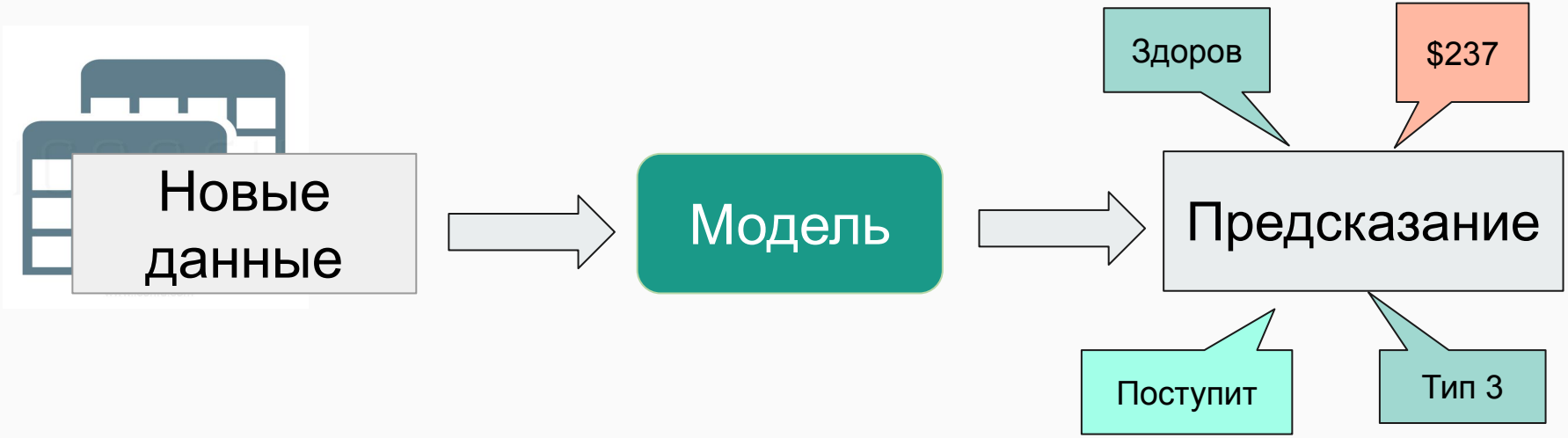
iML и XAI

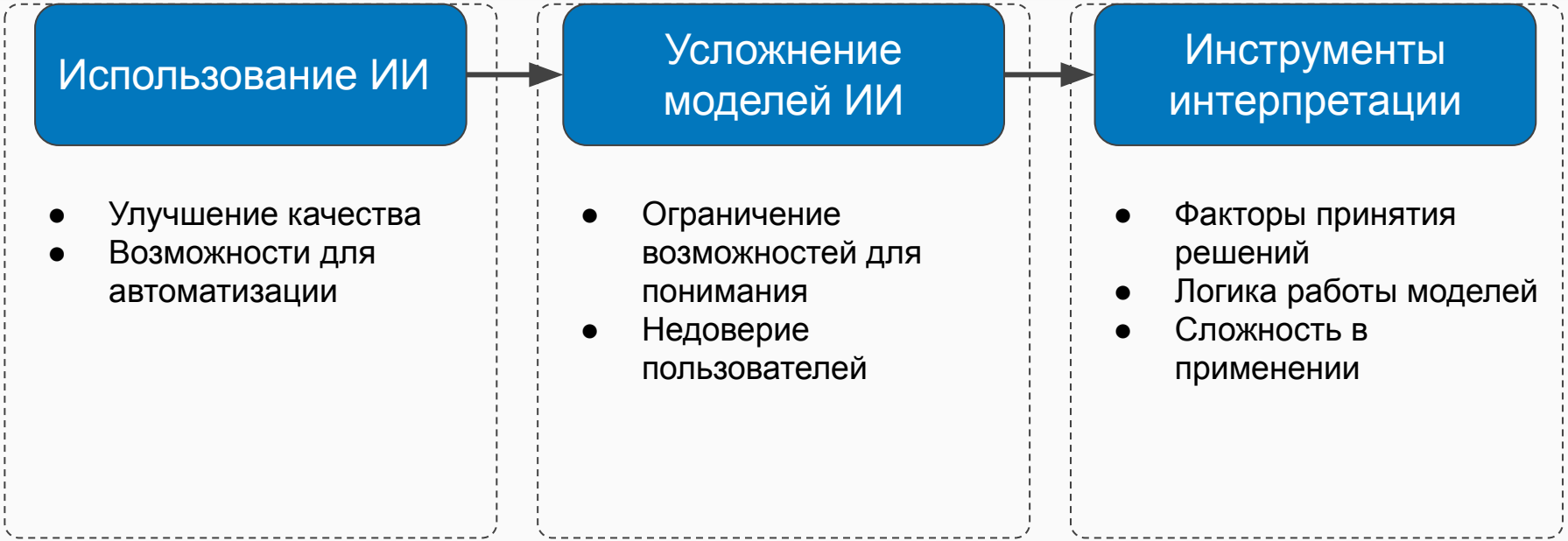
что может пойти не так

НУГ “Машинное обучение и социальный компьютинг” (сентябрь, 2020), проект 20-04-024

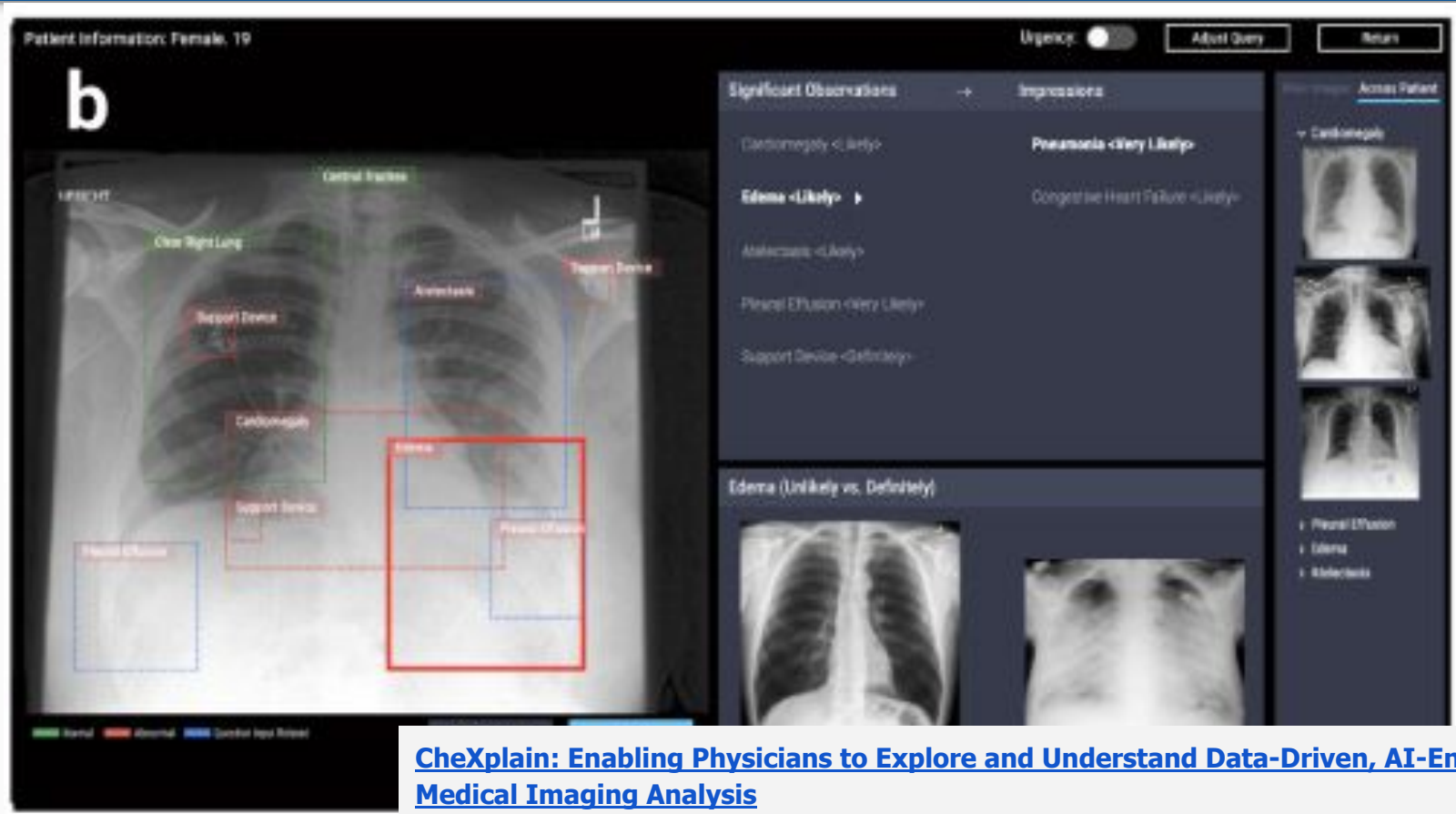
“Черный ящик”


Новые данные
Модель
Здоров
$237
Предсказание
Поступит
Тип 3

Зачем

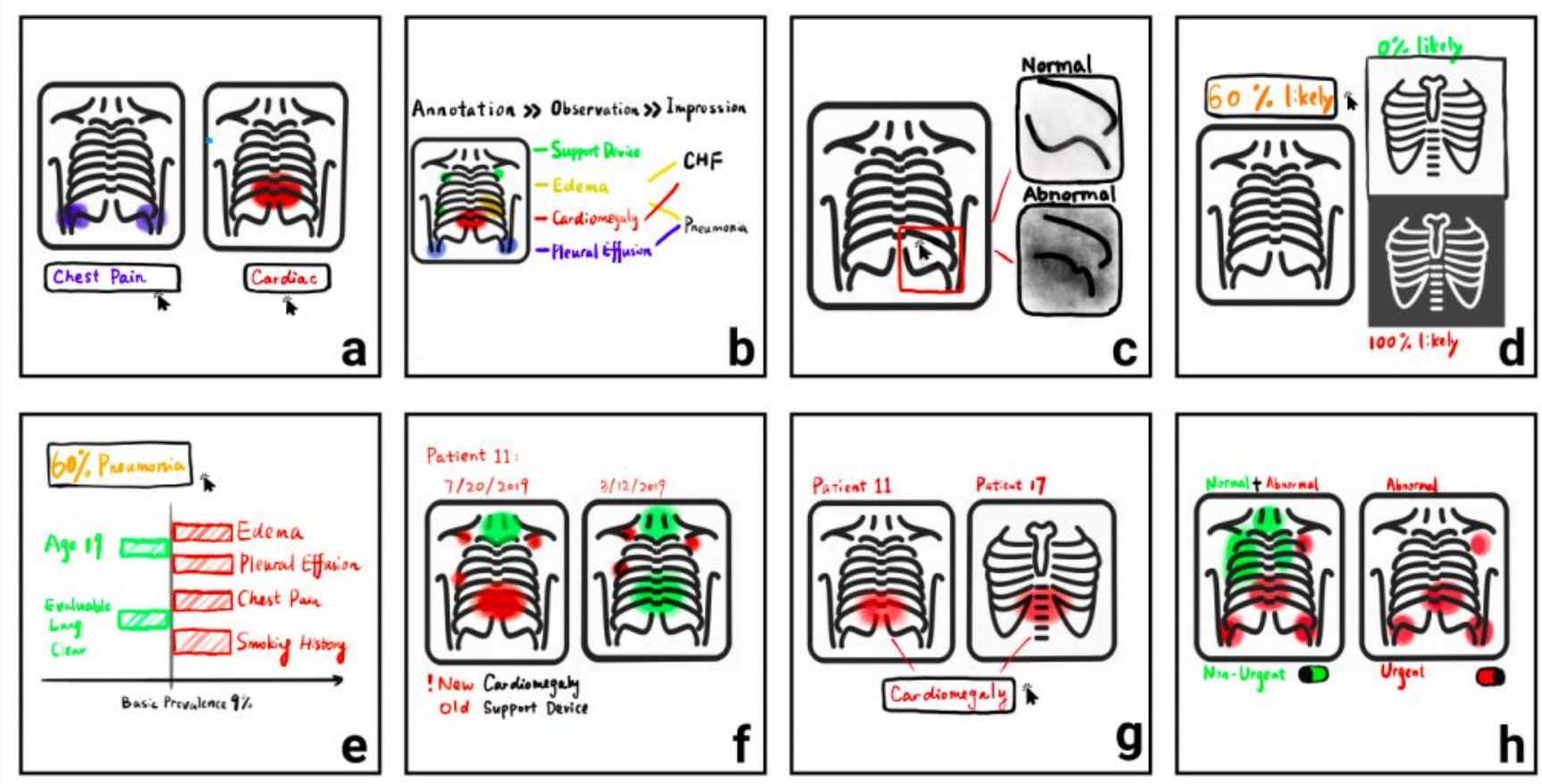

Использование ИИ
Усложнение моделей ИИ
Инструменты интерпретации
Улучшение качества
Возможности для автоматизации
Ограничение возможностей для понимания
Недоверие пользователей
Факторы принятия решений
Логика работы моделей
Сложность в применении

CheXplain: объяснение предсказаний на снимках (CHI 2020)



[CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis]

CheXplain: объяснение предсказаний на снимках (CHI 2020)



Подходы к интерпретации

Класс методов	**Примеры методов**	**Вопросы**
Объяснение модели	Значимость признаков (SHAP), динамика эффекта (PDP)	Как
Объяснение предсказания	Значимость признаков (LIME, SHAP), извлечение правил	Почему
Основанные на контрпримерах	Что-Если, индивидуальное условное ожидание (ICE)	Что если, Как получить/сохранить значение, Почему
Основанные на примерах	Контрпримеры, исторические данные	Почему, Как получить/сохранить значение

*основано на лит. обзоре ВКР Смирновой Анны, 2020, включая Liao V. et al. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences, DOI: https://doi.org/10.1145/3313831.3376590

Не только LIME

- Accumulated Local Effects (ALE) Plot
- SHAP (SHapley Additive exPlanations)
- Anchors (от авторов LIME, но результат в виде правил ЕСЛИ-ТО)
- Контр-примеры (“если я изменю признак X, то предсказание изменится на противоположное”)
- Похожие примеры
- Влиятельные наблюдения
- ...

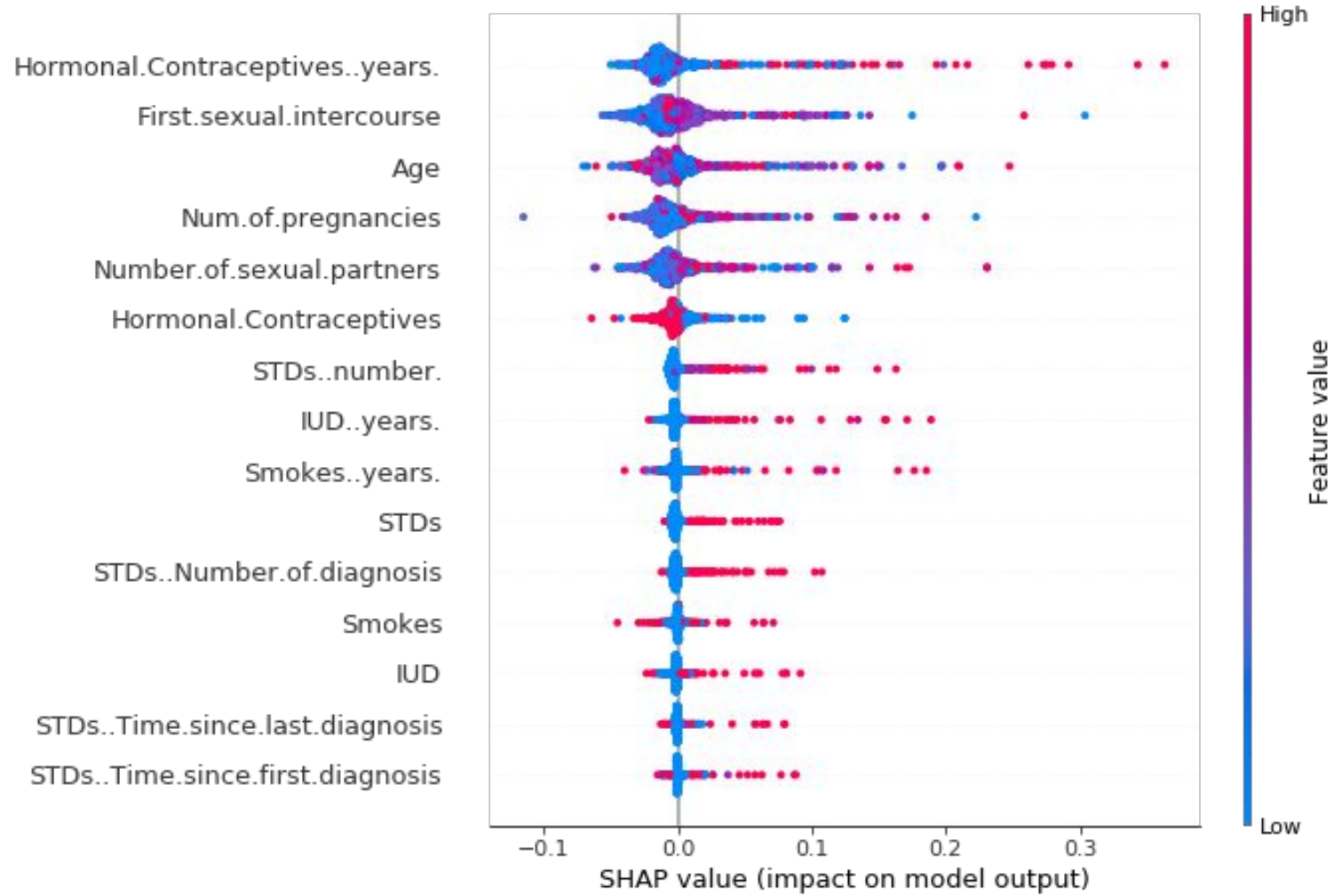
win-win: все отлично

- строим сложный многоуровневый ансамбль с высокой точностью (или другой целевой метрикой)
- поверх строим интерпретацию, чтобы понять, что на что влияет
- победа!

Зачем нам тогда простые модели?
И зачем нам всякие диагностики и т.д.?

Что может пойти не так

- эти методы интерпретируют **модель**, не реальность (другая модель -- другие выводы)
- приближение к приближению (“сломанный телефон”)
- модели могут быть плохими (низкое качество) и модели интерпретации моделей тоже могут быть плохими
- **неправильные выводы из модели интерпретации**
  - локальные методы дают локальную интерпретацию, нельзя делать выводы в общем
  - нет понимания, что именно показывает тот или метод / визуализация



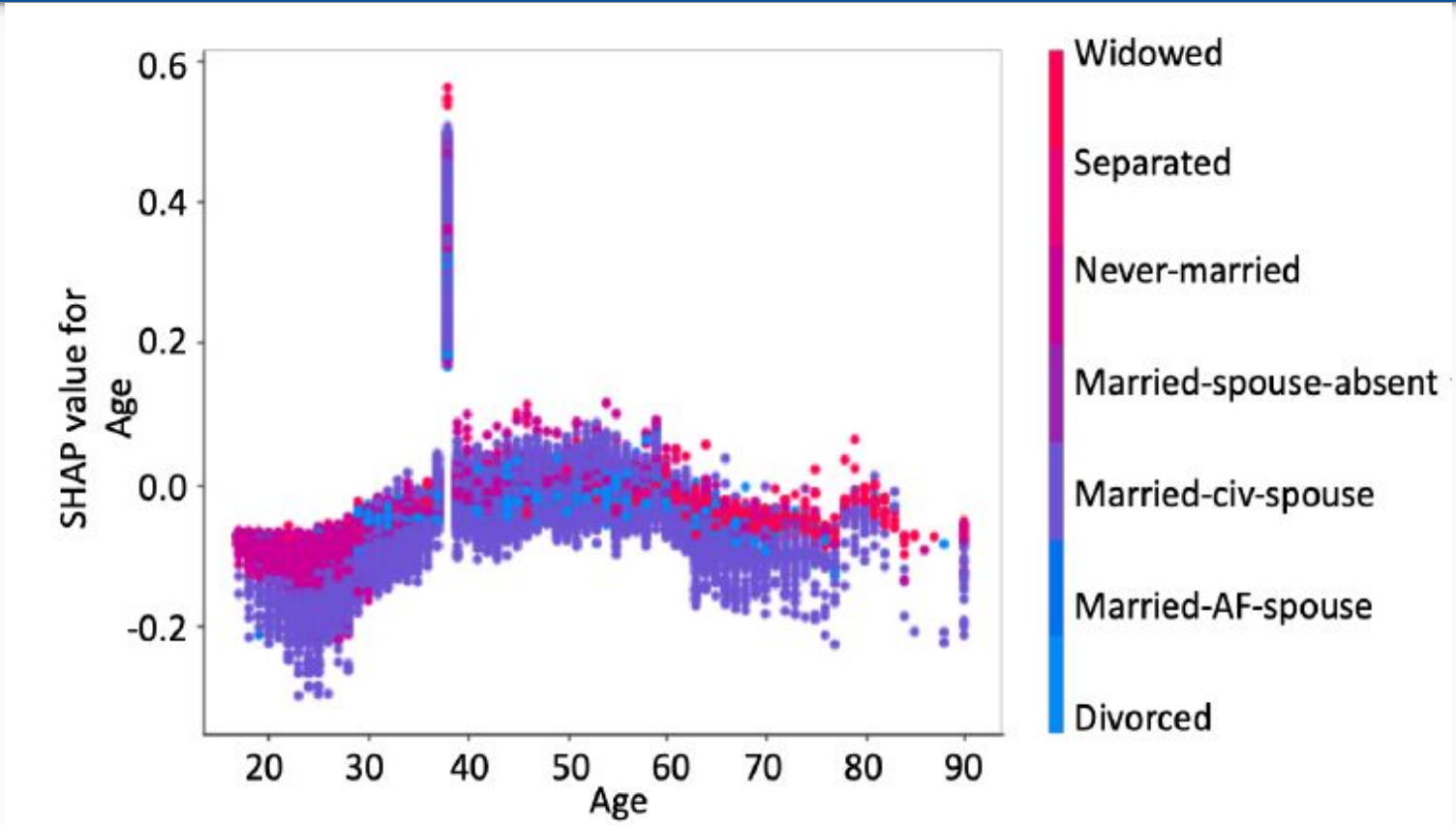
Пример из книги [Interpretable Machine Learning]

Использование iML

Согласно Kaur et al. (Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning, CHI 2020), аналитики используют инструменты не так, как это планировалось их разработчиками

- склонны слишком доверять подобным моделям, даже не понимая принципы их работы, если результаты представлены в “научном” формате, с визуализацией и/или подкреплены ссылками на публикации
- склонны рационализировать отклонения -- если есть правдоподобное объяснение наблюдению, то оно не считается смещением или ошибочным предсказанием
- люди с большим опытом работы в ML правильнее интерпретируют визуализации, но и критичнее к ним относятся

Пример визуализации в исследовании



Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning

Почему HCI

- модели существуют не сами по себе, ими пользуются люди
  - разработчику нужно находить смещения в модели и/или данных
  - эксперту нужно понять, адекватна ли модель
  - PM нужно понять, стоит ли внедрять модель
  - конечный пользователь хочет знать, почему предсказание именно такое
- как упростить взаимодействие с системой и избежать ошибок
- как пользователь меняет систему, отличается ли реальное использование от запланированного

Результаты исследования: Ранжирование

Результаты исследования: Кластеризация

КЛАСТЕР 0

- Отсутствие явных предпочтений
- Предпочтение **направленным** моделям с **известной точностью**
- Отсутствуют разработчики

КЛАСТЕР 1

- Предпочтительная модель: **контрпримеры**
- Редко отдавали предпочтение **ненаправленным** моделям
- Преимущественно менеджеры и разработчики

КЛАСТЕР 2

- Предпочтительные модели: **динамика эффекта** и **значимость факторов**
- Редко выбирали модели с **численными** значениями, известной **точностью** и **методом**

КЛАСТЕР 3

- Предпочтительные модели: контрпримеры и **динамика эффекта**
- Выбирали **ненаправленные** модели

КЛАСТЕР 4

- Предпочтительная модель: **динамика эффектов**
- Выбирали **глобальные**, **ненаправленные** подходы
- Преимущественно менеджеры, разработчики и аналитики

КЛАСТЕР 5

- Предпочтительная модель: **Что-Если**
- Предпочтение **локальным** моделям
- Нет специалистов по данным

Полезные ссылки

- Readling list of CHI2020 papers on human-AI interaction (не только iML)