ML Governance: First Steps

June 23, 2021

Andy Craze
WMF Machine Learning Team

WIKIMEDIA FOUNDATION

Overview

1. Background
2. Related Work
3. Governance Considerations
4. Defining a path to production
5. Next Steps

ML Model Governance

What is it?:

1. The management of production models
    a. Community-developed models
    b. WMF Staff-developed models
2. Implement policies related to
    a. Model lifecycle
    b. Harm Reduction & Remediation
    c. Maintenance
3. Compliance for Privacy laws
    a. GDPR, CCPA, etc.

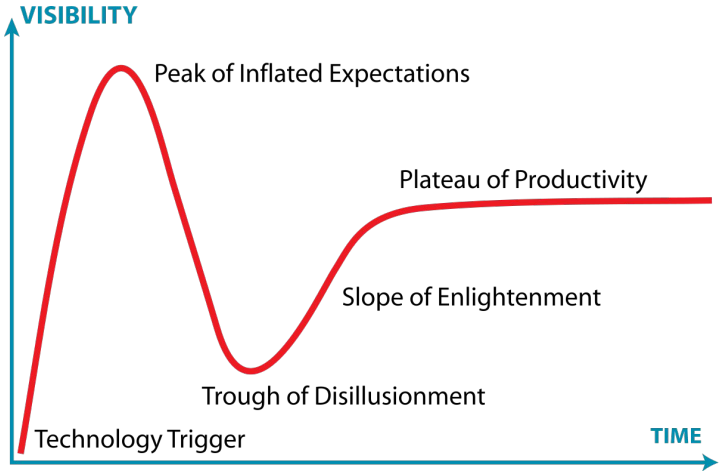
ML Model Governance

Why are we talking about this?:

1. Past the hype peak
   a. Applied ML is valuable
   b. Focus on building Responsible systems

2. “Algorithms are just opinions embedded in code”
   - Cathy O’Neil, *Weapons of Math Destruction*

3. “All models are wrong, but some models are useful.”
   - George Box



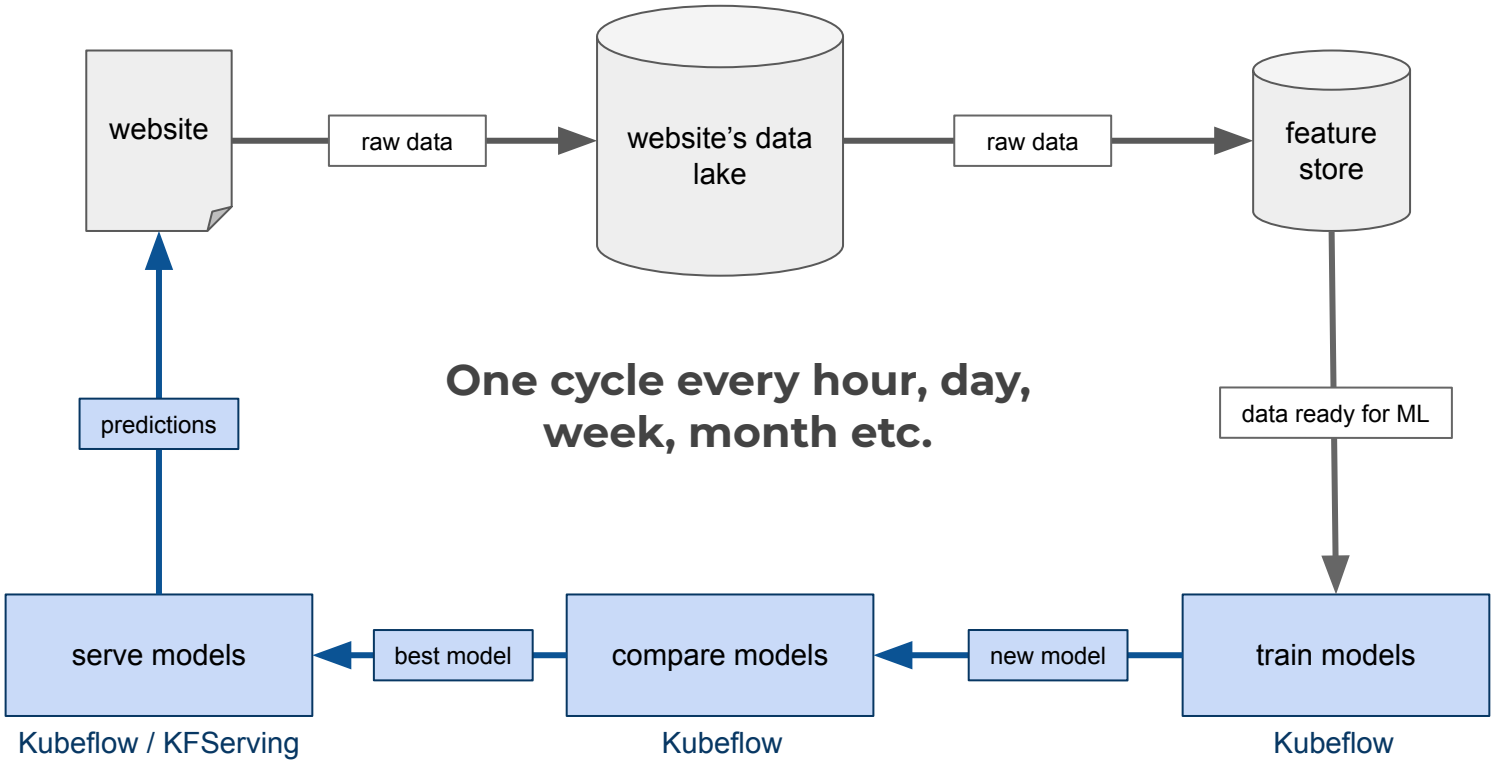
ML Model Governance

Why are we talking about this?:

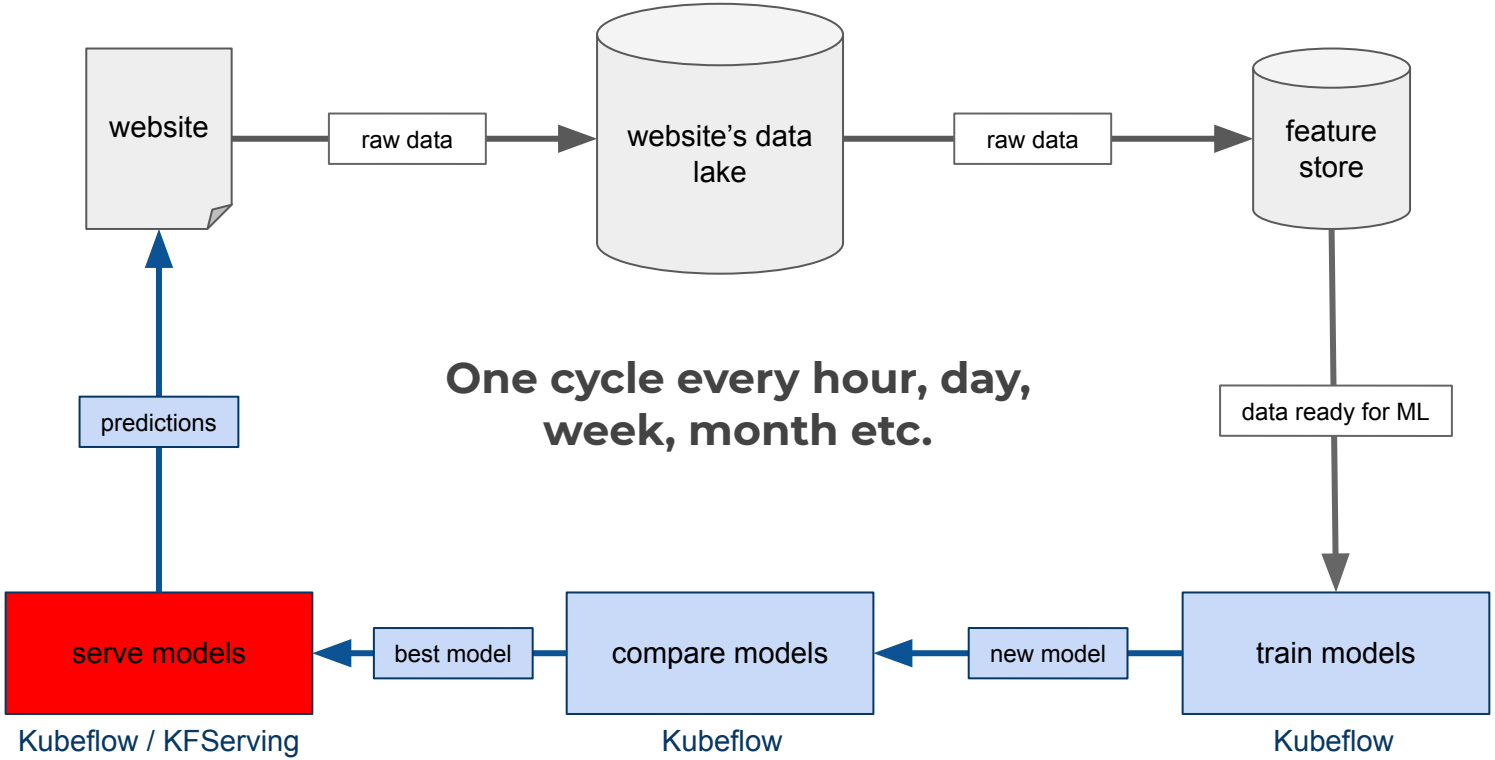
1. The WMF Machine Learning team wants to understand:
    a. Ethical Responsibilities
        i. Transparency
        ii. Accountability
        iii. Fairness
    b. Legal Responsibilities
        i. Data Privacy Laws
        ii. Remediation

2. Embed our values into the governance process
    a. Implicit → Explicit
    b. Balance Short-term gain vs. Long-term value

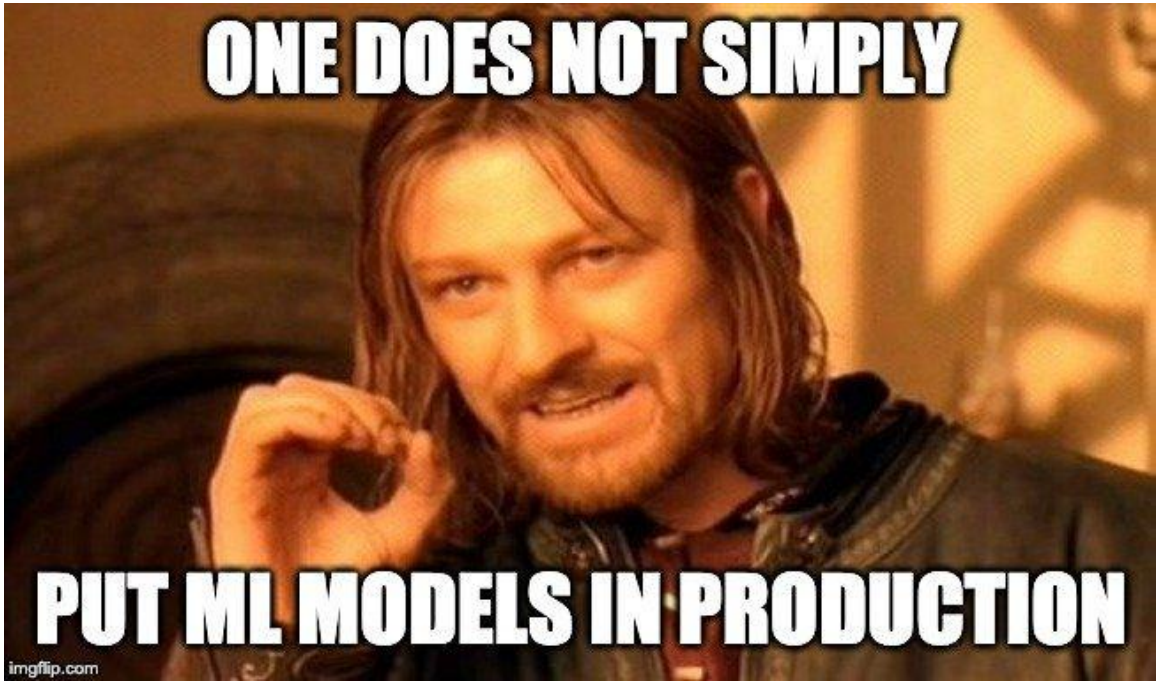
Modern ML Development Cycle



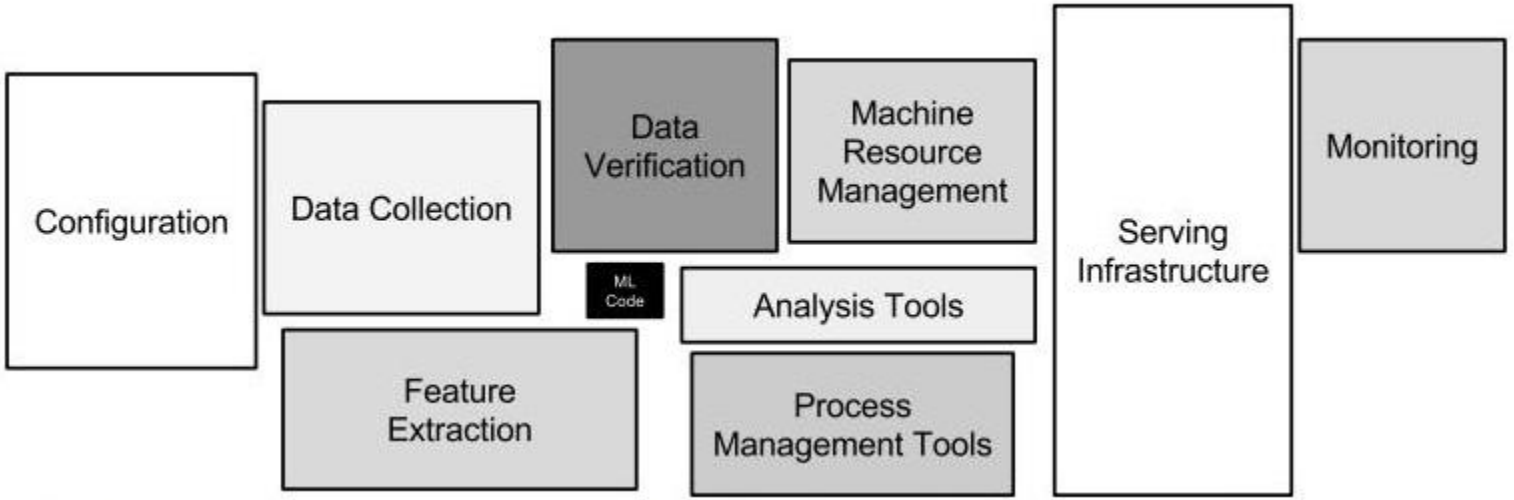
Lift Wing


website
raw data
website's data lake
raw data
feature store
One cycle every hour, day, week, month etc.
predictions
data ready for ML
serve models
best model
compare models
new model
train models
Kubeflow / KFServing
Kubeflow
Kubeflow

ML Governance

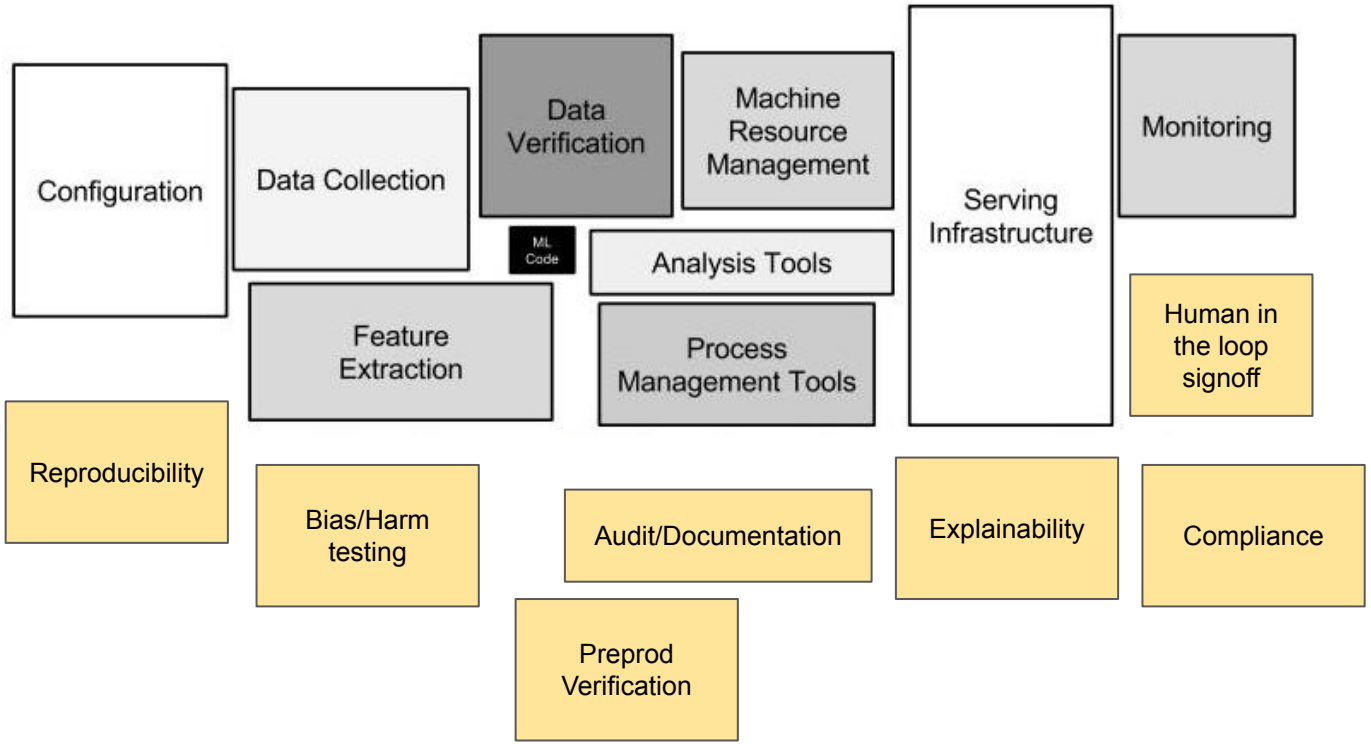


MLOps



From Hidden Technical Debt in Machine Learning Systems - Scully et. al 2015

MLOps Governance



Related Work

Other WMF research & projects:

- Ethical & Human Centered AI - Wikimedia Research 2030. Jonathan T. Morgan, 2019.
  - https://meta.wikimedia.org/wiki/Research:Ethical_and_human-centered_AI#Process_proposals

- ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia, Halfaker, A. & Geiger, S. CSCW'20 Minneapolis, MN, USA (2020).
  - https://arxiv.org/abs/1909.05189

Ethical & Human Centered AI

Process proposals [edit]

Checklists [edit]

> ""Checklists connect principle to practice. Everyone knows to scrub down before the operation. That's the principle. But if you have to check a box on a form after you've done it, you're not likely to forget. That's the practice. And checklists aren't one-shots. A checklist isn't something you read once at some initiation ceremony; a checklist is something you work through with every procedure."[3]

Overview

An ethical AI checklist consists of a list of important steps that must be taken, or questions that must be answered, at each stage of the product development. Checklists work best when the process of working through the checklist is performed consistently, transparently, and collaboratively among team members.

Pros and cons

Pros	Cons
Aids in identification of hidden assumptions, potential negative impacts	Need to be flexible enough to work across products and team workflows, but standardized enough to ensure a baseline level of due diligence
Can cover both concrete requirements ("do this") and softer requirements ("have a conversation about this before proceeding")	Example AI checklists exist, but few have been vetted/tested in actual product development contexts
Facilitates broader participation in decision-making among team members	Binary outcome ("we talked about FOO") may encourage rubber-stamping
Makes it easier for any member of the product team to "flag" missed steps or considerations without fear of reprisal	
Encourages articulation of audience, purpose, and context; success metrics and thresholds	
Increases process consistency between and across teams	
Tracks progress towards goals	

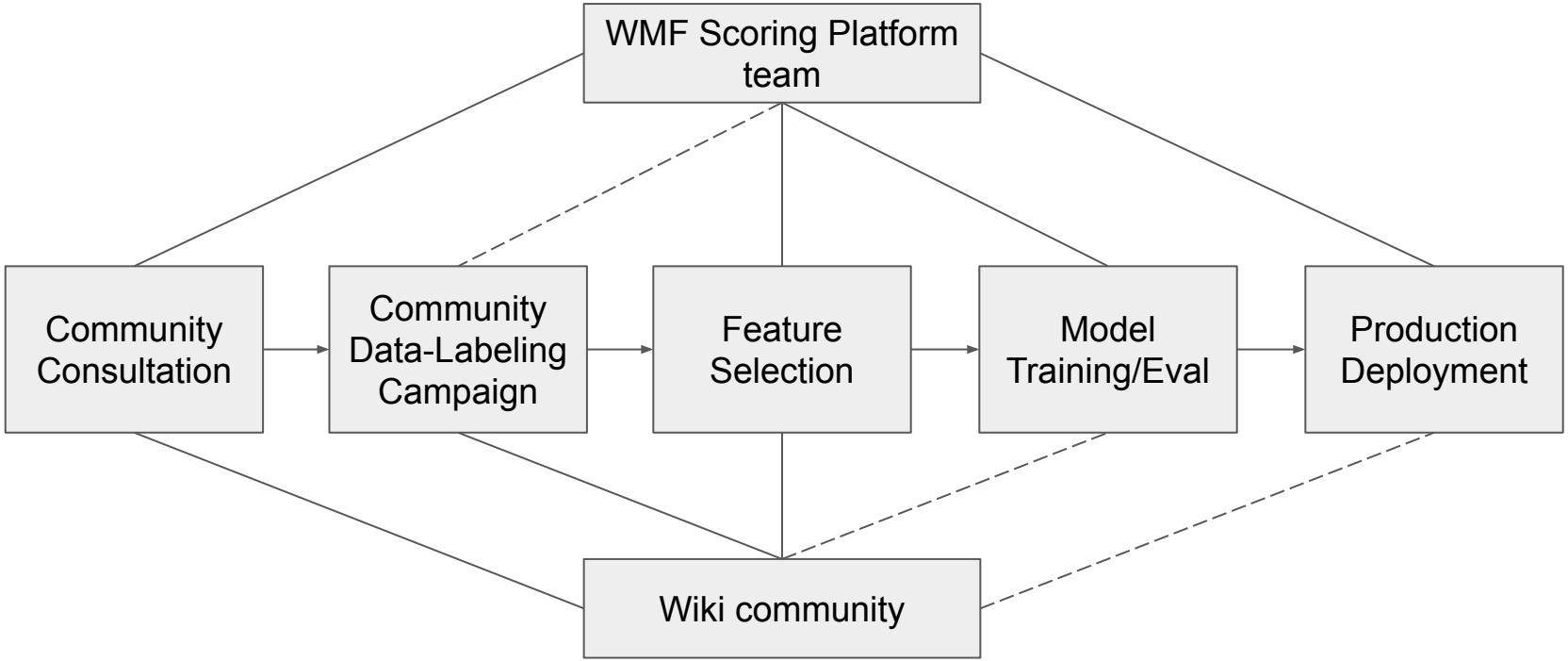
Further reading

1. Of oaths and checklists[3]
2. Care about AI ethics? What you can do, starting today[4]
3. DEON: An Ethics Checklist for Data Scientists[5]
4. Ethical OS Toolkit[6]

Impact assessments [edit]

ORES

Community-based Machine Learning



ORES support checklist

The Objective Revision Evaluation Service (ORES) is a web service that provides machine learning as a service for Wikimedia Projects. The system is designed to help automate critical wiki-work -- for example, vandalism detection and removal. This service is developed as a project of the Wikimedia Scoring Platform team.

Source code for this checklist.

Wikis

Last updated on 10 March 2020 19:52:58 UTC

Wiki	edit quality			article quality				topic routing	
	Basic support	Advanced support		`articlequality`		`draftquality`		`drafttopic`	`articletopic`
		Labeling campaign	model	Labeling campaign	model	Labeling campaign	model	model	model
arwiki	n/a		✓						✓
azwiki		0%							
bawiki		0%							
bnwiki	✓	37%							
bnwikisource		1%							
bswiki	n/a		✓						
cawiki	n/a		✓						
cswiki	n/a		✓						✓
dewiki	n/a		✓						
elwiki	✓								
enwiki	n/a	28%	✓		✓		✓	✓	✓
enwiktionary	✓	0%							
eswiki	n/a		✓						
eswikibooks	n/a		✓						
eswikiquote	n/a		✓						

Governance Considerations

What should a policy cover?

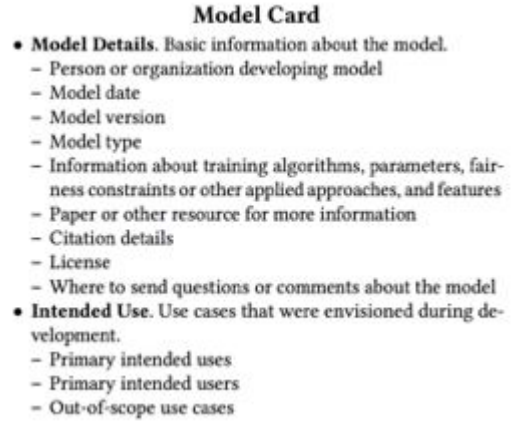
1. **Reproducibility**
2. **Audit & Documentation**
3. **Human-in-the-loop sign off**
4. Pre-prod verification
5. **Transparency & Explainability**
6. **Bias/harm testing**
7. Production Requirements
8. Production Monitoring
9. **Data quality & compliance**

Reproducibility

Audit & Documentation

1. Full logs of changes / decisions made
    a. Experiments ran
    b. Rationale for technical decisions

2. Model Cards (Mitchell et. al 2019)
    a. Living Documentation
    b. Balancing information density
        i. Inclusive for non-experts
3. Dataset documentation
    a. Training data -> *a priori* knowledge
    b. Archival methods (Jo et. al 2020)



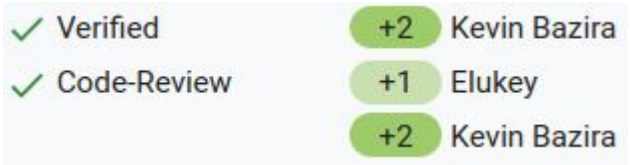
Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases

Human-in-the-loop sign off

Multiple stakeholders should sign off on:

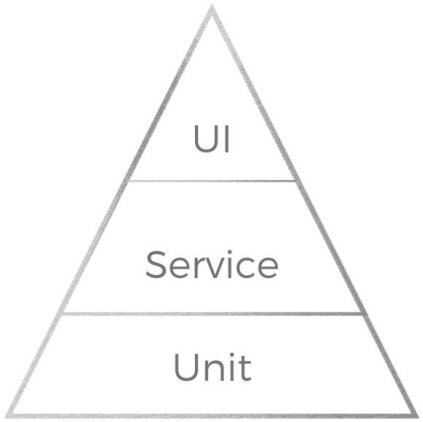
1. Going from test -> production
    a. Similar to code review (Gerrit)
    b. Should include non-technical perspectives
        i. How do we do this consistently?



Preprod Verification

Can we verify that the model software works?

1. Automated checks
    a. Reproducible build
    b. Lint
    c. Code style
2. Testing
    a. Unit Testing
    b. E2E Tests



Transparency & Explainability

Can we explain how the model works?

1. Explainable AI (XAI)
    a. Tree-based models can provide maximum explainability
        i. Feature importance, proof by counterexample
    b. Regression models can use SHAP, LIME, etc.

2. Can we be transparent with opaque models?
    a. Neural Networks
    b. Gradient Boosted algorithms

3. Inclusive to non-experts
    a. Are visual explanations helpful?

Bias and Harm testing

How can we avoid perpetuating harm?

1. Adversarial “red team” testing
    a. Creating baselines on protect variable classes
2. Automated bias/fairness checking on various models
    a. How do we define fairness in a global context?
    b. Open source tools
        i. General Awareness Tool: LIME
        ii. Debiaising tool: AIF360
3. Detecting undocumented training data
    a. Datasheets for Datasets (Gebru et al. 2018)

Production requirements

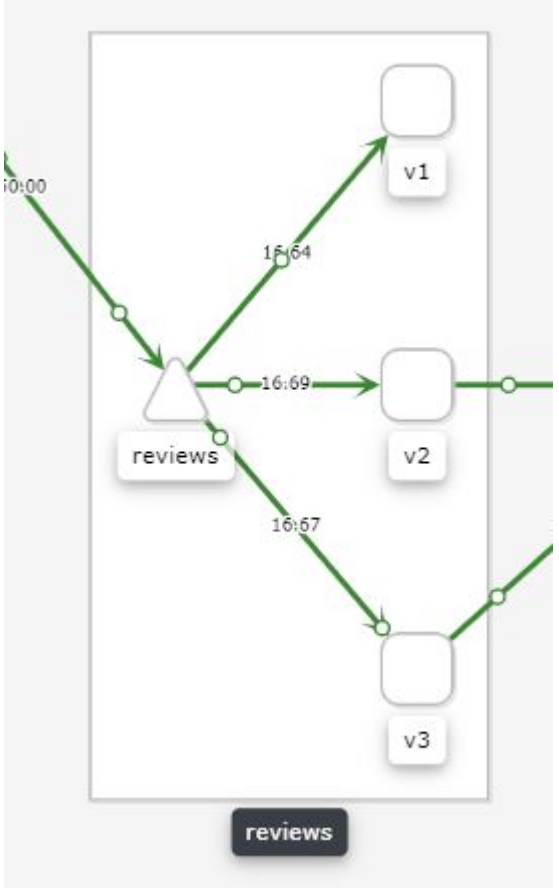
What resources does the model need?

1. Containerized deployment with scalable high-availability
    a. High-throughput/high-traffic models
    b. Batch-scoring models
2. Automated stress/load testing prior to deployment
    a. Handle high or burst traffic
3. Smaller/low-powered models might only need a single prod service.

Production Monitoring

How can we observe the live model?

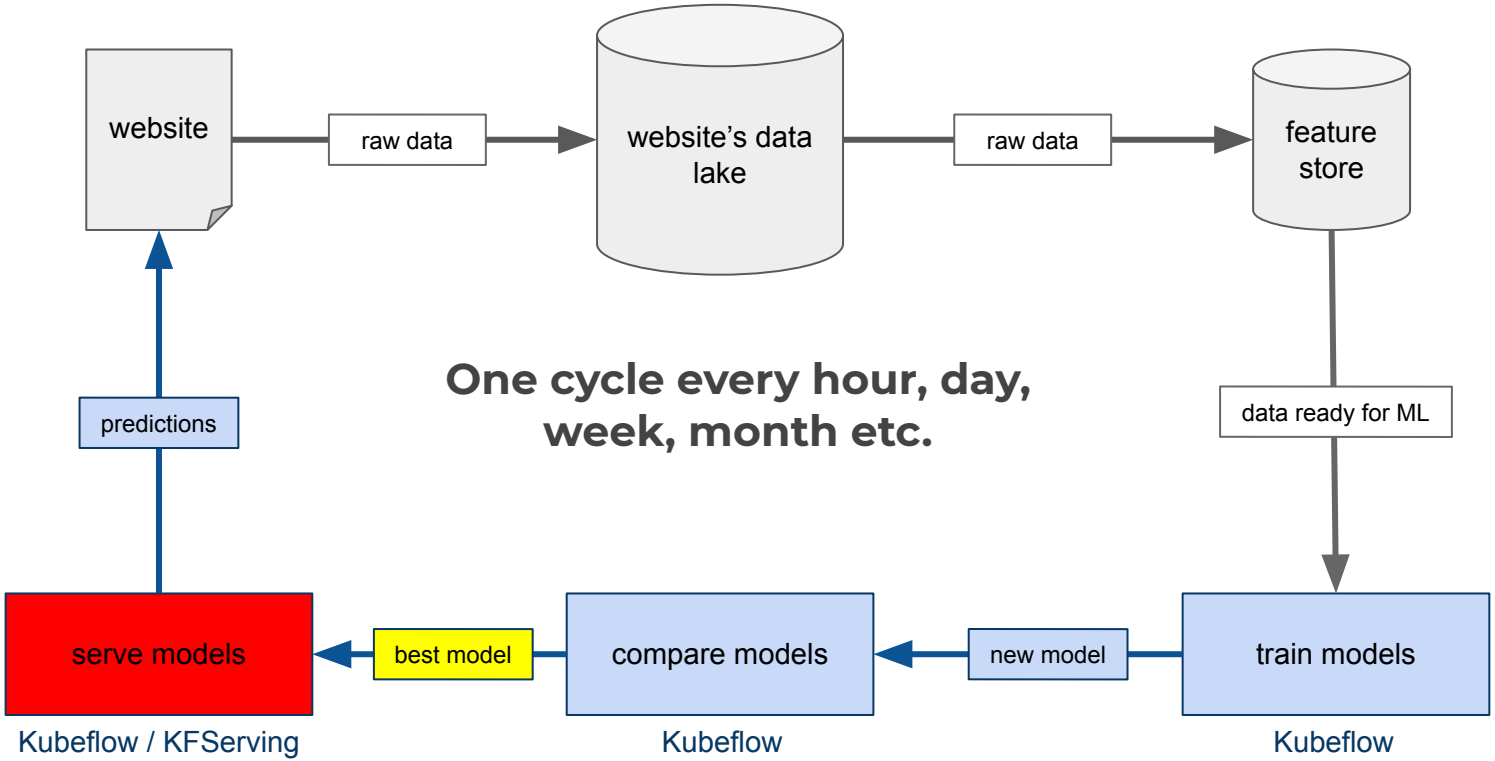
1. Real-time error alerting
   a. High-throughput/high-traffic models
   b. Batch-scoring models
2. Model Balancing
   a. Handle high or burst traffic
3. Automated nightly retraining
4. Model Evaluation & redeployment pipeline



Data quality & compliance

1. Drift Detection
    a. All models degrade over time
2. PII considerations
3. Data quality
    a. Is data still valid/appropriate?
4. Legal compliance
    a. GDPR
        i. Right to Explanation
        ii. Right to be forgotten
        iii. Minimize Discriminatory effects

Defining a path to production



Policy Integration

Can we embed our values into our MLOps Process?

1. How can we put principles into practice?
    a. Checklists to determine model suitability
2. How do we keep track of models in production
    a. Registry with information about each suitable model against each governance consideration
3. Get more non-technical contributors involved.
    a. Continue to lower the barriers to participation
    b. Focus more on documentation
        i. Datasets
        ii. Model cards

Next Steps

Join the conversation!

1. Create Draft Model Deployment Guidelines
    a. https://phabricator.wikimedia.org/T276598

2. Model Reporting
    a. https://phabricator.wikimedia.org/T276397
    b. Model Cards
        i. On-wiki experiment
        https://phabricator.wikimedia.org/T276398

References

1. Morgan, Jonathan. *Ethical & Human Centered AI*. Apr. 2019. *figshare.com*, doi:10.6084/m9.figshare.8044553.v1.
2. Halfaker, Aaron, and R. Stuart Geiger. “ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia.” *ArXiv:1909.05189 [Cs]*, Aug. 2020. *arXiv.org*, http://arxiv.org/abs/1909.05189.
3. Mitchell, Margaret, et al. “Model Cards for Model Reporting.” *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Jan. 2019, pp. 220–29. *arXiv.org*, doi:10.1145/3287560.3287596.
4. Gebru, Timnit, et al. “Datasheets for Datasets.” ArXiv:1803.09010 [Cs], Mar. 2020. arXiv.org, http://arxiv.org/abs/1803.09010.
5. Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. InProceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 306–316.
6. Treveil, Mark, et al. Introducing MLOps: How to Scale Machine Learning in the Enterprise. 2021. Open WorldCat, http://proquest.safaribooksonline.com/?fpi=9781492083283.

Thanks!

Andy Craze
WMF Machine Learning team
acraze@wikimedia.org