TensorFlow Serving

- System to productionize the inference part of ML
- Unify many disparate approaches at Google
- Codify best practices for performance & robustness
- 10M+ inferences/sec @ Google
- Open-source; used at Hortonworks, IBM, SAP, ...
- Extensible

Christopher Olston