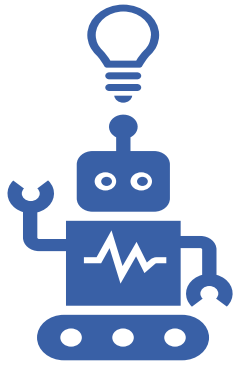
DevOps for Data Science/DataOps

Bianca Furtuna, @Fur_Bi

Applied Machine Learning Scientist

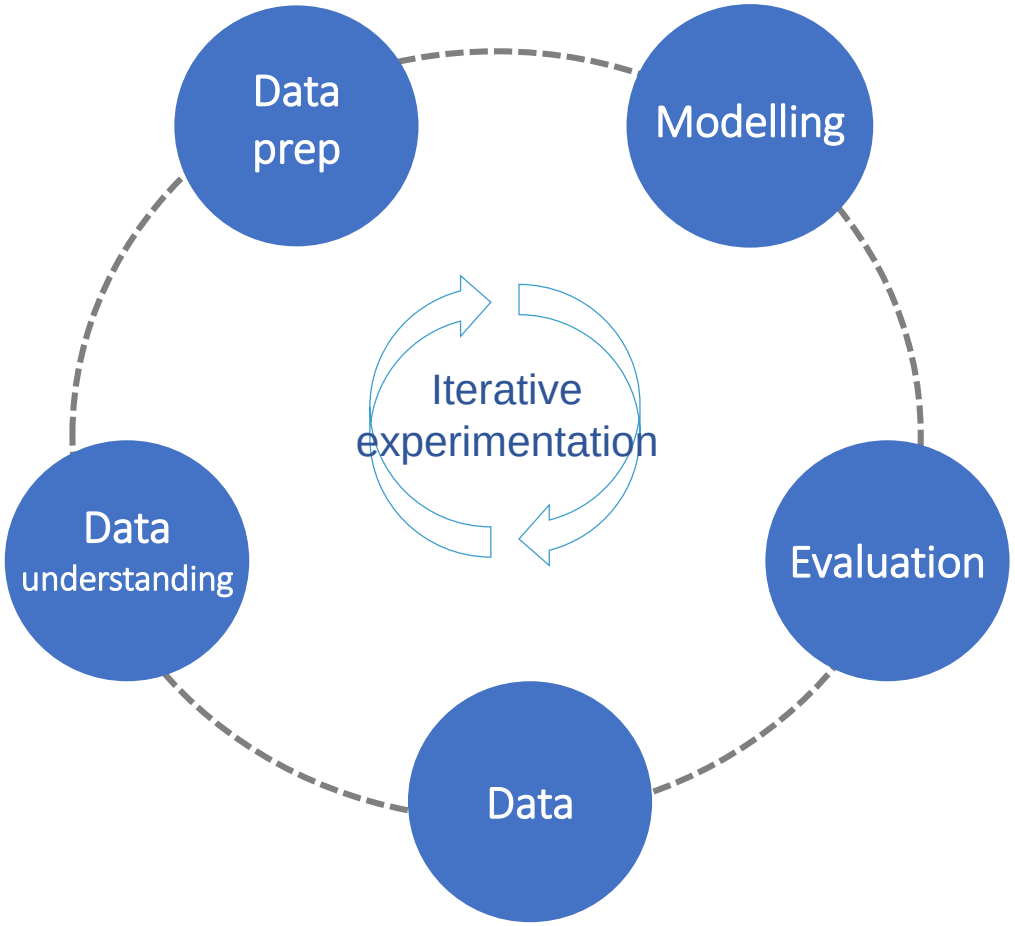
Machine Learning Revolution

- Data-focused companies
- Data-driven decision and automation
- New type of software components that need to be managed



Teams and new roles

Typical Machine Learning Process

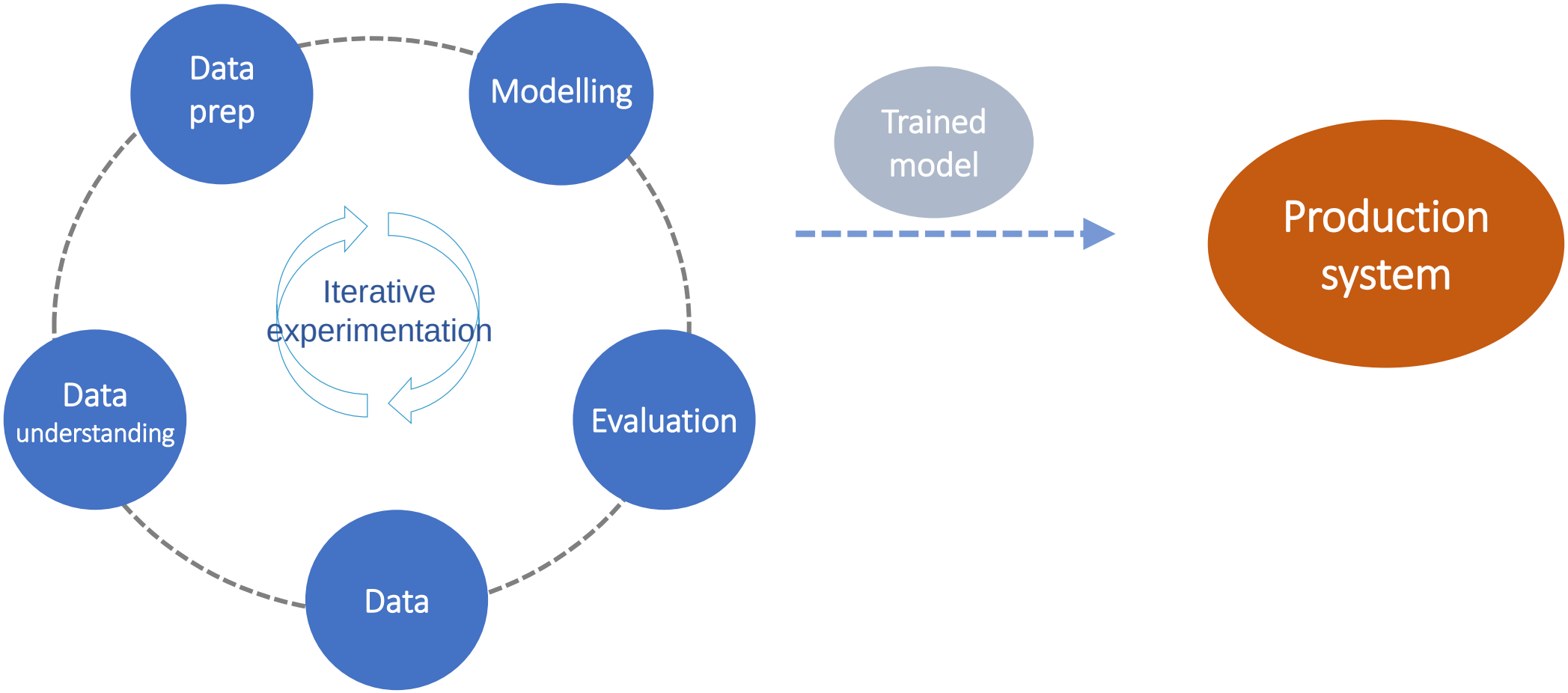


Modelling + Evaluation

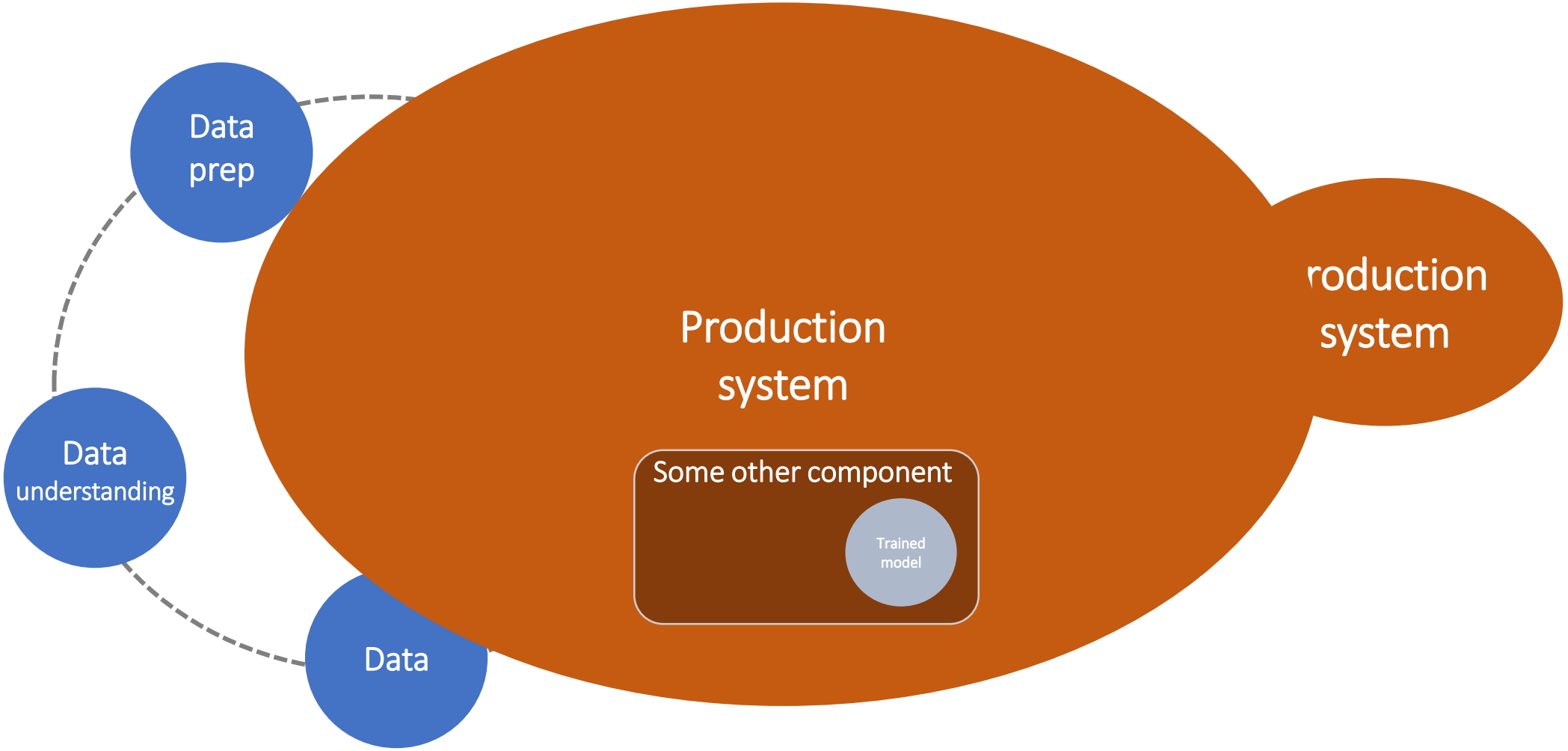
- Most talked about – most exciting
- Experimentation and research happens here
- Some data scientist develop this completely in isolation

- Model versioning and source control
- Need to understand the context of how it is going to be used in order to have the right input/output
- Where does it fit?

Typical Machine Learning Process



Typical Machine Learning Process



Production steps for ML models

- Define model interface (input/output)
- Store the model and the environment requirements (versioning)
- Wrap the model up as a consumable service (e.g. API – Flask app)
- Deploy this service as part of your system
- Run it along side previous solutions to compare performance

Where does DevOps fit?

- CI/CD
- Build data pipelines
- Build and test (API level)
- Monitor and learn
- Automate retraining
- Source control for code, notebooks, model

Once in production …

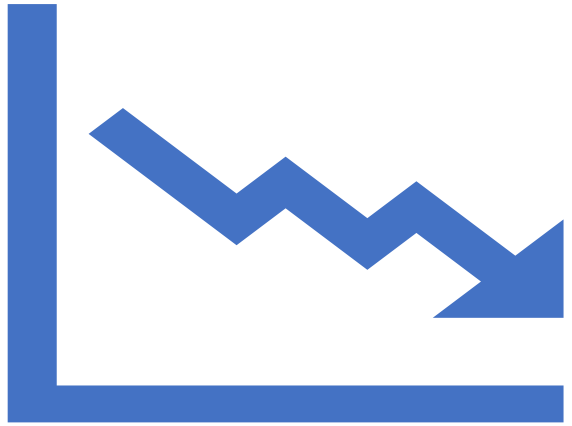
Monitor model performance

Validate live results

Automatic retraining

Data quality monitoring

- ☑ Ensure the data types are consistent
- ☑ Ensure the data is within expected values
- ☑ Ensure no NAs are flowing through the system



Data Validation at each stage of your data pipeline

Conclusions

- Culture change
- Where does it fit?
- Data validation is important
- Follow DevOps principles for ML artifacts
- New teams and new roles -> new processes and tools

Thank you!