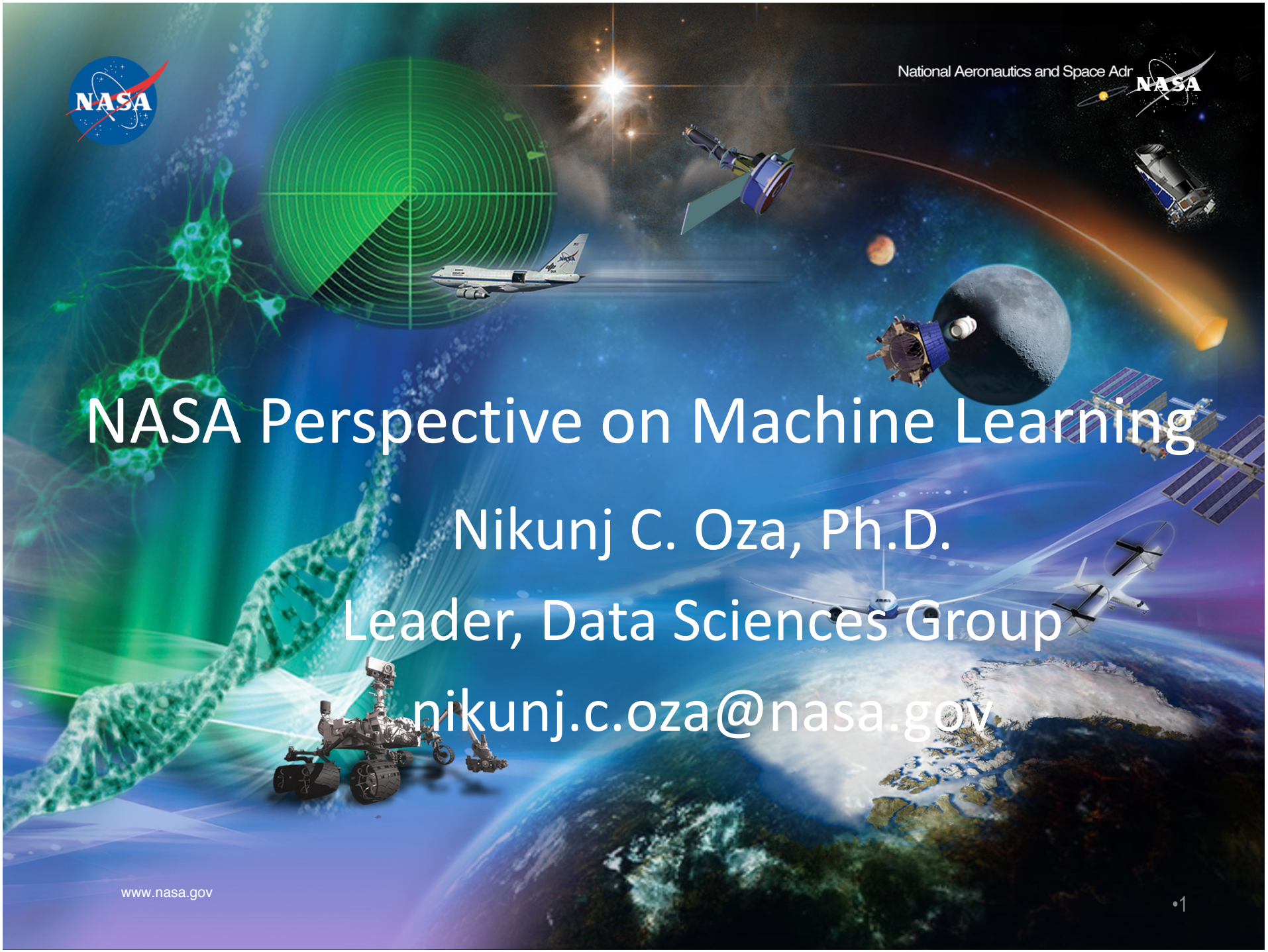


NASA Perspective on Machine Learning

Nikunj C. Oza, Ph.D.

Leader, Data Sciences Group

nikunj.c.oza@nasa.gov

The Data Sciences Group at NASA Ames

Data Mining Research and Development (R&D) for application to NASA problems (Aeronautics, Earth Science, Space Exploration, Space Science)

Group Members

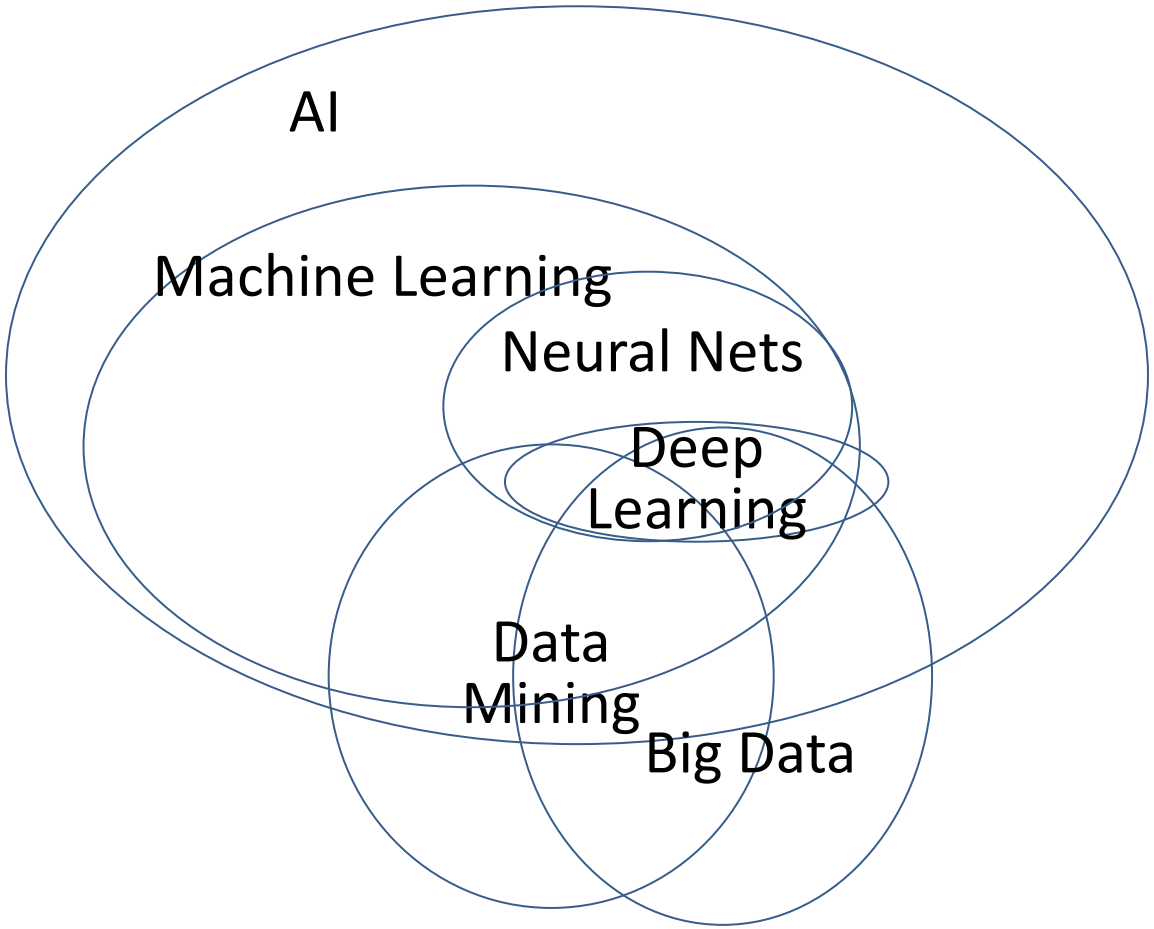
Ilya Avrekh
Kamalika Das, Ph.D.
Dave Iverson
Vijay Janakiraman, Ph.D.
Rodney Martin, Ph.D.
Bryan Matthews
Nikunj Oza, Ph.D.
Veronica Phillips
John Stutz
Hamed Valizadegan, Ph.D.
+ summer students

Funding Sources

- NASA Science: AIST, CMAC programs
- NASA Aeronautics: ATD, SMART-NAS
- NASA Engineering and Safety Center
- NASA Human Space Exploration
- Aero seedling funds, Center Innovation Fund
- Non-NASA: DARPA, DoD

Team Members are NASA Employees, Contractors, and Students.

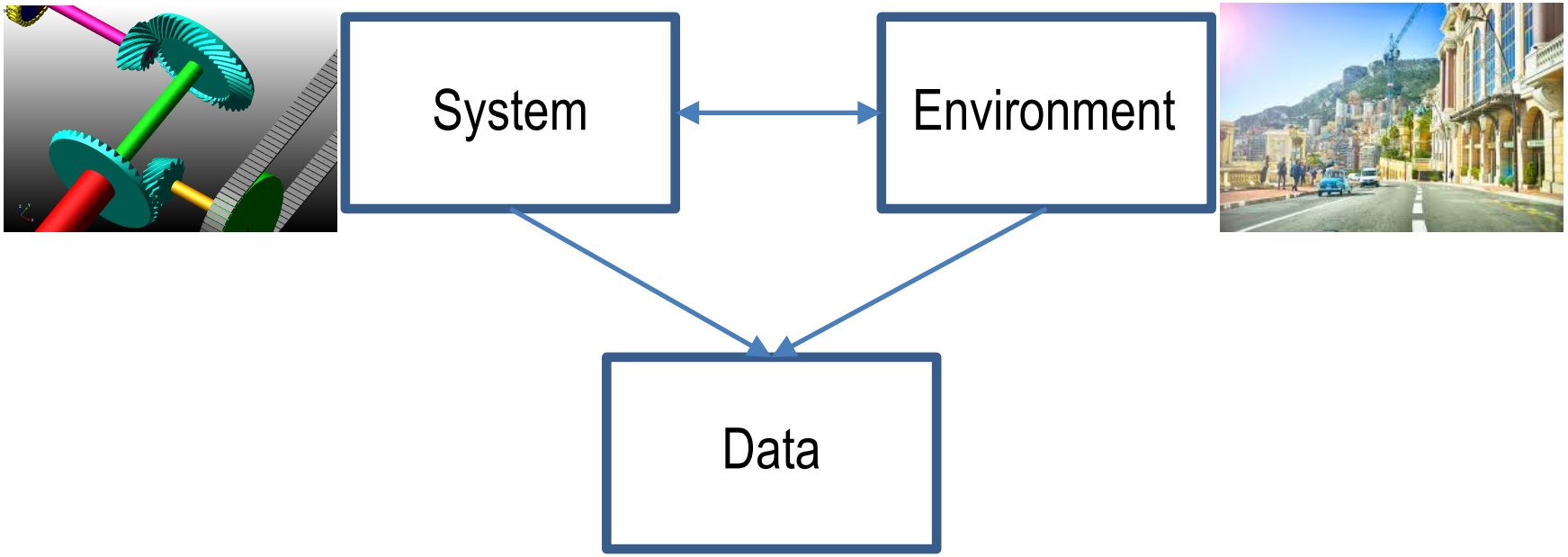
Our Domain


AI
Machine Learning
Neural Nets
Deep Learning
Data Mining
Big Data

Collaborators

- Universities: Basic research in data sciences, domains
- Industry: Data sources, baseline methods, domain expertise
- NASA: Apply basic research, develop for NASA's needs, domain expertise, funding programs
- Other government: funding, domain expertise, data sources

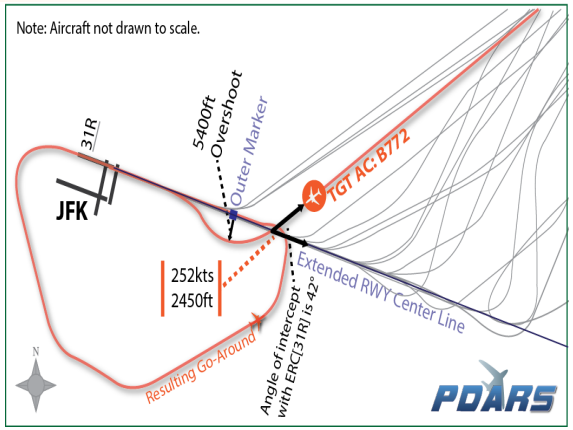
Data Sciences

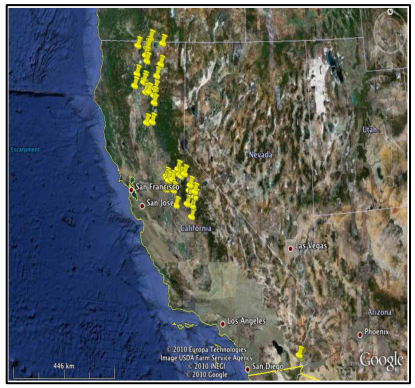


- Data are produced by system operating in an environment
- Data Sciences: Reverse-engineer system and environment
- Understand how system *really* works, correct system model errors, understand true impact of environment

Example NASA Machine Learning Problems

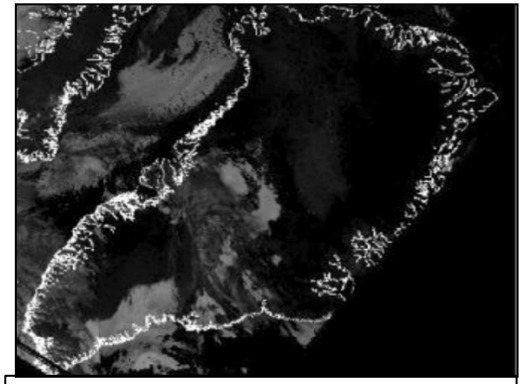
- Aeronautics
  - Anomaly Detection
  - Precursor Identification
  - text mining
- Earth Science
  - Filling in missing measurements
  - anomaly detection
  - teleconnections
  - climate understanding
- Space Science: Kepler planet candidates
- Space Exploration
  - system health management
  - astronaut health



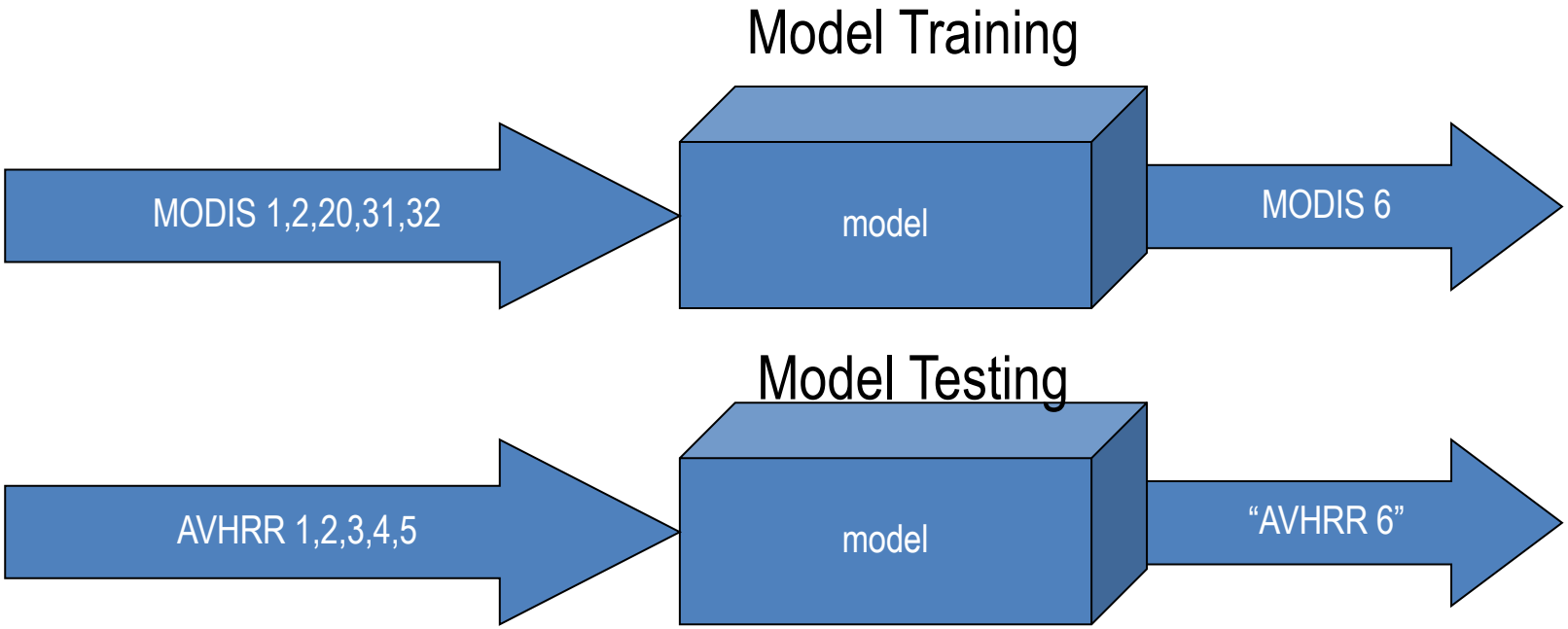


Data Mining for Earth Science Examples

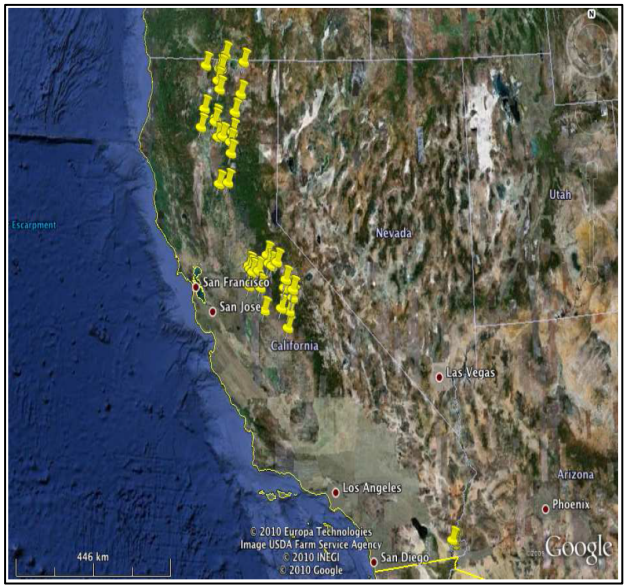
- Virtual Sensors
  - Regression to fill in missing or noisy sensor values, anomaly detection
  - Estimated MODIS channel 6 for older instrument (AVHRR)



Estimating MODIS channel 6 (useful for distinguishing clouds over snow and ice covered regions).



Data Mining for Earth Science Examples

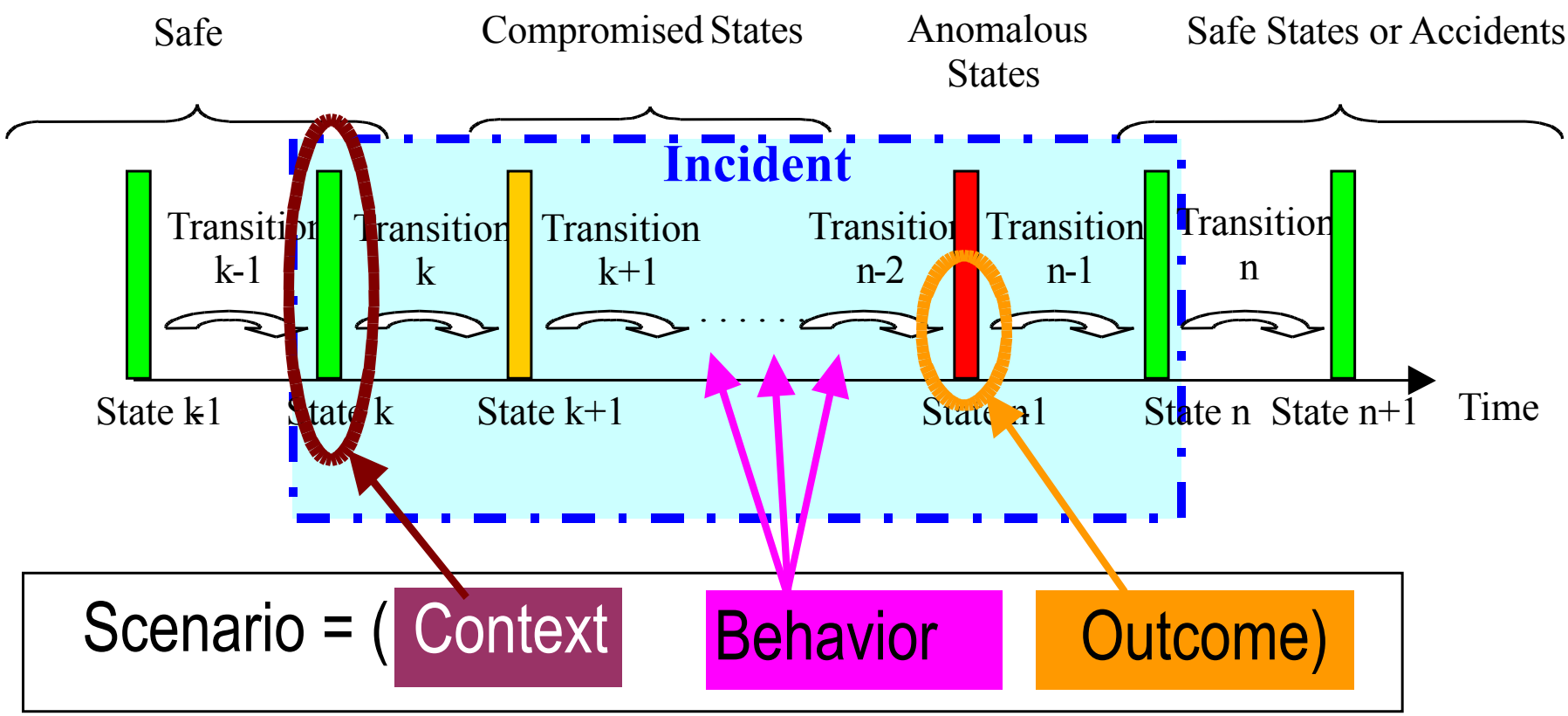


Top few outliers (yellow pins) identified by distributed 1-class SVM-based outlier detection algorithm in the California MODIS data.

Distributed Algorithms For Earth Science

- For large scale data where centralization is impractical
- Developed distributed 1-class SVM for anomaly detection.
  - 99% of the accuracy of centralized algorithm
  - 1% communication overhead running time (relative to embarrassingly parallel runs)

The Anatomy of an Aviation Safety Incident



From Irving Statler, Aviation Safety Monitoring and Modeling Project

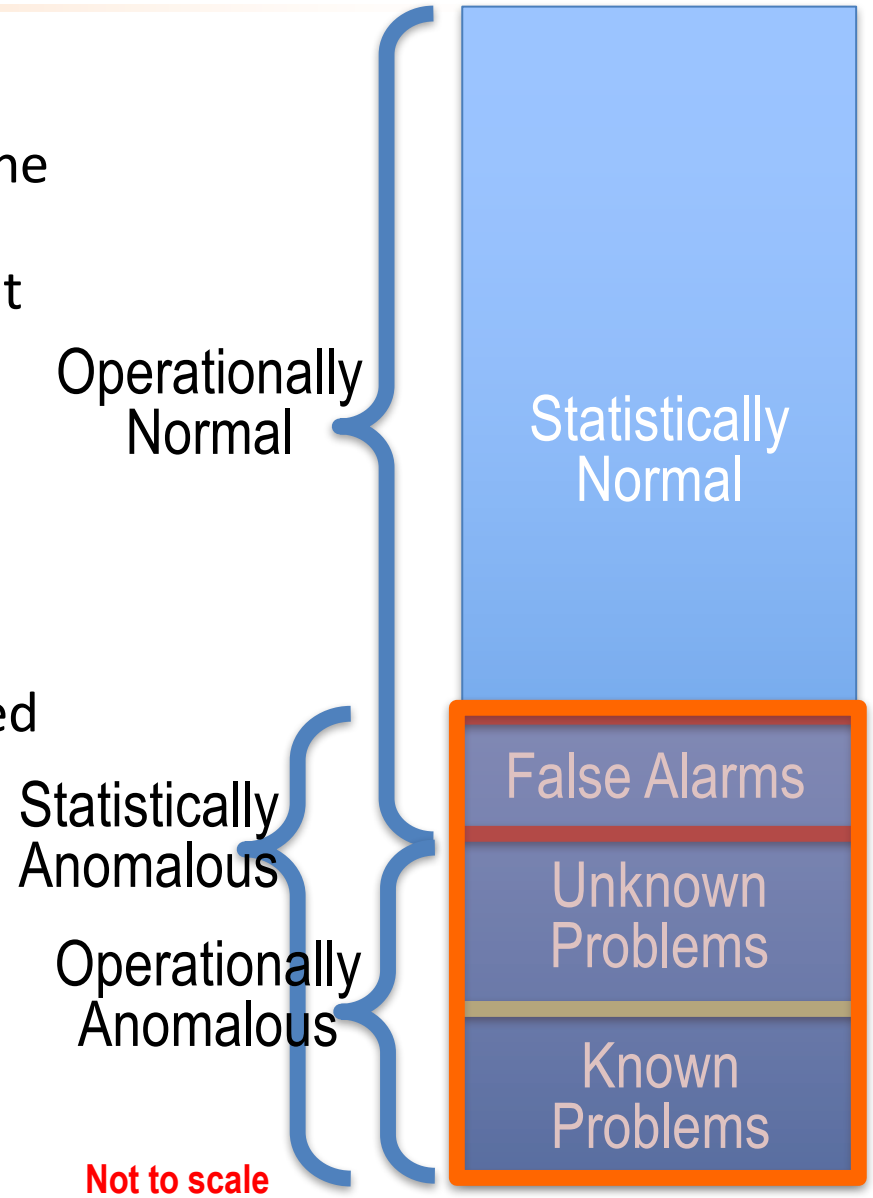
Current Methods of Finding Issues

Exceedance-Based Methods

- Known anomalies/safety issues
- Conditions over 2-3 variables (e.g., speed > 250 knots, altitude = 1000 ft, landing)
- Cannot identify unknown anomalies
- Low false positive rate, high false negative (missed detection) rate.

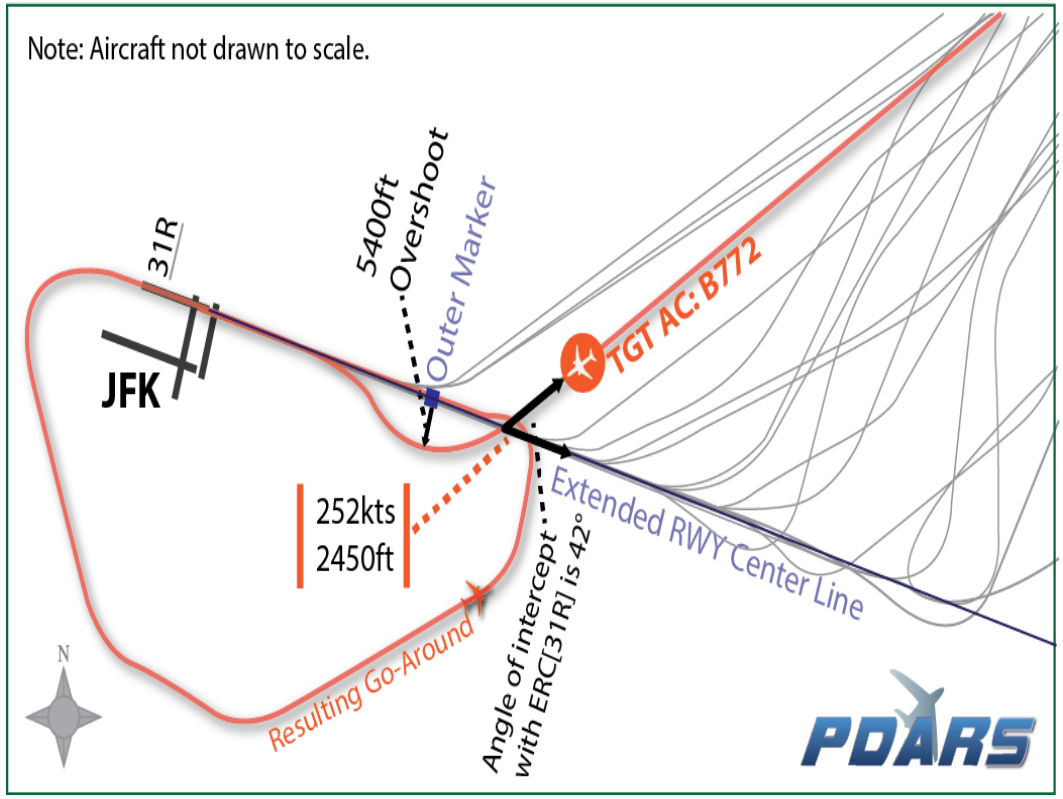
Data-Driven Methods

- DISCOVER anomalies by
  - learning **statistical properties** of the data
  - finding which data points do not fit (e.g., far away, low probability)
- Complementary to existing methods
  - Lower false negative (missed detection) rate
  - Higher false positive rate (identified points/flights unusual, but not always operationally significant)
- Data-driven methods -> insights -> modification of exceedance detection



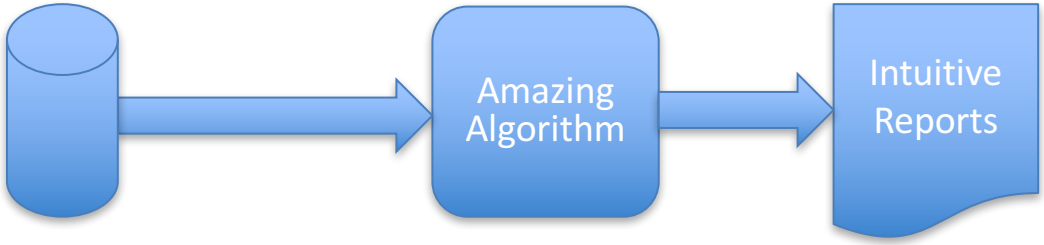
High Speed Go-Around

- Overshoots Extended Runway Centerline (ERC) by over 1 SM
- Over 250 Kts @2500 Ft.
- Angle of intercept > 40°
- Overshoots 2nd approach



Four V's of Big Data

- Volume[1]
  - Radar Tracks: 47 facilities (1 year)
    - ~423 GB (Compressed)
    - ~3.2 TB (CSV)
  - Weather and Forecast (Entire NAS)
    - CIWS ~2.8 TB
- Velocity
  - Radar Tracks: 47 Facilities
    - ~35 GB/month (compressed).
    - ~268 GB/month (uncompressed)
  - Weather and Forecast (Entire NAS)
    - CIWS ~233 GB/month
- Veracity
  - Data drop outs
  - Duplicate tracks
  - Track ending in mid air
  - Reused flight identifiers
- Variety
  - Numerical (continuous/binary)
  - Weather (forecast/actual)
  - Radar/Airport meta data
  - ATC Voice
  - ASRS text reports (Pilot/Controller)



[1]But not always the right kind!

How do we get the Word Out?

DASHlink

disseminate. collaborate. innovate.
https://dashlink.ndc.nasa.gov/

DASHlink is a collaborative website designed to promote:

- Sustainability
- Reproducibility
- Dissemination
- Community building

Users can create profiles

- Share papers, upload and download open source algorithms
- Find NASA data sets.

Ongoing and Future Work

- So far: desktop, HPC. offline, desktop
- Ongoing: in-time for online monitoring
- Future
  - usability for analytics deployments
  - embedded systems, autonomous systems
  - Use all platforms, in best way possible, on the fly

Reading

Contact: **nikunj.c.oza@nasa.gov**

Papers

[1] B. Matthews, D. Nielsen, J. Schade, K. Chan, and M. Kiniry, Comparative Study of Metroplex Airspace and Procedures Using Machine Learning to Discover Flight Track Anomalies, 34th DASC, 2015.

[2] B. Matthews, D. Nielsen, J. Schade, K. Chan, and M. Kiniry, Automated Discovery of Flight Track Anomalies, 33rd DASC, 2014.

[3] S. Das, B. Matthews, N. Oza, and A. Srivastava, Multiple Kernel Learning for Heterogeneous Anomaly Detection: Algorithm and Aviation Safety Case Study, *KDD* 2010.

[4] V. Janakiraman, B. Matthews, and N. Oza, Discovery of Precursors to Adverse Events Using Time Series Data, *(to be presented) SIAM International Conference on Data Mining,* May 5-7, 2016, Miami, FL.