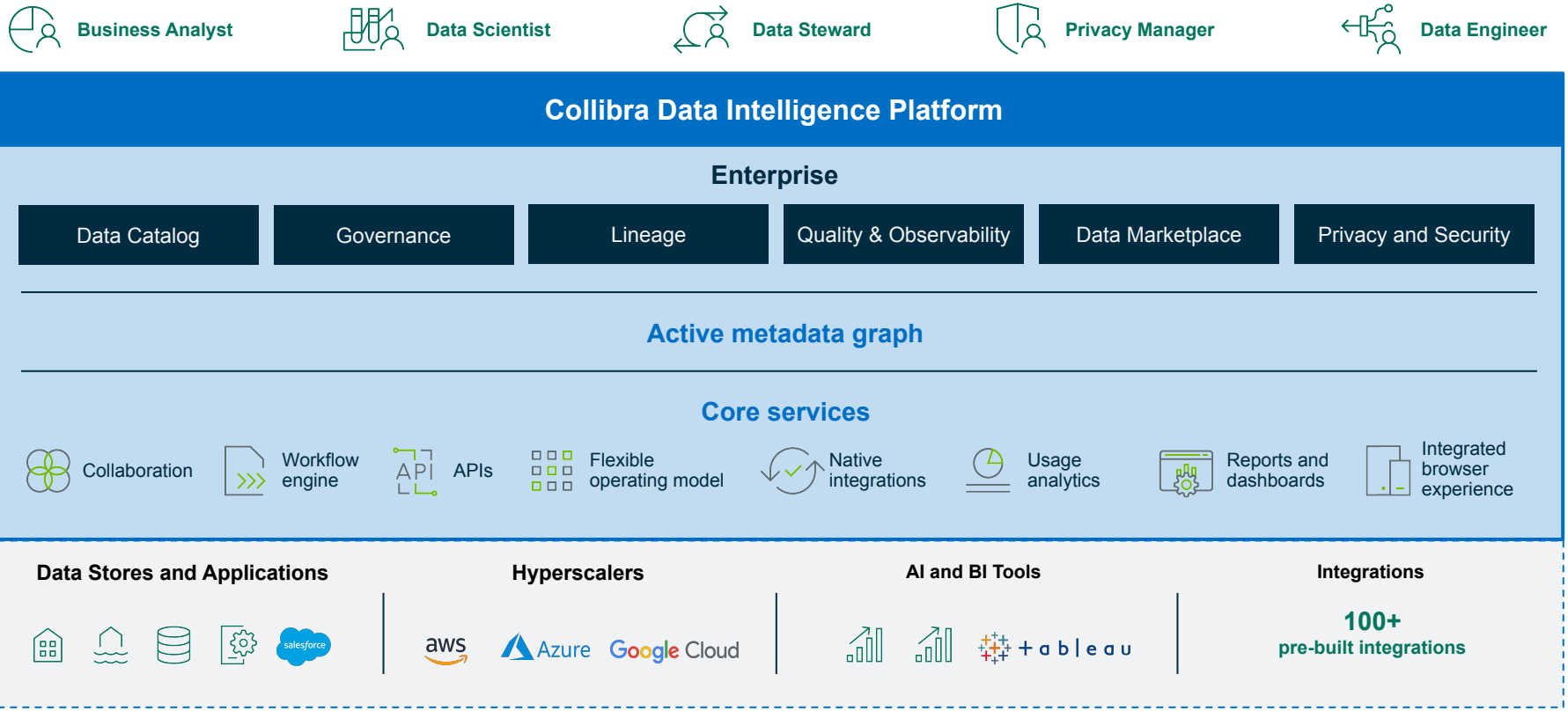
Using AI, Mathematics, and Statistics to Find Similar Data in Massive Data Ecosystems

Collibra

Eric Warner, PhD
Senior Manager, AI Engineering

What is Collibra? | Collibra is the system of engagement for data



Who? | Collibra Artificial Intelligence Team

Eric Warner
Senior AI Manager

Larry Hau
Senior Director, Product

Vicky Froyen
Staff Product Manager

Discovery

Gretel De Paepe
Principal Data Scientist

Delivery

Anna Filipiak
Staff AI Engineer

Kelsey Schuster
Senior AI Engineer

Lillian Neff
AI Engineer

Nick Evers
Senior AI Engineer

Operations

Cristian Caraballo
Senior Ops Eng

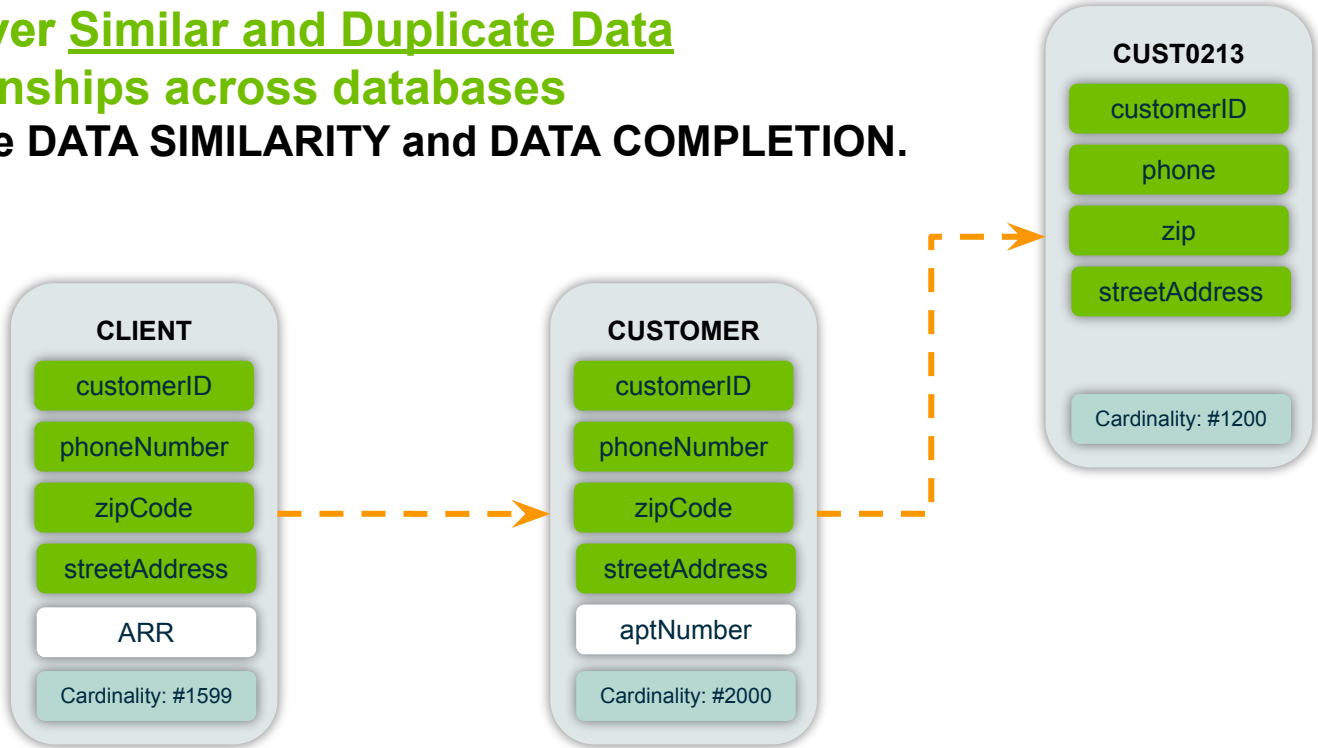
Charlie Ang
Staff Ops Eng

Collibra

What? | Data Similarity

Discover Similar and Duplicate Data relationships across databases

To drive DATA SIMILARITY and DATA COMPLETION.



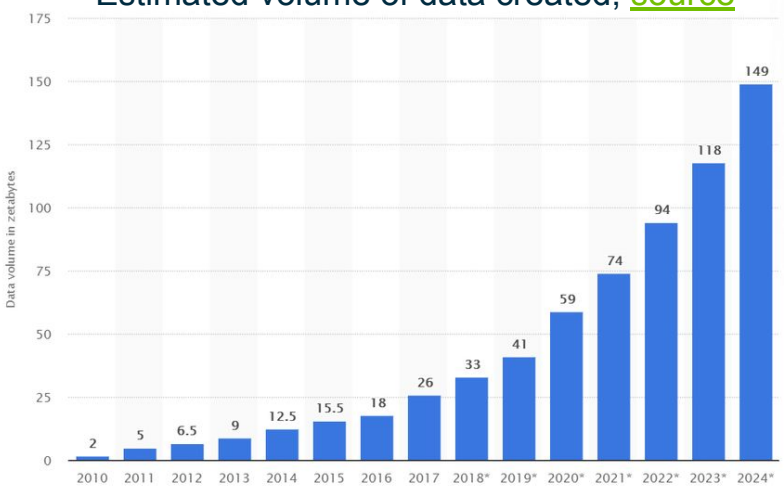
Why? | Data Similarity

Data is a $273T market[1]

A company with a thousand data workers wastes time equivalent to $5.7 million a year trying to find relevant data[2]

By 2026, 60% of GenAI applications will fail due to bad, wrong, or not enough data[3]

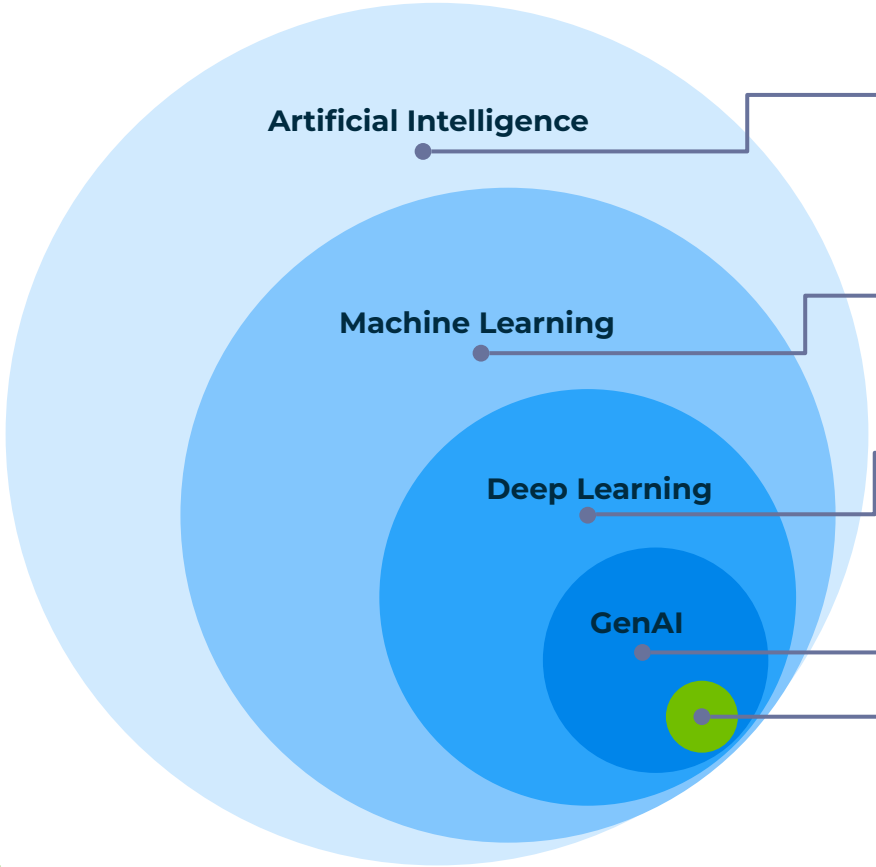
Estimated volume of data created, source



[1,2,3]Source: IDC

Collibra

AI | Riding the Tidal Wave of Hype



The theory and methods to **build machines that think and act like humans**

Ability for computers to **learn like humans**, i.e., spam filtering, fraud detection

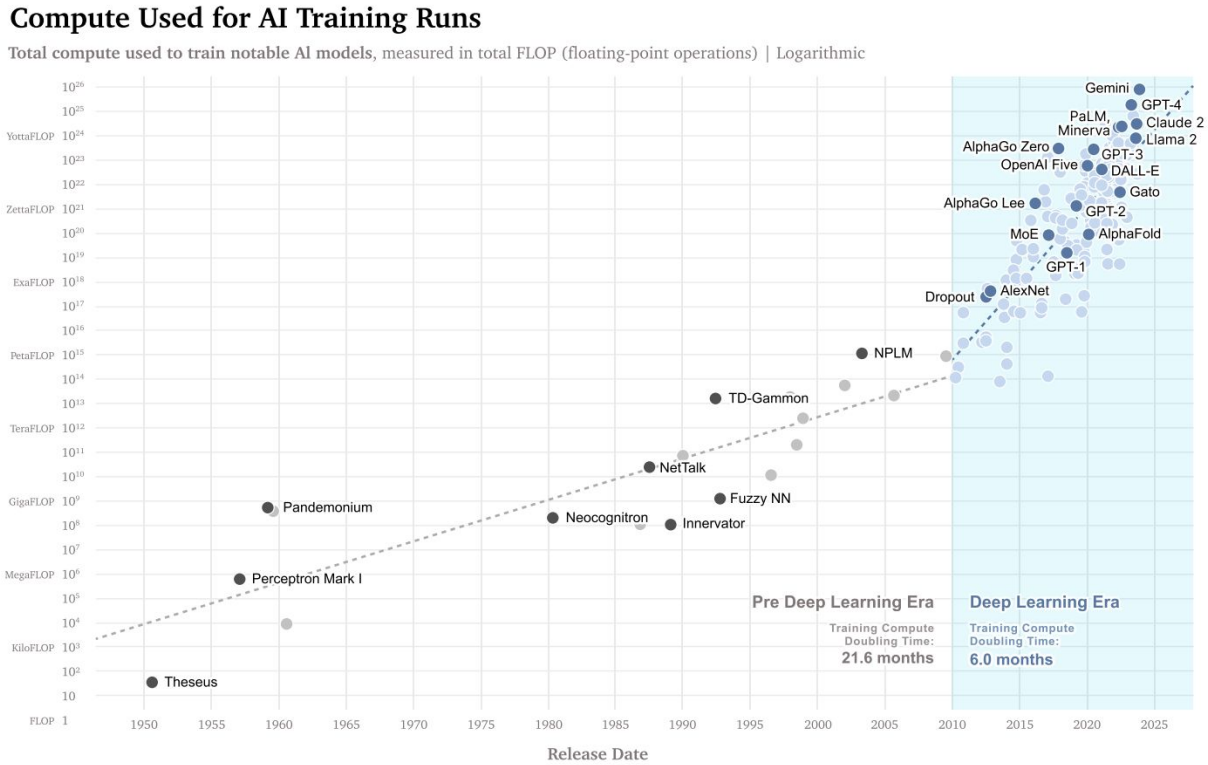
Mimic the human brain with many connections, i.e., transformers, image recognition, face recognition

Ability to generate new text, audio, images, etc. based on training data

ChatGPT & other LLMs, such as Google PaLM / Gemini / Titan ...

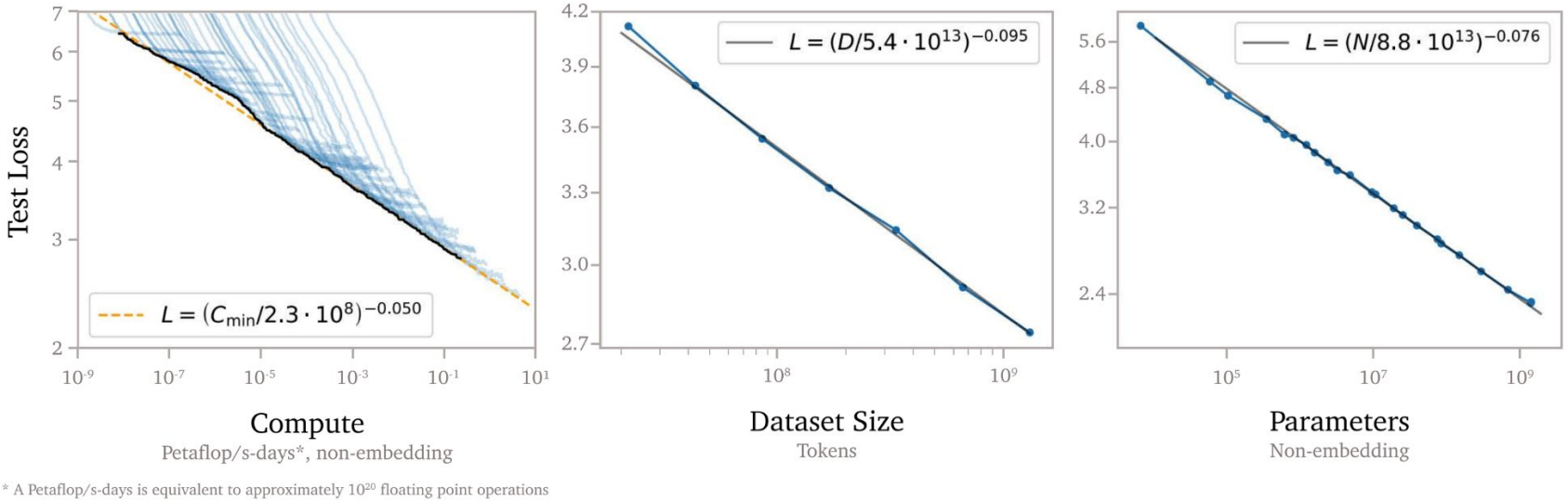
Collibra

LLMs do not come without cost…



[Source]

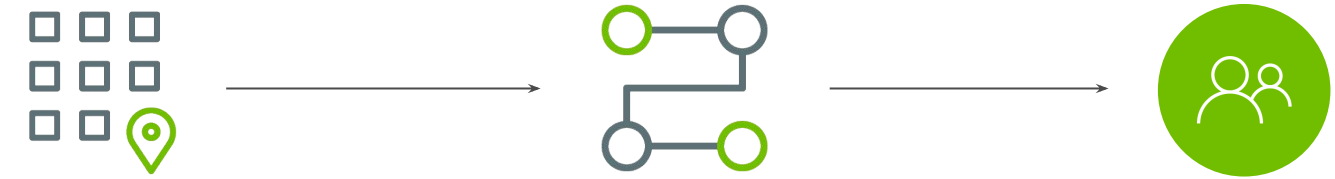
…or upside



* A Petaflop/s-days is equivalent to approximately 10^{20} floating point operations

Source

Navigating the Keys to Data Science

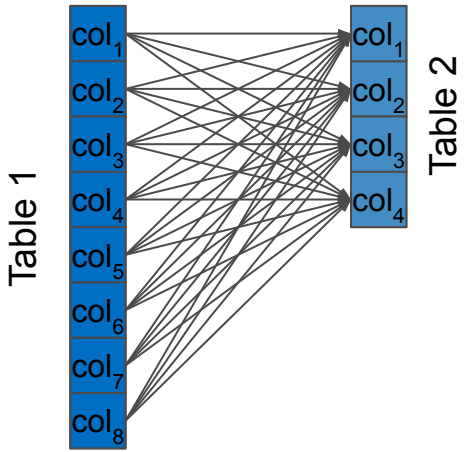


Collibra

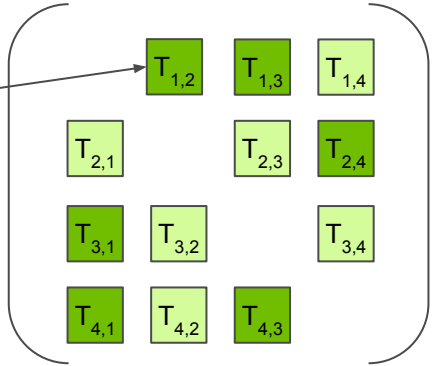
At a high level, how could this (inefficiently) work?

Algorithm Selection

Comparing an average of *n* columns per table and p entries per column…



… for *m* Tables



Compute Efficiency: $O(p^2m^2n^2)m^2)$*

Storage Requirements: 300.02 TB

(100,000 tables, 100 cols per table, 50 entries per column)

**250 quadrillion computations in the above example or ~79 years if GPU can do 100 million comparisons per second*

Collibra

Clearly, this is not feasible

So how are we actually achieving this?

Let’s start at the outcome

Constraints Drive Solutions

Directed Use-Case

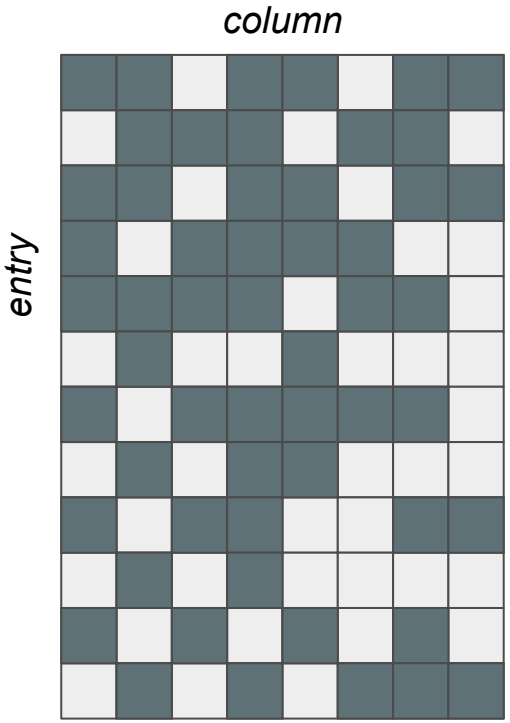
Challenge | Our algorithm

- Must run on customer hardware
- Cannot phone home to an internal or 3rd-party API
- Must actually be able to determine similar/duplicate data

Collibra

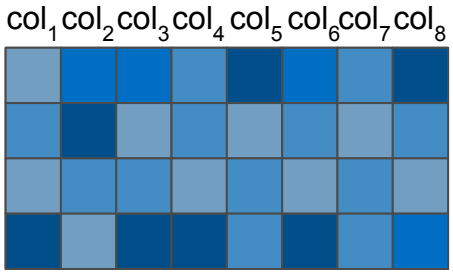
Data Source Constraints

Data Awareness



Given a series of columns…

…randomly sampled and streamed in batches…



…we must compress…

…and add noise…

..while achieving >90% accuracy. How??

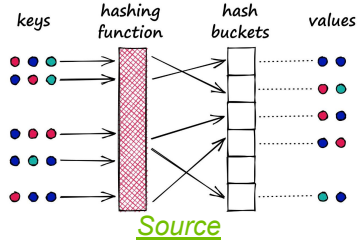
Collibra

PATENT PENDING

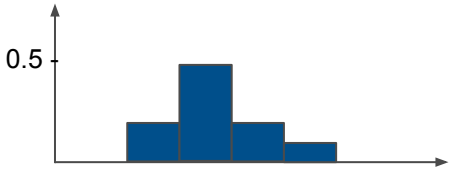
Compression | Our Secret Sauce to Performance

Algorithm Selection

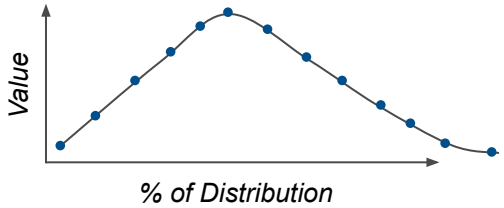
String/Date/Categorical Data
{‘MA’, ‘CT’, ‘MI’, ‘CO’, ‘MI’, ‘NH’}



Integer Data
{1, 2, 2, 2, 3, 2, 2, 3, 1, 4}



Float Data
{ -0.63, -0.41, 0.77, 0.08, 0.28, -0.96, 0.23, 0.06, -0.18, -0.07 }



Our schema aims to maintain as much information as possible while maintaining security and legal boundaries

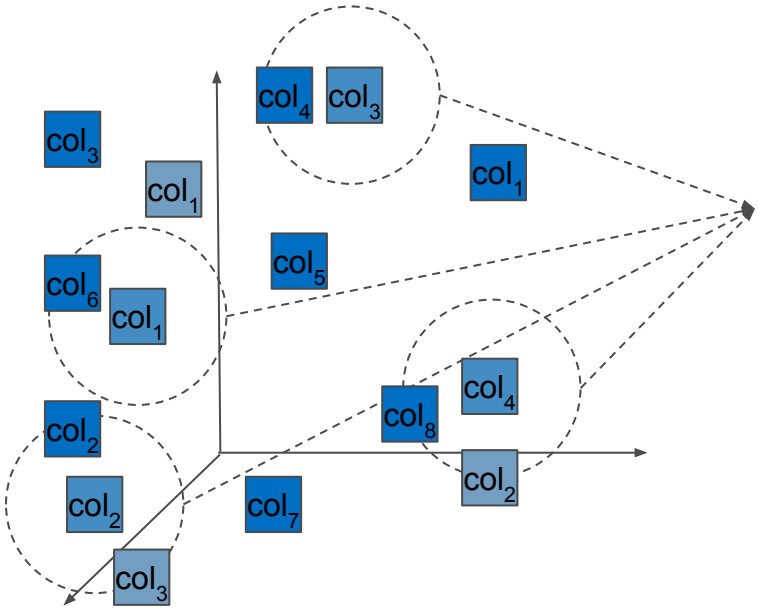
Collibra

PATENT PENDING

At a high level, how does this actually work?

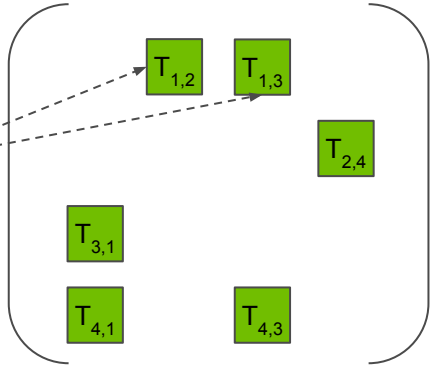
Algorithm Selection

Comparing an average of *n* columns per table and p entries per column…



Combiner

… for *m* Tables



Compute Efficiency: O(p*m*n+m)

Storage Requirements: 90.002 GB

(100,000 tables, 100 cols per table, 50 entries per column)

0.03% of brute-force

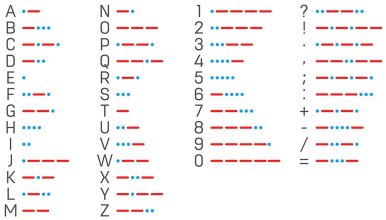
**~100 million computations in the above example if 20% of tables have a duplicate (0.00000004% of brute-force)*

Collibra

Compression

This is not new….

Morse Code (1840's)



source

Shannon Coding (1940's)

i	p_i	l_i	$\sum_{n=0}^{i-1} p_n$	Previous value in binary	Codeword for a_i
1	0.36	2	0.0	0.0000	00
2	0.18	3	0.36	0.0101...	010
3	0.18	3	0.54	0.1000...	100
4	0.12	4	0.72	0.1011...	1011
5	0.09	4	0.84	0.1101...	1101
6	0.07	4	0.93	0.1110...	1110

source

…and very relevant to current trends in AI….

Google DeepMind

Language Modeling Is Compression

Grégoire Delétang[*1], Anian Ruoss[*1], Paul-Ambroise Duquenne[2], Elliot Catt[1], Tim Genewein[1], Christopher Mattern[1], Jordi Grau-Moya[1], Li Kevin Wenliang[1], Matthew Aitchison[1], Laurent Orseau[1], Marcus Hutter[1] and Joel Veness[1]

[*]Equal contributions, [1]Google DeepMind, [2]Meta AI & Inria

source

Collibra

Burning Question | Similar vs Duplicate?

Directed Use-Case

This is not a record-to-record match, but rather a **column-to-column statistical comparison**

PROS

- Computational time
- Spatial requirements
- Does not store raw customer data at rest (lower risk)

CONS

- **Sample Accuracy** | Exact row comparisons are not done
- Column comparisons are statistical comparisons, not exact matching
- We see a maximum number of rows

Bottom Line | We believe that this method improves at volume, i.e., with more tables & columns comes more accuracy & utility

Constraints Drive Solutions

Directed Use-Case

Challenge | Our algorithm

- Must run on customer hardware
- Cannot phone home to an internal or 3rd-party API
- Must actually be able to determine similar/duplicate data

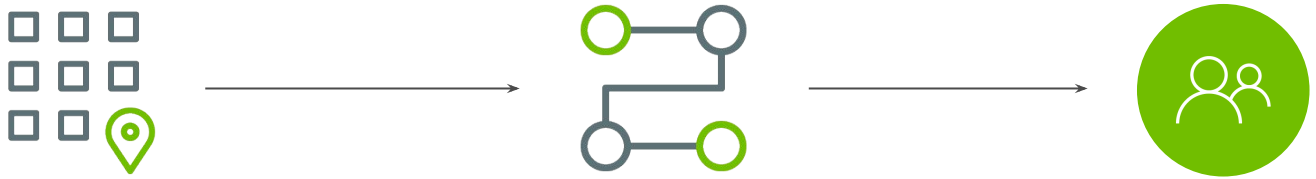
Solution | Let's go back to basics

- Decompose data into a series of statistical properties which minimize compression loss
- Build an algorithm which fits the needs of the end-user

source

Collibra

Navigating the Keys to Data Science



Data Awareness

Sampled, Compressed, & Noisy

Algorithm Selection

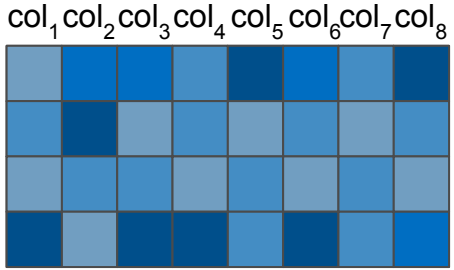
Pragmatic & Performant

Directed Use-Case

Volume-Driven

Prolonged Value | Technology Solution

Pipeline to acquire data



Compressed Data Representation

Context surrounding the data

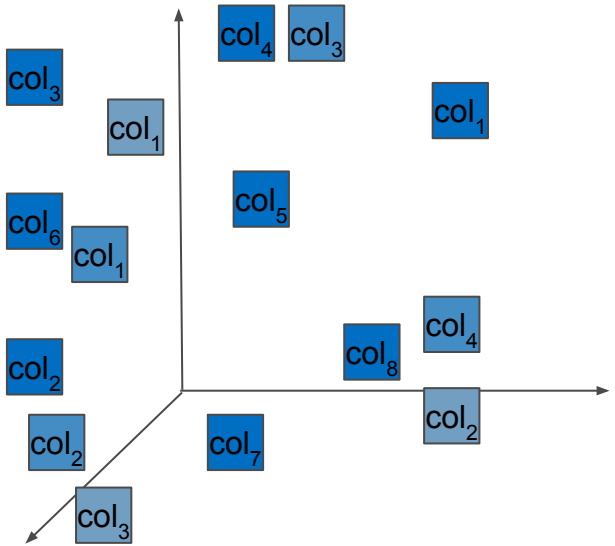
Collibra

Prolonged Value | Technology Solution

Pipeline to acquire data

Compressed Data Representation

Context surrounding the data

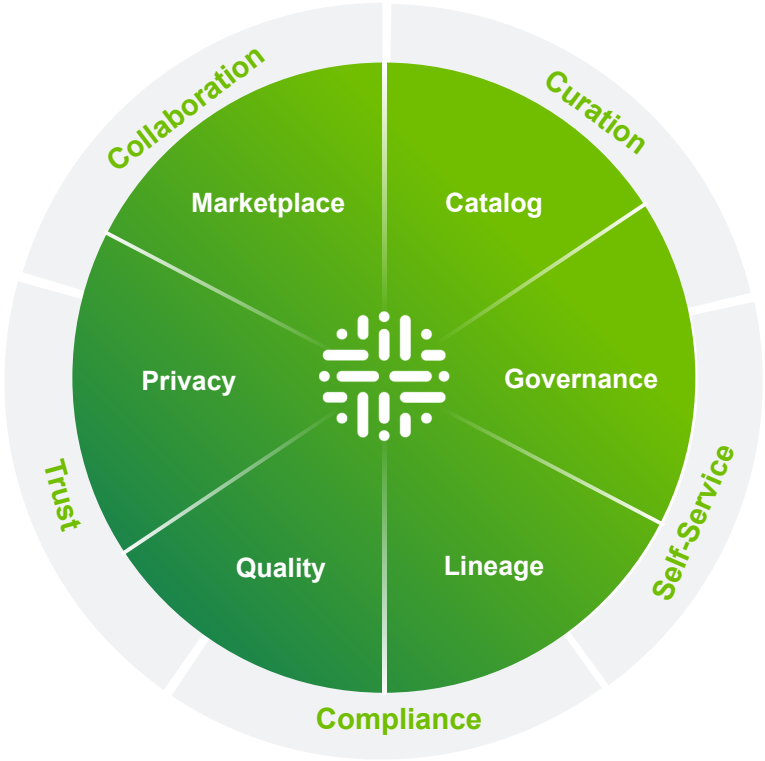


Prolonged Value | Technology Solution

Pipeline to acquire data

Compressed Data Representation

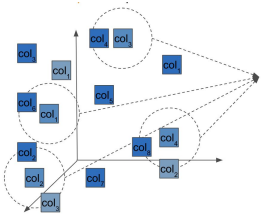
Context surrounding the data

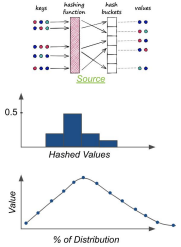


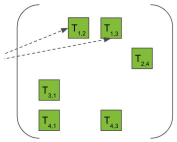
Collibra

PATENT PENDING

Expanding the Use-Case | What is RAG Anyways?







Data Layer

Compression Layer

Model Layer

Relational

Unstructured

Blob

Vector DB

Data Prep

Retrieval Layer

Third-Party LLMs

User Query

Model Response

Collibra

Concluding Remarks

- Sometimes, you need a sledgehammer. Other times, you need a scalpel

- Finding the right people to navigate the various avenues of…
  - Mathematics
  - Statistics
  - ML
  - Neural Networks
  - LLMs

  …will drive pragmatic, performant solutions

- At Collibra, the AI team navigated those various bounds to deliver a data similarity solution which walks the tightrope of performance, compute, and scalability

Imagine.art ‘s interpretation of ‘Data Council’

Thank you!