



Financial Data Professional (FDP)

The *Global Designation* for Finance Professionals in a Data-Driven Industry

Intro To Alt Data



Ronan Crosson, CFA
Director, Data
Strategy
and Analytics



Thomas Combes
Head of Data
Science



Keith Black,
Ph.D., CAIA, CFA,
FDP
Managing
Director, Content
Strategy, CAIA



Mirjam Dekker
PM
FDP Institute





CAIA
ASSOCIATION®

FDP
INSTITUTE™

The Global Designation
for Finance Professionals
in a Data-Driven Industry

The FDPi was created by CAIA to

- ✓ Provide financial professionals with the knowledge necessary to succeed in an industry disrupted by the advent of big data and machine learning.
- ✓ Advocate for the highest levels of professional ethics and standards.
- ✓ Establish the FDP Charter as a global professional designation in the area of financial data science.



EARN YOUR FDP DESIGNATION

A globally-recognized charter is awarded to FDP charter holders



TWO ONLINE CLASSES*

Choose either Python or R
Can be completed before or after the exam



ONE COMPREHENSIVE EXAM

Offered twice per year
March & November

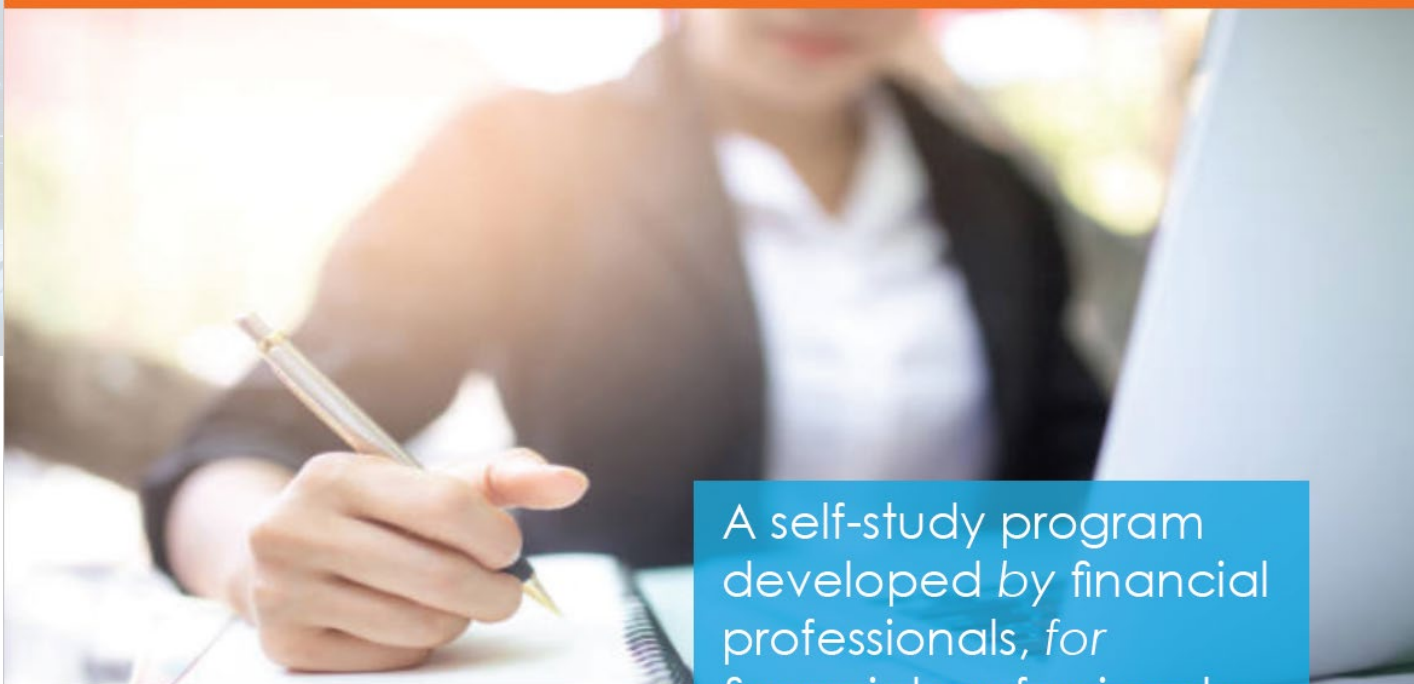


VALUE ADD

Employers increasingly seek to find professionals to have the skills to apply data science tools to solve their most challenging problems

FDP EXAM

1. Introduction to Data Science & Big Data
2. Machine Learning: Introduction to Algorithms
3. Machine Learning: Regression, Support Vector Machine & Time Series Models
4. Machine Learning: Regularization, Regression Trees, Random Forest & Overfitting
5. Machine Learning: Classification & Clustering
6. Machine Learning: Performance Evaluation, Backtesting & False Discoveries
7. Data Mining & Machine Learning: Naïve Bayes & Text Mining
8. Big Data & Machine Learning: Ethical & Privacy Issues
9. Big Data & Machine Learning in the Financial Industry



A self-study program developed by financial professionals, for financial professionals.

Exam candidates now **have one of two test options** for the FDP exam

- Option 1:** At a Prometric test center near you
Exam window: October 12 – November 8, 2020
- Option 2:** Through Remote proctoring from your home/office
Exam Day: December 1, 2020

Learn more at www.fdpinstitute.org





Financial Data Professional (FDP)

The *Global Designation* for Finance Professionals in a Data-Driven Industry

Intro To Alt Data



Ronan Crosson, CFA
Director, Data
Strategy
and Analytics



Thomas Combes
Head of Data
Science



Keith Black,
Ph.D., CAIA, CFA,
FDP
Managing
Director, Content
Strategy, CAIA



Mirjam Dekker
PM
FDP Institute



Agenda

- About Eagle Alpha
- The Growth in Alternative data
- Building an Alternative Data Strategy
- Data Quality:
 - Challenges of Working with Alternative Data
 - When data formats break your parser
 - Eagle Alpha's Backend Process
 - Demo of Findings on Alternative Datasets
- Closing Remarks

Speakers



Ronan Crosson, CFA
Director, Data Strategy and Analytics



Investment experience: 16.5 yrs
Alternative data experience: 5.5 yrs

Ronan leads the analyst team and oversees Eagle Alpha's Data Strategy solution.

Ronan's background is as a senior analyst in State Street Global Advisors and he has a postgraduate diploma in Data Analytics.

ronan.crosson@eaglealpha.com



Thomas Combes
Head of Data Science



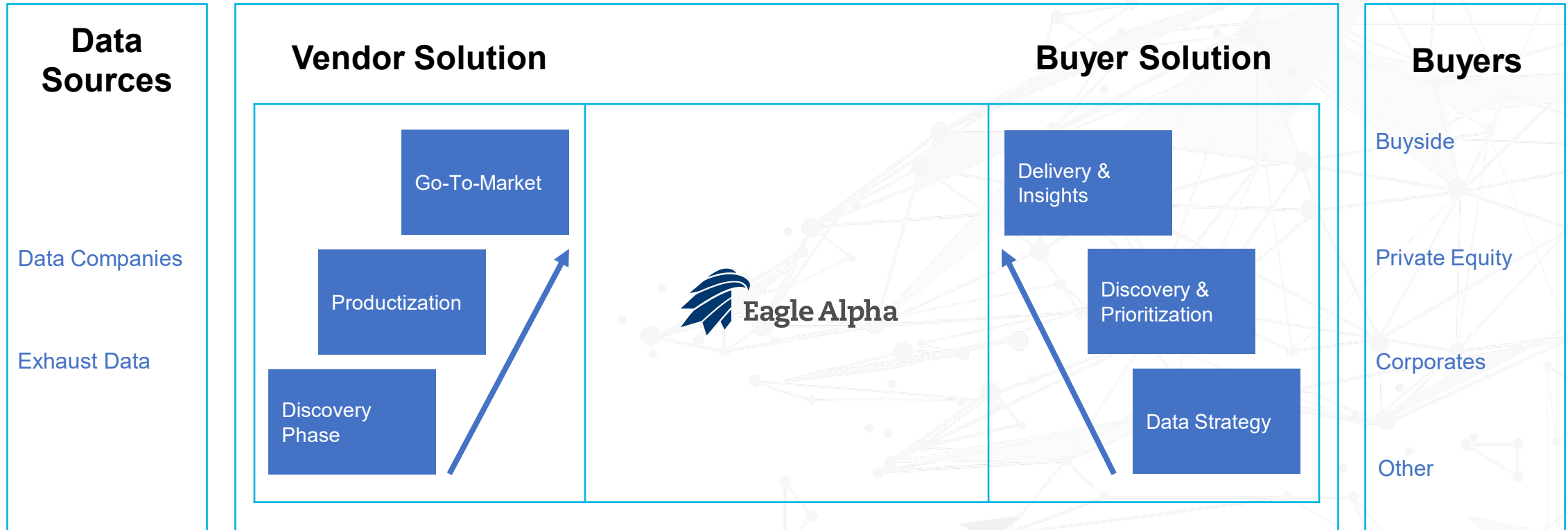
Aerospace Engineer: 6.5 yrs
Data Scientist: 3 yrs

Thomas leads the Data Science team, which builds data products and tests 3rd party datasets for quality and robustness.

Thomas previously worked at Boeing developing software for real-time, large-scale analysis of flight test data.

Thomas.combes@eaglealpha.com

Eagle Alpha Has Been A Pioneer In The Alternative Data Space Since 2012



Eagle Alpha 's Data Credentials are Validated By Leading Wall Street Firms ...

May 2020



Man Group views our database of alternative datasets as market leading. It selected Eagle Alpha as a co-author on a paper entitled 'The Data on Data. Extract: "in today's day and age, data is often referred to as the oil that fuels the investment machine. Some go as far as tagging the search for alternative data as the new oil rush. We look at the data on data to find out what really matters in this new boom".

2019 & 2020

J.P.Morgan

J.P Morgan's quantitative research team views our proprietary data quality testing tool as world-class. It invited us to give workshops to its clients in Boston, New York, London and Sydney between October 2019 and March 2020

October 2019



J.P. Morgan quantitative research team has published two major reports on big data. Both feature Eagle Alpha.

[Link](#)

Since 2018



Since 2018 J.P. Morgan's prime broking team has been the lead partner on all our conference (New York, London, Singapore, Hong Kong, Sydney, virtual).

March 2017



Citi published the first primer on alternative data. It contained a 10-page profile of Eagle Alpha.

[Link](#)

...And Buyers



Eagle Alpha introduced Jupiter to the world of alternative data. They navigated us through the taxonomy, helped us think about how we might use alternative data across our business, introduced us to a wide range of data providers and educated our investment teams in possible ways of extracting value from alternative data. They even helped us think through what would be needed to evolve our own data science capability within the firm. Eagle Alpha's insights, experience and contacts have been invaluable to us.

- Magnus Spence, Head of Investments (Alternatives)

<https://www.jupiteram.com/>



Eagle Alpha has provided substantial insight into new frontiers of alternative data and through our partnership with them we have been able to dramatically scale our footprint in this critical component of our research.

- Mani Mahjouri, CEO (#12 hire at AQR)

<https://www.blueshift.am/>



We recognize the increasing importance of alternative data to the trading industry, and feel that Eagle Alpha is prime placed to help firms capitalize on this. Emmett and his team have a broad and integrated business plan that allows clients to get the best from the space, across a number of areas. We feel that Eagle Alpha is also marked out by the high quality of its personnel, and we have been impressed with the strength and depth of their team and the value they can bring to the investment process. We have been actively partnering with the EA business for some time, and are looking forward to this next step and continuing to watch the business succeed.

- Chris Udy, CEO

<https://www.tibra.com/>



Eagle Alpha has been an invaluable ally as my firm has integrated alternative data into our private equity investment process. The due diligence solution in particular is an innovative way to meet our alternative data needs, which are different from public markets investors. I've appreciated Eagle Alpha's active engagement with us to understand our needs and develop the solution. It's the only solution of this type I've seen in the market.

- Wesley Barnes, CEO

<https://www.brightrivercapital.com/>



The Growth of Alternative Data

Alternative data is defined as non -traditional data that can be used to augment decision making.

- Eagle Alpha was the first company to create a taxonomy of “alternative” data.
- There are currently 26 categories of data in Eagle Alpha’s taxonomy: 24 non-traditional data categories and 2 traditional categories.

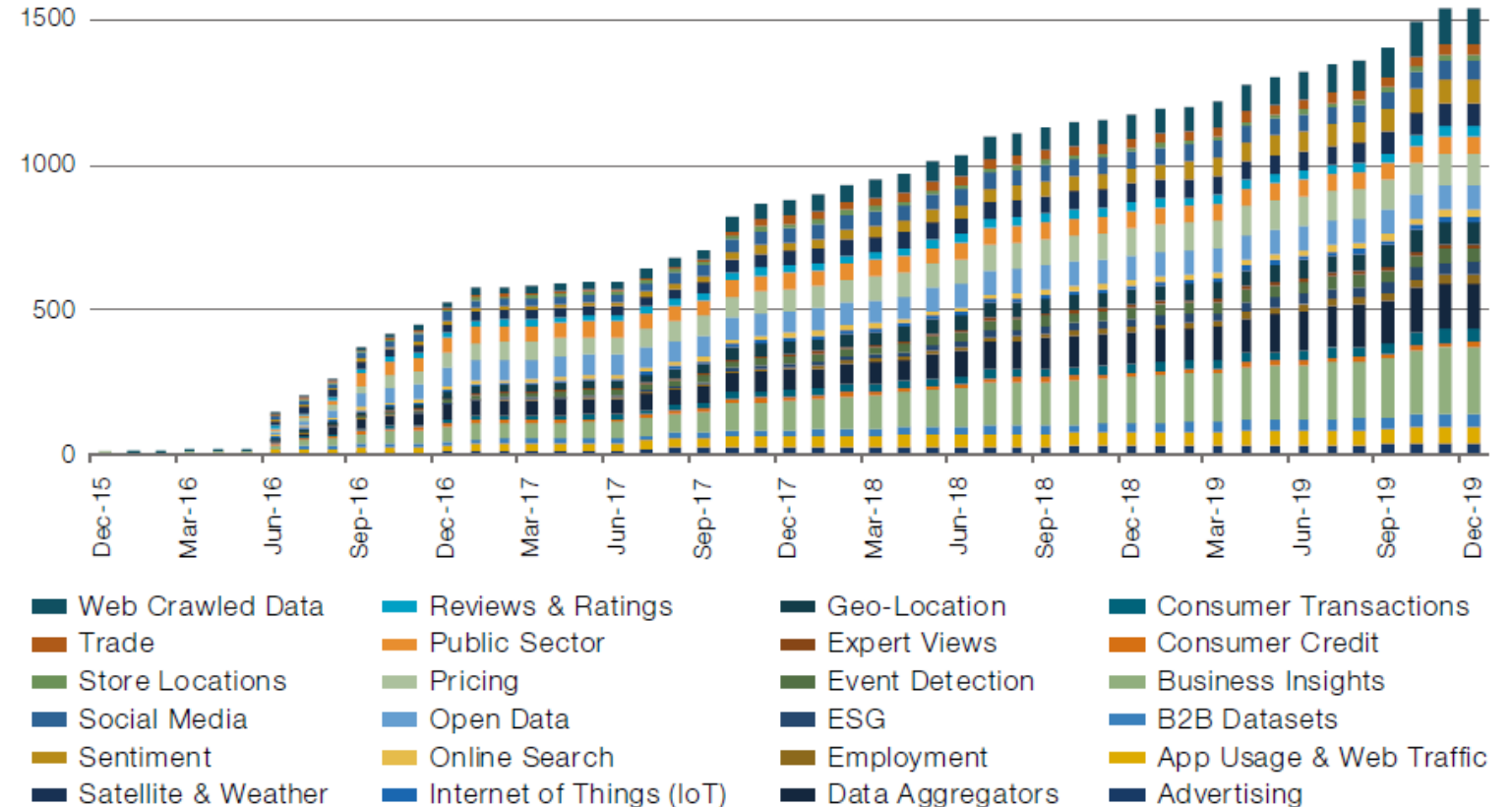


Note: numbers in the chart reflect the number of dataset profiles in the category as at 13 July 2020. A dataset may be included in more than one dataset category.

There are currently over 1,355 in Eagle Alpha's database. This is expected to grow to 5,000 by 2024.

- Geographical distribution of datasets: US: 55%, EMEA: 30% and APAC: 15%.

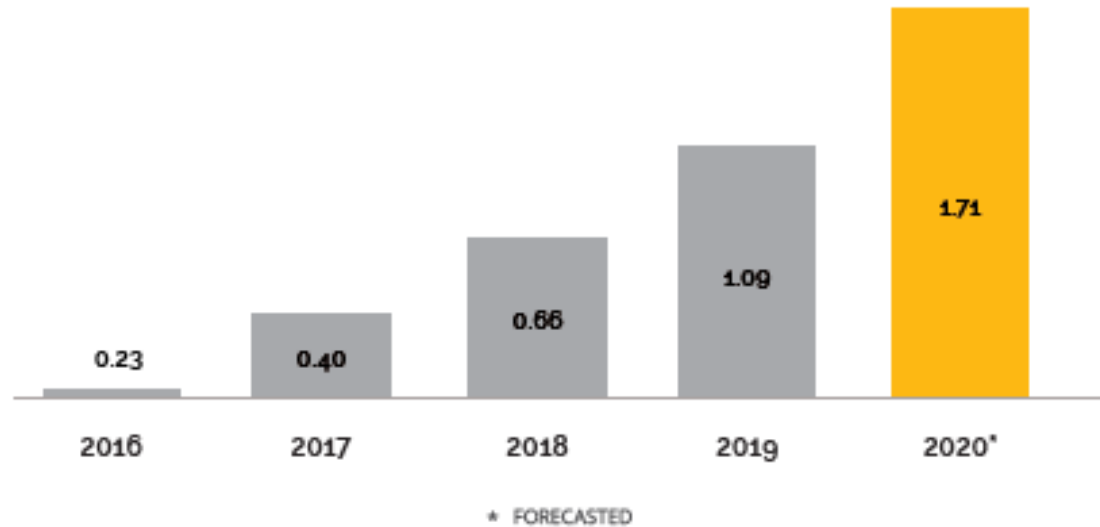
Figure 2. Number of Datasets Over Time



Source: Eagle Alpha; As of December 2019. Datasets are split into Eagle Alpha categories¹.

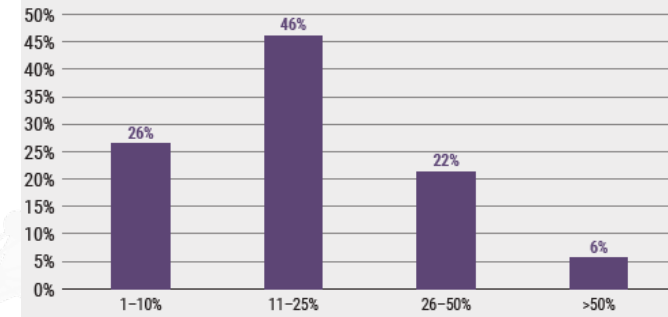
Alternative Data Spend Is Estimated At >\$1bn In 2019; Budgets Growing Rapidly But Adoption <50%

ALTERNATIVE DATA SPEND
(\$BILLIONS)



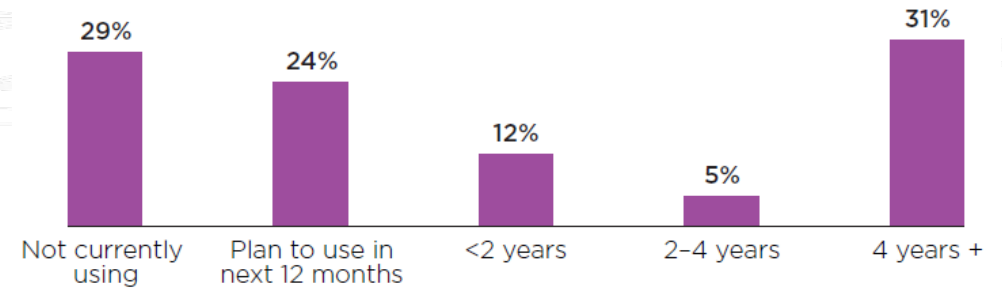
Source: Element 22/UBS "ANALYTICS POWER 2019" report

By what percentage does your organization plan to increase its budget?



Source: Lowenstein Sandler "Alternative Data = Better Investment Strategies, But Not Without Concerns" report (2019)

LENGTH OF TIME USING ALTERNATIVE DATA



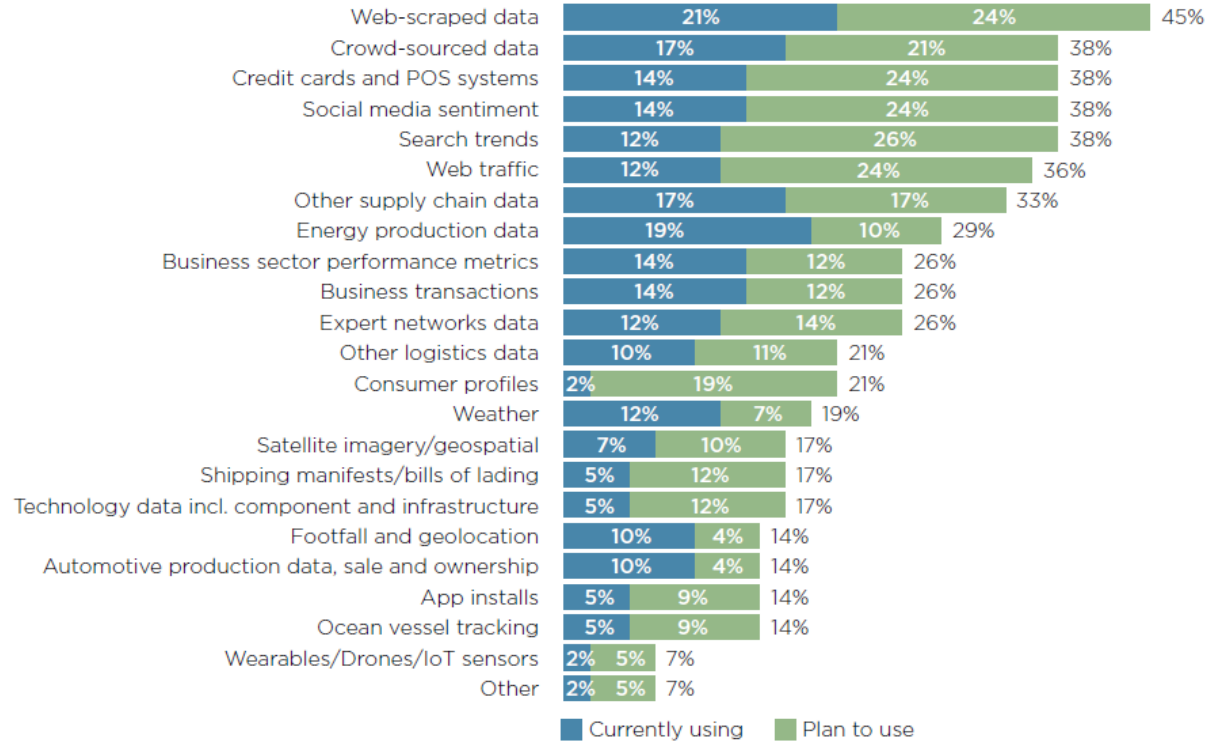
Source: Greenwich Associates 2019 Alternative Data Study

The Ranking Of Most Popular Datasets Varies By Manager Style

Discretionary Funds

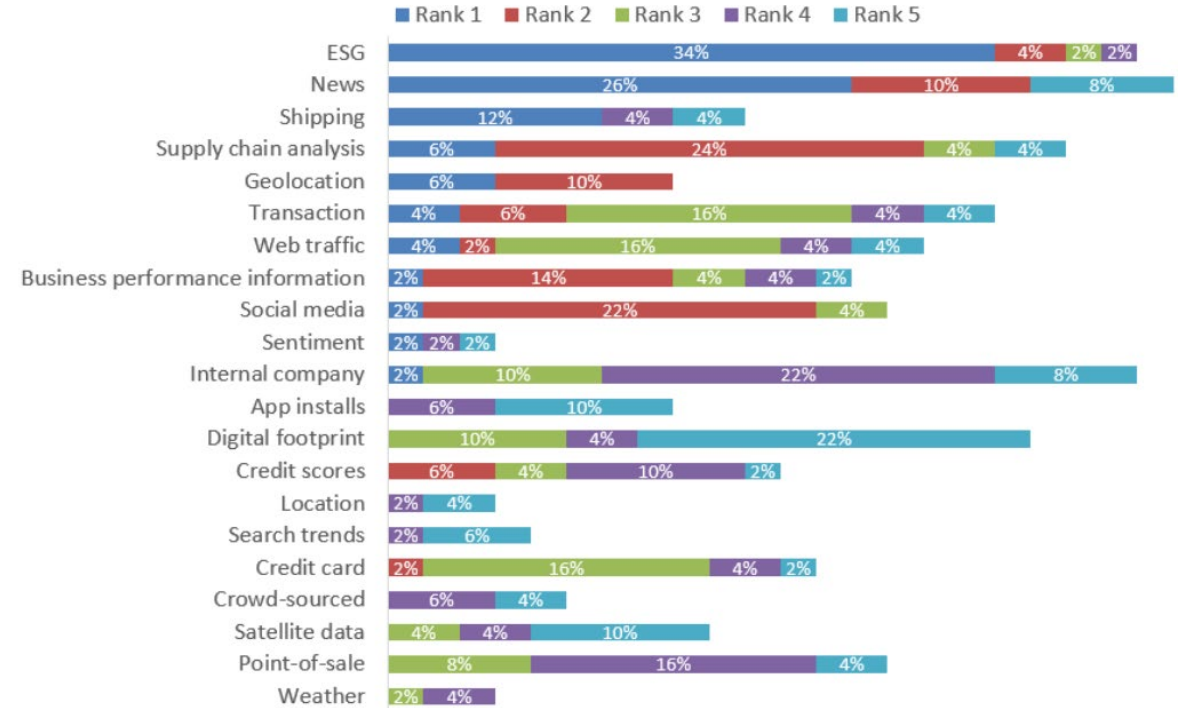
Quant Funds

USAGE OF ALTERNATIVE DATA SETS



Source: Greenwich Associates 2019 Alternative Data Study

Alternative Content Rank by Value (Future)



Source: <https://insight.factset.com/the-future-of-quantitative-analysis-data-sources>

Alternative Data Is Typically Used As A Complementary Input Into An Investment Process

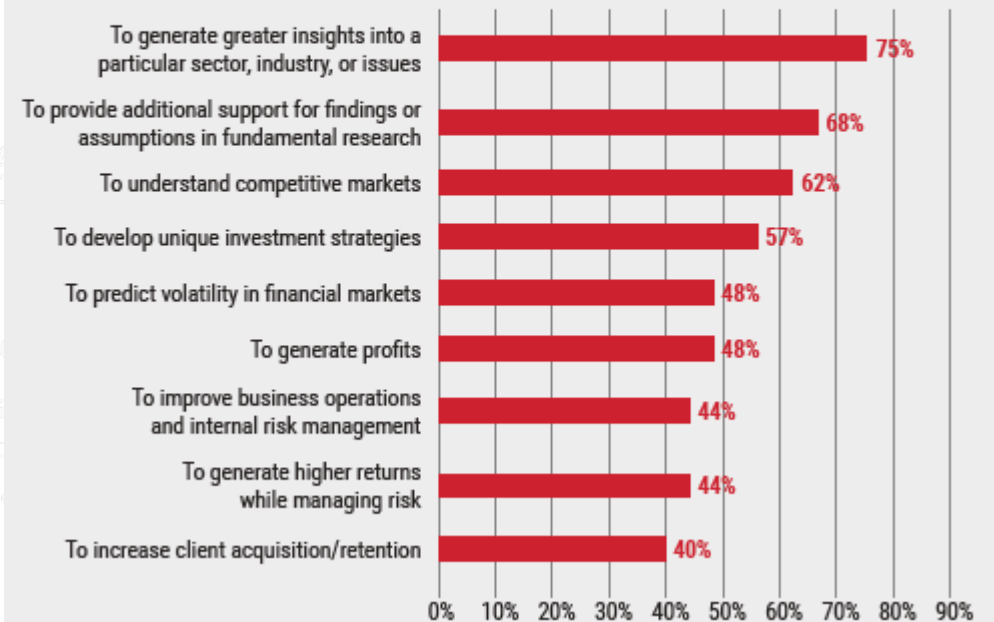
MAIN USES OF ALTERNATIVE DATA



Source: AIMA "Casting the Net - How Hedge Funds are Using Alternative Data" report (2019)

For which of the following purposes do you use alternative data?

(Select all that apply.)



Source: Lowenstein Sandler "Alternative Data = Better Investment Strategies, But Not Without Concerns" report (2019)

The ROI On Alternative Data Is Not Straight Forward To Measure

- ✓ Market spend on alternative data is growing at 50-60%¹
- ✓ Buyside data budgets are growing 15-20%²
- ✓ Renewal rates on datasets are >80%³
- ✓ European hedge funds using AI returned almost triple the global industry average in the three years through May 2020. ⁴

1 - Element 22/UBS "ANALYTICS POWER 2019" report

2 - Lowenstein Sandler "Alternative Data = Better Investment Strategies, But Not Without Concerns" report (2019)

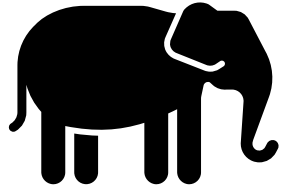
3 - Eagle Alpha analysis

4 - Cerulli Associates research: <https://www.institutionalinvestor.com/article/b1mssrsw1mpr0/AI-Powered-Hedge-Funds-Vastly-Outperformed-Research-Shows>



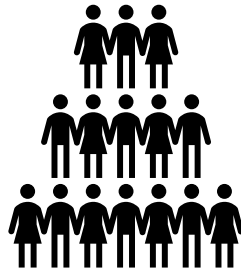
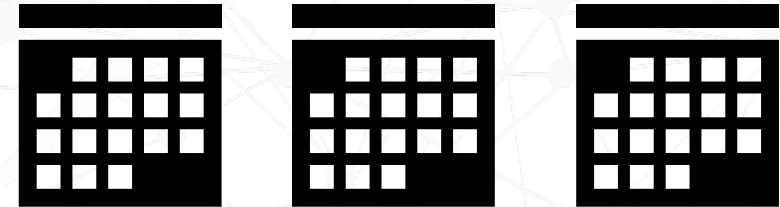
Building an Alternative Data Strategy

Lessons From Alternative Data Early Adopters



Senior Management Buy-In

Long-Term Investment



Involve Multiple Stakeholders

Recommended First Steps in Alternative Data



1. Audit



2. Data Lead



3. Working Group



4. Define Success

Learn From Others' Mistakes

Over-Analysing



Under-Investing



Unrealistic Expectations



Narrow Focus



Quitting Too Early



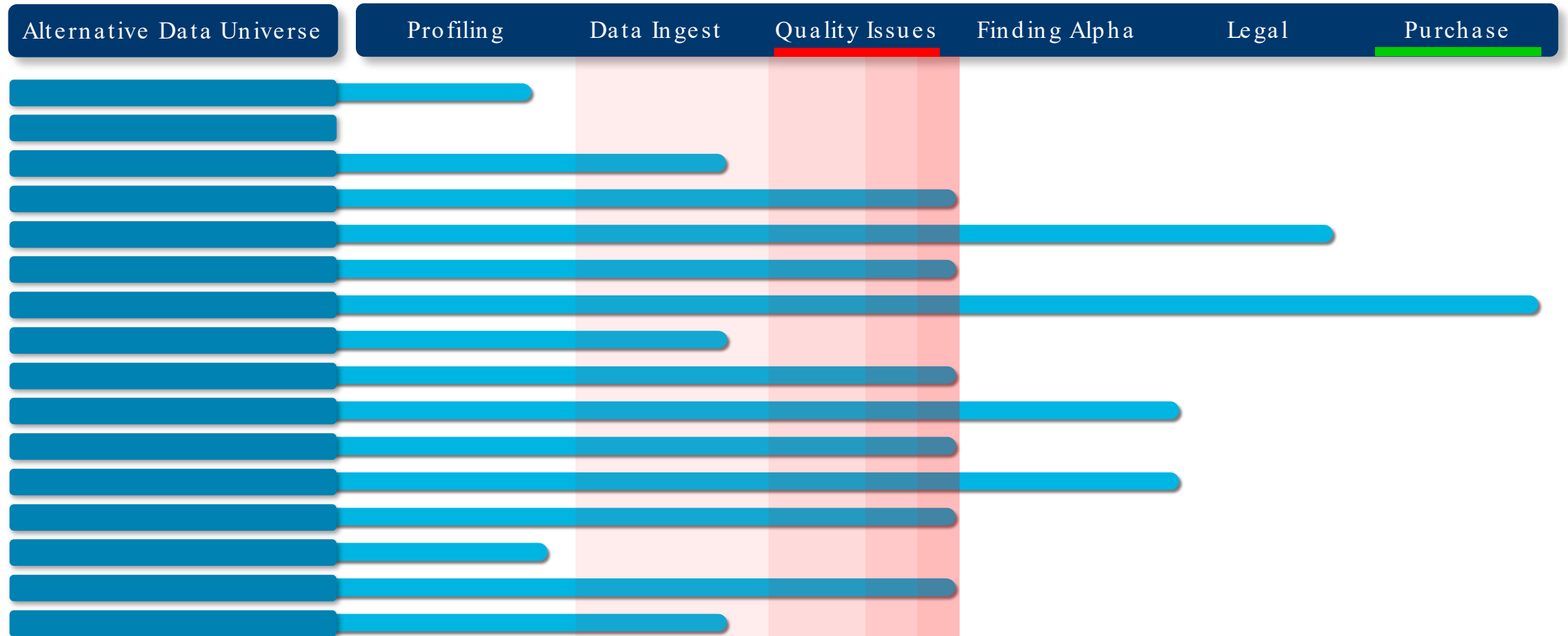
Poor Hiring





Data Quality: Challenges of Working with Alternative Data

Knowing what you don't know – Data Quality

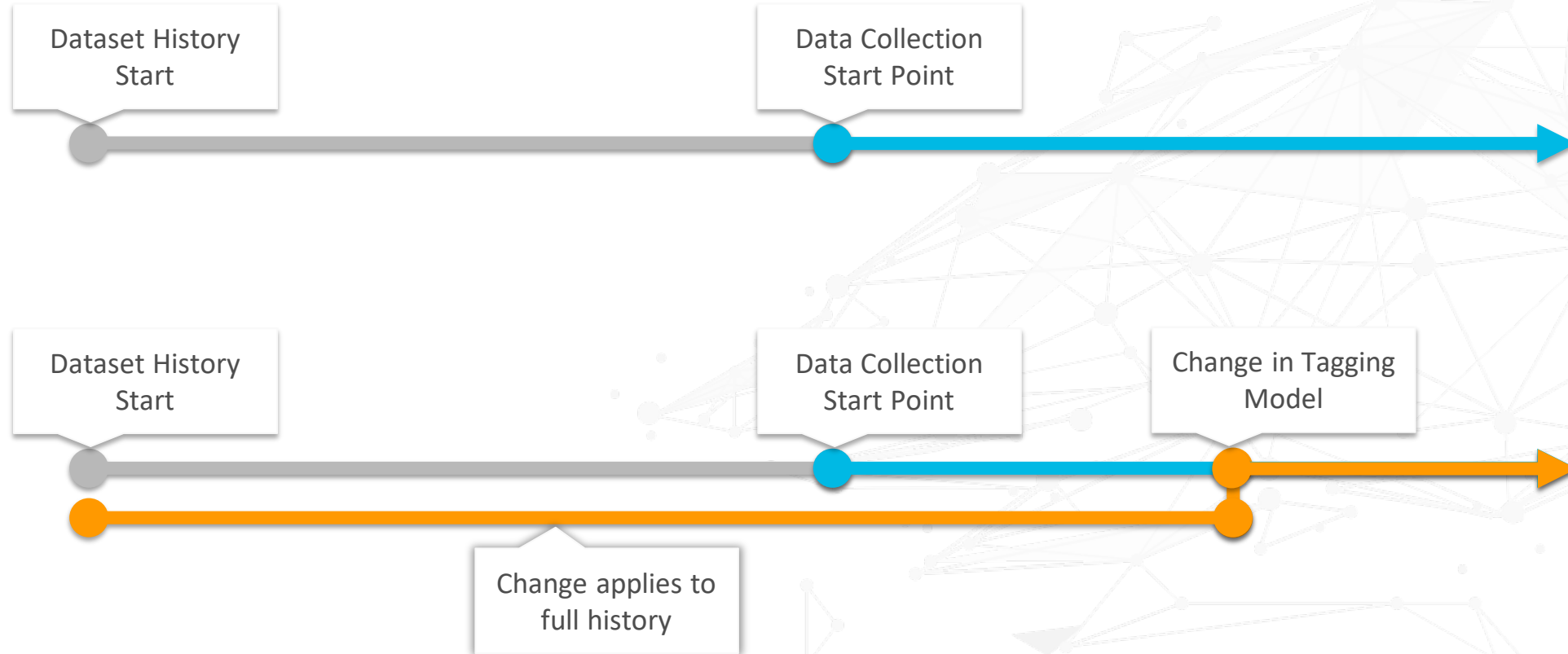


Data Assessment: the Trial Process

- Trials typically use live data for a period of 3 months
- Data ingest and exploration often consume the **first third of the trial period**



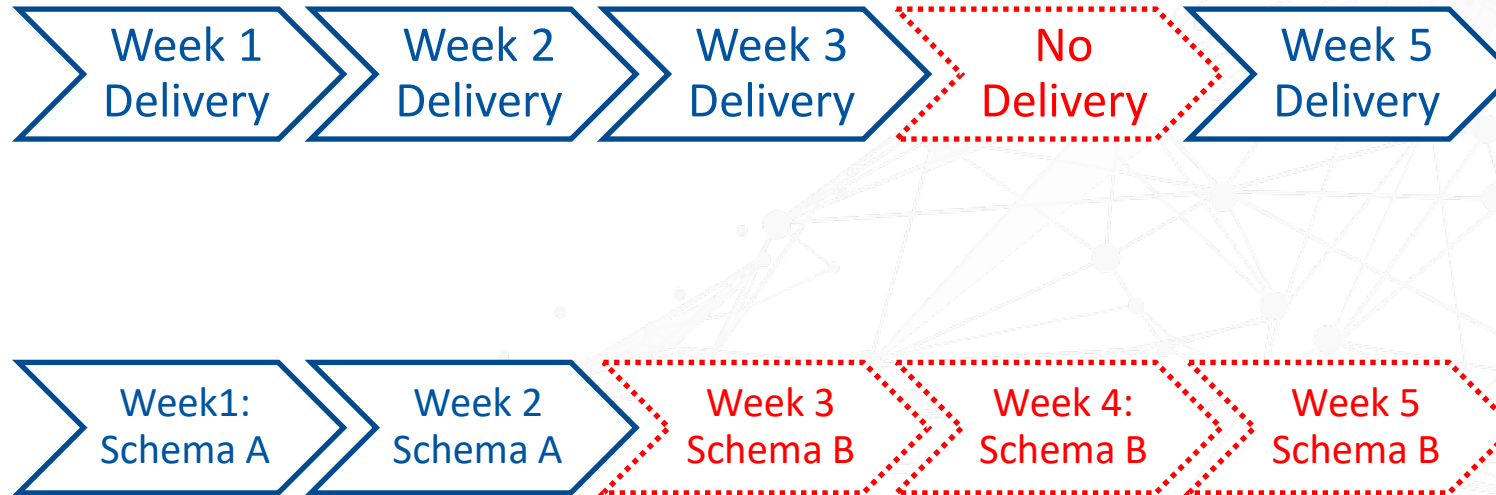
Data Buyer Challenges: Point -in -Time



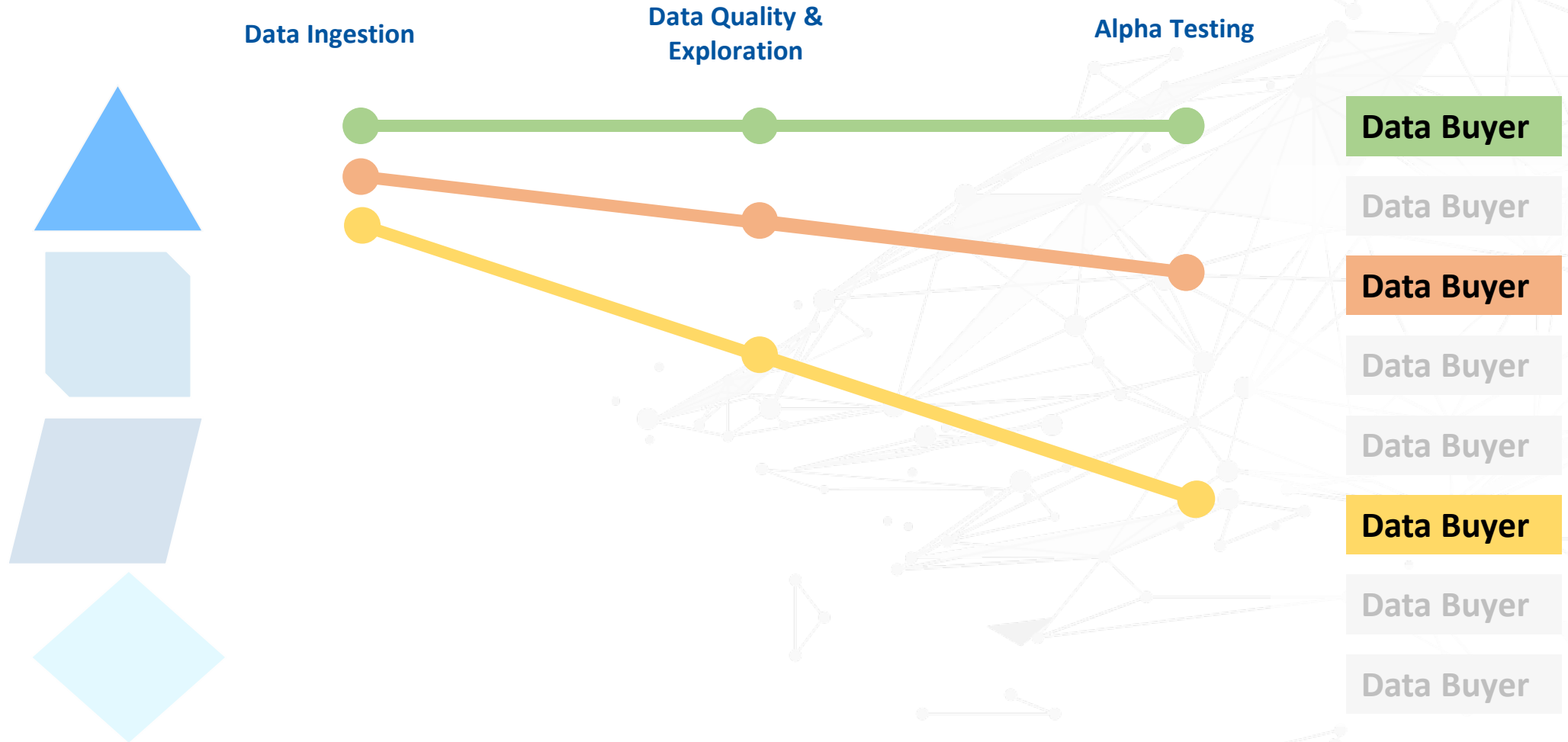
Data Buyer Challenges: Mapping Entities

Name	Ticker	Exchange Code
<input type="text" value="Filter"/>	<input type="text" value="Filter"/>	<input type="text" value="Filter"/>
TESCO PLC	TSCO	XL
TESCO PLC	TSCO	XF
TRACTOR SUPPLY COMPANY	TSCO	UB
TESCO PLC	TSCO	BQ
TESCO PLC	TSCO	XA
TESCO PLC	TSCO	SW
TRACTOR SUPPLY COMPANY	TSCO	AV
TRACTOR SUPPLY COMPANY	TSCO	UA

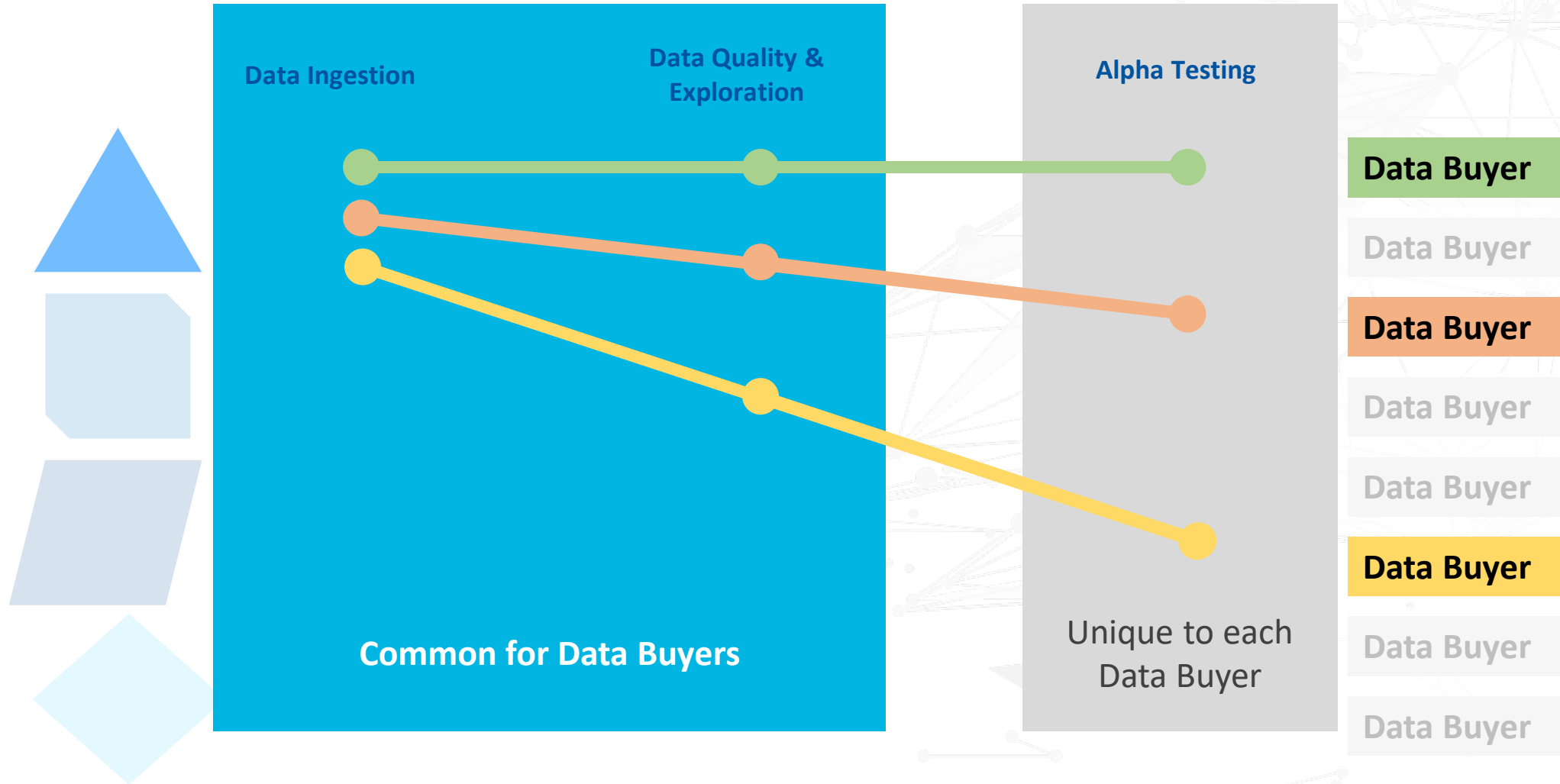
Data Buyer Challenges: Data Delivery Issues



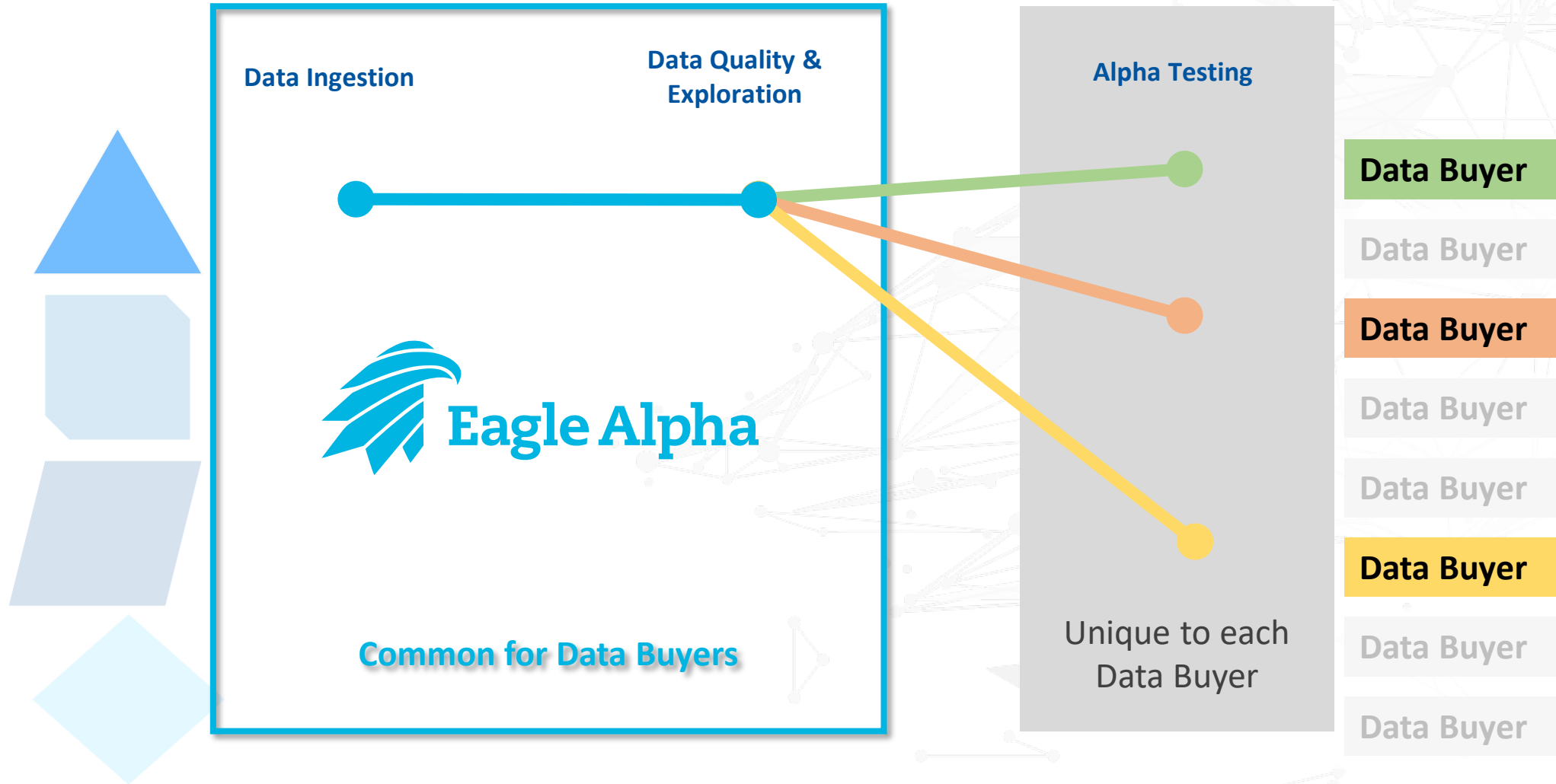
Data Assessment Today



Analysis Commonality



Addressing a Common Problem in One Place





**Data Quality:
When data formats
break your parser**

Pros and Cons of Different Dataset File Formats

Raw Text Formats

- PSV Pipe delimited value file
- TSV Tab delimited value file
- CSV Comma delimited value file

**When raw text appears in one of these formats, often you can accidentally push data into incorrect fields or onto new rows*

XML: eXtensible Markup Language

- Require lots of memory, can be cumbersome – stay away

JSON

- JSON: JavaScript Object Notation
- Can requires lots of memory, JSON lines is better in this case

Columnar File Formats: Parquet, ORC

- Benefit from a defined association between a column and the data in that column (no confusion - see above)
- Can be highly compressible if lots of repetition of column values such as categorical columns

Data Structure Horror Stories

```
b'AGO\t199001\t2\t32\t1\t0\t1\t0\t24\t[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,  
b'AGO\t199002\t2\t32\t1\t0\t1\t0\t25\t[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,  
b'AGO\t199003\t2\t32\t1\t0\t1\t0\t26\t[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,  
b'AGO\t199004\t3\t37\t2\t0\t2\t0\t18\t[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,  
b'AGO\t199005\t3\t37\t2\t0\t2\t0\t19\t[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,  
b'AGO\t199006\t3\t37\t2\t0\t2\t0\t20\t[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,  
b'AGO\t199007\t3\t37\t2\t0\t2\t0\t21\t[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,  
b'AGO\t199008\t3\t37\t2\t0\t2\t0\t22\t[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,  
b'AGO\t199009\t3\t37\t2\t0\t2\t0\t23\t[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,  
b'AGO\t199010\t4\t34\t3\t0\t3\t0\t18\t[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
```

Tab Delimited

Comma Delimited

File extension: TSV

Dates and Times: What time is it?



- ISO 8601 is great – but few vendors appear to follow it:
https://en.wikipedia.org/wiki/ISO_8601
- Date types we have seen in datasets (equal to April 10th 2011):
 - European: 10/4/2011
 - United States: 4/10/2011
 - ISO 8601 Lexicographically sortable: 2011-04-10
 - ISO no dashes: 20110410
 - Year + Month spelled out: 2011Apr
 - Timestamp with timezones
 - Variable timezones with no reference to a fixed timezone (i.e UTC)
 - Unix time: 1302393600
 - Unix time milliseconds: 1302393600000
 - Year + Month + Day in multiple columns
 - Week of Year
- Never edit a CSV in Excel that has a time or date column
- Is this capture time, model prediction time, or publication time?



Data Quality: Eagle Alpha's Backend Process

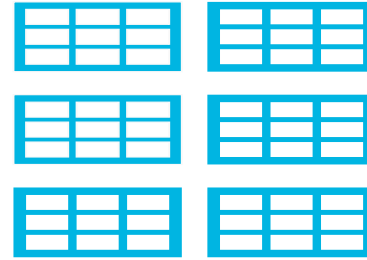
Eagle Alpha Data Quality Testing Process

We **ingest** full raw datasets from vendors, often >1TB in size



We've built a highly-automated **column metadata tagging system** to capture rich dataset traits

We aggregate the raw vendor data into **small, consumable tables**



Our **query generation system** leverages this tagged metadata to build tailored dataset-specific assessment methods for data quality and exploratory analysis



We populate notebooks using a **toolkit of post-processing visualizations and analysis** we are always adding to



Host access to **notebooks and tables** using our JupyterHub and our own data API for collaboration and dissemination



Data on Data: why column -based metadata tagging is useful



- **Input type identification** : *beyond string, int, and float*
 - Categorical data
 - Ticker, Product name, user ID?
 - Temporal data
 - Master time or a Start / Stop time?
 - Metric data
 - Price, lat/long coordinates, unstructured text?
- **Join logic** between primary and reference tables
- **Filtering logic** to rank / limit values in a column
- **Tag = classification** for training machine learning systems

Constructing Training Sets by using Common Data Models

Field: `comp_name`
Tagged as **COMPANY NAME**

Field: `ticker`
Tagged as **TICKER SYMBOL**

Field: `product_desc`
Tagged as **PRODUCT NAME**

<code>comp_name</code>	<code>ticker</code>	<code>product_desc</code>
Apple Inc.	AAPL	Iphone X
Tesla Inc.	TSLA	Model S



Company Name "Apple Inc." with Ticker symbol "AAPL" has a product named "Iphone X"

Company Name "Tesla Inc." with Ticker symbol "TSLA" has a product named "Model S"

Cloud Query as a Service models: Pros and Cons



Amazon
Athena



Google
BigQuery

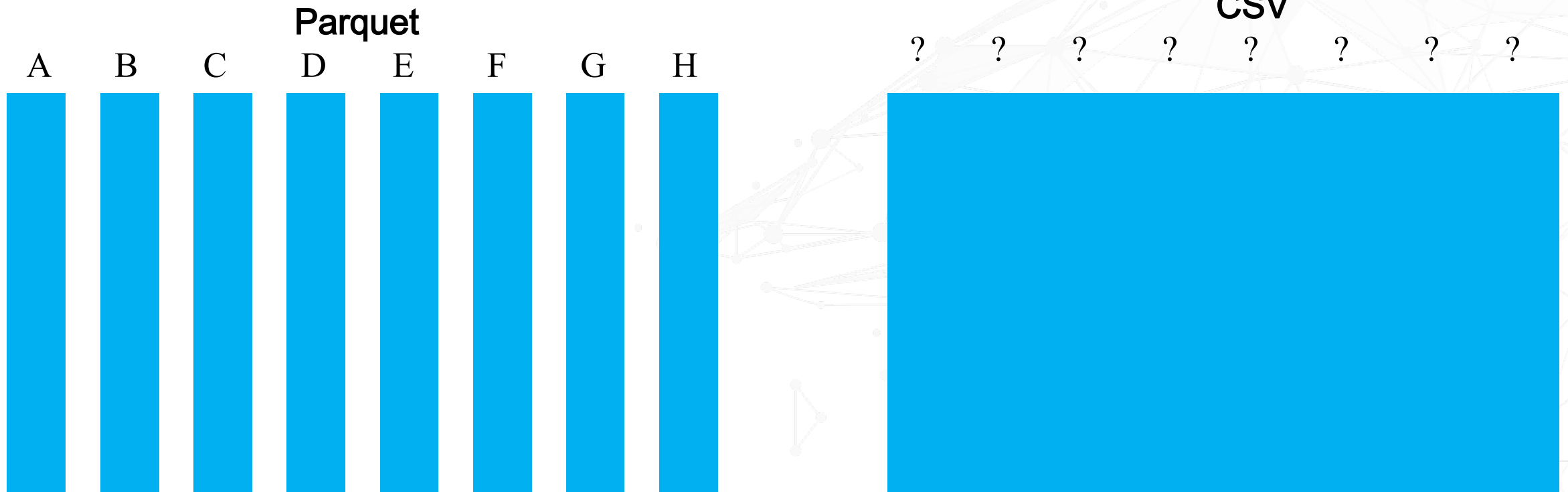


Azure
Synapse
Analytics

- Pros:
 - useful for ad-hoc or non-recurring queries scanning large data sources
 - Data stays in its original form
 - Focus on queries instead of the infrastructure coordination
- Cons:
 - Most cloud-based services are some form of pay-by-data-scanned, which means scrutiny of each query is required so as not to explode costs

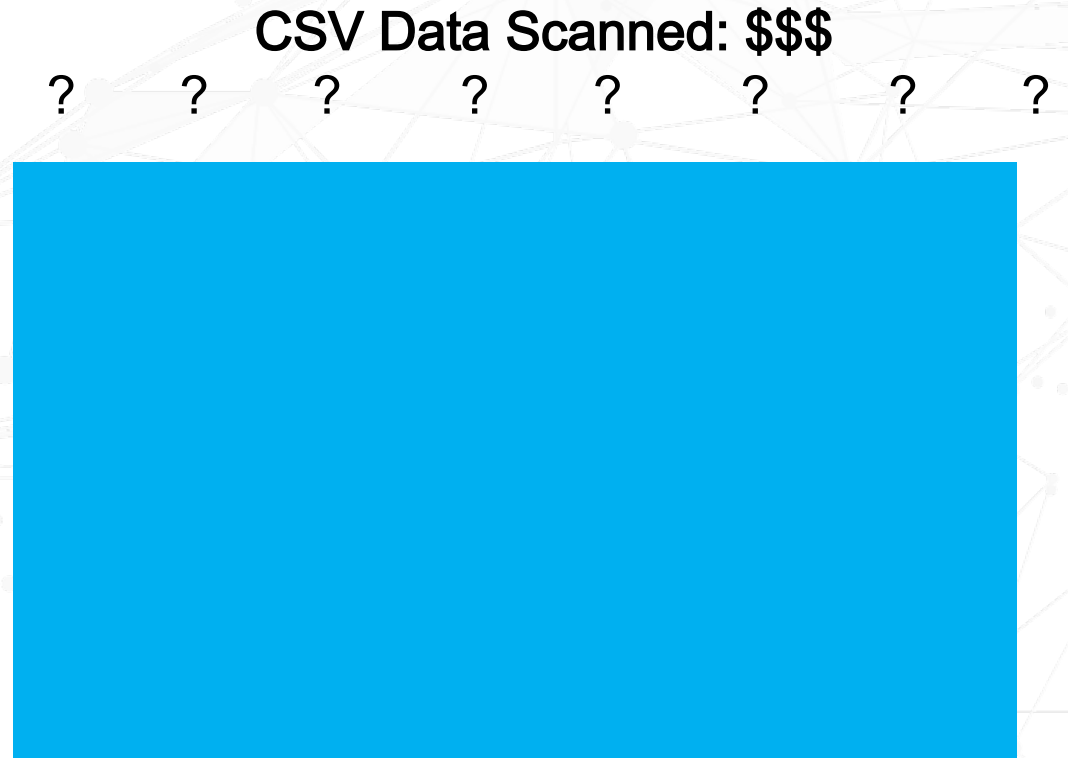
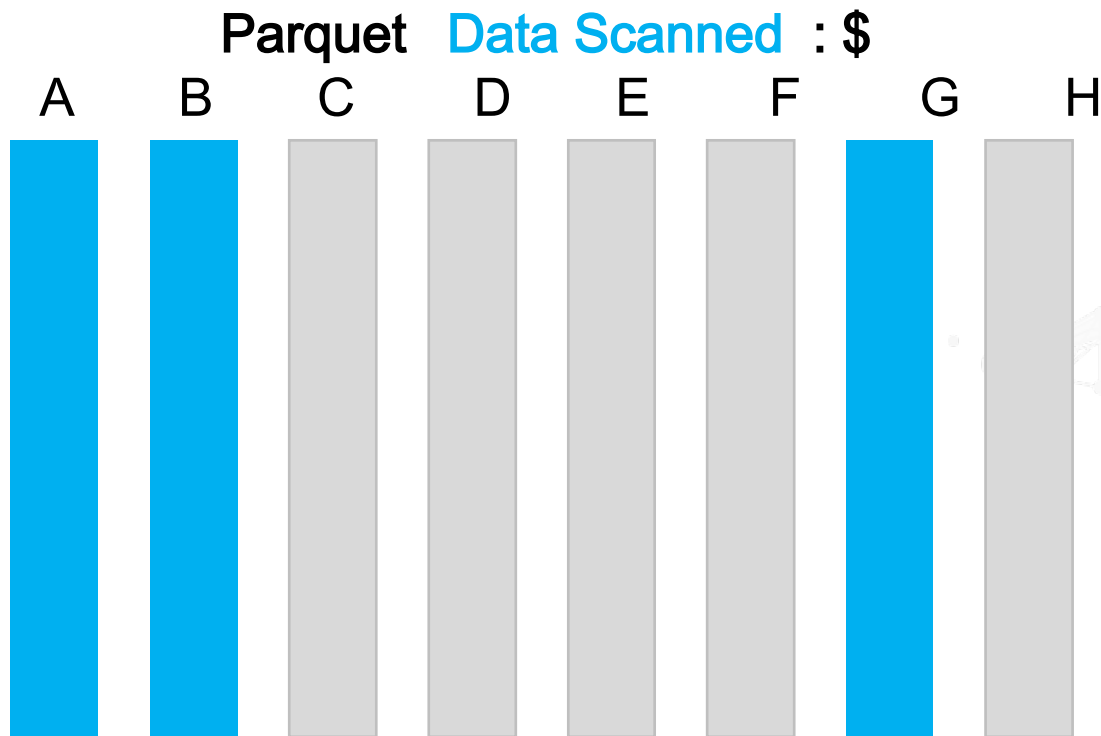
Example Query Strategy: Columnar Stored (Parquet) vs. CSV

Query: Group By A, B, average G



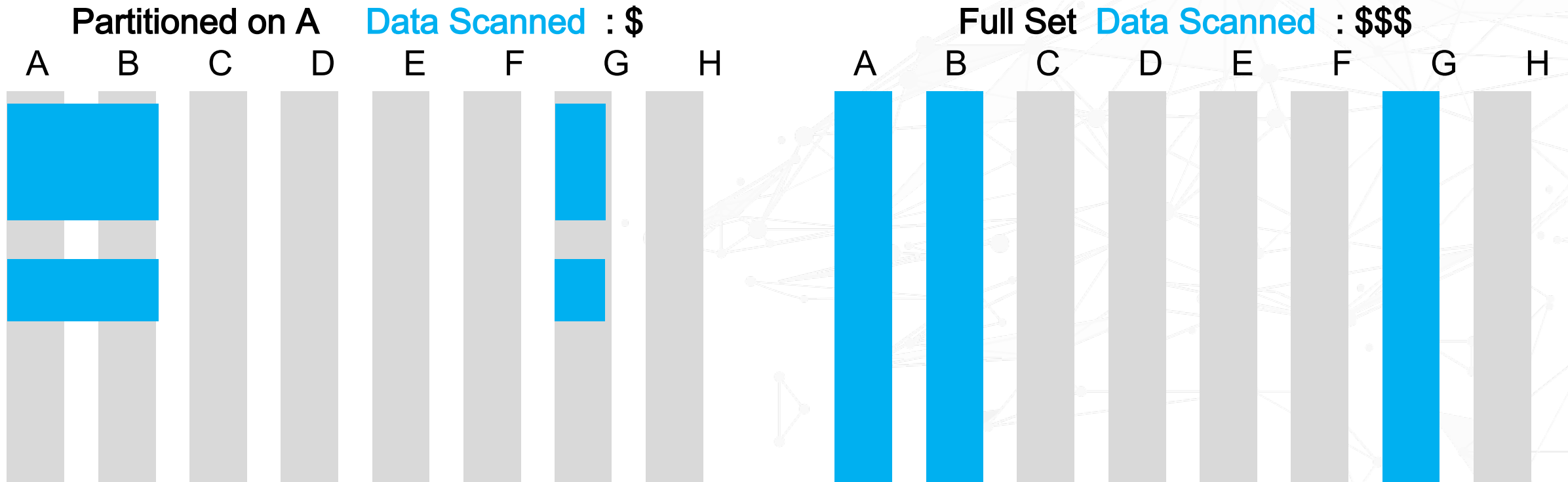
Example Query Strategy: Columnar Stored (Parquet) vs. CSV

Query: Group By A, B, average G

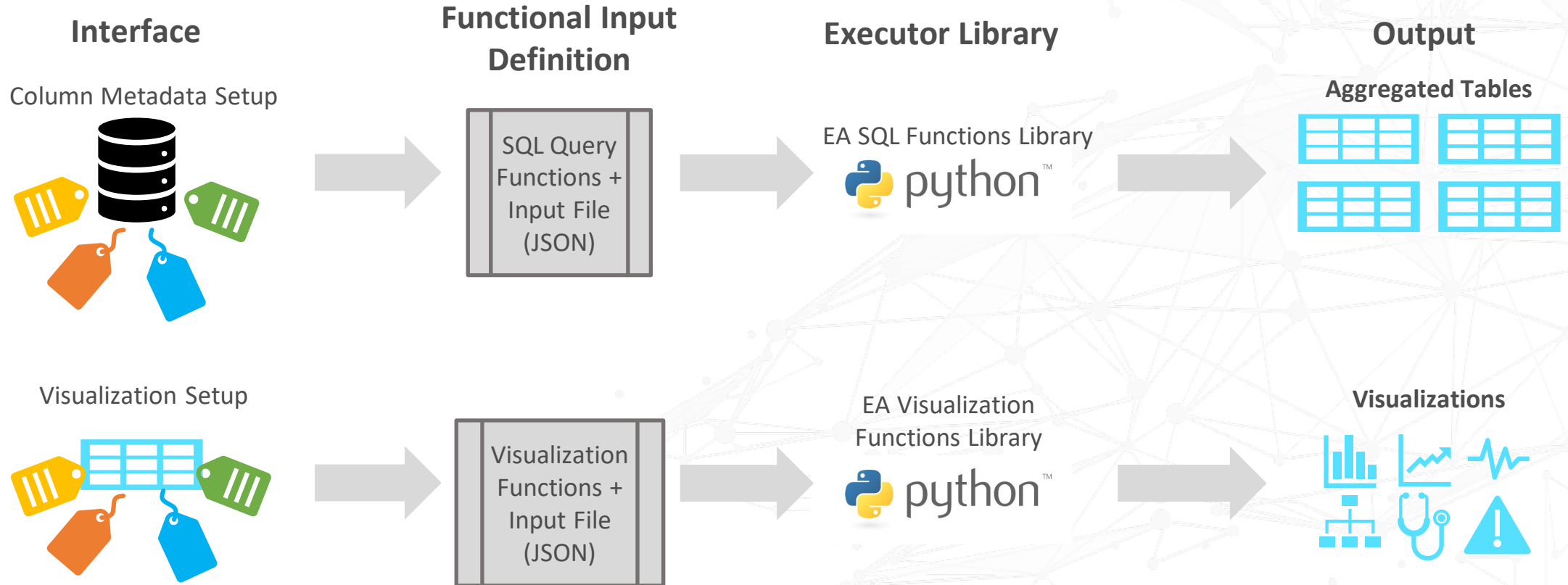


Example Query Strategy: Partitioning Data before Querying

Query: Group By A, B, average G where A=Alpha



Reproducible Queries, Analysis, and Visualizations





Data Quality: Demo of Findings on Alternative Datasets



Closing Remarks

Closing Remarks

- Alternative data is here to stay!
- Alternative data complements existing inputs
- Having an alt data plan is all important for success
- Not all data is created equal. Investing time in robust data quality testing is essential for alternative data.



- Join [Eagle Alpha's Virtual Data Conference September 8th - 10th](#)

Q & A

Add your questions to the chat room

Eagle Alpha
Lead Sponsor **J.P.Morgan**

Virtual Data Conference

3 Half Days - 50 Alternative Data Industry Leading Experts
Book 1-on-1 Meetings With Over 100 Vendors

September 8th - 10th 2020

[REGISTER TO ATTEND](#)

AWS Data Exchange snowflake®

INTEGRATING DISRUPTIVE INNOVATION INTO AN ORGANIZATION'S DNA

"Innovation is the Key to Growth"
-ARK Investment

Cathie Wood,
CEO & Founder, Ark Investment

Investing at the Pace of INNOVATION,
Identifying LONG-TERM GROWTH,
Bringing Together AGILE, DIVERSE TALENT

September 9th @ 4pm EDT

September 9th @ 4pm EDT

FDP: THE GLOBAL DESIGNATION FOR FINANCE PROFESSIONALS IN A DATA-DRIVEN INDUSTRY

Live webinar with our Candidate Relations and Curriculum Team.
September 16 @ 1pm EDT

Candidate Orientation

Strategy → Execution → Success

September 16th @ 1pm EDT

The Ethical Use of Machine Learning in Financial Markets - Myth or Miracle?

Daniel Liebau
Founder
Lightbulb Capital

Michael Weinberg
Managing Director,
Head of Hedge Funds
and Alternative Alpha,
APG Asset Management

October 2, 2020
@ 11:00 AM EDT

October 2nd @ 11 am EDT



In Closing

Registration for any FDP
webinars are free.

Click on the webinar of your
choice for your registration.

Eagle Alpha
Lead Sponsor **J.P.Morgan**

Virtual Data Conference

3 Half Days - 50 Alternative Data Industry Leading Experts
Book 1-on-1 Meetings With Over 100 Vendors

September 8th - 10th 2020

[REGISTER TO ATTEND](#)

**INTEGRATING
DISRUPTIVE
INNOVATION
INTO AN
ORGANIZATION'S
DNA**

"Innovation is the Key to Growth"
-ARK Investment

Cathie Wood,
CEO & Founder, Ark Investment

Investing at the Pace of INNOVATION,
Identifying LONG-TERM GROWTH,
Bringing Together AGILE, DIVERSE TALENT

September 9th @ 4pm EDT

September 9th @ 4pm EDT

FDP:
THE **GLOBAL
DESIGNATION**
FOR FINANCE
PROFESSIONALS
IN A DATA-DRIVEN
INDUSTRY

**Candidate
Orientation**

Live webinar with our Candidate Relations and Curriculum Team.
September 16 @ 1 pm EDT

September 16th @ 1pm EDT

Daniel Liebau
Founder
Lightbulb Capital

**The Ethical Use of
Machine Learning
in Financial Markets
- Myth or Miracle?**

Michael Weinberg
Managing Director,
Head of Hedge Funds
and Alternative Alpha,
APG Asset Management

October 2, 2020
@ 11:00 AM EDT

October 2nd @ 11 am EDT

More details on webinars *and* webinar recordings can be found on

www.fdpinstitute.org/webinars

www.caia.org/caia-infoseries

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

