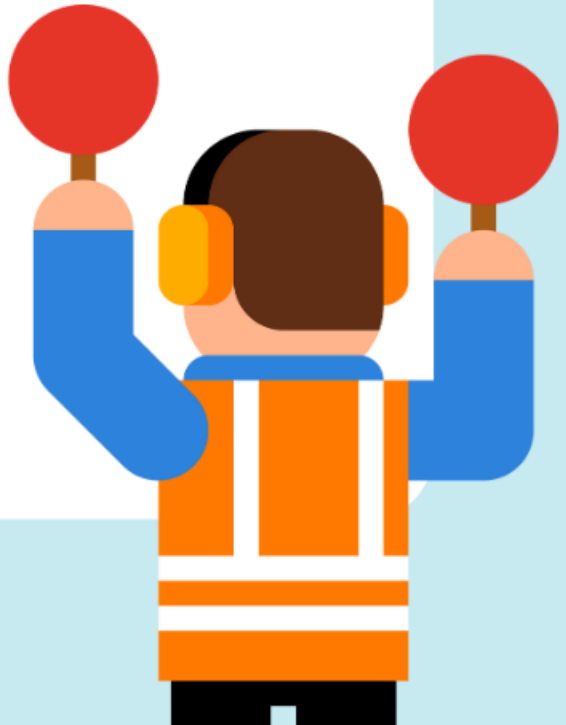
Crafting Flight Simulations:

The Art of Synthetic Data Generation

Dave Ruijter

Session feedback



sqlbits ***2024***

Dave Ruijter

Solution Architect & Engineer
Blue Rocket IT

dave@blue-rocket.it
@DaveRuijter
linkedin.com/in/DaveRuijter
ModernData.ai

Dave Ruijter

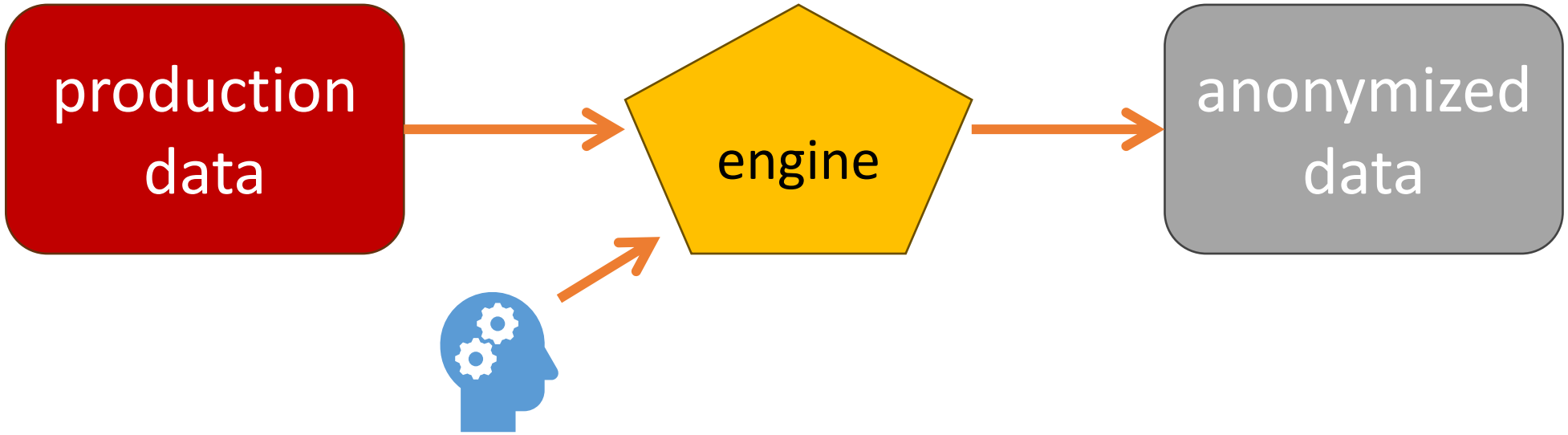
Solution Architect & Engineer
Blue Rocket IT

dave@blue-rocket.it
@DaveRuijter
linkedin.com/in/DaveRuijter
ModernData.ai

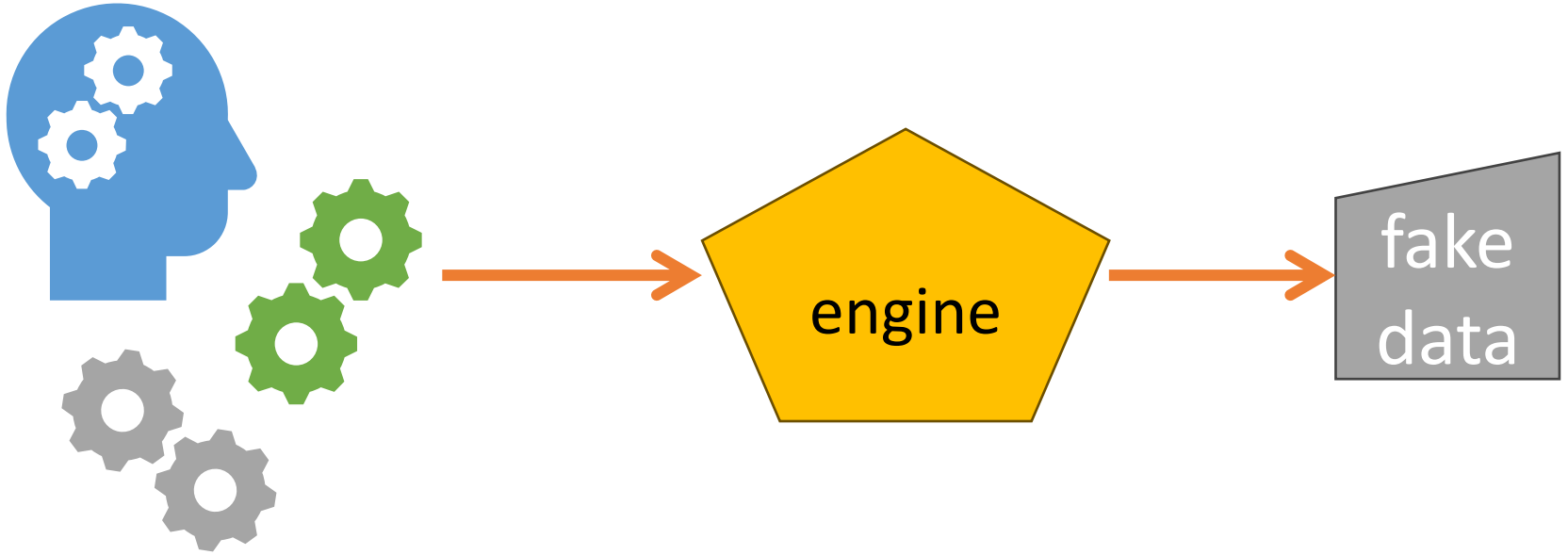
sqlbits
2024

sqlbits
2024

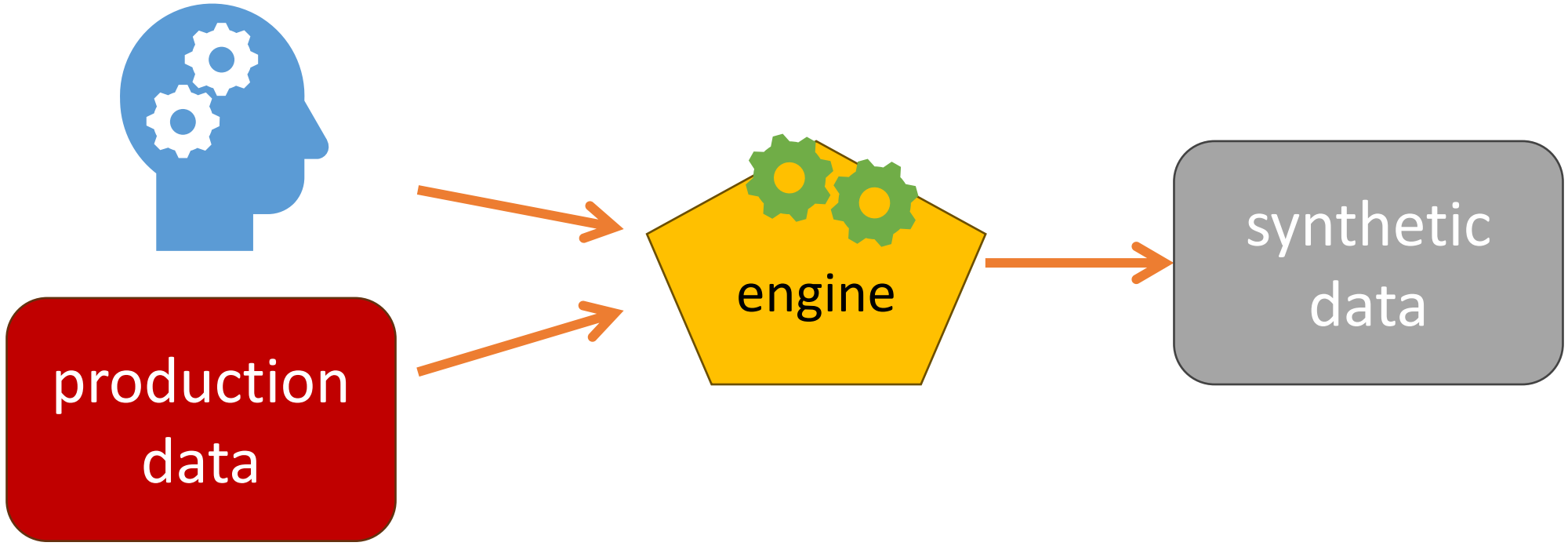
Anonymized real data



Generated fake data



Synthetic data generation



Anonymized vs Fake vs Synthetic

Definition	**real data** processed to remove or alter identifying information	deliberately **fabricated** and does not correspond to real individuals or events	algorithmically **generated to mimic** the statistical properties of real data but does not directly correspond to real data
Purpose	protect individual privacy while still allowing data to be **used for analysis and decision-making**	often limited to **testing and development**	overcome limitations of fake and anonymized data, a **balance** between privacy protection and the utility of real data
Characteristics	very useful, but with **risks**. Anonymization can be reversed, leading to potential privacy breaches	does not replicate the **statistical** properties of real data closely	designed to be statistically **similar to the original**, allowing for accurate analysis and model training, yet privacy safe

The Need for Synthetic Data

- ✓ Privacy and security
- ✓ Utility of real data
- ✓ Bridge data gaps in underrepresented areas

- ✓ ***Machine learning model development***

The Need for Synthetic Data

- ✓ Privacy and security
- ✓ Utility of real data
- ✓ Bridge data gaps in underrepresented areas

- ✓ ***Machine learning model development***

Applications & Benefits

- Enhance software testing and quality assurance
- Accelerate Machine learning and AI development
- Support data privacy regulatory compliance
- Foster innovation in data-scarce fields
- Enable comprehensive scenario analysis and decision making

Algorithms and Techniques

1. Generative Adversarial Network (GAN)
2. Variational Autoencoder (VAE)
3. Decision Trees and Random Forests
4. Bayesian Networks
5. Synthetic Minority Over-sampling Technique (SMOTE)

Tools and platforms

- MOSTLY.AI
- Synthesized
- Hazy
- Sogeti
- Gretel
- Datomize
- CVEDIA
- Rendered.ai
- Oneview
- MDClone

DEMO

Using **MOSTLY.AI**

DEMO

Using **Synthetic Data Vault**

https://docs.sdv.dev/sdv

Challenges and limitations

- Reliability of the data
- Replicating outliers
- Requires expertise, time, and effort
- User acceptance
- Quality check and output control

Q&A

Session feedback