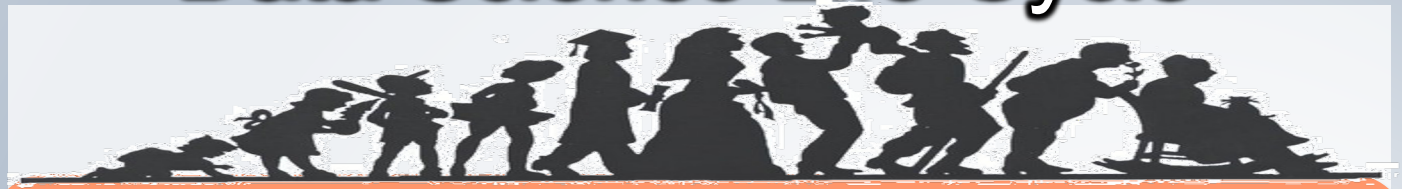
Operational Recommendations for Capturing History and Infusing Data Science (ORCHIDS)

Preparing your mission data for future analysis

Jack Lightholder

Jet Propulsion Laboratory
California Institute of Technology

Data Science Life Cycle



Formulation
Acquisition
Transport
Operations
Calibration
Curation
QC
Analysis
Knowledge
Archiving

Historically…

Real-time reaction

- Humans in the loop
- Fast & loose hand calculation
- Opportunity cost

Thorough Analysis

- Manual rigor
- Exhaustive exploration
- Human-time limited

Data Science Life Cycle

Formulation
Acquisition
Transport
Operations
Calibration
Curation
QC
Analysis
Knowledge
Archiving

- Quick Analysis
  - Quick Calibration
  - Quick QC
  - Quick Analysis
  - Quick Knowledge
- Annotate
- Advise / React
- Repeat

Quick-Look products enable earlier reaction

Advisory systems direct focus

What is Machine Learning?

Algorithms that inductively self-assemble from examples.

Expert Examples
(e.g. Estimates 1-10 of attractiveness of picture)

Training Data
(Encompasses domain of responsibility)

Functional System
- Interpretable as required
- Fixed code (V&V ready)
- Not ever-learning (mostly)
- **EASILY UPDATED**

Strength:
Don't specify rules Don't explain "how"

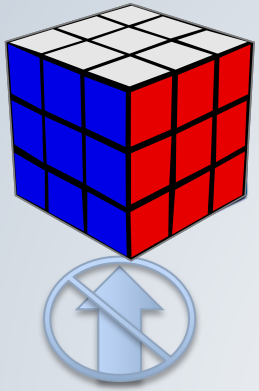
Machine Learning simplifies & systematizes the building and updating of Autonomy / Advisory systems

What is Machine Learning?

Algorithms that inductively self-assemble from examples.

Expert Examples
(e.g. Estimates 1-10 of attractiveness of picture)

Training Data
(Encompasses domain of responsibility)



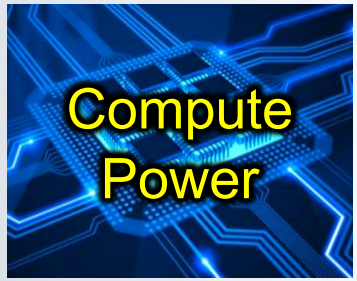
Functional System
- Interpretable as required
- Fixed code (V&V ready)
- Not ever-learning (mostly)
- **EASILY UPDATED**

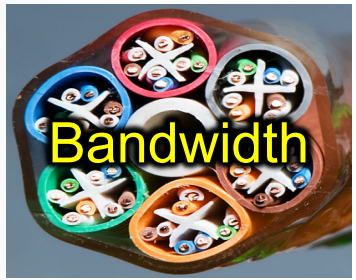
Strength:
Don't specify rules Don't explain "how"

Machine Learning simplifies & systematizes the building and updating of Autonomy / Advisory systems

When to Use ML / Autonomy?

Reaction Time





ML ≠ Expert Replacement

- Eliminates drudgery
- Operates impossibly fast
- Focuses experts on interesting cases
- Enables **larger human feats**

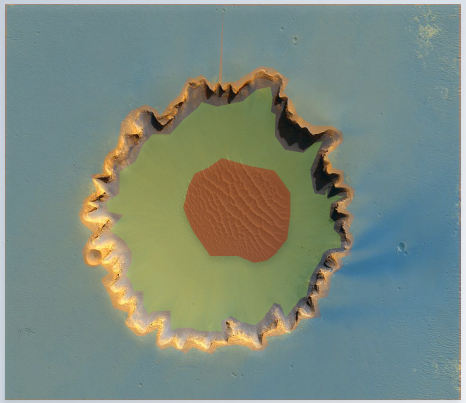
Data Science asks: “Would you like to have the same output with $^1/_6$ the experts or x6 the output with your current experts?”

Summarization Technology

Scene-Wide Labels

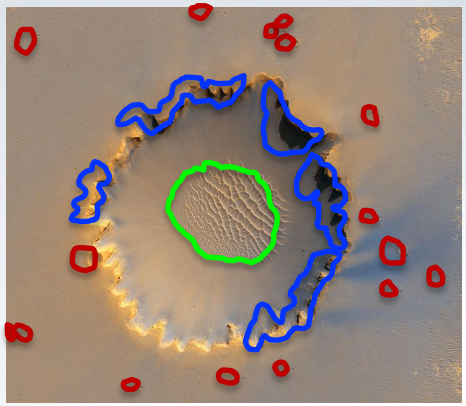
Scene Feature	Present
Dunes	**Yes**
Barchan Dunes	No
Small Craters	**Yes**
Large Craters	**Yes**
Fresh Impacts	No
RSL	No

Terrain Classification



Terrain Type	Image %
Flat Plain	50%
Crater Slope	25%
Dune Field	10%
Ridges	15%

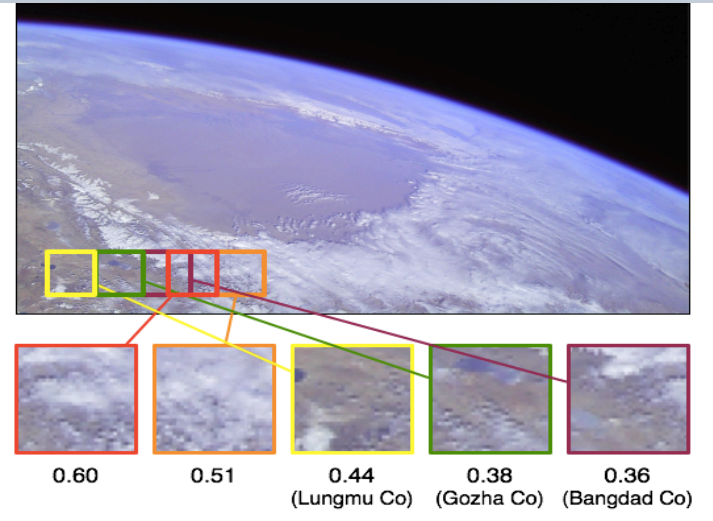
Landmark Identification



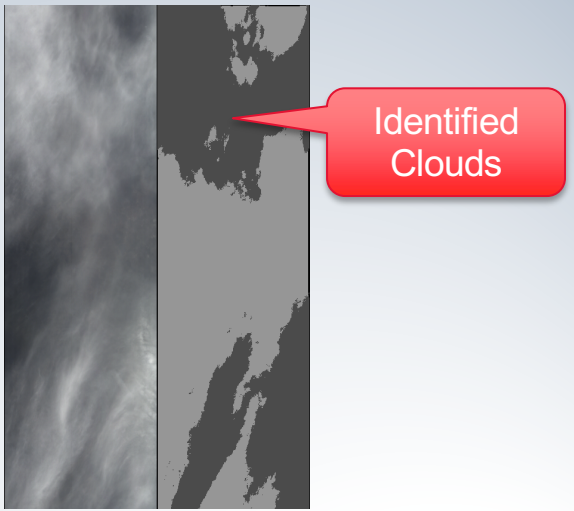
Landmark Type	Number
Small Craters	16
Ridges	4
Dunes	1

JPL

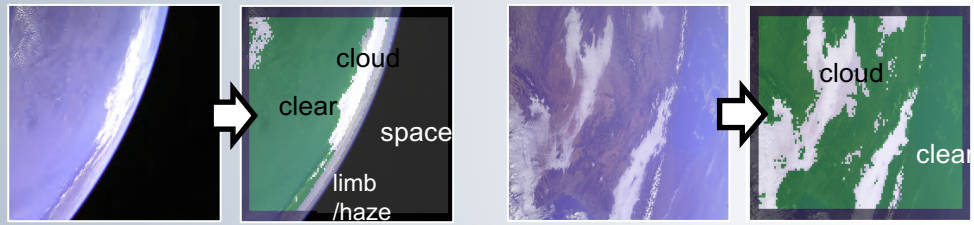
Detecting Features of Interest



Visual Salience: Identified areas of the image that differ from surrounding areas.



Preliminary Cloud Classification results from EO-1



TextureCam: Pixel classification for cloud screening, downlink prioritization

[Chien et al. 2016 JAIS; Thompson et al., i-SAIRAS 2012; Wagstaff et al., *GRL* 2013; Bekker et al., *Astrobiology* 2014; Altinok et al. JFR 2015, Thompson et al. TGARS 2011]

Scene Labeling

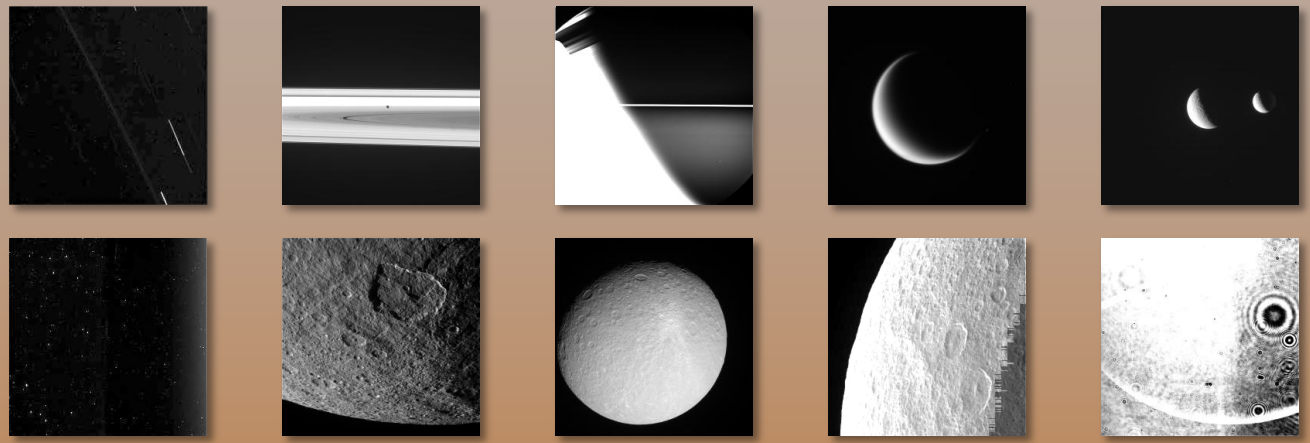
Drs. Alphan Altinok
Brian Bue
Alice Stanboli
Kiri Wagstaff

“Scalable Scene Analysis” System

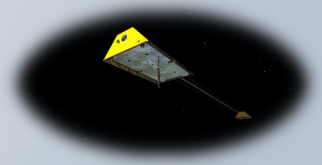
- Convolutional Neural Network
- Implemented on PDS Atlas
- Currently trained for Cassini & MSL Images

craters transients rings surface horizon clouds plume
sky view starfield body types multiple objects phases
artifact eclipse haze over exposure noise ripple camera distance

19 categories – 53 labels



Motivation: Trapped Mission Data, Limited Analysis



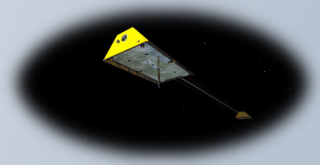
Operating Spacecraft

Science + Engineering

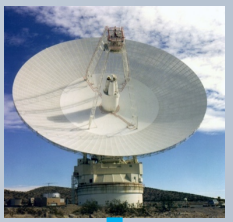
Ground Station

- Spacecraft downlink both Science and Engineering Data
- Traditional Operations treat these flows very differently

Motivation: Trapped Mission Data, Limited Analysis



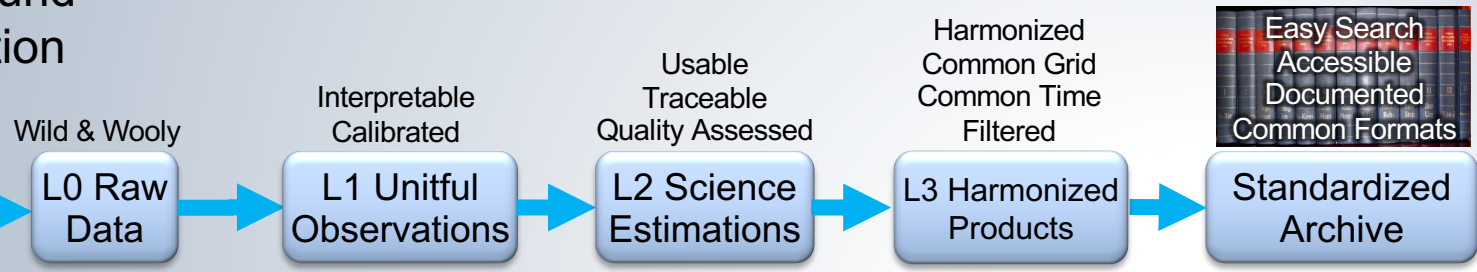
Operating Spacecraft



Ground Station

Science Data

- Science data is destined for exploration
- L0-L3 processing prepares data for analysis
- Standardized archives (DAAC's) serve data
- Search abilities permit researcher access
- Quality assessment and metadata captured
- Excellent document is assumed and required

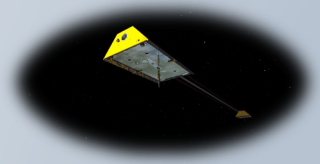


Science Pipeline is Healthy & Optimized for Analysis

(because that's the point of science data)

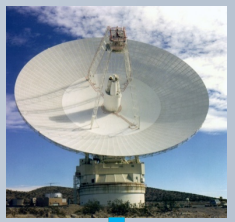
JPL

Motivation: Trapped Mission Data, Limited Analysis



Operating Spacecraft

- Engineering data is for “expert consumption only”
- Shallow or emergency analysis assumed
- Wake the person who understands this stuff!
- New personnel overwhelmed with missing details
- Even basic analysis is stymied, let alone data-driven science



Ground Station

Science Data

Wild & Wooly
L0 Raw Data

Interpretable
Calibrated
L1 Unitful Observations

Usable
Traceable
Quality Assessed
L2 Science Estimations

Harmonized
Common Grid
Common Time
Filtered
L3 Harmonized Products

Easy Search
Accessible
Documented
Common Formats
Standardized Archive

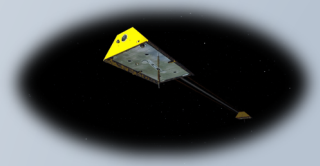
Eng. Data

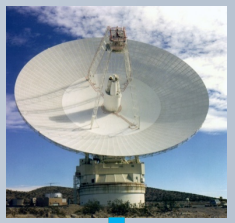
Wild & Wooly
“L0” Raw Data

Pathological Formats
Siloed Access
Obscure/missing Docs
Nonstandard Archive

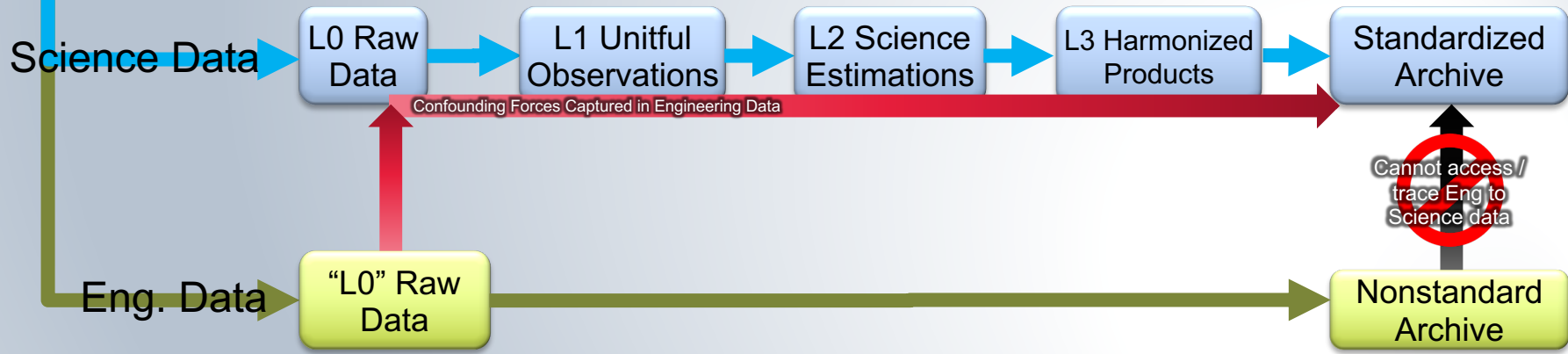
Motivation: Science Anomaly Situation

For GRACE missions, spacecraft IS the instrument!

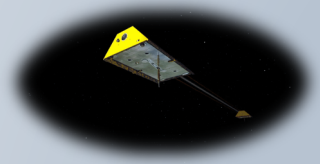




- (this really happens!)
- Scientists trace L2/3 anomalies back to L0
- Issue is on instrument! Linked to spacecraft itself
- No worries… scientists will just look at Eng data
- What do you mean it can't be easily harmonized?
- What do you mean I can't easily make graphs?
- Advanced analysis thwarted… spin up Ops effort
- False dichotomy between Science/Eng data maintained



Motivation: Ops Anomaly Situation



Panic! Streetlight effect forces minimum effort paths

- Look at the eng data we CAN get at
- Check only the most likely relationships
- Confirmational analysis only… exploration hindered
- System interactions are hardest to explore

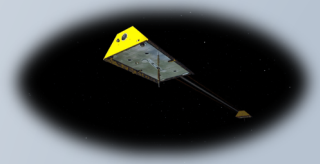
Eng. Data

Wild & Wooly

"L0" Raw Data

Nonstandard Archive

Motivation: Multi-Mission Comparison



Try to compare data from multiple missions or ATLO to Flight

- Totally different archive, formats
- Mature missions no longer have original personnel!
- Requires research project just to harmonize
- Neither archive captured sufficient info for cross-comparison
- May not actually be possible within reasonable time/cost

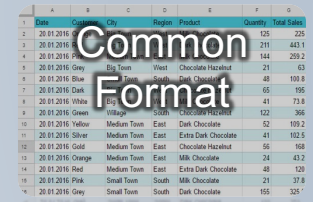
Part of solution already known!

NASA's Analysis Ready Data (ARD) concept: the data we all want & need

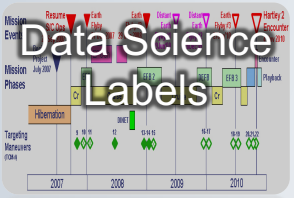
Analyst can just start work

Easily select and interpret data

Event Logs, Known Issues
Faulty Channels, Design Intentions,
Expert Knowledge



Any program can read
No special loaders



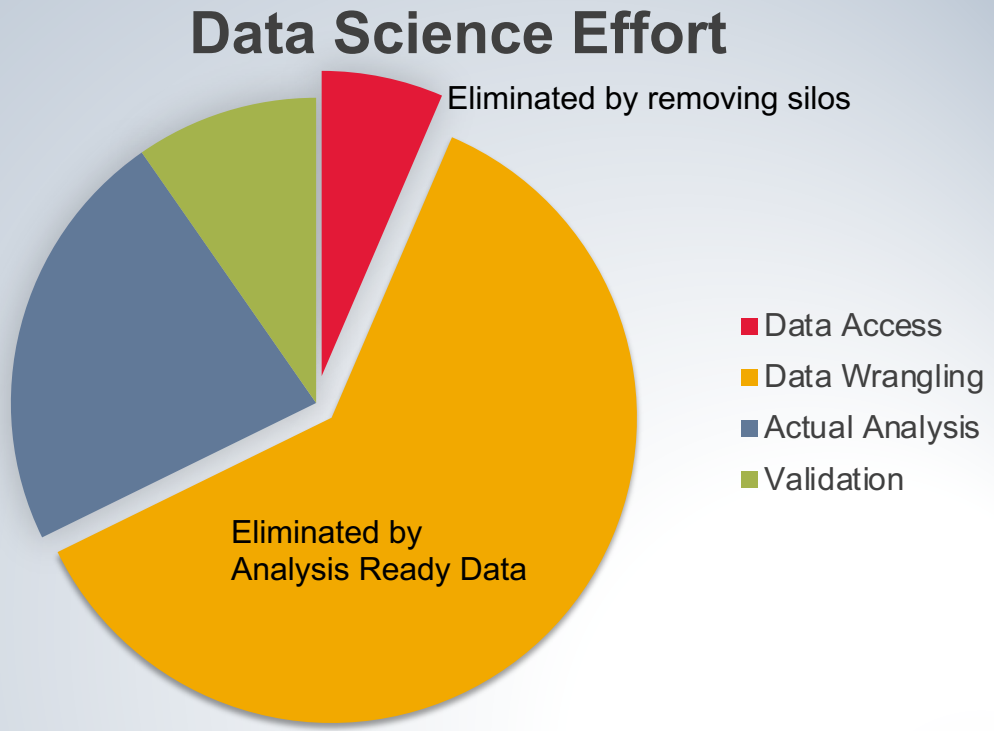
Timelines, events, channel information stored as proper labels

Simple, searchable, standard, scalable

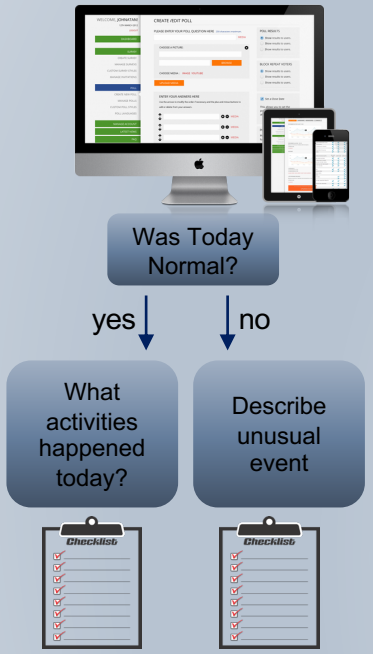
Removing Barriers

Data access, format conversion, time-alignment and missing data point management consumes > 60% of time spent on analysis projects.

Analysts must chase SME's to overcome incomplete / esoteric documentation.

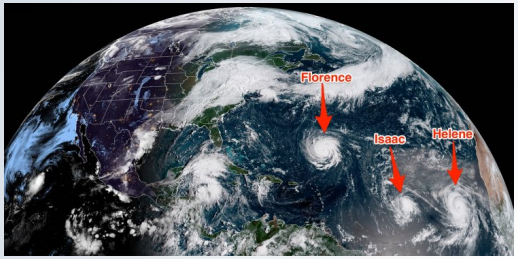


Continuous Knowledge Capture Tools – Observational Science on Ops!



- Simple, easy, high-level daily questions
- Gives overall daily context as labels
- Ever-growing list of standard categories
- Machine-readable database

- Time range annotation (labels)
- What happened when, with categories & notes
- Important patterns & relationships between channels

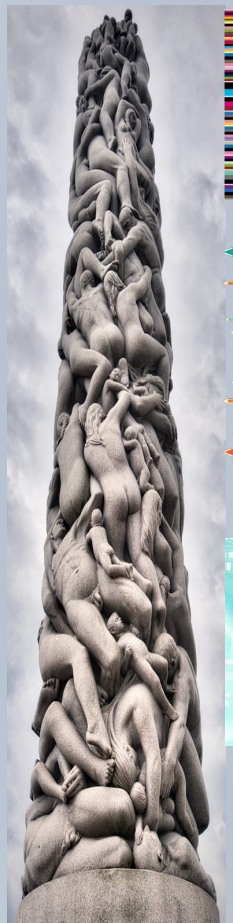


- Anomalies should be treated like Hurricanes
- Each has formal name, trajectory, details
- Start & stop time clearly delineated
- Characteristic features specifically highlighted
- Minimize freeform text!

Delivered solutions – Parallel pipeline for zero disruption

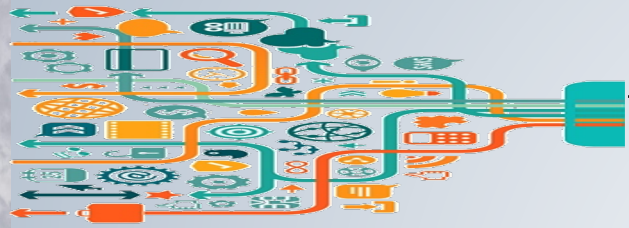
New Pipeline

Established Ops Workflow
(it's made of people!)



Translation Library

- Converts current formats to Data Science formats
- Decodes Ops-specific datatypes (e.g. on-change)



Data Alignment / Harmonization System

- Common time bins, densification (L3 equivalent)
- Merges Metadata

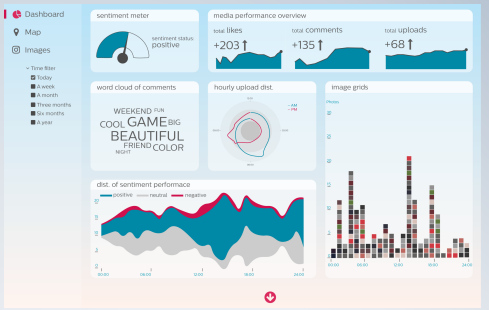
Analytic Database

- On-demand data regridding, filtration
- On-demand data computation results
- On-demand featurization for DS use cases

Tools Enabled by New Pipeline

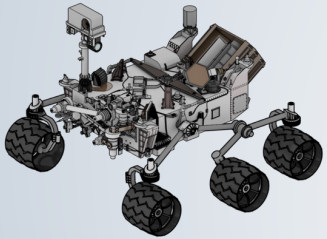
Fast Exploration of ARD

Anomaly detection / Focus of Attention

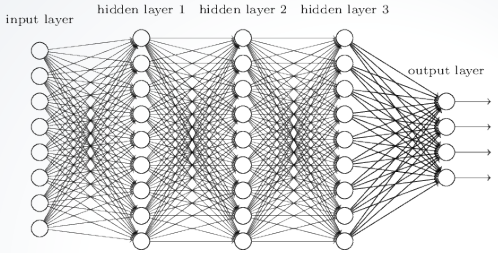


Interactive Data Dashboarding

Expert Labeling Tools



Multi-Subsystem Trending Tools



Machine Learning Model Infusion

Jet Propulsion Laboratory
California Institute of Technology