



# Azure MachineLearning (ML) und R

Was, wie und warum? Ein Crash-Kurs.

Mario Schnalzenberger



AN ACP GROUP COMPANY

# About me – Mario Schnalzenberger



- Informatiker, Statistiker und Volkswirt
- Forscher (Uni Linz)
  - Forschung im Bereich Gesundheit, Alterung, Pensionen und vielem mehr
  - Veröffentlichungen in Klinischer Forschung, Economics und Econometrics
- Bei cubido unterstütze ich Kunden im Bereich
  - DWH und Business Intelligence
  - Predictives in Richtung Industrie 4.0 und Marketing Intelligence
  - Big Data und verwandten Themen
  - SQL Server, Cubes, MDX, R, C#, SAP Infinite Insights, MCSA, MCSE BI, uvm.
- [m.schnalzenberger@cubido.at](mailto:m.schnalzenberger@cubido.at)

# Agenda

Was?  
Big Data?

Wie?  
Demo

Warum?  
Use Cases

Ausblick!  
Mit Cubido

Was?

Big Data - nur größer



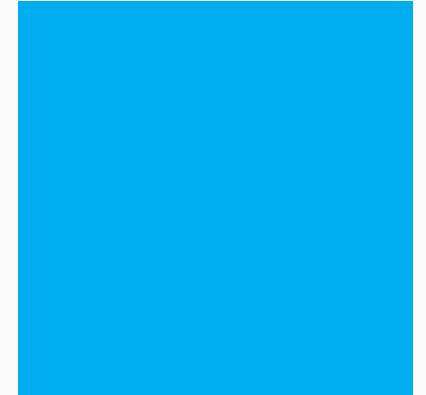
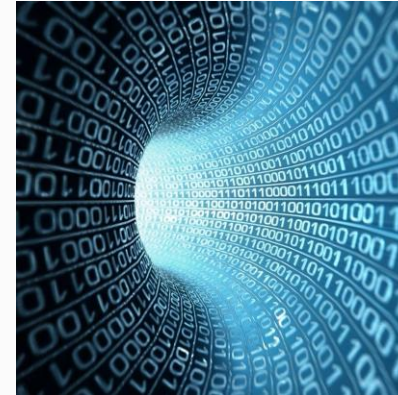
AN ACP GROUP COMPANY



# Sammeln und Aufbereiten – in der Cloud?

## Big Data

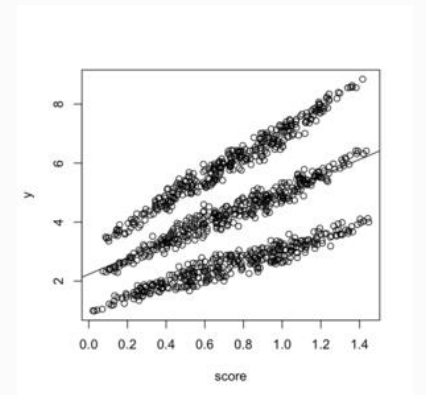
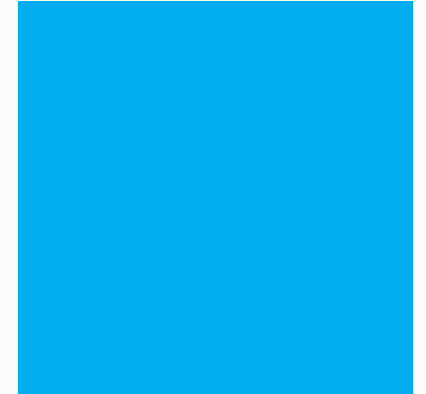
- Entsteht durch das Sammeln von Daten  
(Cloud als einfacher Partner zum Scale out)
- Sammeln und das IoT ist erst im entstehen  
(Cloud - anpassungsfähig und hochverfügbar)
- Aufbereitung der Daten für die Analyse bedarf  
(kurzfristig) hoher Kapazitäten  
(Hadoop o.ä. Technologien, Cloud?)
- Analyse und Synthese mit Rechenkapazitäten



# Analytics – Das haben wir doch schon?

## Analytics?

- Berichte, die descriptive Analytics, meist keine statistische Analyse, *Daten haben keinen zusätzlichen Wert, wenn Sie niemand sieht*
- Advanced Analytics – Statistik schon im frühen Stadium. Vieles ist möglich.
  - Automatisierung
  - Fokussierung
  - Interaktion





# Agility – Prozesse verändern



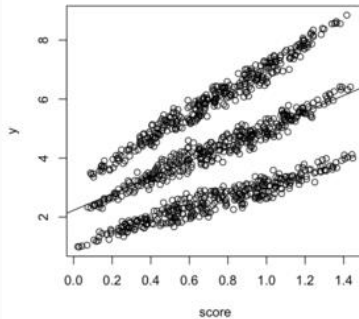
**Vorzeitig mit  
Analyse beginnen**



**Analytiker /  
Statistiker frühzeitig  
einbinden**



**Sammeln –  
Daten und Visionen!**



**Cloud als einfachen  
Partner sehen**



**Langfristig die  
Prozesse optimieren  
und aus eigenen  
Daten lernen**

# Wie – eine Quasi-Demo



AN ACP GROUP COMPANY

# Schon mal geflogen?

- Welchen Flug?
- Welche Airline?
- Welcher Flughafen?
- Welche Uhrzeit?
- Welche Anschlussflüge?



	11:16 A	CANCELLED
5A	10:30 A	CANCELLED
5A	10:15 A	CANCELLED
7A	6:50 A	DELAYED
7A	7:20 A	DELAYED
	10:00 A	CANCELLED
17A	10:10 A	DELAYED

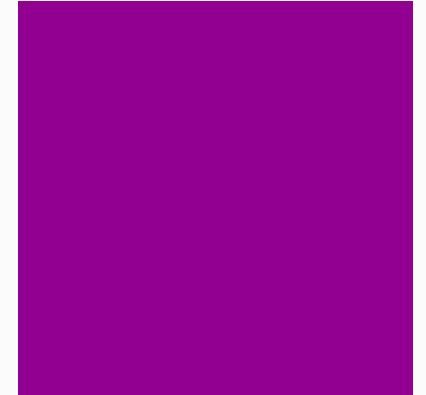
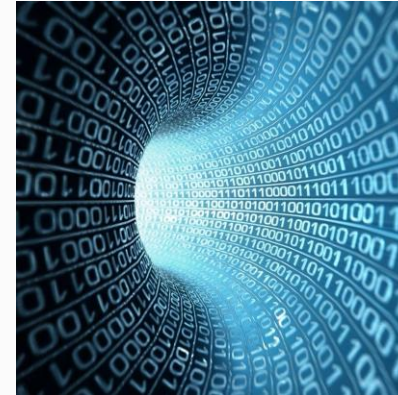
# Konzeptuelle Voraussetzungen

## Was brauchen wir? Womit arbeiten wir?

- Daten
- Experimente (Modelle)
- Web Services
- (Microsoft Account)

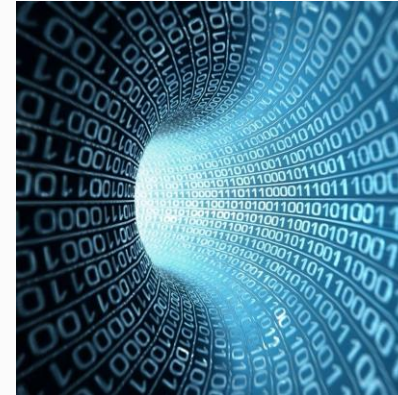
## Werkzeuge

- Azure Portal und ML-Workspaces
- ML Studio



# ML Studio

- GUI mit vielen Funktionen (Copy/Paste)
- Daten für Experimente hochladen
- Experimente (Modelle) entwickeln und prüfen (validieren)
- Web Services veröffentlichen und freigeben



# Machine Learning in Azure Portal

Microsoft Azure | GUTHABENSTATUS

machine learning **VORSCHAU**

NAME	STATUS	BESITZER	ABONNEMENT	SPEICHERORT
MLTest	✓ Online	m.schnalzenberger@cub...	Visual Studio Premium...	USA (Mitte/Süden)

**1** MACHINE LEARNING

**2** + NEU

**3**

SCHLÜSSEL VERWALTEN | IN STUDIO ÖFFNEN | LÖSCHEN

1 ?

**UBIDO**  
AN ACP GROUP COMPANY

# MachineLearning (ML) Studio

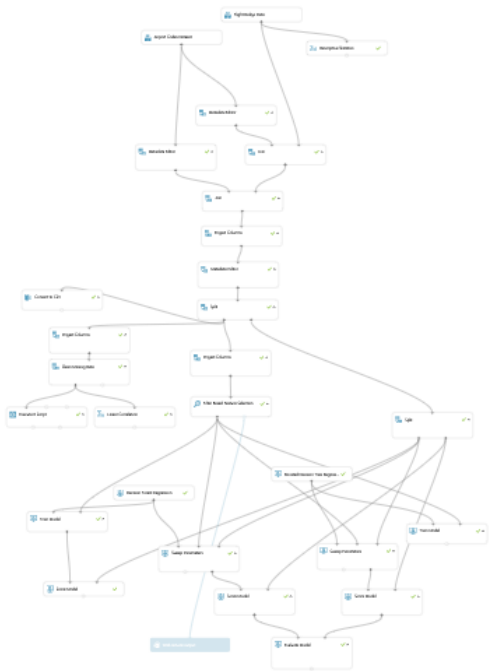


- EXPERIMENTS
- WEB SERVICES
- SETTINGS

## experiments

ALL EXPERIMENTS SAMPLES

	NAME	AUTHOR	STATUS	LAST...↓	🔍
<input checked="" type="checkbox"/>	My First Fligh...	m.schnalzenb...	Draft	2/12/2015 8:1...	
<input type="checkbox"/>	WebService	m.schnalzenb...	Finished	2/11/2015 3:0...	
<input type="checkbox"/>	MyWebService	m.schnalzenb...	Finished	2/11/2015 12:...	



+ NEW

🗑 DELETE

# Experiments

Microsoft Azure Machine Learning | Home Studio

## My First Flight Delay Experiment

In draft  
Draft saved at 08:43:40

Search experiment items

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

**1**

**2** Flight Delays Data

**3** Airport Codes Dataset

**4** Metadata Editor  
Origin Airport Information

**5** Properties  
**Airport Codes Dataset**  
SUBMITTED BY: Microsoft C...  
SIZE: 16.6 KB  
FORMAT: GenericCSV  
CREATED ON: 2/7/2015 6...  
[View dataset](#)

**6** RUN

VIEW RUN HISTORY | SAVE | SAVE AS | DISCARD CHANGES | REFRESH | CANCEL | PUBLISH WEB SERVICE | CREATE SCORING EXPERIMENT



# Experiments

The screenshot displays the Microsoft Azure Machine Learning Studio interface for an experiment titled "My First Flight Delay Experiment". The interface includes a left-hand navigation pane, a central workspace with a data flow diagram, and a right-hand "Properties" panel.

**Navigation Pane (Annotation 1):** A purple circle highlights the "Metadata Editor" option under the "Data Transformation" > "Manipulation" menu.

**Workflow Diagram (Annotations 2, 3, 4, 5, 6):** The diagram shows a sequence of operations:

- Annotation 2:** A purple circle highlights the "Metadata Editor" node for "Origin Airport Information".
- Annotation 3:** A large purple circle highlights the "Metadata Editor" node for "Origin Airport Information" and the "Properties" panel on the right.
- Annotation 4:** A purple circle highlights the "Join" node for "Join Origin Airport Data".
- Annotation 5:** A purple circle highlights the "Join" node for "Join Destinating Airport Information".
- Annotation 6:** A purple circle highlights the "Join" node for "Join Destinating Airport Information".

**Properties Panel (Annotation 3):** The "Metadata Editor" properties are shown, including:

- Column:** Selected columns: city, state, name
- Column names:** city, state, name
- Launch column selector:** Button
- Data type:** Unchanged
- Categorical:** Unchanged
- Fields:** Unchanged
- New column names:** OriginCity, OriginState, O
- START TIME:** 2/11/20...
- END TIME:** 2/11/20...
- ELAPSED TIME:** 0:00:00.0...
- STATUS CODE:** Finished
- STATUS DETAILS:** Task output was present in

**Bottom Bar:** Contains standard application controls: NEW, VIEW RUN HISTORY, SAVE, SAVE AS, DISCARD CHANGES, REFRESH, CANCEL, RUN, PUBLISH WEB SERVICE, and CREATE SCORING EXPERIMENT.

# Visualize the Sources and Outputs

My First Flight Delay Experiment > Score Model > Scored dataset

rows 1550068  
columns 18

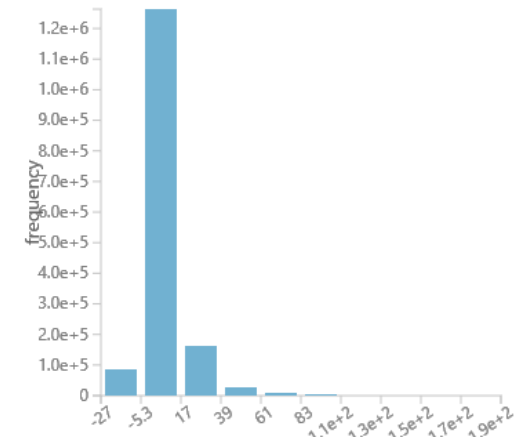
Month	DayofMonth	DayOfWeek	Carrier	CRSDepTime	DepDelay	DepDel15	CRSArrTime	ArrDelay	ArrDel15	Cancelled	Scored Label Mean	Scored Label Standard Deviation
26	7	DL	1340	-4	0	1457	-8	0	0	0	-4.32455	19.675623
18	5	US	1800	-8	0	1915	-8	0	0	0	-2.916969	18.722552
25	7	OO	627	22	1	723	17	1	0	0	-1.520425	17.27603
23	7	MQ	715	-9	0	950	-16	0	0	0	8.044899	49.963153
15	2	WN	1700	10	0	1800	11	0	0	0	-1.439296	20.535412
3	5	WN	840	-5	0	935	-8	0	0	0	0.142071	22.785005
25	2	VX	1840	363	1	2150	356	1	0	0	90.110313	106.616985
11	4	WN	1500	8	0	1510	17	1	0	0	8.548382	30.016789
19	3	UA	935	0	0	1131	16	1	0	0	0.855321	26.883996
13	2	DL	2125	1	0	544	-1	0	0	0	-2.41632	23.844859
24	4	WN	1540	18	1	1705	12	0	0	0	11.497718	29.919001
17	1	WN	1510	-1	0	1900	-17	0	0	0	19.854124	43.148306
19	5	WN	1710	-4	0	1830	-10	0	0	0	10.893888	34.347628
24	3	DL	1030	7	0	1405	-1	0	0	0	1.16815	28.306764
20	2	AA	600	-3	0	740	-19	0	0	0	-3.679286	19.488942
8	6	AA	930	-2	0	1155	-10	0	0	0	-1.851285	22.558995
11	2	UA	1259	117	1	2130	97	1	0	0	15.173034	42.938912
26	4	EV	600	-6	0	716	2	0	0	0	-3.64327	26.246258

## Statistics

Mean	6.4836
Median	3.6408
Min	-27.3379
Max	193.2833
Standard Deviation	12.252
Unique Values	1462406
Missing Values	0
Feature Type	Numeric Score

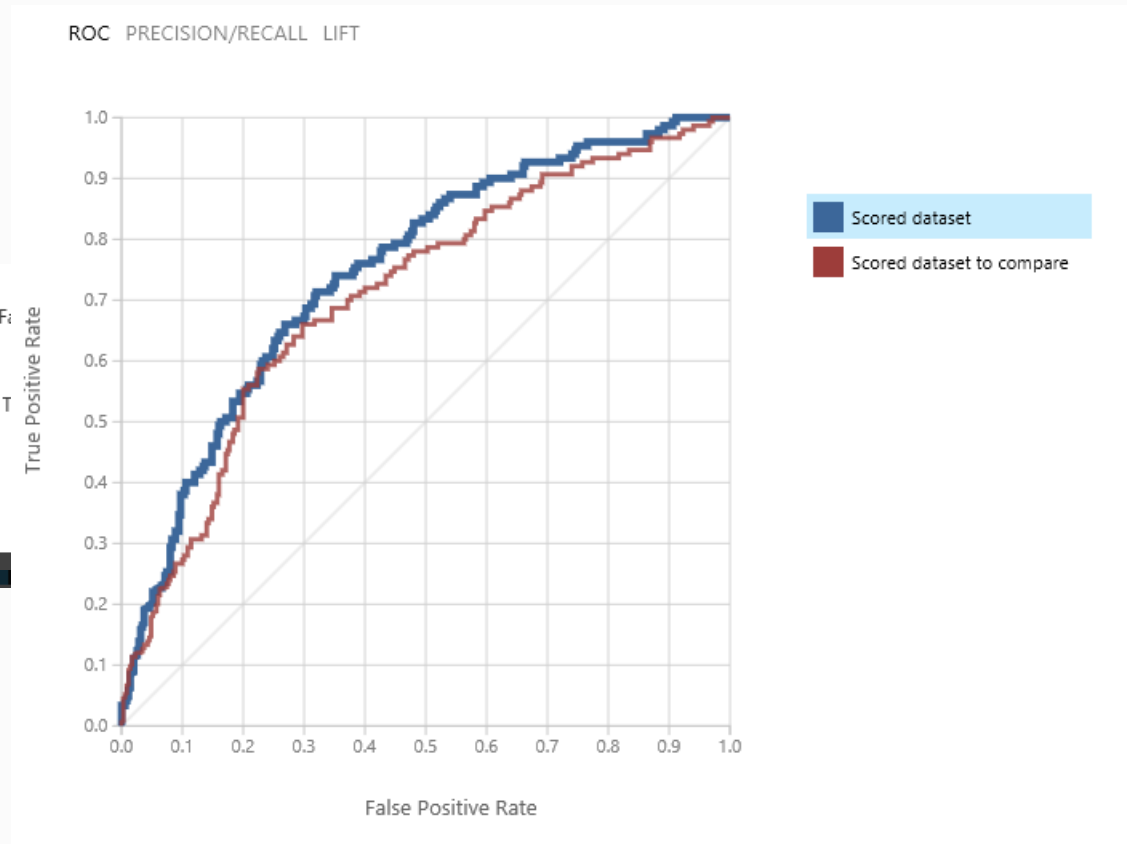
## Visualizations

Scored Label Mean  
Histogram



# Visualize Results – Ergebnisse anzeigen

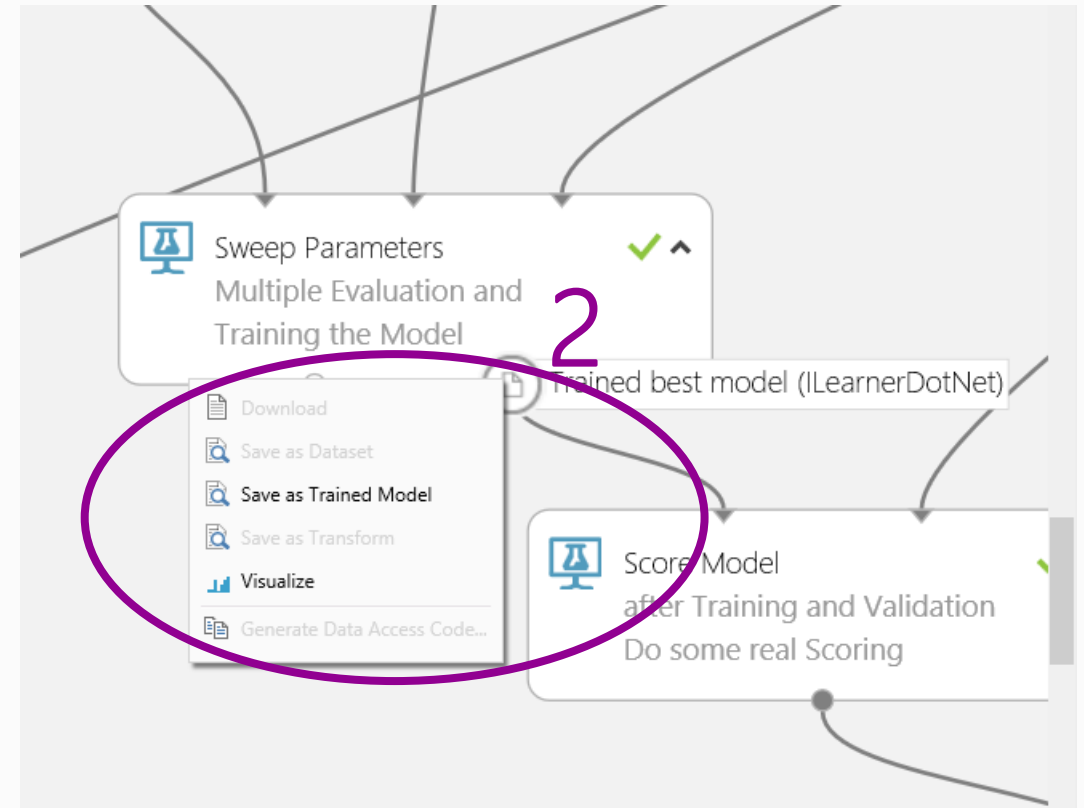
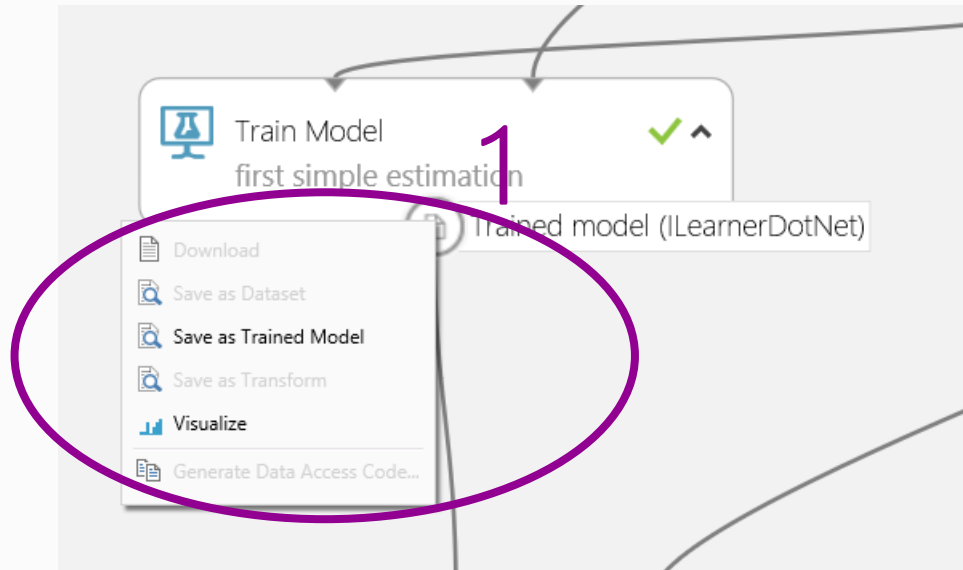
Score Bin	Position	Score Bin	Position
True Positive	11382	True Positive	21425
False Negative	64	False Negative	64
False Positive	23450	False Positive	53177
True Negative	324	True Negative	324



AUC  
0.36

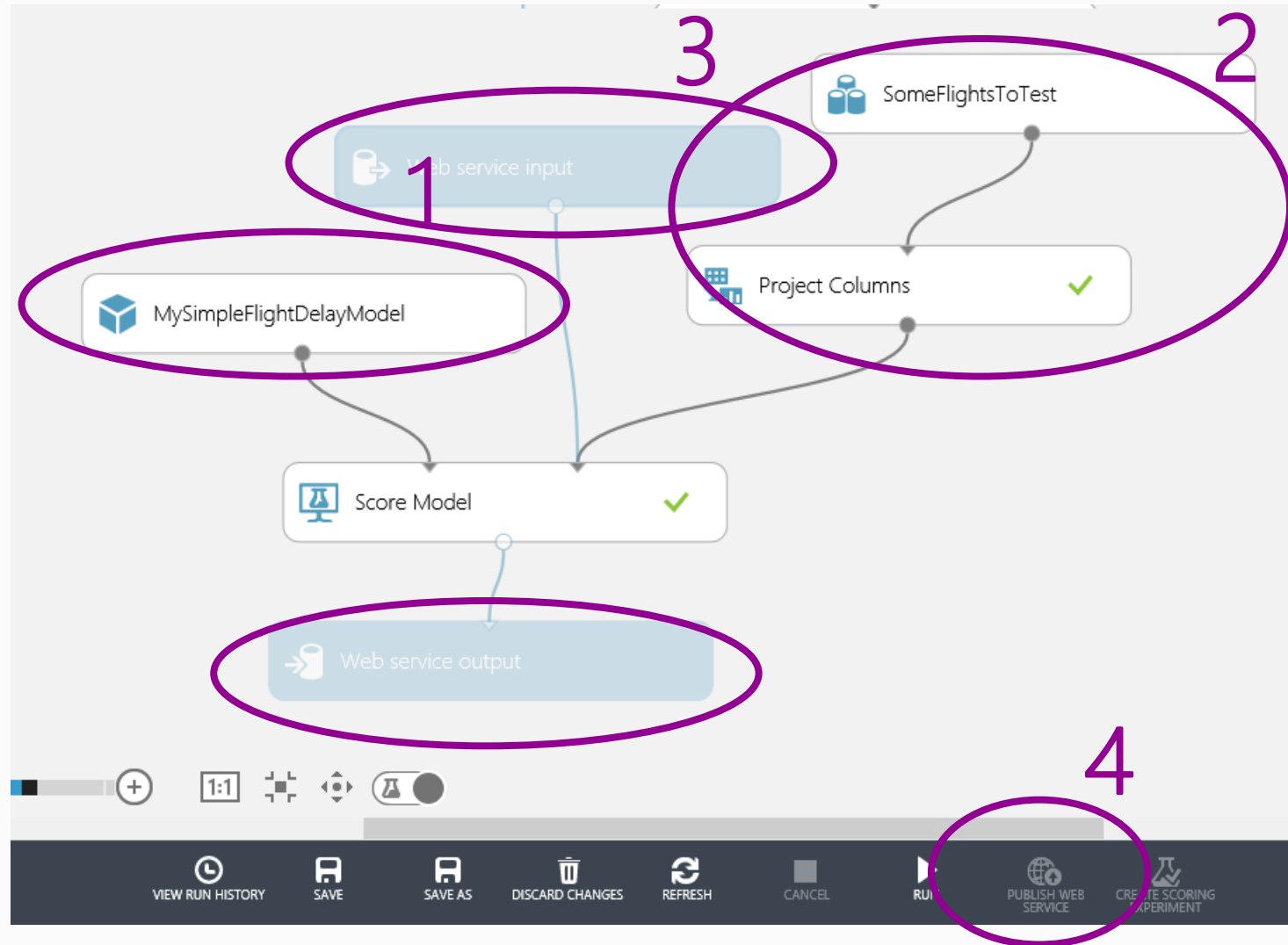
Negative Precision  
Relative AUC

# Training-Ergebnisse speichern



# „Deploy a Webservice“

Ein neues Experiment:  
Experiment:



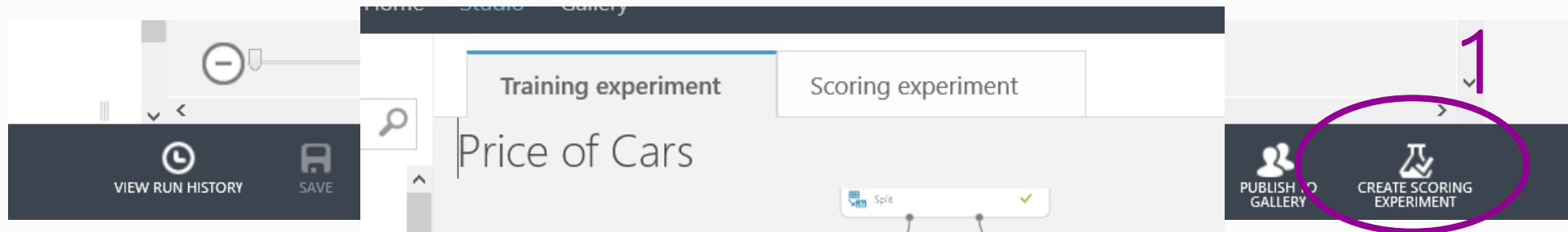
# MachineLearning (ML) Studio

The screenshot displays the Machine Learning Studio interface for a web service named "webservice". The left sidebar contains navigation options: "DASHBOARD" and "CONFIGURATION", with "WebService" selected. The main area shows a "Test WebService Service" dialog box with the prompt "Enter data to predict" and a text input field labeled "DESTCITY".

A dark notification banner at the bottom of the main area displays the test result: `'WebService' test returned ["Las Vegas","NV","Atlanta","GA","2013","5","20","7","US","1000",-4.03629446029663"]`. Below this, a green checkmark indicates a successful result with a JSON structure: `Result: {"Results":{"output1":{"type":"table","value":{"ColumnNames":["DestCity","DestState","OriginCity","OriginState","Year","Month","DayofMonth","DayOfWeek","City","CPSDepTime","Scored Labels"],"ColumnTypes":["String","String","String","String","Int32","Int32","Int32","Int32","String","Int32","Double"],"Values:[["Las Vegas","NV","Atlanta","GA","2013","5","20","7","US","1000",-4.03629446029663]]}}}}`. The value `-4.03629446029663` is circled in purple.

Below the notification banner, the "BATCH EXECUTION" section is visible, featuring a "NEW" button, a "DELETE" button, and a "YEAR" input field with a dropdown arrow and a checkmark icon.

# Jetzt noch einfacher



# Was ist R?

- Freie Programmiersprache für statistisches Rechnen
- Viele Möglichkeiten auch grafisch auszuwerten
- Interpreter
- Für komplexe Modelle die so noch nicht in Azure enthalten sind

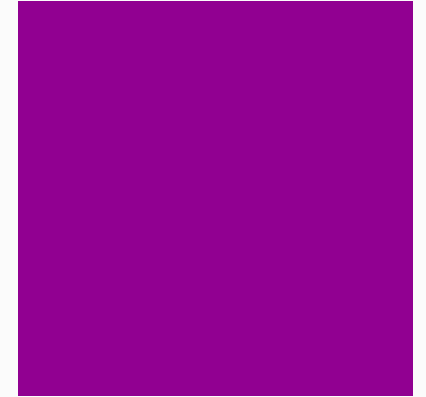


# Beispiel und eine (kleine) Demo

Autopreise

R

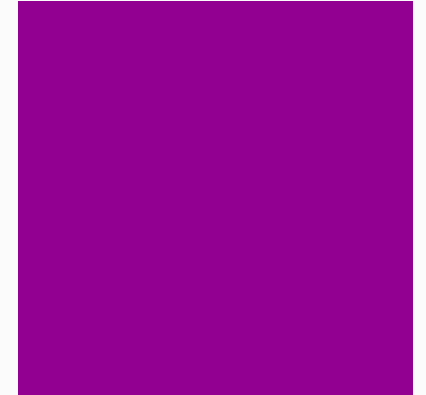
Und?



# Experimenteditor – Best Practice

## Design einer Analyse

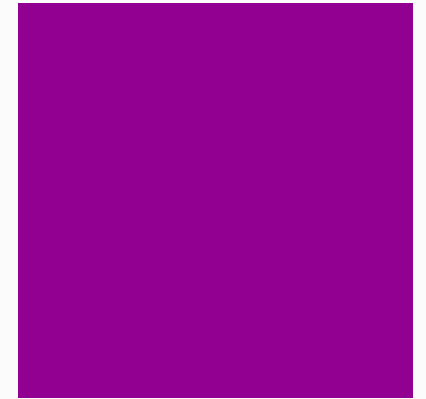
1. (Roh-)Daten zusammen aufbereiten (viel Arbeit!)
2. Daten ins Azure hochladen (direkt im Editor)
3. Analyse Designen (Statistik)
4. Verschiedene Methoden vergleichen
5. Bestes Modell für weitere Verwendung "speichern"



# Offene Punkte

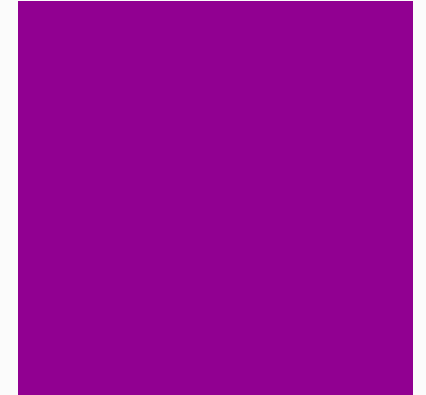
## Was würde ich mir wünschen ...

1. Ergebnisse im Detail anzeigen
2. Analyse von Modelle im Detail ermöglichen (Statistik Tools)
3. Vergleich Modelle von ML mit R ?
4. Modelle analysieren alle Covariate im Dataset (complex)
5. Predictions von Modellen mit mehr als "nur" Score und SD ...
6. Mehr Pakete für R



# Bitte fundierte statistische Kenntnisse

- Plausibilität
- mathematische Verteilungsaspekte
- mögliche richtige Modelle
- Modelle gegeneinander rechnen lassen
- Modellauswahl treffen

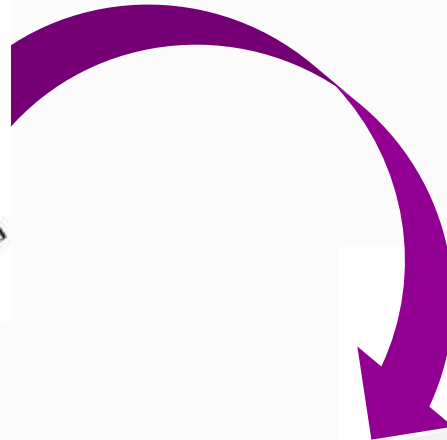


# Warum? – Use Cases

**CUBIDO**

AN ACP GROUP COMPANY

# Data Driven Business



# Use Cases

## Problem und Glaubhaftigkeit

Alle Daten können analysiert werden, daher können wir alles machen?

- Marketing verbessern (bis in einzelne Details)
- Produktion überwachen (Maintenance)
- Qualität überwachen, Fehler verhindern
- Adaptive Lagerhaltung/Logistik (Wetterdaten für Eisladen)
- Recommender
- Churn Analysis (Kundenabgang verhindern)



# Ausblick – Mit Cubido



AN ACP GROUP COMPANY



# Business Intelligence



# Individuelle Softwareentwicklung



# Mobile Solutions



Fragen?

