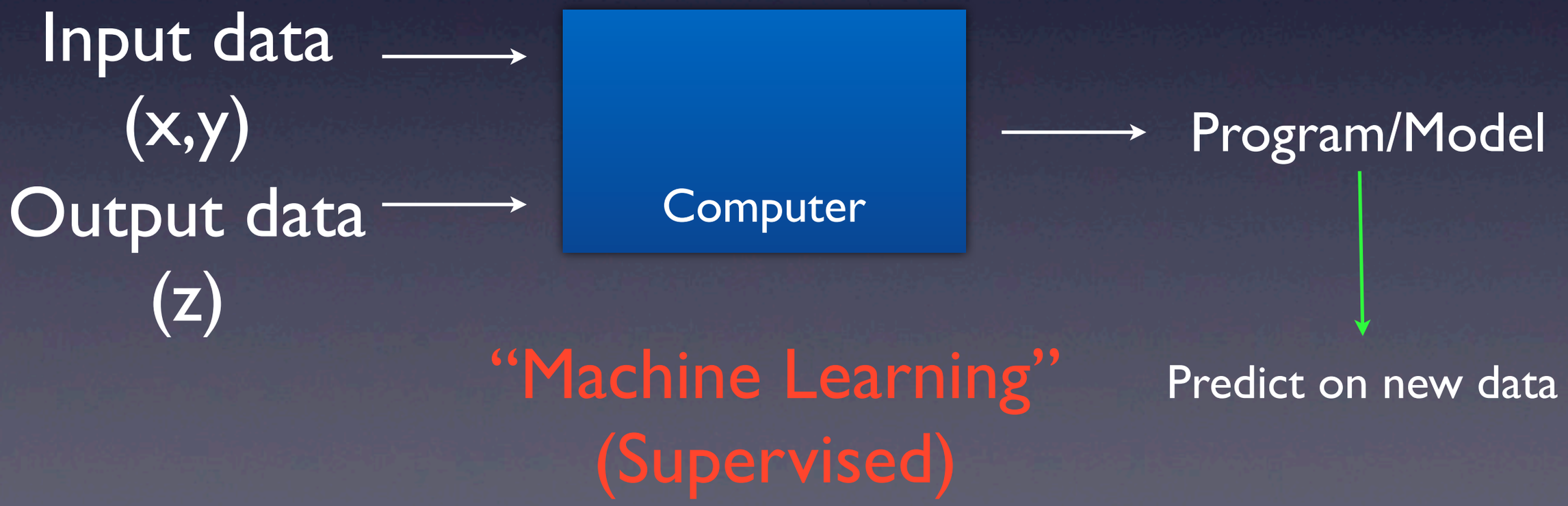
Machine Learning Introduction

Kristen Menou

What is ML?

"Standard Programming"



"Machine Learning"
(Supervised)

Definitions/Conventions

- Unsupervised learning: categorize/characterize/find trends in arbitrary dataset (e.g.: clustering in x,y plane, in the absence of z)
- Supervised learning: predict after learning from a set of example datapoints: z=f(x,y)
- Regression: supervised learning with continuous, ordered response (e.g. 0-1000)
- Classification: supervised learning with categorical response (e.g. yes/no, black/white)

A variety of algorithms

Supervised Regression

- Simple and multiple linear regression
- Decision tree, random forest
- Artificial Neural networks
- Nearest neighbor methods (e.g., k-NN or k-Nearest Neighbors)
- ...

Supervised Two-class & Multi-class Classification

- Logistic regression and multinomial regression
- Artificial Neural networks
- Decision tree, random forest
- SVM (support vector machine)
- Bayesian classifiers (e.g., Naive Bayes)
- Nearest neighbor methods (e.g., k-NN or k-Nearest Neighbors)
- ....

Unsupervised

- K-means clustering
- PCA (principal component analysis)
- ...

MACHINE INTELLIGENCE LANDSCAPE 2016

Predictive Analytics

AGENTS

PROFESSIONAL
Howdy! · x.ai · KASISTO · DigitalGenius · OVERLAP.CC · meekan · clara · PRIMER

PERSONAL
facebook M · XIAOICE · large · fuse|machines · assistant.ai · nestor · @wesome · Magic

OS INTERFACES
Siri · Cortana · VIV · maluuba · api.ai · CogNEA · Google Now

AUTONOMOUS SYSTEMS

AIR
3DR · dji · PROJECT LOON (Google) · VERTICAL · DroneDeploy · AIRDOG · SKYCATCH · SKYDIO · Airware · LILY

GROUND
Google · UBER · TESLA · CRUISE · MOBILEYE · COMMA · AdasWorks

SEA
LIQUID ROBOTICS · bluefin data · OPENROV · BluHaptics

INDUSTRIAL
KIVA Systems · fetch · HARVEST AUTOMATION · CLEARPATH · AVIDBOTS · ENERGID · rethink robotics · GREYORANGE · OSARO

ENTERPRISE

SECURITY / FRAUD
Sentinel · graphistry · BITSIGHT · feedzai · AREA 1 · drawbridge · sift science · CYLANCE · Brighterion

HR / RECRUITING
textio · hiQ · gild · SpringRole · entelo · unitive · GIGSTER

SALES
6sense · infer · people pattern · Preact · Prism · AVISO · Vidora · sentient · salespredict · Gainsight

MARKETING
LiftIgniter · RADIUS · brightfunnel · retention SCIENCE · AIRPR

CUSTOMER SUPPORT
CLARABRIDGE · QUANTI · loop Labs · ACTION · DigitalGenius · FRAM

INTERNAL INTEL
Alation · ADATAO · Palantir · sapho · SKIPFLAG · lucid · Rainbird · Digital Reasoning · Narrative Science

MARKET INTEL
Quid · mattermark · DataFox · bottlenose · PREMISE · enigma · CB INSIGHTS

PLATFORMS

RESEARCH / AGI
vicarious · Cycorp · Google DeepMind · Numenta · nnaisense · CURIOUS · SCALED INFERENCE · GEOMETRIC INTELLIGENCE

FULL STACK
context relevant · CognitiveScale · NVIDIA · TERADEEP · QUALCOMM · nervana SYSTEMS

MACHINE LEARNING
rapidminer · bigml · AYASDI · cortical.io · amazon web services · Azure Machine Learning · narrative logics · blue yonder · SKYTREE

INDUSTRIAL IOT
ThingWorx · UPTAKE · IMUBIT · Preferred Networks · Alluvium · xively · PLANET OS

AUDIO
Grid · nexidia · VOCO · NUANCE · Expect · popUP archive

VISION
ORBITAL INSIGHT · Descartes Labs · DEXTRO · cortica · clarifai · PLANET LABS · MetaMind

DATA ENRICHMENT
diffbot · Paxata · TRIFACTA · iDIBON · CrowdFlower · WorkFusion

INDUSTRIES

ADTECH
ADTHEORENT · dstillery · BEYONDVERBAL · METAMARKETS · TAPAD · rocketfuel · affectiva

AGRICULTURE
BLUE RIVER · tule · TerrAvion · mavrx · THE CLIMATE CORPORATION · CERES TECHNOLOGIES · HONEYCOMB

FOR GOOD
Conservation Metrics · DataKind · DATA POP · thorn · BAYES IMPACT

RETAIL FINANCE
inVenture · affirm · earnest · MIRADOR · Lendo · zesty.finance · LendUp

LEGAL
Everlaw · RAVEL · LEGAL ROBOT · seal · BEAGLE · ROSS · Lex Machina

MATERIALS & MFG
zymergen · AUGMATE · GINKGO BIOWORKS · ETRINE TECHNOLOGIES · SIGHT MACHINE · CALCULARIO · Eigen Innovations

HEALTHCARE
deep genomics · 3SCAN · enlitic · Calico · Atomwise · Recombine · color · METABIOTA · GRAND ROUNDS · Google Life Sciences · IBM Watson Health

INDUSTRIES (CONT'D)

EDUCATION
KNEWTON · coursera · turnitin · gradescope · UDACITY · KHANACADEMY

TRANSPORT & LOGISTICS
NAUTO · taleris · PRETECKT · clearmetal

INVESTMENT FINANCE
Bloomberg · Quantopian · Dataminr · KENSHO · iSENTIUM · NEURENSIC · alphasense

TECH USER TOOLS

DATA SCIENCE
DOMINO · kaggle · Sentenai · sense · yseop · Outlier · ŷhat · DataRobot

MACHINE LEARNING
Cortana Analytics · AlchemyAPI · glowfi.sh · IBM Watson Platform · Anodot · MonkeyLearn · (h [s]) HyperScience · fuzzy.io · SIGOPT · 0xdata H2O · SPARKBEYOND · indico

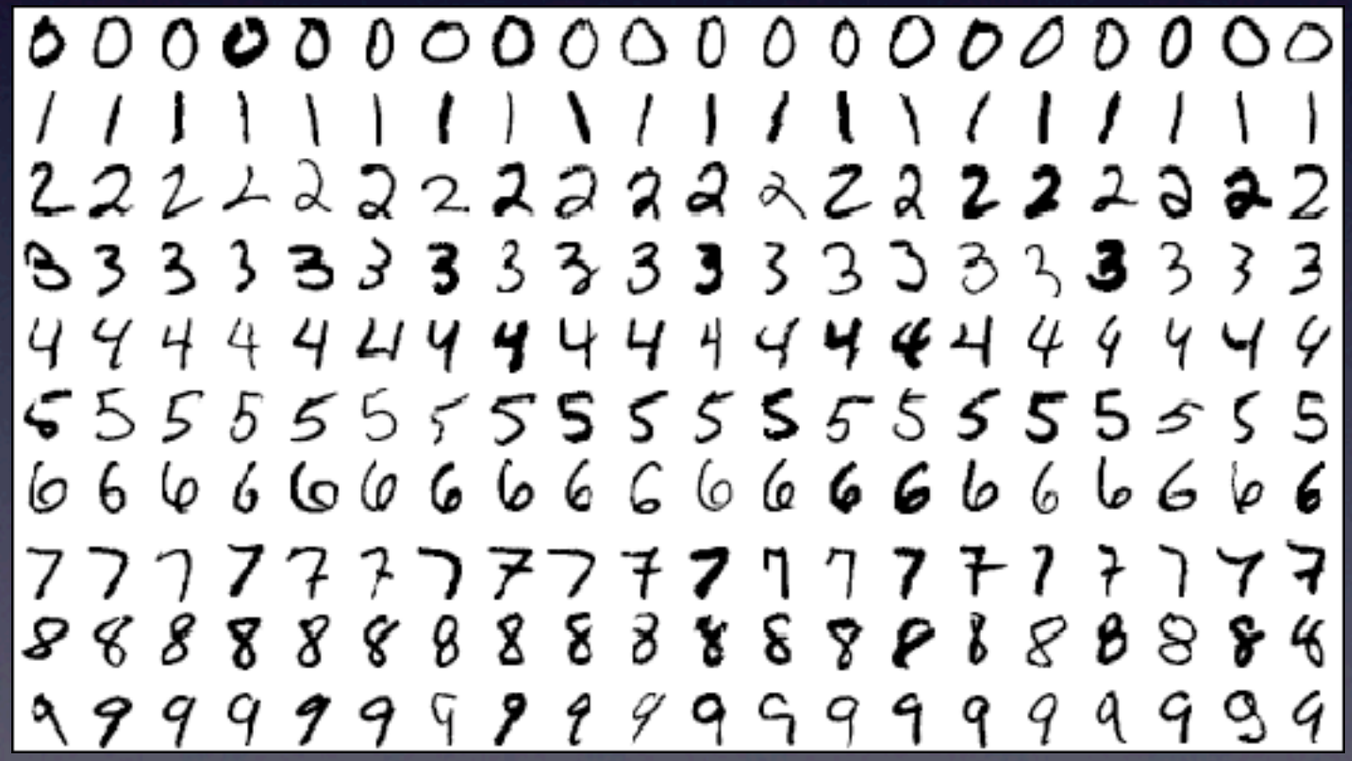
OPEN SOURCE
SKYMIND · TensorFlow · DL4J · Caffe · theano · Microsoft DMTK · Spark MLlib · spaCy · SciKit · CGT

SHIVONZILIS.COM/MACHINEINTELLIGENCE

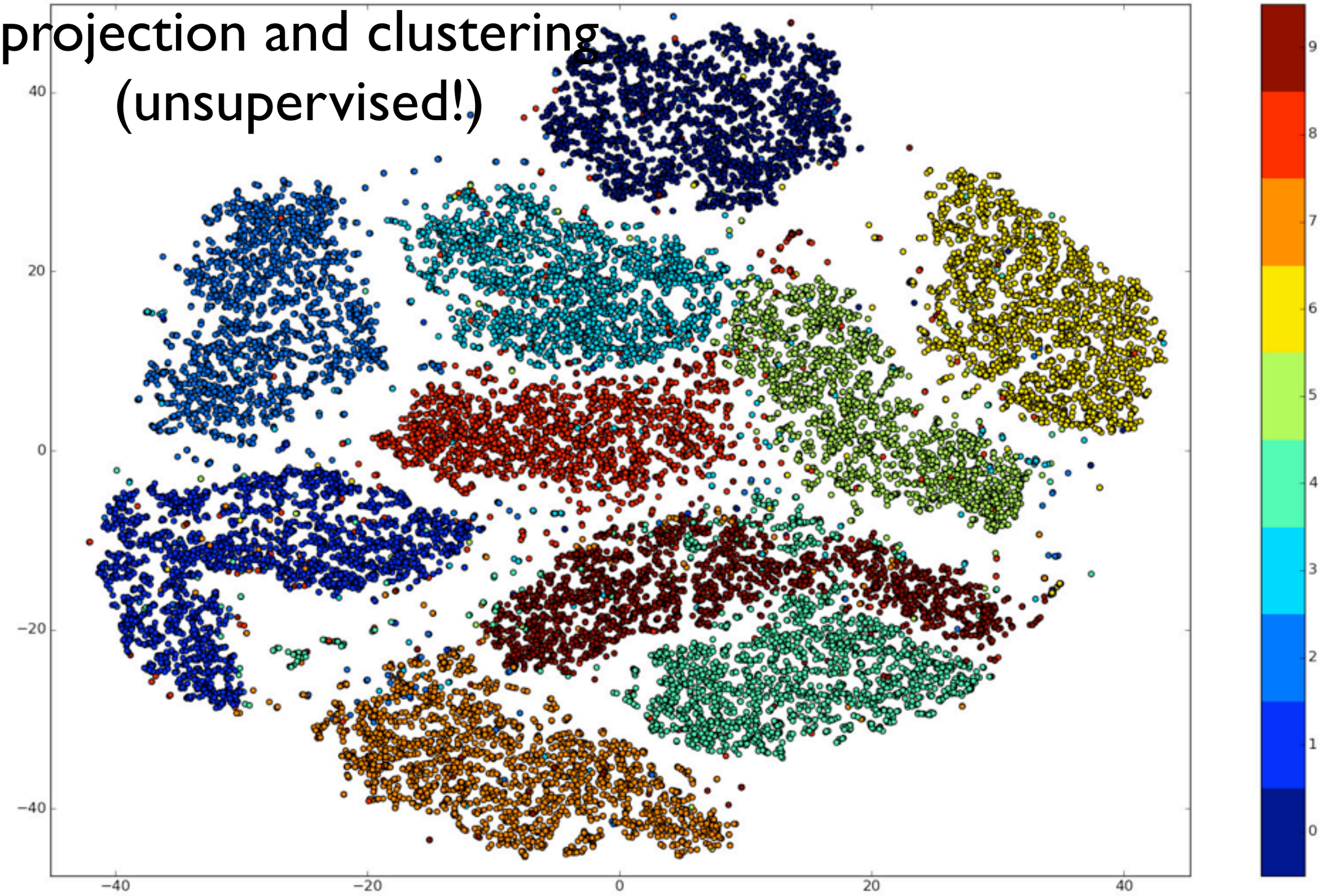
Bloomberg BETA

State-of-the-Art

MNIST data: hand-written digits
(10-class classification)



t-SNE algorithm: 2D projection and clustering (unsupervised!)



ML in practice

- It is unclear what constitutes the best ML solution on a given problem
- Data science competitions provide useful comparisons (+ near-optimal solutions)
- Competition & collaboration both help

Kaggle: ML training camp

- Dataset & ML Competition Host
- Competitions: company provides data, crowd builds ML models, predict on "unseen/unlabeled" test data, best predictive model wins $$$
- Company gets crowd-sourced, near-optimal ML solution for their specific data science problem
- ~500,000 kagglers, typically ~500-3000 participate in a given competition
- Key issues: feature engineering, overfitting, ML algorithms (choice + optimization)

Active Competitions

All Competitions

Active Competitions

Second Annual Data Science Bowl
Transforming How We Diagnose Heart Disease

7.0 days
755 teams
$200,000

Santander Customer Satisfaction
Which customers are happy customers?

56 days
674 teams
241 scripts
$60,000

Home Depot Product Search Relevance
Predict the relevance of search results on homedepot.com

49 days
1380 teams
1009 scripts
$40,000

BNP Paribas Cardif Claims Management
Can you accelerate BNP Paribas Cardif's claims management process?

42 days
1833 teams
697 scripts
$30,000

March Machine Learning Mania 2016
Predict the 2016 NCAA Basketball Tournament

5.0 days
528 teams
332 scripts
$25,000

Yelp Restaurant Photo Classification
Data Mining Engineer at **Yelp**
San Francisco, CA

36 days
171 teams
73 scripts
Jobs

1539 scripts

$60,000 • 674 teams

Santander Customer Satisfaction

Merger and 1st Submission Deadline

Wed 2 Mar 2016 — Mon 2 May 2016 (56 days to go)

Dashboard

Home
- Data
- Make a submission

Information
- Description
- Evaluation
- Rules
- Prizes
- Timeline

Forum

Scripts
- New Script
- New Notebook

Leaderboard

My Team

My Submissions

Public Leaderboard

1. BreakfastPirate
2. NxGTR
3. anokas
4. DS.RESEARCH

Competition Details » Get the Data » Make a submission

Which customers are happy customers?

From frontline support teams to C-suites, customer satisfaction is a key measure of success. Unhappy customers don't stick around. What's more, unhappy customers rarely voice their dissatisfaction before leaving.

Santander Bank is asking Kagglers to help them identify dissatisfied customers early in their relationship. Doing so would allow Santander to take proactive steps to improve a customer's happiness before it's too late.

In this competition, you'll work with hundreds of anonymized features to predict if a customer is satisfied or dissatisfied with their banking experience.

$60,000 • 674 teams

Santander Customer Satisfaction

Merger and 1st Submission Deadline

Wed 2 Mar 2016 — Mon 2 May 2016 (56 days to go)

Dashboard

Public Leaderboard - Santander Customer Satisfaction

This leaderboard is calculated on approximately 50% of the test data.
The final results will be based on the other 50%, so the final standings may be different.

See someone using multiple accounts?
Let us know.

#	Δ1d	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	BreakfastPirate *	0.841667	20	Mon, 07 Mar 2016 19:25:16 (-2.2d)
2	—	NxGTR *	0.841416	17	Mon, 07 Mar 2016 06:32:43 (-3.2d)
3	—	anokas *	0.841367	21	Mon, 07 Mar 2016 21:45:07 (-45.7h)
4	↑1	DS.RESEARCH	0.841221	25	Mon, 07 Mar 2016 06:01:14 (-0.3h)
5	↓1	Babar16	0.841218	8	Sun, 06 Mar 2016 21:36:32
6	—	Dimitris Leventis	0.841136	25	Mon, 07 Mar 2016 11:38:55 (-3.1d)
7	—	carl	0.841116	14	Sun, 06 Mar 2016 21:38:37 (-24h)
8	—	Florian	0.841112	7	Fri, 04 Mar 2016 07:07:45
9	—	Kim Quy	0.841085	6	Sun, 06 Mar 2016 09:29:59
10	—	Robert Martin	0.841060	30	Mon, 07 Mar 2016 21:28:55 (-0.1h)
11	↑400	YaronBlinder	0.840953	4	Mon, 07 Mar 2016 19:56:11

$60,000 • 674 teams

Santander Customer Satisfaction

Merger and 1st Submission Deadline

Wed 2 Mar 2016 — Mon 2 May 2016 (56 days to go)

Dashboard

Competition Forum

New topic | Start Watching

Search

1 2 >

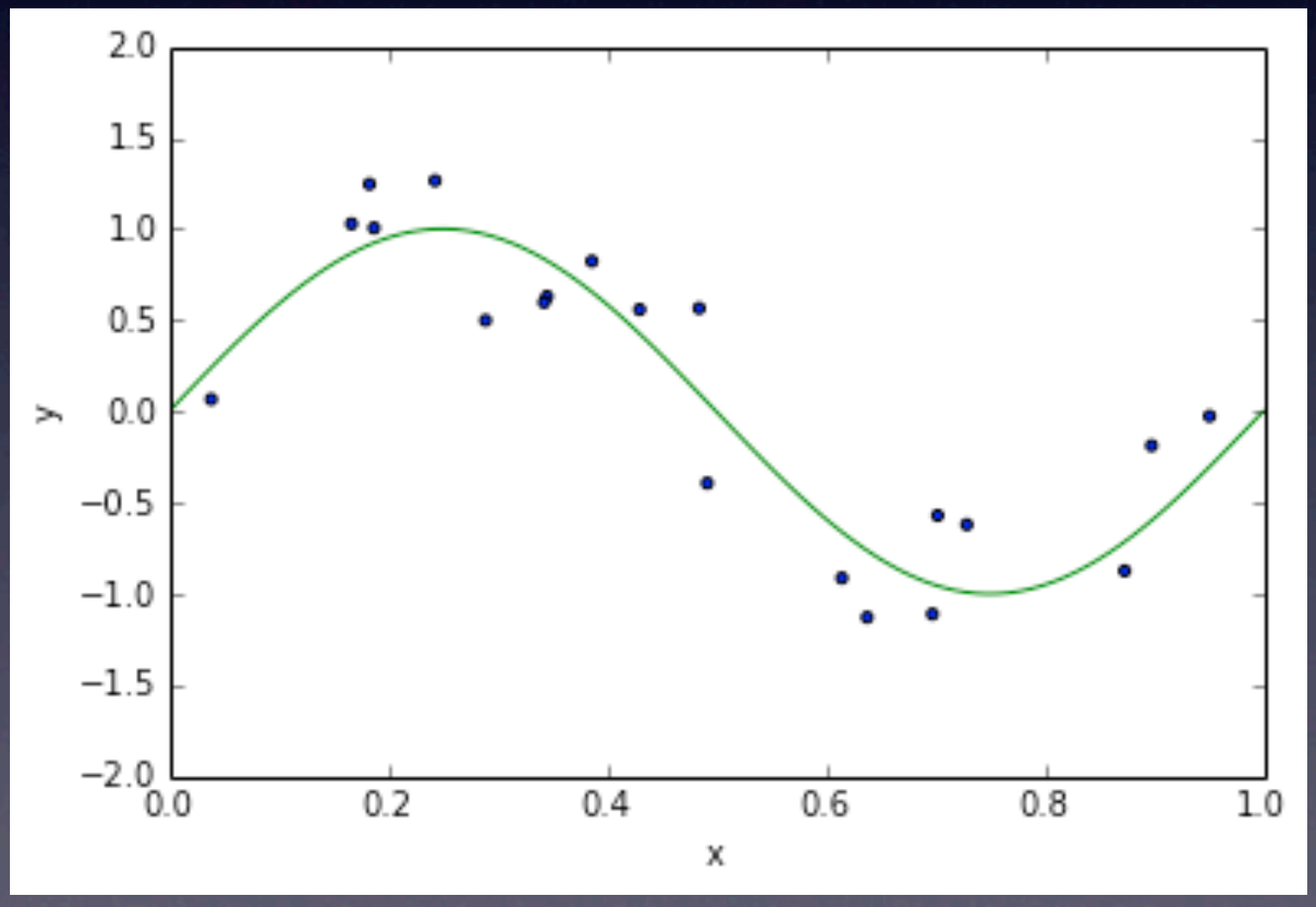
Votes	28 topics, 121 posts	Replies	Views	Last Post
0	PCA visualization *by* Tuomas Tikkanen, 2 days ago	0	0	Ben Hamner 22 minutes ago
0	Highly Correlated Variables *by* Vivek Sarma, yesterday	0	0	Ben Hamner 28 minutes ago
0	var3 *by* EvgenyZavalnyuk, 3 hours ago	4	62	BreakfastPirate 28 minutes ago
0	LB ~0.84 for starters *by* _孙鬼龙_, 2 hours ago	0	0	_孙鬼龙_ 1 hour ago

ML project: general steps

- Data selection and pre-processing
- Data splitting (cross-validation)
- Feature selection (remove) and feature engineering (add)
- Model selection & optimization
- Deployment/Prediction phase

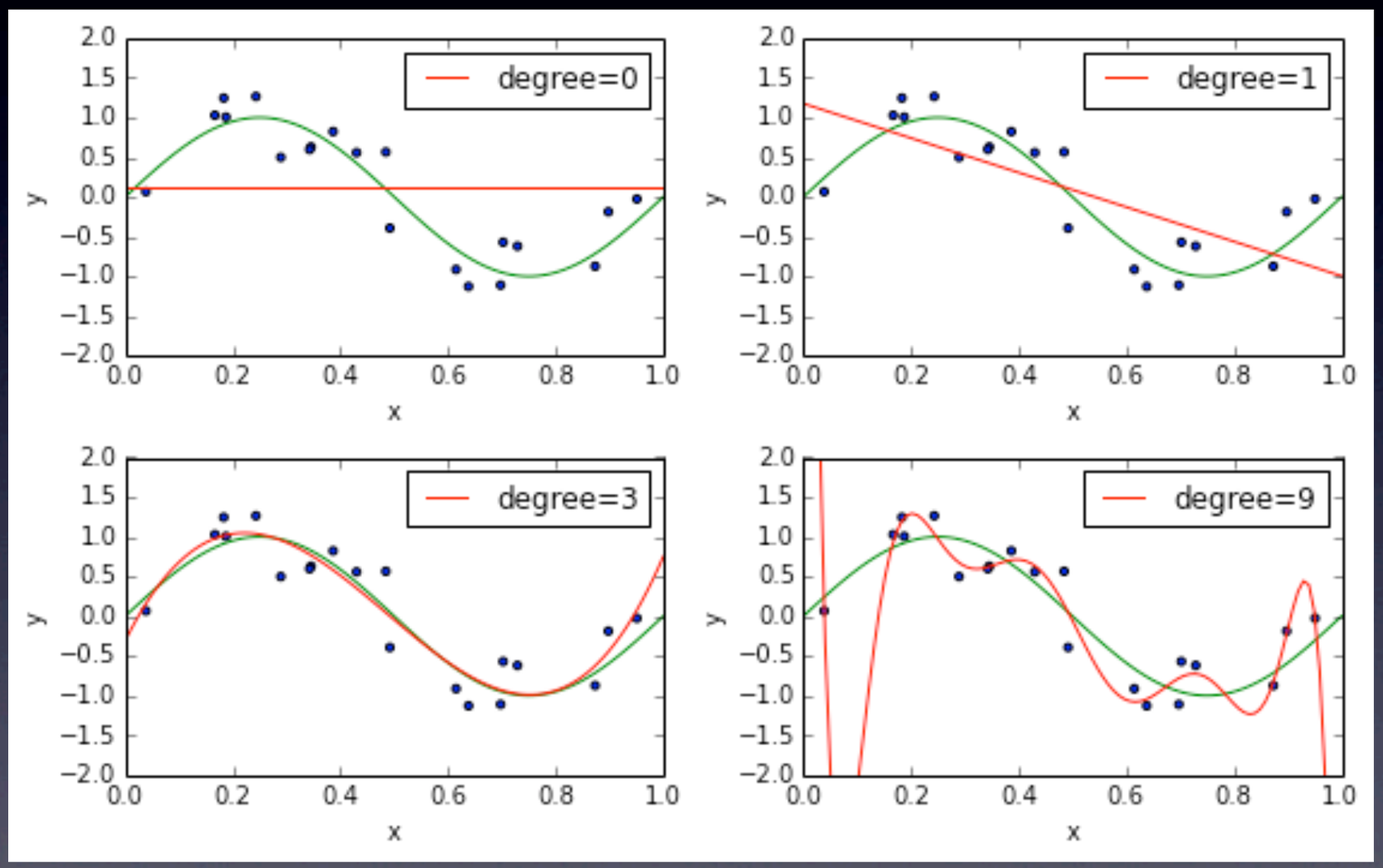
Synthetic Data

Generate 20 samples from sinusoid + gaussian noise: y=f(x)



Fitting choices

Best fit polynomials of various degrees (minimizing the squared residuals)

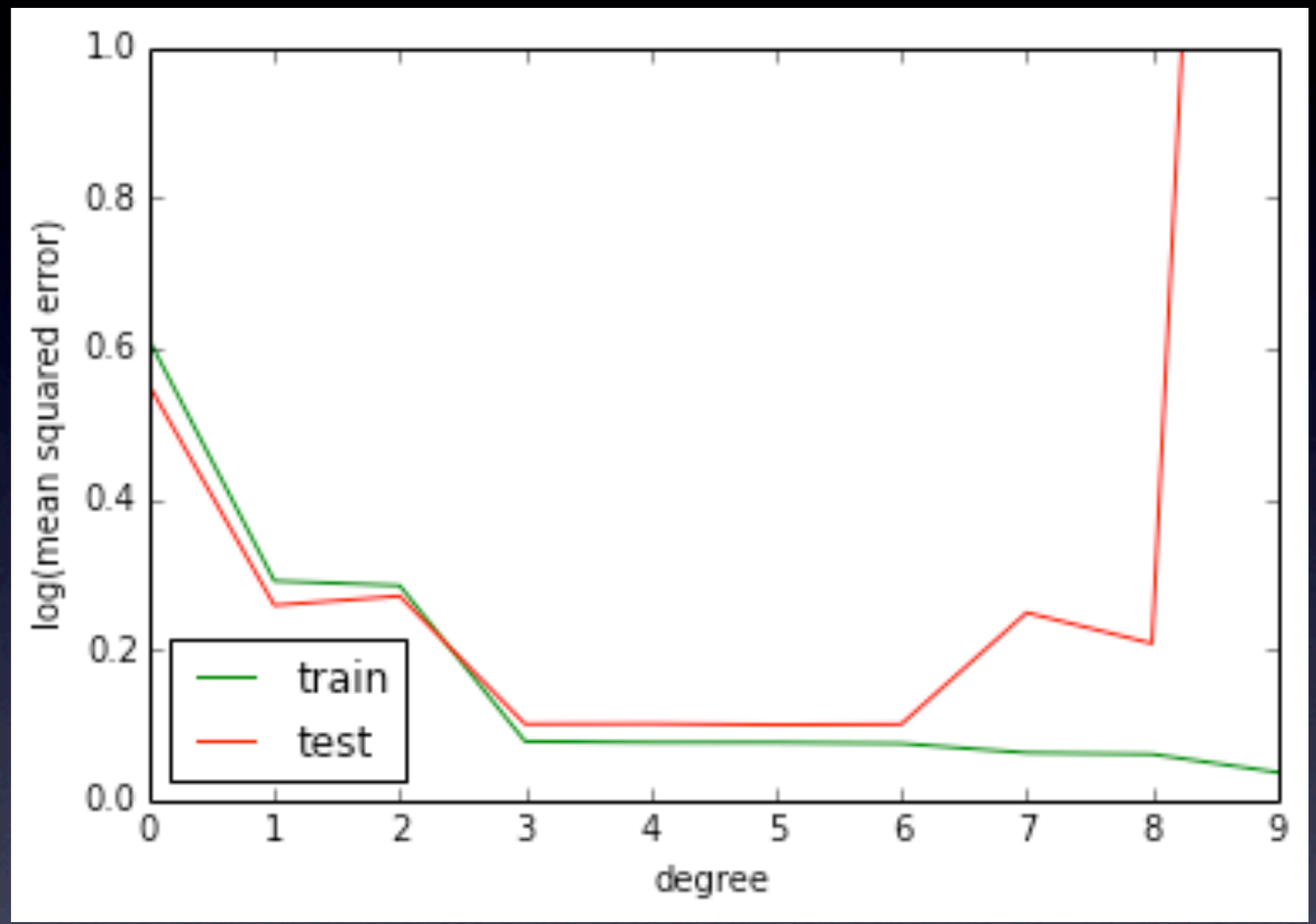


Good?

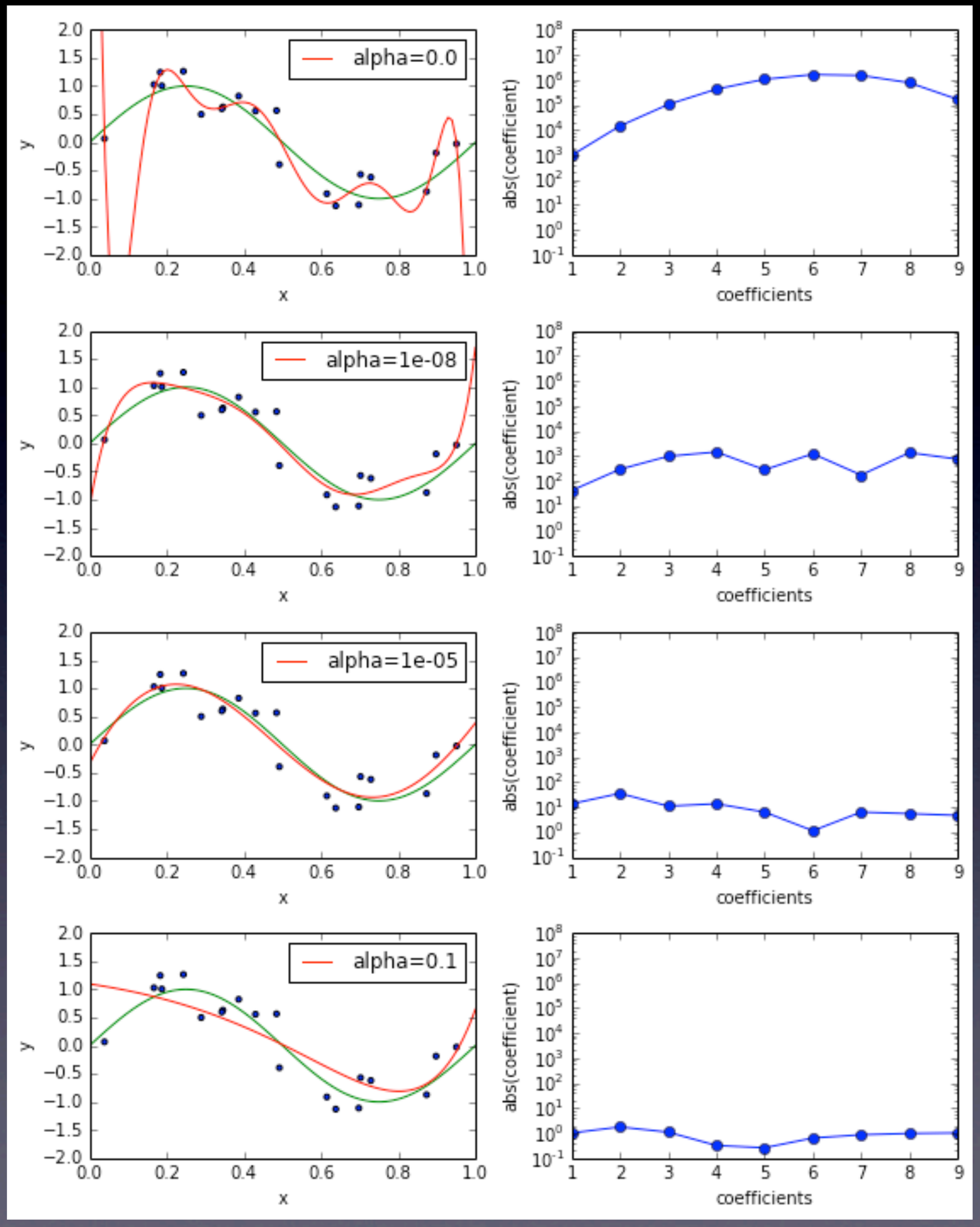
Fits the noise?
"Overfitting"

Train-Test Split

Split data points in train (say 2/3) and test (1/3)



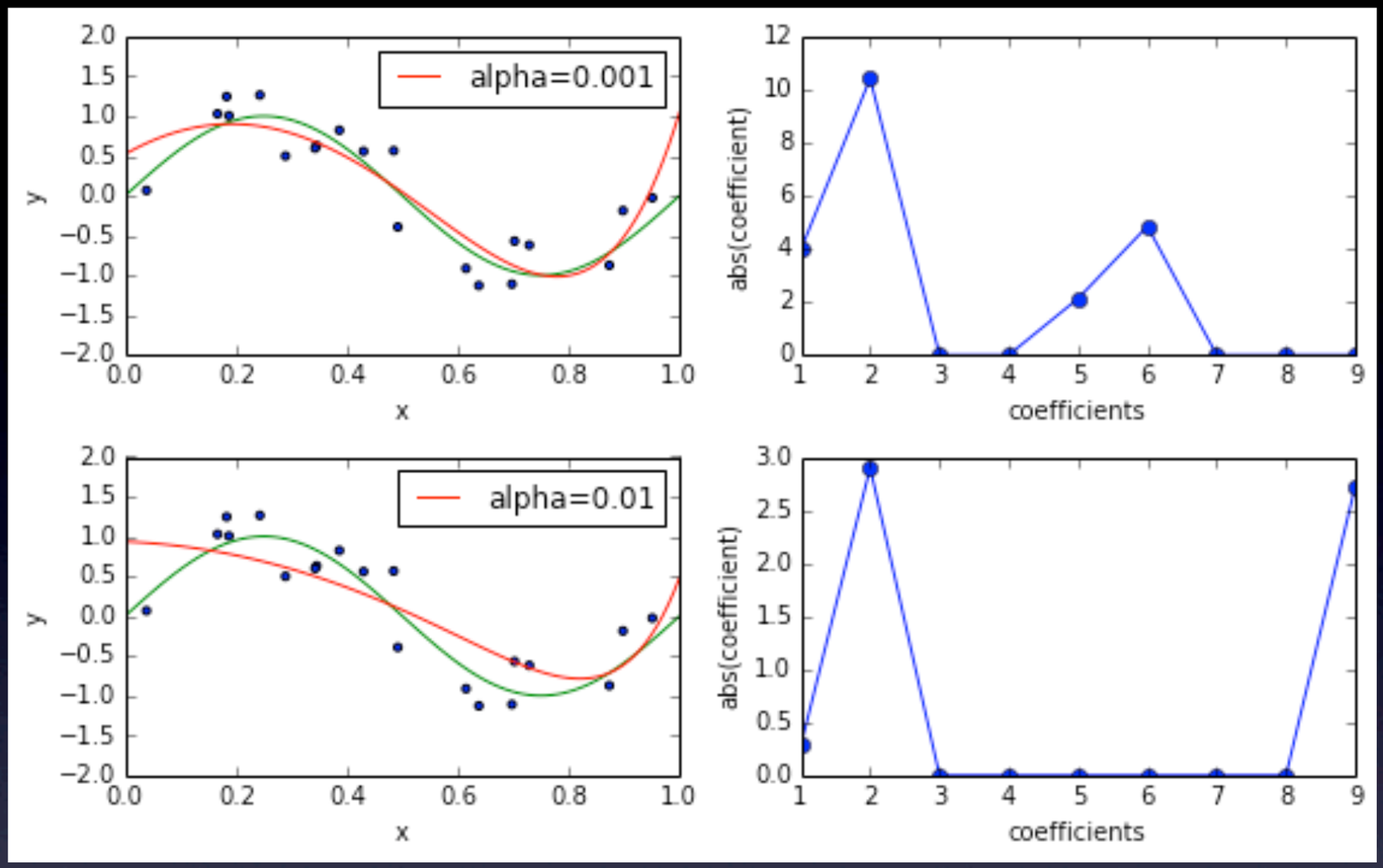
Fit/Learn/Train on train set, predict on test set (mean squared error). Best model will "generalize" best on the test data (rather than "fitting the noise" in the train data)



Strategy 1:
limit complexity

Strategy 2:
regularization.

Damp coefficients of polynomial fit. Adjust free parameter with train/test validation.



Sparse regularization: zero-out coefficients preferentially (only 3-4 non-zero).

Again train/test validation required for model evaluation.

A kind of automated feature selection!