Scalable Machine Learning Pipelines for Enterprise Data Mastering

Big Data Conference Europe 2020

Topics we will cover today

- Data Mastering and its needs

Topics we will cover today

- Data Mastering and its needs
- Challenges in data mastering

Topics we will cover today

- Data Mastering and its needs
- Challenges in data mastering
- Machine Learning for data mastering

Data Mastering – Why and What?

Customer Analytics

is a process used to build

a

single, accurate and trusted view

of

customer behavior

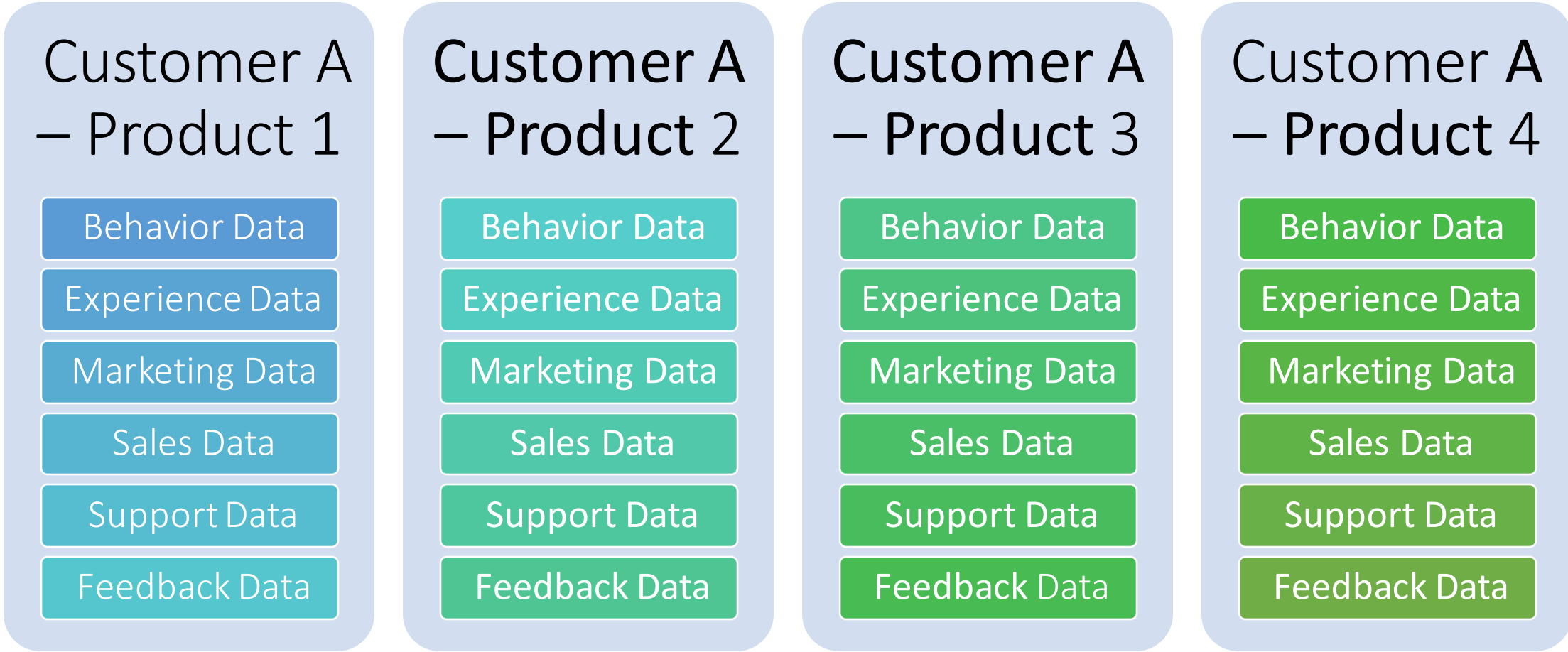
Customer Analytics

- Customer Life Time Value
- Net Promoter Score
- Customer Satisfaction
- Customer Effort Score

Customer Analytics Outcomes

- Remove process bottlenecks
- Timely and relevant messaging
- Improve brand perception and recall

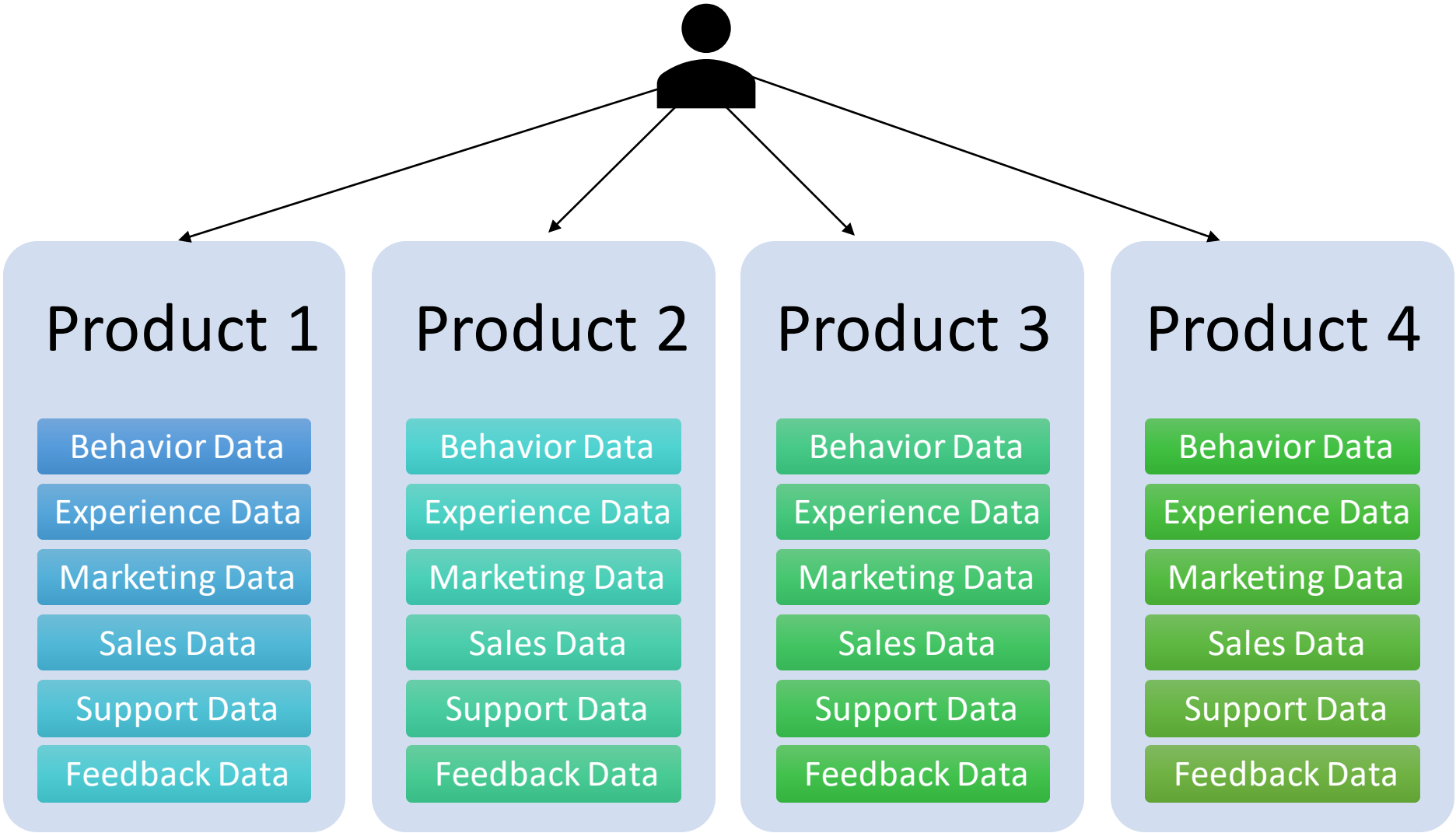
Customer Product Centric View



Data Mastering

- Identify core entities across different business processes

Customer Centric View



Suppliers in Procurement

- Vendor Registration
- Requirements Specification
- Inspection
- Contract
- Delivery

Supplier Data Mastering

- Trusted unified view of supplier across procurement systems
- Link to experience, behavior, contract, legal, purchase and other data

Supplies Data Mastering

The
Whole
is
greater than
the
sum
of its
parts

Data Mastering

Integrating the data silos to understand

- Customer
- Product
- Vendor
- Partners
- Supplies
- Employees…..

Data Mastering – Benefits

- Increase Revenue
- Reduce Costs
- Reduce risks
- Stay compliant

Challenges in Data Mastering

Business Challenges

- Ownership
- Coordination
- Governance
- Processes

Technical Challenges in Data Mastering

1. Lots of source systems and formats

Technical Challenges in Data Mastering

1. Lots of source systems and formats
2. Different schemas

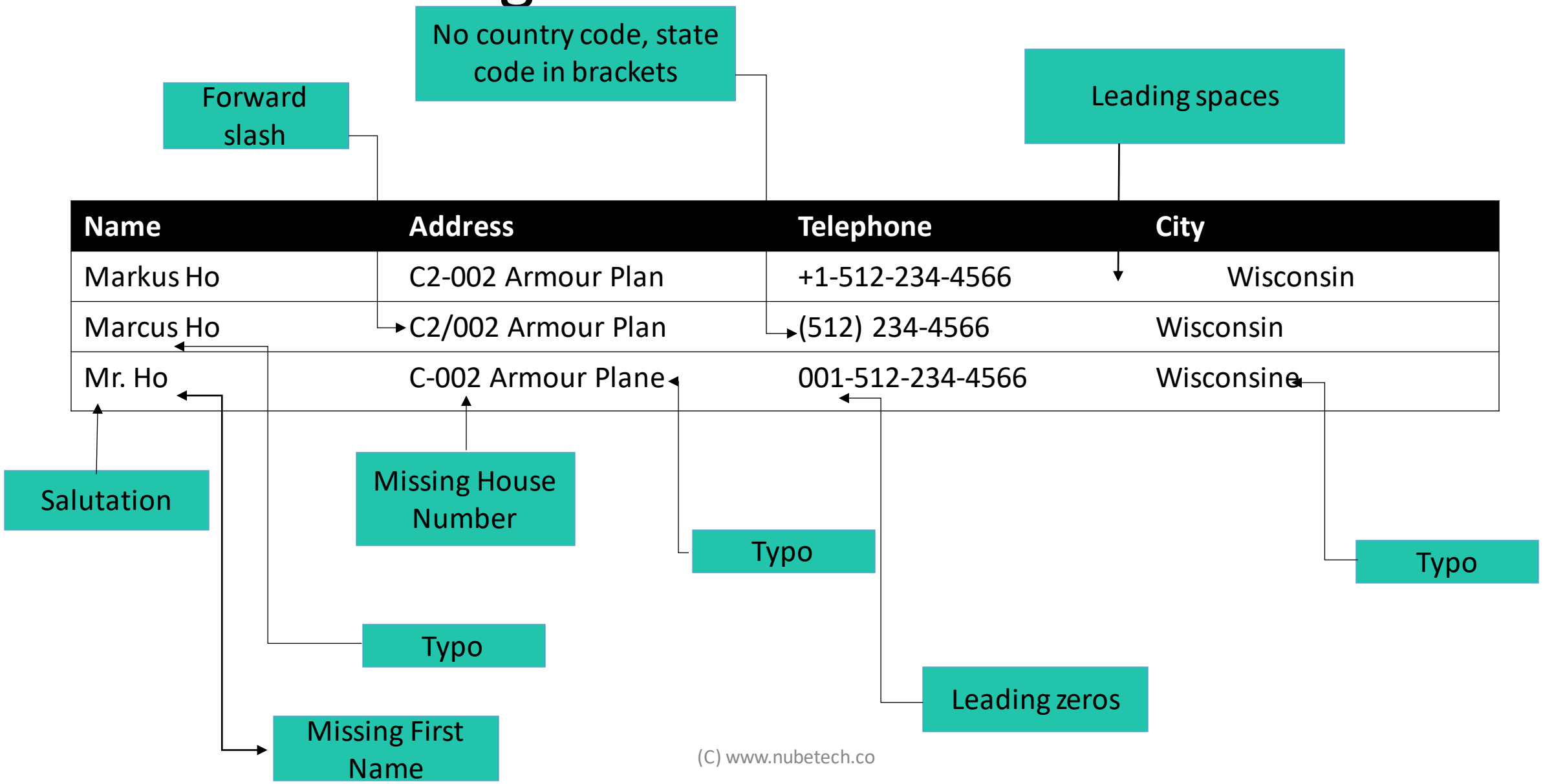
Technical Challenges in Data Mastering

1. Lots of source systems and formats
2. Different schemas
3. Record level variations

Technical Challenges in Data Mastering

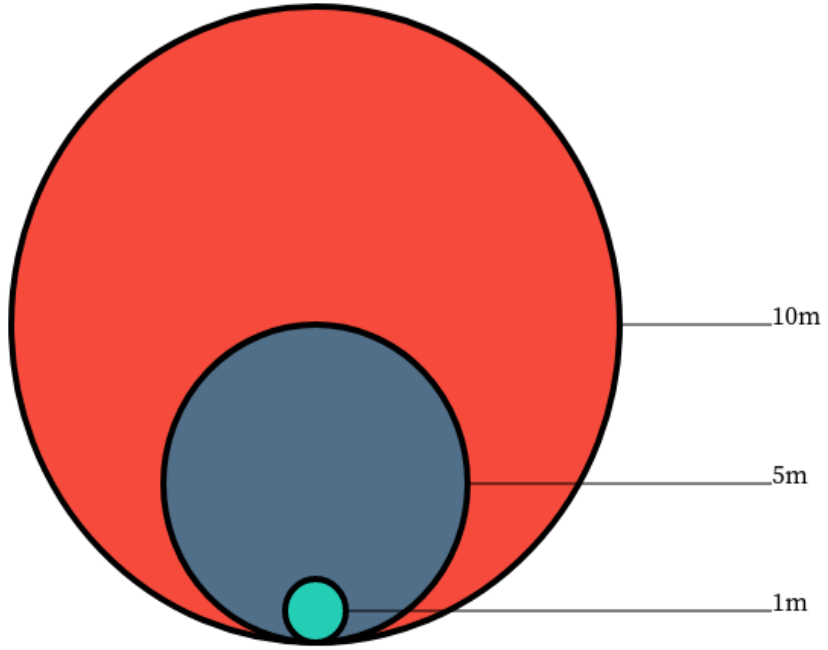
1. Lots of source systems and formats
2. Different schemas
3. Record level variations
4. Multifaceted relationships

Data Matching



Data Matching At Scale

Number of Records	Comparisons	Number of Comparisons
10,000	(10,000*9,999)/2	49,995,000
100,000	(100,000*99,999)/2	4,999,950,000
1,000,000	(1,000,000*999,999)/2	499,999,500,000



Data Mastering has a scale problem

- Variety of mastered entities
- Sheer number of records
- Number of data sources

Applying AI to Solve Data Mastering Challenges

Applying AI

- Get some training data to learn from.
- Build representations from that data. .
- Build a model that fits the data.

Build a data pipeline which derives the features from the raw data and applies the model and gets the prediction.

Word of Caution

- Is the problem really an AI problem?
- Is it painful and large enough to justify our effort and cost ?
- Is the data representative enough or something over which we can learn ?
- Do we have training data?
- Do we have the expertise and interest to manage production AI?

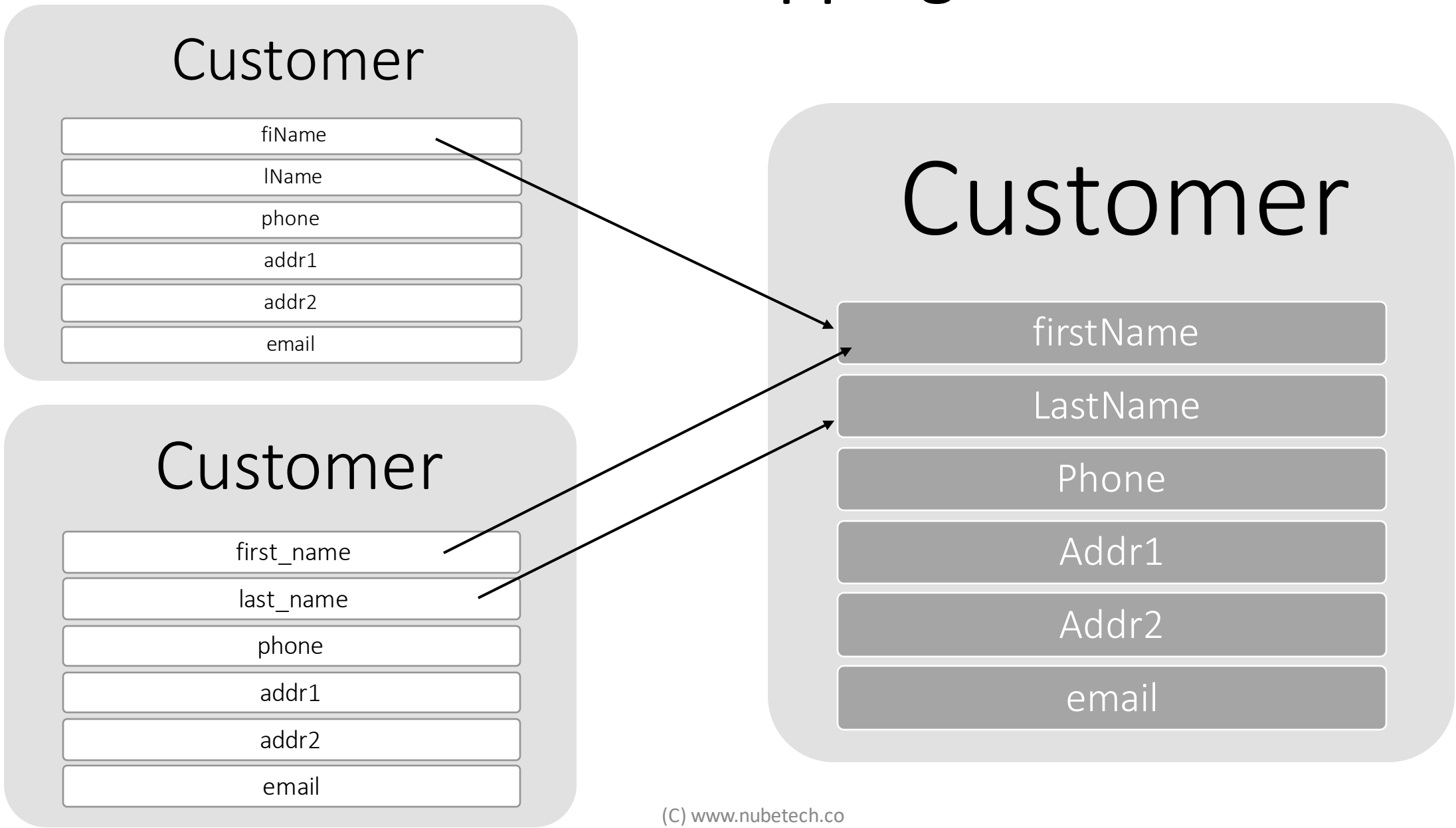
Training Data

- From previous efforts
- Active Learning
- Weak Supervision

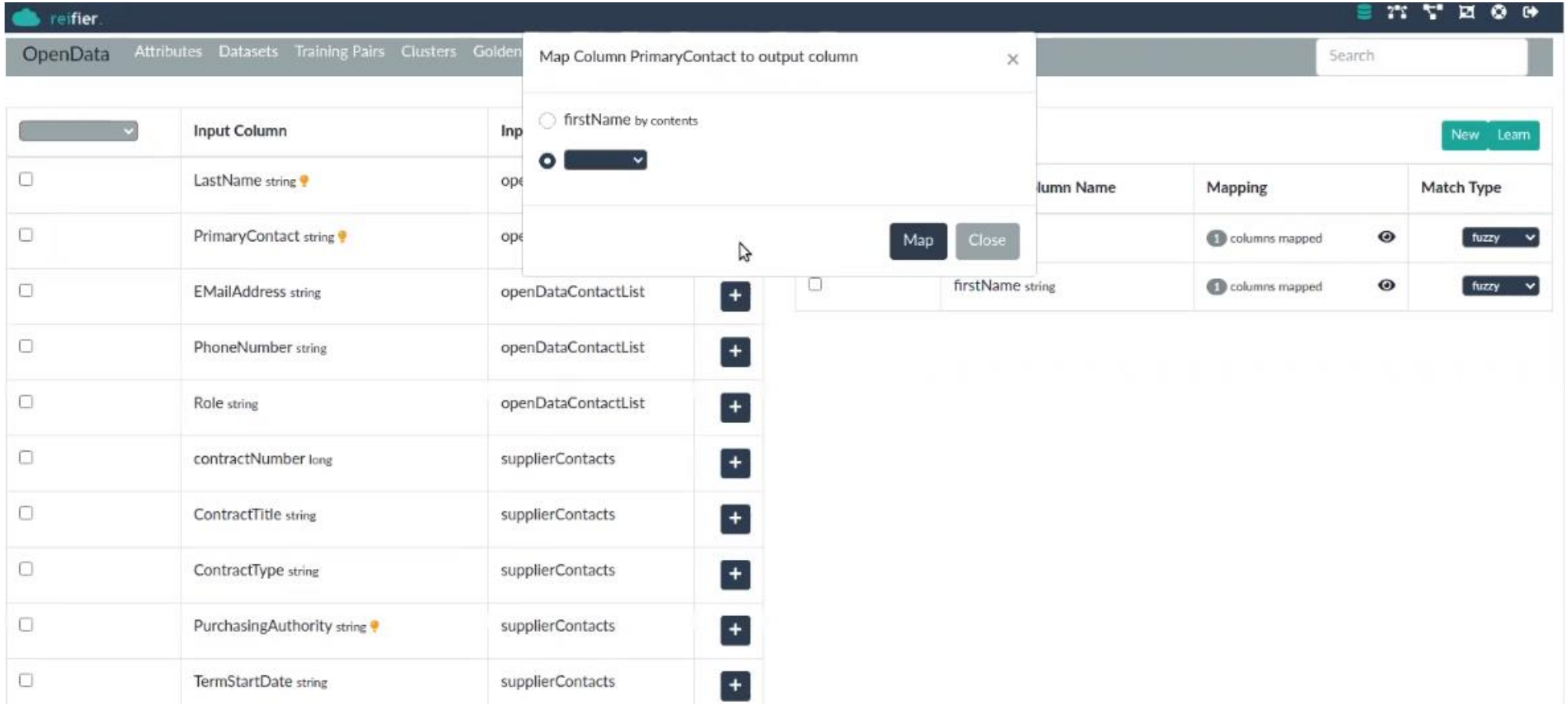
Data Mastering With AI At Scale

- Schema integration
- Data Matching/Deduplication/Entity consolidation
- Product and supplies classification
- Improving data quality

Schema Mapping



Schema Mapping with AI



Data Matching – Rule Based

- Data Profiling
- Data standardization
- Define Fuzzy Match Key
- Define Fuzzy Match Process
- Database Tuning

Data Matching With AI

- Clustering

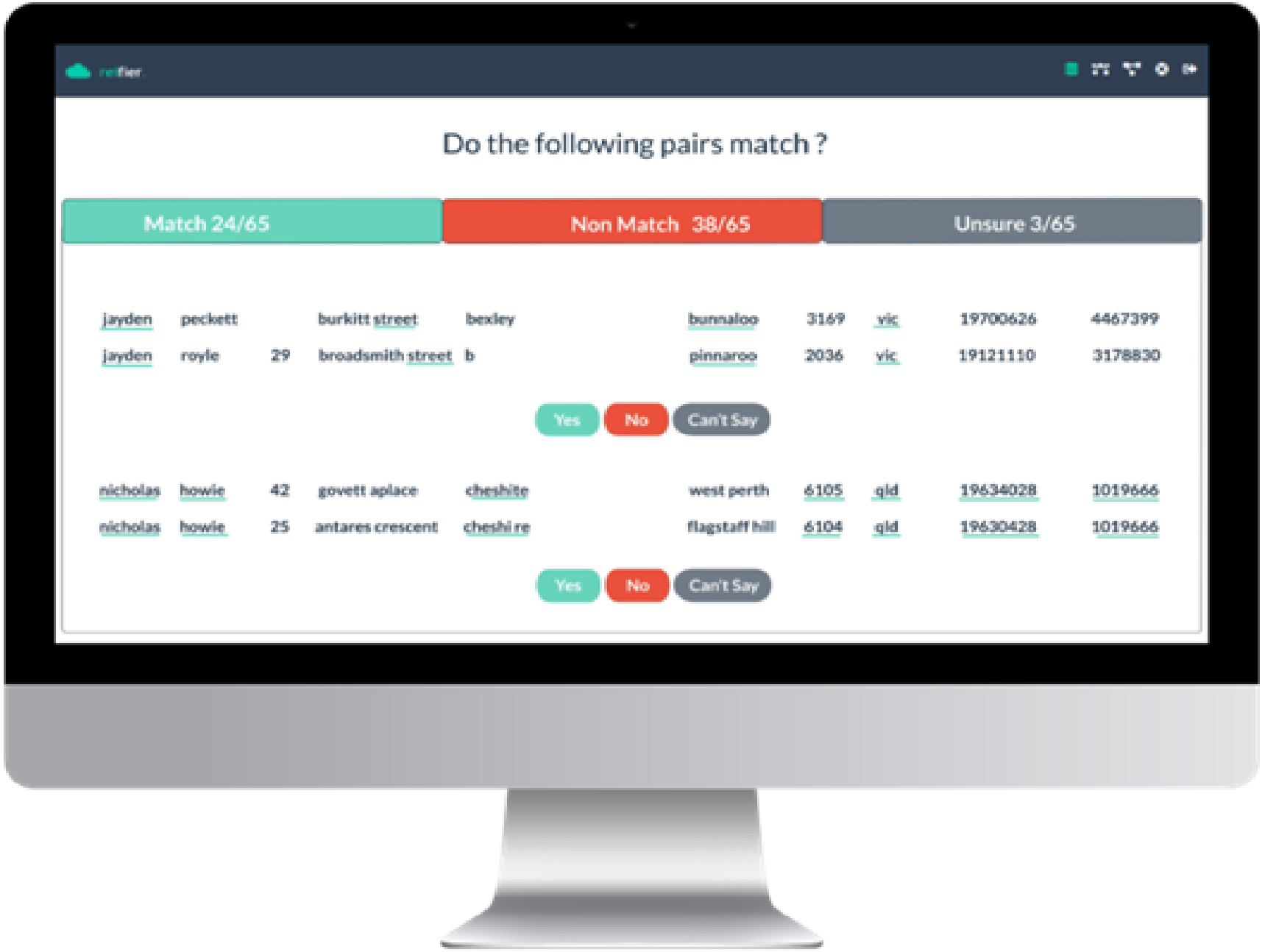
- Classification

- Active Learning

- Answer few yes no questions to build training data
- Tune feature weights and importance of columns based on user feedback

Data Matching With AI

- ~~Data Profiling~~
- ~~Data standardization~~
- ~~Define Fuzzy Match Key~~
- ~~Define Fuzzy Match Process~~
- ~~Database Tuning~~



Other areas

- Data Quality Anomaly Detection
- Automated Transformations
- Unstructured Mastering

Questions?

Thank you!

Get in touch with me at sonal@nubetech.co to discuss more!