inovex

ML Deployment

Vom Prototyp zur Produktion

Marcel Spitzer

Karlsruhe, 15.3.2018

Marcel Spitzer
Big Data Scientist @ inovex

Wi.-Mathematik (B.Sc.), Wi.-Informatik (M.Sc.)

- Data Science mit Python und R
- Deployment von ML Anwendungen
- Big Data, Hadoop, Spark

Beispiel 1: Batch Szenario
Demand Prediction

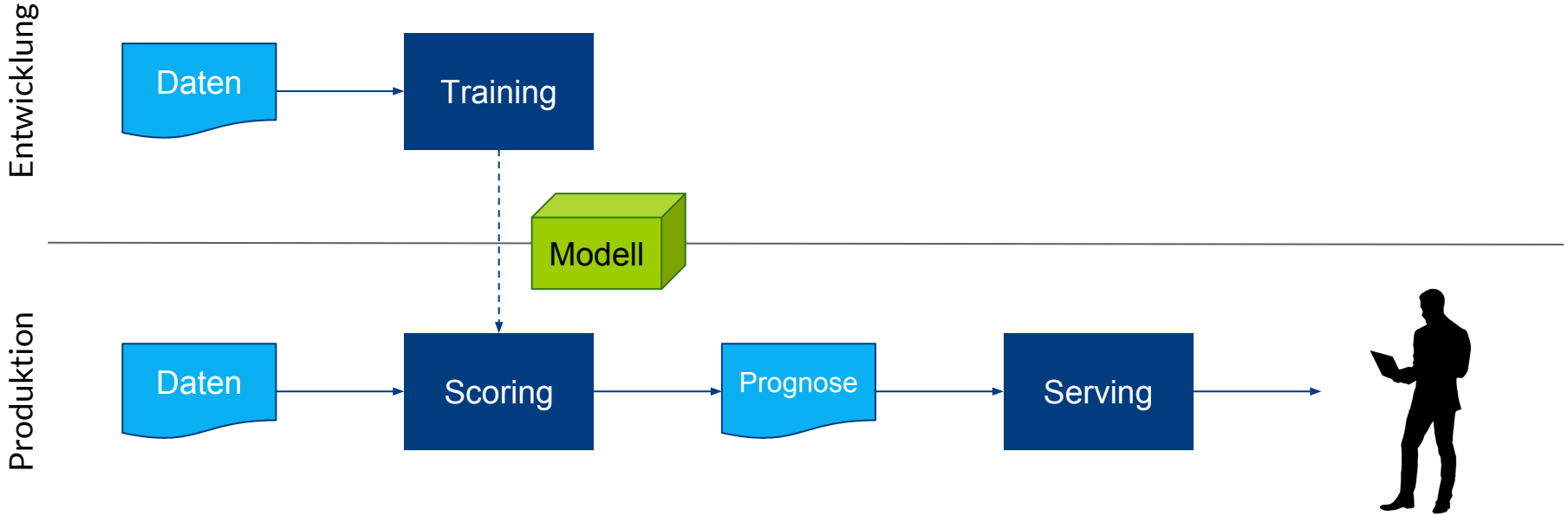
- Vorhersage von Produkt-nachfrage auf Tagesbasis
- Bereitstellung in Datenbank
- Zugriff über Analysetools und Dashboards

https://pixabay.com/photo-2860753/

Beispiel 2: (Near-)Realtime Szenario
Recommender System

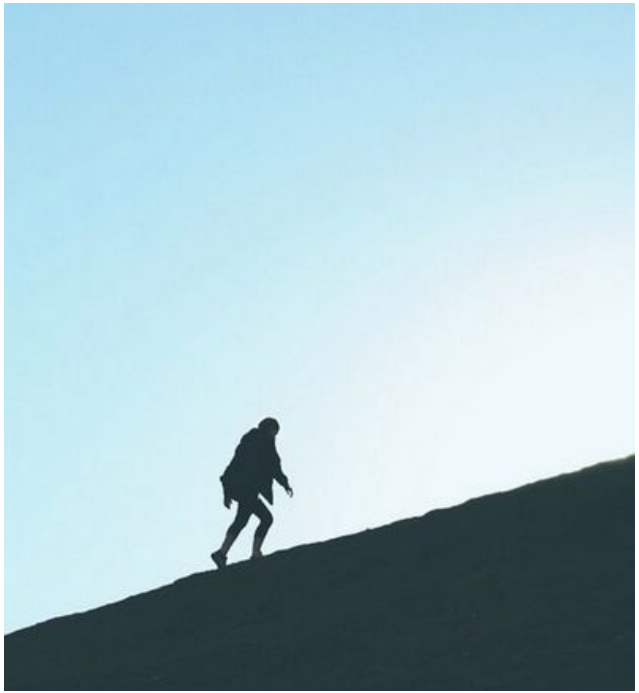
- Produktempfehlungen auf Basis von Nutzerverhalten
- Bereitstellung als Webservice
- Zugriff über REST API

ML Deployment Prozess



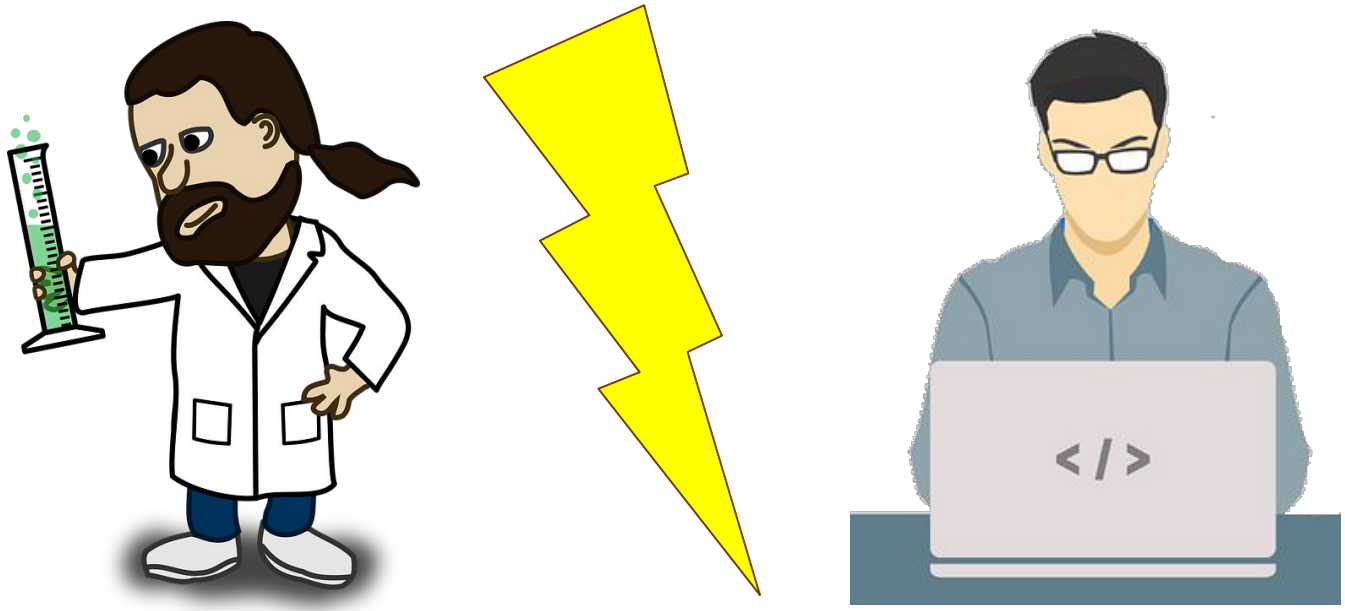
Herausforderungen

- Hand-Over in Produktion
- Skalierbarkeit vs. Komplexität
- Monitoring und Modellmanagement
- Inkrementelle Verbesserung
- Hohe Verfügbarkeit



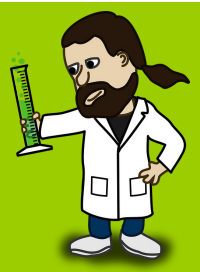
https://unsplash.com/photos/wrfj-SRaB1Q

Data Science und SW Engineering



Data Science und SW Engineering

Data Science



- Experimentell
- Rapid Prototyping, Ad-Hoc Analysen
- Reports, Dashboards

SW Engineering

- Systematisch
- Clean Code, Test-Driven Development, CI/CD
- Wartbare und funktionale Software

Data Science und SW Engineering

- Data Science braucht Iteration
- Prototyping oft in einer Sandbox-Umgebung
- In Produktion müssen ML Anwendungen realen Bedingungen standhalten

Data Science und SW Engineering

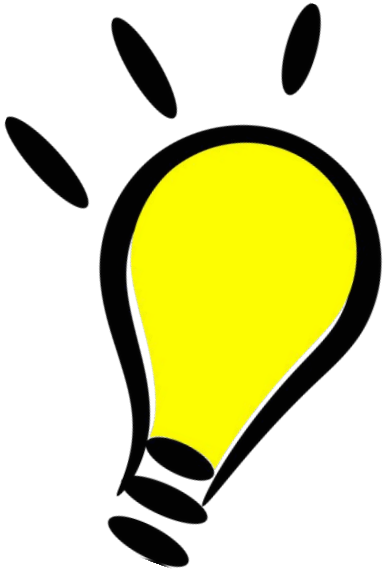
DS kann *chaotisch* sein, deshalb

- Versionskontrolle
- Dependency-Management
- Automatisierter CI-Prozess

https://xkcd.com/1838/

Data Science und SW Engineering

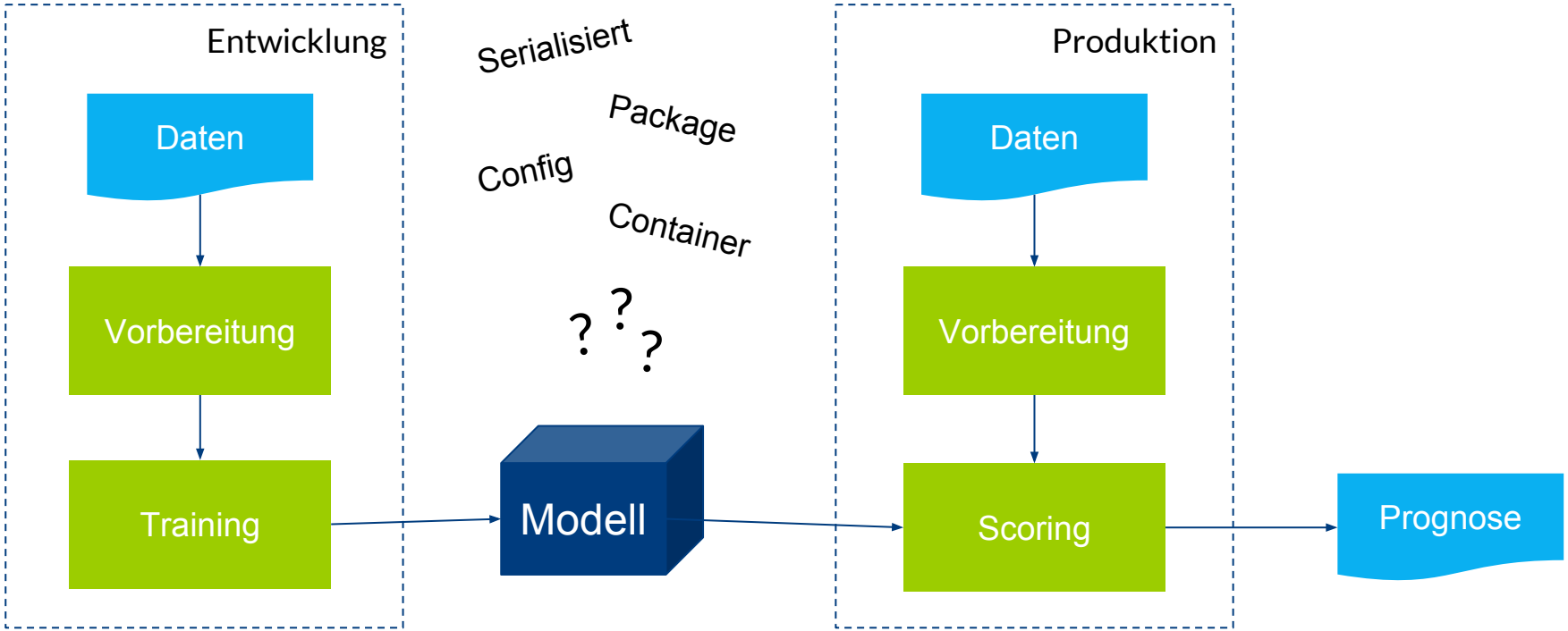
- ML Deployment erfordert Engineering Skills
- Frühzeitige Einbindung von SW Engineers
- Praktikable Lösungen und Etablierung von Standards



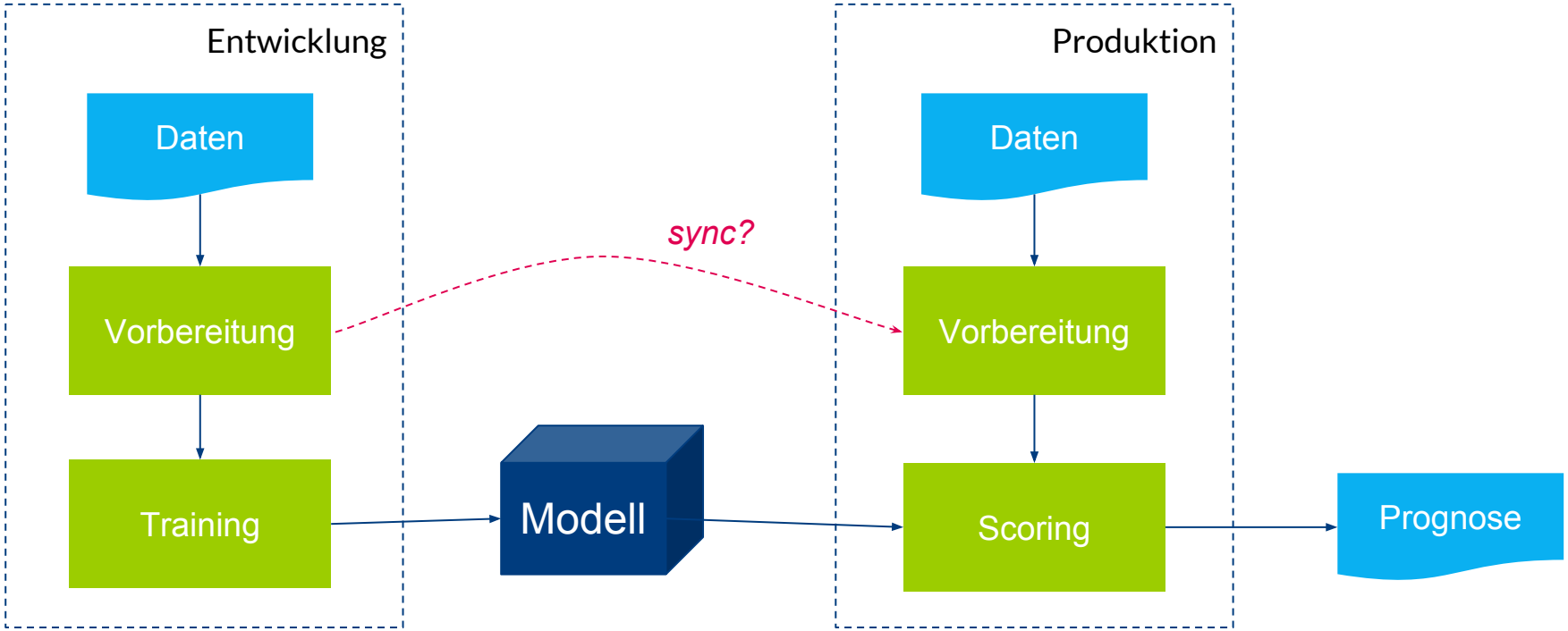
https://pixabay.com/photo-1294877/

Hand-Over von ML Modellen

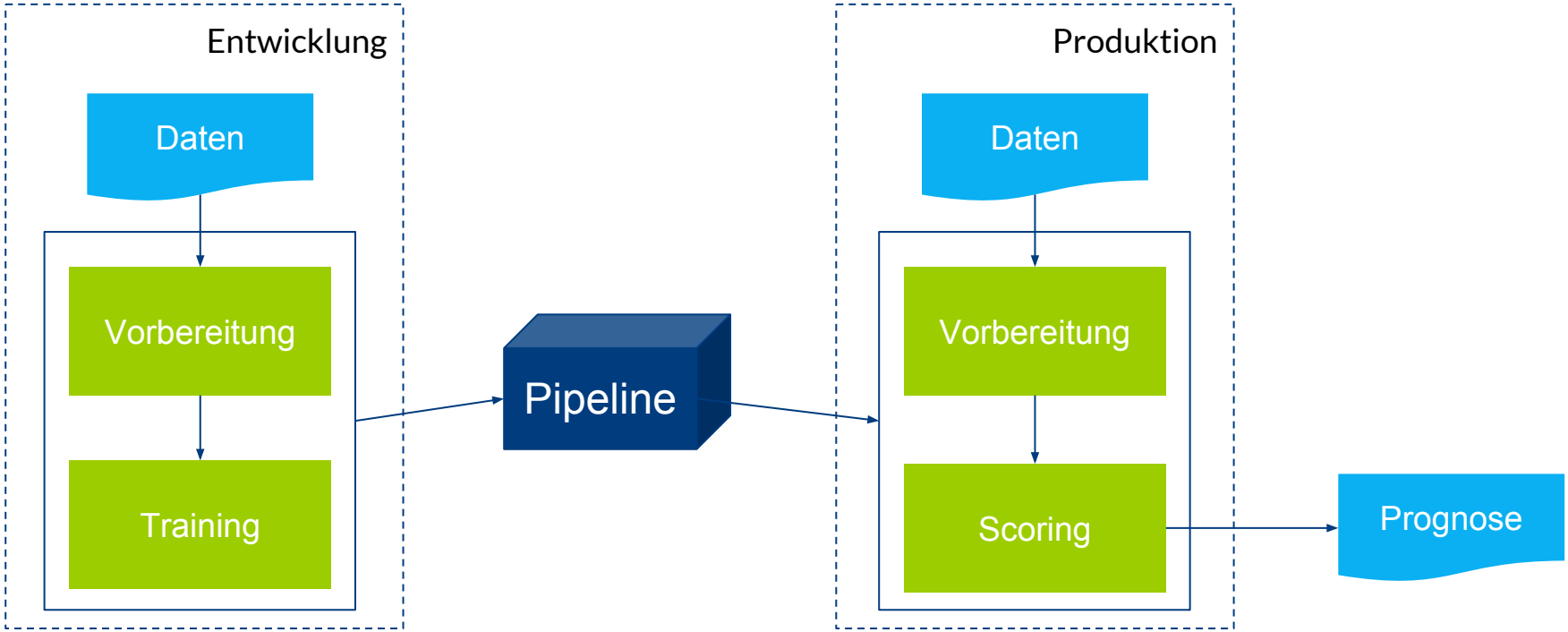
Hand-Over von ML Modellen



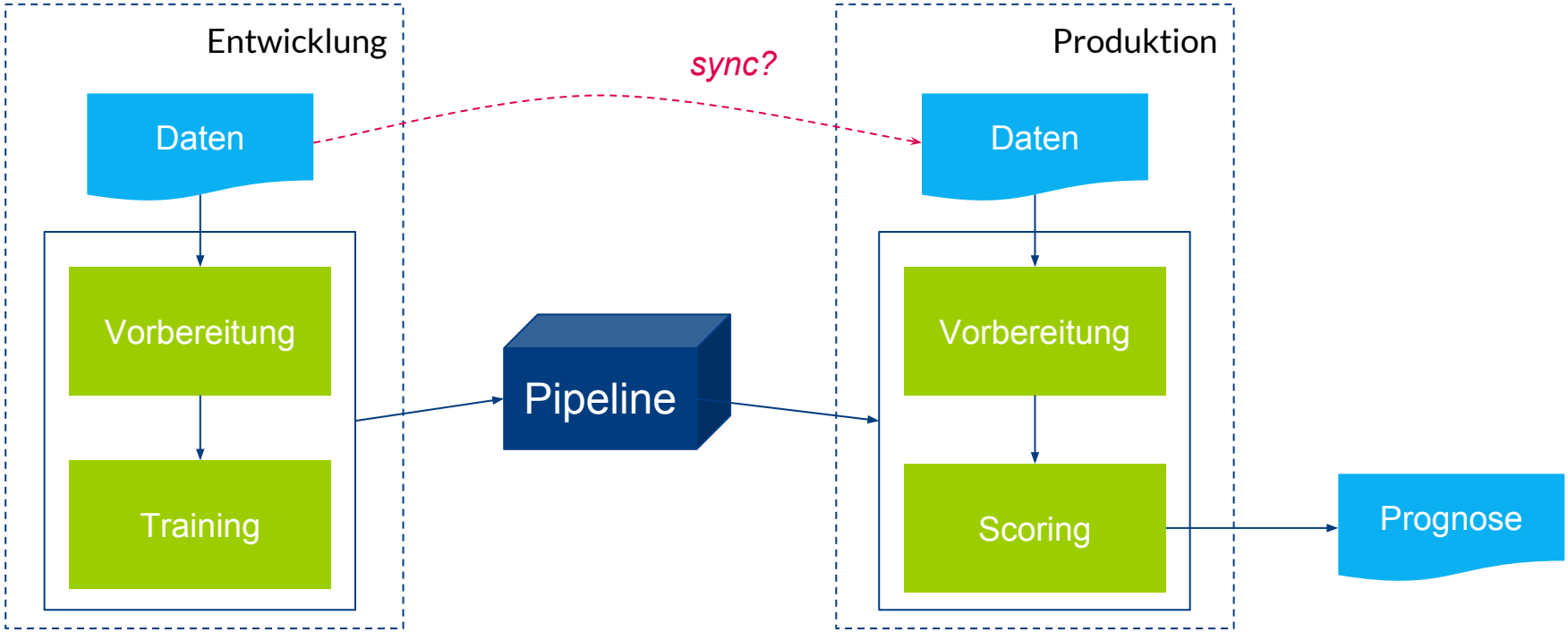
Hand-Over von ML Modellen



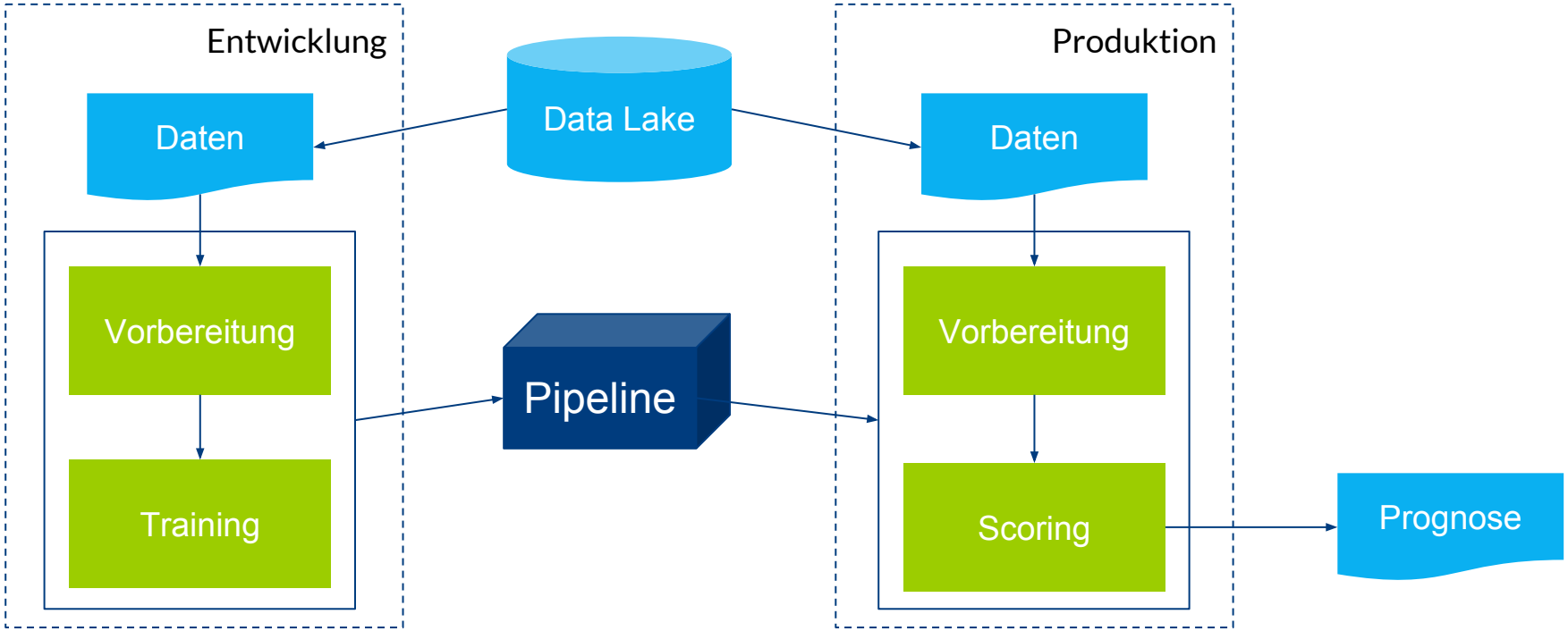
Hand-Over von ML Modellen



Hand-Over von ML Modellen

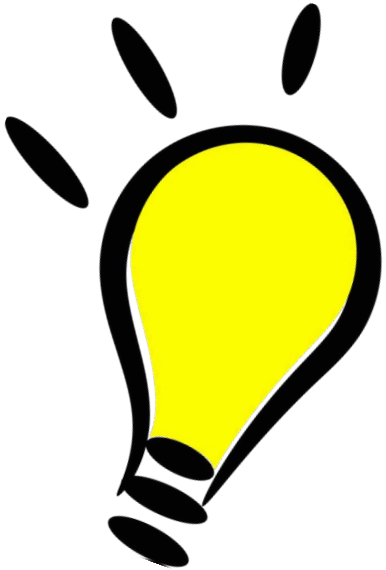


Hand-Over von ML Modellen



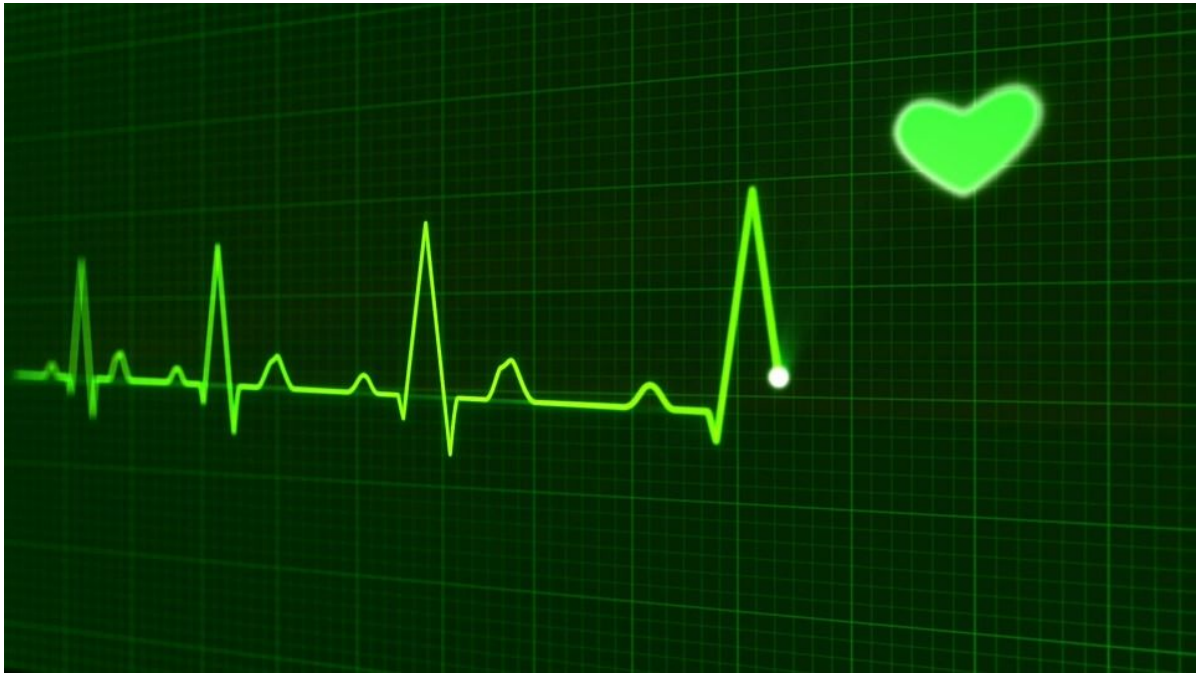
Hand-Over von ML Modellen

- Serialisierung erfordert konsistenten SW Stack
- Container bieten größtmögliche Flexibilität
- Pipeline-Konzept erleichtert Reproduzierbarkeit
- Data Lake für konsistente und aktuelle Daten



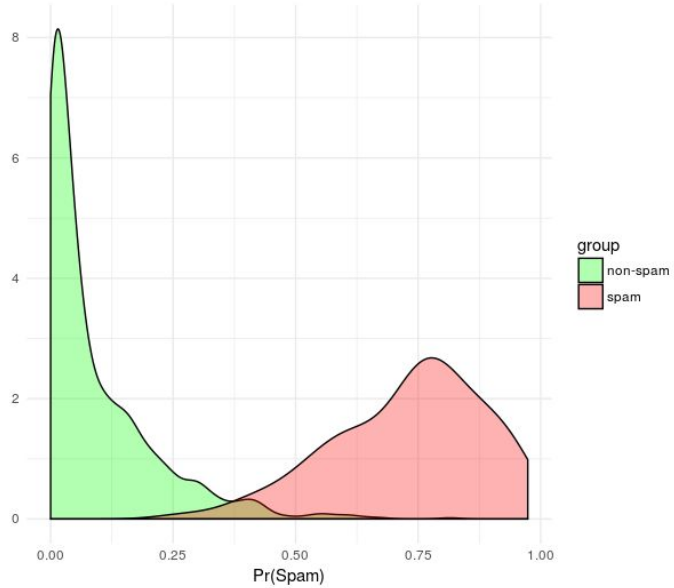
https://pixabay.com/photo-1294877/

Monitoring und Modellmanagement



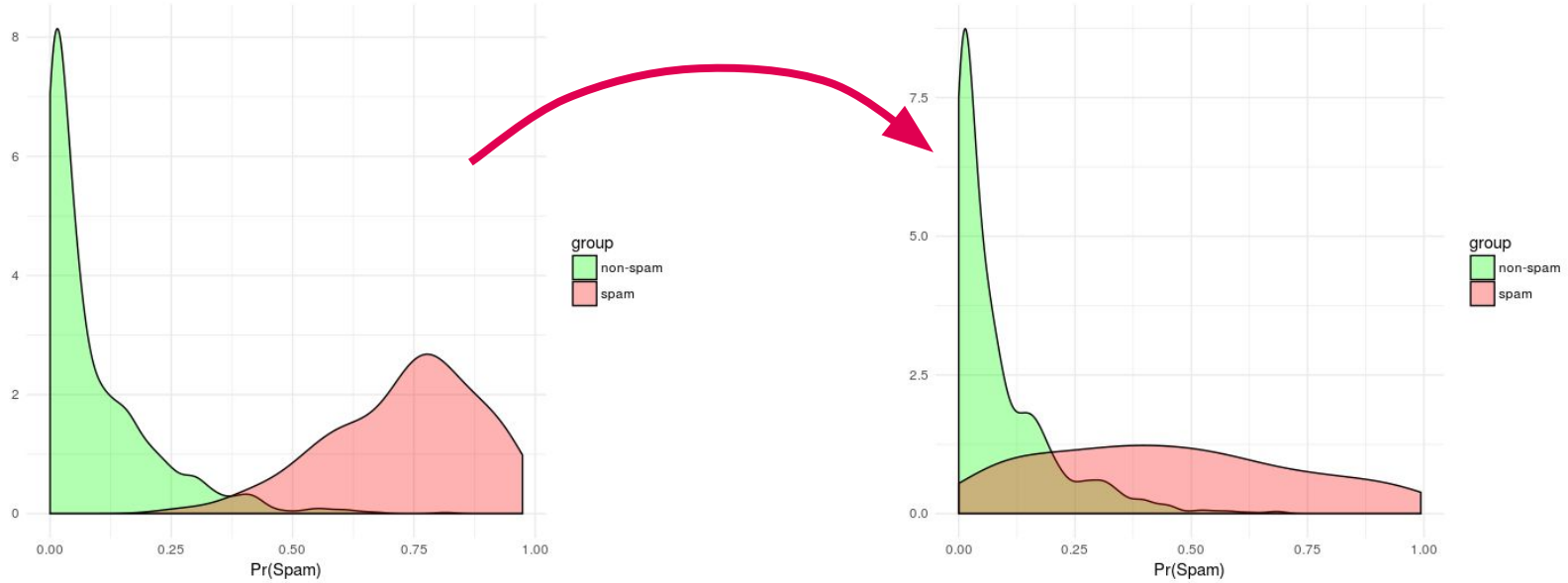
Monitoring und Modellmanagement

Beispiel: Spam-Klassifikation



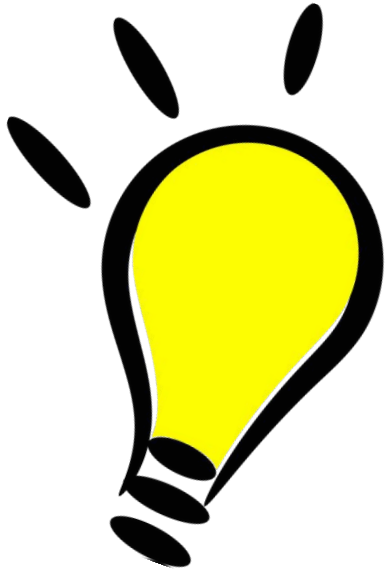
Monitoring und Modellmanagement

Beispiel: Spam-Klassifikation

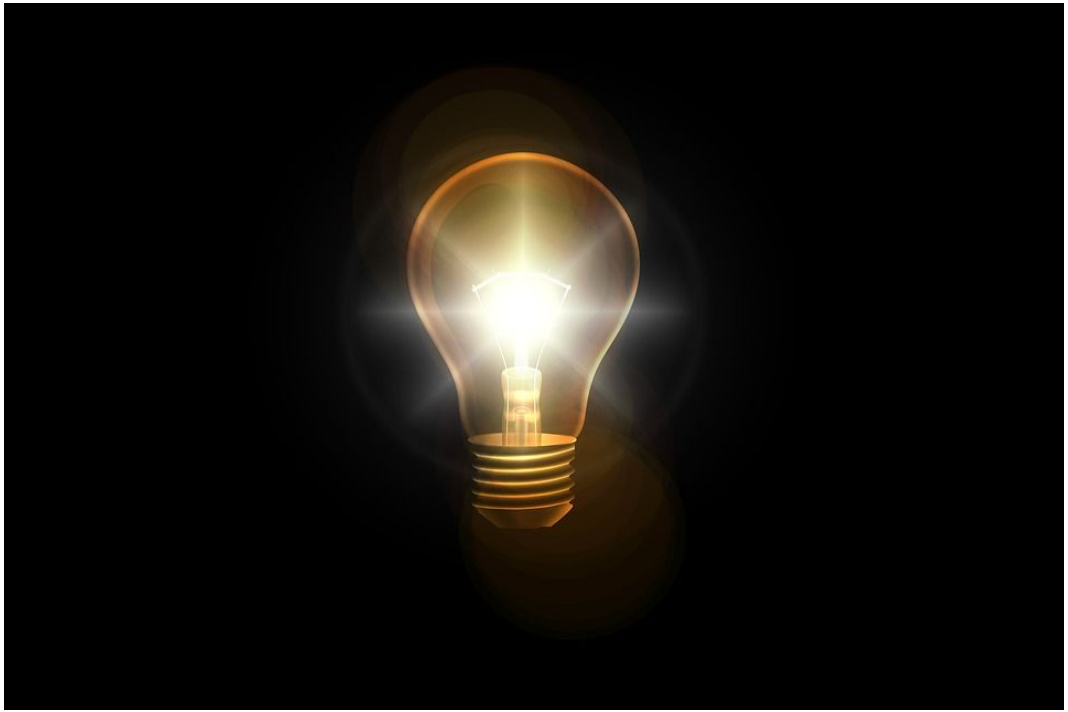


Monitoring und Modellmanagement

- Verhaltensmuster ändern sich mit der Zeit
- ML Anwendungen müssen adaptierbar sein
- Erfordert Re-Training auf aktuellen Daten
- Automatisiertes Deployment für schnelle Reaktionsfähigkeit



Deployment-Strategien



Strategie 1: Re-Implementierung

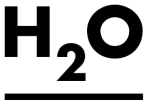
Modell liegt prototypisch in Python oder R vor und wird auf Stack des SW Engineers reimplementiert

- Ineffizient und fehleranfällig
- Korrekte Replikation der Modelle schwierig
- Nur für einfache Modelle realisierbar

Strategie 2: ML Frameworks

Verwendung dedizierter ML Frameworks in Entwicklung und Produktion

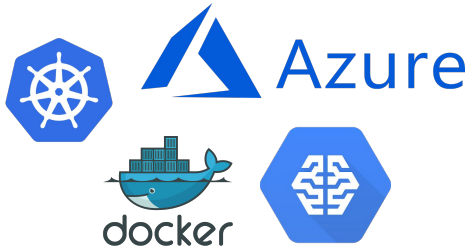
- Ende-zu-Ende Abdeckung des ML Workflows
- Lokale und Remote Entwicklung
- Multi-Language Support
- Horizontale Skalierbarkeit
- Eingeschränkter Funktionsumfang



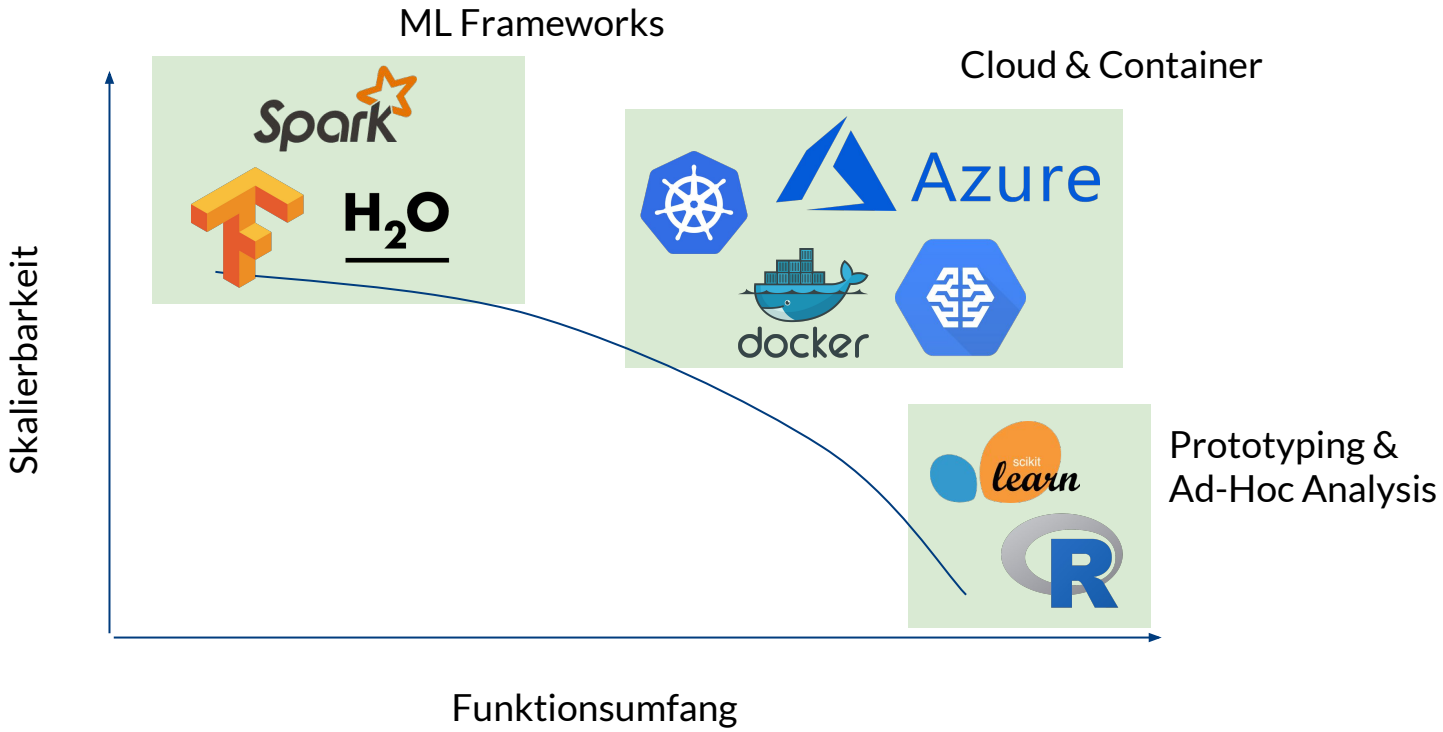
Strategie 3: Microservices

Modell wird in Python oder R trainiert und als Webservice in einem Container deployt

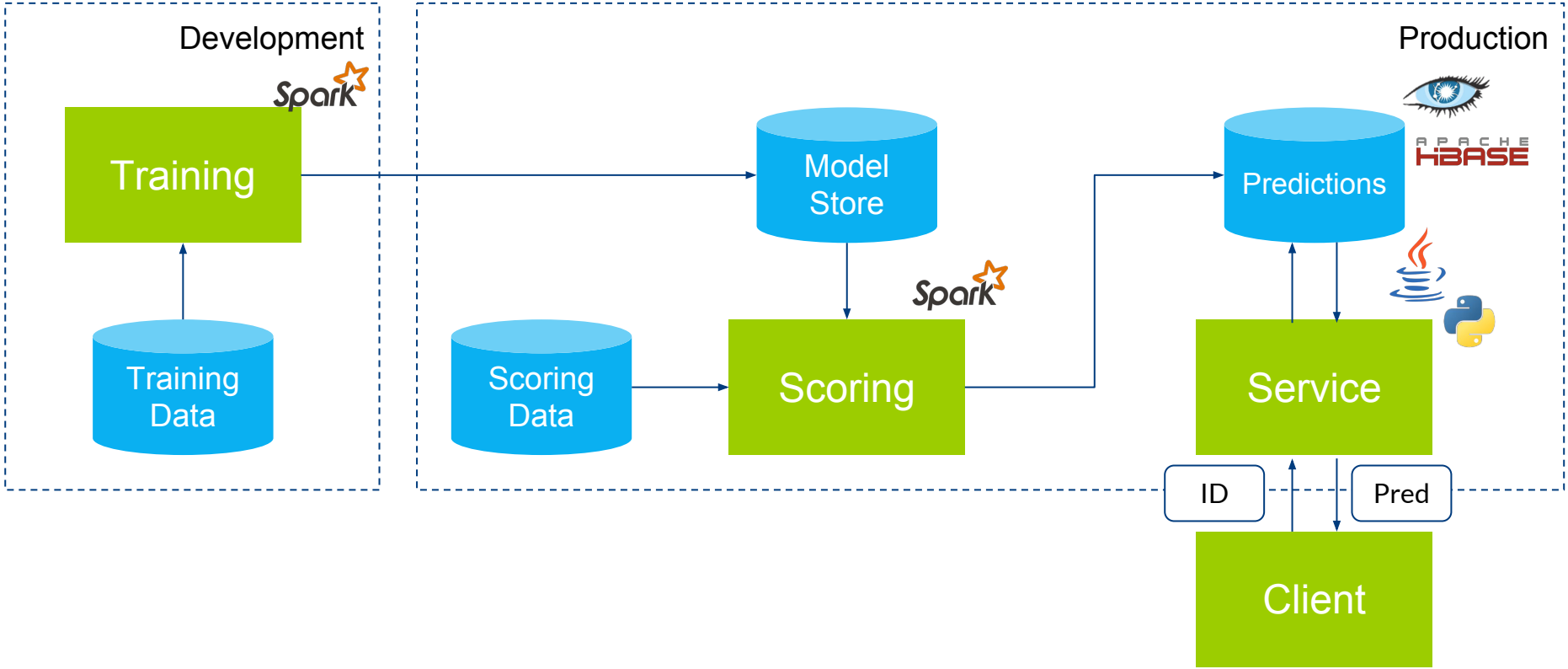
- Voller Funktionsumfang von Python/R
- Isolation von Abhängigkeiten
- Horizontale Skalierbarkeit
- On-Premise oder Cloud



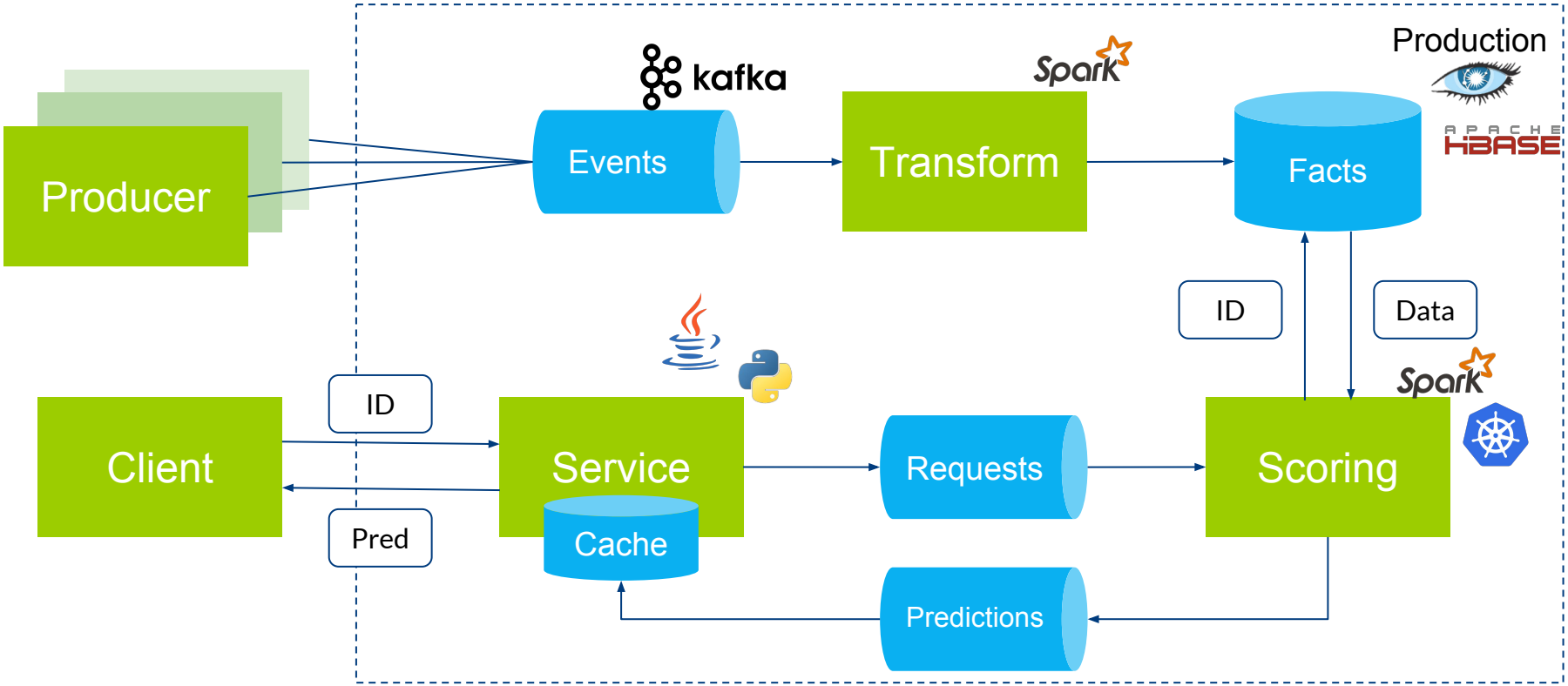
Skalierbarkeit vs. Funktionsumfang



Beispielarchitektur 1: Batch-Scoring

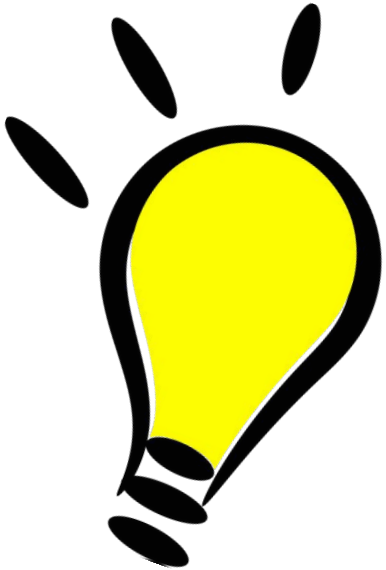


Beispielarchitektur 2: Realtime-Scoring



Zusammenfassung

- ML Deployment umfasst den gesamten Lebenszyklus
- Engineering ist essenziell für produktionsreifes ML
- Pipeline-Konzept erleichtert Reproduzierbarkeit
- Kontinuierliches Performance-Monitoring
- ML Frameworks und Container-basierte Ansätze



Vielen Dank

Marcel Spitzer
Big Data Scientist

inovex GmbH
Ludwig-Erhard-Allee 6
76131 Karlsruhe

marcel.spitzer@inovex.de