Congress on Harnessing Big Data for the Environment: *Private Sector Capabilities*

Kristin M. Tolle, Ph.D.

Director, Program Management Advanced Analytics Training and Ecosystem Development

December 6th, 2016

Agenda

- General Azure Overview
- Case Study: Predictive Analytics for Flooding
- Case Study: Energy Demand Forecasting
- Case Study: Deep Learning for Massive Datasets – Galaxy Classification

The Cloud is Ubiquitous

34

Azure Regions

The Cloud is Secure

More certifications than any other cloud provider

Industry leader for customer advocacy and privacy protection

Unique data residency guarantees

Choice + Flexibility

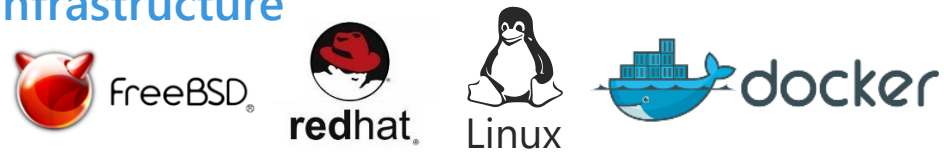
Management

Applications

App Frameworks

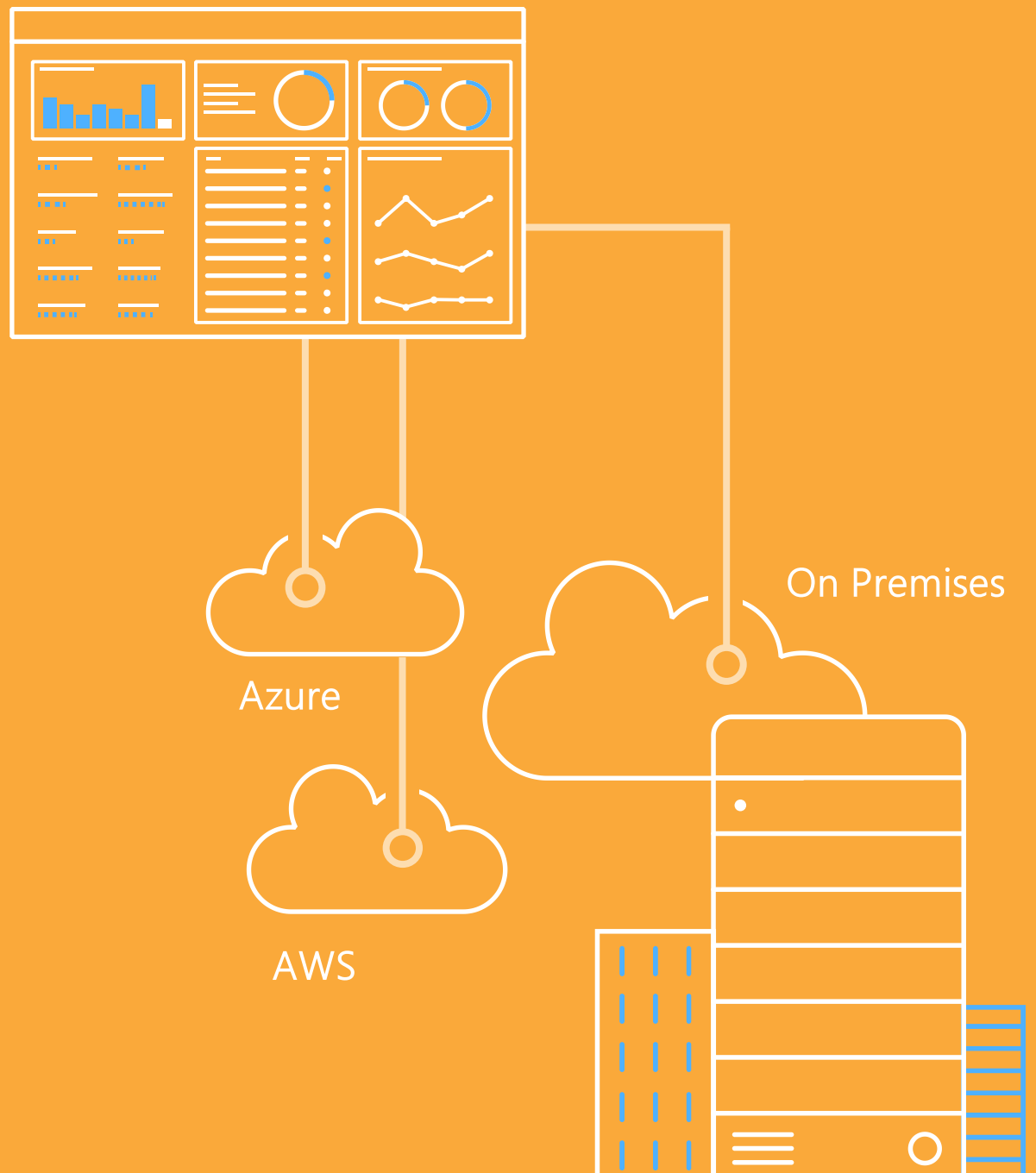
Databases & Middleware

Infrastructure



Hybrid Management and Security

- Log analytics
- Automation
- Backup
- DR and data protection
- Security



Cortana Intelligence Suite

Transform data into intelligent action

Data Sources

Apps

Sensors and devices

Information Management

- Data Factory
- Data Catalog
- Event Hub

Big Data Stores

- Data Lake Store Blob Store
- SQL Data Warehouse

Machine Learning and Analytics

- Machine Learning
- Data Lake Analytics
- HDInsight (Hadoop and Spark)
- Stream Analytics

Intelligence

- Cognitive Services
- Bot Framework
- Cortana

Business Scenarios

- Recommendations, customer churn, forecasting, etc.

Dashboards & Visualizations

- Power BI

People

Apps

- Web
- Mobile
- Bots

Automated Systems

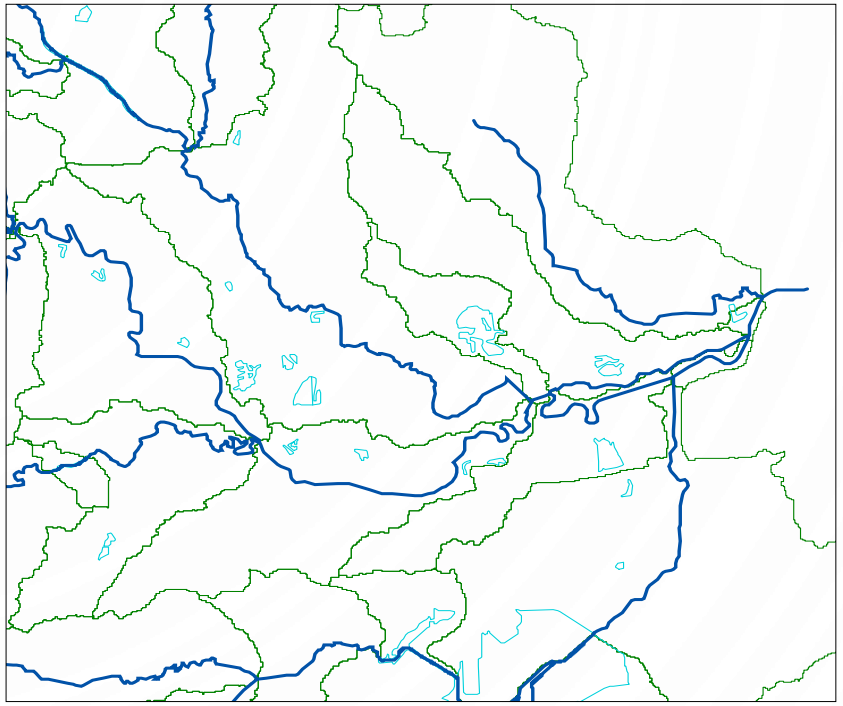
Data → Intelligence → Action

Flood Forecasting

National Flood Interoperability Experiment:

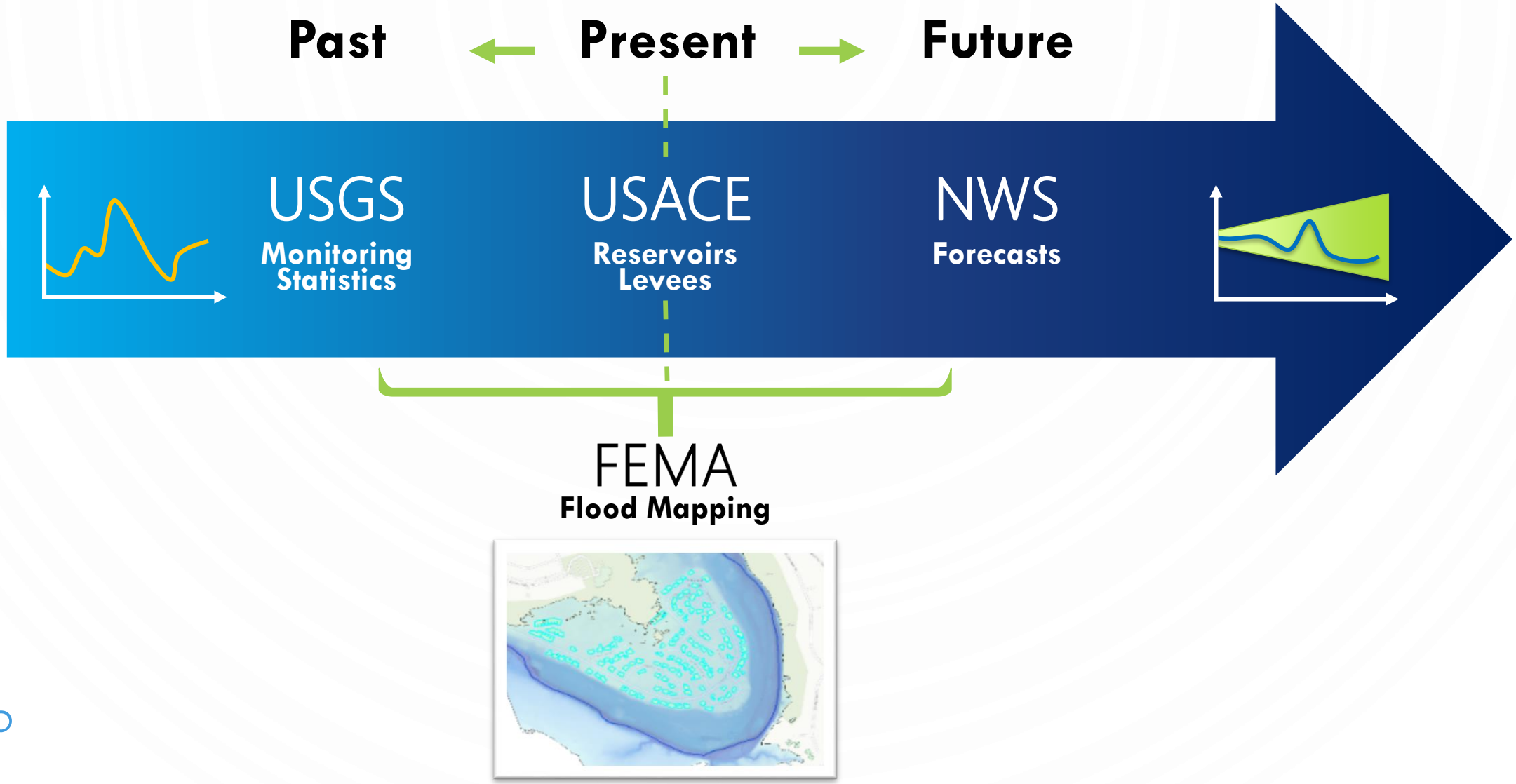
Close the gap between national flood forecasting and local emergency response

…improve information flow



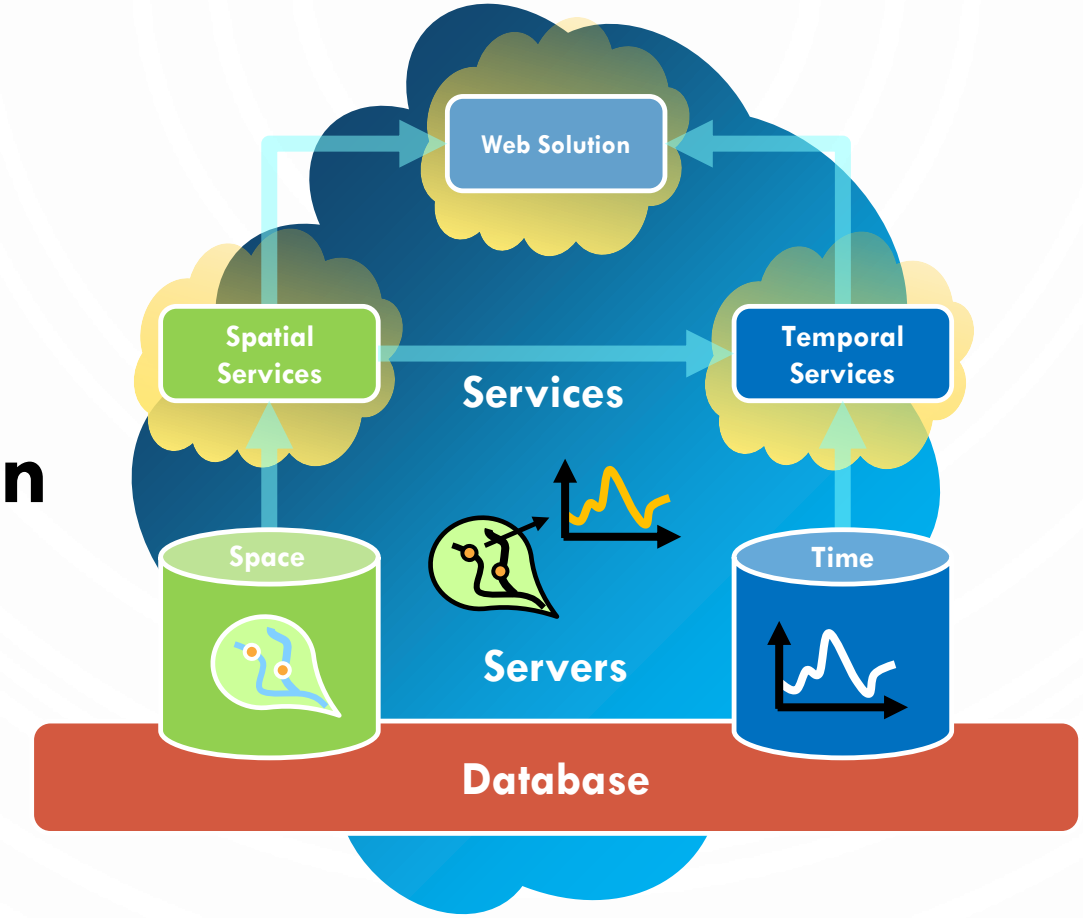
…national forecasting at stream reach scale

Authoritative Data Sources



Challenge: Integrating Space and Time

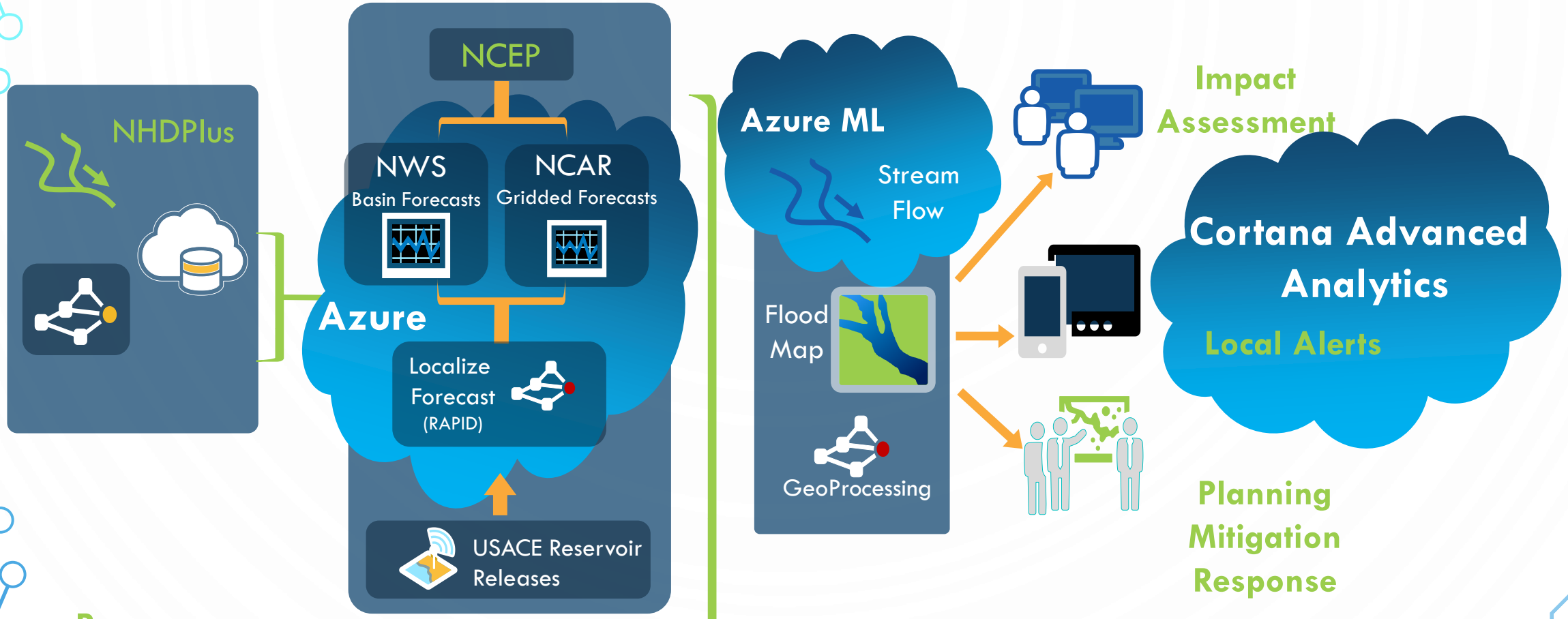
Spatial Aggregation



Temporal Aggregation

Bringing the pieces together…

Advanced flood modeling on consistent spatial domain

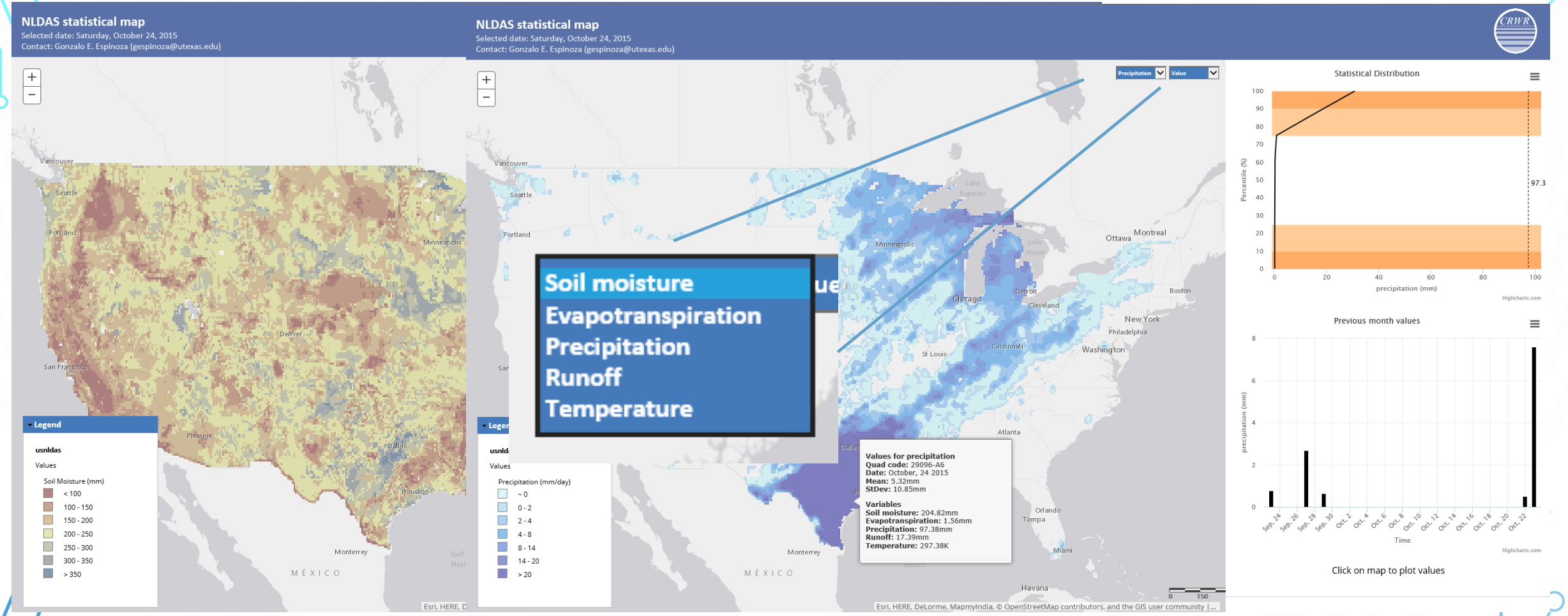


Process Geospatial Data

Model Stream Flows

Create Flood Data

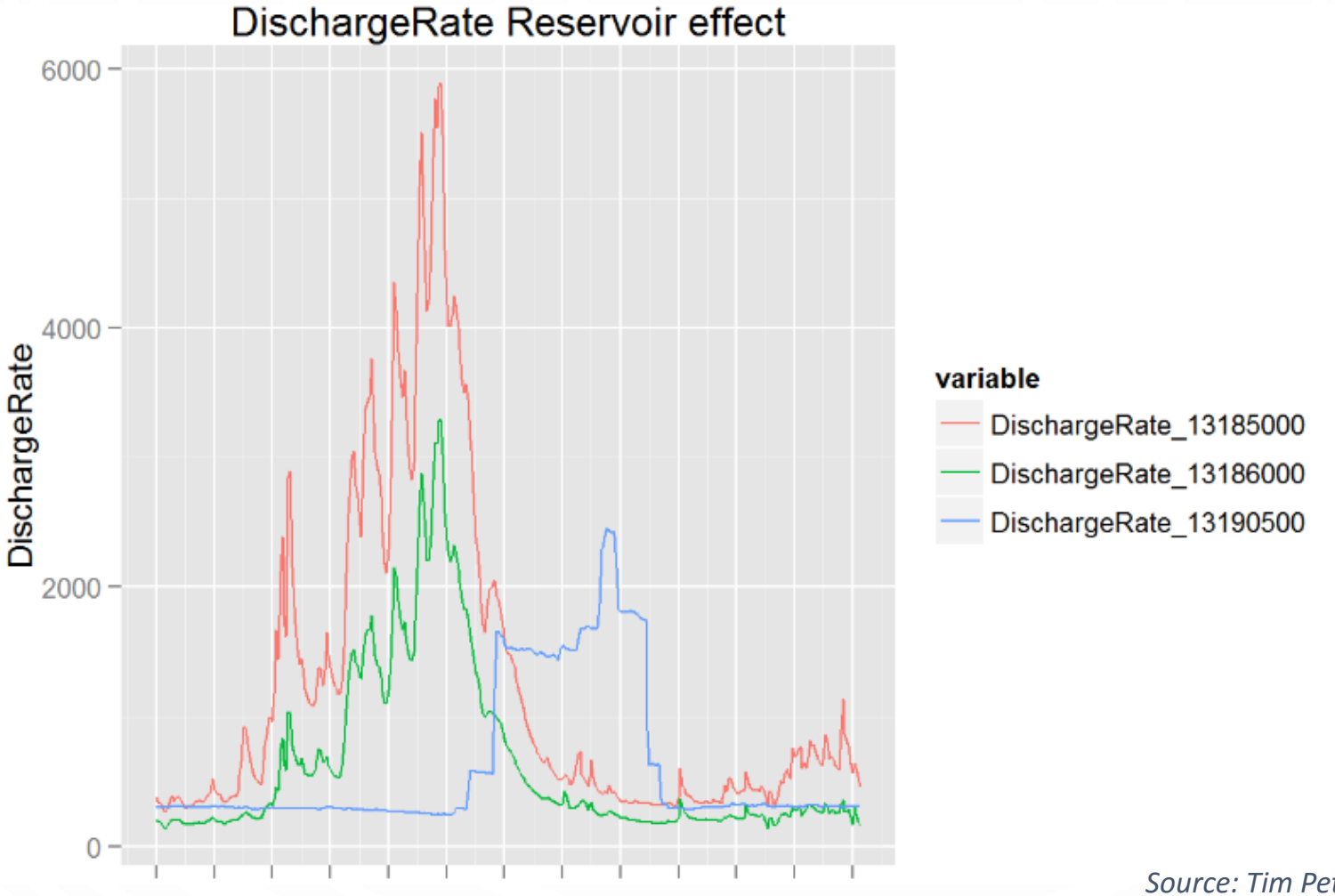
Calculating Soil Moisture Models



http://statsnldas.azurewebsites.net

Source: Gonzalo Espinoza Davalos

Machine Learning to Estimate Missing Values



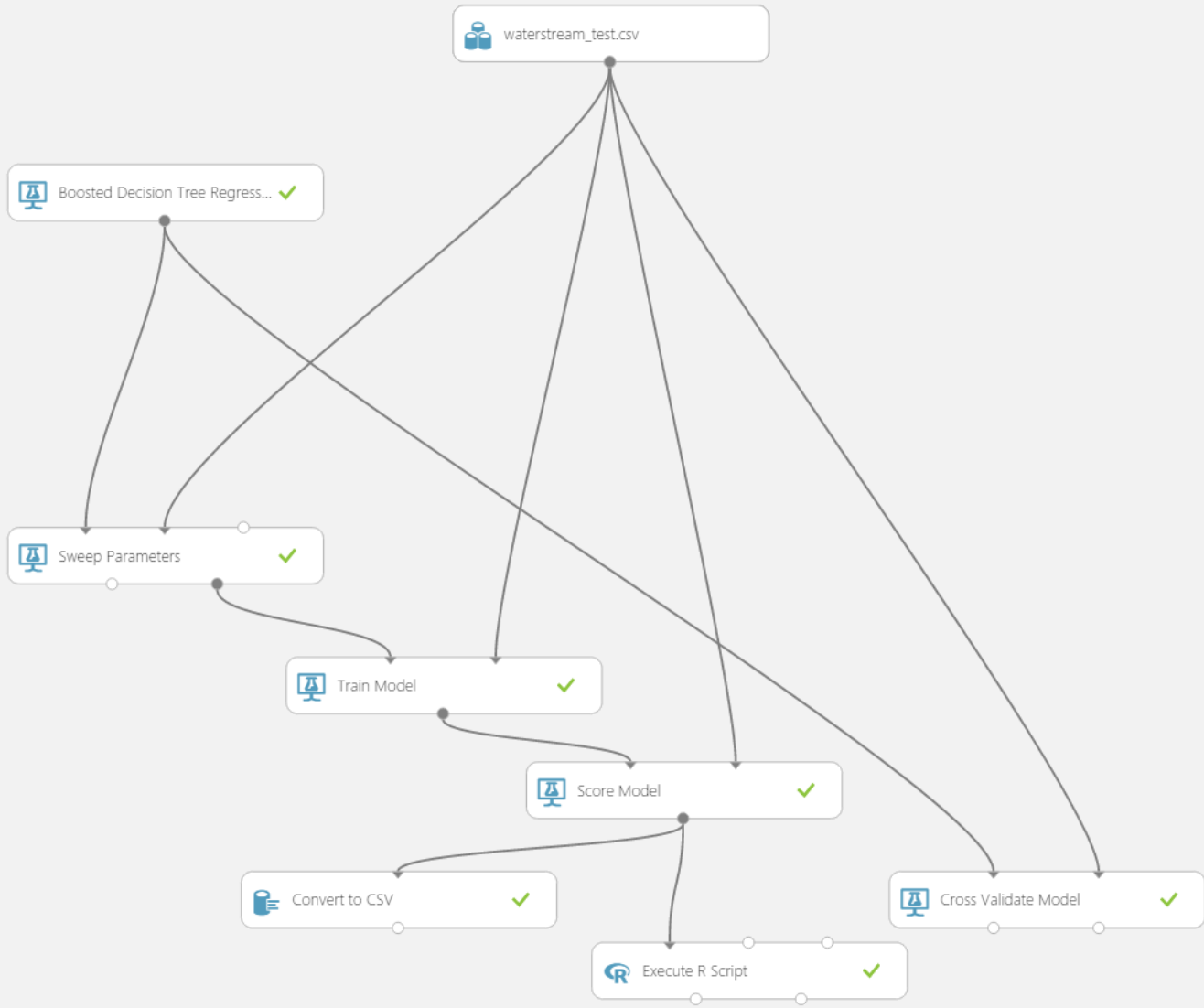
Source: Tim Petty and Prashant Dhingra

Microsoft Azure Machine Learning | Home Studio Gallery PREVIEW

Water_Stream_Prediction

Search experiment items

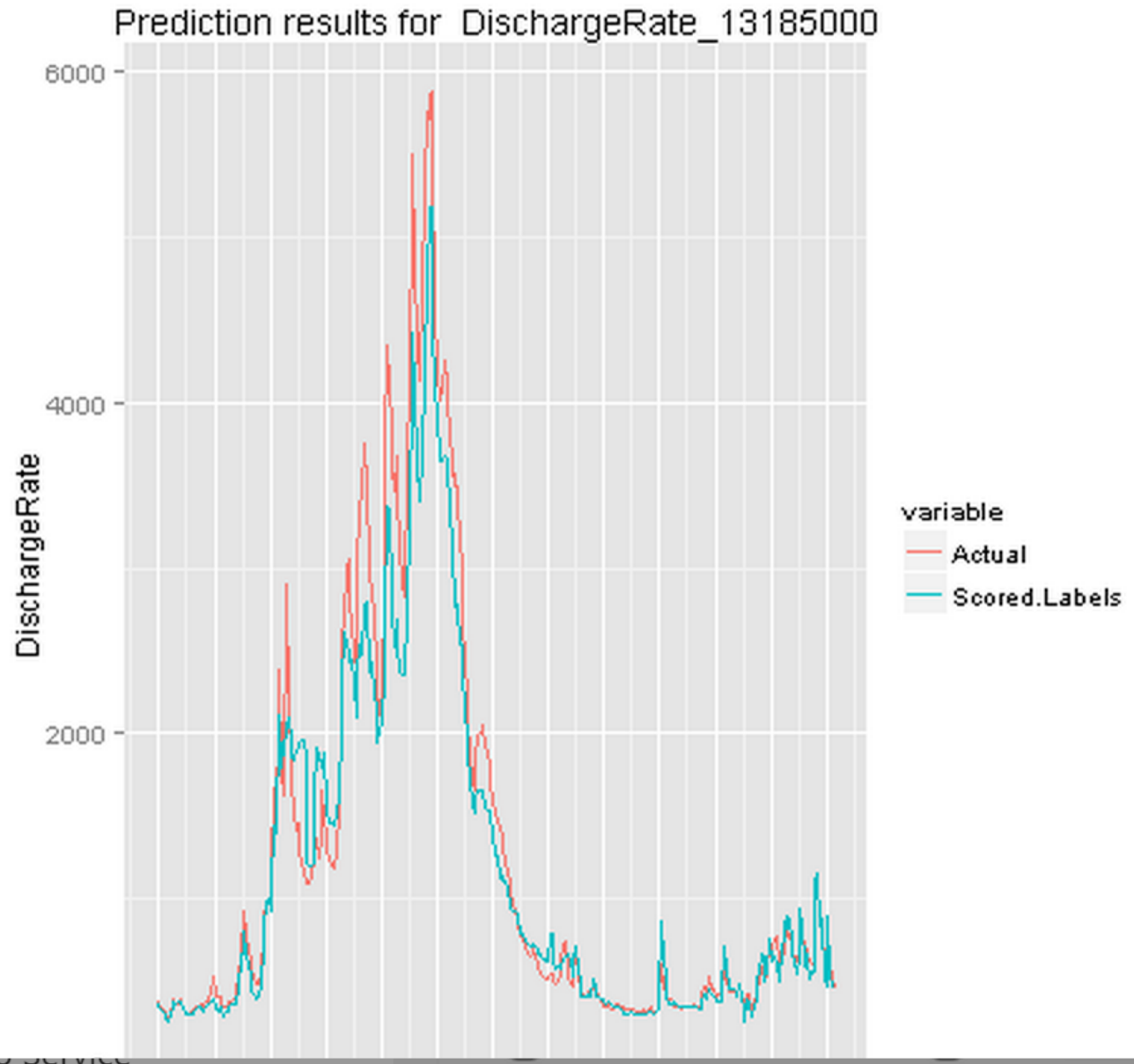
- Saved Datasets
- Data Format Conversions
- Data Input and Output
- Data Transformation
- Feature Selection
- Machine Learning
- OpenCV Library Modules
- Python Language Modules
- R Language Modules
- Statistical Functions
- Text Analytics
- Web Service



Water_Stream_Prediction

Search

Water_Stream_Prediction › Execute R Script › R Device

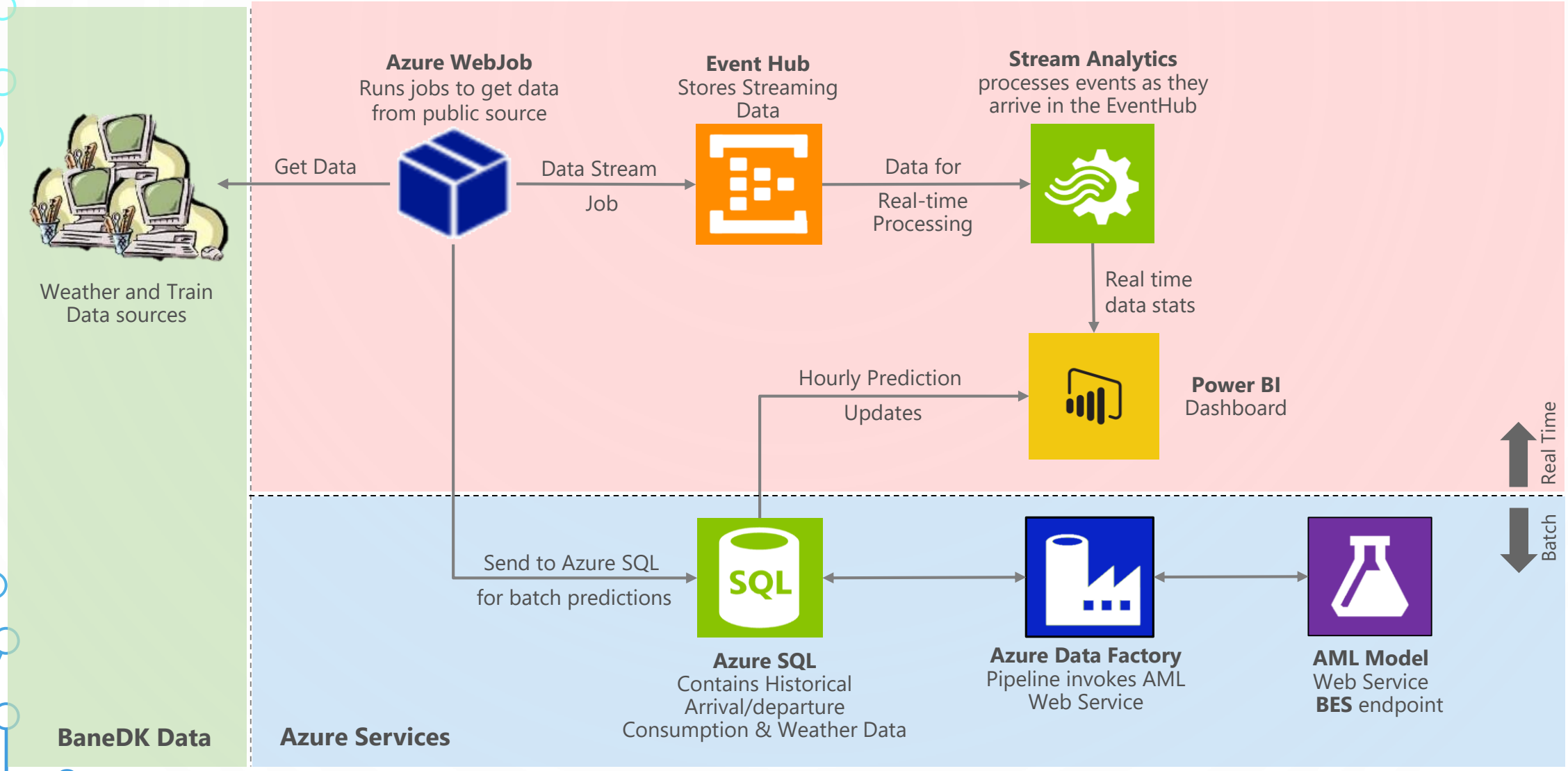


Next step… Scale to the US

Calculate ~3 million river reaches

Energy Forecasting

ARCHITECTURE



Azure Stream Analytics

Microsoft Azure

Subscriptions yijichen@microsoft.com

EnergyForecast_5...
EnergyForecastV2_...
pmaircraftstreama...
pmaircraftstreama...
pmaircraftstreamd...
pmrawsensors
rbaEnergyForecast...

energyforecast_5minsdata_asa

DASHBOARD MONITOR INPUTS QUERY OUTPUTS SCALE CONFIGURE

Need help with your query? Check out some of the most common Stream Analytics query patterns here.

query

Query can't be edited while a job is running.

Power BI is designed for cases where Azure Stream Analytics does a significant data load reduction by aggregating on a hopping or tumbling window.

```
SELECT a.TimeStamp as Time, a.TimeZone, a.Name as Region, a.PTID, b.Latitude, b.Longitude, a.Load as Demand
INTO outputPBI
FROM
(SELECT TimeStamp, TimeZone, Name, PTID, Load FROM InputEventHub) a
join
(SELECT PTID, Name, Latitude, Longitude FROM InputBlobRefData) b
on a.PTID=b.PTID
;
```

Missing some language constructs? Let us know! (Powered by UserVoice - Privacy Policy)

Test Rerun

NEW START STOP

Azure Machine Learning

Microsoft Azure Machine Learning | Home | Studio | Gallery PREVIEW | EnergyForecastAML

Search experiment items

- Saved Datasets
- Data Format Conversions
- Data Input and Output
- Data Transformation
- Feature Selection
- Machine Learning
- OpenCV Library Modules
- Python Language Modules
- R Language Modules
- Statistical Functions
- Text Analytics
- Web Service
- Deprecated

EnergyDemandForecast-24hr-withT — Finished running

Reader
Get demand and temperature data from SQL

Execute R Script
Data cleaning & feature engineering

Execute R Script
Data for training

Execute R Script
Data for forecasting

Boosted Decision Tree Regr...

Train Model

Score Model
Make forecast

Execute R Script
Select the column we need

Writer
Write forecast data to SQL

Properties

Reader

Data source: Azure SQL Database

Database server name: ghdb7wg5iw.database.windows.net

Database name: EnergyForecastSQLDatabase

Server user account name: energydemouser

Server user account password: ••••••••••••

Accept any server certificate (insecure)

Database query

```
1 select b.localTime as TimeStamp, b.PTID, a.HourAvgL
2 (select * from DemandHistoryHourly
3 where convert(varchar(10),TimeStamp,110)>=convert(v
4 and PTID=61761) a
5 right join
```

Quick Help

Load data from sources such as the Web, Azure SQL database, Azure table, Hive table, or Windows Azure BLOB storage
(more help...)

NEW | RUN HISTORY | SAVE | DISCARD CHANGES | RUN | SET UP WEB SERVICE | PUBLISH TO GALLERY

Azure Machine Learning

EnergyDemandForecast-24hr-withTemp - Demo › Reader › Results dataset

rows 8090 columns 4

view as

TimeStamp	PTID	HourAvgLoad	temperature
2012-09-12T00:00:00	61761	5090.6	59.61
2012-09-12T01:00:00	61761	4797.5	60.52
2012-09-12T02:00:00	61761	4624.4	60.14
2012-09-12T03:00:00	61761	4544.7	59.59
2012-09-12T04:00:00	61761	4546.1	58.47
2012-09-12T05:00:00	61761	4786.3	57.85
2012-09-12T06:00:00	61761	5383.9	57.79
2012-09-12T07:00:00	61761	6037.4	59.12
2012-09-12T08:00:00	61761	6573.9	63.09
2012-09-12T09:00:00	61761	6978.5	67.45
2012-09-12T10:00:00	61761	7214.9	69.9
2012-09-12T11:00:00	61761	7395.9	71.72
2012-09-12T12:00:00	61761	7524.2	72.1

Statistics

Mean	5604.3624
Median	5651.45
Min	2859.6001
Max	10610.9004
Standard Deviation	946.7895
Unique Values	7099
Missing Values	48
Feature Type	Numeric Feature

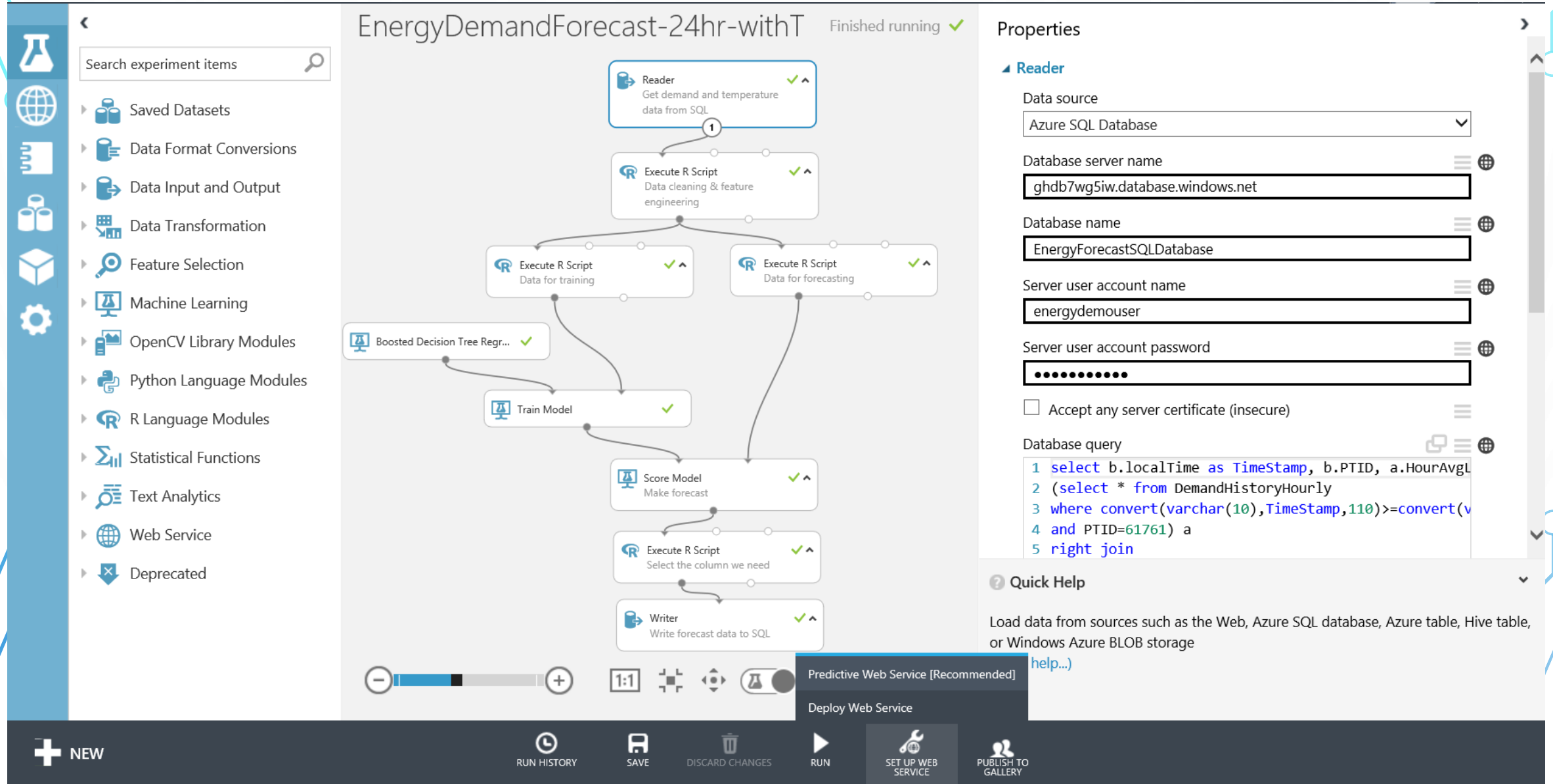
Visualizations

HourAvgLoad

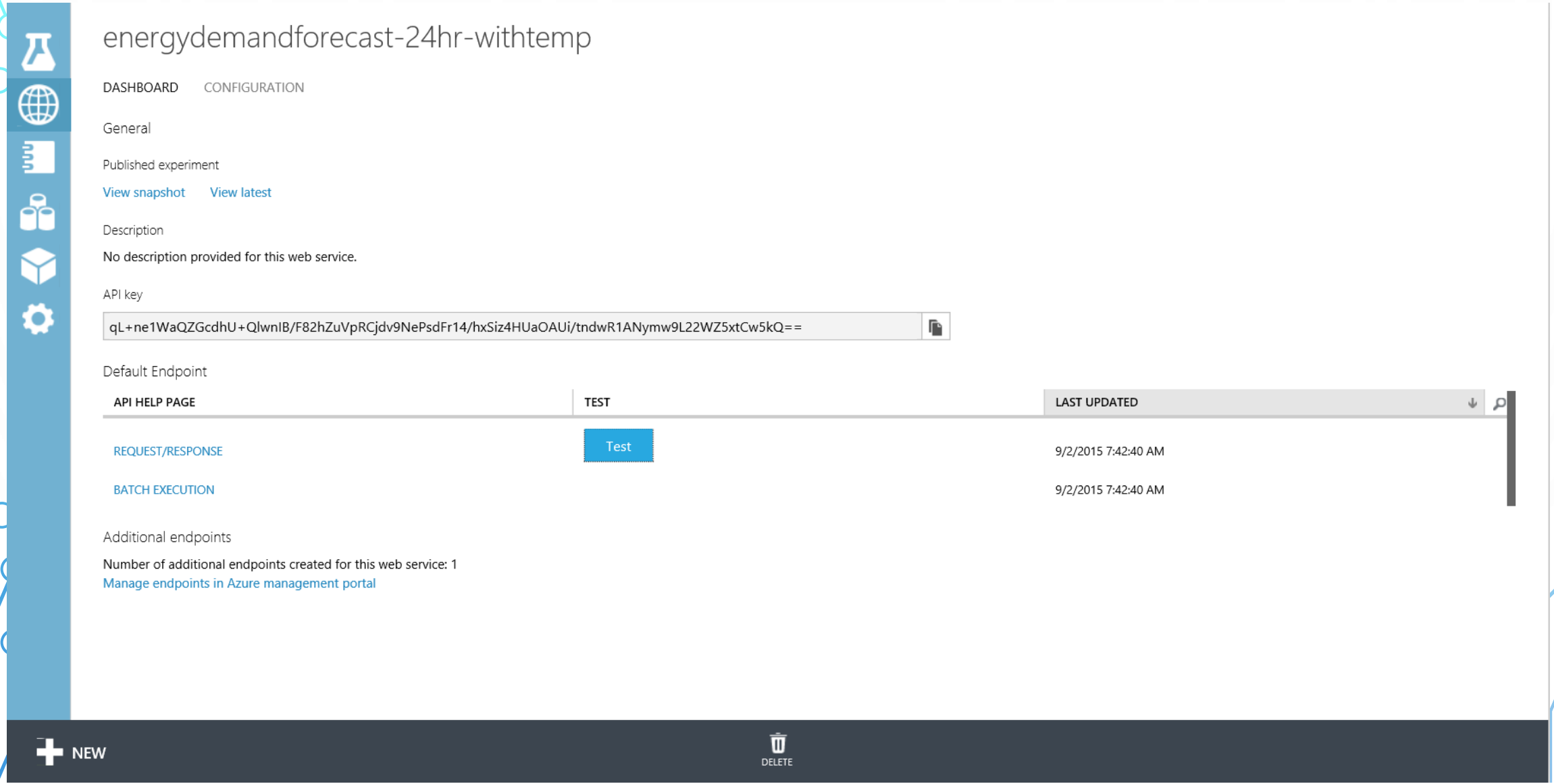
Histogram

compare to None

Azure Machine Learning



Azure Machine Learning



Azure Data Factory

Internal | Microsoft Azure > EnergyForecastADF24hrsModel > Diagram > sql_output52 > Data slices (by slice time) > Filter | Report bug

NEW
HOME
BROWSE ALL
NOTIFICATIONS
ACTIVE
BILLING

EnergyForecastADF24hrsModel
Data factory

Delete | Move

Diagram
EnergyForecastADF24hrsModel

Essentials

Resource group
EnergyForecastADFResourceGroup

Location
WestUS

Subscription name
ADS Demo Subscription

Provisioning state
Succeeded

Subscription id
9f156ff1-0bac-4c28-adcd-60bd97ff0cfc

All settings →

Summary

Author and deploy | Diagram | Sample pipelines

Contents

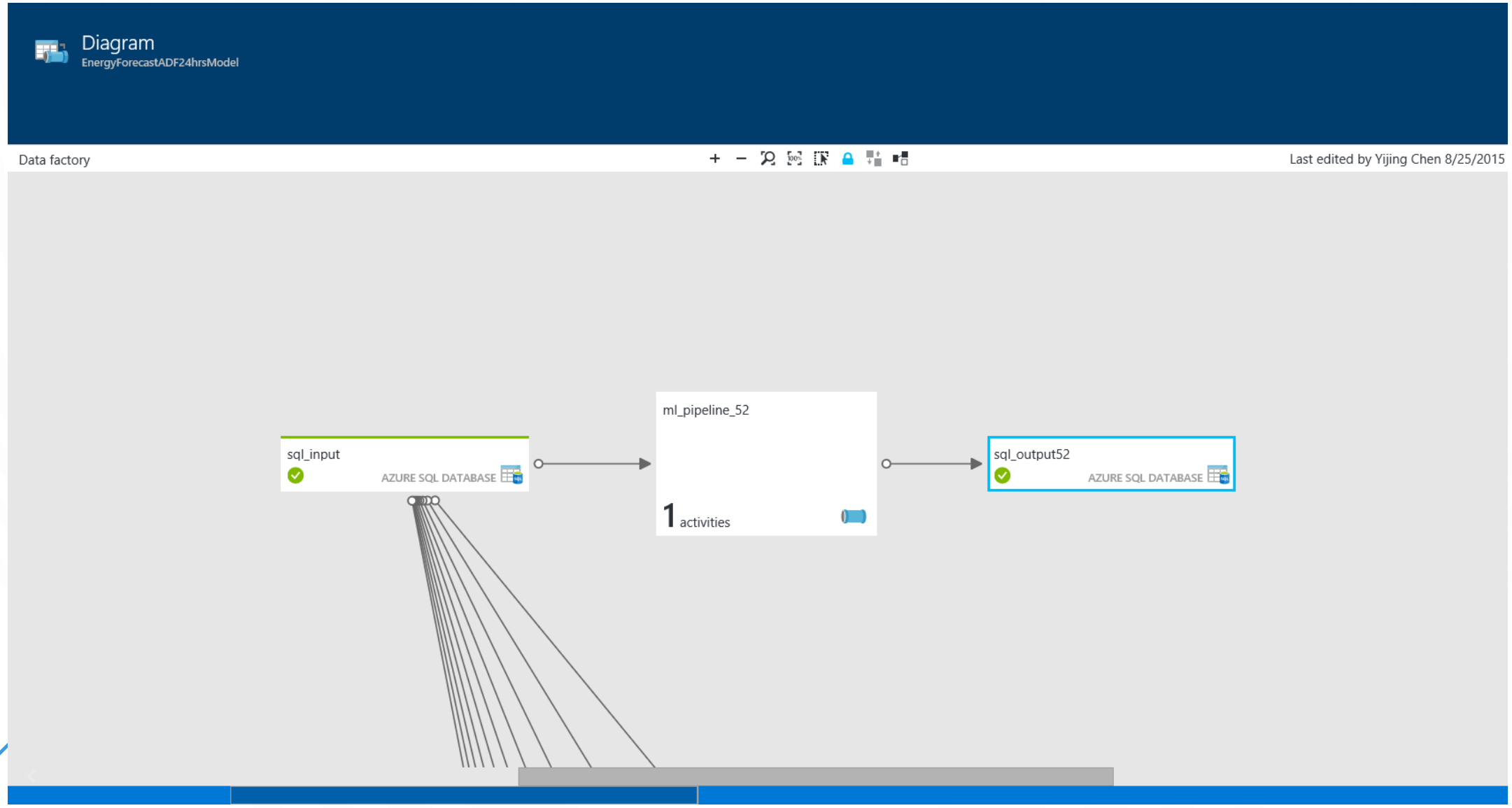
Datasets	Pipelines	Linked services
12	11	2
With errors 0	ml_pipeline_52	Data Stores 2
	ml_pipeline_53	Data Gateways 0
	...	...

Operations

Data factory

sql_input
AZURE SQL DATABASE

Azure Data Factory



Azure SQL Database

EnergyForecastSQLDatabase

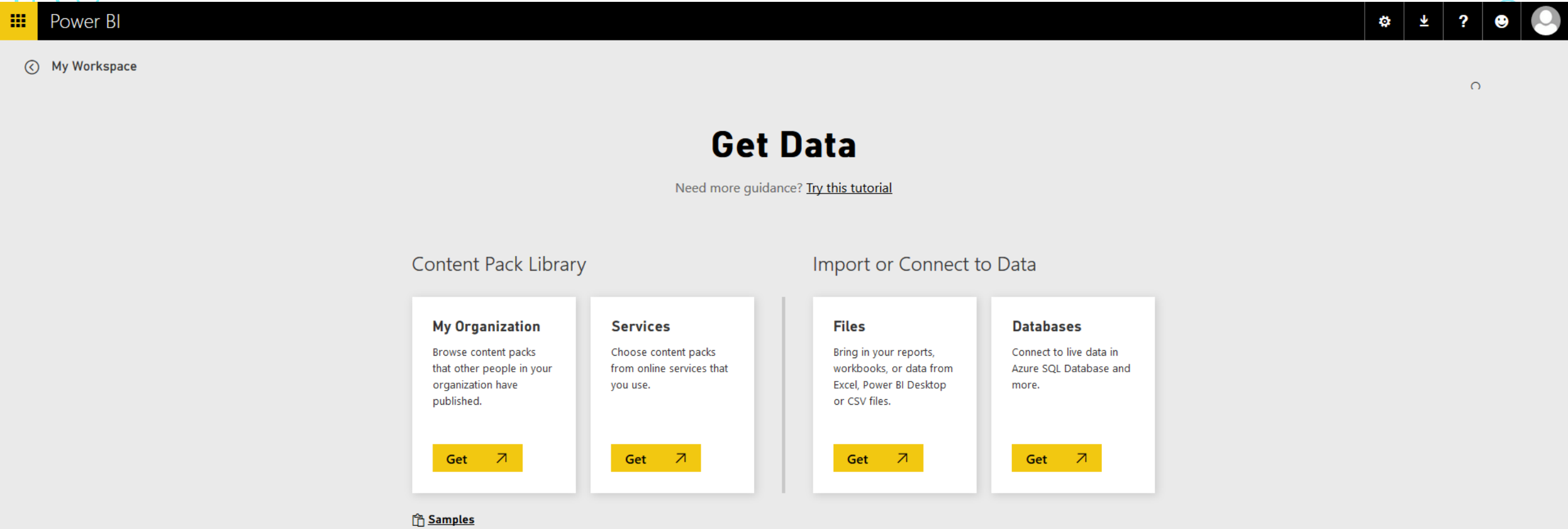
```
select TimeStamp,PTID,Name,Demand,Forecast,Hi95,Lo95,APE from v_24hrsV2ForecastPointErrorOngoing;
```

100 %

T-SQL | Results | Message

	TimeStamp	PTID	Name	Demand	Forecast	Hi95	Lo95	APE
1	2015-09-09 00:00:00.000	61752	WEST	1977.308	1945.6005859375	2067.62565239257	1823.57551948243	1.60358214631213
2	2015-09-09 00:00:00.000	61753	GENESE	1274.492	1236.68481445313	1339.22356717661	1134.14606172964	2.96642848194266
3	2015-09-09 00:00:00.000	61754	CENTRL	1906.825	1903.62426757813	2027.50057235469	1779.74796280156	0.167854086017595
4	2015-09-09 00:00:00.000	61755	NORTH	466.2583	480.521423339844	501.986668504588	459.0561781751	3.05905355112562
5	2015-09-09 00:00:00.000	61756	MHK VL	980.6583	996.900024414063	1078.255265899	915.544782929122	1.65620367476263
6	2015-09-09 00:00:00.000	61757	CAPITL	1600.533	1576.87219238281	1675.10120658335	1478.64317818228	1.47832803228981
7	2015-09-09 00:00:00.000	61758	HUD VL	1329.575	1282.88610839844	1373.3138754749	1192.45834132198	3.51156155072596
8	2015-09-09 00:00:00.000	61759	MILLWD	355.5917	334.907653808594	368.824894346224	300.990413270963	5.81679000428074
9	2015-09-09 00:00:00.000	61760	DUNWOD	923.9	858.163269042969	922.319857476789	794.006680609148	7.11513731291259
10	2015-09-09 00:00:00.000	61761	N.Y.C.	7634.192	7665.3486328125	8111.29290089669	7219.40436472831	0.408120973531843
11	2015-09-09 00:00:00.000	61762	LONGIL	2989.058	2947.04565429688	3173.26100377251	2720.83030482124	1.40554952090465

PowerBI


Power BI
My Workspace
Get Data
Need more guidance? Try this tutorial
Content Pack Library
Import or Connect to Data
My Organization
Browse content packs that other people in your organization have published.
Get
Services
Choose content packs from online services that you use.
Get
Files
Bring in your reports, workbooks, or data from Excel, Power BI Desktop or CSV files.
Get
Databases
Connect to live data in Azure SQL Database and more.
Get
Samples

Demo: Live Dashboard

PowerBI

Power BI

CAW EnergyForecast Dashboard

CAW EnergyForecast Dashboard

Share Dashboard

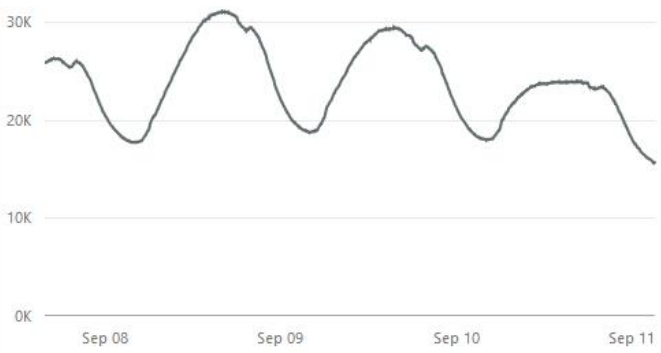
Ask a question about the data on this dashboard

Data

Number of Region

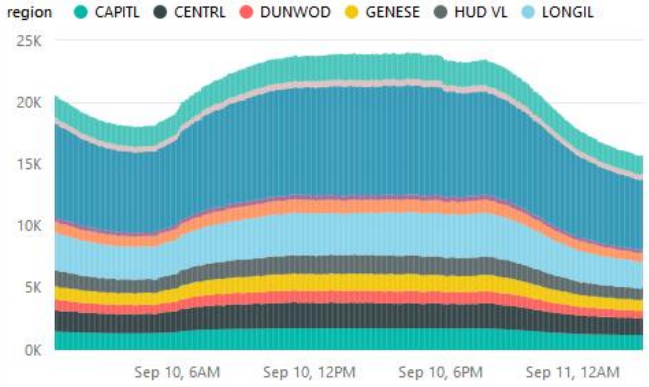
11

Total Demand by Time



Regional Demand by Time

BY TIME, REGION



Demand Raw Data

time	region	demand
09/11/15 03:10:00 AM	CAPITL	1,212.90
09/11/15 03:10:00 AM	CENTRL	1,414.70
09/11/15 03:10:00 AM	DUNWOD	594.40
09/11/15 03:10:00 AM	GENESE	863.30
09/11/15 03:10:00 AM	HUD VL	958.20
09/11/15 03:10:00 AM	LONGIL	2,159.40
09/11/15 03:10:00 AM	MHK VL	712.80
09/11/15 03:10:00 AM	MILLWD	213.20
09/11/15 03:10:00 AM	N.Y.C.	5,660.50
09/11/15 03:10:00 AM	NORTH	436.20
09/11/15 03:10:00 AM	WEST	1,499.90
09/11/15 03:05:00 AM	CAPITL	1,200.50
09/11/15 03:05:00 AM	CENTRL	1,357.70

Insights

Average Demand in Last 24Hrs

2K

Average Demand by Region

Last 24Hrs Peak Demand

time	demand
09/10/15 04:30:0...	24,056.80
09/10/15 03:50:0...	24,010.90
09/10/15 04:45:0...	24,005.10
09/10/15 04:35:0...	23,990.70

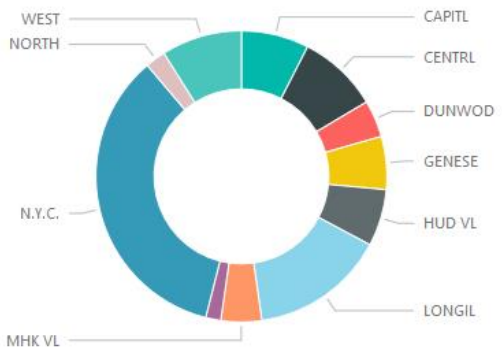
Last 24Hrs Off-Peak Demand

time	demand
09/11/15 03:05:0...	15,665.60
09/11/15 02:55:0...	15,684.50
09/11/15 03:00:0...	15,722.90
09/11/15 03:10:0...	15,725.50

Total Demand

Average Demand

Total Demand by Region



PowerBI

Power BI

CAW EnergyForecast Dashboard

CAW EnergyForecast Dashboard

Share Dashboard

Ask a question about the data on this dashboard

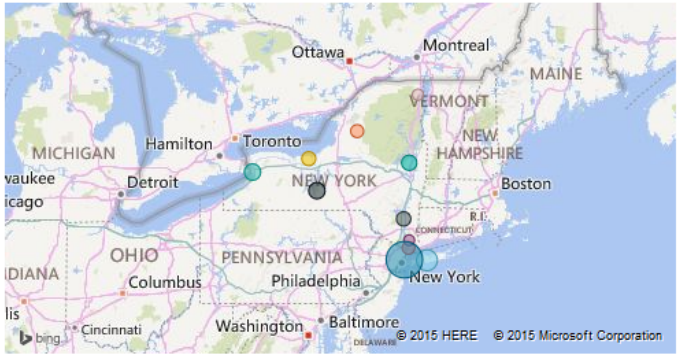
Insights

Average Demand in Last 24Hrs

2K

Average Demand by Region

CAPITL CENTRL DUNWOD GENESE HUD VL LONGIL MHK VL



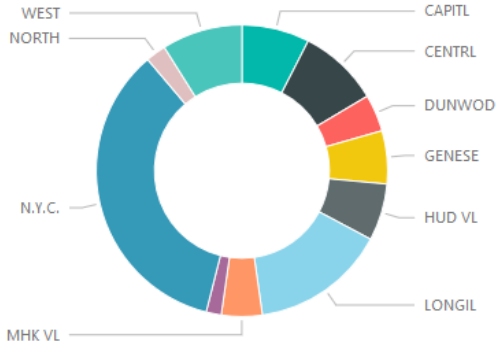
Last 24Hrs Peak Demand

time	demand
09/10/15 04:30:0...	24,056.80
09/10/15 03:50:0...	24,010.90
09/10/15 04:45:0...	24,005.10
09/10/15 04:35:0...	23,990.70

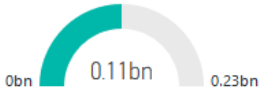
Last 24Hrs Off-Peak Demand

time	demand
09/11/15 03:05:0...	15,665.60
09/11/15 02:55:0...	15,684.50
09/11/15 03:00:0...	15,722.90
09/11/15 03:10:0...	15,725.50

Total Demand by Region



Total Demand



Average Demand

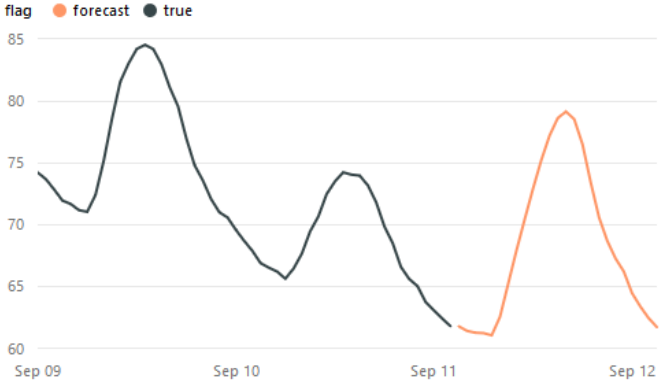
Forecast

Average Error Rate (%)

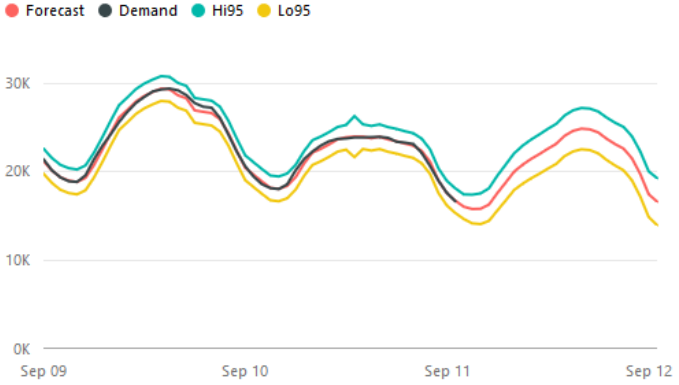
2.00

Average Temperature by Time

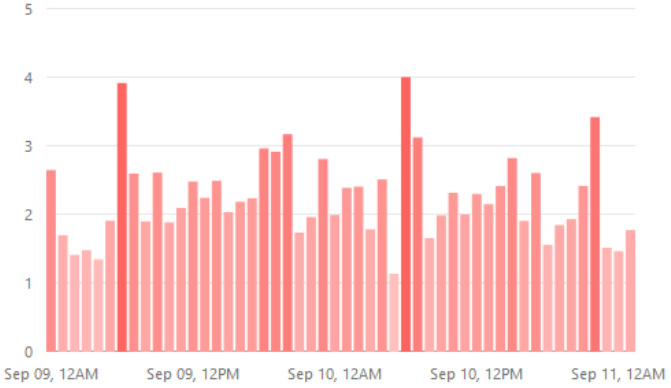
POWERED BY FORECAST.IO



Demand vs Forecast with Confidence Interval



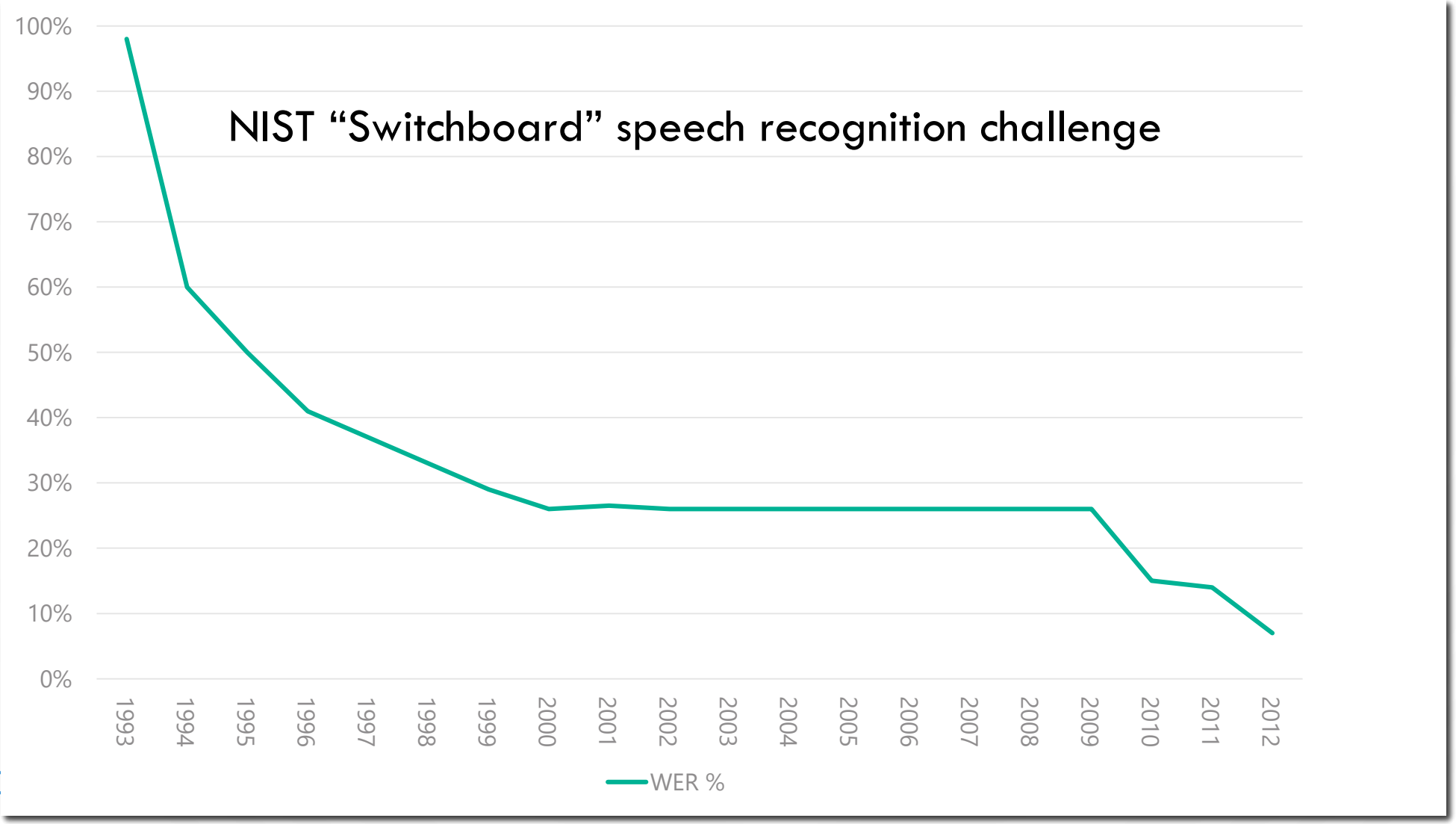
Error Rate by Time (%)



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.

Satya Nadella,
CEO, Microsoft

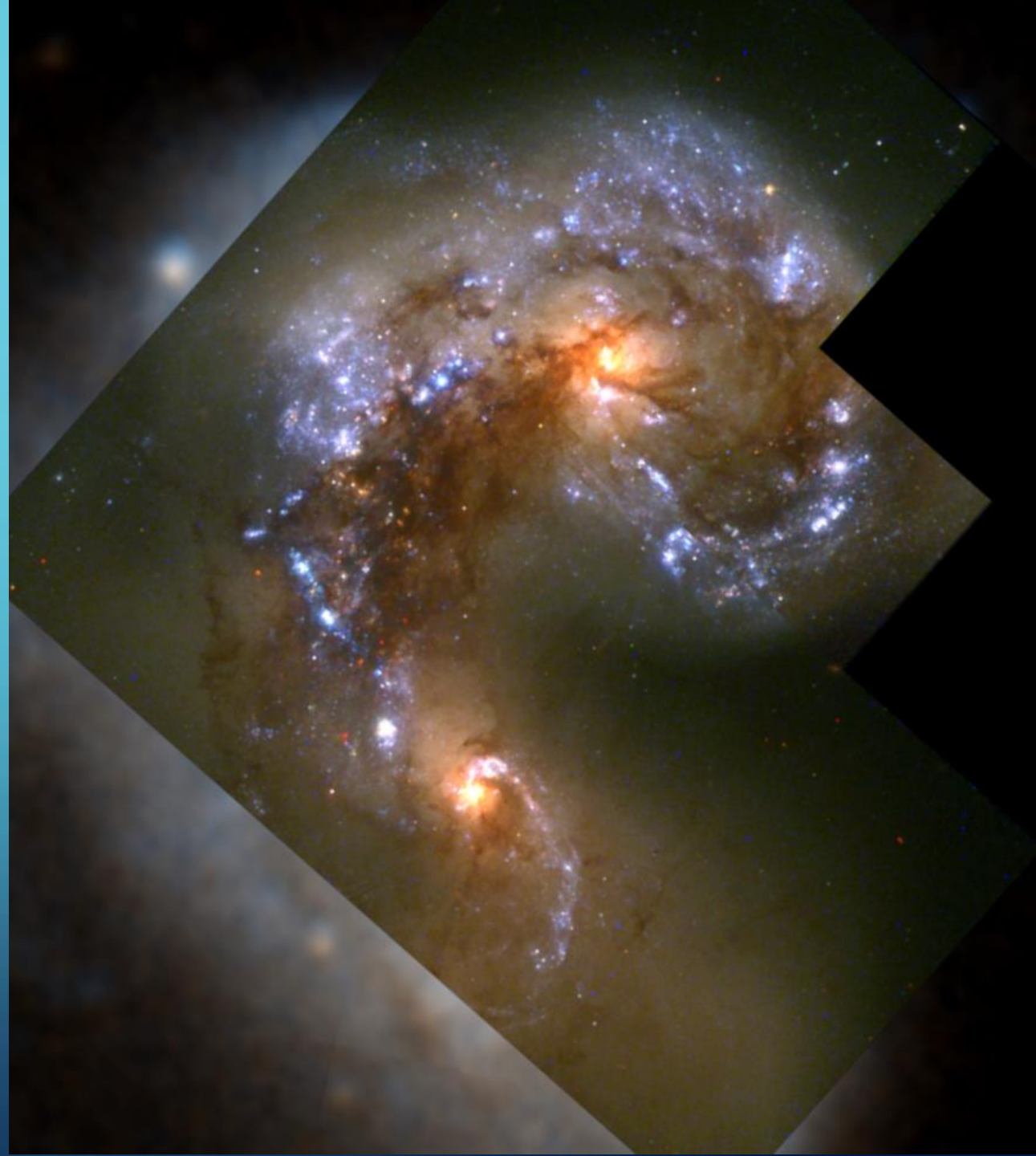
Deep Learning's First Major Success:
Speech Recognition




skype
I was wondering what you are going to do later?
Me preguntaba lo que vas a hacer después?
Going to the pub, do you want to join us? I think you met some of the team at the party in April.
Al pub , ¿quieres unirte a nosotros? Creo que conociste a algunos miembros del equipo en la fiesta en abril .
Can I join you guys after my meeting?
¿Puedo unirme a ustedes después de mi reunión?
Type a message in English here

http://www.worldwidetelescope.org/

Classifying galaxy structure using neural networks with R



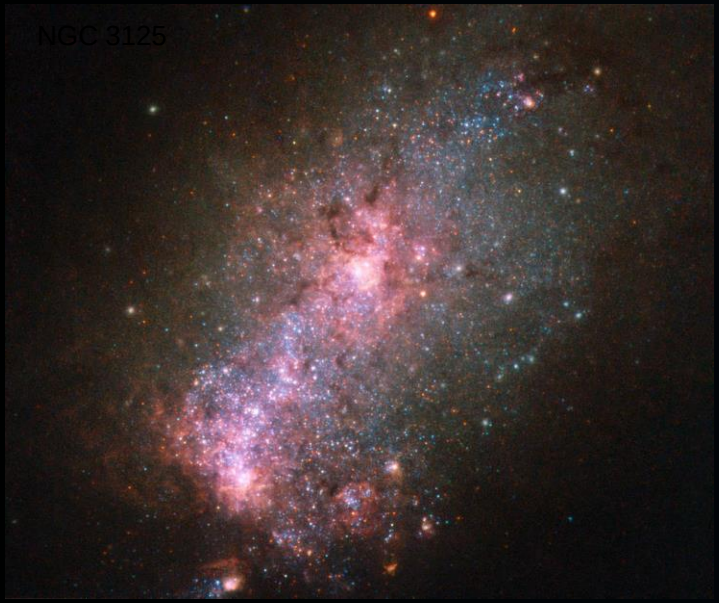
2 trillion Hubble ultra deep

~~200 billion~~ Hubble deep field

~~100 Billion~~ Galaxies in observable universe

http://www.nasa.gov/feature/goddard/2016/hubble-reveals-observable-universe-contains-10-times-more-galaxies-than-previously-thought

Galaxy shape tells us about evolution



Forming

Ancient

Image: http://www.nasa.gov/image-feature/goddard/2016/hubble-views-a-galaxy-fit-to-burst

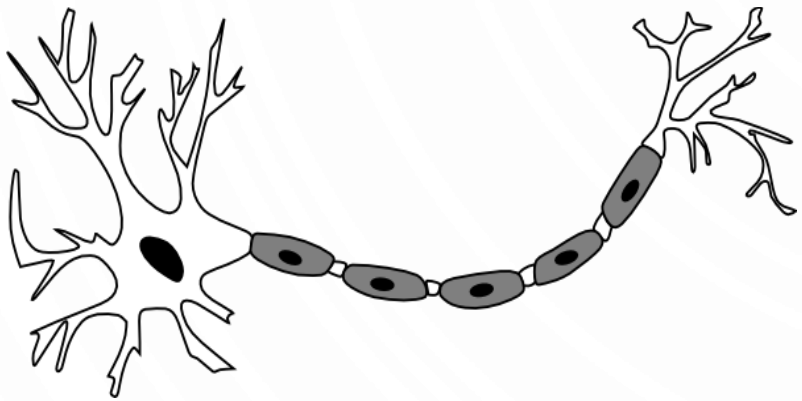
Galaxy shape tells us about evolution

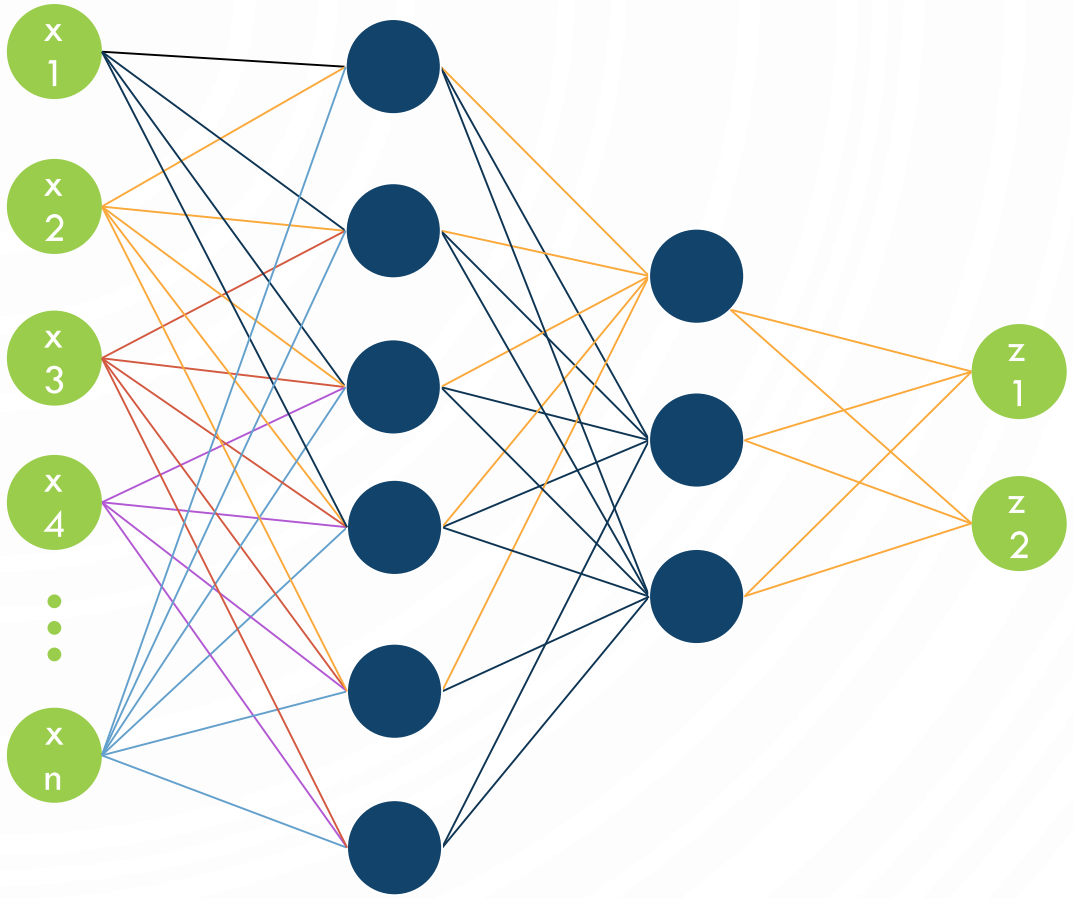
Spiral galaxies

Elliptical galaxies

NASA, ESA

ARTIFICIAL NEURAL NETWORKS



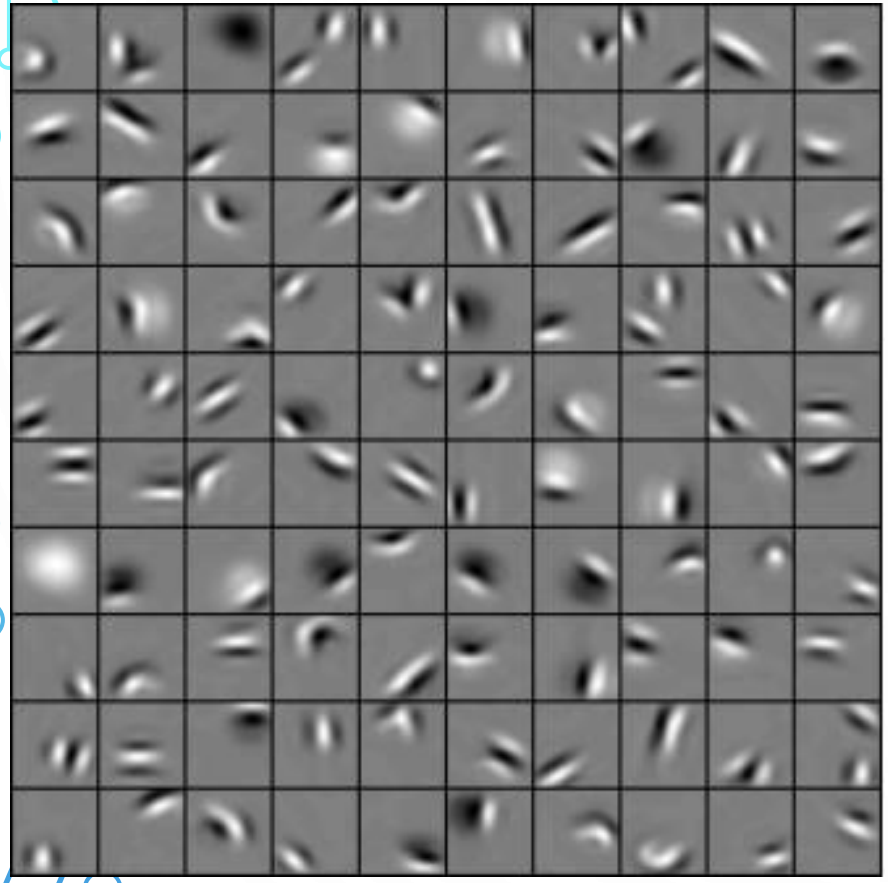


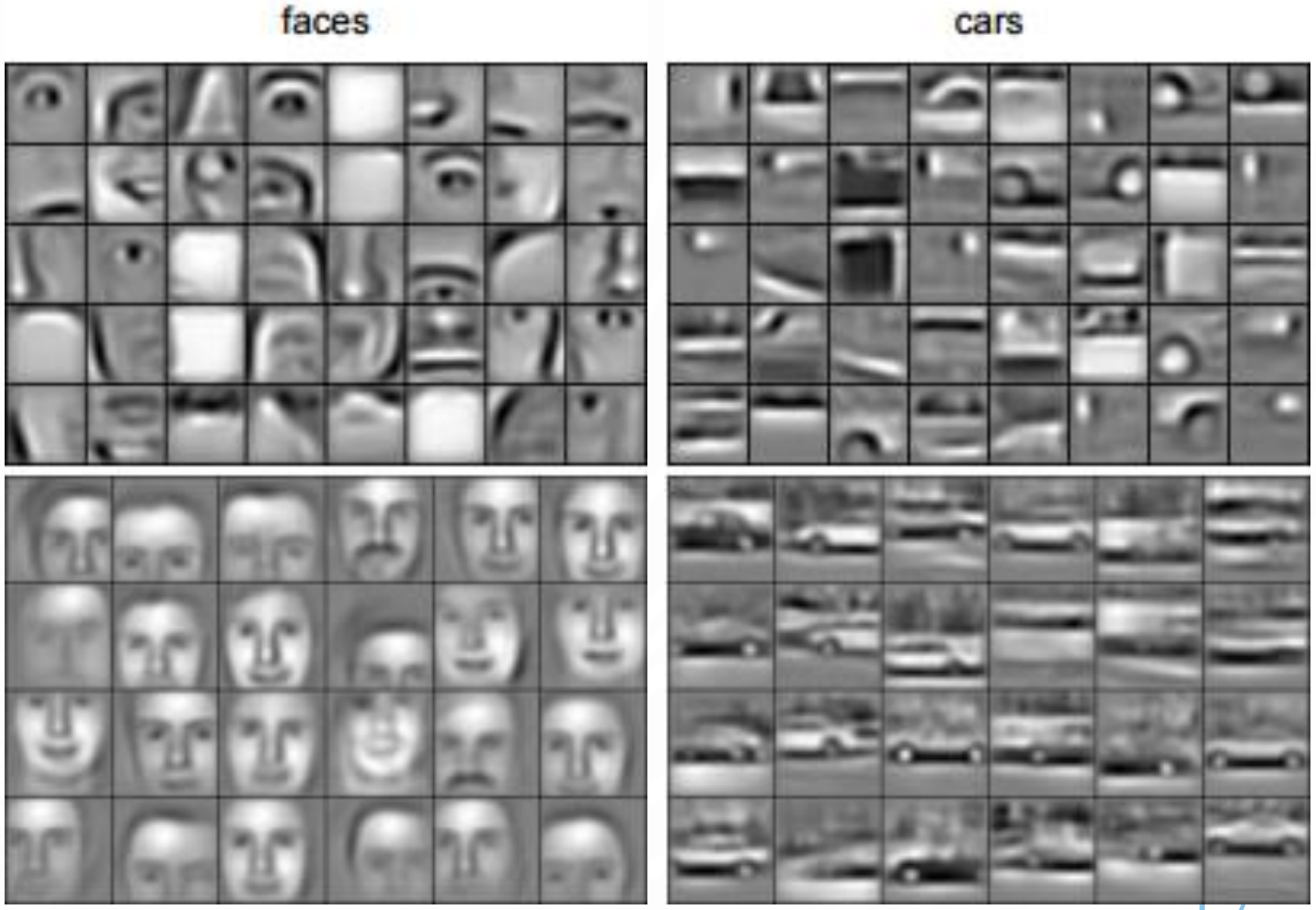
Data

Hidden layer(s)

Outcome

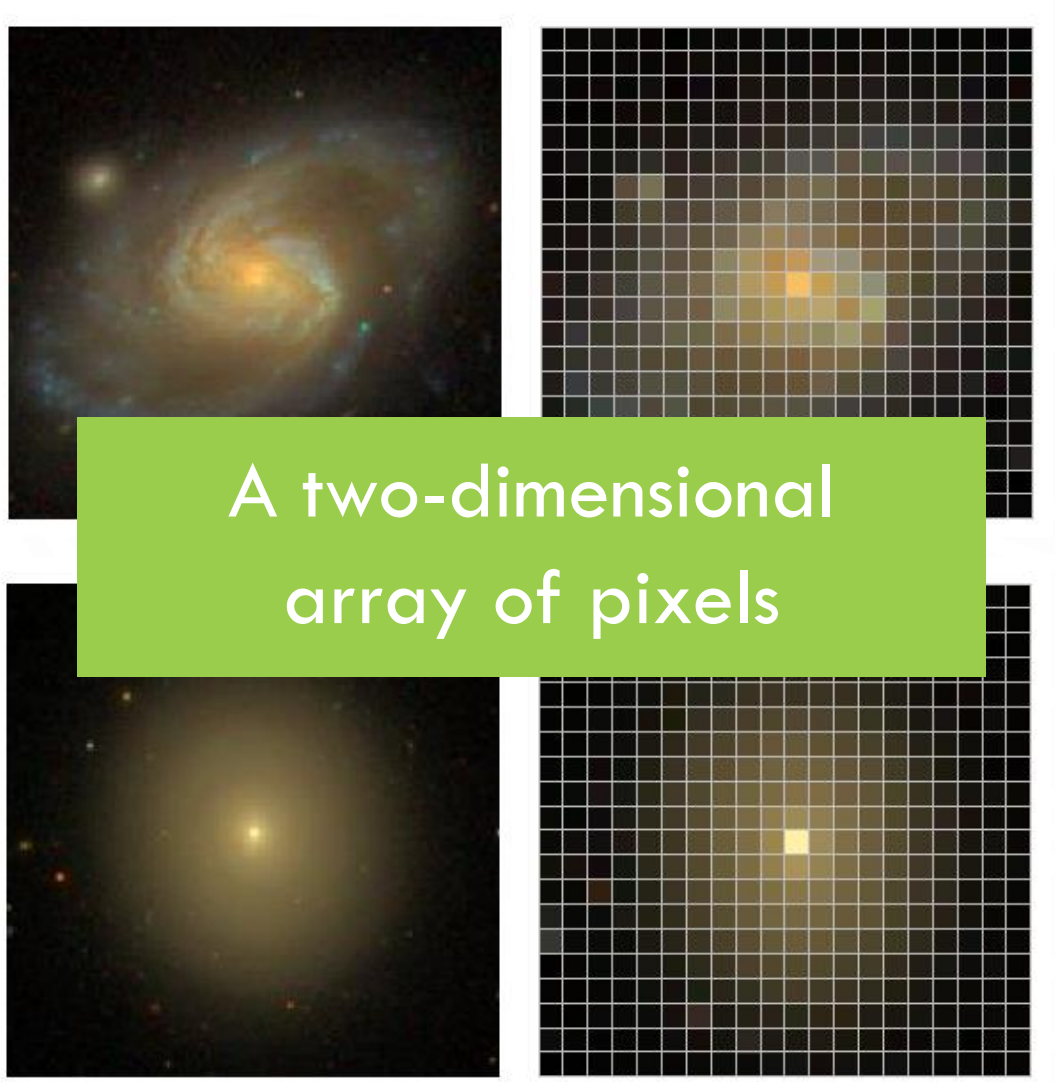
How do computers "see"?



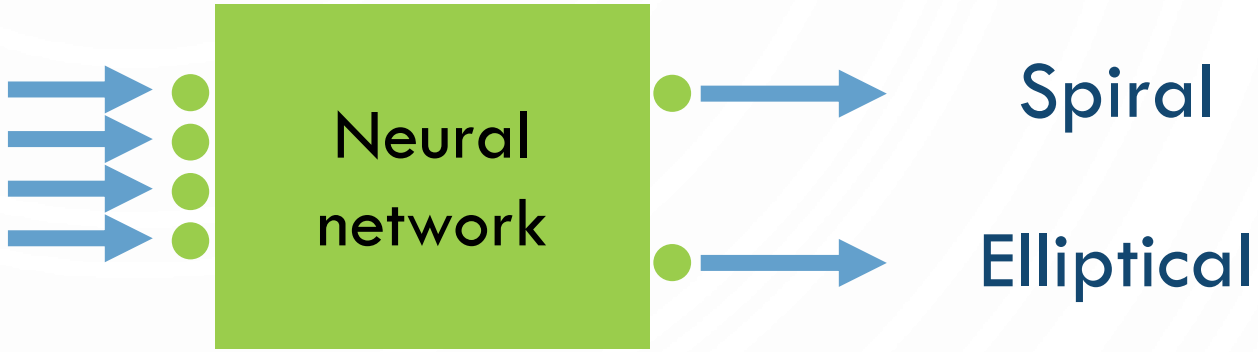


- Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representation
- Honglak Lee, Roger Grosse, Rajesh Ranganath, Andrew Y. Ng

WHAT COMPUTERS "SEE"



A two-dimensional array of pixels



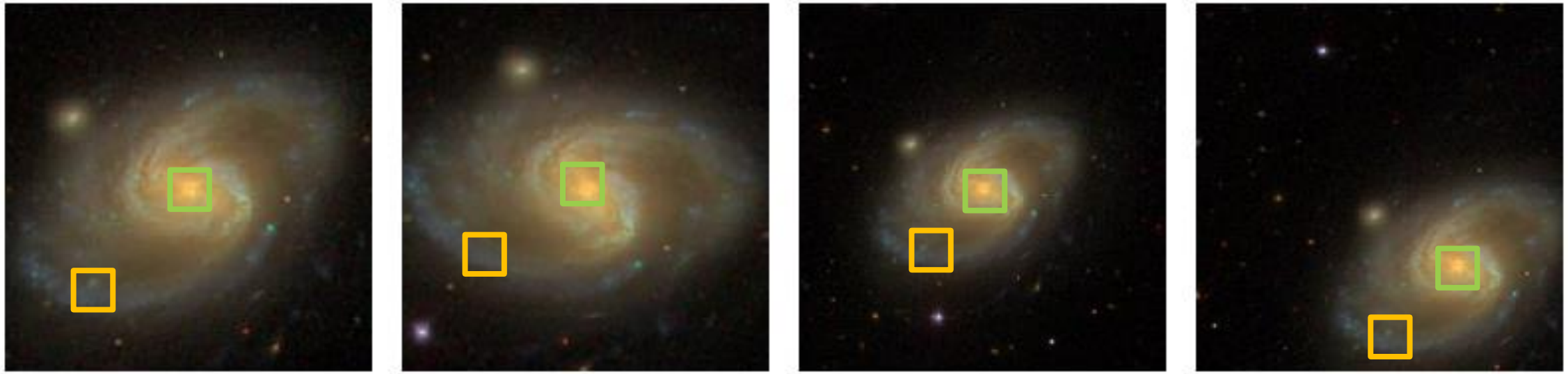
THE COMPLICATION

rotation scaling translation

Neural network

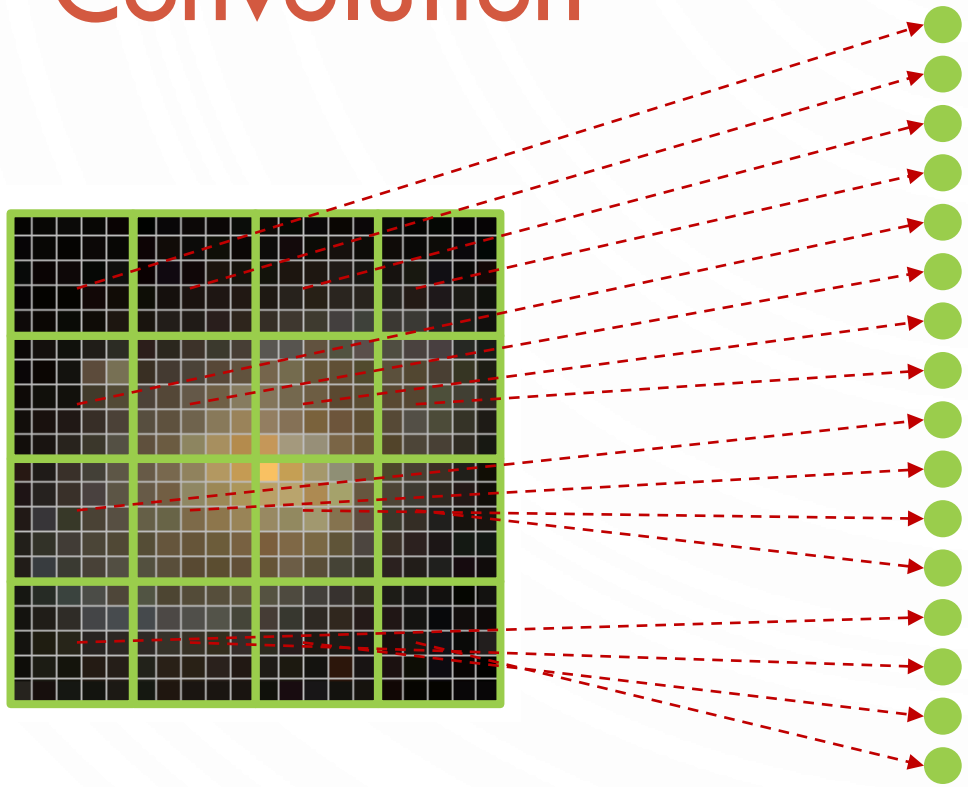
THE SOLUTION

Match pieces of the image



Then repeat across the entire image

Convolution

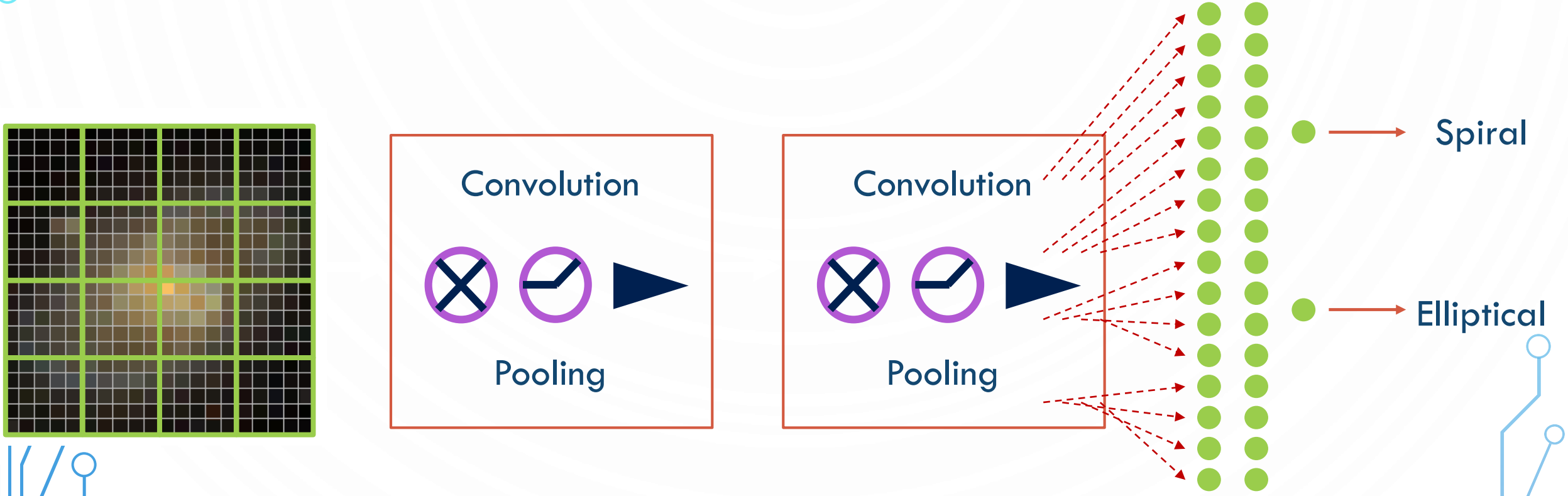


Matches specific shape (kernel) across entire image

Automatic feature generation

Deep stacking

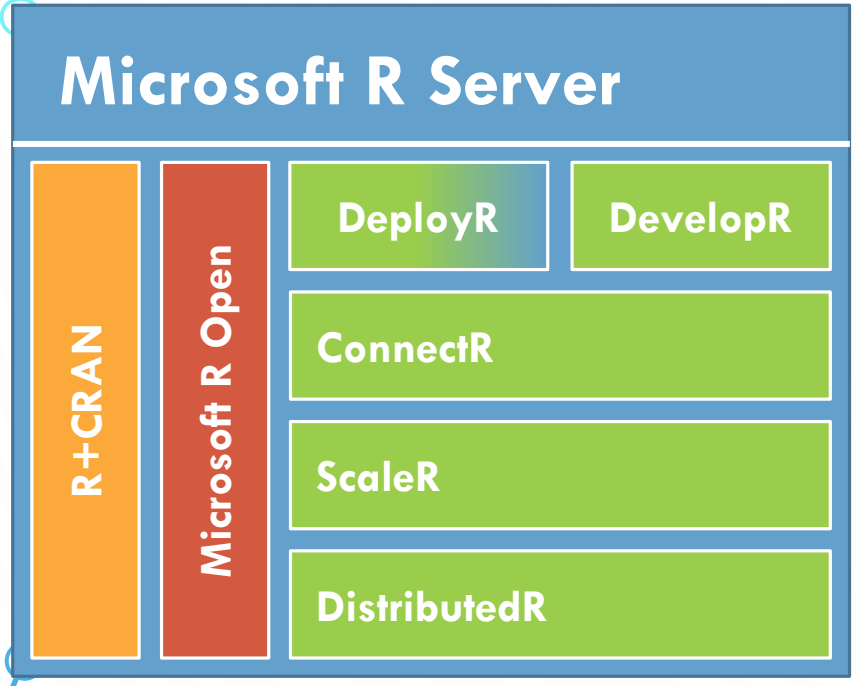
Layers can be repeated several (or many) times



R Server

Microsoft R Server – R for Big Data

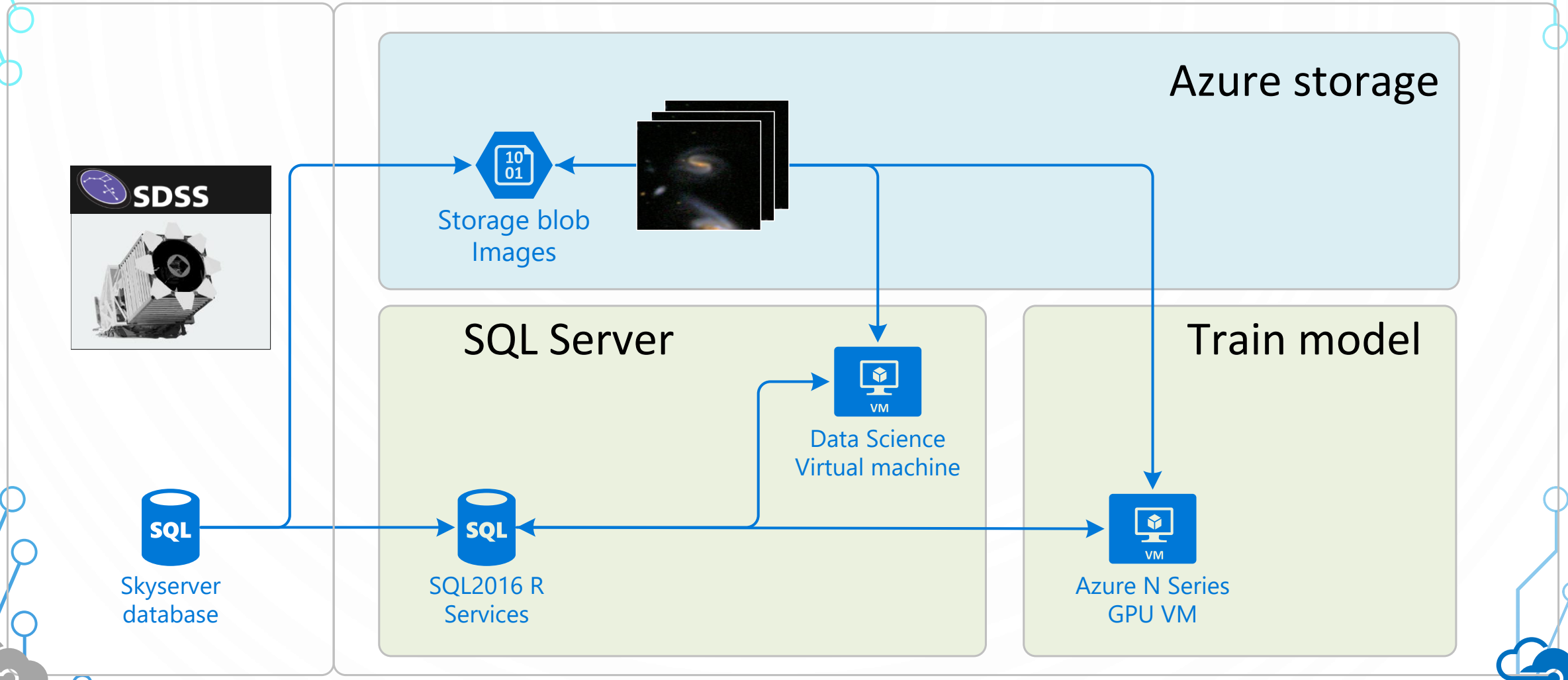
The All-Inclusive Big Data Big Analytics Platform



High-performance open source R:

- Scale & Performance
  - Scales from workstations to large clusters
  - Scales to large data sizes
  - Growing portfolio of Parallelized algorithms
- Write Once Deploy Anywhere - multiple platforms
- IDE for data scientists and developers

Training on Azure



Web

Azure

Results

100K * 3	Training images, augmented with rotation
8	Layers in deep network
176K	Weights to compute in network
2.5B	Weight updates per second
1.8 hours	Computing time on Azure N series GPU
88%	Overall accuracy - training data
55%	Overall accuracy - test data (91% micro-average)

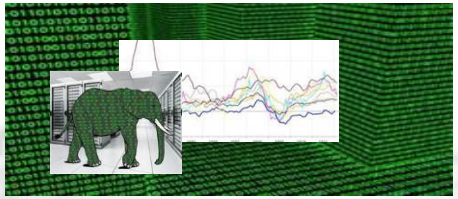
The technique works, but has scope for improvement!

"Big Data for Healthy Cities: Using Location-aware Technologies, Open Data and 3D Urban Models to Design Healthier Built Environments"

H. Miller and K. Tolle

Transforming Science to Being Service-centric

Gather data from devices



Transform data into information and insight

Monitor processes and assets remotely

Connectivity, open data access and interoperability transforms changes scientific focus from reactive to proactive

Enable inclusive global cross-functional collaboration

Take corrective action anytime, anywhere