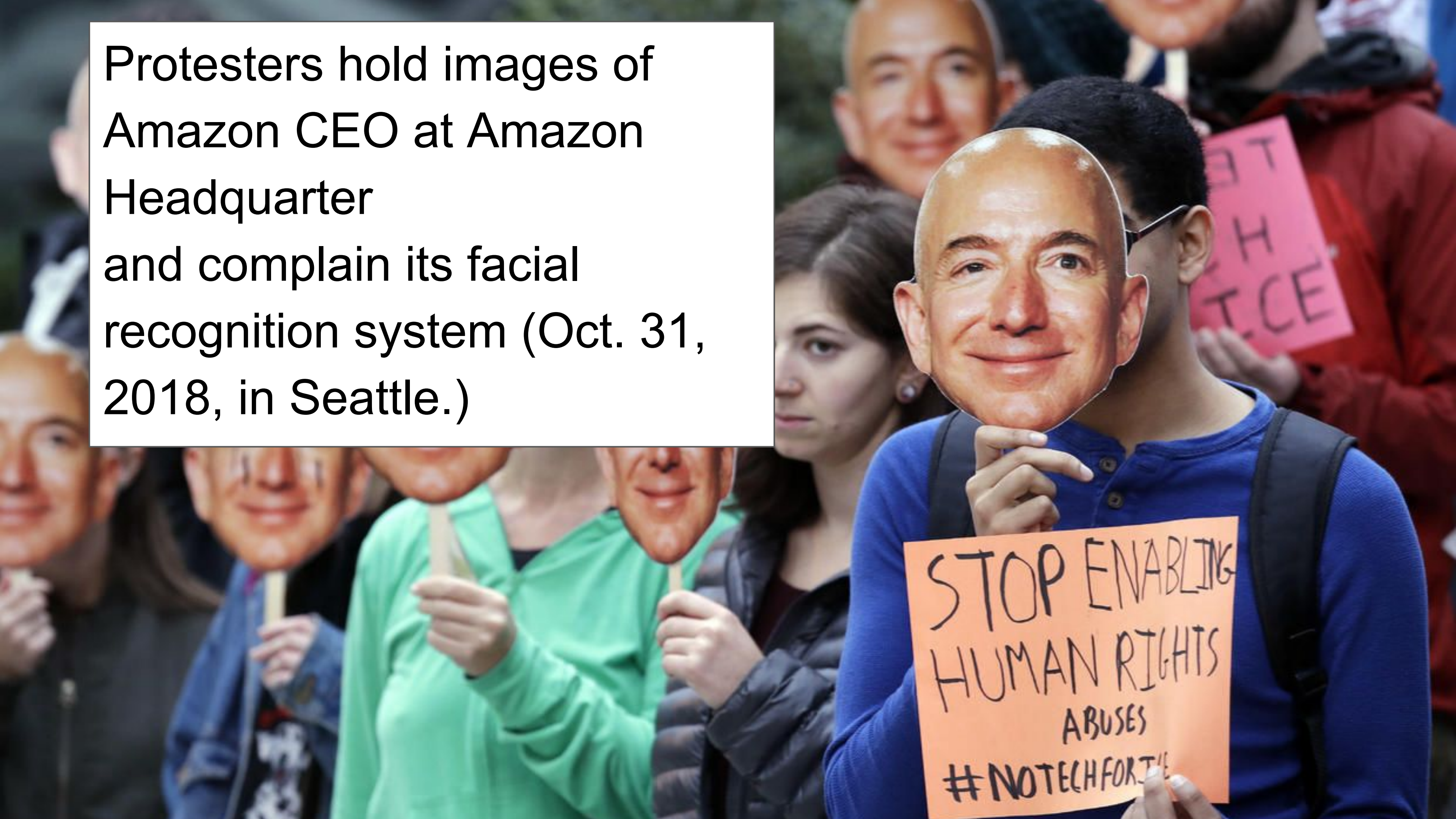
# Racial and Gender Bias in AI

Junhong Wang



Protesters hold images of Amazon CEO at Amazon Headquarter
and complain its facial recognition system (Oct. 31, 2018, in Seattle.)

| Subject | Error Rate (misclassified as opposite sex) |
|---|---|
| Darker-skinned women | 31% |
| Lighter-skinned women | 7% |
| Darker-skinned men | 1% |
| lighter-skinned men | 0% |

AI can learn biased from human creators. What potential risks can this cause?

Biased AI can be a threat to civil liberties. Biased AI can be abused to promote gender and racial discrimination.

# Roadmap

1. How does machine learning work?
2. How does facial recognition work?
3. What does “AI is biased” mean?
4. How to fix biased AI?

# Roadmap

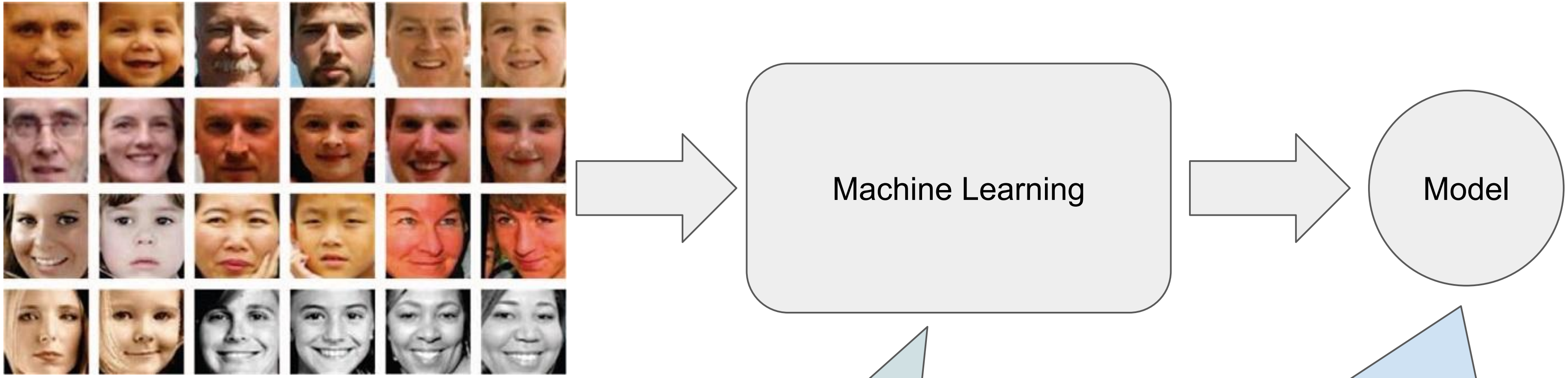
1. **How does machine learning work?**

2. How does facial recognition work?

3. What does “AI is biased” mean?

4. How to fix biased AI?

# How does machine learning work?



Machine Learning

Model

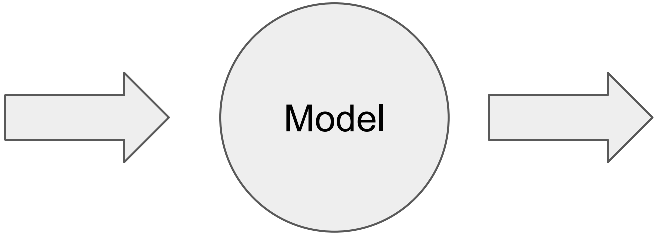
Collect the data we want the machine to learn. Each picture should be labeled as “male” or “female”.

The machine tries to figure out the pattern of the faces and the label. (e.g. many females have larger eyes, many males have bears)

positive point for female feature.
negative point for male feature.

| | |
|---|---|
| Big mouth | : -1 |
| Bear | : -1 |
| Beautiful skin | : +1 |
| Large eyes | : +1 |
| ...etc | |

# How does machine learning work?



Model

Female!

|  |  |
|---|---|
| Big mouth | : -1 |
| Bear | : -1 |
| Beautiful skin | : +1 |
| Large eyes | : +1 |

Hmm… okay, you have a big mouth (-1),
you have a beautiful skin (+1),
you don't have a bear (+1).
Total point is +1, which is greater than 0.
You are a female!

# Roadmap

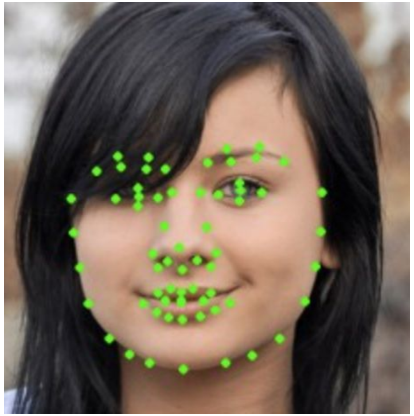
1. How does machine learning work?

2. **How does facial recognition work?**

3. What does “AI is biased” mean?

4. How to fix biased AI?

# How does facial recognition work?



X coord of nose

Y coord of nose

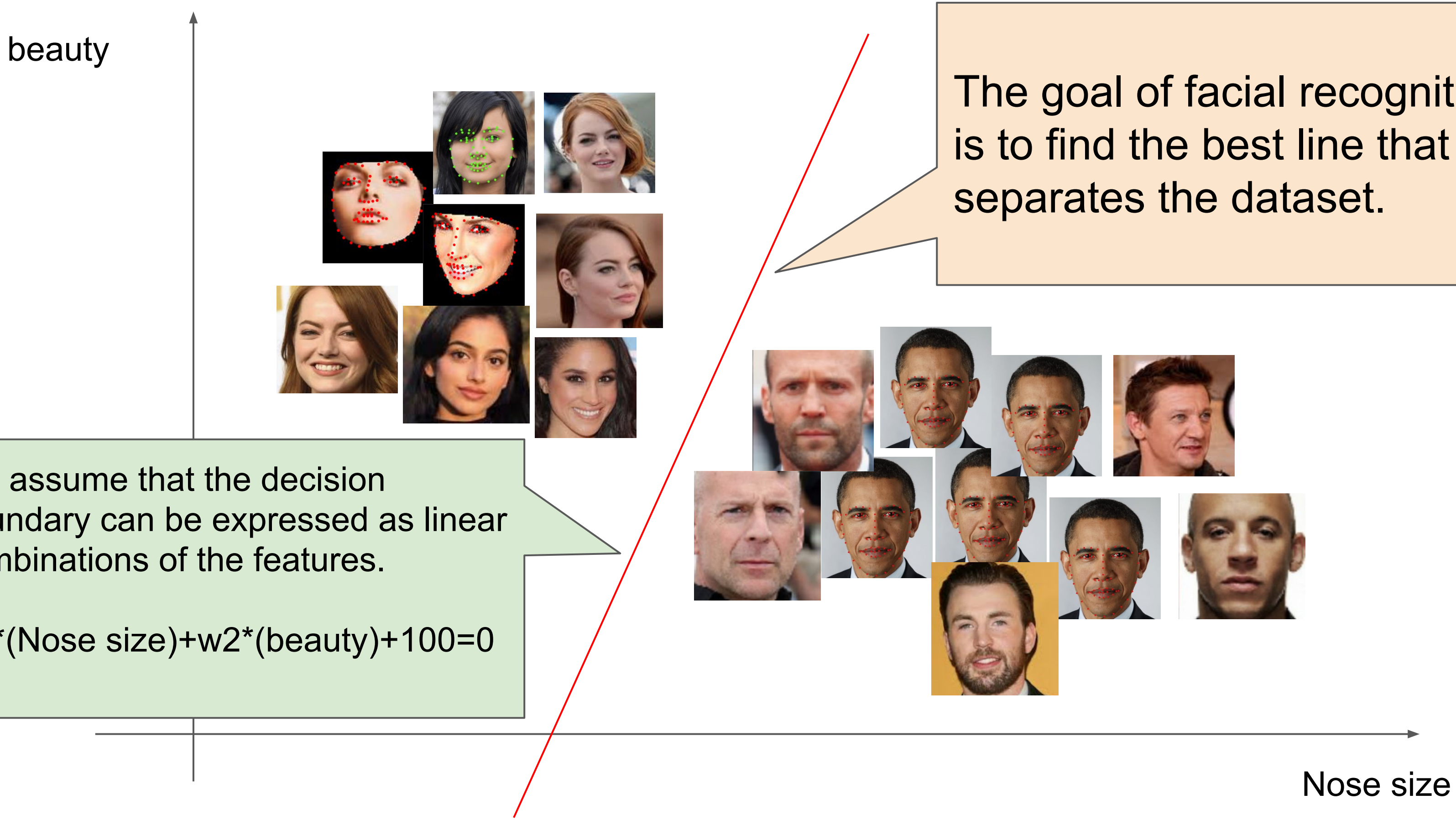
⋮

Y coord of mouth

For a computer, image is just a collection of pixels. It doesn’t know what an eye, nose or mouth are. We will convert the image to a vector, where each component represent a feature of the face.

A better machine learning algorithm such as neural network can further transform the features of coordinates into more abstract feature (e.g. size of nose, shape of face, etc.)

# How does facial recognition work?

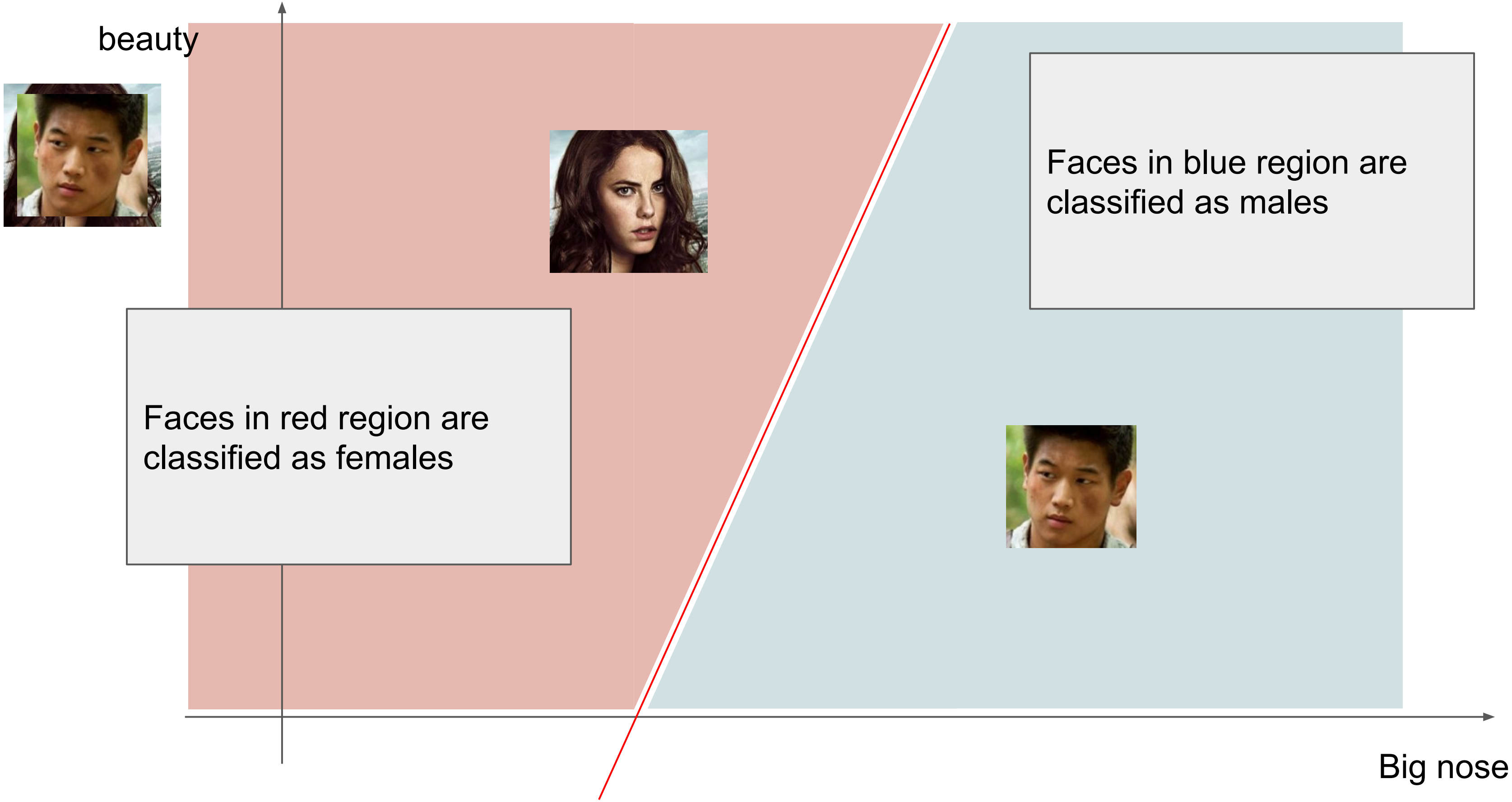


The goal of facial recognition is to find the best line that separates the dataset.

We assume that the decision boundary can be expressed as linear combinations of the features.

w1*(Nose size)+w2*(beauty)+100=0

# How does facial recognition work?



Faces in red region are classified as females

Faces in blue region are classified as males

# Roadmap

1. How does machine learning work?

2. **How does facial recognition work?**

3. What does “AI is biased” mean?

4. How to fix biased AI?

# Roadmap

1. How does machine learning work?
2. **The math behind the magic**
3. What does “AI is biased
4. How to fix biased AI?

It is okay to completely ignore this part. The point of this section is to briefly show you how computer can figure out the best fitting line.

# The math behind the magic

Ok
as

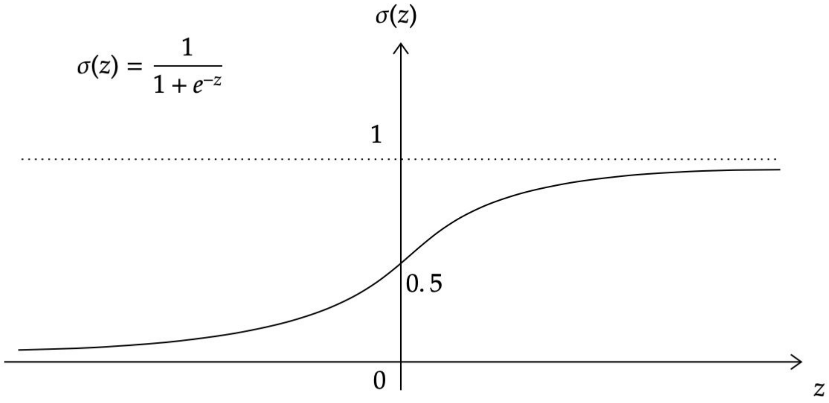
Gi
it's

We assume y can be represented by linear combination of the features x

$$y = \theta^T x$$

T
e
x

Is there a nice function that maps from any real value to a value between 0 and 1? => Sigmoid



Probability to be female $= \sigma(\theta^T x)$

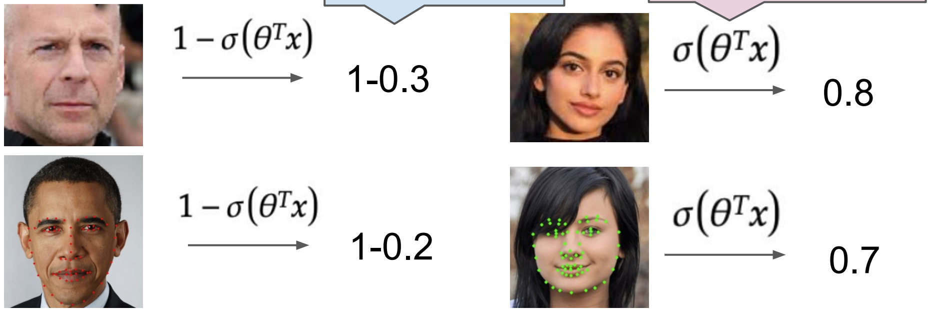
Probability to be male $= 1 - \sigma(\theta^T x)$

# The math behind the magic

Likelihood to be male

Likelihood to be female

We have lots of dataset. We want to maximize the likelihood to observe what we observe.



$\xrightarrow{1-\sigma(\theta^T x)}$ 1-0.3

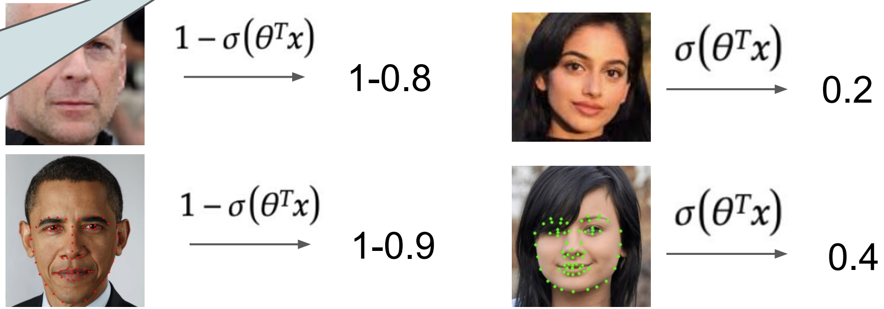
$\xrightarrow{\sigma(\theta^T x)}$ 0.8

$\xrightarrow{1-\sigma(\theta^T x)}$ 1-0.2

$\xrightarrow{\sigma(\theta^T x)}$ 0.7

Likelihood = 0.7*0.8*0.8*0.7

We want to find the value of theta to get the **maximum** likelihood. This type of problem is called optimization problem.



$\xrightarrow{1-\sigma(\theta^T x)}$ 1-0.8

$\xrightarrow{\sigma(\theta^T x)}$ 0.2

$\xrightarrow{1-\sigma(\theta^T x)}$ 1-0.9

$\xrightarrow{\sigma(\theta^T x)}$ 0.4

Likelihood = 0.2*0.1*0.2*0.4

# The math behind the magic

y=0

$\xrightarrow{1-\sigma(\theta^T x)}$ 1-0.3

$\xrightarrow{\sigma(\theta^T x)}$ 0.8

y=1

$$P(y^{(1)};\ x,\ \theta) = \sigma(\theta^T x^{(1)})^{y^{(1)}} \left(1-\sigma(\theta^T x^{(1)})\right)^{1-y^{(1)}}$$

$$\arg\max \prod_{i=1}^{n} \left(\sigma(w^T x^{(i)})\right)^{y^{(i)}} \left(1-\sigma(w^T x^{(i)})\right)^{1-y^{(i)}}$$

# The math behind the magic

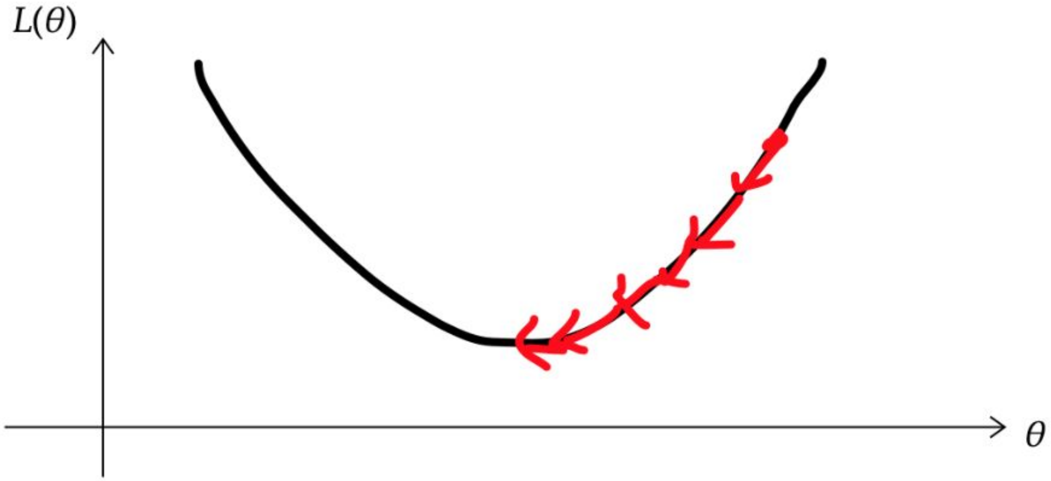
$$\arg\max \prod_{i=1}^{n} \left(\sigma\left(w^T x^{(i)}\right)\right)^{y^{(i)}} \left(1 - \sigma\left(w^T x^{(i)}\right)\right)^{1-y^{(i)}}$$

$$= \arg\max \sum_{i=1}^{n} y^{(i)} \log\left(\sigma\left(w^T x^{(i)}\right)\right) + \left(1 - y^{(i)}\right)\log\left(1 - \sigma\left(w^T x^{(i)}\right)\right)$$

$$= \arg\min \; - \sum_{i=1}^{n} y^{(i)} \log\left(\sigma\left(w^T x^{(i)}\right)\right) + \left(1 - y^{(i)}\right)\log\left(1 - \sigma\left(w^T x^{(i)}\right)\right)$$

# The math behind the magic

$$\arg\min \; -\sum_{i=1}^{n} y^{(i)} \log\left(\sigma\left(w^T x^{(i)}\right)\right) + \left(1 - y^{(i)}\right) \log\left(1 - \sigma\left(w^T x^{(i)}\right)\right)$$



We want to move theta to opposite direction of the gradient!

# The math behind the magic

$$\frac{\delta J}{\delta \theta} = -\sum_{i=1}^{n} \frac{\delta}{\delta \theta}\left(y^{(i)} \log\left(\sigma\left(\theta^T x^{(i)}\right)\right) + \left(1 - y^{(i)}\right)\log\left(1 - \sigma\left(\theta^T x^{(i)}\right)\right)\right)$$

$$\frac{\delta J}{\delta \theta} = \sum_{i=1}^{n}\left(\sigma\left(\theta^T x^{(i)}\right) - y^{(i)}\right)x^{(i)}$$

$$= -\sum_{i=1}^{n}\left(y^{(i)} - \sigma\left(\theta^T x^{(i)}\right)\right)x^{(i)}$$

$$= \sum_{i=1}^{n}\left(\sigma\left(\theta^T x^{(i)}\right) - y^{(i)}\right)x^{(i)}$$

```
Randomly set w

Loop (a lot of times):
    Compute the gradient
    w = w - 0.01 * gradient

// you found the best w!
```

# Roadmap

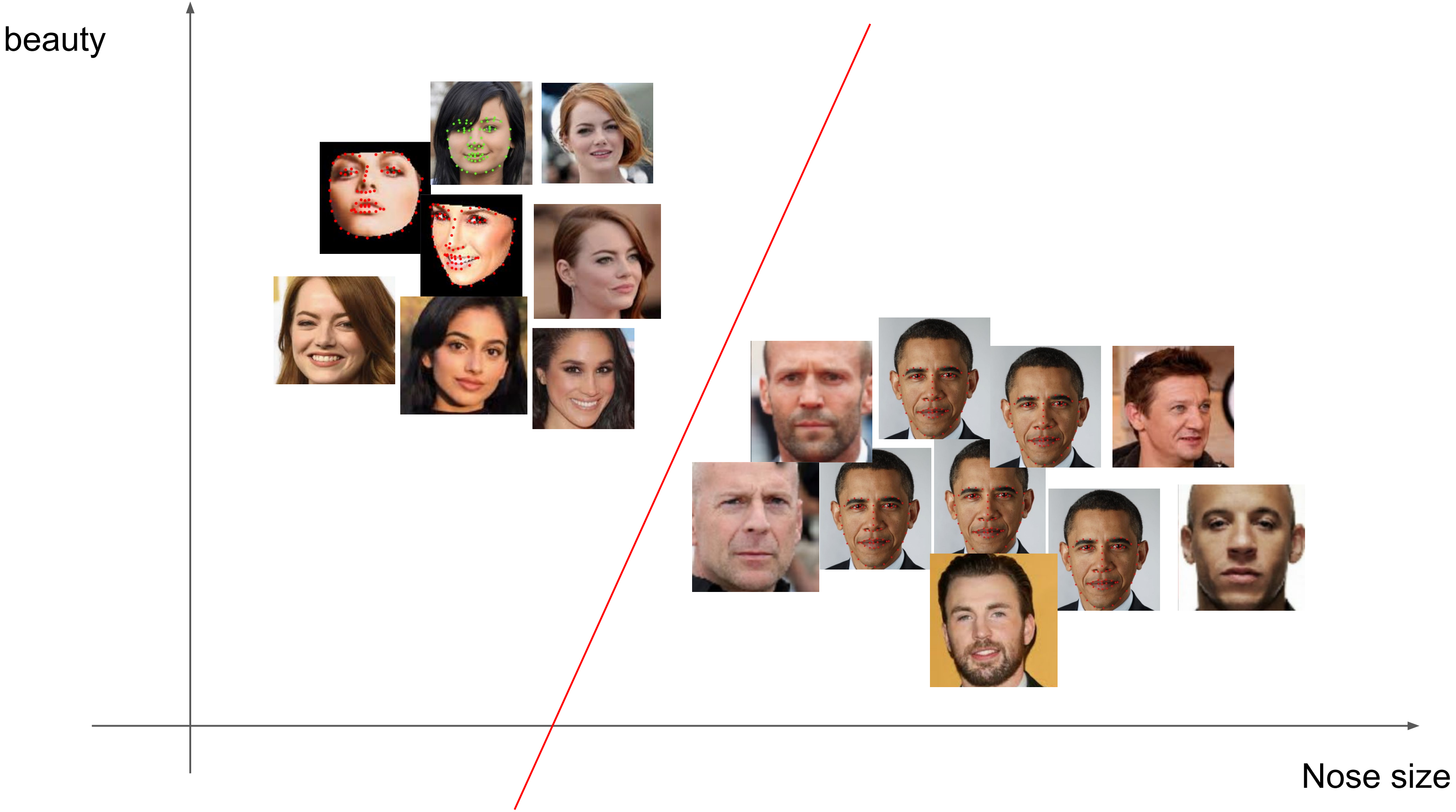
1. How does machine learning work?

2. How does facial recognition work?

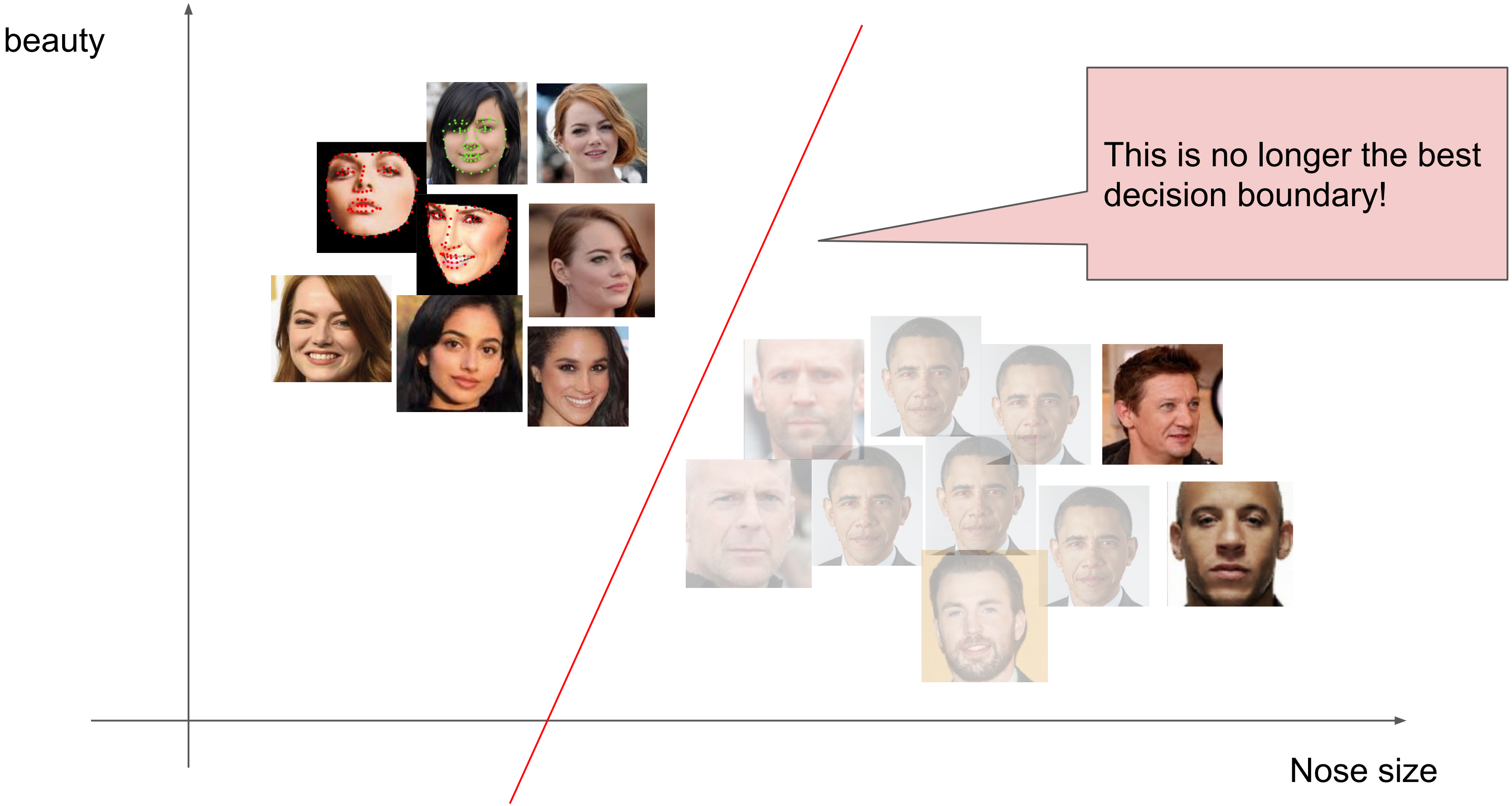
3. **What does “AI is biased” mean?**

4. How to fix biased AI?

# What does “AI is biased” mean?

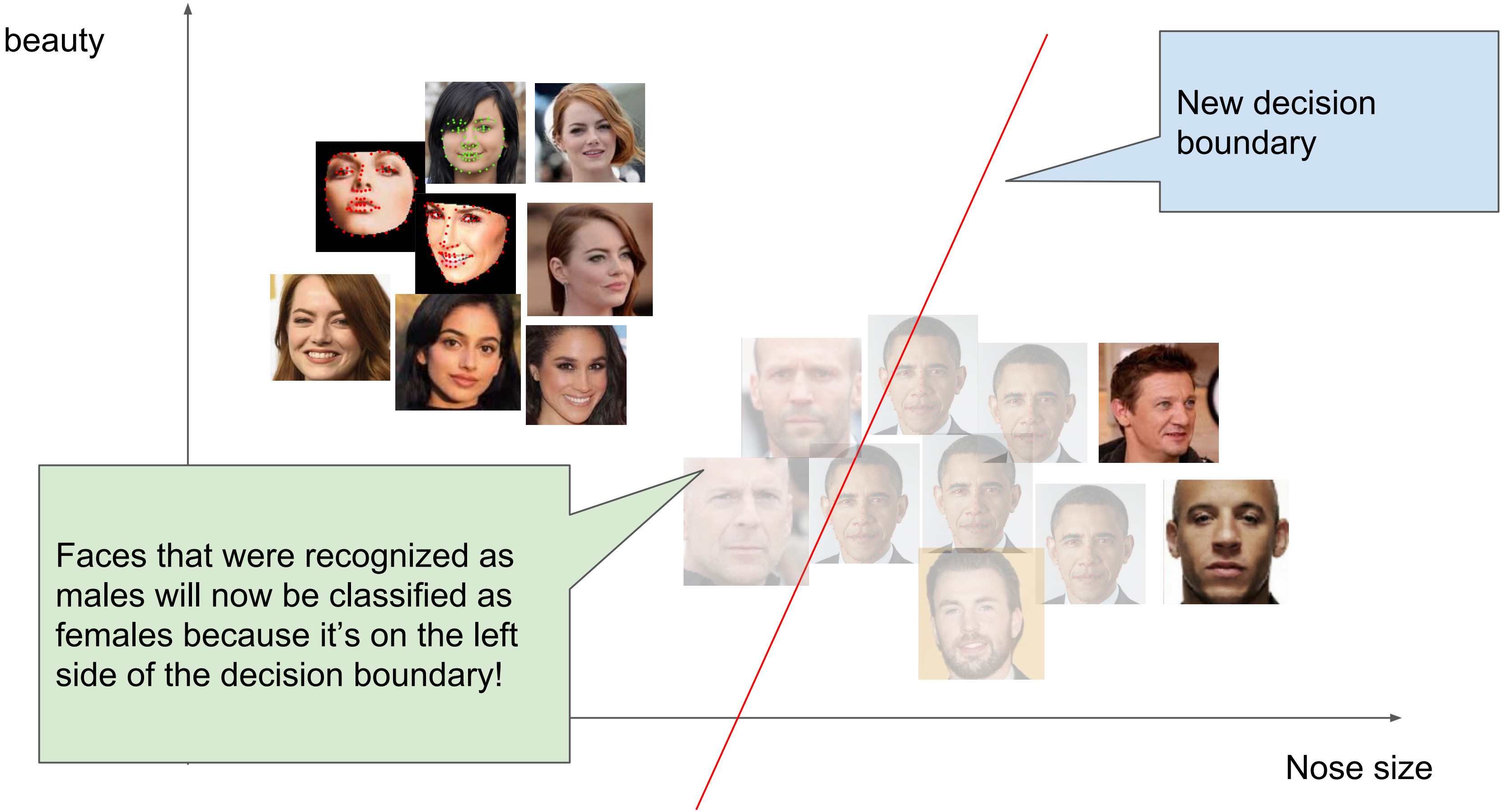


# What does “AI is biased” mean?



This is no longer the best decision boundary!

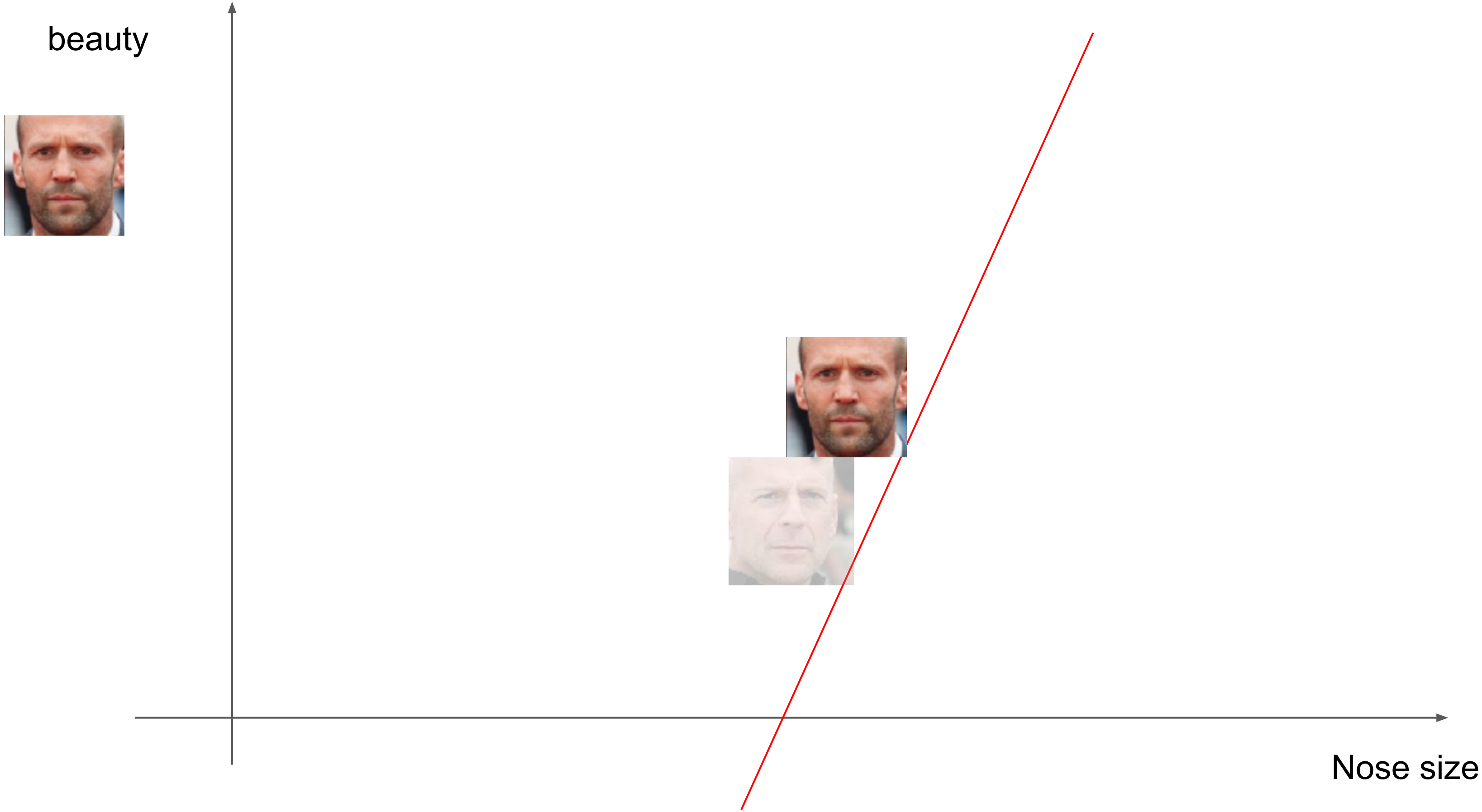
# What does “AI is biased” mean?



New decision boundary

Faces that were recognized as males will now be classified as females because it’s on the left side of the decision boundary!

# What does “AI is biased” mean?



# Roadmap

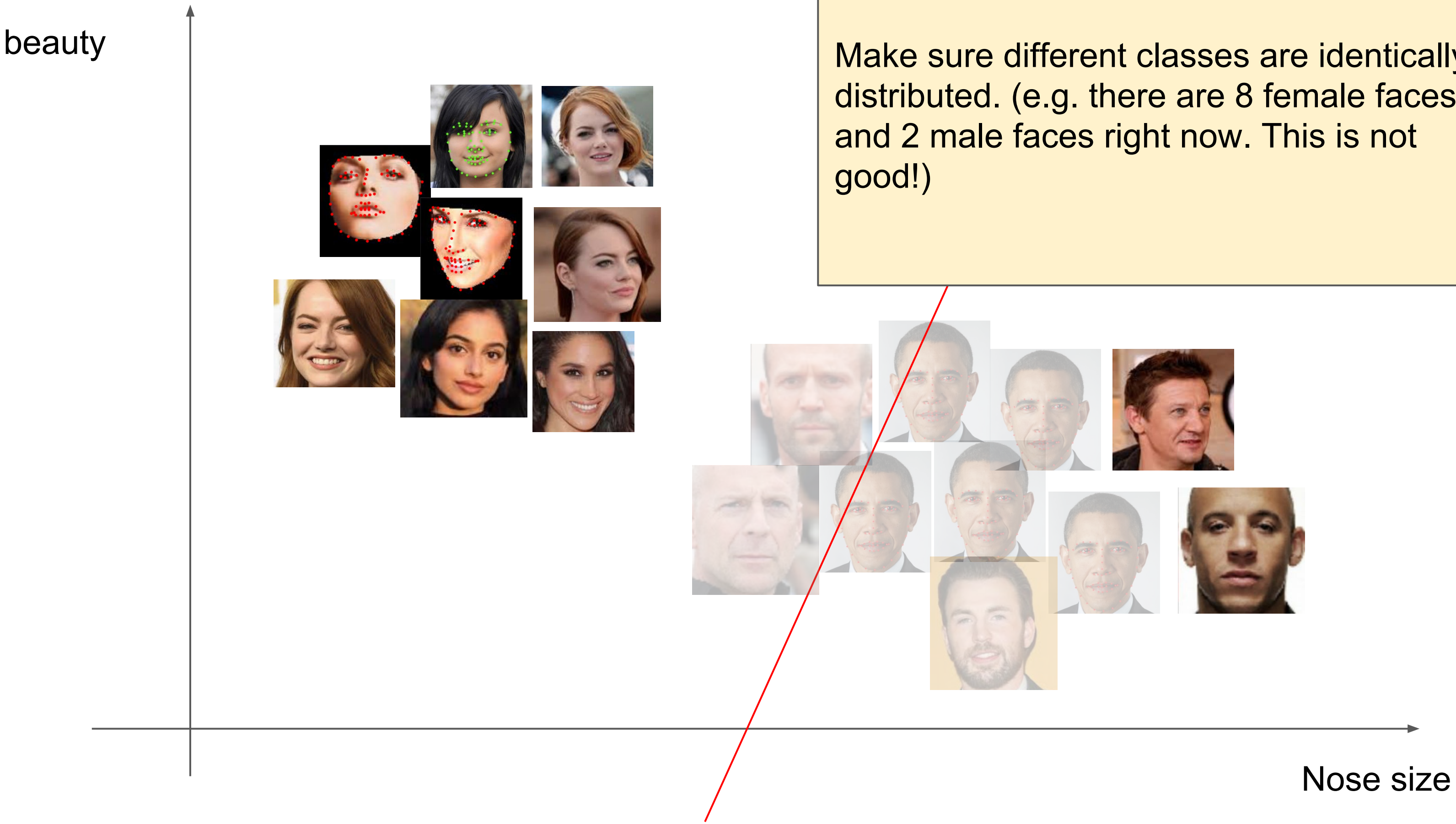
1. How does machine learning work?

2. How does facial recognition work?

3. What does “AI is biased” mean?

4. **How to fix biased AI?**

# How to fix biased AI?



Make sure different classes are identically distributed. (e.g. there are 8 female faces and 2 male faces right now. This is not good!)

## References

1. “Amazon Face-Detection Technology Shows Gender and Racial Bias, Researchers Say.” CBS News, CBS Interactive, 26 Jan. 2019, www.cbsnews.com/news/amazon-face-detection-technology-shows-gender-racial-bias-researchers-say/.
2. Daumé, Hal. “A Course in Machine Learning.” A Course in Machine Learning, http://ciml.info/dl/v0_9/ciml-v0_9-ch06.pdf.
3. Namee, Mac B et al.“The problem of bias in training data in regression problems in medical decision support.” www.scss.tcd.ie/publications/tech-reports/reports.00/TCD-CS-2000-58.pdf.