Features

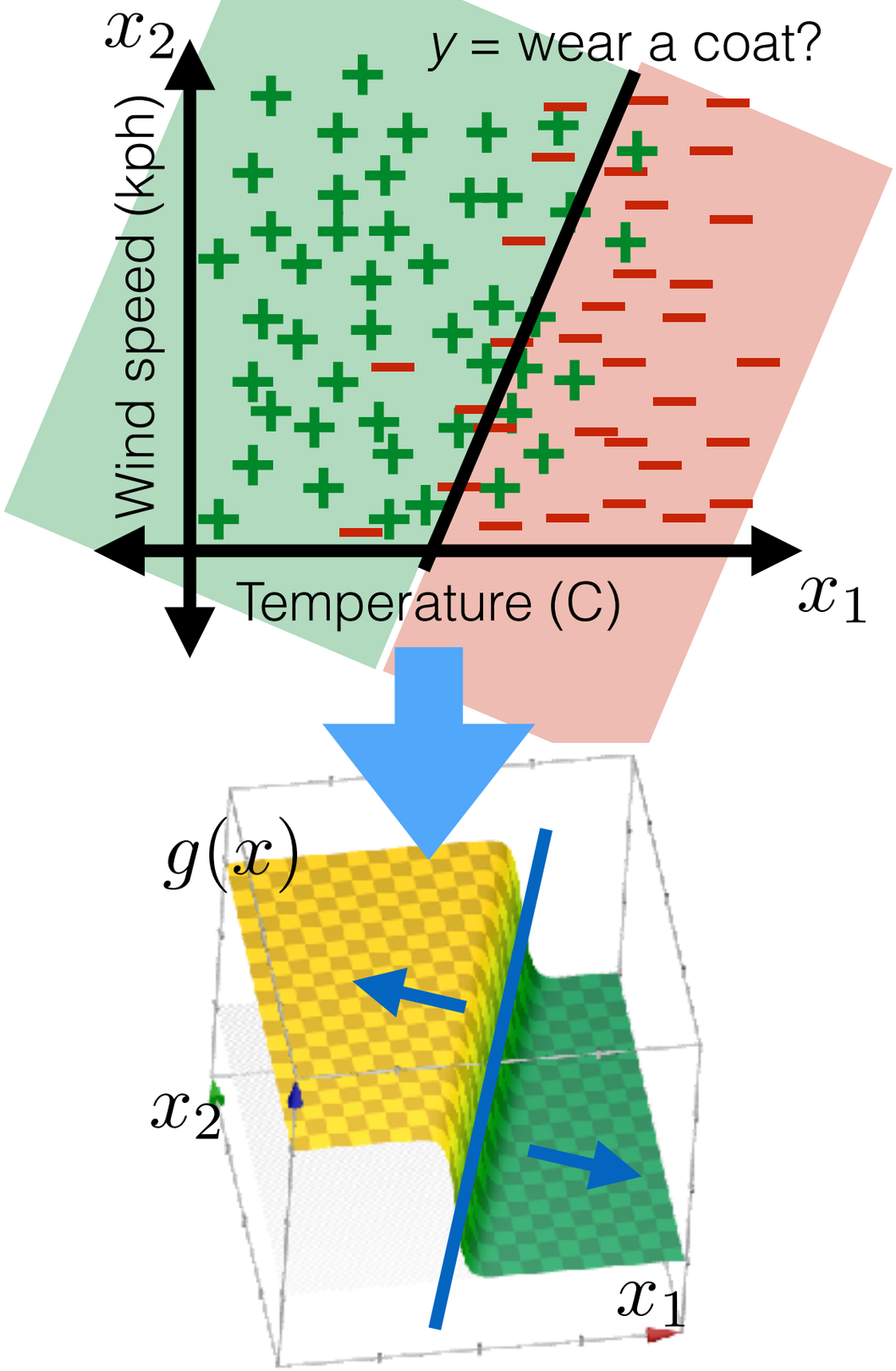
Prof. Tamara Broderick

Edited From 6.036 Fall21 Offering

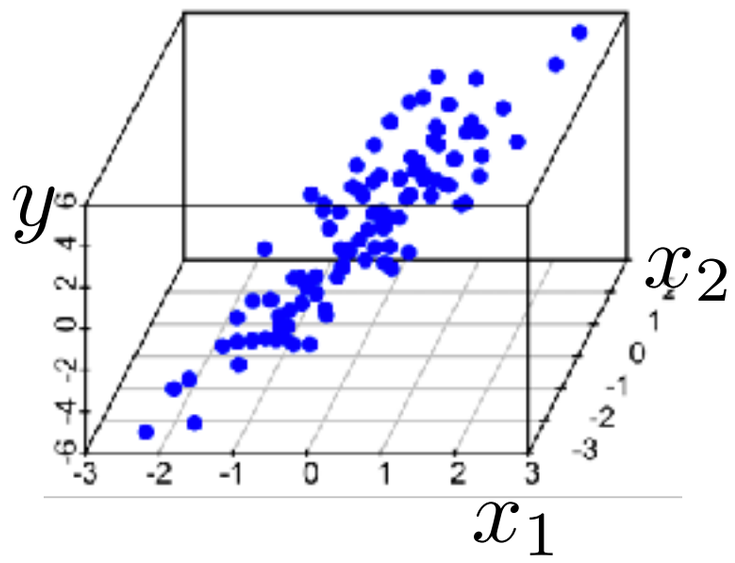
Recall

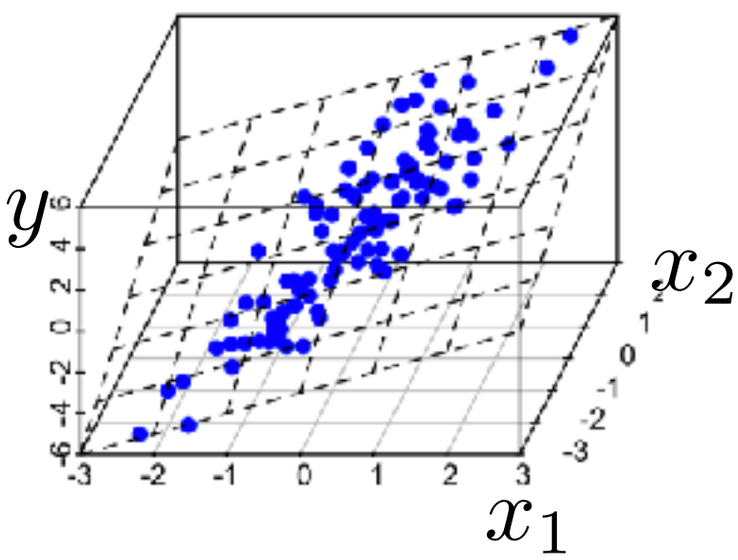
classification

- Logistic regression



- Linear regression





A more-complete ML analysis

1. Establish a goal & find data
   - Example goal: diagnose whether people have heart disease based on their available information
2. Encode data in useful form for the ML algorithm
3. Choose a loss & a regularizer. Write an objective function to optimize.
   - Example: logistic regression
     - Loss: negative log likelihood
     - Regularizer: ridge penalty (squared norm)
4. Optimize the objective function & return a hypothesis
   - Example: Gradient descent or SGD
5. Evaluation & interpretation

A machine learning (ML) analysis

- First, need goal & data. E.g. diagnose whether people have heart disease based on their available information
- Next, put data in useful form for learning algorithm

	has heart disease?	resting heart rate (bpm)	pain?	job	medicines	age	family income (USD)
1	no	55	no	nurse	pain	40s	133000
2	no	71	no	admin	beta blockers, pain	20s	34000
3	yes	89	yes	nurse	beta blockers	50s	40000
4	no	67	no	doctor	none	50s	120000

A machine learning (ML) analysis

- First, need goal & data. E.g. diagnose whether people have heart disease based on their available information
- Next, put data in useful form for learning algorithm

	has heart disease?	resting heart rate (bpm)	pain?	job	medicines	age	family income (USD)
1	no	55	no	nurse	pain	40s	133000
2	no	71	no	admin	beta blockers, pain	20s	34000
3	yes	89	yes	nurse	beta blockers	50s	40000
4	no	67	no	doctor	none	50s	120000

A machine learning (ML) analysis

- First, need goal & data. E.g. diagnose whether people have heart disease based on their available information
- Next, put data in useful form for learning algorithm

	has heart disease?	resting heart rate (bpm)	pain?	job	medicines	age	family income (USD)
1	no	55	no	nurse	pain	40s	133000
2	no	71	no	admin	beta blockers, pain	20s	34000
3	yes	89	yes	nurse	beta blockers	50s	40000
4	no	67	no	doctor	none	50s	120000

Encode data in usable form

- Identify the labels and encode as real numbers

	has heart disease?
1	no
2	no
3	yes
4	no

$$\text{`yes'} \leftrightarrow 1$$
$$\text{`no'} \leftrightarrow 0$$

1	0 $= y^{(1)}$
2	0
3	1
4	0

- Depending on your algorithm, might instead use $\{+1, -1\}$
- Save mapping to recover predictions of new points

Encode data in usable form

- Identify the features and encode as real numbers
- Feature: any function of the data (except labels)
- Today, old features: x; new features: $\phi(x)$

$(x^{(1)})^\top$

	resting heart rate (bpm)	pain?	job	medicines	age	family income (USD)
1	55	no	nurse	pain	40s	133000
2	71	no	admin	beta blockers, pain	20s	34000
3	89	yes	nurse	beta blockers	50s	40000
4	67	no	doctor	none	50s	120000

Encode data in usable form

- Identify the features and encode as real numbers
- Feature: any function of the data (except labels)
- Today, old features: x; new features: $\phi(x)$

	resting heart rate (bpm)	pain?	job	medicines	age	family income (USD)
1	55	no	nurse	pain	40s	133000
2	71	no	admin	beta blockers, pain	20s	34000
3	89	yes	nurse	beta blockers	50s	40000
4	67	no	doctor	none	50s	120000

Encode data in usable form

- Identify the features and encode as real numbers
- Feature: any function of the data (except labels)
- Today, old features: x; new features: $\phi(x)$

	resting heart rate (bpm)	pain?	job	medicines	age	family income (USD)
1	55	no	nurse	pain	40s	133000
2	71	no	admin	beta blockers, pain	20s	34000
3	89	yes	nurse	beta blockers	50s	40000
4	67	no	doctor	none	50s	120000

Encode data in usable form

- Identify the features and encode as real numbers
- Feature: any function of the data (except labels)
- Today, old features: x; new features: $\phi(x)$

	resting heart rate (bpm)	pain?	job	medicines	age	family income (USD)
1	55	0	nurse	pain	40s	133000
2	71	0	admin	beta blockers, pain	20s	34000
3	89	1	nurse	beta blockers	50s	40000
4	67	0	doctor	none	50s	120000

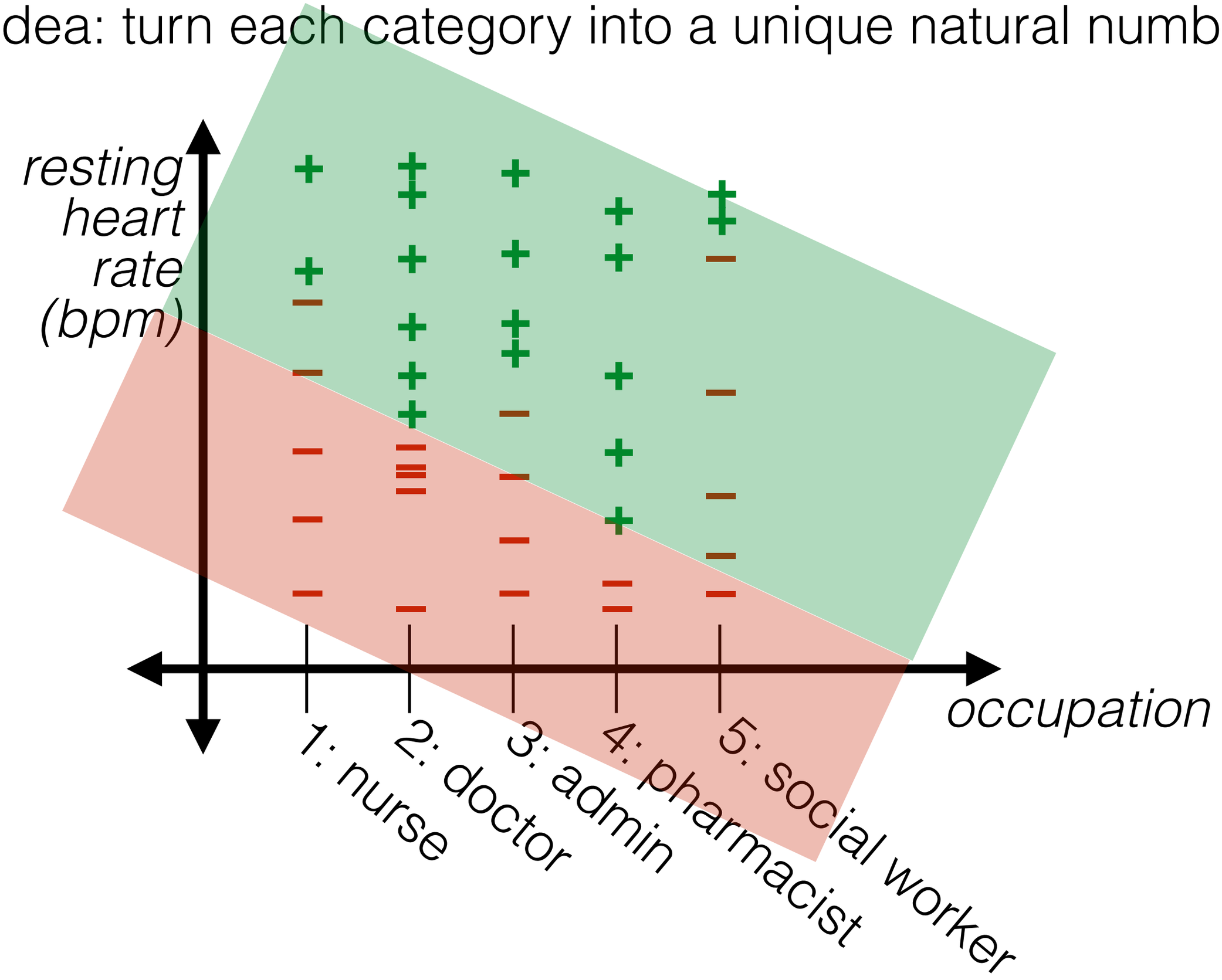
Encode data in usable form

- Identify the features and encode as real numbers
- Feature: any function of the data (except labels)
- Today, old features: x; new features: $\phi(x)$

	resting heart rate (bpm)	pain?	job	medicines	age	family income (USD)
1	55	0	nurse	pain	40s	133000
2	71	0	admin	beta blockers, pain	20s	34000
3	89	1	nurse	beta blockers	50s	40000
4	67	0	doctor	none	50s	120000

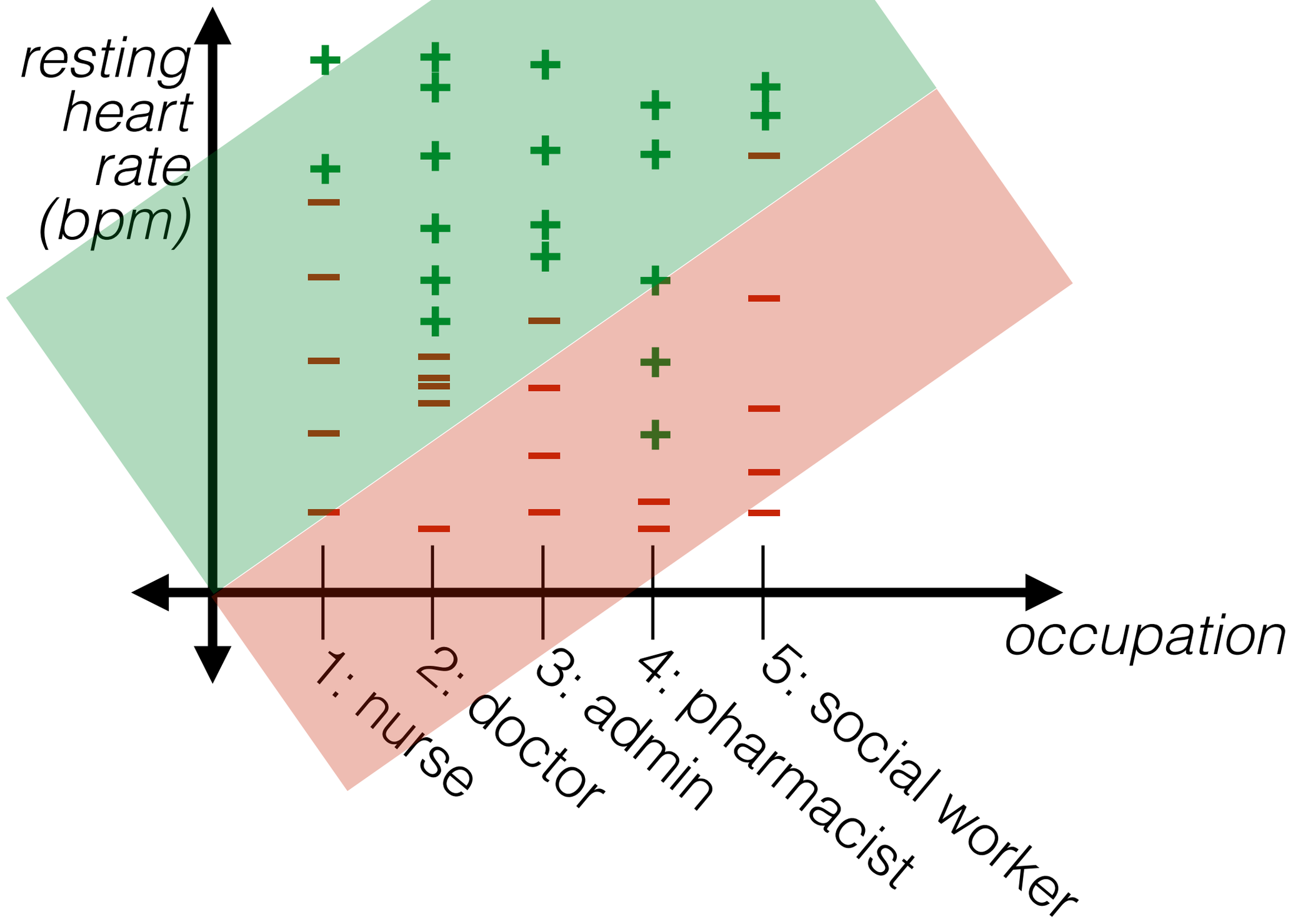
Encode categorical data

- Idea: turn each category into a unique natural number



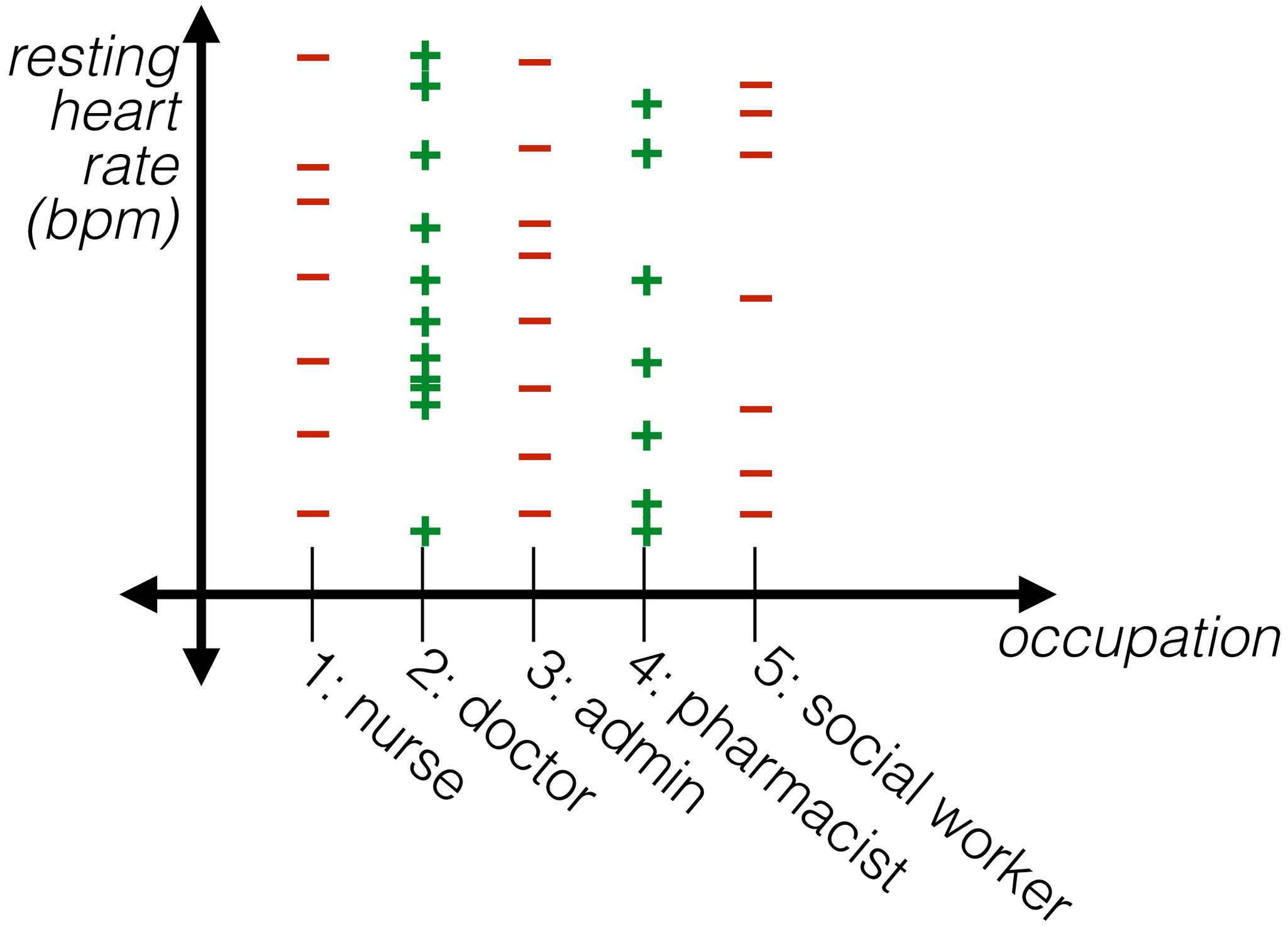
Encode categorical data

- Idea: turn each category into a unique natural number



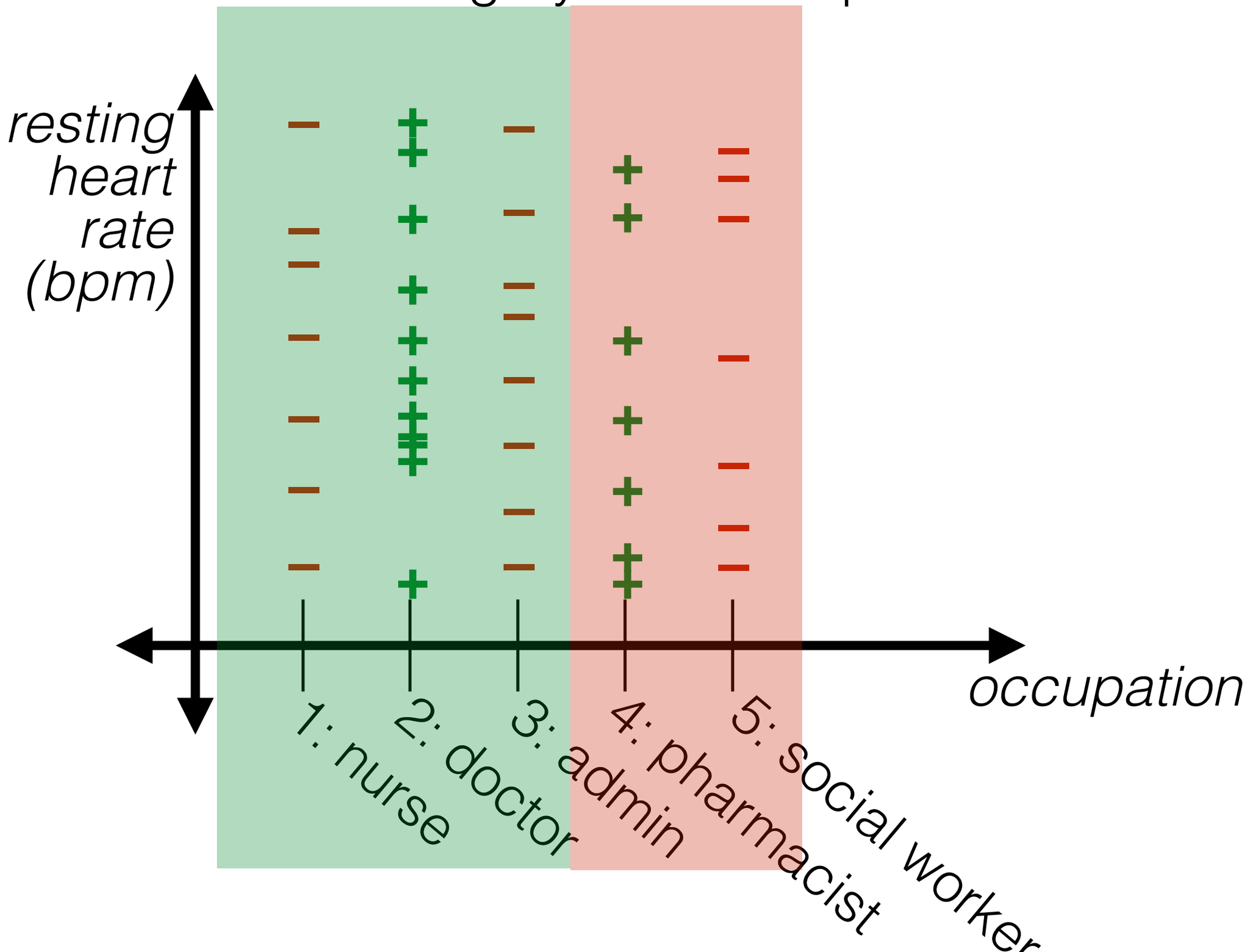
Encode categorical data

- Idea: turn each category into a unique natural number



Encode categorical data

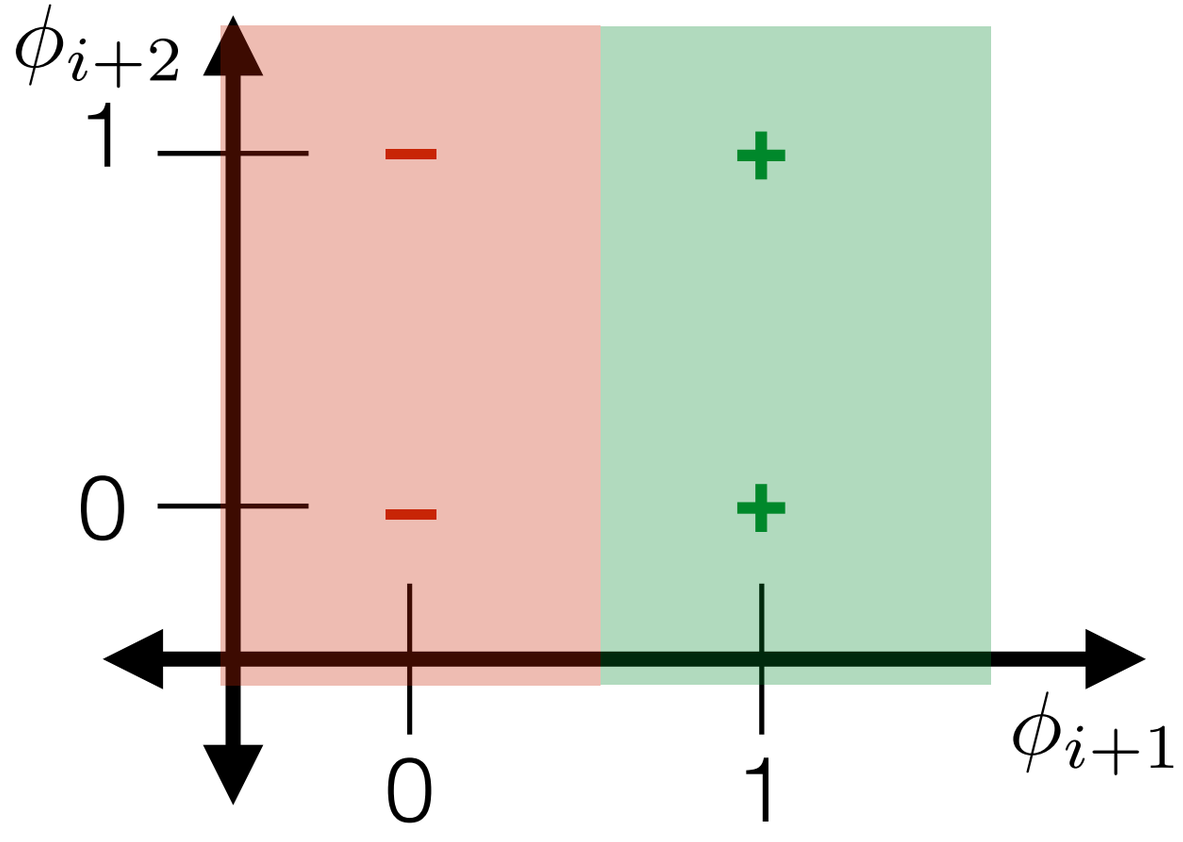
- Idea: turn each category into a unique natural number



Encode categorical data

- Idea: turn each category into a unique binary number

	ϕ_i	ϕ_{i+1}	ϕ_{i+2}
nurse	0	0	0
admin	0	0	1
pharmacist	0	1	0
doctor	0	1	1
social worker	1	0	0

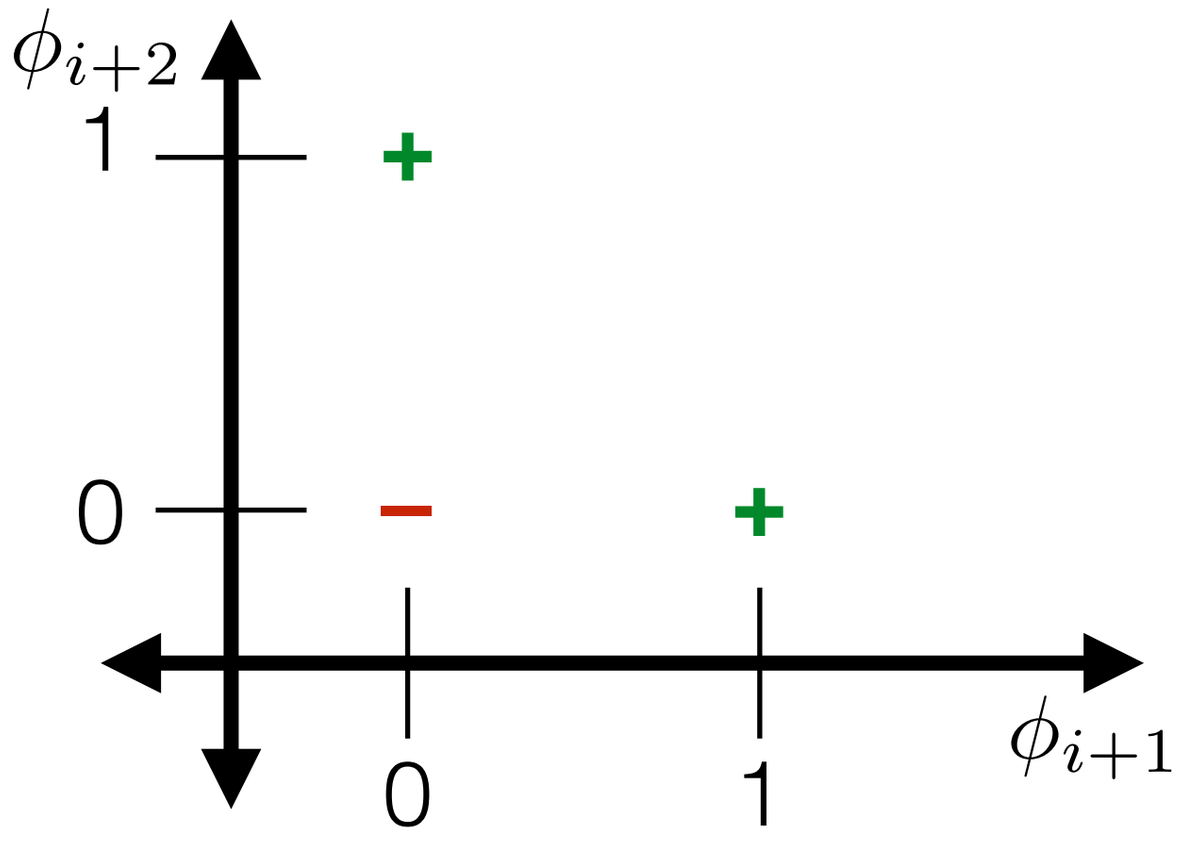


Encode categorical data

- Idea: turn each category into own unique 0-1 feature

	ϕ_i	ϕ_{i+1}	ϕ_{i+2}	ϕ_{i+3}	ϕ_{i+4}
nurse	1	0	0	0	0
admin	0	1	0	0	0
pharmacist	0	0	1	0	0
doctor	0	0	0	1	0
social worker	0	0	0	0	1

- "one-hot encoding"



Encode data in usable form

- Identify the features and encode as real numbers

	resting heart rate (bpm)	pain?	job	medicines	age	family income (USD)
1	55	0	nurse	pain	40s	133000
2	71	0	admin	beta blockers, pain	20s	34000
3	89	1	nurse	beta blockers	50s	40000
4	67	0	doctor	none	50s	120000

Encode data in usable form

- Identify the features and encode as real numbers

	resting heart rate (bpm)	pain?	j1,j2,j3,j4,j5	medicines	age	family income (USD)
1	55	0	1,0,0,0,0	pain	40s	133000
2	71	0	0,1,0,0,0	beta blockers, pain	20s	34000
3	89	1	1,0,0,0,0	beta blockers	50s	40000
4	67	0	0,0,0,1,0	none	50s	120000

Encode data in usable form

- Identify the features and encode as real numbers

	resting heart rate (bpm)	pain?	j1,j2,j3,j4,j5	medicines	age	family income (USD)
1	55	0	1,0,0,0,0	pain	40s	133000
2	71	0	0,1,0,0,0	beta blockers, pain	20s	34000
3	89	1	1,0,0,0,0	beta blockers	50s	40000
4	67	0	0,0,0,1,0	none	50s	120000

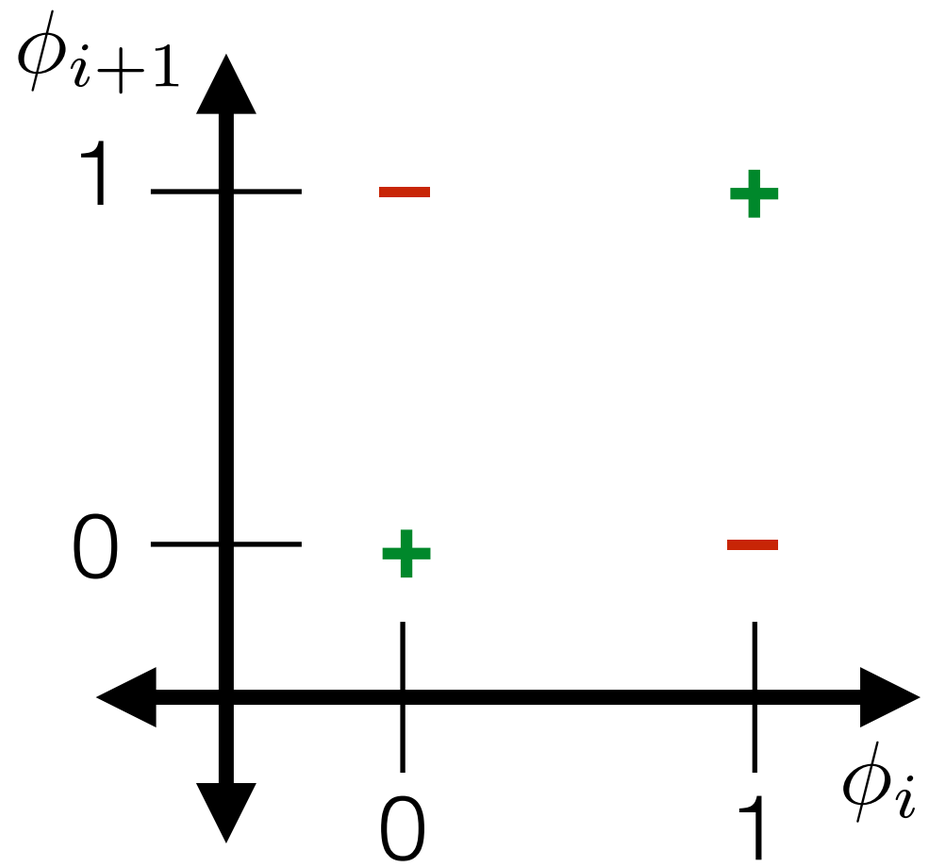
Encode categorical data

- Should we use one-hot encoding?

	ϕ_i	ϕ_{i+1}	ϕ_{i+2}	ϕ_{i+3}
pain	1	0	0	0
pain & beta blockers	0	1	0	0
beta blockers	0	0	1	0
no medications	0	0	0	1

- Idea: factored encoding

	ϕ_i	ϕ_{i+1}
pain	1	0
pain & beta blockers	1	1
beta blockers	0	1
no medications	0	0



Encode data in usable form

- Identify the features and encode as real numbers

	resting heart rate (bpm)	pain?	j1,j2,j3,j4,j5	medicines	age	family income (USD)
1	55	0	1,0,0,0,0	pain	40s	133000
2	71	0	0,1,0,0,0	beta blockers, pain	20s	34000
3	89	1	1,0,0,0,0	beta blockers	50s	40000
4	67	0	0,0,0,1,0	none	50s	120000

Encode data in usable form

- Identify the features and encode as real numbers

	resting heart rate (bpm)	pain?	j1,j2,j3,j4,j5	m1, m2	age	family income (USD)
1	55	0	1,0,0,0,0	1,0	40s	133000
2	71	0	0,1,0,0,0	1,1	20s	34000
3	89	1	1,0,0,0,0	0,1	50s	40000
4	67	0	0,0,0,1,0	0,0	50s	120000

Encode data in usable form

- Identify the features and encode as real numbers

	resting heart rate (bpm)	pain?	j1,j2,j3,j4,j5	m1, m2	age	family income (USD)
1	55	0	1,0,0,0,0	1,0	40s	133000
2	71	0	0,1,0,0,0	1,1	20s	34000
3	89	1	1,0,0,0,0	0,1	50s	40000
4	67	0	0,0,0,1,0	0,0	50s	120000

Encode data in usable form

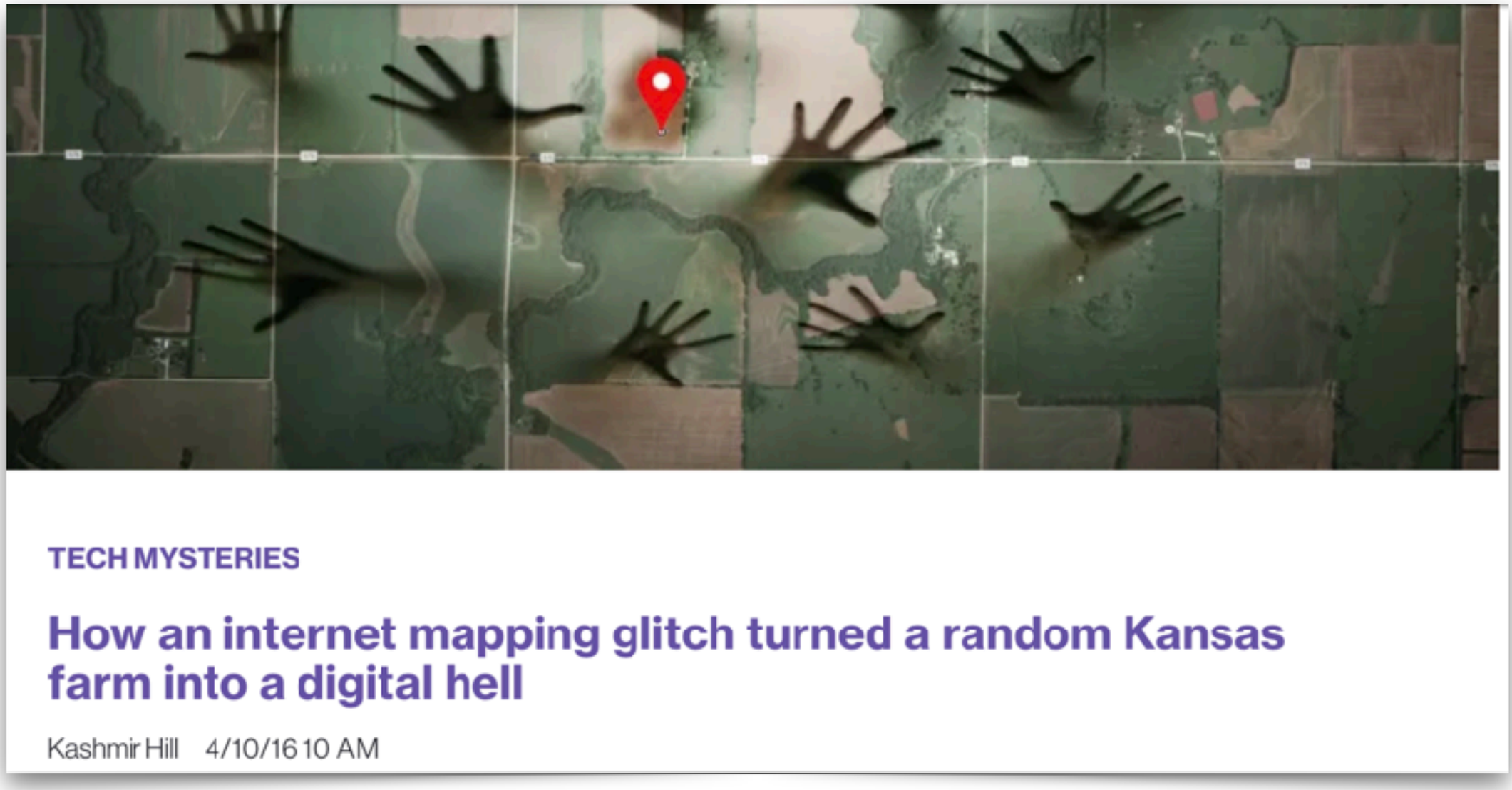
- Identify the features and encode as real numbers

	resting heart rate (bpm)	pain?	j1,j2,j3,j4,j5	m1, m2	age	family income (USD)
1	55	0	1,0,0,0,0	1,0	45	133000
2	71	0	0,1,0,0,0	1,1	25	34000
3	89	1	1,0,0,0,0	0,1	55	40000
4	67	0	0,0,0,1,0	0,0	55	120000

Using a representative # for a range

- Potential pitfall: level of detail might be treated as meaningful (by you or others using the data)
- A way to diagnose many problems: plot your data!

age
45
25
55
55



TECH MYSTERIES

How an internet mapping glitch turned a random Kansas farm into a digital hell

Kashmir Hill 4/10/16 10 AM

[https://fusion.tv/story/287592/internet-mapping-glitch-kansas-farm/?utm_medium=social]

Encode data in usable form

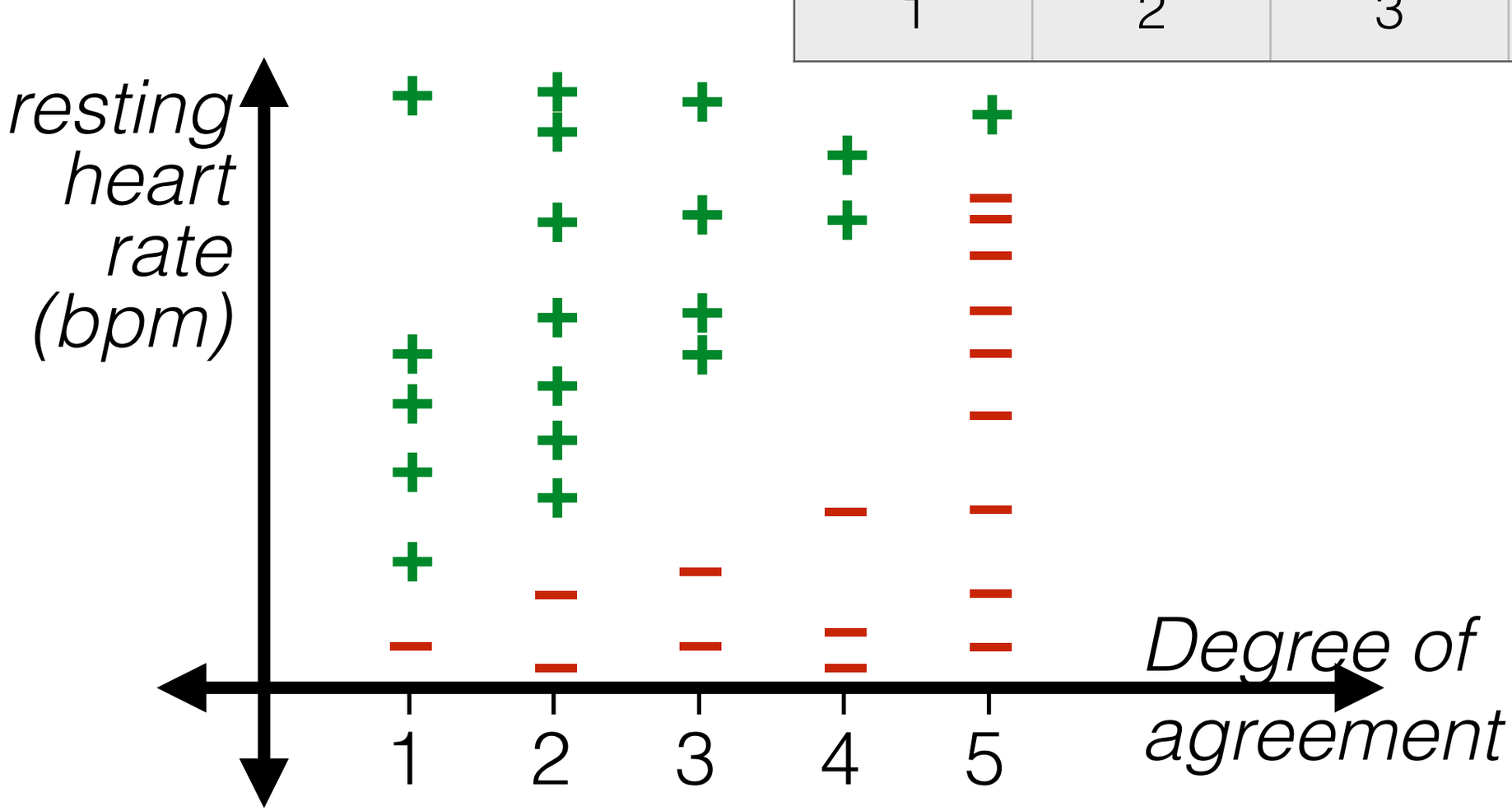
- Identify the features and encode as real numbers

	resting heart rate (bpm)	pain?	j1,j2,j3,j4,j5	m1, m2	decade	family income (USD)
1	55	0	1,0,0,0,0	1,0	4	133000
2	71	0	0,1,0,0,0	1,1	2	34000
3	89	1	1,0,0,0,0	0,1	5	40000
4	67	0	0,0,0,1,0	0,0	5	120000

Encode ordinal data

- Numerical data: order on data values, and differences in value are meaningful
- Categorical data: no order on data values
- Ordinal data: order on data values, but differences not meaningful
  - E.g. Likert scale:

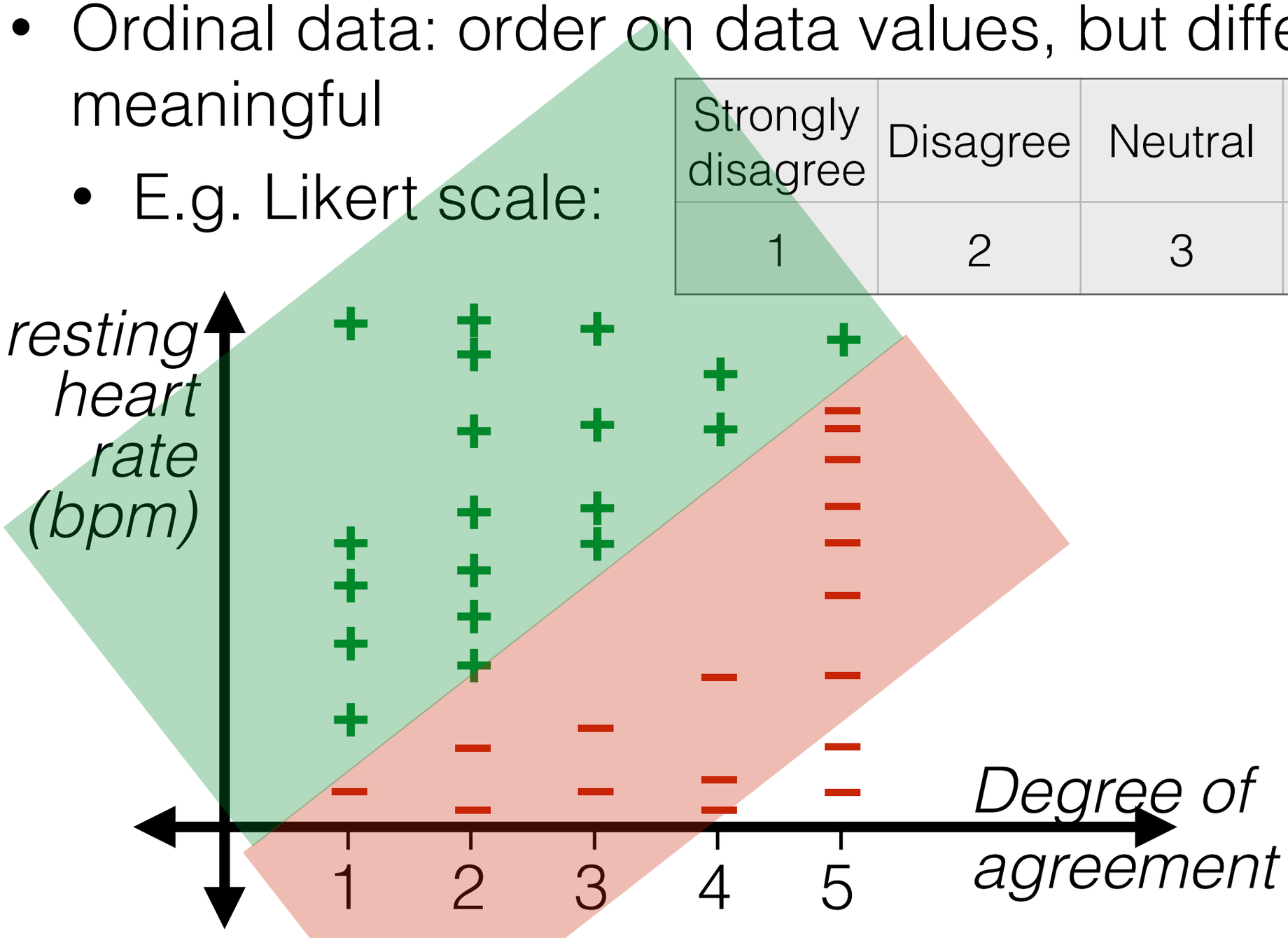
Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1	2	3	4	5



Encode ordinal data

- Numerical data: order on data values, and differences in value are meaningful
- Categorical data: no order on data values
- Ordinal data: order on data values, but differences not meaningful
  - E.g. Likert scale:

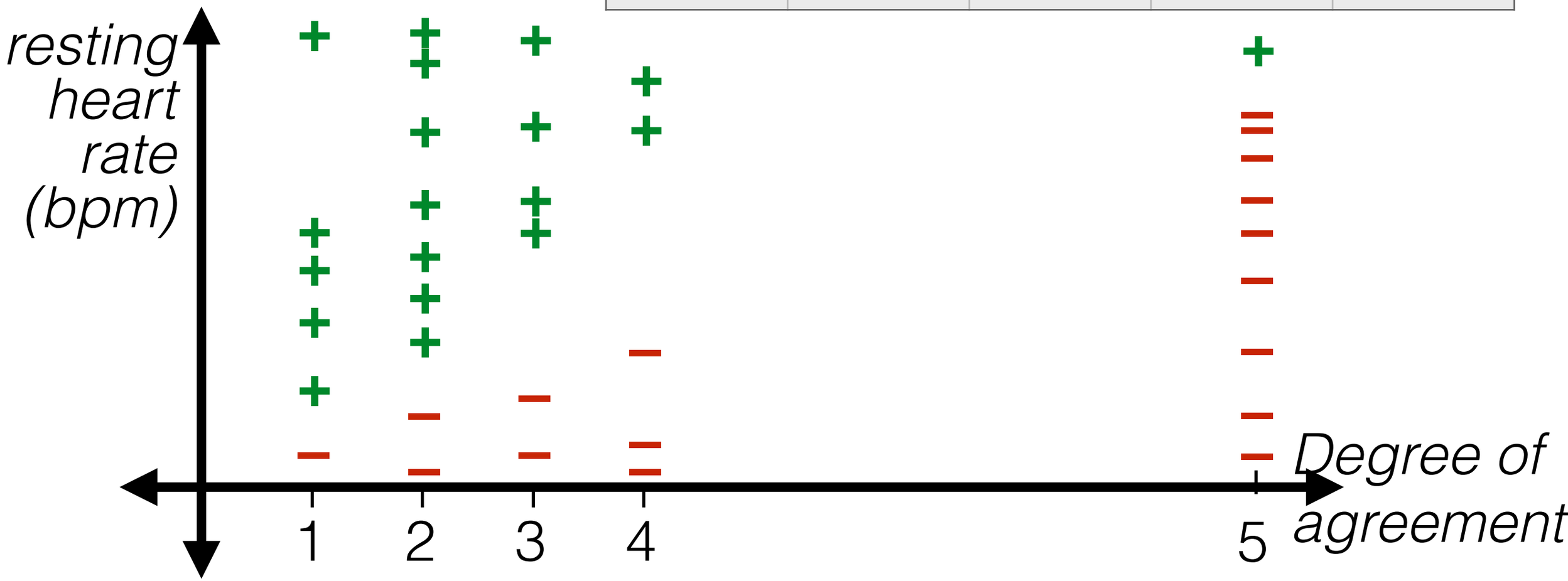
Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1	2	3	4	5



Encode ordinal data

- Numerical data: order on data values, and differences in value are meaningful
- Categorical data: no order on data values
- Ordinal data: order on data values, but differences not meaningful
  - E.g. Likert scale:

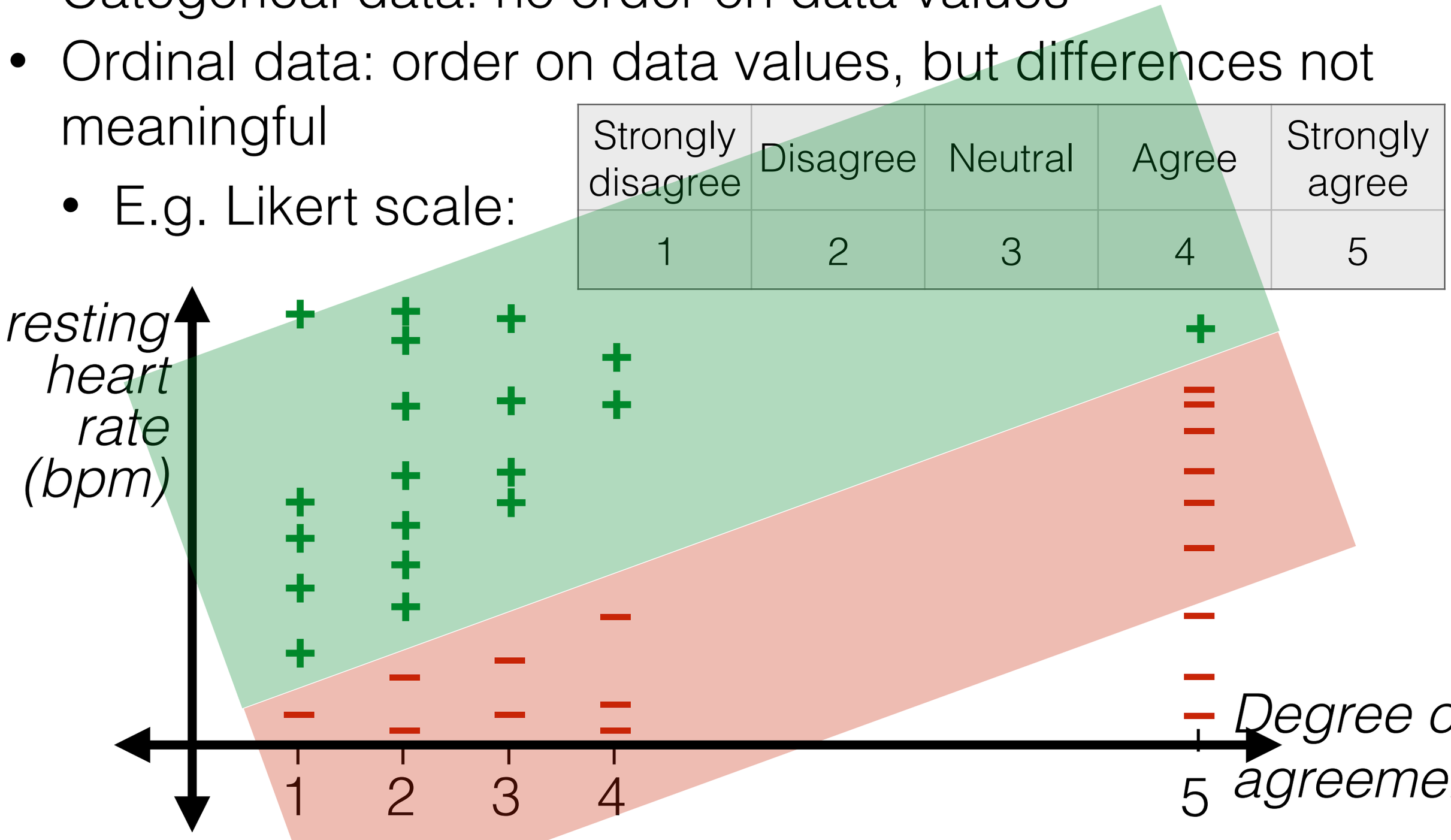
Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1	2	3	4	5



Encode ordinal data

- Numerical data: order on data values, and differences in value are meaningful
- Categorical data: no order on data values
- Ordinal data: order on data values, but differences not meaningful
  - E.g. Likert scale:

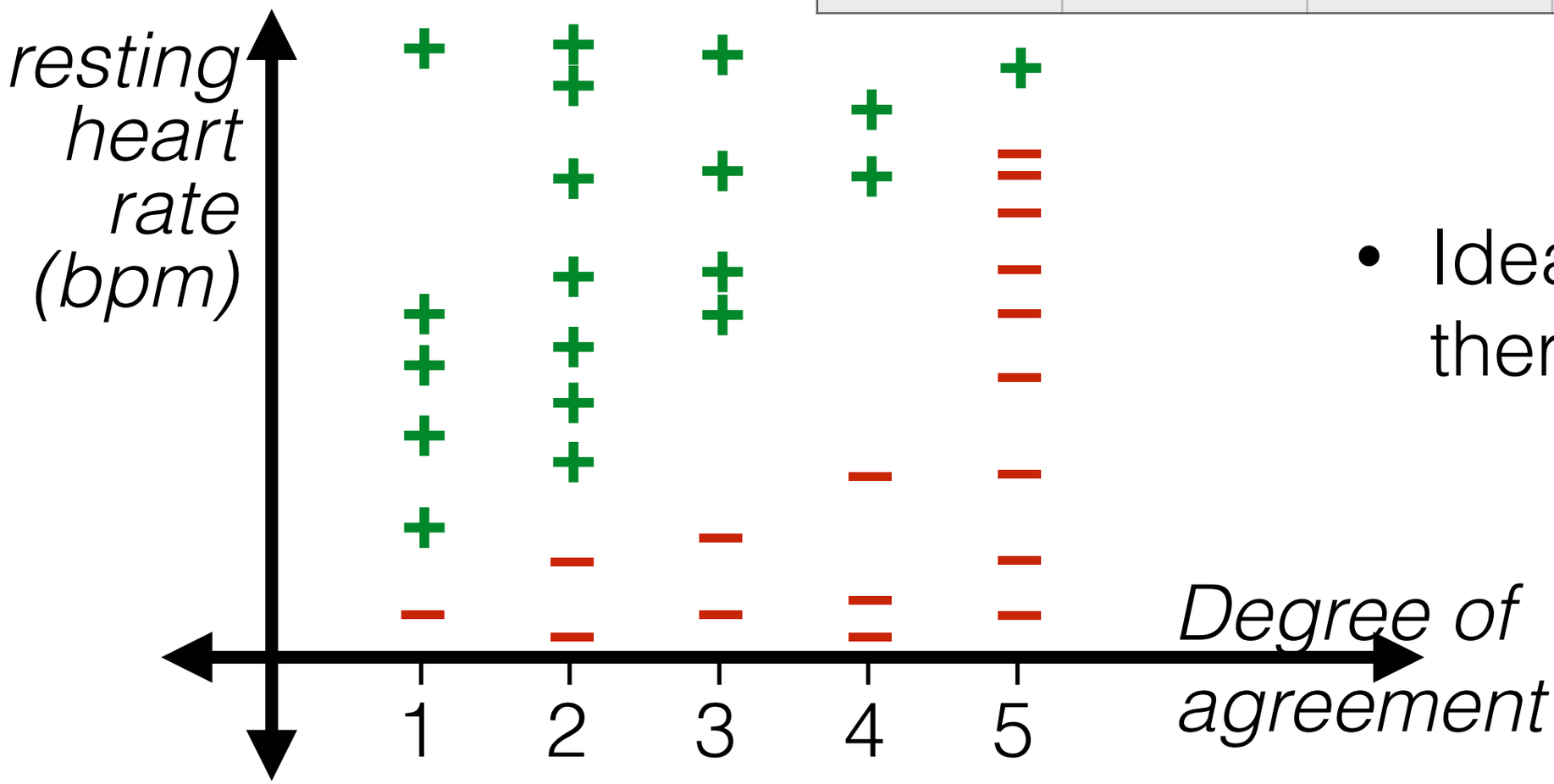
Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1	2	3	4	5



Encode ordinal data

- Numerical data: order on data values, and differences in value are meaningful
- Categorical data: no order on data values
- Ordinal data: order on data values, but differences not meaningful
  - E.g. Likert scale:

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1	2	3	4	5

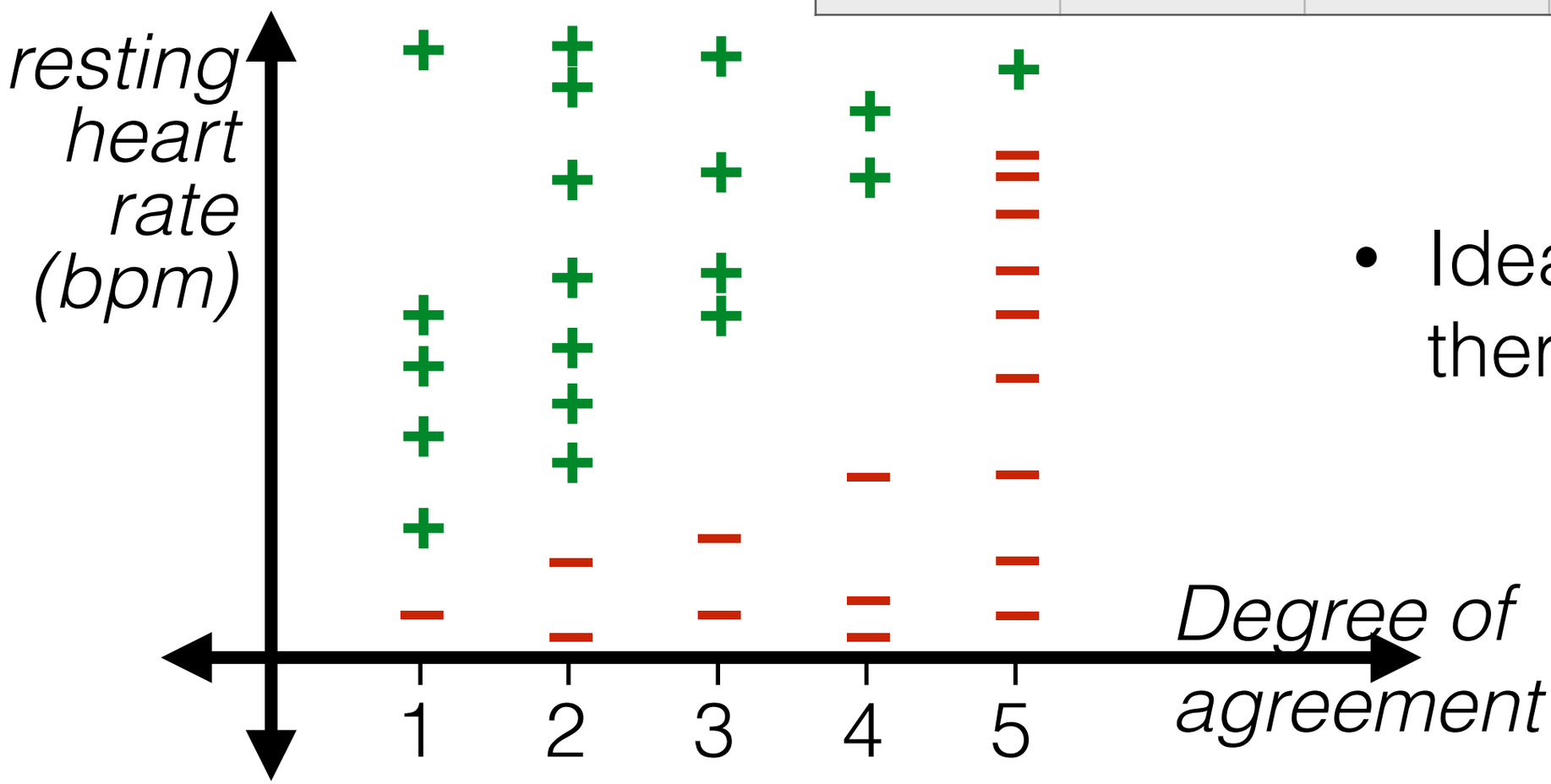


- Idea: Unary/ thermometer code

Encode ordinal data

- Numerical data: order on data values, and differences in value are meaningful
- Categorical data: no order on data values
- Ordinal data: order on data values, but differences not meaningful
  - E.g. Likert scale:

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1,0,0,0,0	1,1,0,0,0	1,1,1,0,0	1,1,1,1,0	1,1,1,1,1



- Idea: Unary/ thermometer code

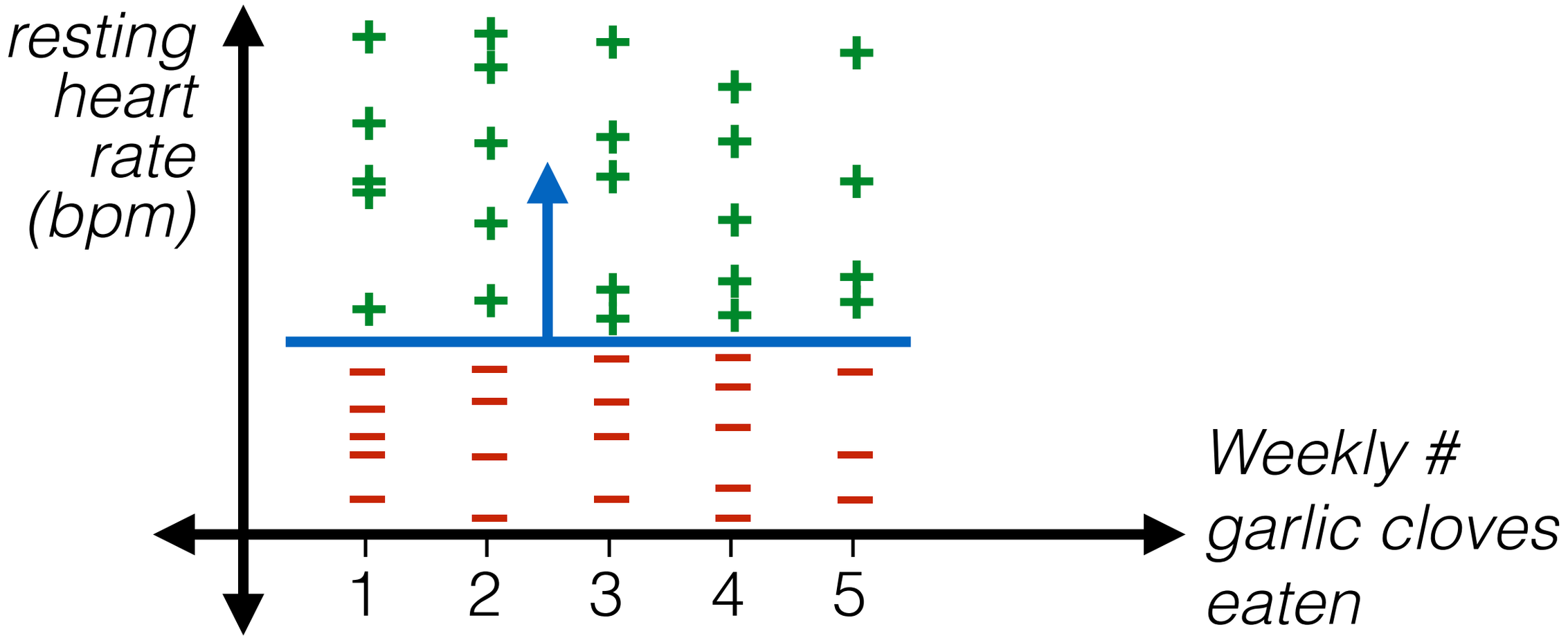
Encode data in usable form

- Identify the features and encode as real numbers

	resting heart rate (bpm)	pain?	j1,j2,j3,j4,j5	m1, m2	decade	family income (USD)
1	55	0	1,0,0,0,0	1,0	4	133000
2	71	0	0,1,0,0,0	1,1	2	34000
3	89	1	1,0,0,0,0	0,1	5	40000
4	67	0	0,0,0,1,0	0,0	5	120000

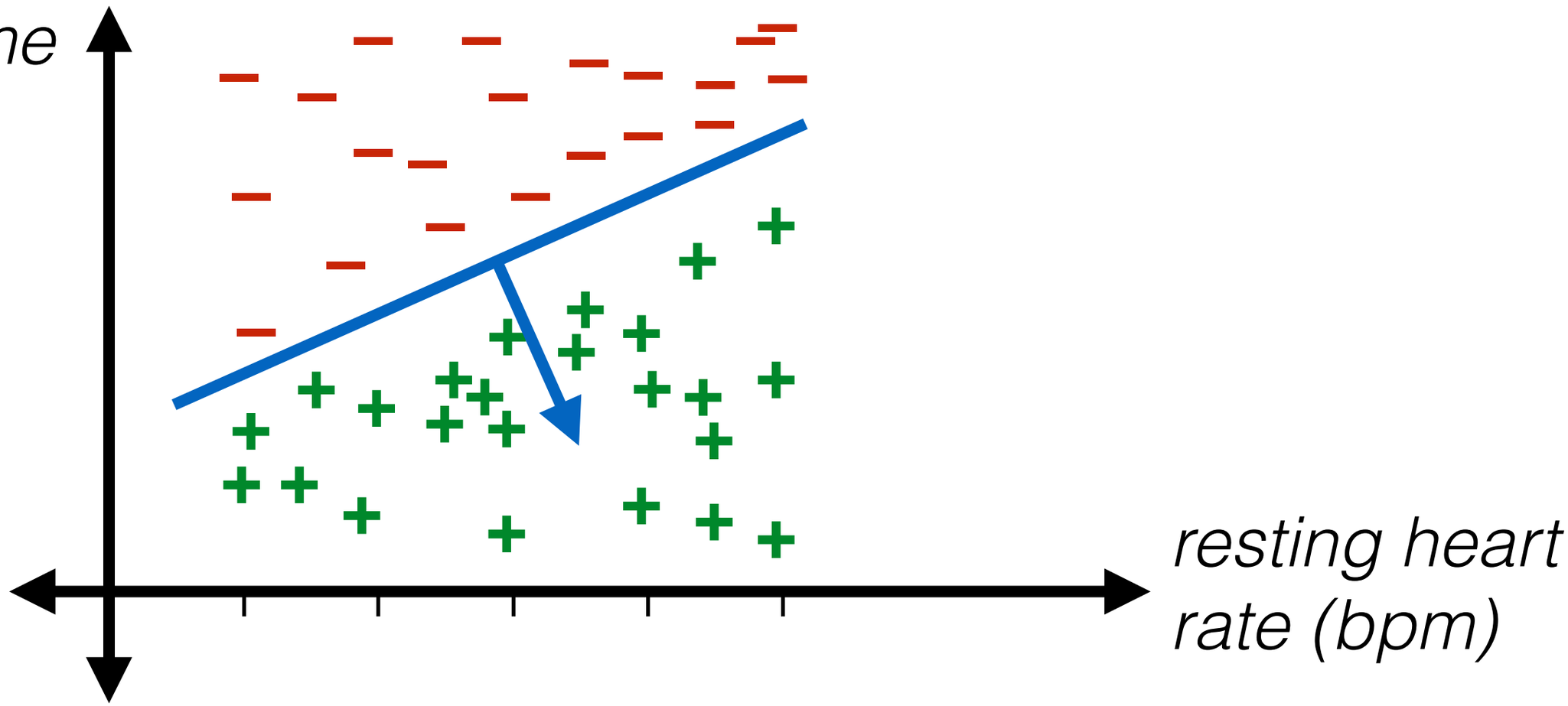
Encode numerical data

- A closer look at the output of a linear classifier



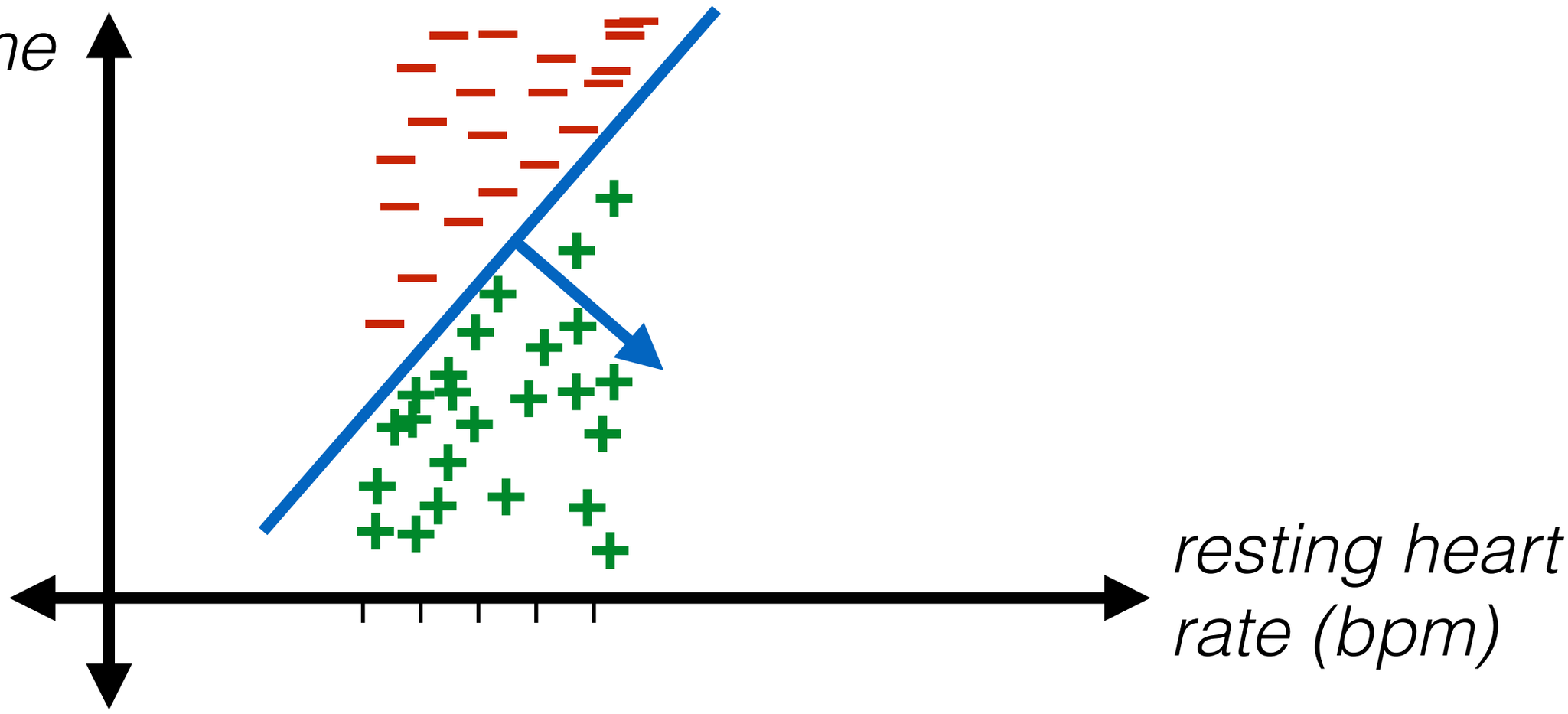
Encode numerical data

- A closer look at the output of a linear classifier



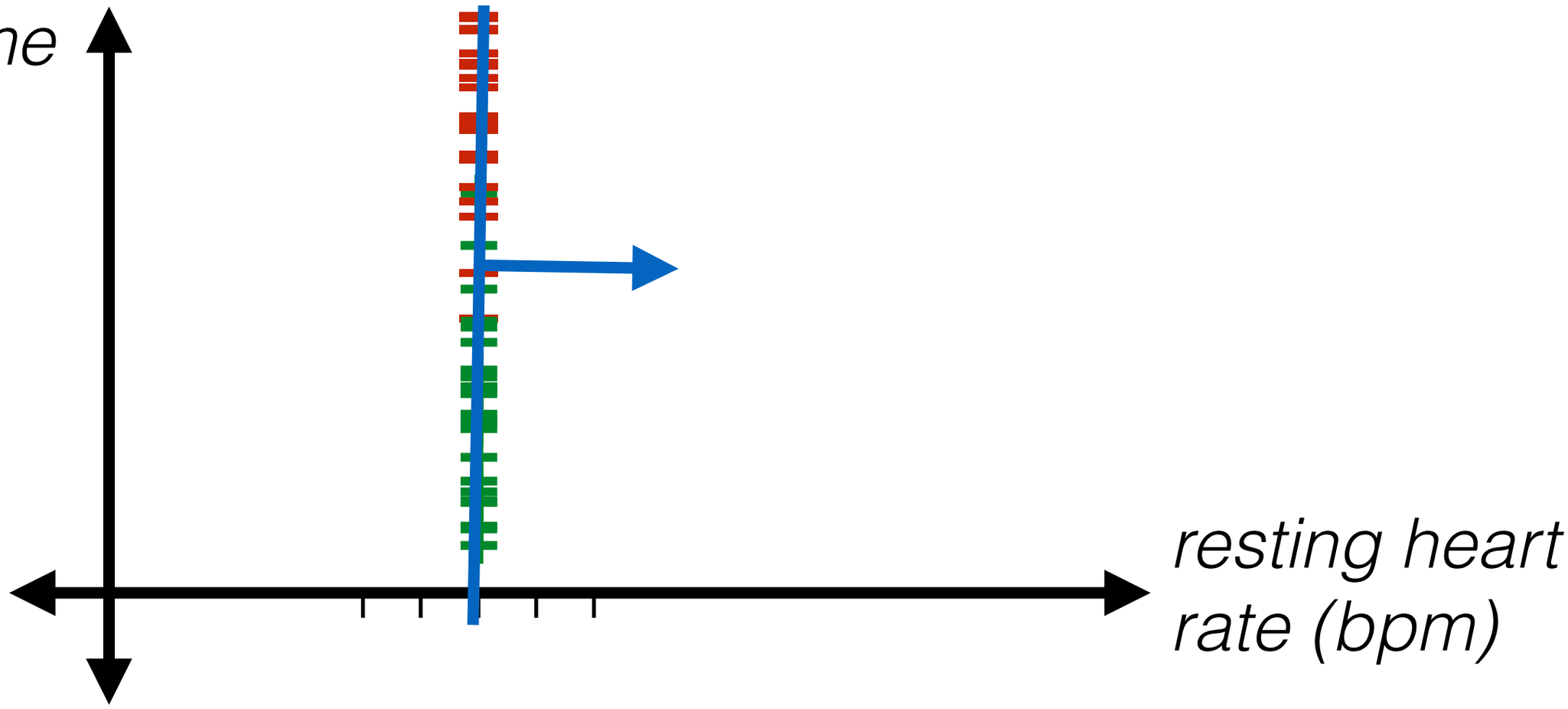
Encode numerical data

- A closer look at the output of a linear classifier



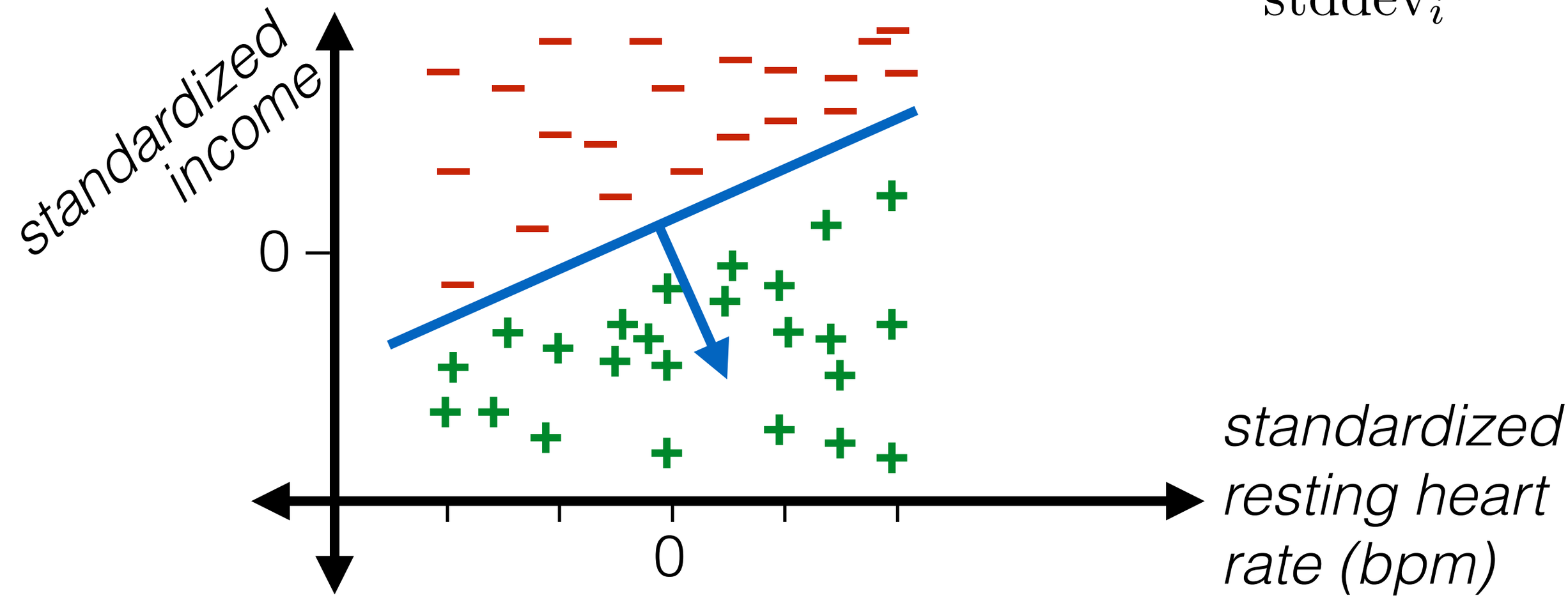
Encode numerical data

- A closer look at the output of a linear classifier



Encode numerical data

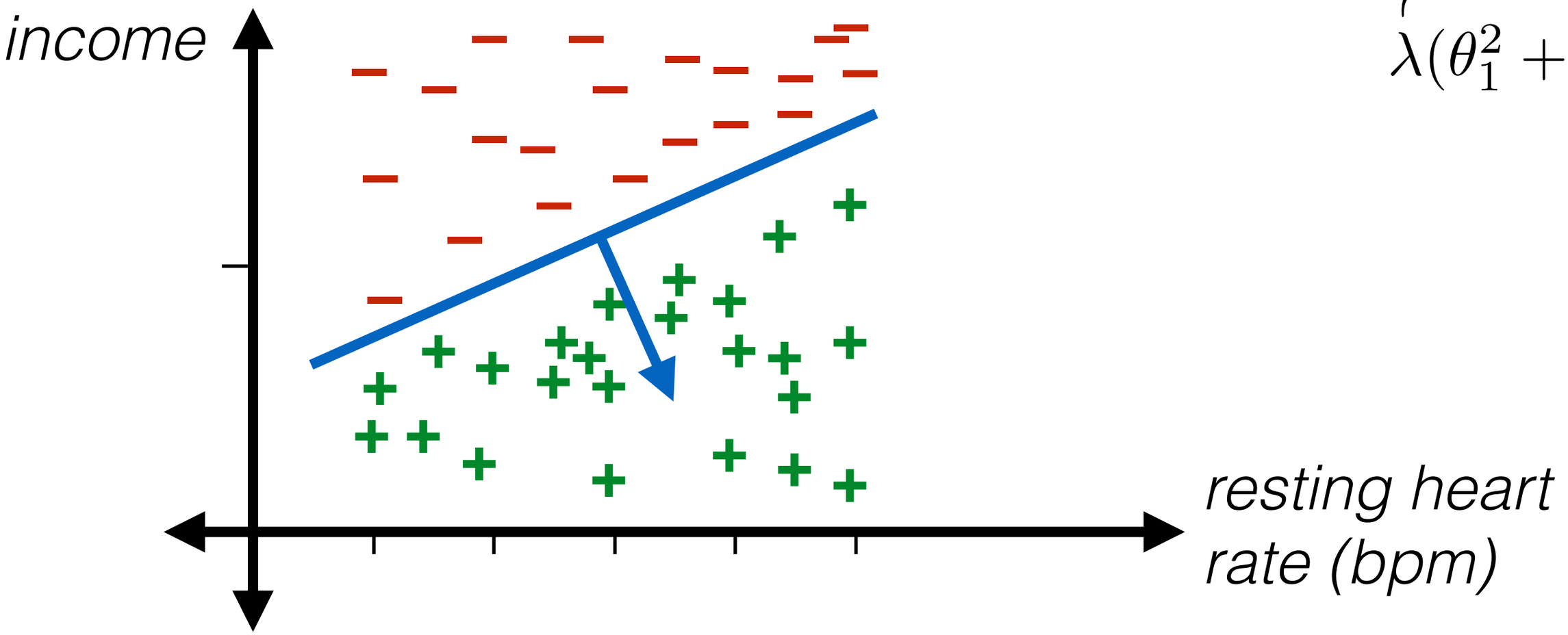
- A closer look at the output of a linear classifier
- Idea: standardize numerical data
  - For *i*th feature and data point *j*: $\phi_i^{(j)} = \frac{x_i^{(j)} - \text{mean}_i}{\text{stddev}_i}$



- Conclusion: it may be easier to visualize and interpret learned parameters if you standardize data

Encode numerical data

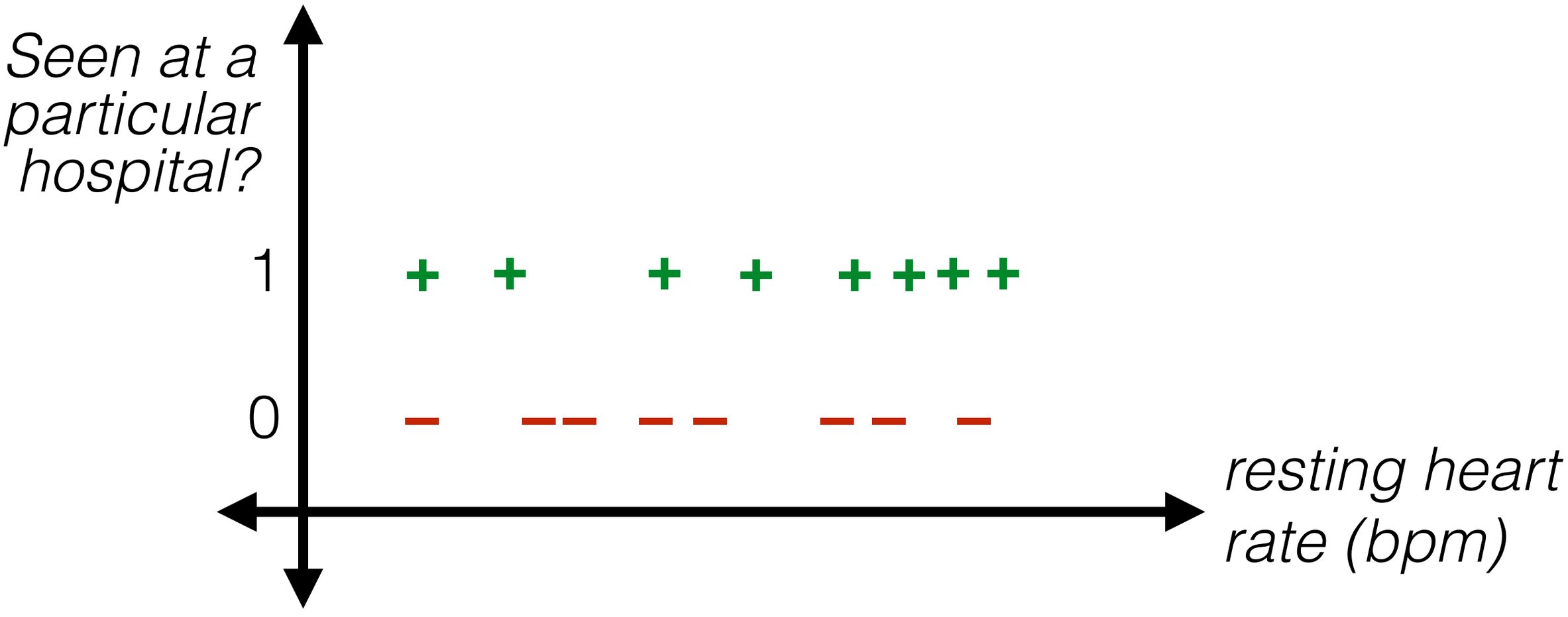
- Standardization can also affect which hypothesis is chosen — e.g. when using a ridge penalty
- Recall: $J_{lr}(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^{n} L_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)}) + \overbrace{\lambda(\theta_1^2 + \theta_2^2)}^{\lambda \|\theta\|^2}$



- If we don't standardize the data, the penalties for different dimensions of θ can be wildly different

More benefits of plotting your data

- And talking to experts



Encode data in usable form

- Identify the features and encode as real numbers
- Standardize numerical features

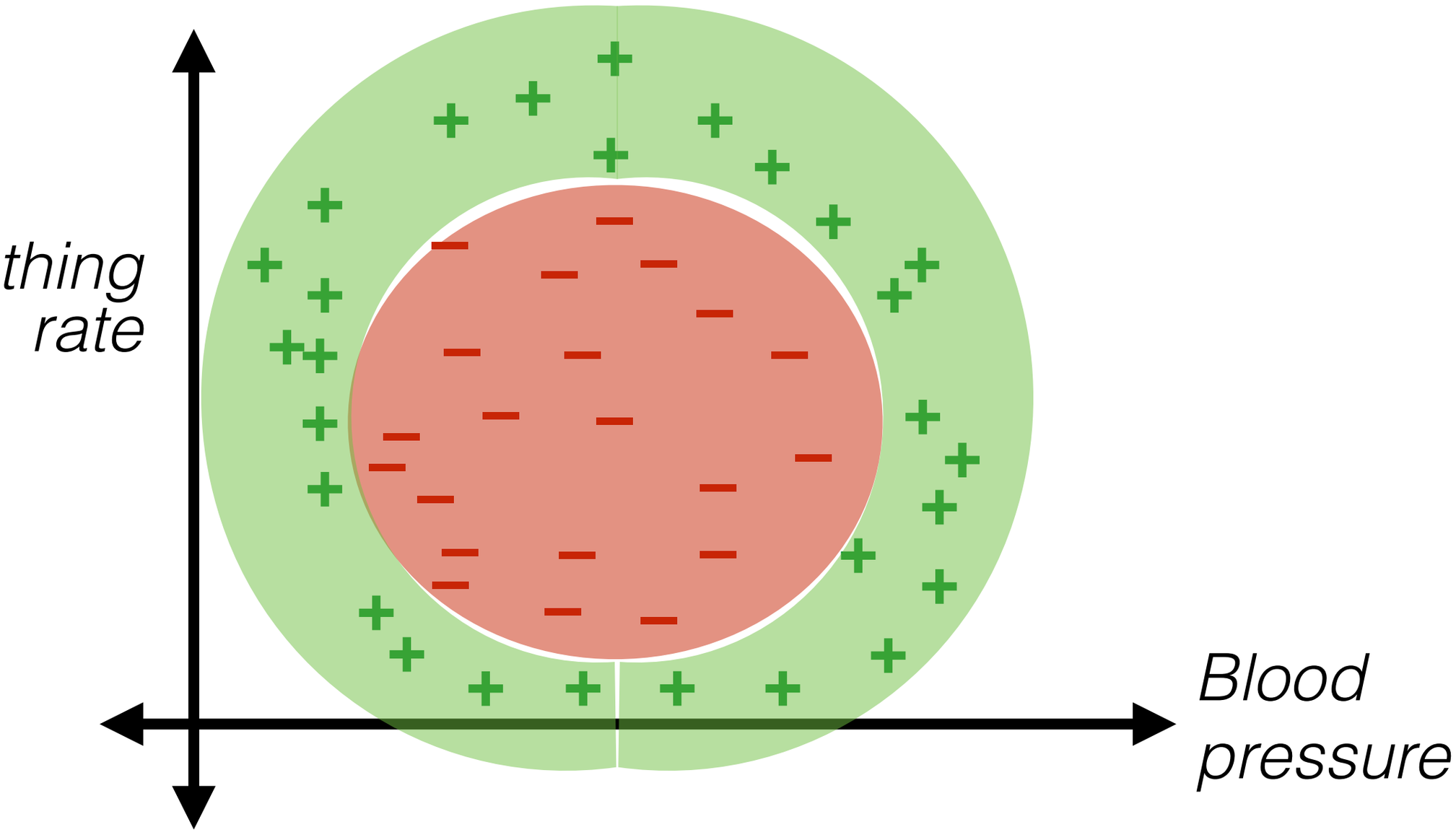
	resting heart rate (bpm)	pain?	j1,j2,j3,j4,j5	m1, m2	decade	family income (USD)
1	55	0	1,0,0,0,0	1,0	4	133000
2	71	0	0,1,0,0,0	1,1	2	34000
3	89	1	1,0,0,0,0	0,1	5	40000
4	67	0	0,0,0,1,0	0,0	5	120000

Encode data in usable form

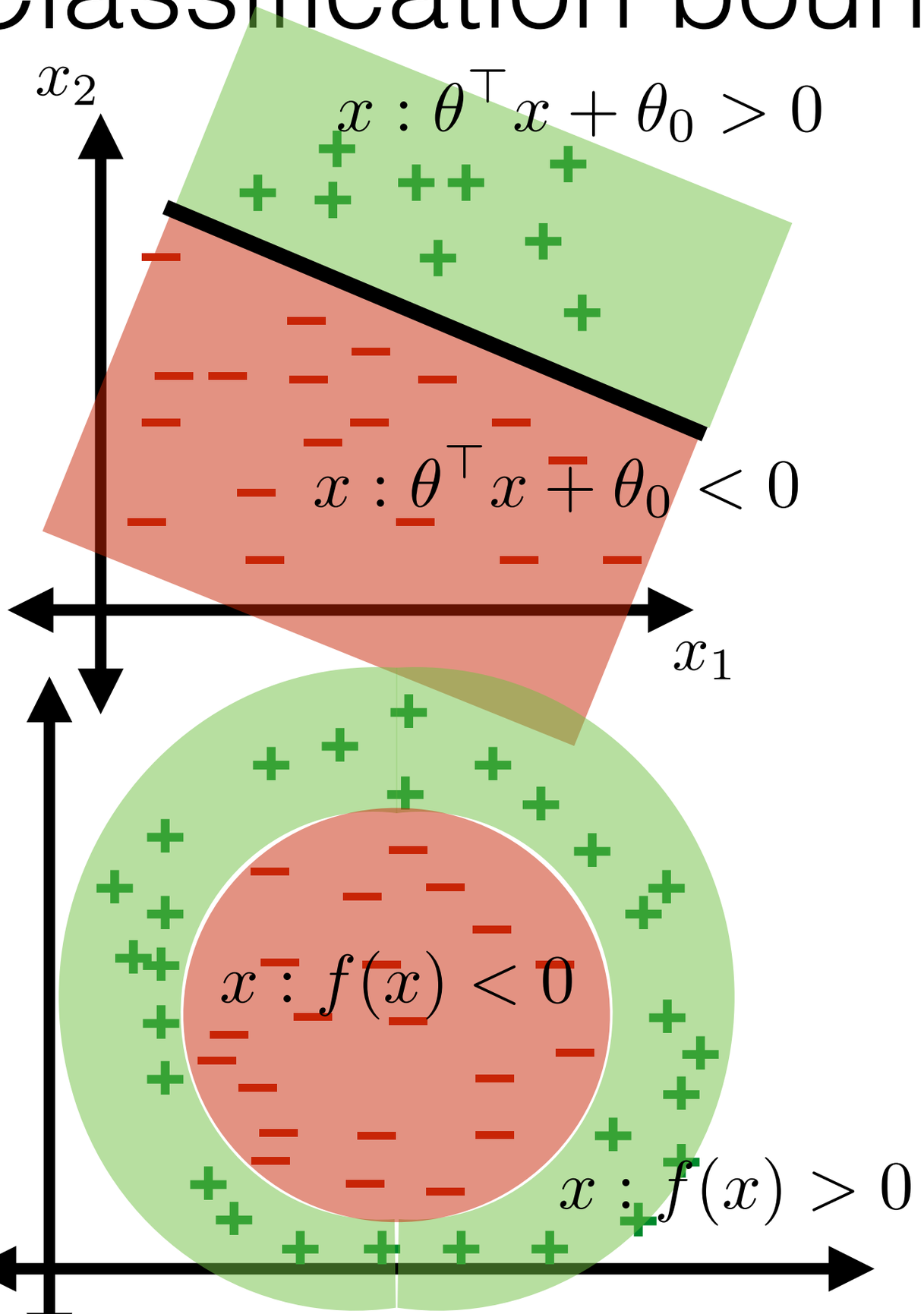
- Identify the features and encode as real numbers
- Standardize numerical features

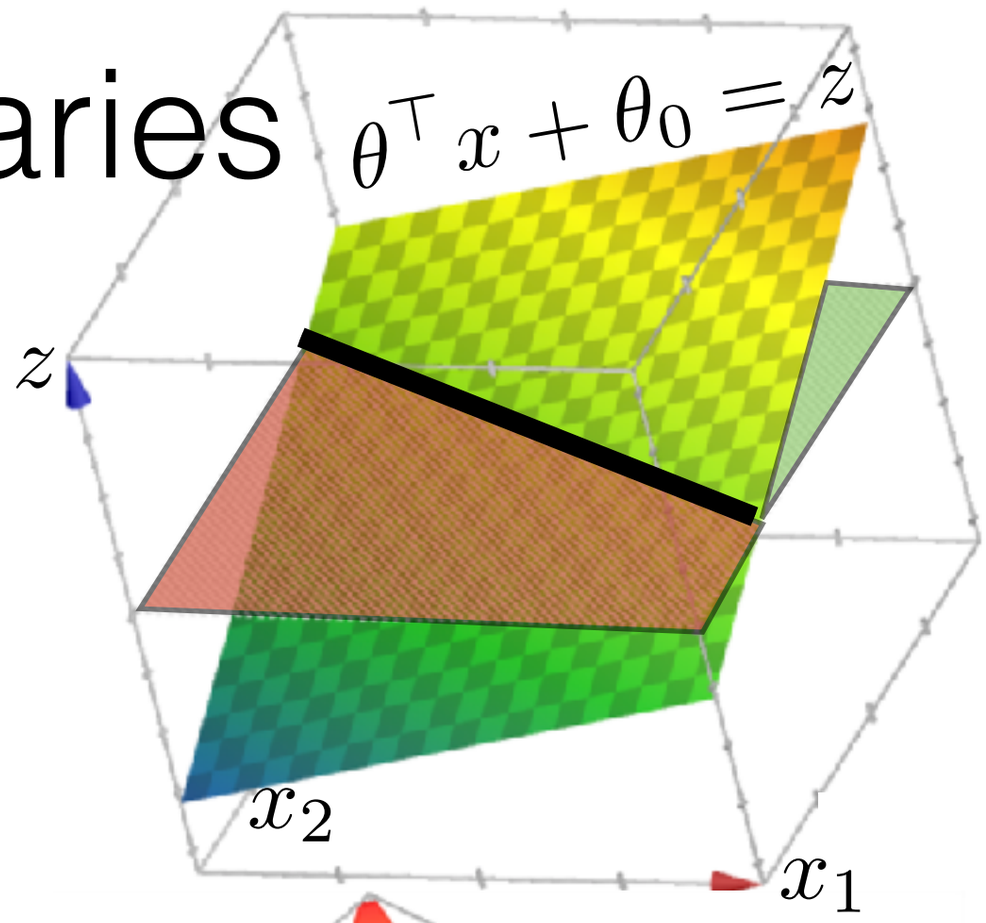
	resting heart rate (bpm)	pain?	j1,j2,j3,j4,j5	m1, m2	decade	family income (USD)
1	-1.5	0	1,0,0,0,0	1,0	1	2.075
2	0.1	0	0,1,0,0,0	1,1	-1	-0.4
3	1.9	1	1,0,0,0,0	0,1	2	-0.25
4	-0.3	0	0,0,0,1,0	0,0	2	1.75

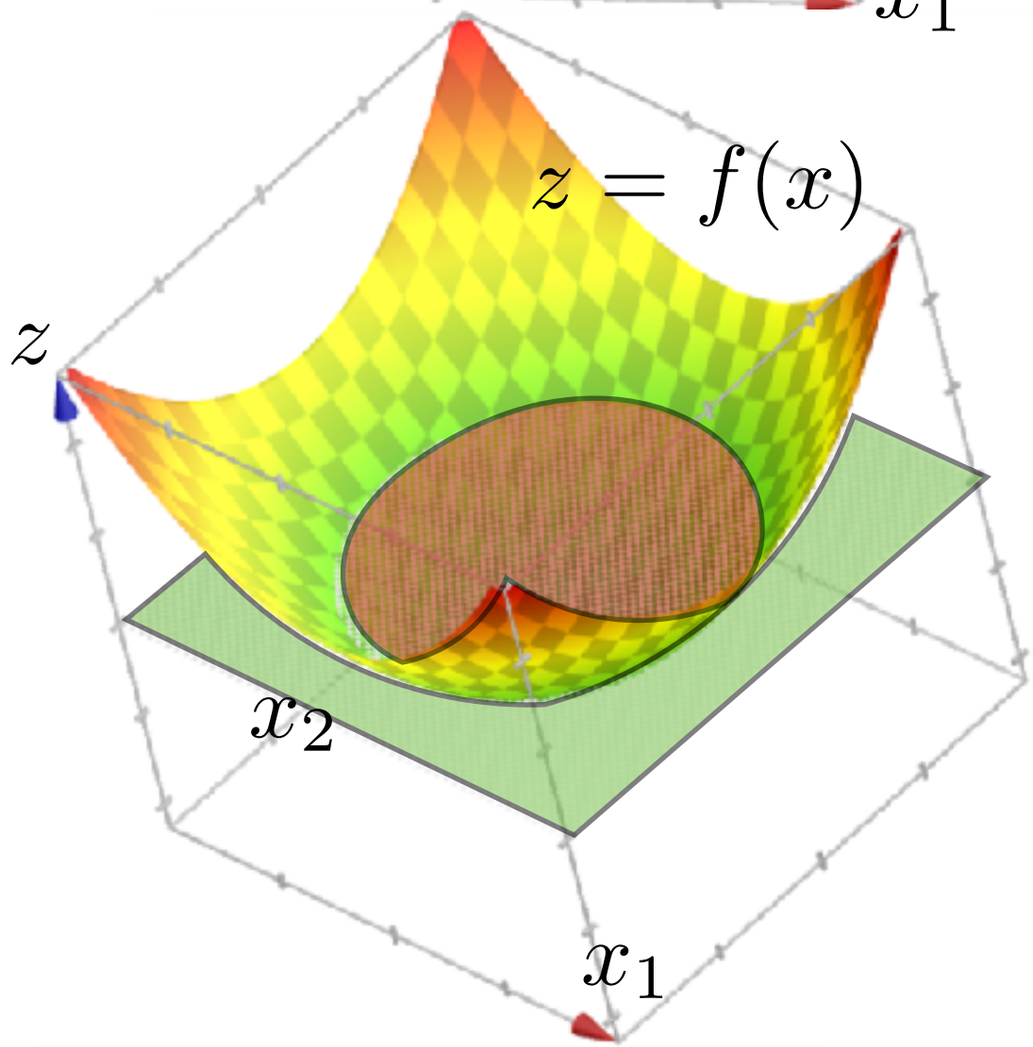
Nonlinear boundaries



Classification boundaries





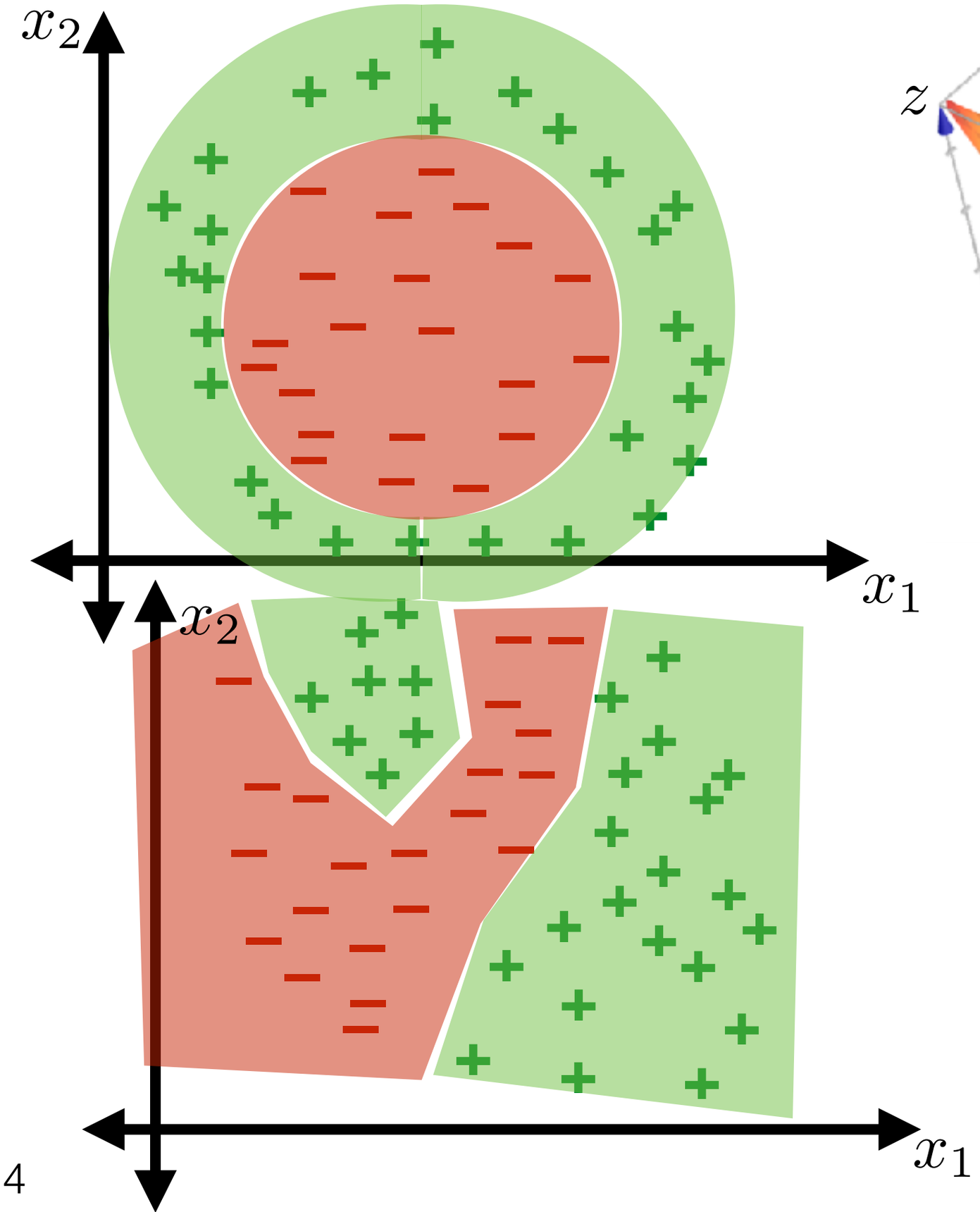


Nonlinear boundaries

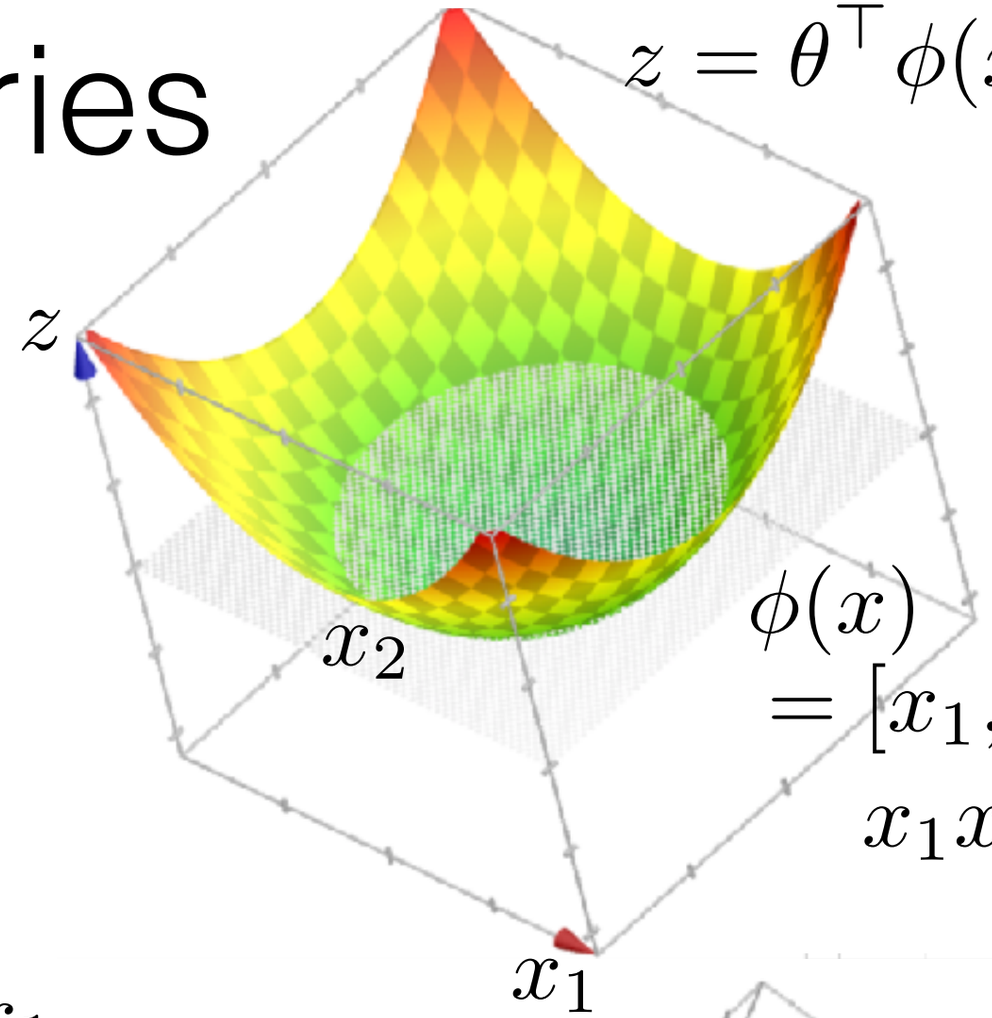
- Idea: can approximate a smooth function with a *k*th order Taylor polynomial (e.g. around 0)

order (*k*)	terms when *d*=1	terms for general d
0	$[1]$	$[1]$
1	$[1, x_1]$	$[1, x_1, \ldots, x_d]$
2	$[1, x_1, x_1^2]$	$[1, x_1, \ldots, x_d,$ $x_1^2, x_1x_2, \ldots, x_{d-1}x_d, x_d^2]$
3	$[1, x_1, x_1^2, x_1^3]$	$[1, x_1, \ldots, x_d,$ $x_1^2, x_1x_2, \ldots, x_{d-1}x_d, x_d^2,$ $x_1^3, x_1^2x_2, x_1x_2x_3, \ldots, x_d^3]$

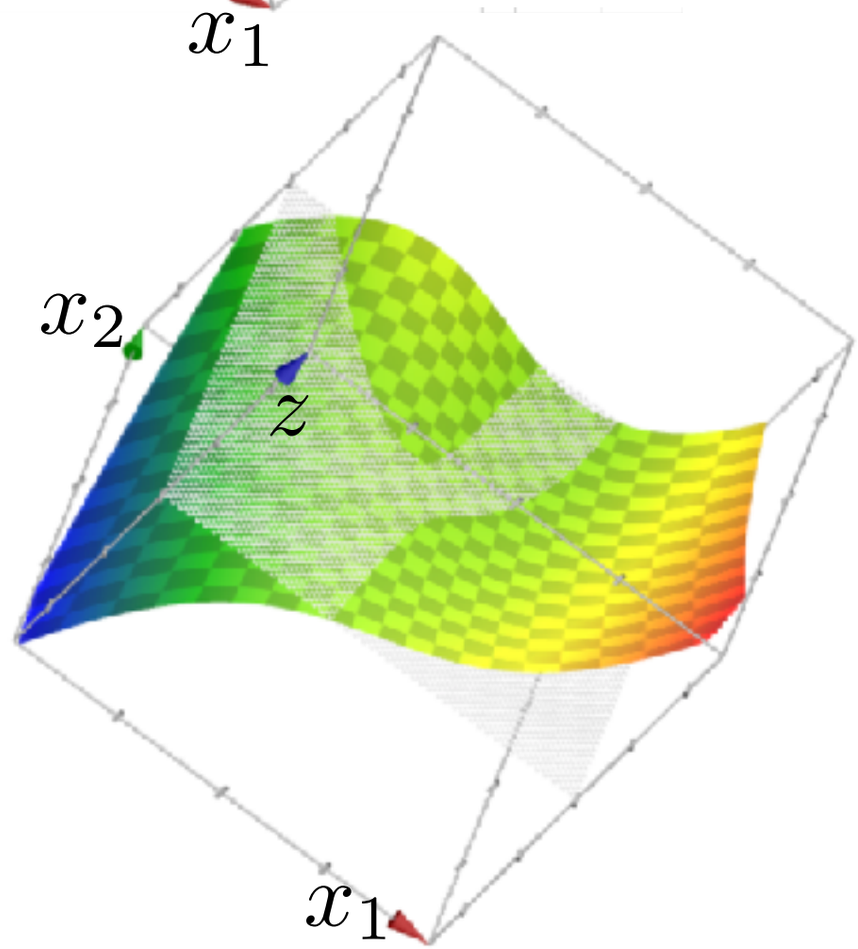
Nonlinear boundaries



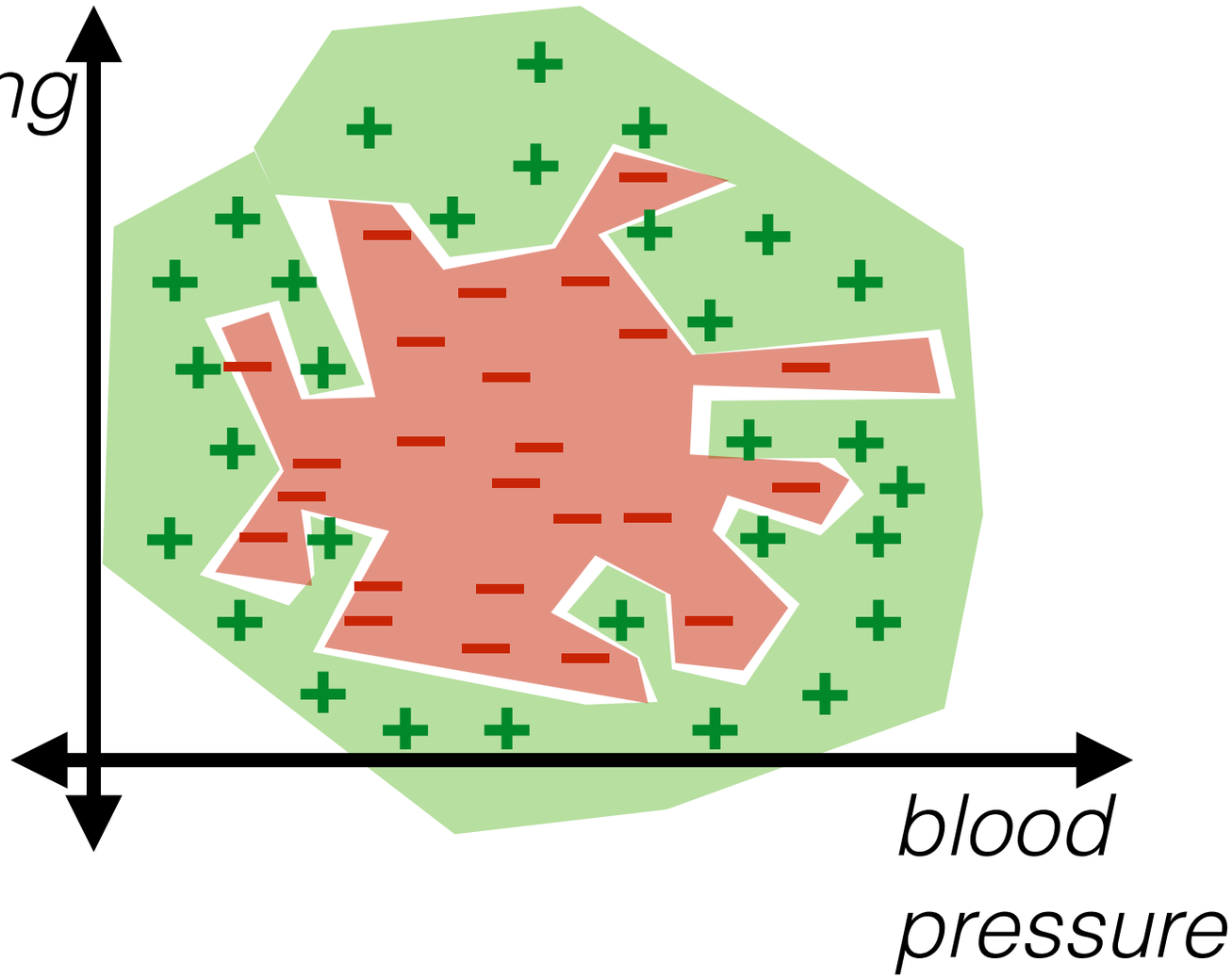
$z = \theta^\top \phi(x) + \theta_0$



$$\phi(x) = [x_1, x_2, x_1^2, x_1 x_2, x_2^2]^\top$$



Nonlinear boundaries



- Training error is 0!
- But seems like our classifier is overfitting

- Benefit of polynomial features: can be super flexible if we use polynomials up to a high degree
- If we use polynomials up to a high degree, we're prone to overfit