Fuse Cyber Data to Supercharge Machine learning and A.I.

Dr. Sameer Joshi

CEO, Datanova Scientific

Datanova scientific

Data Unifier

Fusion Hub

Today’s agenda

Top level understanding of data for Machine Learning and A.I.

Data variety is the biggest obstacle for Machine Learning in cyber.

This obstacle can be overcome with data fusion.

Today’s takeaway – Data Fusion is an indispensable part of machine learning & A.I.

Automated data utilization is key to cyber dominance

More Data

- Sensors and other sources are generating more data everyday
- Far too much for the human cognitive threshold

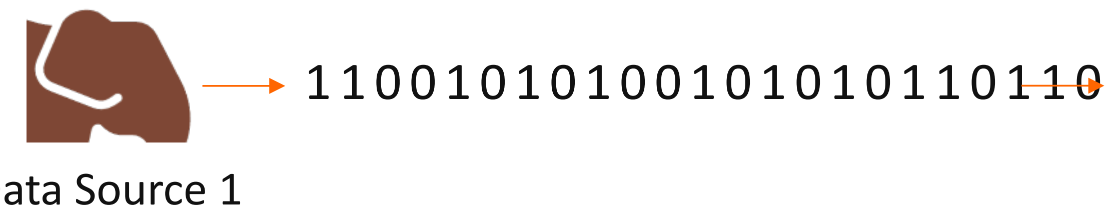
More Coverage

- More of the domain is reflected in the data
- More opportunities for discovery and optimization

More Problems

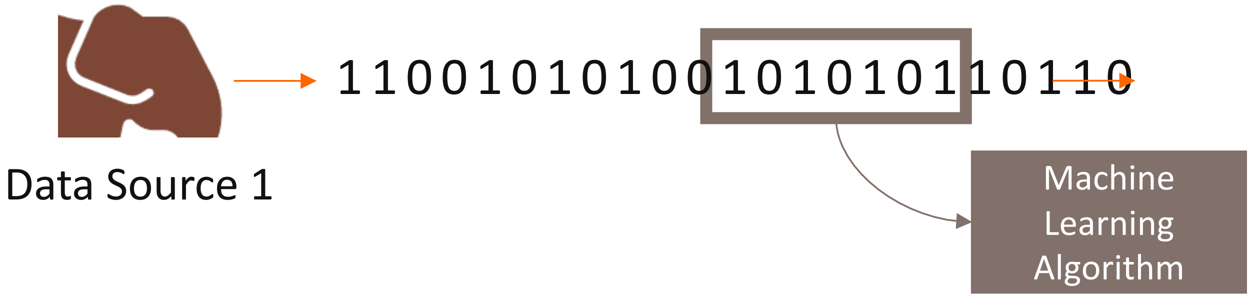
- The size and speed of the data are well understood issues.
- The variety of the data is the challenge of the day

Training machine learning



Let’s consider a single cyber source

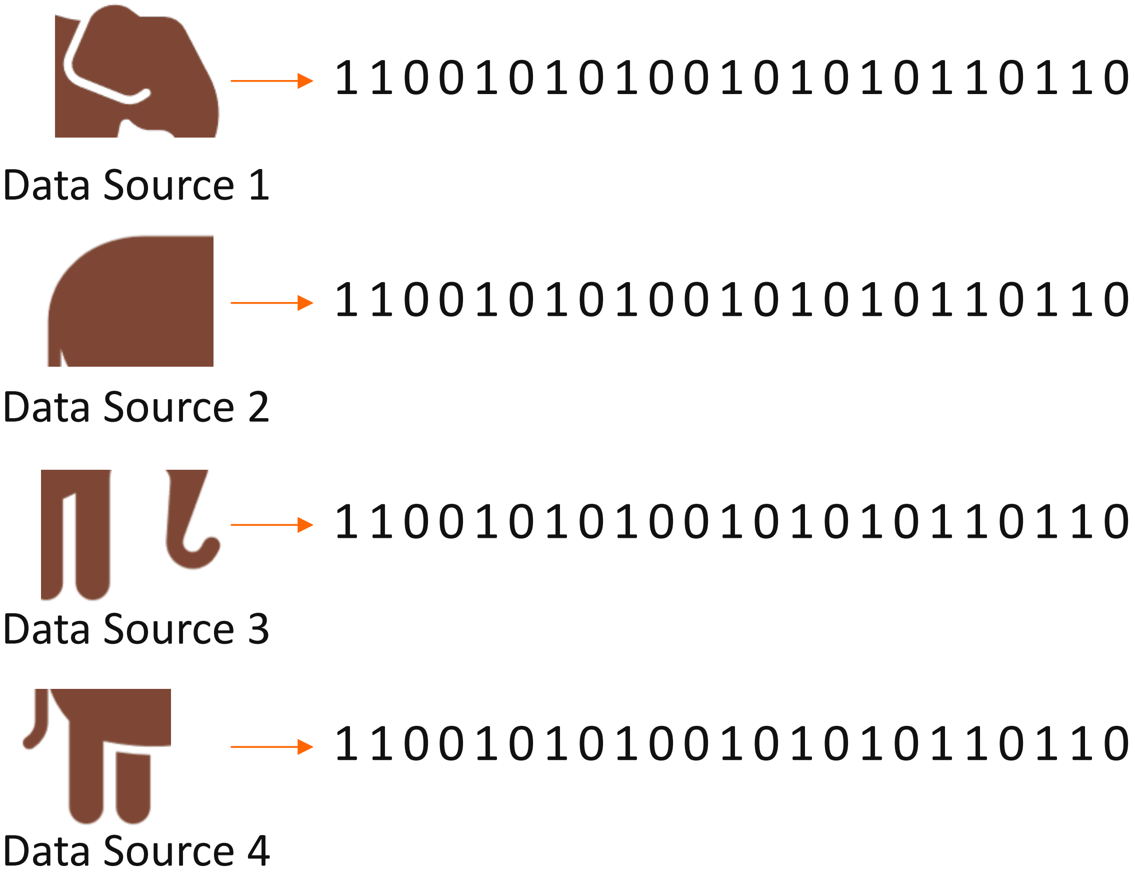
Training machine learning



Training on a single source

- Very common, and very well understood
- The primary challenges are the speed, richness, and quality of data
- Insights are limited to what this single source can offer

The real world is more complex



1100101010010101011 0110

Data Source 1

11001010100101010110110

Data Source 2

11001010100101010110110

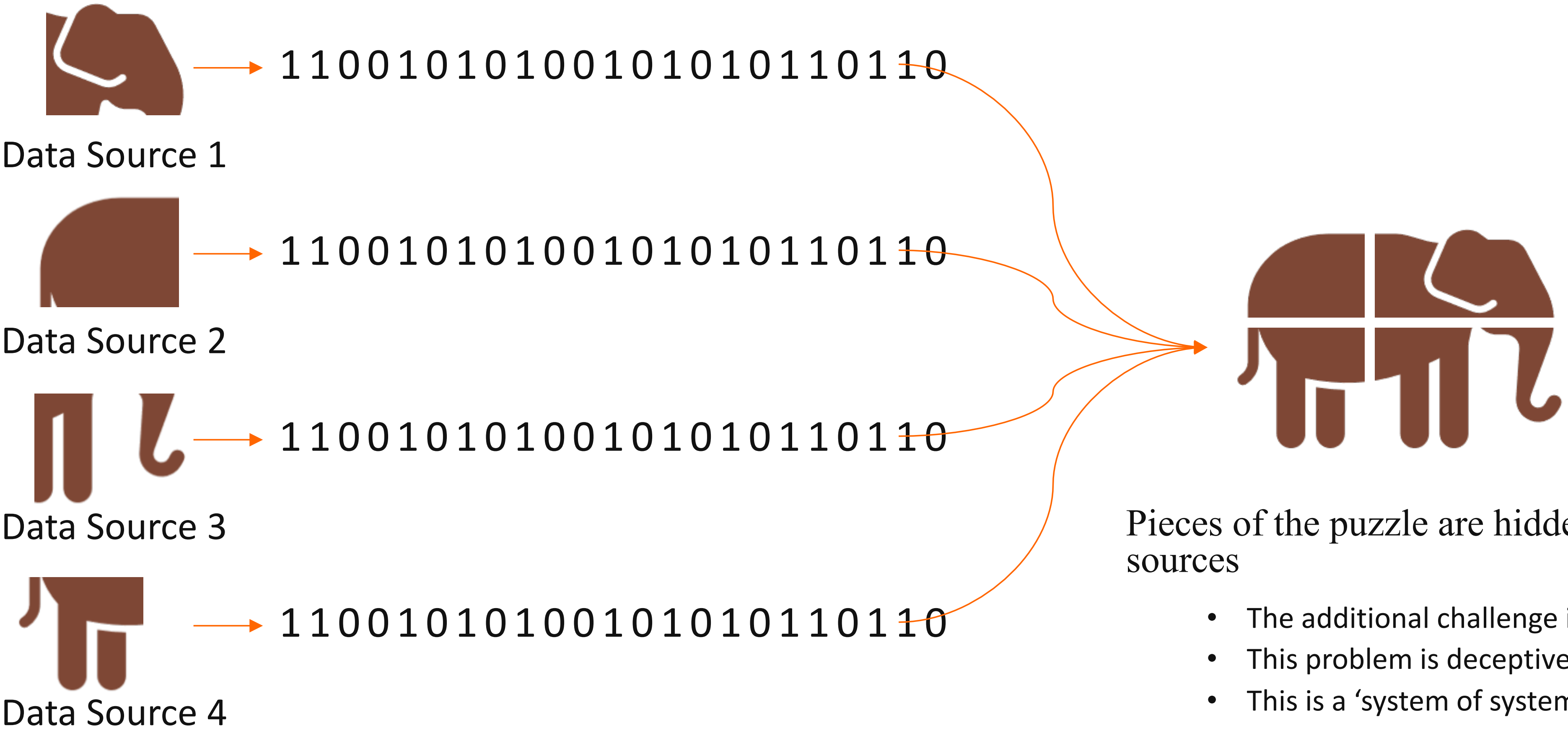
Data Source 3

11001010100101010110110

Data Source 4

There are many different data sources, and….

The real world is more complex



Data Source 1

Data Source 2

Data Source 3

Data Source 4

Pieces of the puzzle are hidden across various data sources

- The additional challenge is the variety of data
- This problem is deceptively non-trivial
- This is a ‘system of system’ with moving parts

Challenges for multiple source learning

Can I just use machine learning to solve this problem as well? ...No

- Very high sample complexity
- No easy way to stitch models together

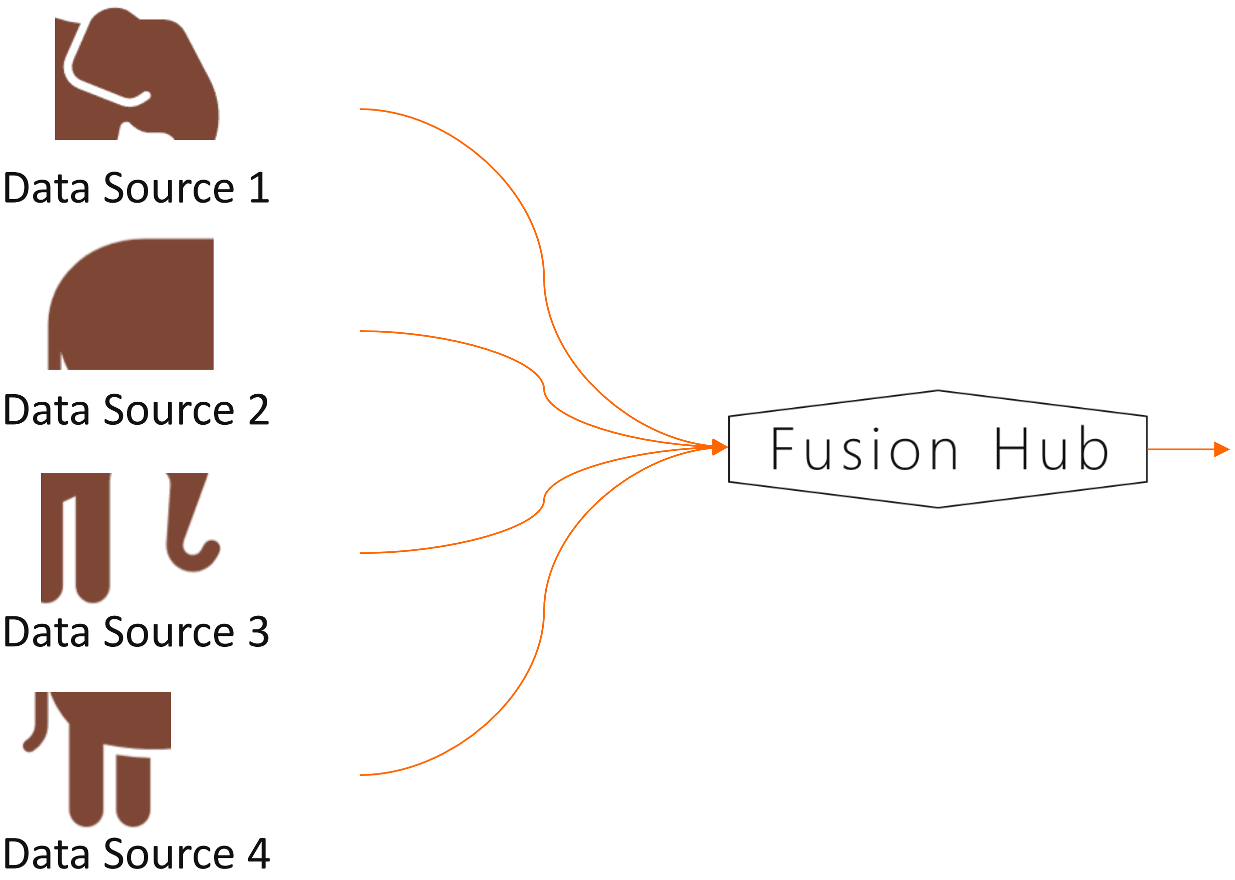
Can I prepare the data by hand? …No,

- Too expensive
- Forbes study* says data scientists spend 80% of their time working data
- This cost is compounded for multiple data sources

* https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says

So what is the path forward?

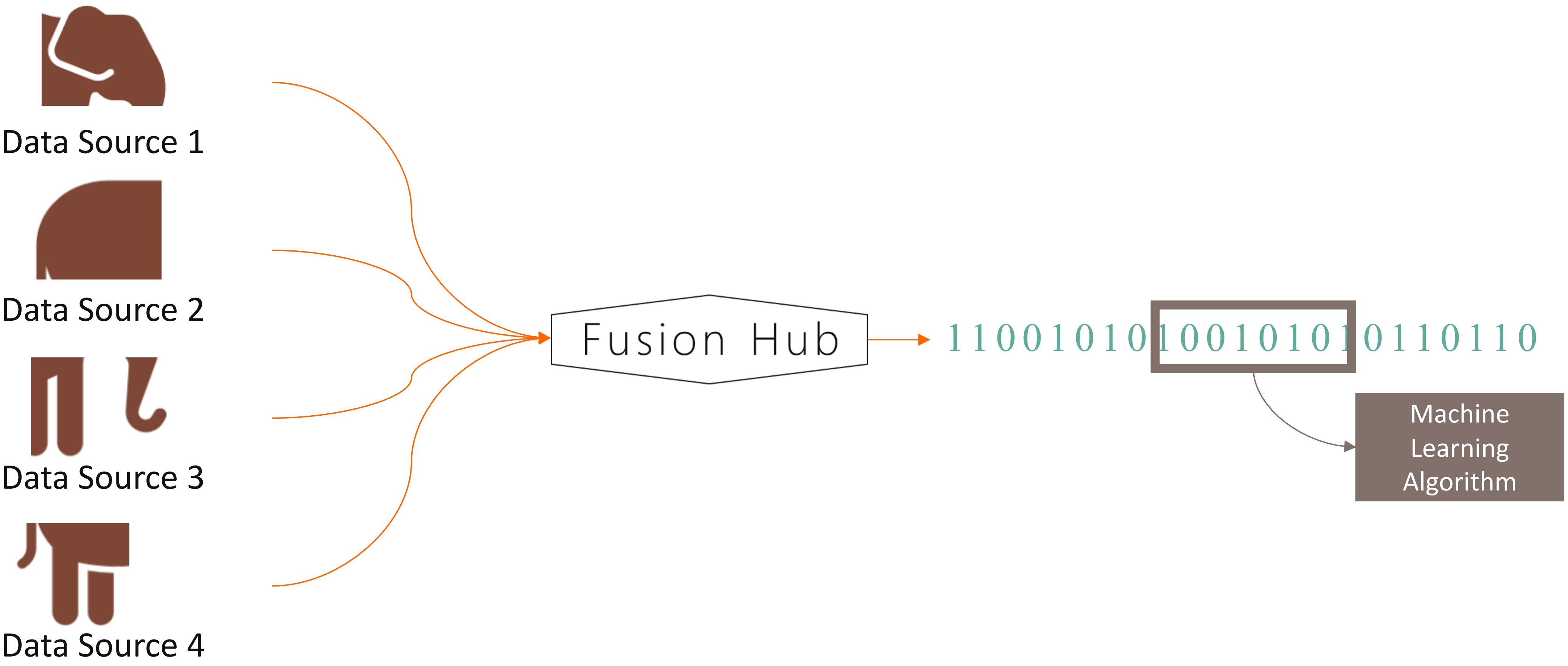
Solution - use Data Fusion to create a single source of truth across all data before learning



Fused data is:

- Standardized in every way (names, types, units, conventions, etc.)
- Has one record for one real-world entity
- Data Trust built-in
- Data Quality built-in

This creates a single, enriched, high quality data baseline for machine learning



Data Source 1

Data Source 2

Data Source 3

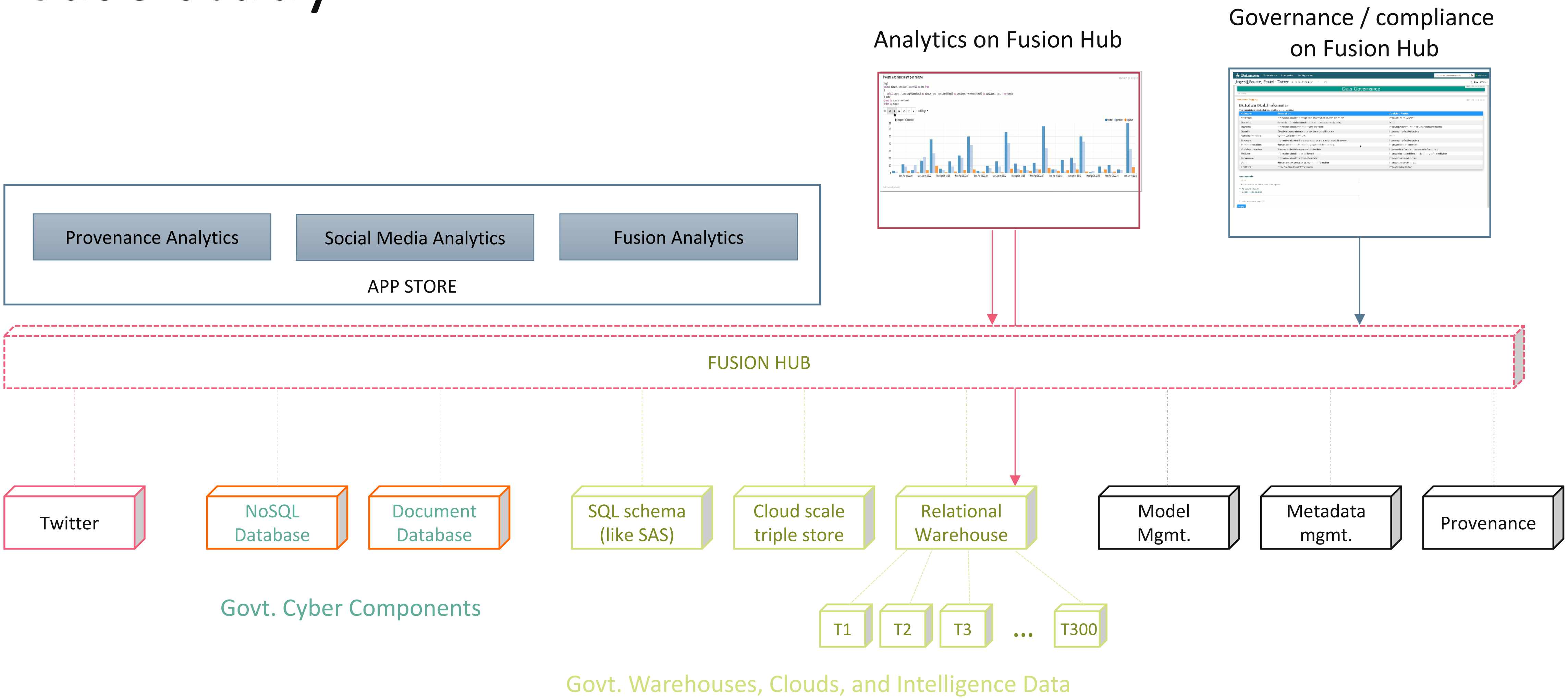
Data Source 4

Fusion Hub

11001010100101010110110

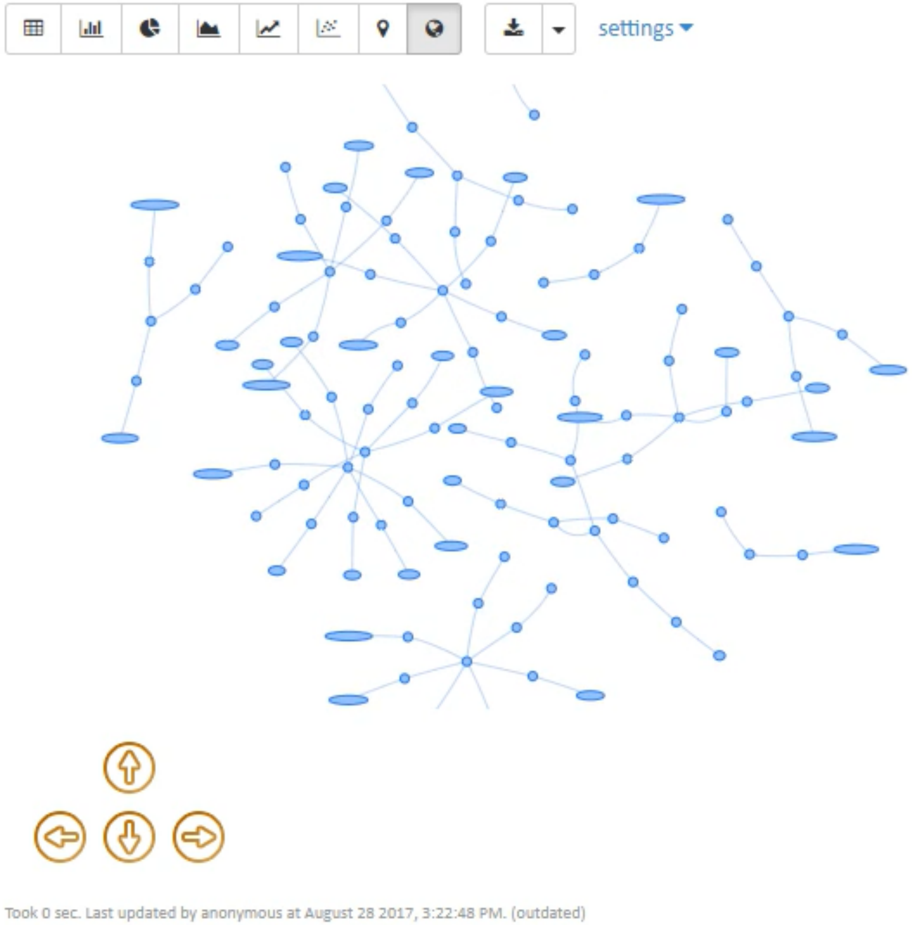
Machine Learning Algorithm

Case study


Analytics on Fusion Hub
Governance / compliance on Fusion Hub
Provenance Analytics
Social Media Analytics
Fusion Analytics
APP STORE
FUSION HUB
Twitter
NoSQL Database
Document Database
SQL schema (like SAS)
Cloud scale triple store
Relational Warehouse
Model Mgmt.
Metadata mgmt.
Provenance
T1
T2
T3
…
T300
Govt. Cyber Components
Govt. Warehouses, Clouds, and Intelligence Data

Start with messy Raw data across sources

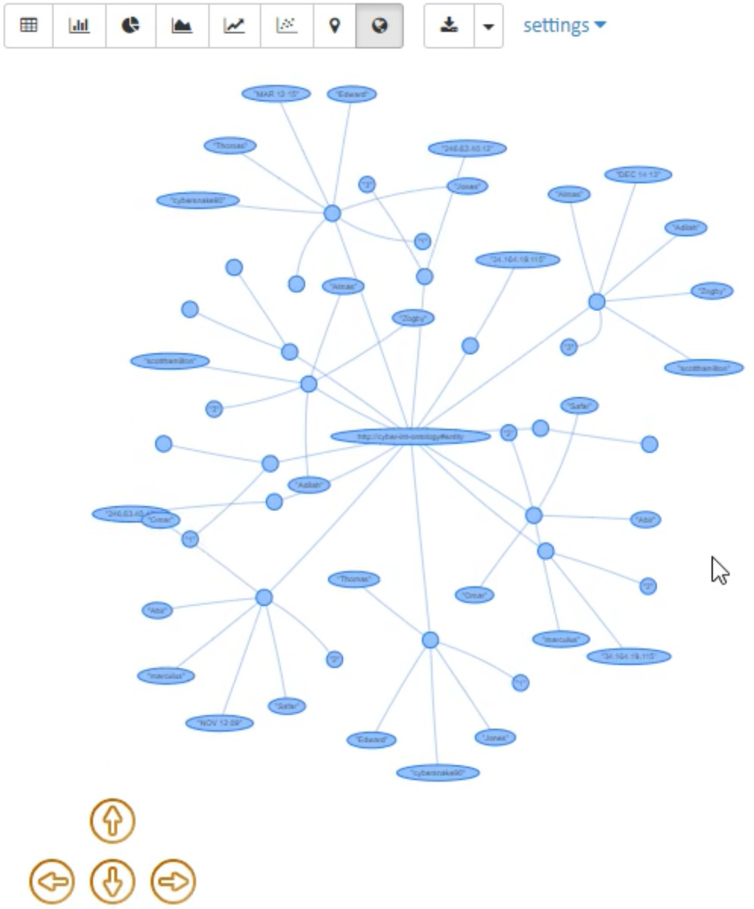
1



Took 0 sec. Last updated by anonymous at August 28 2017, 3:22:48 PM. (outdated)

Unified data is orderly, but still voluminous

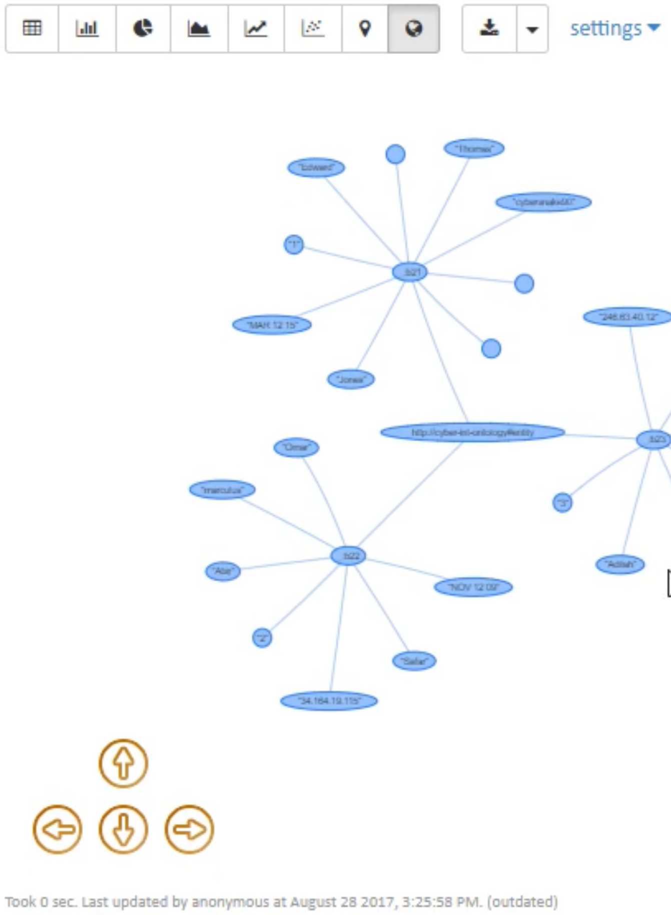
2



Took 1 sec. Last updated by anonymous at August 28 2017, 3:22:58 PM. (outdated)

Fused data is perfect for machine learning

3



Took 0 sec. Last updated by anonymous at August 28 2017, 3:25:58 PM. (outdated)

Benefits of fusion for machine learning & A.I.

- Reduces sample complexity dramatically (i.e., makes machine learning possible)
- High-quality and consistent data baseline
- Eliminates repetitive data prep
- Auto-discovers new data sources