NYU STERN

NEW YORK UNIVERSITY · LEONARD N. STERN SCHOOL OF BUSINESS

CENTER FOR BUSINESS ANALYTICS

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!”

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**

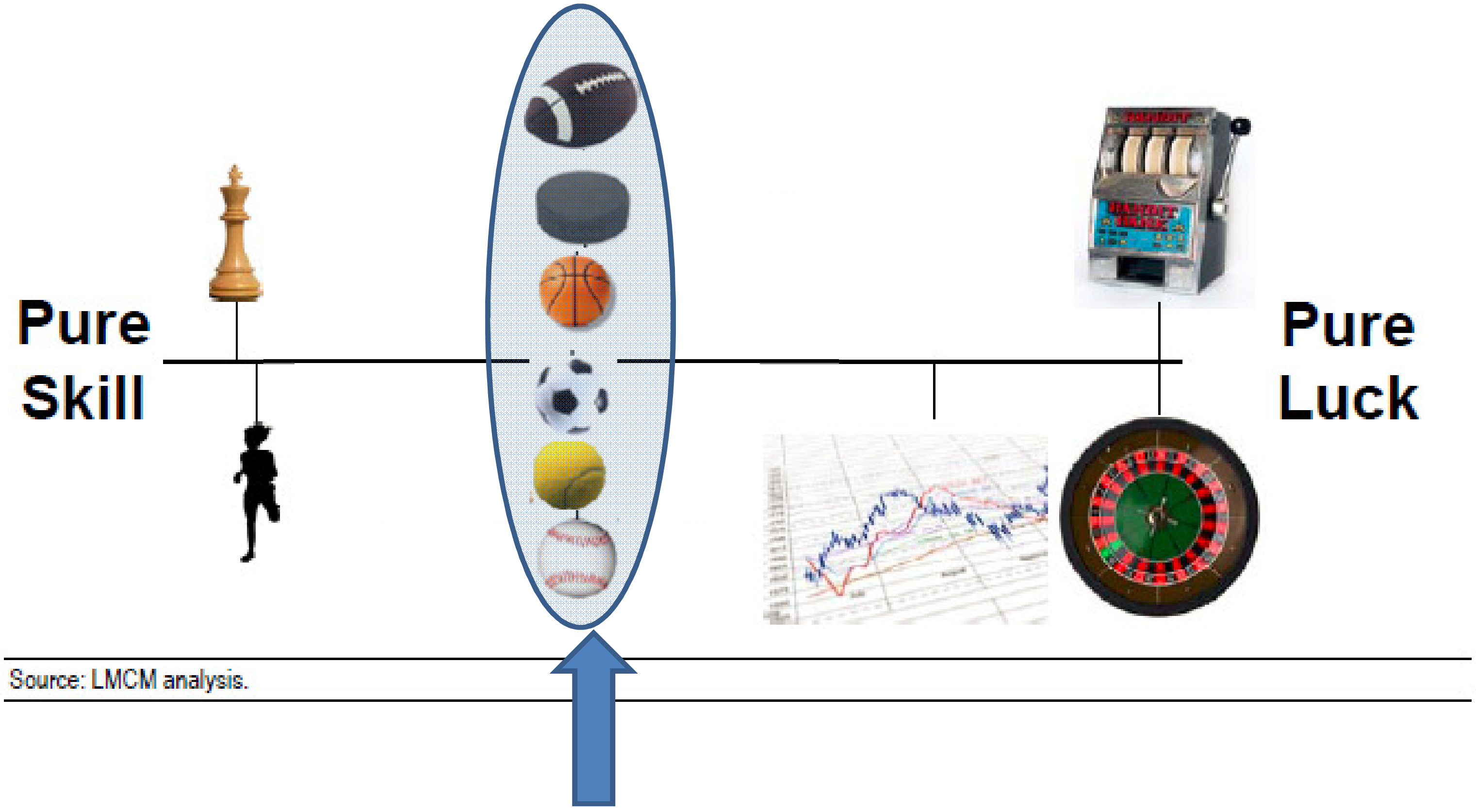
***Dhar, V., Data Science and Prediction, Communications of the ACM, Vol. 56 No. 12, December 2013.**
http://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltex

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.

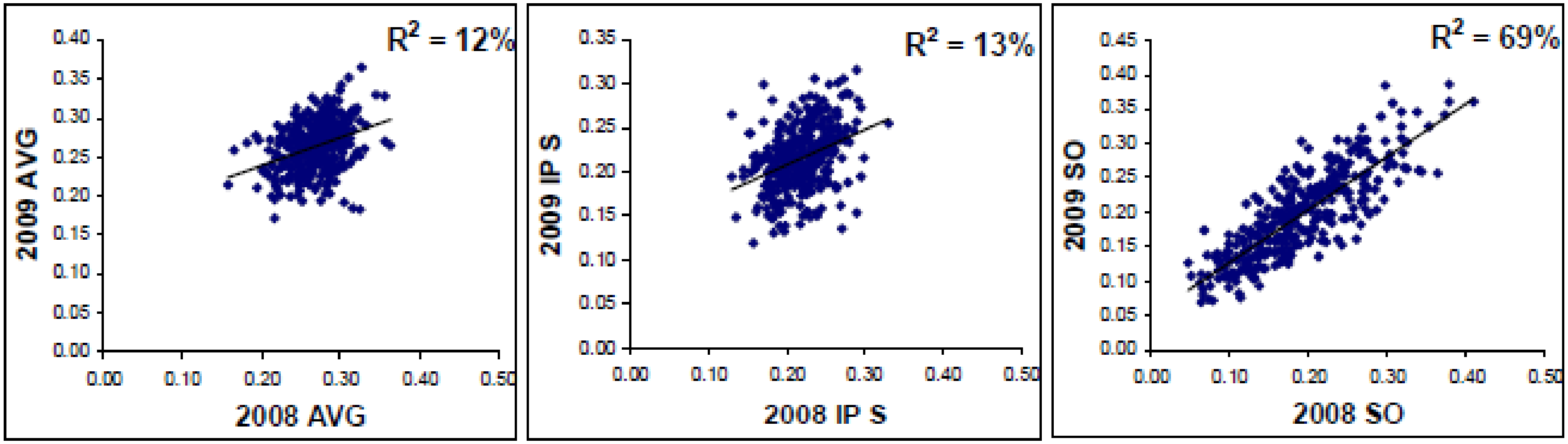
What is Noise Anyway?



Source: LMCM analysis.

How would you order these on the continuum?

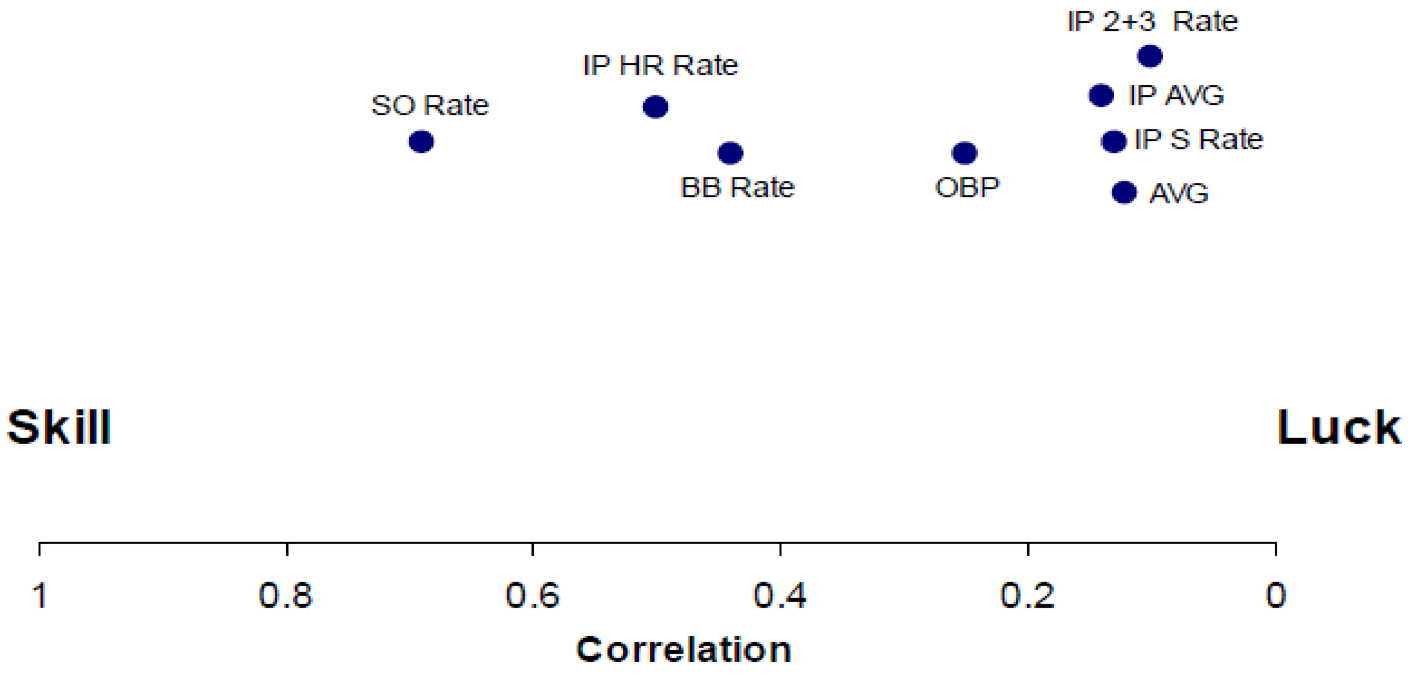
Skill and Luck in Baseball: Batting Avg, Singles, and Strikeout YoY Correlation



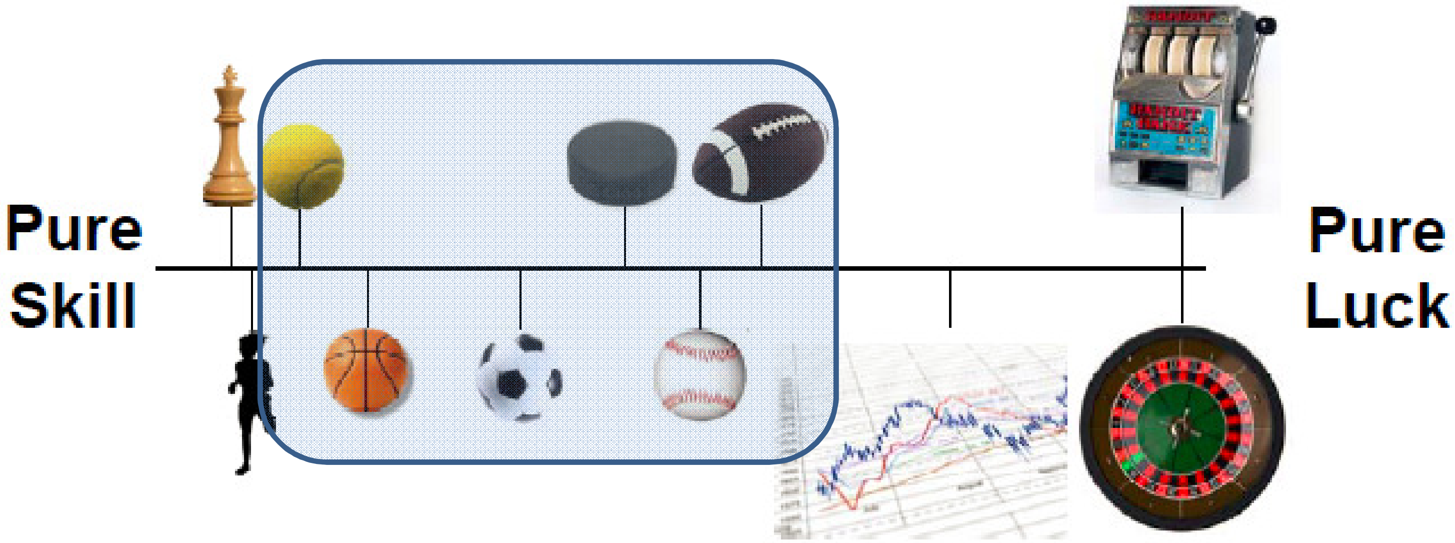
Source: LMCM analysis based on Jim Albert, "A Batting Average: Does It Represent Ability or Luck?" *Working Paper*, April 17, 2004.

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?



Source: LMCM analysis.

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

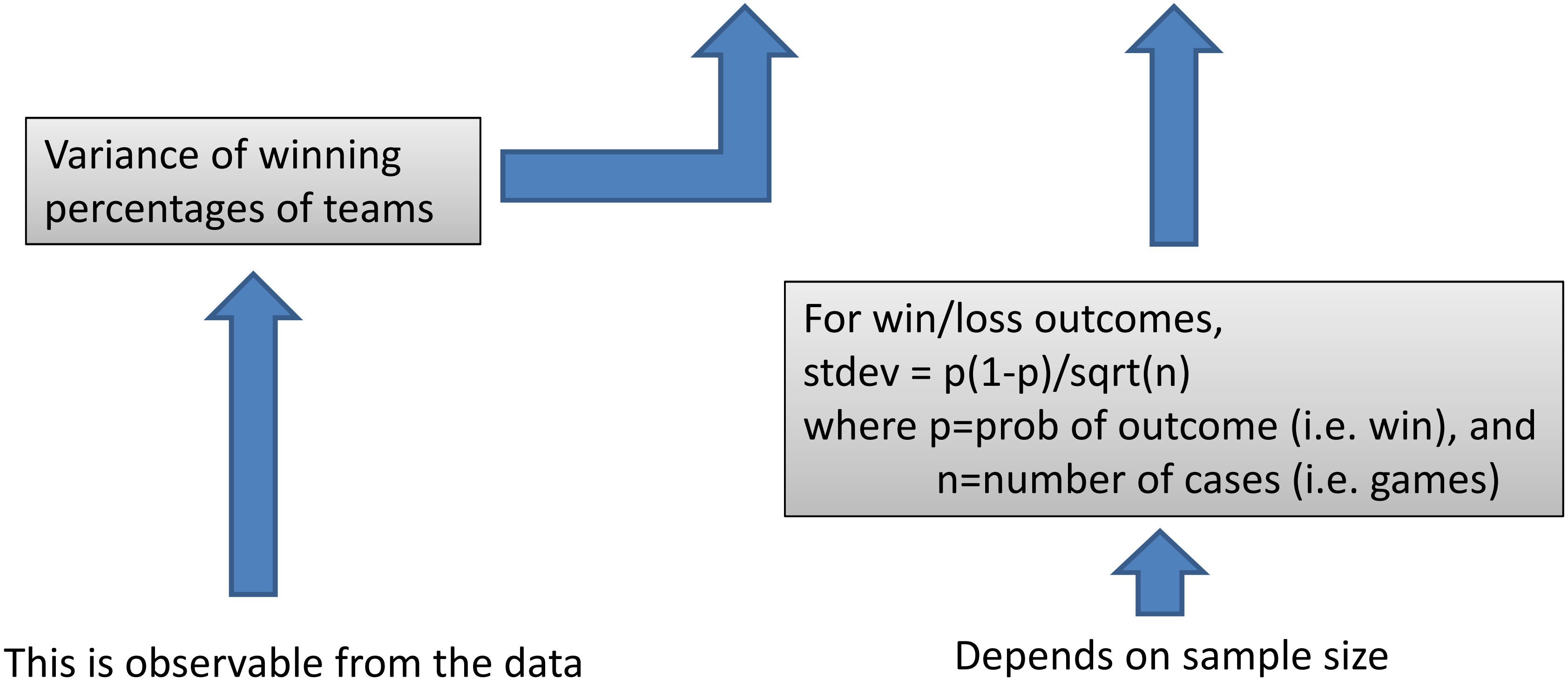
Knowing where we lie in the continuum allows us to anticipate outcomes

Illusion of control is a factor in luck situations!

Disentangling Skill and Luck: The Formula

Variance(observed) = Variance (skill) + Variance (luck)

Variance(skill) = Variance (observed) – Variance (luck)



Variance of winning percentages of teams

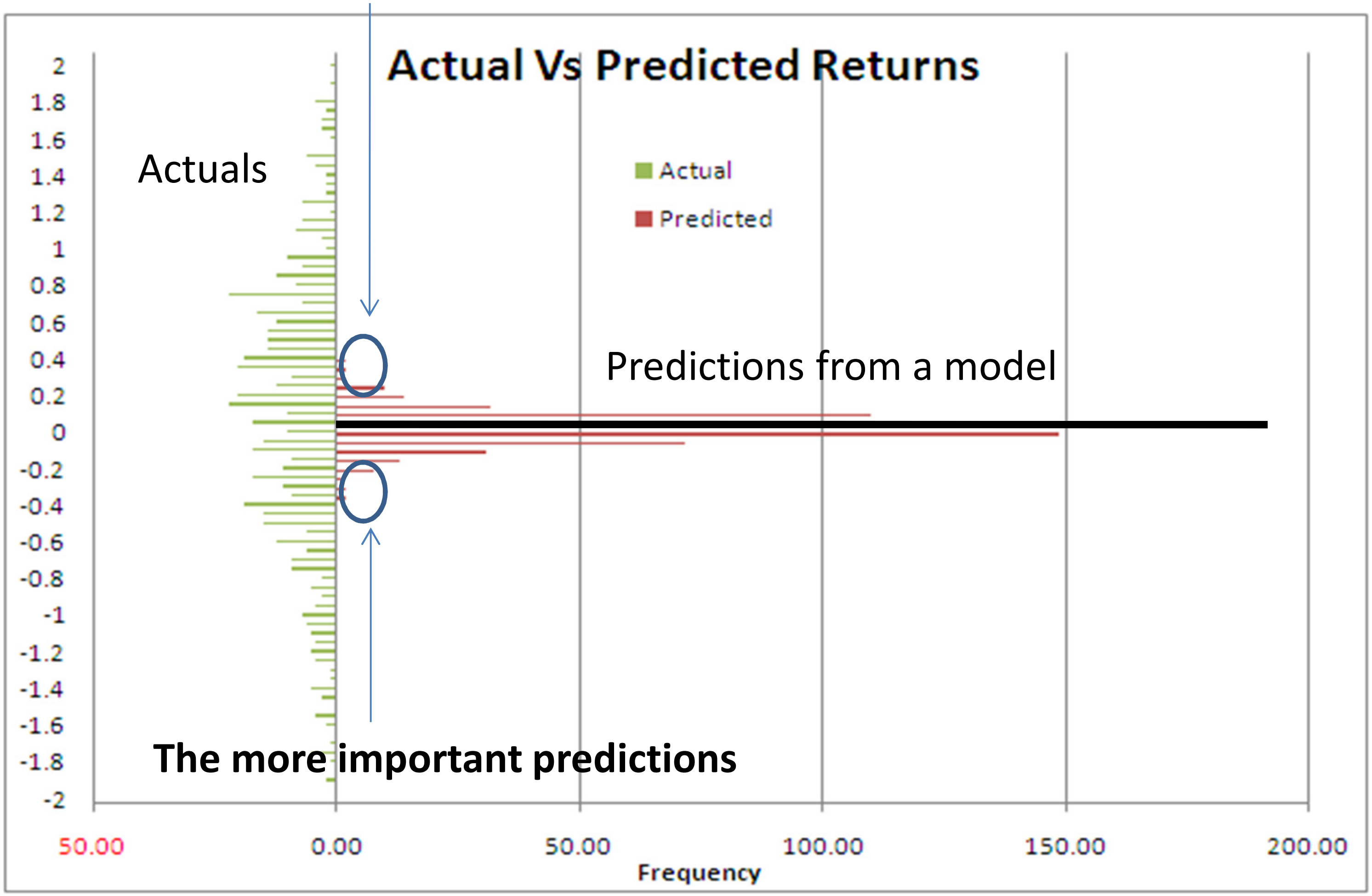
For win/loss outcomes,
stdev = p(1-p)/sqrt(n)
where p=prob of outcome (i.e. win), and
n=number of cases (i.e. games)

This is observable from the data

Depends on sample size

Prediction in Noisy Domains (Markets)*

The more important predictions



Actuals

Predictions from a model

The more important predictions

*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

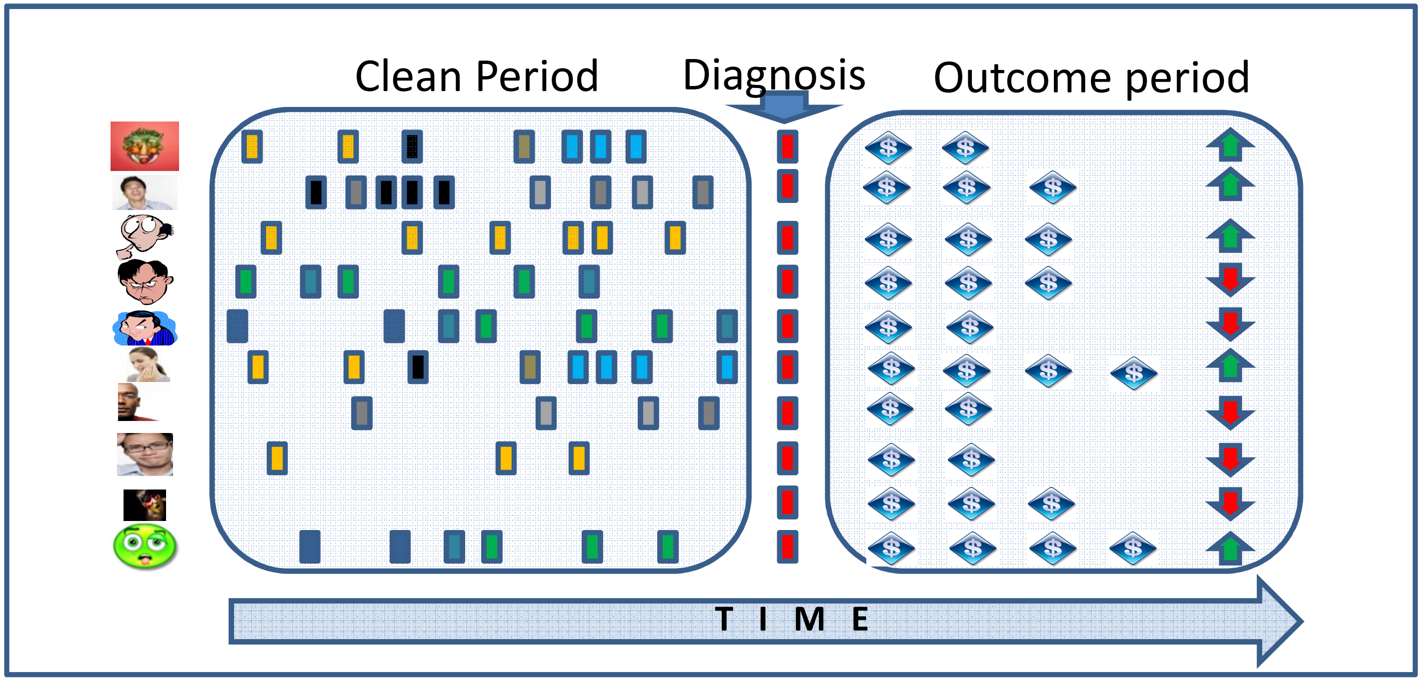
2. Asking the Right Question

“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?

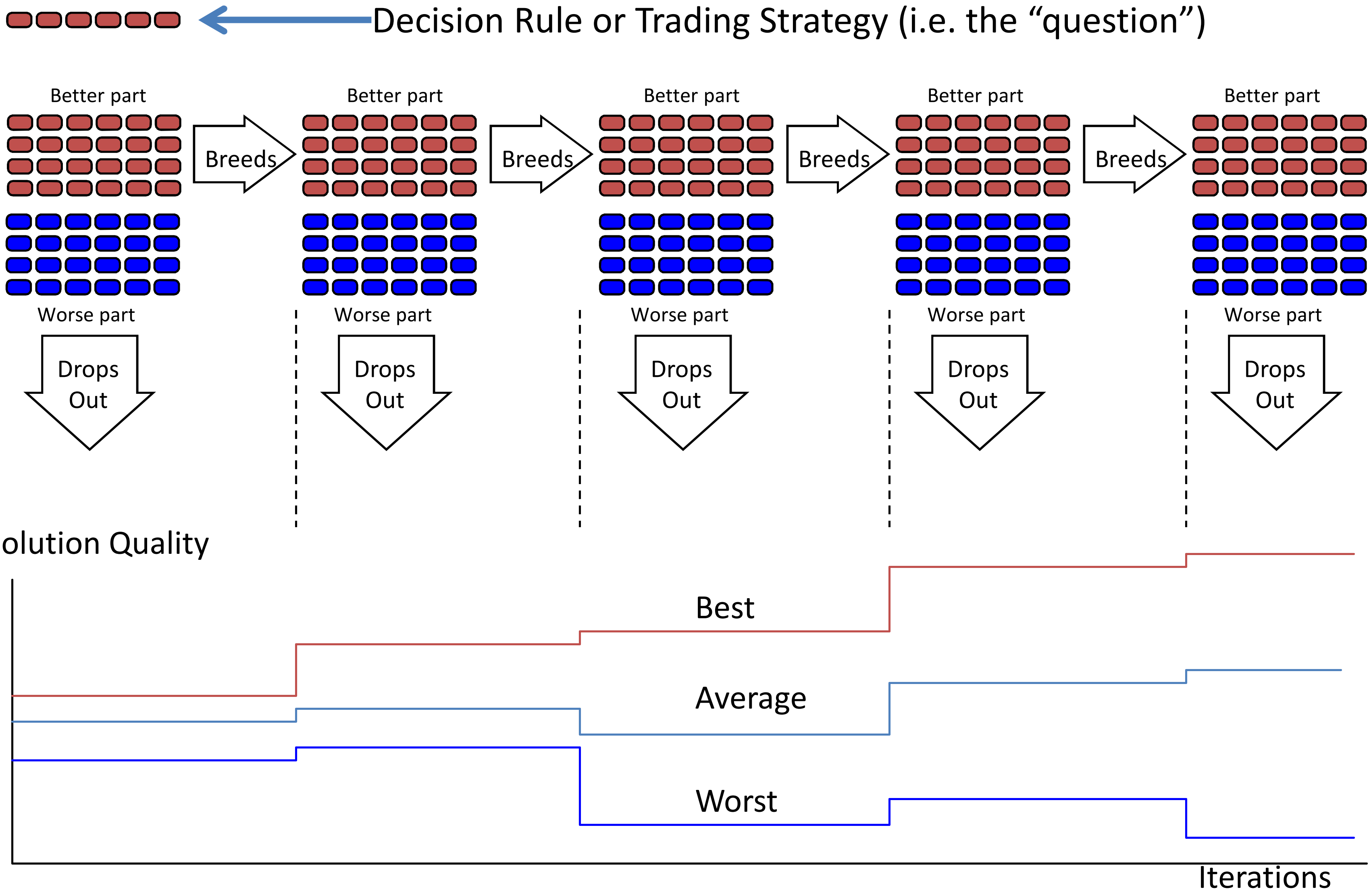
What is the Right Question Here?*



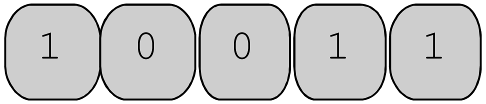
Are complications associated with the yellow meds?
Or with the gray meds?
Or the yellows in the absence of the blues?
Or is it more than three yellows or three blues?
Or is it the greens in “quick succession?”
Or does it have to do with “lifestyle choices?!” (i.e. Bias? Gather mo data?

***Dhar, V., Data Science and Prediction, Communications of the ACM, Vol. 56 No. 12, December 2013.**
http://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltex

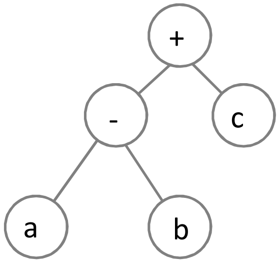
High Level View of Model Discovery



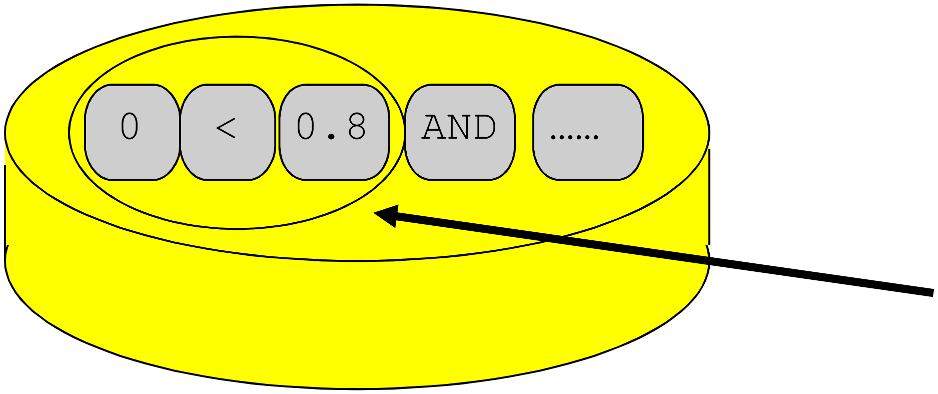
Solutions Can Represent Arbitrary Data Structures


1 0 0 1 1

Arrays and sequences



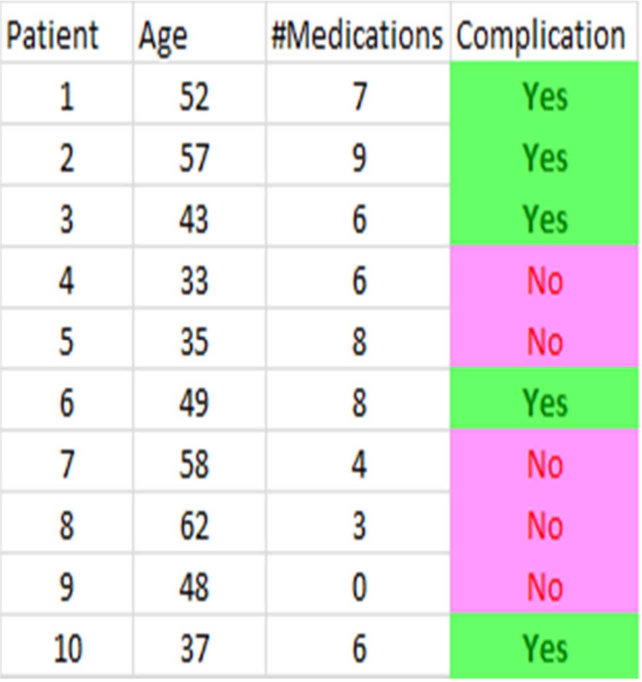
Trees


0 < 0.8 AND ……

Boolean Expressions

I.e. X1 between 0 and 0.8

An Interesting Pattern?



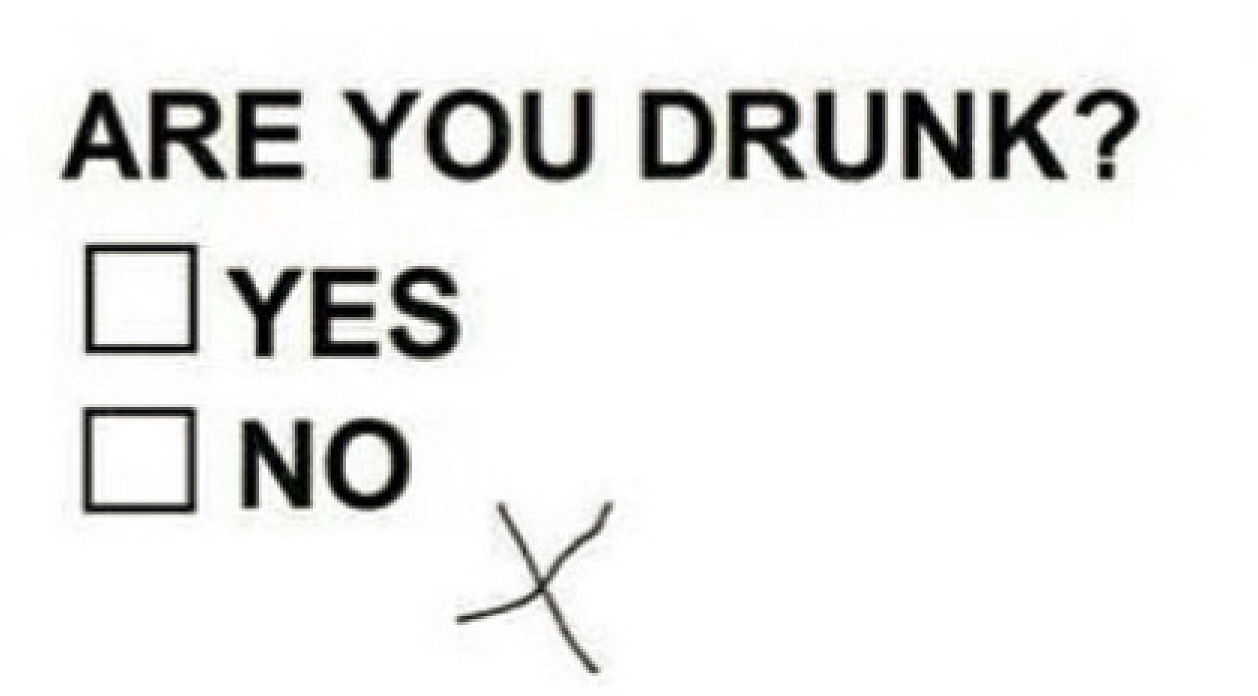
Patient	Age	#Medications	Complication
1	52	7	Yes
2	57	9	Yes
3	43	6	Yes
4	33	6	No
5	35	8	No
6	49	8	Yes
7	58	4	No
8	62	3	No
9	48	0	No
10	37	6	Yes

Age >= 37
AND
#Medications >= 6
→
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

Is More Always Better?*

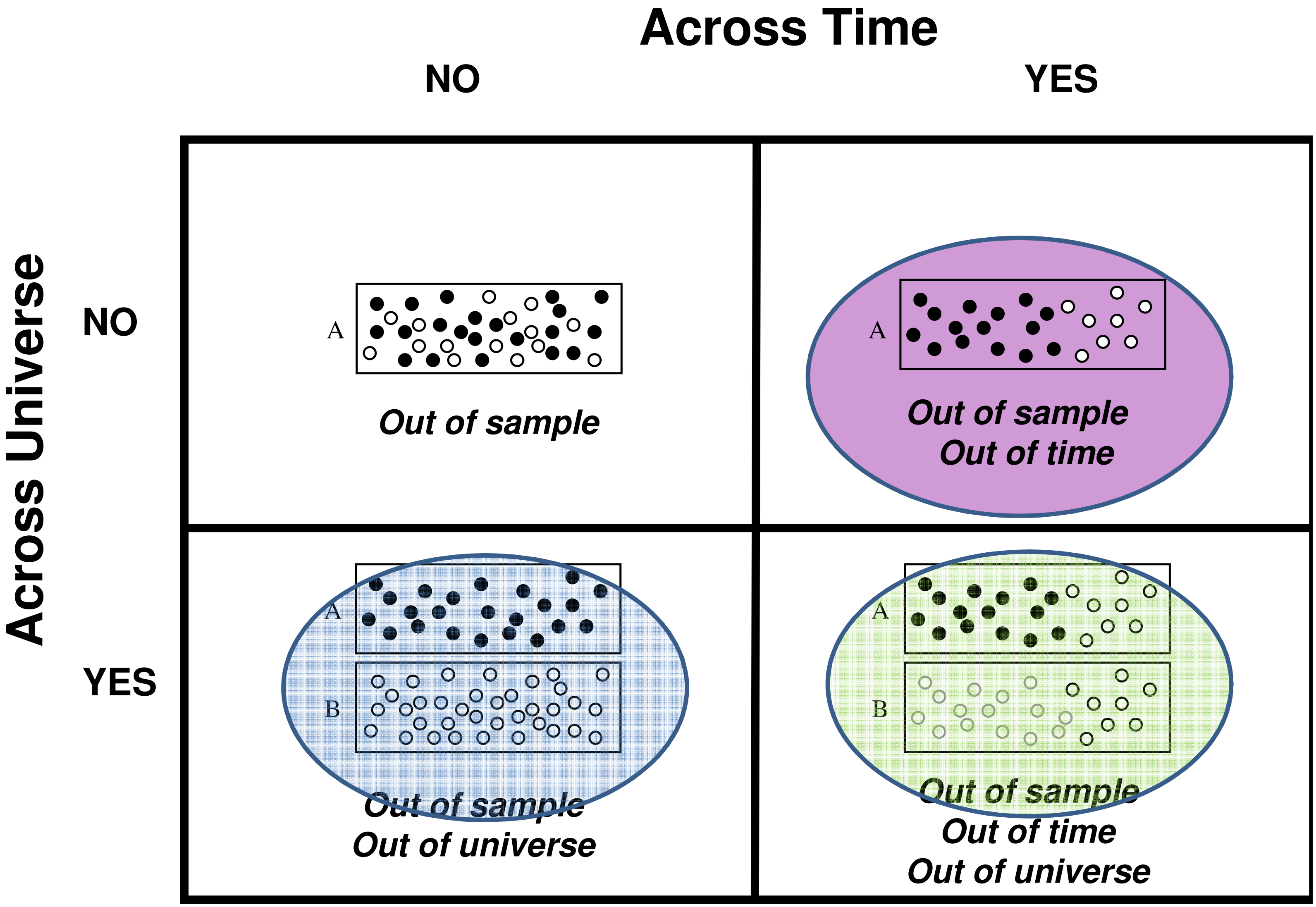
- Is more data always better?
- How much you should pay for “external” data?**

*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

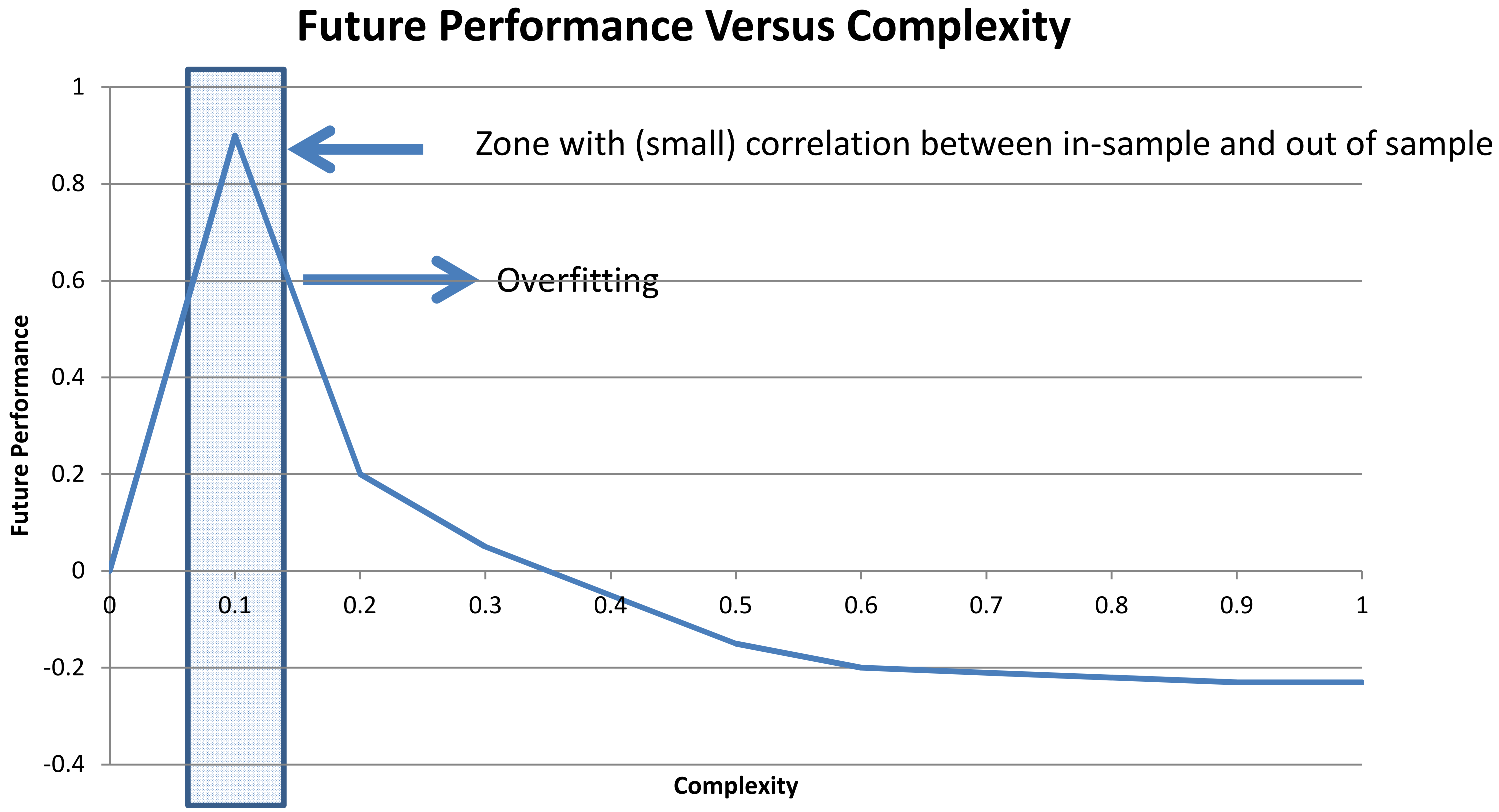
…In assessing whether new knowledge is “actionable” for decision making: **predictive power**, not just ability to **explain the past**

Validation Framework for Prediction



Across Universe \ Across Time	NO	YES
NO	*Out of sample*	*Out of sample* *Out of time*
YES	*Out of sample* *Out of universe*	*Out of sample* *Out of time* *Out of universe*

Prediction Vs Explanation: Varying Model Complexity



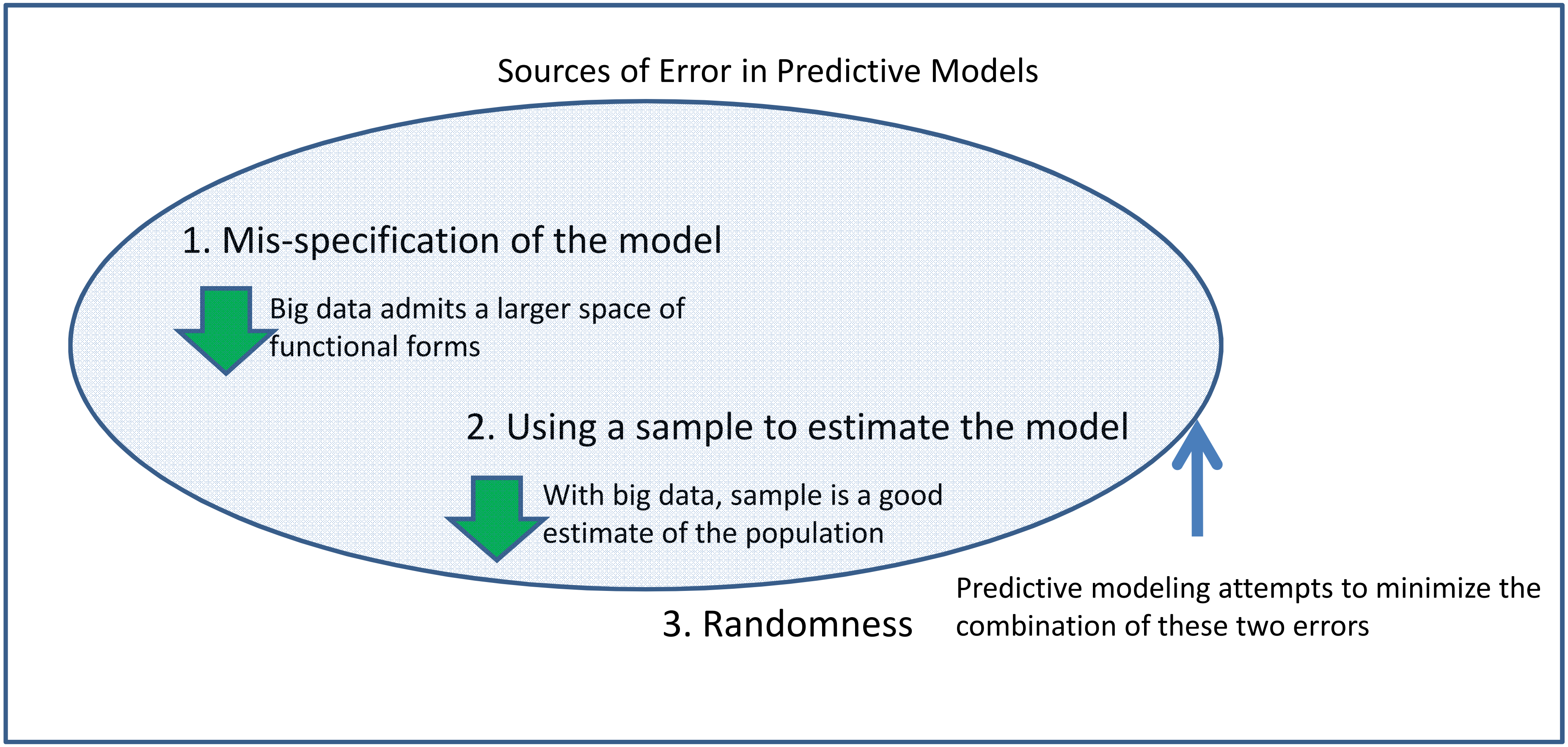
Future Performance Versus Complexity

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

Overfitting

Note: Complexity measure is illustrative only, based on the learning method used

Prediction Errors and Big Data(*)



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

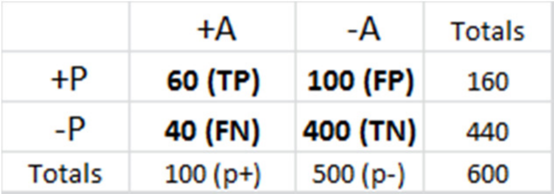
3. Randomness

Predictive modeling attempts to minimize the combination of these two errors

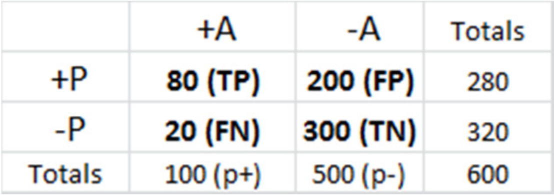
*Adapted from Shmueli, G. To explain or to predict? *Statistical Science 25*, 3 (Aug. 2010), 289–310.)

Making Predictions Usable in Noisy Domains*

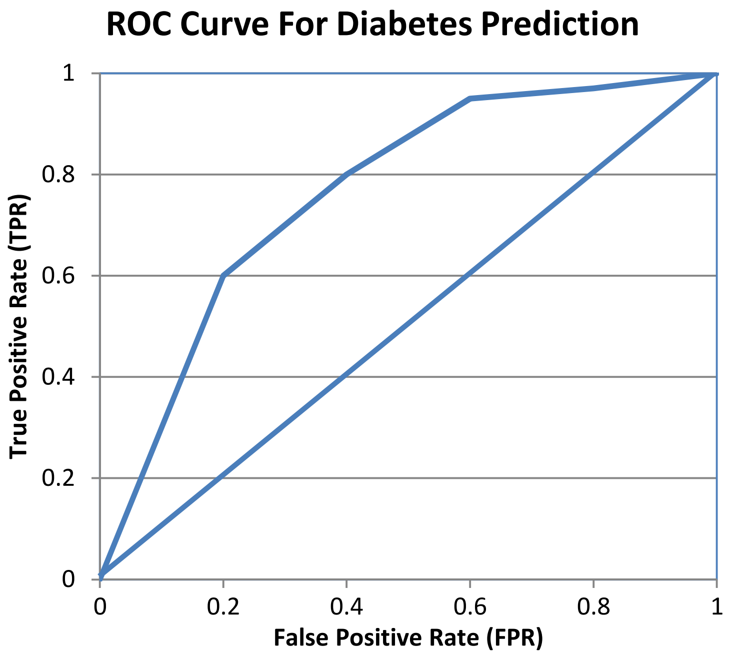
…why bother making predictions? i.e. with so many false positives?



	+A	-A	Totals
+P	**60 (TP)**	**100 (FP)**	160
-P	**40 (FN)**	**400 (TN)**	440
Totals	100 (p+)	500 (p-)	600



	+A	-A	Totals
+P	**80 (TP)**	**200 (FP)**	280
-P	**20 (FN)**	**300 (TN)**	320
Totals	100 (p+)	500 (p-)	600



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?

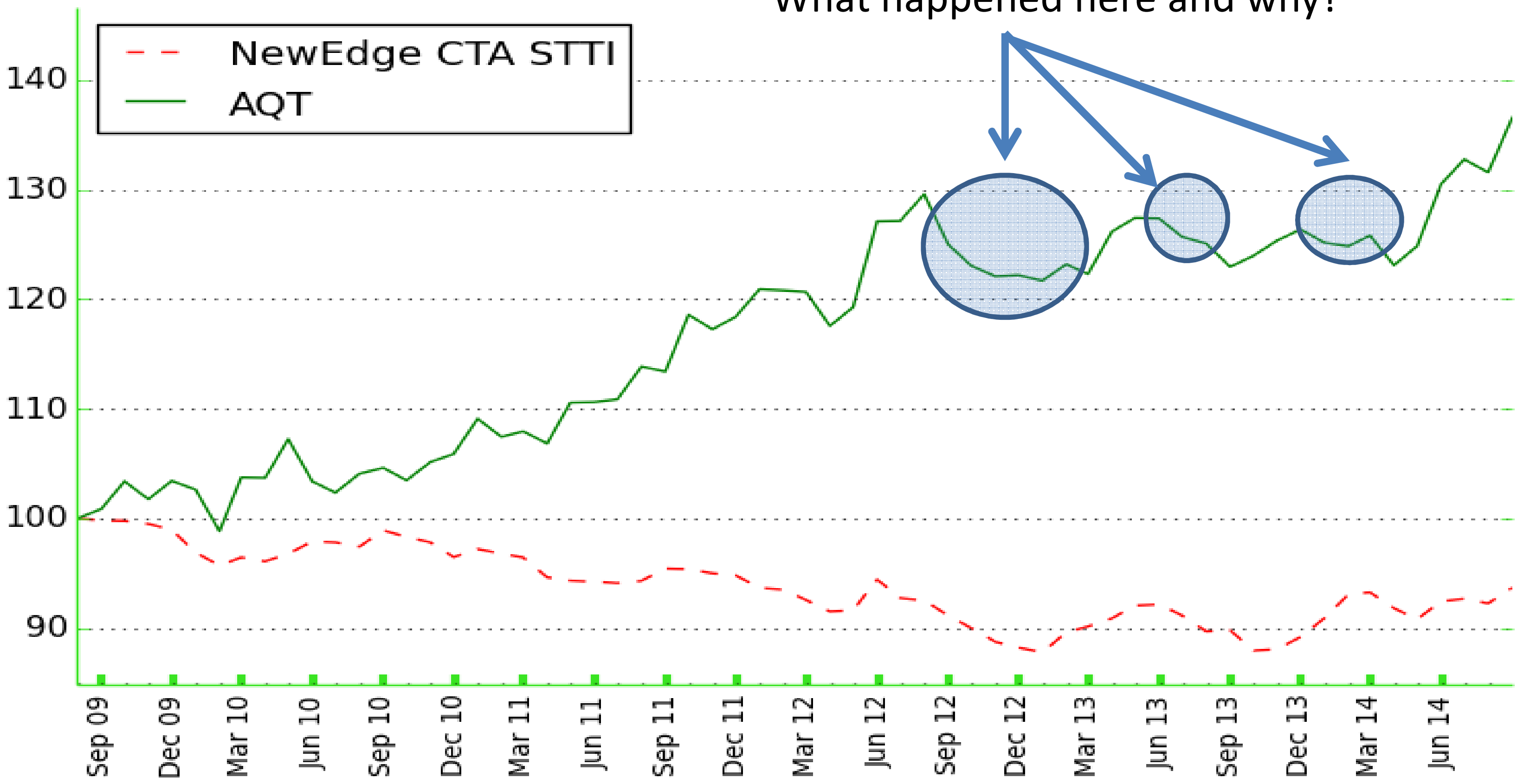
Transparency in Complex Systems

“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!