figure eight FEDERAL

Data is King: Let's Talk About it

December 2021

Vinay Malkani
Chief of Engineering
Vinay.Malkani@F8-Federal.com

Jack Hild
Figure Eight Federal Team Member, former CIO of Digital Globe, NGA Hall of Famer
Jackhildgeo@gmail.com

Table of Contents

Our Experience

Among others

Figure Eight Federal is critical in the creation of the highest quality decision-grade AI for leaders engaged in advancing America's security and competitive position

Experience

Domain expertise

- 15 Years+ enabling AI projects
- 13 Billion+ human annotation judgements
- *Commercial: Apple, Oracle, eBay, Adobe, IBM, Boeing, Raytheon, etc.*

On prem and cloud deployment options available

- Robust API Structure
- Ready-to-use infrastructure

Computer Vision: FMV, SAR, SYERS,EO/IR, WAMI , Tiled Imagery

Natural Language Processing (NLP) supported for 180+ languages

Supports all data types

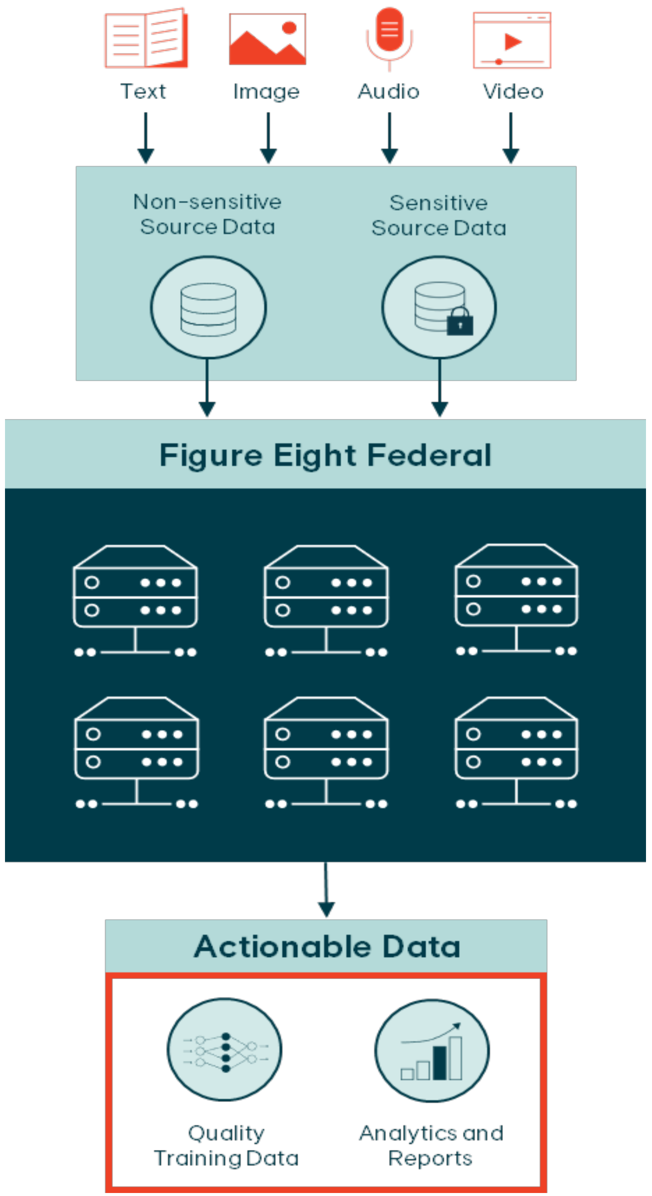
- Text, image, audio, video, 3D, etc.
- 500k+audio hours processed
- 100M+ images labeled

High quality machine learning training data generation via:

- Labeling unstructured data
- Ingest and peer review of existing labels
- Relabeling of low confidence predictions, and more

Machine Learning enhanced workflow

- Automation of multistep annotation projects
- Pre-classification/Pre-annotation



Our Offerings

Services we provide

Prelabelled Datasets	Data Enrichment	Data Annotation & Synthetic Data Generation	Model Development and Testing
Kickstart your AI project with prelabelled datasets including synthetic	Leverage platform to provide meta data infusion; acquire high quality unbiased data	Provide Platform and Crowd to accurately and efficiently label training data	Validate real-world model performance across a range of use cases and demographics

Data types we support

- Text
- Image
- Audio
- Video
- Point cloud
- Multi-modal

AI use cases we support

- Social media
- Healthcare
- Security
- Targeting Solutions
- AR/VR
- ISR
- Document processing
- Autonomous vehicles
- Vision/ GEOINT

Our products

Data annotation platform

- Image and Video Annotation and Transcription
- Text Annotation and Translation
- Audio Annotation and Transcription
- Data Collection and Enrichment
- Data Classification
- Point Cloud Annotation
- Machine Learning Assisted Smart Labelling
- Workflows
- Secure workspace
- Quality management

Workforce options

- F8F Global Crowd
- F8F Secure workforce
- F8F U.S only custom crowd

Service options

Data Pipeline Management, Self-managed via API, Model Repository

Design Engineering Services, Managed service

Data

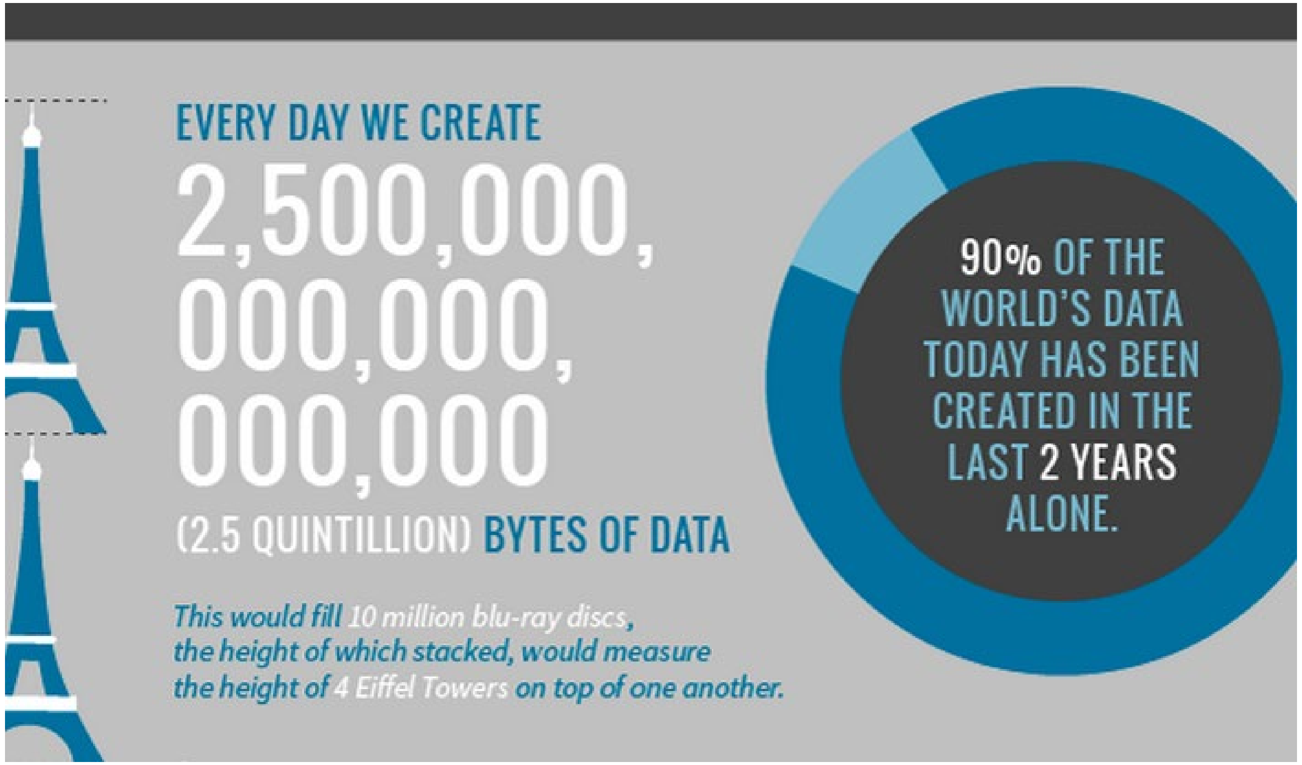
DIKW
Data Transformation
Data Quality
Data Fusion

Data is Everywhere

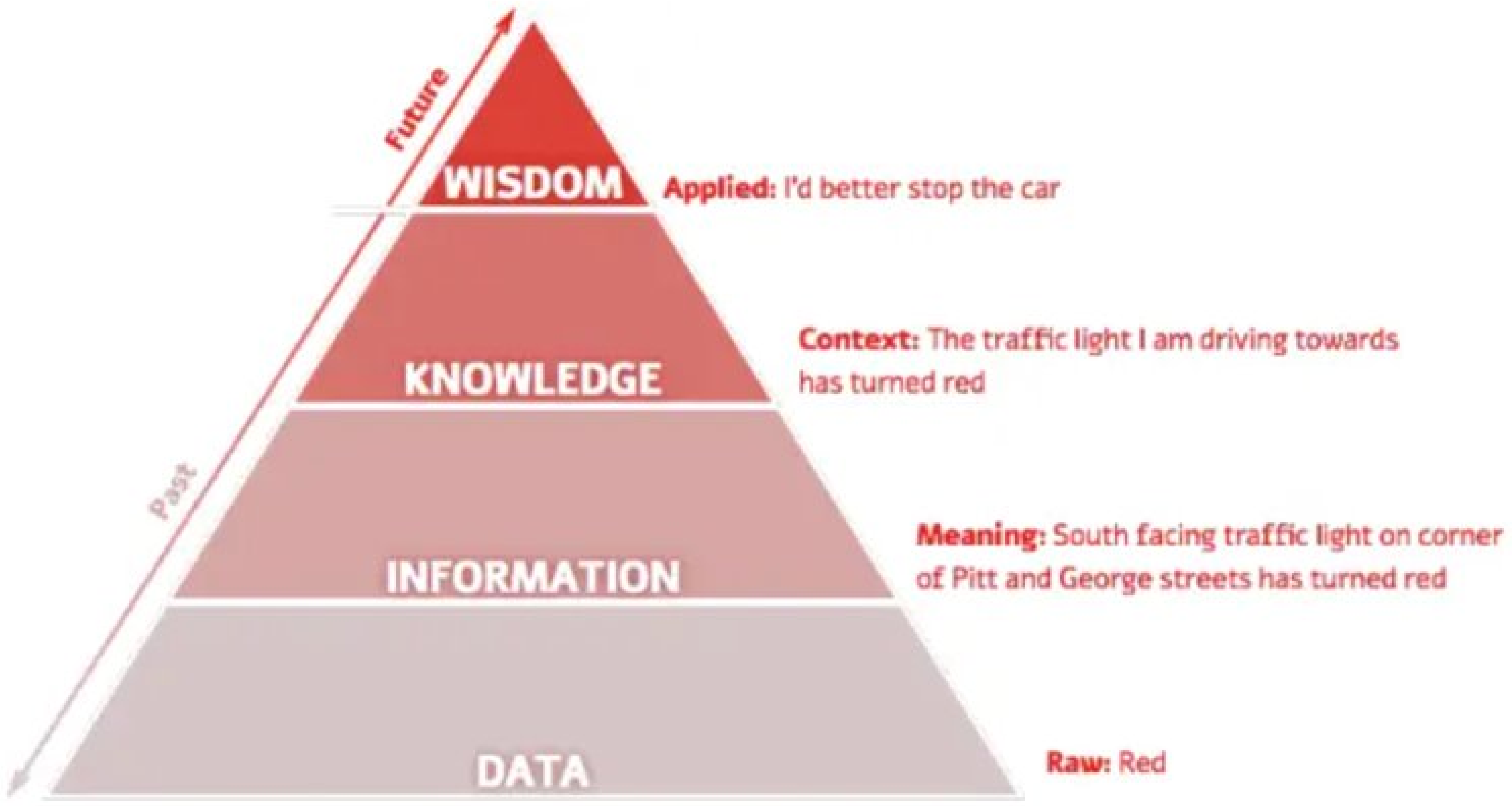
- We created 2.5 quintillion data bytes daily in 2020. *(Forbes)*

- 15% of the content on Facebook is video *(Social Insider)*

- 463 exabytes of data will be generated each day by people as of 2025. *(Raconteur)*

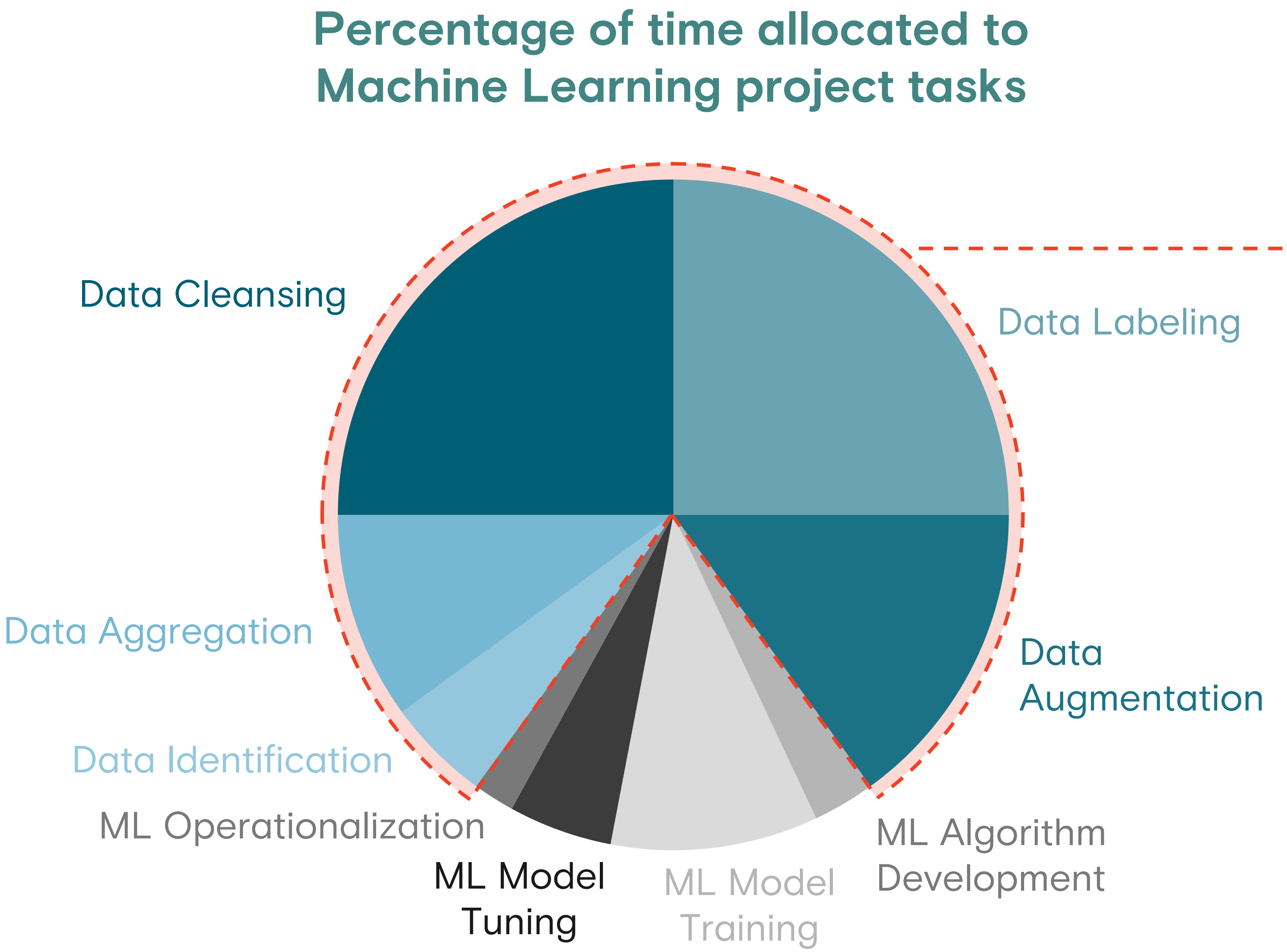


DIKW Model



A traditional data-information-knowledge-wisdom pyramid – source Mushon

Data Transformation



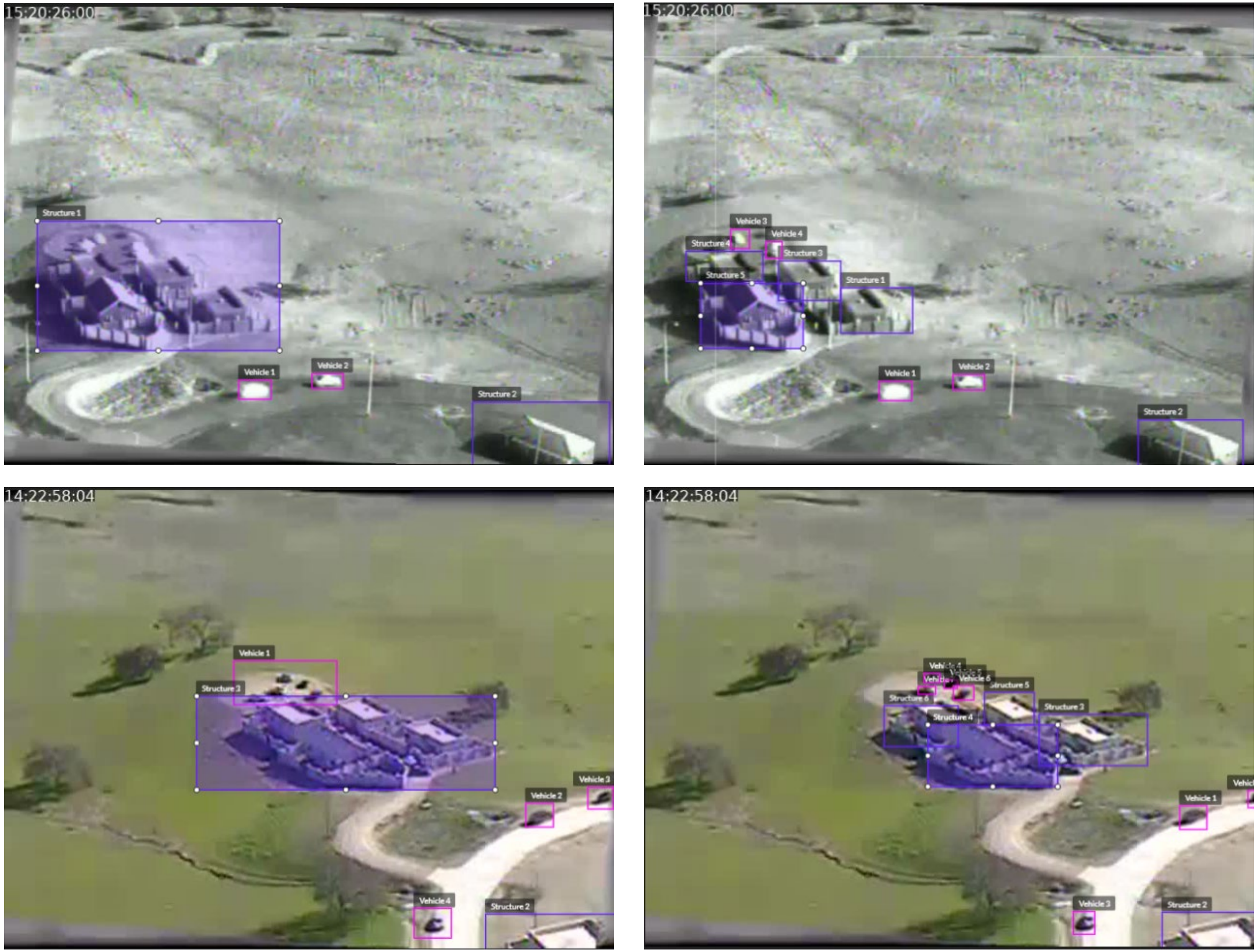
"The hardest part of data science is getting good, clean data. Cleaning data is often 80% of the work"

DJ Patil- 2016 US Chief Data Scientist

"We've trained the model on a particular training data set. But that data set is not representative of global terrain or global information…so when you think of the diversity…the training data set from a testing and representative perspective is so important."

Nand Mulchandani - 2020 CTO, JAIC

Data Quality



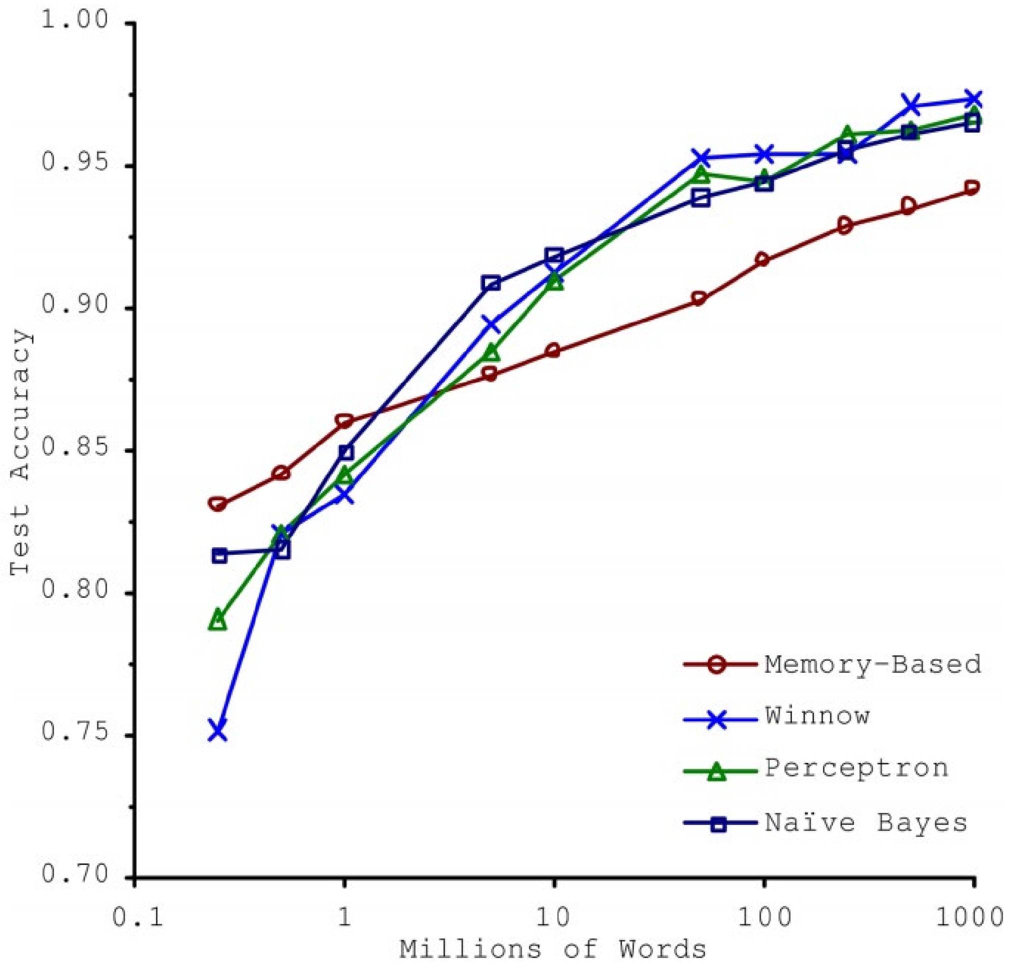
According to an IBM study, poor data quality cost the United States 3.1 trillion dollars.

Left: Poor Data, Right: Good Data

Data Quality Issues: Insufficient Data

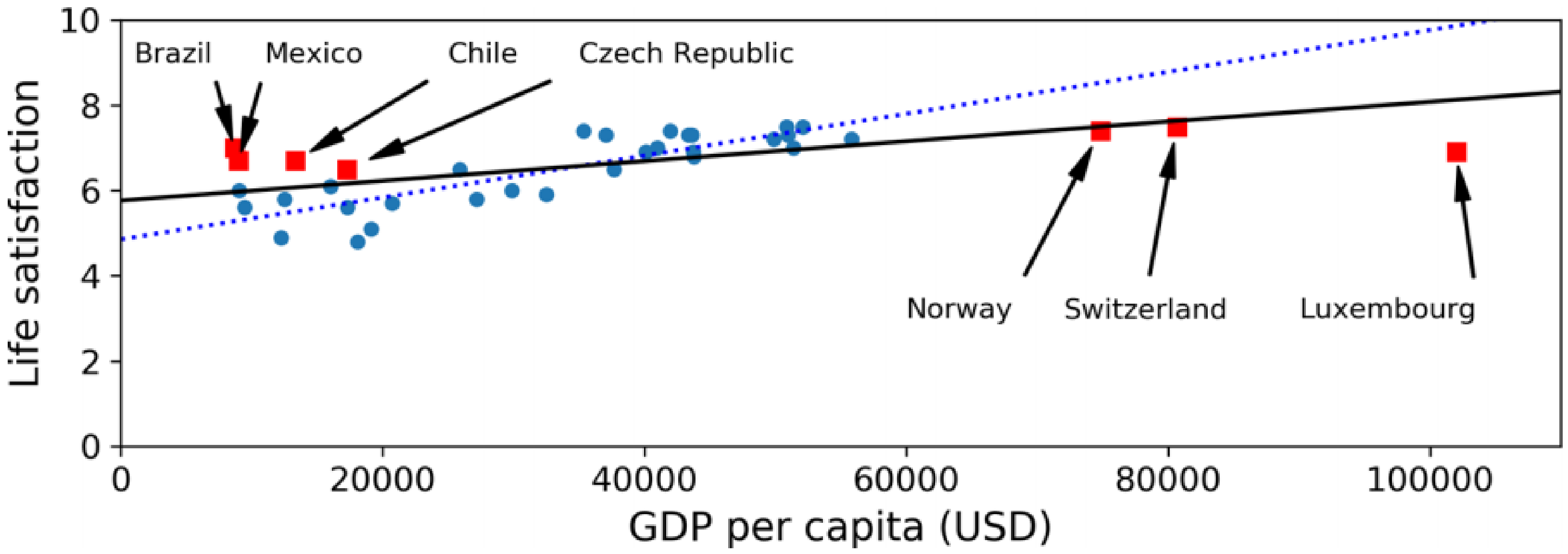
It takes a lot of data for most Machine Learning algorithms to work properly.

Even for very **simple problems** you typically need **thousands** of examples, and for **complex problems** such as image or speech recognition, you may need **millions** of examples (unless you can reuse parts of an existing model).



Data Quality Issues: Non-Representative Data

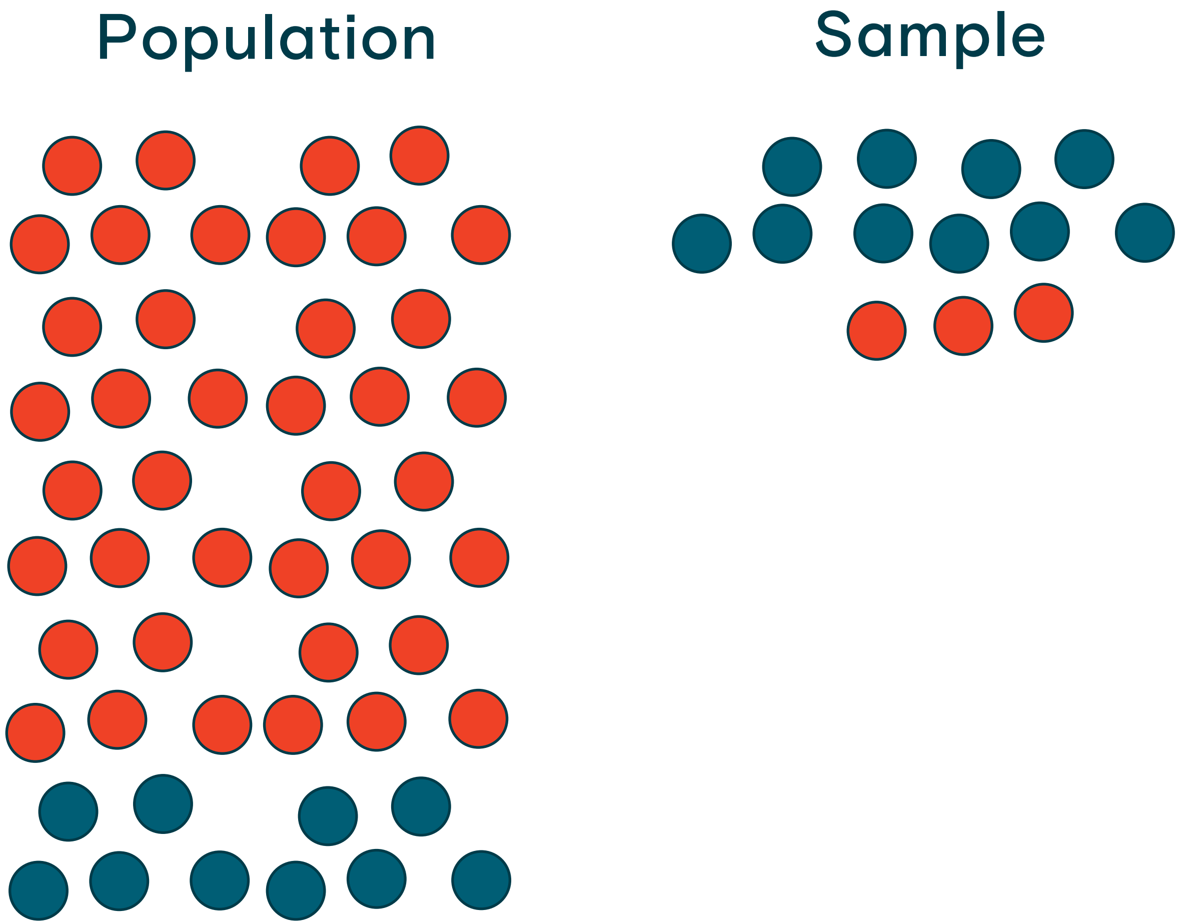
In order to ***generalize*** well, it is crucial that your training data be ***representative*** of the new cases you want to generalize to. This is true whether you use instance-based learning or model-based learning



Solid line: linear model on new data, Dotted Line: old model

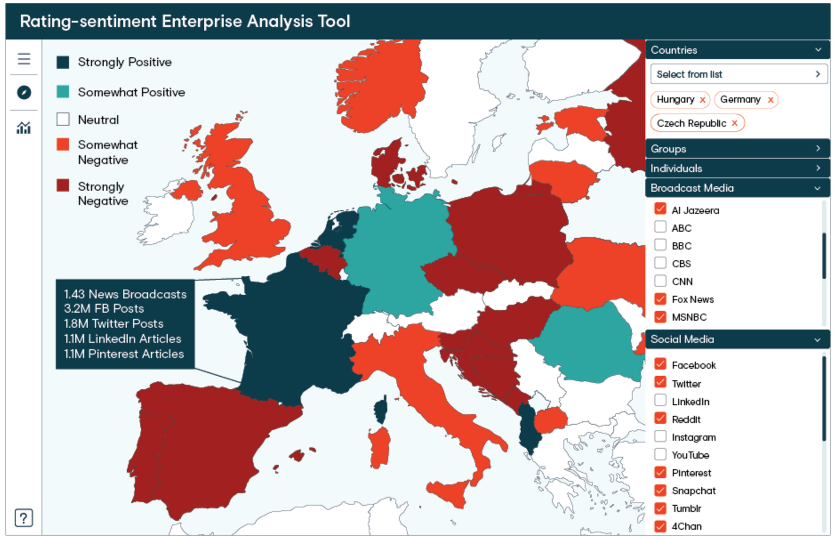
Data Quality Issues: Bias

If the sampling method of the data is flawed, samples can be nonrepresentative. This is will create bias (sampling bias).



Data Fusion

Data fusion techniques combine data from multiple sensors and related information from associated databases to achieve improved accuracy and more specific inferences than could be achieved using a single sensor alone



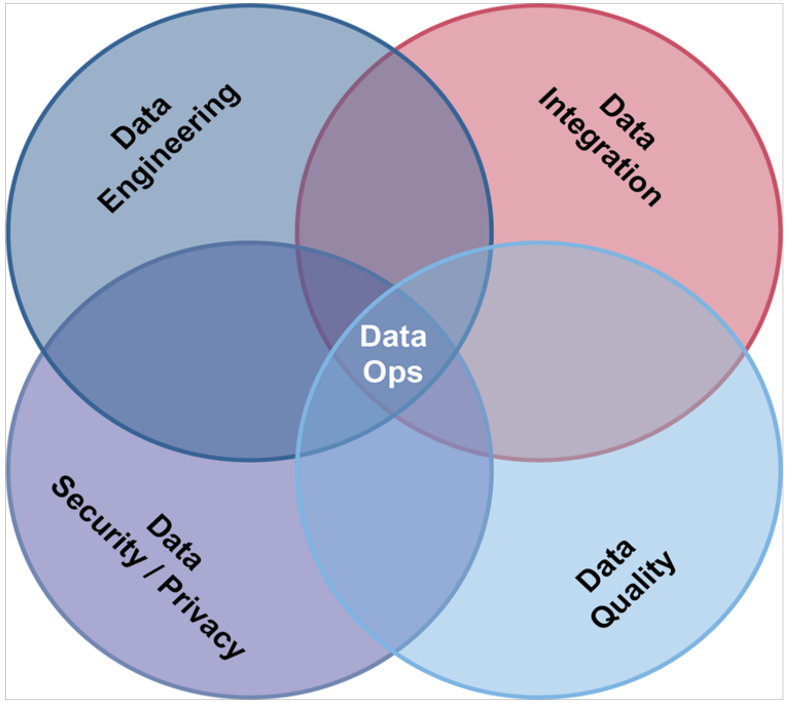
This allows you drive up confidence in your data is transformed into knowledge.

figure eight FEDERAL

Data Automation

Data Automation

Begin with Data Automation In Mind



PRESSURE FROM BOTH ENDS OF THE STACK!

From the top of the stack, more users want access to more data in more combinations. And from the bottom of the stack, more data is available than ever before — some aggregated, much of it not.

The only way for data professionals to deal with pressure of heterogeneity from both the top and bottom of the stack is to embrace a new approach to managing data that blends operations and collaboration to organize and deliver data from many sources to many users reliably with the provenance required to support reproducible data flows.

-Andy Palmer, 2015

Data Engineering for Machine Learning Best Practices

Appoint a Data Custodian

- Data custodians are responsible for the safe custody, transport, storage of data and implementation of business rules.

Know your data types including:

- Available input formats
- Desired output formats
- Average, maximum and minimum file sizes

Plan your data pipeline and capture data provenance

- Create a data catalogue
- Create a naming convention for your data and ensure that the convention includes markers for tracing data

Create an audit log for source data with traceable lineage to final data format

- Blockchain could be useful if you need immutability, but simpler options may be sufficient for your business needs
- Create hashes of datafiles for unique identifiers

Use extensible tooling for data labeling

- Your tools should not require an engineer to make changes in how you interact with your data

Automate your data pipeline

- Your labels should be traceable back to your source data and should be immediately recognizable

Continued

Use a small sampling of your data and test your pipeline before running large batches

Mark your batches of data

- Similar to consumer products, data batches should be easily identifiable

Verify your batches through personal review

Understand your quality requirements and thresholds for data labeling

- Define how you measure the quality of your data
- Set success criteria based on desired outcomes, not the data itself

Retraining your model is a pipeline itself

- Consider at least two paths
- Ensure ability to identify and relabel data
- Define a specified path to send the marked data to your vendor

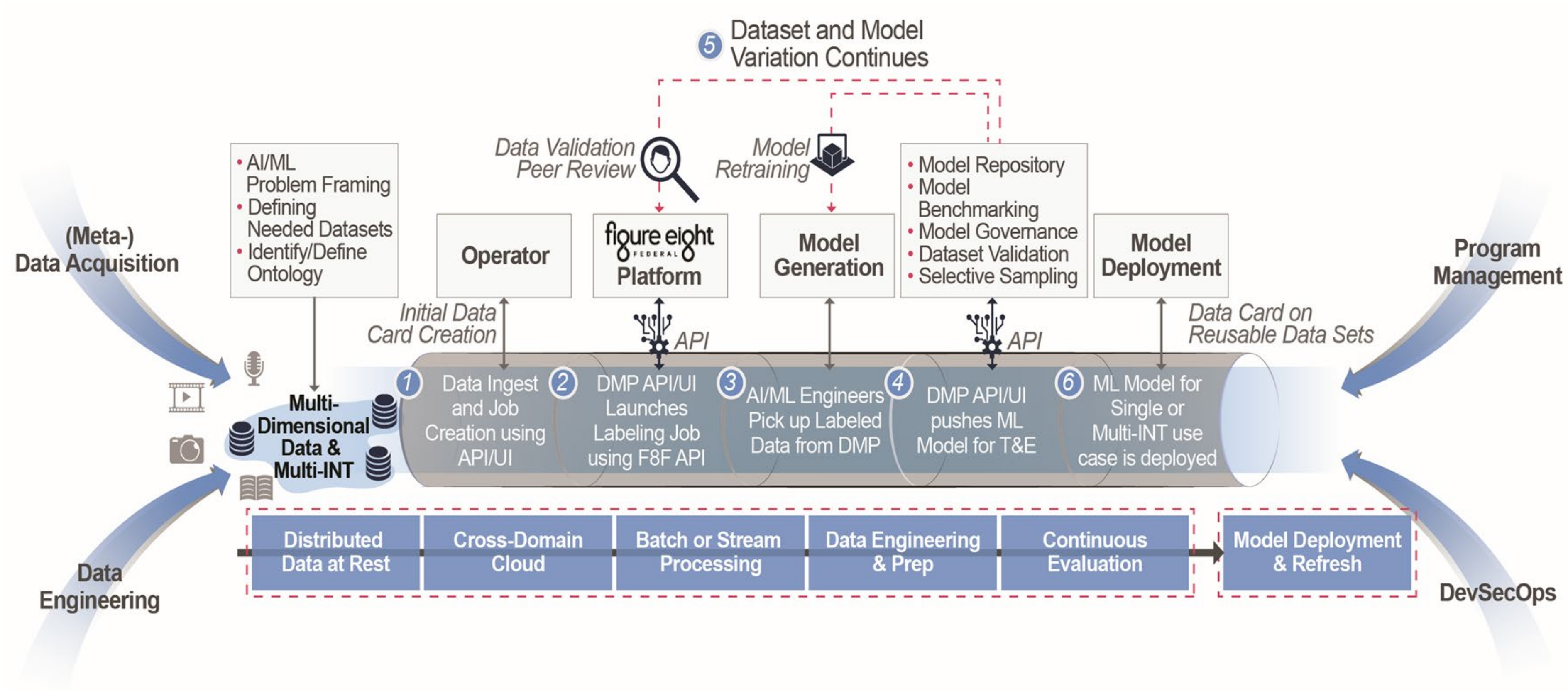
Understand that bad data is worse than no data

- Poor data is costly to fix, and may require starting from scratch

Store your ML weights in a datastore that can handle historical testing

- This allows for you to map progress as you train and retrain your model using your data
- It can be stored with your code using git LFS

Figure Eight Federal Approach

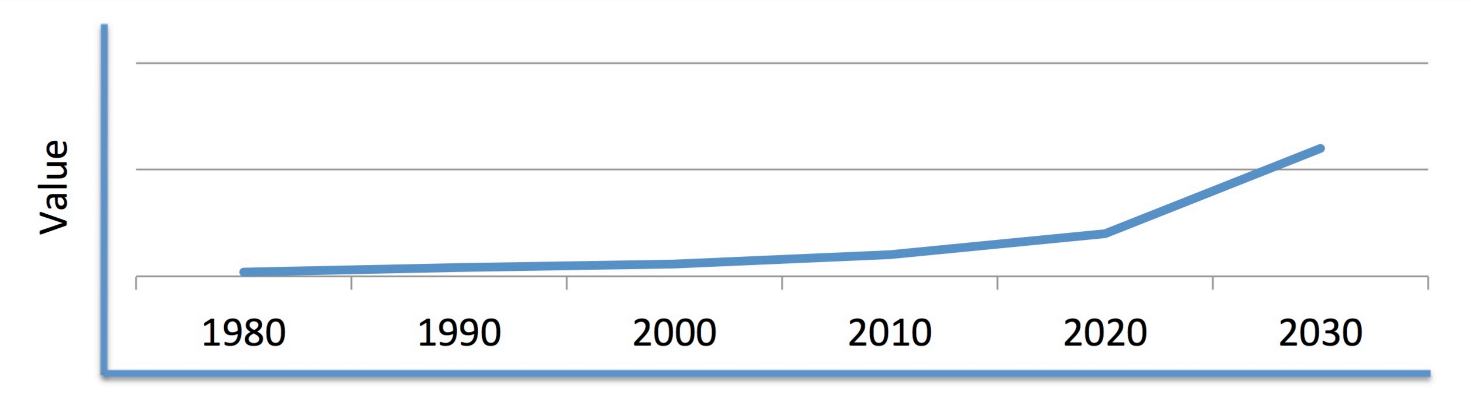


So What?.... a GEOINT Perspective

- **Data Relevance** – Getting the right data right
- **Context** – Bringing the pieces together.... Knowledge, Understanding, Wisdom
- **DIKW Latency** – Great Data + Speed = Operational Success

Historical Context – GEOINT and ML/AI

- 1980s
  - AFE - Automatic Feature Extraction
  - ATR - Automatic Target Recognition
- 1990s
  - Beyond hardcopy imagery and maps
  - GPS set free
  - "The Tsunami of Data" in anticipation of the expected output of commercial imagery
- 2000s
  - The War on Terrorism and the explosion of manned and unmanned motion imagery
  - Google Earth and location data on the internet
  - Gen 1 commercial imagery
  - Open-source spatial data – HLS, Navigation, Ag, Energy, Disaster Mgt....
- 2010s
  - More of the above
  - Small satellite commercial imagery
  - Commercial SAR
- 2020s
  - Much more of all the above
  - HSI



Operational value of ML/AI initiatives

Data Relevance

- Getting the right data…for GEOINT
  - Still a challenge even with the Tsunami
  - Open-source PoI greatly improved, but is it current enough?
  - GEOINT needs more training data
  - Spatially relevant text and audio a growth area

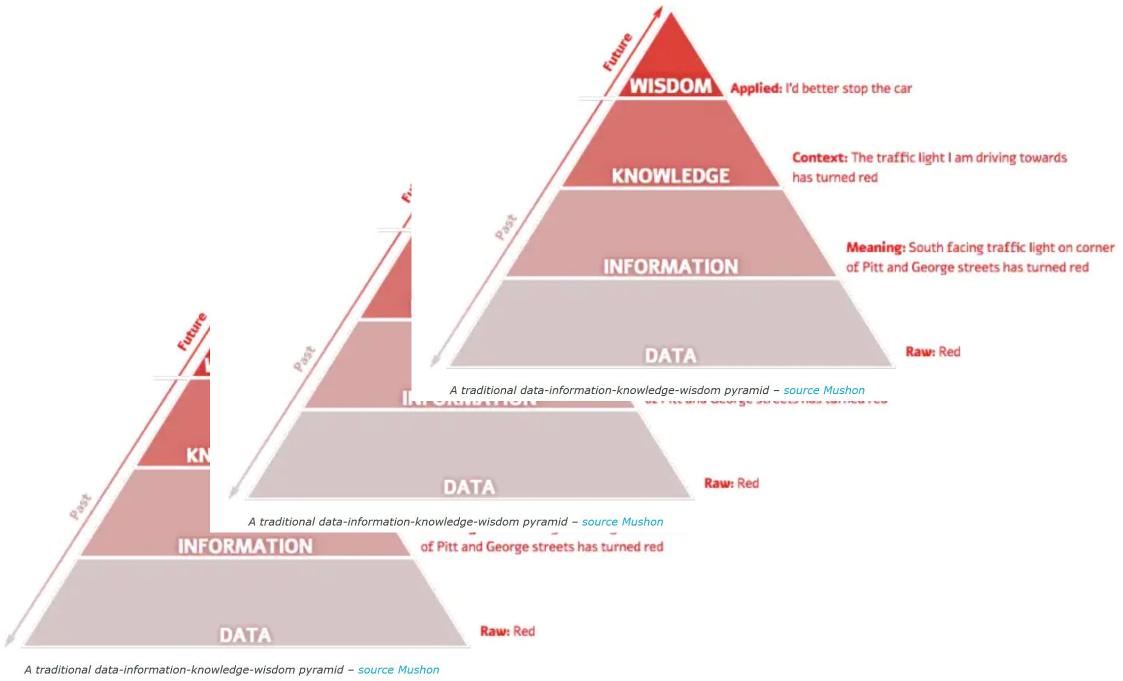
- Getting the right data right
  - Data isn't king…quality data with the right processes and people gets you into the Royal Family
  - Advancing the statistical definition of "right"
  - But do the all the pieces fit together?

- Mapping and Intelligence missions

Context

- More context = more assurance that the data is right
  - Object detection + Geospatial data
  - Multiple objects

- More data assurance = greater operational acceptance



Even greater wisdom and insight

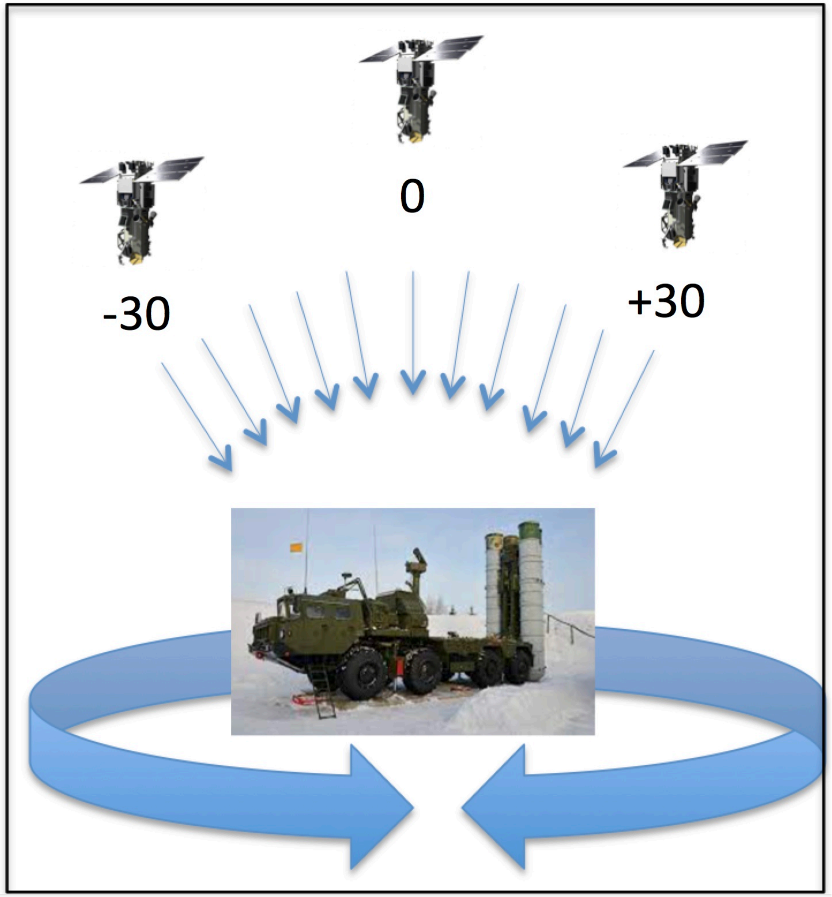
Operational Relevance

Latency

- Getting the *right data right* is useless if it arrives too late
- Advances in compute and coms have opened the door to provide more operationally relevant data
- Delivering AI/ML output…knowledge and wisdom into the hands of users is key.
- Must gain trust through deep proof that the algorithms work and then get those answers to operators.

Challenges

- More Data
  - SAR and HSI
  - ...and all the other spatial data
  - Where are the cyber concerns?
- Better algorithms
  - ...driven by more and better training data
  - What about other phenomenologies?
    - MSI
    - SAR
    - HSI
  - How much can 3D contribute to 'better' data?
  - Are there more combinations?
- More Speed
  - Making the algorithms operationally relevant
  - G-EGD example....What can be next?
  - Op Center Dashboards



Look angle challenge
@1 second increments,
60*60*60*360*60*60 "looks"
280,000,000,000
And then wet/dry, sun/shade, etc

figure eight FEDERAL

Questions & Comments?
Please reach out.

Vinay Malkani

Chief of Engineering

Vinay.Malkani@F8-Federal.com

Jack Hild

Jackhildgeo@gmail.com