Elements of an AI/ML Architecture for NASA

Brian Thomas
HQ/OCIO
11/27/2018
Outline

● About Agency Data Analytics Team
  ○ Our scope of work, sampling of projects

● The Problem: Enabling the best return on ML/AI technologies
  ○ Definition of ML, scope
  ○ Ingredients of ML/AI Data + Processing power
  ○ Understanding key aspects of ML in practice
  ○ Problems with doing ML @NASA

● Solutions / Elements
  ○ Data & Processing

● Summary
About Me

- **What is an “Agency Data Scientist”?**

  “To Understand the Challenges and Capabilities of NASA in Data Science, Big Data and Data Analytics”

- Meet with folks around the agency
- Meet with vendors
- Targeted, pathfinding prototypes
- Inform Data Strategy for Agency
Agency Data Analytics Team

- “Agency” + “Analytics/Data Science” + “Technology & Innovation”
- **Core Activities:**
  - Text Analytics
  - Machine Learning
  - Statistical Modelling
  - Data Visualization
  - Technology Evaluation
  - Strategic Policy

Website: [https://analytics.nasa.gov](https://analytics.nasa.gov)
Some Machine Learning Projects

- Email Classification/Records Management
- Scientific Document Tagging
- Speech to Text
- Network Traffic Anomaly Detection
- ESD Ticket Analysis
What is Machine Learning?

"Field of study that gives computers the ability to learn without being explicitly programmed".

- Arthur Samuel, 1959
Why ML now?

- Data
- Processing Power
- Easier Tooling
Implementing ML

Diagram:
- **Data Providers**
  - Raw Data
  - Raw Data
  - Raw Data

- **Select Data**
- **Pre-Processing**
- **Structured Data**
- **Learning Algorithm**
  - Machine Learning Algorithms
  - Iterate to get best model

- **Golden Model**
- **Deploy Selected Model**
- **Candidate Model**
- **Applications**
Derived Implications

- **Trained from Data means:**
  - Bespoke solution
  - Specialists Needed
    - Cleaning & Feature Extraction Matters
    - Training time vs Execution time tradeoffs
    - Try multiple models/solutions to maximize

- **Big Data ML means:**
  - It's not generally feasible to simply share code and someone compile it to get a solution
  - May need to build off someone else’s solution to create your own
Problems/Roadblocks for ML

**Discovery**
- Word of mouth/“I know a gal/guy”

**Access**
- Siloed Data
- Excessive Restrictions
- Difficult interface (ex. db connector + hidden/mysterious schema/interface)

**Understanding**
- Documentation is human readable; Humans must explain schema

**insufficient Processing power**
- Cost and/or access

**difficult to Share results or build on prior work**
- Hard to replicate, lots of time spent engineering the solution
Solutions

Mixture of Technology, Policy and Culture Changes
Policy & Platforms: Improving Access to Data

Proper Data Governance

- There are no “Data Owners”
- Leverage/Crowdsource Expertise: Data Stewards
- Metadata
  - Publish Data Dictionaries
  - Publish Clear Rules for Access

Attribute-based Data Access

- NAMS Integration?
- Fewer Hoops (NAMS workflows)
- Clearer traceability of who has access to what
Reuse of ML Solutions: Source Code Repositories

- Agency Solution(s)
  - Visible to all agency workers
  - Code pushed up from local repositories
  - Nice adds: issue tracking, Pull Requests, Plugin support, ..
Tackling Sharing Bespoke Code: Containers

Using containers, everything required to make a piece of software run is packaged together in one place.

FROM ubuntu:14.04  
RUN apt-get update  
RUN apt-get install -y apache2  

Example Configuration File
Tackling Code Reuse/Data Access: APIs
Reusing Functionality/Data: Networks of APIs!
Tackling Cost Effective (Cloud) Processing

- **Cloud Computing Efforts**
  - Center-based cloud environments (e.g., ARC, GSFC, MSFC, LRC)
  - Agency cloud moderate environments (better enable cross-center teams)

- **Agency ITIF (w/ WSO)**
  - Template security plan
  - Faster provisioning, clear costing
  - AWS (Google Cloud, Azure?)
Putting it together: Agency Architecture Components

- **Data**: Modern Data Architecture
  - **Finding**:
    - Data Governance Platform (DGP), Container and API Registries
  - **Understanding**:
    - DGP (Data Dictionaries, taxonomies and mappings, provenance and other metadata help quantify data quality)
  - **Access**:
    - API infrastructure, DGP, TBD system(s) for Attribute-based Access
Putting it together: Agency Architecture Components

- **Processing**: Ubiquitous, affordable, processing resources and shared code/containers/products
  - Cost effective cloud computing for the agency
    - Easy interface for use and understandable costing
  - Promote shared computing
    - Agency level source code repositories
    - Service to provide validated container images
    - API Registry, A&A*, OAuth2* (* w/ ICAM)
Who Helps?

- **OCIO**
  - Information Management Program
  - Open Innovation Team
  - Agency Data Analytics and Data Management teams
  - Applications Program
  - Computing
  - Security

- **Center CIO**
  - Data Science and Data Management Teams
  - Security

- **Teams at missions** (you?)
End