Watsonx.ai Proof of experience (PoX) education

Synthetic data

Felix Lee
felix@ca.ibm.com

Seller guidance and legal disclaimer

IBM and Business Partner
Internal Use Only

Slides in this presentation marked as **"IBM and Business Partner Internal Use Only"** are for IBM and Business Partner use and should not be shared with clients or anyone else outside of IBM or the Business Partners' company.

The information contained in this publication is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this publication, it is provided AS IS without warranty of any kind, express or implied. In addition, this information is based on IBM's current product plans and strategy, which are subject to change by IBM without notice. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this publication or any other materials. Nothing contained in this publication is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.

References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. Product release dates and/or capabilities referenced in this presentation may change at any time at IBM's sole discretion based on market opportunities or other factors and are not intended to be a commitment to future product or feature availability in any way. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth, or other results.

All client examples described are presented as illustrations of how those clients have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by client.

Content

Watsonx.ai synthetic data

- Need for synthetic data generation
- Synthetic data use cases
- Synthetic data generation
  - Manual schema specification
  - Mimic existing schema
- Synthetic data output options

Generative or Traditional AI – both need lots of data

- Model development
  - Tuning ML models
  - Prompt tuning generative AI models
- Testing models

The more realistic the data, the better the model will perform in a PoX as well as in actual production.

Data is now so essential to the modern economy that demand for real, high-quality data has grown exponentially. At the same time, stricter data privacy rules and ever-larger AI models have made gathering and labeling real data increasingly difficult or impractical.

Gartner predicts (link resides outside ibm.com) that by 2024, 60% of the data used in training AI models will be synthetically generated.

The problem

Real data = risk, costs, & delays

Costs[2]

59%

of AI budgets on average are spent on training data.

Heavy penalties[1]

$2.3b

in cumulative GDPR fines since 2017, with 50% attributable to non-compliance with general data processing principles.

Transformation delays

4-6weeks +

for teams to get access to production data, setting back project timelines & delaying progress. This can extend to months for certain industries and data types, like financial institutions and healthcare.

IBM watsonx.ai synthetic data

Issues, and risks of real data in a PoX

- Real client data can provide the best validation in a PoX...
- But using real client data in a PoX is often not possible/desirable
  - Risk and privacy/security issues
  - Too time-consuming to obtain (usually tightly controlled with strict approval processes in place) in a PoX

Synthetic data to the rescue

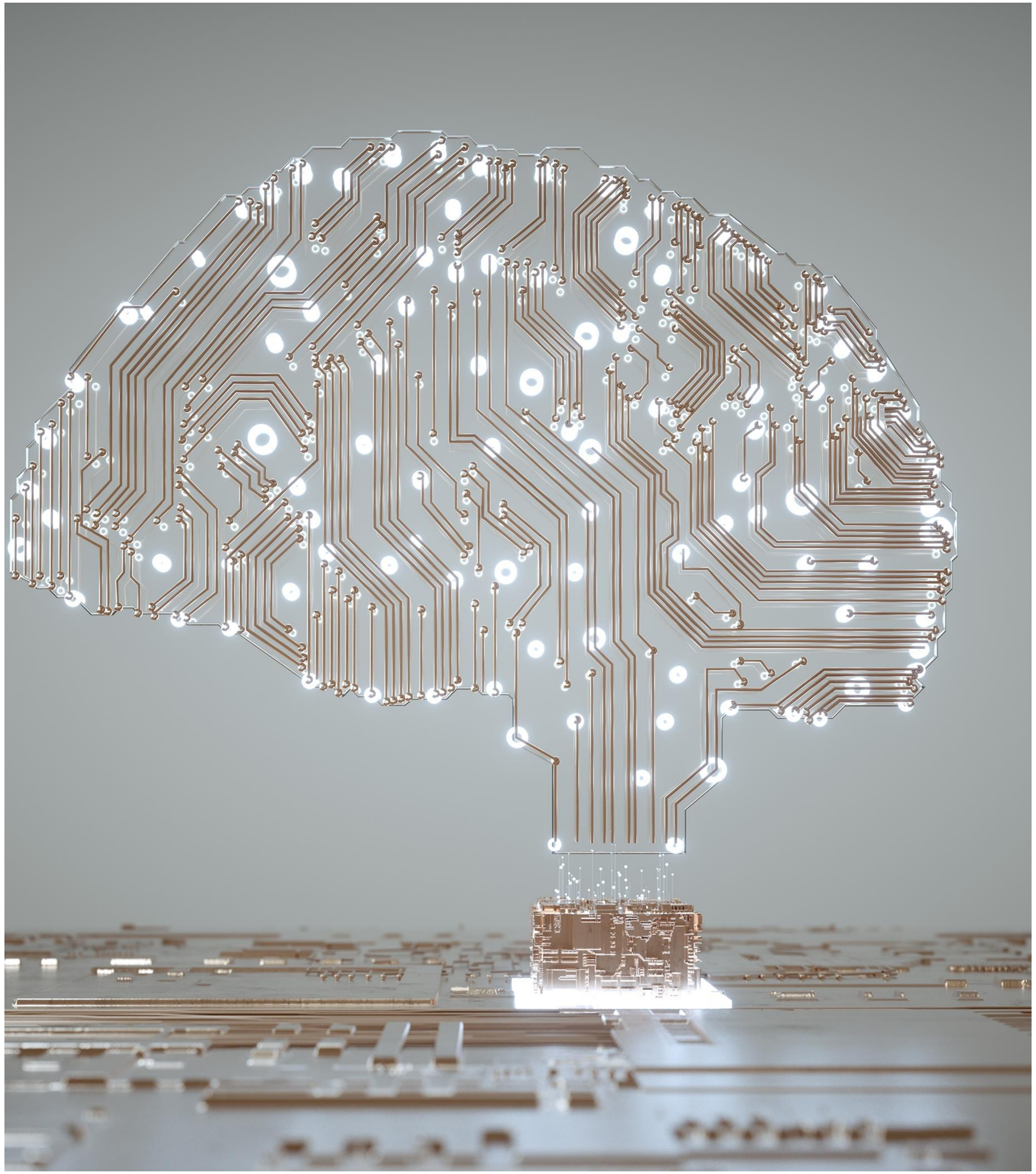
- Clients can use watsonx.ai to generate synthetic data that closely resembles real data
- Synthetic data can be generated in 2 ways:
  - By a manually specified schema and data distribution
  - By mimicking a small example of existing data – with the ability to mask/anonymize sensitive information

Synthetic adoption[1]

Forbes predicts that 60% of all data used for the development of AI and analytics projects will be synthetically generated, by 2024

5 Benefits of synthetic data

1. Innovation/GTM speed
2. Minimal risk
3. Reduced costs
4. Scale
5. Sharing & monetization



Common use cases for synthetic tabular data

Client demo data

Creating synthetic data to tailor demos for clients/industries before real client data becomes available

Employee training assets

Generate data needed to improve the realism of internal training programs

AI model training

Generating more data or edge case data to combine with real data to improve predictive accuracy of AI models

Monetize/share externally

Generating more data and sharing it externally with business partners, or monetizing with little privacy concerns

Extract insights

Create 1-for-1 synthetic copy of sensitive data, to share internally for insight extraction and strategic analysis

Application test data

High-fidelity, synthetic test data to expedite test cases and validation of software functionality, performance, and reliability

What-if assessments

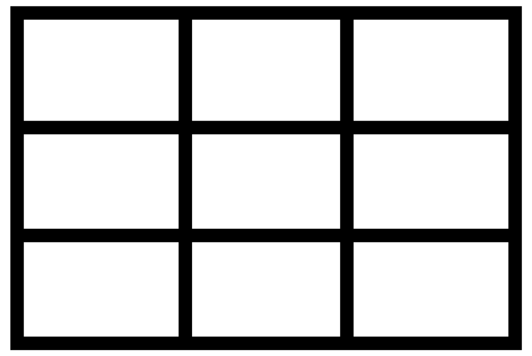
Simulate how synthetic agents' individual decisions impact macro-level metrics, like fraud, sales, or patient diagnoses

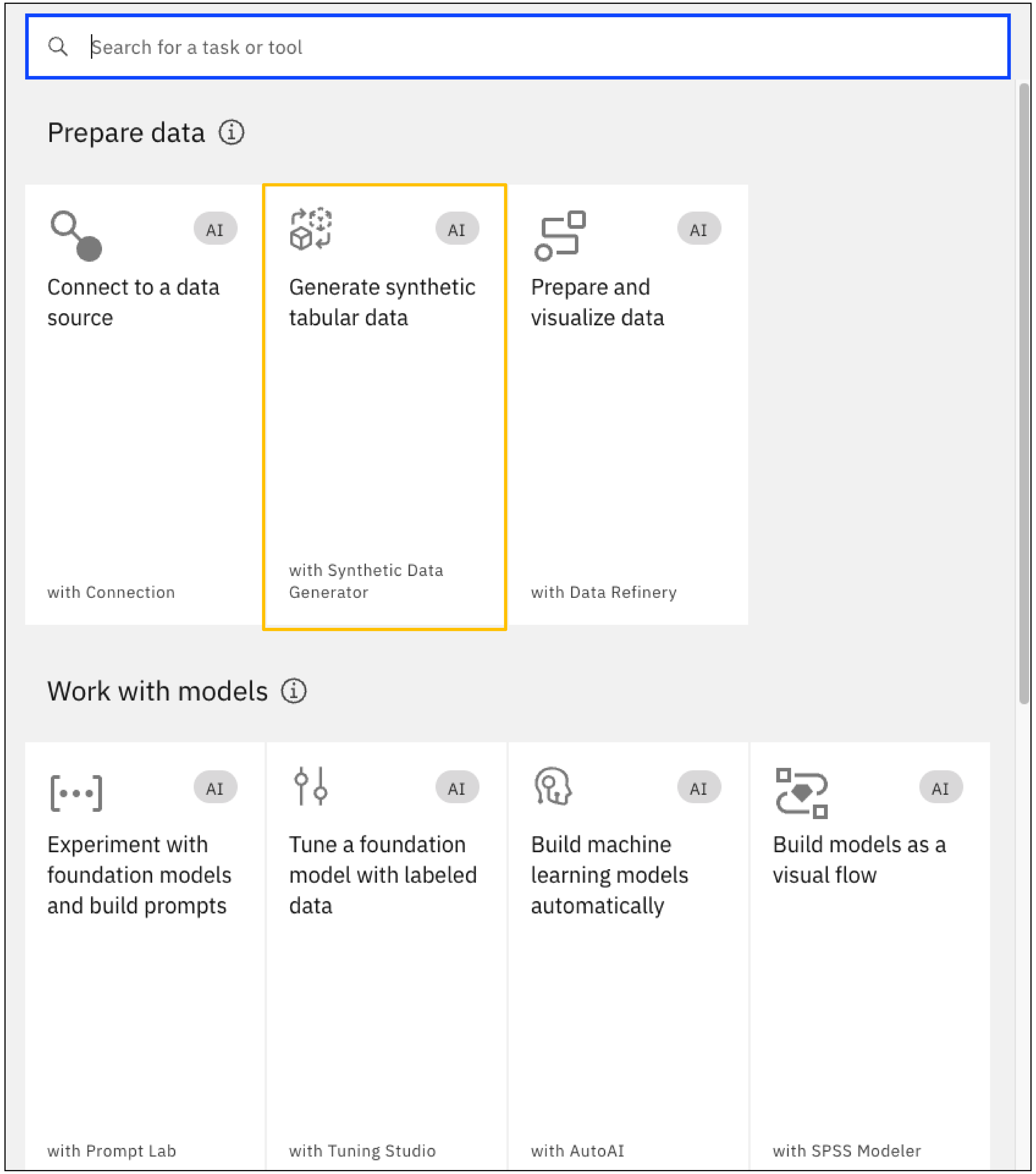
Synthetic data in watsonx.ai

Structured synthetic data

Data that has a standardized format, typically tabular with rows and columns that clearly define data attributes

Classic examples are CSV files or tables in relational databases





Two ways to generate synthetic data with watsonx.ai

Define custom schema and distribution

Clients can define:

- Table column data type and characteristics
  - String, Integer, Real, Time, Date, and Timestamp
  - Maximum, minimum, mean, spread
- Data distribution
  - Normal, uniform, and others
  - Discrete distribution such as Bernoulli, Category (more on this later)
- Correlation (if any) among columns
- Number of output rows

Mimic existing data schema and distribution

Clients can provide a set of sample data

- Watsonx.ai generates synthetic data by mimicking the schema and distribution of the input data

Users can define:

- Various columns to be anonymized
- Number of output rows
- How the output data can be best fitted to the input data
- Applying differential privacy for more protection

Custom schema versus mimic

Custom schema

- Much easier to acquire and manage than the client’s data, especially in a PoX
- No privacy concerns
- Clients have complete control over how the generated data would look
- Care must be taken so that generated data has desired characteristics, and not just a statistical behavior

Mimic

- Can be difficult and time-consuming to acquire a sample data set
- Can still have privacy concerns
- Can use a small set of data to generate a huge set of data that closely resembles reality
- No need to manually determine what distribution to use to generate data with particular characteristics
- Care must be taken to ensure the set of sample data is the right set for the use case

Model data distribution – in custom synthetic data generation

Determines how the output data will be distributed

Generated data has use case requirements

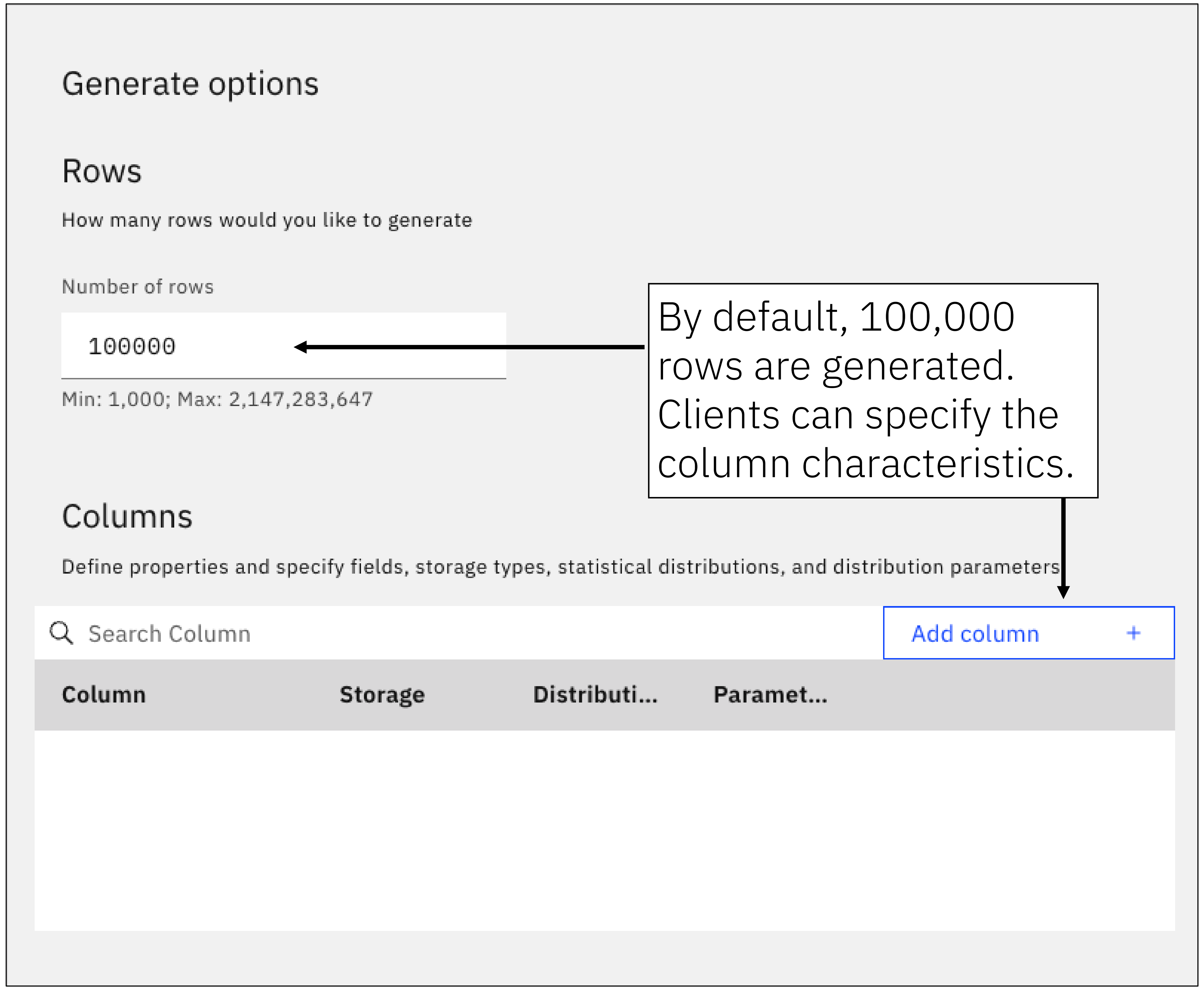
- Human characteristics (age, height, etc.) are often well-represented in a Normal Distribution
- Some data (such as the ratio of male to female in an employee table) may need to adhere to local demographic distribution
- Others (such as departments) have a finite set of discrete values
- Business logics may dictate correlations
- Data scientists are best able to determine how a generated data set in a use case should be distributed

Watsonx.ai synthetic data generation controls

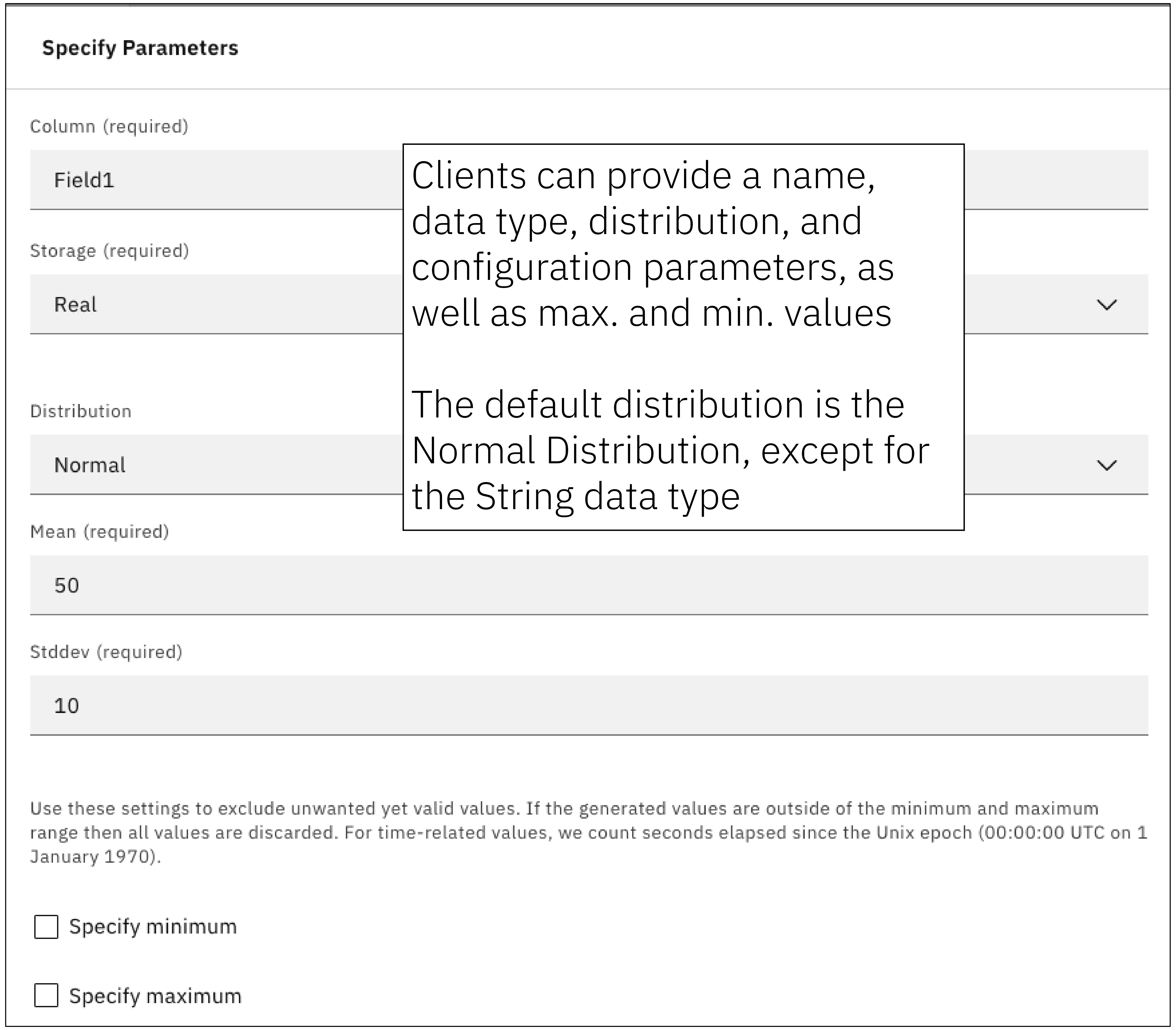
- Watsonx.ai offers many distributions to simulate the desired output
- Clients can specify values for distribution configuration parameters such as:
  - Mean, Stddev, Max and Min values for Normal Distribution
  - Beginning/End and Probability values for ranges in a Range Distribution
  - Explicit probability for each discrete category of a Category Distribution (or other discrete distributions)

Custom schema configuration

Fitting options



Column options



Model data distribution – in a mimic use case

Output data distribution should match input data

Generated data needs to mimic input data

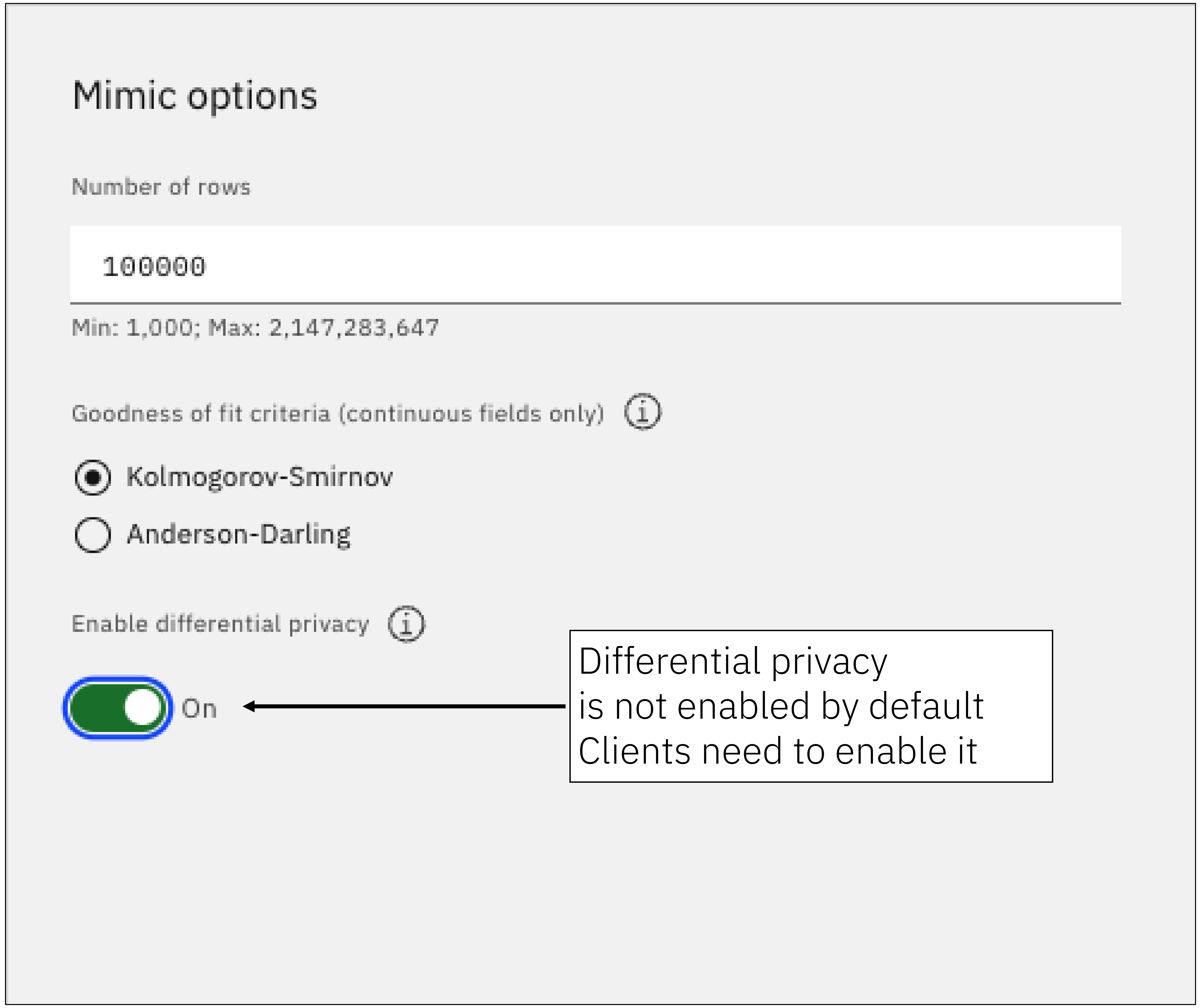
- Real-life data rarely (if ever) matches up closely against a statistical distribution
- IBM watsonx.ai determines how to adjust the parameters of distributions to find the closest fit via goodness-of-fit tests
  - Kolmogorov-Smirnov (default)
  - Anderson-Darling
- Any input columns can be anonymized

Applying differential privacy

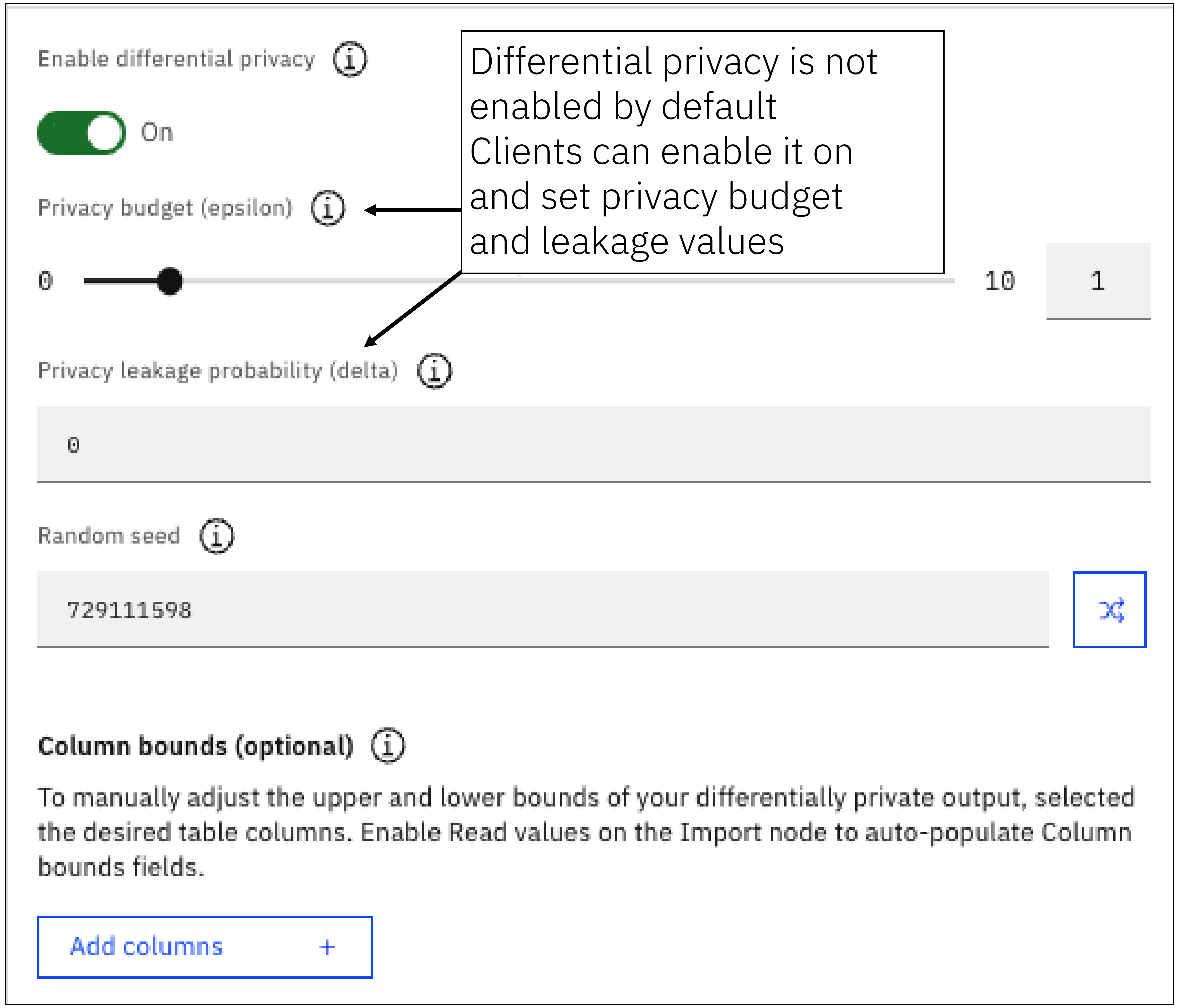
- Clients can enable differential privacy
  - Apply to input data before generating synthetic data
  - Ensure that no user-sensitive data is exposed in the generated data output
- A trade-off between privacy and accuracy
  - Higher the privacy, lower the accuracy (looks less like the input data)
- Clients can control the level of privacy

Mimic configuration

Fitting options – enabling differential privacy



Differential privacy options



Output format

Output file location and format

- Clients can export output to:
  - A watsonx.ai project
  - A connection (to Db2, AWS RDS, Azure SQL database, and many more data sources)
- Clients can select from many popular formats:
  - CSV or another delimited format
  - Excel
  - JSON, Parquet, SAV, or XML
- Clients can select a compression codec to use

Output file data format

- Clients can specify:
  - Decimal format
  - Date format
  - Time format
  - Timestamp format
  - NULL value

IBM