wealthport

Automated Data Curation at Scale

Bernhard Bicher (CEO)
Dr. Noah S. Bieler (Principal Data Scientist)

Winterthur, 12th of June 2015

Data Preparation Today

Data Scientists spend up to 80% of their time preparing data.

Data Preparation is no self-service activity without IT involvement.

Semi-automatic integration of more than 25 data sources is unfeasible.

Data origins and lineage are frequently lost during processing.

Three Options

Manual ✖

Hire work force

Unreliable

Not sustainable

Expensive

Rule-based ✖

ETL

High Maintenance

Completeness

Needs expensive IT guy

Probabilistic ✔

Use statistics, NLP, ML

Choosing and combining the right algorithms

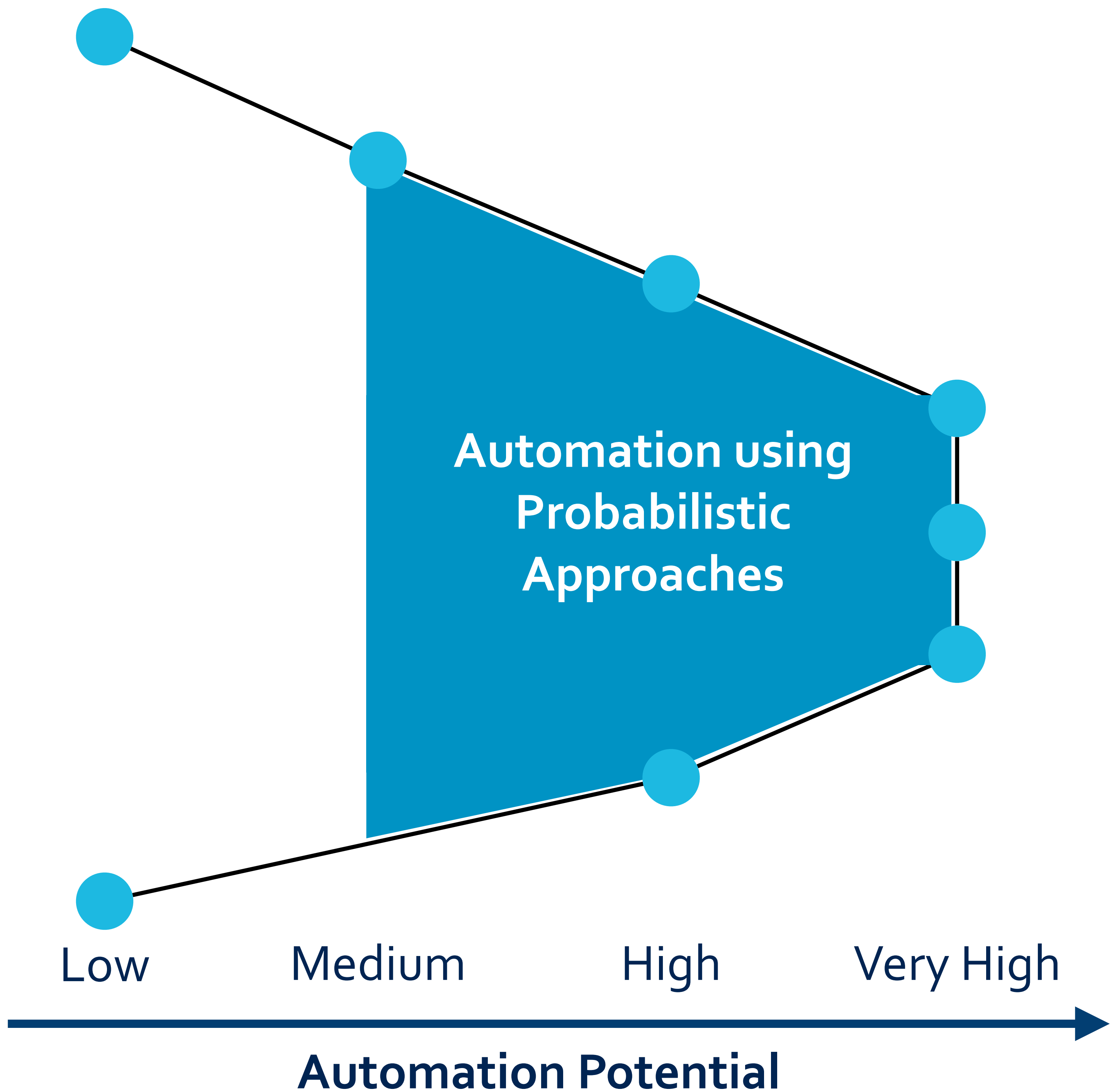
Only approximate results

ETL Extract Transform Load
NLP Natural Language Processing
ML Machine Learning

The Art of Data Integration

Identify Sources
Profile Data
Clean Data
Normalise Data
Identify Joins
Entity Resolution
Deduplication
Post-Processing

Integrated Data



wealthport

Probabilistic Methods and Approaches

Identify Sources

Profile Data Outlier Detection, Authoritative Data, Type Detection

Clean Data

Encoding Errors Fixing, Pattern Mining, Column Swap

Normalise Data

Probability Distribution, Entropy Measurement

Identify Joins

Naive vs. Advanced ML Approaches

Entity Resolution

Computational Complexity Reduction

Deduplication

Post-Processing

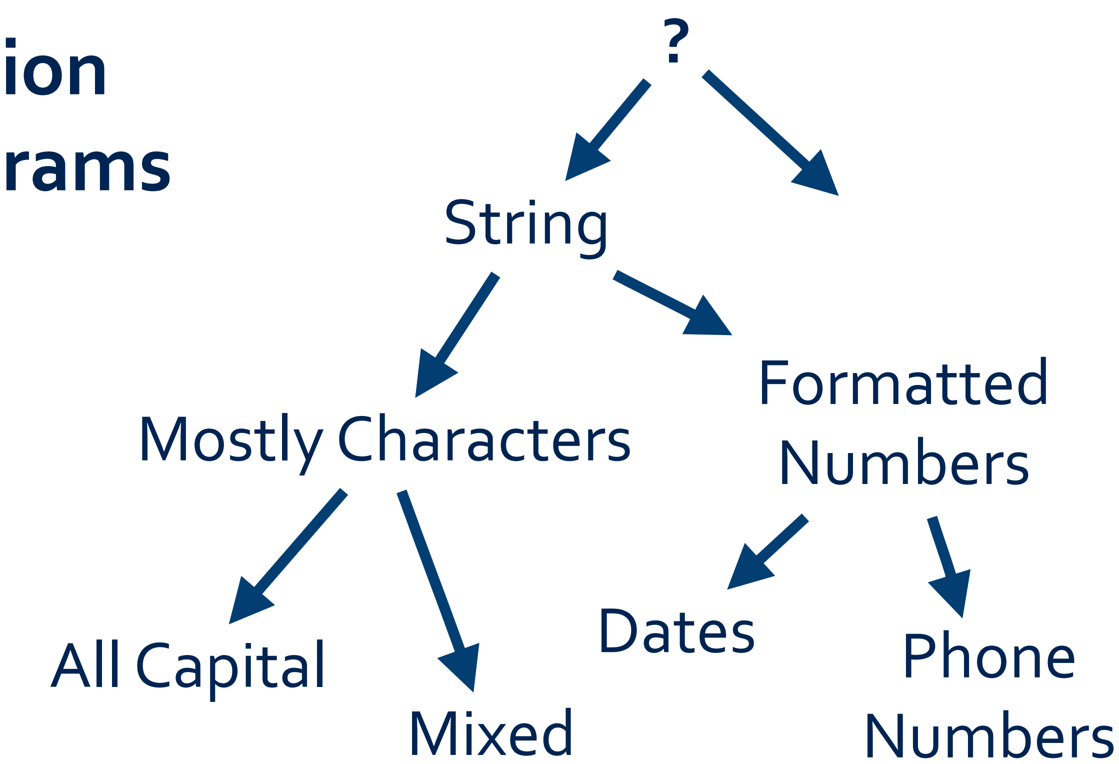
Profile Data

Example: Probabilistic Schema Detection

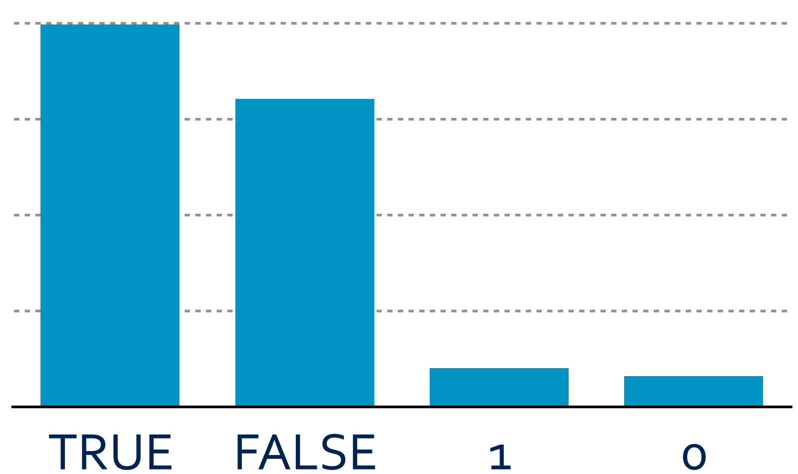
First Name	Last Name	Premium	City	Country
Hans	Müller	TRUE	Winterthur	N/A
Hans	Mueller	1	Winterthur	CH
Jan	Muster	FALSE	Windisch	CH

Identify Missing Values

Content Detection using Decision Trees



Outlier Detection based on Histograms



Profiling based on Authoritative Data

Last Name
Müller
Mundt
Muster
...

Clean, Normalise and Impute Data

First Name	Last Name	Premium	City	Country
Max	Morgenthal	TRUE	Winterthur	
Hans	M♦ller	TRUE	Winterthur	CH
Hans	Mueller	1	CH	Winterthur
Jan	Muster	FALSE	Windisch	CHE

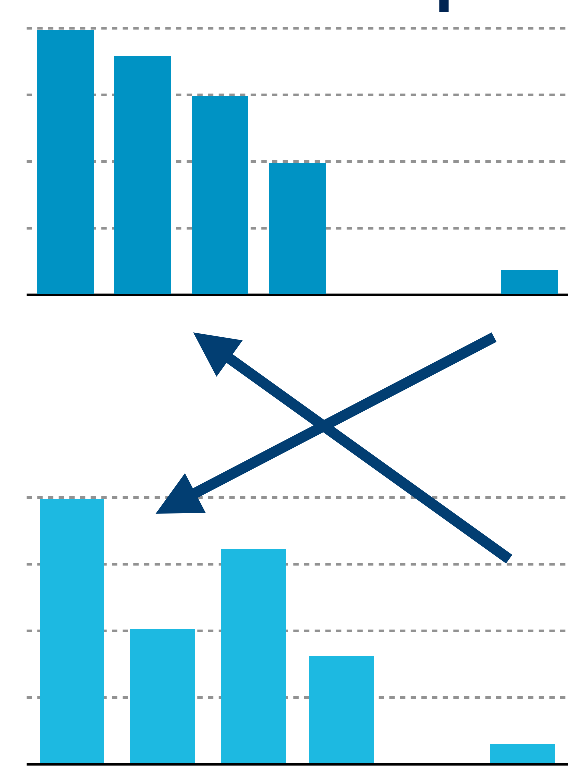
Pattern Mining

```
city == "Winterthur"
implies Country = "CH"
```

Fix Encoding Errors

M♦ller → Müller

Column Swap



Normalisation according to a Synonym Table

ISO2	ISO3	Name	....
CH	CHE	Schweiz	
DE	DEU	Deutschland	
FR	FRA	Frankreich	

Identify Join Columns

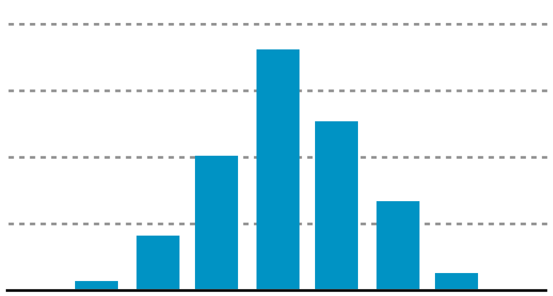
Comparison of Probability Distribution

Datasilo 1

FirstName	ClientID	Premium	....
Martin	1028934-1	TRUE	
Sara	7462946-5	TRUE	
Anna	9471991-3	FALSE	

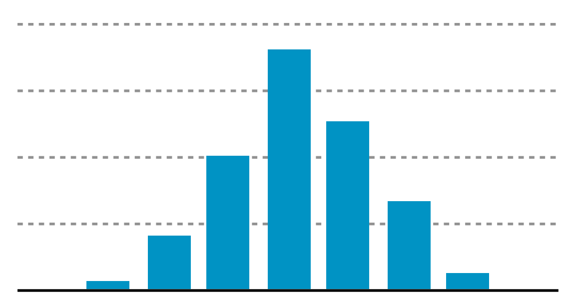
Datasilo 2

CID	ProductName	ProductID	....
C-9471991	Monitor LCD	6413	
C-7462946	Mouse Laser	5433	
C-1028934	Keyboard QWERTY	961	

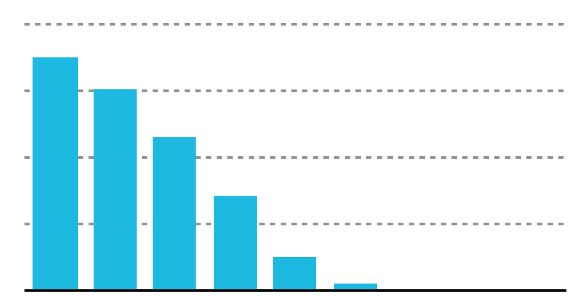


μ_1

similiar



μ_1'



μ_2

wealthport

Entity Resolution & Deduplication

Naive Approach

First Name	Last Name	Premium	City	Country
Hans	Müller	TRUE	Winterthur	
Hans	Mueller	1	Winterthur	CH
Jan	Muster	FALSE	Windisch	CH

All weights w_i are the same.

$w_i = \{0.2, 0.2, 0.2, 0.2, 0.2\}$

$$s = \sum_i w_i s_i$$

Advanced Approach

First Name	Last Name	Premium	City	Country
Hans	Müller	TRUE	Winterthur	CH
Hans	Müller	TRUE	Winterthur	CH
Jan	Muster	FALSE	Windisch	CH

Adapt the weights w_i using ML and optimise similarity calculations.

$w_i = \{0.3, 0.3, 0.1, 0.2, 0.1\}$

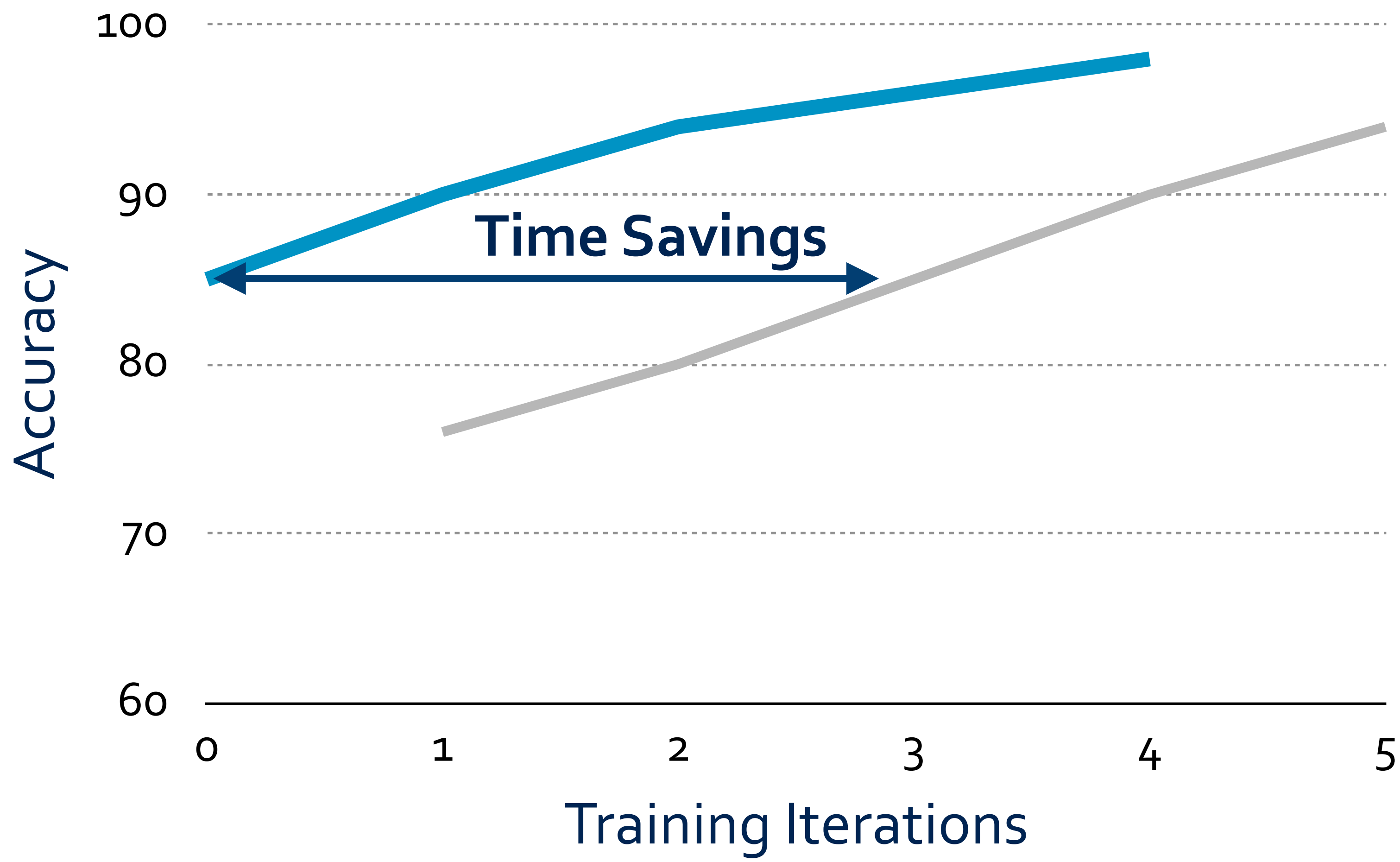
De-Noising and normalisation helps to compare entities.

User feedback is incorporated into the estimate of the weights $\{w_i\}$ using ML.

Cleaned data

wealthport

Example: Deduplication of 1M records



ML & pre-estimated Weights

Pure ML Approach
3 iterations with 1'000 manual feedback à 30 sec required to achieve same accuracy

→ Effort: 3.1 days

Better out-of-the-box precision using ML and pre-estimated weights.

Start by initialising weights according to the column content.

For some cases, this can even eliminate the need for training at all.

Tackling Complexity in Deduplication

Clustering

$n = 10^6$

$k = 10^2$

$m = 50$

$n^2 \longrightarrow 10^{12}$

$0.5n^2 \longrightarrow 0.5 \cdot 10^{12}$

$k \cdot n \cdot m + 0.5 \cdot k(n/k)^2 \longrightarrow 10^{10}$

n Number of data records

k Number of clusters

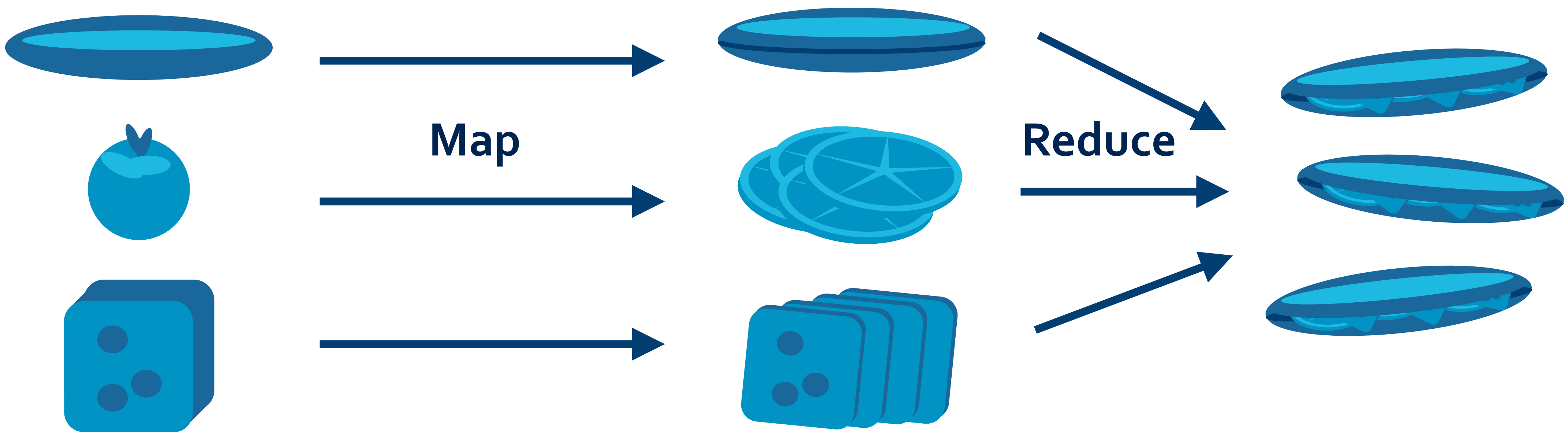
m Number of iterations

Better scalability leads to faster execution.

Higher data locality, a “triangle” can run on a single node.

State-of-the-Art Infrastructure

Map-Reduce style using Apache Spark



Scalable: runs on a single Laptop as well as on a 10k-node Cluster.

Programmed in Scala: functional and object-oriented.

Supports streaming, and provides MLlib and GraphX for machine learning and graph algorithms.

Summary

1. **Probabilistic methods save precious time**
   Decide on trade-off between fast data integration and precision

2. **Leverage machine learning**
   Use business expert feedback to improve system precision and degree of automation.

3. **Broad data analysis**
   Mine over 100 instead of just 25 data sources.

Wealthport AG
Rütistrasse 16
CH-8952 Schlieren
+41 76 420 67 68

info@wealthport.ch
www.wealthport.ch
Twitter: @wealthport

Join us at www.meetup.com/spark-zurich!

Empowering organisations to unlock their wealth of data