Appsilon
DATA SCIENCE

Traits of a world class Data Scientist

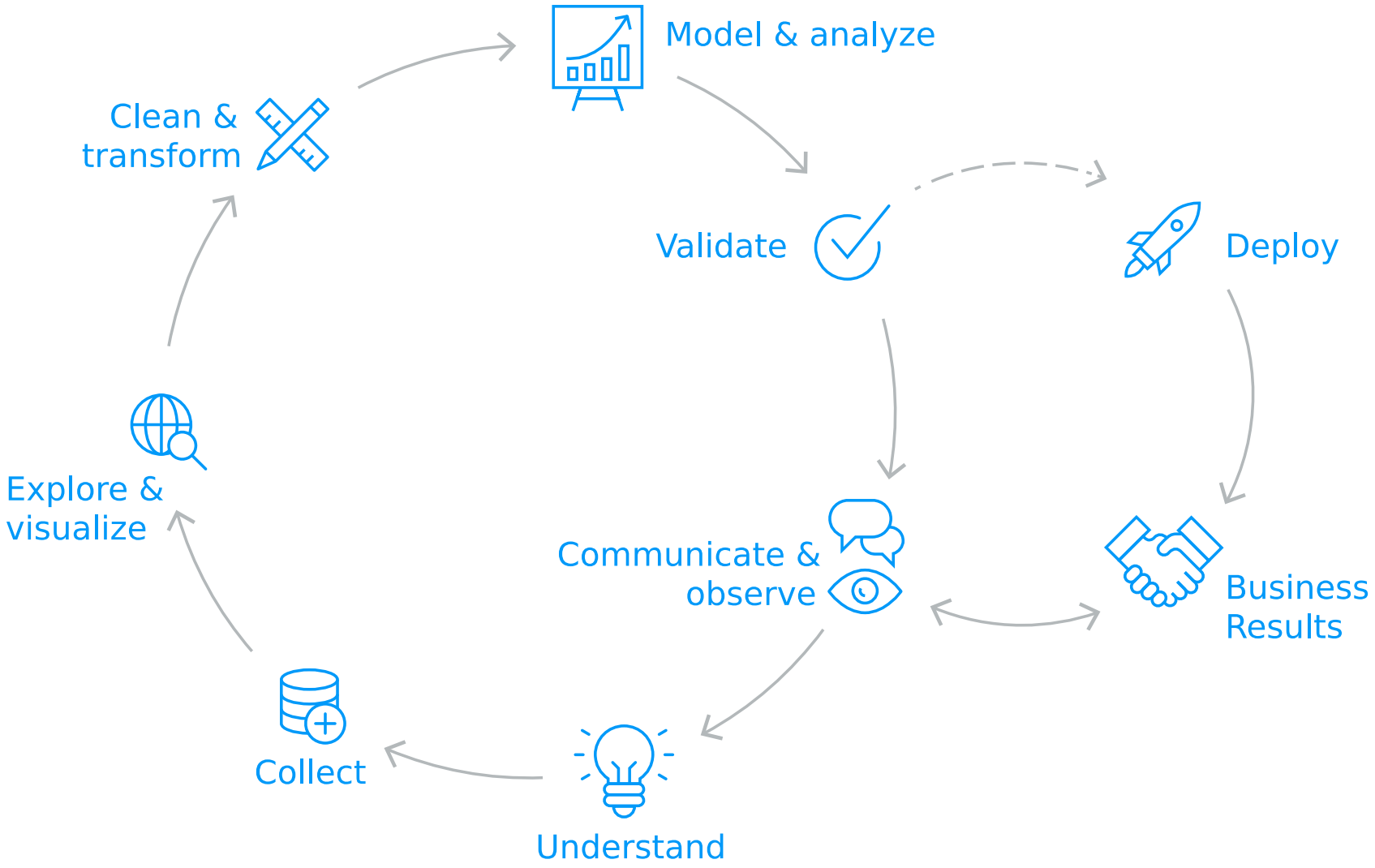
WhyR 2019 | Olga Mierzwa-Sulima | 09 2019
Senior Data Scientist

@olga_mie


German telecom case
about churn

Appsilon
DATA SCIENCE

What does a Data Scientist do?

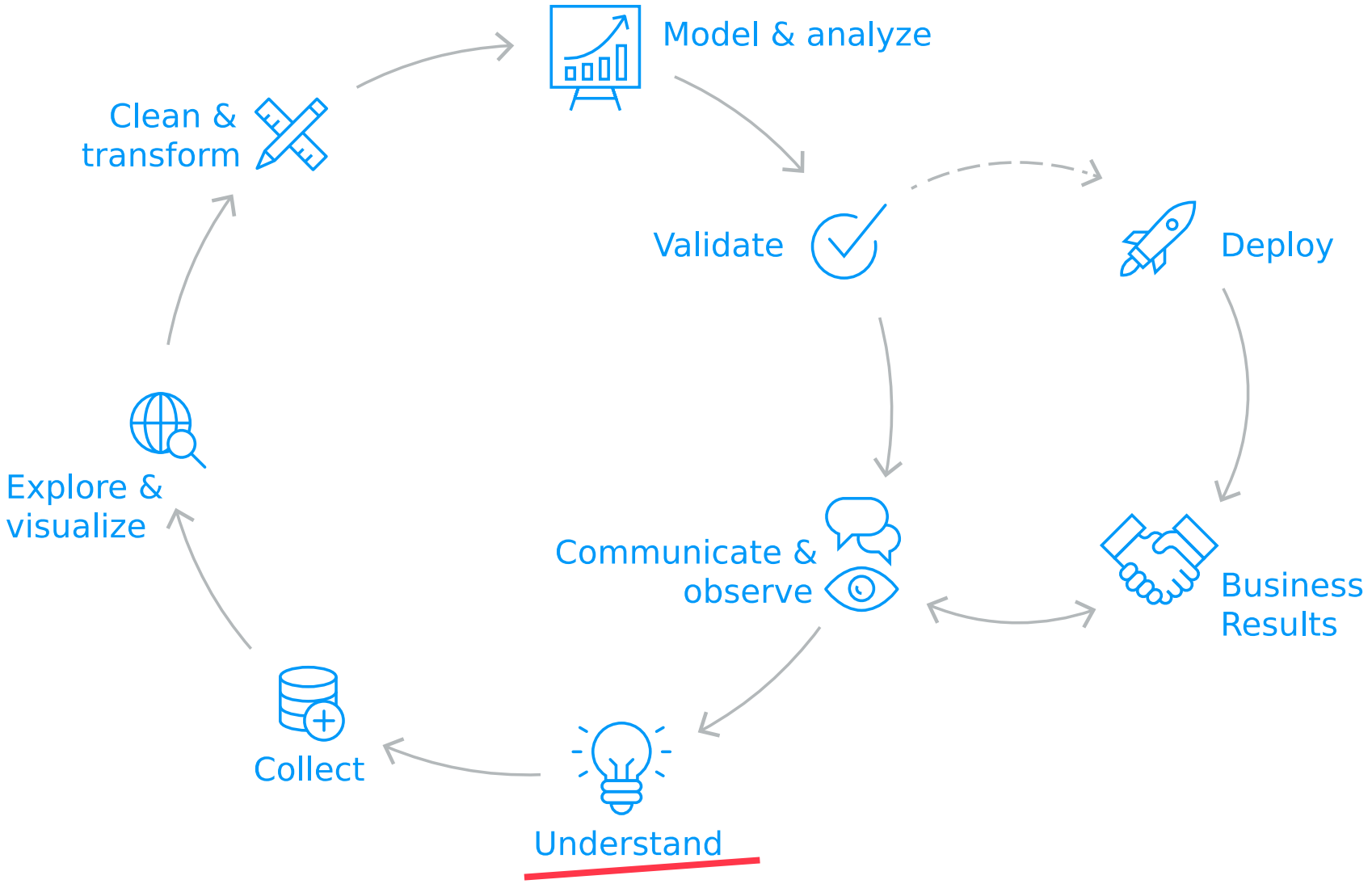


Inspired by Szilard Pafka chart

@olga_mie

Appsilon
DATA SCIENCE

What does a Data Scientist do?



Inspired by Szilard Pafka chart

Data Science Value

Whether:
what's **answerable** is **valuable**
what's **unknown** is **answerable**

Appsilon
DATA SCIENCE

François Chollet
@fchollet

Follow

There is an ongoing misconception that AI/ML are intrinsically valuable, and that therefore working in the field is bound to make you rich.

A ML model is only as valuable as the problem it solves. ML without an application isn't worth anything (beyond intellectual curiosity).

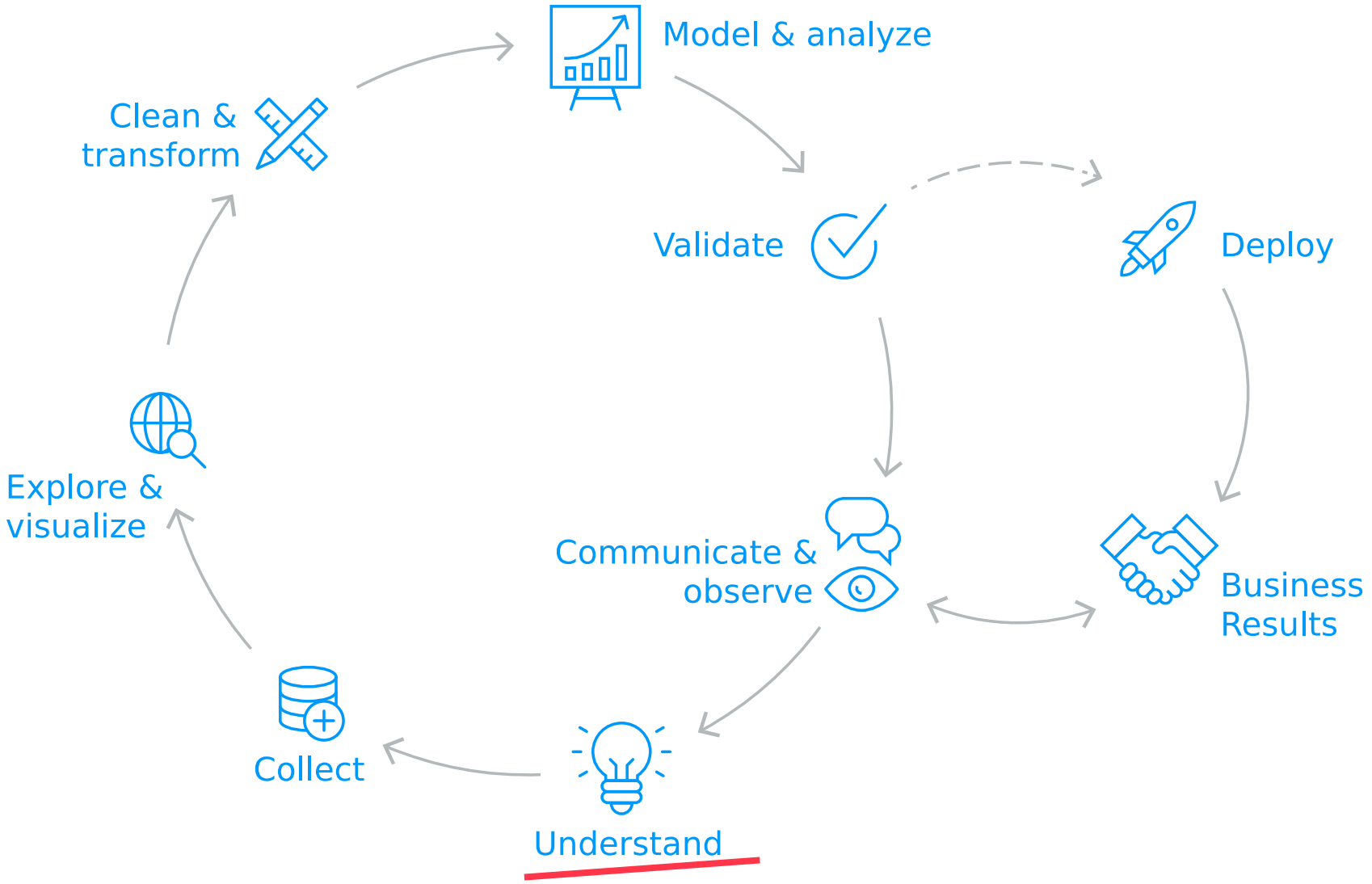
11:43 PM - 18 May 2019

788 Retweets **2,789** Likes

@olga_mie

Appsilon
DATA SCIENCE

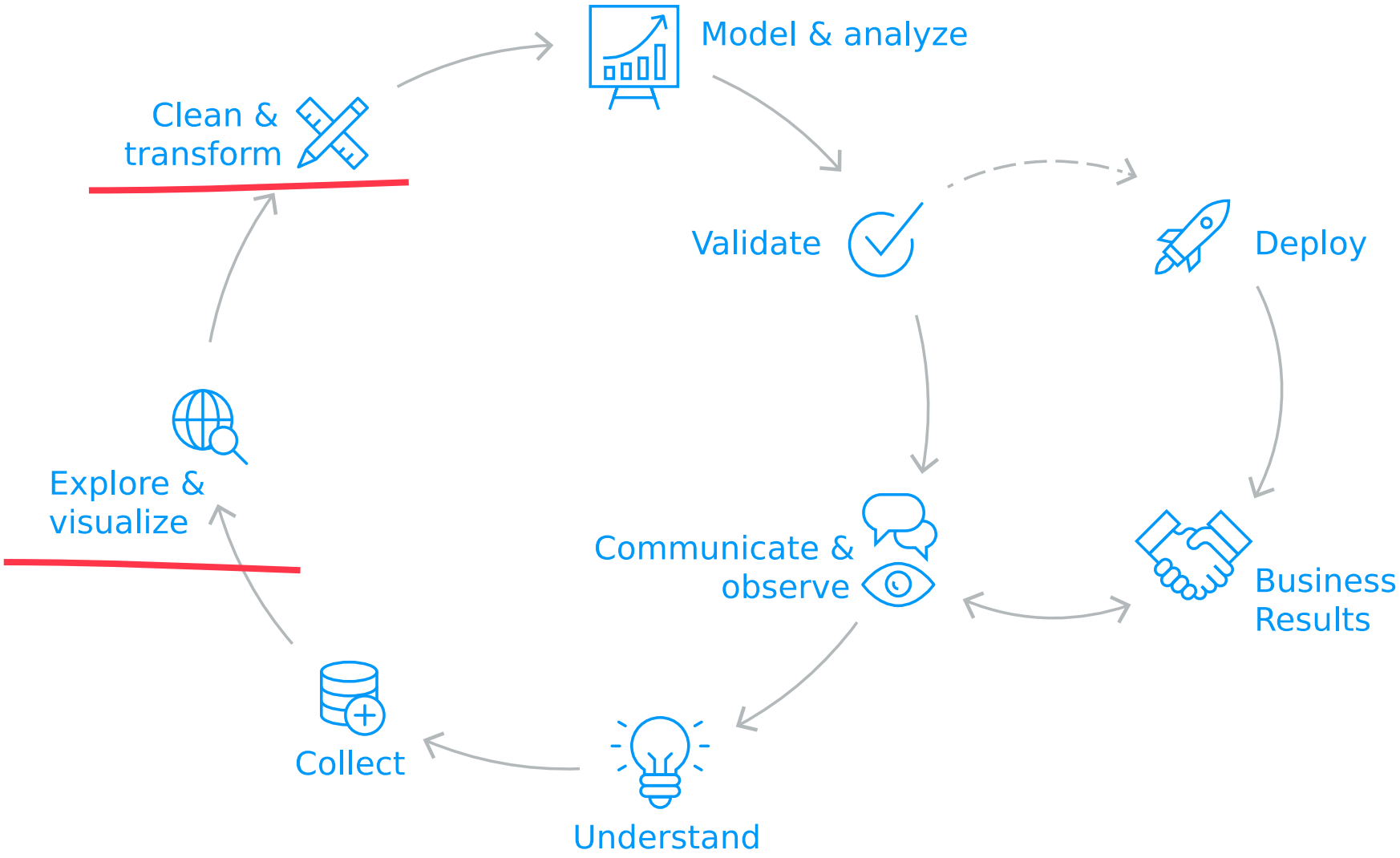
What does a Data Scientist do?



Inspired by Szilard Pafka chart

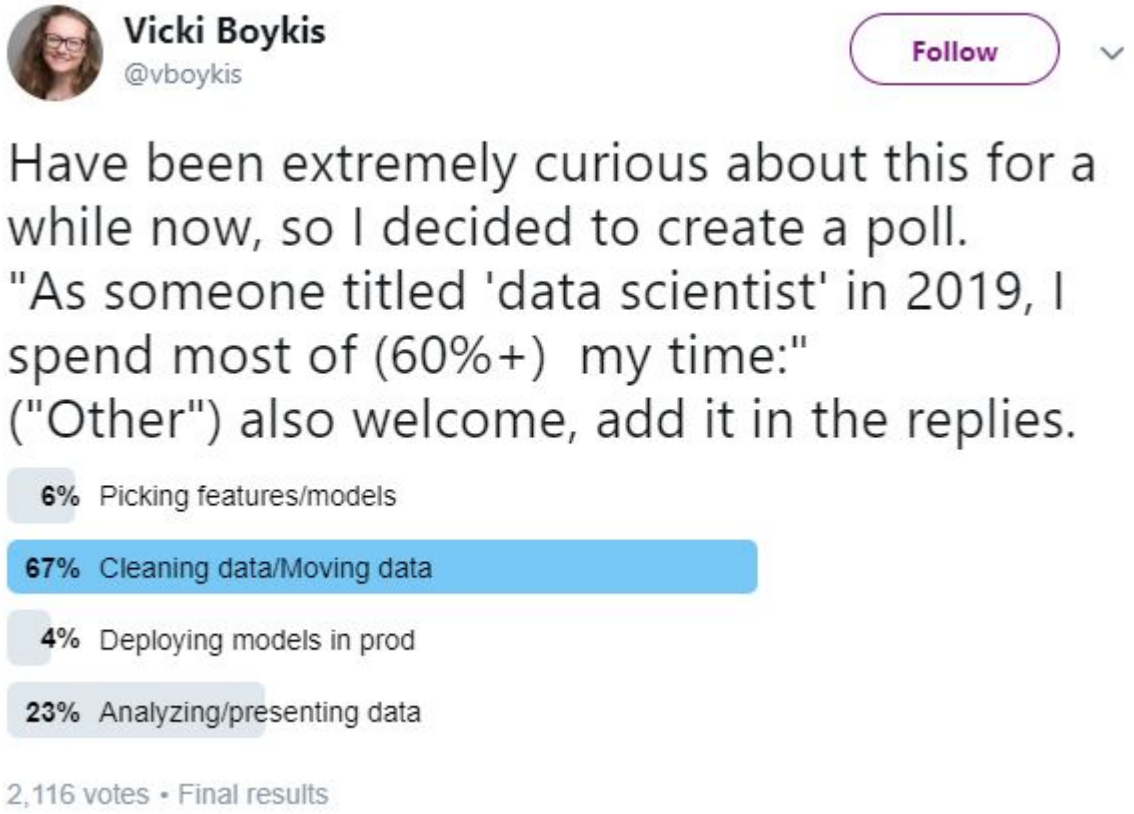
Appsilon
DATA SCIENCE

Everyone complains about



Appsilon
DATA SCIENCE

Truth about data science?



Appsilon
DATA SCIENCE

Data preparation

Takes at least 60% of the time

Prepares the data and the modeler

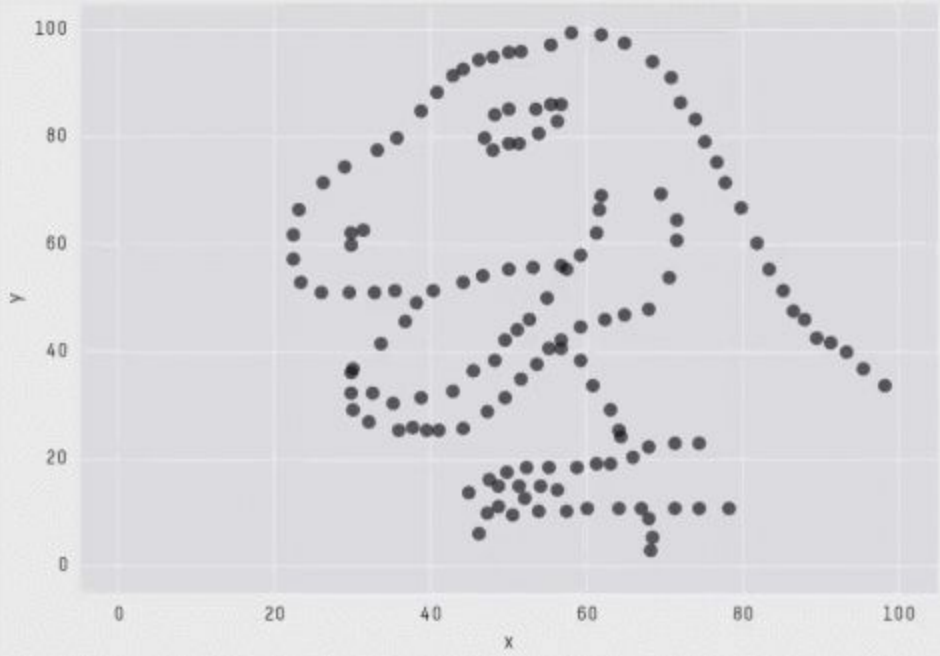
Teaches to be curious

Has high impact on any further work incl. model quality

Why would you want to plot?

Appsilon
DATA SCIENCE

Meet DataSaurus



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Appsilon
DATA SCIENCE

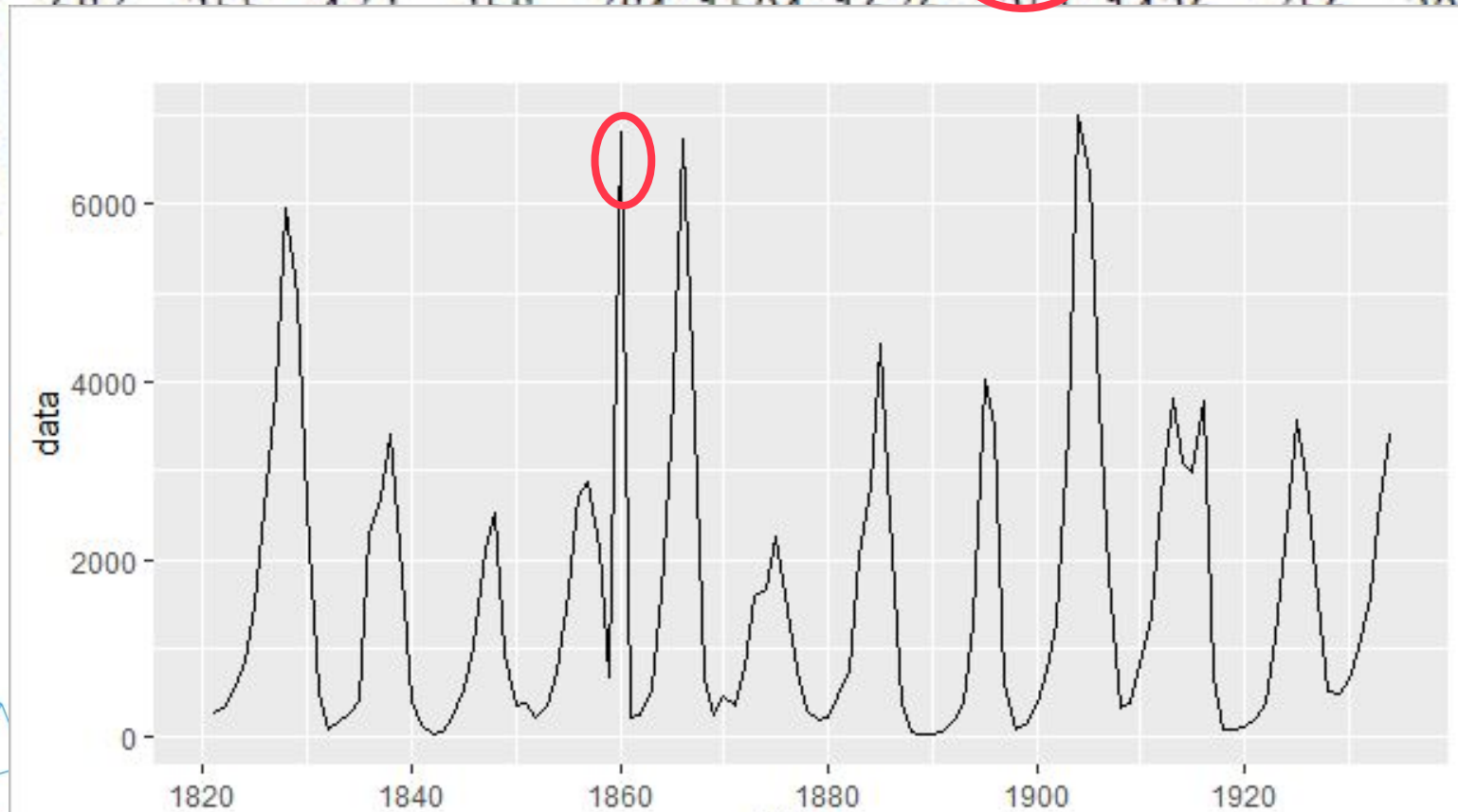
Outlier hunt?

```
Time Series:
Start = 1821
End = 1934
Frequency = 1
  [1]  269  321  585  871 1475 2821 3928 5943 4950 2577  523   98  184  279  409
 [16] 2285 2685 3409 1824  409  151   45   68  213  546 1033 2129 2536  957  361
 [31]  377  225  360  731 1638 2725 2871 2119  684 6800  236  245  552 1623 3311
 [46] 6721 4254  687  255  473  358  784 1594 1676 2251 1426  756  299  201  229
 [61]  469  736 2042 2811 4431 2511  389   73   39   49   59  188  377 1292 4031
 [76] 3495  587  105  153  387  758 1307 3465 6991 6313 3794 1836  345  382  808
 [91] 1388 2713 3800 3091 2985 3790  674   81   80  108  229  399 1132 2432 3574
[106] 2935 1537  529  485  662 1000 1590 2657 3396
```

Appsilon
DATA SCIENCE

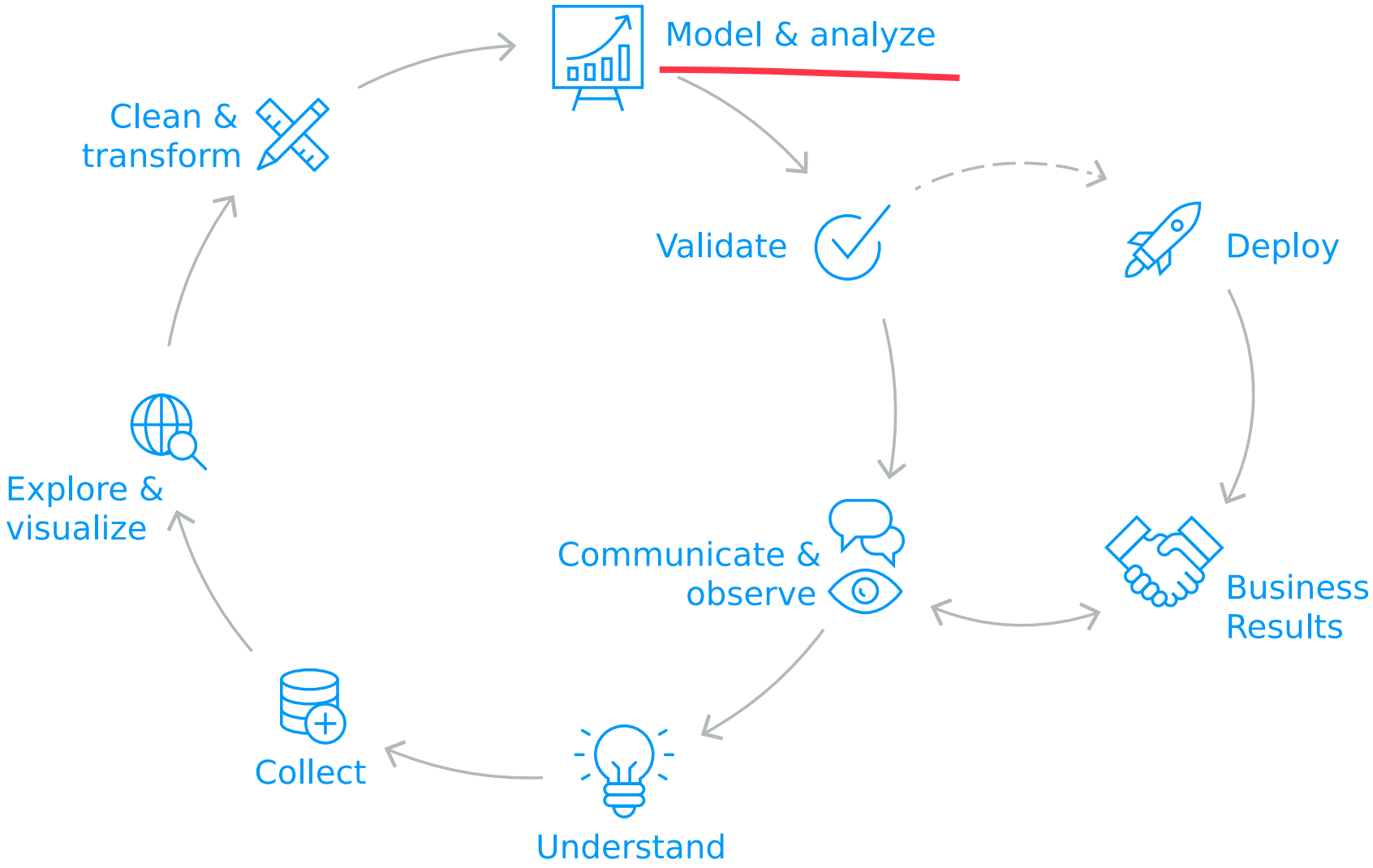
Outlier hunt?

```
Time Series:
Start = 1821
End = 1934
Frequency = 1
  [1]  269  321  585  871 1475 2821 3928 5943 4950 2577  523   98  184  279  409
 [16] 2285 2685 3409 1824  409  151   45   68  213  546 1033 2129 2536  957  361
 [31]  377  225  360  731 1638 2725 2871 2119  684 6800  236  245  552 1623 3311
 [46] 6721 4254                                                  9  201  229
 [61]  469  736                                                  7 1292 4031
 [76] 3495  587                                                  5  382  808
 [91] 1388 2713                                                  2 2432 3574
[106] 2935 1537
```



Appsilon
DATA SCIENCE

Everyone wants to...



Appsilon
DATA SCIENCE

Modeling

Hard: need to understand the Math and the algo

Know pros and cons of different algorithms

Leverage domain knowledge

Use the right tools

Appsilon
DATA SCIENCE

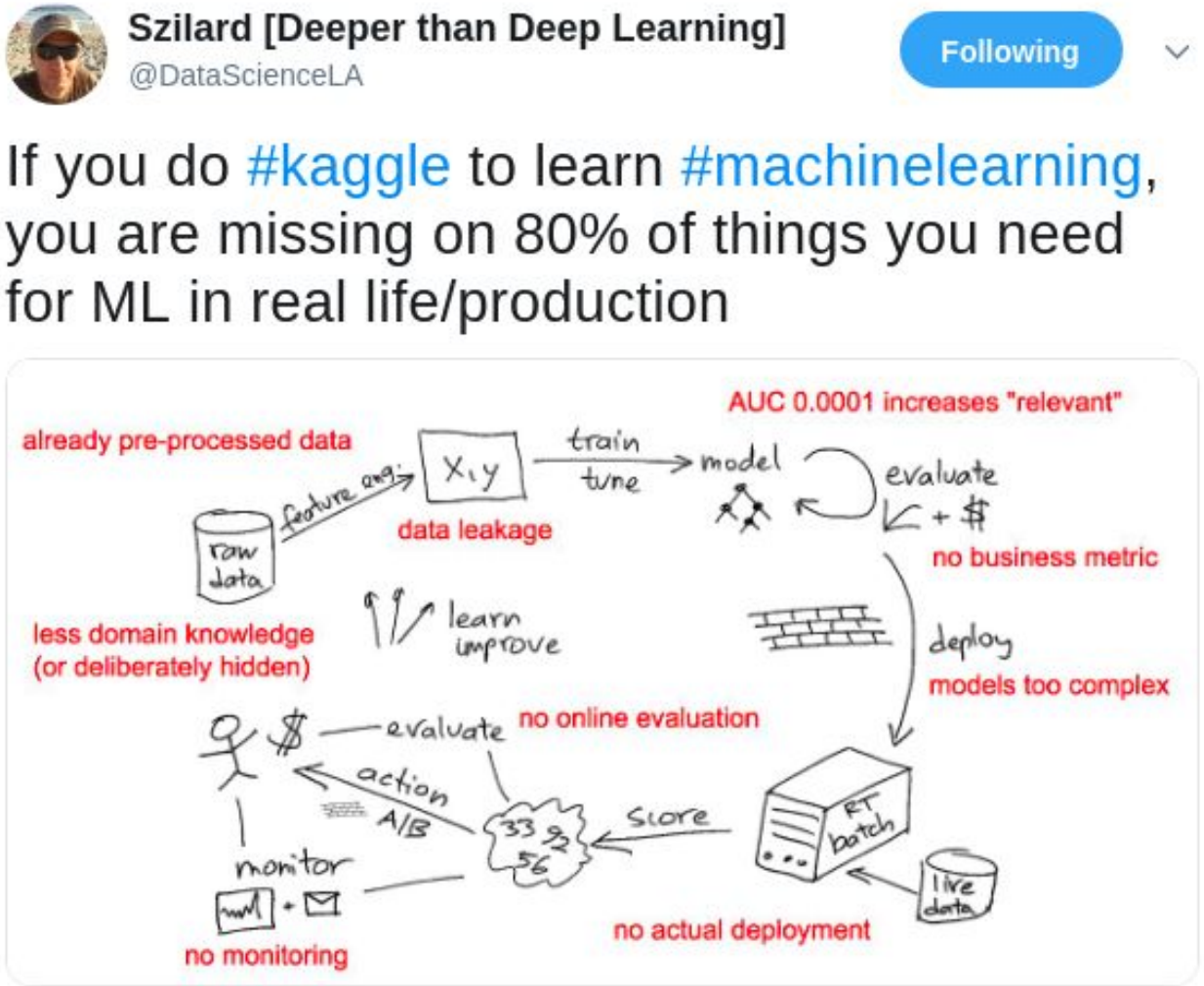
Tools: Performance R GBM packages

	xgboost	lightgbm	h2o
easy R install	cran	linux OK	java_cran
maintained	yes	yes	yes
preprocessing	1-hot	1-hot/categ int	not needed
new cats scoring	no	no	yes
early stopping	yes	yes	yes
speed (CPU)	**ok**	**fastest**	**slow (small data)**
GPU supported	yes	yes	via xgboost
speed GPU	**fastest**	**ok/slow**	**indirectly/slower**
REST scoring	**no**	**no**	**yes**
other algos	RF	RF	RF/GLM/NN
best for	**Kaggle**	**Kaggle**	**prod/real-time**

Source : https://github.com/szilard/GBM-perf

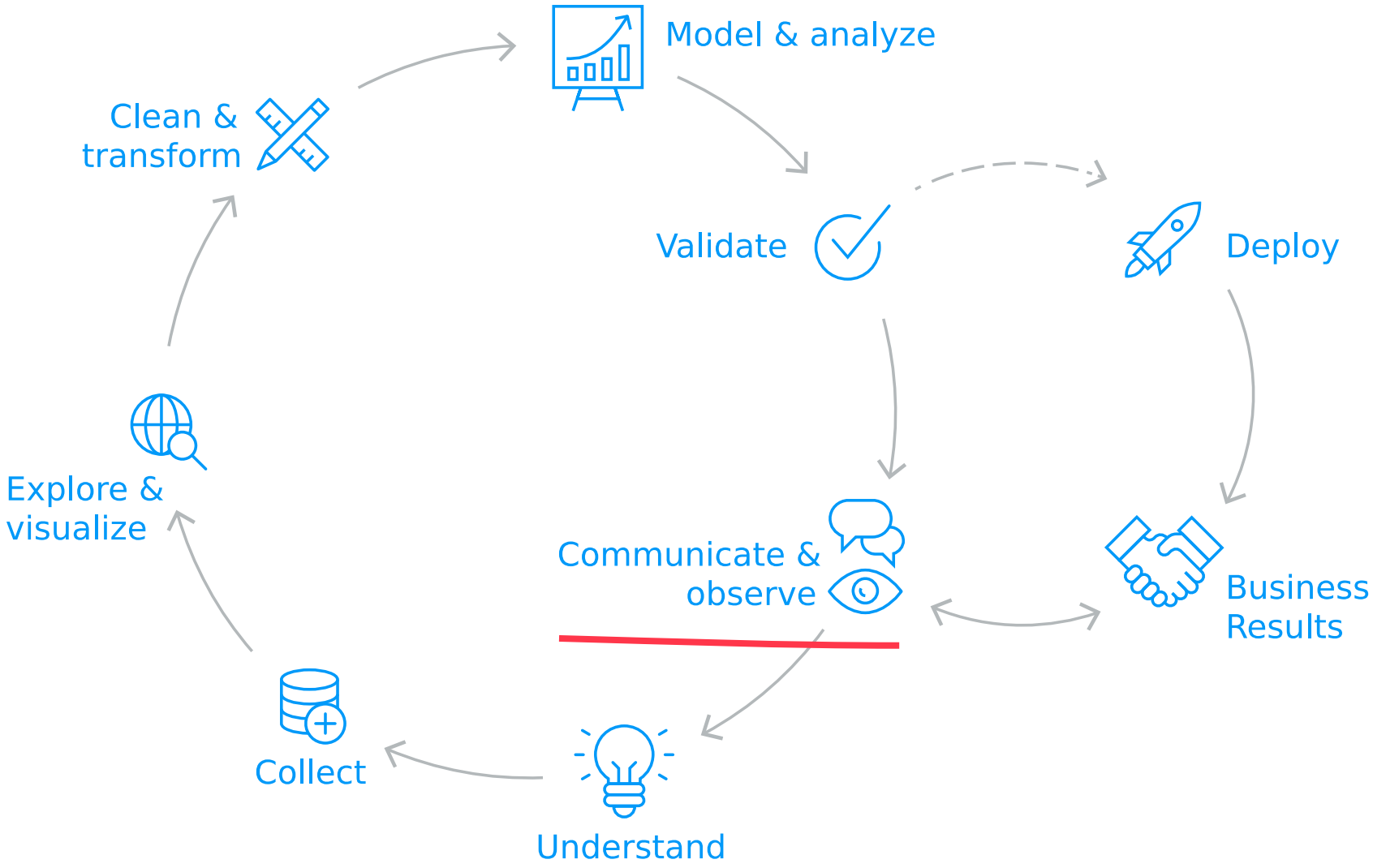
Appsilon
DATA SCIENCE

Kaggle != ML



Appsilon
DATA SCIENCE

Everyone should master



@olga_mie

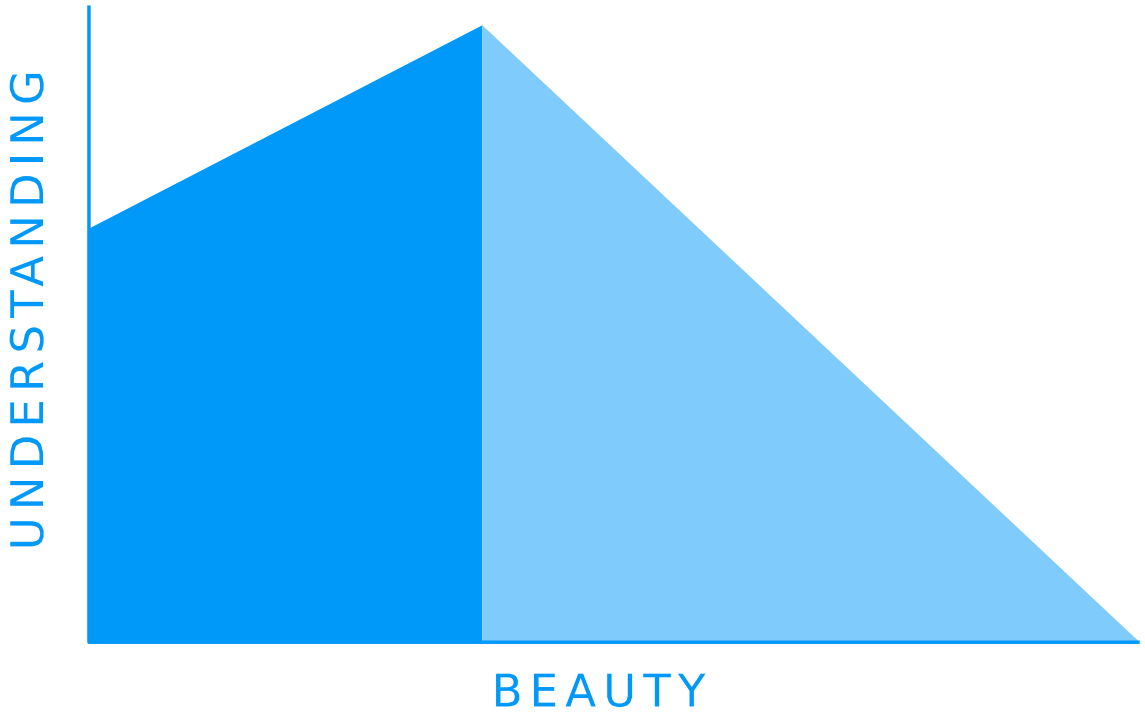
Appsilon
DATA SCIENCE

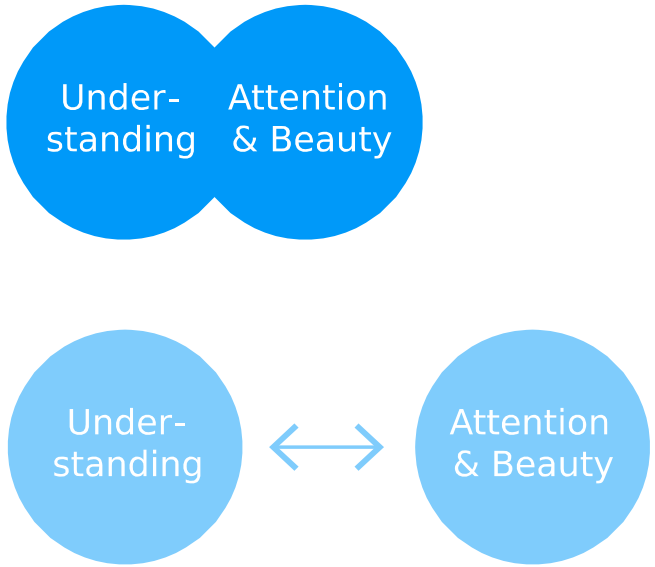
Communicate ➡ Sell

- **Understand** your audience/user/client
  - A role of listening, empathy and exploration
- Make it beautiful yet understandable
- Make it impactful and business worthy

Appsilon
DATA SCIENCE

Make it beautiful yet understandable





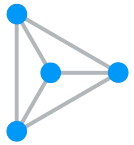
Appsilon
DATA SCIENCE

Communicate ➡ Sell

- Understand your audience/user/client
  - Role of listening, empathy and exploration
- Make it beautiful yet understandable
- Make it impactful and business worthy

Kill it quickly

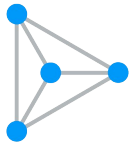
- Move on to a new idea / hypothesis
- Advice on the best next step
- Hand-off



Appsilon
DATA SCIENCE

World class data scientist ships value

Model without an application is worth nothing



Appsilon
DATA SCIENCE

Questions?

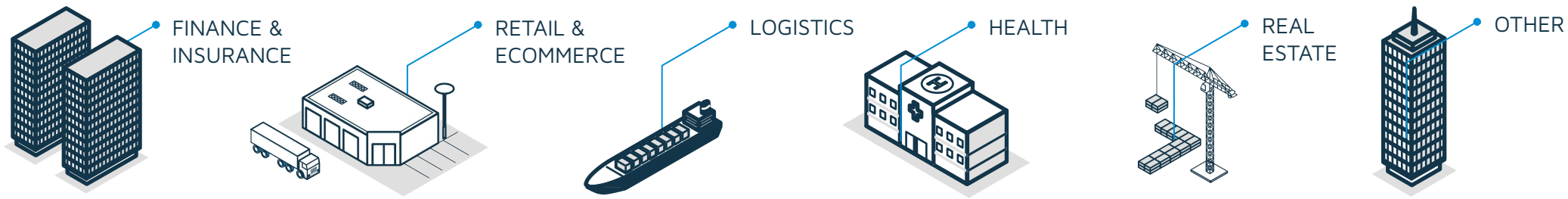
olga@appsilon.com

Appsilon
DATA SCIENCE

What mark will you leave on the world?

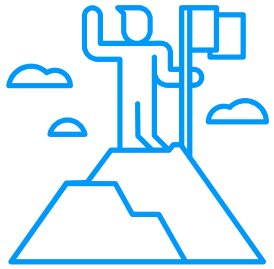
appsilon.com/careers

fb.com/appsilondatascience



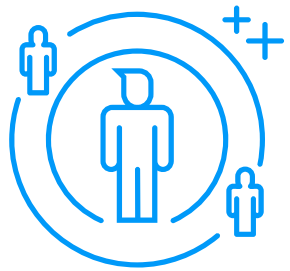
Appsilon
DATA SCIENCE

Not only tech skills



Ownership

Drive to ship project success
Own deliverables



Influence

Communicates clearly
Teamwork
Building and maintaining relations

Appsilon
DATA SCIENCE

Performance R GBM packages

	gbm (R pkg)	xgboost	lightgbm	h2o
easy R install	cran	cran	linux OK	java_cran
maintained	no	yes	yes	yes
preprocessing	not needed	1-hot	1-hot/categ int	not needed
new cats scoring	yes	no	no	yes
early stopping	no	yes	yes	yes
speed (CPU)	**1 core**	**ok**	**fastest**	**slow (small data)**
GPU supported	no	yes	yes	via xgboost
speed GPU	**NA**	**fastest**	**ok/slow**	**indirectly/slower**
REST scoring	no	no	no	yes
other algos	no	RF	RF	RF/GLM/NN
best for	-	**Kaggle**	**Kaggle**	**prod/real-time**