FAT-SG

A Characterization of Monotone Influence Measures for Data Classification

Jakub Sliwinski, Martin Strobel, Yair Zick

NUS National University of Singapore

Some news headlines in the recent months:

Just like humans, artificial intelligence can be sexist and racist

Princeton University study finds machine learning copies human prejudices when learning language

Machine Bias

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

AI programs exhibit racial and gender biases, research reveals

tl;dr

-We want a general use influence measure, not based on domain knowledge

-Not requiring black-box access, just some dataset produced by the classifier

-We state the properties we want, and determine what satisfies them

-We show an experiment

Bob’s loan application was approved.
We want to be able to say this:

Listen Bob:
-It was important that your salary is **high**
-You are **not too old**
-Your height **doesn't matter**
...

Bob’s loan application was approved.
We want to be able to say this:

Listen Bob:

-It was important that your salary is **high**

-You are **not too old**

-Your height **doesn't matter**

…

Capture that numerically:

Salary: **5**

Age: **-1**

Height: **0.05**

…

Bob’s loan application was approved.
We want to be able to say this:

Listen Bob:
-It was important that your salary is **high**
-You are **not too old**
-Your height **doesn't matter**
…

Capture that numerically:

Salary: **5**
Age: **-1**
Height: **0.05**
…

We want a **vector** to express how influential the features were for Bob’s classification - an **influence measure**.

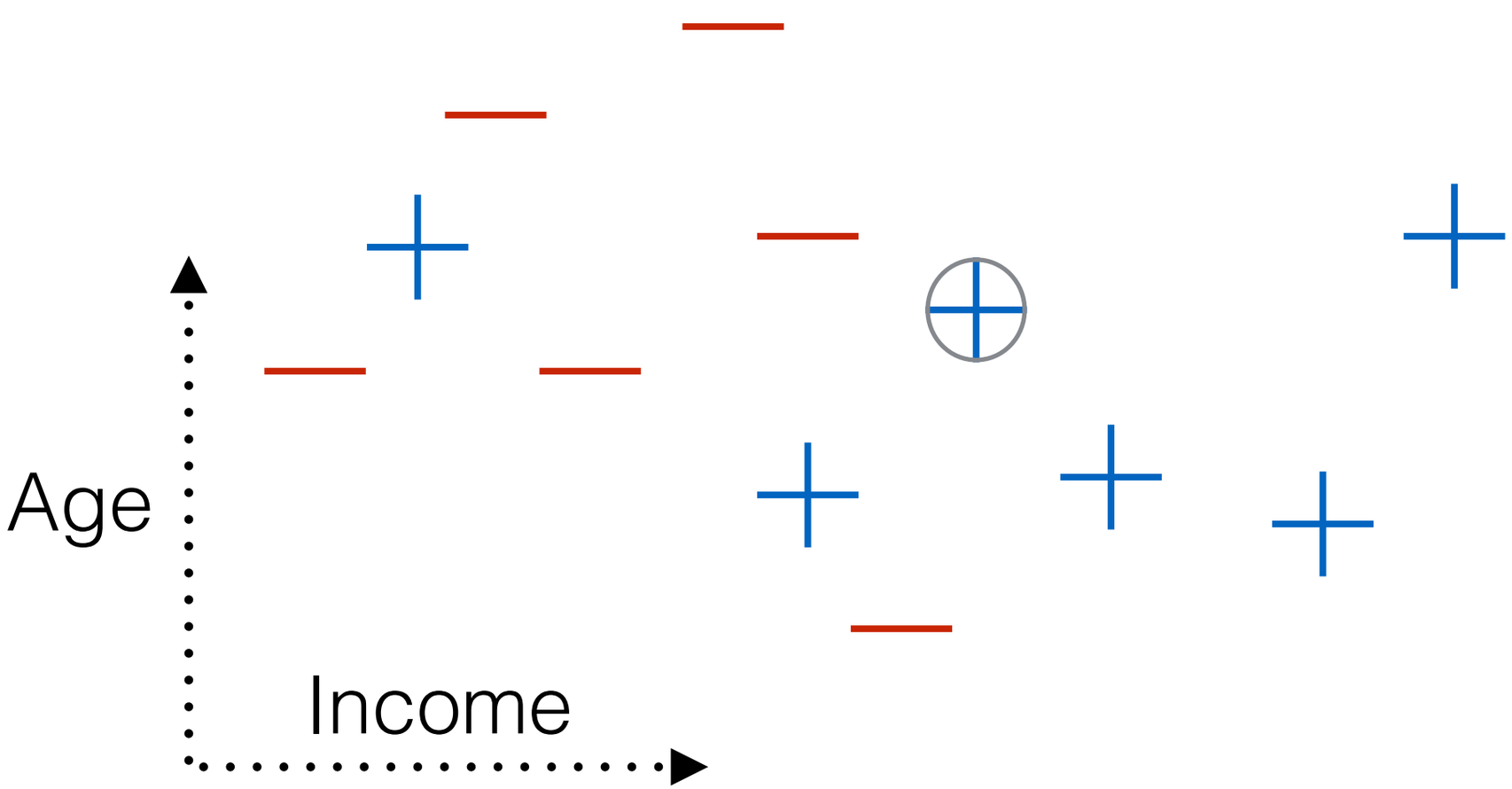
What we have:

1) A dataset:

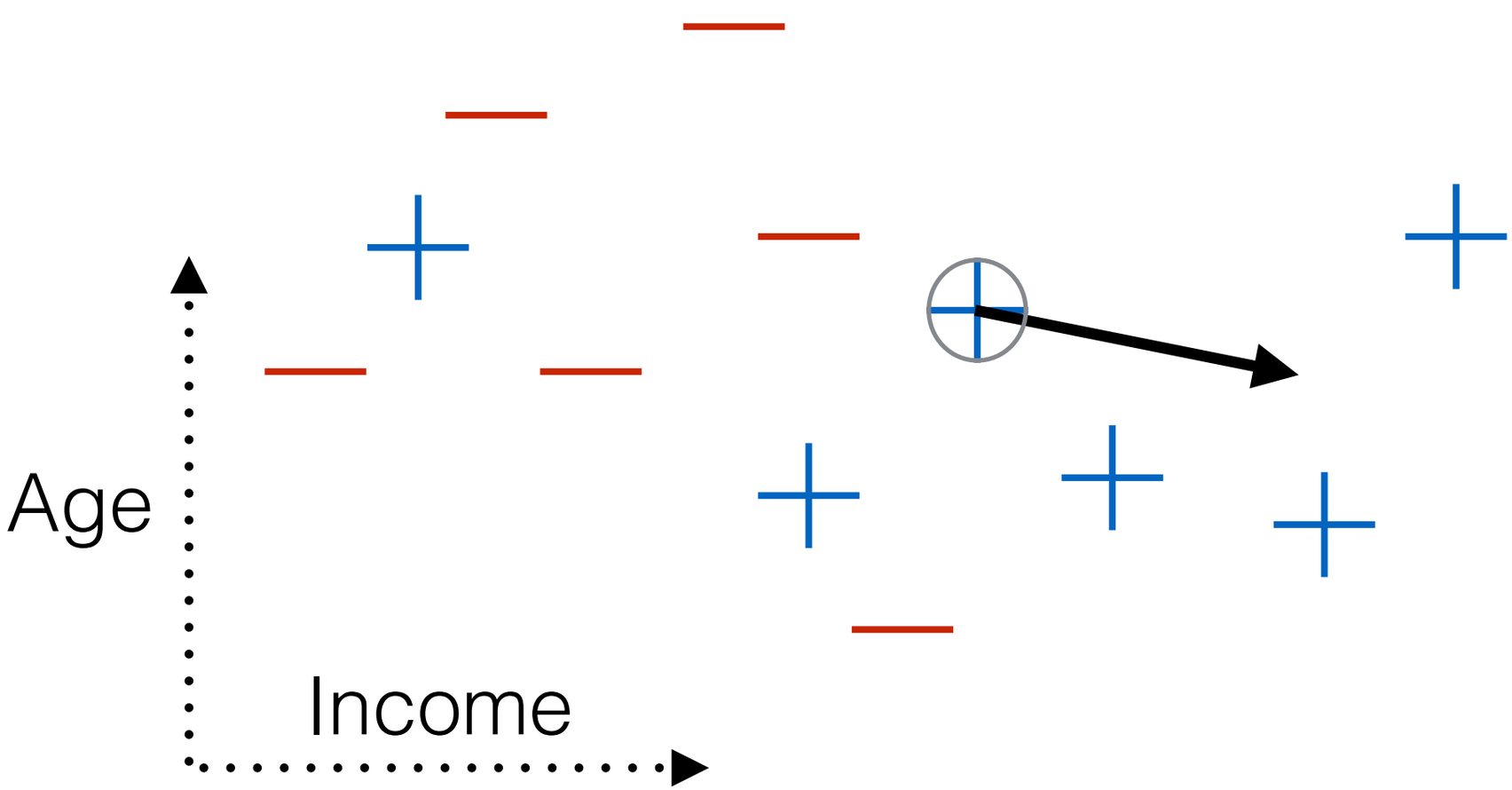
ID	Salary	Age	…	Loan granted?
Alice	3000	28	…	−
Bob	8000	40	…	+
Carol	5000	24	…	+
…	…	…	…	

2) One record is to be explained - Bob

We want a function from the dataset to the influence vector.



We want a function from the dataset to the influence vector.



$$\phi(\vec{x}, \mathcal{X}) \rightarrow \mathbb{R}^n$$

How do we approach the problem: axioms

1) Shift Invariance
2) Rotation and Reflection Faithfulness
3) Continuity
4) Flip Invariance
5) Monotonicity
6) Random Labels

How do we approach the problem: axioms

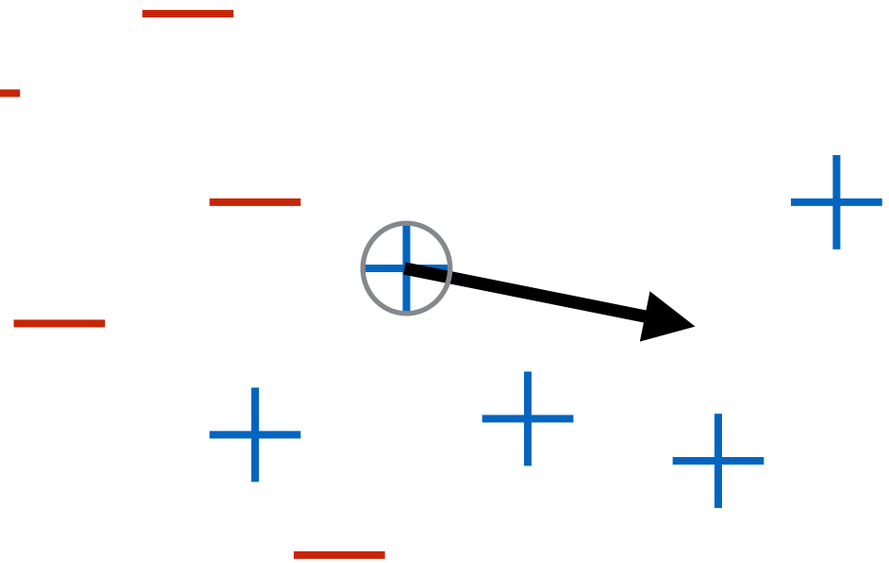
1) Shift Invariance
2) Rotation and Reflection Faithfulness
3) Continuity
4) Flip Invariance
5) Monotonicity
6) Random Labels

Random Labels:

If we view the dataset as a random variable, and the labels are assigned i.i.d. uniformly at random, then for any $\vec{x}$ and dataset we have:

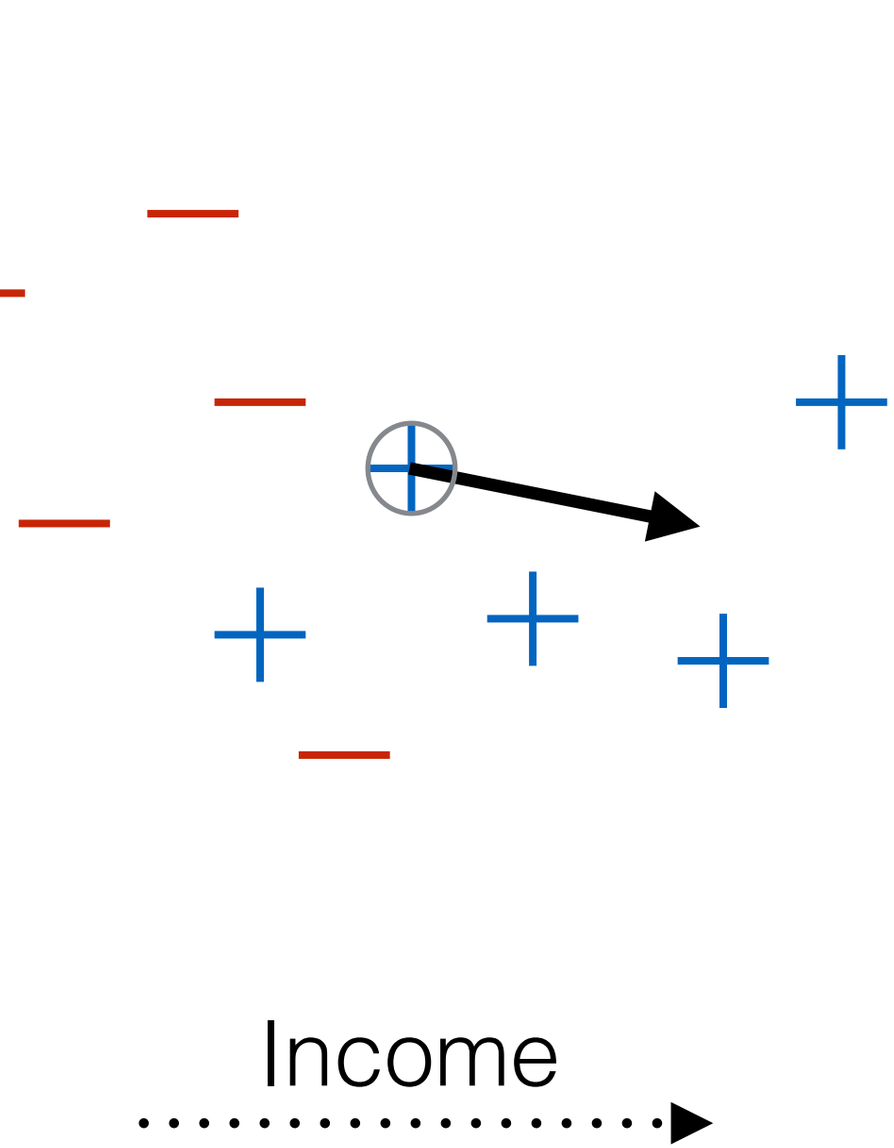
$$\mathbf{E}[\phi(\vec{x}, \mathcal{X})] = \vec{0}$$

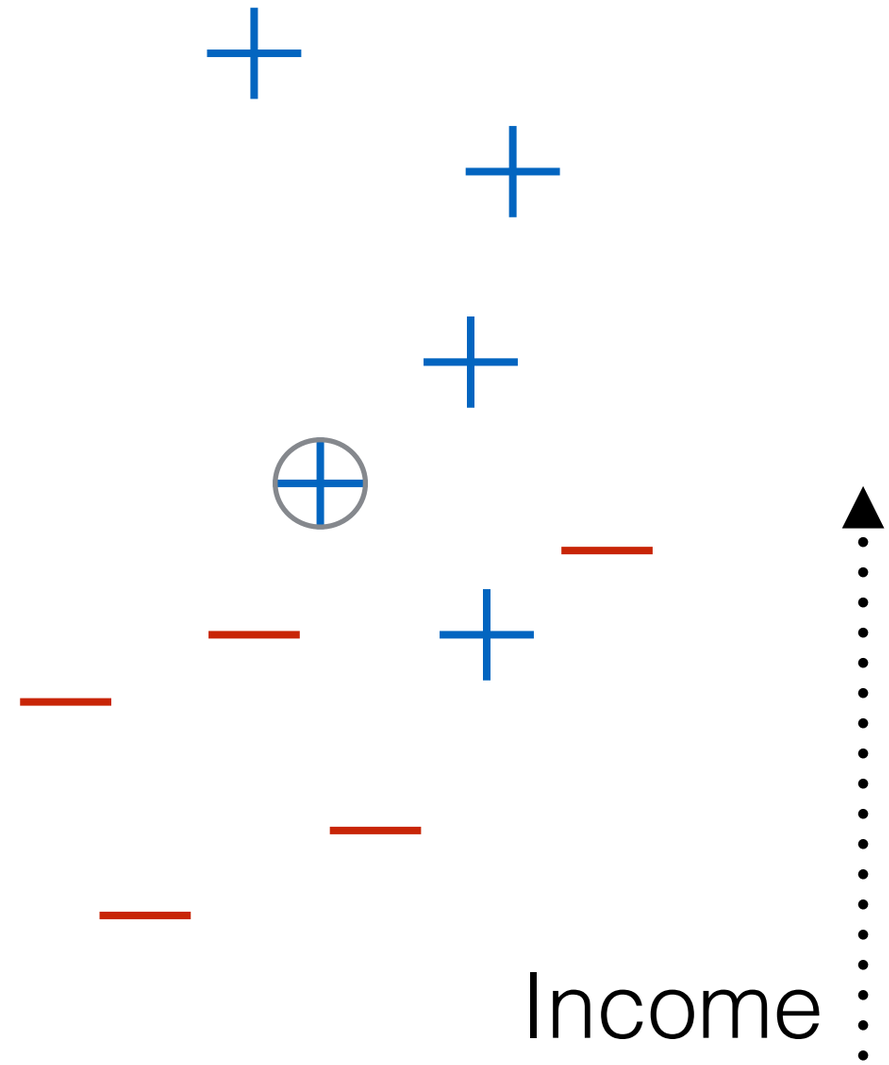
Axiom: Rotation faithfulness



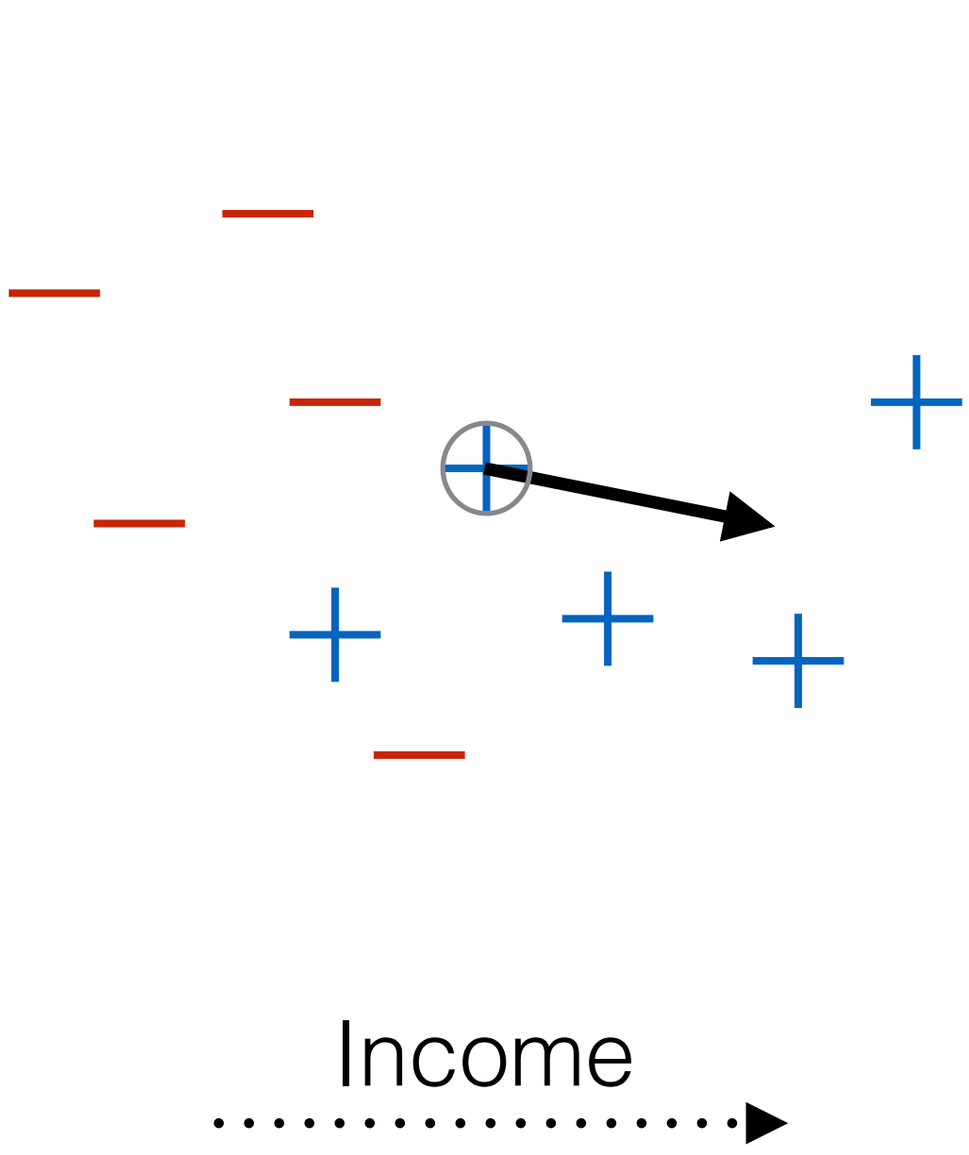
Income

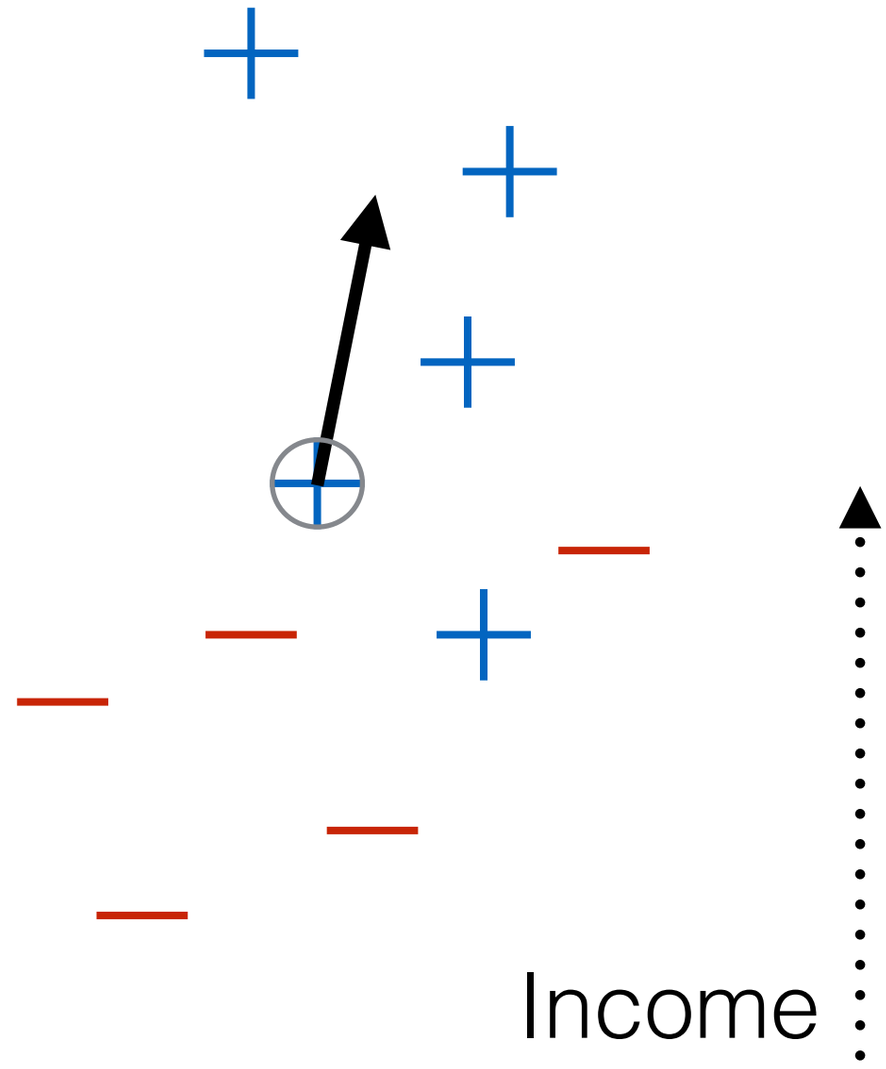
Axiom: Rotation faithfulness



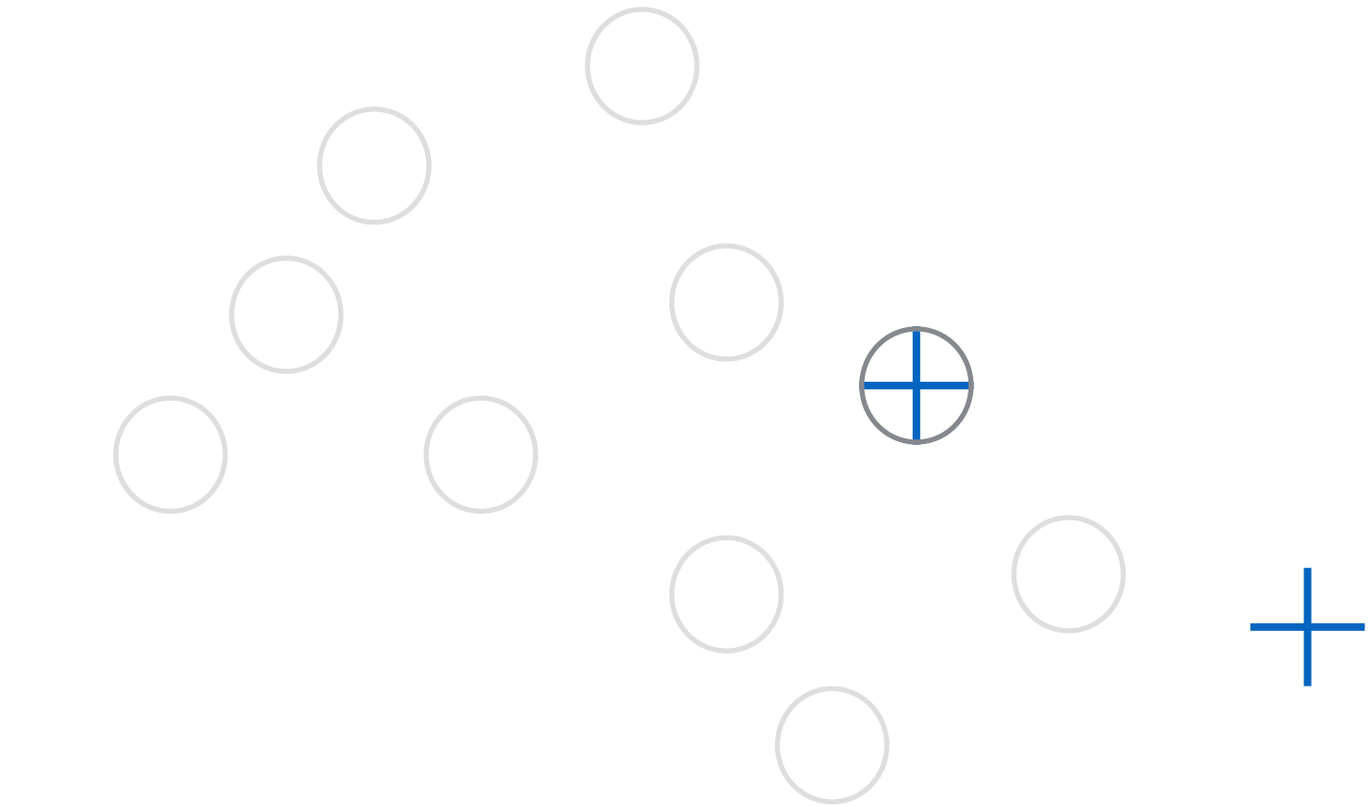


Axiom: Rotation faithfulness



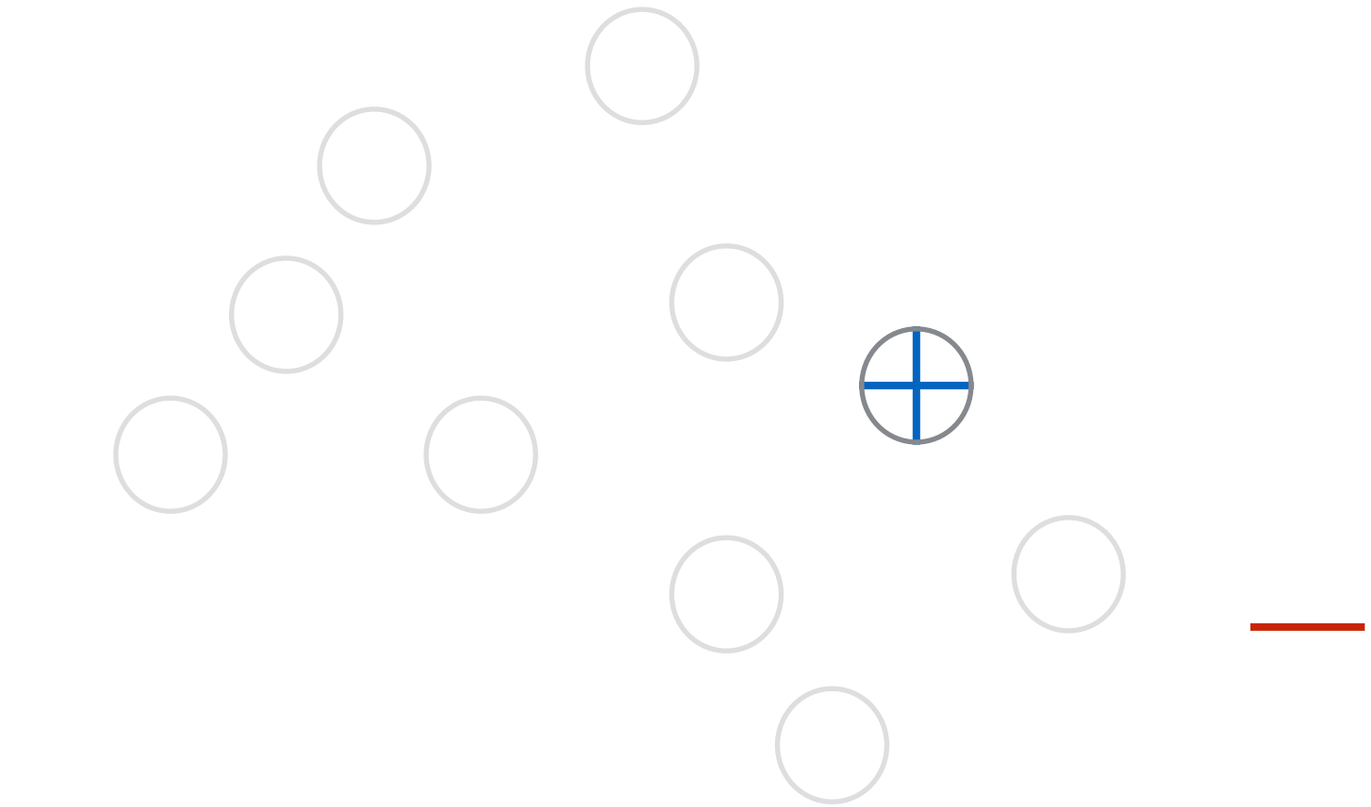


Crucial axiom: monotonicity



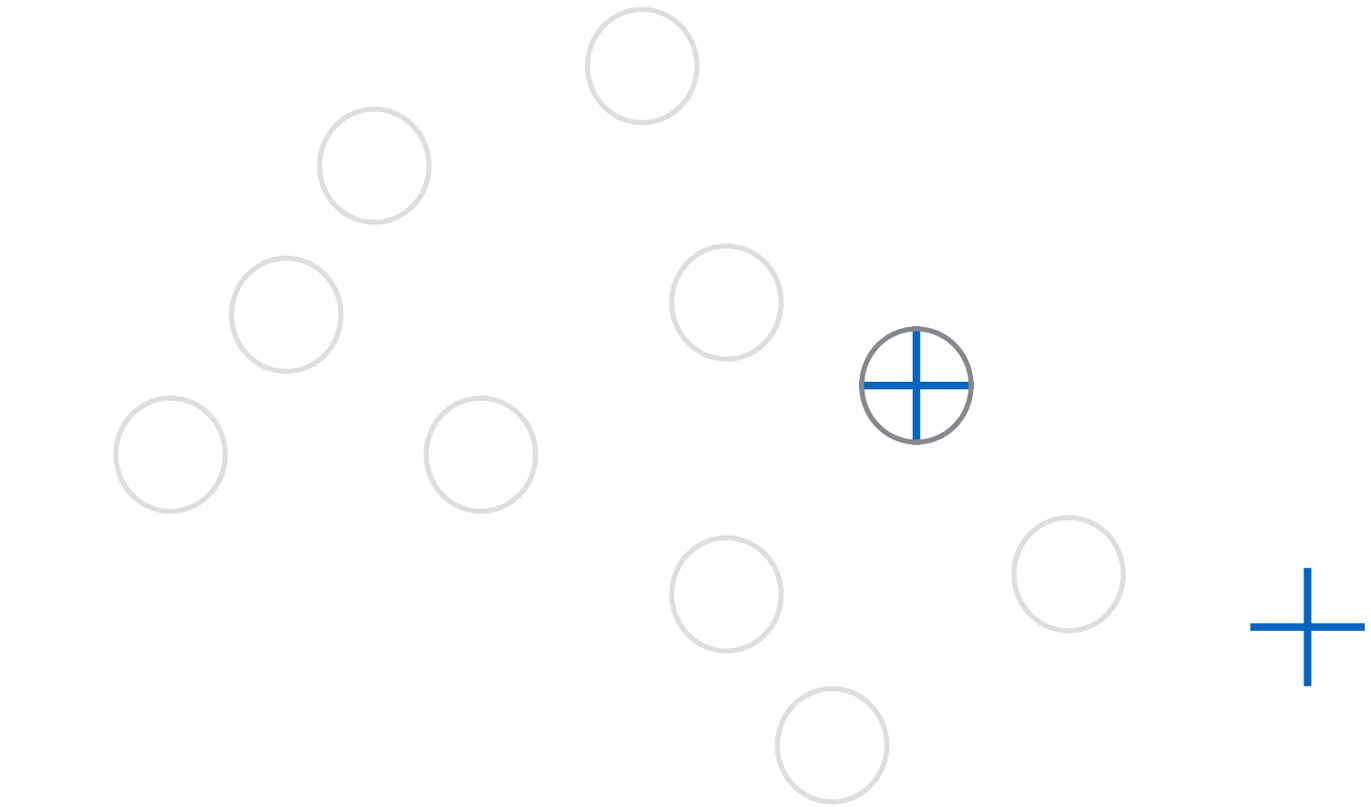
Income

Crucial axiom: monotonicity



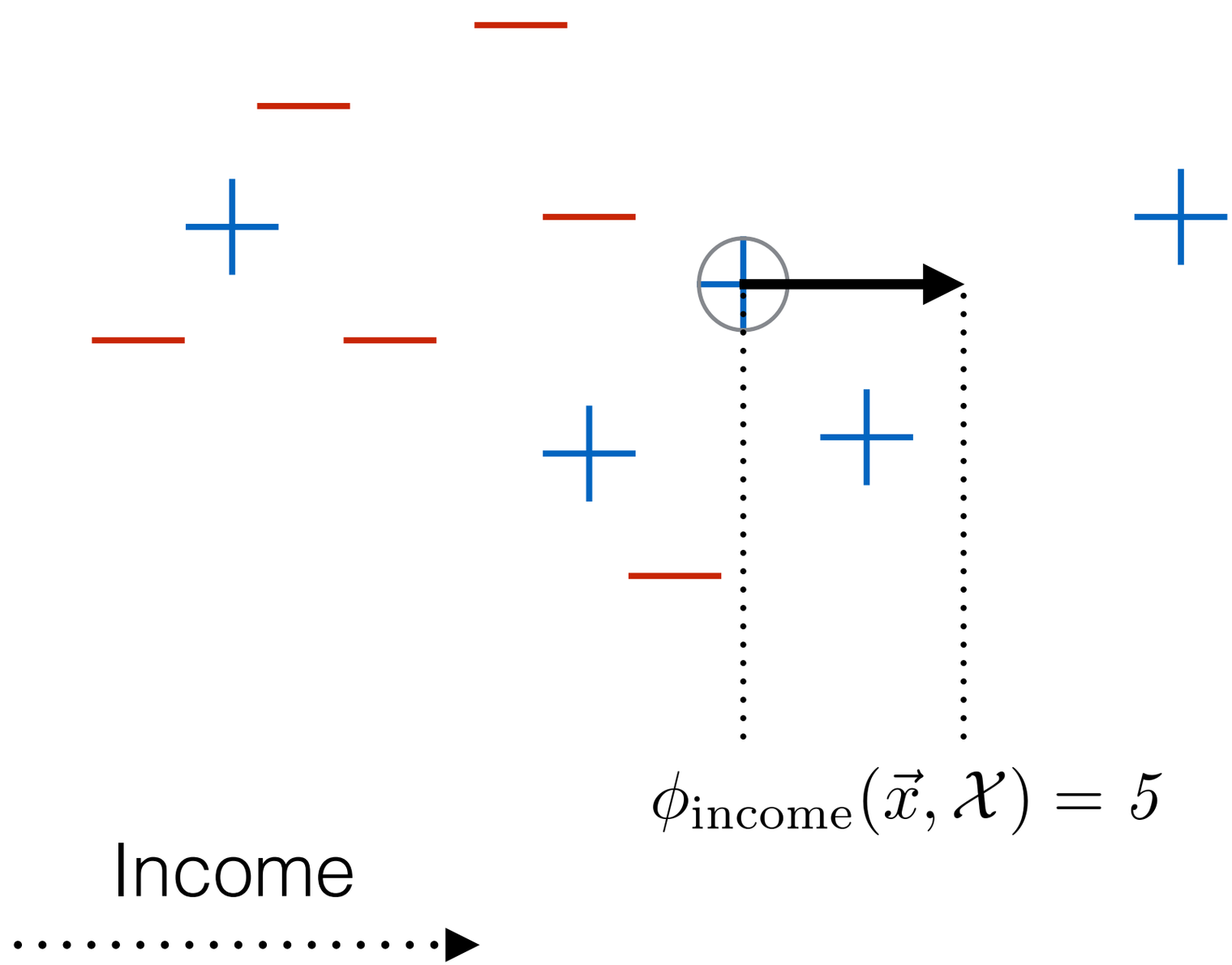
Income

Crucial axiom: monotonicity

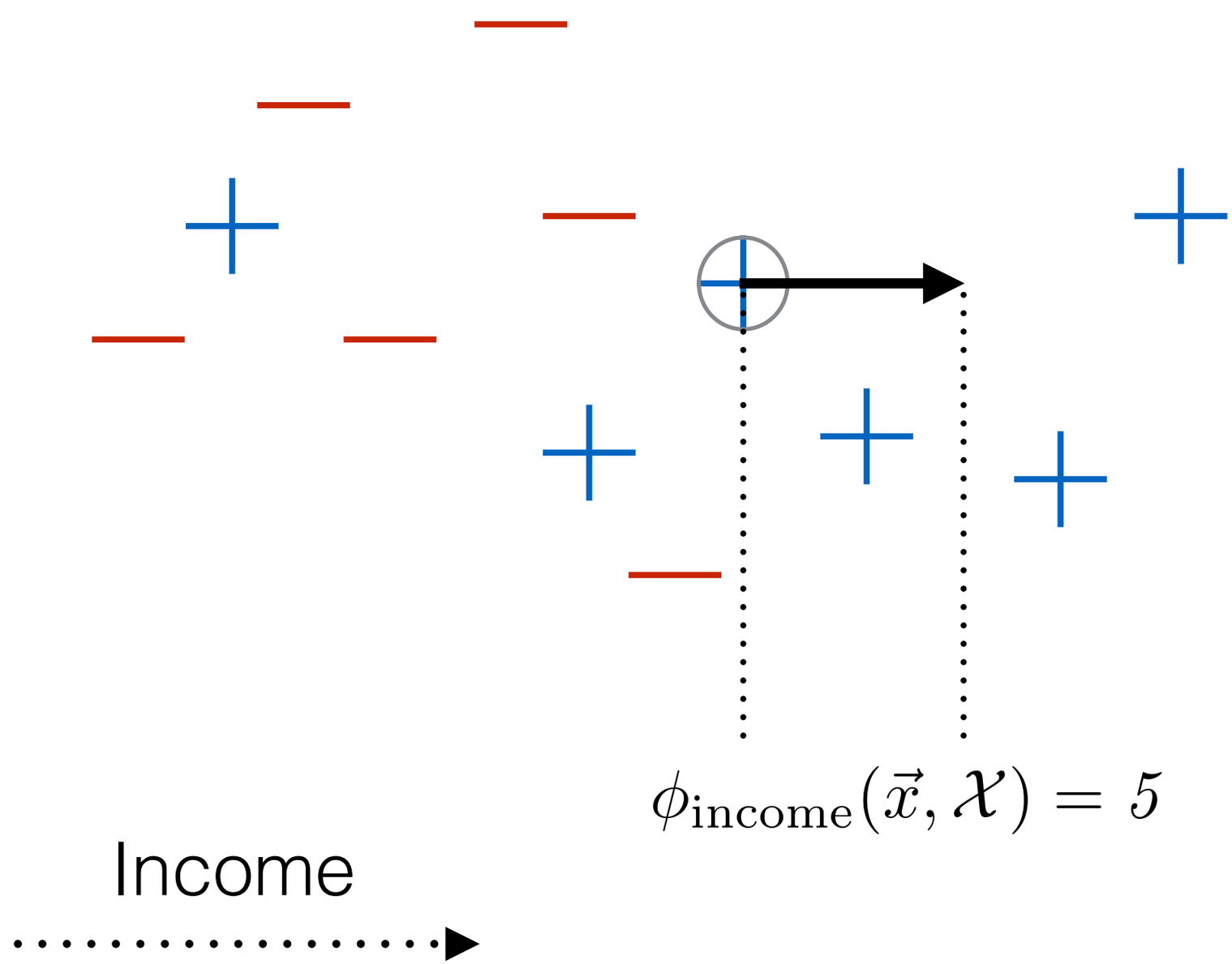


Income

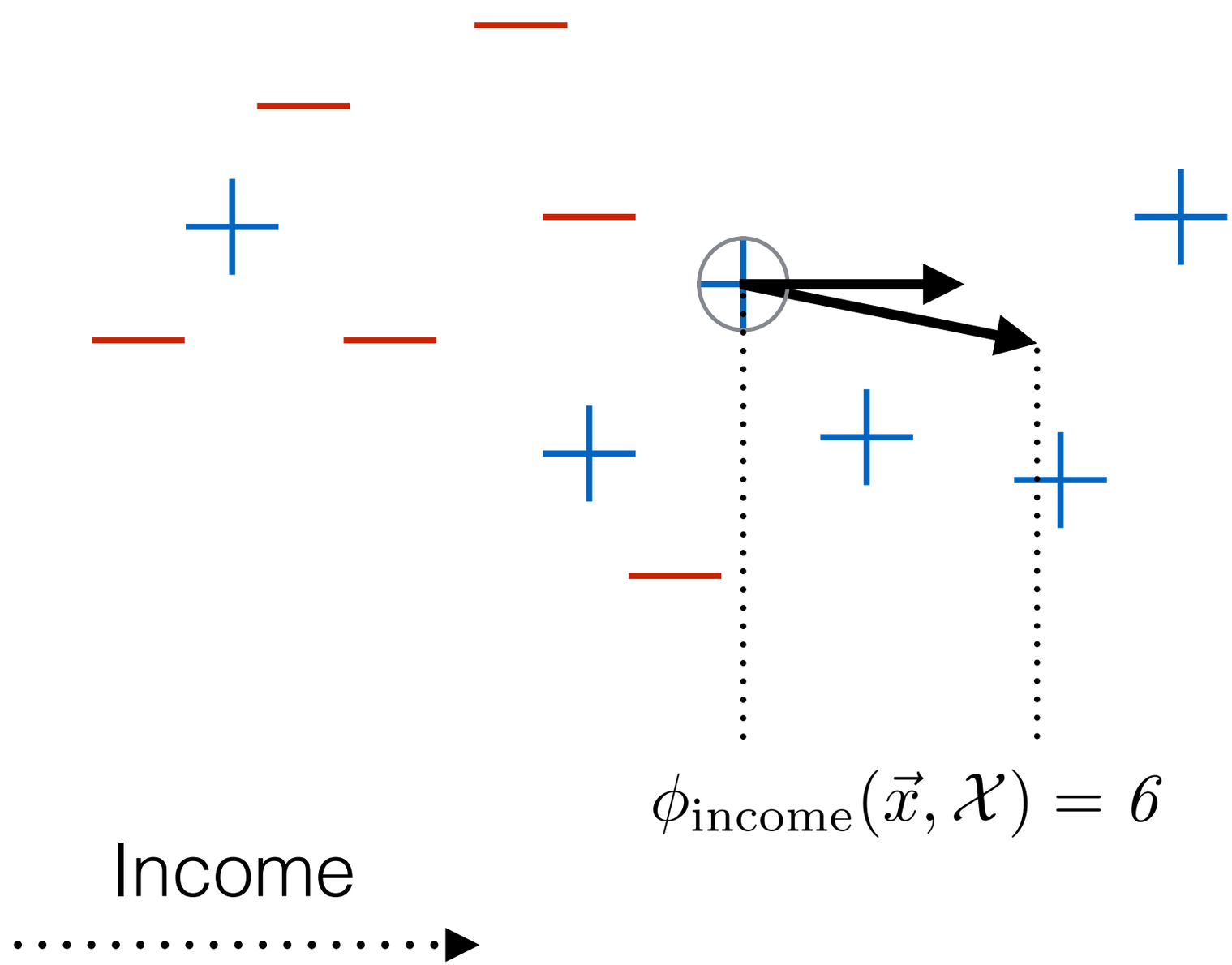
Crucial axiom: monotonicity



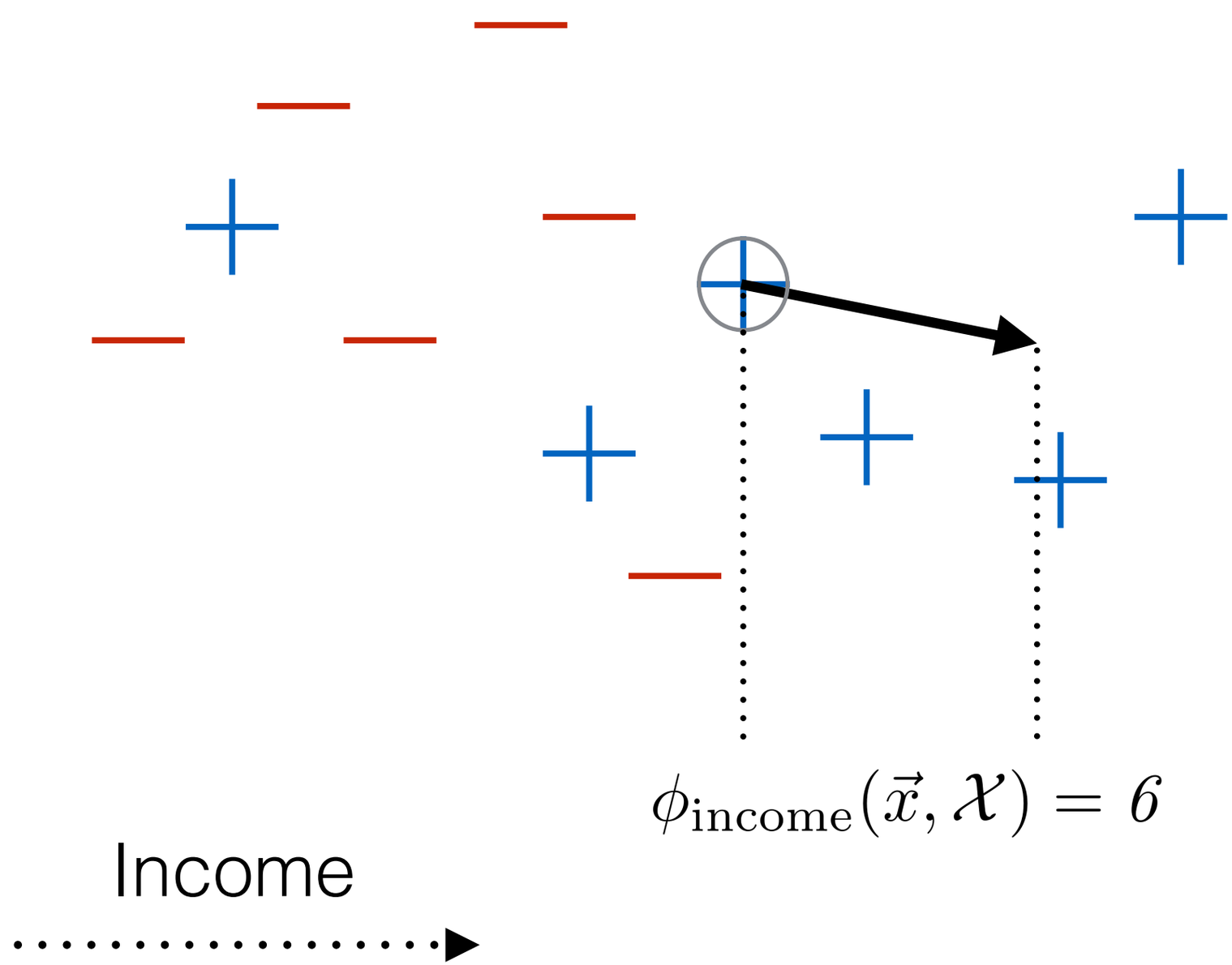
Crucial axiom: monotonicity



Crucial axiom: monotonicity



Crucial axiom: monotonicity



Theorem: The influence measure has to be of this form to satisfy the axioms:

$$\phi(\vec{x}, \mathcal{X}) = \sum_{\vec{y} \in \mathcal{X} \setminus \vec{x}} (\vec{y} - \vec{x}) \alpha(\|\vec{y} - \vec{x}\|_2) 1_{c(\vec{x})=c(\vec{y})}$$

Theorem: The influence measure has to be of this form to satisfy the axioms:

$$\phi(\vec{x}, \mathcal{X}) = \sum_{\vec{y} \in \mathcal{X} \setminus \vec{x}} (\vec{y} - \vec{x}) \alpha(\|\vec{y} - \vec{x}\|_2) 1_{c(\vec{x})=c(\vec{y})}$$

Sum over datapoints

Theorem: The influence measure has to be of this form to satisfy the axioms:

$$\phi(\vec{x}, \mathcal{X}) = \sum_{\vec{y} \in \mathcal{X} \setminus \vec{x}} (\vec{y} - \vec{x}) \alpha(\|\vec{y} - \vec{x}\|_2) 1_{c(\vec{x})=c(\vec{y})}$$

Sum over datapoints

Point towards the datapoint

Theorem: The influence measure has to be of this form to satisfy the axioms:

$$\phi(\vec{x}, \mathcal{X}) = \sum_{\vec{y} \in \mathcal{X} \setminus \vec{x}} (\vec{y} - \vec{x}) \alpha(\|\vec{y} - \vec{x}\|_2) 1_{c(\vec{x})=c(\vec{y})}$$

Sum over datapoints

Point towards the datapoint

A function that tells how important (for Bob) a datapoint is based on how far from Bob it is.
For example, we will go with:

$$\alpha(x) = \frac{1}{x^2}$$

Theorem: The influence measure has to be of this form to satisfy the axioms:

$$\phi(\vec{x}, \mathcal{X}) = \sum_{\vec{y} \in \mathcal{X} \setminus \vec{x}} (\vec{y} - \vec{x}) \alpha(\|\vec{y} - \vec{x}\|_2) 1_{c(\vec{x})=c(\vec{y})}$$

Sum over datapoints

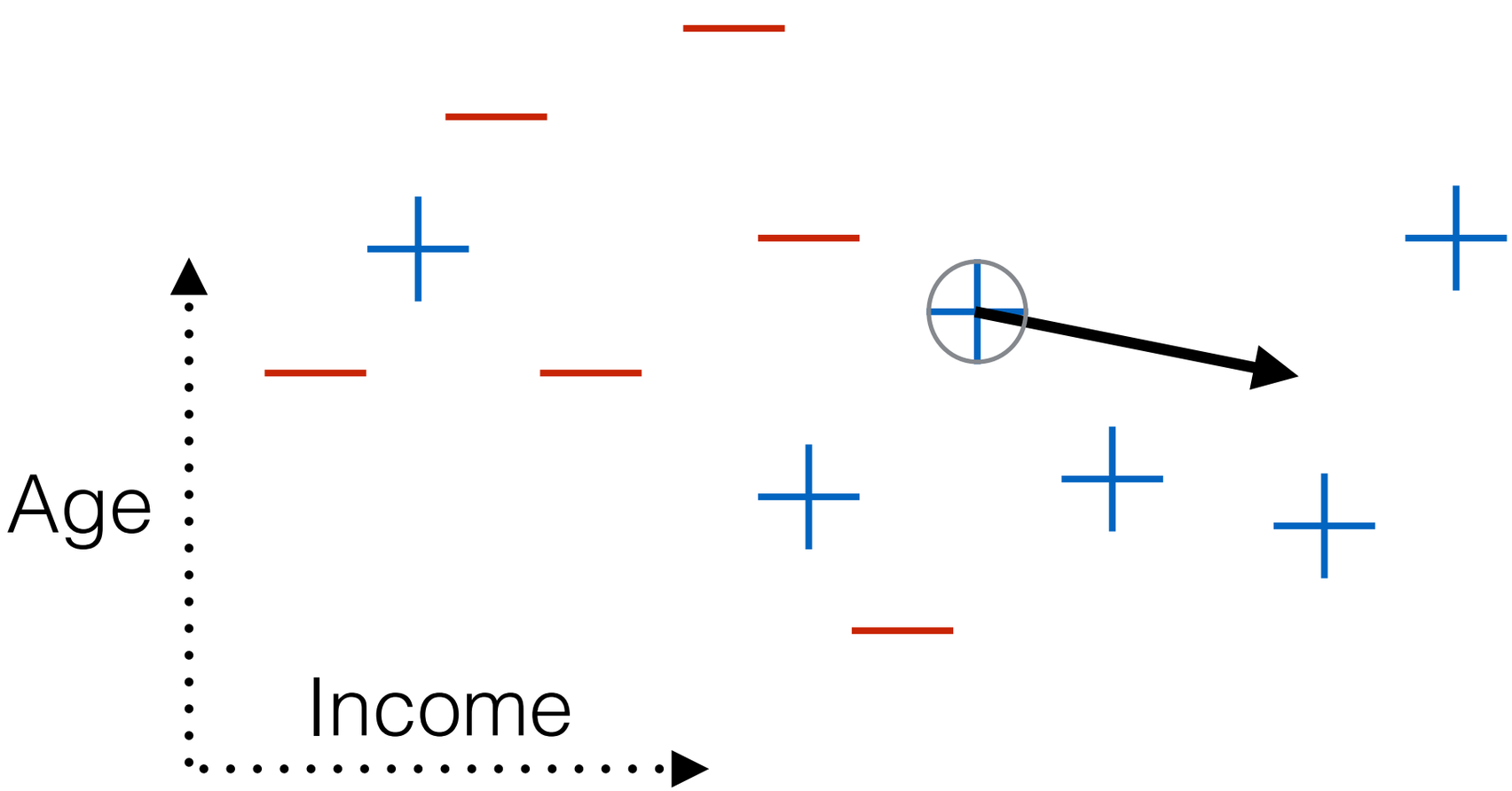
Point towards the datapoint

A function that tells how important (for Bob) a datapoint is based on how far from Bob it is.
For example, we will go with:

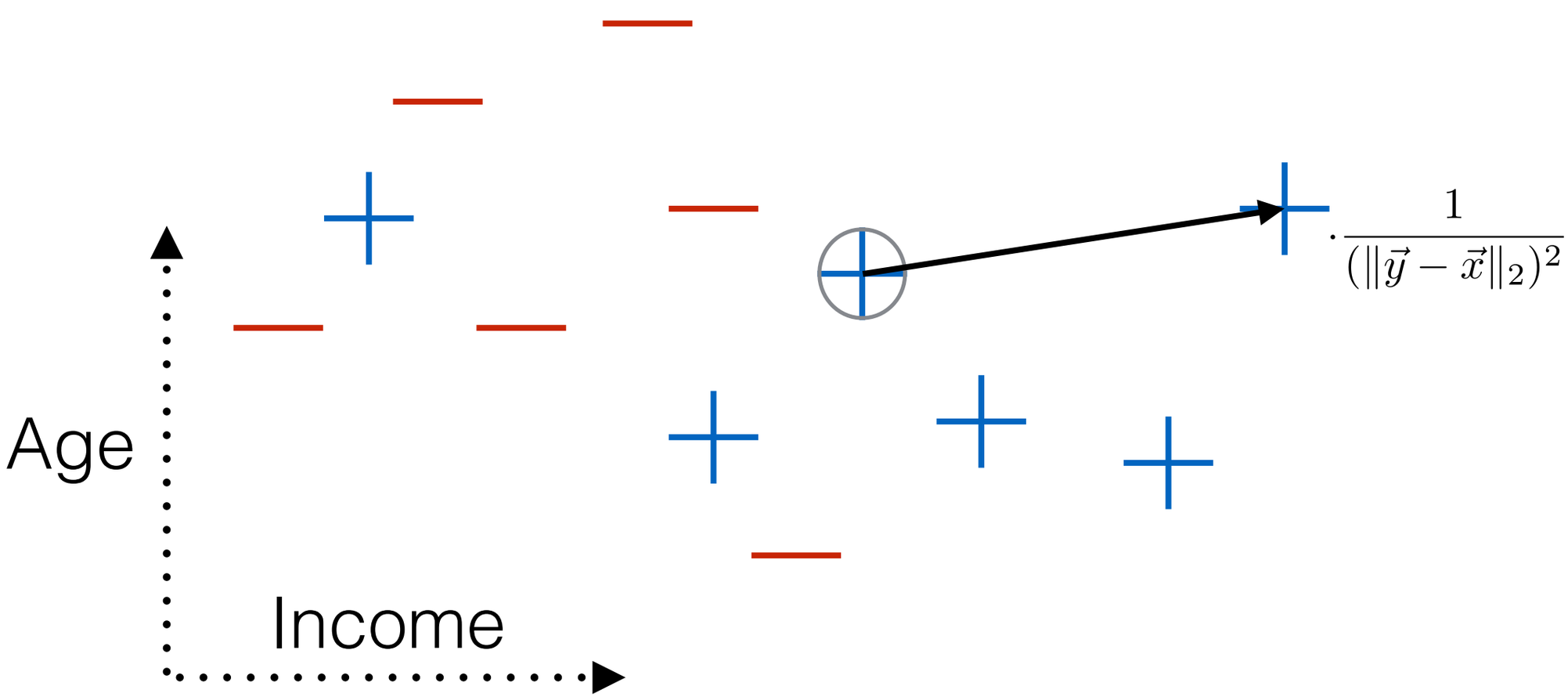
$$\alpha(x) = \frac{1}{x^2}$$

{1,-1}-valued indicator
"Is this point's label the same as Bob's?"

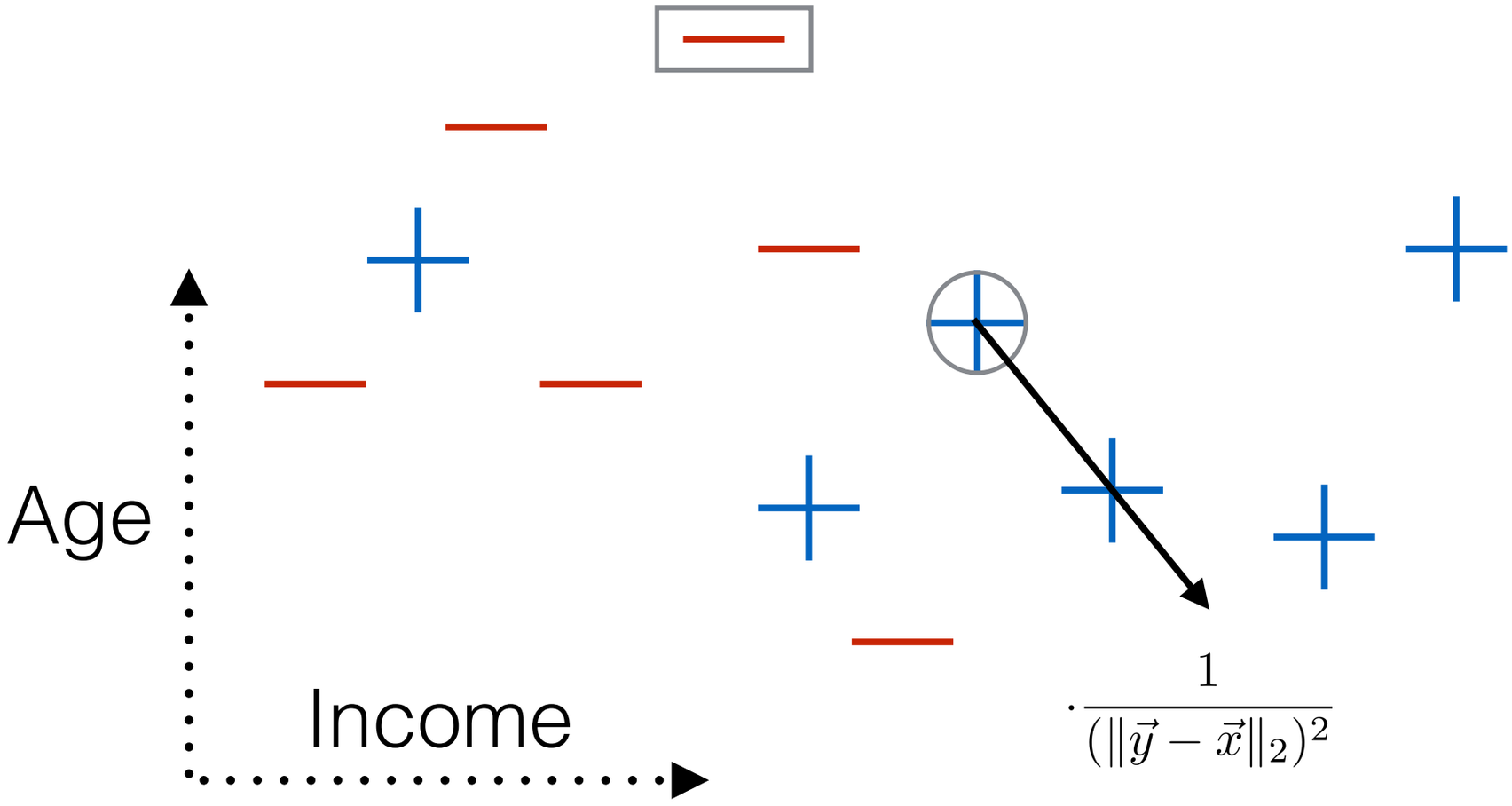
$$\phi(\vec{x}, \mathcal{X}) = \sum_{\vec{y} \in \mathcal{X} \setminus \vec{x}} (\vec{y} - \vec{x}) \alpha(\|\vec{y} - \vec{x}\|_2) 1_{c(\vec{x})=c(\vec{y})}$$



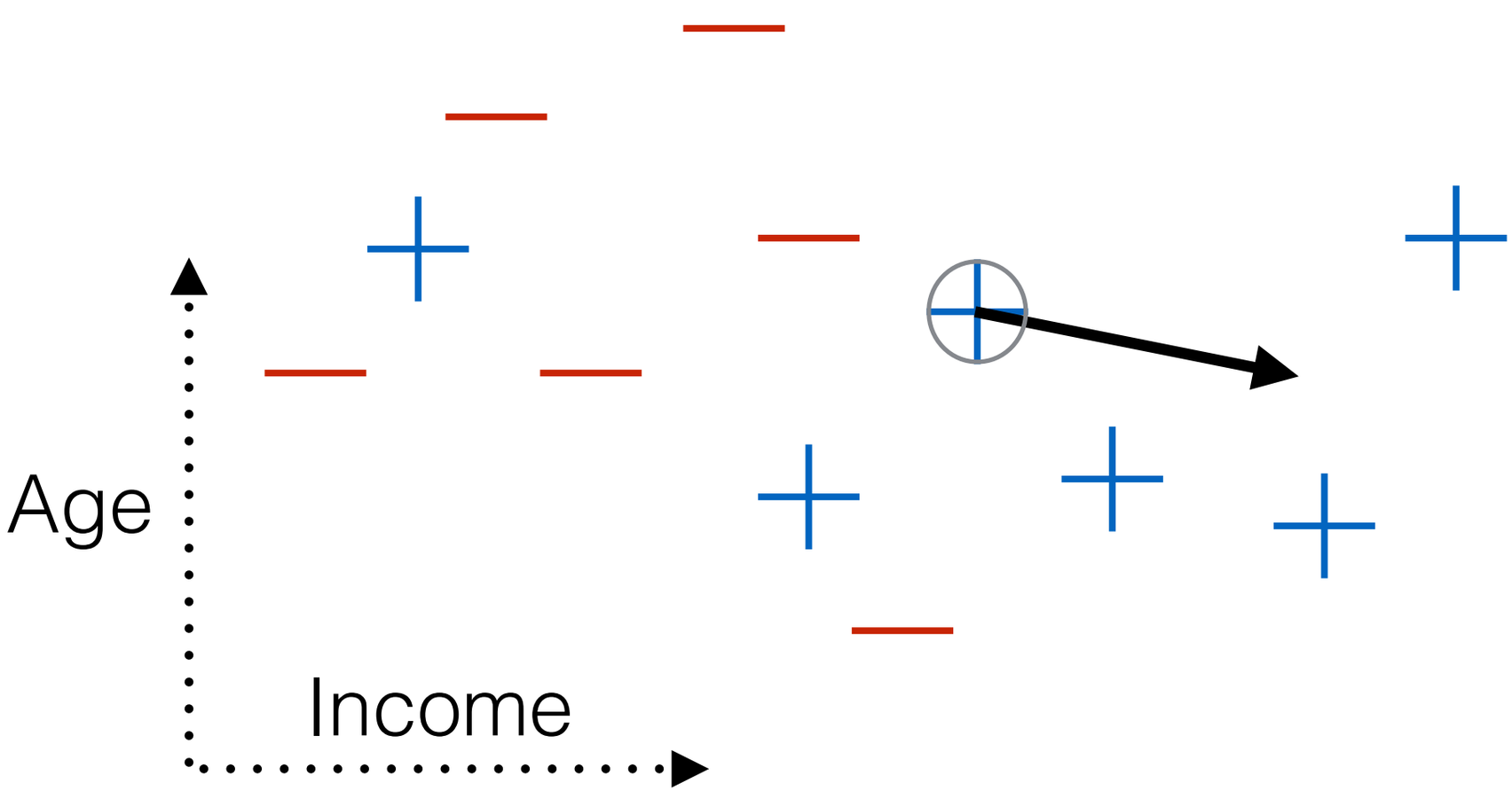
$$\phi(\vec{x}, \mathcal{X}) = \sum_{\vec{y} \in \mathcal{X} \setminus \vec{x}} (\vec{y} - \vec{x}) \alpha(\|\vec{y} - \vec{x}\|_2) 1_{c(\vec{x})=c(\vec{y})}$$



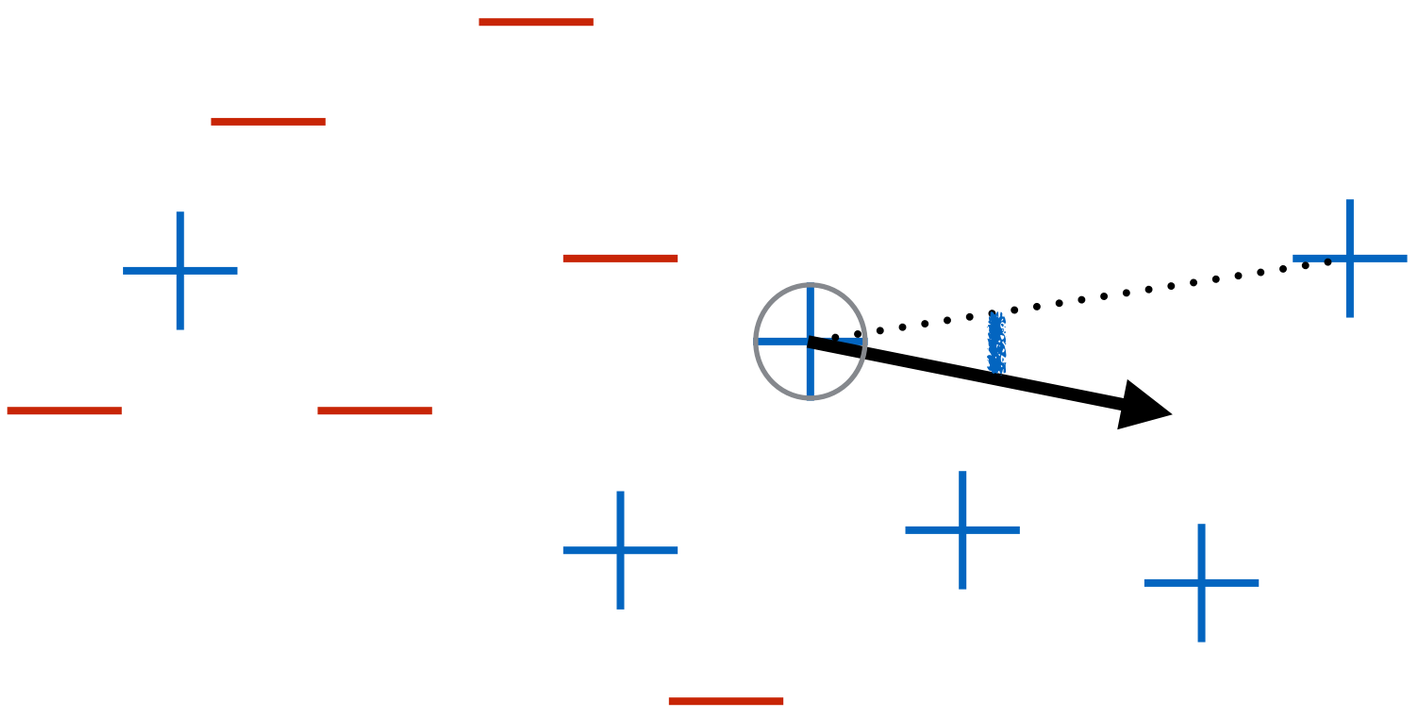
$$\phi(\vec{x}, \mathcal{X}) = \sum_{\vec{y} \in \mathcal{X} \setminus \vec{x}} (\vec{y} - \vec{x}) \alpha(\|\vec{y} - \vec{x}\|_2) 1_{c(\vec{x})=c(\vec{y})}$$



$$\phi(\vec{x}, \mathcal{X}) = \sum_{\vec{y} \in \mathcal{X} \setminus \vec{x}} (\vec{y} - \vec{x}) \alpha(\|\vec{y} - \vec{x}\|_2) 1_{c(\vec{x})=c(\vec{y})}$$

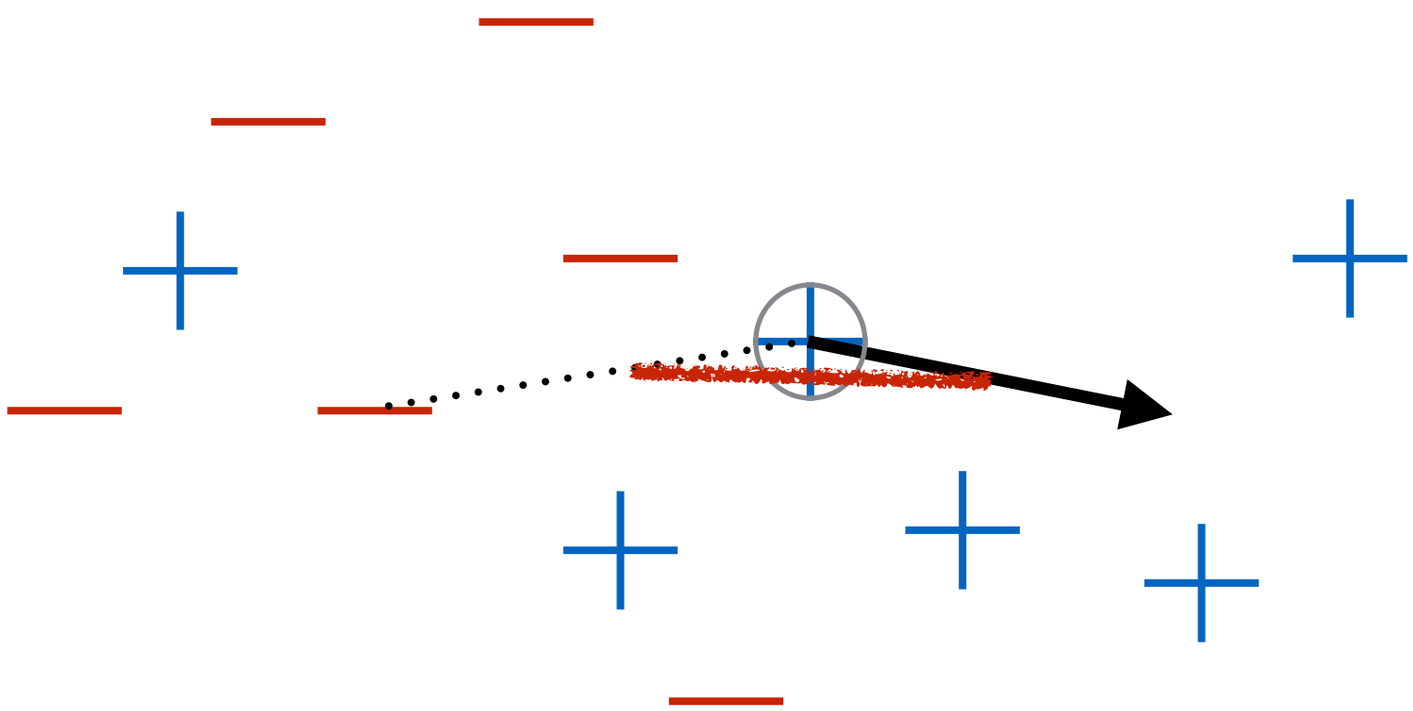


Theorem: The solution to the following optimisation problem is the same function:



$$\phi(\vec{x}, \mathcal{X}) := \arg\max_{\phi \in \mathbb{R}^n} \left(\sum_{\vec{y} \in \mathcal{X}, c(\vec{y}) = c(\vec{x})} \cos(\vec{y}, \phi) \alpha(\|\vec{y}\|_2) - \sum_{\vec{y} \in \mathcal{X}, c(\vec{y}) \neq c(\vec{x})} \cos(\vec{y}, \phi) \alpha(\|\vec{y}\|_2) \right)$$

Theorem: The solution to the following optimisation problem is the same function:



$$\phi(\vec{x}, \mathcal{X}) := \arg\max_{\phi \in \mathbb{R}^n} \left(\sum_{\vec{y} \in \mathcal{X}, c(\vec{y}) = c(\vec{x})} \cos(\vec{y}, \phi) \alpha(\|\vec{y}\|_2) - \sum_{\vec{y} \in \mathcal{X}, c(\vec{y}) \neq c(\vec{x})} \cos(\vec{y}, \phi) \alpha(\|\vec{y}\|_2) \right)$$

Experiment

Dataset of ~12000 images divided equally into happy and sad facial expressions.

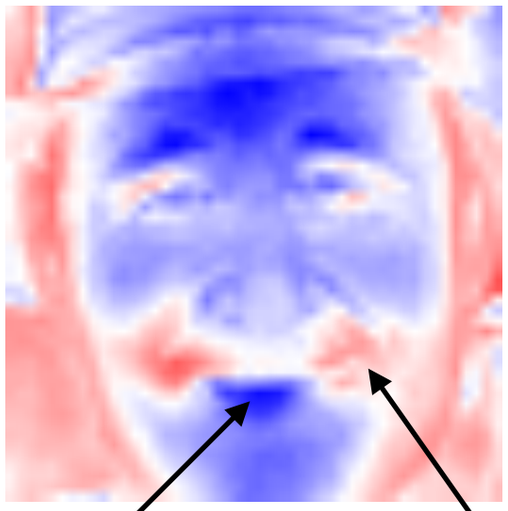
Images are 28x28 greyscale.

Every pixel’s brightness value is a feature.

Example result

Image:

Influence vector:



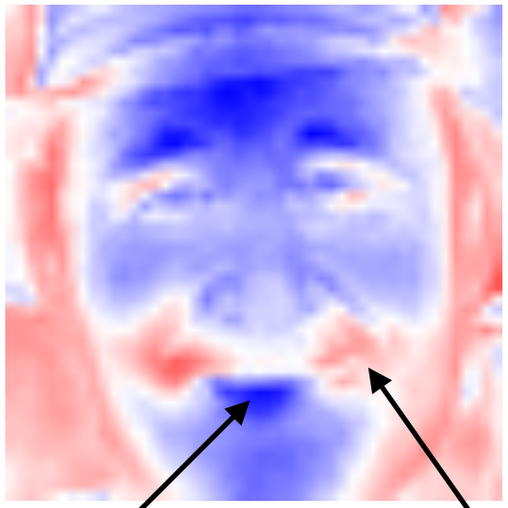
Classified as happy because it's bright enough here

and dark enough here

Example result

Image:

Influence vector:



Overlay:

Classified as happy because
it's bright enough here

and dark enough here

Example problem

Things can go wrong if the locality of the datapoint is not described well by the data

Image:

Influence vector:

Example problem

Things can go wrong if the locality of the datapoint is not described well by the data

Image:

Influence vector:

Overlay:

Summary

-Axiomatic formulation

-Based on the actual data rather than access to the classifier

-Alternative formulation as cosine similarity

-Interesting properties within Game Theory

-Related to mathematical formulations of responsibility and blame