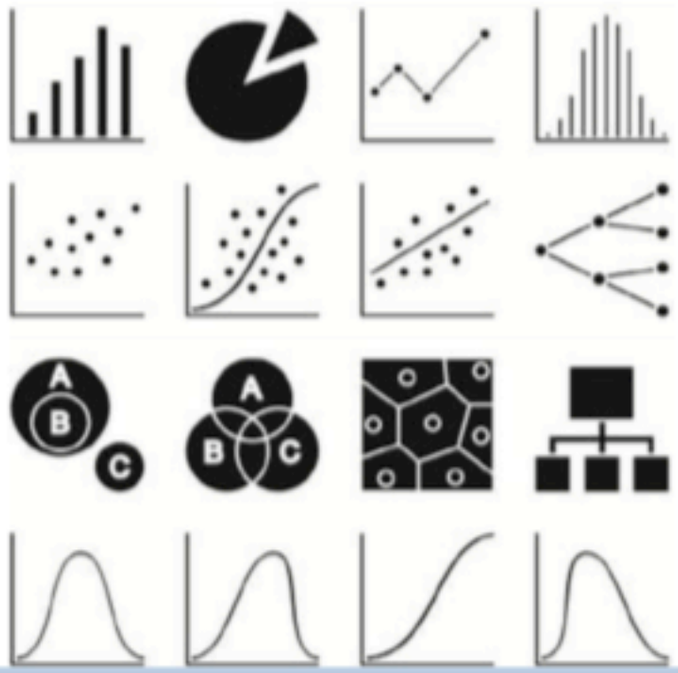
J.P.Morgan

May 2017



Big Data and AI Strategies

Machine Learning and Alternative Data Approach to Investing

I: INTRODUCTION AND OVERVIEW

Junho Song

I: INTRODUCTION AND OVERVIEW

Table of Contents

I: INTRODUCTION AND OVERVIEW 6

Summary 7
Introduction to Big Data and Machine Learning 9
Classification of Alternative Data Sets 12
Classification of Machine Learning Techniques 16
Positioning within the Big Data Landscape 21

II: BIG AND ALTERNATIVE DATA 25

Overview of Alternative Data 26
Data from Individual Activity 30
Data from Business Processes 38
Data from Sensors 42

III: MACHINE LEARNING METHODS 51

Overview of Machine Learning Methods 52
Supervised Learning: Regressions 57
Supervised Learning: Classifications 77
Unsupervised Learning: Clustering and Factor Analyses 93
Deep and Reinforcement Learning 102
Comparison of Machine Learning Algorithms 117

IV: HANDBOOK OF ALTERNATIVE DATA 135

Table of contents of data providers 136
A. Data from Individual Activity 137
B. Data from Business Processes 147
C. Data from Sensors 176
D. Data Aggregators 189
E. Technology Solutions 191

APPENDIX 214

Techniques for Data Collection from Websites 215
Packages and Codes for Machine Learning 226
Mathematical Appendices 231
References 254
Glossary 270

May, 2017

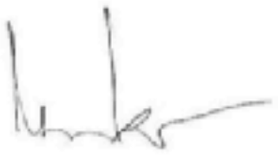
Dear Investor,

Over the past few years, we have witnessed profound changes in the marketplace with participants increasingly adopting quantitative investing techniques. These include Risk Premia investing, algorithmic trading, merging of fundamental and quantitative investment styles, consumption of increasing amounts and differentiated types of data, and adoption of new methods of analysis such as those based on Machine Learning and Artificial Intelligence.

In fact, over the past year, the exponential increase of the amount and types of data available to investors prompted some to completely change their business strategy and adopt a 'Big Data' investment framework. Other investors may be unsure on how to assess the relevance of Big Data and Machine Learning, how much to invest in it, and many are still paralyzed in the face of what is also called the 'Fourth Industrial Revolution'.

In this report we aim to provide a framework for Machine Learning and Big Data investing. This includes an overview of types of alternative data, and Machine Learning methods to analyze them. Datasets are at the core of any trading strategy. For this reason, we first **classify and analyze the types of alternative datasets.** We assess the relevance of various datasets for different types of investors and illustrate the use of Big Data in trading strategies. Datasets covered include **data generated by individuals** (e.g. social media), **data generated by business processes** (e.g. commercial transactions) and **data generated by machines** (e.g. satellite image data). After focusing on Datasets, we explain and **evaluate different Machine Learning methods** which are necessary tools to analyze Big Data. These methods include **Supervised Machine Learning**: regressions, classifications; **Unsupervised Machine Learning**: clustering, factor analyses; as well as methods of **Deep and Reinforcement Learning**. We provide theoretical, practical (e.g. codes) and investment examples for different Machine Learning methods, and compare their relative performance. The last part of the report **is a handbook of over 500 alternative data and technology providers,** which can be used as a rough roadmap to the Big Data and Artificial Intelligence landscape.

We hope this guide will be educative for investors new to the concept of Big Data and Machine Learning, and provide new insights and perspectives to those who already practice it.



Marko Kolanovic, PhD
Global Head of Macro Quantitative and Derivatives Strategy
J.P.Morgan Securities LLC

…the exponential increase of the amount and types of data available to investors prompted some to completely change their business strategy and adopt a **'Big Data' investment framework**….

Datasets covered include **data generated by individuals** (e.g. social media), **data generated by business processes** (e.g. commercial transactions) and **data generated by machines** (e.g. satellite image data).

The last part of the report **is a handbook of over 500 alternative data and technology providers,** which can be used as a rough roadmap to the Big Data and Artificial Intelligence landscape.

Summary

Big Data and Machine Learning ‘revolution’:

- Real time : number of customers visiting, transactions, **agricultural yields, activity of oil rigs**

- useful data are not readily available and one needs to purchase, **organize and analyze alternative datasets** in order to extract tradeable signals.

- Machine Learning techniques requires some theoretical knowledge and **a lot of practical experience** in designing quantitative strategies.

Datasets and Methodologies:

- **acquiring, understanding AND technologies, methods**

- **data generated by individuals** (social media posts, product reviews, search trends, etc.),

- **data generated by business processes** (company exhaust data, commercial transaction, credit card data, etc.)

- **data generated by sensors** (satellite image data, foot and car traffic, ship locations, etc.).

<== these datasets need a level of analysis before they can be used in a trading strategy.

- **different asset classes**

- **different investment styles** (e.g. macro, equity long-short, etc.).

Summary

Fear of Big Data and Artificial Intelligence:

- Strategies based on Machine Learning and Big Data also require **market intuition**, **understanding** of economic drivers **behind data**, and **experience** in designing tradeable strategies.

How will Big Data and Machine Learning change the investment landscape?

- **We think the change will be profound.**

- Eventually, 'old' datasets will lose most predictive value and **new datasets** that capture 'Big Data' will **increasingly become standardized.**

(old data : quarterly corporate earnings, low frequency macroeconomic data, etc.)

- Machine Learning techniques will **become a standard tool** for quantitative investors and perhaps some fundamental investors too. (risk premia, trend followers, equity long-short quants, etc.)

- datasets that have **high Sharpe ratio signals** (viable as a standalone funds) will **disappear.**

- the **bulk of Big Data signals** will not be viable as stand-alone strategies, but will still be very valuable in the **context of a quantitative portfolio**.

Potential Pitfalls of Big Data and Machine Learning:

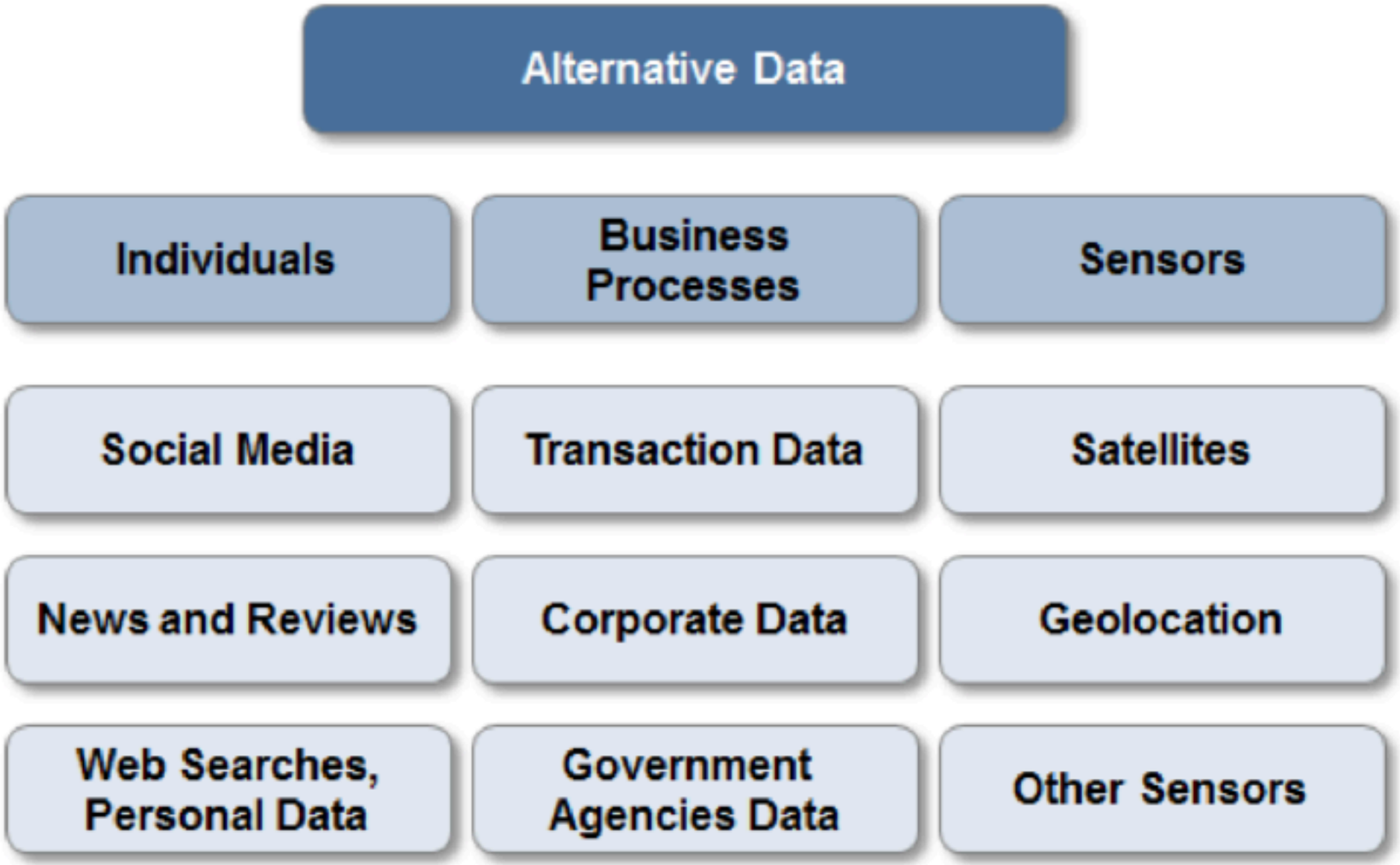
- **Certain types of data** may lead into blind alleys (datasets that **don't contain alpha, signals** that have too little investment capacity, decay quickly, or are simply too expensive to purchase)
- **Managers** may invest ; **complex models and architecture** that don't justify marginal performance improvements.
- **Machine Learning algorithms** cannot entirely replace human intuition. (overfit, underfoot)
- it is more important to understand the **economics behind data and signals**, than to be able to develop complex technological solutions. (**not lead to viable trading strategies.**)

Roles of Humans and Machines:

- for **short term trading,** such as high frequency market making, humans already play a very small role.
- On a **medium term** investment… Machines have the ability to quickly analyze news feeds and tweets, process earnings statements, scrape websites, and trade on these instantaneously. These strategies are already eroding the advantage of **fundamental analysts, equity long-short managers and macro investors.**
- On a **long term** horizon, machines **will likely not be able to compete with strong macro and fundamental human investors**.

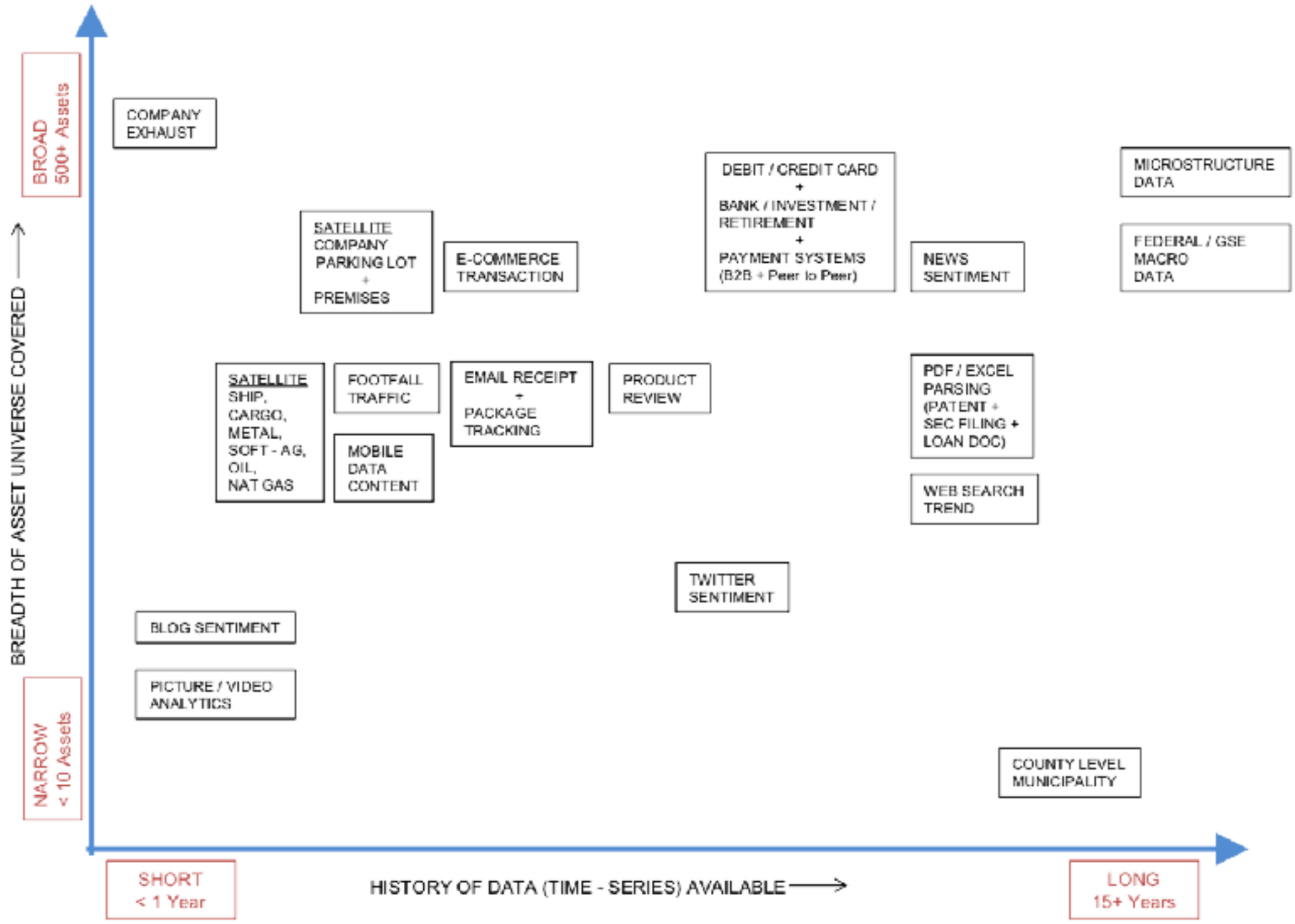
Classification of Alternative Data Sets

Figure 3: Classification of Big/Alternative Data sources



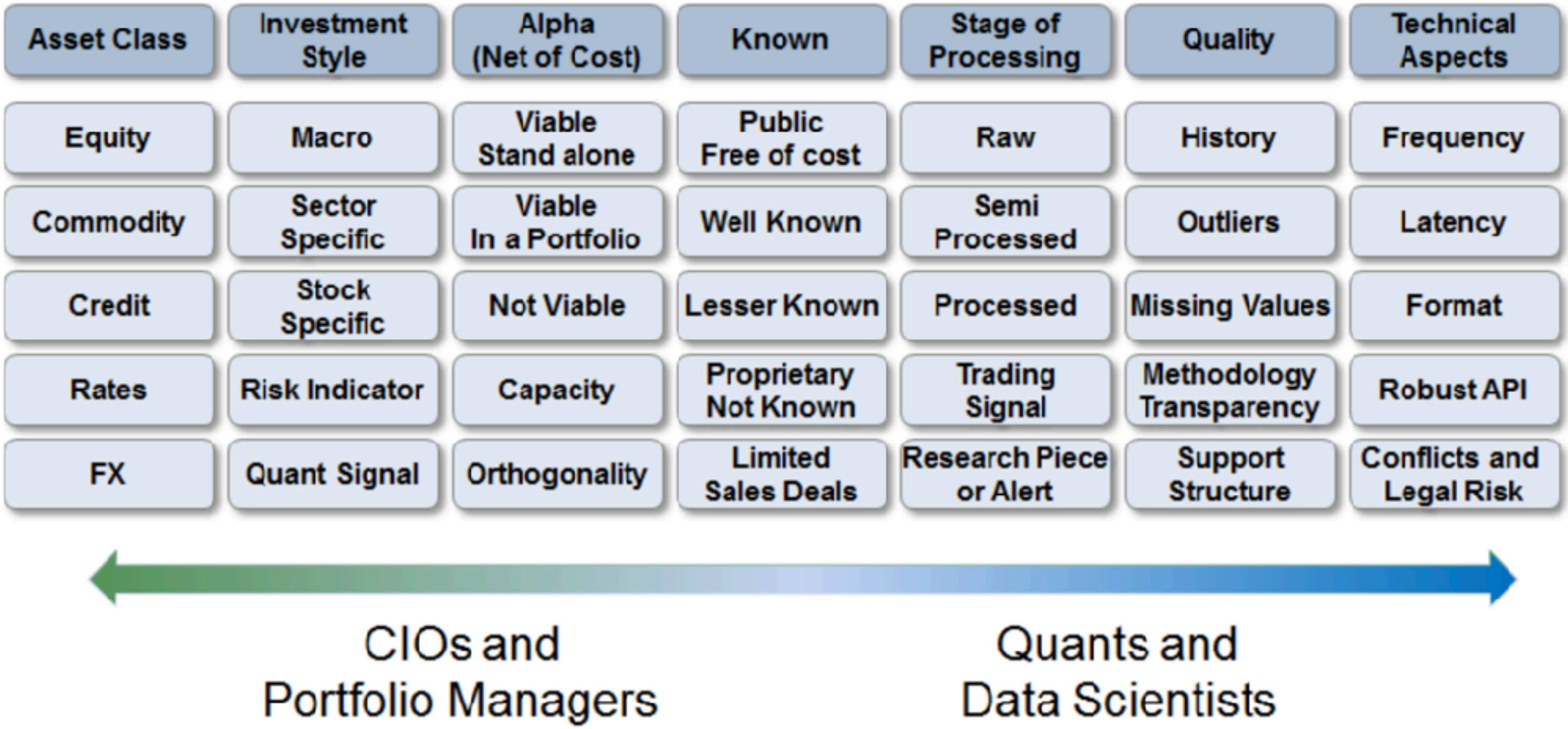
Classification of Alternative Data Sets

Figure 11: Typical length of history for alternative data sets



Classification of Alternative Data Sets

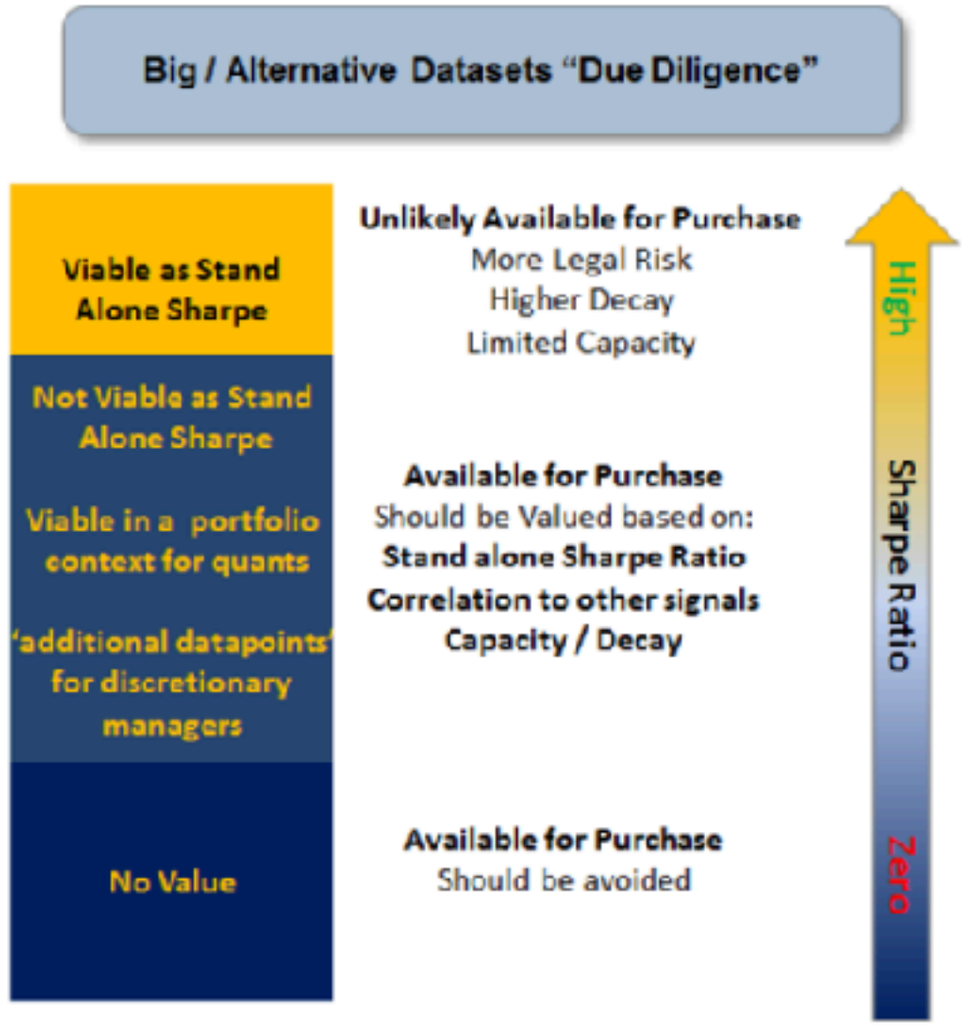
Figure 4: Attributes of an alternative data set



Classification of Alternative Data Sets

1) **Asset class**. Most Big Data are still focused on equities and commodities.
2) **Investment style** – most data are sector and stock specific and relevant for equity long- short investors.
3) **Alpha content**. Alpha content has to be analyzed in the context of the price to purchase and implement the dataset. Trading strategies based on alternative data are tested, and Alpha is estimated from a backtest. These tests can find whether a dataset has enough alpha to make it a **viable standalone** trading strategy.

Figure 5: Information content of an alternative data set

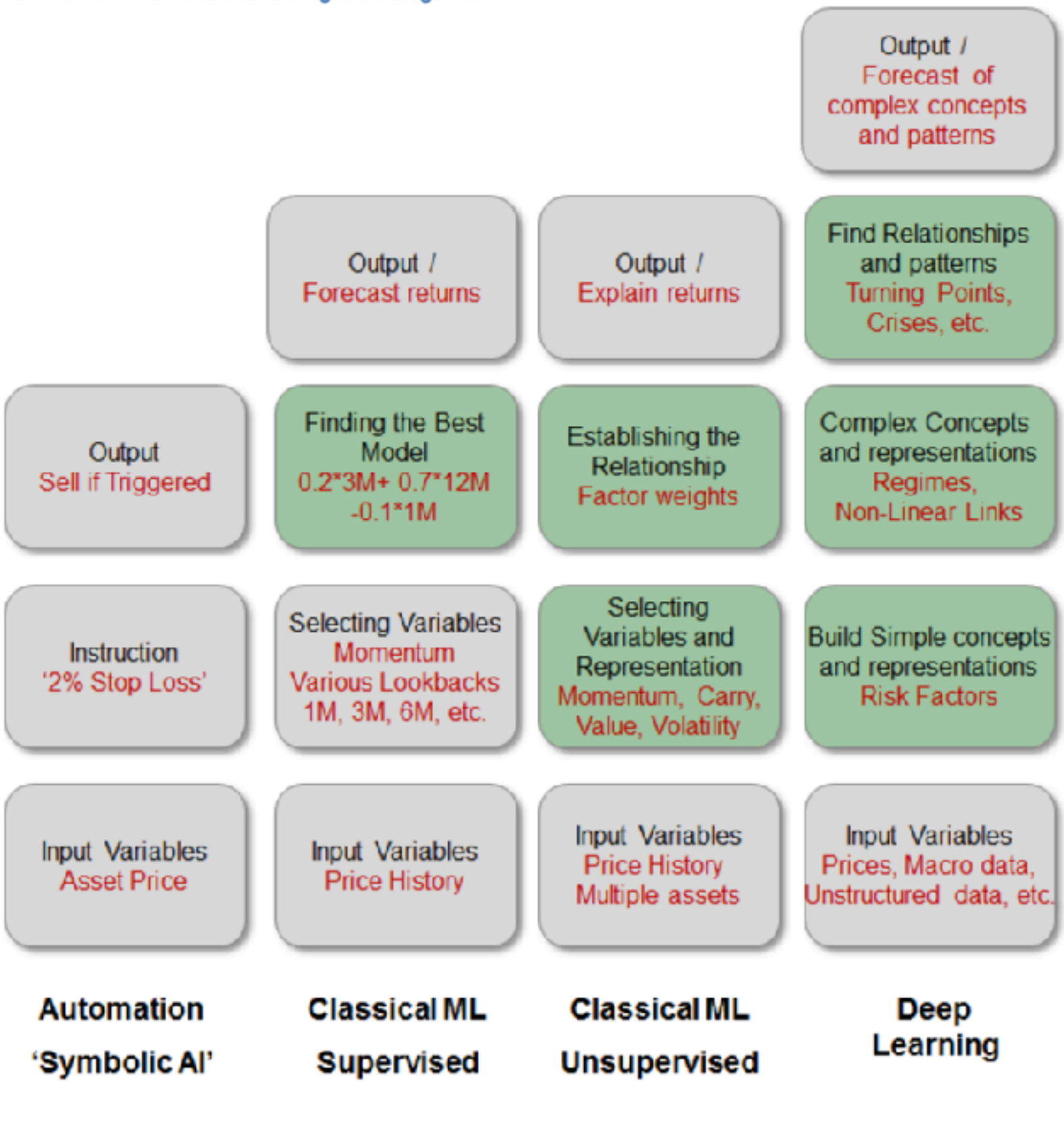


Classification of Alternative Data Sets

4) **How well-known is a dataset**. The more broadly a dataset is known, the less likely it is to lead to a stand-alone strategy with a strong Sharpe ratio. **Well-known public datasets** such as financial ratios (P/E, P/B, etc.) likely have fairly low alpha content and are not viable as a standalone strategies.
5) **Stage of processing** of data when acquired. Fundamental investors prefer processed signals and insights instead of a large amount of raw data. The highest level of data processing happens when data is presented in the form of **research reports**, alerts or trade ideas.
6) **Quality** of data is another important feature, especially so for data scientists and quants.
7) **Technical Aspects** of a big and alternative datasets.

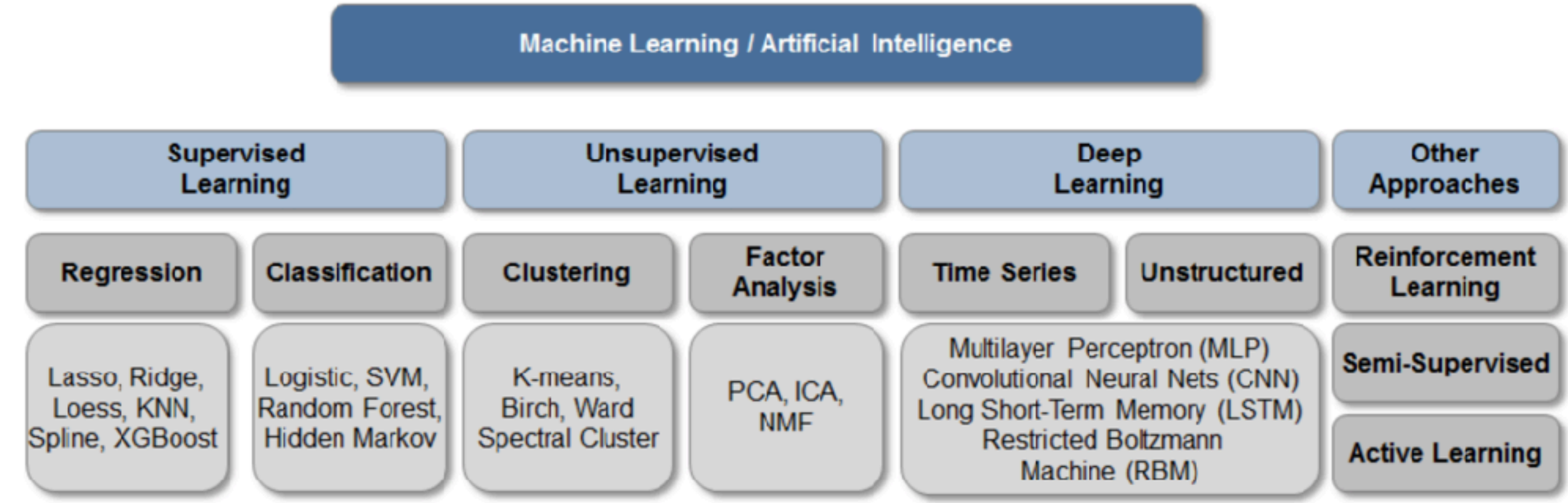
Classification of Machine Learning Techniques

Figure 6: Illustration of different Machine Learning / AI categories



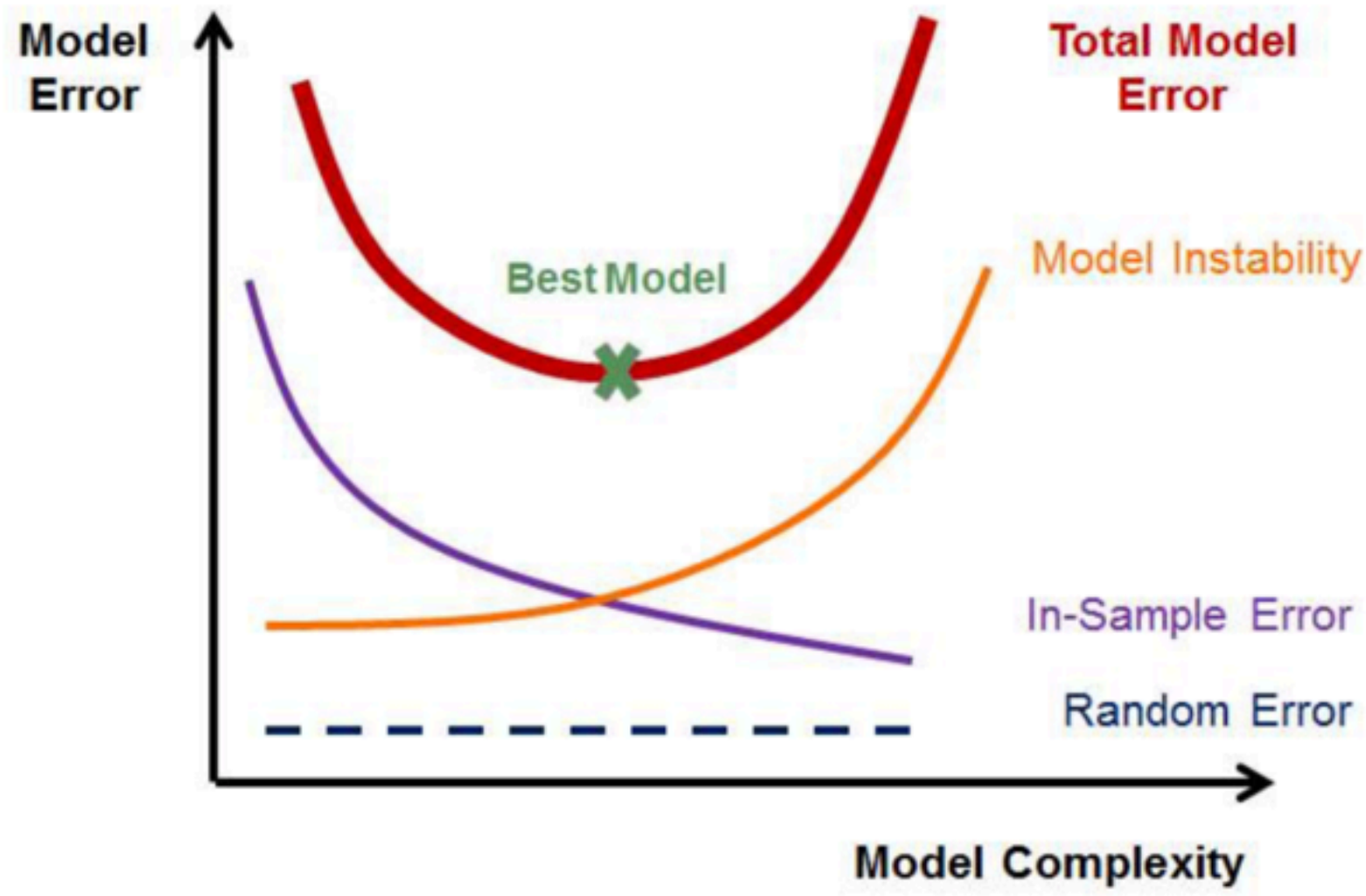
Classification of Machine Learning Techniques

Figure 7: Classification of Machine Learning techniques



Classification of Machine Learning Techniques

Figure 8: Tradeoff between 'model bias' and 'model variance'



Positioning within the Big Data Landscape

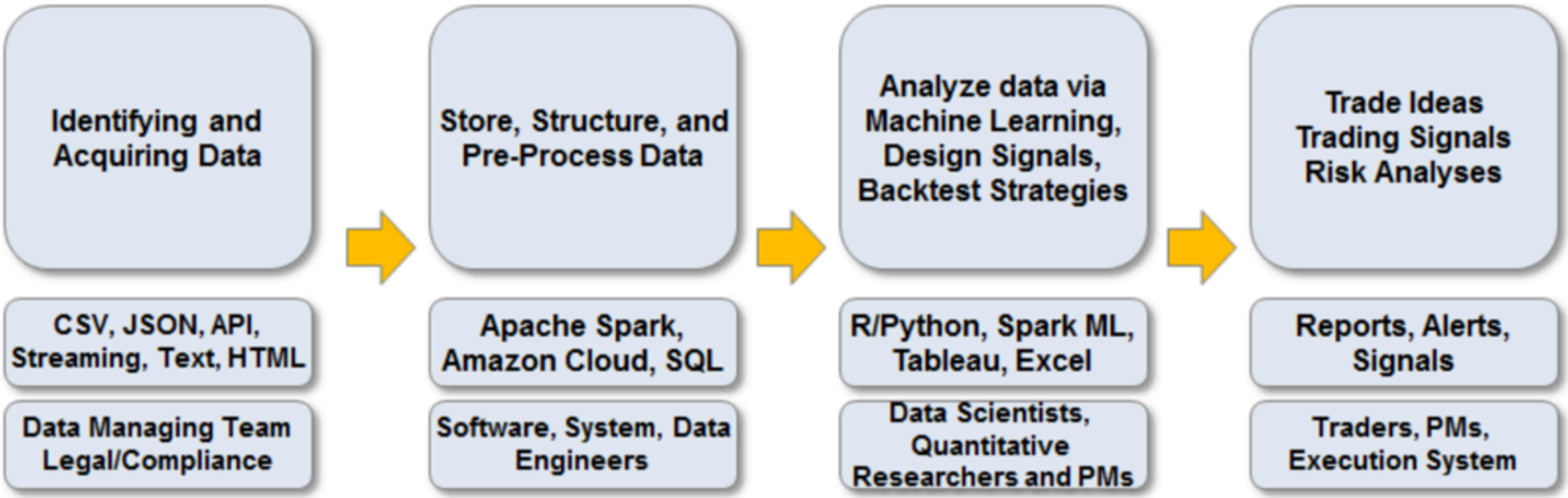



Figure 9: Big Data workflow for investment managers