A Conversation with …
Ganesh Mani

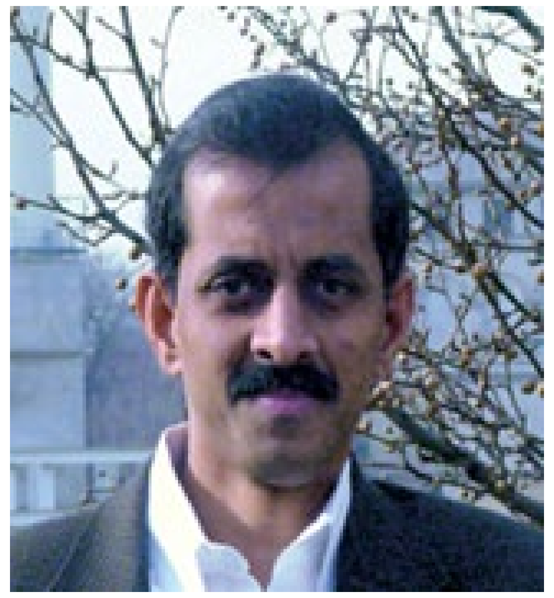
Managing the Data Supply Chain

Ganesh Mani Adj. Faculty Carnegie Mellon
Mehrzad Mahdavi, Executive Director, FDP Institute
Kathy Wilkens, Senior Advisor, FDPI Curriculum
Mirjam Dekker, Project Manager, FDP Institute

www.fdpinstitute.org
April 7, 2020

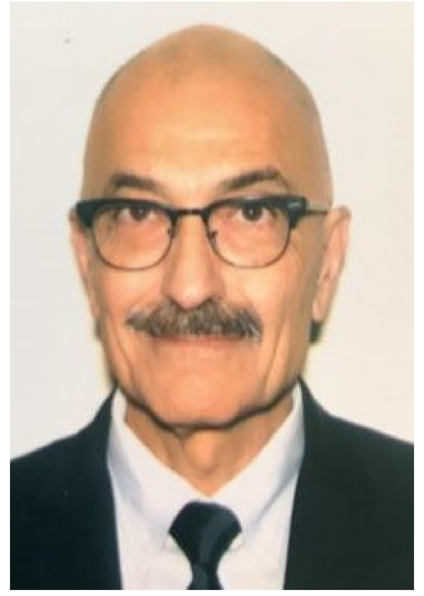
Agenda

Download the thought leadership paper here



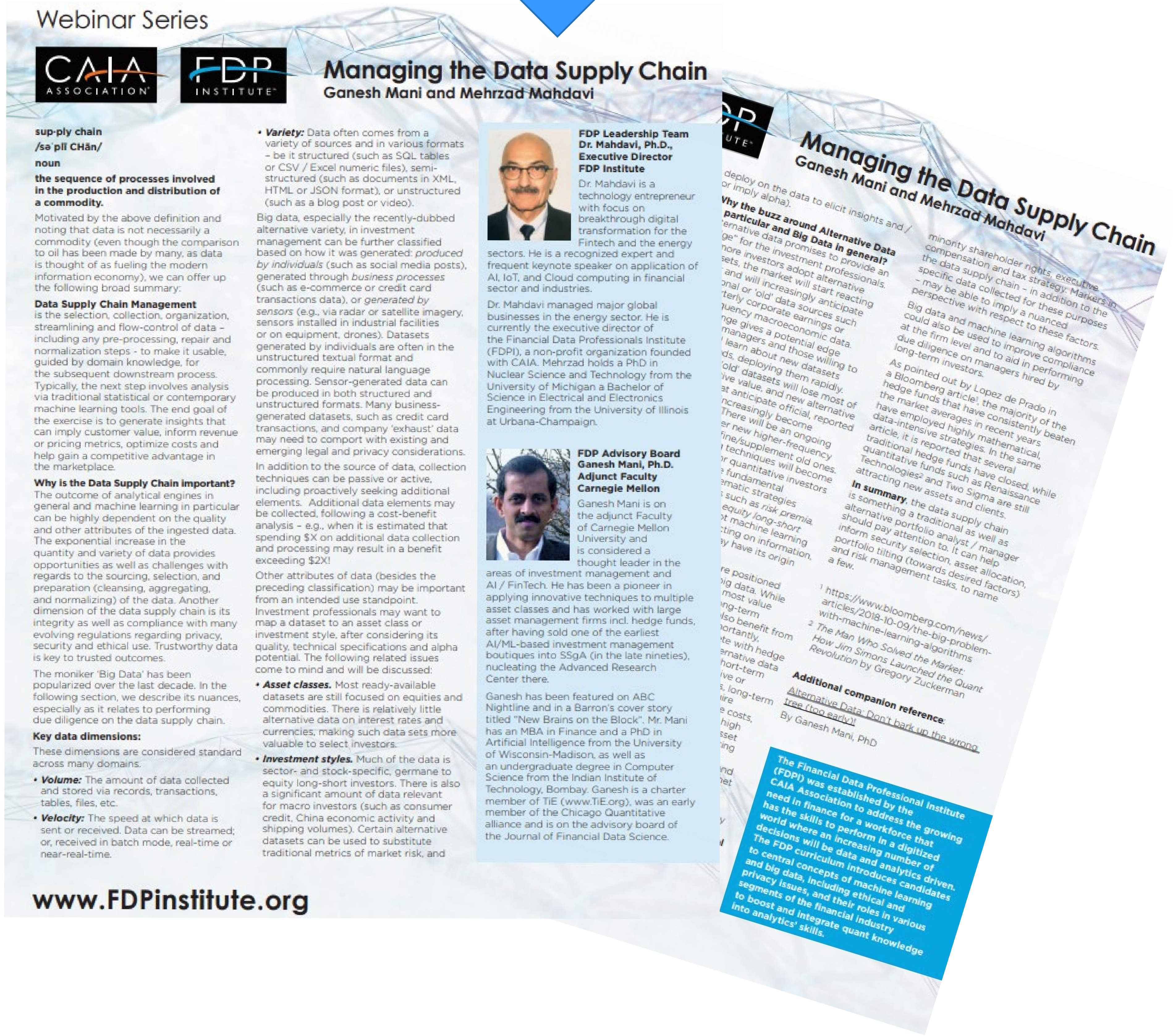
Ganesh Mani
Adj. Faculty
Carnegie
Mellon

Katherine
Wilkens
Sr. Curriculum
Advisor FDPI



Mehrzad Mahdavi
Executive Director
FDPI

Mirjam
Dekker
PM
FDPI



Webinar Series

CAIA ASSOCIATION FDP INSTITUTE

Managing the Data Supply Chain

Ganesh Mani and Mehrzad Mahdavi

sup·ply chain
/sə´plī CHān/
noun
the sequence of processes involved in the production and distribution of a commodity.

Motivated by the above definition and noting that data is not necessarily a commodity (even though the comparison to oil has been made by many, as data is thought of as fueling the modern information economy), we can offer up the following broad summary:

Data Supply Chain Management is the selection, collection, organization, streamlining and flow-control of data – including any pre-processing, repair and normalization steps - to make it usable, guided by domain knowledge, for the subsequent downstream process. Typically, the next step involves analysis via traditional statistical or contemporary machine learning tools. The end goal of the exercise is to generate insights that can imply customer value, inform revenue or pricing metrics, optimize costs and help gain a competitive advantage in the marketplace.

Why is the Data Supply Chain important? The outcome of analytical engines in general and machine learning in particular can be highly dependent on the quality and other attributes of the ingested data. The exponential increase in the quantity and variety of data provides opportunities as well as challenges with regards to the sourcing, selection, and preparation (cleansing, aggregating, and normalizing) of the data. Another dimension of the data supply chain is its integrity as well as compliance with many evolving regulations regarding privacy, security and ethical use. Trustworthy data is key to trusted outcomes.

The moniker 'Big Data' has been popularized over the last decade. In the following section, we describe its nuances, especially as it relates to performing due diligence on the data supply chain.

Key data dimensions:

These dimensions are considered standard across many domains.

- ***Volume:*** The amount of data collected and stored via records, transactions, tables, files, etc.
- ***Velocity:*** The speed at which data is sent or received. Data can be streamed; or, received in batch mode, real-time or near-real-time.
- ***Variety:*** Data often comes from a variety of sources and in various formats – be it structured (such as SQL tables or CSV / Excel numeric files), semi-structured (such as documents in XML, HTML or JSON format), or unstructured (such as a blog post or video).

Big data, especially the recently-dubbed alternative variety, in investment management can be further classified based on how it was generated: *produced by individuals* (such as social media posts), generated through *business processes* (such as e-commerce or credit card transactions data), or *generated by sensors* (e.g., via radar or satellite imagery, sensors installed in industrial facilities or on equipment, drones). Datasets generated by individuals are often in the unstructured textual format and commonly require natural language processing. Sensor-generated data can be produced in both structured and unstructured formats. Many business-generated datasets, such as credit card transactions, and company 'exhaust' data may need to comport with existing and emerging legal and privacy considerations.

In addition to the source of data, collection techniques can be passive or active, including proactively seeking additional elements. Additional data elements may be collected, following a cost-benefit analysis – e.g., when it is estimated that spending $X on additional data collection and processing may result in a benefit exceeding $2X!

Other attributes of data (besides the preceding classification) may be important from an intended use standpoint. Investment professionals may want to map a dataset to an asset class or investment style, after considering its quality, technical specifications and alpha potential. The following related issues come to mind and will be discussed:

- ***Asset classes.*** Most ready-available datasets are still focused on equities and commodities. There is relatively little alternative data on interest rates and currencies, making such data sets more valuable to select investors.
- ***Investment styles.*** Much of the data is sector- and stock-specific, germane to equity long-short investors. There is also a significant amount of data relevant for macro investors (such as consumer credit, China economic activity and shipping volumes). Certain alternative datasets can be used to substitute traditional metrics of market risk, and

FDP Leadership Team
Dr. Mahdavi, Ph.D.,
Executive Director
FDP Institute

Dr. Mahdavi is a technology entrepreneur with focus on breakthrough digital transformation for the Fintech and the energy sectors. He is a recognized expert and frequent keynote speaker on application of AI, IoT, and Cloud computing in financial sector and industries.

Dr. Mahdavi managed major global businesses in the energy sector. He is currently the executive director of the Financial Data Professionals Institute (FDPI), a non-profit organization founded with CAIA. Mehrzad holds a PhD in Nuclear Science and Technology from the University of Michigan a Bachelor of Science in Electrical and Electronics Engineering from the University of Illinois at Urbana-Champaign.

FDP Advisory Board
Ganesh Mani, Ph.D.
Adjunct Faculty
Carnegie Mellon

Ganesh Mani is on the adjunct Faculty of Carnegie Mellon University and is considered a thought leader in the areas of investment management and AI / FinTech. He has been a pioneer in applying innovative techniques to multiple asset classes and has worked with large asset management firms incl. hedge funds, after having sold one of the earliest AI/ML-based investment management boutiques into SSgA (in the late nineties), nucleating the Advanced Research Center there.

Ganesh has been featured on ABC Nightline and in a Barron's cover story titled "New Brains on the Block". Mr. Mani has an MBA in Finance and a PhD in Artificial Intelligence from the University of Wisconsin-Madison, as well as an undergraduate degree in Computer Science from the Indian Institute of Technology, Bombay. Ganesh is a charter member of TiE (www.TiE.org), was an early member of the Chicago Quantitative alliance and is on the advisory board of the Journal of Financial Data Science.

www.FDPinstitute.org

Managing the Data Supply Chain

Ganesh Mani and Mehrzad Mahdavi

minority shareholder rights, executive compensation and tax strategy. Markers in the data supply chain – in addition to the specific data collected for these purposes – may be able to imply a nuanced perspective with respect to these factors.

Big data and machine learning algorithms could also be used to improve compliance at the firm level and to aid in performing due diligence on managers hired by long-term investors.

As pointed out by Lopez de Prado in a Bloomberg article[1], the majority of the hedge funds that have consistently beaten the market averages in recent years have employed highly mathematical, data-intensive strategies. In the same article, it is reported that several traditional hedge funds have closed, while quantitative funds such as Renaissance Technologies[2] and Two Sigma are still attracting new assets and clients.

In summary, the data supply chain is something a traditional as well as alternative portfolio analyst / manager should pay attention to. It can help inform security selection, asset allocation, portfolio tilting (towards desired factors) and risk management tasks, to name a few.

[1] https://www.bloomberg.com/news/articles/2018-10-09/the-big-problem-with-machine-learning-algorithms

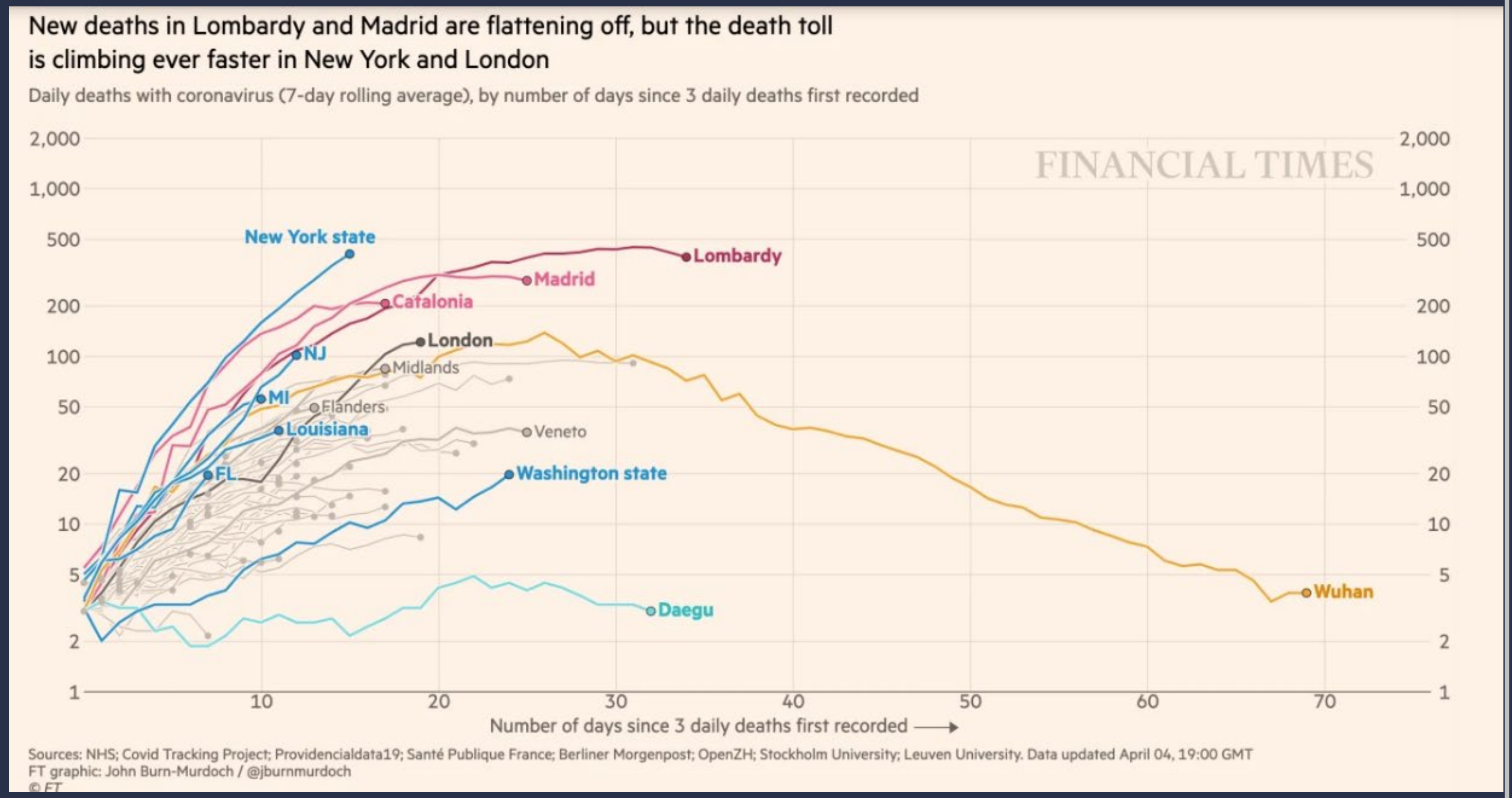
[2] *The Man Who Solved the Market: How Jim Simons Launched the Quant Revolution* by Gregory Zuckerman

Additional companion reference:
Alternative Data: Don't bark up the wrong tree (too early)!
By Ganesh Mani, PhD

The Financial Data Professional Institute (FDPI) was established by the CAIA Association to address the growing need in finance for a workforce that has the skills to perform in a digitized world where an increasing number of decisions will be data and analytics driven. The FDP curriculum introduces candidates to central concepts of machine learning and big data, including ethical and privacy issues, and their roles in various segments of the financial industry to boost and integrate quant knowledge into analytics' skills.

↑ Previous

By Region View (as of late April 4th)

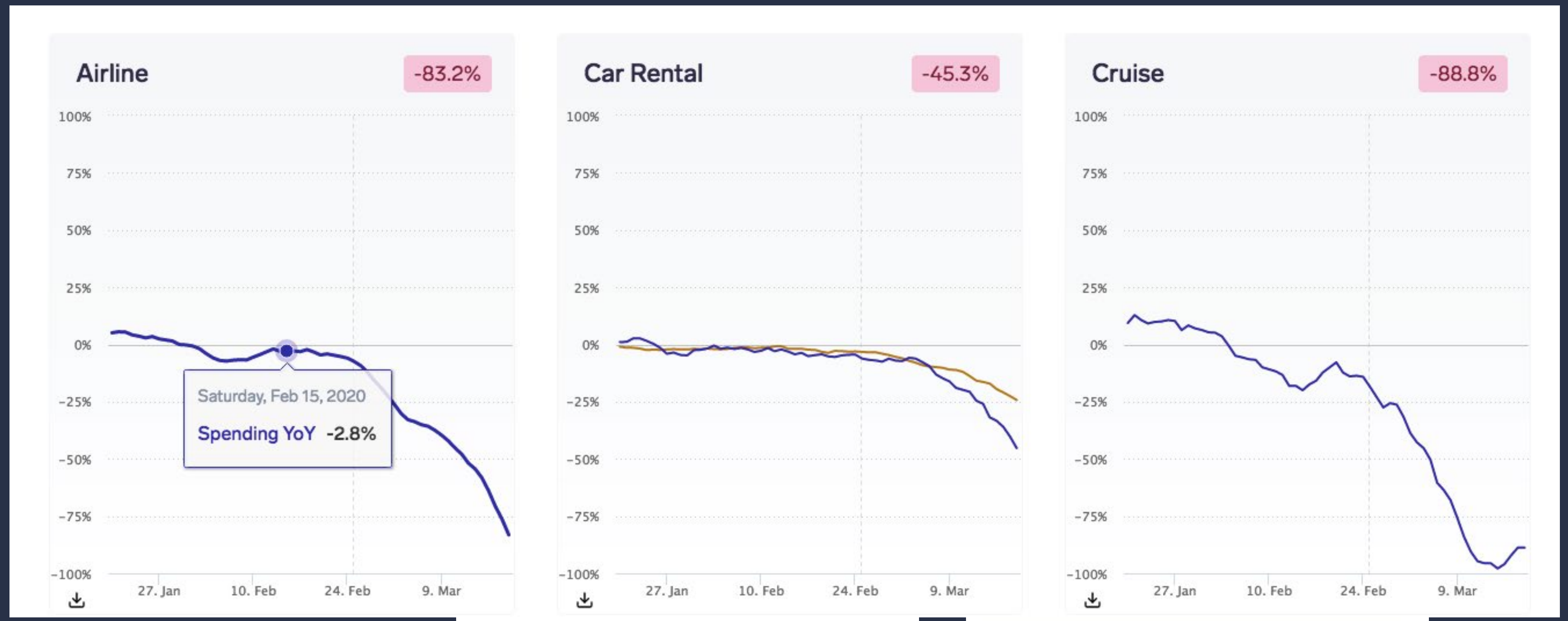


New deaths in Lombardy and Madrid are flattening off, but the death toll is climbing ever faster in New York and London

Daily deaths with coronavirus (7-day rolling average), by number of days since 3 daily deaths first recorded

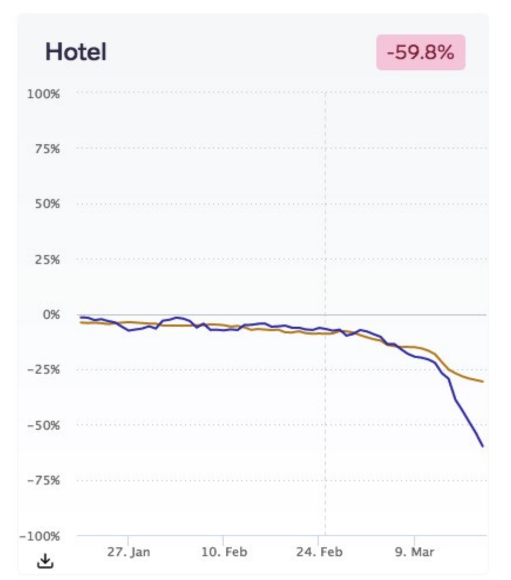
Sources: NHS; Covid Tracking Project; Providencialdata19; Santé Publique France; Berliner Morgenpost; OpenZH; Stockholm University; Leuven University. Data updated April 04, 19:00 GMT
FT graphic: John Burn-Murdoch / @jburnmurdoch
© FT

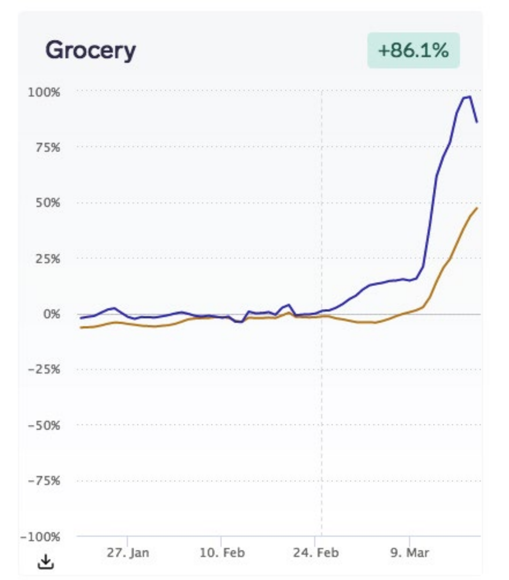
Source: Financial Times



US Demand: Travel vs. Grocery

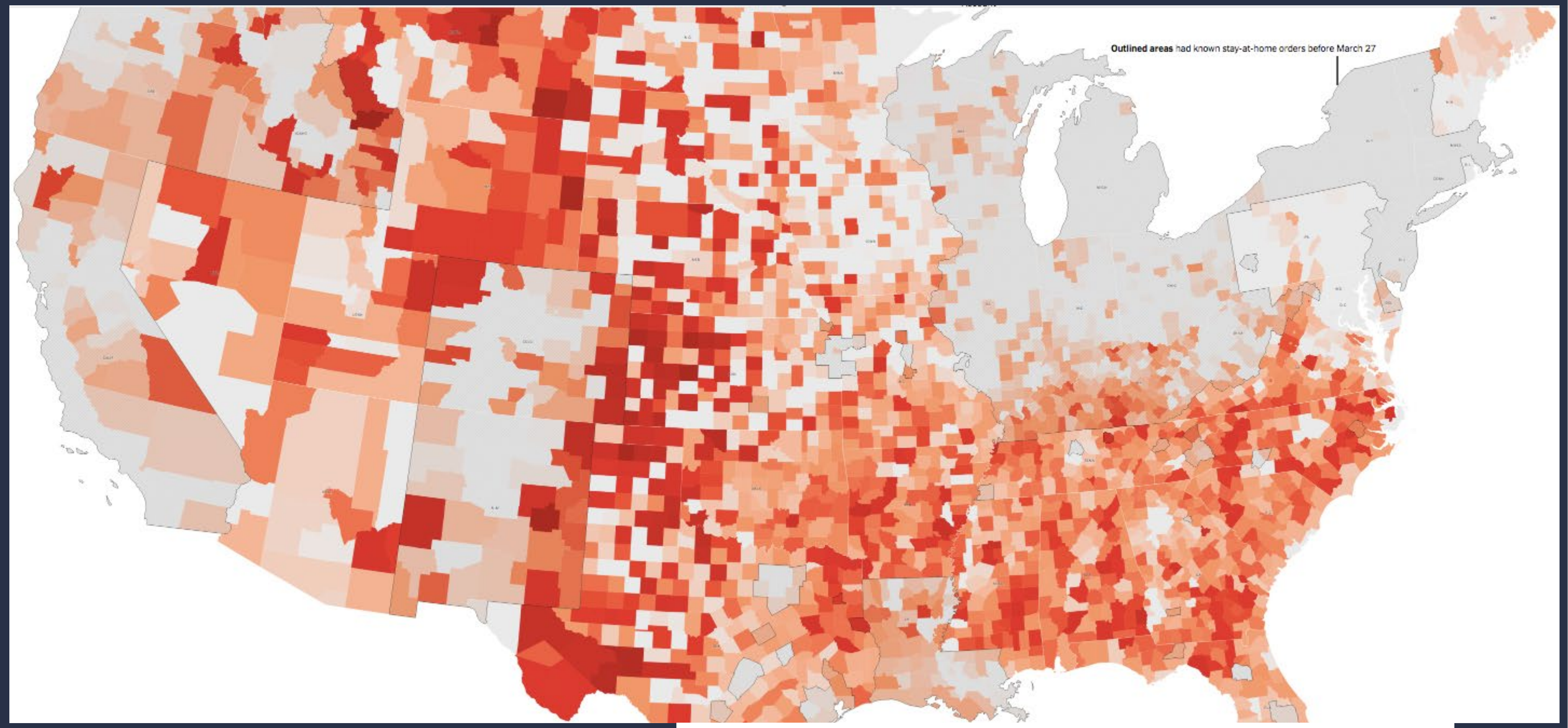
(Late Jan – Early Mar)





Source: Exabel

Social Distancing Compliance


Outlined areas had known stay-at-home orders before March 27

Where people were still traveling

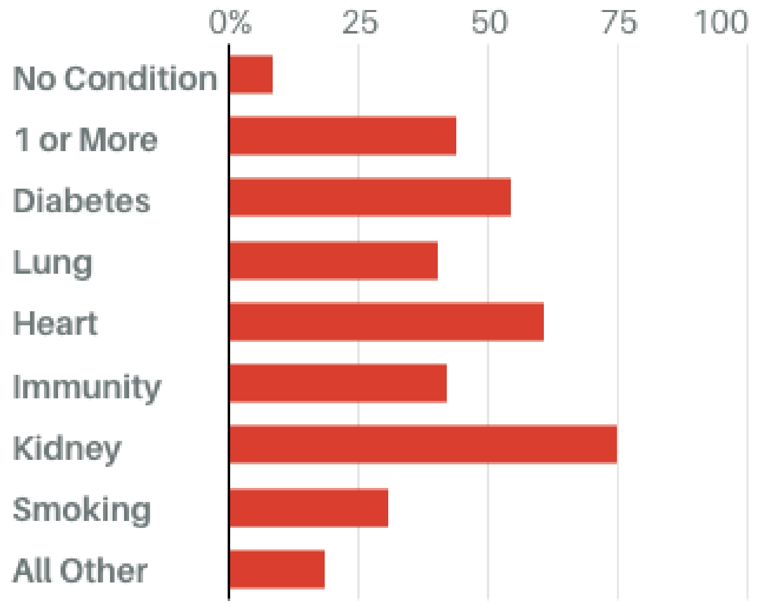
Percent change in average travel for the week of March 23, compared with travel before the coronavirus outbreak.

Source: NYTimes & CubeIq

Health

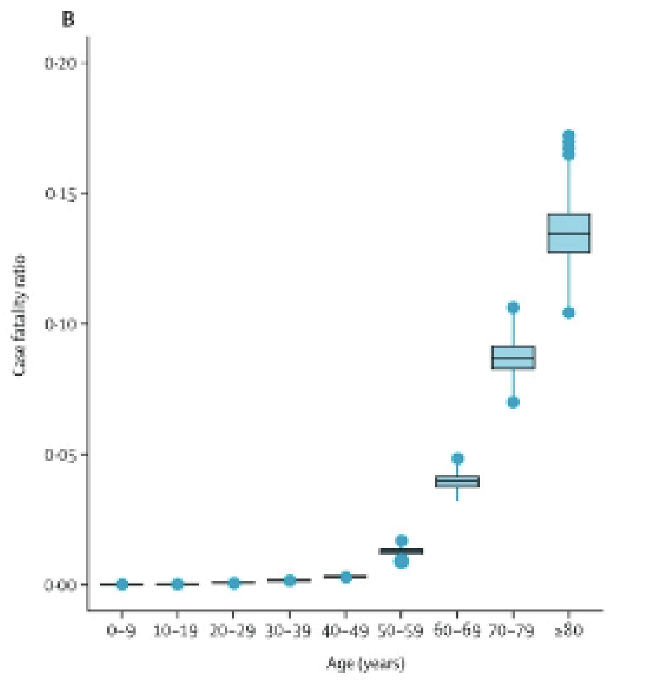
Age category (binary), years				
	<60	194	30 763	0·631% (0·545–0·726)
	≥60	829	13 909	5·96% (5·57–6·37)

CDC Hospitalization Data By Underlying Health Conditions



Note: From April 3 update. Data through March 38. Study based on 7,162 cases with completed information.
Source: Centers for Disease Control and Prevention

Source: Barron's



Chinese fatality data (by age)

Source: (Mar 30, 2020) Lancet, Verity et al.

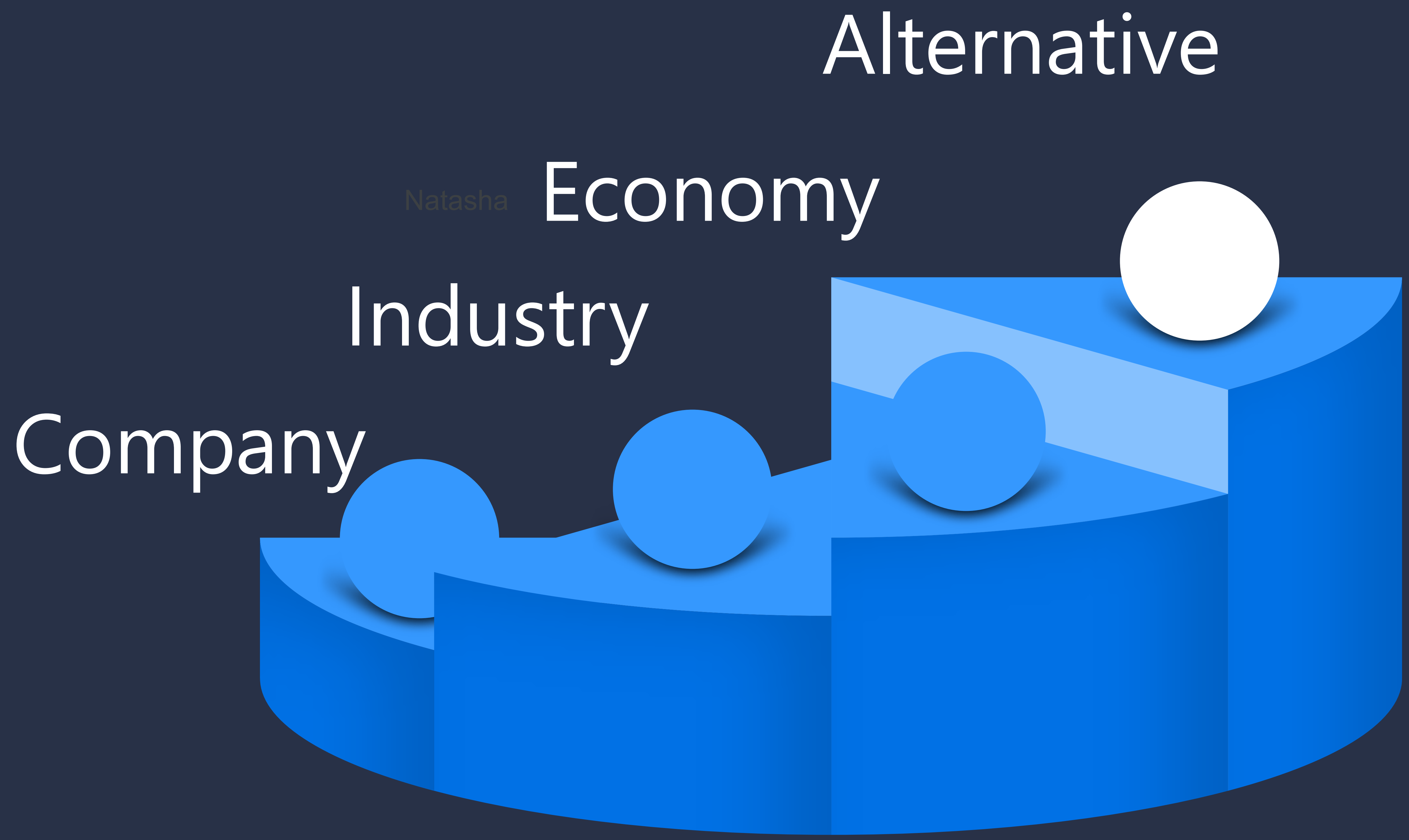
Next ↓

Carnegie Mellon University

Ganesh Mani, PhD, MBA
Adjunct Faculty

Ganesh has designed and developed many innovative alpha-generation frameworks based on disparate data (both traditional and alternative) and AI/ML techniques. Clients include leading asset management firms (incl. hedge funds) and plan sponsors.

AI / ML:
Useful for Analysis as well as for Data scrubbing / repair.

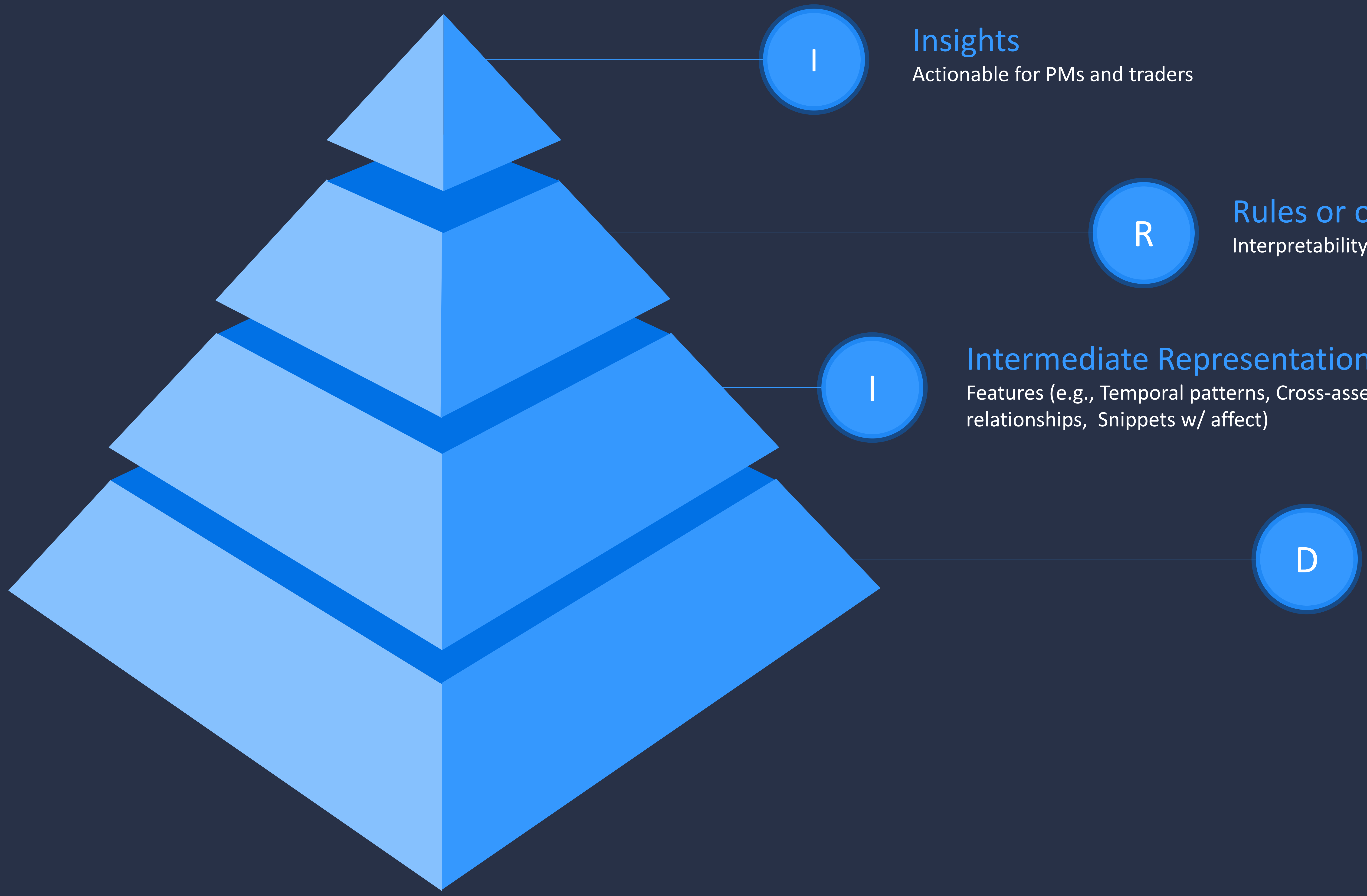


Previous
Next

Deck Template: Thanks to YouExec

Reduce Complexity

AI / ML provides a rich toolbox



Insights

Actionable for PMs and traders

Rules or other summaries

Interpretability trade-offs

Intermediate Representations

Features (e.g., Temporal patterns, Cross-asset relationships, Snippets w/ affect)

Data

Large amounts often w/ noise

Diving for insights

Making information actionable

Data

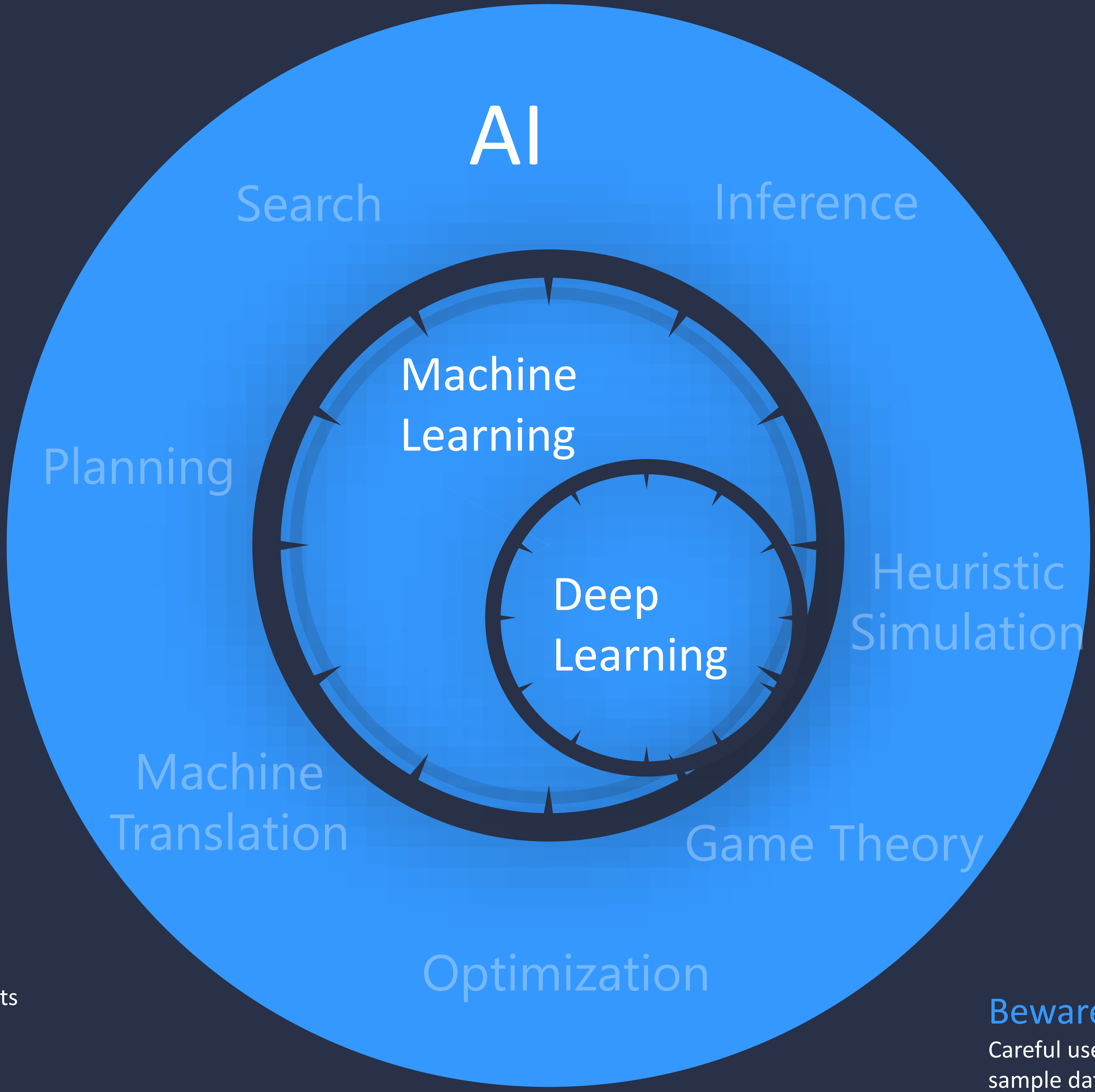
Volume, Velocity & Variety
Provenance, Vintage & Noise-level
Traditional vs. Alternative

Augmented Intelligence

Man with bicycle analogy

Opinion vs. Fact

AI / ML provides strong opinions or hints



Models

Choice should be dictated by data characteristics and goals

Beware of overfitting

Careful use of cross-validation and out-of-sample data

↑ Previous

Next ↓

Test and Deploy

Drift?

Monitor errors and type!

Is new data similar to training data?

Validate

Carefully

Time-series data may pose additional challenges

Retrain

How often?

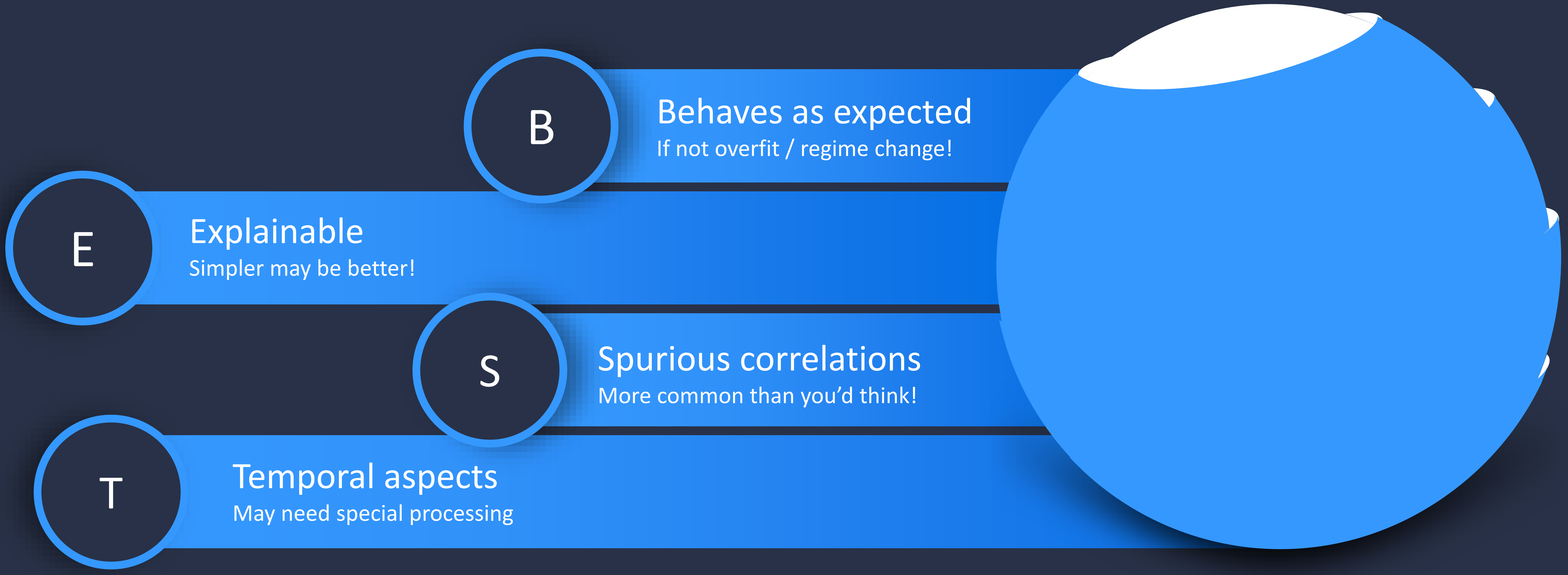
Many factors (e.g., lookback) go into this decision

↑ Previous

Next ↓

↑ Previous

Multiple models / solutions possible



B

Behaves as expected

If not overfit / regime change!

E

Explainable

Simpler may be better!

S

Spurious correlations

More common than you'd think!

T

Temporal aspects

May need special processing

Try to be creative!

Whether you use a simpler technique or a more complex one, try to expose insights.

Next ↓

Kind reminders of upcoming webinars as we go through the Q & A.

Q & A

CAIA ASSOCIATION FDP INSTITUTE

WEBINAR SERIES

A Conversation With…

Rick Roche, CAIA

Man. Dir. Little Harbor Advisors

“Evolution of Machine Learning in Investment Mgmt.”

April 22, 2020

1 pm EST

CAIA ASSOCIATION FDP INSTITUTE

WEBINAR SERIES

A Conversation With…

Imir Arifi

Trusted AI Evangelist

@DataRobot

Machine learning Models in Credit Risk Analysis

April 29th – 1pm EDT

www.fdpinstitute.org/webinars

In Closing

- New curriculum available at the end of April
- Registration for the October 26 – November 8th exam opens May 10th
- For a recent candidate webinar go to www.fdpinstitute.org/webinars

Learn more about the FDP Institute at
www.fdpinstitute.org