National Aeronautics and Space Administration

Data Science @ OCIO/NASA

Office of the Chief Information Officer

Brian Thomas

June 29, 2016

www.nasa.gov

Outline

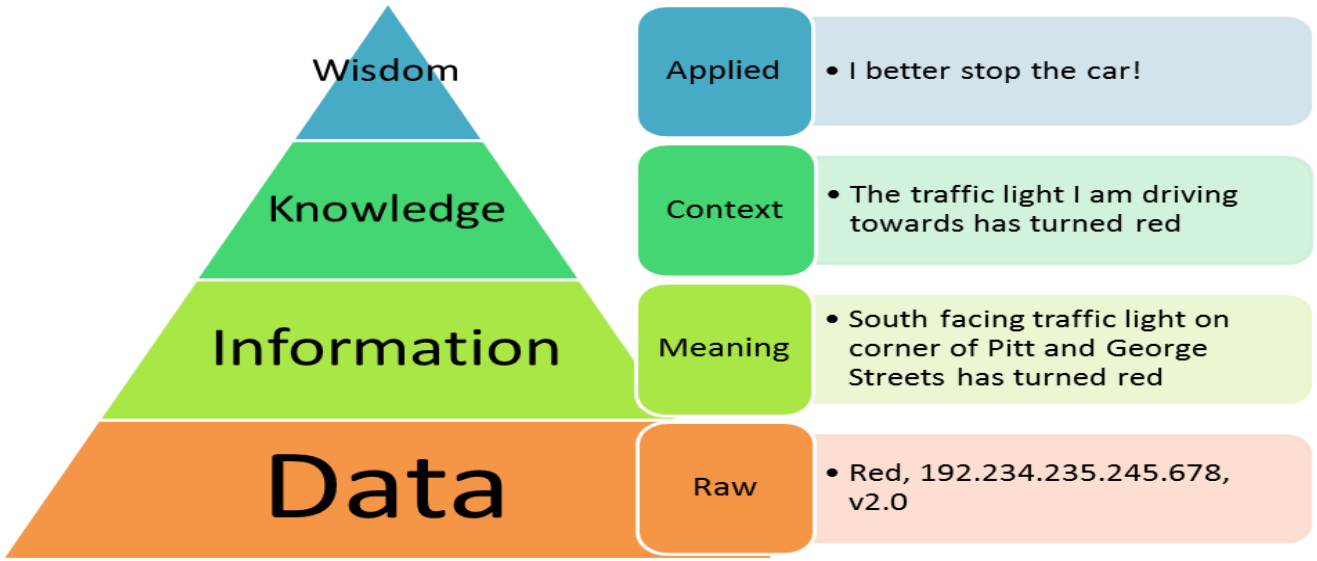
- What is Data Science?
- Data Science activities at OCIO/NASA
- Big Data Issues for Data Science
- Summary

Data Science? What's that?

Wikipedia :

> *"**Data Science is an interdisciplinary field** about processes and systems **to extract knowledge** or insights **from data** in various forms, either structured or unstructured."*

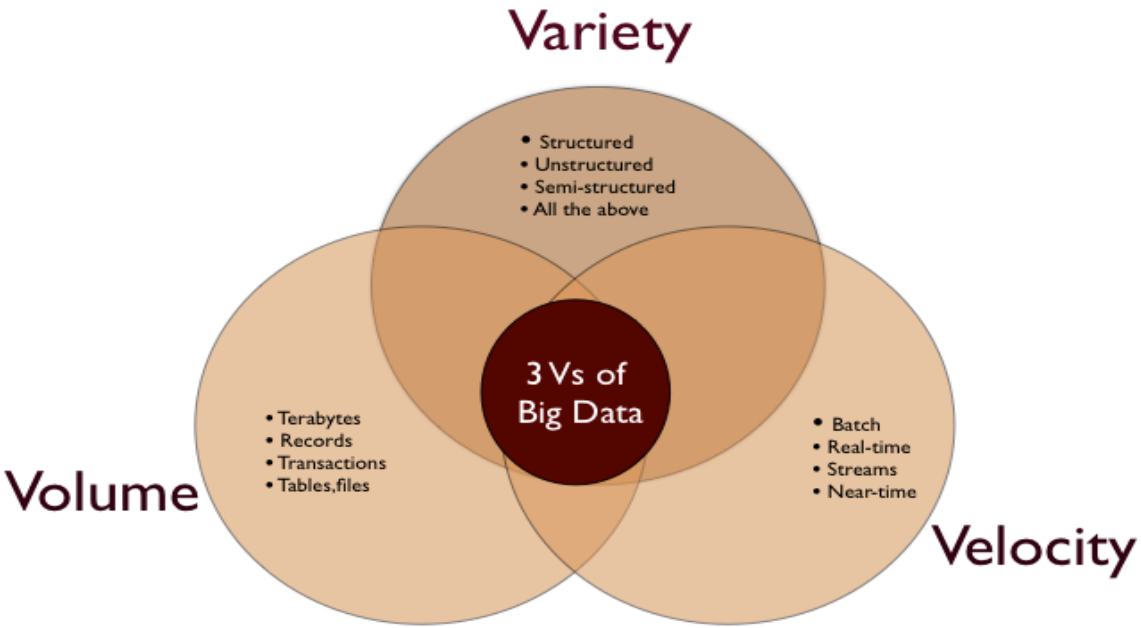
Data -> Knowledge (and beyond!)



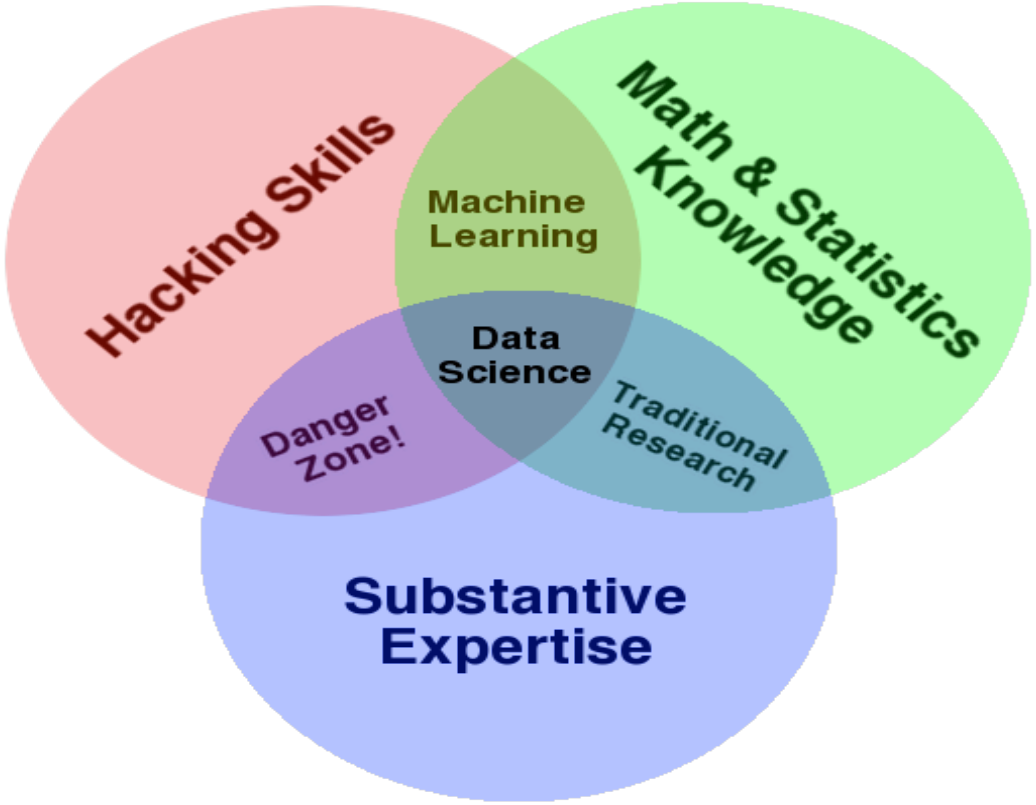
© 2011 Angus McDonald

“Problem Datasets”?

“Problem Datasets”? - Characteristics



Data Scientists: Distinguishing Technical Skills



Data Science @ NASA/CIO : the “Data Team”

Brian Thomas
Agency Data Scientist

Nick Skytland
NASA Data Evangelist

Sandeep Shetye
Chief Information Architect

Plus a team of talented Data Scientists, Software Engineers and Interns (located at various NASA campuses)

Tech & Innovation Division--What We Do:

ARCH

Information Architecture

The development and use of models, policies, rules or standards that govern how information is collected, stored, arranged, integrated, and put to use in data systems and in organizations.

Standards and interoperability
IA Reference Architecture
Data Mining
Data Contract Language
Mission Support Tech Consulting

Big Data
NIAM.NASA.Gov

DATA

Data Science

The collection, management and analysis of data in order to produce information and drive decision making.

Data Strategy
Data Governance
Data Lifecycle Management
Data Analytics Lab
Data Fellows Program
Data Stewards

</> OPEN

Open Innovation

The development of new innovation frameworks and techniques, and the development and delivery of machine-readable instructions to access, arrange, and apply data.

Agency Open Data Mgmt
Digital Strategy Reporting
Space Apps Challenge
Innovation Incubator
Data Innovation Pipeline
Women in Data
Open.NASA, Github/NASA
Data.NASA, Code.NASA
API.NASA

DIGITAL

Digital Integration

The delivery of enhanced digital capability to support data interoperability and accessibility to enable data insights and discovery.

Data inventory/Registry
Tagging/discoverability
Data usability/APIs
Computing/Coding
Mission-focused tech applications

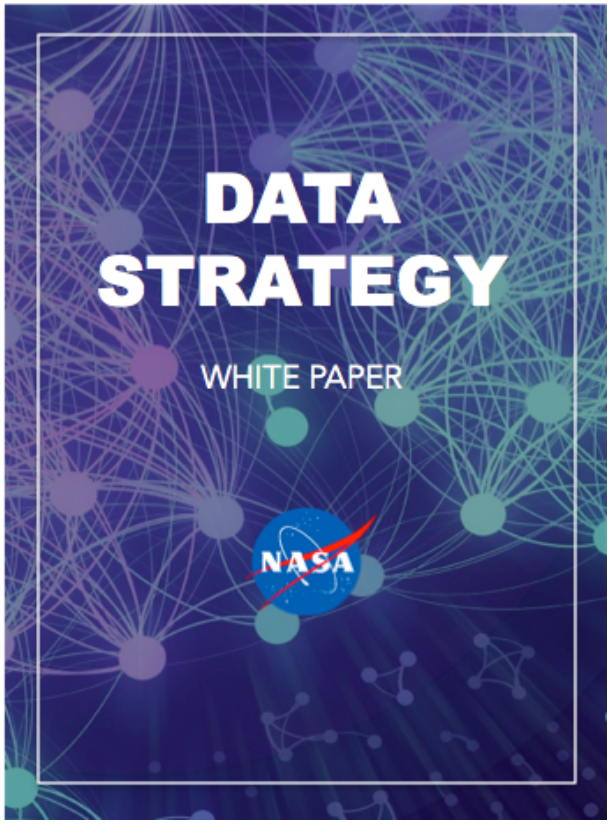
TECH

Tech Infusion

Research, prototype and assess the creation, modifications and usage of processes or tools to solve a problem, generate or perform a specific function to meet stakeholder needs

Data Centric Architecture
Internet of Things
Software as a Service
Virtual Desktop Infra
Collaboration (Google Apps/Secure Dropbox, NASATube)

Our Plan



- Data Management
- Unified Data Lifecycle
- Data Governance
- Data Analytics Lab
- Data Fellows Program
- Data Stewards

Our Collaborators

Current and Past (OCIO Data Team) Project Partners include:

NASA Missions/Centers

- Ames Research Center
- Marshall Space Flight Center
- Johnson Space Center

Agency

- Office of Chief Information Officer
- Office of General Council
- Office of Human Capital Management
- Office of the Chief Engineer
- Office of the Chief Scientist

Academia

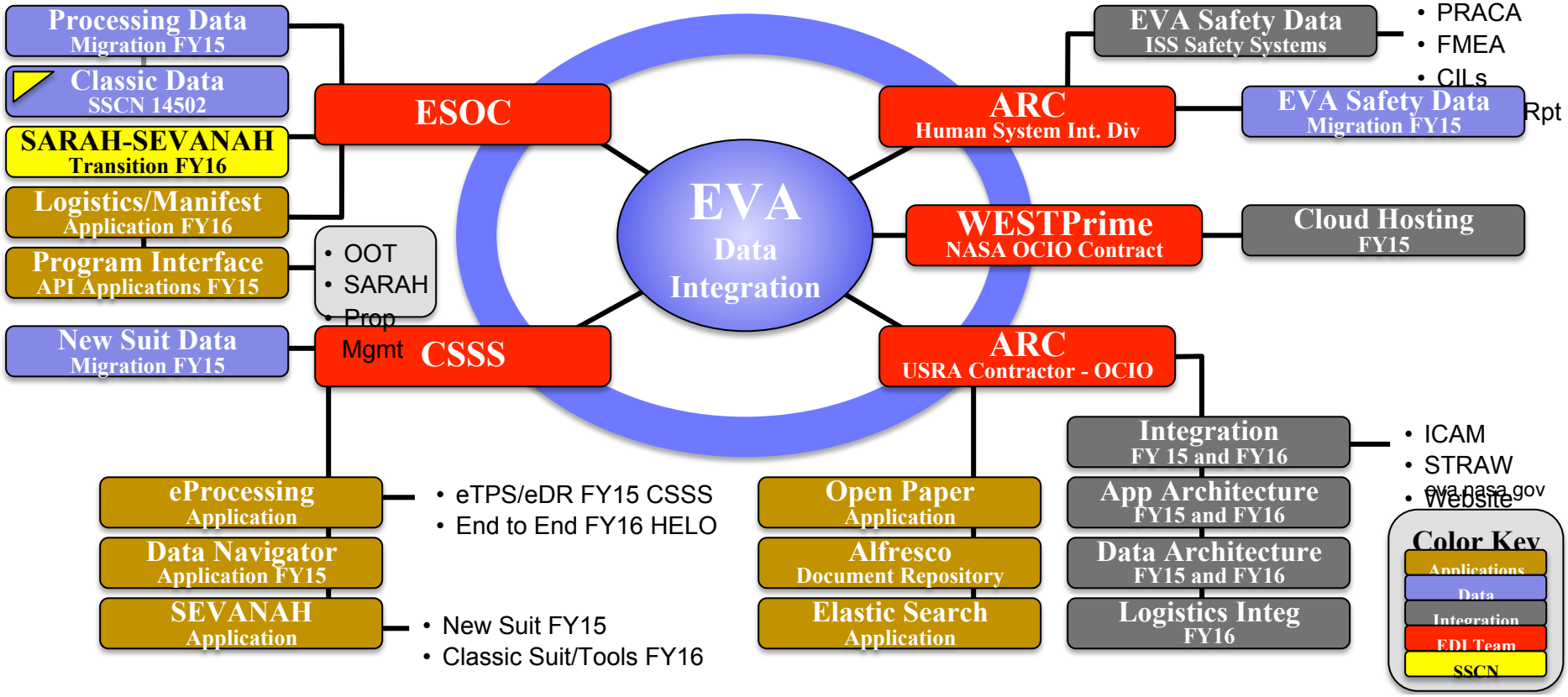
- University of Maryland UC

OCIO Data Science Summary Projects Roster

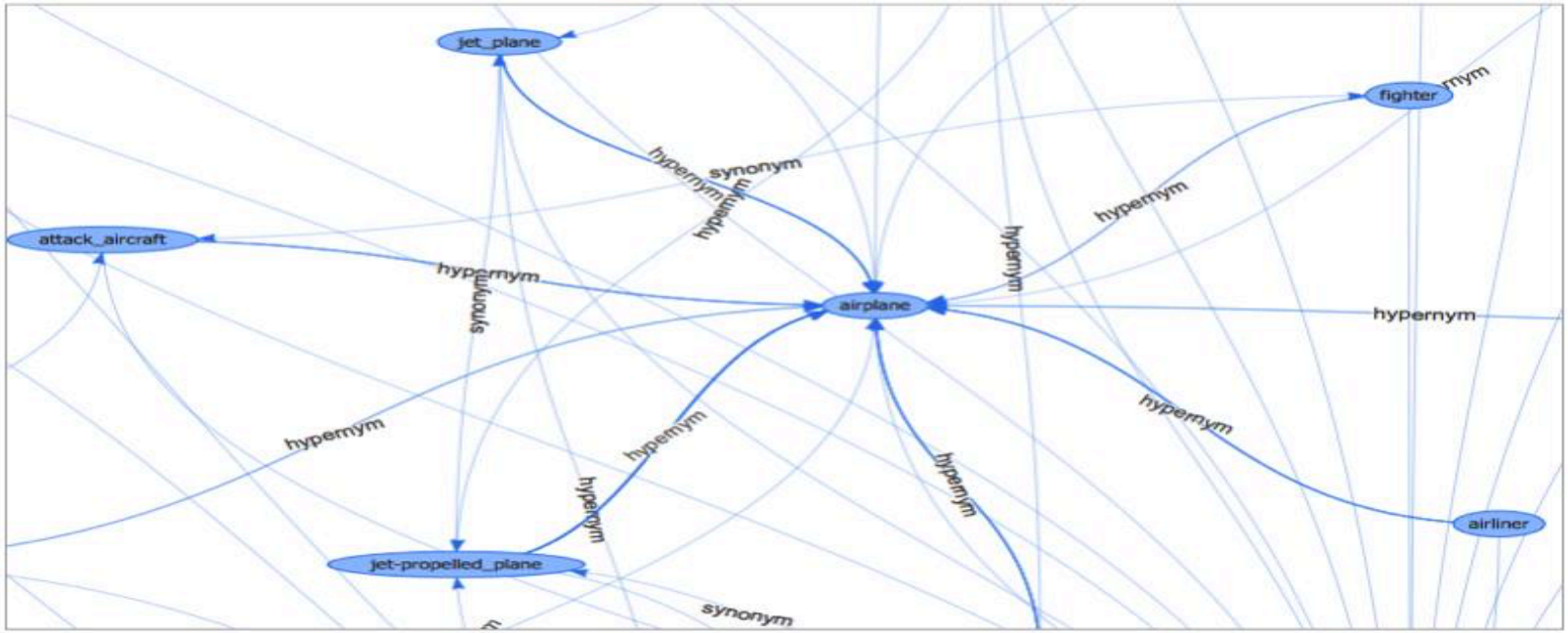
Prototypes/Research into methods/tools/platforms in:

- **Data Analytics/Deep Learning**
  - » ExMC-telemedicine, Mapping Data Universe, Security
- **Search Analytics**
  - » Helping optimize strategies for agency document search
- **Document Classification/Content Tagging**
  - » Security, Open Data, Aerospace, Publications, Records Management
- **Quantum Computing**
  - » API for analytics
- **Financial Analytics**
  - » OCIO, CFO
- **Data Integration/Advanced Data Management**
  - » ISS, MBSE, Orion, EVA, Publications, Records
  - » Microservices (OCIO, Open Data)

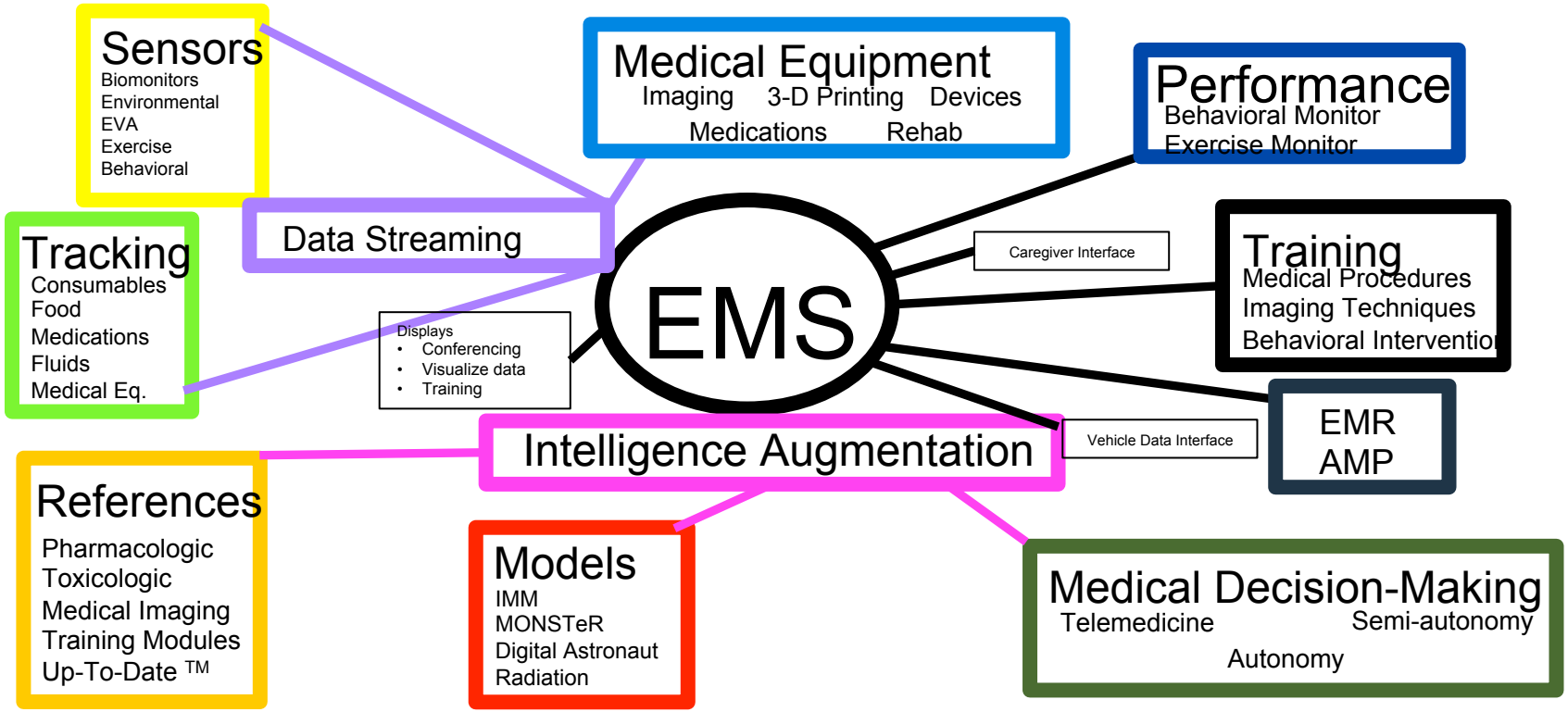
Sampling of Current Projects



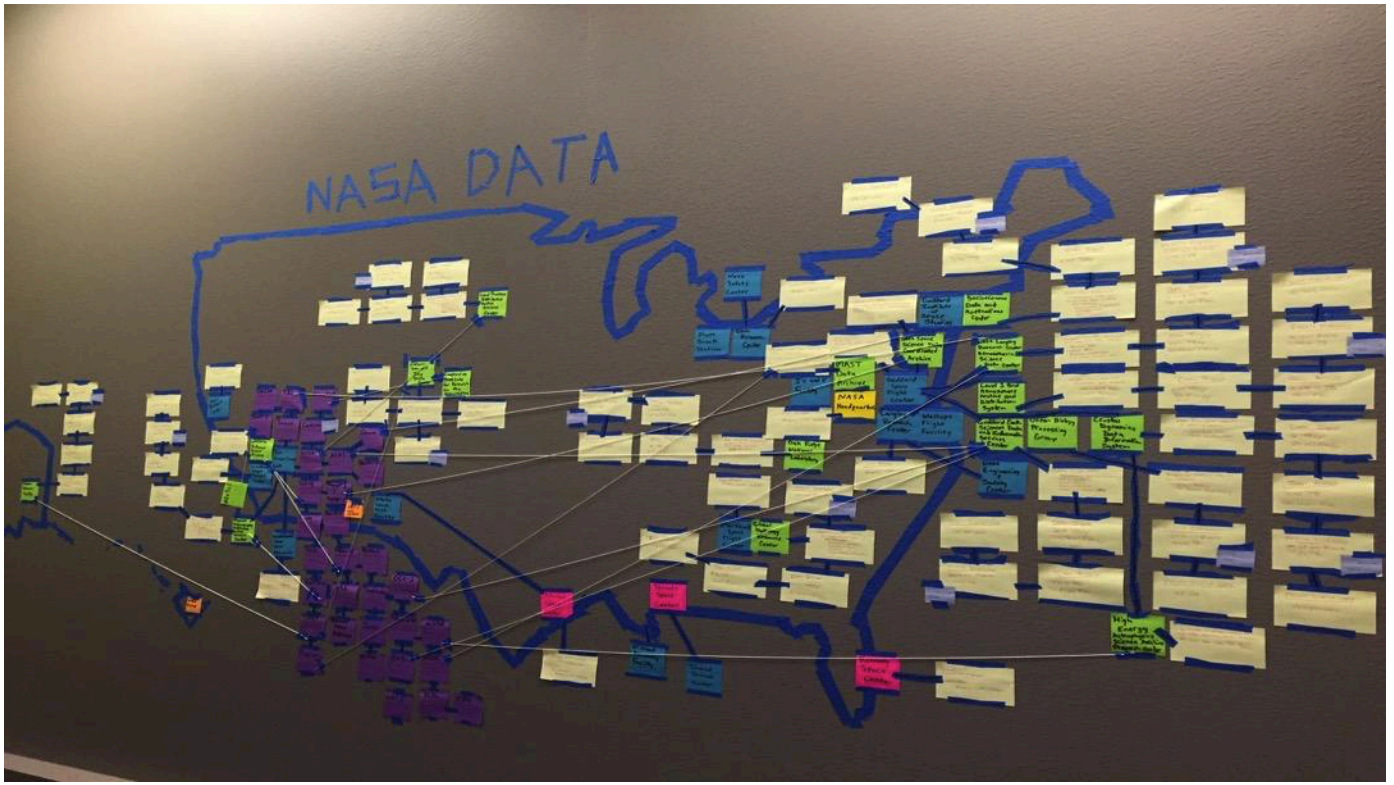
Sampling of Current Projects



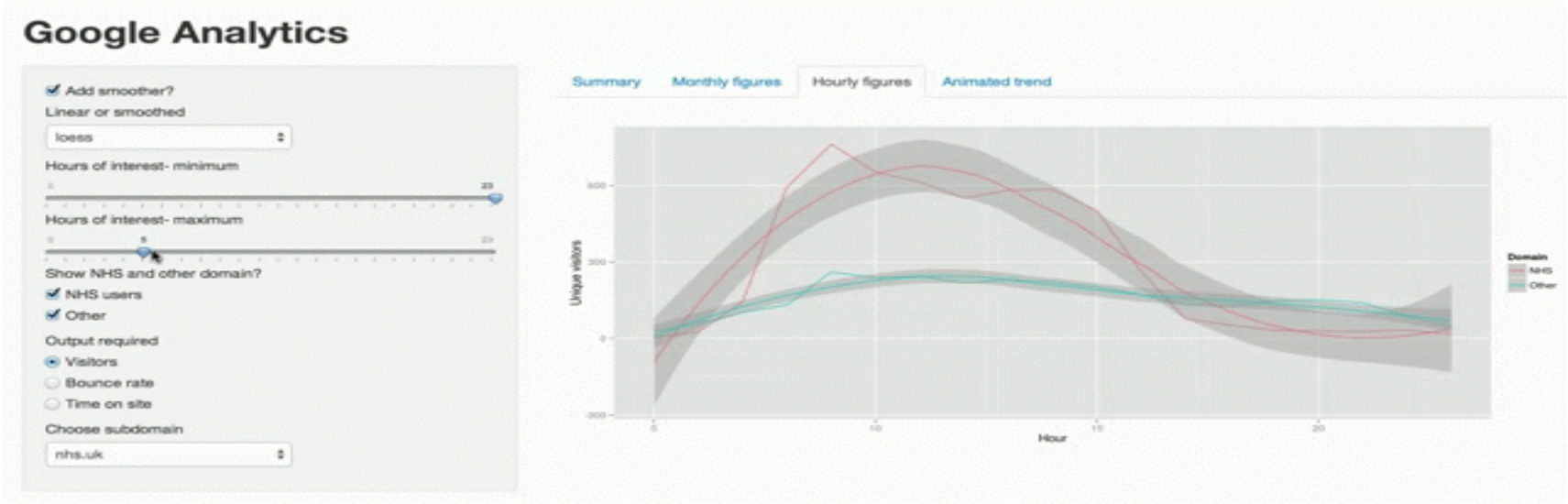
Sampling of Current Projects



Sampling of Current Projects



Sampling of Current Projects



Building Microservices

- **Providing easier data access** (files, databases) (NASA aggregate dictionary: http://nasa-dictionary.heroku.com)
- **Creating higher value data** (mash-ups, APOD : http://api.nasa.gov/planetary/apod)
- **Providing platform-independent capability** (http://nasa-tagging.heroku.com)

- **NASA has all types of "V's" problems**
  - » "Variety" is the hardest to solve from a technical standpoint
- **NASA implementation of data analytics is uneven**
  - » Science doing the best with their data
  - » Business units are doing the poorest
- **Need for much greater Data Sharing**
  - » Microservices (and other solutions) promise an easy way to overcome silos but will create new challenges for NASAs information architecture.
  - » Variety problem will become acute
  - » Adequate Governance/Access/Security are critical
  - » Provenance is also important

Big Data Issues for Data Science

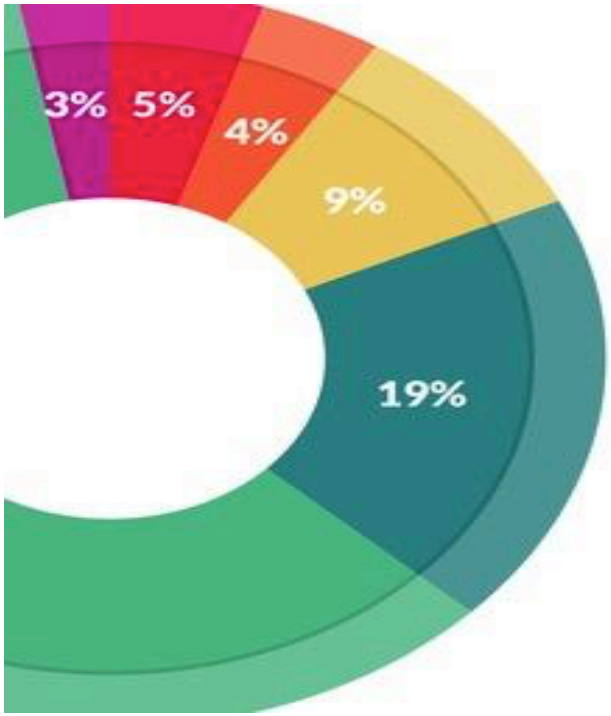
- **Network Throughput**
  - » As we enable greater exchange of data between former silos
- **“Code Shipping” ability**
  - » we will never be able to move all of the data
- **Cost effective cloud computing**
  - » Microservices, by their nature, are likely to arise from small teams
  - » PAAS which will support ‘off the shelf/ready to go’ data science
- **Infrastructure support for Governance/Quality**
  - » Who can access, original or derived data?
  - » Provenance of creation? transforms?
- **Discovery of Data at the Agency**
  - » Data in all areas still rapidly growing
  - » Combining datasets is powerful -- more important than ever to facilitate the finding of data across NASA
  - » Enabling better collaboration and return on data

Summary

- **NASA has been doing "Data Science" for a long time**
  - »Many different missions/centers are engaged in Data Science activities
- **Uneven distribution of experience/uptake**
  - »Science areas @NASA are doing the best job with Big Data, Business units least well
- **Hard issues to solve with the "variety" problem**
- **Data Silos and Data Provenance are pressing issues**
- **Big Data will often require shipping analysis to the data**
- **Need to support 21st century environment for NASA**

Spare Slides

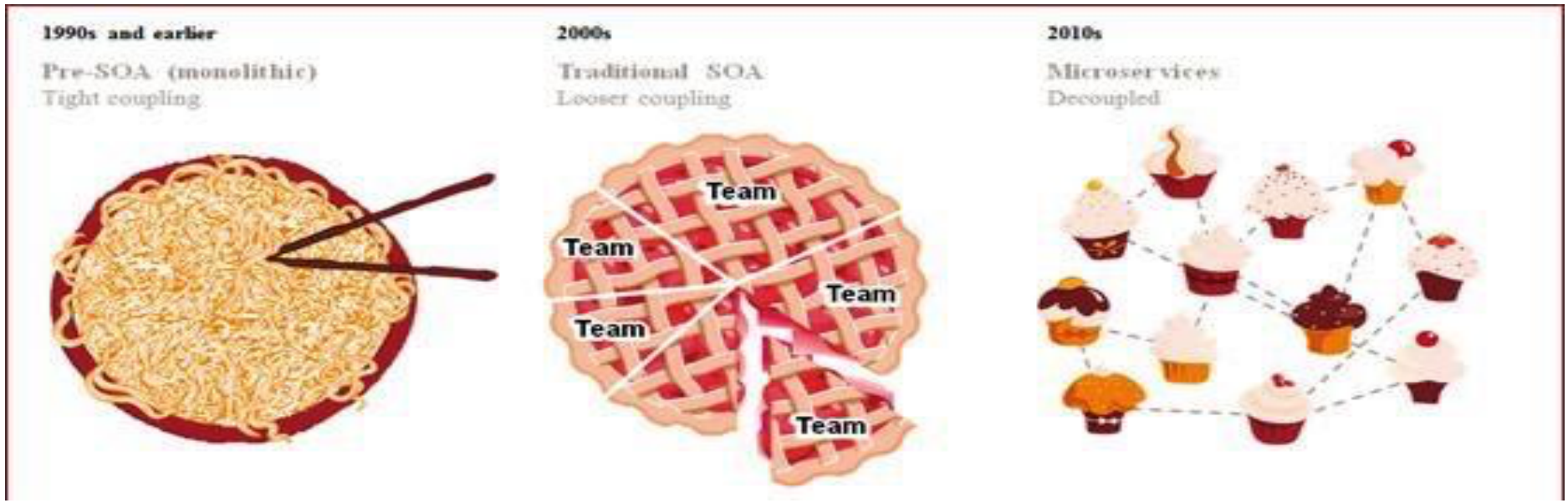
Why APIs?



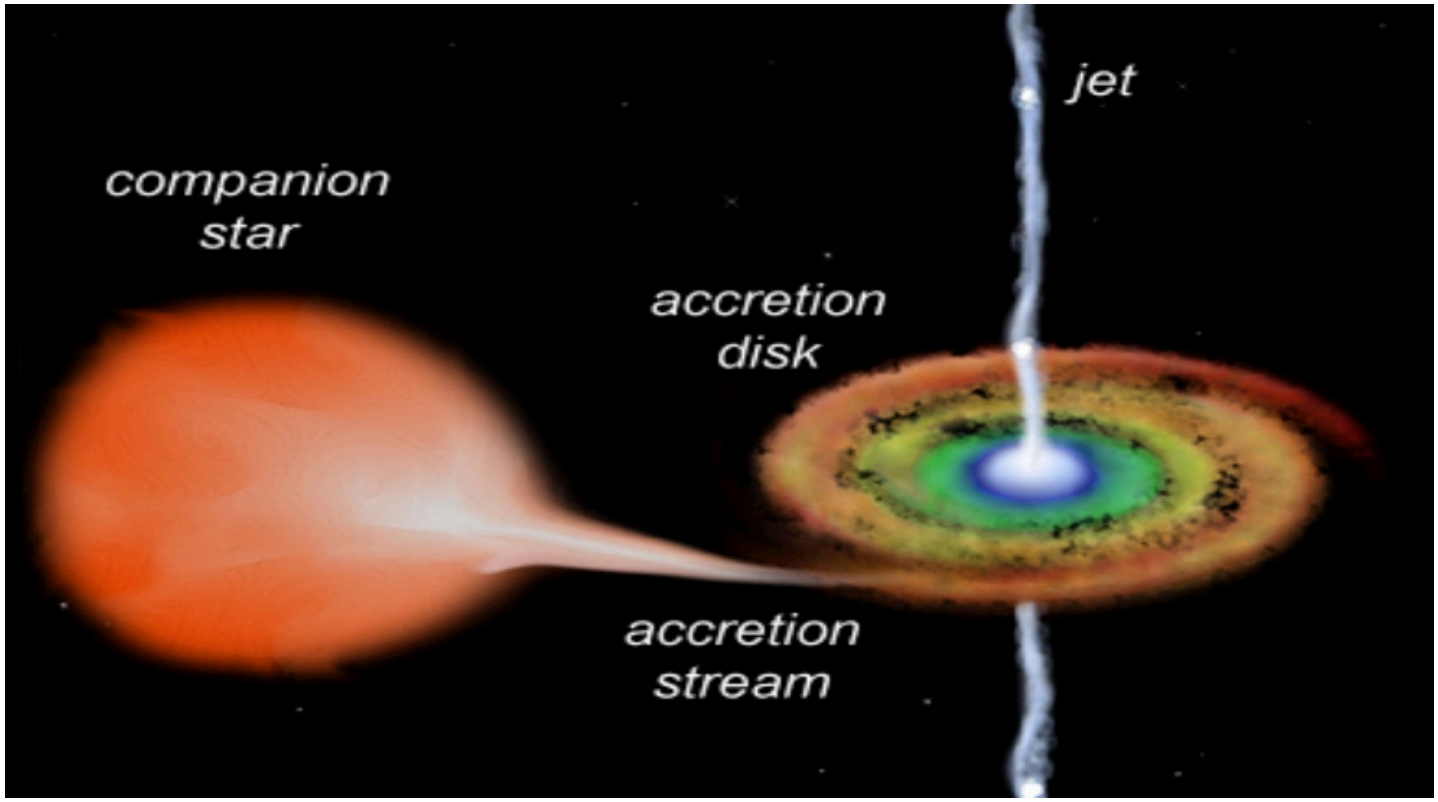
What data scientists spend the most

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Sampling of Current Projects



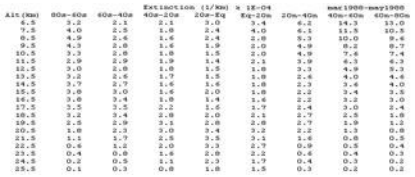
Field of Study: High Energy Astronomy



Finding, Munging & Analyzing the Data

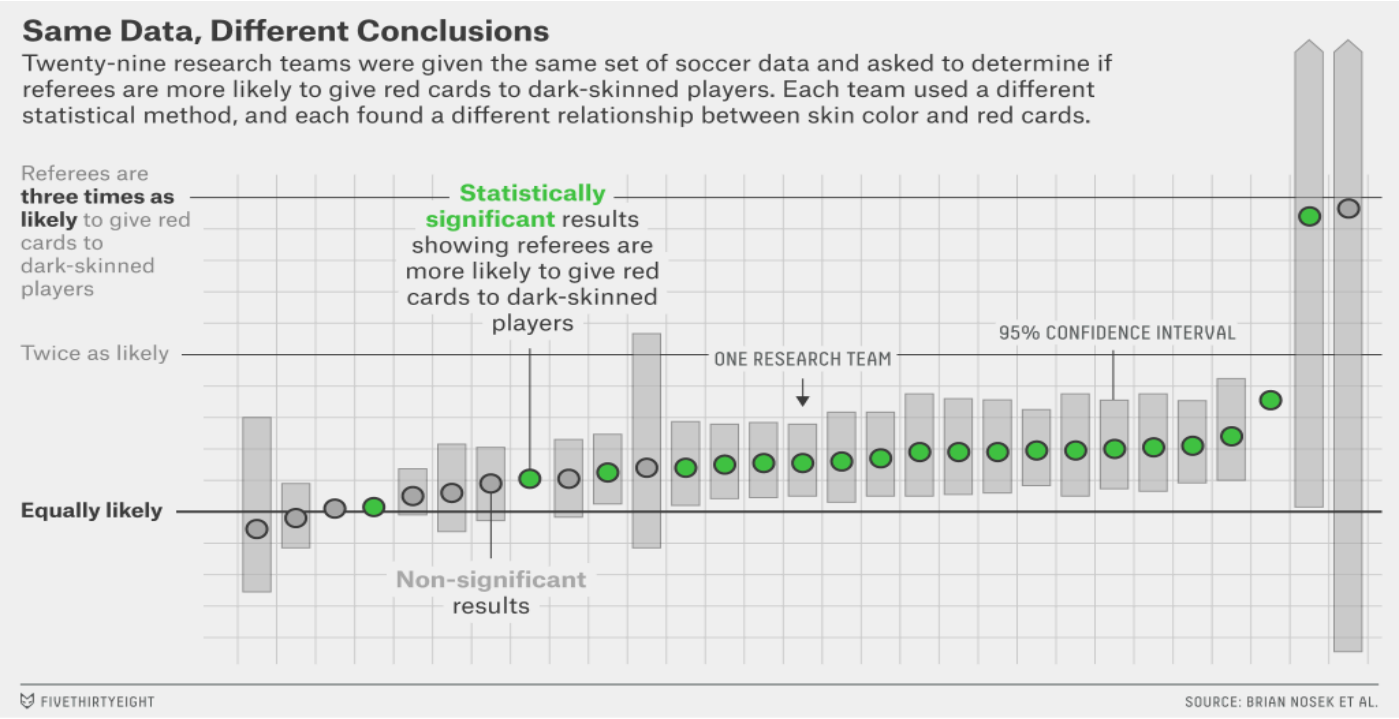
Space-based Data

Useable Data



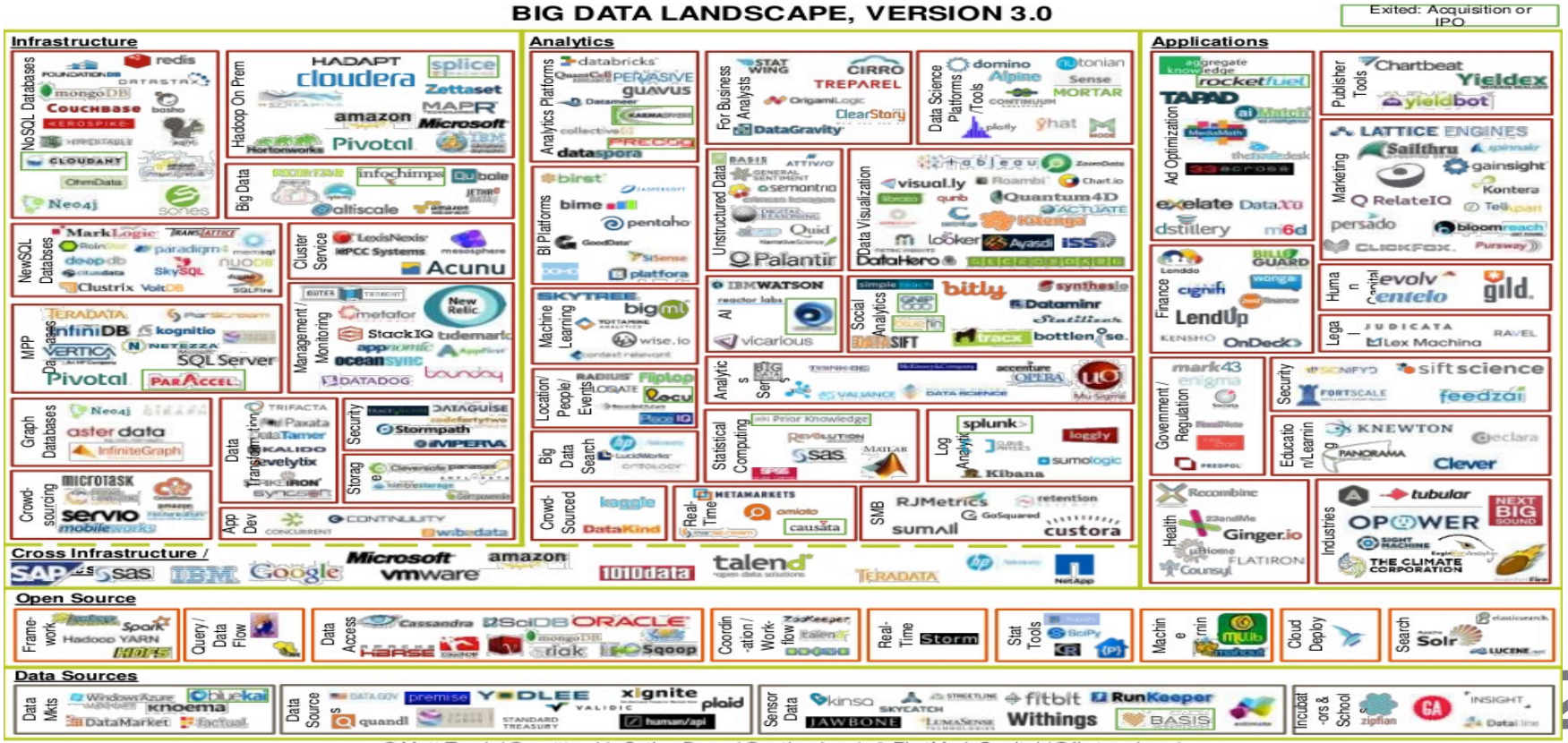
Ground-based Data

Example of the Danger Zone "Work"



http://fivethirtyeight.com/features/science-isnt-broken

Tools & Technologies
Big Data/Data Science Landscape



© Matt Turck (@mattturck), Sutian Dong (@sutiandong) & FirstMark Capital (@firstmarkcap)

Optimizing Processes