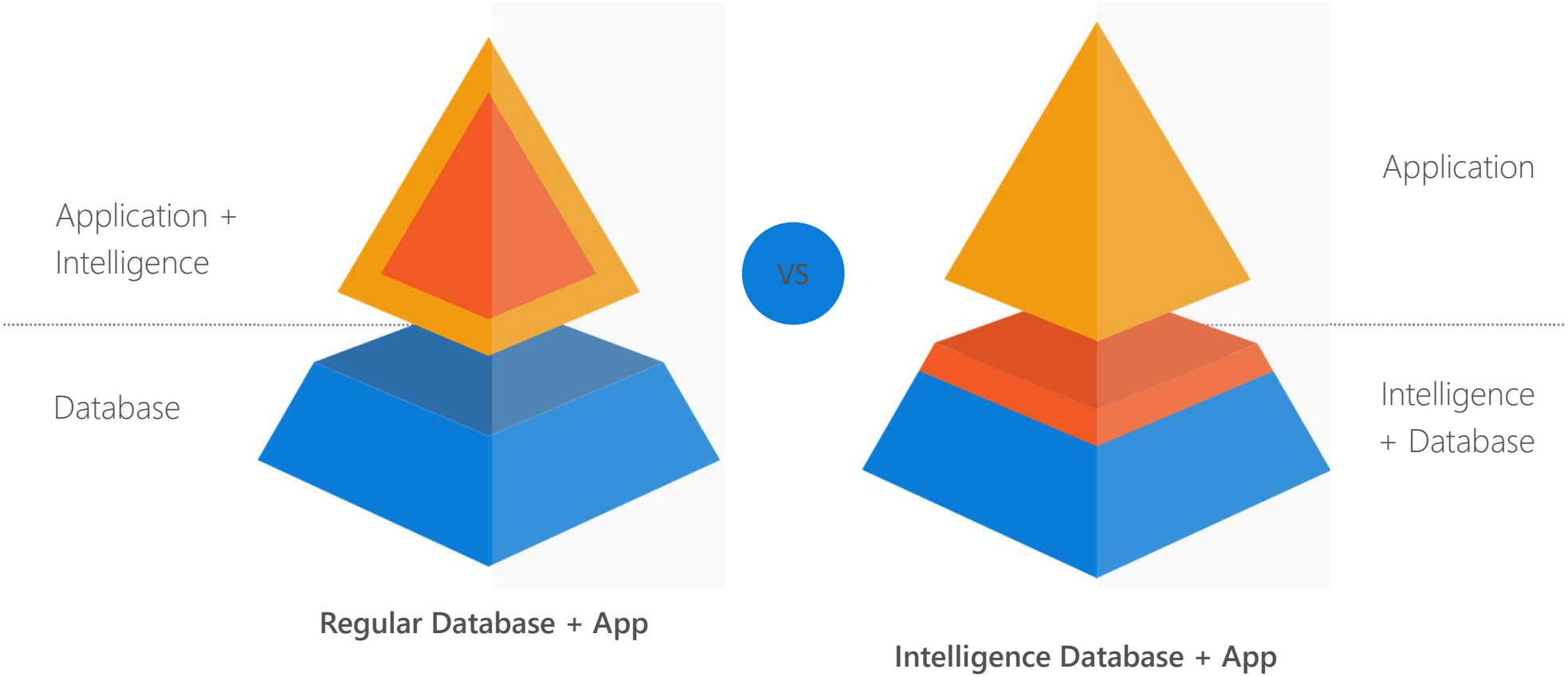
SQL Server Machine Learning Services

In-Database Machine Learning in SQL Server

Agenda

- Why ML In-Db
- How does it work?
- Security

Bringing Intelligence to where Data Lives

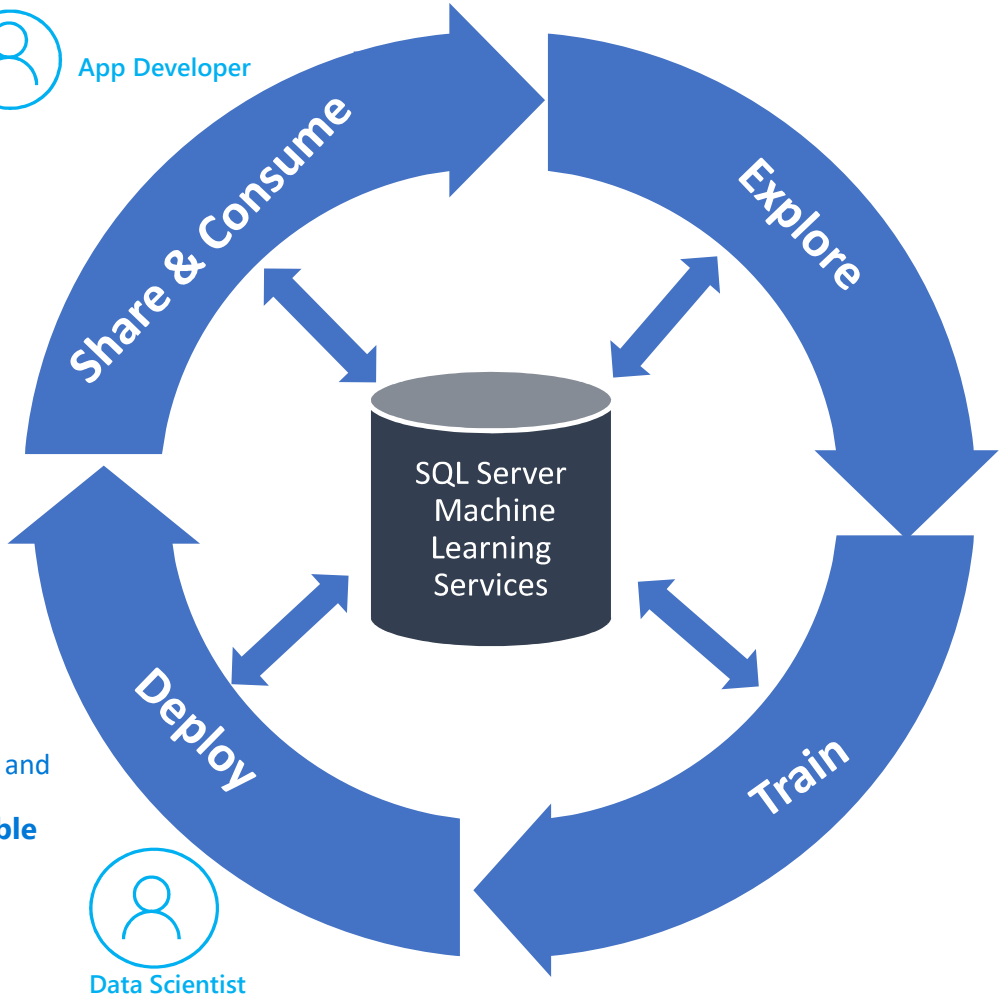


http://nielsberglund.com

Why In-database ML with SQL Server?

- Better Collaboration and Insights Sharing
- Streamlined Deployment of R/Python Scripts and Models
- Faster Time to Insights
- Better Security and Compliance

Better Collaboration and Insights Sharing



App Developer

Make apps intelligent - Consume machine learning models from any app with a **simple stored proc call**

Explore and experiment with SQL Server data in your favorite IDE - use R, Python, T-SQL.

Data Scientist

SQL Developer/DBA

Deploy and schedule training and prediction scripts in-DB - make models and predictions **accessible to any application**

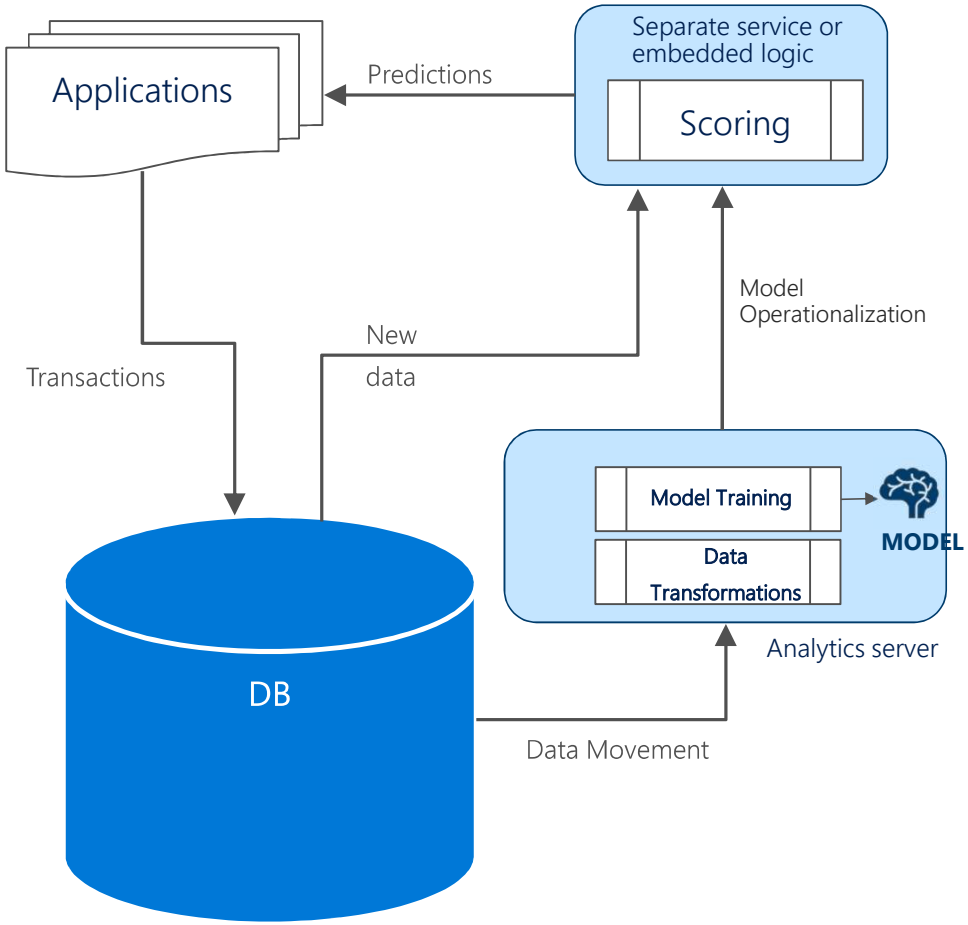
Manage model version in DB

Train models remotely on SQL Server from your IDE

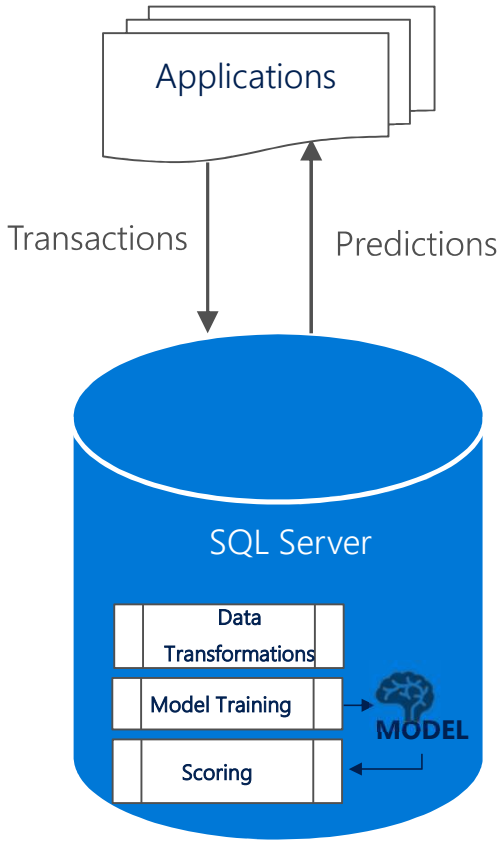
Train models using R/Python code embedded in T-SQL stored proc.

Data Scientist

Streamline Productivity and Simplify Deployment

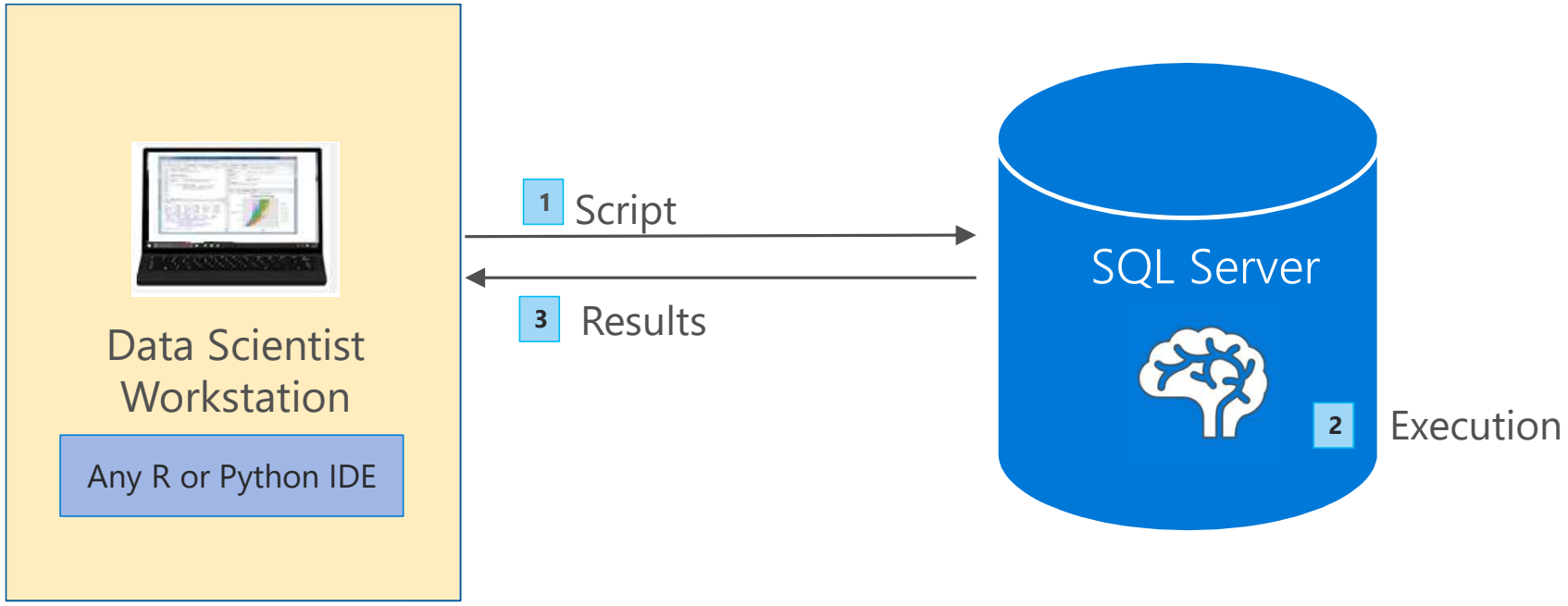


Machine Learning outside of DB



In-DB Machine Learning

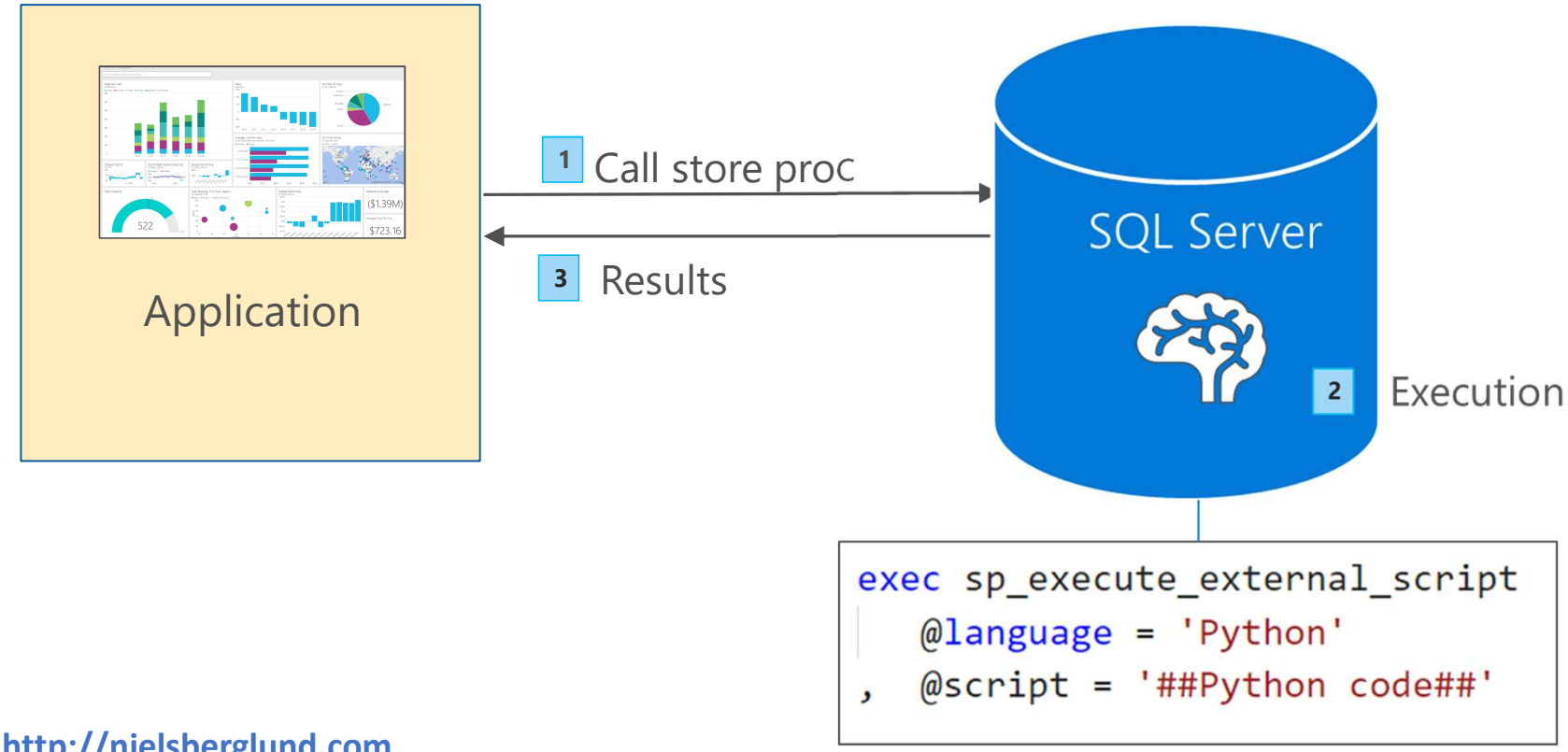
Data Scientists - Data Exploration and Model Development



Value for Data Scientists

- Work with full datasets (not samples) without moving data out of secure, compliant boundaries
- Work from your favorite IDE; remotely leverage the power of SQL Server
- Work with ANY open source package in-database
- Leverage scalable, fast MS algorithms and pre-trained models when needed – combine the best of open source and MS options
- Streamlined model deployment without developer dependence
- Model version management in-database

Data & App. Developer - Model Deployment & Consumption



Value for Data and Application Developers

- App Developers – make existing and new apps intelligent
  - Consume models easily by simply calling a T-SQL stored procedure
  - No knowledge of models or model conversion into other languages needed
- Data Engineers and DBAs – leverage the power of R/Python
  - General purpose data processing
  - Create powerful data visualizations.
- DBAs – securely enable their organizations to do machine learning and AI on SQL Server
  - Manage, govern and secure the resources

Faster Time to Insights

- Integration with SQL query execution
  - Parallel query pushing data to multiple external processes / threads
  - Use in-memory technology and Columnstore Indexes alongside your ML scripts
- Streaming mode execution
  - Stream data in batches to the R/Python process to scale beyond available memory
- Train and Predict using parallelism
  - Leverage RevoScaleR/revoscalepy and scale your R and Python scripts using multi-threading and parallel processing
- Native scoring for faster real-time predictions (New in 2017)

ML Services (In-Database) is Secure

Reduced surface area and isolation

- 'external scripts enabled' required
- R/Python script execution outside of SQL Server process space

Script execution requires explicit permission

- sp_execute_external_script requires EXECUTE ANY EXTERNAL SCRIPT for non-admins
- SQL Server login/user required and db/table access

R/Python processes have limited privileges

- R/Python processes run under local user accounts in the SQLRUserGroup
- Each execution is isolated. Different users with different accounts
- Windows firewall rules to block outbound traffic

Run R and Python In-Db

```
1  EXEC sp_execute_external_script
2  @language =N'R',
3  @script=N'
4  OutputDataSet <- InputDataSet;
5  ',
6  @input_data_1 =N'SELECT 1 AS hello'
7  WITH RESULT SETS (([hello] int not null));
8  GO
9
```

RESULTS

	hello
1	1

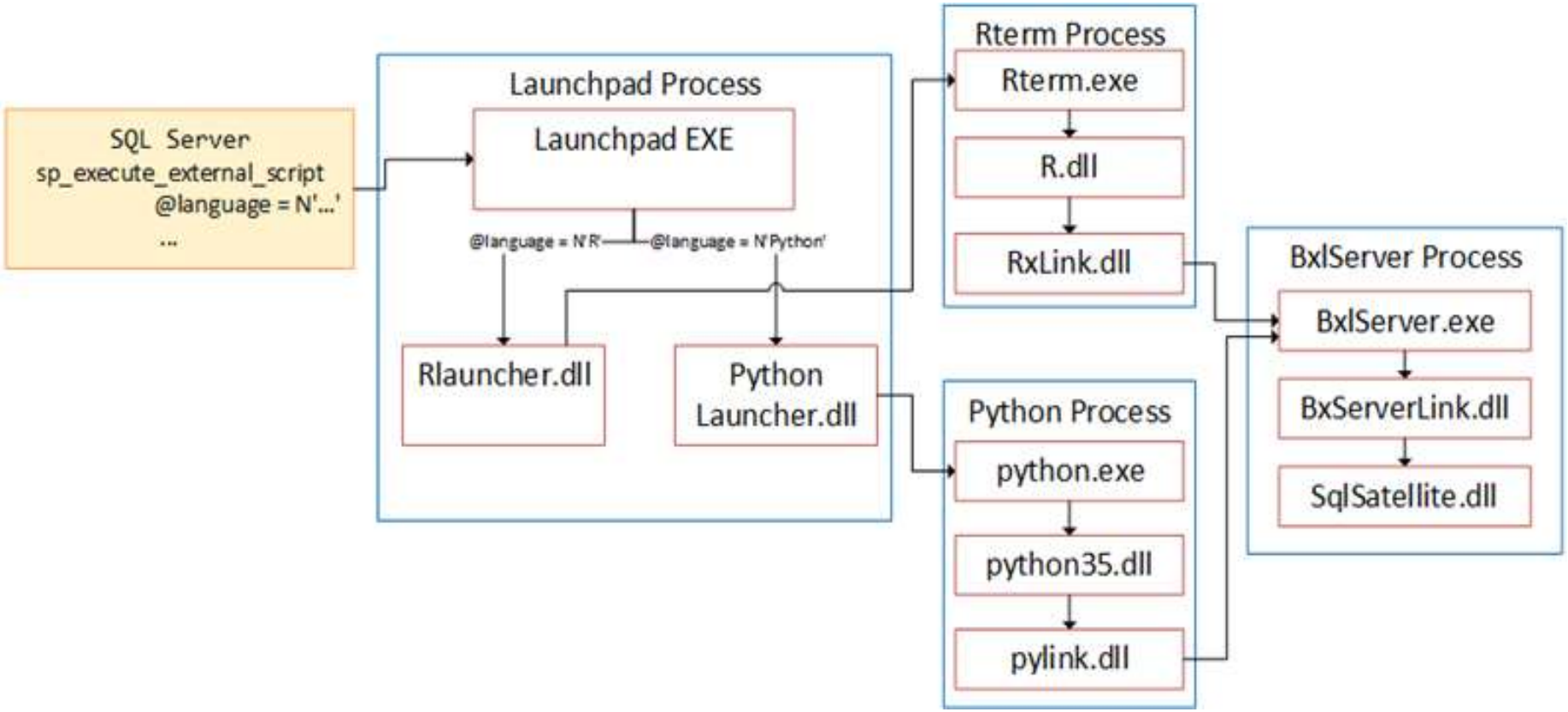
```
1  EXEC sp_execute_external_script
2  @language =N'Python',
3  @script=N'
4  OutputDataSet = InputDataSet;
5  ',
6  @input_data_1 =N'SELECT 1 AS hello'
7  WITH RESULT SETS (([hello] int not null));
8  GO
9
```

RESULTS

	hello
1	1

http://nielsberglund.com

Architecture



http://nielsberglund.com

Roadmap

- Azure SQL Database
  - R support, followed by Python
- ML Services in SQL Server on Linux
- Additional algorithms and pre-trained models
- Failover cluster support
- Partitioning support for input data
- Native Scoring for more models
- Support for more languages ???