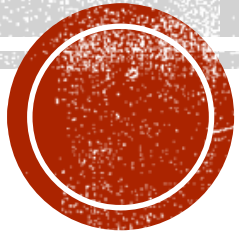
AZURE MACHINE LEARNING (AML) WALKTHROUGH

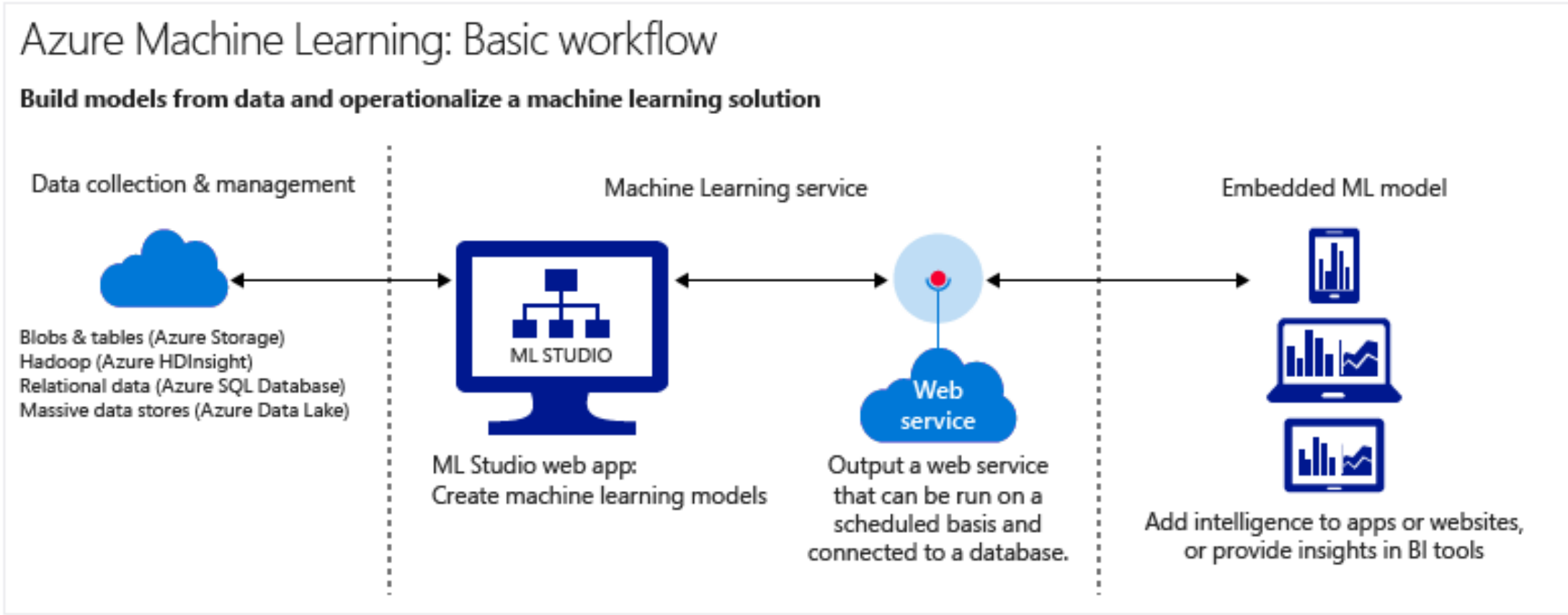
What it is, how much it costs, grants, and a simple demo



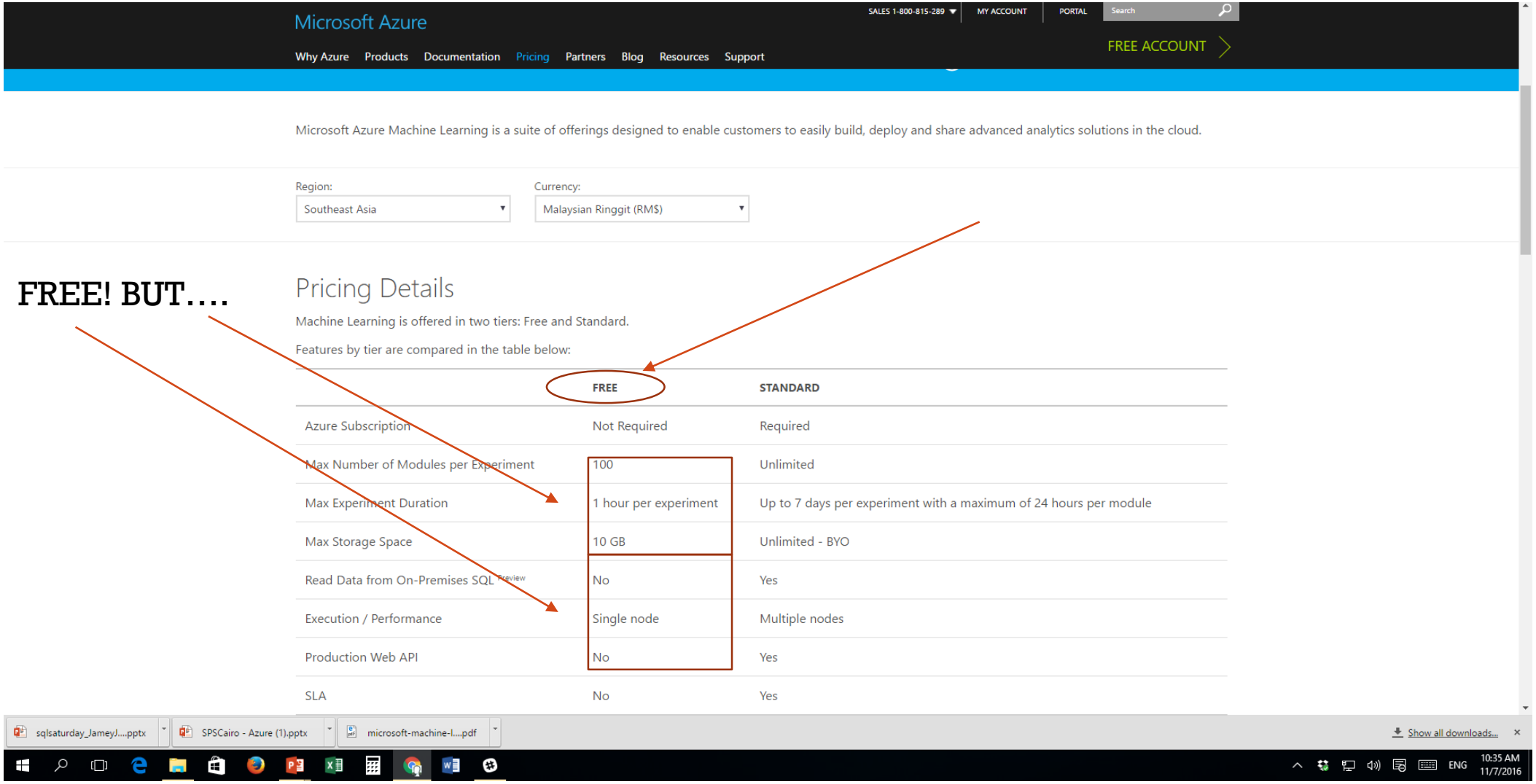
WHAT IS AML?

- Azure Machine Learning is a powerful cloud-based predictive analytics service that makes it possible to quickly create and deploy predictive models as analytics solutions.
- Azure Machine Learning not only provides tools to model predictive analytics, but also provides a fully-managed service you can use to deploy your predictive models as ready-to-consume web services. Azure Machine Learning provides tools for creating complete predictive analytics solutions in the cloud: Quickly create, test, operationalize, and manage predictive models. You do not need to buy any hardware nor manually manage virtual machines.
- (this is the official Microsoft blurb)

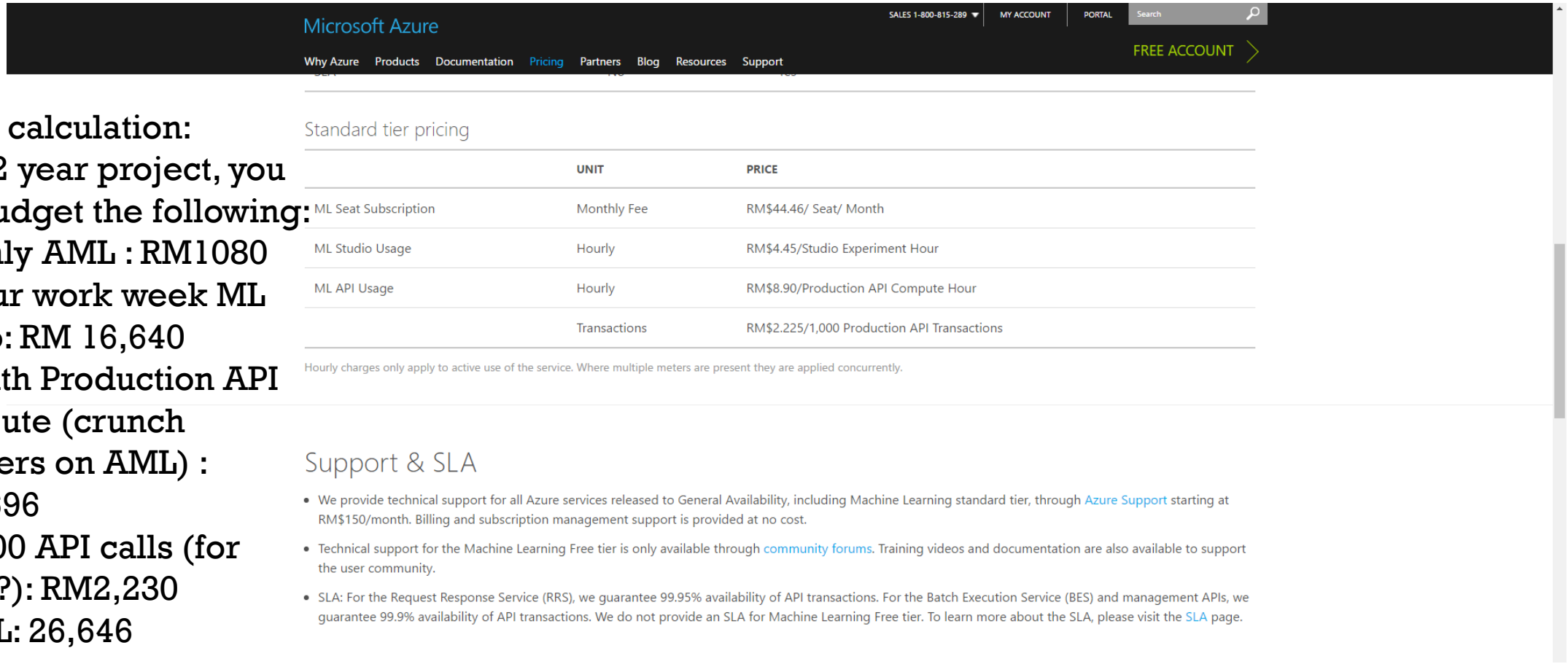
AML BASIC WORKFLOW



HOW MUCH WILL IT COST ME?



BUT IF YOU HAVE THE FUNDS TO PAY?(SERVICES, RENTAL UNDER GRANTS)



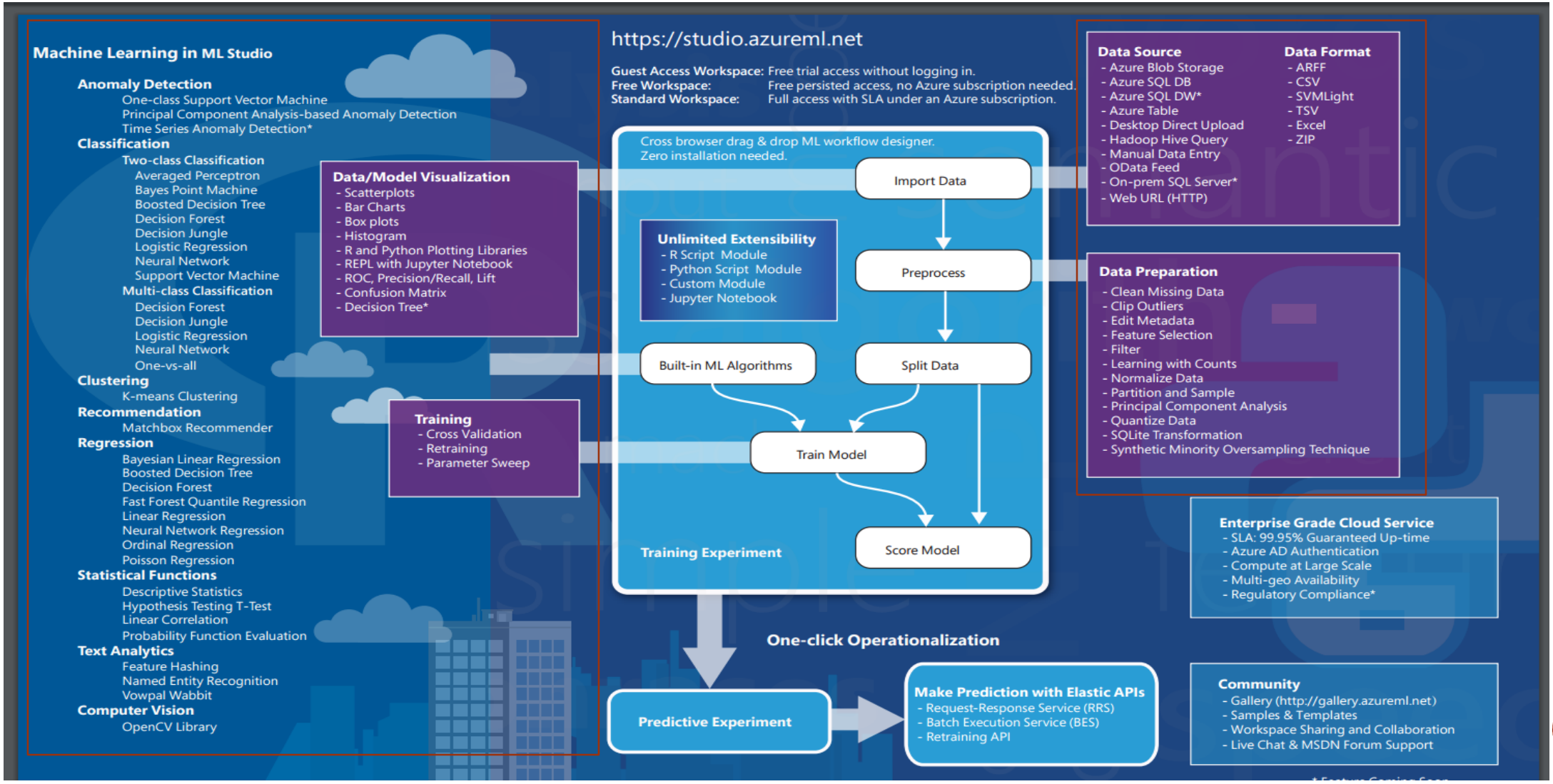
Quick calculation:
For a 2 year project, you can budget the following:
Monthly AML : RM1080
40 hour work week ML Studio: RM 16,640
1 month Production API Compute (crunch numbers on AML) : RM6,696
200,000 API calls (for demo?): RM2,230
TOTAL: 26,646

AML STUDIO (ANYTHING STUDIO IS A MS PRODUCT…)

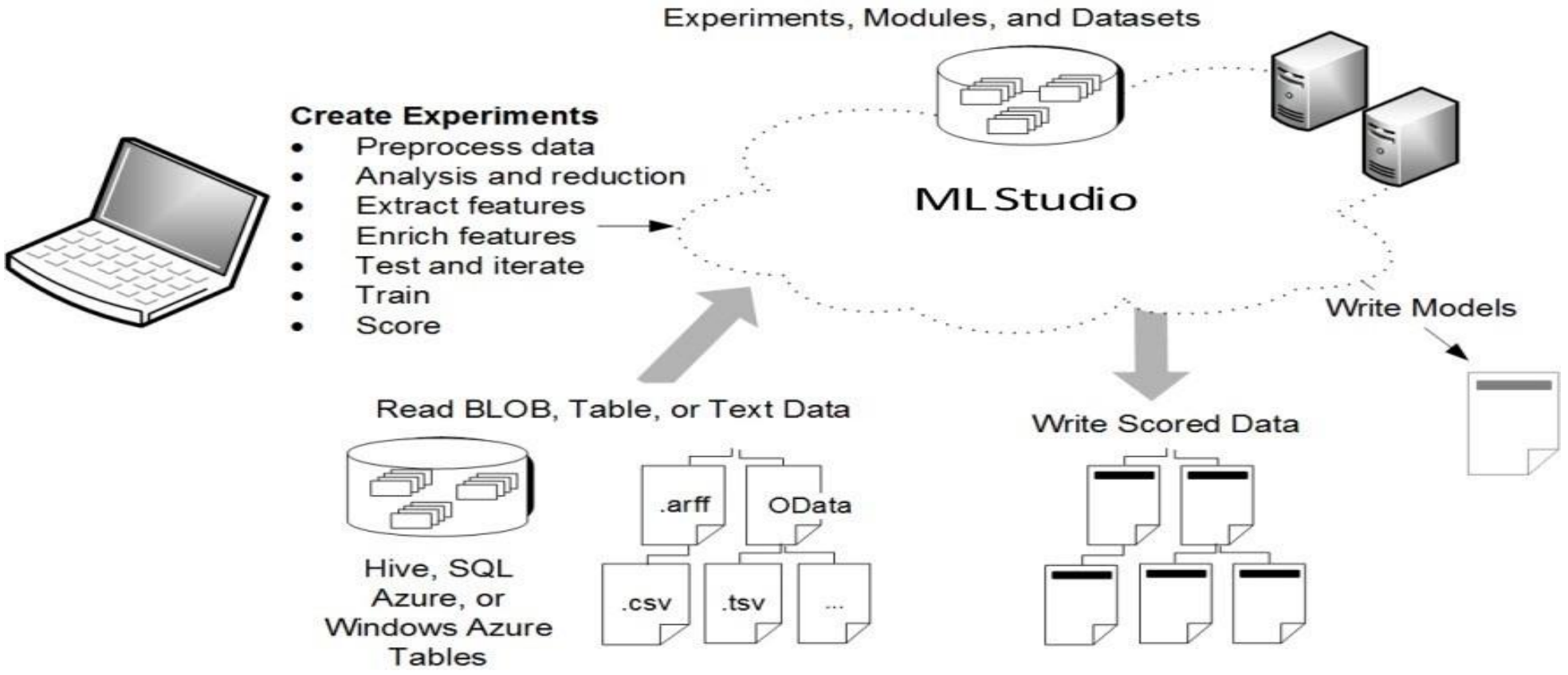
- **Azure Machine Learning Studio** gives you an interactive, visual workspace to easily build, test, and iterate on a predictive analysis model. You drag-and-drop ***datasets*** and analysis ***modules*** onto an interactive ***canvas***, connecting them together to form an ***experiment***, which you ***run*** in Machine Learning Studio. To iterate on your model design, you ***edit*** the experiment, ***save*** a copy if desired, and run it again. When you're ready, you can convert your ***training experiment*** to a ***predictive experiment***, and then ***publish*** it as a ***web service*** so that your model can be accessed by others.

Useful as many funding agencies wants to incorporate output of research into operational systems – web services, APIs, packages and libraries will achieve this

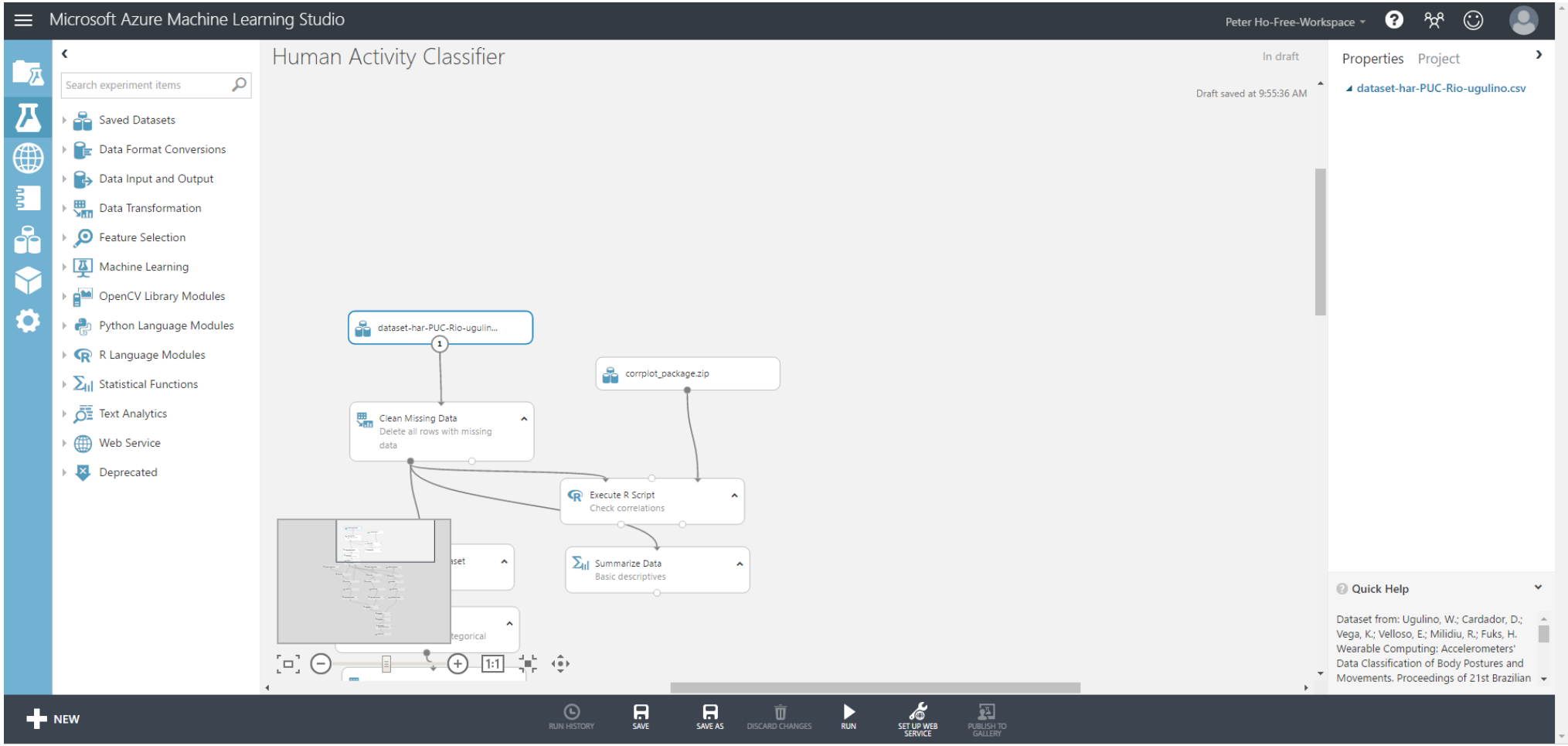
ALGOS SUPPORTED IN AML STUDIO (SUPPORTS EXTENSION VIA R AND PYTHON)



DRAG AND DROP ML (LIKE WEKA AND RAPIDMINER)



AML STUDIO IDE



COMPONENTS OF AN AML EXPERIMENT

- An experiment consists of datasets that provide data to analytical modules, which you connect together to construct a predictive analysis model. Specifically, a valid experiment has these characteristics:
- The experiment has at least one dataset and one module
- Datasets may be connected only to modules
- Modules may be connected to either datasets or other modules
- All input ports for modules must have some connection to the data flow
- All required parameters for each module must be set

DATASETS (10 GB LIMIT, NOT QUITE BIG DATA READY)

- A dataset is data that has been uploaded to Machine Learning Studio so that it can be used in the modeling process.
- The following types of data can expand into larger datasets during feature normalization, and are limited to less than 10 GB:
  - Sparse
  - Categorical
  - Strings
  - Binary data (including images from BLOB storage)
- The following modules are limited to datasets less than 10GB:
  - Recommender modules
  - SMOTE module
  - Scripting modules: R, Python, SQL
  - Modules where the output data size can be larger than input data size, such as Join or Feature Hashing.
  - Cross-validation, Tune Model Hyperparameters, Ordinal Regression and One-vs-All Multiclass, when number of iterations is very large.
- For datasets larger than a few GB (FAQ sais a couple), you should upload data to Azure storage or Azure SQL Database or use HDInsight, rather than directly uploading from local file.

MODULES

- A module is an algorithm that you can perform on your data. Machine Learning Studio has a number of modules ranging from data ingress functions to training, scoring, and validation processes.
- 4 modules can be executed in parallel (good for comparing different classifiers/algorithms)
- Supported algorithms has been shown in the previous slide, but here's the official blurb from Microsoft
- Machine Learning Studio provides state of the art algorithms, such as Scalable Boosted Decision trees, Bayesian Recommendation systems, Deep Neural Networks, and Decision Jungles developed at Microsoft Research. Scalable open-source machine learning packages like Vowpal Wabbit are also included. Machine Learning Studio supports machine learning algorithms for multiclass and binary classification, regression, and clustering

DEMO (SOME STATS BEFORE WE RUN IT FOR REAL)

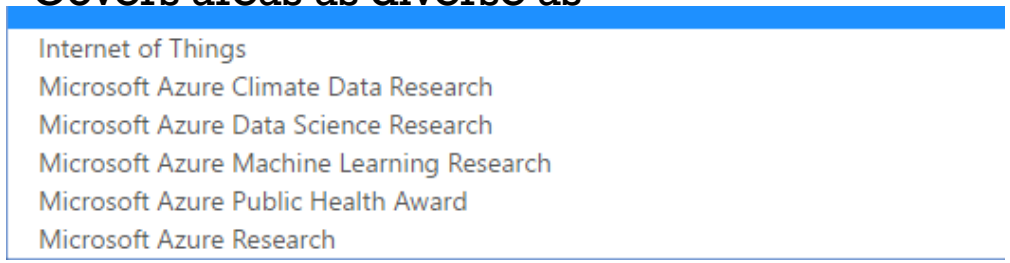
- Data size: 14.25 MB
- Experiment size : 0.03 GB out of 10 GB
- Runtime : 2 minutes 20 second
- Objectives :
- 1.To classify actions performed by subjects in experiments as sitting, standing, walking, standingup or sitting down
- 2.To publish scored probabilities and class label as a web service
- 3. To consume published web service via R (should be easily done via C# and Python too)

PERSONAL IMPRESSIONS

- AML can be a useful tool to expose non data scientists / programmers to solving their problems using ML
- Learning curve is not difficult and it would be quick easy to pick up for someone with experience with similar drag and drop ML tools such as Weka, RapidMiner, Knime etc. Exposing your FYP students to it would be helpful to build confidence in ML before weaning them off AML and working on your favorite ML language
- On paper, the ability to expose the AML experiment as a webservice will be useful in integrating the solution into a production system. The cost incurred needs to be factored in when considering AML as a solution.
- Experiments are essentially repeatable , and can be validated by expert audience. Experiments can also be published to a public gallery
- Current drawbacks would be the lack of support for large datasets, which may incur additional costs for data storage on a AML friendly platform

OPPORTUNITY!

- Microsoft periodically issues a Call for Proposal for Microsoft Azure for Research Award or a Azure for Education (if you want to teach…)
- All it needs is a 3 pager proposal with 1000 word limit
- Covers areas as diverse as



- Undergrads and PG can apply through their supervisors
- The size of the awards is large. A request for **20 TB of storage and 200,000 hours of compute time is reasonable**. A request for 1 GB of storage and 20 hours of compute time is too small for an award.
- Next deadline is August 15, 2016. This can supplement your own research resources

USEFUL LINKS

- https://studio.azureml.net/
- http://www.kdnuggets.com/2014/12/ibm-watson-analytics-microsoft-azure-machine-learning-p1.html
- https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-cheat-sheet/
- http://www.kdnuggets.com/2014/11/microsoft-azure-machine-learning.html
- https://www.microsoft.com/en-us/research/academic-program/microsoft-azure-for-research/
- https://www.microsoft.com/en-us/research/academic-program/microsoft-azure-for-research/