Microsoft Garage: Modernizing Data Processing at the Museum of Science

Nicholas Bradford | Tim Petri | Himanshu Sahay

A Major Qualifying Project submitted to Worcester Polytechnic Institute.

Presented 14 December 2016.
Hall of Human Life

- Opened in late 2013
- Fifteen interactive kiosks (link stations) in 5 categories
- Wristband with unique barcode enables a cross-kiosk experience
- Additional exploration from the web browser at home
Existing System
Objectives

- Make the complete data set available in Azure
- Provide insights into visitor usage patterns and exhibit health
- Introduce the idea of anomalous data and monitoring for hardware malfunction
Moving Data to the Cloud

- Set up a SQL database in Azure, similar to the on-premise solution
  - Allows to scale performance on the fly (adding resources)
  - Created with future integration in mind
  - Ready-made integrations with tools such as Power BI, and Azure Machine learning
- Moved full historical data set into Azure
  - 600,000+ visitors and almost 10,000,000 visitor answers
- Created custom views to support dashboard and machine learning models
Rule-Based Outlier Detection

- Found several incorrect data points
- Adopted a rule-based approach to flag incorrect (“outlier”) data
- Tested kiosks in person to force outliers and generate acceptable bounds for each question*
- Recorded in database
- Ran all data through rules to retroactively flag as inlier or outlier

* questions accepting numeric answers
Dashboards

- Set of visualizations and demographic filters
  - Age
  - Gender
  - Time of visit
  - Date of visit

- Live connection between Azure SQL database and Power BI, near real time

- Data processing
  - Relationships between views
  - Conditional columns

- 2 dashboards: exhibit overview and detail view

- Completed 2 rounds of reviews with primary users
Hardware Failure Detection: Motivation

Automatically flag potential hardware failures even when data falls within the outlier bounds.

*Rule-based approach in action.*

*Rules fail if relationships or distribution change.*
Anomaly Model: Multivariate Gaussian

Detect more subtle “anomalies” by fitting a normal distribution and considering covariance.

Contamination = 0%
(trains on 100% of inlier data)

Contamination = 5%
(trains on best 95% of inlier data)
Historical Model: Univariate Gaussian

Set a threshold for acceptable anomaly rate for each kiosk (2 standard deviations above mean).

Typical distribution.  A reasonable cutoff appears.  100% anomalies: probably bad.
Hardware Failure Detection: Azure ML

1. Training data (past year)
2. Extraction (per kiosk)
3. Anomaly Model (find anomalies)
   - \( \uparrow \text{contam.} = \uparrow \text{strict} \)
4. Historical Model (judge anomaly rate)
   - \( \uparrow \text{threshold} = \downarrow \text{alerts} \)
5. Log results (in DB & email)
Putting it All Together: Architecture

Future Work
● Integration with existing Hall of Human Life system
● Testing hardware failure detection system
Dashboard Demo
Thank you!
References

(1) Museum of Science: Image from Hall of Human Life http://exhibits.mos.org/
(2) Cloud database icon: https://www.caspio.com/wp-content/uploads/2015/05/caspio-features-illustr_cloud-data_3_2x.png
Hall of Human Life Overview

Today's Date
Wednesday, December 14

Visitors Today
601

Top Exhibit Today
Balance
238

Visitors Per Hour
60

Hall of Human Life Overview

Visitors per Kiosk

% Outlier Data

Gender Distribution

Age Distribution

Average Dwell Time

Filters

Time of Day
- Select All
- 09:00 - 09:59
- 10:00 - 10:59
- 11:00 - 11:59
- 12:00 - 12:59
- 13:00 - 13:59
- 14:00 - 14:59
- 15:00 - 15:59
- 16:00 - 16:59
- 17:00 - 17:59
- 18:00 - 18:59
- 19:00 - 19:59
- 20:00 - 20:59

Age Bucket
- Select All
- 0-9
- 10-19
- 20-29
- 30-39
- 40-49
- 50-59
- 60-69
- 70-99

Gender
- Select All
- female
- male

Date Picker
5/16/2016
5/25/2016
Sharing Reports

Share dashboard

Share
Access

Grant access to
Enter email addresses

Include an optional message...

- Recipients will have access to the same data, reports, and workbooks as you have in this dashboard, unless their access is restricted by row-level security defined for the dataset. [Learn more]

- Allow recipients to share your dashboard
- Send email notification to recipients

Share | Cancel

- Save as
  Save a copy of this report

- Print
  Print current page

- Publish to web
  Embed this report for public access by anyone on the Internet

- Embed in SharePoint Online
  Get a link to securely embed this report in a SharePoint page

- Export to PowerPoint (Preview)
  Export this report as a PowerPoint presentation

- Download report (Preview)
  Download a .pbix copy