Intro to ML concepts

Aarti Singh

Machine Learning 10-315
Sept 2, 2020

Logistical update

- Canvas fixed
  - Zoom links for lecture/recitation and office hours available on Canvas
  - Recording of lectures and recitations available at Zoom tab on Canvas
  - Piazza login directly

- Recitation on Friday Sept 4 – Probability distributions + optimization review and hands-on exercises

- QnA1 to be released TODAY

What is Machine Learning?

Design and Analysis of algorithms that

- improve their performance
- at some task
- with experience

(experience) **(performance on task)**

Tasks, Experience, Performance

Machine Learning Tasks

Broad categories -

- **Supervised learning**

  Classification, Regression

- **Unsupervised learning**

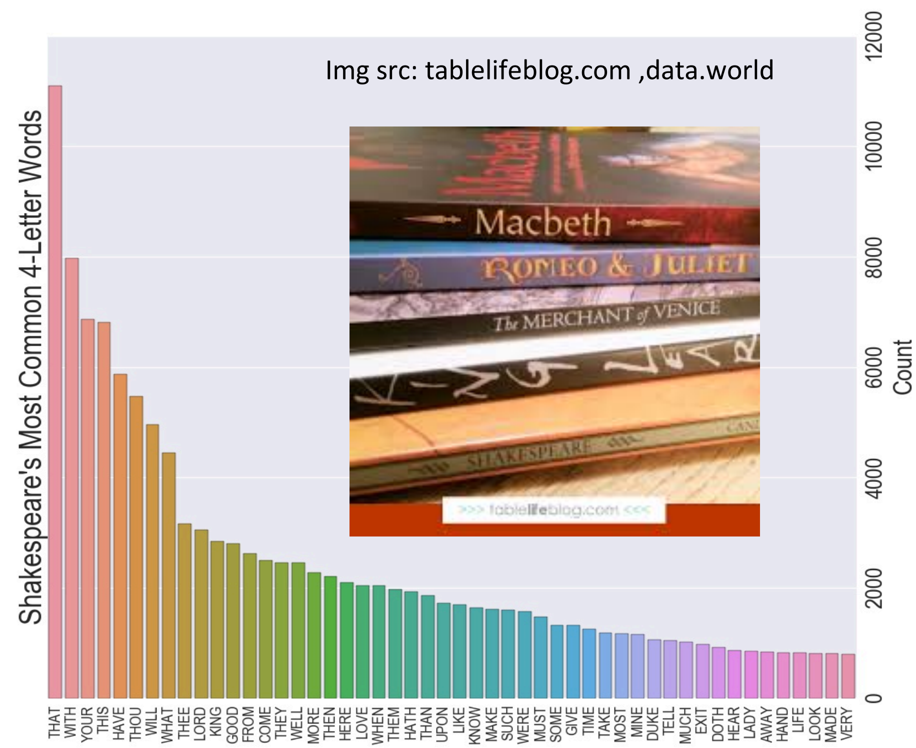
  Density estimation, Clustering, Dimensionality reduction

- Semi-supervised learning
- Active learning
- Reinforcement learning
- Many more …

Unsupervised Learning

Learning a Distribution

Bias of a coin



Distribution of words in text

- What other distribution would be interesting to learn?

Unsupervised Learning

Clustering - Group similar things e.g. images [Goldberger et al.]

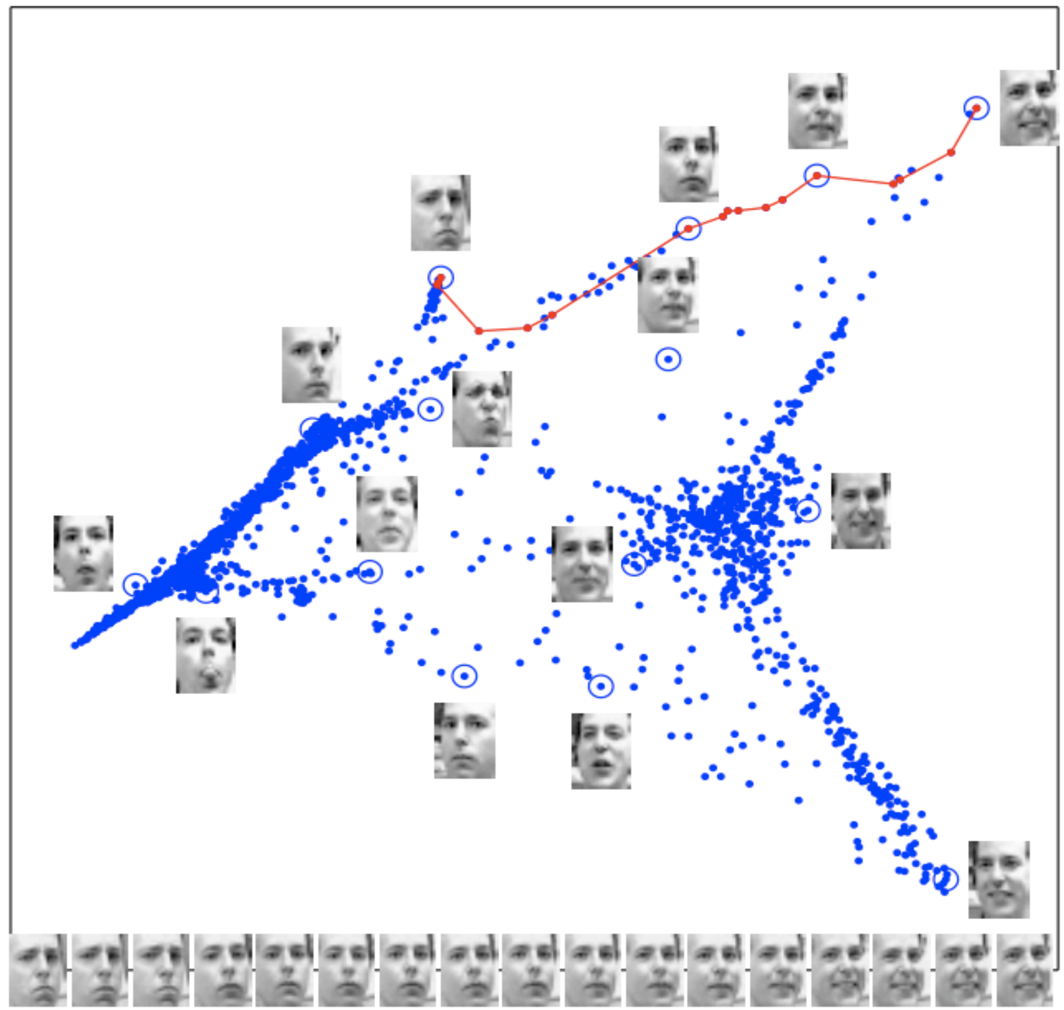
Unsupervised Learning

Dimensionality Reduction/Embedding

[Saul & Roweis '03]

Images have thousands or millions of pixels.

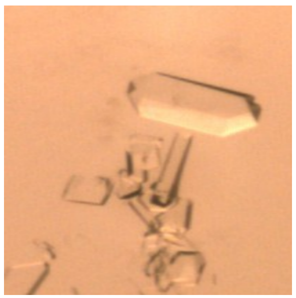
Can we give each image a small set of coordinates, such that similar images are near each other?



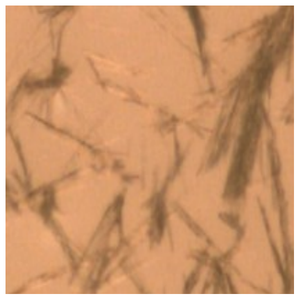
Tasks, **Experience**, Performance

Experience = Training Data

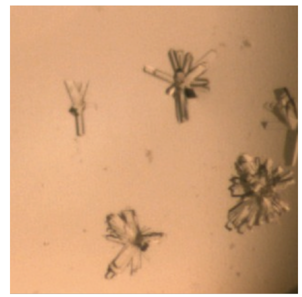
Task: Learning stage of protein crystallization



Crystal

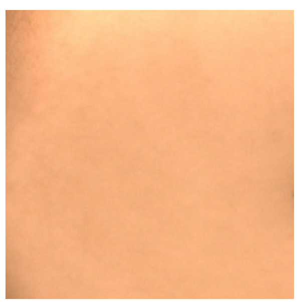


Needle

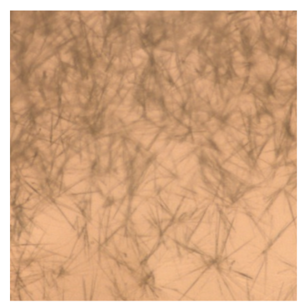


Tree

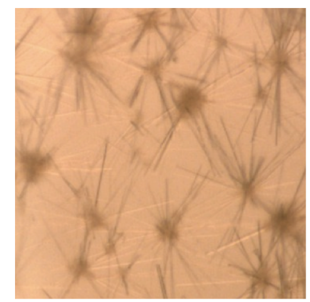
Tree



Empty



Needle



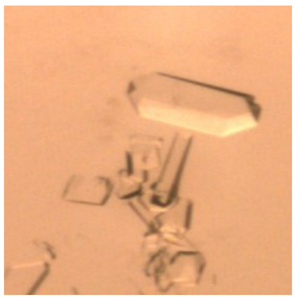
?

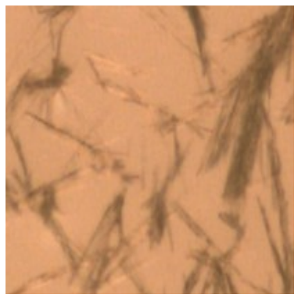
Experience

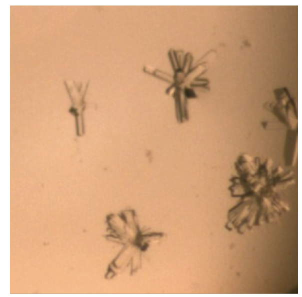
Performance

Training Data vs. Test Data

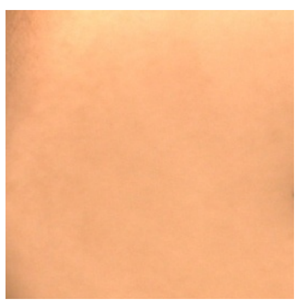
Task: Learning stage of protein crystallization

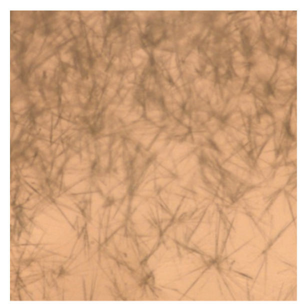

Crystal


Needle

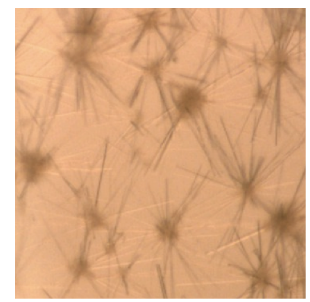

Tree


Tree


Empty

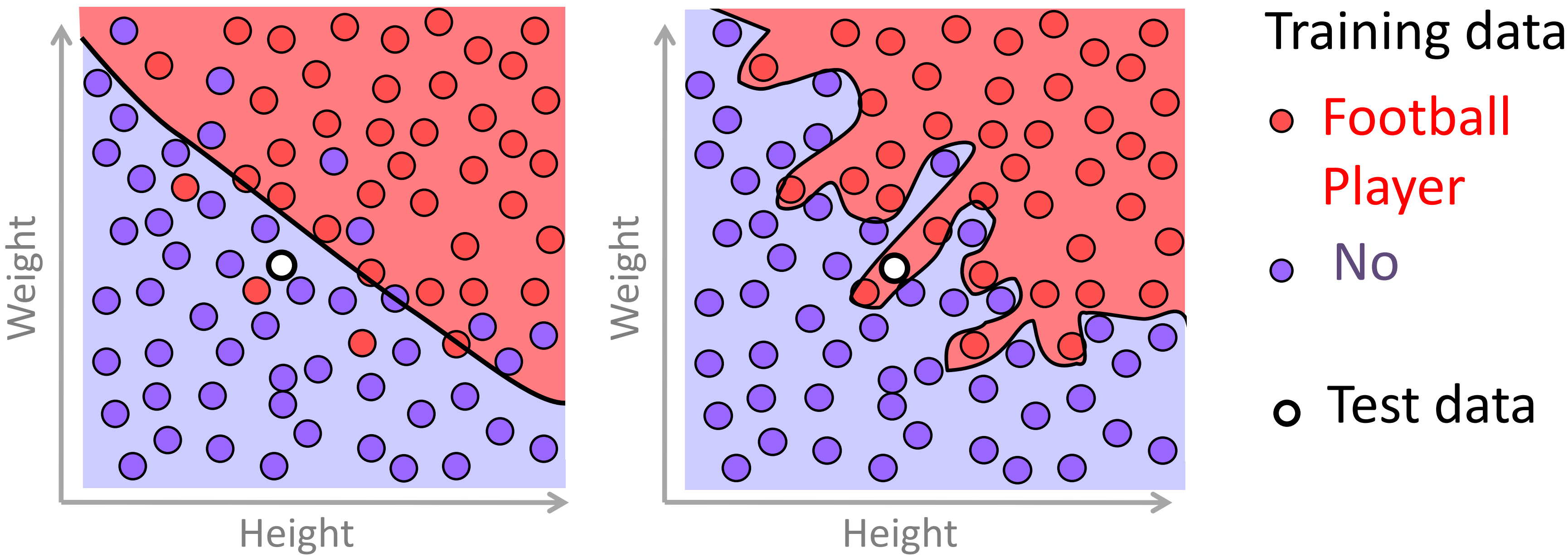

Needle

Experience


?

Performance

Training Data vs. Test Data



- A good machine learning algorithm
  - **Generalizes** aka performs well on test data
  - ~~Does not **overfit** training data~~

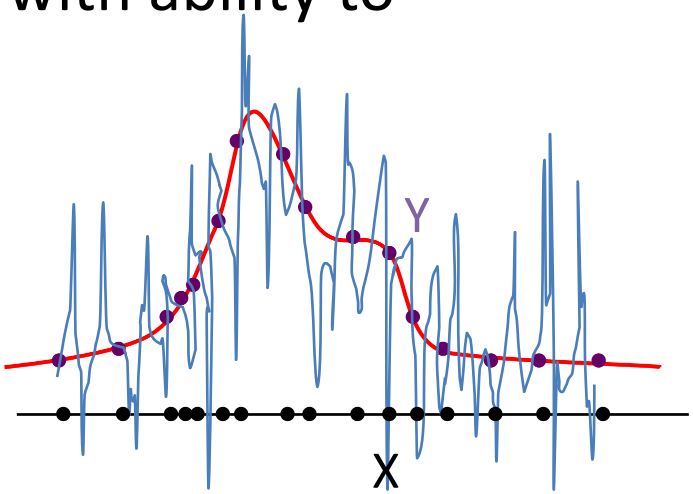
Memorizing vs. Learning

- Is it okay to **overfit** training data?
- Is it okay to **memorize** training data?

Sometimes yes (e.g. if labels are noiseless)

BUT needs to be accompanied with ability to generalize

➢ Which fit is better (Red/Blue)?



- What is learning really?

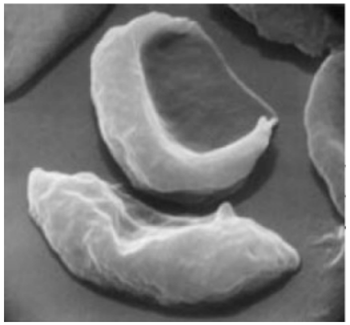
Can algorithm **generalize** aka perform well on test data

Tasks, Experience, Performance

Performance Measure

Performance:

$\text{loss}(Y, f(X))$ - Measure of closeness between label *Y* and prediction *f*(*X*) for test data X

X	Diagnosis, *Y*	*f*(*X*)	$\text{loss}(Y, f(X))$
	"Anemic cell"	"Anemic cell"	0
		"Healthy cell"	1

$$\text{loss}(Y, f(X)) = 1_{\{f(X) \neq Y\}}$$ **0/1 loss**

Performance Measure

Performance:

$\text{loss}(Y, f(X))$ - Measure of closeness between label *Y* and prediction *f*(*X*) for test data X

X	Share price, *Y*	*f*(*X*)	$\text{loss}(Y, f(X))$
Past performance, trade volume etc. as of Sept 8, 2010	"\$24.50"	"\$24.50"	0
		"\$26.00"	1?
		"\$26.10"	2?

$$\boxed{\text{loss}(Y, f(X)) = (f(X) - Y)^2 \quad \textbf{squared loss}}$$

Performance Measure

For test data X, measure of closeness between label *Y* and prediction *f*(*X*)

Binary Classification $\text{loss}(Y, f(X)) = 1_{\{f(X) \neq Y\}}$ **0/1 loss**

Regression $\text{loss}(Y, f(X)) = (f(X) - Y)^2$ **squared loss**

Lets think of unsupervised tasks next.

Performance Measure

For test data X, measure how good is the learnt distribution, clustering or embedding $f(X)$

Learning a distribution

Clustering

Groups 1-10: Jamboard_1_10
Groups 11-20: Jamboard_11_20

Dimensionality reduction

- What performance measure would you use for each task?

Performance Measure

For test data X, measure how good is the learnt distribution, clustering or embedding $f(X)$

Learning a distribution

"Likelihood"

- What performance measure would you use for each task?

Performance Measure

For test data X, measure how good is the learnt distribution, clustering or embedding $f(X)$

Clustering

- What performance measure would you use for each task?

Performance Measure

For test data X, measure how good is the learnt distribution, clustering or embedding $f(X)$

Dimensionality reduction

- What performance measure would you use for each task?

Glossary of Machine Learning

- Task
- Supervised learning
  - Classification
  - Regression
- Unsupervised learning
  - Learning distribution
  - Clustering
  - Dimensionality reduction/Embedding
- Input, X
- Label, Y
- Prediction, f(X)
- Experience = Training data
- Test data
- Overfitting
- Generalization
- Performance
- Likelihood
- Loss – 0/1, squared, negative log likelihood

Why is ML not …

- Interpolation?
  - Noise, stochasticity, transfer across domains, …
- Statistics?
  - care about computationally efficiency (feasible, at least polynomial time in input size but typically much faster)
- Optimization?
  - Don’t know true objective function, only stochastic version computed using data samples
- Data mining?
  - Generalization on new unseen data
- Your question?

ML common sense

- Training vs Testing accuracy
- Baselines
- Mean vs Best accuracy
- Standard deviation
- Underlying goal/purpose

ML common sense

- Training vs Testing accuracy
- Baselines
- Mean vs Best accuracy
- Standard deviation
- Underlying goal/purpose

Critical to report testing and NOT training accuracy

Regression example: Blood samples were collected for 100 subjects who were administered a covid-19 vaccine.

An ML algorithm was trained to predict the number of antibodies in the blood of these 100 subjects given their profiles.

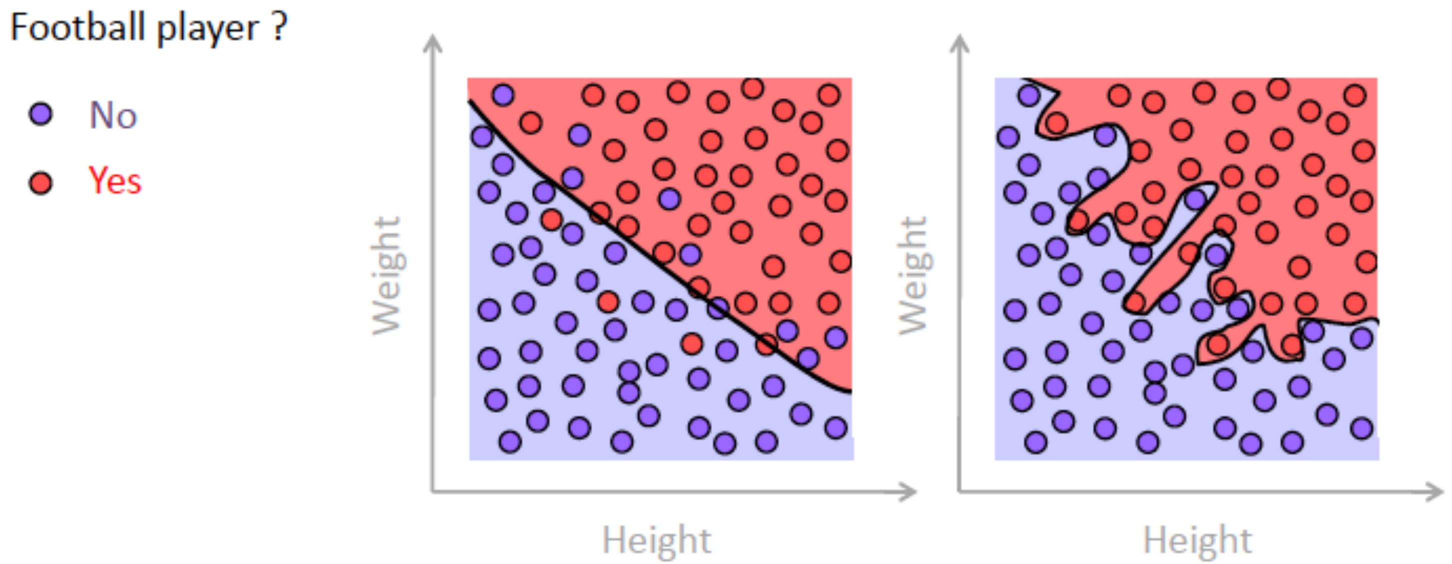
The normalized mean square error of the trained model was 0.001 for predicting the antibodies in these 100 subjects.

- Is this a good model?

10 more subjects were then recruited and the normalized mean square error of the model's predictions of antibodies for these 10 subjects was 0.35.

Critical to report testing and NOT training accuracy

Classification example:



Regression example: Training error 0.3 in predicting activity at one brain region using activity in other brain regions Test error 0.9

Model fit example: Training likelihood 0.99, Testing likelihood 0.3

ML common sense

- Training vs Testing accuracy
- Baselines
- Mean vs Best accuracy
- Standard deviation
- Underlying goal/purpose

Baselines are extremely important: biased classes

Accuracy of classifier

- Are these good classifiers?

	Test accuracy
• Classifier 1	92%
• Classifier 2	87%

Test dataset had 9300 normal patients and 700 patients with cancer

Baselines are extremely important: multiple classes

Accuracy of classifier

➢ Are these good classifiers?

	Test accuracy
• Classifier 1	52%
• Classifier 2	44%

Test dataset 10000 images: 2 classes, 5000 images each

Test dataset 10000 images: 10 classes, 1000 images each

Baselines are extremely important: regression

Accuracy of regressor

➢ Are these good predictors?

	Test Mean Squared Error
• Regressor 1	25
• Regressor 2	100

Standard deviation of test data ~7

MSE vs $R^2 := 1 - \text{MSE}/\text{Variance}$

(Fraction of variance explained by predictor)

ML common sense

- Training vs Testing accuracy
- Baselines
- ➢ Mean vs Best accuracy
- ➢ Standard deviation (Std)
- Underlying goal/purpose

Best run test accuracy doesn't make a classifier better

Accuracy of classifier

	Mean	Best run
• Classifier 1	92%	97%
• Classifier 2	87%	100%

High mean test accuracy doesn't make a classifier better

Accuracy of classifier

	Mean
• Classifier 1	92%
• Classifier 2	87%

High mean test accuracy doesn't make a classifier better

Accuracy of classifier

	Mean	Std
• Classifier 1	92%	15%
• Classifier 2	87%	5%

High mean test accuracy doesn't make a classifier better

Accuracy of classifier

	Mean	Std	Range
• Classifier 1	92%	15%	77-100
• Classifier 2	87%	5%	82-92

ML common sense

- Training vs Testing accuracy
- Baselines
- Mean vs Best accuracy
- Standard deviation

➢ Underlying goal/purpose

Purpose often dictates validity of classifier

Accuracy of classifier

	Mean	Std	Range
• Classifier 1	92%	15%	77-100
• Classifier 2	87%	5%	82-92

- ➢ Which classifier would you choose when recommending movies?
- ➢ Which classifier would you choose when diagnosing serious illness?

Purpose often dictates validity of regressor

Accuracy of regressor

➢ Are these good predictors?

	MSE
• Regressor 1	25
• Regressor 2	0.0001

Purpose often dictates validity of regressor

Accuracy of regressor

- Are these good predictors?

	MSE	Task
• Regressor 1	25	Predict age of a person
• Regressor 2	0.0001	Predict proportion of lead in water

MS(quared)E vs. MA(bsolute)E

Units important

End of Lecture