Manage Your Datasets

Eng Teong Cheah
MVP Visual Studio &
Development Technologies

Agenda

Categorizing your data
Importing data
Exploring & transforming data

Categorizing your data

Data structure

Growth in the amount of unstructured data typically dwarfs that for structured data, resulting in significant challenges for the IT departments that must look after all this data.

Machine Learning can import data from variety of data sources with varying degrees of structure, such as Excel, Blob Storage, big data, SQL Server and even simple CSV and text files.

Structured vs. Unstructured data

Structured data refers to information with a high degree of organization, such that inclusion in a relational database is seamless and readily searchable by simple, straightforward search engine algorithms or other search operations; whereas unstructured data is essentially the opposite.

Structured vs. Unstructured data

The lack of structure makes compilation a time and energy-consuming task. It would be beneficial to a company across all business strata to find a mechanism of data analysis to reduce the costs unstructured data adds to organization.

Big Data Challenges

Most experts define big data in terms of the three Vs. You have big data if your data stores have the following characteristics:

- Volume
- Velocity
-Variety

Big Data Challenges

These 3 characteristics cause many of the challenges that organizations encounter in their big data initiatives. Some of the most common of those big data challenges include the following:

1. Dealing with data growth
2. Generating insights in a timely manner
3. Recruiting and retaining big data talent
4. Integrating disparate data sources
5. Validating data
6. Securing big data
7. Organization resistance

Importing data

Importing Data

To use your own data in Machine Learning Studio to develop and train a predictive analytics solution, you can:

- Upload data from a **local file** ahead of time from your hard drive to create a dataset module in your workspace
- Access data from one of several **online data sources** while your experiment is running using Import Data module.

Data formats and data types supported

You can import a number of data types into your experiment, depending on what mechanism you use to import data and where it's coming from:

- Txt, CSV, TSV, Excel
- Azure table, Hive table
- SQL database table
- OData
- SVMLight
- ARFF, zip

Data formats and data types supported

The following **data types** are recognized by Machine Learning Studio:

- String
- Integer
- Double
- Boolean
- DateTime
- TimeSpan

Exploring and transforming data

Exploring and Transforming data

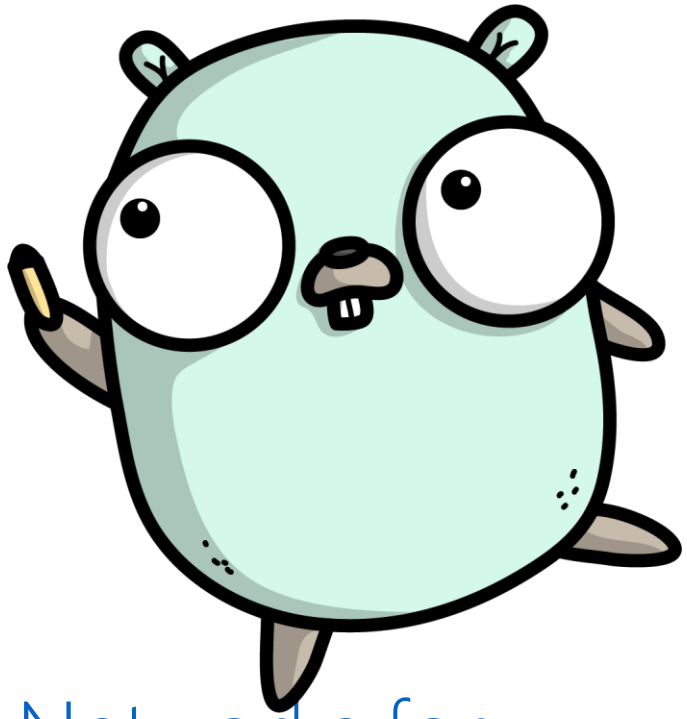
4 different storage environments that are typically used in the Data Science Process:

- **Azure blob container** data is explored using the Pandas Python package
- **SQL Server** data is explored by using SQL and by using a programming language like Python
- **Hive table** data is explored using Hive queries.
- **Azure Machine Learning (AML) Studio** data is explored using AML modules..

Demo

Managing datasets

Resources



TutorialsPoint

Microsoft Docs

Lecture Collection | Convolutional Neural Networks for Visual Recognition(Spring 2017)

Python Numpy Tutorial

Image Credits: @ashleymcnamara

Thank you

Eng Teong Cheah
Microsoft MVP Visual Studio & Development Technologies
Twitter: @walkercet
Github: https://github.com/ceteongvanness
Blog: https://ceteongvanness.wordpress.com/
Youtube: http://bit.ly/etyoutubechannel