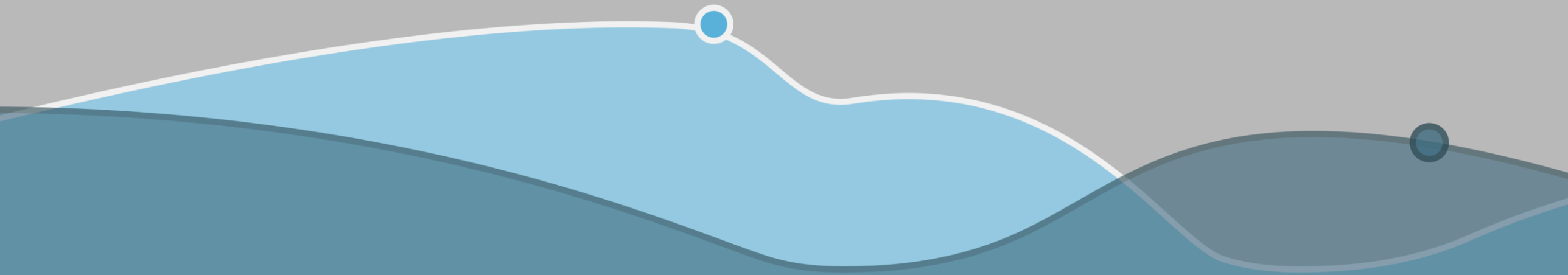
Microsoft

Cortana Analytics Workshop

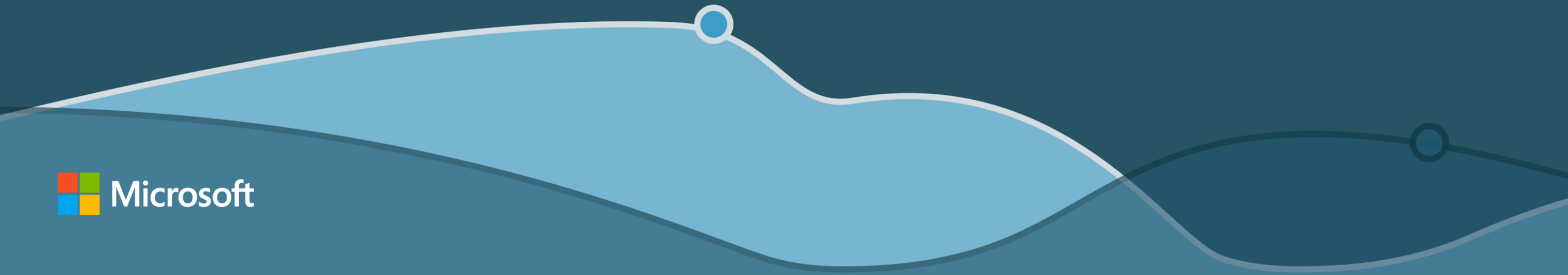
Sept 10 – 11, 2015 • MSCC



Real-World Data Collection for Cortana Analytics

Spyros Sakellariadis
Principal Program Manager, MS Applied Technologies, Azure Machine Learning

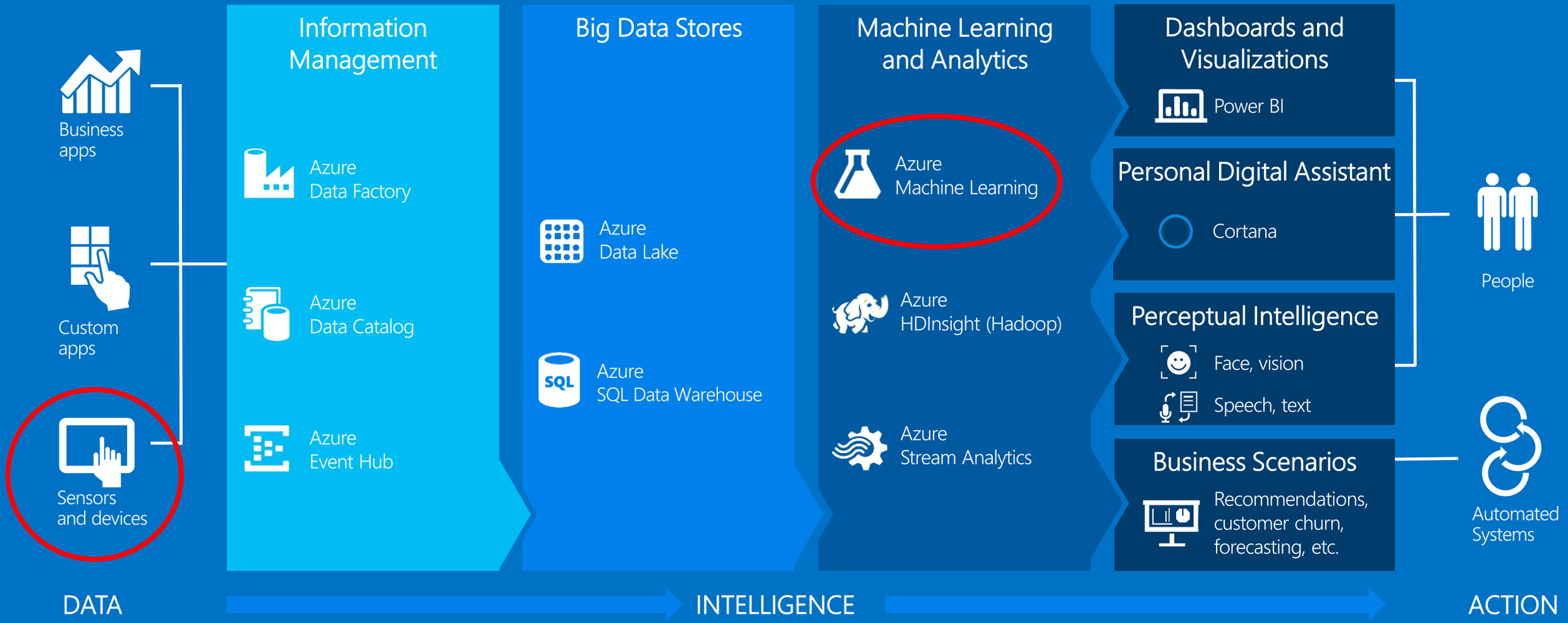
Joshua Peschel
Assistant Professor, Civil & Environmental Engineering, University of Illinois



Microsoft

Cortana Analytics Suite

Transform data into intelligent action



Synopsis

Why Cortana Analytics?

Cortana Analytics makes it possible for more people to do very complex analyses quickly

Caveat emptor

Avoid the temptation to jump into the analysis too quickly…

… Consider a few anecdotes of faulty conclusions due to insufficient attention to data acquisition

What we are working on

A current sensor data project starting to push 4.3M records/month to Cortana, projecting 1.1B/month

What we hope you'll take away from this session

You need to pay particular attention to design of data collection phase in experiment

The stuff of nightmares

Scientists using theoretical models,
That get turned into recommendations,
That become policy, and
Leave the real world behind
in the rear view mirror

Photo Wendelland Carolyn/E+/Getty Images

Putting the ‘Real’ into ‘Real-World’
Working with data before it’s a dataset

First, dig a hole

Photo stevecoleimages/E+/Getty Images

Or, climb a tree

Photo Hero Images/DigitalVision/Getty Images

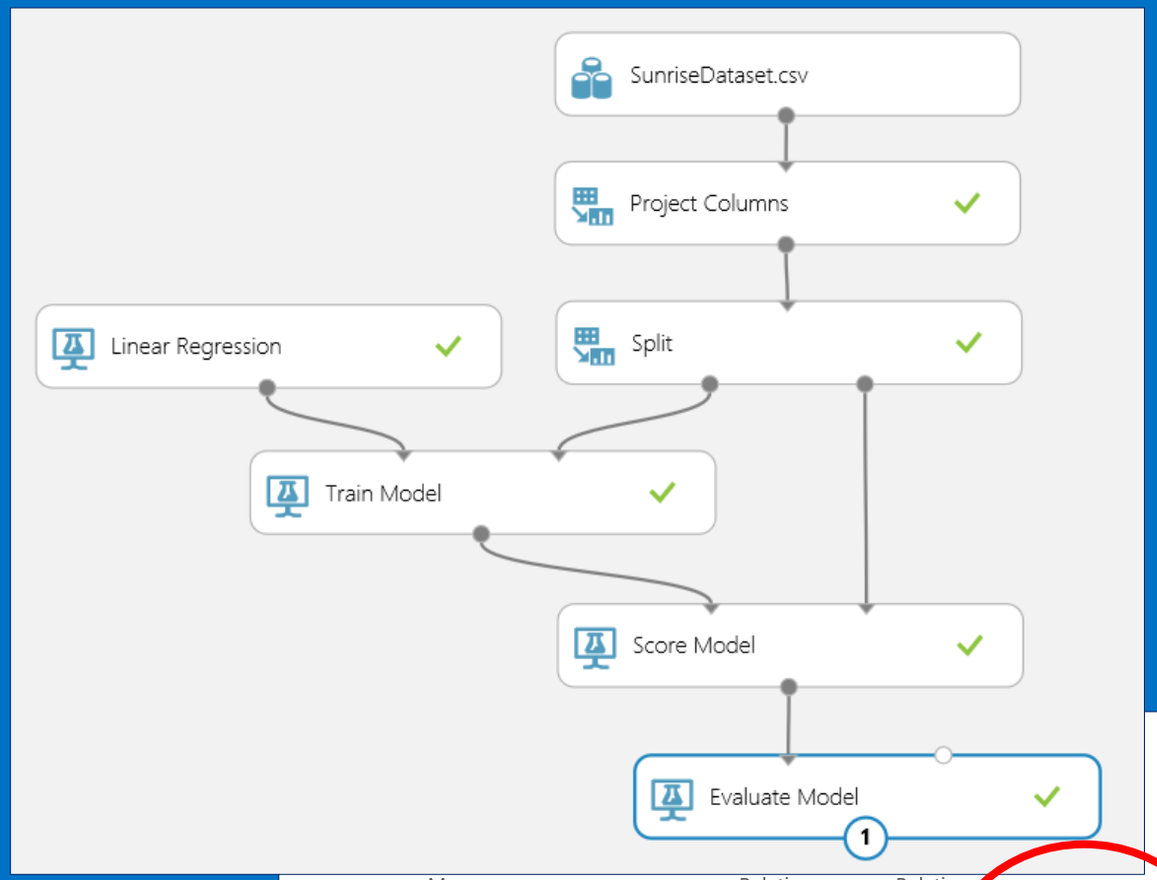
Let’s start with a few anecdotes
Tons of data leading to problematic conclusions

1. Bertrand Russell’s chicken
2. My VM usage metrics
3. IoT developers’ preferences
4. Juvenal’s black swan

Bertrand Russell's chicken

What's the probability that the sun will rise tomorrow?

Month	Day	Sunrise	Date	Time	Sunrose
1	1	757	1/1/2014	7:57 AM	Yes
1	2	757	1/2/2014	7:57 AM	Yes
1	3	757	1/3/2014	7:57 AM	Yes
1	4	757	1/4/2014	7:57 AM	Yes
1	5	757	1/5/2014	7:57 AM	Yes
1	6	756	1/6/2014	7:56 AM	Yes



	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
view as					
	0	0	0	0	1

Alas

"The man who has fed the chicken every day throughout its life at last wrings its neck instead, showing that more refined views as to the uniformity of nature would have been useful to the chicken."

The Problems of Philosophy, Bertrand Russell, 1912

Improving the chicken's reasoning

Chicken should add *fed* variable

Correlation coefficient of sunrise and being fed is high
Still not enough – chicken still ends up on dinner table

Should add other chickens' fates

If there is only one chicken on farm, would need to add other farms and chickens' fates
But chicken has no concept of a farm or even other chickens

Key learning – high scores not enough

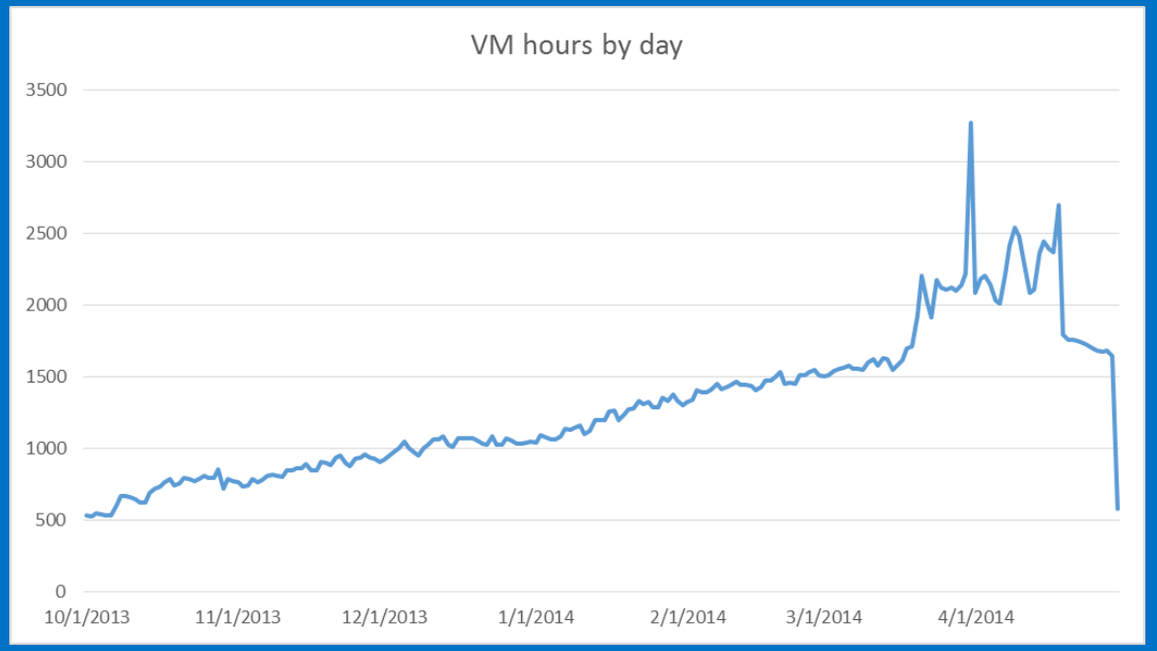
Looking at dataset alone may not be sufficient to predict outcomes
Significant domain knowledge needed in order to decide if dataset contains right data

My VM usage metrics

How many compute hours are used by my VMs?

All internal cross-checks good, dataset passed engineering scrutiny

Analyst opinion – interest in product was declining precipitously



Further investigation showed decline due to change in data acquisition process, not decline of VM usage – dataset that looked good was in fact bad

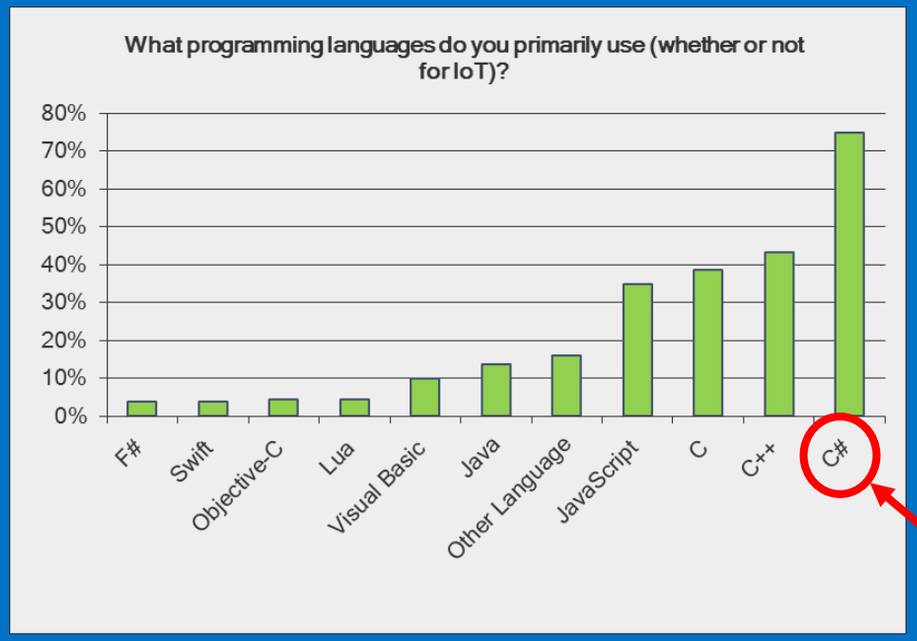
IoT developers' preferences

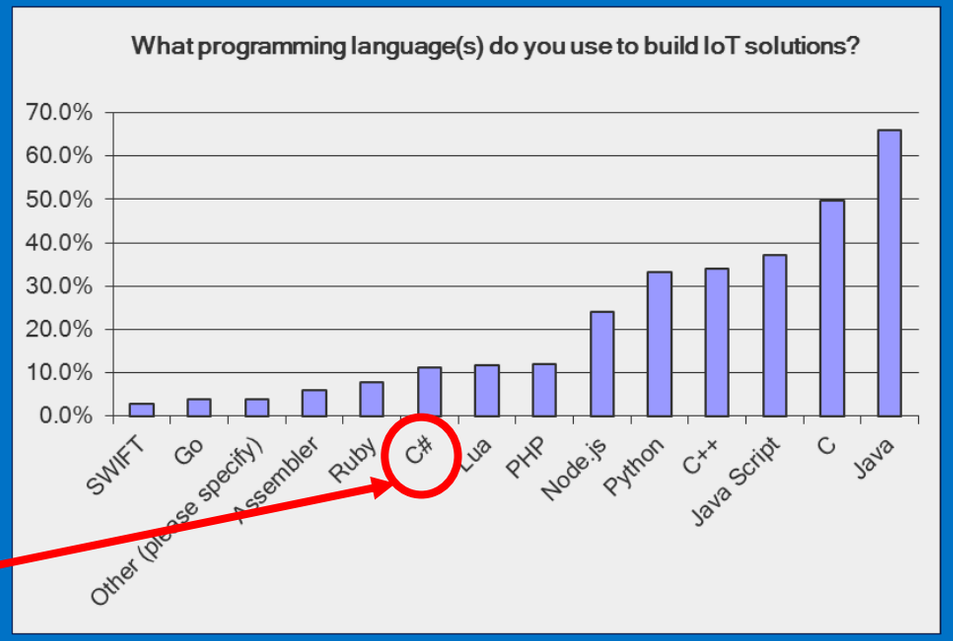
What programming languages do IoT developers use?

Two datasets from two IoT survey producers looked good, but led to widely different analyses

Difference due to selection bias or difference in wording of question?

What's the role of the domain expert in the experimental design?





http://eclipse.org/org/community_survey/SurveySummary_03312015.xls

Juvenal's black Swan

Are all swans white?

Only white swans had been observed (Juvenal Satires, 1st century AD), leading to the hypothesis that "All swans are white"

Discovery of black swans in Australia (Willem de Vlamingh, 1697) fueled the debate over inductivism and scientific methodology that can be traced back to the allegory of the cave (Plato, *Republic*, 4th century BC) as a metaphor for ignorance

What's the role of the data scientist or the domain expert in creating or blessing a dataset?

Photo Ian Gethings/Moment/Getty Images

Key takeaways

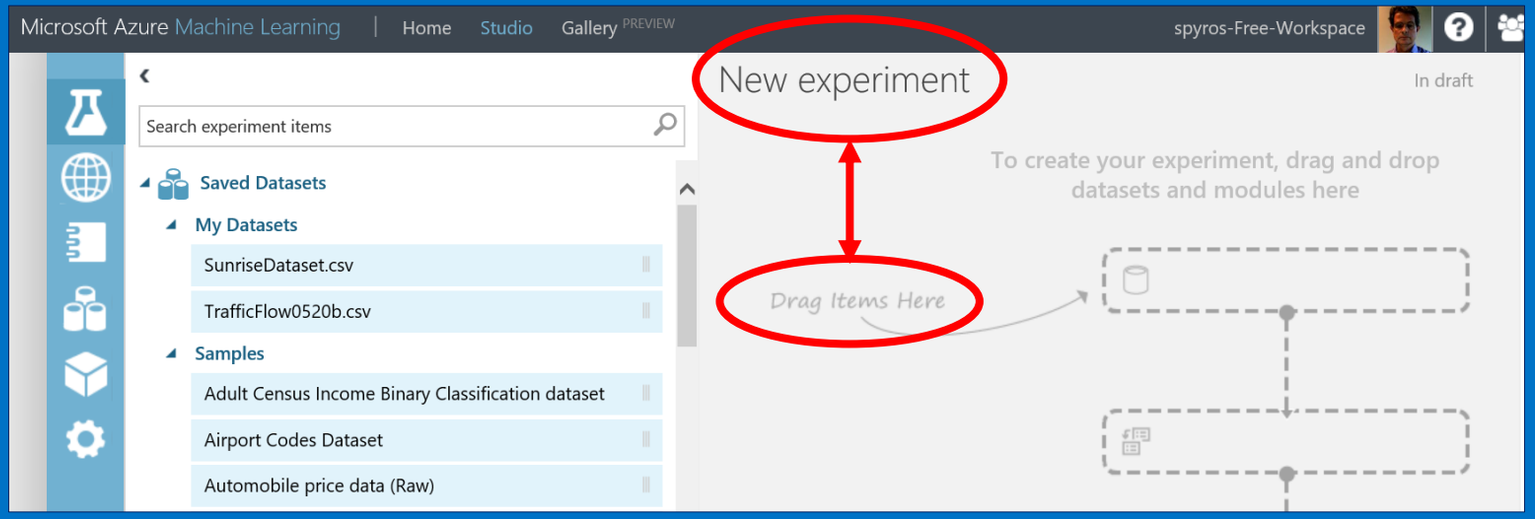
Domain subject matter knowledge important

Conclusions seemed fine from the perspective of the dataset

Needed hindsight and domain expertise to realize the problems of the dataset

Translation

Lots of science involved before we can drag a dataset for a new experiment



Double-clicking on data acquisition

In a typical data science project, we spend up to 75 to 80% of time in data acquisition and preparation.[1]

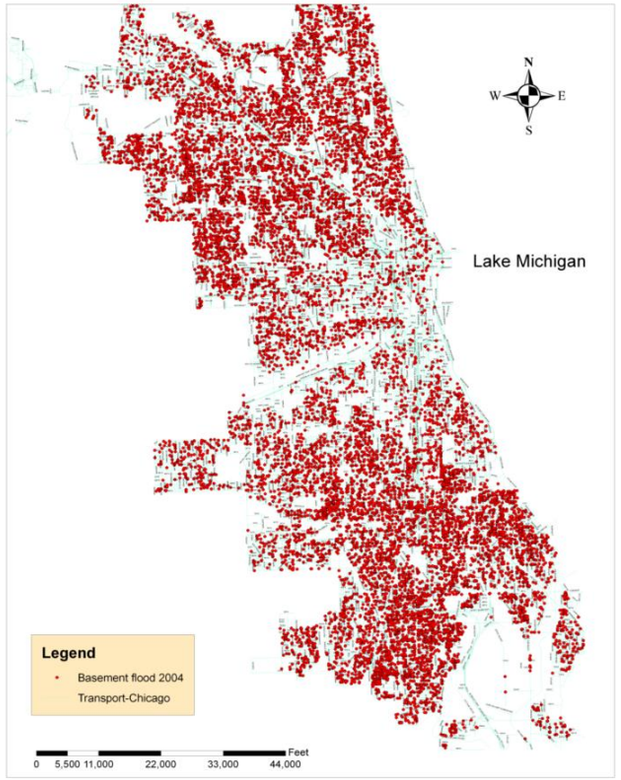
Finding a good set of features for creating a predictive model requires experimentation and knowledge about the problem you want to solve.[2]

Photo by John A. Kelly, USDA Natural Resources Conservation Service Creative Commons

[1] *Predictive Analytics with Azure Machine Learning*, Roger Barga , Valentine Fontama , Wee Hyong Tok http://www.apress.com/9781484204467
[2] *Create your first experiment in Azure Machine Learning Studio*, Gary Ericson https://azure.microsoft.com/en-us/documentation/articles/machine-learning-create-experiment/

The problem at hand

It floods in Chicago …

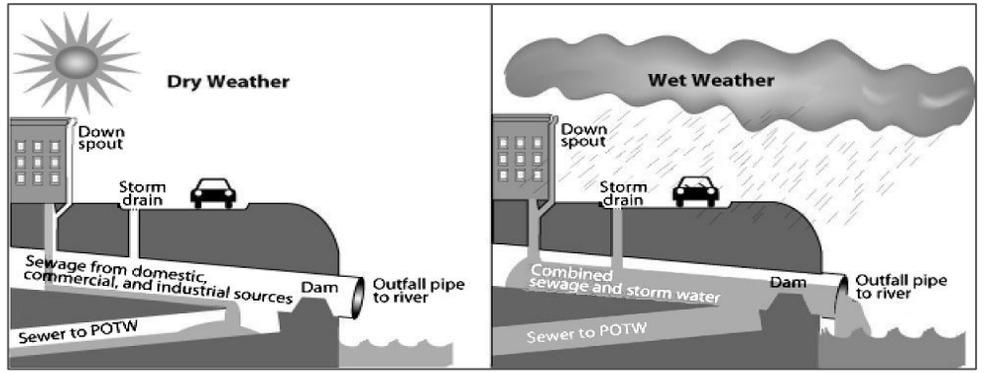


Basement flooding complaints
documented in Chicago during 2004

Photo AP Photo/M. Spencer Green

The problem

… and in over 700 cities with similar infrastructure



Images by Environmental Protection Agency (EPA), Washington, D.C.

In most urban areas, stormwater is drained through engineered collection systems and discharged into nearby waterways; trash, bacteria, heavy metals, and other pollutants from the urban landscape can all be included in the discharge.

Hypothesis

Improving stormwater handling can reduce flooding


Photos by Christopher Zurcher, Creative Commons

Vegetation, soils, and natural processes to manage water and create healthier urban environments.

Permeable Pavers

Infiltration Planters

Infiltration Bump outs

Images by Environmental Protection Agency (EPA), Washington, D.C.

Coming clean with water

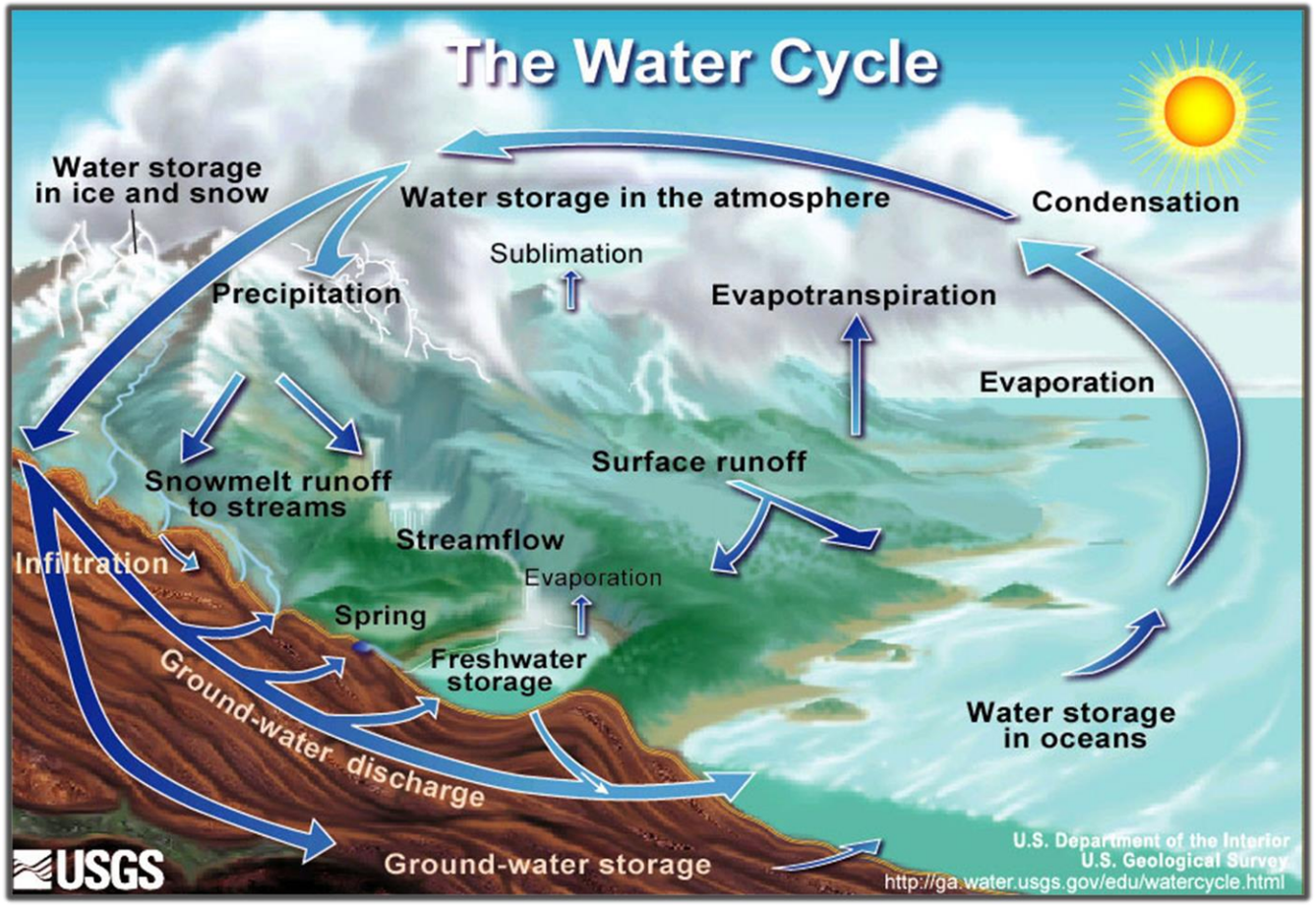


Image by United States Geological Survey (USGS), Washington, D.C.

Data acquisition requirements

Need to understand

How water gets in to the system
Once it gets in there, how is it stored
How it gets out

Minimal model inputs

Precipitation
Pipe network geometry and connectivity
Subsurface infiltration properties
Inflows (street runoff, drains)
Land cover (trees and plants)

Model predictions

Stormwater conveyance system outflows (cfs)
Water depths

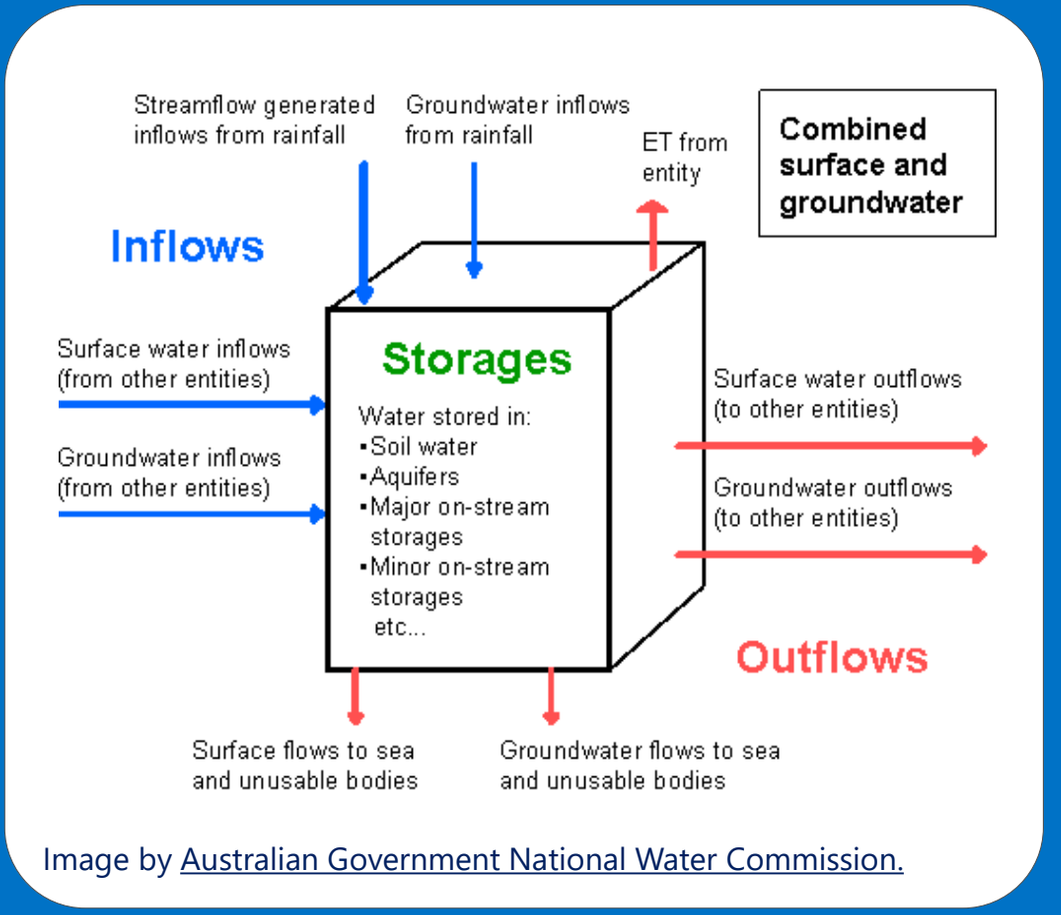


Image by Australian Government National Water Commission.

Things that can mess up a basic model

Photo by Keith Mountain, Creative Commons

The real world is really messy

Where do we put sensors to avoid ending up with a model that does not match reality?

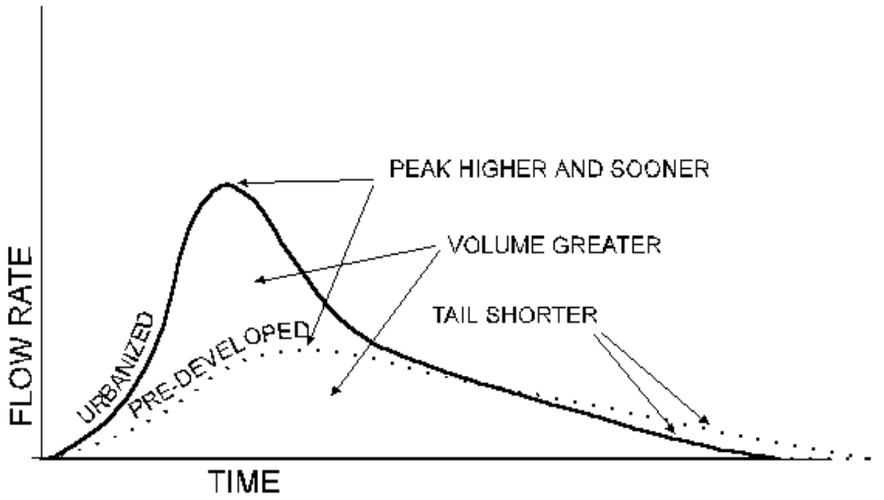


Image by Environmental Protection Agency (EPA), Washington, D.C.

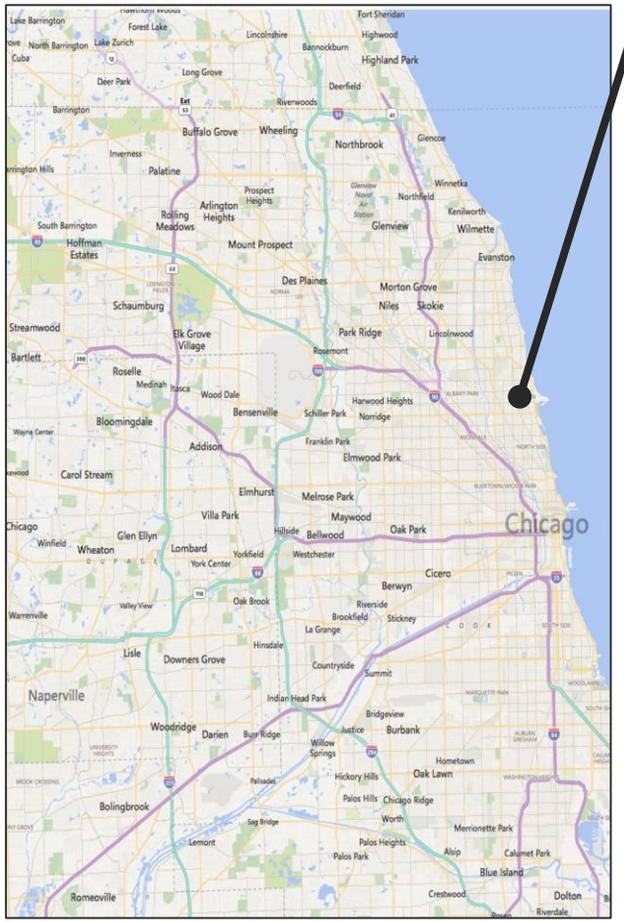
What if we go the other way

Image by City of Vancouver, “Greenest City Action Plan 2020”

Turning cities green can be just as messy

Where do we put sensors to get a model for an environment that has never existed before?

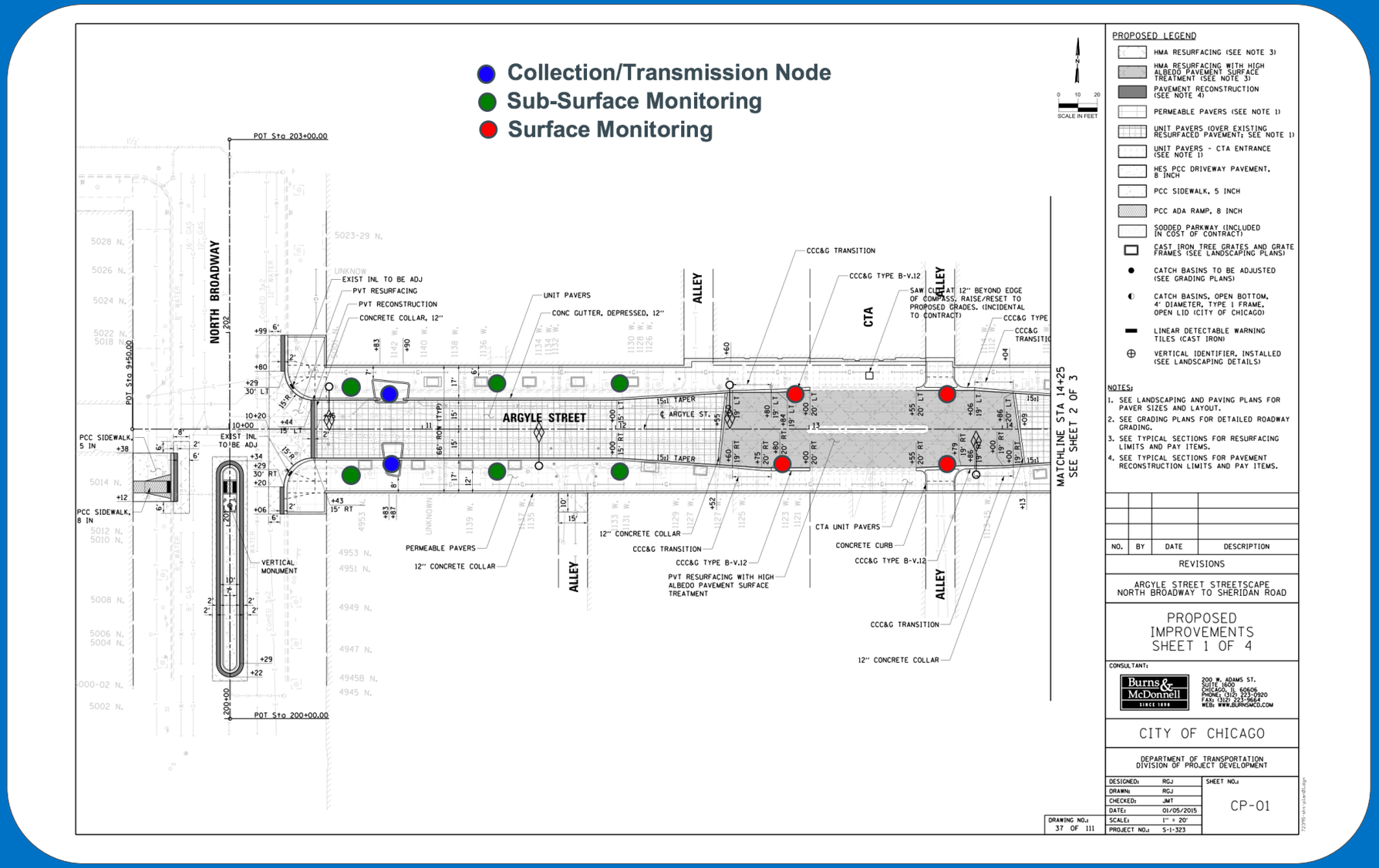
Initial monitoring location example



Commercial
- permeable pavers
- infiltration planters

Street 1

Local streetscape-level sensing



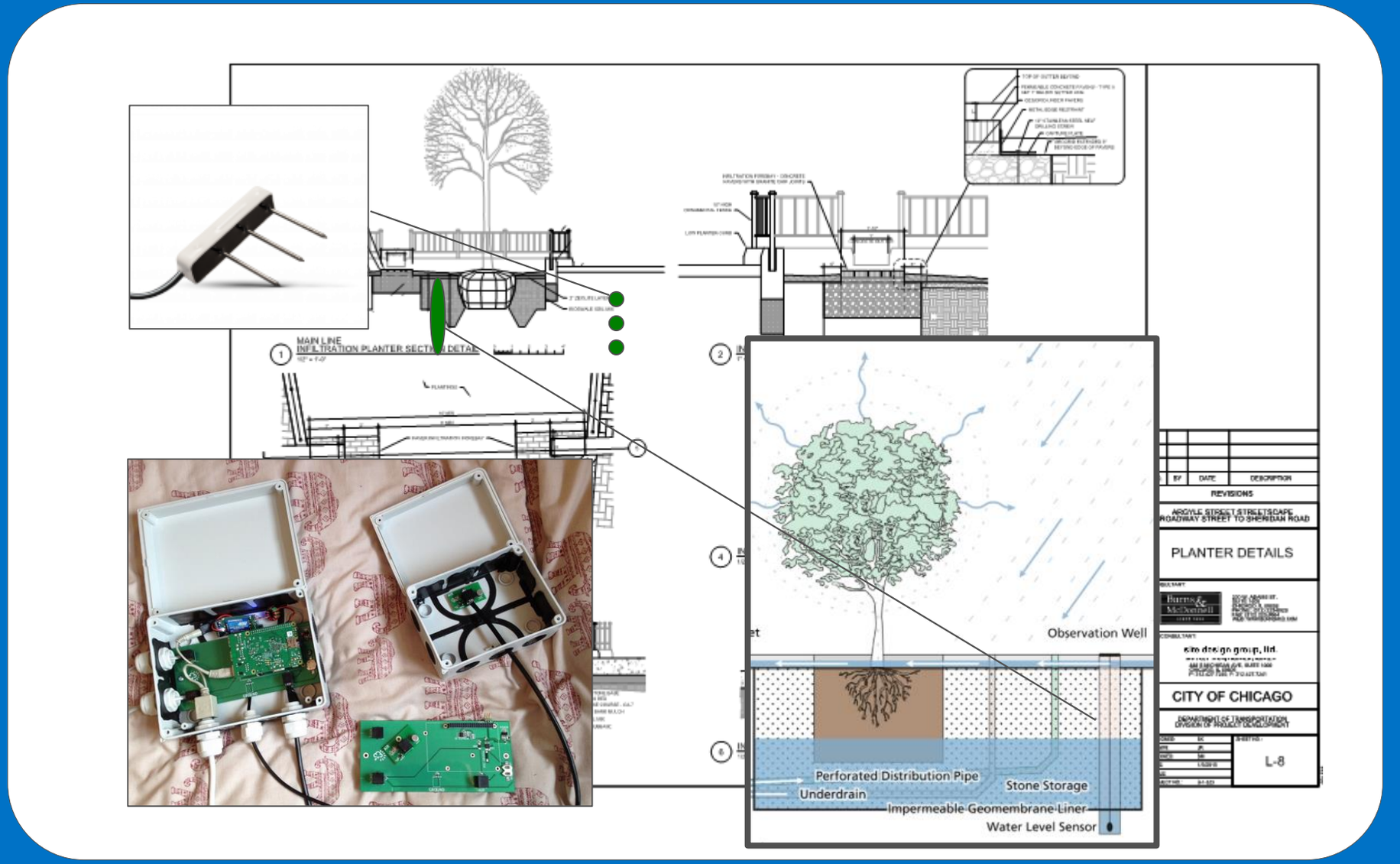
Types of streetscape sensors

Sensor

Decagon GS3

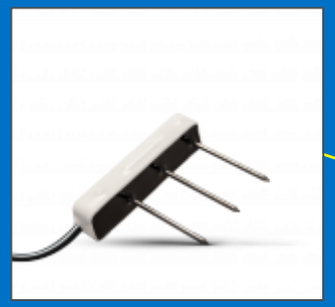
Measures

Soil moisture (θ)
Temperature (T)
Electrical conductivity

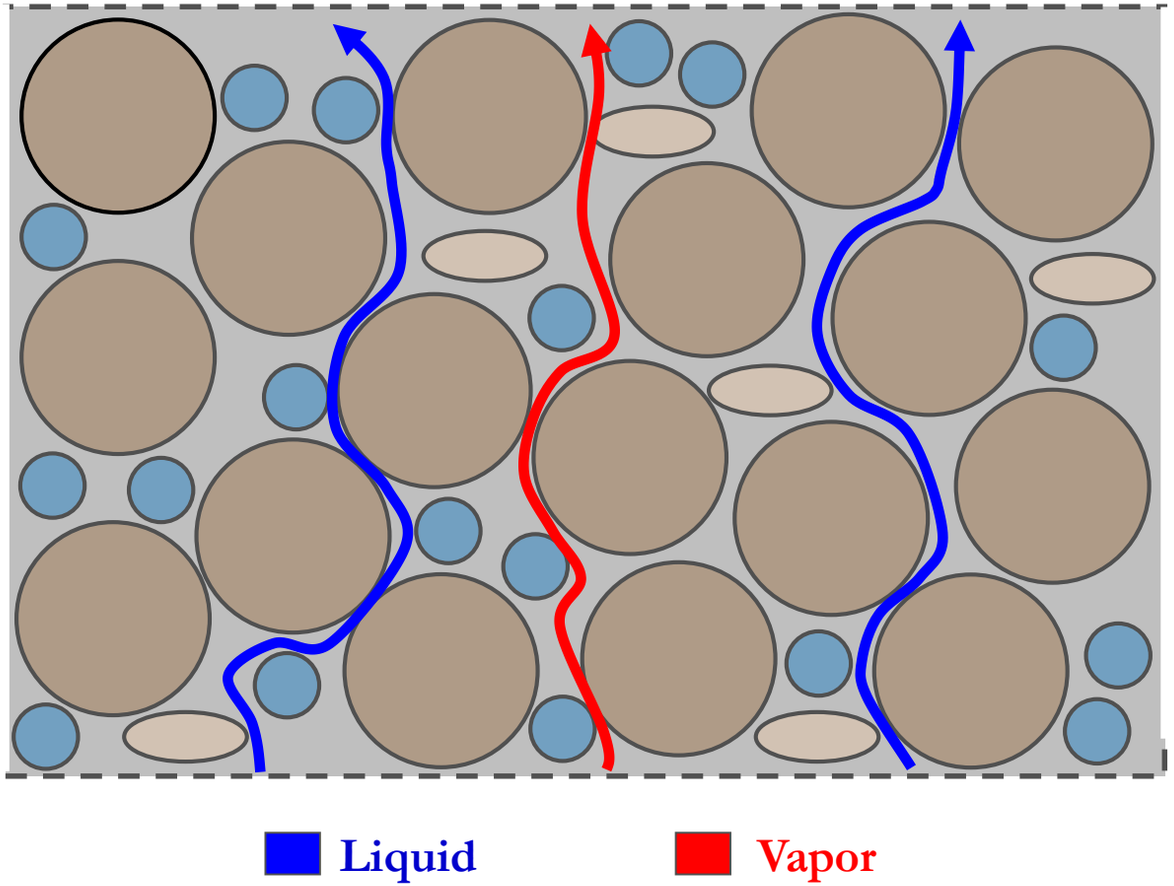


What it really looks like

Installing of 100 sensors started this week
100 sensors sending 1 reading/minute –
4.3M readings/month
25,000 sensors planned within 5 years –
1.1B readings/month



So how do you model dirty?



Isothermal Transport

Liquid Flux

$\theta_2 > \theta_1$

Vapor Flux

Nonisothermal Transport

Liquid Flux

$T_2 > T_1$

Vapor Flux

$$\frac{\partial \theta}{\partial t} = \nabla\left(D_T \nabla T\right) + \nabla\left(D_\theta \nabla \theta\right) + \frac{\partial K}{\partial z}$$

Larger set of streetscape sensors

Data Stream	Category	Monitoring	Data Type	Example Device	Comm Protocol	Transmit Frequency	Synchrony
Precipitation	Hydrology	Continuous	Float	Decagon ECRN-100 High Resolution Rain Gauge	I2C or Serial	1-minute (adaptive)	Synchronous
Humidity	Climate	Continuous	Float	Decagon VP-3	I2C or Serial	1-minute (adaptive)	Synchronous
Temperature	Climate	Continuous	Float	Decagon VP-3 Temp	I2C or Serial	1-minute (adaptive)	Synchronous
Solar Radiation	Climate	Continuous	Float	Decagon VP-3 Solar Radiation	I2C or Serial	1-minute (adaptive)	Synchronous
Wind Direction	Climate	Continuous	Float	Sparkfun SEN-08942	I2C or Serial	1-minute (adaptive)	Synchronous
Wind Speed	Climate	Continuous	Float	Sparkfun SEN-08942	I2C or Serial	1-minute (adaptive)	Synchronous
Soil Moisture	Hydrology	Continuous	Float	Decagon GS3 TDR	I2C or Serial	ideally sub-minute (adaptive)	Synchronous
Soil Temperature	Hydrology	Continuous	Float	Decagon GS3 Soil Temperature	I2C or Serial	Ideally sub-minute (adaptive)	Synchronous
Electrical Conductivity	Water Quality	Continuous	Float	Decagon GS3 Soil EC	I2C or Serial	1-minute (adaptive)	Synchronous
Lysimeters	Hydrology	Continuous	Float	Decagon G3 System	I2C or Serial	1-hour (adaptive)	Asynchronous
Total Suspended Solids	Water Quality	Discrete	Float	Capture for Lab	N/A	1-day	Asynchronous
Total Nitrogen	Water Quality	Discrete	Float	Capture for Lab	N/A	1-day	Asynchronous
Microbial Activity	Water Quality	Discrete	String; Float	Capture for Lab	N/A	1-day	Asynchronous
Phosphorus; COD	Water Quality	Discrete	String; Float	Capture for Lab	N/A	1-day	Asynchronous
Visual Sensing	All	Continuous	Int	Raspberry Pi	I2C or Serial	sub-minute (adaptive)	Synchronous
Sediment Rates	Water Quality	Continuous	Float	Capture for Lab	N/A	1-day	Asynchronous
Pipe Flow	Hydrology	Continuous	Float	TBD	I2C or Serial	sub-minute (adaptive)	Synchronous

There is no take-back once the sensors go in!

Image by Illinois Department of Transportation

Storm Water Management Model (SWMM)

Inputs

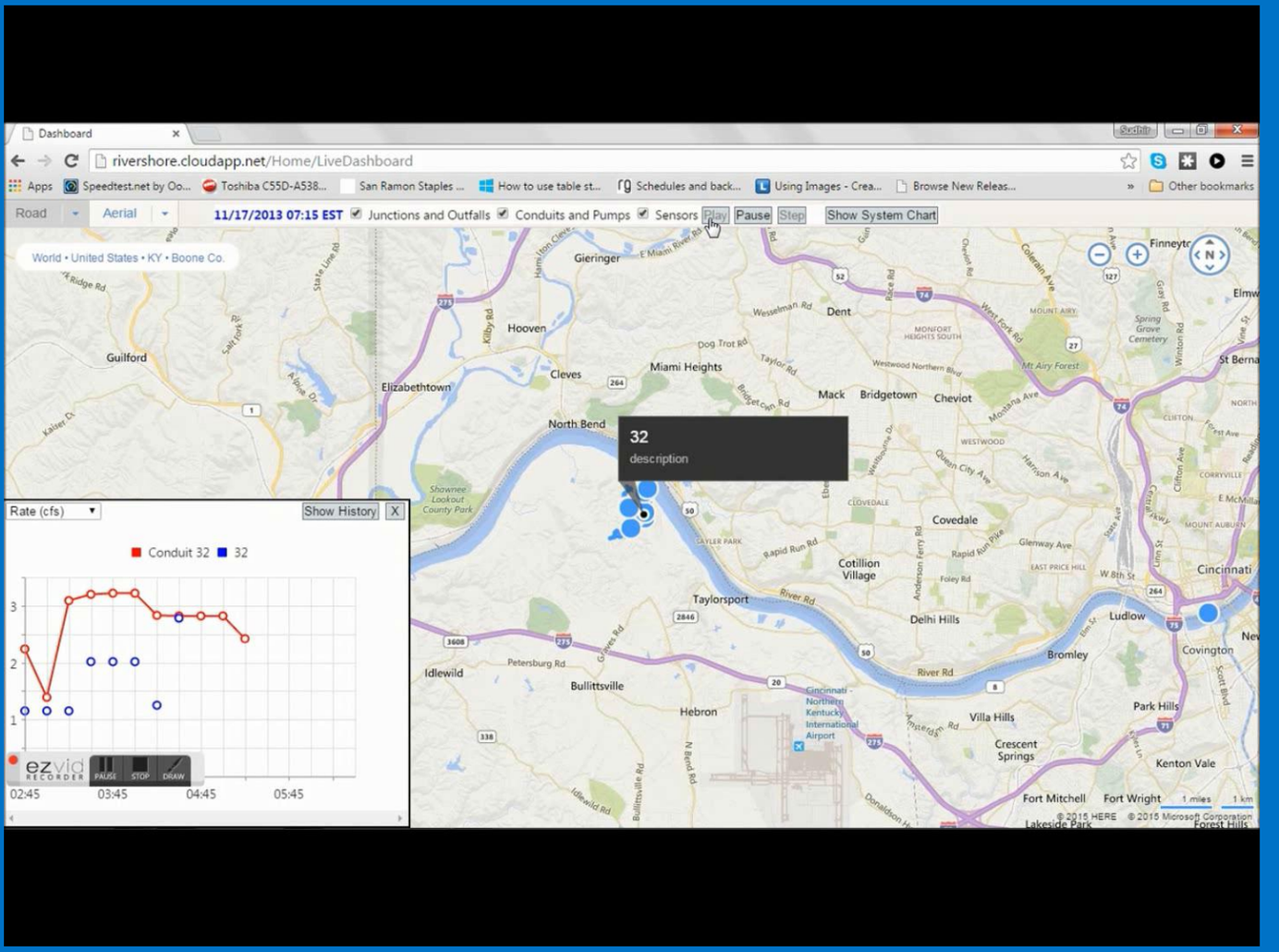
Precipitation
Pipe network geometry and connectivity
Subsurface infiltration properties
Inflows (street runoff, drains)
Land cover (trees and plants)

Prediction

Stormwater outflow (red line)

Actuals

Water flow (blue dots)



State of environmental modelling

Issues with current models such as SWMM

Difficulty of adding new inputs and outputs
Oversimplification - trading statistics for physics
Access – all are desktop applications
Computational speed - may take weeks to run

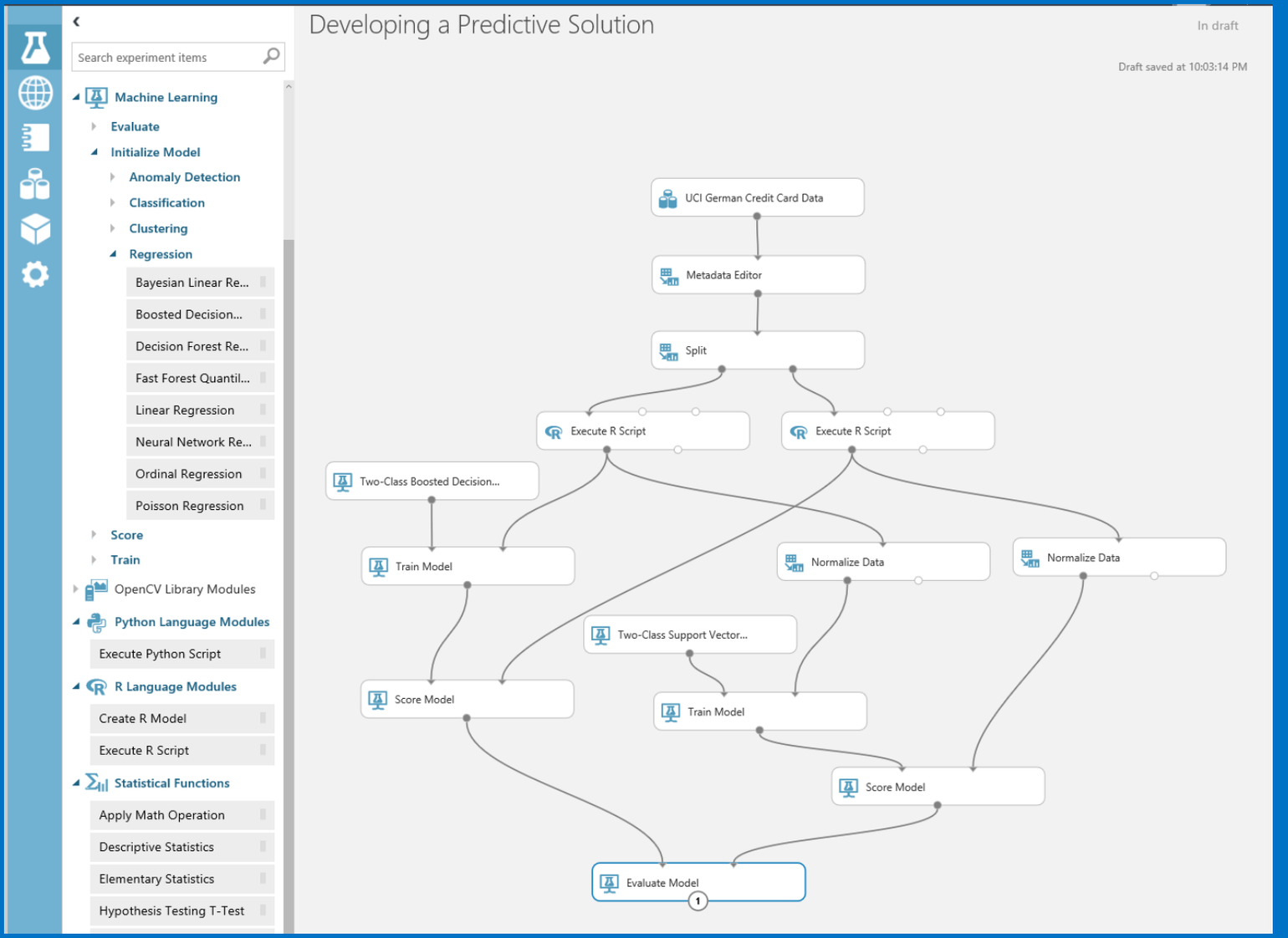
Enhancements

Transitioning multiple monolithic desktop models to published web services
Anomaly detection for hydrologic time series

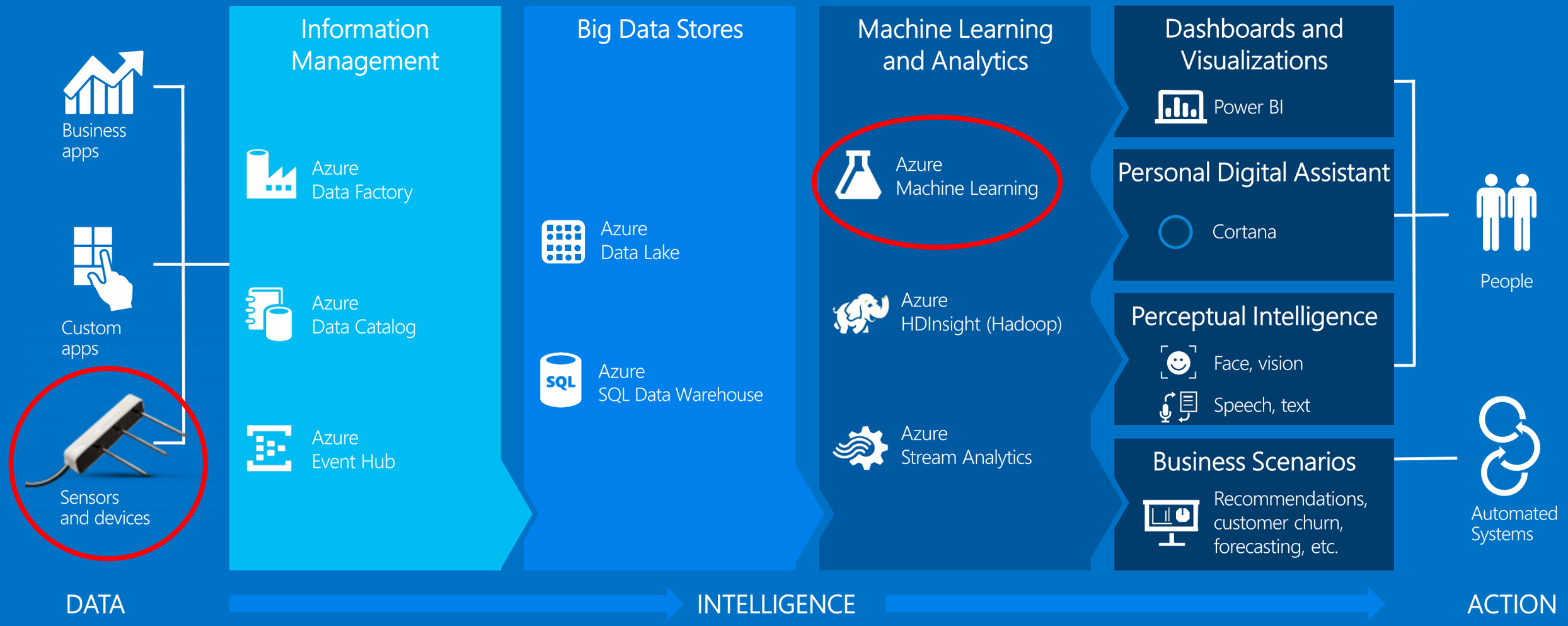
Azure Machine Learning

Features

Data input and output
Data transformations
Machine Learning models
Built-in statistical functions
R script execution
Python script execution
Model training
Model evaluation
Model scoring
Publishing to web service



Cortana Analytics Suite



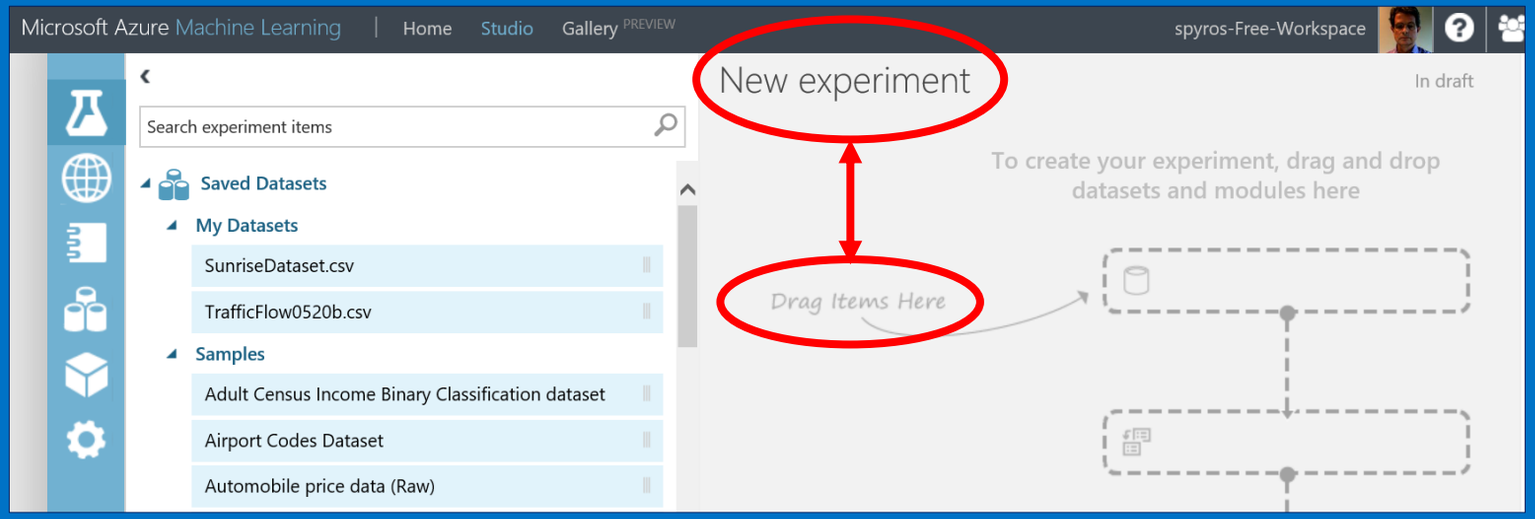
Key takeaways

Domain subject matter knowledge important

Statistical models can be created that don't reflect reality
Need deep domain expertise to define the inputs of the model
Need deep domain expertise to develop the statistical model

Translation

Lots of science involved before we can drag a dataset for a new experiment



Guidance

Question the validity of the dataset

Involve domain expert in design of data acquisition

Keep the Science in Data Science!

Microsoft