Automated Data Quality Assurance with Machine Learning

4th European Conference On Artificial Intelligence in Finance and Industry
Winterthur, 05.09.2019

Martin Müller-Lennert
Senior Data Scientist
martin.mueller-lennert@incubegroup.com

Milica Petrović
Senior Data Scientist
milica.petrovic@incubegroup.com

InCube
Digitization - Data - Intelligence

Talk Outline

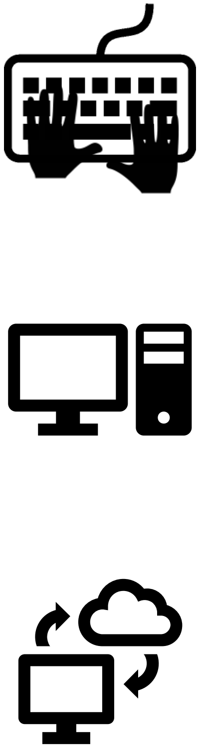
1. **What's Wrong with Data Quality?**
2. **Error Detection using ML**
3. **Demo & Features**
4. **Error Remediation using RPA**

Talk Outline

1. **What's Wrong with Data Quality?**
2. Error Detection using ML
3. Demo & Features
4. Error Remediation using RPA

Data Quality Today

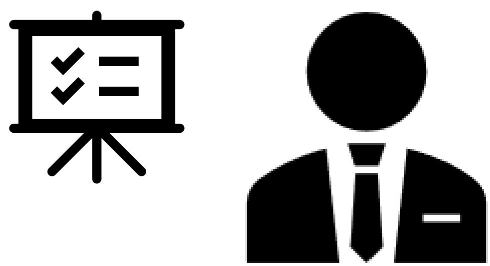
What’s wrong with it?



Data Sources

Data Quality

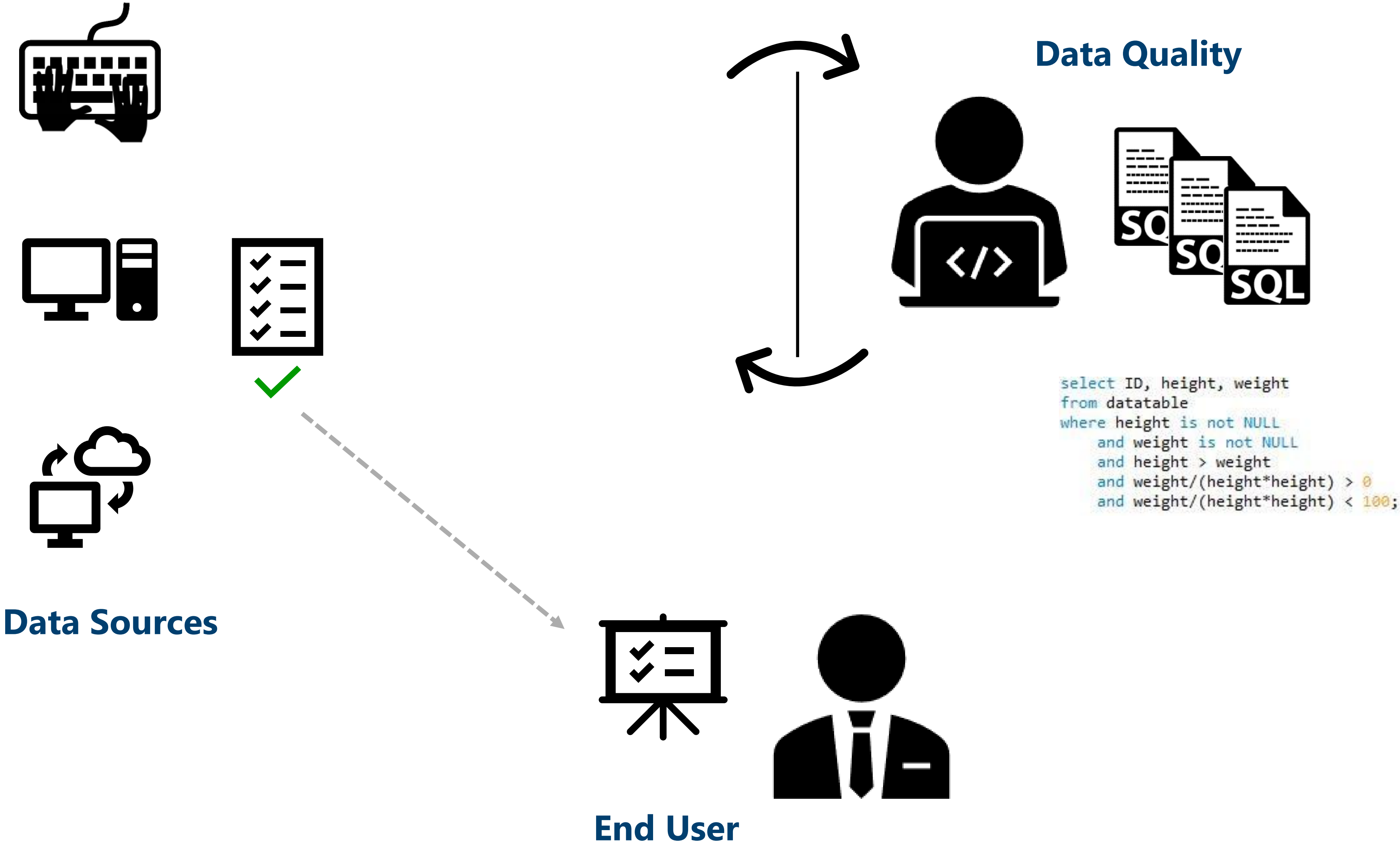

```
select ID, height, weight
from datatable
where height is not NULL
    and weight is not NULL
    and height > weight
    and weight/(height*height) > 0
    and weight/(height*height) < 100;
```



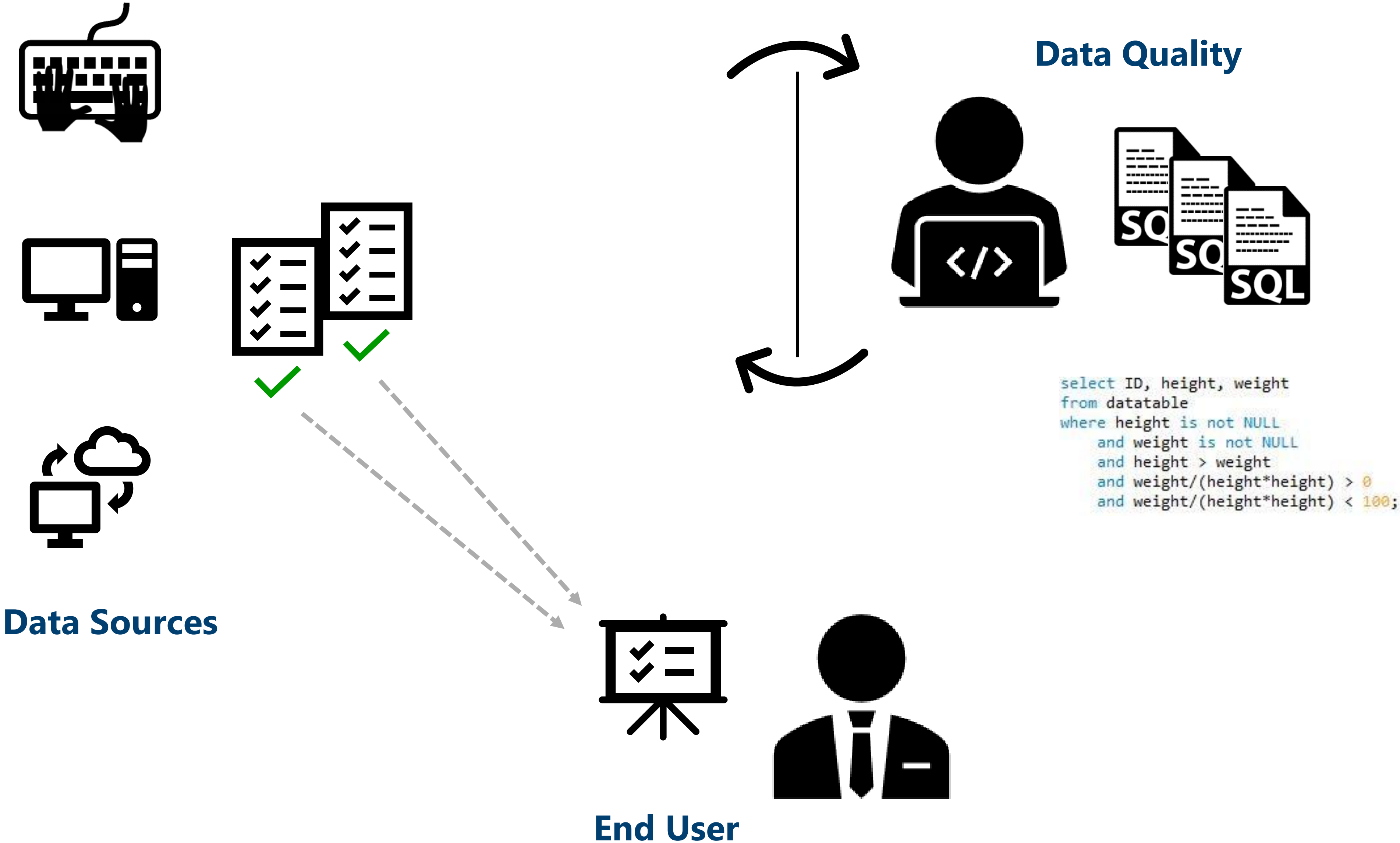
End User

Data Quality Today

What's wrong with it?



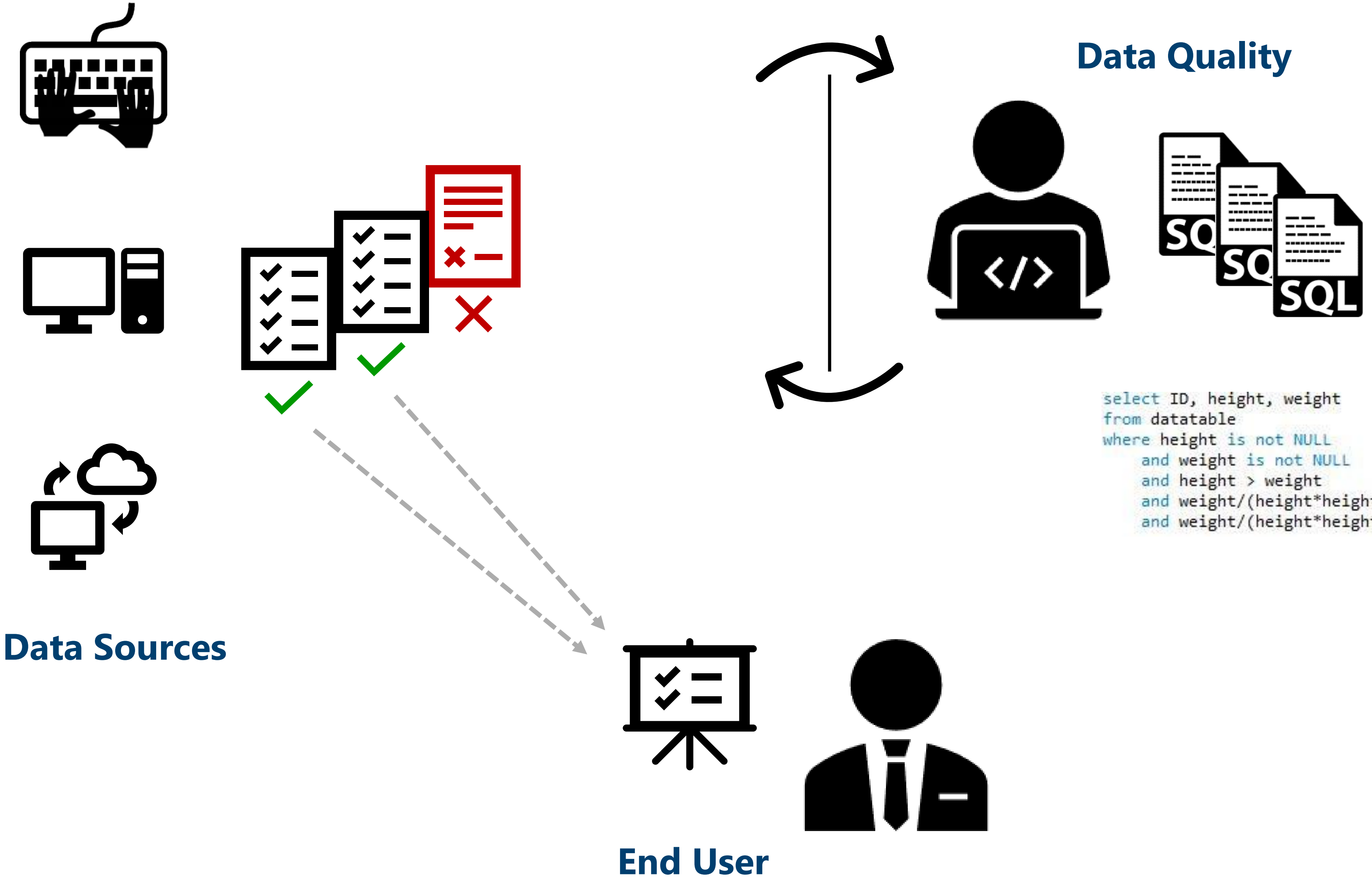
Data Quality Today
What's wrong with it?




```sql
select ID, height, weight
from datatable
where height is not NULL
    and weight is not NULL
    and height > weight
    and weight/(height*height) > 0
    and weight/(height*height) < 100;
```

Data Quality Today

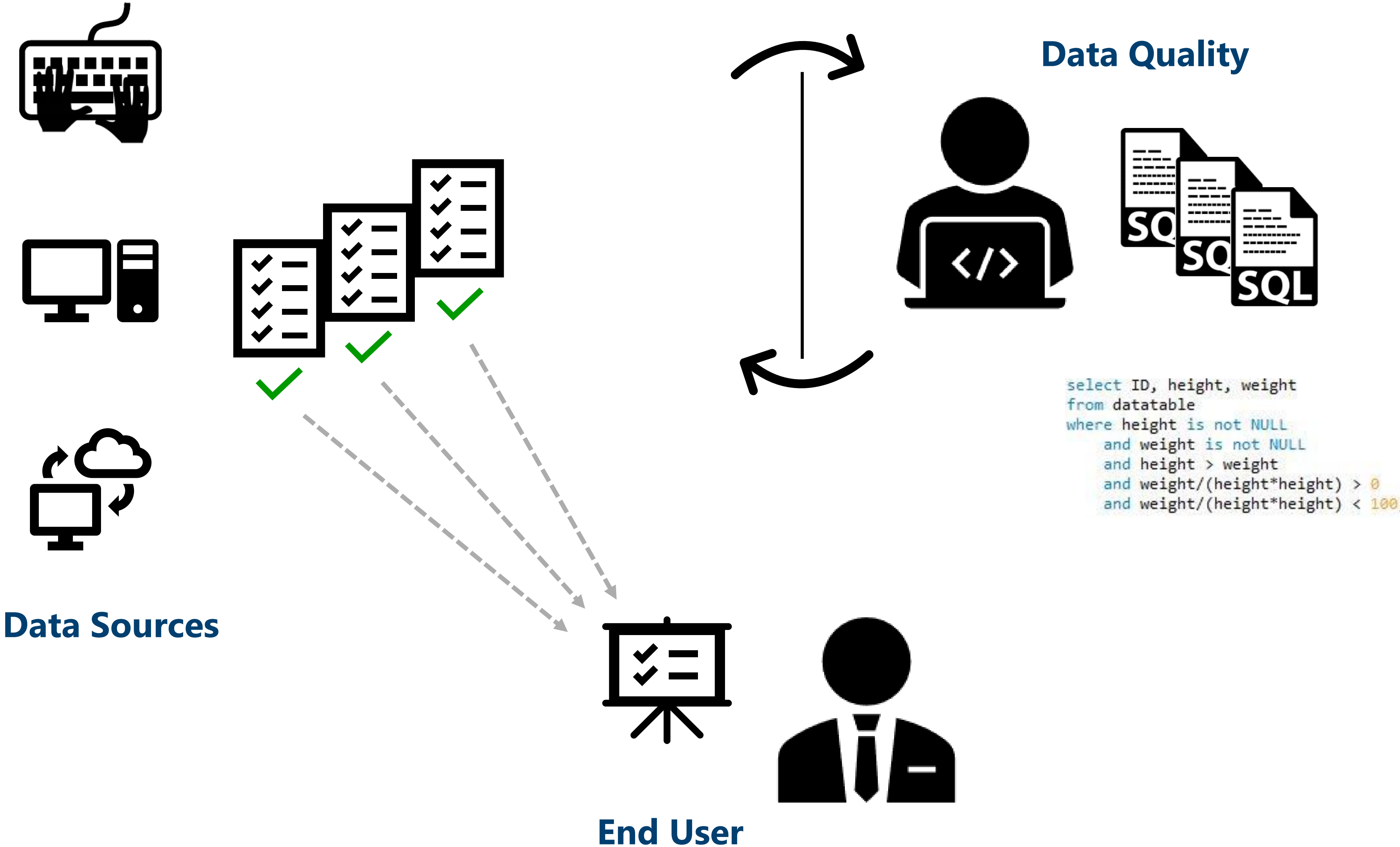
What's wrong with it?




```
select ID, height, weight
from datatable
where height is not NULL
    and weight is not NULL
    and height > weight
    and weight/(height*height) > 0
    and weight/(height*height) < 100;
```

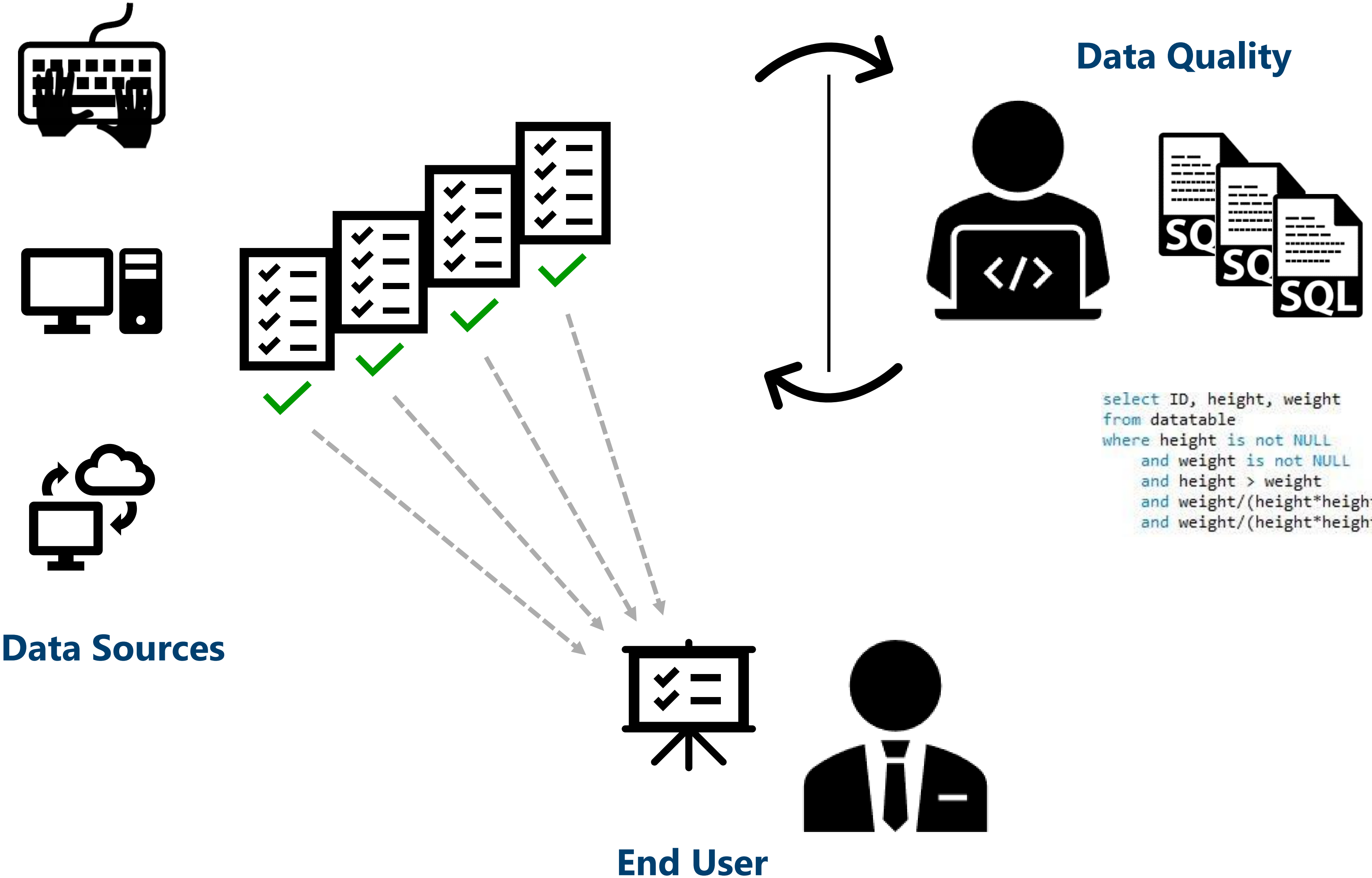
Data Quality Today

What's wrong with it?



Data Quality Today

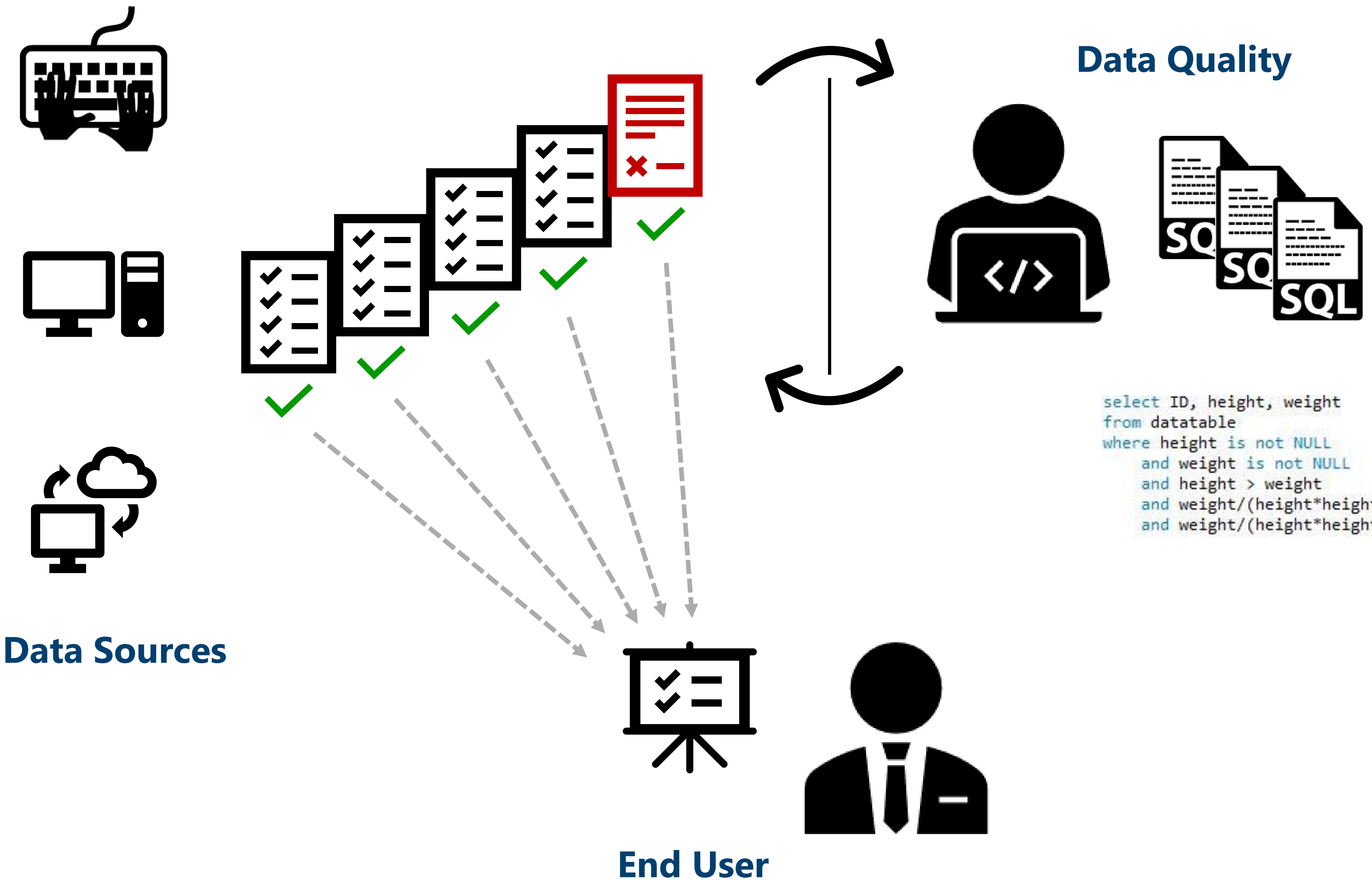
What's wrong with it?




```sql
select ID, height, weight
from datatable
where height is not NULL
    and weight is not NULL
    and height > weight
    and weight/(height*height) > 0
    and weight/(height*height) < 100;
```

Data Quality Today

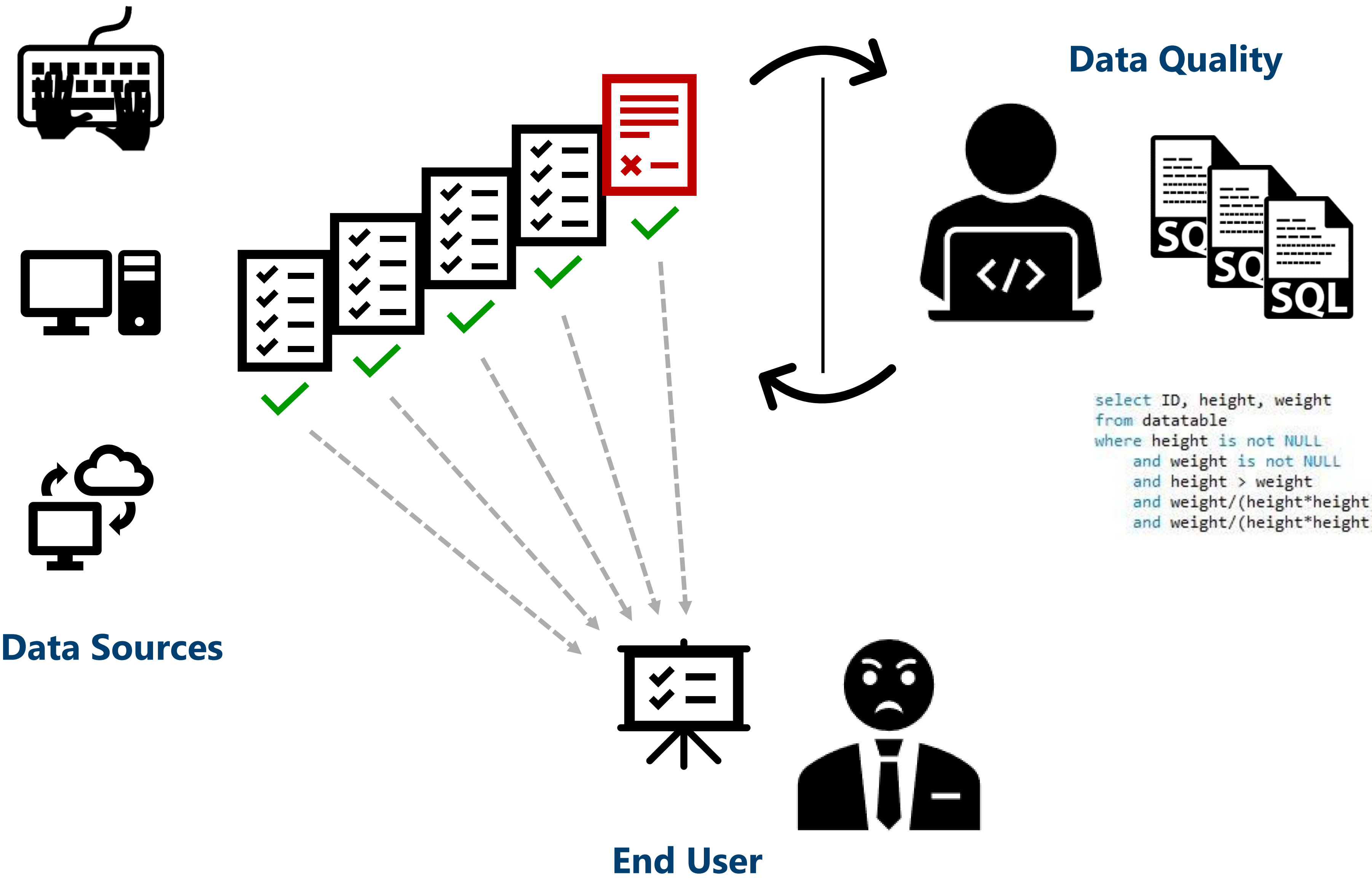
What’s wrong with it?




```sql
select ID, height, weight
from datatable
where height is not NULL
    and weight is not NULL
    and height > weight
    and weight/(height*height) > 0
    and weight/(height*height) < 100;
```

Data Quality Today

What's wrong with it?




```
select ID, height, weight
from datatable
where height is not NULL
    and weight is not NULL
    and height > weight
    and weight/(height*height) > 0
    and weight/(height*height) < 100;
```

Data Quality Today

Our Take at a Solution

Data Quality Today

- Manually coded SQL rules
- Uni-/bi-variate checks

Challenges

- Too much data
- Too few rules
- Too narrow focus
- Too late

Solutions with Machine Learning

- **Automate**: simultaneous error detection & faster process
- **Reusability**: tailored ML algorithms reused for fields of similar type
- **Deep dive**: discovery of new types of errors based on multivariate relationships

Autoencoders

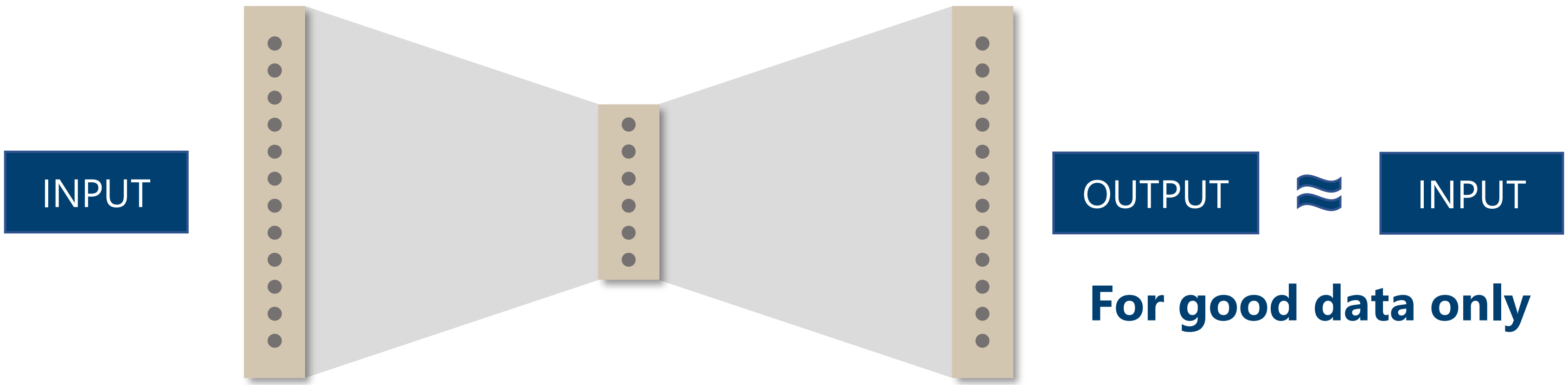
- **Unsupervised**
- Capture multivariate relationships

Talk Outline

1. What's Wrong with Data Quality?
2. **Error Detection using ML**
3. Demo & Features
4. Error Remediation using RPA

Autoencoders for Data Quality

Architecture and Training



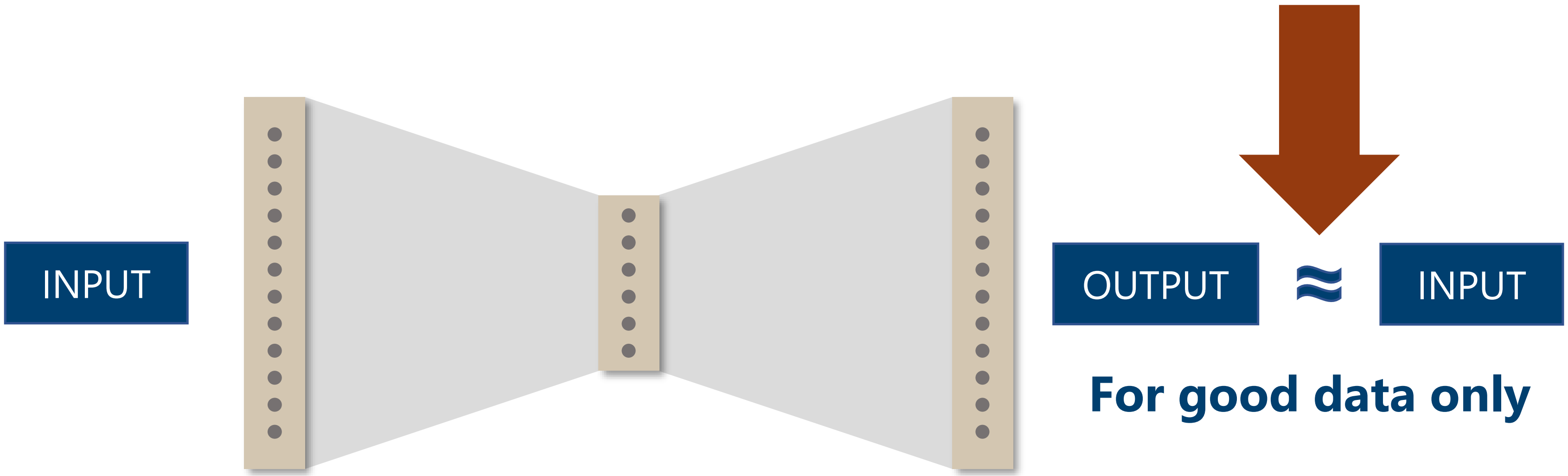
For good data only

Target: Reconstruct input

Bottleneck: Enforced by architecture or regularization

Ensures network learns structure of input data

Autoencoders for Data Quality
Architecture and Training



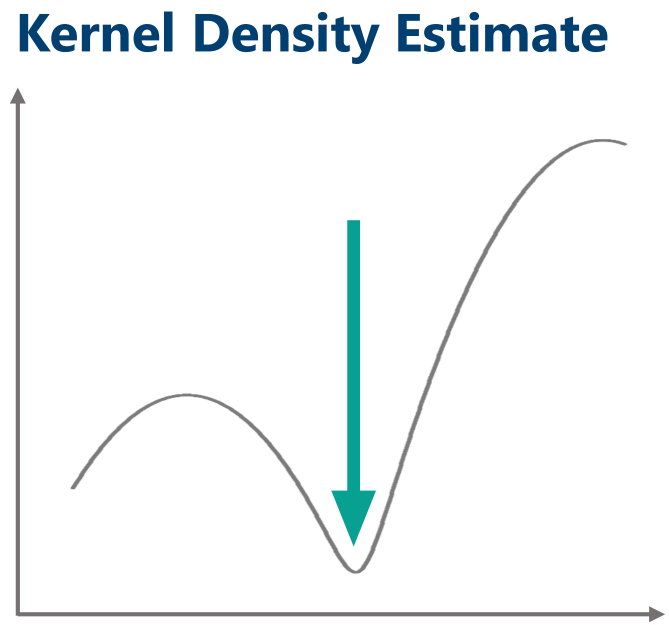
Training on imperfect data: Requires large share of good data

Limits potency of network: More layers not always better

Discriminating Good and Bad Data Records

Clustering the Reconstruction Errors

SAMPLE DATA

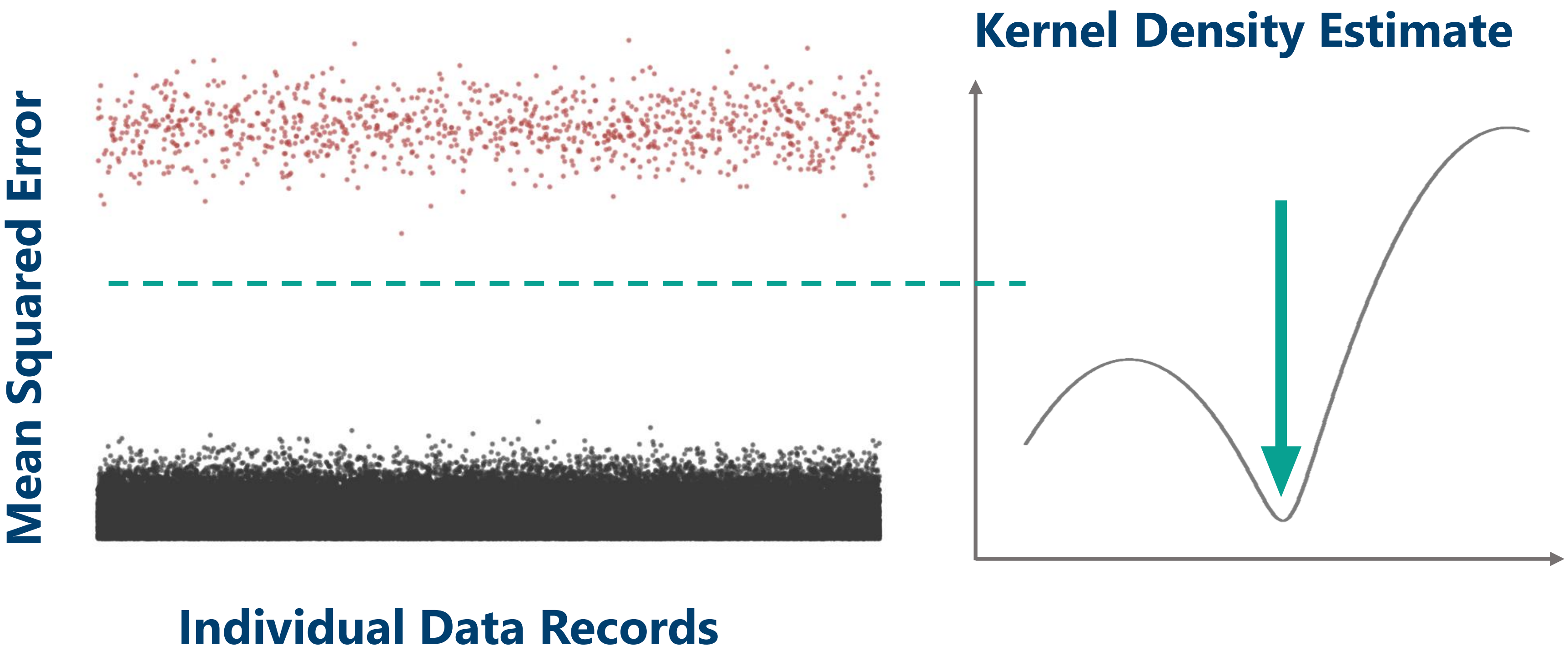


Challenge: Many data points and potentially extreme class imbalance

Discriminating Good and Bad Data Records

Clustering the Reconstruction Errors

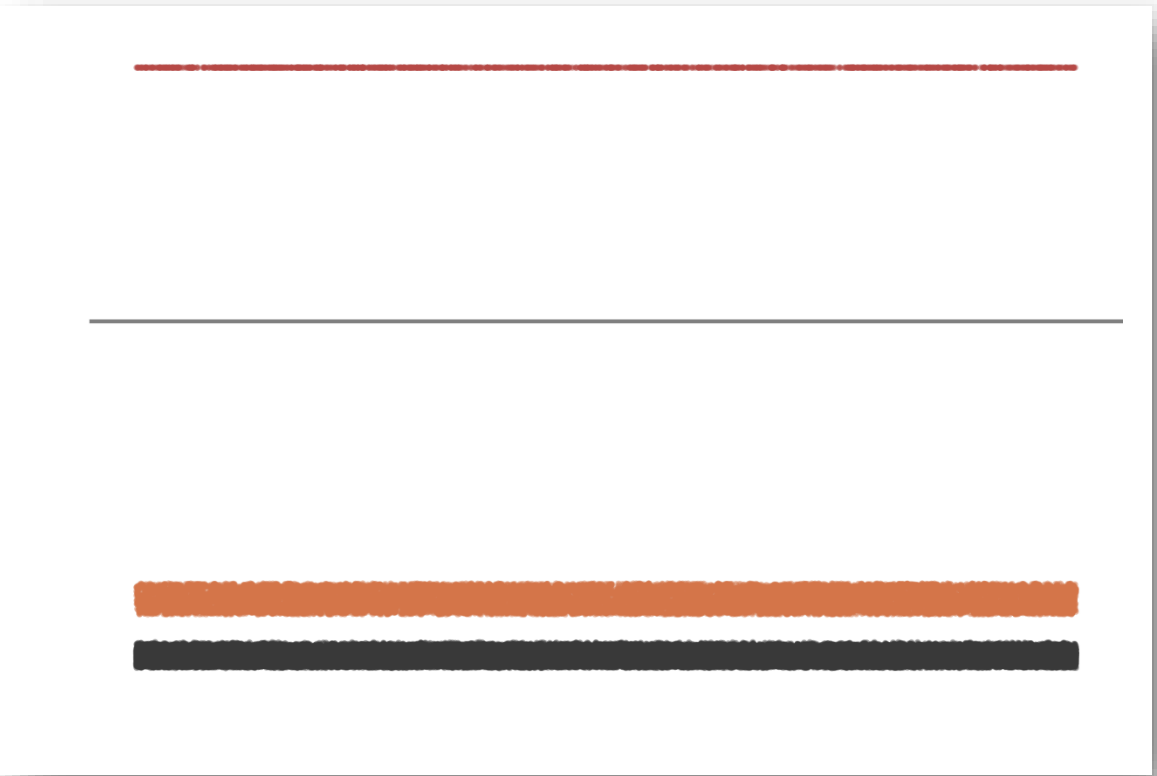
SAMPLE DATA



Challenge: Many data points and potentially extreme class imbalance

Discriminating Good and Bad Data Records

Sequence of Autoencoders



1st iteration

Remove Detected Anomalies

Keep Rest of Data

Challenge: Magnitude of reconstruction error varies across data error types

Discriminating Good and Bad Data Records

Sequence of Autoencoders

Discriminating Good and Bad Data Records

Sequence of Autoencoders

Discriminating Good and Bad Data Records
Sequence of Autoencoders

Across iterations: Increase model complexity

Stopping: When threshold separates large chunk of data

Talk Outline

1 What's Wrong with Data Quality?

2 Error Detection using ML

3 **Demo & Features**

4 Error Remediation using RPA

Demo
Birth date

DQ with Autoencoders | Demo | Info

Select Field for Anomaly Detection

- ( ) First name
- ( ) Last name
- ( ) Company name
- (•) Birth date
- ( ) Income
- ( ) Tax class
- ( ) Revenue

Analyze | Reset Data

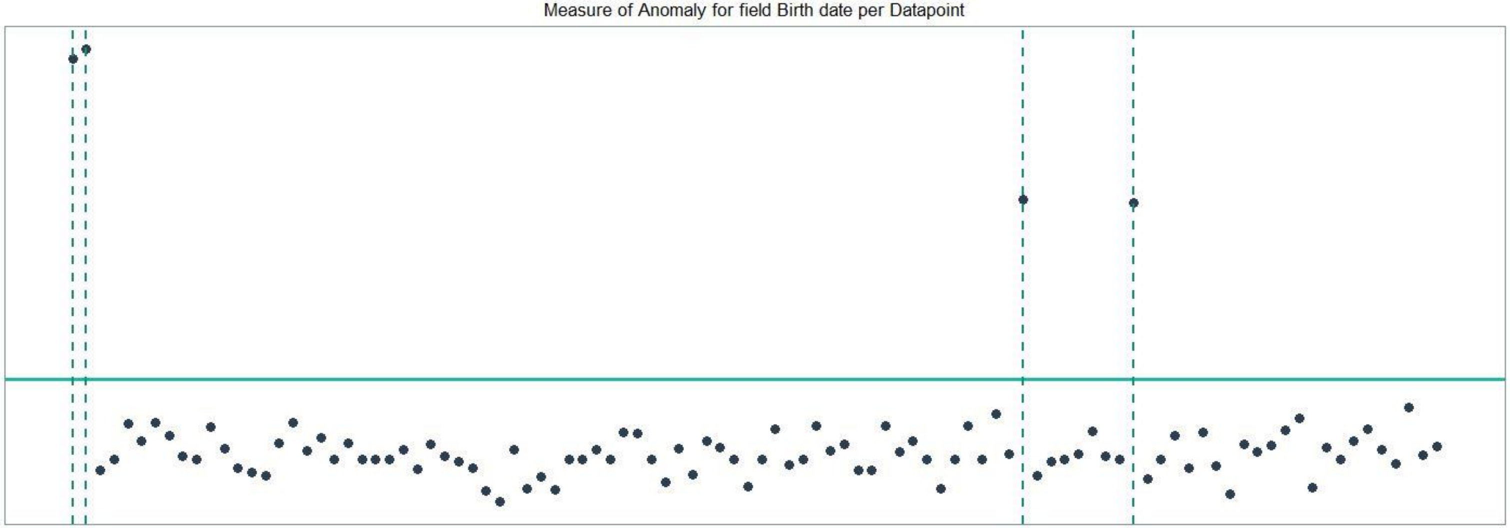
KPIs for selected field

True Positives	True Negatives
4 FILTER TABLE	96 FILTER TABLE

False Positives	False Negatives
0 FILTER TABLE	0 FILTER TABLE

	Customer type	First name	Last name	Company name	Birth date	Income	Income class	Tax class	Child
1	private	Mrs Georgia	Kingsley		11/19/1875	540000	high	full	no
2	private	Demetria	Harmon		08/19/2020	270000	high	full	yes
3	private	Homer	Carr		10/12/1984	180000	medium	full	no
4	company			FreeSeas Inc.					
5	private	Littleton	Dean		01/22/1957	135000	medium	full	no
6	private	Aliana	Snider		01/30/1976	112500	medium	reduced	yes
7	private	Ralph	Jones		06/28/1953	135000	medium	full	no
8	private	Jimmy	Jordahl		04/21/1958	67500	low	reduced	no
9	private	Amos	Miller		01/23/1982	126000	medium	full	no
10	company			Peregrine Pharmaceuticals Inc.					
11	private	Elgin	Howell		02/24/1952	81000	low	reduced	no
12	private	Adriana	Bailey		07/09/1980	90000	low	reduced	no
13	private	Jamin	Sakaguchi		08/05/1971	90000	low	reduced	no
14	private	Charles	Partin		10/27/1977	90000	low	reduced	yes
15	private	Jeanetta	Clark		08/03/1966	292500	high	full	no

Measure of Anomaly for field Birth date per Datapoint



Demo
Birth date

SAMPLE DATA

DQ with Autoencoders | Demo | Info

Select Field for Anomaly Detection

- First name
- Last name
- Company name
- Birth date
- Income
- Tax class
- Revenue

Analyze | Reset Data

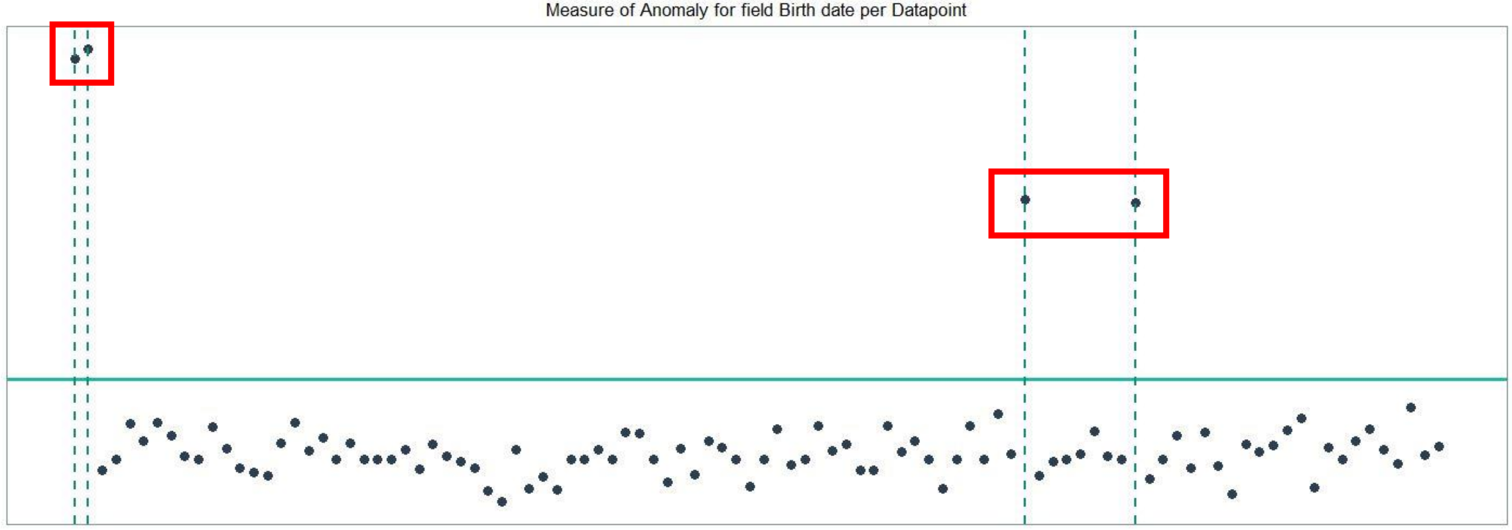
KPIs for selected field

True Positives	True Negatives
4 REMOVE FILTER	96 FILTER TABLE

False Positives	False Negatives
0 FILTER TABLE	0 FILTER TABLE

	Customer type	First name	Last name	Company name	Birth date	Income	Income class	Tax class	Children	Education	Revenue	
1	private	Mrs Georgia	Kingsley		11/19/1875	540000	high	full	no	Higher education		
2	private	Demetria	Harmon		08/19/2020	270000	high	full	yes	Incomplete higher		
70	company			Palmetto Bancshares, Inc. (SC)	02/15/1943						299504	1
78	private	Rupert	Carty			135000	medium	full	no	Higher education		

Measure of Anomaly for field Birth date per Datapoint



Demo

First name

SAMPLE DATA

DQ with Autoencoders | Demo | Info

Select Field for Anomaly Detection

- (•) First name
- ( ) Last name
- ( ) Company name
- ( ) Birth date
- ( ) Income
- ( ) Tax class
- ( ) Revenue

Analyze | Reset Data

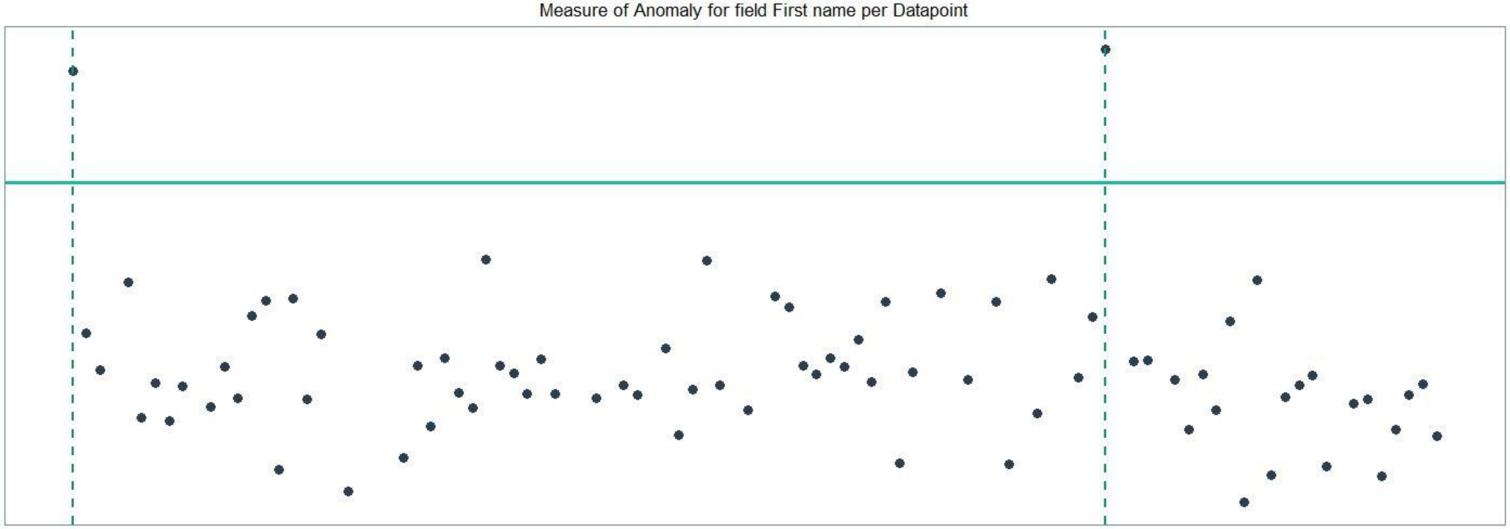
KPIs for selected field

True Positives	True Negatives
2 FILTER TABLE	98 FILTER TABLE

False Positives	False Negatives
0 FILTER TABLE	0 FILTER TABLE

	Customer type	First name	Last name	Company name	Birth date	Income	Income class	Tax class	Child
1	private	Mrs Georgia	Kingsley		11/19/1875	540000	high	full	no
2	private	Demetria	Harmon		08/19/2020	270000	high	full	yes
3	private	Homer	Carr		10/12/1984	180000	medium	full	no
4	company			FreeSeas Inc.					
5	private	Littleton	Dean		01/22/1957	135000	medium	full	no
6	private	Aliana	Snider		01/30/1976	112500	medium	reduced	yes
7	private	Ralph	Jones		06/28/1953	135000	medium	full	no
8	private	Jimmy	Jordahl		04/21/1958	67500	low	reduced	no
9	private	Amos	Miller		01/23/1982	126000	medium	full	no
10	company			Peregrine Pharmaceuticals Inc.					
11	private	Elgin	Howell		02/24/1952	81000	low	reduced	no
12	private	Adriana	Bailey		07/09/1980	90000	low	reduced	no
13	private	Jamin	Sakaguchi		08/05/1971	90000	low	reduced	no
14	private	Charles	Partin		10/27/1977	90000	low	reduced	yes
15	private	Jeanetta	Clark		08/03/1966	292500	high	full	no

Measure of Anomaly for field First name per Datapoint



Demo
First name

SAMPLE DATA

DQ with Autoencoders | Demo | Info

Select Field for Anomaly Detection

- (•) First name
- ( ) Last name
- ( ) Company name
- ( ) Birth date
- ( ) Income
- ( ) Tax class
- ( ) Revenue

Analyze | Reset Data

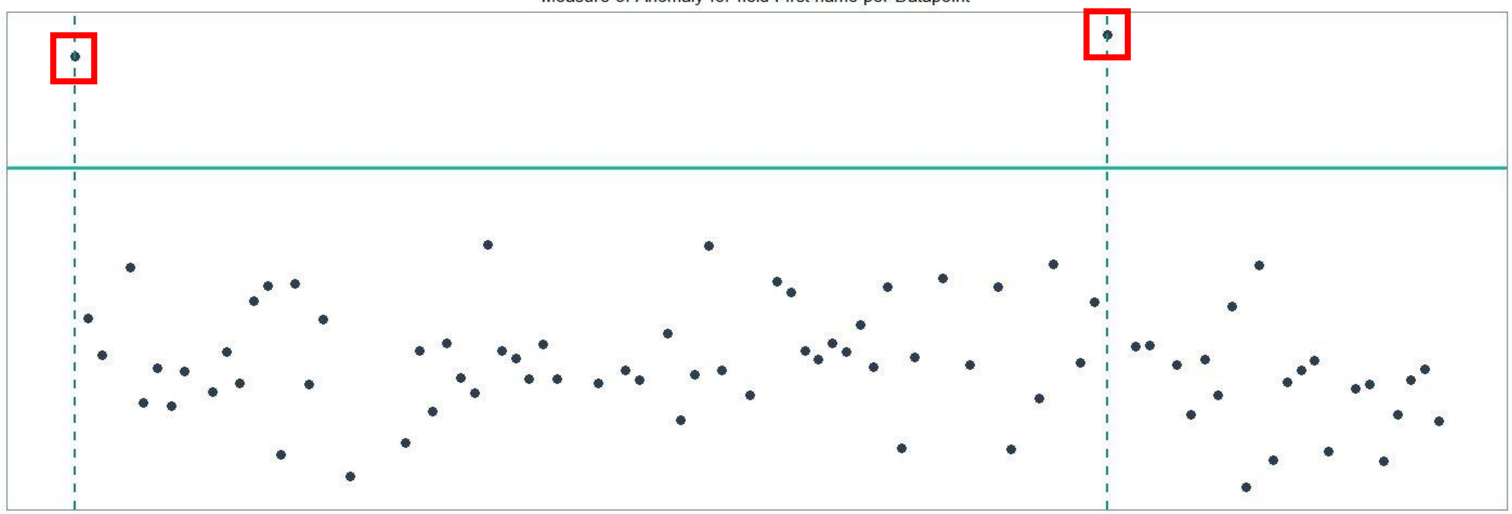
KPIs for selected field

True Positives	True Negatives
2 REMOVE FILTER	98 FILTER TABLE

False Positives	False Negatives
0 FILTER TABLE	0 FILTER TABLE

	Customer type	First name	Last name	Company name	Birth date	Income	Income class	Tax class	Children	Education	Revenue	Emplo
1	private	Mrs Georgia	Kingsley		11/19/1875	540000	high	full	no	Higher education		
76	private	James (the incredible)	Hamilton		04/21/1961	135000	medium	full	no	Lower secondary		

Measure of Anomaly for field First name per Datapoint



Demo

Revenue

SAMPLE DATA

DQ with Autoencoders | Demo | Info

Select Field for Anomaly Detection

- ( ) First name
- ( ) Last name
- ( ) Company name
- ( ) Birth date
- ( ) Income
- ( ) Tax class
- (•) Revenue

Analyze | Reset Data

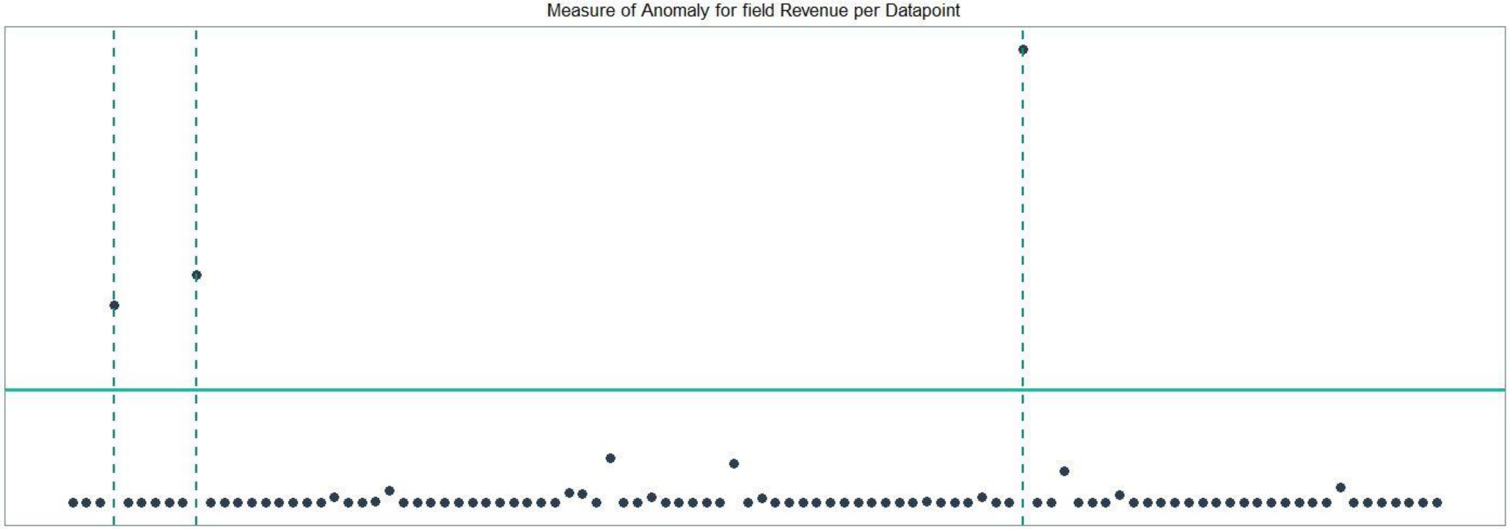
	Company name	Birth date	Income	Income class	Tax class	Children	Education	Revenue	Employees	Credit
1		11/19/1875	540000	high	full	no	Higher education			900000
2		08/19/2020	270000	high	full	yes	Incomplete higher			1546020
3		10/12/1984	180000	medium	full	no	Secondary / secondary special			1029658.5
4								3092300	31	1631698
5		01/22/1957	135000	medium	full	no	Higher education			521280
6		01/30/1976	112500	medium	reduced	yes	Higher education			229500
7		06/28/1953	135000	medium	full	no	Higher education			343800
8		04/21/1958	67500	low	reduced	no	Secondary / secondary special			563269.5
9		01/23/1982	126000	medium	full	no	Lower secondary			916470
10	ceuticals Inc.							1400258	1400	10000000
11		02/24/1952	81000	low	reduced	no	Secondary / secondary special			132444
12		07/09/1980	90000	low	reduced	no	Lower secondary			225000
13		08/05/1971	90000	low	reduced	no	Secondary / secondary special			942300
14		10/27/1977	90000	low	reduced	yes	Higher education			50940
15		08/03/1966	292500	high	full	no	Secondary / secondary special			450000

KPIs for selected field

True Positives	True Negatives
3 FILTER TABLE	97 FILTER TABLE

False Positives	False Negatives
0 FILTER TABLE	0 FILTER TABLE

Measure of Anomaly for field Revenue per Datapoint



Demo

Revenue

SAMPLE DATA

DQ with Autoencoders | Demo | Info

Select Field for Anomaly Detection

- First name
- Last name
- Company name
- Birth date
- Income
- Tax class
- Revenue (selected)

Analyze | Reset Data

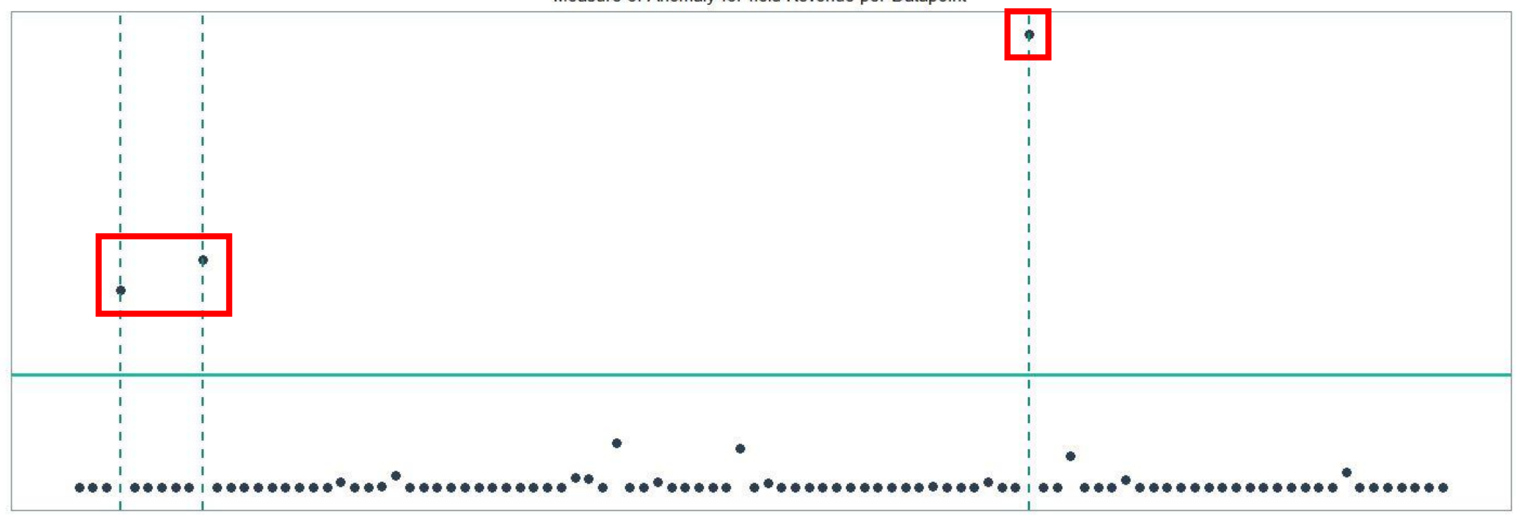
	e	First name	Last name	Company name	Birth date	Income	Income class	Tax class	Children	Education	Revenue	Employees	Credit
4				FreeSeas Inc.							3092300	31	1631698
10				Peregrine Pharmaceuticals Inc.							1400258	1400	10000000
70				Palmetto Bancshares, Inc. (SC)	02/15/1943						299504	1	2638759

KPIs for selected field

True Positives	True Negatives
3 REMOVE FILTER	97 FILTER TABLE

False Positives	False Negatives
0 FILTER TABLE	0 FILTER TABLE

Measure of Anomaly for field Revenue per Datapoint



Reusability of Pre-Processing and Model Setup

Setup per feature type

Type of variable	Pre-processing	Model
Character	One-hot encoding of characters	Variational autoencoders with LSTM cells
Categorical	One-hot encoding	Complete autoencoder with regularization
Date	Numerical features from digits	Complete autoencoder with regularization
	Normalization	
Numerical	Normalization	Undercomplete autoencoder with custom loss

Talk Outline

1. What's Wrong with Data Quality?
2. Error Detection using ML
3. Demo & Features
4. **Error Remediation using RPA**

Identifying Where DQ Exception Occurred

Searching for Correct Data using the ID

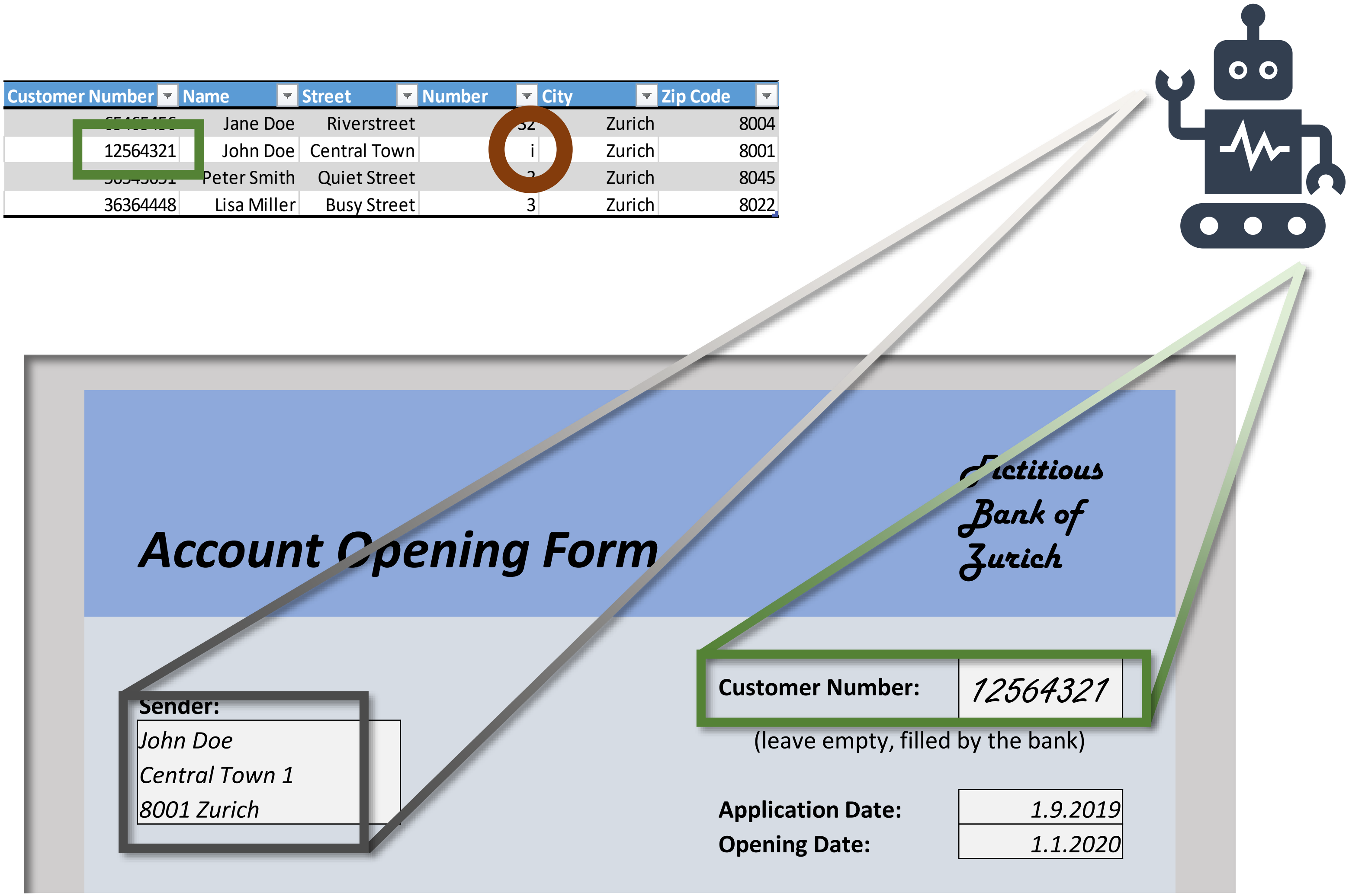
Example: CRM data at a bank

Customer Number	Name	Street	Number	City	Zip Code
65465456	Jane Doe	Riverstreet	32	Zurich	8004
12564321	John Doe	Central Town	i	Zurich	8001
36543651	Peter Smith	Quiet Street	2	Zurich	8045
36364448	Lisa Miller	Busy Street	3	Zurich	8022

AUTOENCODER

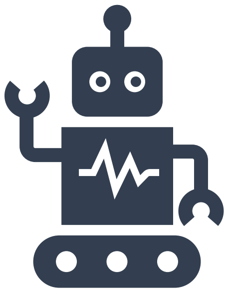
Finding Correct Data via “Intelligent” Robots

Computer Vision or NLP Help Find and Read Correct Data



Remediation Proposal

RPA Robot Proposes DQ Exception Correction

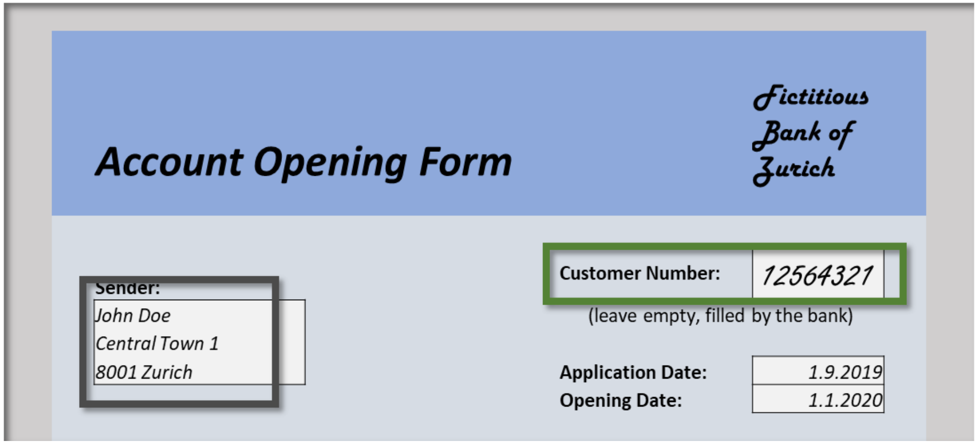


SUGGESTS THE CHANGE

Customer Number	Name	Street	Number	City	Zip Code
[illegible]	Jane Doe	Riverstreet	32	Zurich	8004
12564321	John Doe	Central Town	i	Zurich	8001
[illegible]	Peter Smith	Quiet Street	2	Zurich	8045
36364448	Lisa Miller	Busy Street	3	Zurich	8022

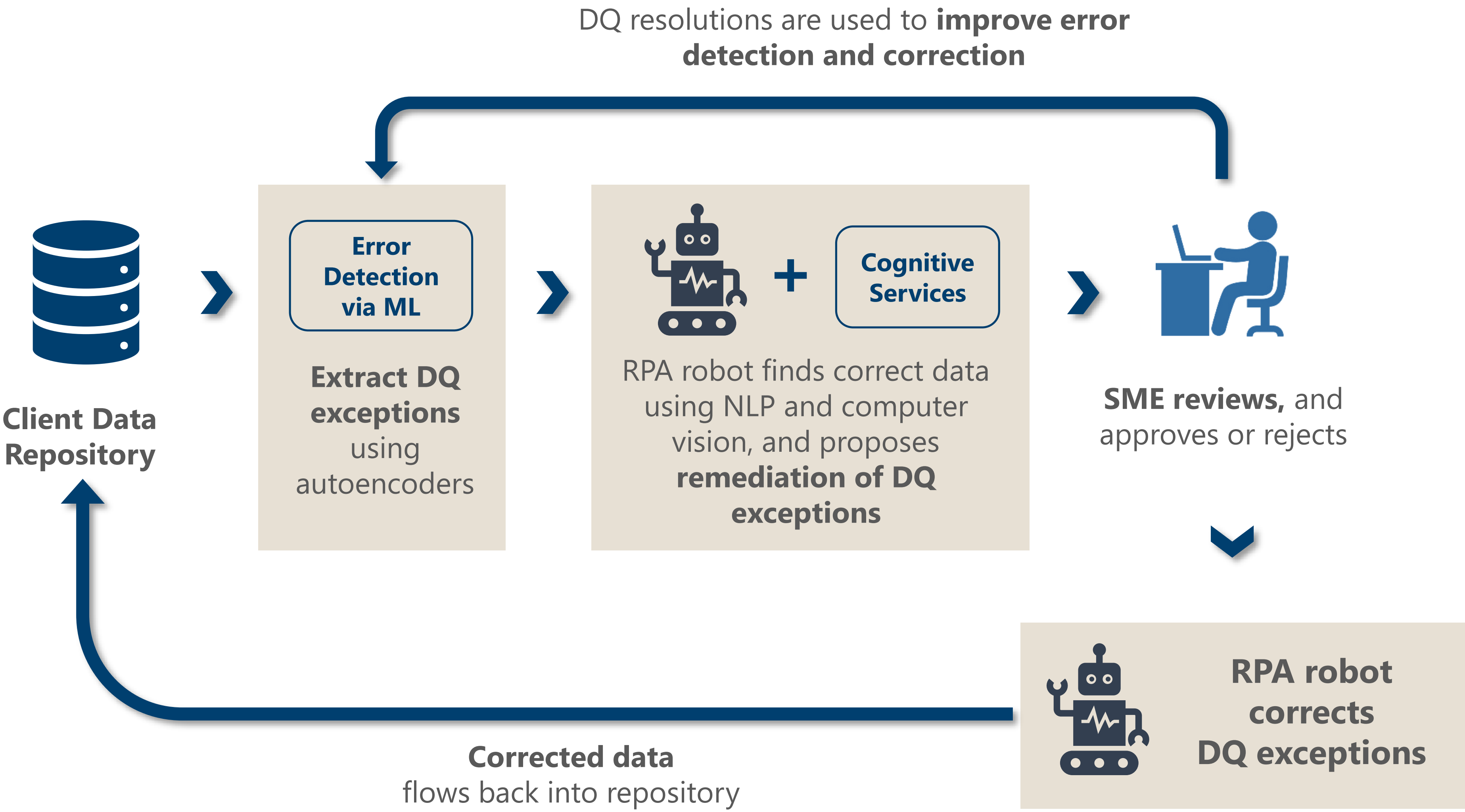
Customer Number	Name	Street	Number	City	Zip Code
[illegible]	Jane Doe	Riverstreet	32	Zurich	8004
12564321	John Doe	Central Town	1	Zurich	8001
[illegible]	Peter Smith	Quiet Street	2	Zurich	8045
36364448	Lisa Miller	Busy Street	3	Zurich	8022

BASED ON



Automation of Data Quality Remediation

Process Overview



Key Findings from Projects and Experience

Autoencoders

- **Extension:** ML can **replicate** rule-based DQ checks **and** find **new** errors
- **Multivariate relationships:** Detection of interdependencies
- **Unsupervised learning!** Training data quality matters
- **High reusability:** Only one-time customization effort per data type

RPA

- **Cost savings:** Automate interactions with existing IT infrastructure
- **High scalability:** Operations can be performed in parallel
- **Competency:** Subject Matter Experts can focus on value-adding tasks