AIM3 – Scalable Data Analysis and Data Mining

Machine Learning in Practice and Technical Debt

Christoph Boden, Sebastian Schelter, Juan Soto,
Volker Markl

Fachgebiet Datenbanksysteme und Informationsmanagement
Technische Universität Berlin

http://www.dima.tu-berlin.de/

Techincal Debt

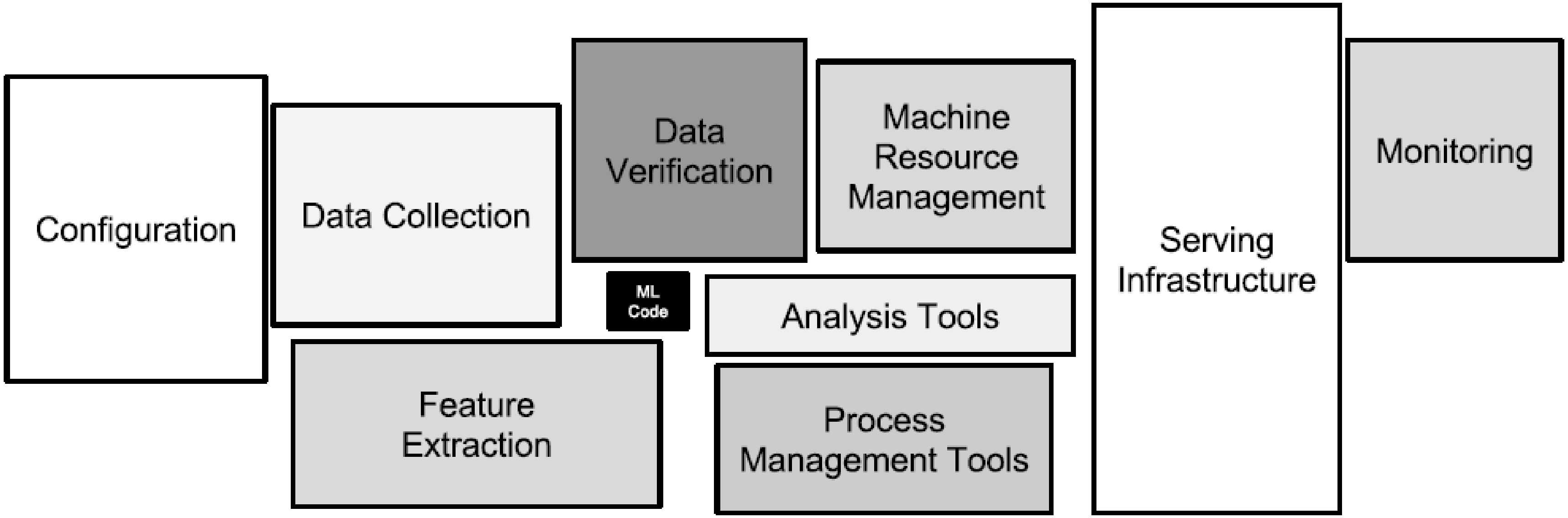
- Complex Models Erode Boundaries
- Entanglement
  - CACE principle: Changing Anything Changes Everything
- Correction Cascades
- Undeclared Consumers.

Data Dependencies Cost More than Code Dependencies

- Unstable Data Dependencies
  - -> Versioning

- Underutilized Data Dependencies
  - Legacy Features
  - Bundled Features
  - ǫ-Features
  - Correlated Features

- Only a small fraction of real-world ML systems is composed of the ML code
- The required surrounding infrastructure is vast and complex



Feedback Loops

- Direct Feedback Loops
- Hidden Feedback Loops

Issues to avoid (Anti Patterns)

- Pipeline Jungles
- Glue Code
- Dead Experimental Codepaths
- Abstraction Debt

Dealing with Changes in the ExternalWorld

- Fixed Thresholds in Dynamic Systems
- Monitoring and Testing
  - Prediction Bias
  - Action Limits
  - Up-Stream Producers.

- Hidden Technical Debt:
  - D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in Machine learning systems. In *Proceedings of the 28th International Conference on Neural Information Processing Systems* (NIPS'15)
  - Tom van der Weide, Dimitris Papadopoulos, Oleg Smirnov, Michal Zielinski, and Tim van Kasteren. 2017. Versioning for End-to-End Machine Learning Pipelines. In *Proceedings of the 1st Workshop on Data Management for End-to-End Machine Learning* (DEEM'17)
  - The Anatomy of a Production-Scale Continuously-Training Machine Learning Platform KDD 2017 (forthcomming)
  - Jimmy Lin and Dmitriy Ryaboy. 2013. Scaling big data mining infrastructure: the twitter experience. *SIGKDD Explor. Newsl.* 14, 2 (April 2013)

  - http://martin.zinkevich.org/rules_of_ml/rules_of_ml.pdf