require an understanding of how the data are being generated
  - Are there “natural experiments” that are reflected in the data or are the data somehow biased through self selection?
  - Is it possible to run natural experiments to get additional data to answer the question?
Have We Collected The Right Data?

ARE YOU DRUNK?

☐ YES
☐ NO

X

X
Is More Always Better?*

• Is more data always better?

• How much you should pay for “external” data?**

*See Provost et.al article in Big Data journal (volume 2, issue 1, 2014)
** See Dalessandro et.al article in Big Data journal (volume 2, issue 2, 2014)
4. Prediction as Epistemic Criterion

...In assessing whether new knowledge is “actionable” for decision making: predictive power, not just ability to explain the past
Validation Framework for Prediction

Across Time

<table>
<thead>
<tr>
<th>Across Universe</th>
<th>NO</th>
<th>YES</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td>![Diagram](Out of sample)</td>
<td>![Diagram](Out of sample Out of time)</td>
</tr>
<tr>
<td>YES</td>
<td>![Diagram](Out of sample Out of universe)</td>
<td>![Diagram](Out of sample Out of time Out of universe)</td>
</tr>
</tbody>
</table>
Prediction Vs Explanation: Varying Model Complexity

Future Performance Versus Complexity

Zone with (small) correlation between in-sample and out of sample

Overfitting

Note: Complexity measure is illustrative only, based on the learning method used
Sources of Error in Predictive Models

1. Mis-specification of the model
   - Big data admits a larger space of functional forms

2. Using a sample to estimate the model
   - With big data, sample is a good estimate of the population

3. Randomness
   - Predictive modeling attempts to minimize the combination of these two errors

*Adapted from Shmueli, G. To explain or to predict? Statistical Science 25, 3 (Aug. 2010), 289–310.)
Making Predictions Usable in Noisy Domains*

...why bother making predictions? i.e. with so many false positives?

<table>
<thead>
<tr>
<th></th>
<th>+A</th>
<th>-A</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>+P</td>
<td>60 (TP)</td>
<td>100 (FP)</td>
<td>160</td>
</tr>
<tr>
<td>-P</td>
<td>40 (FN)</td>
<td>400 (TN)</td>
<td>440</td>
</tr>
<tr>
<td>Totals</td>
<td>100 (p+)</td>
<td>500 (p-)</td>
<td>600</td>
</tr>
</tbody>
</table>

It all depends on the costs of being wrong...and the aggregate pickup*

*From “Big Data and Predictive Analytics in Healthcare,” Big Data Journal, volume 2, Number 3, Sep 2014
http://online.liebertpub.com/doi/pdfplus/10.1089/big.2014.1525
If Patterns Emerge Before Reasons Are Apparent...

...does it imply you’re better off rebuilding the model periodically?

...when is it necessary to understand the reasons for the patterns? (i.e. causality)
If Prediction is Important...

...is problem formulation different than if the purpose is explanation?

...is it about the right tradeoff between the structure of the problem and the complexity of the model?

...how do you ensure against common mistakes such as “data snooping” or “leakage” in building predictive models?
Transparency in Complex Systems

“How can you describe the learned model?”

Is it possible to envision when the model will and won’t perform well?
What’s the Explanation?

What happened here and why?
Current Areas of Inquiry with Big Data

• Unstructured data
  – WATSON: leveraging human curated data and AI to *synthesize* answers
  – Deciding data what to keep, aggregate, etc
• New risk factors based on text and other data at higher frequency
  – News and social media
  – “Digital IQ” of companies
• Understanding the relationship between noise, system behavior, expected predictive accuracy, and performance bounds
• Understanding how “actionable results” will change future data making future insights harder and less actionable!
Thank You!