worldusabilityday

Using Human-Centered Machine-Learning (HCML) to Improve Data Quality & Data Governance Projects

Edward F. O’Connor

ManTech®

Certified Section 508 Compliant

11.12.2020

Human-Centered Design Center of Excellence

Agenda

1) Basics
2) Breakdown
3) Techniques & Examples
4) References & Resources

Part 1: Basics

What is Human-Centered Machine Learning

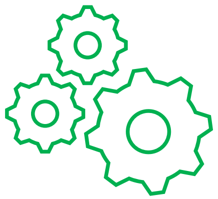
<u>Must</u> include **"reframing machine learning workflows"** (Gillies et al., 2016) around real-world activities and involving the people impacted by the system in its definition and ongoing operations.

Human-Centered Design Center of Excellence

Why ML needs Governance & HCD

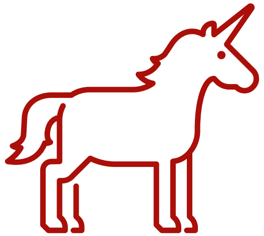
Monolith vs. Exposed Activities

Prototype ML systems have a limited number of sub-systems which expand dramatically in production. It is critical to make each sub-system transparent and map to operational staff early.



+

+

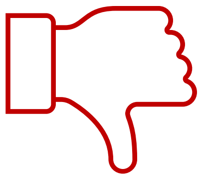


Unrealistic Skillset Combo vs. Balanced Teams

ML projects revolve around data scientists and engineering experts with unrealistic skillsets. Teams need to decompose the work in a way that is makes it easier to hire and sustain staff.

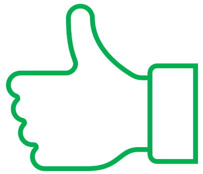
=

=



Failure to Sustain vs. Ongoing Value

Many ML projects don’t get to production, are financially unsustainable, or regarded as long-shot investments. Maintenance costs can balloon, and users often don’t see value.



Wait… what are (some of) the “basics”?

… basic statistics, probability, multivariable calculus, linear algebra, regression, SVM, KNN, decision-trees, KMeans …

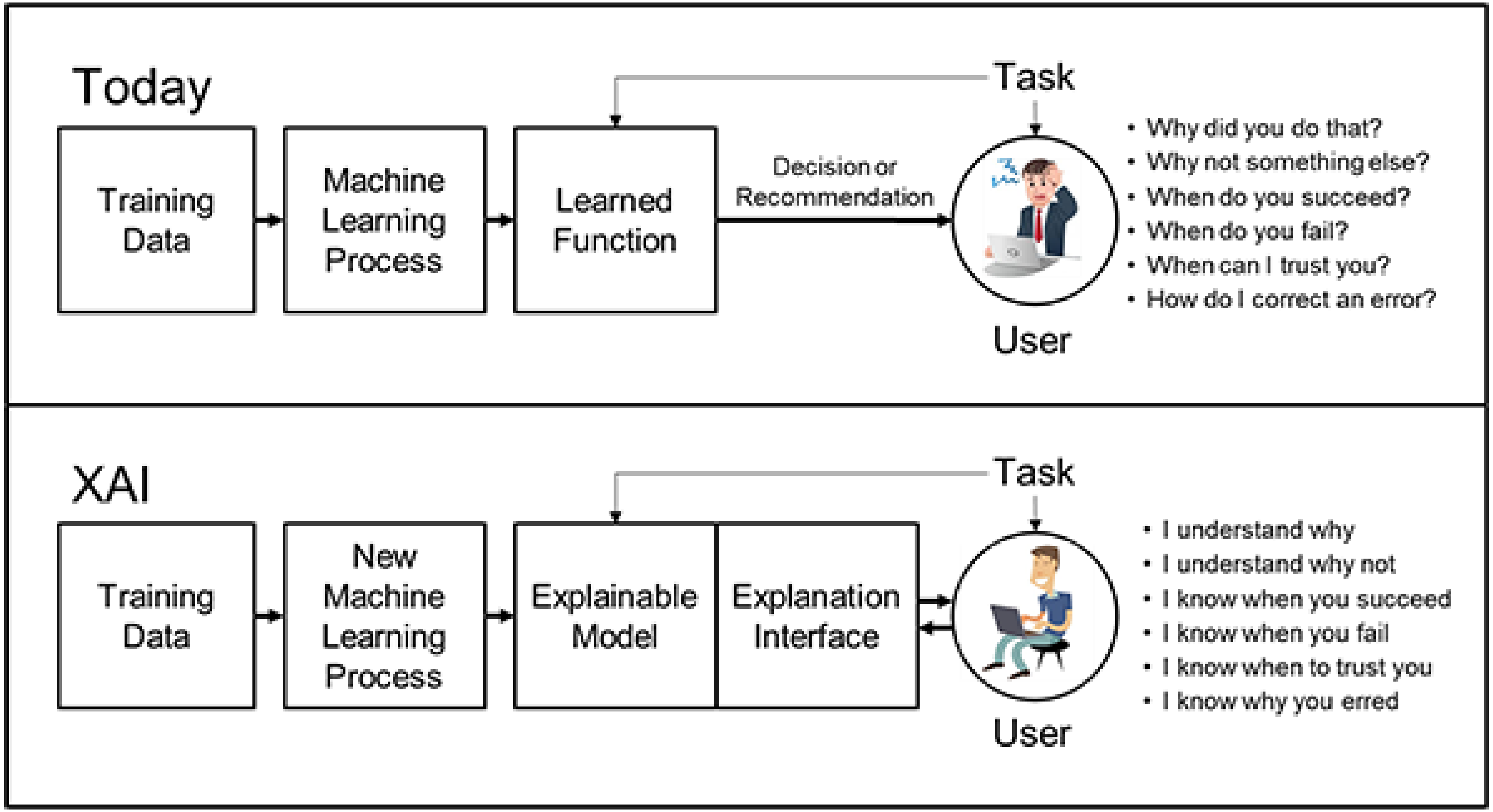
… data wrangling, imputation, encoding, transformation (PCA, LDA, etc.), Python (Numpy, Pandas, Matplotlib, Seaborn, scikit-learn), R, visualization …

… access control, lineage, quality, explainability, transparency, compliance, security, interpretability, measurement …

… ethics, collaboration, law, performance testing, infrastructure, industry experience, HCI, UXD …

Human-Centered Design
Center of Excellence

Key Lesson for HCML – Reject the Black Box



Explainable artificial intelligence (XAI) – what is different according to DARPA (Turek, 2018).

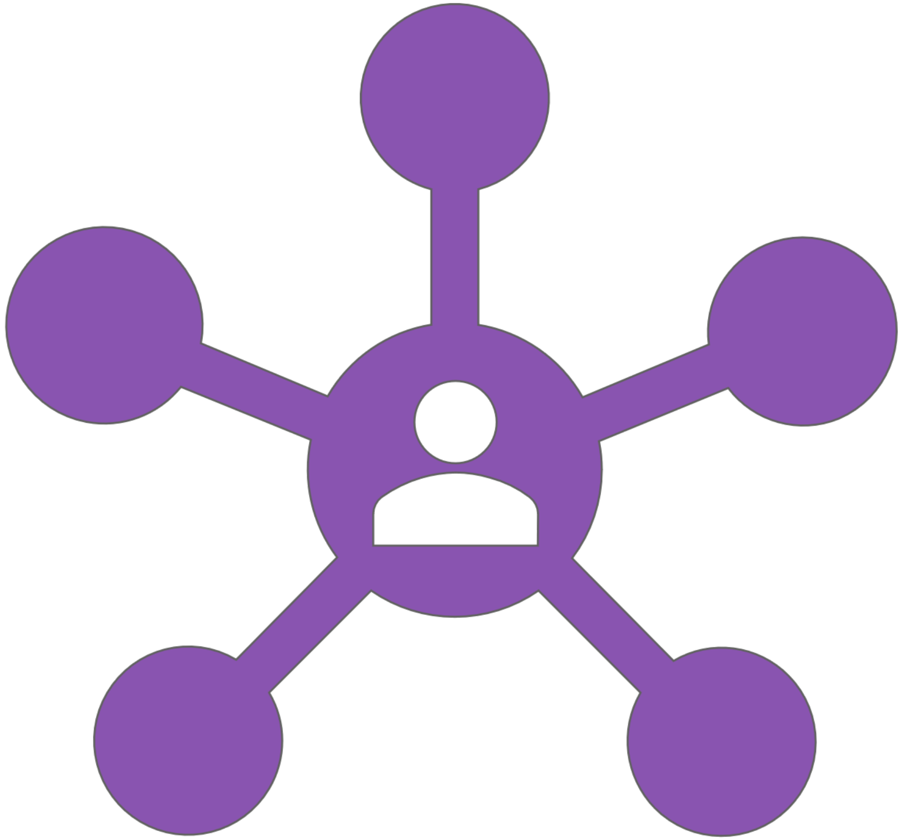
Takeaway: Operational Definitions Matter

- Human Centered Design or Centred (ISO 9241, 2019)?
- Where exactly do people get involved as the system evolves over time (Amershi et al., 2014)?
- What is the relationship between explainability, interpretability, and completeness? (Gilpin et al., 2018)

Don't be part of the next big industry disconnect:

An old story: "This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems." – (Breiman, 2001)

Part 2: Breakdown



Simplest(?) View of ML Activities

Deployment & Integration

Data Gathering & Improvement

Model Monitoring & Testing

Model Training & Selection

HCML Pieces – Deployment & Integration

Early Automation

Define the CI/CD pipeline early and automate it from the get-go. Prioritize avoiding technical debt related to operationalization (Sculley et al., 2015).

UI Prototyping

How will the system intersect with users? Review user workflow, mock it up, and measure UI integration difficulty.

Appeals & Feedback

How will users tell you when something is (probably) wrong? How will feedback loops work? What are the legal/regulatory requirements (Hacker et al., 2020)?

Human-Centered Design
Center of Excellence

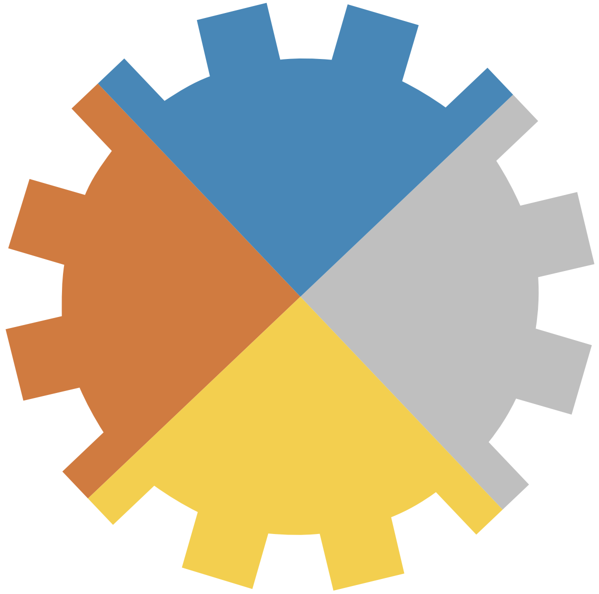
HCML Pieces – Model Monitoring

Auto Explanations & Expert Review

Why was this decision made? Black-box models do not belong in operational healthcare.

Broaden Key Performance Indicators (KPIs)

KPIs that look at accessibility, fairness, frequency and speed of use, impact to decision-making, and feedback-loop and appeals usage are key.



Testing, Assessment, & Diagnostics

Q-Q and lift curves are the basics – add in residual analysis, sensitivity and other testing.

HCML Pieces – Model Training & Selection

Simplify Models

Try different approaches to defining features and constrain. Avoid the black box and measure KPI impact of additions (Rudin & Radin, 2019).

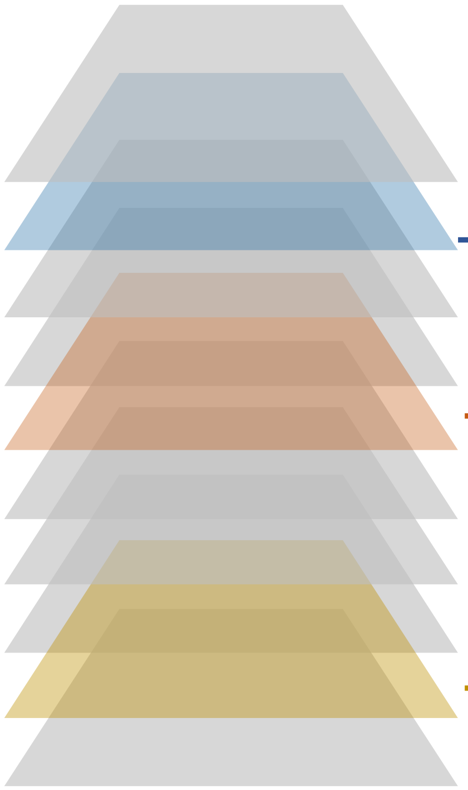
Focus on Interpretability

Prioritize interpretability early and factor it into model decisions.

Baseline and Measure Improvements

Utilize ALL your KPIs as guidance for model tuning and selection. Take documented baselines and create timeline dashboards on how the model has changed over time.

HCML Pieces – Data Gathering & Improvement



Measure the Impact to KPIs of Data Quality Improvement

Measure the impact of each data improvement task against downstream KPIs. Could that KPI be moved more effectively in another part of the overall process?

Sampling, De-Identification, and Synthetic Data

What is the difference between de-identified data and synthetic data created by reviewing the parameters of a data-set?

Define an Outreach Strategy

How will you influence quality improvement in your data sources? Can you incentivize, offer to assist, setup easy to use data-checking tools and propagate them? Do not give up on improving upstream data even if it seems impossible.

Takeaways on the HCML Life-Cycle

Less Technical

- Balance your team
- Prioritize explainability and interpretability
- Code from requirements, go old-school on roles
- Define process for appeals and quality feedback loops

More Technical

- Architect around activities
- Emphasize set-based architecture, learn on OSS before final tool selection, use throwaway prototypes
- Parallel-deploy, performance test, and don’t skip any steps

Part 3: Techniques & Examples

Context Behind the Discussion

Loosely coupled integrated delivery system (IDS) in the safety-net sector including a hospital, clinic system, and key community services.

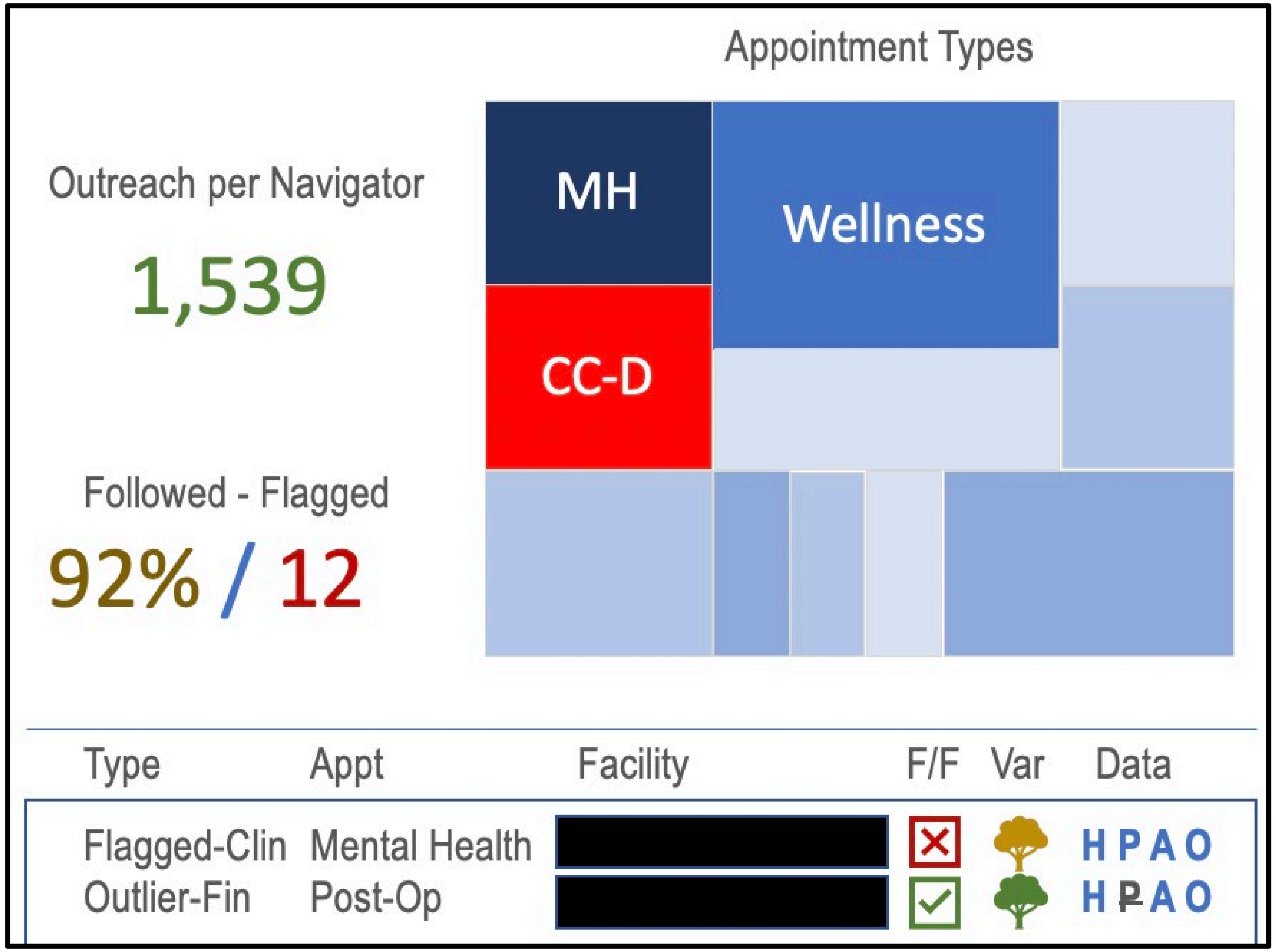
The IDS is using ML to:

1. Analyze patient data and drive prioritization for care coordination, patient navigation, or medical management.
2. Evaluate incoming (batch or heavy stream) as it comes in and providing a real-time feedback loop to data providers.
3. Review data-sets for anomalies and making recommendations for audit/review.

Using Dashboards to Support Governance

Dashboards facilitate easy viewing of all decisions:

- Model used
- Data involved
- KPI relationships
- Assets describing why a decision was made
- And a LOT more…

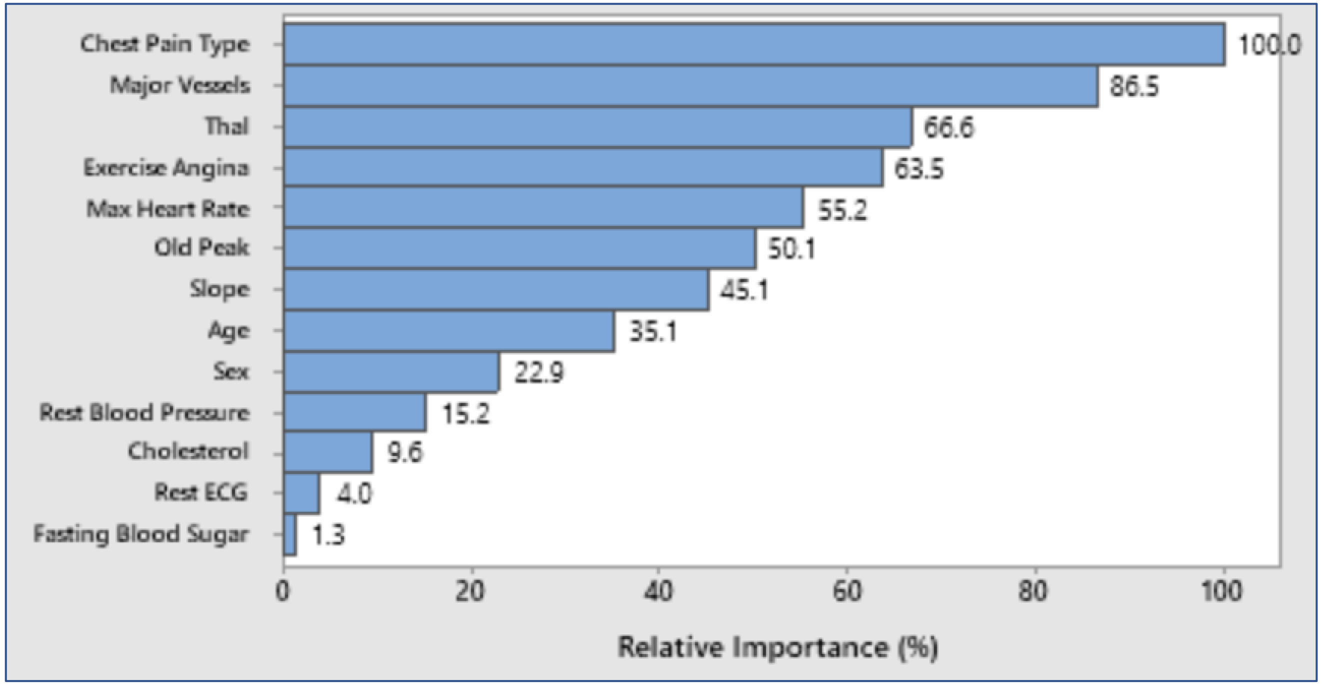


Human-Centered Design
Center of Excellence

Common Assets for Explainability

Assets for interpretability and explainability should be auto-generated when possible – but can include anything created by the system OR team:

Relative Variable Importance



(Minitab, n.d., as discussed by Chauncey et al., 2012).

Bayesian Rules List

- **if** hemiplegia and age > 60
  - **then** stroke risk 58.9% (53.8%–63.8%)
- **else if** cerebrovascular disorder
  - **then** stroke risk 47.8% (44.8%–50.7%)
- **else if** transient ischaemic attack
  - **then** stroke risk 23.8% (19.5%–28.4%)
- **else if** occlusion and stenosis of carotid artery without infarction
  - **then** stroke risk 15.8% (12.2%–19.6%)
- **else if** altered state of consciousness and age > 60
  - **then** stroke risk 16.0% (12.2%–20.2%)
- **else if** age ≤ 70
  - **then** stroke risk 4.6% (3.9%–5.4%)
- **else** stroke risk 8.7% (7.9%–9.6%)

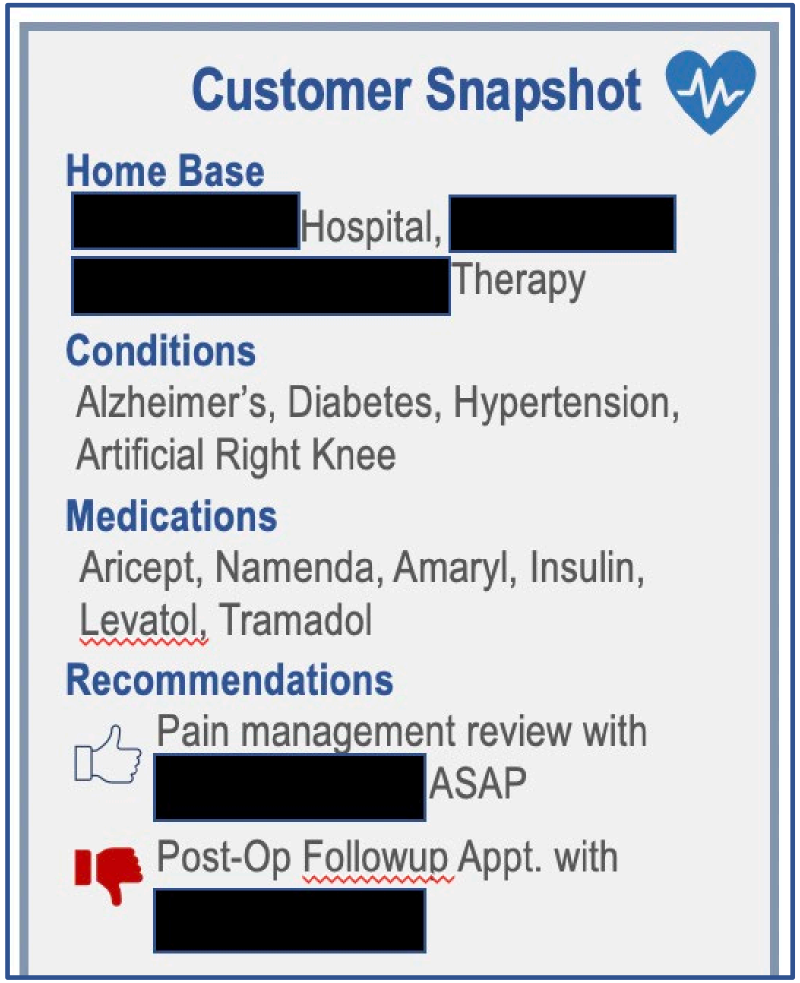
(Letham et al., 2015, as cited in Gunning, 2017).

Gather Feedback in Real-Time

In addition to ML-specific techniques for monitoring the model's performance – also gather feedback from users however possible.

Techniques include:

- Immediate rating gathering
- Noting follow, ignore, or variance
- Identifying themes including user types, facilities, encounters, etc.
- Try to do it fast (near or real-time)



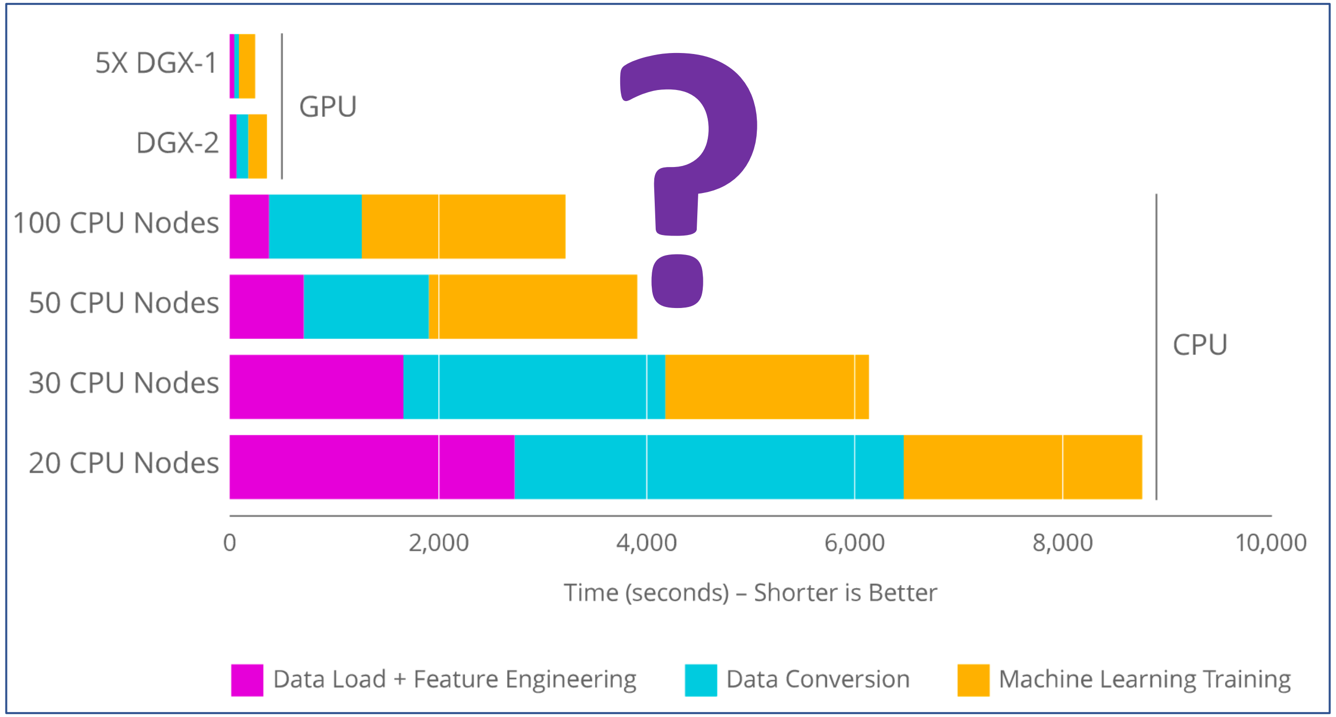
Wrap Model-Gen Rules with Expert Rules

- We knew that models could (and should) be constrained using expert-provided deterministic rules
- Exactly **how** you implement this is worth considering early, let's talk about **why**
- Understanding what is inside your platform can help. OSS-BSD licensed tools are often inside proprietary platforms – make sure they implemented rule-combining as well as you could on your own (Haas, 2020)

Rework for Speed

You will hit a point where parts of your pipeline are too slow. Have a culture of:

- Set-based design
- Throwaway prototyping
- Code porting
- Use case audit



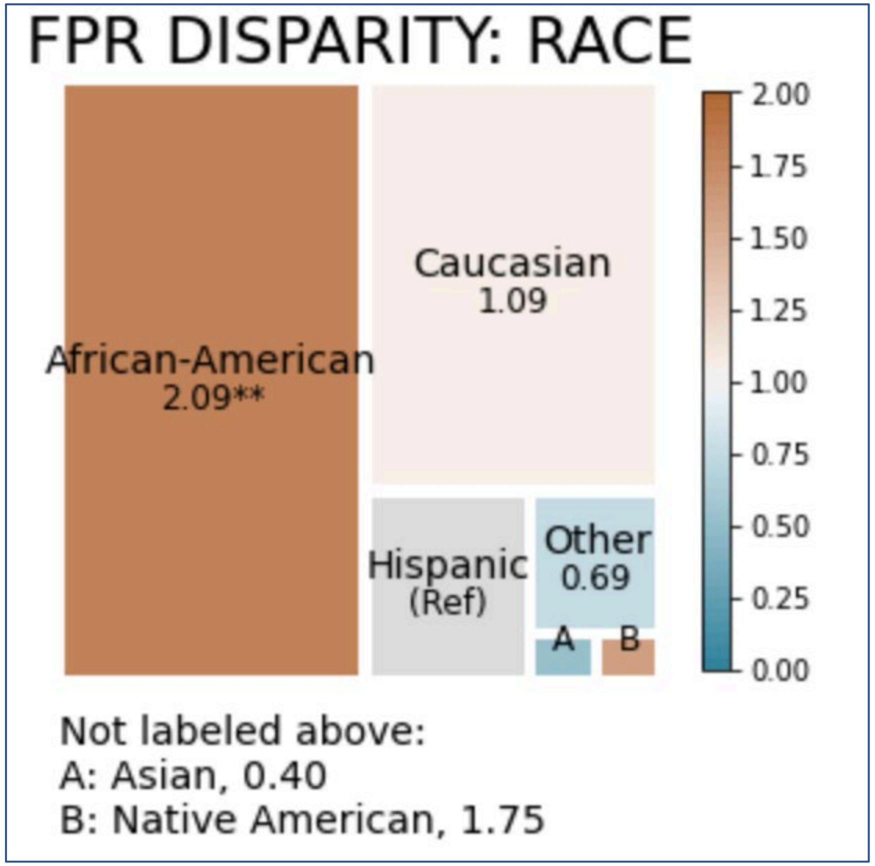
Performance claims/estimates from NVIDIA RAPIDS (www.rapids.ai).

Or: Slow Things Down and Modularize

Aequitas is an open-source auditing tool focused on identifying discrimination and bias. The IDS needed to report metrics on facility coverage to a political audience – and utilized both:

- Proportional encounter analysis
- Auto-generated audits from Aequitas

Use of the tool evolved from: 1) manual off-line use, to 2) automated daily reports by facility, to 3) integration into KPIs and (pending?) live dashboards.



COMPAS Analysis Demo (Aeguitas, n.d., supporting Angwin et al., 2016)

Takeaways from Real-World Systems

- Distinguish between data wrangling and using ML for spotting data quality issues; leverage the former to support/automate the latter
- Healthcare use-cases often require multi-expert input – joining clinical, claims, and community data is commonly required
- Use cases often need very fast processing and UI-embedding
- Packaged software is improving quickly – but can need help from or lag behind open-source options (see OSS resources, slides 29-30)
- Anti-patterns are still being understood and probably include:
    - (False?) - Spending 80% of your time in wrangling is required
    - (Mostly False?) - Lowering complexity will decrease accuracy

CONTACT ME

DETAILS

ManTech®

www.ManTech.com

Edward F. O’Connor

Edward.OConnor@ManTech.com

www.linkedin.com/in/eddieoak/

Human-Centered Design
Center of Excellence

References

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza. (2014). *Power to the people: The Role of humans in interactive machine learning.* AI Magazine. https://doi.org/10.1609/aimag.v35i4.2513

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks.* Pro Publica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Breiman, L. (2001). Statistical modeling: The Two cultures. *Statistical Science, 16*(3), 199-231. http://dx.doi.org/10.1214/ss/1009213726

Chancellor, S., Baumer, E., & De Choudhury, M. (2019). Who is the "human" in human-centered machine learning: The Case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction, Article 147.* https://dl.acm.org/doi/10.1145/3359249

Gillies, M., Fiebrink, R., Tanaka, A., Garcia, J., Bevilacqua, F., Heloir, A., Nunnari, F., Mackay, W., Amershi, S., Lee, B., d'Alessandro, N., Tilmanne, J., Kulesza, T., & Caramiaux, B. (2016). Human-centred machine learning. *CHI EA '16: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems,* 3558 – 3565. https://doi.org/10.1145/2851581.2856492

Chauncey, D., Inozu, B., Kamataris, V., & Mount, C. (2012). *Performance improvement for healthcare: Leading change with lean, six sigma, and constraints management.* McGraw-Hill.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. doi:10.1109/dsaa.2018.00018

Gunning, D. (2017). *Explainable artificial intelligence (XAI).* Defense Advanced Research Projects Agency (DARPA). http://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

Haas, L. (2020). *Hybrid rule-based machine learning with scikit-learn.* Towards Data Science (Medium). https://towardsdatascience.com/hybrid-rule-based-machine-learning-with-scikit-learn-9cb9841bebf2

Hacker, P., Krestel, R., Grundmann, S., & Naumann, F. (2020). Explainable AI under contract and tort law: legal incentives and technical challenges. *Artificial Intelligence and Law*, 1-25. https://link.springer.com/content/pdf/10.1007/s10506-020-09260-6.pdf

Hall, P. & Gill, N. (2018). *An Introduction to machine learning interpretability* (2nd ed.). O'Reilly. https://www.oreilly.com/library/view/an-introduction-to/9781492033158/

International Organization for Standardization. (2019). *Ergonomics of human-system interaction – part 210: Human-centred design for interactive systems.* (ISO Standard No.9241-210:2019). https://www.iso.org/standard/77520.html

Letham, B., Rudin. C., McCormick, T., and Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics 2015, 9*(3), 1350-1370.

Minitab (n.d.). *Relative variable importance chart for CART® Classification.* https://support.minitab.com/en-us/minitab/19/help-and-how-to/statistical-modeling/predictive-analytics/how-to/cart-classification/interpret-the-results/all-statistics-and-graphs/relative-variable-importance-chart/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Roy, A., Nikolitch, K., McGinn, R., Jinah, S., Klement, W., & Kaminsky, Z. A. (2020). A machine learning approach predicts future risk to suicidal ideation from social media data. *NPJ Digital Medicine*, *3*, 78. https://doi.org/10.1038/s41746-020-0287-6

Rudin, C. & Radin, J. (2019). *Why are we using black box models in AI when we don't need to? A Lesson from an explainable AI competition.* Harvard Data Science Review. https://doi.org/10.1162/99608f92.5a8a3a3d

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in neural information processing systems* (pp. 2503-2511).

Turek, M. (2018). *Explainable artificial intelligence (XAI).* Defense Advanced Research Projects Agency. https://www.darpa.mil/program/explainable-artificial-intelligence

Open Source Software Resources

Apache Arrow – A cross-language development platform for in-memory analytics [Computer Software]. https://github.com/apache/arrow

Aequitas - Bias & Fairness Audit [Computer Software]. https://github.com/dssg/aequitas

AI Fairness 360 (AIF360) [Computer Software]. https://github.com/Trusted-AI/AIF360

CleverHans – A Python library to benchmark machine learning systems' vulnerability to adversarial examples [Computer Software]. https://github.com/tensorflow/cleverhans

cuML – GPU machine learning algorithms (NVIDIA RAPIDS) [Computer Software]. https://github.com/rapidsai/cuml

GoAi – GPU Open Analytics Initiative [Computer Software]. https://github.com/gpuopenanalytics

LIME: Explaining the predictions of any machine learning classifier [Computer Software]. https://github.com/marcotcr/lime

Mlflow: A Machine learning lifecycle platform [Computer Software]. https://github.com/mlflow/mlflow

RuleFit – Fit Lasso model to binary rules created from tree ensembles [Computer Software]. https://github.com/Zelazny7/rulefit

Open Source Software Resources - Continued

SHAP is a game theoretic approach to explain the output of any machine learning model [Computer Software]. https://github.com/slundberg/shap

Skater – Python Library for Model Interpretation/Explanations [Computer Software]. https://github.com/oracle/Skater

Themis – Software fairness tester [Computer Software]. https://github.com/LASER-UMASS/Themis

TreeInterpreter – Package for interpreting scikit-learn's decision tree and random forest predictions [Computer Software]. https://github.com/andosa/treeinterpreter

What-If Tool – Interface for expanding understanding of a black-box classification or regression ML model [Computer Software]. https://github.com/PAIR-code/what-if-tool

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable [Computer Software]. https://github.com/dmlc/xgboost