▪ What is Data Science?
▪ Data Science activities at OCIO/NASA
▪ Big Data Issues for Data Science
▪ Summary
“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!)

Data

- Raw
  - Red, 192.234.235.245.678, v2.0

Information

- Meaning
  - South facing traffic light on corner of Pitt and George Streets has turned red

Knowledge

- Context
  - The traffic light I am driving towards has turned red

Wisdom

- Applied
  - I better stop the car!
“Problem Datasets”?

LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
“Problem Datasets”? - Characteristics

Variety
- Structured
- Unstructured
- Semi-structured
- All the above

Volume
- Tera-bytes
- Records
- Transactions
- Tables/files

Velocity
- Batch
- Real-time
- Streams
- Near-time

3 Vs of Big Data
Data Scientists: Distinguishing Technical Skills

- Hacking Skills
- Math & Statistics Knowledge
- Substantive Expertise
  - Danger Zone!
  - Traditional Research
  - Machine Learning
  - Data Science
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)
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 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 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 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

Key Recommendations:

• Data Management
• Unified Data Lifecycle
• Data Governance
• Data Analytics Lab
• Data Fellows Program
• Data Stewards
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
Prototypes/Research into methods/tools/platforms in:

- **Data Analytics/Deep Learning**
  » ExMC-telemmedicine, 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

- Processing Data Migration FY15
- Classic Data SSCN 14502
- SARAH-SEVANAH Transition FY16
- Logistics/Manifest Application FY16
- Program Interface API Applications FY15
- New Suit Data Migration FY15

EVA Data Integration

ESOC

- eProcessing Application
- Data Navigator Application FY15
- SEVANAH Application

- OOT
- SARAH Prop Mgmt

CSSS

- eTPS/eDR FY15 CSSS
- End to End FY16 HELO

ARC

- Human System Int. Div
- USRA Contractor - OCIO

- Integration FY15 and FY16
- App Architecture FY15 and FY16
- Data Architecture FY15 and FY16
- Logistics Integ FY16

WESTPrime NASA OCIO Contract

- EVA Safety Data ISS Safety Systems
- EVA Safety Data Migration FY15

Cloud Hosting FY15

EVA NASA OCIO Contract

- ICAM
- STRAW
- NASA OCIO Website

Color Key

- Applications
- Data
- Integration
- EDI Team
- SSCN

New Suit FY15

- Classic Suit/Tools FY16

EVA Safety Data Migration FY15

EVA Safety Data

ISS Safety Systems

EVA Safety Data

Migration FY15
Sampling of Current Projects
Sampling of Current Projects

Sensors
- Biomonitor
- Environmental
- EVA
- Exercise
- Behavioral

Tracking
- Consumables
- Food
- Medications
- Fluids
- Medical Eq.

Data Streaming

Medical Equipment
- Imaging
- 3-D Printing
- Devices
- Medications
- Rehab

Performance
- Behavioral Monitor
- Exercise Monitor

EMS

Intelligence Augmentation

Models
- IMM
- MONSTeR
- Digital Astronaut
- Radiation

References
- Pharmacologic
- Toxicologic
- Medical Imaging
- Training Modules
- Up-To-Date™

Medical Decision-Making
- Telemedicine
- Semi-autonomy
- Autonomy

Training
- Medical Procedures
- Imaging Techniques
- Behavioral Intervention

Vehicle Data Interface
Caregiver Interface

Displays
- Conferencing
- Visualize data
- Training
Sampling of Current Projects
Google Analytics

- Add smoother?
- Linear or smoothed: loess
- Hours of interest: minimum
- Hours of interest: maximum
- Show NHS and other domains?
- NHS users
- Other
- Output required:
  - Visits
  - Bounce rate
  - Time on site
- Choose subdomain: nhs.uk

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)
Observed Trends

- 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
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
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

1990s and earlier
Pre-SOA (monolithic)
Tight coupling.

2000s
Traditional SOA
Looser coupling

2010s
Microservices
Decoupled
Field of Study: High Energy Astronomy
Finding, Munging & Analyzing the Data

Space-based Data

Useable Data

Ground-based Data
Same Data, Different Conclusions
Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.

http://fivethirtyeight.com/features/science-isnt-broken
Tools & Technologies
Big Data/Data Science Landscape

<table>
<thead>
<tr>
<th>Infrastructure</th>
<th>Analytics</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoSQL Databases</td>
<td>Apache Hadoop</td>
<td>BI Platforms</td>
</tr>
<tr>
<td>Big Data</td>
<td>IBM Analytics</td>
<td>Data Source Platforms</td>
</tr>
<tr>
<td>Cloudera</td>
<td>Predictive Analytics</td>
<td>Data Science Tools</td>
</tr>
<tr>
<td>Hadoop On Prem</td>
<td>SAS</td>
<td>Predictive Analytics</td>
</tr>
<tr>
<td>MapR</td>
<td>SAS</td>
<td>Predictive Analytics</td>
</tr>
<tr>
<td>MarkLogic</td>
<td>Teradata</td>
<td>Predictive Analytics</td>
</tr>
<tr>
<td>Neo4j</td>
<td>Alteryx</td>
<td>Predictive Analytics</td>
</tr>
<tr>
<td>MongoDB</td>
<td>TIBCO</td>
<td>Predictive Analytics</td>
</tr>
<tr>
<td>Neo4j</td>
<td>TIBCO</td>
<td>Predictive Analytics</td>
</tr>
<tr>
<td>Graph Databases</td>
<td>SQL Server</td>
<td>Big Data Technologies</td>
</tr>
<tr>
<td>Neo4j</td>
<td>SQL Server</td>
<td>Big Data Technologies</td>
</tr>
<tr>
<td>Neo4j</td>
<td>SQL Server</td>
<td>Big Data Technologies</td>
</tr>
<tr>
<td>Neo4j</td>
<td>SQL Server</td>
<td>Big Data Technologies</td>
</tr>
<tr>
<td>Neo4j</td>
<td>SQL Server</td>
<td>Big Data Technologies</td>
</tr>
</tbody>
</table>

Cross Infrastructure / Open Source

<table>
<thead>
<tr>
<th>Data Sources</th>
<th>Data Sources</th>
<th>Data Sources</th>
<th>Data Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWS</td>
<td>AWS</td>
<td>AWS</td>
<td>AWS</td>
</tr>
<tr>
<td>Azure</td>
<td>Azure</td>
<td>Azure</td>
<td>Azure</td>
</tr>
<tr>
<td>Google</td>
<td>Google</td>
<td>Google</td>
<td>Google</td>
</tr>
<tr>
<td>IBM</td>
<td>IBM</td>
<td>IBM</td>
<td>IBM</td>
</tr>
<tr>
<td>Microsoft</td>
<td>Microsoft</td>
<td>Microsoft</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Oracle</td>
<td>Oracle</td>
<td>Oracle</td>
<td>Oracle</td>
</tr>
<tr>
<td>Salesforce</td>
<td>Salesforce</td>
<td>Salesforce</td>
<td>Salesforce</td>
</tr>
<tr>
<td>Twitter</td>
<td>Twitter</td>
<td>Twitter</td>
<td>Twitter</td>
</tr>
</tbody>
</table>

© Matt Turck (@matturck), Susan Dong (@sudong) & FirstMark Capital (@firstmarkcap)
## Optimizing Processes

### Legacy System Subsystem
- NBL Training Tool Tracking (NB)
- GSE Calib & Per. Maint (CL)
- TPS Routing and WO (WO)
- Configuration Mgmt (PM)
- Procedure Deviation (PD)
- Work Breakdn Struc (WB)
- Anthropometric (AN)

### New Common Functions
- Physical Asset Service (ship, recv, inventory)
- Logical Asset Service (Items, BOM, Event planning)
- Engineering and Fabrication Services
- Quality Management Service (procedure deviation, ...)
- Purchasing Service

### Administrative Services
- Authorization Administration Service
- Menu System Maint (ME)

### New Common Functional Groups
- Logistics
- Configuration Management
- Design Drafting
- Design Engineering
- Quality Assurance
- Purchasing
- CSSS PO&C
- Admin PP&B

<table>
<thead>
<tr>
<th>Legacy Subsystem</th>
<th>New Common Functions</th>
<th>EVA Functional Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBL Training Tool Tracking (NB)</td>
<td>Physical Asset Service (ship, recv, inventory)</td>
<td>Logistics</td>
</tr>
<tr>
<td>GSE Calib &amp; Per. Maint (CL)</td>
<td>Logical Asset Service (Items, BOM, Event planning)</td>
<td>Configuration Management</td>
</tr>
<tr>
<td>TPS Routing and WO (WO)</td>
<td>Engineering and Fabrication Services</td>
<td>Design Drafting</td>
</tr>
<tr>
<td>Configuration Mgmt (PM)</td>
<td>Quality Management Service (procedure deviation, ...)</td>
<td>Design Engineering</td>
</tr>
<tr>
<td>Procedure Deviation (PD)</td>
<td>Purchasing Service</td>
<td>Quality Assurance</td>
</tr>
<tr>
<td>Work Breakdn Struc (WB)</td>
<td>Authorization Administration Service</td>
<td>Purchasing</td>
</tr>
<tr>
<td>Anthropometric (AN)</td>
<td></td>
<td>CSSS PO&amp;C</td>
</tr>
<tr>
<td>Physical Inventory (PH)</td>
<td></td>
<td>Admin PP&amp;B</td>
</tr>
<tr>
<td>Receiving (RE)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document Maint. (DM)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bill of Materials (BM)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Event Bill of Materials (ER)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fabrication Apps. (FB)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engineering Apps (EN)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Readiness Rev (TR)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H/W Limited Life (LL)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality Records (QR)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purchase Requisition (PR)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Application Utilities (AU)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Cleanup Utilities (DC)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SQLs to Menu Dr. Rpts (SQ)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>