Data Science and Its Implications for Business and Society

Vasant Dhar*
Professor

Stern School of Business & NYU Center for Data Science / Editor-in-Chief, Big Data

@vasantdhar

*This talk is based on:

Dhar, V., The Future of AI. Big Data, volume 4, number 1, March 2016
Dhar, V., Should You Trust Your Money to a Robot? *Big Data,* volume 2, number 2, June 2015
Dhar, V., Data Science and Prediction, *Communications of the ACM*, volume 56, number 12, December 2013

MY PERSPECTIVE

- Professor at NYU since 1983
- Worked on Wall Street for several years in the 90s
- Created first pure machine-learning based automated hedge fund in operation in 2002 that has been on autopilot for almost 7 years
- Editor-in-Chief, Big Data journal, March 2014 to present
- Director, PhD Program in Data Science, NYU Center for Data Science
- Working with big data in Finance, Healthcare, Education, Sports

DATA SCIENCE: WHAT IS IT?

"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**

A key goal is for the decision making algorithm to "scale gracefully" and "adapt" to a changing environment in parallel with its operation

***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

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.”

Machine learning enables machines to ask the right questions for us

MY BIG RESEARCH QUESTION

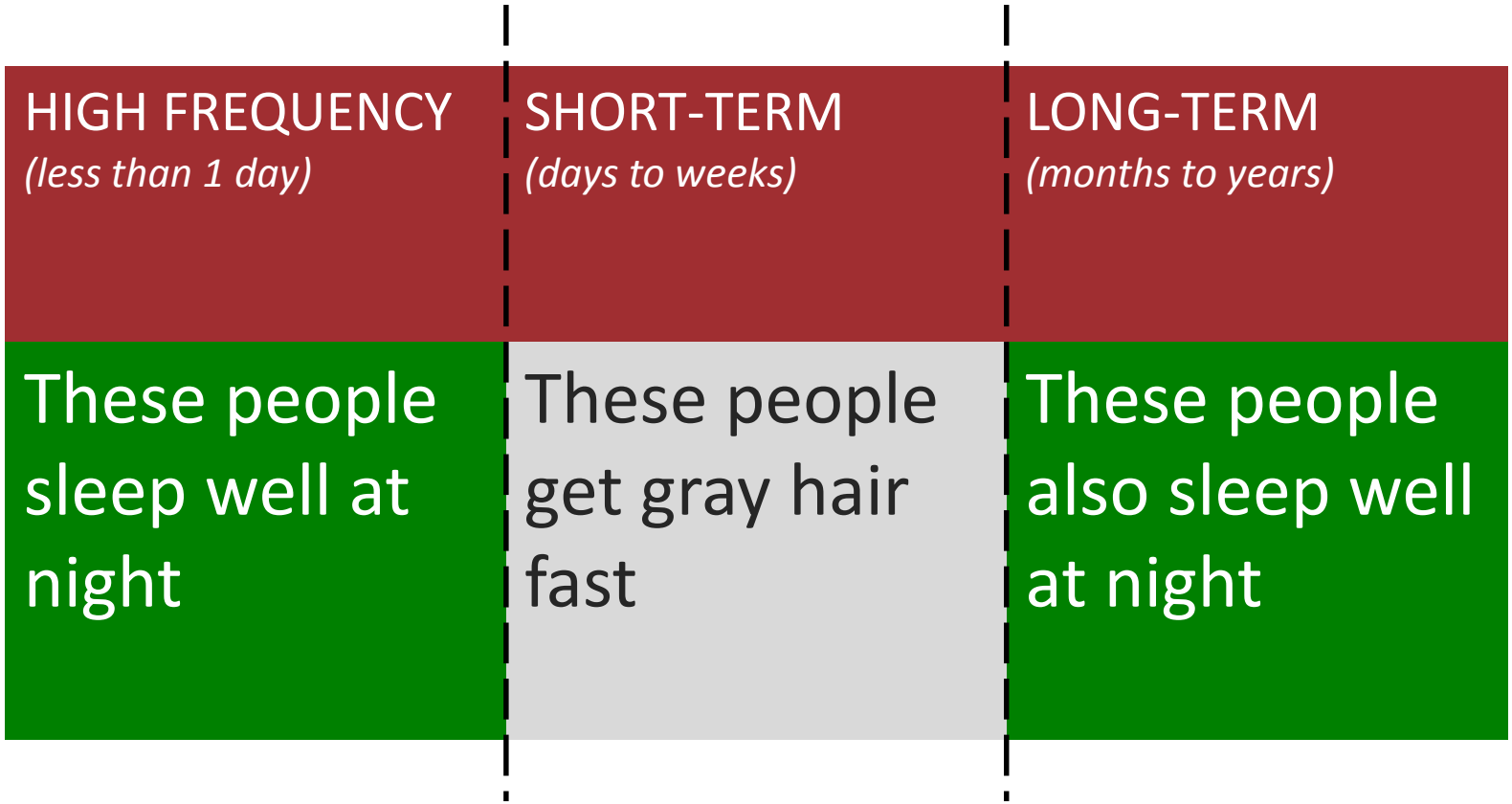
When Should We Trust Robots With Decisions?

MY RESEARCH PROJECTS

- Finance (Should You Trust Your Money to a Robot?)
- Healthcare (Should You Trust Your Healthcare to a Robot?)
- Sports (Should You Trust Your Team to a Robot?)
- Education (Should You Trust Your Child's Education to a Robot?)

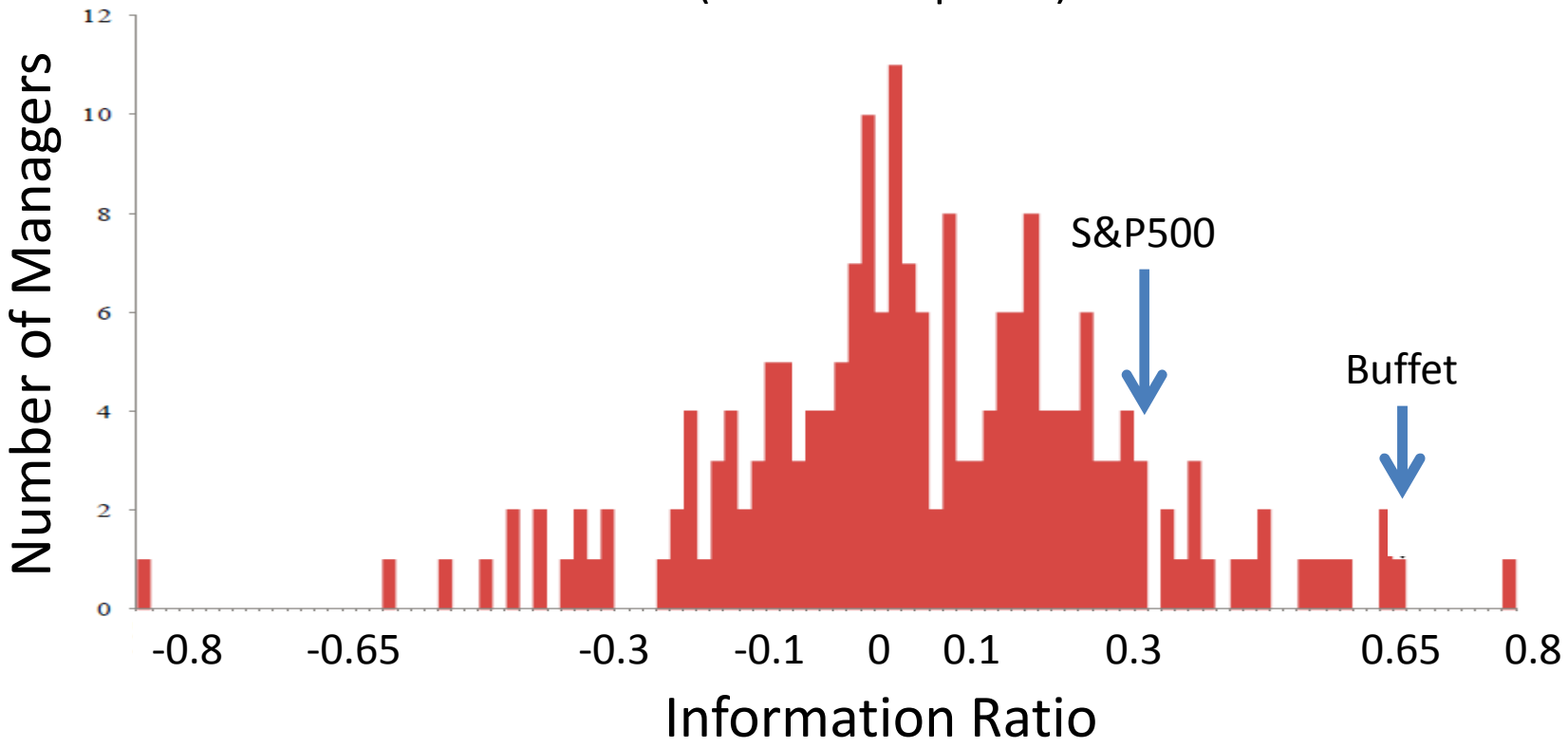
FINANCE

THE INVESTMENT LANDSCAPE



HUMANS AS INVESTORS

Performance Distribution of Managers with 30 Years of track record (survivorship bias)



It is difficult for humans to outperform the "best companies of the industry."
It takes a long time to know whether someone really has skill.

Source: Franzzini et. Al, 2010, Buffet's Alpha

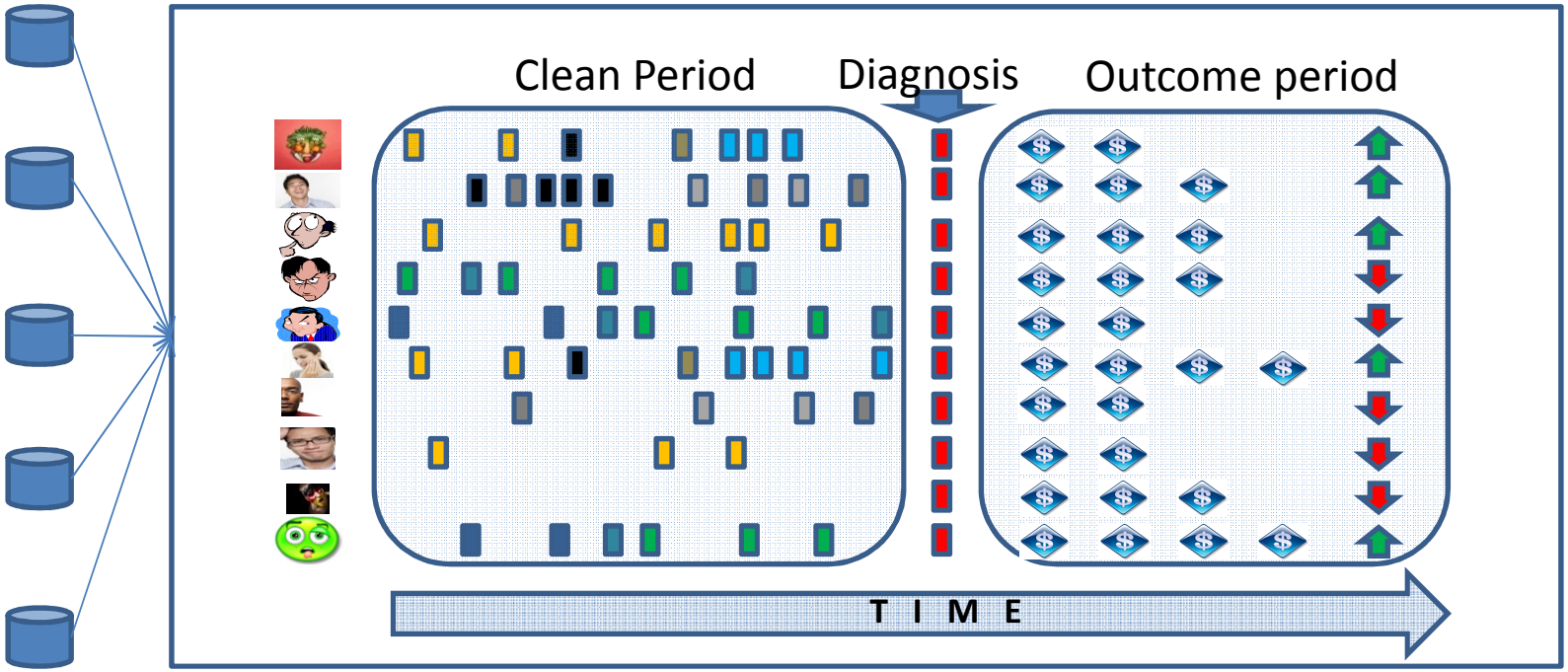
HOW WELL DO ROBOTS DO?

TOWARDS THE LEFT	IN THE MIDDLE	TOWARDS THE RIGHT
humans don't stand a chance *(read Flash Boys by Michael Lewis)*	there's a lot of interest in the middle Out of 43 mostly systematic futures trading managers on a platform run by DB with at least a 5 year track record until April 30 2015, the average Information Ratio BEFORE FEES was 0.6; two thirds fall in the range of 0.3 to 0.9*	computers don't have a good basis… but humans do poorly

*See Dhar, V., Should You Trust Your Money to a Robot? Big Data journal, Volume 2, No 2, June 2015

HEALTHCARE

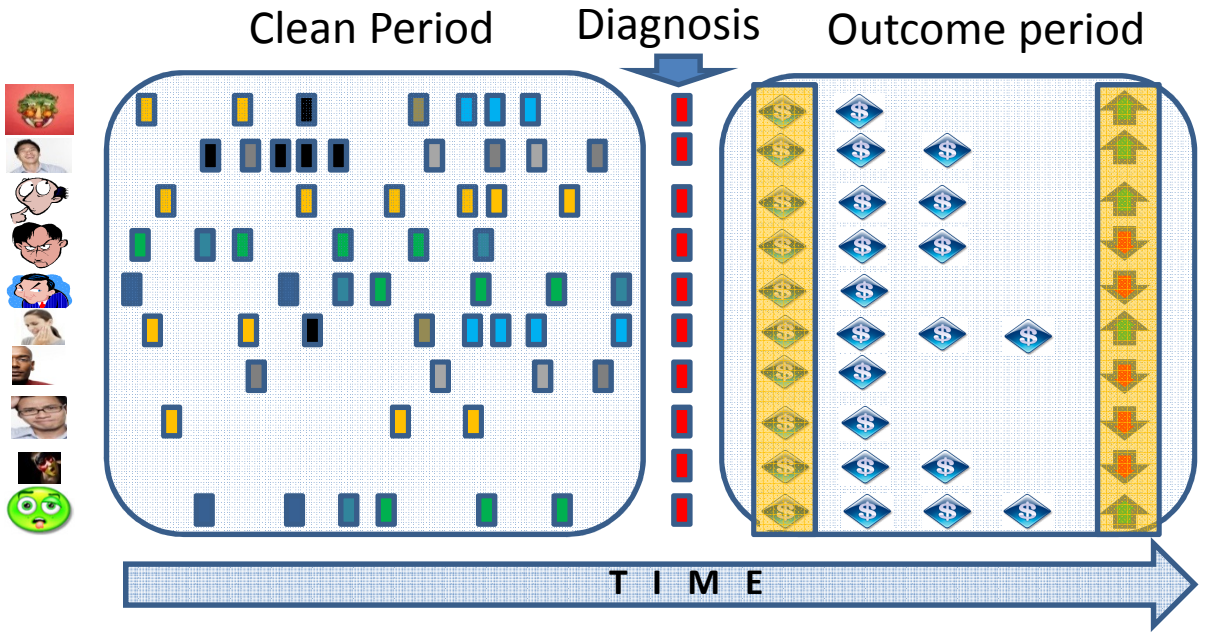
WHAT IS THE RIGHT QUESTION HERE?



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

DIABETES: PREDICT AS EARLY AS POSSIBLE



Many problems are “2 class” **classification** types: will someone develop a disease or not?

Others are **regression** types: how much will it cost, how long could it take?

DIABETES PREDICTION

Classification Errors are of 2 types: False Positives and False Negatives

Suppose we build a predictive model that results in outputs of the type below

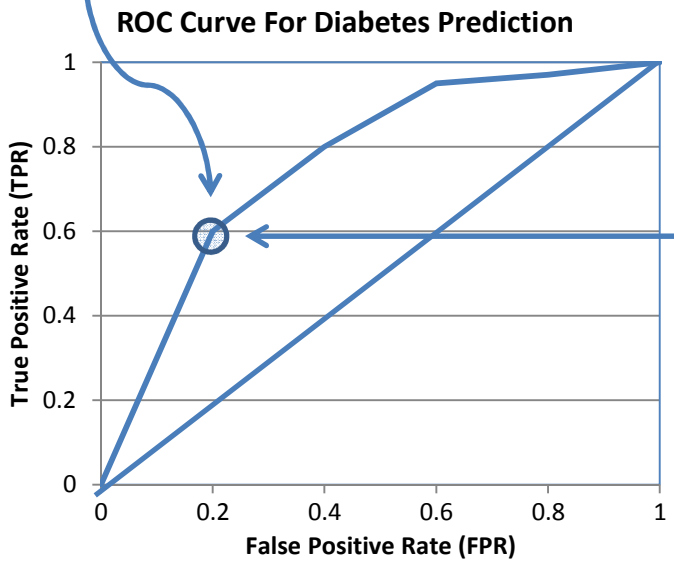
	+A	-A	Totals
+P	**60 (TP)**	**100 (FP)**	160
-P	**40 (FN)**	**400 (TN)**	440
Totals	100 (p+)	500 (p-)	600

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

Which is a better scenario?

EXPECTED VALUE FOR 2 CLASS PROBLEM*

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



p(+): probability of "+" = 100/600 = 0.1666
p(-) : probability of "-" = 500/600 = 0.8333

TPR : true positive rate = TP/(TP+FN) 60/100: 0.6
FPR: false positive rate = FP/(FP+TN) 100/500: 0.2
(1-FPR): true negative rate: 0.8
(1-TPR): false negative rate: 0.4

B(+P,+A): benefit of predicting "+" correctly
B(-P, -A): benefit of predicting "-" correctly
C(-P,+A): cost of predicting "-" incorrectly
C(+P,-A): cost of predicting "+" incorrectly

(false negatives)

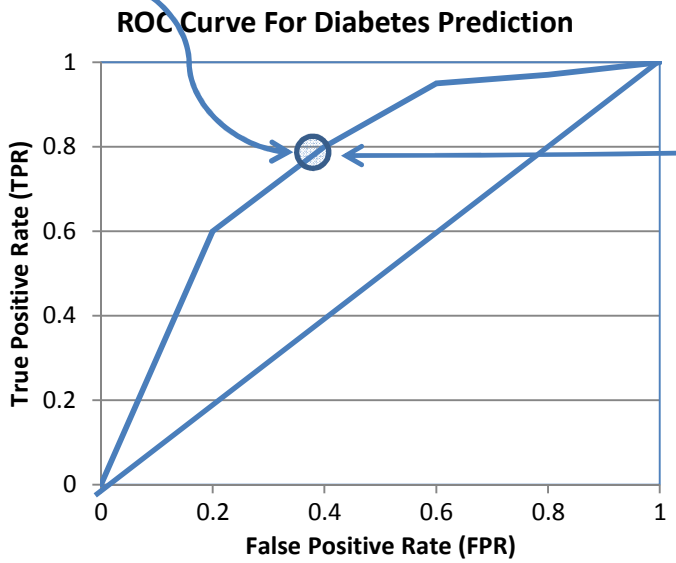
Expected Value = p(+)*[TPR***B**(+P,+A) - (1-TPR)***C**(-P,+A)] + p(-)*[(1-FPR)***B**(-P,-A) - FPR***C**(+P,-A)]

(false positives)

*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

EXPECTED VALUE FOR 2 CLASS PROBLEM

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



p(+): probability of "+" = 100/600 = 0.1666
p(-) : probability of "-" = 500/600 = 0.8333

TPR : true positive rate = TP/(TP+FN) 80/100: 0.8
FPR: false positive rate = FP/(FP+TN) 200/500: 0.4
(1-FPR): true negative rate: 0.6
(1-TPR): false negative rate: 0.2

B(+P,+A): benefit of predicting "+" correctly
B(-P, -A): benefit of predicting "-" correctly
C(-P,+A): cost of predicting "-" incorrectly
C(+P,-A): cost of predicting "+" incorrectly

Expected Value = p(+)*[TPR***B**(+P,+A) - (1-TPR)***C**(-P,+A)] + p(-)*[(1-FPR)***B**(-P,-A) - FPR***C**(+P,-A)]

HOW ARE RESULTS USED IN PRACTICE?

Differentiate the overall population into finer segments according to the risk levels predicted by the model and tune the outreach at the individual level according to the risk level

For lower risk segments, conduct lighter outreach actions such as promoting awareness and caution in the population predicted to be at risk, realizing that a majority of these cases will be false positives

For individuals predicted to be at higher risk, conduct more aggressive outreach and testing, including tracking compliance of people on medication, especially those exhibiting comorbidity

Key: machine is currently used to support not automate: could this change?

EDUCATION

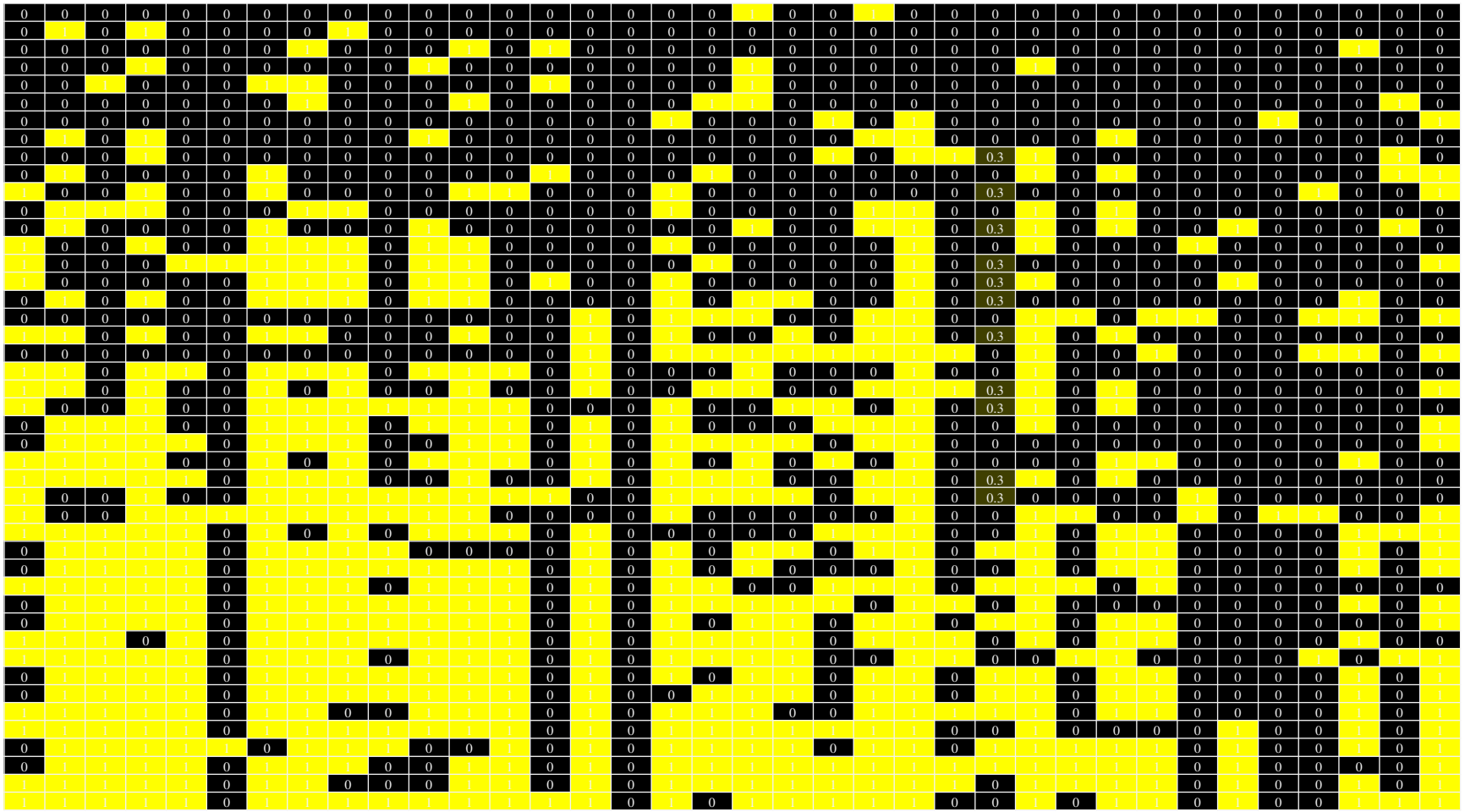
REPRESENTATION OF CONCEPTS IN PRIMARY EDUCATION

knows

Concepts

doesn't know

Students



Goal: move yellow part of the heatmap to the right

SPORTS

USE OF BIG DATA IN SPORTS

- Who should I hire given who I already have?
- What should my strategy be against Team X?
- What will happen if I rest Tim Duncan against the Golden State Warriors?
- Based on footage analysis of my last N games, what are my team's strengths and weaknesses?
- What are my strengths and weaknesses against teams of type X?

WHAT'S DIFFERENT NOW IN AI?

AI Has Come of Age in Pattern Recognition

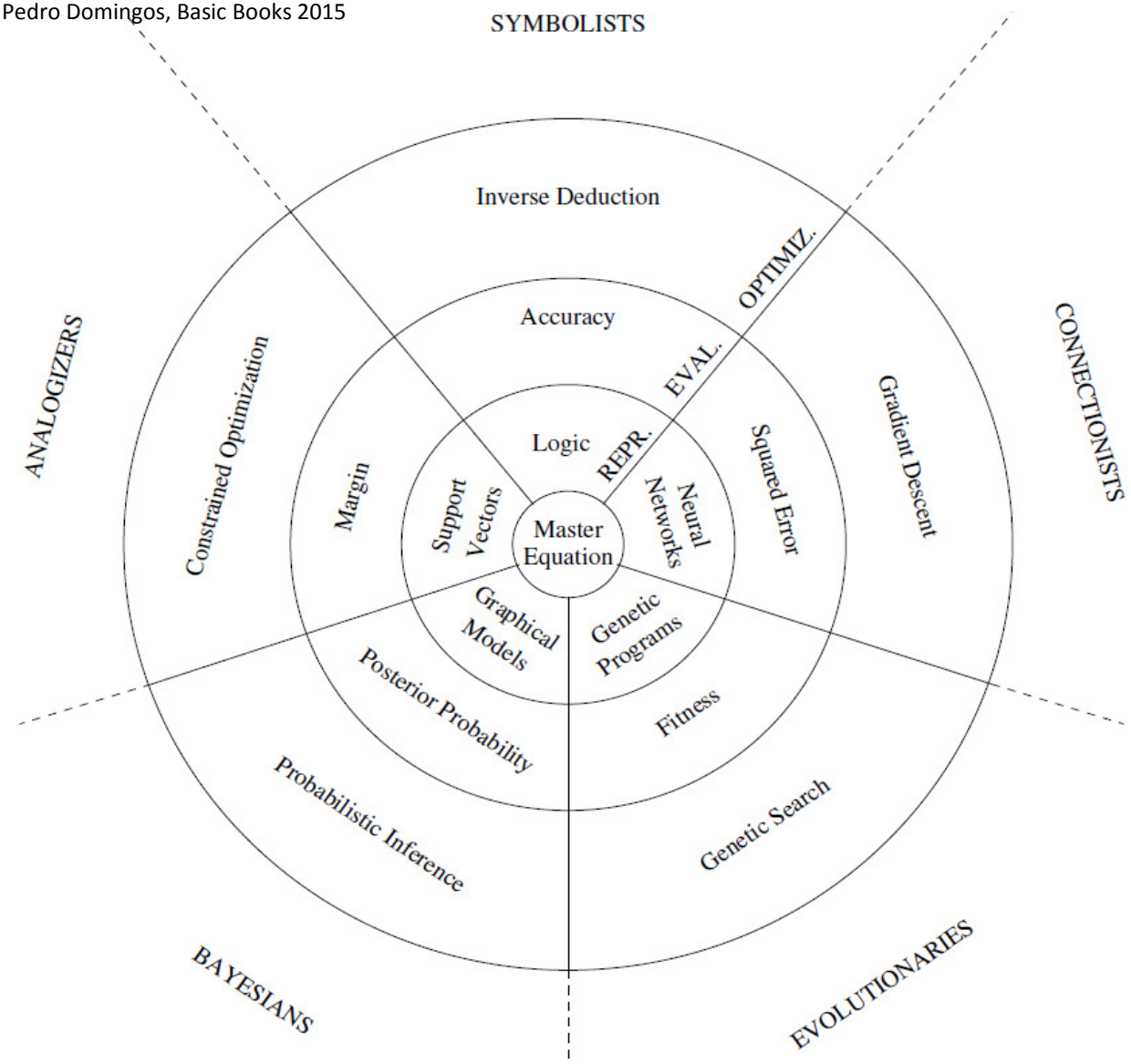
- We've come a long way in solving PERCEPTION
  - Previously we had to input the REPRESENTATION of the problem into the computer
  - Now, the machine can deal with the input directly and construct its own internal representation
    - Similar to automated "feature construction"
  - This moves the intelligence of the system "upstream"
    - Eliminates potential human bias
    - Eliminates need for labeling data/examples manually
- Solving perception has lead to some key breakthroughs in supervised and reinforcement learning

AI Has Come of Age in Pattern Discovery

- Machines are becoming increasingly capable of doing certain types of science due to big data
  - They can ask questions
  - They can test hypotheses automatically
  - They can apply epistemic criteria to determine what counts as knowledge
- This moves the intelligence “downstream” into decision making
  - Creates “actionable” knowledge
  - Creates potential for task automation
    - What factors determine whether a task can be automated?

MACHINE LEARNING PARADIGMS

From The Master Algorithm, Pedro Domingos, Basic Books 2015



CHALLENGES AND UNSOLVED PROBLEMS*

*Dhar, V., The Future of Artificial Intelligence, Big Data, volume 4, number 1, 2016

http://online.liebertpub.com/doi/full/10.1089/big.2016.29004.vda

HOW CAN/SHOULD WE CONTROL SYSTEMS THAT ARE SMARTER THAN US?

HOW CAN/SHOULD WE CONTROL SYSTEMS THAT ARE SMARTER THAN US BUT WE DON'T UNDERSTAND?

If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively…we had better be quite sure that the purpose put into the machine is the purpose we really desire and not merely a colorful imitation of it?

Norbert Wiener, Some Moral and Technical Consequences of Automation, Science, volume 131, No 3410, May 1960.

Even if we are clear about the purpose we desire we cannot be sure we will be happy with the consequences of the machine's behavior

SHOULD THERE BE AN OBJECTIVE FUNCTION FOR AI SYSTEMS?

- Optimality?
  - Could be good for perception but what about cognition?
- Should we deal with this as we go along?

WILL AI DESTROY MORE JOBS THAN IT CREATES?

WHAT IS THE FUTURE OF EMPLOYMENT?

- Many human jobs require perception as a basic requirement
- What happens when machines become facile at seeing, hearing, and reading?

IS OUR CURRENT REGULATORY SYSTEM ADEQUATE FOR DEALING WITH ROBOTS?

- The short answer is no!
- (…All nine justices who decided *DC v. Heller* agreed that the Second Amendment reads: "A well regulated militia, being necessary to the security of a free state, the right of the people to keep and bear arms, shall not be infringed." They disagreed quite sharply about what those words mean and how they relate to each other. (Legal experts even disagree on what the *commas* in the Second Amendment mean
- Does this mean that designers of AI systems will make inadvertent decisions about the interpretation of laws?

THANK YOU!