Bias and Fairness in ML

Hector Corrada Bravo
CMSC 643

Many Cars Tone Deaf To Women's Voices

Female voices pose a bigger challenge for voice-activated technology than men's voices

http://www.autoblog.com/2011/05/31/women-voice-command-systems/

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

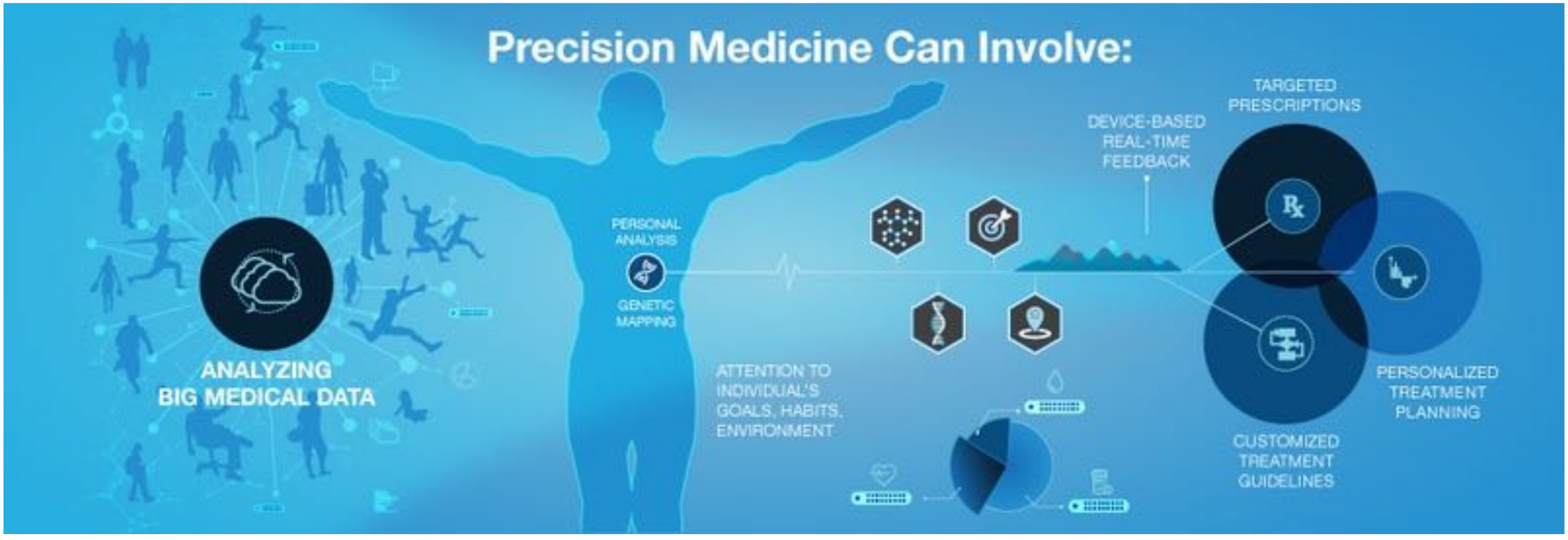
Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Precision Medicine



American Heart Association
https://professional.heart.org/idc/groups/ahamah-public/@wcm/@sop/@rsch/documents/image/ucm_486549.jpg

Precision Medicine

- Inherently a prediction problem
  - From genotype to risk
  - From genotype to therapeutic response

DNA Sequence Variation in a Gene Can Change the Protein Produced by the Genetic Code

Gene A from Person 1

GCA AGA GAT AAT TGT...

Ala (1) Arg (2) Asp (3) Asn (4) Cys (5) ...

Protein Products

Gene A from Person 2

Codon change made no difference in amino acid sequence

GCG AGA GAT AAT TGT...

Ala (1) Arg (2) Asp (3) Asn (4) Cys (5) ...

Gene A from Person 3

Codon change resulted in a different amino acid at position 2

GCA AAA GAT AAT TGT...

Ala (1) Lys (2) Asp (3) Asn (4) Cys (5) ...

OR

Y-GA 98-649

Health or Disease?

DNA Sequence

Person 1 A A A T T T

Normal protein

Some DNA variations have no negative effects

Person 2 A A T T T T

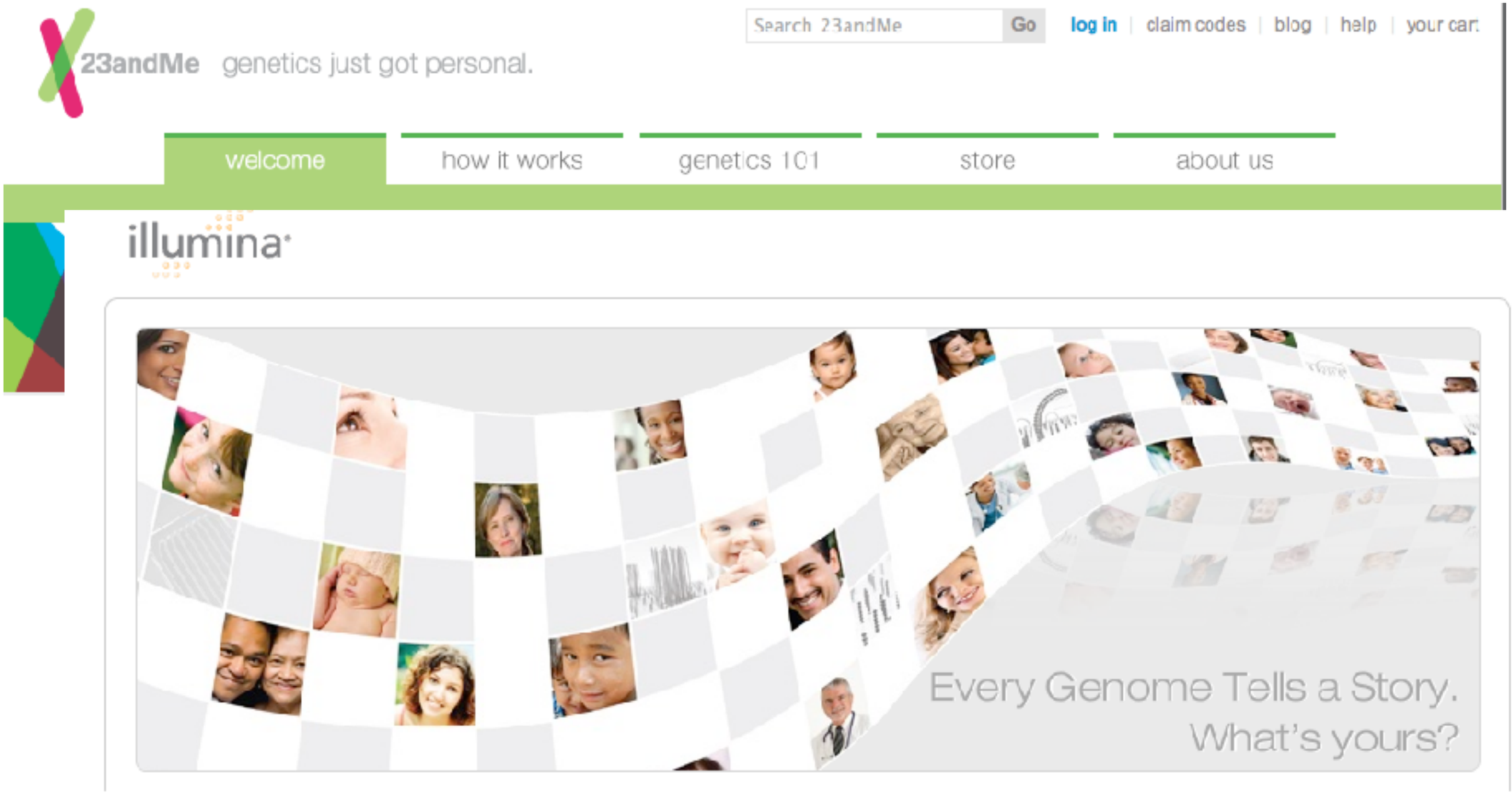
Low or nonfunctioning protein

Person 3 A A C T T T

Other variations lead to disease (e.g., sickle cell) or increased susceptibility to disease (e.g., lung cancer)

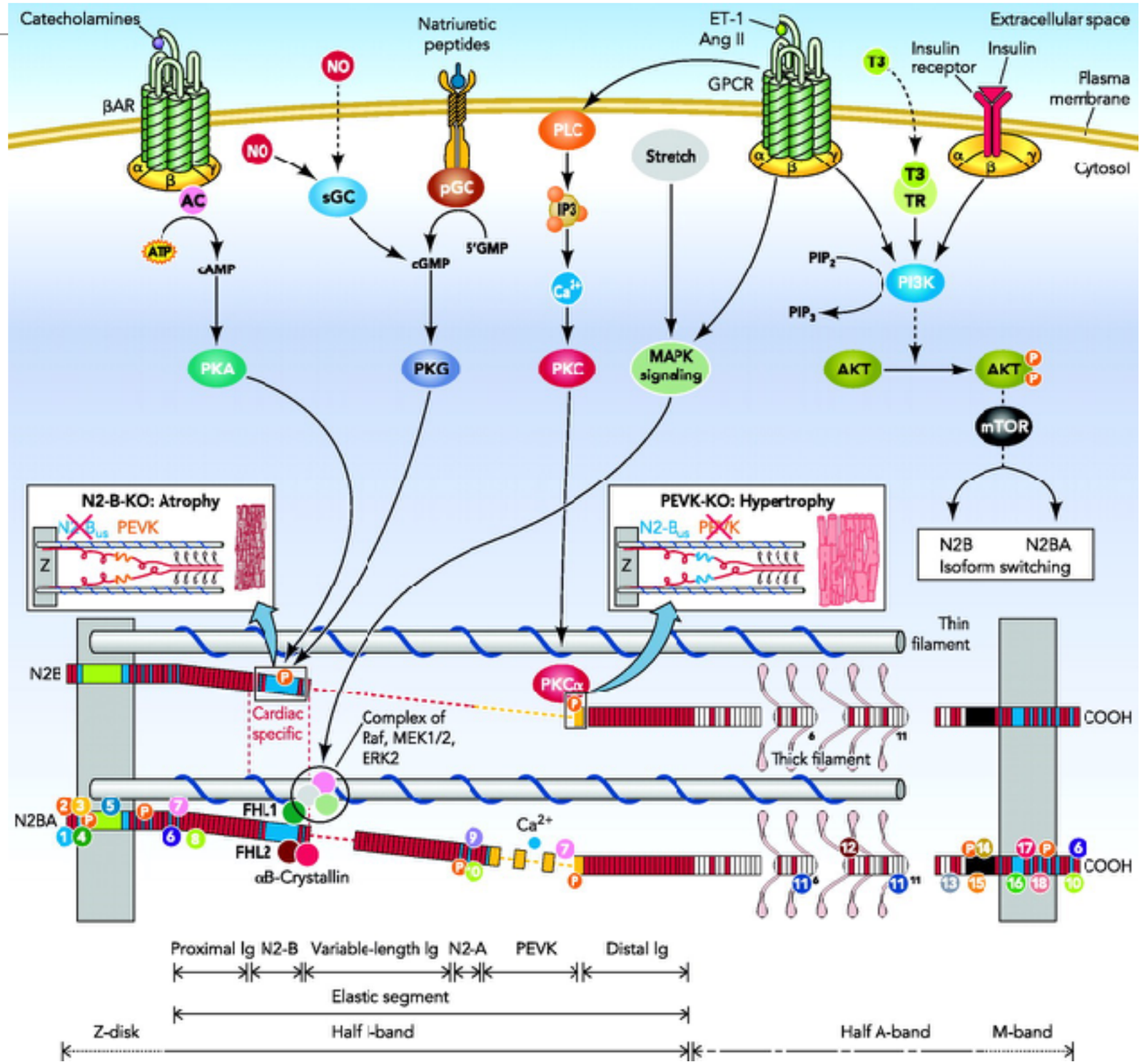
YGG-00-0480

Personal Genomics



Approaches

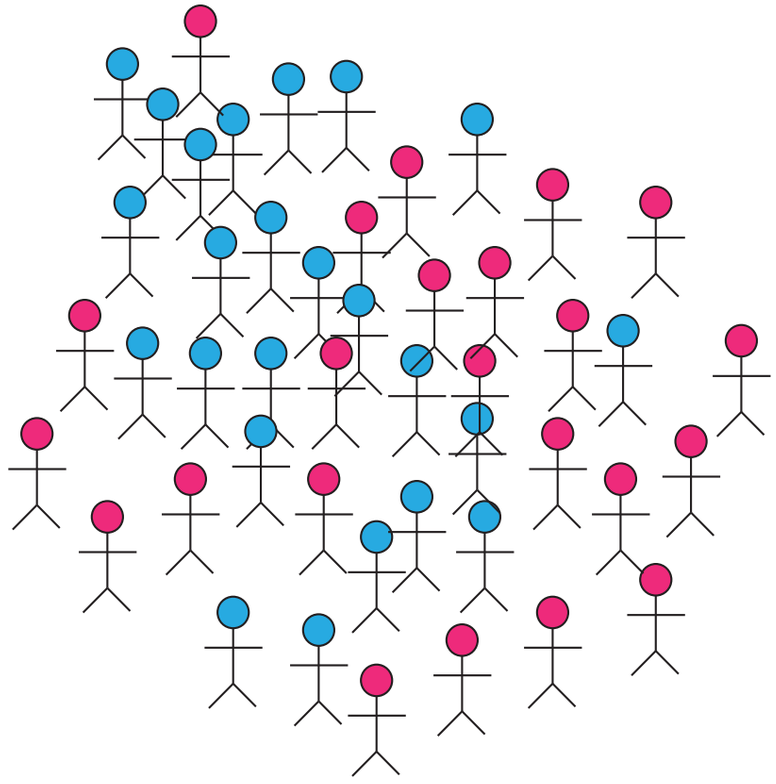
- Mechanistic models

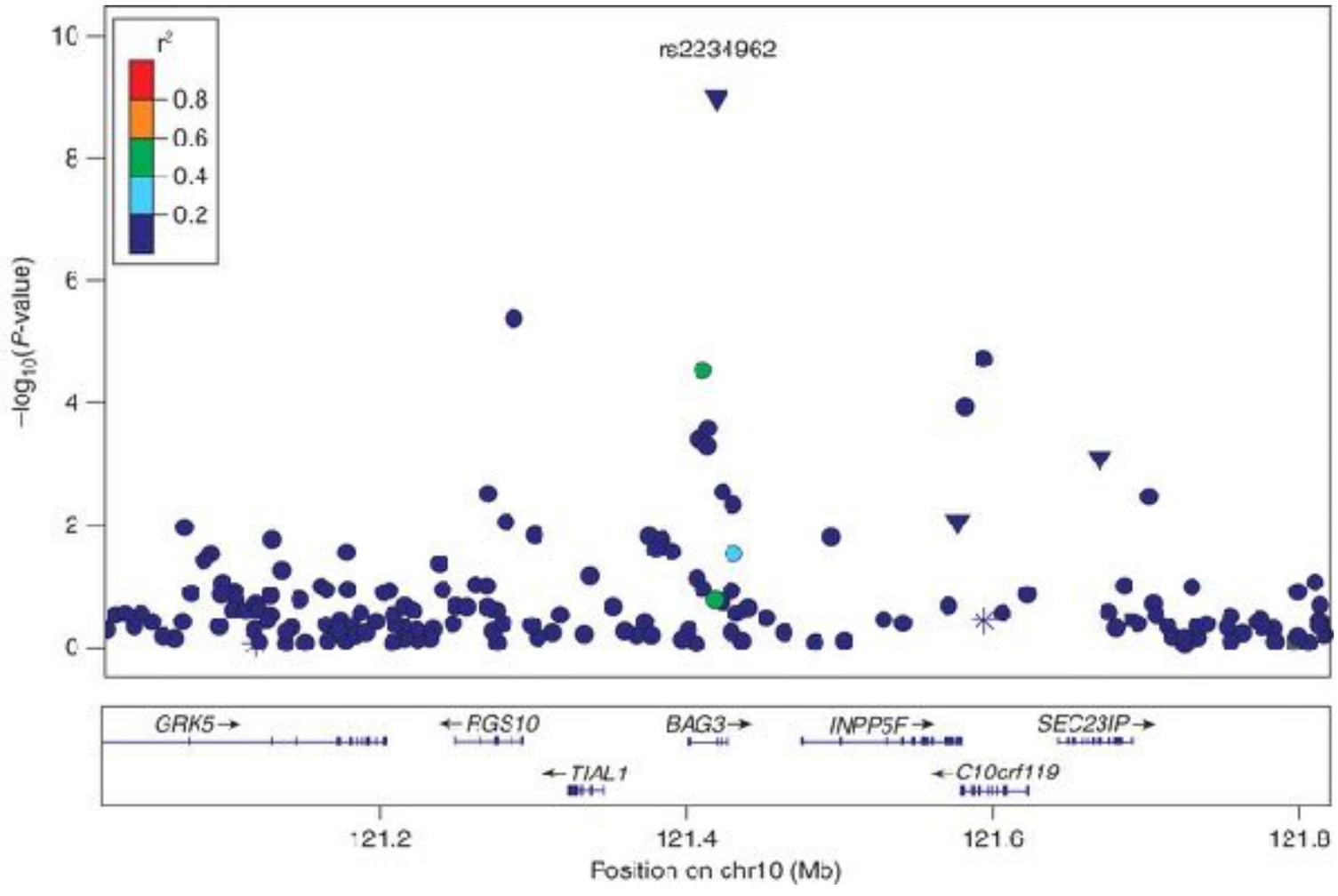


Linke and Kruger, 2010
http://physiologyonline.physiology.org/content/25/3/186

Approaches

- Observational association modeling



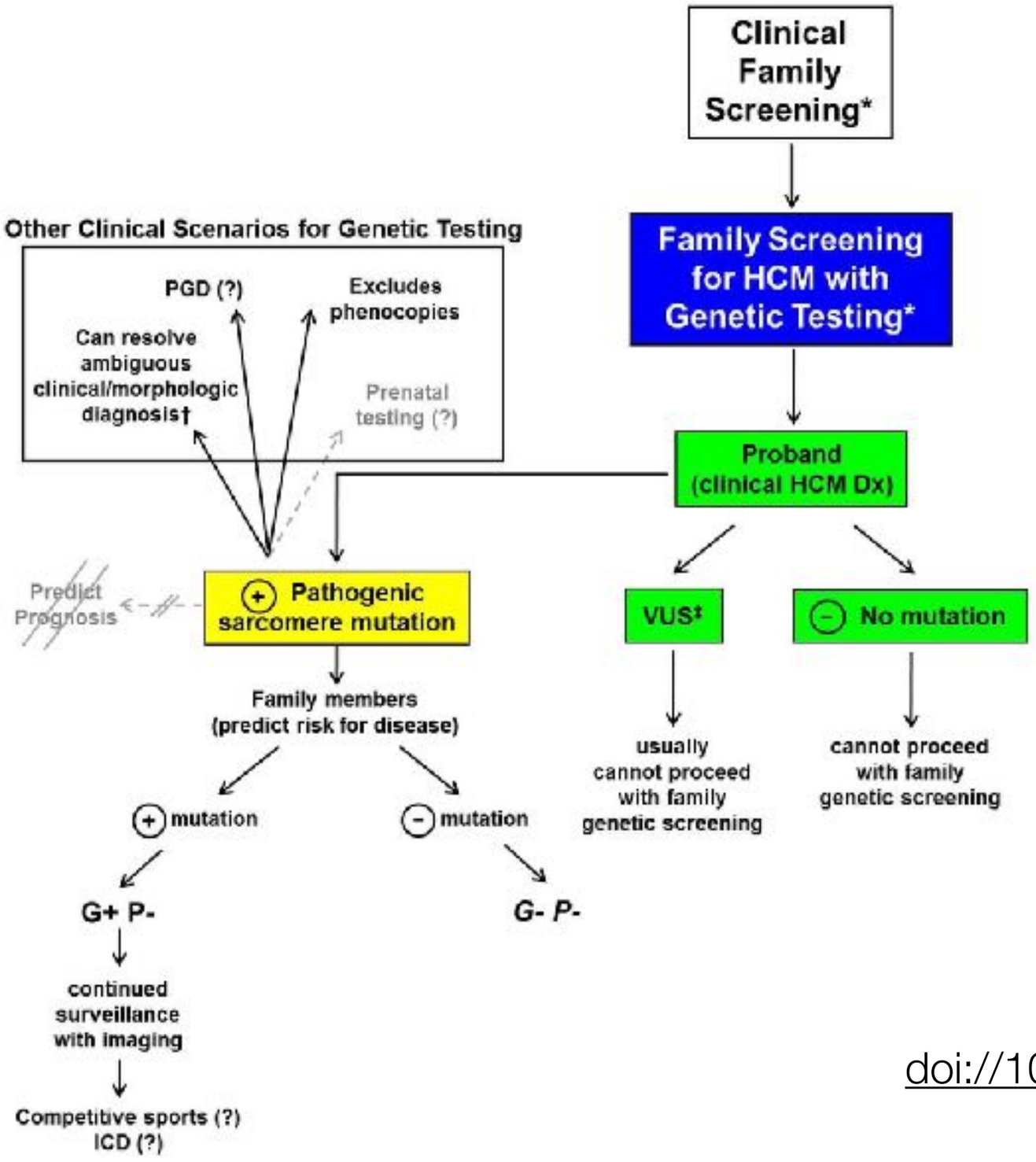


hypertrophic cardiomyopathy

control

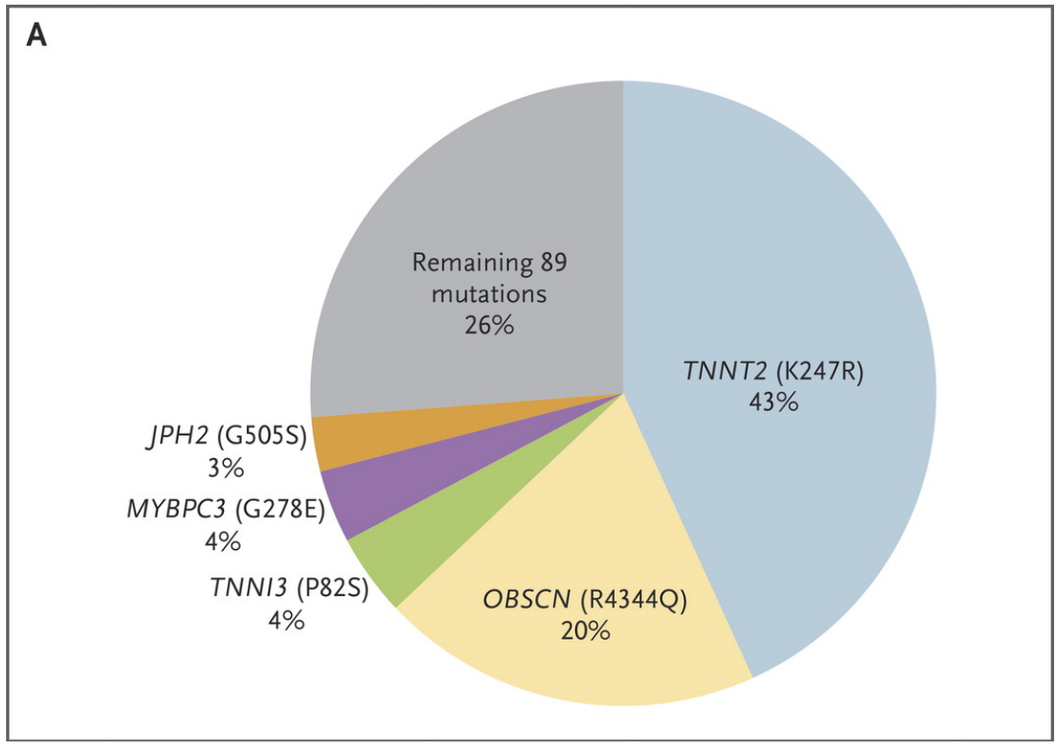
Villard, et al., 2011
http://eurheartj.oxfordjournals.org/content/32/9/1065

From association study to risk prediction



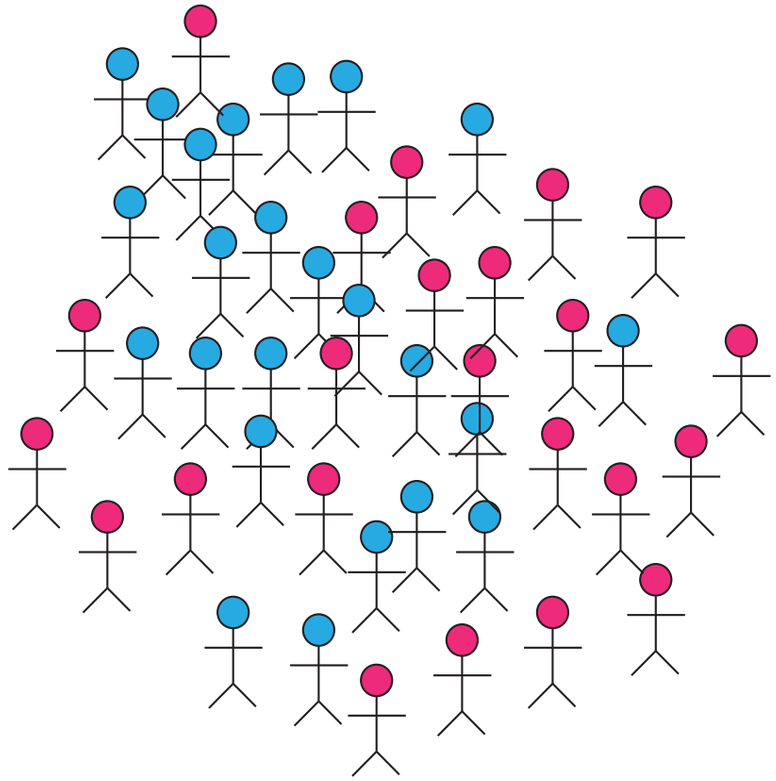
Maron, et al., 2011
doi://10.1016/j.jacc.2012.02.068

How can this lead to health disparities?



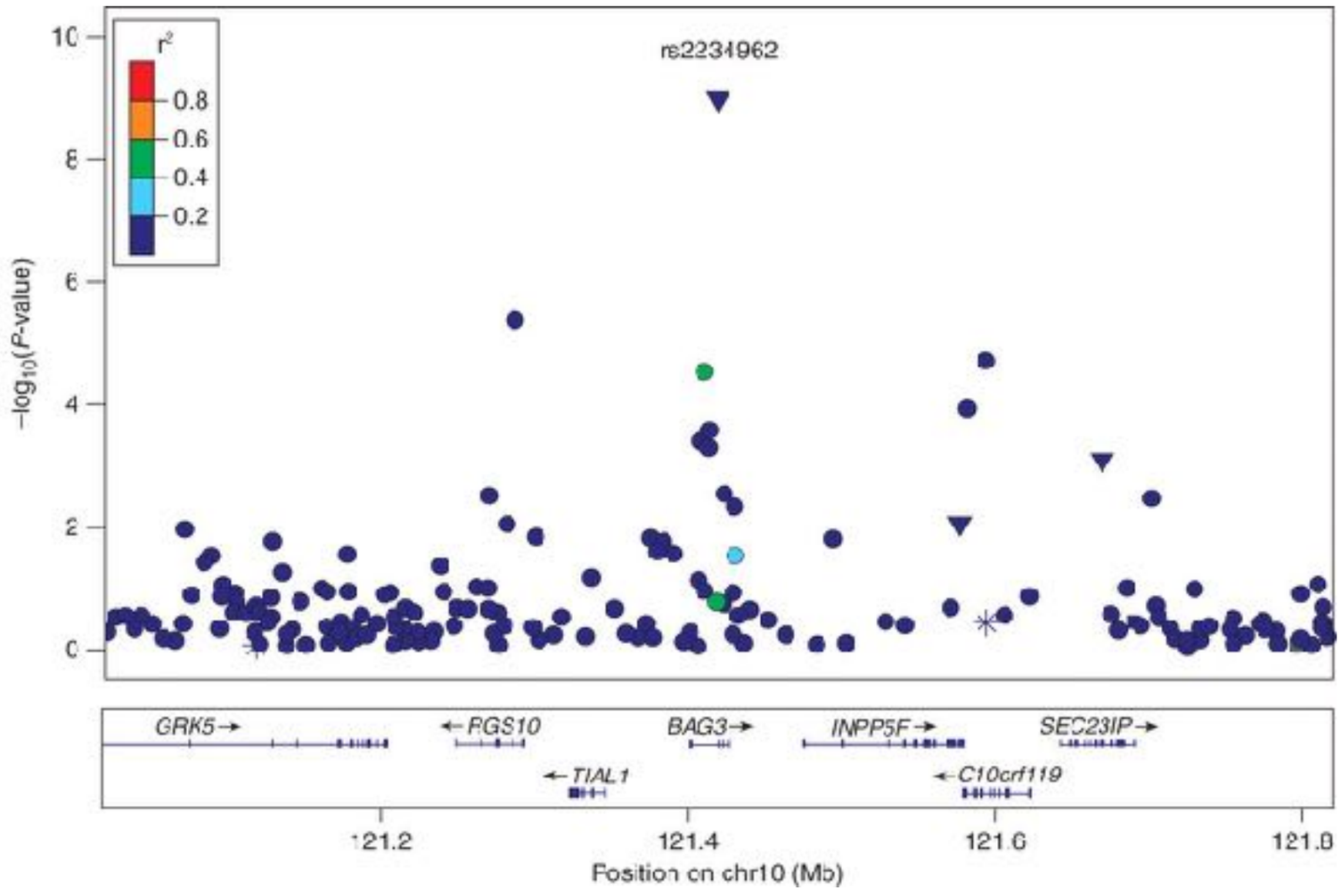
Approaches

- Observational association modeling



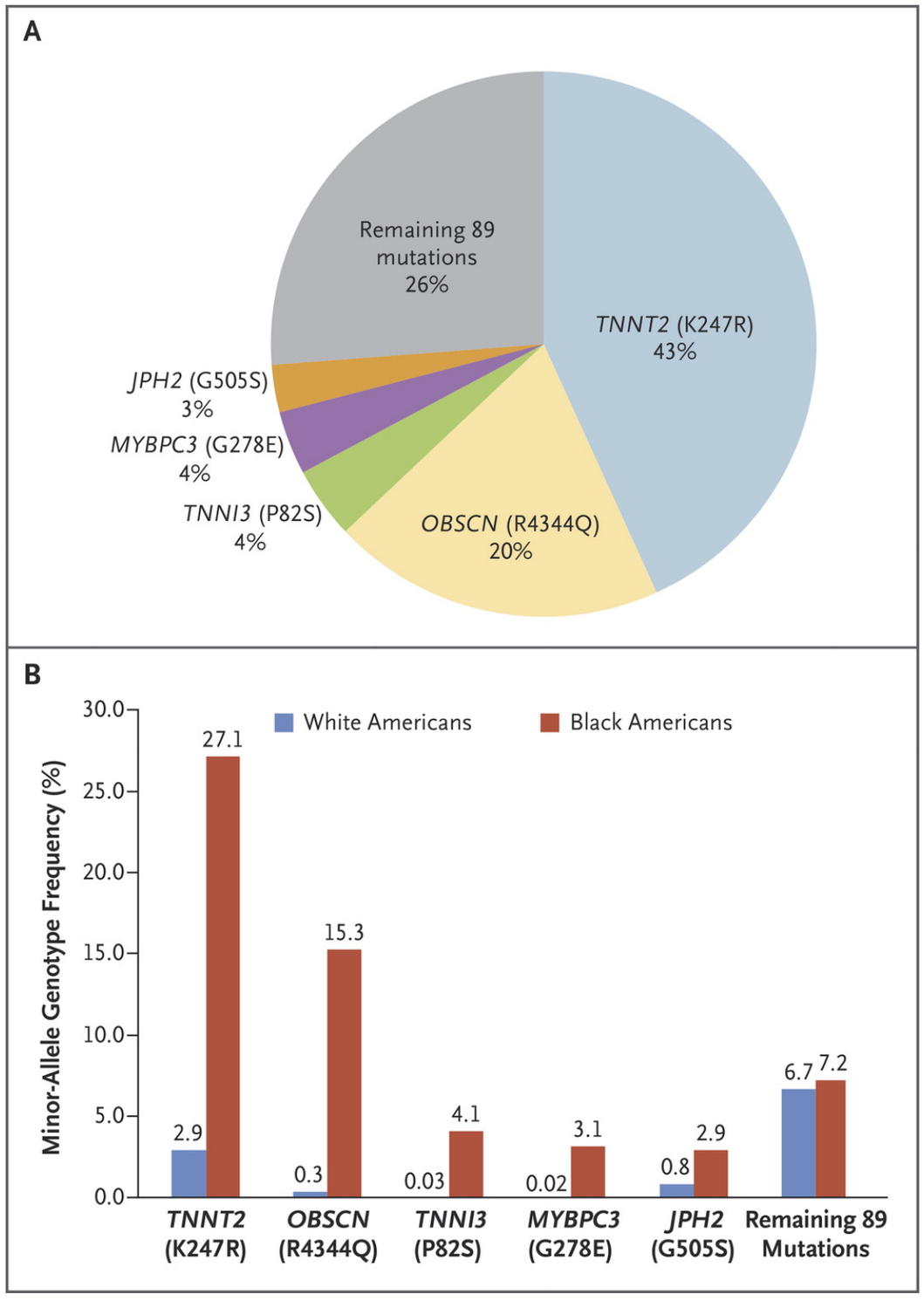
hypertrophic cardiomyopathy

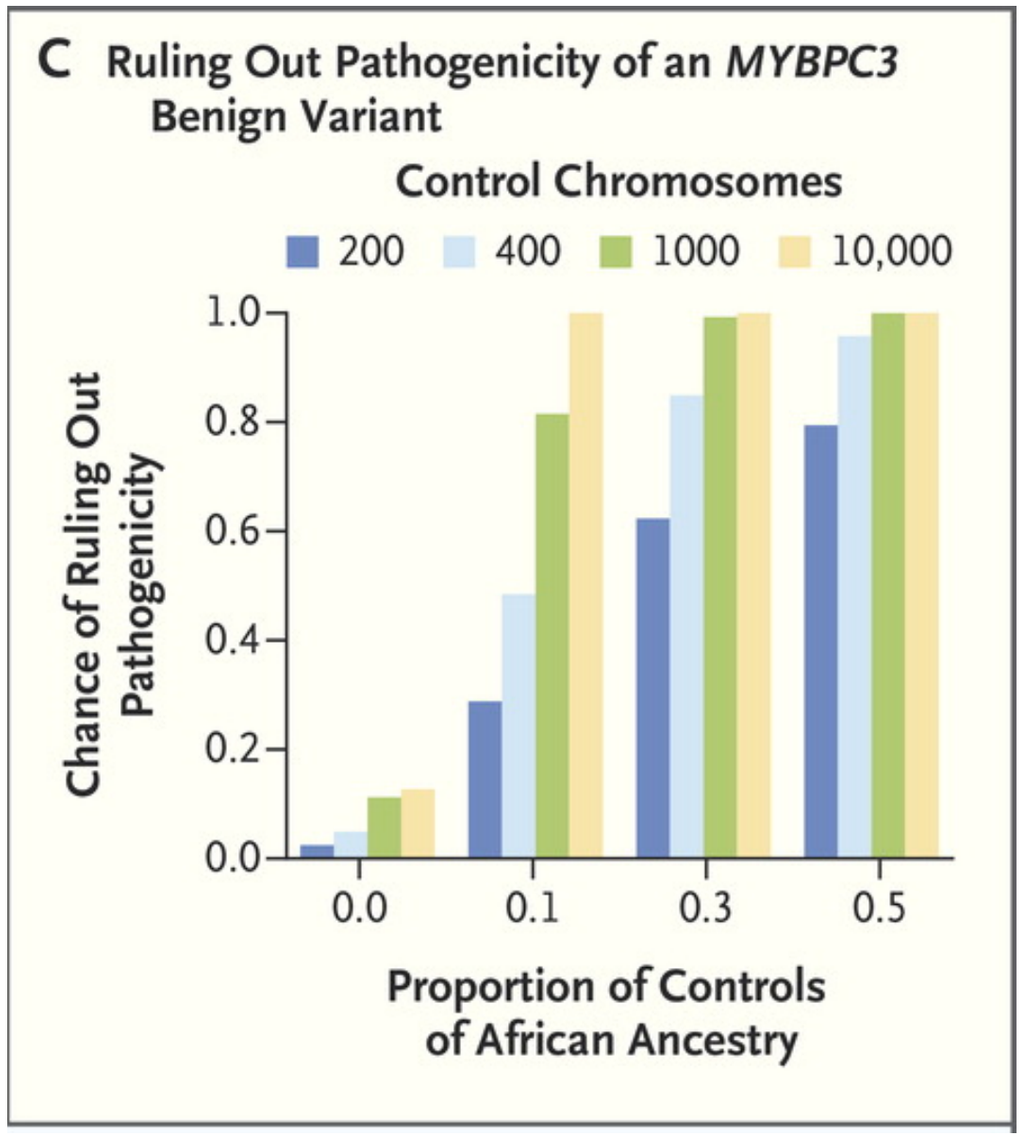
control



Villard, et al., 2011
http://eurheartj.oxfordjournals.org/content/32/9/1065

How can this lead to health disparities?





Recall: Formal Definition of Binary Classification (from CIML)

TASK: BINARY CLASSIFICATION

Given:

1. An input space $\mathcal{X}$
2. An unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$

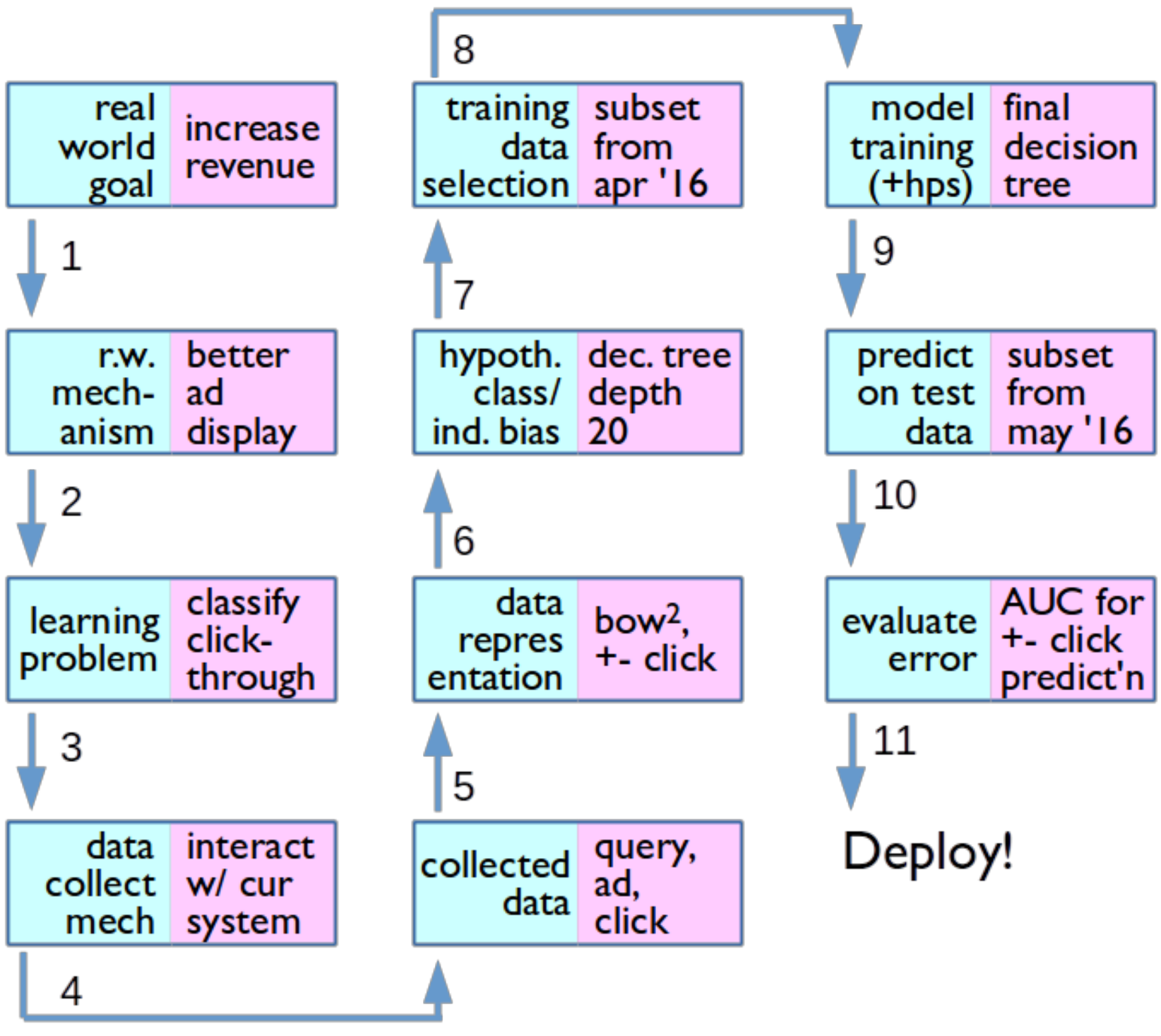
Compute: A function f minimizing: $\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[f(\boldsymbol{x}) \neq y\right]$

Train/Test Mismatch

- When working with real world data, training sample
  - reflects human biases
  - is influenced by practical concerns
    - e.g., what kind of data is easy to obtain

- Train/test distribution mismatch is frequent issue
  - aka covariate shift, sample selection bias, domain adaptation

Typical Design Process for an ML Application



Bias is pervasive

- Bias in the labeling
- Sample selection bias
- Bias in choice of labels
- Bias in features or model structure
- Bias in loss function
- Deployed systems create feedback loops

Data collection

- What data should (not) be collected
- Who owns the data
- Whose data can (not) be shared
- What technology for collecting, storing, managing data
- Whose data can (not) be traded
- What data can (not) be merged
- What to do with prejudicial data

[Fung, 2016]

Data Modeling

- Data is biased (known/unknown)
  - Invalid assumptions
  - Confirmation bias
  - Publication bias
  - Badly handling missing values

[Fung, 2016]

Deployment

- Spurious correlation / over-generalization
- Using “black-box” methods that cannot be explained
- Using heuristics that are not well understood
- Releasing untested code
- Extrapolating
- Not measuring lifecycle performance (concept drift in ML)

[Fung, 2016]

ACM Code of Ethics

"To minimize the possibility of indirectly harming others, computing professionals must minimize malfunctions by following generally accepted standards for system design and testing. Furthermore, it is often necessary to assess the social consequences of systems to project the likelihood of any serious harm to others. If system features are misrepresented to users, coworkers, or supervisors, the individual computing professional is responsible for any resulting injury."

https://www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct

Data Science guiding principles

- Start with clear user need and public benefit
- Use data and tools which have minimum intrusion necessary
- **Create robust data science models**
- Be alert to public perceptions
- Be as open and accountable as possible
- Keep data secure

[UK cabinet office]

Domain Adaptation

- What does it mean for 2 distributions to be related?
- When 2 distributions are related how can we build models that effectively share information between them?

Unsupervised adaptation

- **Goal:** learn a classifier f that achieves low expected loss under new distribution $\mathcal{D}^{new}$

- Given labeled training data from old distribution $\mathcal{D}^{\text{old}}$ $(x_1, y_1), \ldots, (x_N, y_N)$

- And unlabeled examples from new distribution $\mathcal{D}^{\text{new}}$: $z_1, \ldots, z_M$

Relation between test loss in new domain and old domain

$$\text{test loss} \tag{8.1}$$

$$= \mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{new}}}\left[\ell(y, f(\boldsymbol{x}))\right] \qquad \text{definition} \tag{8.2}$$

$$= \sum_{(x,y)} \mathcal{D}^{\text{new}}(\boldsymbol{x}, y)\ell(y, f(\boldsymbol{x})) \qquad \text{expand expectation} \tag{8.3}$$

$$= \sum_{(x,y)} \mathcal{D}^{\text{new}}(\boldsymbol{x}, y)\frac{\mathcal{D}^{\text{old}}(\boldsymbol{x}, y)}{\mathcal{D}^{\text{old}}(\boldsymbol{x}, y)}\ell(y, f(\boldsymbol{x})) \qquad \text{times one} \tag{8.4}$$

$$= \sum_{(x,y)} \mathcal{D}^{\text{old}}(\boldsymbol{x}, y)\frac{\mathcal{D}^{\text{new}}(\boldsymbol{x}, y)}{\mathcal{D}^{\text{old}}(\boldsymbol{x}, y)}\ell(y, f(\boldsymbol{x})) \qquad \text{rearrange} \tag{8.5}$$

$$= \mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{old}}}\left[\frac{\mathcal{D}^{\text{new}}(\boldsymbol{x}, y)}{\mathcal{D}^{\text{old}}(\boldsymbol{x}, y)}\ell(y, f(\boldsymbol{x}))\right] \qquad \text{definition} \tag{8.6}$$

How can we estimate the ratio between Dnew and Dold?

Fixed base distribution

S = selection variable

$$\frac{\mathcal{D}^{\text{new}}(\boldsymbol{x},y)}{\mathcal{D}^{\text{old}}(\boldsymbol{x},y)} = \frac{\frac{1}{Z^{\text{new}}}\mathcal{D}^{\text{base}}(\boldsymbol{x},y)p(s=0\mid\boldsymbol{x})}{\frac{1}{Z^{\text{old}}}\mathcal{D}^{\text{base}}(\boldsymbol{x},y)p(s=1\mid\boldsymbol{x})} \quad \text{definition} \quad (8.9)$$

$$= \frac{\frac{1}{Z^{\text{new}}}p(s=0\mid\boldsymbol{x})}{\frac{1}{Z^{\text{old}}}p(s=1\mid\boldsymbol{x})} \quad \text{cancel base} \quad (8.10)$$

$$= Z\frac{p(s=0\mid\boldsymbol{x})}{p(s=1\mid\boldsymbol{x})} \quad \text{consolidate} \quad (8.11)$$

$$= Z\frac{1-p(s=1\mid\boldsymbol{x})}{p(s=1\mid\boldsymbol{x})} \quad \text{binary selection} \quad (8.12)$$

$$= Z\left[\frac{1}{p(s=1\mid\boldsymbol{x})}-1\right] \quad \text{rearrange} \quad (8.13)$$

We can estimate P(s=1|x) using a binary classifier!

Algorithm 23 SELECTIONADAPTATION($\langle (\boldsymbol{x}_n, y_n) \rangle_{n=1}^{N}$, $\langle \boldsymbol{z}_m \rangle_{m=1}^{M}$, $\mathcal{A}$)

$D^{dist} \leftarrow \langle (\boldsymbol{x}_n, +1) \rangle_{n=1}^{N} \bigcup \langle (\boldsymbol{z}_m, -1) \rangle_{m=1}^{M}$ // assemble data for distinguishing // between old and new distributions

$\hat{p} \leftarrow$ train logistic regression on D^{dist}

$D^{weighted} \leftarrow \left\langle \left(\boldsymbol{x}_n, y_n, \frac{1}{\hat{p}(\boldsymbol{x}_n)} - 1 \right) \right\rangle_{n=1}^{N}$ // assemble weight classification // data using selector

return $\mathcal{A}(D^{weighted})$ // train classifier

Supervised adaptation

- **Goal:** learn a classifier f that achieves low expected loss under new distribution $\mathcal{D}^{new}$

- Given labeled training data from old distribution $\mathcal{D}^{\text{old}}$ $\left\langle \boldsymbol{x}_n^{(\text{old})}, y_n^{(\text{old})} \right\rangle_{n=1}^{N}$

- And labeled examples from new distribution $\mathcal{D}^{\text{new}}$: $\left\langle \boldsymbol{x}_m^{(\text{new})}, y_m^{(\text{new})} \right\rangle_{m=1}^{M}$

One solution: feature augmentation

- Map inputs to a new augmented representation

shared old-only new-only

$$\boldsymbol{x}_n^{(\text{old})} \mapsto \left\langle \boldsymbol{x}_n^{(\text{old})}, \boldsymbol{x}_n^{(\text{old})}, \underbrace{0, 0, \ldots, 0}_{D\text{-many}} \right\rangle$$

$$\boldsymbol{x}_m^{(\text{new})} \mapsto \left\langle \boldsymbol{x}_m^{(\text{new})}, \underbrace{0, 0, \ldots, 0}_{D\text{-many}}, \boldsymbol{x}_m^{(\text{new})} \right\rangle$$

One solution: feature augmentation

- Transform Dold and Dnew training examples
- Train a classifier on new representations
- Done!

One solution: feature augmentation

- Adding instance weighting might be useful if N >> M

- Most effective when distributions are “not too close but not too far”
  - In practice, always try “old only”, “new only”, “union of old and new” as well!

Theorem 9 (Unsupervised Adaptation Bound). *Given a fixed representation and a fixed hypothesis space $\mathcal{F}$, let $f \in \mathcal{F}$ and let $\epsilon^{(best)} = \min_{f^* \in \mathcal{F}} \frac{1}{2}\left[\epsilon^{(old)}(f^*) + \epsilon^{(new)}(f^*)\right]$, then, for all $f \in \mathcal{F}$:*

$$\underbrace{\epsilon^{(new)}(f)}_{\text{error on } \mathcal{D}^{new}} \leq \underbrace{\epsilon^{(old)}(f)}_{\text{error on } \mathcal{D}^{old}} + \underbrace{\epsilon^{(best)}}_{\text{minimal avg error}} + \underbrace{d_{\mathcal{A}}(\mathcal{D}^{old}, \mathcal{D}^{new})}_{\text{distance}} \tag{8.27}$$