Reflections

On the meaningful understanding of the logic of automated decision-making

Eric Horvitz

Berkeley Center for Law & Technology
Privacy Law Forum: Silicon Valley
March 24, 2017

horvitz@microsoft.com
http://erichorvitz.com
Microsoft Research

General Data Protection Regulation

(EU) 2016/679, 25 May 2018

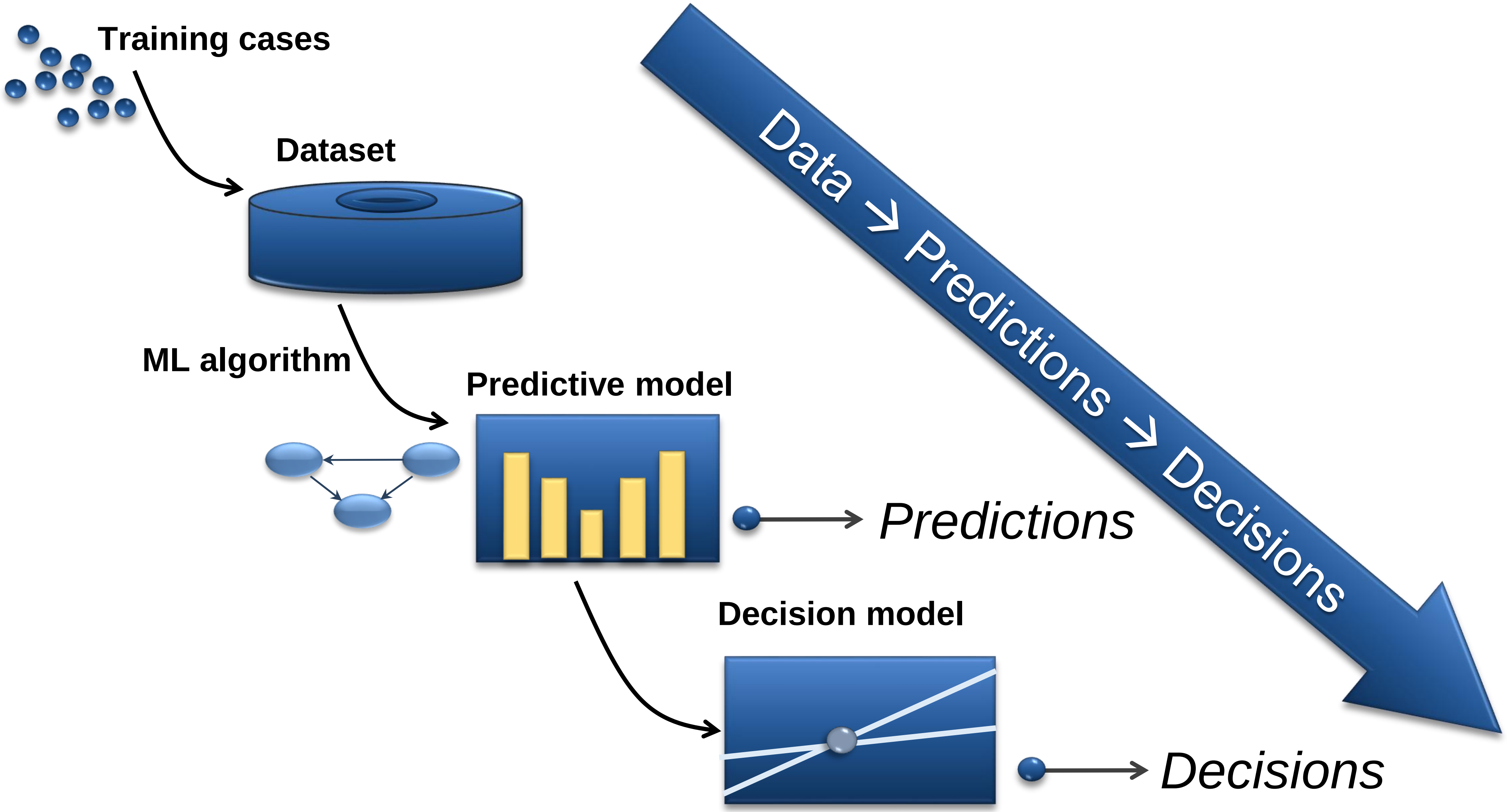
Article 15 (1)(h):

"existence of automated decision-making…and …**meaningful information about the logic involved**

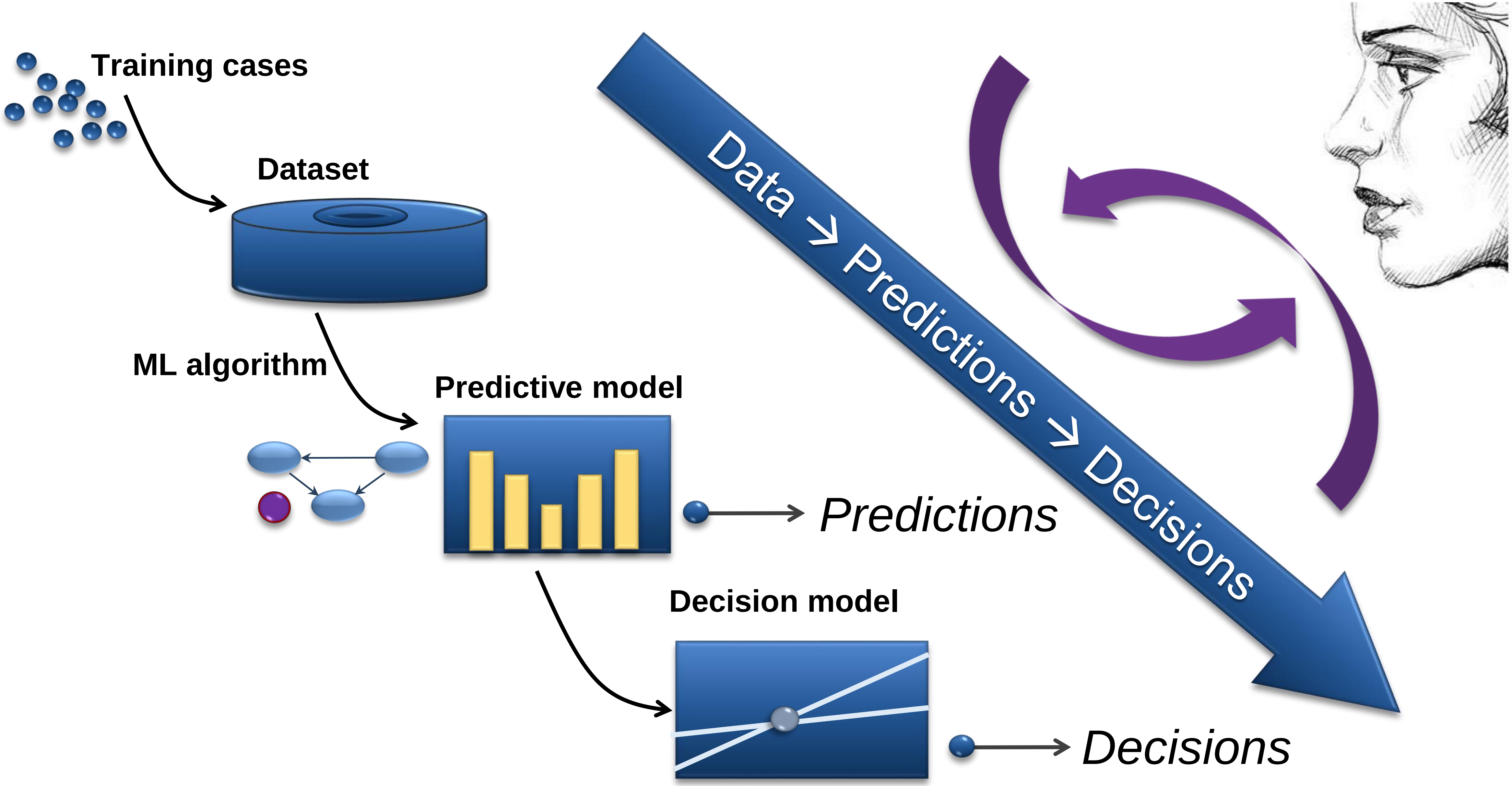
Many questions about interpreting Article.

Multiple aspects of AI decision-making pipeline.

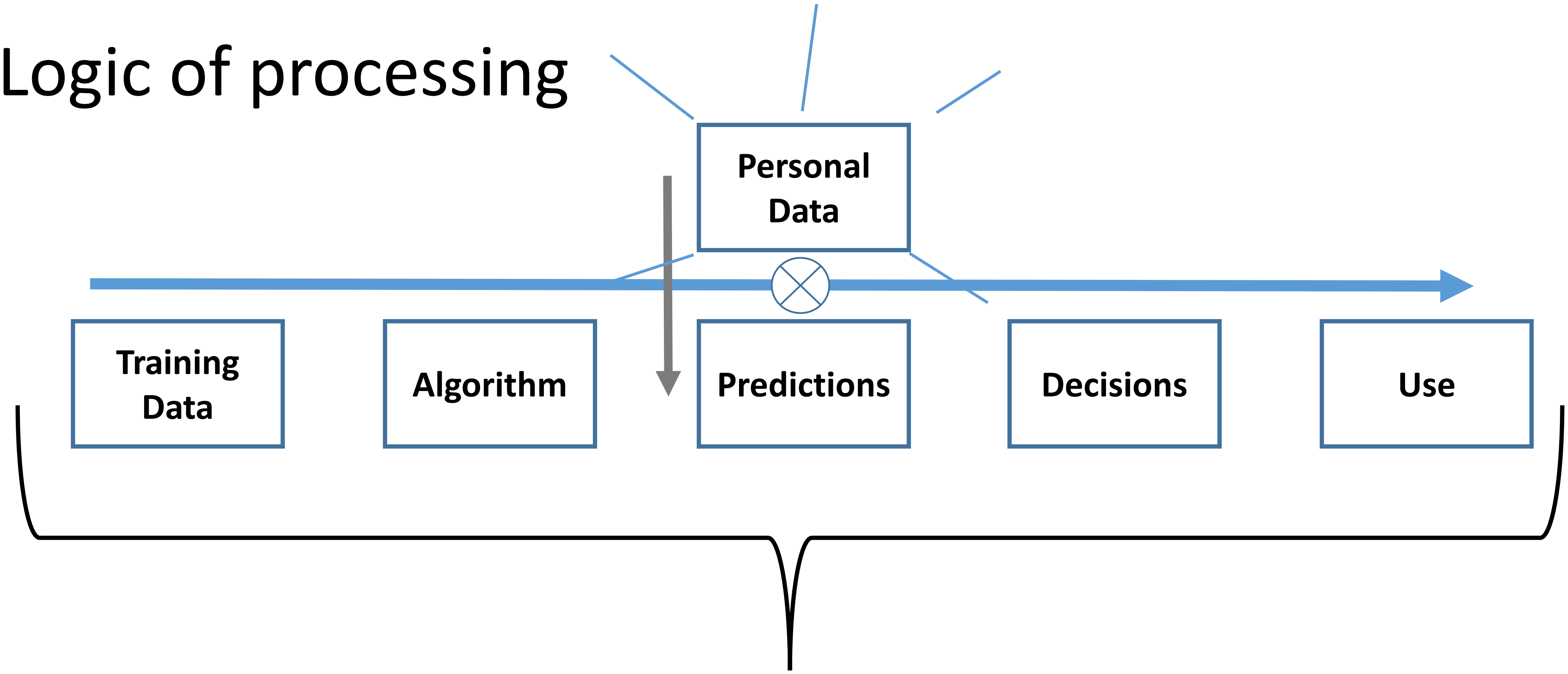
Data → Predictions → Decisions



Data → Predictions → Decisions

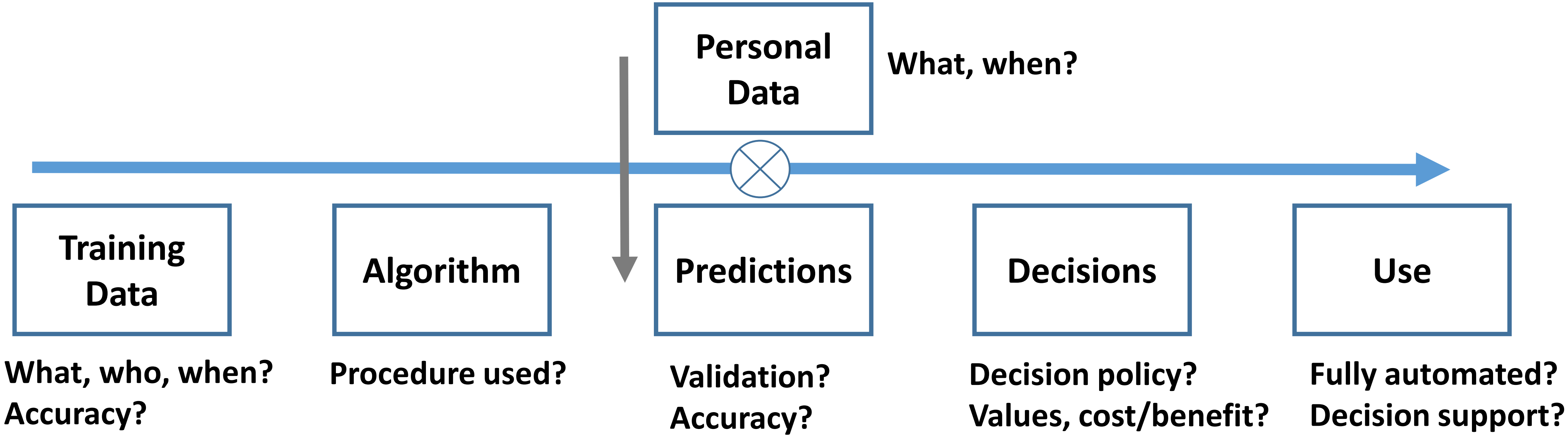


Logic of processing

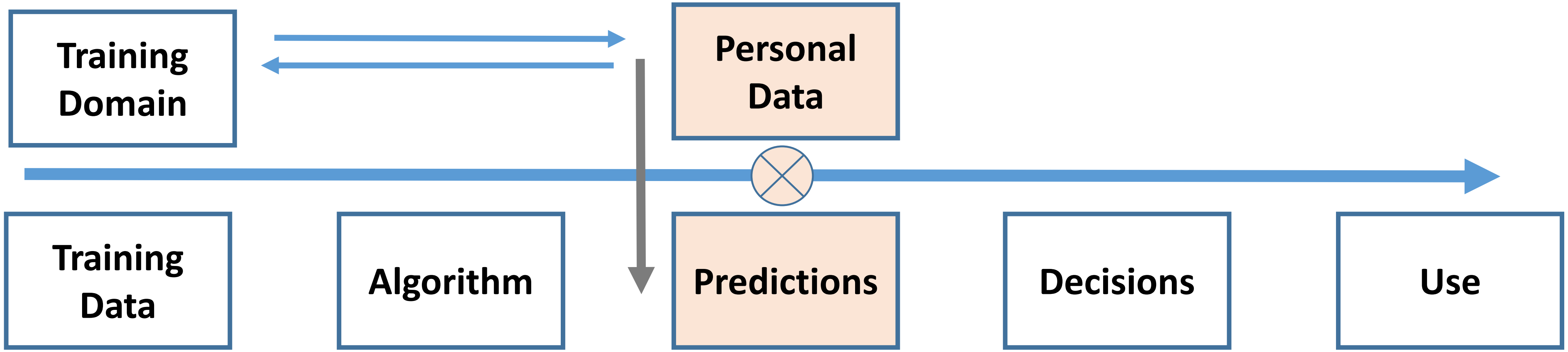


meaningful information about the logic involved

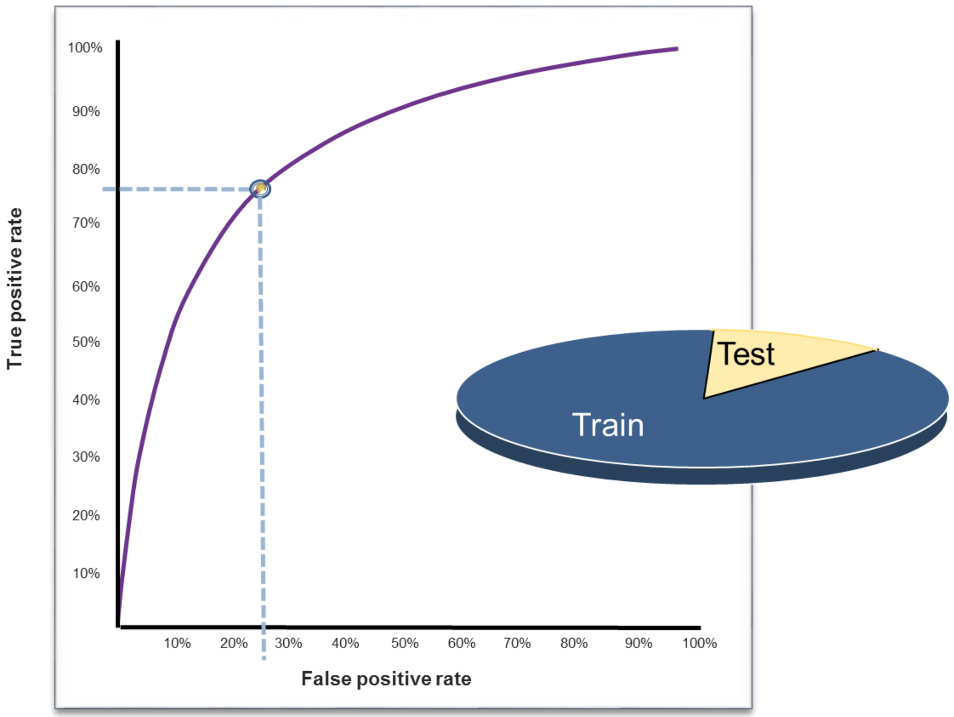
Logic of processing



Localization & Personalization



Validation?
Accuracy?



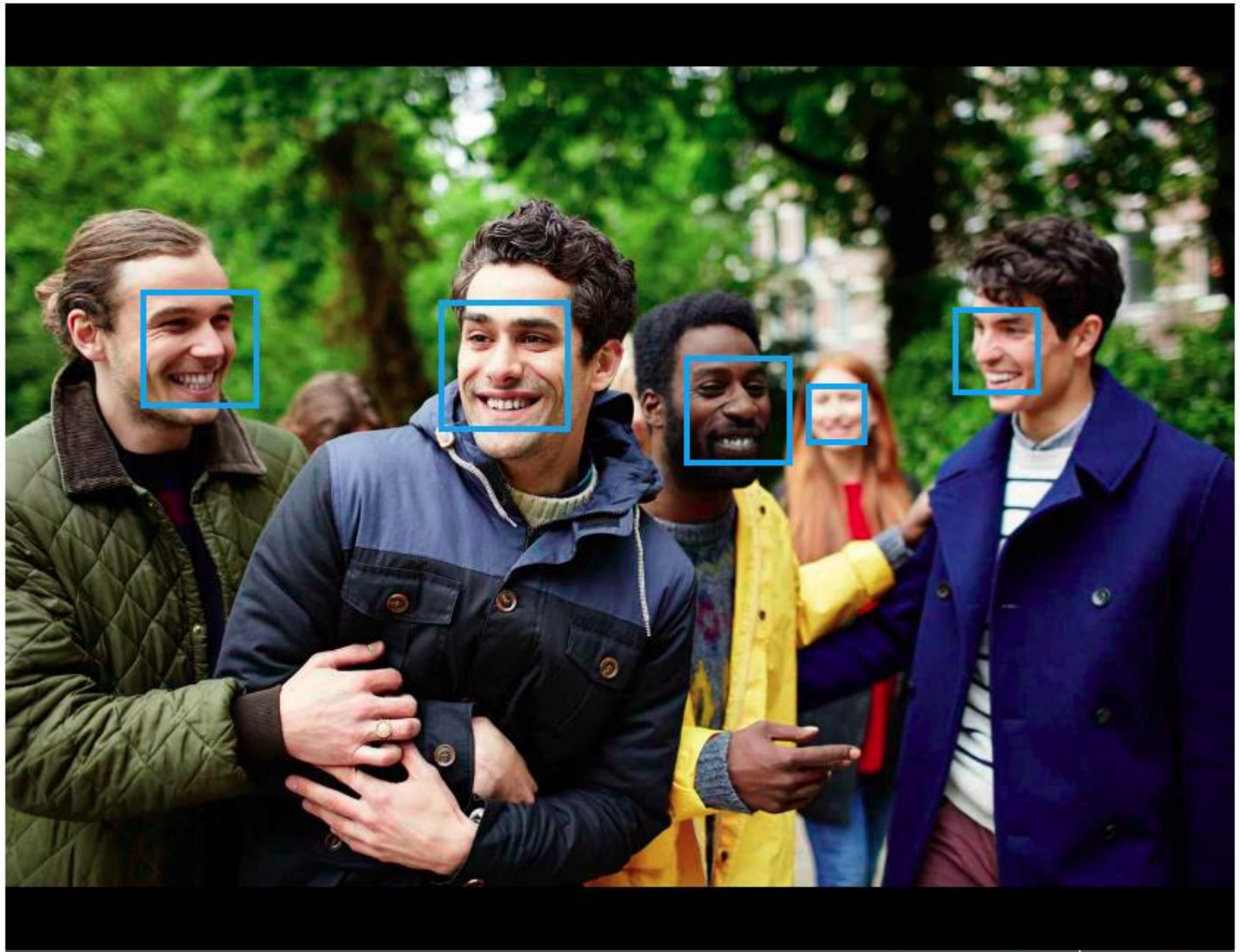


Image URL

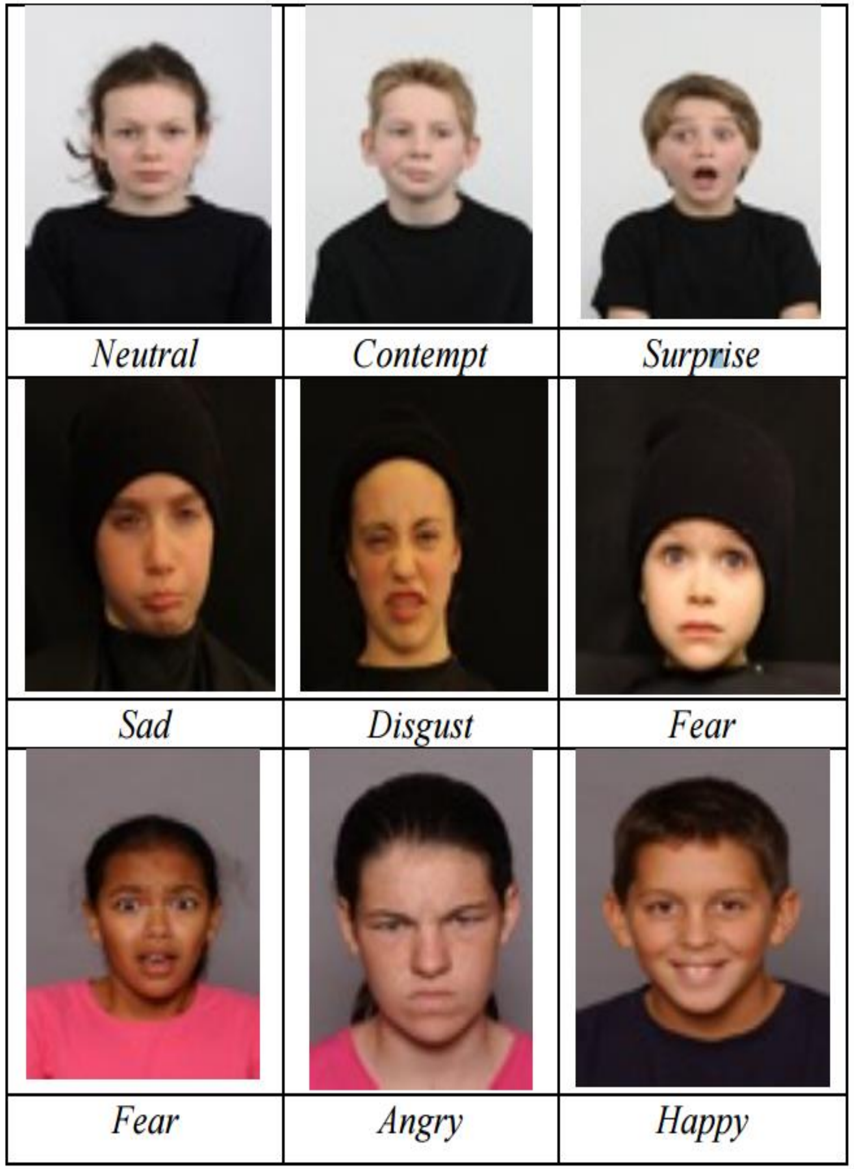
Detection Result:
5 faces detected

JSON:
[
  {
    "faceRectangle": {
      "left": 488,
      "top": 263,
      "width": 148,
      "height": 148
    },
    "scores": {
      "anger": 9.075572e-13,
      "contempt": 7.048959e-9,
      "disgust": 1.02152783e-11,
      "fear": 1.778957e-14,
      "happiness": 0.9999999,
      "neutral": 1.31694478e-7,
      "sadness": 6.04054263e-12,
      "surprise": 3.92249462e-11
    }
  },
  {
    "faceRectangle": {






Localization & Personalization



Addressing Bias in Machine Learning Algorithms: A Pilot Study on Emotion Recognition for Intelligent Systems

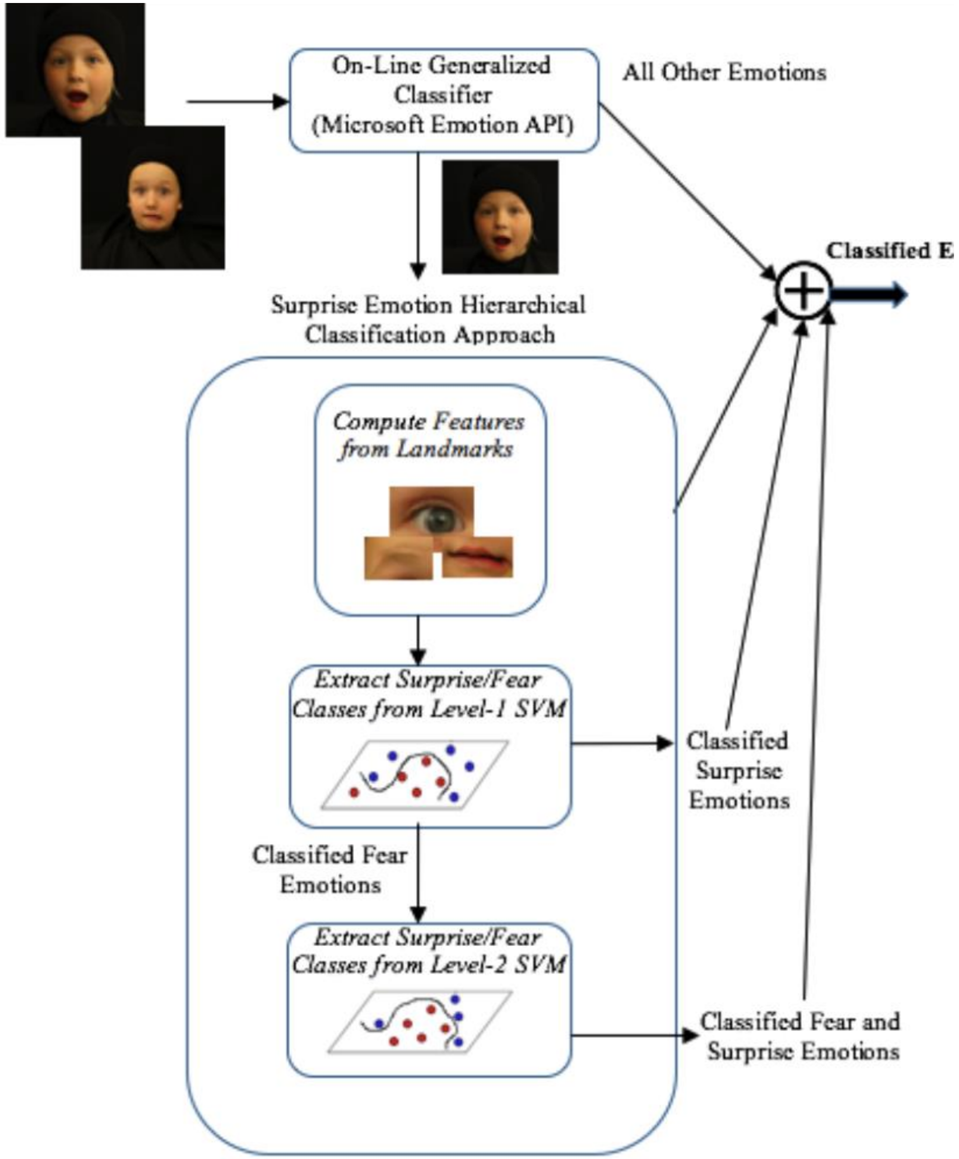
Ayanna Howard[1*], Cha Zhang[2], Eric Horvitz[2]

March 2017

Machine learning “contact lens” for children

A. Howard, C. Zhang, E. Horvitz (2010). [Addressing Bias in Machine Learning Algorithms: A Pilot Study on Emotion Recognition for Intelligent Systems](). IEEE Workshop on Advanced Robotics and its Social Impacts.

Localization & Personalization



Addressing Bias in Machine Learning Algorithms: A Pilot Study on Emotion Recognition for Intelligent Systems

Ayanna Howard[1*], Cha Zhang[2], Eric Horvitz[2]

March 2017

Machine learning "contact lens" for children

A. Howard, C. Zhang, E. Horvitz (2010). [Addressing Bias in Machine Learning Algorithms: A Pilot Study on Emotion Recognition for Intelligent Systems](). IEEE Workshop on Advanced Robotics and its Social Impacts.

Moving into High-Stakes Arenas

Localization & Personalization

Readmissions Manager

Reducing Hospital Readmissions is an Impending Priority

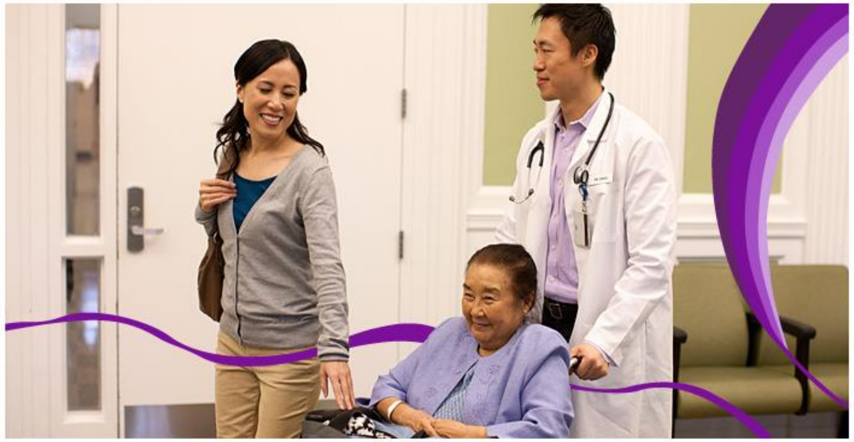
Overview

One in five Medicare inpatients is readmitted within 30 days. The Centers for Medicare and Medicaid Services (CMS) considers 40%-75% of these readmissions to be preventable.

In October 2012, CMS will begin to track readmission and impose financial penalties on hospitals with higher–than–expected readmission rates for certain conditions. Other payers will certainly follow.

It is clear that hospital admissions and readmissions are becoming a critical parameter for tracking care delivery from both a financial and quality perspective.

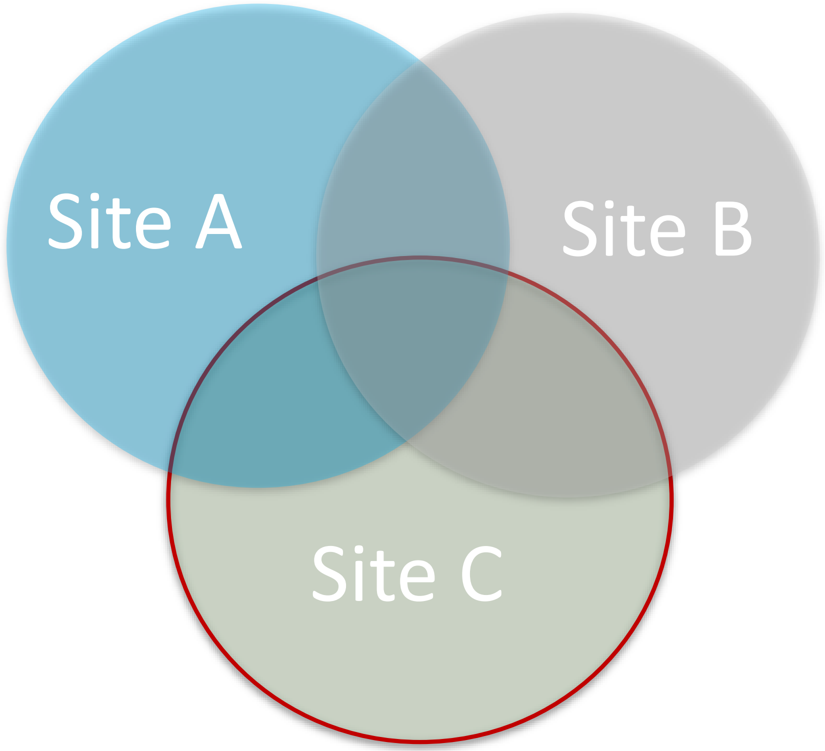
Readmissions Manager for Microsoft Amalga is an innovative solution to help organizations address this very important business need.



Readmissions Manager Targets Avoidable Hospital Readmissions

PROB_NUM_%	FACTORS_PRO_READMISSION
37.9	Num past 6m visits = 6 to 10 / Patient had dx = Disorders of fluid, electrolyte, an
32.72	stayed <1 day in the hospital / Patient had dx = Disorders of fluid, electrolyte, and
30.83	Patient had dx = Chronic renal failure / 44 < Age < 60
29.05	Patient had dx = Disorders of fluid, electrolyte, and acid-base balance / Patient ha
28.54	
27.36	Patient had dx = Acute renal failure / Patient had dx = Chronic renal failure
18.05	Patient had dx = Other personal history presenting hazards to health / Patient ha
16.57	stayed <1 day in the hospital
16.18	Patient had dx = Disorders of fluid, electrolyte, and acid-base balance / Patient ha
15.52	
14.53	stayed <1 day in the hospital / Ave gap of past yr visits = between 15 and 30 days
14.42	stayed <1 day in the hospital / Patient had dx = Other personal history presentinc
14.39	stayed <1 day in the hospital
13.59	stayed <1 day in the hospital / 44 < Age < 60
13.36	stayed <1 day in the hospital / Hour of visit = 00
12.44	stayed <1 day in the hospital

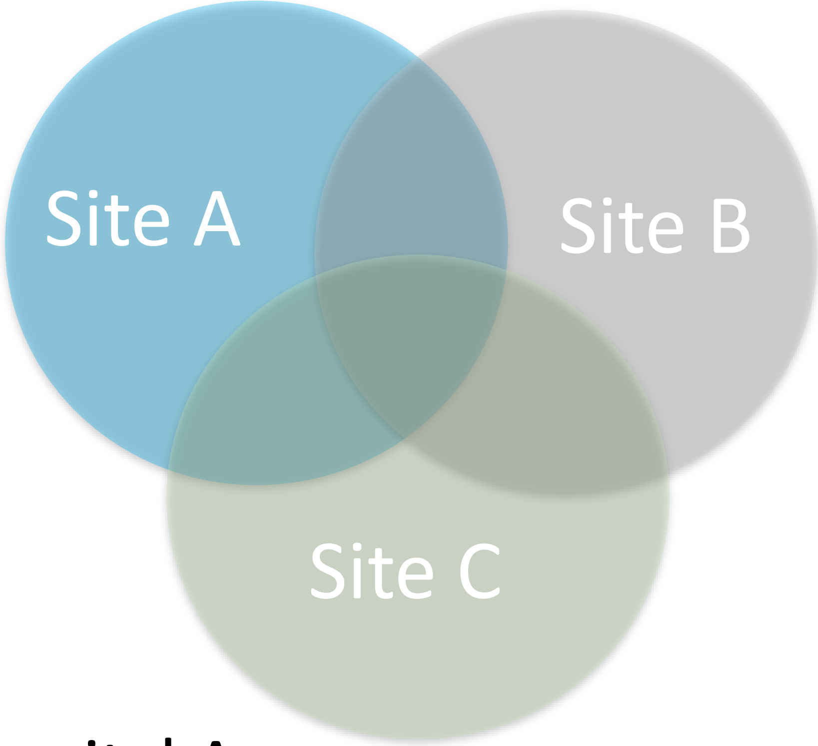
M. Bayati, M. Braverman, M. Gillam, K.M. Mack, G. Ruiz, M.S. Smith, E. Horvitz (2014). Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study. PLOS One Medicine.



Site-specific evidence

Site-specific pts, prevalencies

Site covariate dependencies



Site-specific eviden

Site-specific pts, pr

Site covariate depe

Hospital A

Community hospital: 180 beds, 10,000 admissions/

Hospital B

Acute care teaching hospital: 250 beds, 15, 000 inpa

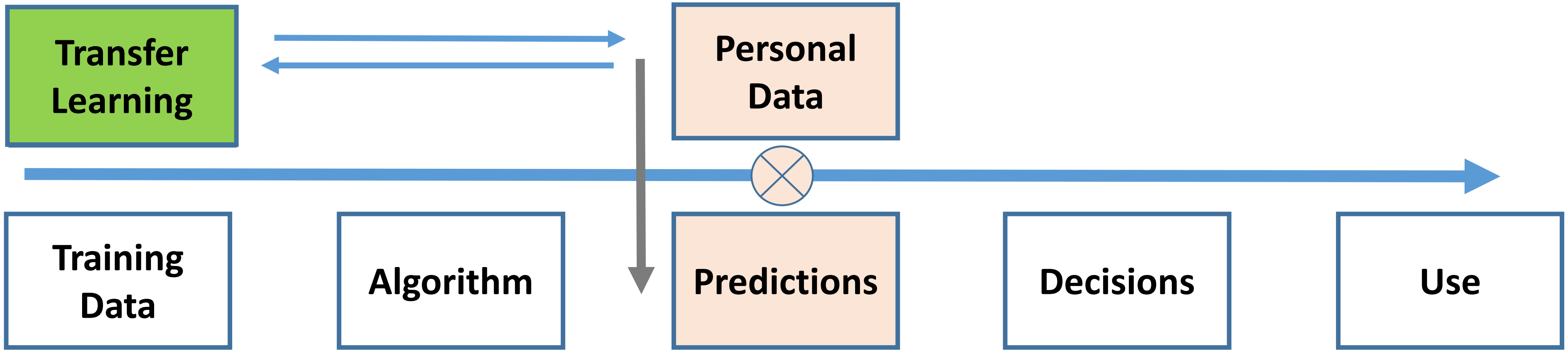
Hospital C

Major teaching & research hospital: 900 beds, 40,0

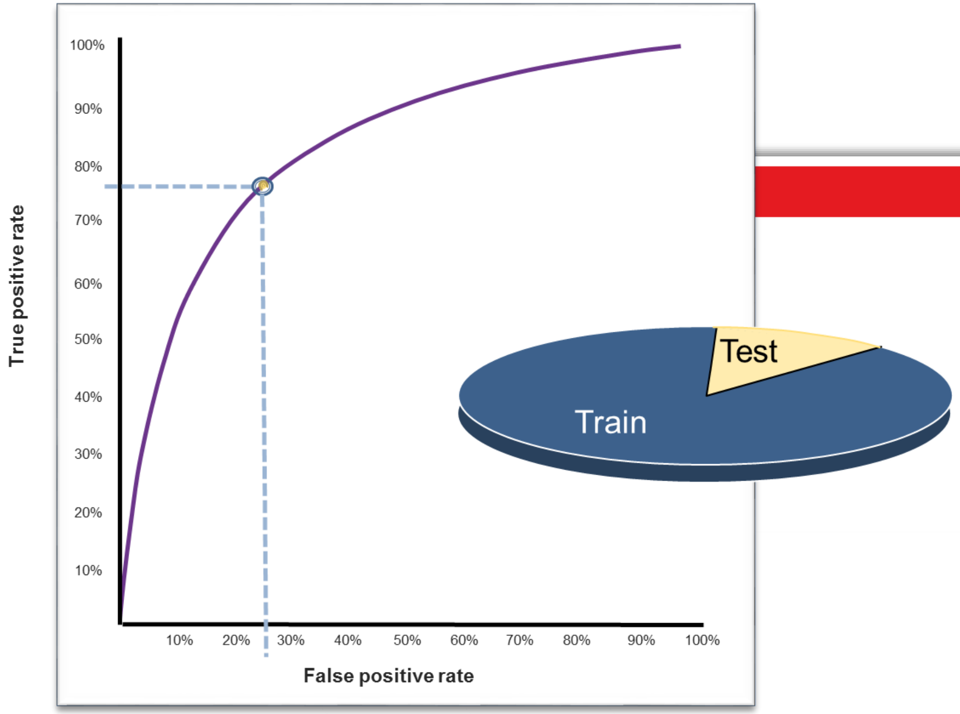
Table 1 Descriptive statistics comparing the study population across the three different institutions

	Hospital A (%) (n=21 959)	Hospital B (%) (n=29 315)	Hospital C (%) (n=81 579)
Female gender	62.34	50.29	55.97
Age:			
[0, 2)	14.38	0.00	9.00
[2, 10)	0.75	0.00	0.00
[10, 15)	0.80	0.07	0.00
[15, 25)	7.23	3.77	6.73
[25, 45)	21.27	15.46	19.05
[45, 60)	21.28	30.98	22.77
[60, 70)	13.16	21.19	16.78
[70, 80)	10.79	15.97	13.74
[80 100)	8.11	10.20	9.24
≥100	2.25	2.36	2.67
Hospital admission type:			
Newborn	13.13	0.00	8.74
Term pregnancy	7.53	0.00	8.89
Routine elective	15.87	31.28	17.39
Urgent	7.53	7.84	11.26
Emergency	10.79	15.97	13.74
Hospital service:			
Medicine	51.18	49.15	40.85
Orthopedics	5.61	18.76	1.54
Surgery	7.53	5.97	10.28
Obstetrics	13.97	0.00	10.09
Cardiology	0.00	2.99	11.36
Newborn	13.15	0.00	9.01
Psychiatry	0.00	13.11	3.70
Hemodialysis	3.06	5.32	6.76
Diabetic	24.44	32.73	33.59
Clostridium difficile	0.80	1.08	1.05
Previous visit in past 90 days	5.87	7.43	5.54

Challenges of Localization & Personalization



Validation?
Accuracy?

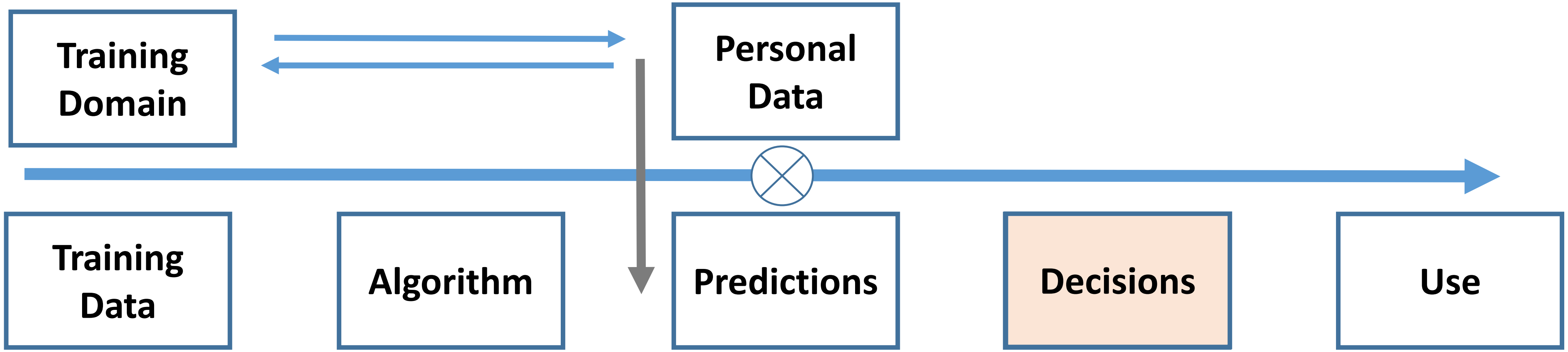



Research and applications

A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions

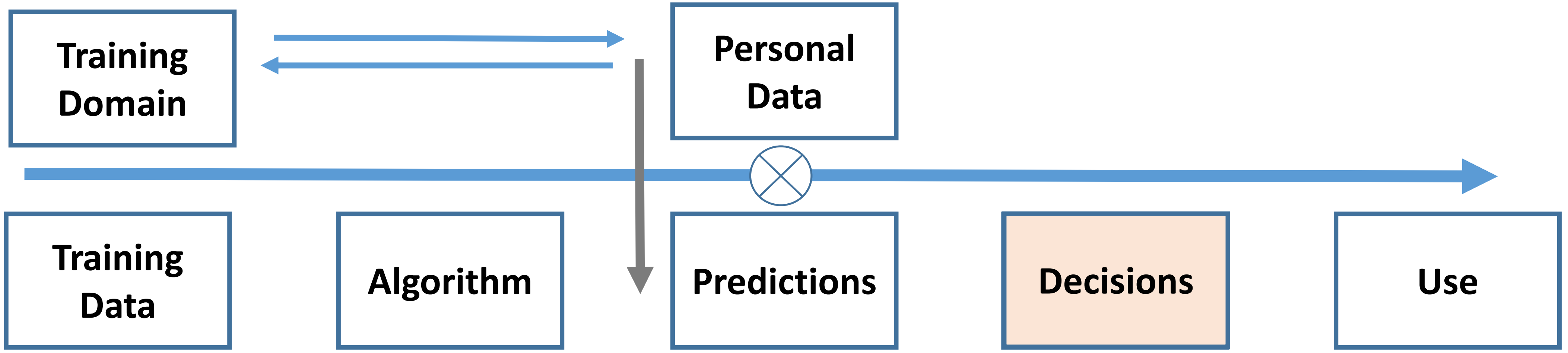
J. Wiens, J. Guttag, and E. Horvitz (2014). [A Study in Transfer Learning: Leveraging Data from Multiple Hospitals to Enhance Hospital-Specific Predictions](), *Journal of the American Medical Informatics Association*.

On Decisions & Values

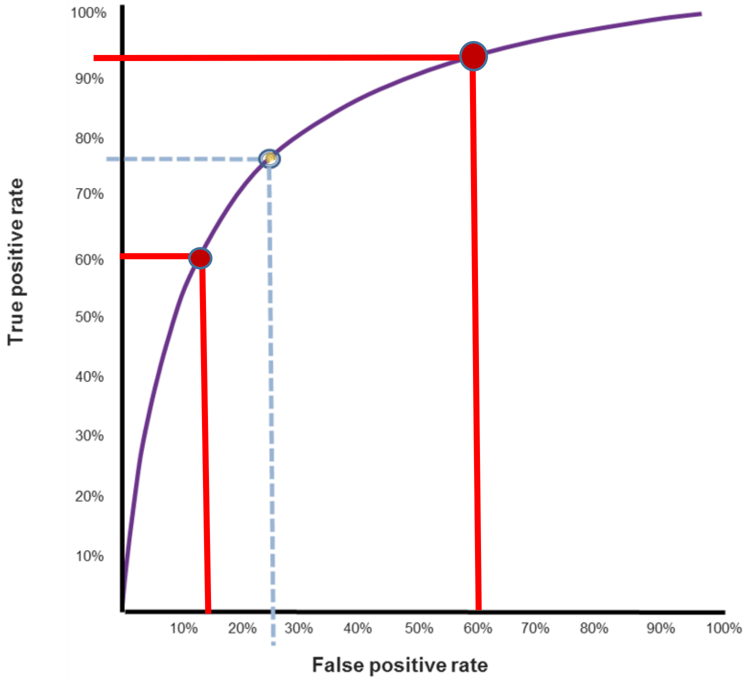


Validation?
Accuracy?

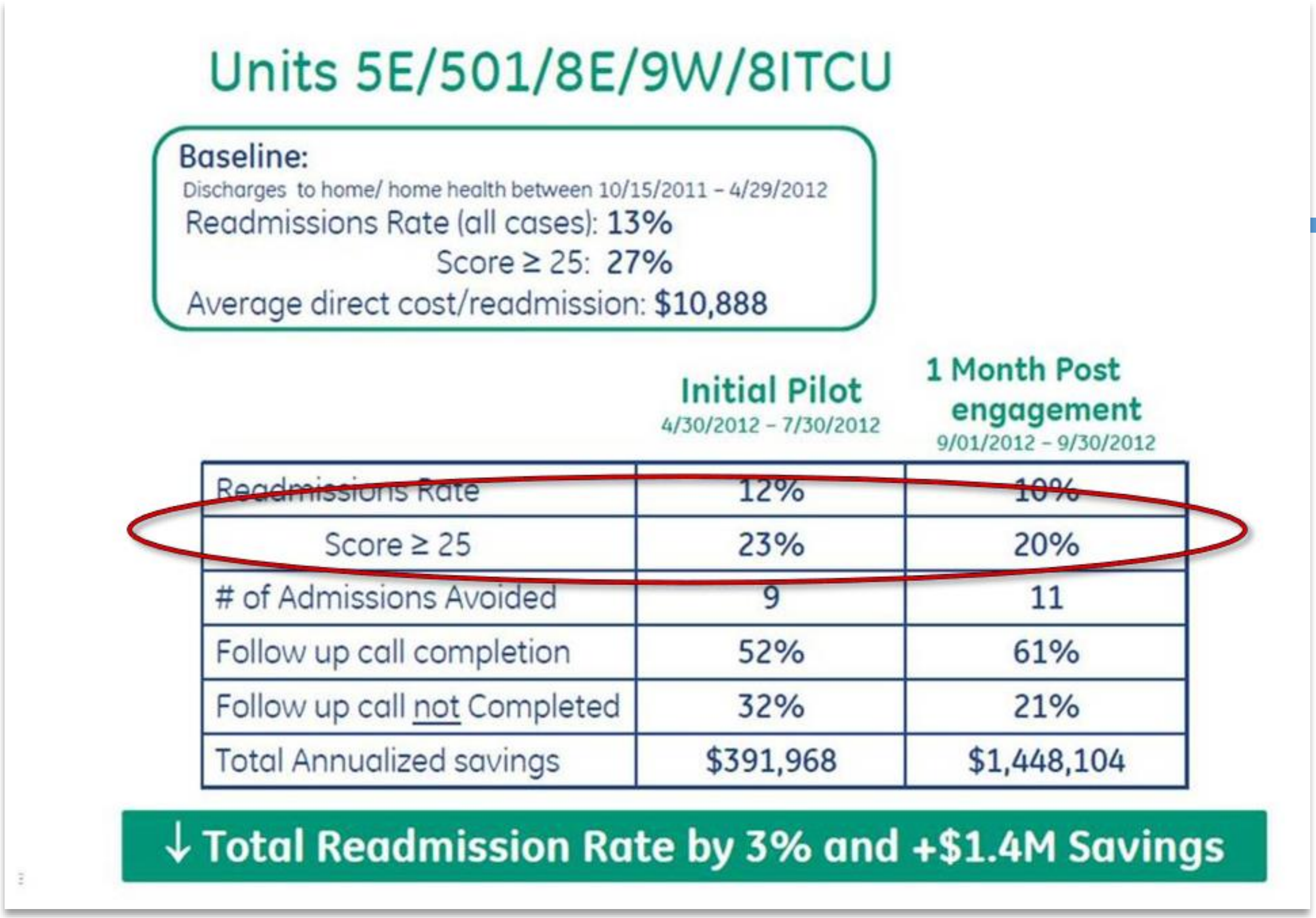
On Decisions & Values



Validation?
Accuracy?



On Decisions & Values



Units 5E/501/8E/9W/8ITCU

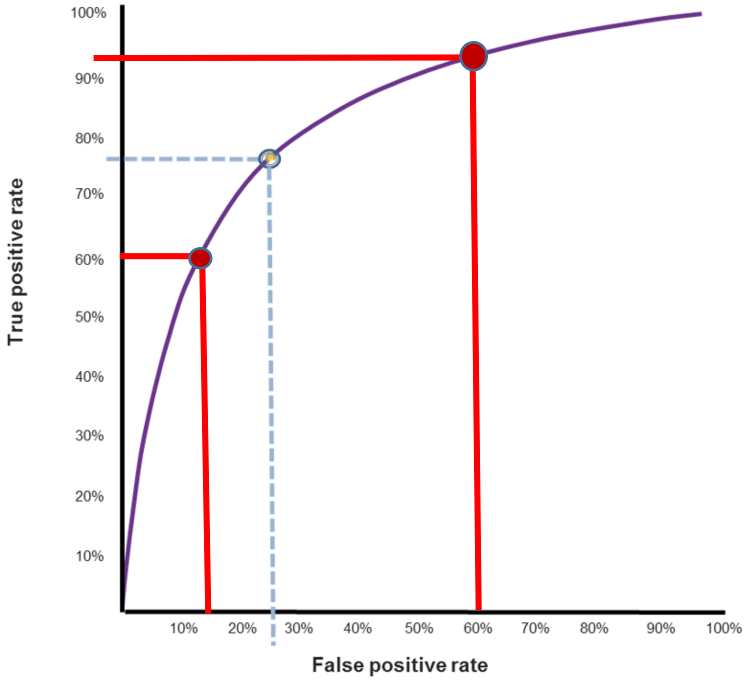
Baseline:
Discharges to home/ home health between 10/15/2011 – 4/29/2012
Readmissions Rate (all cases): **13%**
Score ≥ 25: **27%**
Average direct cost/readmission: **$10,888**

	Initial Pilot 4/30/2012 – 7/30/2012	1 Month Post engagement 9/01/2012 – 9/30/2012
Readmissions Rate	12%	10%
Score ≥ 25	23%	20%
# of Admissions Avoided	9	11
Follow up call completion	52%	61%
Follow up call not Completed	32%	21%
Total Annualized savings	$391,968	$1,448,104

↓ Total Readmission Rate by 3% and +$1.4M Savings

Decisions

Use



On Decisions & Values

Units 5E/501/8E/9W/8ITCU

Baseline:
Discharges to home/ home health between 10/15/2011 – 4/29/2012
Readmissions Rate (all cases): **13%**
Score ≥ 25: **27%**
Average direct cost/readmission: **$10,888**

	Initial Pilot 4/30/2012 – 7/30/2012	1 Month Post engagement 9/01/2012 – 9/30/2012
Readmissions Rate	12%	10%
Score ≥ 25	23%	20%
# of Admissions Avoided	9	11
Follow up call completion	52%	61%
Follow up call not Completed	32%	21%
Total Annualized savings	$391,968	$1,448,104

↓ Total Readmission Rate by 3% and +$1.4M Savings

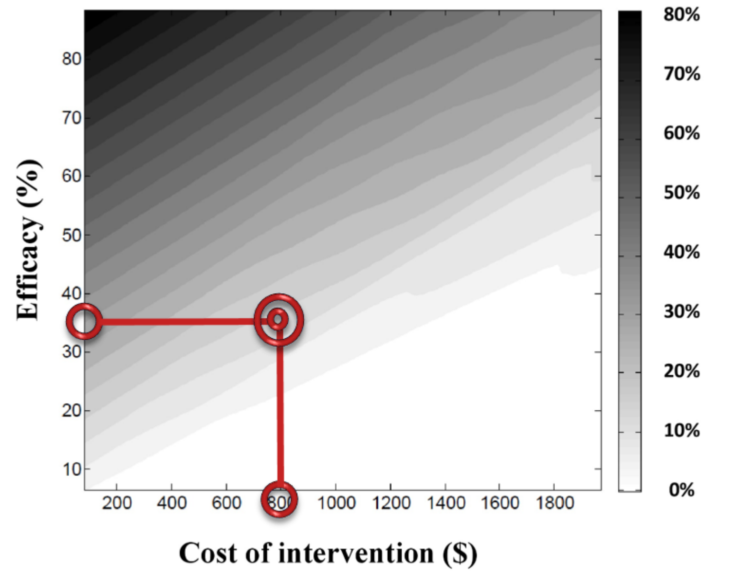
Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study

Mohsen Bayati, Mark Braverman, Michael Gillam, Karen M. Mack, George Ruiz, Mark S. Smith, Eric Horvitz

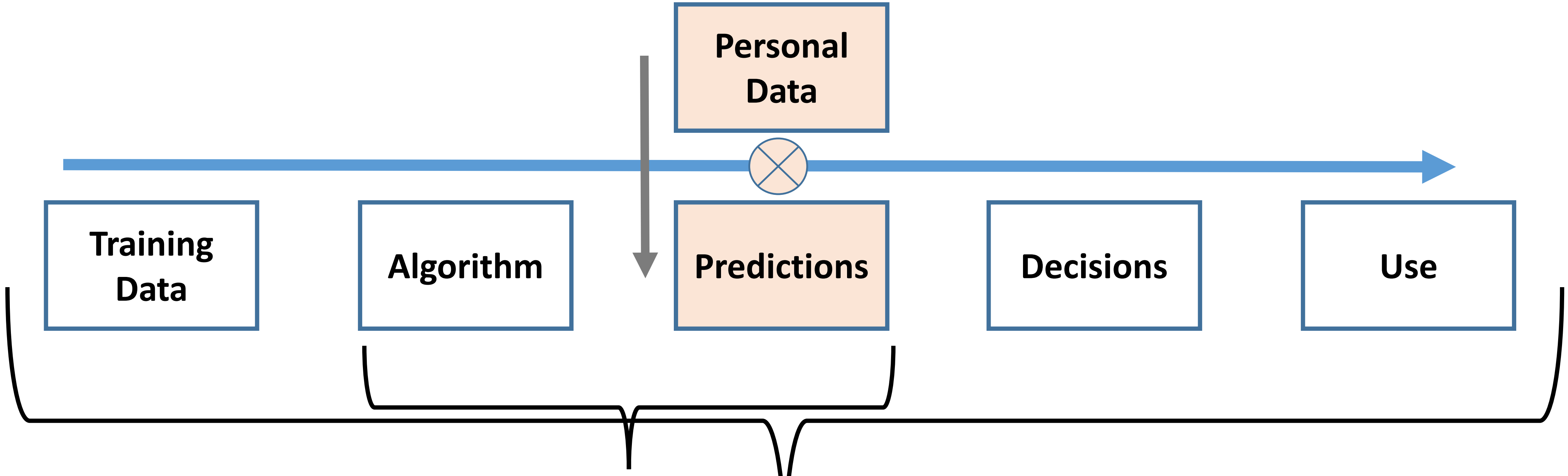
Decisions → **Use**

$800 intervention @ 35% efficacy?

↓31.4% readmissions ↓$13.2%.



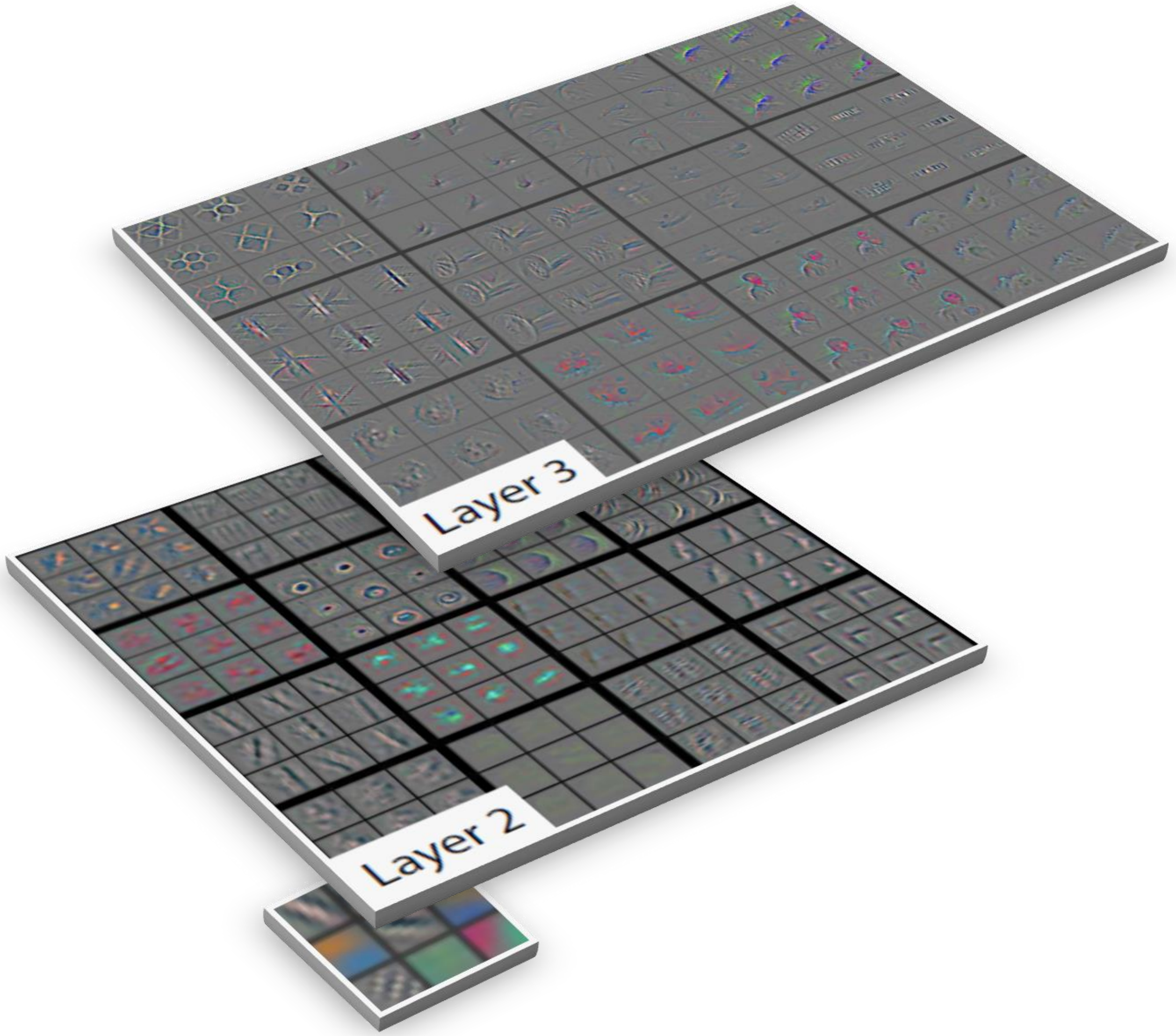
Narrowing Focus

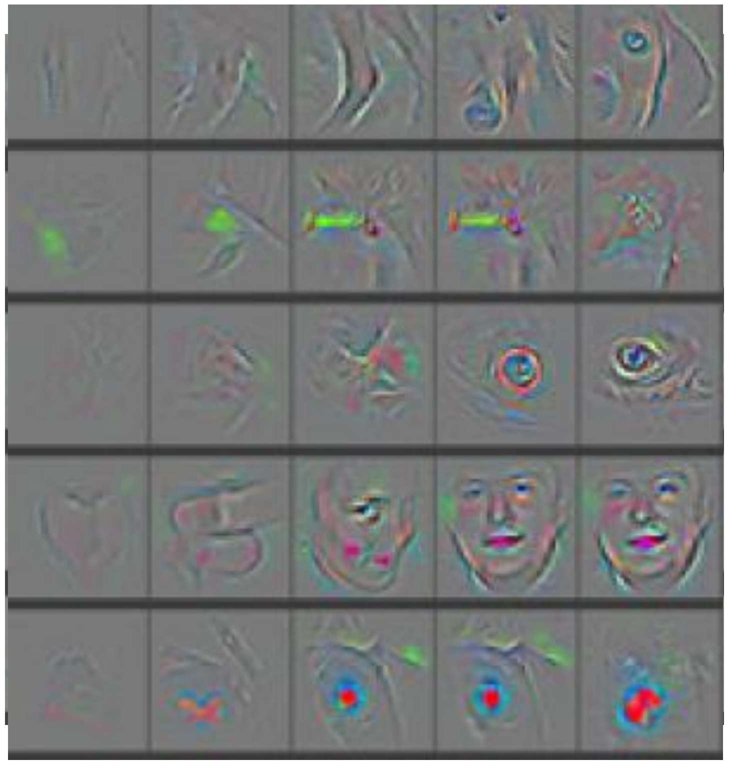


Interpretability & explainability

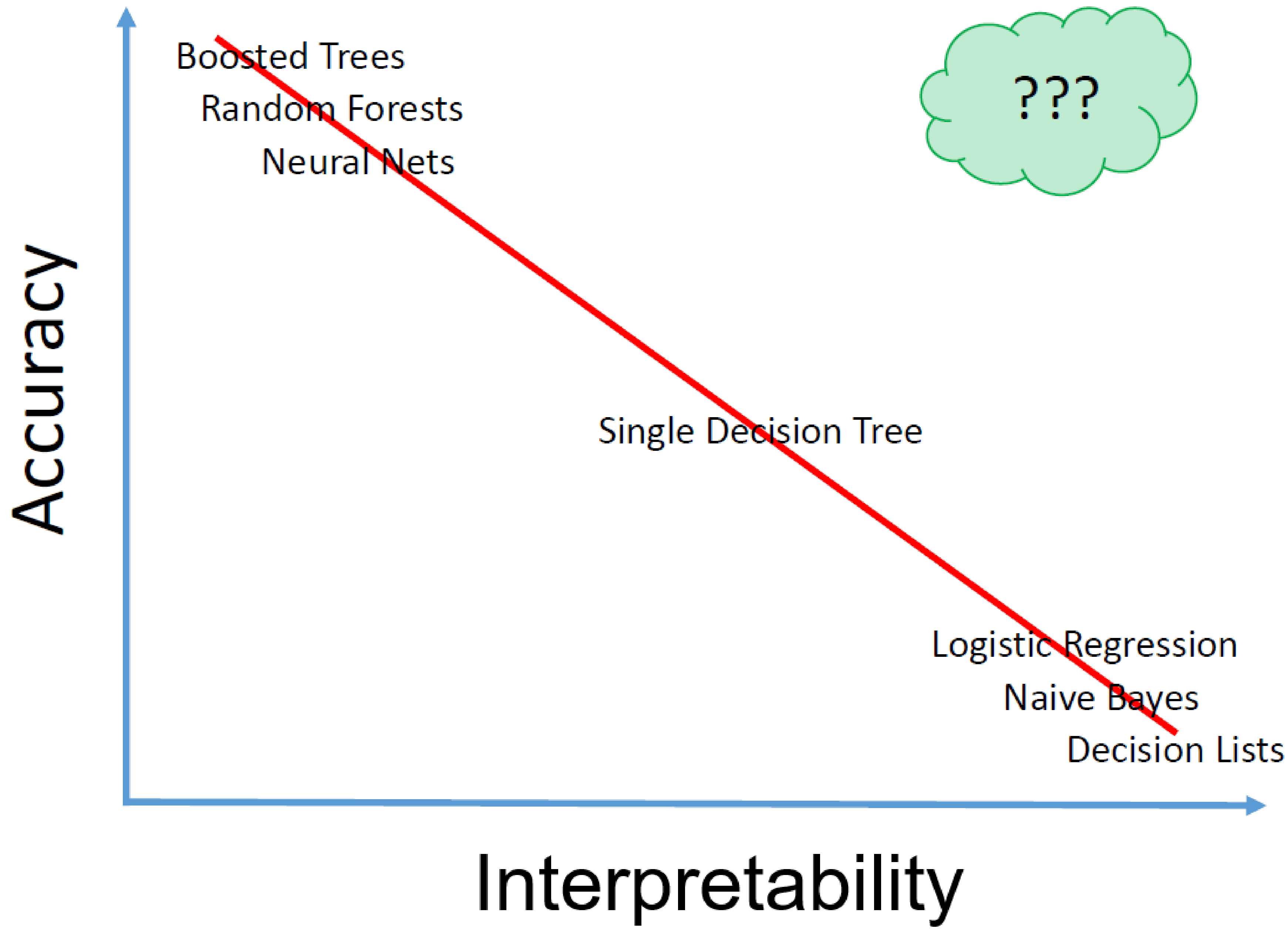
meaningful information about the logic involved

Meaningful Information about ML Procedure?





M. Zeiler, R. Fergus Visualizing and Understanding Convolutional Networks, Arxiv (2013)


Boosted Trees
Random Forests
Neural Nets
???
Single Decision Tree
Logistic Regression
Naive Bayes
Decision Lists
Accuracy
Interpretability

Interpretability & Explanation of Machine Learning

Rich & open-area of research

One approach: Enable end users to understand contribution of individual features

What influence does changing observations x have if other values are not changed?

Interpretability-Power Tradeoff

$$y = f(x_1, \ldots, x_n)$$

Neural networks

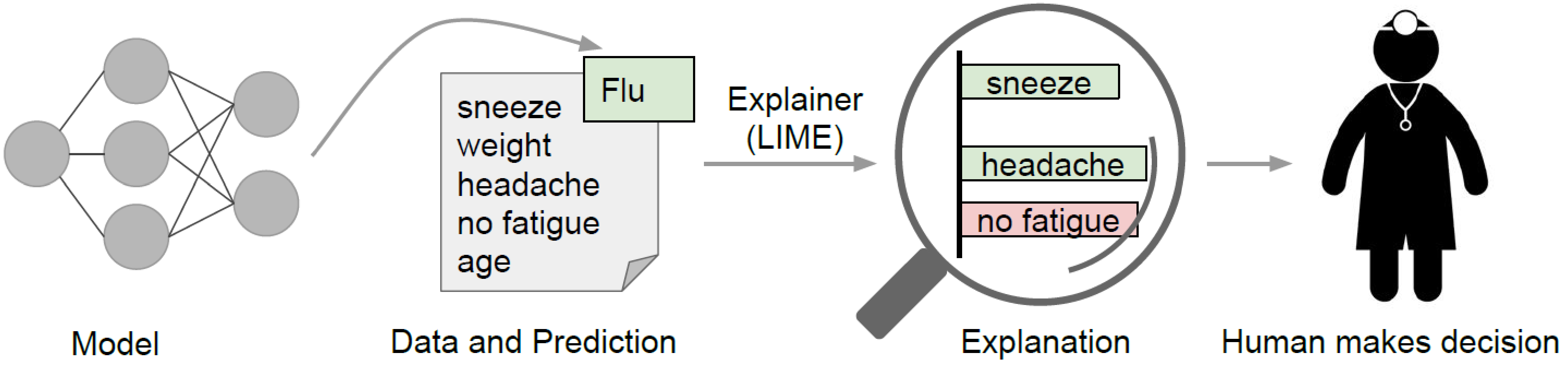
$$y = f_1(x_1) + \ldots + f_n(x_n)$$

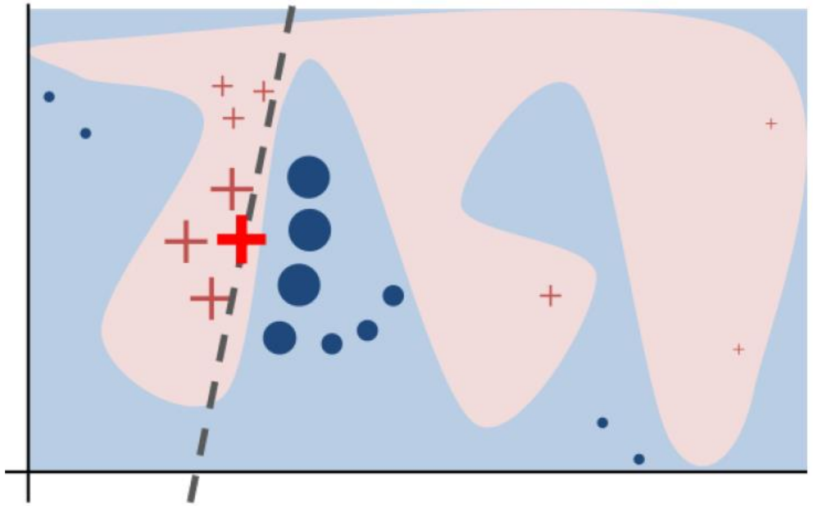
Promising space

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

Logistic regression

Interpretability & Explanation





"Why Should I Trust You?"
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

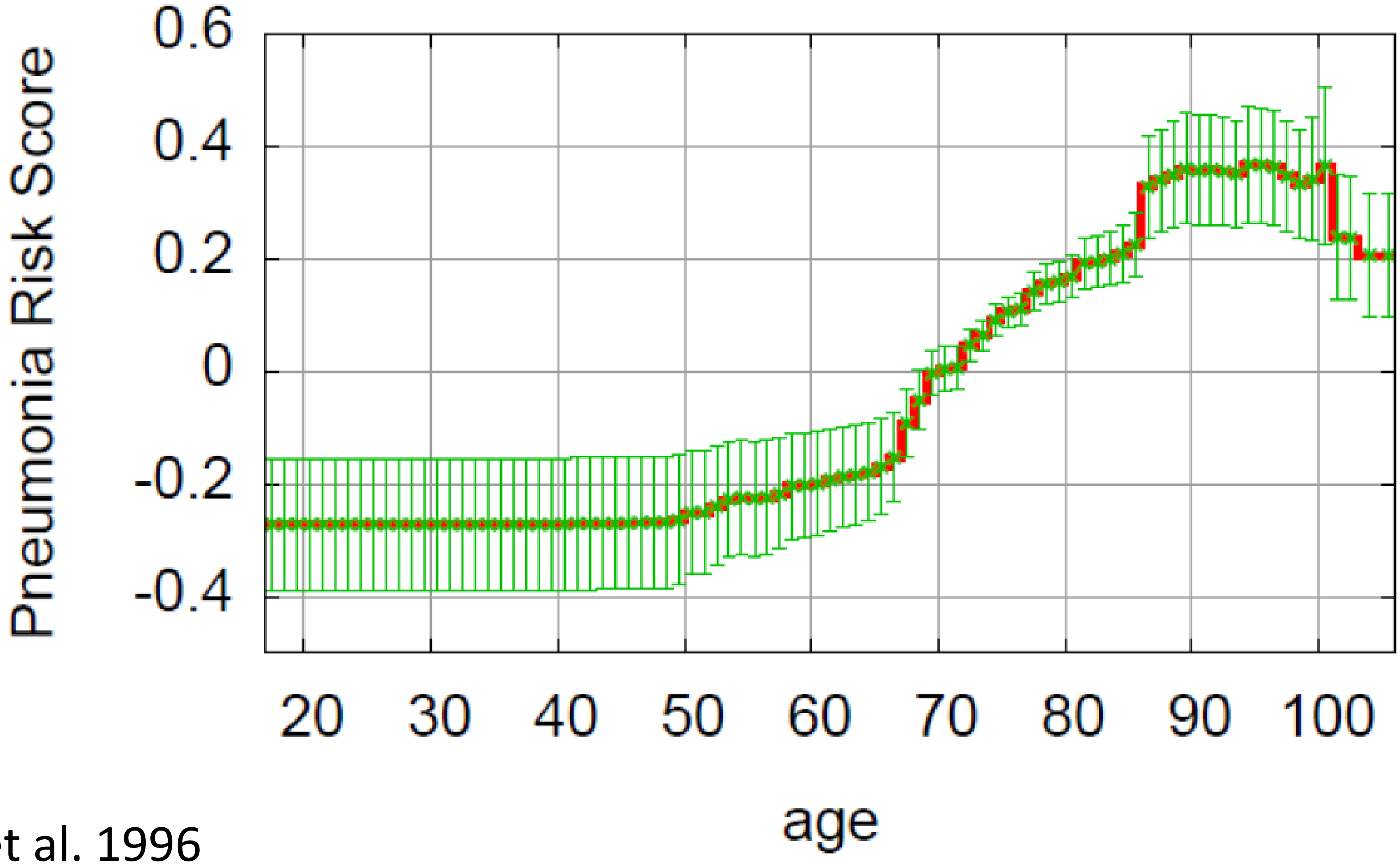
Insights—and Errors

An evaluation of machine-learning methods for predicting pneumonia mortality

Gregory F. Cooper[a,*], Constantin F. Aliferis[a], Richard Ambrosino[a], John Aronis[b], Bruce G. Buchanan[b], Richard Caruana[c], Michael J. Fine[d], Clark Glymour[e], Geoffrey Gordon[c], Barbara H. Hanusa[d], Janine E. Janosky[f], Christopher Meek[e], Tom Mitchell[c], Thomas Richardson[e], Peter Spirtes[e]

AI Journal 1996

Inspection & Troubleshooting



Cooper, et al. 1996

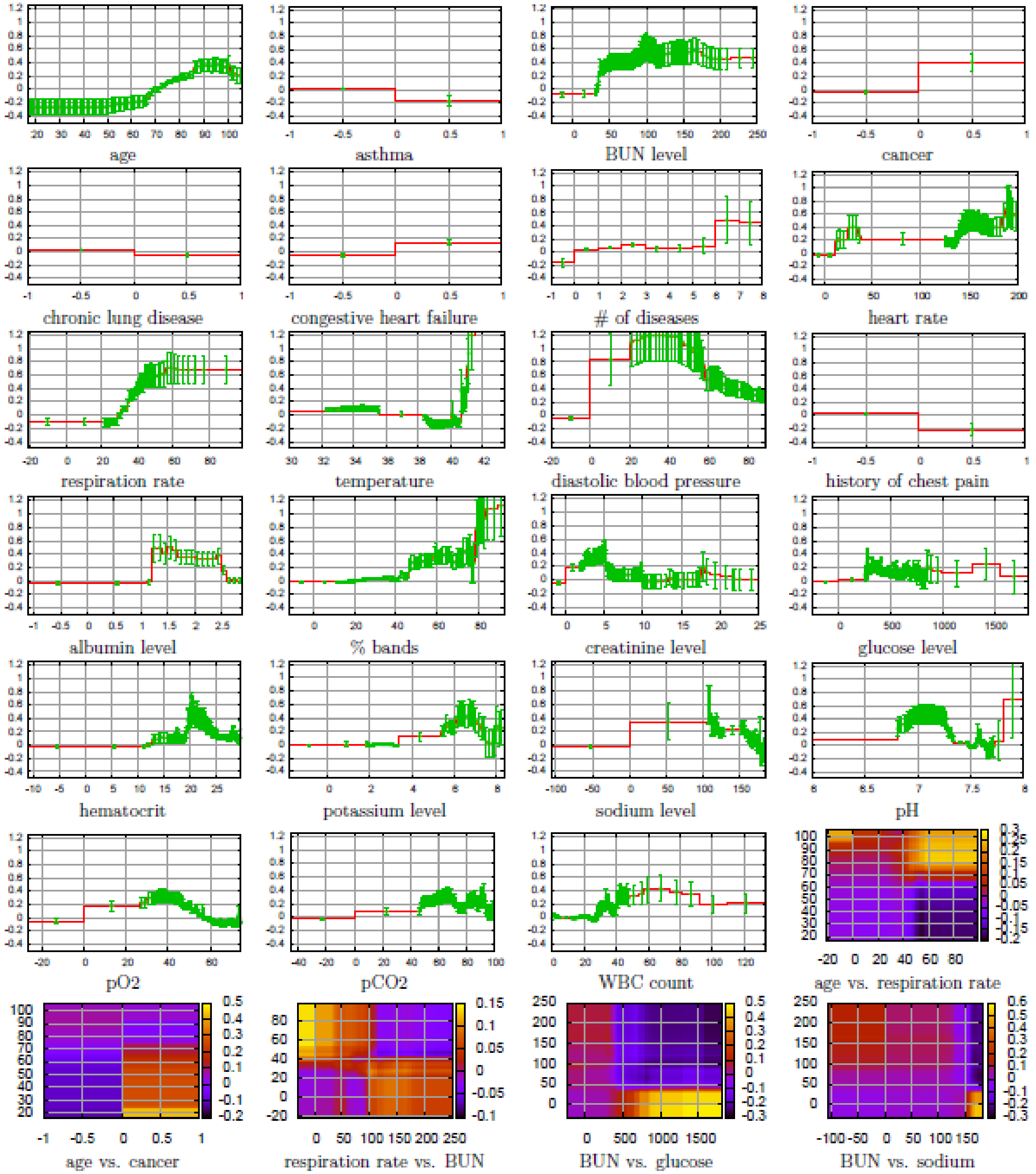
Inspection & Troubleshooting

HasAsthma(x) => LessRisk(x) (!)

True pattern in data:

- asthmatics presenting with pneumonia considered very high risk
- receive agressive treatment and often admitted to ICU
- history of asthma means often get to healthcare sooner
- treatment lowers risk of death compared to general population
- if we use model for admission decision, could hurt asthmatics

Having our Cake and…



Caruana, et al. 2015

Directions

- Research on understandability & explanation
- Best practices, norms, standards on comprehension
- What aspects of AI pipelines?
- How much detail is “meaningful” understanding?
- For whom? when? What uses & contexts?