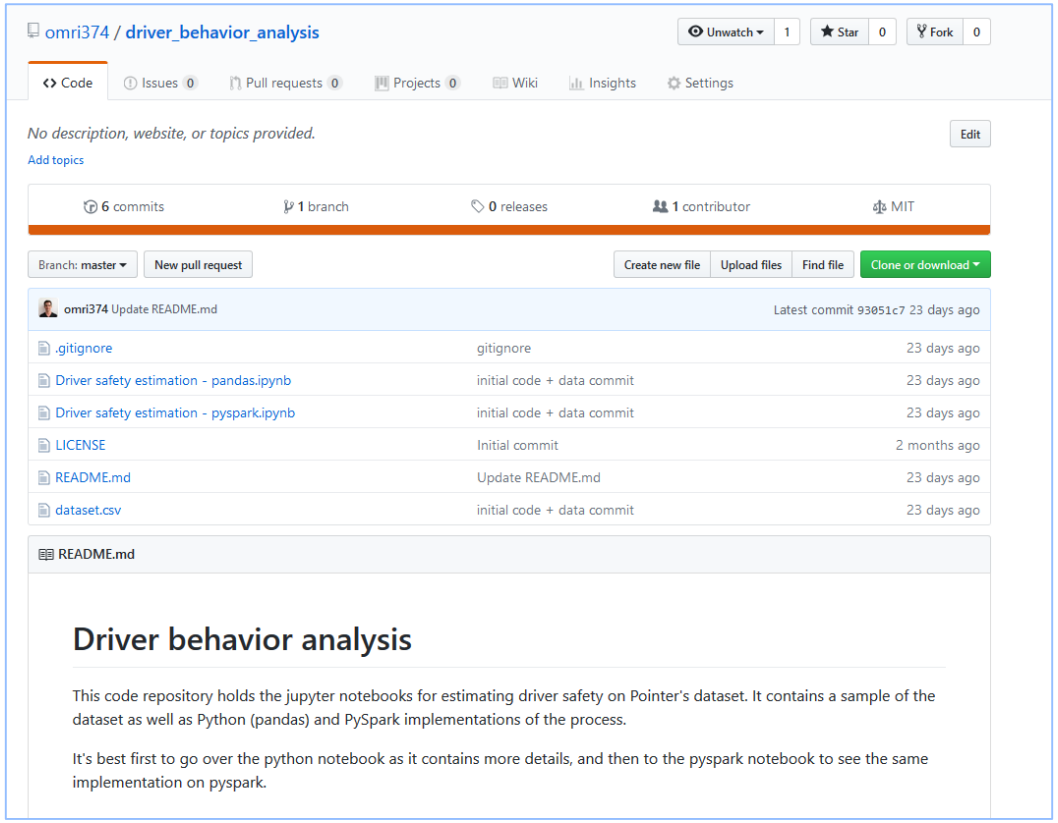
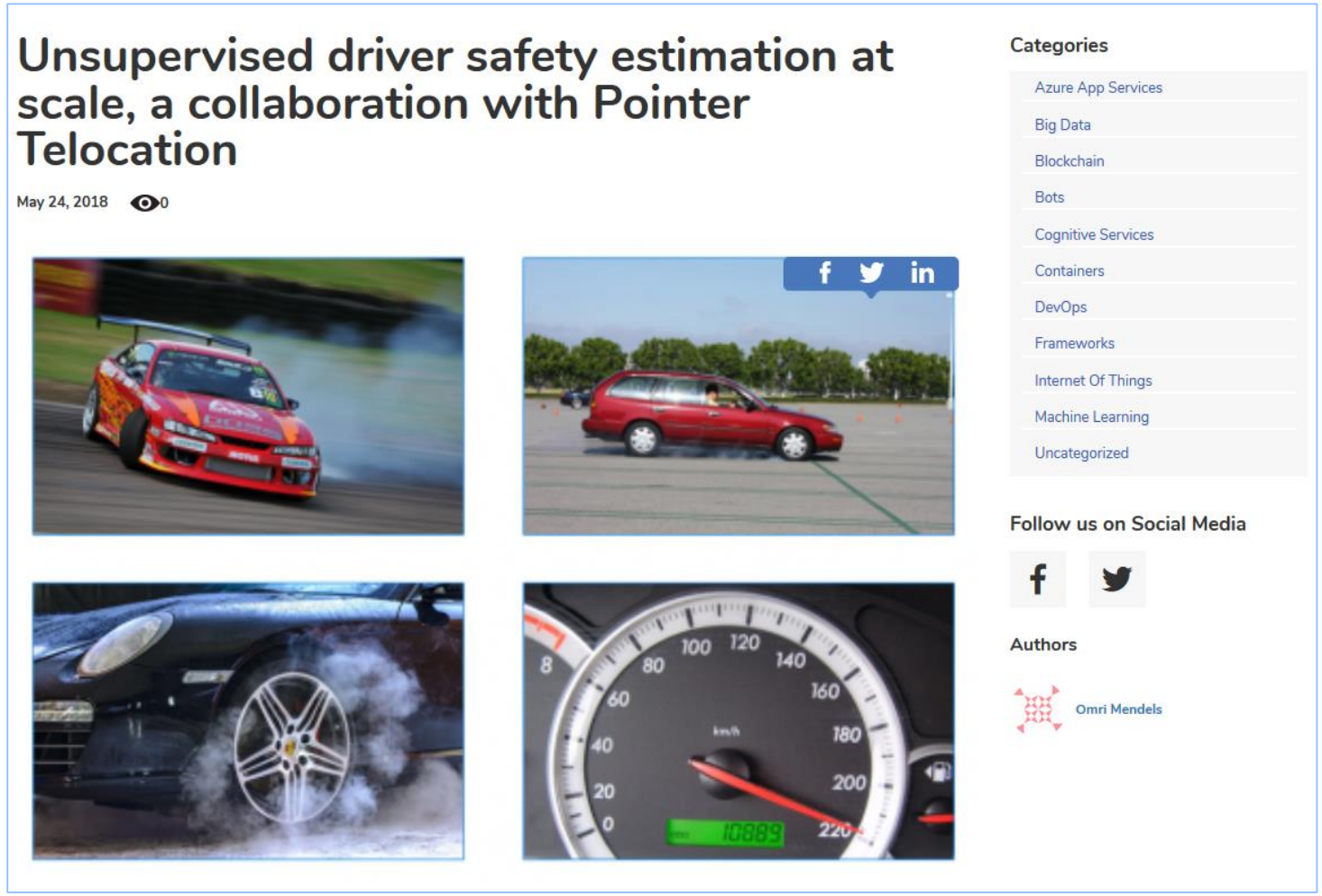
R on the cloud

Omri Mendels
omri.mendels@Microsoft.com

About me

- Part of a team that work with partners on novel projects
- Write **open source code** and **blogs** with the solutions
- *#RealLifeCode*





We'll cover today:

Using the Data Science Virtual Machine on Ubuntu

R Studio Server

Jupyter Notebooks and Azure Notebooks

R on Spark

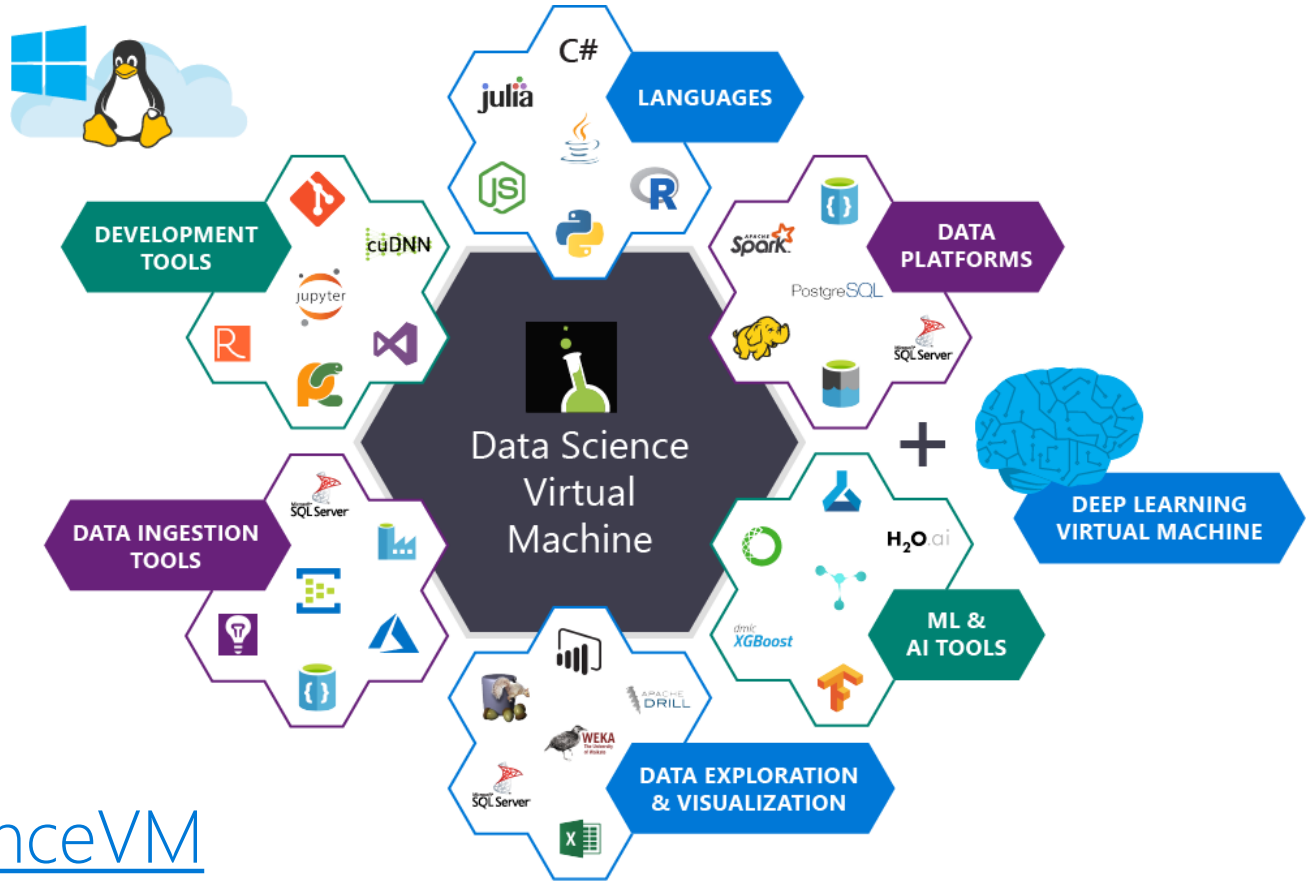
Links and Resources

* Many of the slides are courtesy of David Smith

Data Science Virtual Machine

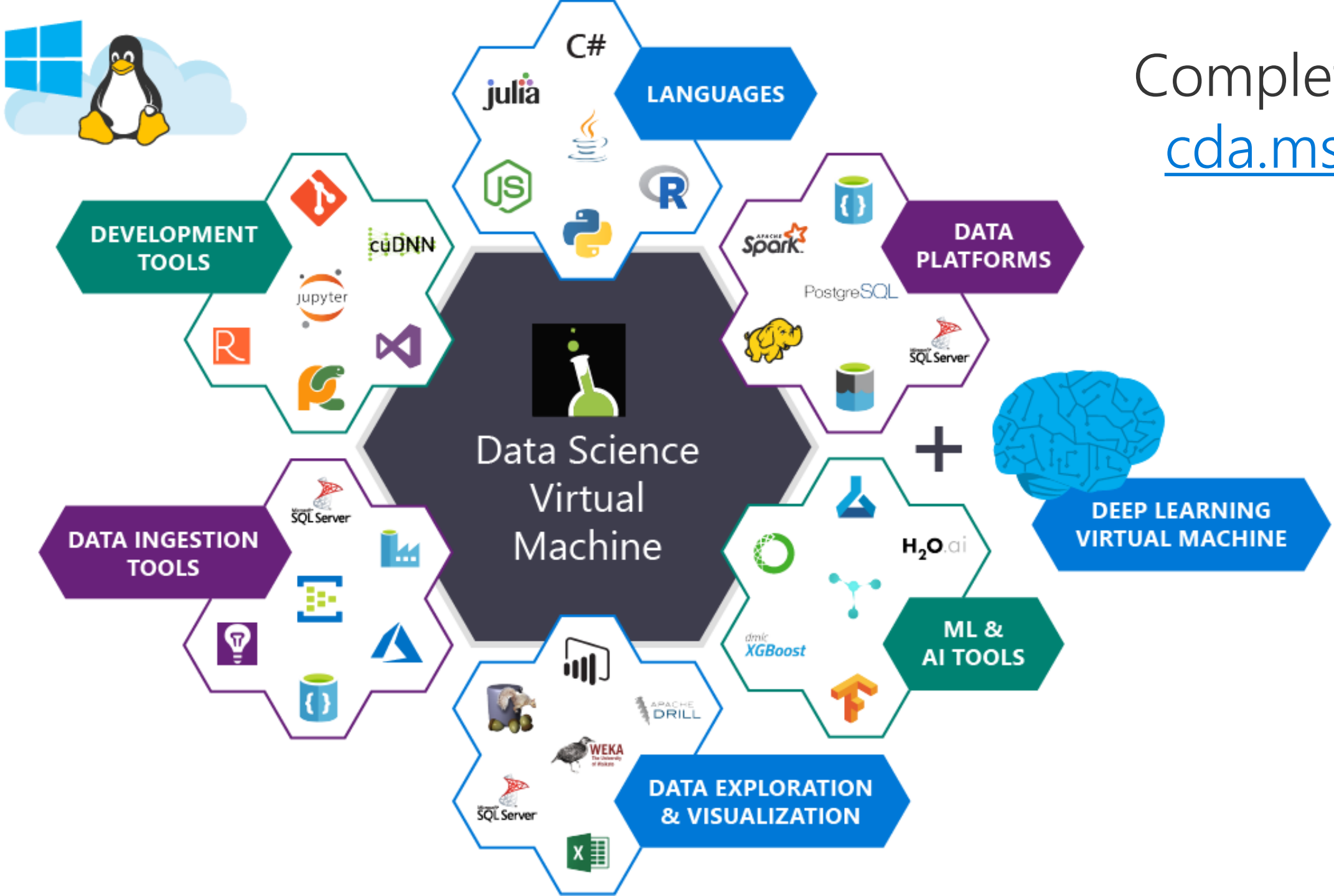
What is the Data Science VM?

- VM image with many Data Science & ML tools pre-installed
- Regularly updated and tested for compatibility
- Windows and Ubuntu
- Pay only standard VM rates



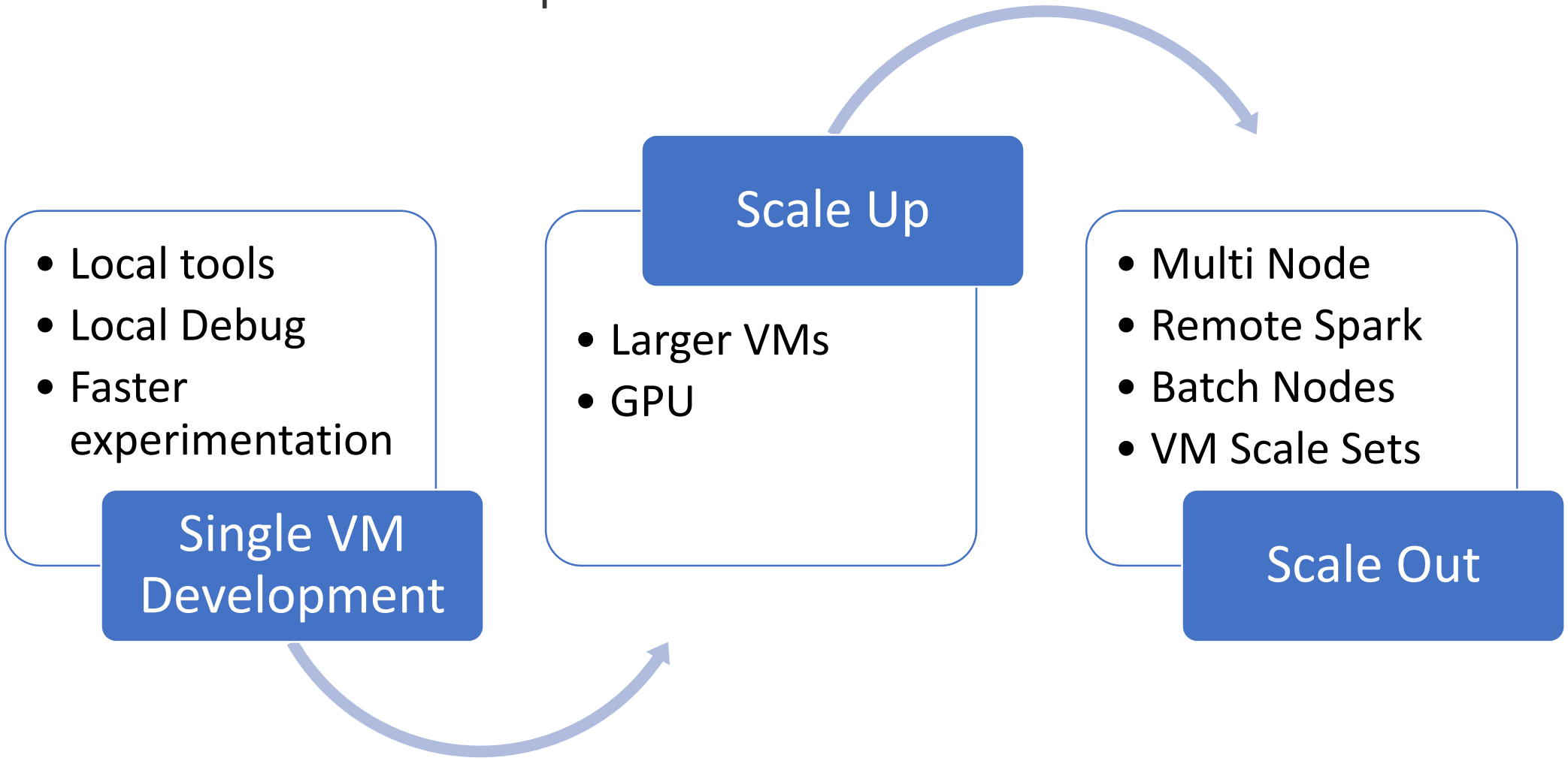
https://aka.ms/dsvm/overview

https://github.com/Azure/DataScienceVM



Complete list:
cda.ms/sN

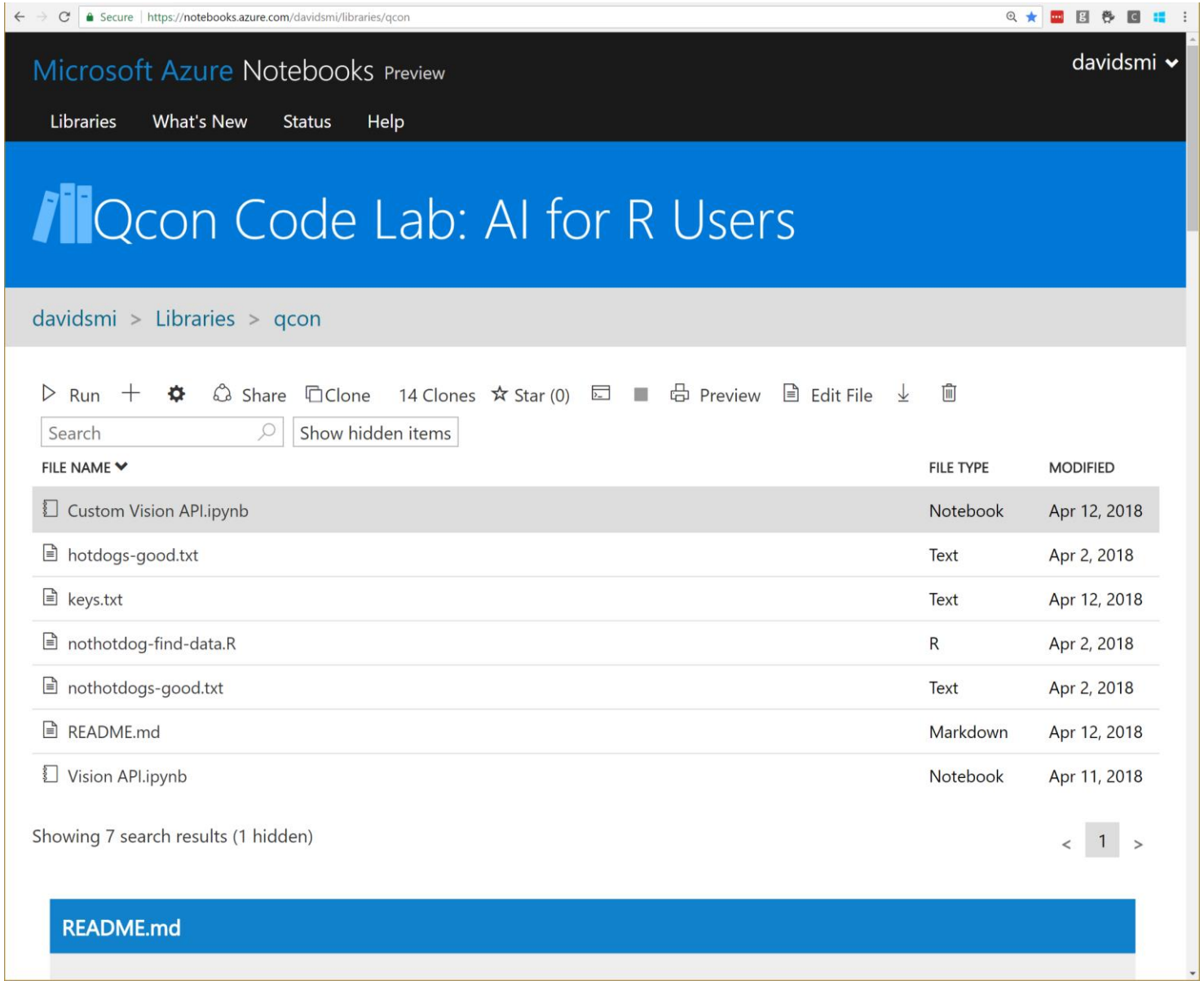
Iterative Development with DSVM



Connecting to the DSVM

- Windows: Remote Desktop
  - Tip: download RDP file for quick connections
- Ubuntu:
  - `ssh`
  - X windows (X2Go client)
- App-Specific Interfaces
  - JupyterHub: https://xxx.xxx.xxx.xxx:8000/
  - RStudio Server: http://xxx.xxx.xxx.xxx:8787/
    - how to launch on DSVM: https://cda.ms/s0

Azure Notebooks: notebooks.azure.com



Distributed ML with Spark and R

What is Spark?

- Distributed data processing engine
  - Store and analyze massive volumes in a robust, scalable cluster
- Successor to Hadoop
  - in-memory engine 100x faster than map-reduce
- Highly extensible, with machine-learning capabilities
  - Supports Scala, Java, Python, R, SQL …
- Largest open-source data project
  - Apache project with 1000+ contributors
- Managed cloud services available
  - Azure Databricks (completely managed analytics environment) & HDInsight (managed cluster)

Spark + R

- Two main packages
  - SparkR (official package)
    - Write R code that translates to
  - sparklyr (unofficial but integrates better with dplyr)

- How to use Spark with R
  - Run Spark locally
  - Provision a cluster in the cloud (either yourself or a managed solution)
  - Use DataBricks – a managed cluster that provides additional cluster/jobs/code management capabilities

Demo: R on spark with sparklyr

- sparklyr: R interface to Spark
  - open-source R package from RStudio
- Move data between R and Spark
- “References” to Spark Data Frames
  - Familiar R operations, including dplyr syntax
  - Computations offloaded to Spark cluster, and deferred until needed
  - CPU/RAM/Disk consumed in cluster, not by R
- Interfaces to Spark ML algorithms
- Integrated with RStudio Server

Demo: Machine Learning with sparklyr

Links and references

Provision new DSVM: https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/dsvm-ubuntu-intro

Azure Notebooks: https://notebooks.azure.com

Sparklyr: http://spark.rstudio.com/

DataBricks: https://databricks.com/

Sparklyr Demo: https://gist.github.com/omri374/90c51349294298bda1161eab8220495d

Thank you!