“Housekeeping”

Twitter: #ACMLearning

- Welcome to today’s ACM Learning Webinar, “**Explainable Models for Healthcare AI”** The presentation starts at the top of the hour and lasts 60 minutes. Audio and video will automatically play throughout the event. On the bottom panel you’ll find a number of widgets, including Twitter, Sharing, and Wikipedia apps.

- If you are experiencing any problems/issues, **refresh** your console by pressing the **F5** key on your keyboard in Windows, **Command + R** if on a Mac, or refresh your browser if you’re on a mobile device; or close and re-launch the presentation. You can also view the Webcast Help Guide, by clicking on the “Help” widget in the bottom dock.

- To control volume, adjust the master volume on your computer. If the volume is still too low, use headphones.

- At the end of the presentation, you’ll see a survey open on your screen. Please take a minute to fill it out to help us improve your next webinar experience. You may also open the survey at any time throughout the presentation from resources window.

- This session is being recorded and will be archived for on-demand viewing in the next 1-2 days. You will receive an automatic email notification when it is available, and check http://learning.acm.org/ in a few days for updates. And check out http://webinar.acm.org for archived recordings of past webcasts.

Explainable Models for Healthcare AI

Ankur Teredesai

Muhammad Aurangzeb Ahmad

Carly Eckert M.D.

Vikas Kumar

KenSci Inc. + University of Washington Tacoma

https://www.kensci.com/explainable-machine-learning/

ACM Highlights

- Learning Center tools for professional development: http://learning.acm.org
  - The Safari Learning Platform featuring the **entire Safari collection of nearly 50,000** technical books, video courses, O'Reilly conference videos, learning paths, tutorials, case studies
  - 1,800+ Skillsoft courses, 4,800+ online books, and 30,000+ task-based short videos for software professionals covering programming, data management, DevOps, cybersecurity, networking, project management, and more; including training toward top vendor certifications such as AWS, CEH, Cisco, CISSP, CompTIA, Oracle, RedHat, PMI.
  - 1,200+ books from Elsevier on the ScienceDirect platform (including Morgan Kaufmann and Syngress titles)
  - Learning Webinars from thought leaders and top practitioners
  - Podcast interviews with innovators, entrepreneurs, and award winners
- Popular publications:
  - Flagship *Communications of the ACM (CACM)* magazine: http://cacm.acm.org
  - *ACM Queue* magazine for practitioners: http://queue.acm.org
- The **ACM Code of Ethics**, a set of principles and guidelines principles and guidelines designed to help computing professionals make ethically responsible decisions in professional practice: https://ethics.acm.org
- **ACM Digital Library**, the world's most comprehensive database of computing literature: http://dl.acm.org
- International conferences that draw leading experts on a broad spectrum of computing topics: http://www.acm.org/conferences
- Prestigious awards, including the **ACM A.M. Turing Award** and ACM Prize in Computing: http://awards.acm.org
- And much more… http://www.acm.org

"Housekeeping"

Twitter: #ACMLearning

- Welcome to today's ACM Learning Webinar, "**Explainable Models for Healthcare AI**" The presentation starts at the top of the hour and lasts 60 minutes. Audio and video will automatically play throughout the event. On the bottom panel you'll find a number of widgets, including Twitter, Sharing, and Wikipedia apps.

- If you are experiencing any problems/issues, **refresh** your console by pressing the **F5** key on your keyboard in Windows, **Command + R** if on a Mac, or refresh your browser if you' re on a mobile device; or close and re-launch the presentation. You can also view the Webcast Help Guide, by clicking on the "Help" widget in the bottom dock.

- To control volume, adjust the master volume on your computer. If the volume is still too low, use headphones.

- At the end of the presentation, you'll see a survey open on your screen. Please take a minute to fill it out to help us improve your next webinar experience. You may also open the survey at any time throughout the presentation from resources window.

- This session is being recorded and will be archived for on-demand viewing in the next 1-2 days. You will receive an automatic email notification when it is available, and check http://learning.acm.org/ in a few days for updates. And check out http://webinar.acm.org for archived recordings of past webcasts.

Talk Back

- Use Twitter widget to Tweet your favorite quotes from today's presentation with hashtag #ACMLearning
- Submit questions and comments via Twitter to @acmeducation – we're reading them!
- Use the sharing widget in the bottom panel to share this presentation with friends and colleagues.
- The ACM Discourse Page is available for post-webinar discussion – https://on.acm.org

Explainable Models for Healthcare AI

Ankur Teredesai

Muhammad Aurangzeb Ahmad

Carly Eckert M.D.

Vikas Kumar

KenSci Inc. + University of Washington Tacoma

https://www.kensci.com/explainable-machine-learning/

KenSci

#DeathVsDataScience

Applied AI for Global Health

Conceived in academia. Supported by the federal Health agencies. Trusted by leading health systems globally.

57 Million Lives

37 Published Papers

Microsoft Partner of the Year 2018 Finalist

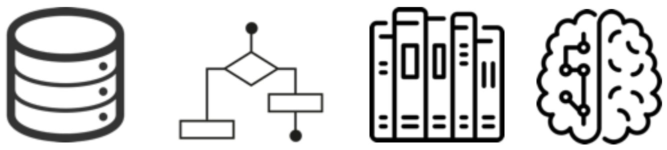
Microsoft Health Innovation Awards 2018 WINNER

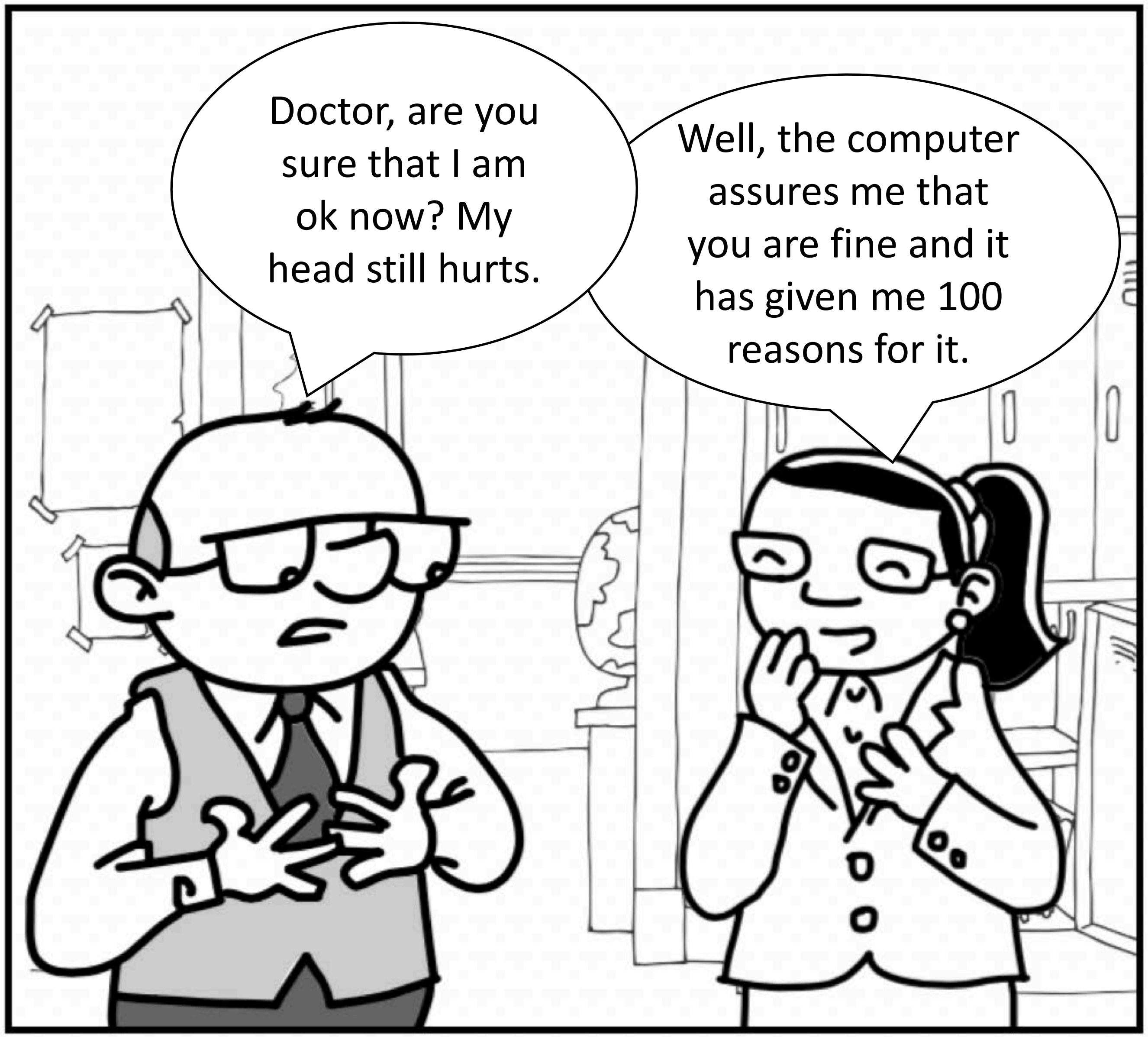
KDD

IEEE

Built by Clinicians, Data Scientists, and Engineers

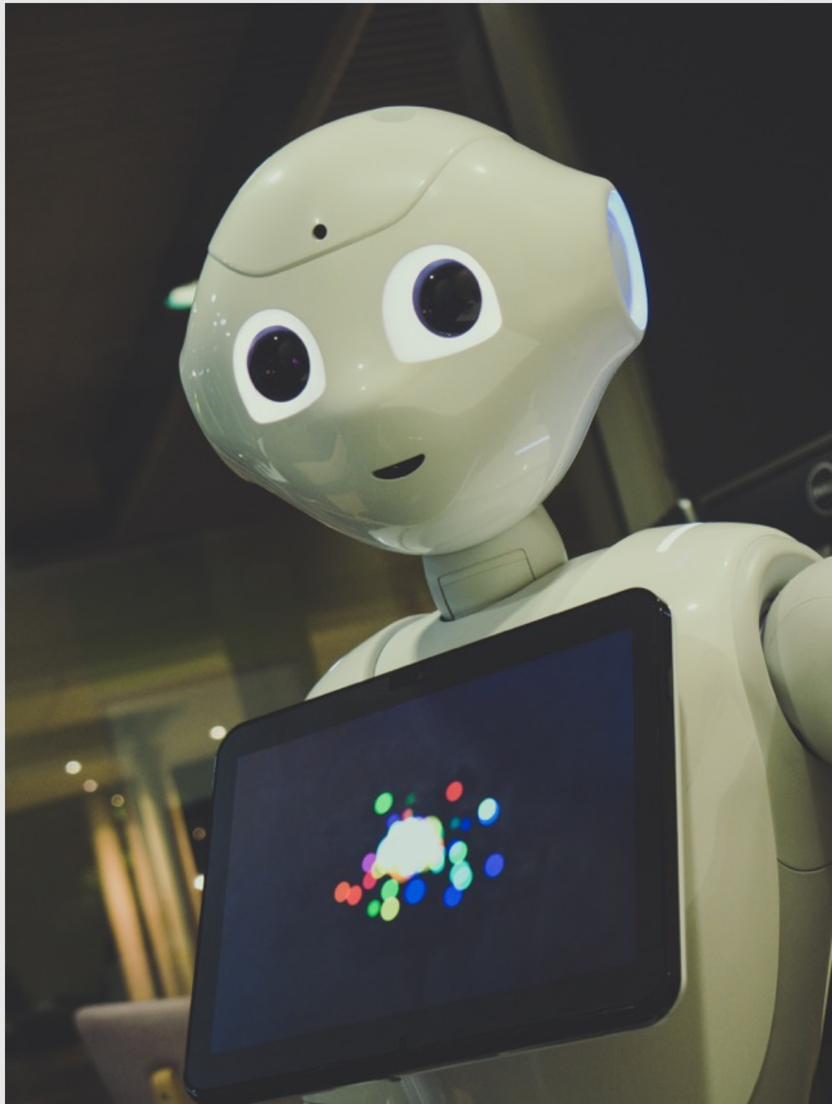
Explainable Models in AI




Doctor, are you sure that I am ok now? My head still hurts.
Well, the computer assures me that you are fine and it has given me 100 reasons for it.

Learning Objectives of the Tutorial

- Why ?
  - The AI & ML needs to provide explanations in Healthcare.
- What ?
  - Type of Explanations do we need in Healthcare
- How ?
  - To select the right machine learning algorithms when explanations are needed
- Where ?
  - Settings and different types of interpretable machine learning models
- When?
  - The past the present and the future of explainable AI in healthcare

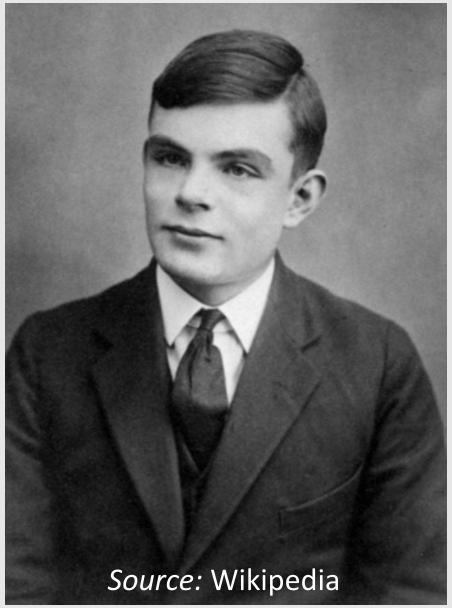


Explainable By Any other Name.. is still explainable

- **Explain:** Make (an idea or situation) clear to someone by describing it in more detail or revealing relevant facts.
- **Interpret:** Explain the meaning of (information or actions)
- **Understand:** Perceive the intended meaning of (words, a language, or a speaker)
- **Comprehend:** Grasp mentally; understand
- **Intelligible:** Able to be understood; comprehensible
  [Dictionary, Oxford English 2018]

A (Brief) History of *Need for Explainations* in Statistics

- Foundations of Logic: Clear and explicit reasoning, explainable to almost anyone
- Bayes: Grounding Probability and inference on solid foundations (18^{th} century) [Bayes 1763]
- Goal of (early) AI: Mimic human reasoning mechanically
- Early Expert Systems in healthcare like MYCIN were explainable systems
- Push around explainability when ensemble methods first came about in the 1980s
- Current interest in explainability is because of widespread adoption and deployment of machine learning systems



Source: Wikipedia

Explainable AI / Interpretable Machine Learning

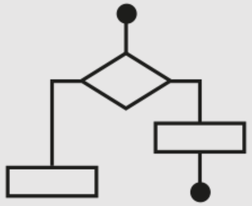
- **Explanations** of AI/machine learning models to **humans** with **domain knowledge**
  [Craik 1967, Doshi-Velez 2014]
- Why is the prediction being made?
- Comprehensible to humans in (i) natural language (ii) easy to understand representations

Standard Model Lagrangian

Explainable AI is More Than Models

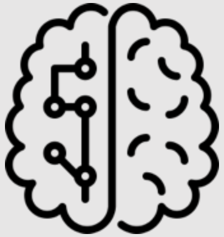
AI Solution

Features



Algorithm

Model Parameters



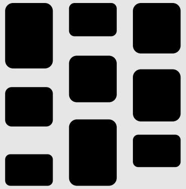
Model

Each element constituent of the solution process needs to be explainable for the solution to be truly explainable [Lipton 2016]

User

Cognitive Capacity

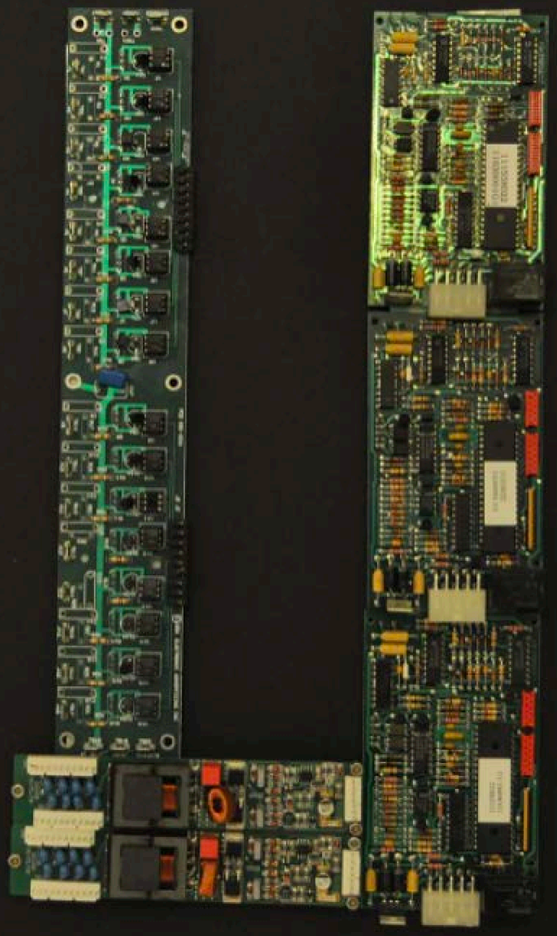
Domain Knowledge



Explanation Granularity

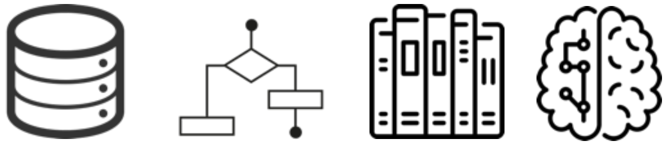
What Explanations Are Not?

- **Explanation vs. Justification**
  - Explaining is giving reasons for the prediction
    [Biran 2014, Biran 2017A, Biran 2017B]
  - Justification is putting the explanation in a context
  - Justification does not have to correspond to how the model actually works
- **Explanation vs. Causality**
  - Explanations are mostly:
    - *Not* Causal
    - *Not* Prescriptive
  - Example: End of life prediction for heart failure patients



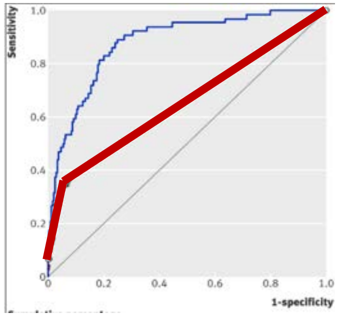
Source: www.aurumahmad.com

Healthcare AI



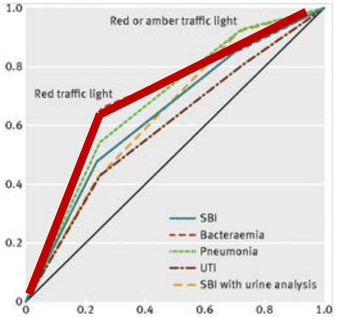
How Are Decisions Made in Healthcare?

Factors: 7 ± 2



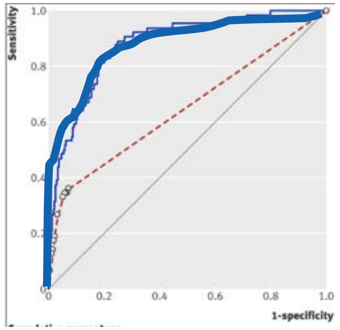
~80% Care Decisions

Factors: 10s



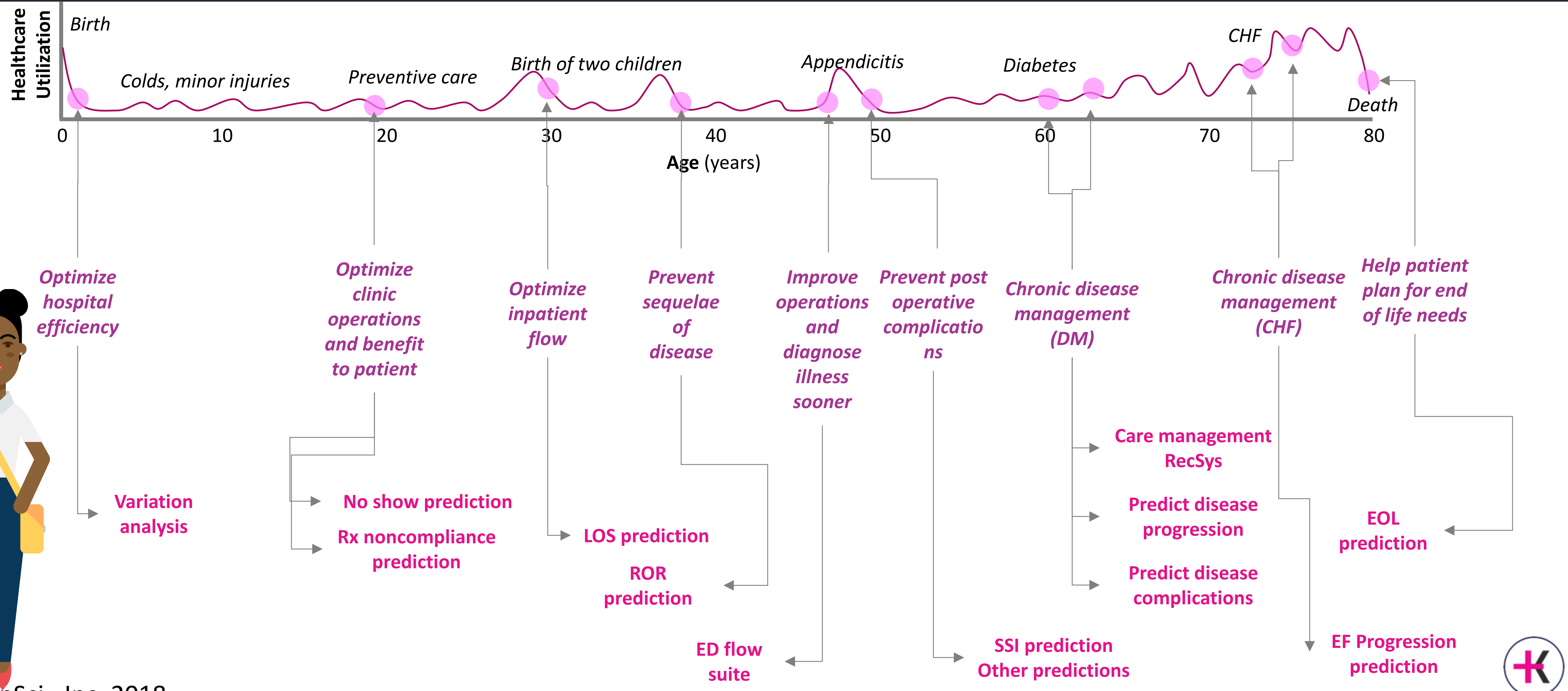
~18% Care Decisions

Factors: 100s

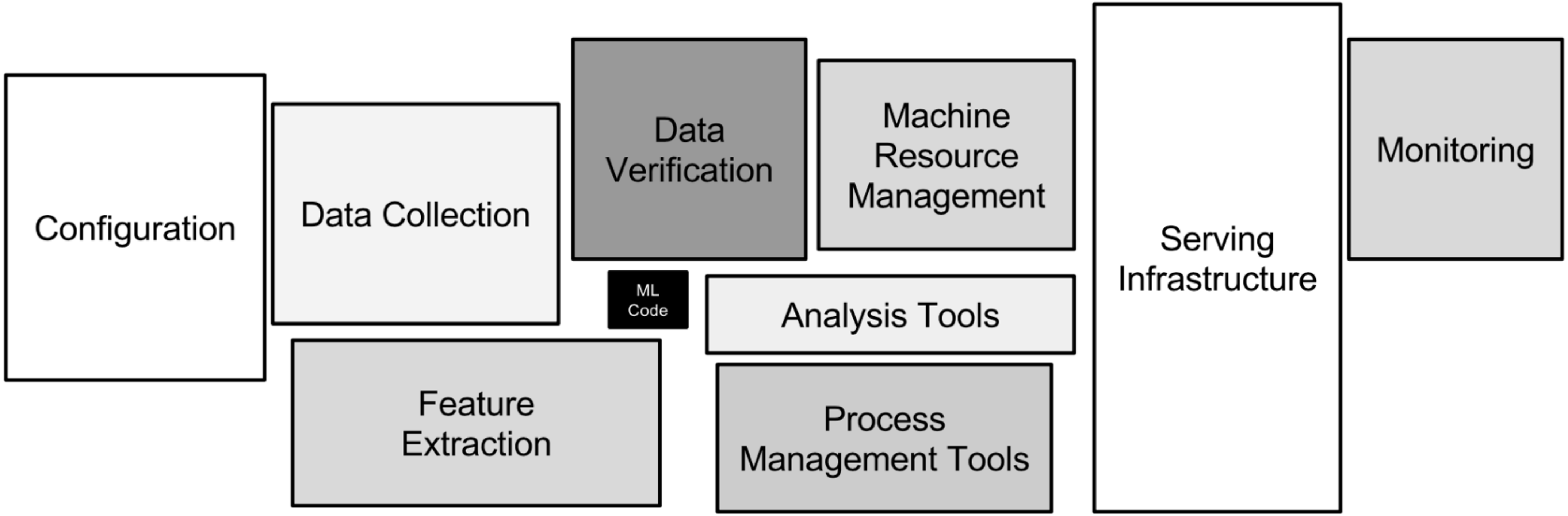


~2% Care Decisions

How Machine Learning Can Improve Healthcare Across the Continuum of Care?



Operationalizing AI in Healthcare

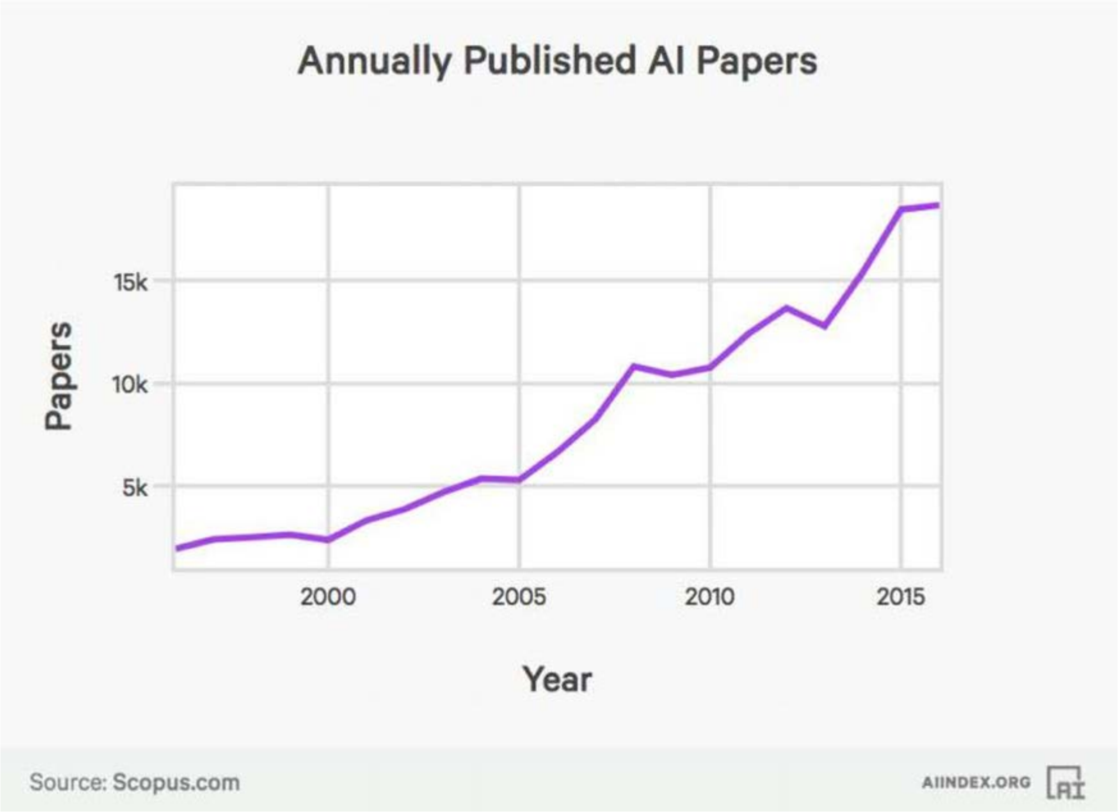


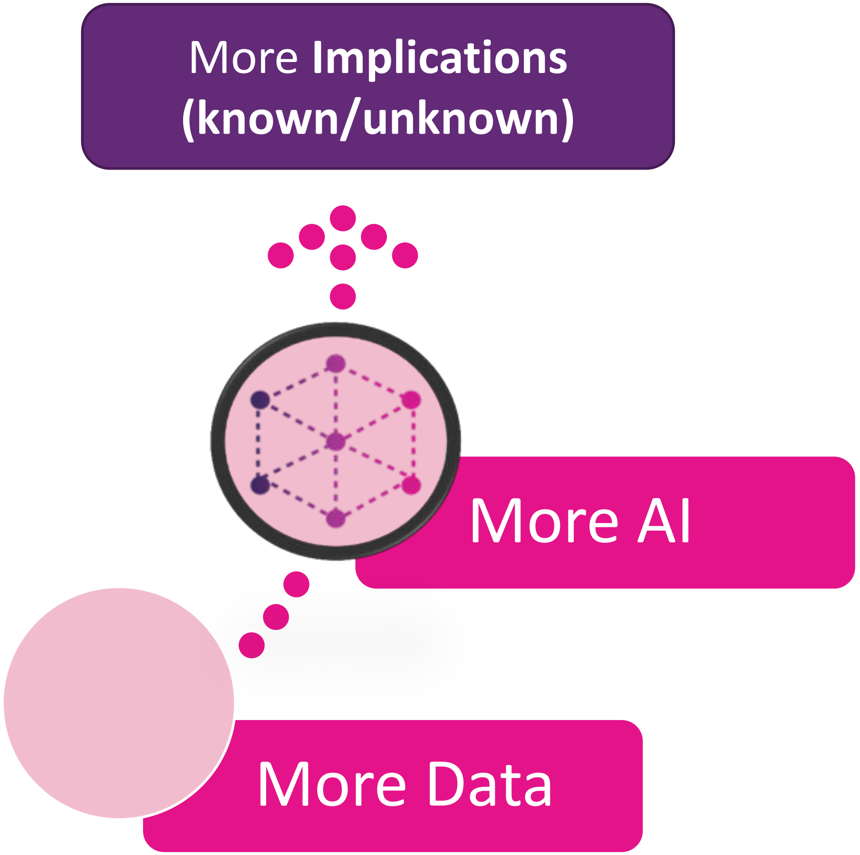
Only a small fraction of real-world machine learning systems actually constitutes machine learning code [Sculley 2015].

Data Complexity in Operationalizing AI in Healthcare

- Syntactic Correctness
  - Is the data in the correct format e.g., if the AI data pipeline requires sex for males as ‘m’ and the input data encodes it as ‘1’
- Morphological Correctness
  - Is the data within the range of possible values e.g., a blood pressure of 500 does not make sense
- Semantic Correctness
  - Do the variables actually correspond to what the semantics that are being ascribed to them e.g., a variable which encodes blood pressure as high vs. low will have different semantics for children as compared to adults

Why Do We Need Explanations in Healthcare AI Now?





Do We ALWAYS Need Explanations in Healthcare AI?

When fairness is critical:

- Any context where humans are required to provide explanations so that people cannot hide behind machine learning models [Al-Shedivat 2017B, Doshi-Velez 2014]

When consequences are far-reaching:

- Predictions can have far reaching consequences e.g., recommend an operation, recommend sending a patient to hospice etc.

When the cost of a mistake is high:

- Ex: misclassification of a malignant tumor can be costly and dangerous

When a new/unknown hypothesis is drawn:

- *"It's not a human move. I've never seen a human play this move."* [Fan Hui]
- Pneumonia patients with asthma had lower risk of dying [Caruana 2015]

Compliance is key:

- GDPR
- Right to Explanation

Predictive performance is not enough

[Doshi-Velez 2017]

Why Do We **Need** Explanations Now?

Risk Prediction with Blackbox Models

Patient ID		*Has Asthma*		*Risk of Death*
84	...	Yes	...	5%
85	...	Yes	...	6%
86	...	No	...	12%
87	...	No	...	15%
...	...	...	...	...

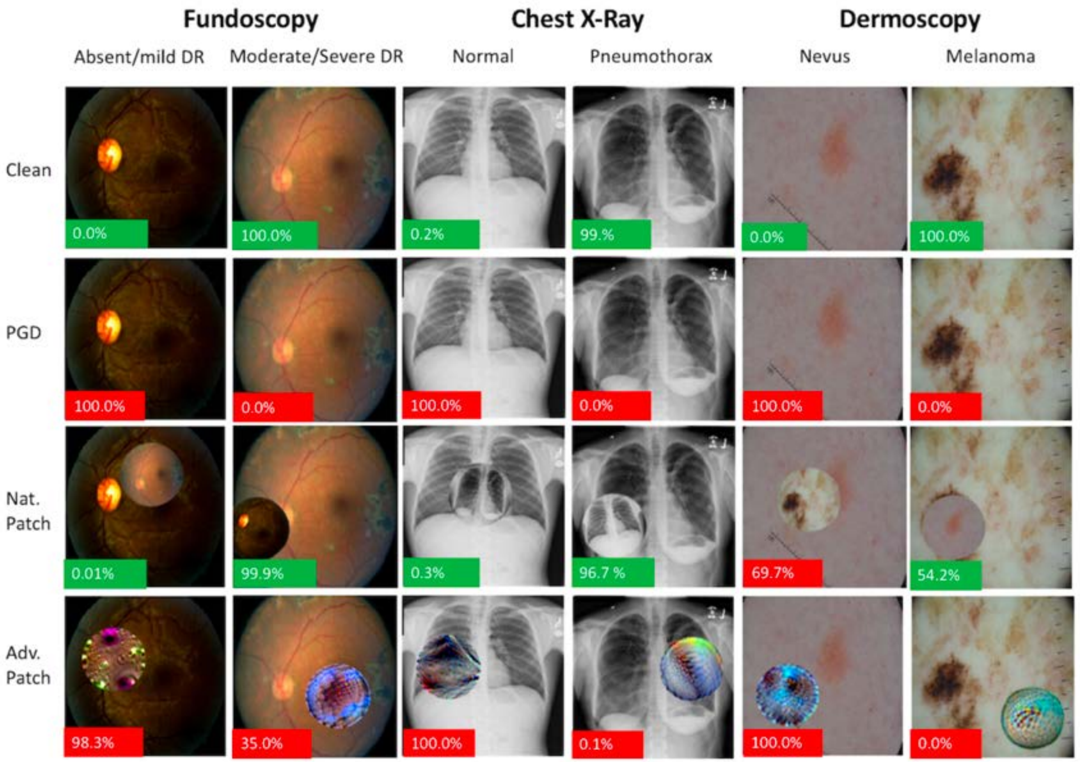
Feature Importance (Higher risk of death): Low High

Feature Importance (Lower risk of death): Low High

With Context: Patients with asthma have a lower risk of death from pneumonia *because they receive more intensive care.*

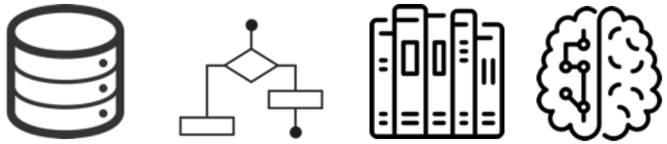
[Caruana 2015, Caruana 2017]

Misdiagnosis via Adversarial Attacks

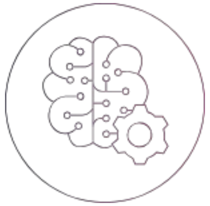


[Finlayson 2018]

Pillars of Explainable AI



7 Pillars of Explainable AI in Healthcare



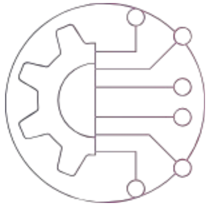
Transparency

Domain Sense

Consistency

Parsimony

Generalizability



Trust/
Performance

Fidelity

Pillar 1: Transparency | Admission Prediction

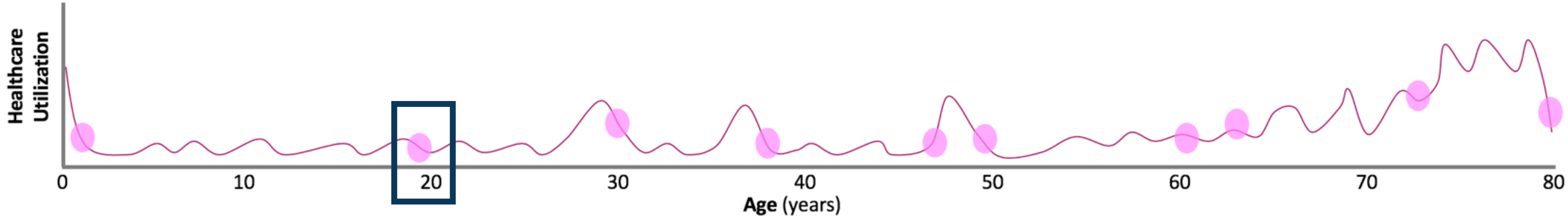
Transparency

Ability of the machine learning algorithm, model, and the features to be understandable by the user of the system.

Admission Prediction

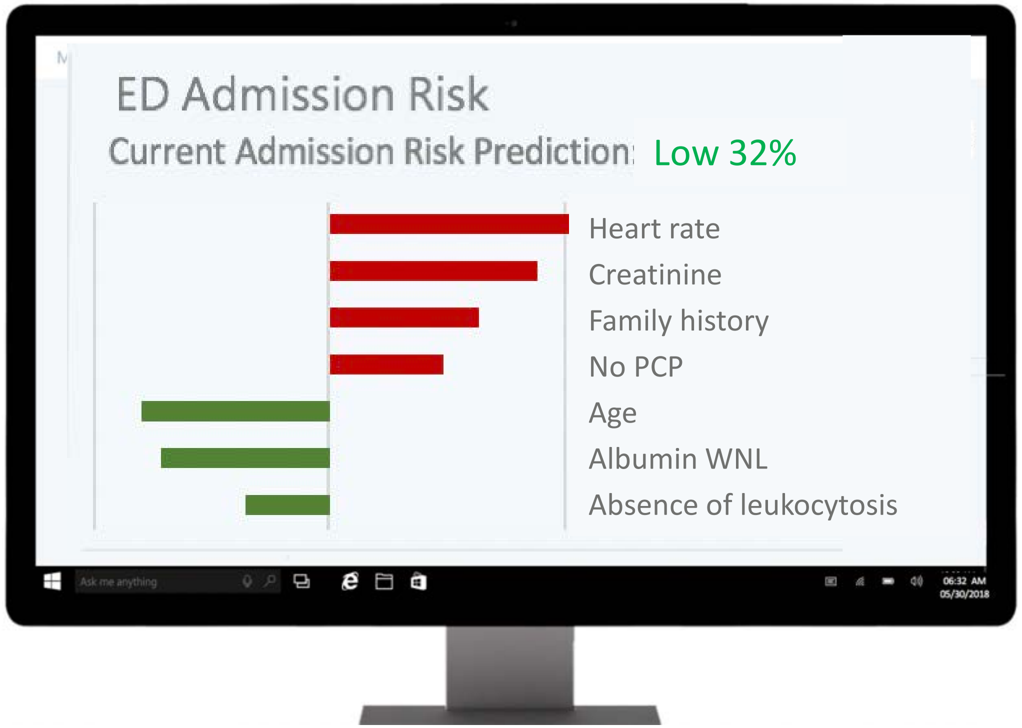
What is the likelihood of the patient being admitted to the hospital from the emergency department.

Katherine presents to the emergency department with severe headaches. She has multiple episodes of vomiting. She is evaluated by the clinical staff and has imaging and laboratory work done. She has very little medical history and considers herself active and healthy.



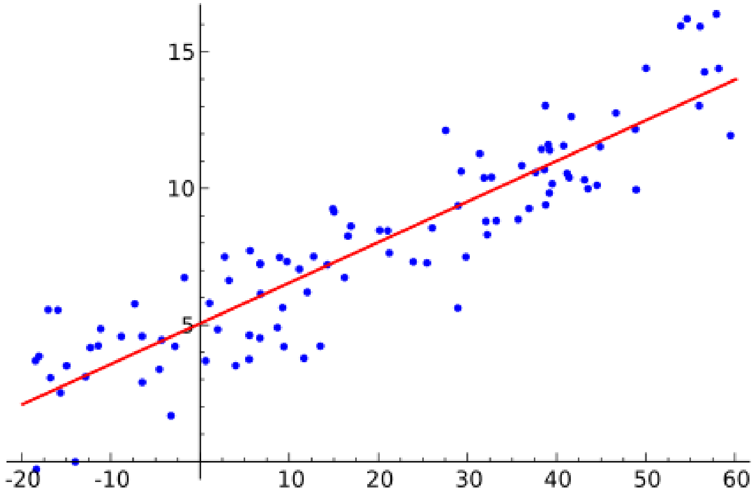
Transparency

- The ML model for predicting Katherine's likelihood of admission gives her a low likelihood (0.32)
- Katherine's physician has noted her age, health history and vital signs and is reassured by her relatively low risk score
- The physician knows that the risk model is a deep learning model so he cannot understand how it is working
- But, he can examines the top factors associated with prediction



Transparent to Whom?

- Transparency may mean different different things to different people
- Understanding Model Outputs:



- Understanding Algorithms:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, ..., n$$

- Understanding the algorithm may not mean be sufficient:

$$a_j^l = \sigma\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right)$$

Transparency | Examples

Transparent

- Falling Rule Lists
- GAM (Generalized Additive Models)
- GA2M (Generalized Additive Models with interactions)
- LIME (Locally Interpretable Model Agnostic Explanations)
- Naïve Bayes
- Regression Models
- Shapley Values

Semi-Transparent

- Shallow Ensembles

Non-Transparent

- Deep Learning
- SVM (Support Vector Machines)
- Gradient Boosting Models

Pillar 2: Domain Sense | ED Census Prediction

Domain Sense

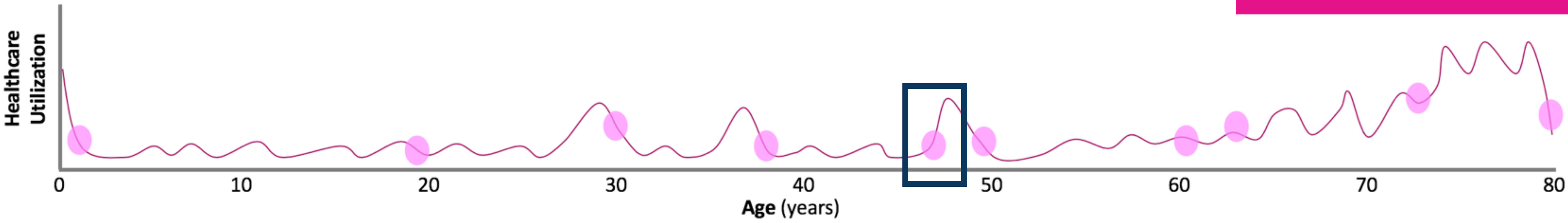
The explanation should make sense in the domain of application and to the user of the system

ED Census Prediction

Predict the number of patients in the emergency department (ED) at a given time

Several years later, Katherine revisits the emergency department due to abdominal pain. She has an elevated temperature and is dehydrated.

She is at the ED on a Friday after work. The ED is very crowded and she must wait several hours to be seen.



Domain Sense | Model Output & Actionability

Making Sense of Model Output

- Interpreting output from machine learning models may also has an element of subjectivity
- *Example:* If a patient has a readmission risk score 0.62, what does that mean?

Actionability

- Actionability does not presume causality but can be used to devise interventions
- Actionability is temporal e.g., flu vaccine administration is a 1 minute task vs. weight loss programs which can take months

Mutability	Interveniability	Actionability	Example
Immutable			Age, Sex, Ethnicity
Mutable	Non-Interveniable		Intrinsic Heart rate variability Marital Status
	Interveniable	Signal	Temperature (in Apendectomy)
		Intervention	Appendicitis
	Post-Interveniable		Immunization

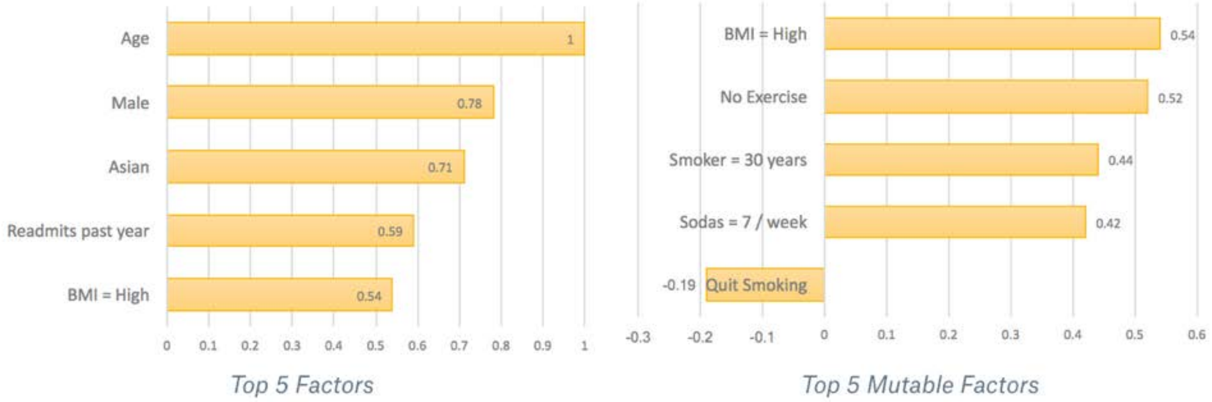


Figure 1: Factors for Predicting Risk of Readmission

Pillar 3: Consistency | LWBS

Consistency

The explanation should be consistent across different models and across different runs of the model

LWBS

Left without being seen refers to a patient leaving the facility without being seen by a physician

~18:00

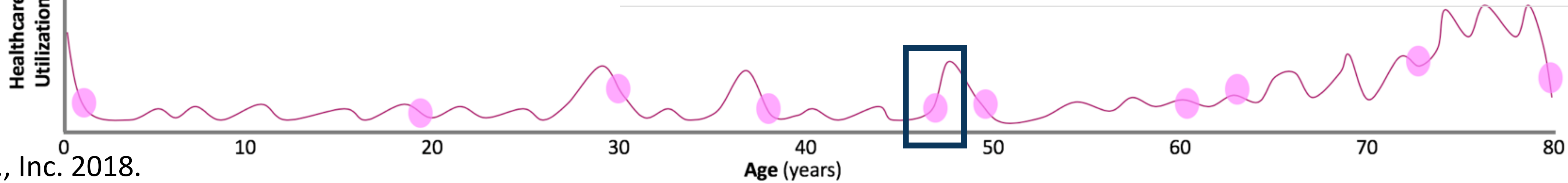
Patients at Risk of Leaving Without Being Seen

Family Name	Given Name	Age	Chief Complaint	PCP Name	Insurance Type	Checked In Time	Elapsed Waiting Time	Past LWBS Occurences	Predicted LWBS Score	Factors
Franklin	Samuel	56	WOUND INFECTION	Overman	Private	18:30	90	2	62	Prior LWBS
Pierce	Jerome	41	ABDOMINAL PAIN	Mark	Uninsured	17:45	135	1	40	Insurance status
Collins	Katherine	37	ABDOMINAL PAIN	Monroe	Private	16:52	68	0	24	Past history

~22:00

Patients at Risk of Leaving Without Being Seen

Family Name	Given Name	Age	Chief Complaint	PCP Name	Insurance Type	Checked In Time	Elapsed Waiting Time	Past LWBS Occurences	Predicted LWBS Score	Factors
Franklin	Samuel	56	WOUND INFECTION	Overman	Private	18:30	90	2	62	Temperature
Pierce	Jerome	41	ABDOMINAL PAIN	Mark	Uninsured	17:45	135	1	40	Insurance status
Collins	Katherine	37	ABDOMINAL PAIN	Monroe	Private	16:52	314	0	24	Chief complaint
Kennedy	Susan	72	DYSPNEA	Lewis	Medicare	19:45	135	1	60	Age



Consistency | Model Multiplicity

- Given the same dataset, multiple machine learning algorithms can be constructed with similar performance
- The explanations from multiple explainable algorithms should be very similar
- Divergent Explanations symptomatic of problem with explanations and/or algorithm(s)
- Evaluation:
    - Human Evaluation: Expert Agreement
    - Machine Evaluation: Ranked List Comparison

Variable	Model A	Model B	Model C
Age	1	1	5
Gender	2	4	6
Diabetic	3	5	1
Race	4	6	4
Smoker	5	2	3
Alcoholic	6	3	2

Variables for Length of Stay Prediction Ranked

$$W = \frac{12S}{m^2(n^3 - n)}$$

Kendall's W: [Kendall 1939]

Pillar 4: Parsimony| Admission Disposition

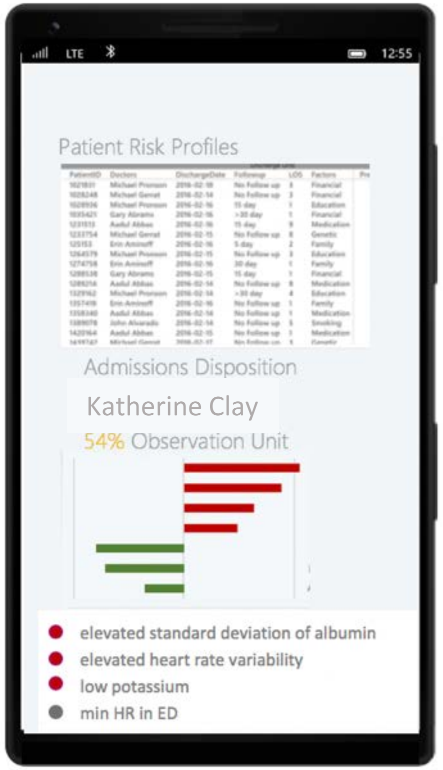
Parsimony

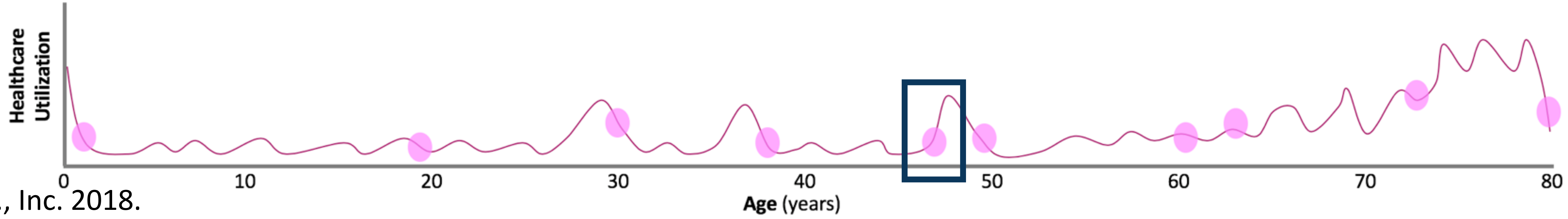
The explanation should be as simple as possible

Applies to both the complexity of the explanation and the number of features provided to explain

Admission Disposition

Where in the hospital the patient should go once they are admitted





Parsimony

- MDL (Minimum Description Length) and Occam's Razor
- **Occam's Razor:** To derive a unifying diagnosis that can explain all of a patient's symptoms
- **Hickam's Dictum:** A man can have as many diseases as he damn well pleases
- Occam's Razor in Machine Learning [Domingos 1999]
  - Occam's First Razor
  - Occam's Second Razor
- The simplest explanation is not always the best explanation

Pillar 5: Generalizability | Length of Stay Prediction

Generalizability

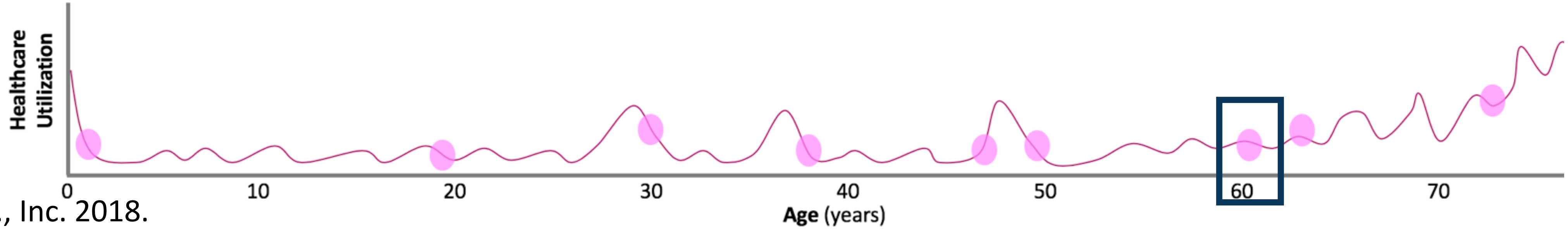
Models and explanations should be generalizable across problem whenever possible

Katherine eventually develops diabetes. It is well controlled and she takes her medications as directed. One afternoon, she is admitted from clinic due to highly elevated glucose levels and a urinary tract infection. Her nurse tells her that based on her illness and other factors, her predicted length of stay is 3 days.

Length of Stay

The time that a patient will spend at a particular healthcare facility

Room Number	Patient ID	Last Name	First Name	Age	Gender	Attending Provider	Chief Complaint	Admission Date	Predicted Discharge Date	Elapsed LOS	Predicted LOS	Explanation
9B	3408738	Mascroft	Pete	53	Male	Thomas Louwers	Dyspnea	5/22/2018	5/28/2018	6	6	Diabetes
15A	20350155	Hookano	Karma	24	Female	Jonathan Miller	Anemia	5/28/2018	5/31/2018	0	3	Diabetes
7C	26059755	Grant	Tien	71	Male	Thomas Louwers	SOB	5/24/2018	5/28/2018	4	4	Diabetes
10A	4664428	Berceir	Carmelia	32	Female	Venessa Overman	Fatigue	5/26/2018	5/29/2018	2	3	Diabetes
14A	3985950	Brea	Tameika	46	Female	Susan Mark	Sepsis	5/27/2018	5/30/2018	1	3	Diabetes
3B	8354235	Lacio	Lien	53	Male	Susan Mark	Abd pain	5/22/2018	6/1/2018	6	10	Diabetes
13A	4205535	Sayegh	Vina	55	Female	Jonathan Miller	Fatigue	5/26/2018	5/31/2018	2	5	Diabetes
11C	42399976	Zeidan	Oma	68	Female	Susan Mark	CKD	5/28/2018	5/30/2018	0	2	Diabetes
12C	59204677	Rile	Joanne	61	Female	Venessa Overman	UTI	5/25/2018	5/29/2018	3	4	Diabetes



Pillar 5: Generalizability | Length of Stay Prediction

Model Generalizability

- *Local Models:* Models that give explanations at the level of an instance e.g., LIME, Shapley Values etc.
- *Cohort Level Models:* A type of global models where the explanations are generated at the level of cohort
- *Global Models:* Models that give explanations e.g., Decision Trees, Rule Based Models etc.

Algorithm Generalizability

- *Model Agnostic Explanations:*
  - Examples: LIME, Shapley Values etc.
- *Model Class Specific Explanations:*
  - Examples: Tree Explainers, Gradient Boost Explainers
- *Model Specific Explanations:*
  - Examples: CENs, Decision Trees, Random Forest Explainer etc.

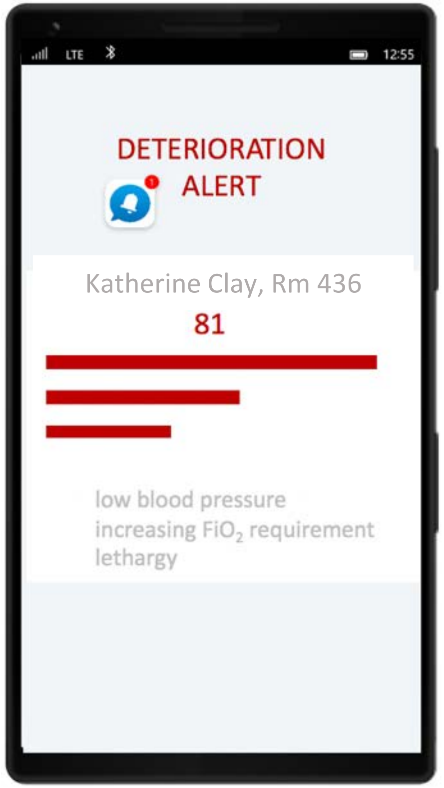
Pillar 6: Trust/Performance | ICU Transfer Prediction

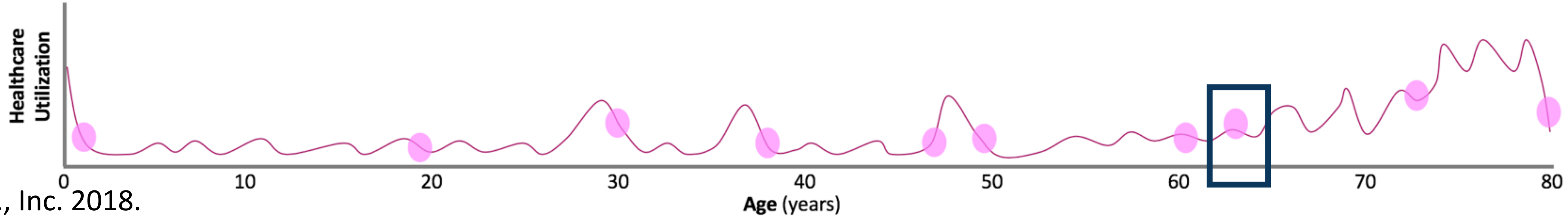
Trust / Performance

- The expectation that the corresponding predictive algorithm for explanations should have a certain performance

ICU Transfer Prediction

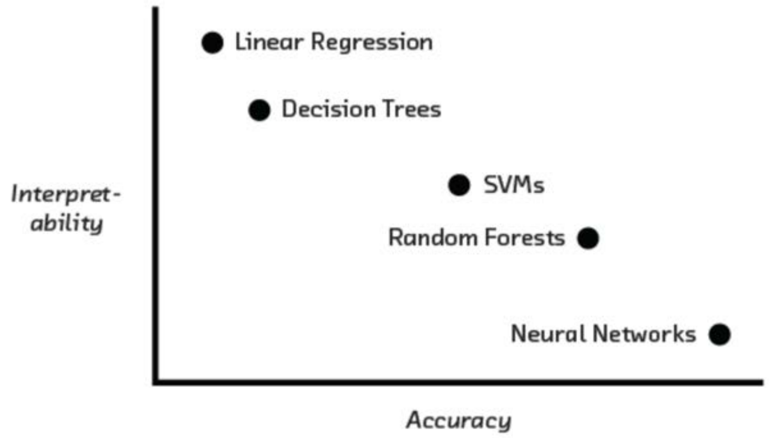
- Predict if a patient on the hospital ward will require transfer to the intensive care unit due to increasing acuity of care needs

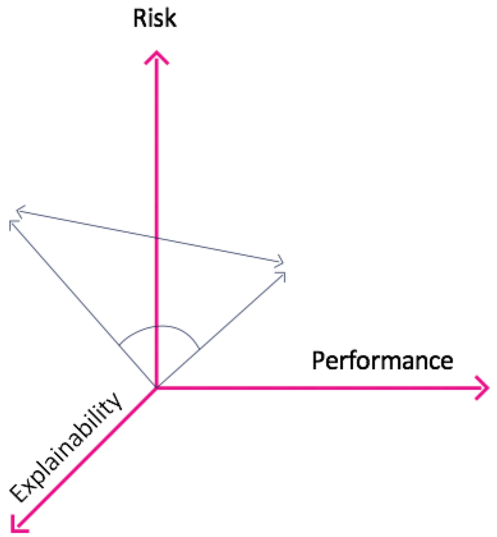




Trust | Prediction Performance

- Expectation that the predictive system has a sufficiently high performance e.g., precision, recall, AUC etc. [Lipton 2016, Hill 2018]
- Explanations accompanied with sub-par predictions can foster distrust
- The model should perform sufficiently well on the prediction task in its intended use
- The model has at least parity with the performance of human practitioners
- Trauma patients: vital signs and lab criteria fulfill criteria to trigger alarm, leads to increasing numbers of false positives [Nguyen 2014]





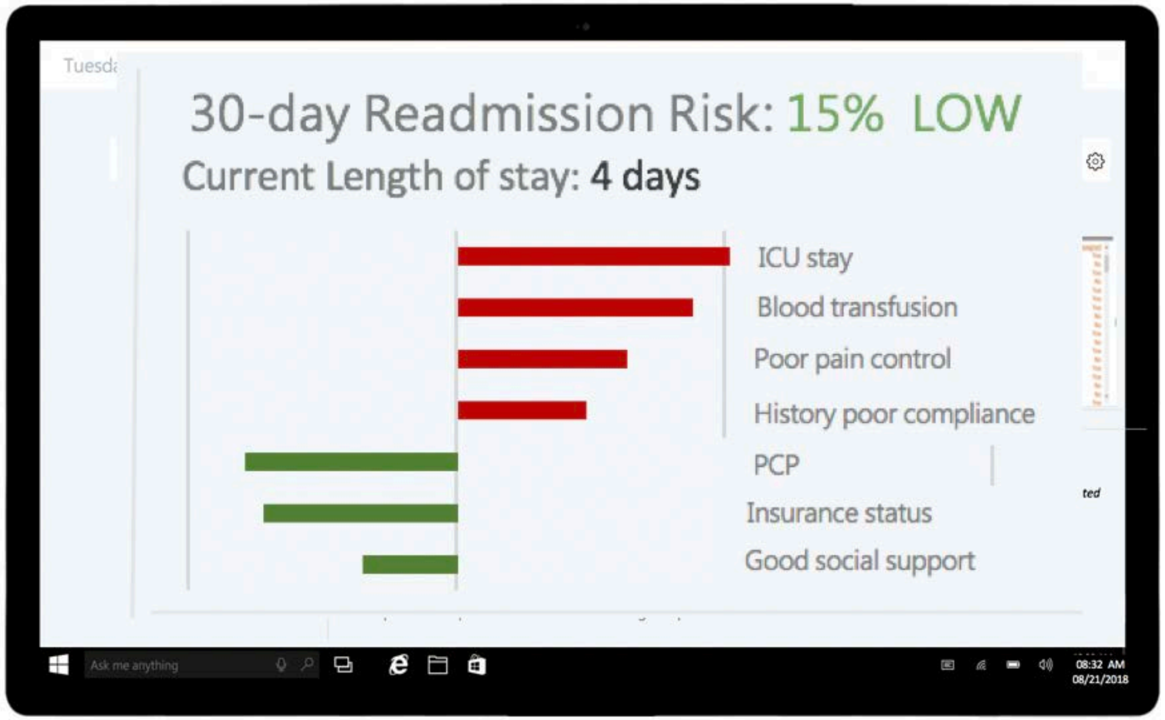
Pillar 7: Fidelity| Risk of Readmission

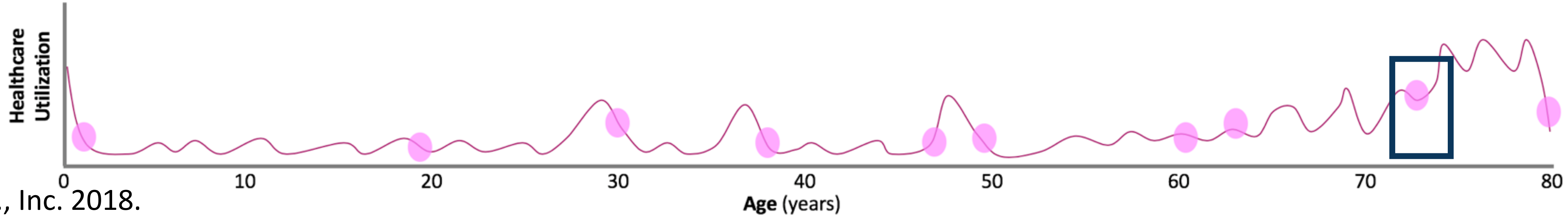
Fidelity

- The expectation that the explanation and the predictive model align well with one another

Risk of Readmission

- Predict if the patient will be readmitted within a particular span in time, i.e. 30 days from time of discharge





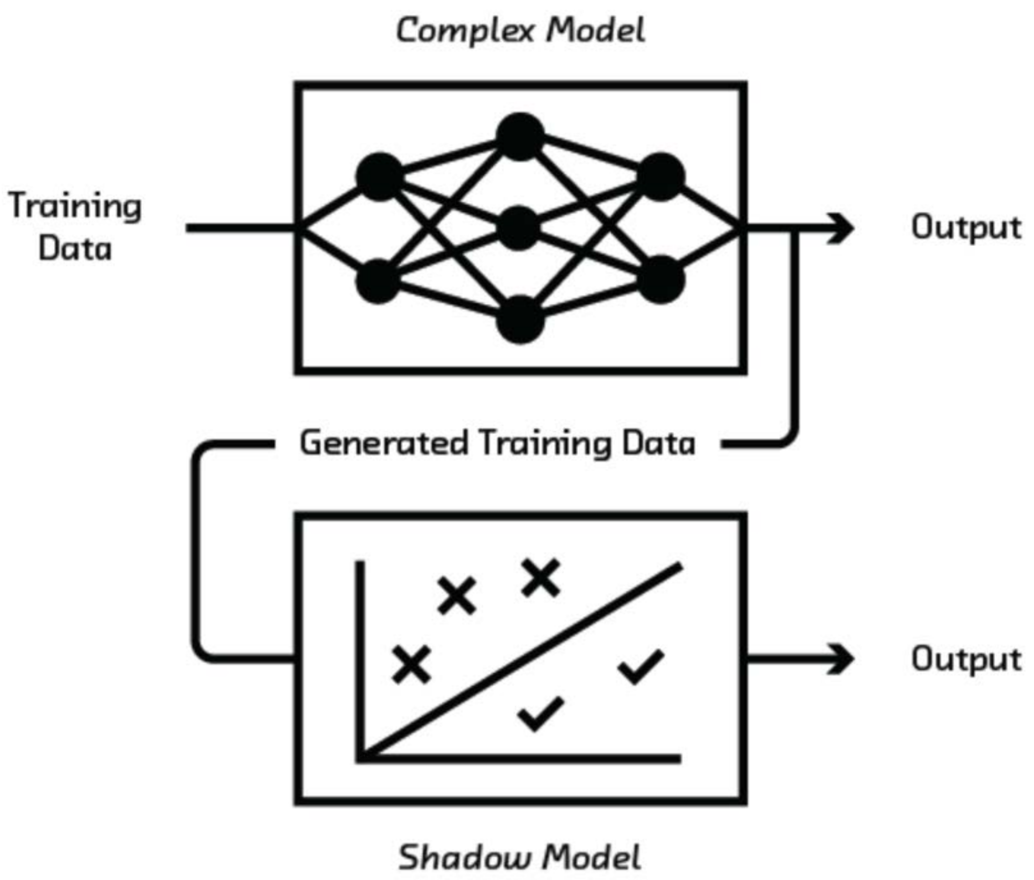
Fidelity | Explanations

- An explanation is **Sound** if it adheres to how the model actually works
- An Explanation is **Complete** if it encompasses the complete extent of the model
- **Ante-Hoc:** Models where the predictive model and the explanation model is the same
- **Post-Hoc Models:** Models where the predictive model and the explanation model are different
- Special Case: **Mimic Models**

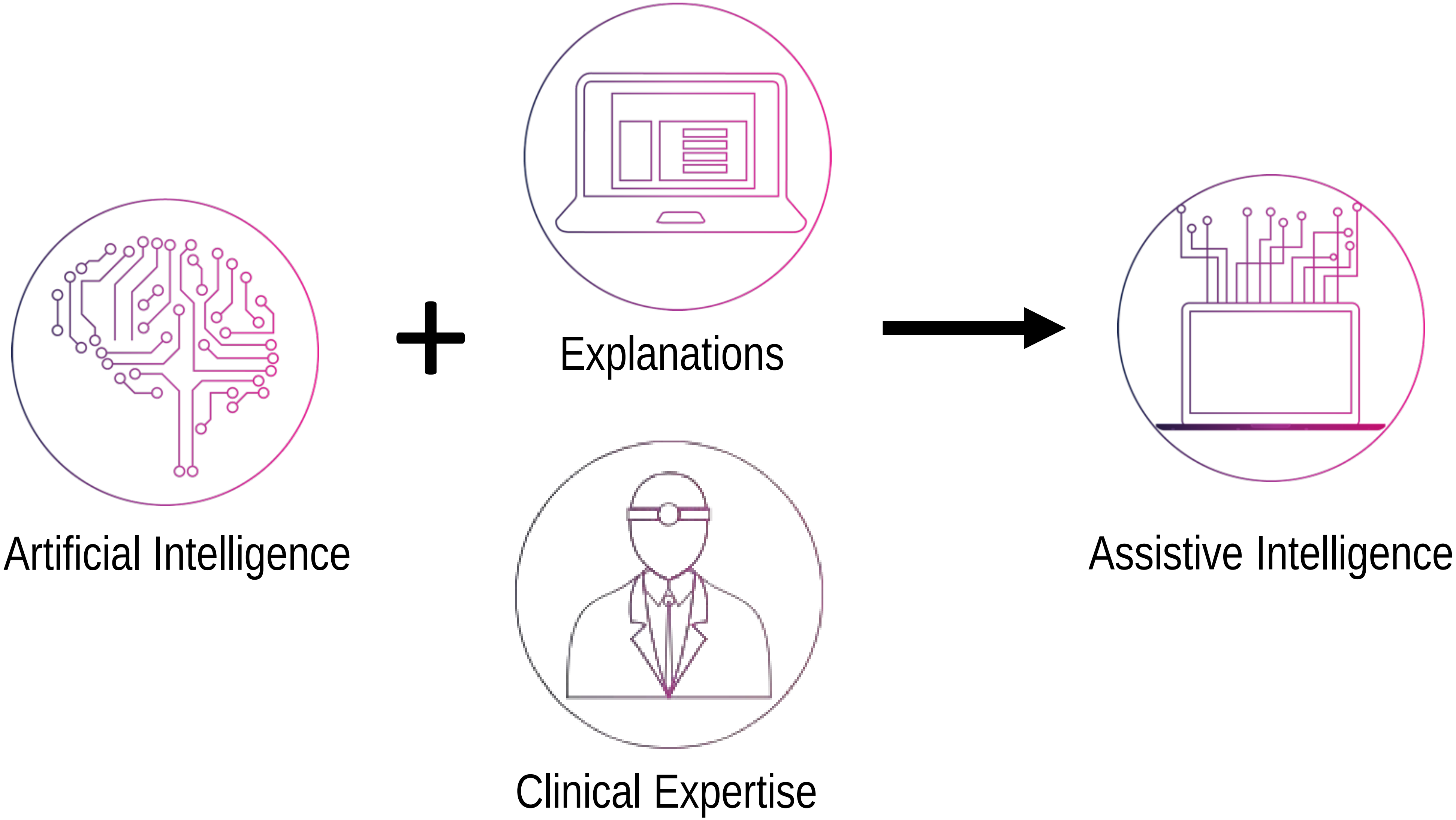
Example: Readmission model which uses lunar cycles as a feature

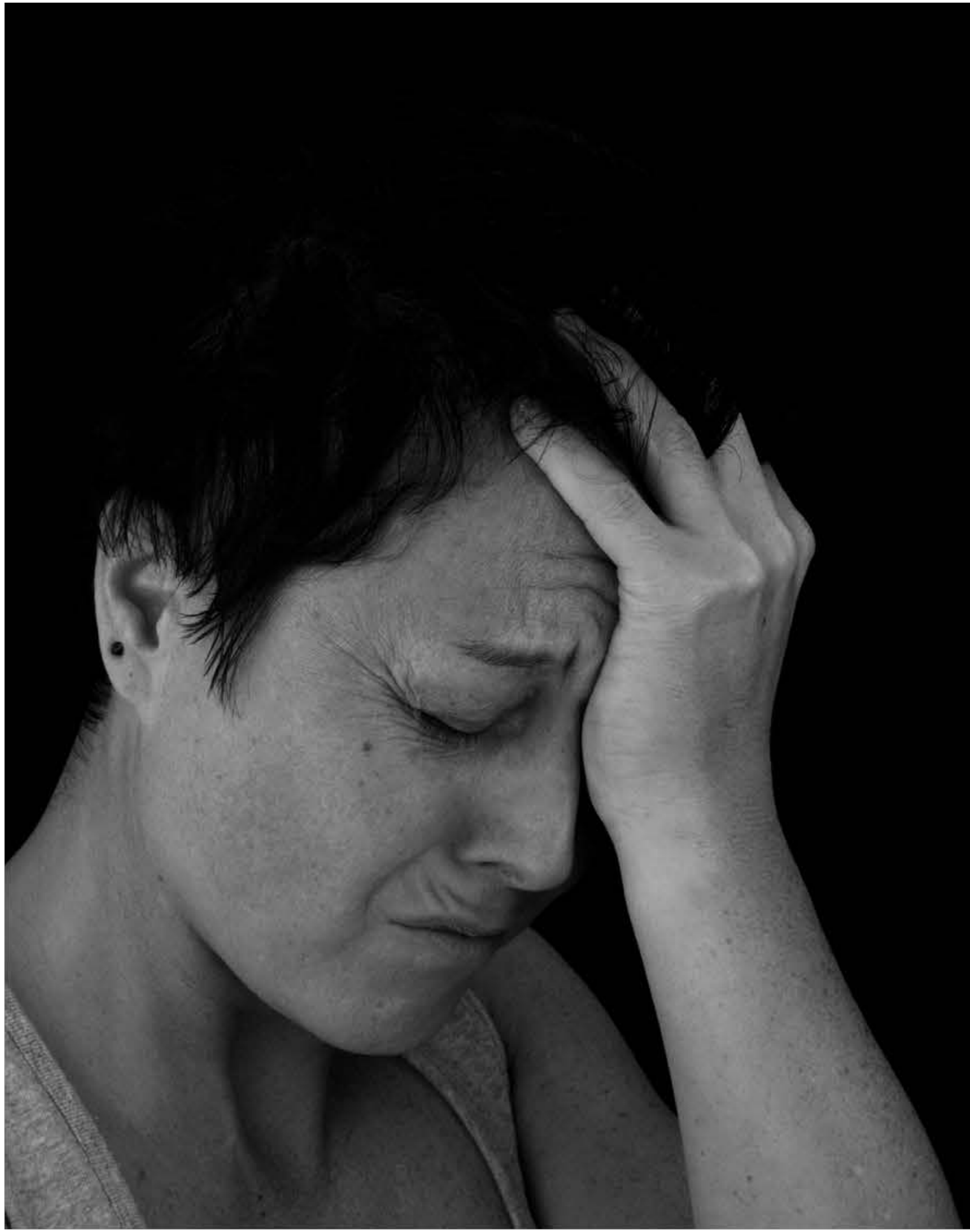
Fidelity | Mimic Models

- Also known as Shadow, Surrogate or Student Models
- Use the output (instead of the true labels) from the complex model and the training data to train an model which is explainable [Bucilua 2006, Tan 2017]
- The performance of the student model is usually quite good
- Example: Given a highly accurate SVM, train a decision tree on the predicted label of the SVM and the original data



Conclusion: Explainable ML in Healthcare AI





HEALTHCARE NEEDS HELP.

AND HOPE.

DEATH VS. DATA SCIENCE

HELP US IN THIS MISSION

@KenSci

Ankur@KenSci.com

References

Ahmad 2018	Muhammad Aurangzeb Ahmad, Carly Eckert, Greg McKelvey, Kiyana Zolfagar, Anam Zahid, Ankur Teredesai. Death vs. Data Science: Predicting End of Life IAAI February 2-6, 2018
Al-Shedivat 2017A	Al-Shedivat, Maruan, Avinava Dubey, and Eric P. Xing. "Contextual Explanation Networks." arXiv preprint arXiv:1705.10301 (2017).
Al-Shedivat 2017B	Al-Shedivat, Maruan, Avinava Dubey, and Eric P. Xing. "The Intriguing Properties of Model Explanations."
Bach 2015	Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." PloS one 10.7 (2015): e0130140
Bayes 1763	Bayes, Thomas, Richard Price, and John Canton. "An essay towards solving a problem in the doctrine of chances." (1763): 370-418.
Biran 2014	Biran, Or, and Kathleen McKeown. "Justification narratives for individual classifications." In Proceedings of the AutoML workshop at ICML, vol. 2014. 2014.
Biran 2017A	Biran, Or, and Kathleen R. McKeown. "Human-Centric Justification of Machine Learning Predictions." In IJCAI, pp. 1461-1467. 2017.
Biran 2017B	Biran, Or, and Courtenay Cotton. "Explanation and justification in machine learning: A survey." In IJCAI-17 Workshop on Explainable AI (XAI), p. 8. 2017.
Bucilua 2006	Buciluǎ, Cristian, Rich Caruana, and Alexandru Niculescu-Mizil. "Model compression." In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 535-541. ACM, 2006.
Caruna 2015	Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1721-1730. ACM, 2015.
Caruna 2017	Caruana, Rich, Sarah Tan, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, Noemie Elhadad et al. "Interactive Machine Learning via Transparent Modeling: Putting Human Experts in the Driver's Seat." IDEA 2017
Chang 2009	Chang, Jonathan, et al. "Reading tea leaves: How humans interpret topic models." Advances in Neural Information Processing Systems. 2009
Craik 1967	Craik, Kenneth James Williams. The nature of explanation. Vol. 445. CUP Archive, 1967.
Craven 1996	Mark W. Craven and Jude W. Shavlik. Extracting tree-structured representations of trained networks. Advances in Neural Information Processing Systems, 1996.
Datta 2016	Datta, Anupam, Shayak Sen, and Yair Zick. "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems." Security and Privacy (SP), 2016 IEEE Symposium on. IEEE, 2016
Doshi-Velez 2014	Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).
Doshi-Velez 2017	Doshi-Velez, Finale, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. "Accountability of AI under the law: The role of explanation." arXiv preprint arXiv:1711.01134 (2017).
Druzdzel 1996	Druzdzel, Marek J. "Qualitiative verbal explanations in bayesian belief networks." AISB QUARTERLY (1996): 43-54.
Farah 2014	Farah, M.J. Brain images, babies, and bathwater: Critiquing critiques of functional neuroimaging. Interpreting Neuroimages: An Introduction to the Technology and Its Limits 45, S19-S30 (2014)
Freitas 2014	Freitas, Alex A. "Comprehensible classification models: a position paper." ACM SIGKDD explorations newsletter 15, no. 1 (2014): 1-10.

References

Friedman 2001	Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. New York: Springer series in statistics, 2001.
Friedman 2008	Friedman, Jerome H., and Bogdan E. Popescu. "Predictive learning via rule ensembles." The Annals of Applied Statistics2, no. 3 (2008): 916-954.
Goldstein 2014	Goldstein, Alex, Adam Kapelner, Justin Bleich, and Emil Pitkin. "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation." Journal of Computational and Graphical Statistics 24, no. 1 (2015): 44-65.
Guidotti 2018	Guidotti, Riccardo, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. "A survey of methods for explaining black box models." arXiv preprint arXiv:1802.01933(2018).
Gunning 2016	David Gunning Explainable Artificial Intelligence (XAI) DARPA/I2O 2016
Gorbunov 2011	Gorbunov, K. Yu, and Vassily A. Lyubetsky. "The tree nearest on average to a given set of trees." Problems of Information Transmission 47, no. 3 (2011): 274.
Hall 2018	Hall, P., Gill, N., Kurka, M., Phan, W. (May 2018). Machine Learning Interpretability with H2O Driverless AI. http://docs.h2o.ai.
Hardt 2016	Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." In Advances in neural information processing systems, pp. 3315-3323. 2016.
Hastie 1990	T. Hastie and R. Tibshirani. Generalized additive models . Chapman and Hall/CRC, 1990.
Herlocker 2000	Herlocker, Jonathan L., Joseph A. Konstan, and John Riedl. "Explaining collaborative filtering recommendations." In Proceedings of the 2000 ACM conference on Computer supported cooperative work, pp. 241-250. ACM, 2000.
Hoffman 2017	Hoffman, Robert R., Shane T. Mueller, and Gary Klein. "Explaining Explanation, Part 2: Empirical Foundations." IEEE Intelligent Systems 32, no. 4 (2017): 78-86.
Hutson 2018	Hutson, Matthew. "Has artificial intelligence become alchemy?." (2018): 478-478.
Koh 2017	Koh, Pang Wei, and Percy Liang. "Understanding black-box predictions via influence functions." arXiv preprint arXiv:1703.04730 (2017).Harvard
Kulesza 2014	Kulesza, Todd. "Personalizing machine learning systems with explanatory debugging." (2014).
Lakkaraju 2016	Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. "Interpretable decision sets: A joint framework for description and prediction." Proc. 22nd ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining. ACM, 2016.
Lei 2017	Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. Journal of the American Statistical Association, 2017
Lipovetsky 2001	Lipovetsky, Stan, and Michael Conklin. "Analysis of regression in game theory approach." Applied Stochastic Models in Business and Industry 17.4 (2001): 319-330
Lipton 2016	Lipton, Zachary C. "The mythos of model interpretability." arXiv preprint arXiv:1606.03490 (2016).
Lipton 2017	Lipton, Zachary C. "The Doctor Just Won't Accept That!." arXiv preprint arXiv:1711.08037 (2017).
Lou 2013	Lou, Yin, Rich Caruana, Johannes Gehrke, and Giles Hooker. "Accurate intelligible models with pairwise interactions." In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 623-631. ACM, 2013.
Lundberg 2017	Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." In Advances in Neural Information Processing Systems, pp. 4765-4774. 2017.
Miller 2017A	Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv preprint arXiv:1706.07269 (2017).
Miller 2017B	Miller, Tim, Piers Howe, and Liz Sonenberg. "Explainable AI: Beware of inmates running the asylum." In IJCAI-17 Workshop on Explainable AI (XAI), vol. 36. 2017.
Montavon 2017	Montavon, Grégoire, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. "Explaining nonlinear classification decisions with deep taylor decomposition." Pattern Recognition 65 (2017): 211-222.

References

Moosavi-Dezfooli 2016	Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a simple and accurate method to fool deep neural networks." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574-2582. 2016.
Morstatter 2016	Fred Morstatter and Huan Liu Measuring Topic Interpretability with Crowdsourcing KDD Nuggets November 2016
Nott 2017	Nott, George "Google's research chief questions value of 'Explainable AI'" Computer World 23 June, 2017
Olah 2018	Olah, Chris, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. "The building blocks of interpretability." Distill 3, no. 3 (2018): e10.
Perez 2004	Pérez, Jesus Maria, Javier Muguerza, Olatz Arbelaitz, and Ibai Gurrutxaga. "A new algorithm to build consolidated trees: study of the error rate and steadiness." In Intelligent Information Processing and Web Mining, pp. 79-88. Springer, Berlin, Heidelberg, 2004.
	Quinlan, J. Ross. "Some elements of machine learning." In International Conference on Inductive Logic Programming, pp. 15-18. Springer, Berlin, Heidelberg, 1999.
Ras 2018	Ras, Gabrielle, Pim Haselager, and Marcel van Gerven. "Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges." arXiv preprint arXiv:1803.07517(2018).
Ribeiro 2016a	Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144. ACM, 2016.
Ribeiro 2016b	Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin Introduction to Local Interpretable Model-Agnostic Explanations (LIME) August 12, 2016 Oriely Media
Ross 2018	Ross, Casey IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show Stat News July 25, 2018
Ross 2017	Ross, Andrew Slavin, Michael C. Hughes, and Finale Doshi-Velez. "Right for the right reasons: Training differentiable models by constraining their explanations." arXiv preprint arXiv:1703.03717 (2017).
Saabas 2014	Saabas, Ando. Interpreting random forests Data Dive Blog 2014
Shrikumar 2017	Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences." arXiv preprint arXiv:1704.02685 (2017)
Strumbelj 2014	Strumbelj, Erik, and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions." Knowledge and information systems 41.3 (2014): 647-665.
Tan 2017	Tan, Sarah, Rich Caruana, Giles Hooker, and Yin Lou. "Detecting Bias in Black-Box Models Using Transparent Model Distillation." arXiv preprint arXiv:1710.06169 (2017).
Turing 1950	Machinery, Computing. "Computing machinery and intelligence-AM Turing." Mind 59, no. 236 (1950): 433.
Ustun 2016	Ustun, Berk, and Cynthia Rudin. "Supersparse linear integer models for optimized medical scoring systems." Machine Learning 102, no. 3 (2016):349-391.
Wang 2015	Wang, Fulton, and Cynthia Rudin. "Falling rule lists." In Artificial Intelligence and Statistics, pp. 1013-1022. 2015.
Weld 2018	Weld, Daniel S., and Gagan Bansal. "Intelligible Artificial Intelligence." arXiv preprint arXiv:1803.04263 (2018).
Weller 2017	Weller, Adrian. "Challenges for transparency." arXiv preprint arXiv:1708.01870 (2017).
Wick 1992	M. R. Wick and W. B. Thompson. Reconstructive expert system explanation. Artificial Intelligence, 54(1- 2):33–70, 1992
Yang 2016	Yang, Hongyu, Cynthia Rudin, and Margo Seltzer. "Scalable Bayesian rule lists." arXiv preprint arXiv:1602.08610 (2016).

ACM: The Learning Continues…

- Questions/comments about this webcast? learning@acm.org
- ACM Code of Ethics: https://ethics.acm.org
- ACM's Discourse Page: http://on.acm.org
- ACM Learning Webinars (on-demand archive): http://webinar.acm.org
- ACM Learning Center: http://learning.acm.org
- ACM SIGKDD: http://www.kdd.org/