



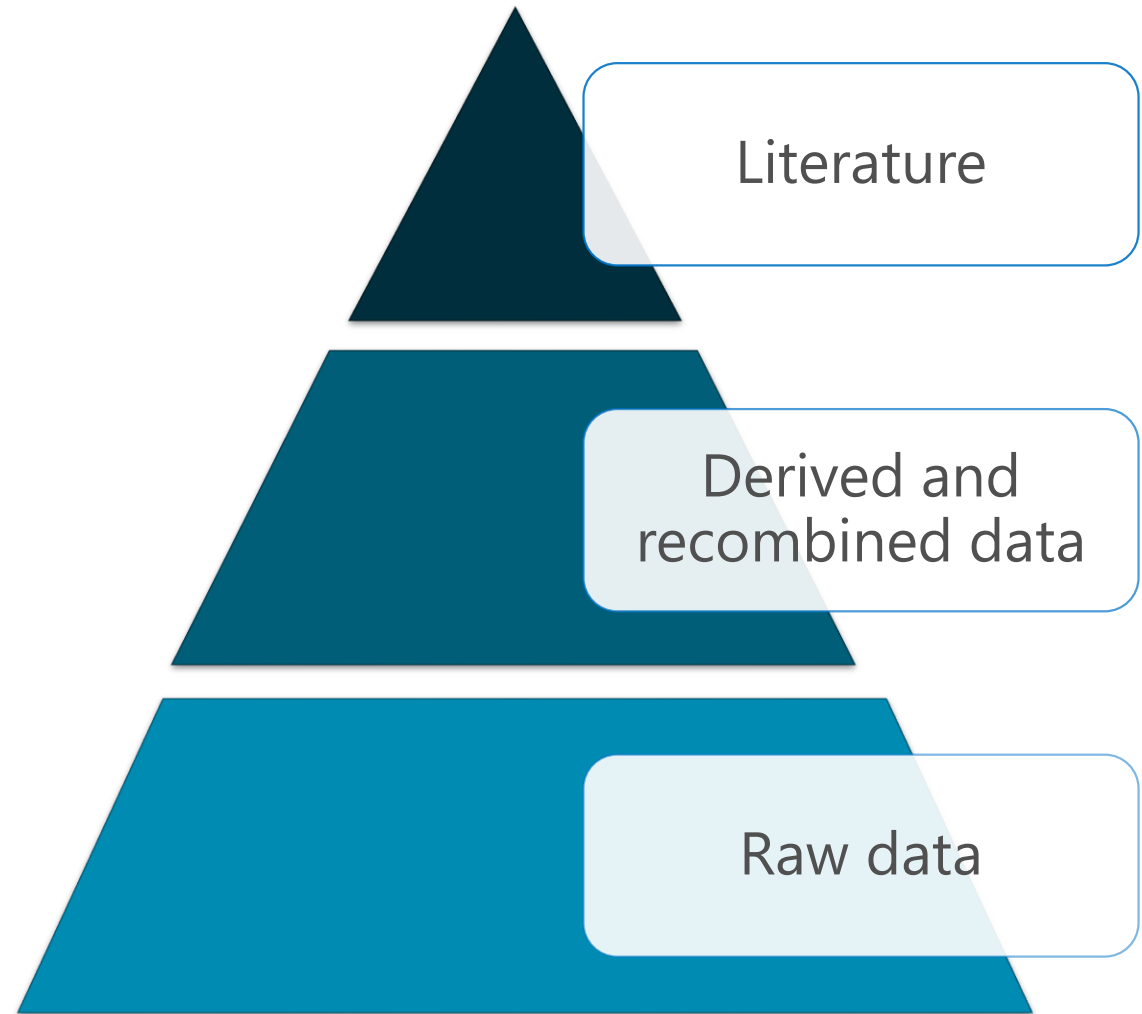
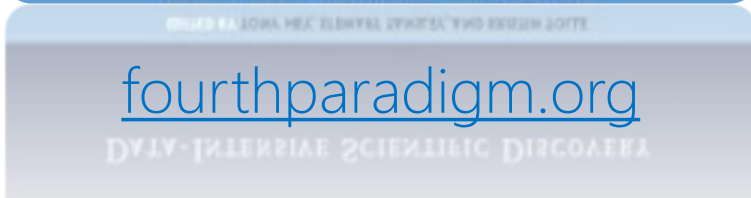
From Data to Decisions:  
*New opportunities for data-driven research  
and machine learning*

ICSTI Annual Conference, 10/20/14 Tokyo

Alex Wade | Director – Scholarly Communication  
Microsoft Research



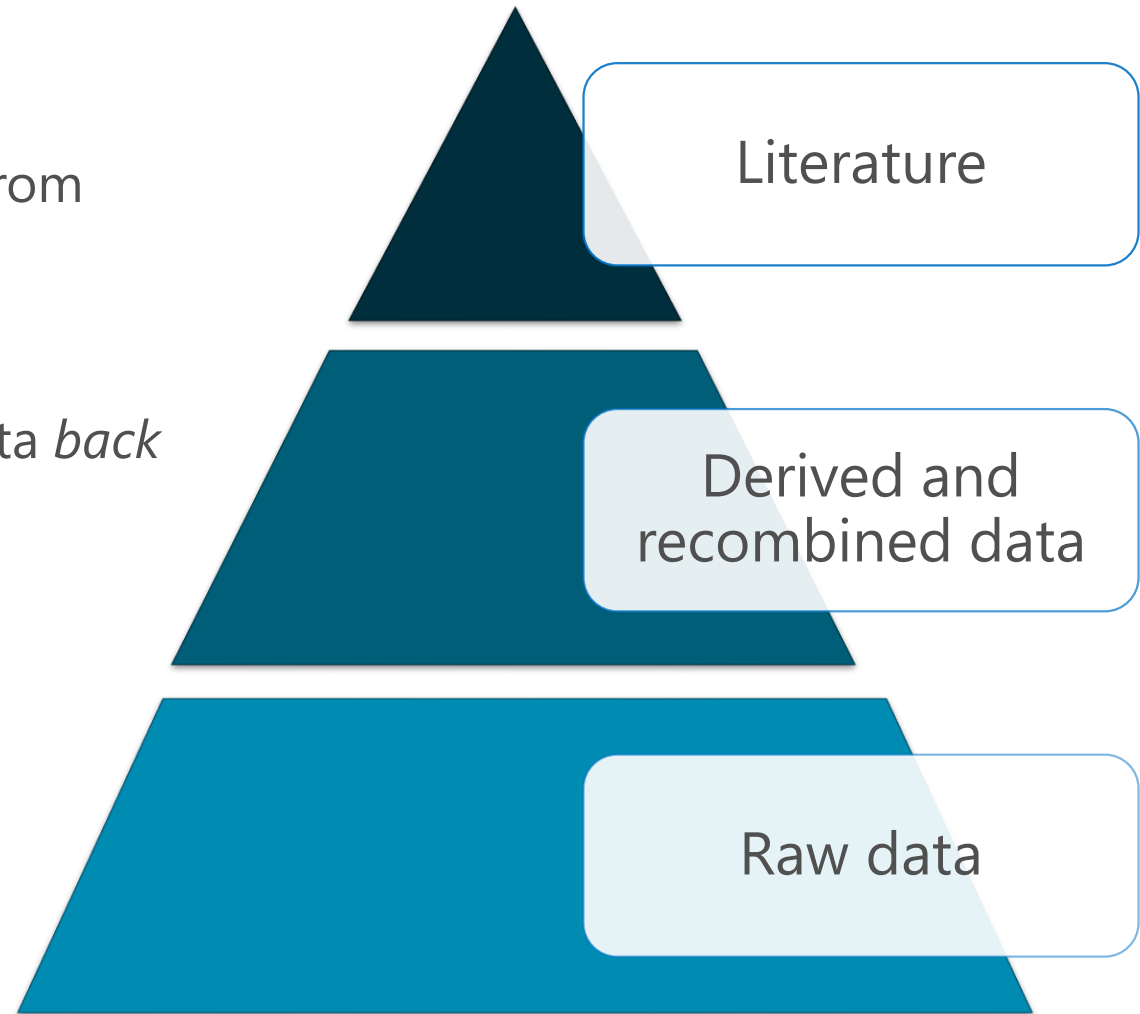
# All Scientific Data Online



*(From Jim Gray's last talk)*

# All Scientific Data Online

- Many disciplines overlap and use data from other sciences.
- Internet can unify all literature and data
- Go from literature *to* computation *to* data *back to* literature.
- Information at your fingertips – for everyone, everywhere
- Increase Scientific Information Velocity
- Increase in Science Productivity



# Objectives

Microsoft Azure Overview

Data Discovery & Reuse

- Azure DataMarket
- Power BI for Excel
- Azure Machine Learning

Azure4Research

Microsoft Azure

# Microsoft Azure

**Microsoft Research**

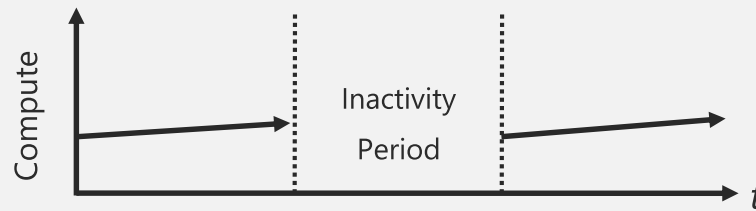
Microsoft Azure for Research





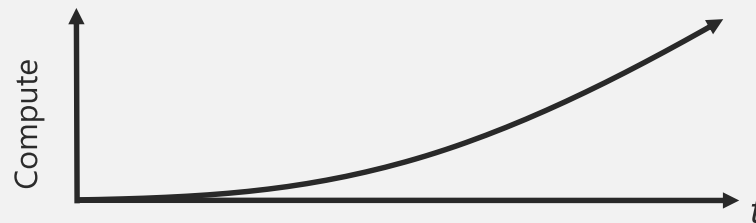
# What is the cloud?

An approach to computing that's about internet scale and connecting to a variety of devices and endpoints



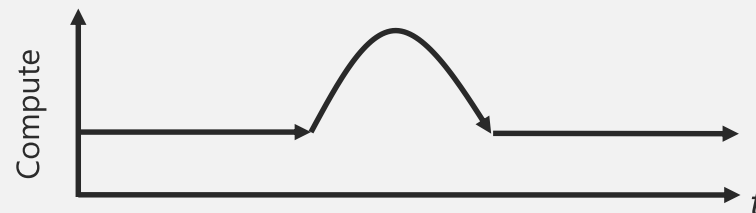
## On and Off

On & off workloads (e.g. batch job)  
Over provisioned capacity is wasted  
Time to market can be cumbersome



## Growing Fast

Successful services needs to grow/scale  
Keeping up w/ growth is big IT challenge  
Cannot provision hardware fast enough



## Bursting

Unexpected/unplanned peak in demand  
Sudden spike impacts performance  
Can't over provision for extreme cases

# 16 regions worldwide in 2014



Azure  
footprint



## Virtual machines (Windows Server & Linux) – IaaS

- You can use remote desktop or SSH and run any workload
- These virtual machines enable you to be admin on the box
- Durable – if you reboot a VM, it is still there with all of your changes and data
- You can deploy VMs and group them so they are their own private network



## Web sites – PaaS (simple)

- Build with .NET, Java, Node.js, PHP, Python. Deploy with TFS/Git/FTP/Mercurial/Dropbox
- Start for free, scale up as your traffic grows
- You don't have to worry or think about VMs, servers, or infrastructure...
- You can simply focus on building and deploying HTTP based applications
- Use any tool and any operating system to build sites: Windows, OS X, and Linux
- Can be registered with a load balancer and scaled out as needed to additional VMs



## Cloud services – PaaS (full)

- Cloud Services is another model we support for building applications
- Build highly scalable apps and services
- You might have a combination of front ends, middle tiers, as well as virtual machines running as part of your solution
- Automated application management



# Azure Summary

Microsoft Azure provides a comprehensive set of services that you can selectively compose to build your cloud apps

## Global Data Center Footprint

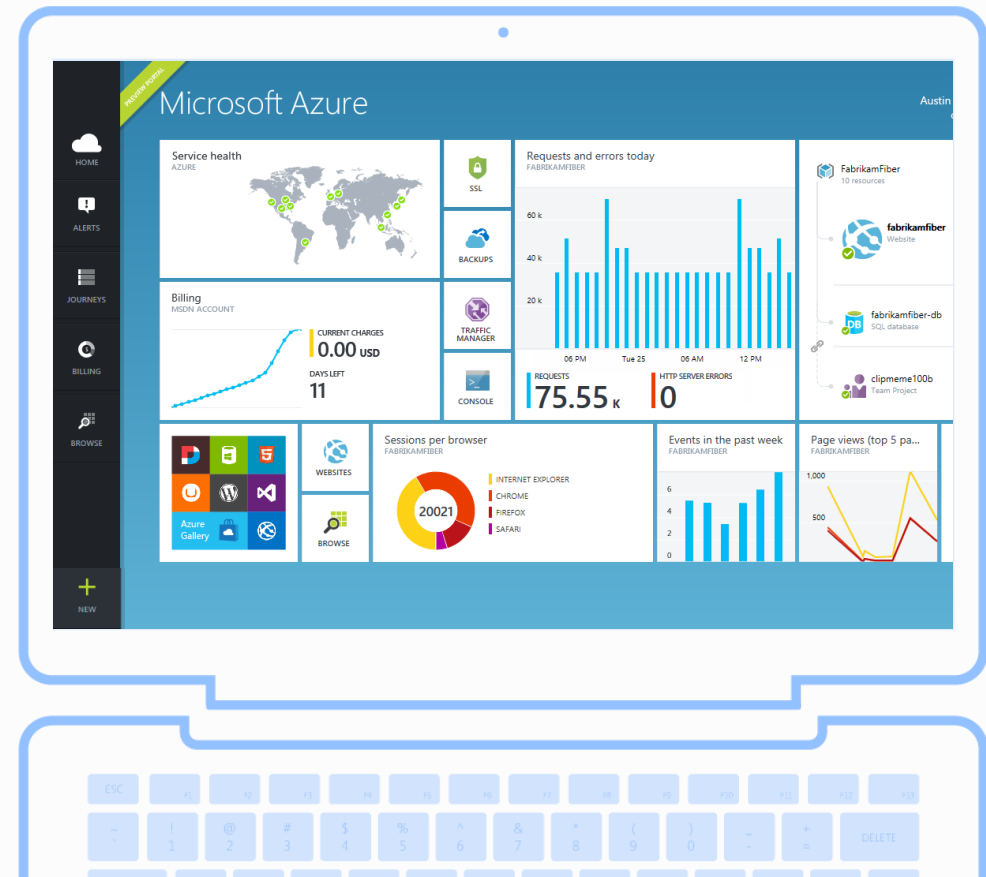
Pay only for what you use.

## Flexible & Open Compute Options

Virtual Machines, Web Sites, & Cloud Services

## Managed Building Block Services

SQL Database, Cache, Service Bus, & more



Microsoft Azure

# Discovery and Reuse of Research Data

**Microsoft Research**



## One-Stop Shop for Premium Data and Applications

Hundreds of Apps, Thousands of Subscriptions, Trillions of Data Points

Discover  
Subscribe  
[Publish](#)

### applications



**Idare-Facturas**  
Idare-facturas es una herramienta para la gestión y almacenamiento de los archivos de facturación...



**CloudAnalyzer**  
CloudAnalyzer is a reporting, analytics and monitoring solution for Azure applications. It's ...



**PushLocker**  
PushLocker is a Software as a service solution to Push Notifications on the Windows Phone and Win...

### data



**Drug Prescribing by GP Practices in England**  
Covering all general practices in England, the data includes figures on the number of prescriptions...



**Millennium Development Goals Database - United Nations Statistics**  
The Millennium Development Goals Database presents official data for more than 60 indicators to m...



**FinViews Fundamentals**  
FinViews specialises in solutions for the financial services industry and investors that enable d...

### announcements

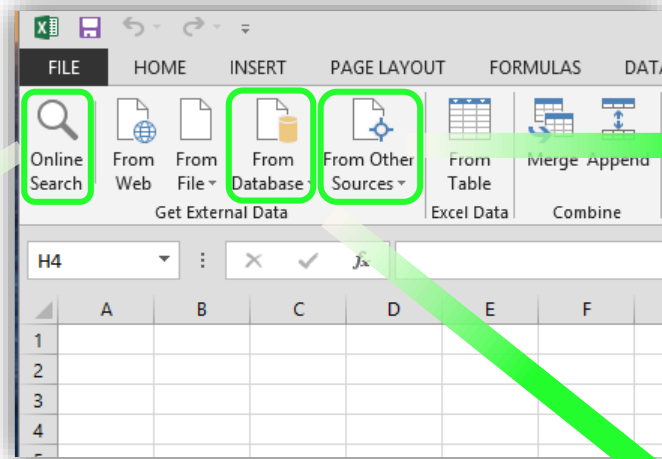
New Bing Speech Recognition Control and Updated Bing OCR and Translator Controls on Windows Azure Marketplace  
[10/21/2013](#)

Changes in Brazil local currency charges  
[10/11/2013](#)

SOC 2 with CSA CCM Attestation Streamlines Security Evaluation for Windows Azure Customers  
[8/22/2013](#)

Windows Azure Marketplace is available in 50 additional countries and features new exciting content

# Excel: Power Query & Power Map



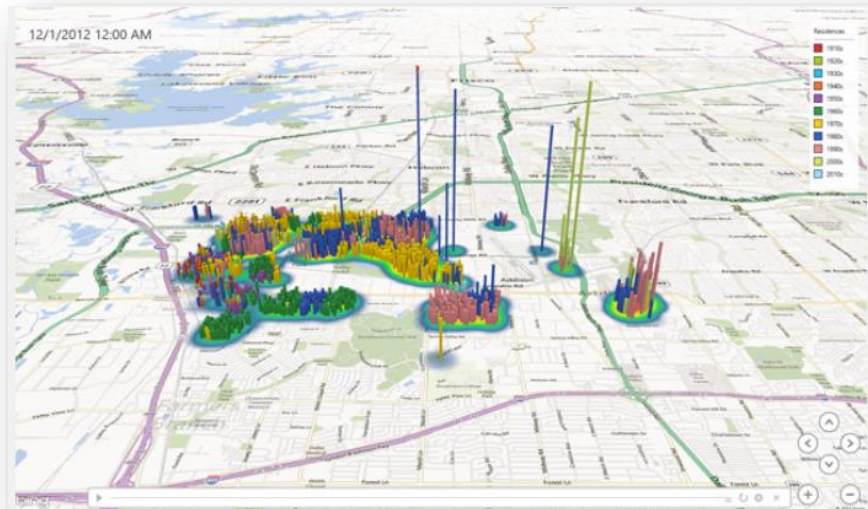
**Online Search**

gdp per capita

2376 results

- Real GDP Per Capita, All States 1...  
From Bureau of Economic Analysis on Febr...  
Real GDP by state is an inflation-adjusted measure of each state's gross...
- Central Intelligence Agency (1993...  
From Wikipedia on January 20, 2014.  
List of countries by GDP (PPP) per capita - Wikipedia, the free encyclopedia
- Methodology - List of countries...  
From Wikipedia on January 20, 2014.  
List of countries by GDP (PPP) per capita - Wikipedia, the free encyclopedia
- World Bank (2005–2012) - Metho...  
From Wikipedia on January 20, 2014.  
List of countries by GDP (PPP) per capita - Wikipedia, the free encyclopedia
- CIA World Factbook (2003–2012)

1 2 3 4 5 Next



- From SharePoint List**  
Import data from a Microsoft SharePoint site.
- From OData Feed**  
Import data from an OData feed.
- From Windows Azure Marketplace**  
Import data from the Microsoft Windows Azure Marketplace.
- From Hadoop File (HDFS)**  
Import data from a Hadoop Distributed File System.
- From Windows Azure HDInsight**  
Import data from Microsoft Windows Azure HD Insight.
- From Active Directory**  
Import data from Microsoft Active Directory.
- From Facebook**  
Import data from Facebook.

- From SQL Server Database**  
Import data from a Microsoft SQL Server database.
- From Windows Azure SQL Database**  
Import data from a Microsoft Windows Azure SQL database.
- From Access Database**  
Import data from a Microsoft Access database.
- From Oracle Database**  
Import data from an Oracle database.
- From IBM DB2 Database**  
Import data from a DB2 database.
- From MySQL Database**  
Import data from a MySQL database.
- From PostgreSQL Database**  
Import data from a PostgreSQL database.
- From Teradata Database**  
Import data from a Teradata database.

Microsoft Azure



# Machine Learning

Computing systems that improve with experience

Microsoft Research



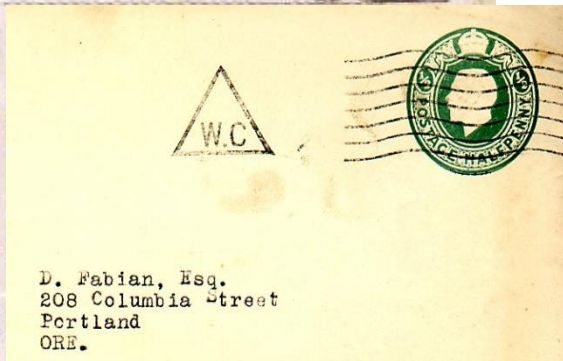
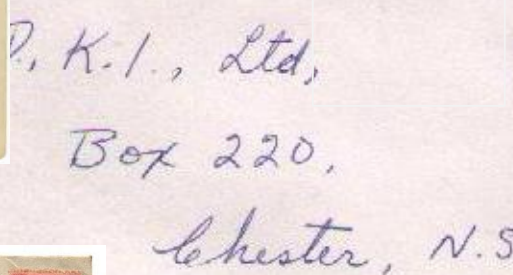
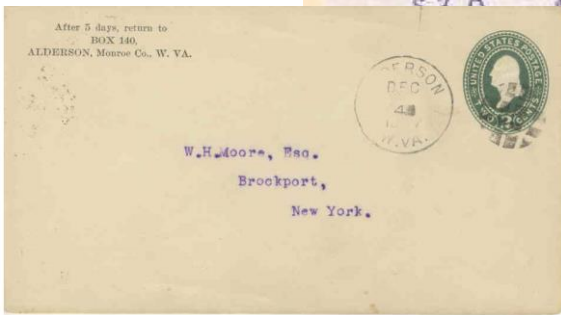


I believe over the next decade computing will become even more ubiquitous and intelligence will become ambient...

This will be made possible by an ever-growing network of connected devices, incredible computing capacity from the cloud, insights from big data, and intelligence from machine learning.

“If you invent a breakthrough in Artificial Intelligence, so **machines can learn**, that is worth 10 Microsofts”





1 1 5 4 3  
7 5 3 5 3  
5 5 9 0 6  
3 5 2 0 0

1	1	5	4	3
7	5	3	5	3
5	5	9	0	6
3	5	2	0	0

Training examples    Training labels



ML System



Accurate digit classifier

2



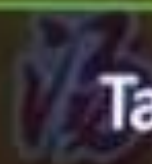
清炒娃娃菜  
Stir fried baby vegetables

清炒油麦菜  
Stir fried lettuce

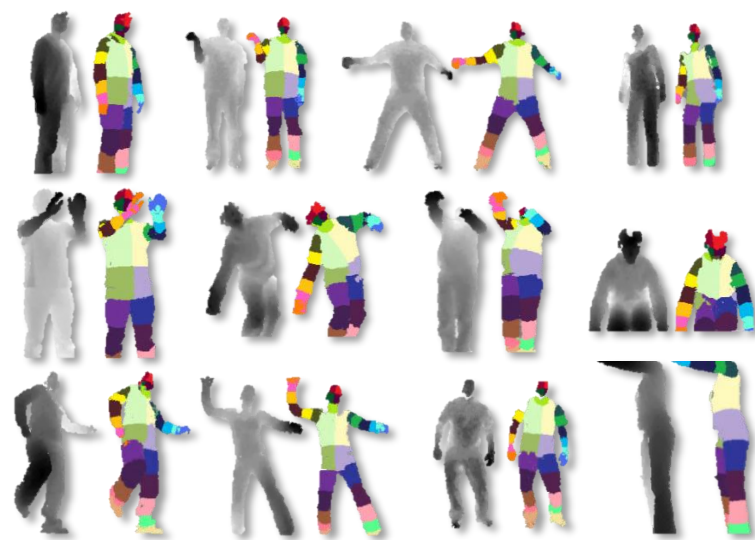
红烧茄子  
Braised Eggplant

老厨白菜  
Old kitchen Chinese cabbage

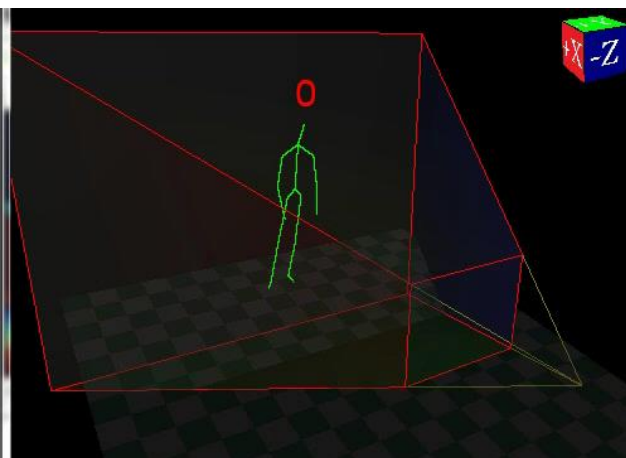
西红柿炒鸡蛋  
Scrambled egg with tomato



Tap to pause.



training data (expensive)



ML system



getting a morgage in seattle

8,140,000 RESULTS Any time

Ads related to getting a morgage in seattle

**15-Year Mortgage Rates | QuickenLoans.com**  
[www.QuickenLoans.com/Rates](http://www.QuickenLoans.com/Rates)  
 Lock Your Rate. 3.500% (3.92% APR) With America's #1 Online Lender.

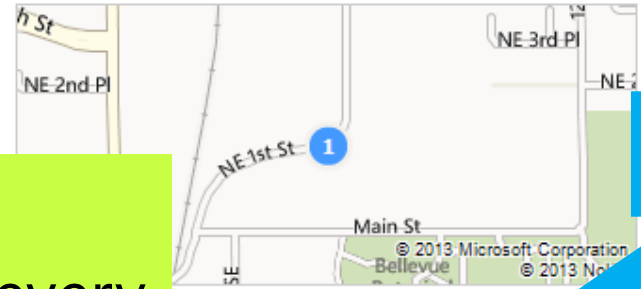
**LendingTree®Official Site - Amazingly Low Mortgage Rates.**  
[LendingTree.com/Get-a-Mortgage](http://LendingTree.com/Get-a-Mortgage)  
 APR from

**TILA**  
 seattle  
 Meet o

**Pre Q**  
[www.w](http://www.w)  
 Estima

Machine learning enables nearly every value proposition of web search.

Seattle's Best Mortgage Inc



St Ste B306 · Bellevue  
 0 · Directions  
[seattlesbm.com](http://seattlesbm.com)

- RELATED SEARCHES
- Getting a First Mortgage
  - Getting a Mortgage Self-Employed
  - Getting a Mortgage Loan Approved
  - Getting a Mortgage On Land
  - Getting a Mortgage in 2013
  - How to Get a Mortgage License
  - How to Get a Mortgage After Bankruptcy
  - Mortgage Calculator

Ads related to getting a morgage in seattle

**Seattle Mortgage Rates**  
[Seattle.BankRateLocator.com](http://Seattle.BankRateLocator.com)  
 Rates Dropped to 3.18%. No Closing Costs! Get Fr  
 Quotes in 30 Seconds

12 Year Mortgage Rates

What language?

Which ads to show, and in what order?

Misspelled?

Which links are most likely to get clicked?

What is the probability of a click on each ad?

What is the intent?

Are any of these pages malicious?

What pages should we index?

What ad pricing will optimize revenue?

# Machine Learning Scenarios

*Machine learning and predictive models are core new capabilities that will touch everything in the new world of intelligent applications and ambient intelligence...*

The screenshot shows the Microsoft Azure Machine Learning Center website. The top navigation bar is dark blue and contains the Microsoft Azure logo, contact information (SALES 1-800-867-1389), account links (MY ACCOUNT, PORTAL), a search bar, and a prominent 'FREE TRIAL' button with a right-pointing arrow. Below the navigation bar, the main content area has a light blue background. The 'Machine Learning Center' title is displayed in large white text. A descriptive paragraph explains that it is a fully-managed cloud service for embedding predictive analytics. A large blue button with white text and a right-pointing arrow says 'Get started with Machine Learning'. To the right, a 'Featured' section lists three items: 'Develop a predictive solution', 'Predictive Modeling with Azure ML', and 'Frequently asked questions', each with a corresponding icon. On the left side of the main content area, there are sections for 'Machine Learning Studio', 'Quick links', and a list of links: 'Management Portal', 'Blog', and 'Forums', each with a right-pointing arrow.

Microsoft Azure

SALES 1-800-867-1389

MY ACCOUNT

PORTAL

Search

Features Pricing **Documentation** Downloads Gallery Community Support

**FREE TRIAL** →

## Machine Learning Center

A fully-managed cloud service that enables data scientists and developers to efficiently embed predictive analytics into their applications, helping organizations use massive data sets and bring all the benefits of the cloud to machine learning.

[Get started with Machine Learning](#) →

Machine Learning Studio ▶

### Quick links

Management Portal ▶

Blog ▶

Forums ▶

### Featured

- Develop a predictive solution
- Predictive Modeling with Azure ML
- Frequently asked questions

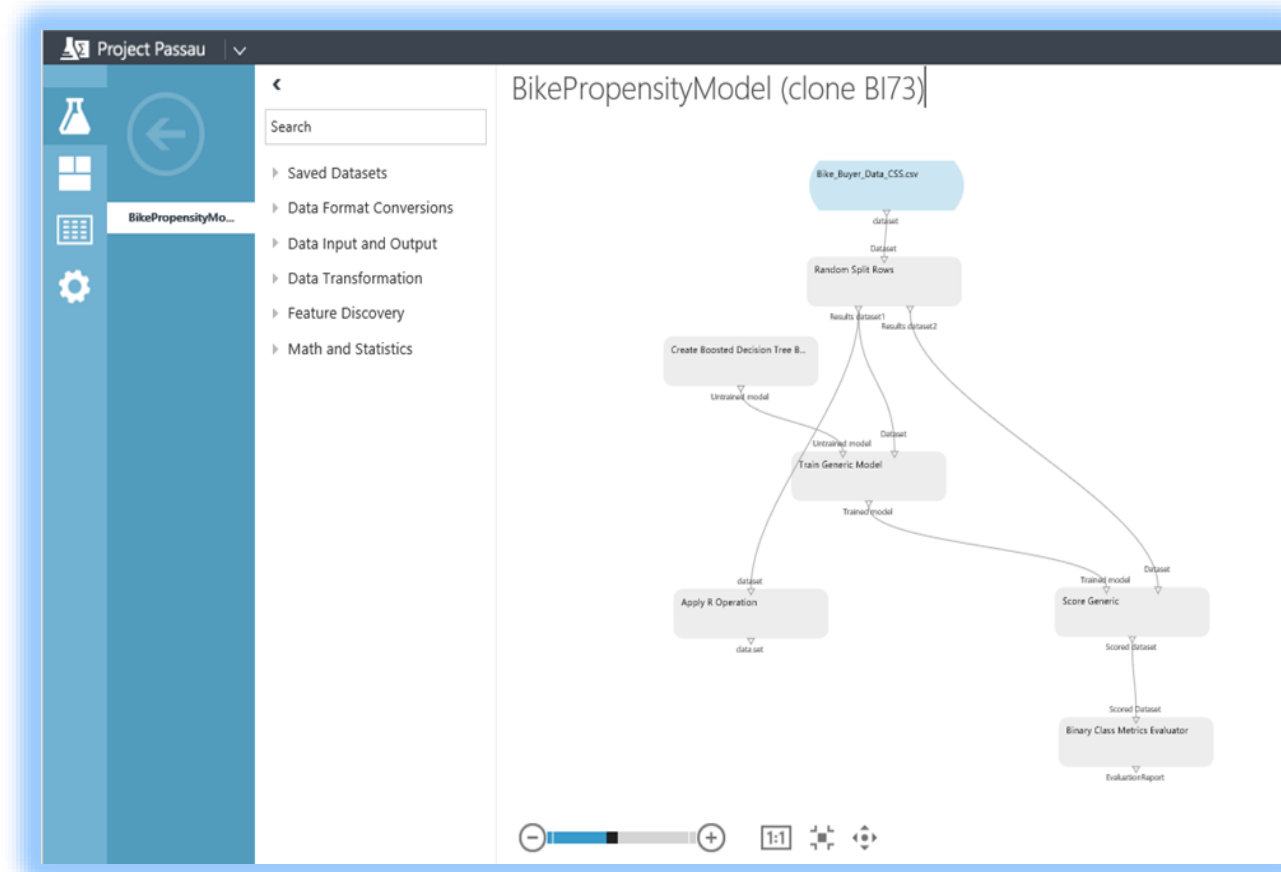
# Microsoft Azure Machine Learning

## Guiding Principles

Reduce complexity to broaden participation



- Accessible through a web browser, no software to install;
- Collaborative, work with anyone, anywhere via Azure workspace
- Visual composition with end2end support for data science workflow;
- Extensible, support for R OSS.

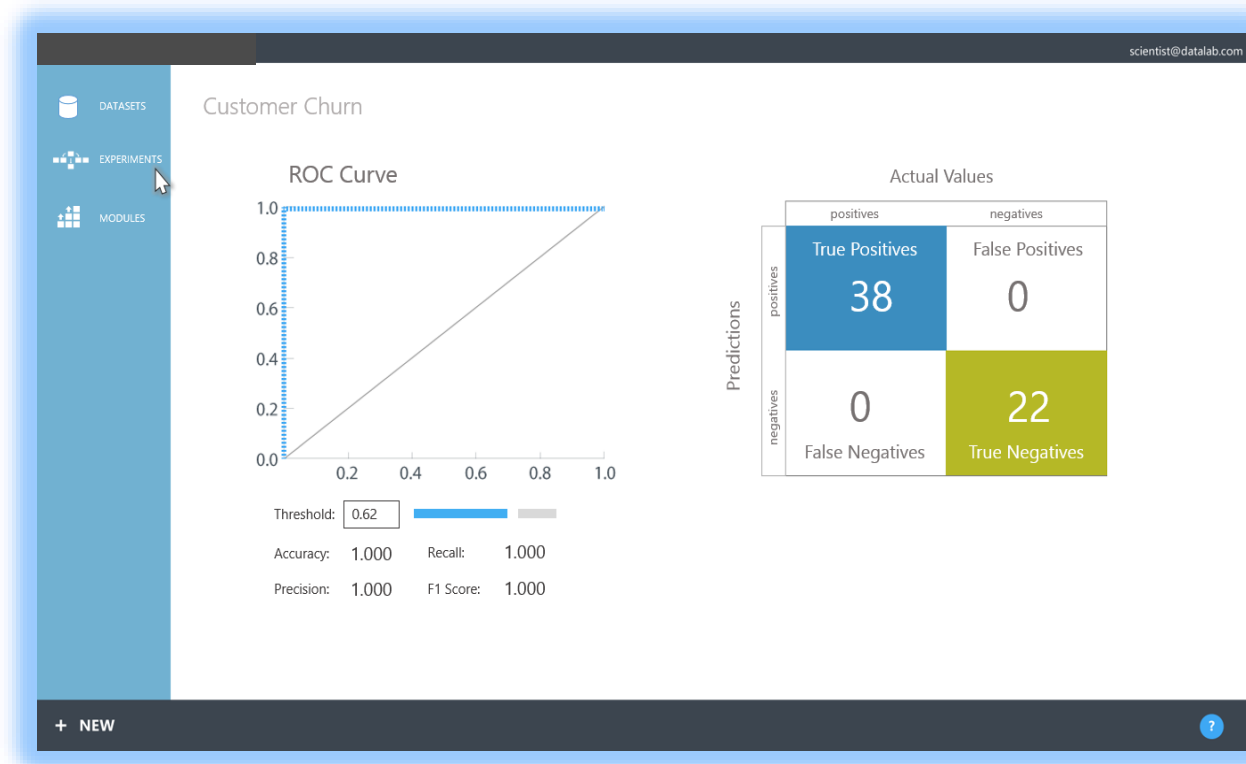


# Microsoft Azure Machine Learning

## Guiding Principles


Rapid experimentation to create a better model

- Rapidly try a range of features, ML algorithms and modeling strategies;
- An immutable library of models, share, search, discover & reuse;
- Quickly deploy model as Azure web service to our ML API service.



Microsoft Azure

Bringing cloud  
computing to  
researchers

Learn about the new  
Azure for Research program 



# Azure4Research

- Training and Webinar series
- Technical papers & curriculum
- Research community engagements
- Azure Research Awards (>400 to date)



[\*Microsoft Azure for Research Group\*](#)



[\*@azure4research\*](#) & [\*#azureresearch\*](#)

Microsoft Research

[www.azure4research.com](http://www.azure4research.com)

# Research Engagements

## PNNL "Global Azure Bootcamp for Diabetes Research"

A crowd sourced experiment with over 1000 people in 57 countries running VMs and PNNL GlyQ-IQ software

## NCAR "Big Data challenges in climate and weather research"

Build a gateway to climate data generated by NCAR simulations.

## Univ of Pittsburgh "Kepler's Conjecture"

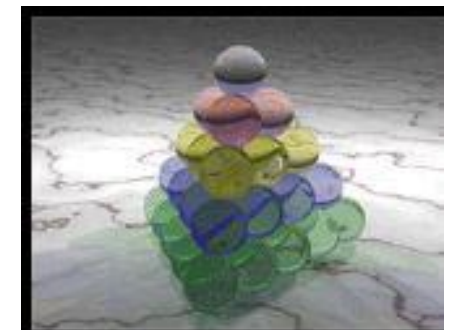
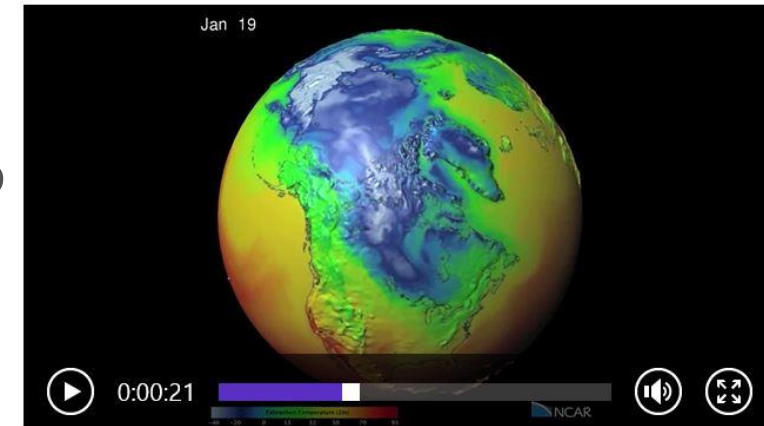
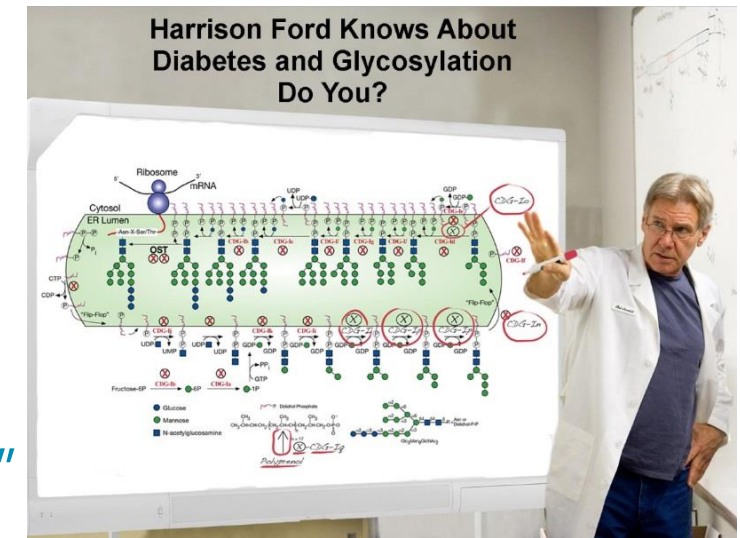
Thomas Hales' proof of the sphere packing conjecture – 30X speedup

## UCLA "Systems Genetics Adipose Tissue Transcriptome for Metabolic Syndrome Traits"

A major GWAS study using Fast-LMM tools to look at cardiovascular disease. Running on 700 cores

## BSC "HDInsight/Hadoop performance on Azure"

Performance tuning using sequence alignment tool (SNAP) Ravi Pandya– SQL Azure values the work





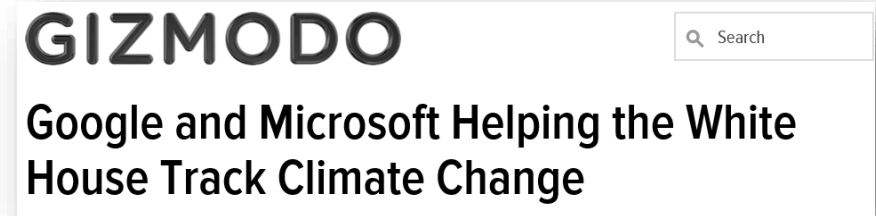
# Special RFPs

## Series of special calls for proposals

- Science Virtual Machines  
8 projects to build VM Images for VMDepot
- Brazilian Initiative – targeted for new Azure Data Center in Brazil
- The Matlab cluster – joint project with Mathworks
- Berkeley CS – support UCB CS researchers

## The White House Climate Data Initiative

- Partnered w/OSTP to provide 40 Azure for Research awards
- Combined with access to the FetchClimate tools
- Coordinating with OSTP and USDA on 2<sup>nd</sup> round of CDI on Food Security

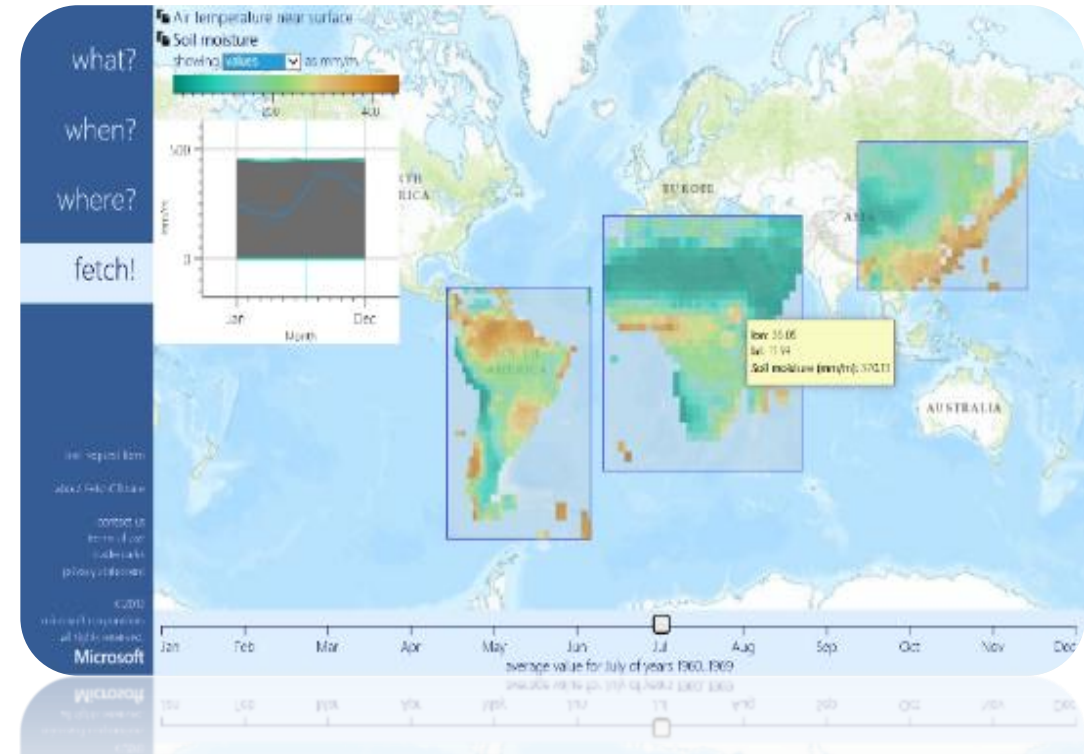


## Google, Intel, Microsoft help build climate change tools

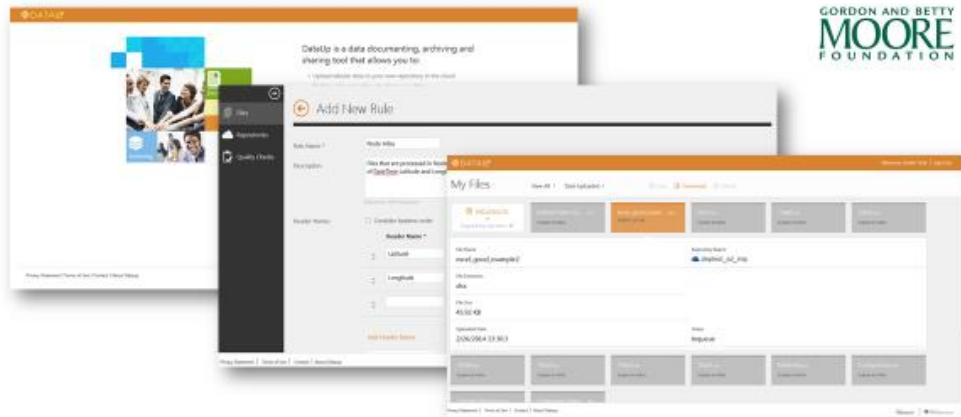
Wendy Koch, USA TODAY 10:52 a.m. EDT March 19, 2014

# FetchClimate

- Intelligent environmental information service
- Automatically:
  - Selects best data source to answer the query
  - Regrids results
  - Calculates uncertainty
- Azure4Research grants for FetchClimate



<http://fetchclimate2.cloudapp.net/>



## Dataverse

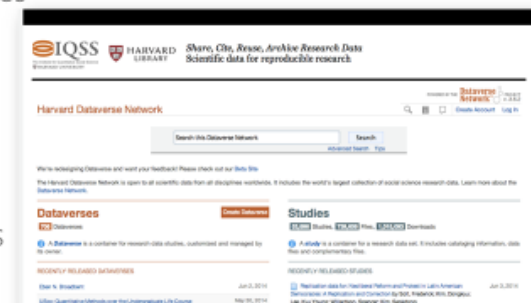


### Dataverse Software

- Framework to publish/cite/preserve research data: <http://thedata.org>
- All data types, multiple disciplines
- Open-source via GitHub

### Dataverse Repository

- Hosted instance **free and open**: <http://thedata.harvard.edu>
- >53,000 datasets, >735,000 files
- Federates with > 10 Dataverse installations around the world



## CKAN Data Repository Software



### Publish & find datasets

Publish datasets via import or through a web interface. Search by keyword or filter by tags. See [dataset information](#) at a glance. Full change [history](#) lets you easily undo changes or view old versions.



### Store & manage data

Store the raw data and metadata. Visualise structured data with interactive tables, graphs and maps. Get statistics and usage metrics for your datasets. Search [geospatial](#) data on a map by area.



### Engage with users & others

Federate networks with other CKAN nodes. Theme with CSS or integrate with a CMS. Build a [community](#) with extensions that allow users to comment on and follow datasets.



### Customise & extend

Use the API's rich programming interface, and benefit from over 60 [extensions](#) including link checking, comments, and analytics. CKAN's [Open Source](#) licence allows you to download and run it for free.

## Research Data Registry and Discovery Service

- UK pilot project
  - Stand up working system
  - Explore metadata harvesting
  - Test metadata harvesting
  - Collect feedback
- UK Data Archive, NERC Data Catalogue, nine universities
- Based on ANDS platform and modified

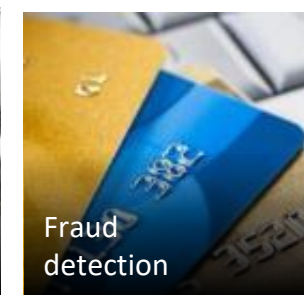
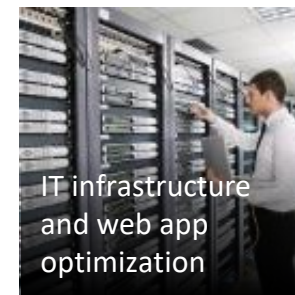
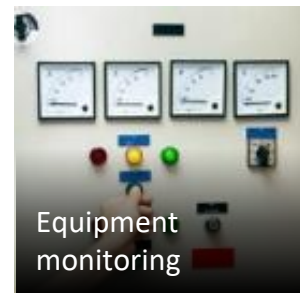
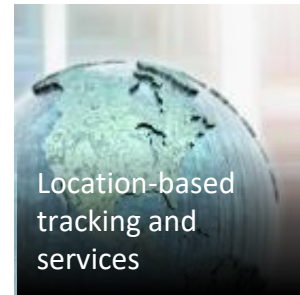
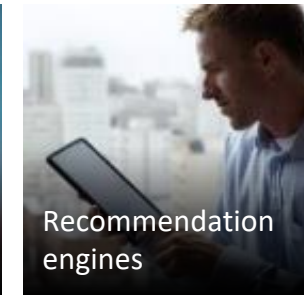


# Azure Machine Learning for Research

*Machine learning with the simplicity and power of the cloud*

Now available for the academic community


- Data science instructional awards
  - Individual account on Azure ML for each student;
  - 500 GB of cloud data storage for each student;
- Shared workspaces for research collaborations
  - 10 TB of cloud data storage, to enable a group of researchers interested in hosting a data collection in Microsoft Azure ML to discover and share predictive models.
- Next deadline is Nov. 15<sup>th</sup>, every two months after that



Apply online at <http://research.microsoft.com/Azure-ML>

# Thank you!

 @alexwade

 @azure4research



© 2014 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries. The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

[www.azure4research.com](http://www.azure4research.com)