Learning from Data
Lecture 5: Practice with new tasks and datasets

Malvina Nissim and Johannes Bjerva
m.nissim@rug.nl, j.bjerva@rug.nl

ESSLLI, 26 August 2016

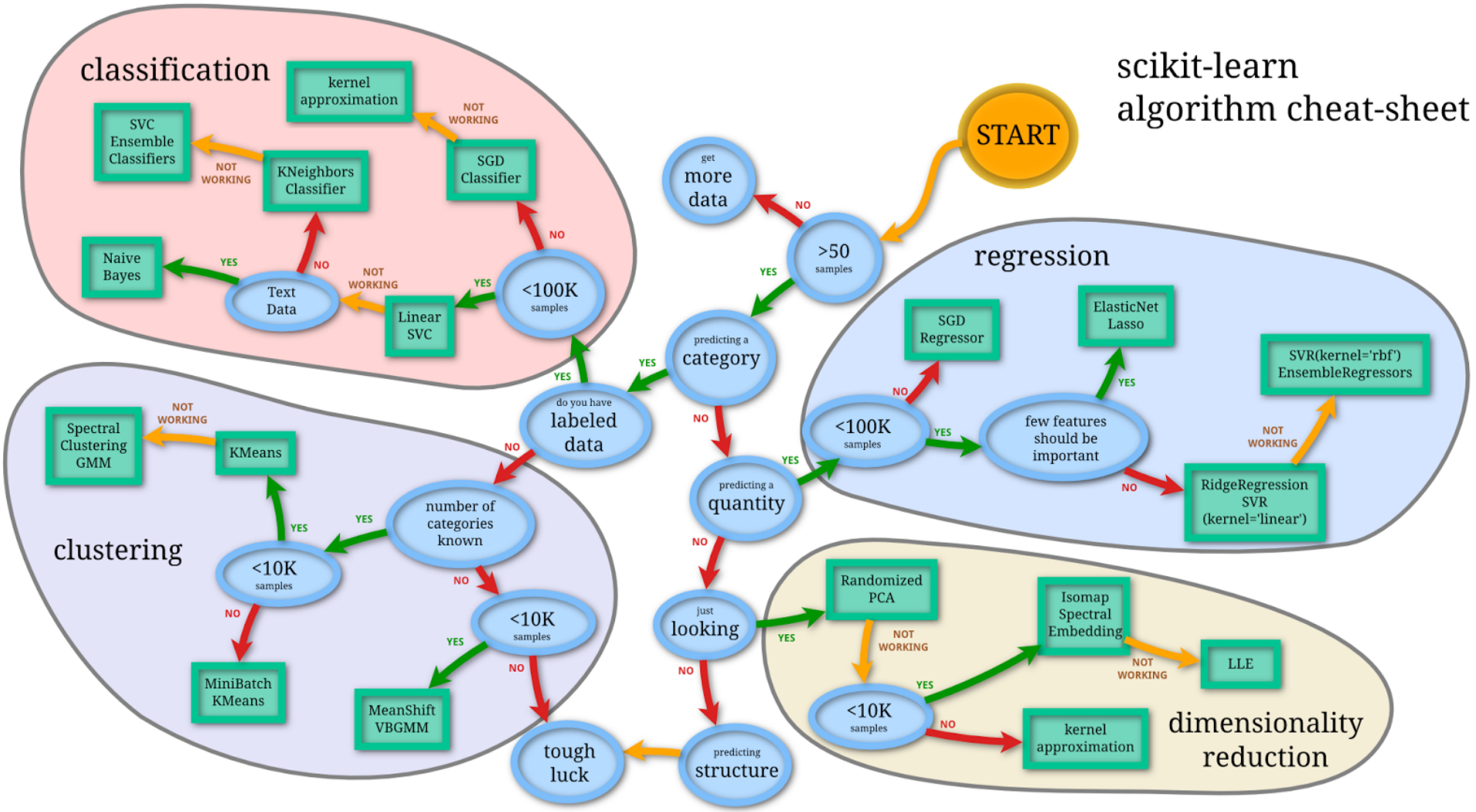
Getting the updated code

START AS SOON AS YOU ENTER THE ROOM! (PLEASE)

IMPORTANT AS ALL THE NEW DATASETS ARE IN

- Approach 1: Use *git* (updateable, recommended if you have *git*)
  1. In your terminal, type: 'git clone https://github.com/bjerva/esslli-learning-from-data-students.git'
  2. Followed by 'cd esslli-learning-from-data-students'
  3. Whenever the code is updated, type: 'git pull'
- Approach 2: Download a zip (static)
  1. Download the zip archive from: `https://github.com/bjerva/esslli-learning-from-data-students/archive/master.zip`
  2. Whenever the code is updated, download the archive again.

how to choose the "right" algorithm?


scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
<100K samples
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
NOT WORKING
SVC Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
NOT WORKING
kernel approximation
NO
SGD Classifier
classification
NO
predicting a quantity
YES
<100K samples
NO
SGD Regressor
YES
few features should be important
YES
ElasticNet Lasso
NO
RidgeRegression SVR (kernel='linear')
NOT WORKING
SVR(kernel='rbf') EnsembleRegressors
regression
NO
number of categories known
YES
<10K samples
YES
KMeans
NOT WORKING
Spectral Clustering GMM
NO
<10K samples
NO
MiniBatch KMeans
YES
MeanShift VBGMM
clustering
NO
just looking
YES
Randomized PCA
NOT WORKING
<10K samples
YES
Isomap Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck

Classification vs Regression

create models of prediction from gathered data

- classification
  the dependent variables are categorical
  - input x: feature vector
  - output: **discrete class label**

- regression
  the dependent variables are numerical
  - input x: feature vector
  - output y: **continuous value**

classification and regression are the most standard ways of doing **supervised learning**

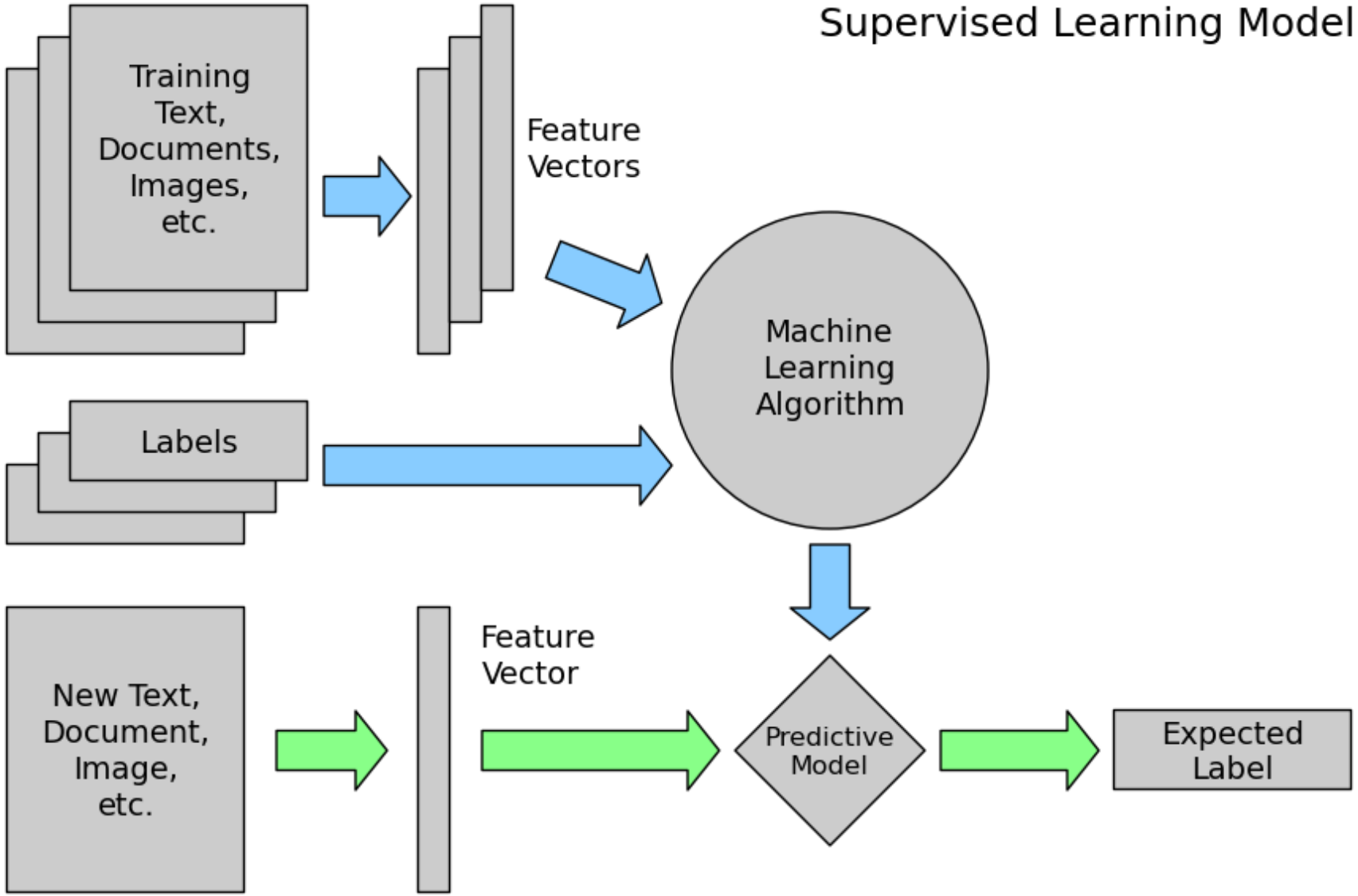
Supervised and Unsupervised Learning

information about the **correct** distribution/label of the training examples

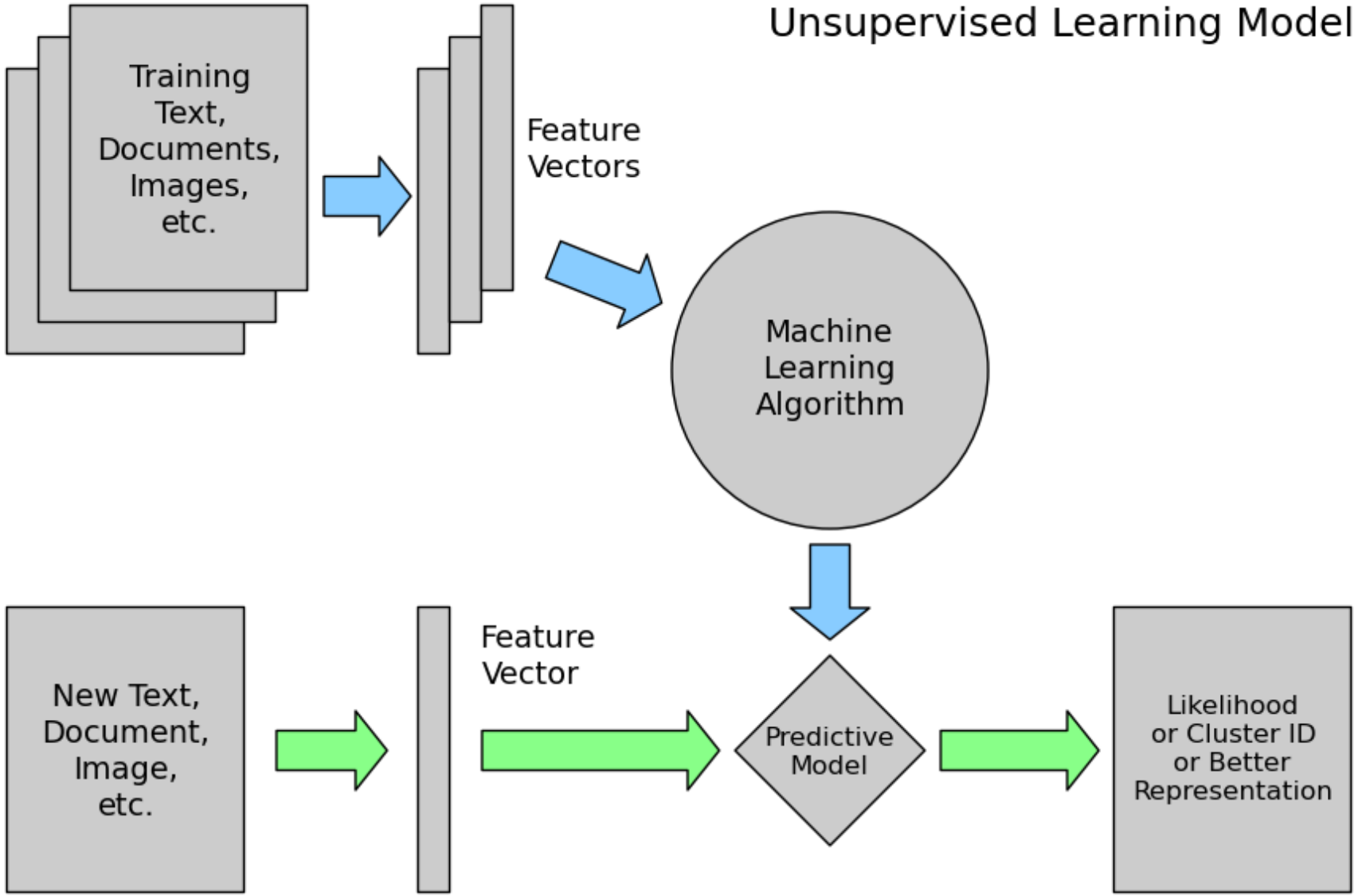
- in supervised learning it is known

  $\rightarrow$ fitting a model to labelled data which has the correct answer associated to it
- in unsupervised learning it is not known

  $\rightarrow$ finding structure in unlabelled data



source: http://www.astroml.org/



source: http://www.astroml.org/

Supervised learning

supervised learning – classification or regression

- in training, instances are associated with their class label
- based on features, the system must search for patterns and build a model
- the model must be able to *predict* the class of previously unseen instances

Unsupervised learning

unsupervised learning – clustering

- partitioning instances into subsets (clusters) that share similar characteristics
- subsets are not predefined
- a system can be told *how many* clusters it should form (K-means algorithm)

practice

datasets

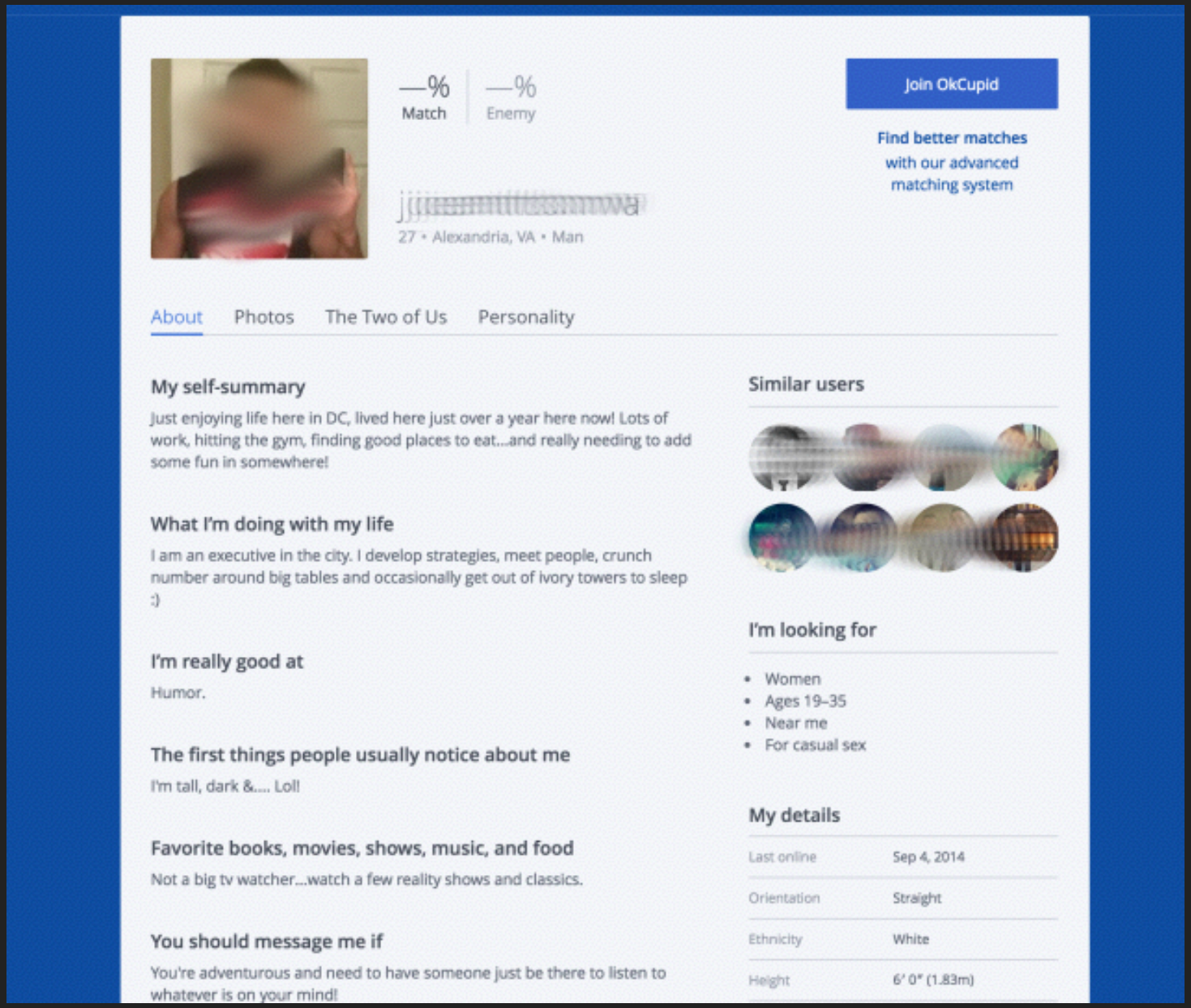
(Your) Datasets

- Ok Cupid (4 tasks)
- Word Sense Disambiguation in Russian (4 words)
- Slovene Regional Language Variants
- Pragmatic conditionals

FILIP KLUBIČKA

LEARNING FROM OKCUPID DATA

THE OKCUPID DATASET



—% Match —% Enemy

Join OkCupid

Find better matches with our advanced matching system

27 • Alexandria, VA • Man

About Photos The Two of Us Personality

My self-summary

Just enjoying life here in DC, lived here just over a year here now! Lots of work, hitting the gym, finding good places to eat...and really needing to add some fun in somewhere!

What I'm doing with my life

I am an executive in the city. I develop strategies, meet people, crunch number around big tables and occasionally get out of ivory towers to sleep :)

I'm really good at

Humor.

The first things people usually notice about me

I'm tall, dark &..... Lol!

Favorite books, movies, shows, music, and food

Not a big tv watcher...watch a few reality shows and classics.

You should message me if

You're adventurous and need to have someone just be there to listen to whatever is on your mind!

Similar users

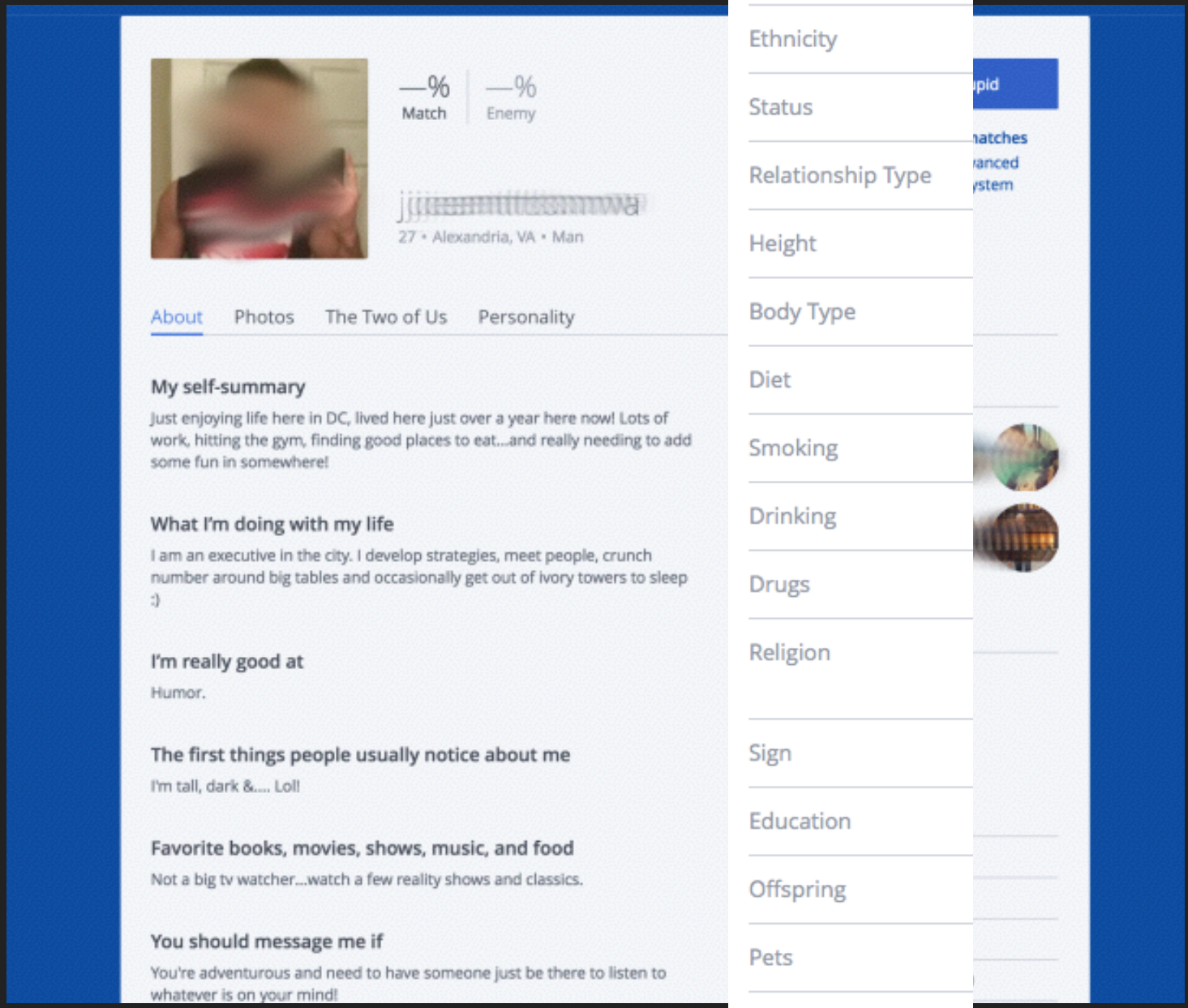
I'm looking for

- Women
- Ages 19–35
- Near me
- For casual sex

My details

Last online	Sep 4, 2014
Orientation	Straight
Ethnicity	White
Height	6' 0" (1.83m)

THE OKCUPID DATASET



EXAMPLE OF USER FEATURE VECTOR

username	age	body_type	diet	drinks	drugs	education	Essay0 – My self-summary	Essay1 – What I'm doing with my life	Essay2 – I'm really good at	Essay3 – The first thing people notice about me
----w----	22	a little extra	vegetarian	socially	never	working on college/university	I would love to think that i was some some kind of intellectual: either the dumbest smart guy, or the smartest dumb guy. can't	currently working as an international agent for a freight forwarding company. import, export, domestic you know the works. online	making people laugh. ranting about a good salting. finding simplicity in complexity, and complexity in simplicity.	the way i look. i am a six foot half asian, half caucasian mutt. it makes it tough not to notice me, and for me to blend in.

Essay4 – Favorite books, movies, shows, music and food	Essay5 – The six things I could never do without	Essay6 – I spend a lot of time thinking about	Essay7 – On a typical Friday night, I am	Essay8 – The most private thing I'm willing to admit	Essay9 – You should message me if
books: absurdistan, the republic, of mice and men (only book that made me want to cry), catcher in the rye, the prince. movies: gladiator,	food. water. cell phone. shelter.	duality and humorous things	trying to find someone to hang out with. i am down for anything except a club.	i am new to california and looking for someone to wisper my secrets to.	you want to be swept off your feet! you are tired of the norm. you want to catch a coffee or a bite. or if you want to talk philosophy.

ethnicity	height	income	job	last_online	location	offspring	orientation	pets	religion	sex	sign	smokes	speaks	status
asian, white	75	-1	student	2012-06-28-20-30	oakland, california	doesn't have kids, but might want them	straight	likes dogs and likes cats	agnosticism and very serious about it	m	gemini	sometimes	english	single

- https://github.com/everett-wetchler/okcupid
- http://www.stephenleefischer.com/posts/scraping-okcupid-will-bot-for-love

FOUR TASKS

- Predict GENDER
- Predict ORIENTATION
- Predict DIET
- Predict SIGN

NB: there is one dataset per task

WSD for Russian nouns

Task: to predict word sense given context

Words: lavka (2), kran (2), kosak (4), ruchka (5)

Contexts:

- 185 - 200 for each word
- from the Internet corpus
- about 10 words before and after the target word

Variations:

- left context / full context
- with lemmatization / without lemmatization

WSD for Russian nouns

lavka

1 

2 

1 

2 

3 

4 

kosak

ruchka

1 

2 

3 

4 

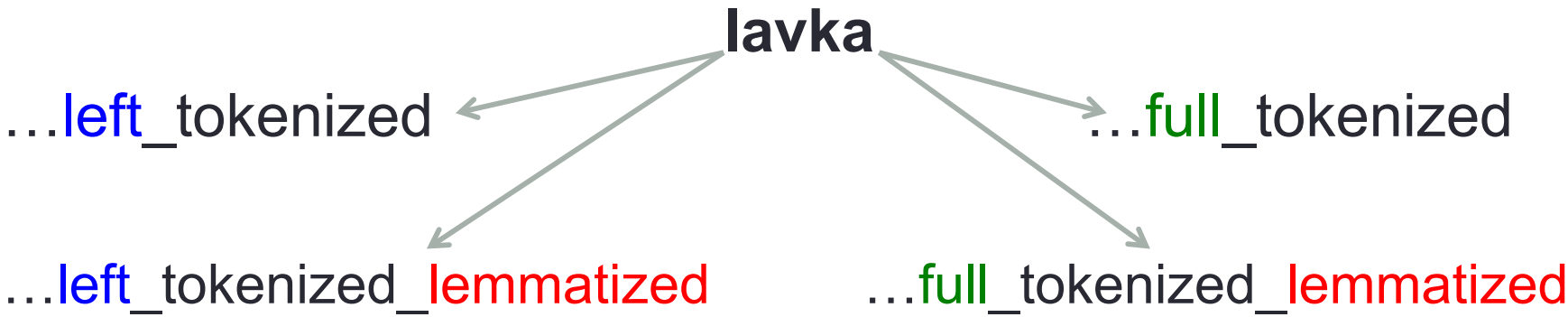
5 

kran

1 

2 

WSD for Russian nouns. Datasets



label	text-cat	targetword-cat
2	одессит игорь который уже привлекался к ответственности за разбой и 35 летний валерий зайдя в ювелирной нанесли тяжкие телесные повреждения продавцу кастетом травмировали лицо и все же девушке удалось	лавке

Classification of Slovene Regional Language Variants

- Dataset: Slovene tweets manually annotated by user's region of origin
  - 4 out of 7 Slovene regions (Gorenjska, Dolenjska, Štajerska, Primorska)
  - 500 tweets per region = 2000 tweets
- Task: build a model to predict the regional language variant of Slovene tweets

498	Dolenjska,@LukaD_ saj bi si še skoraj verjel ... jeba, hokej je za zvečer. Sporoč, da ni nič hujšga.
499	Dolenjska,@klavdijaactual sam da ni zamrzjeno ;)
500	Dolenjska,@matija10 @GregaBorinc @lisjakm ravno danes ponoč sem izgubil svoj dormeo, ker so moje pur
501	Dolenjska,@PristovnikB okol jurja ja, mal manj mogoce. Pr nas so bl hribcki. Urca pa pol pa se to se mi ni da
502	Gorenjska,Razturil ene 5 ljudi kot pobegli vlak. Carski filing, ko mors iz sebe spravt vso sranje, ki se nabere te
503	Gorenjska,@sivanosoroginja Je pa nekej,de mam jst vč šampanca,kt Ti jastoga. . . pa dobr'tek ;)
504	Gorenjska,@rjutri ce bi na glas povedala bi me zihr dal na koruzo hihi
505	Gorenjska,@JsSmRenton sej, jaz morm tud kaj zaslužiti... Zato pa za pol gnara, se prsparas. D
506	Gorenjska,@StellarGirl_ ce kdo rabi pojasnilo, potem je robot! Fuck 'em all! :*
507	Gorenjska,@KoMelita jah, men so ble ušeč, zato sm jih tut kupu. pojamram pa zato k niso ble lih zastonj, pa

Pragmatic conditionals: 'if p, q' (Chi-Hé Elder, c.elder@uea.ac.uk)

Categories

- res = resultative ('if you take the class you will learn a lot')
- inf = inferential ('if it's 6.30 the class must be over')
- pch = propositional content hedge ('if I remember rightly...')
- ifh = illocutionary force hedge ('...if you see what I mean')
- tm = topic marker ('if you think about conditionals, they usually start with 'if'')
- dir = directive ('if you could just pay attention...')

Features

- 'Primary meaning' (the main message communicated)
  - Bare form 'if p, q'; p only; q only; enriched forms p', q', etc; completely overridden logical form r
- Speech act type (A = assertive, D = directive, C = commissive)
- Conditionality of p and q (Y/N)

Practical session

- running experiments in small groups (2-3)
- reporting on experiments (a couple of minutes per group, depending on how many groups there are)
  - task
  - dataset
  - set up
  - features
  - classifier
  - results
  - any reflections

General commands

general options

- `--max-train-size N` (maximum number of training samples to look at)
- `--nchars N --nwords N --features X Y Z`

visualisation options

- `--cm` (print confusion matrix + classification report)
- `--plot` (shows CM)

algorithm-specific options

- K-Nearest Neighbor (`knn`): `--k N`
- Decision Tree (`dt`): `--max-nodes N --min-samples N`

example run

```
python run_experiment.py --csv data/trainset-sentiment-extra.csv
--nchars 1 --algorithms svm --cm
```

Look into `data` folder for datasets' names (tasks with more datasets have own dir under data)

Datasets names

- Ok Cupid (4 tasks)
  - `okcupid/okcupid_data_diet.csv`
  - `okcupid/okcupid_data_orientation.csv`
  - `okcupid/okcupid_data_sex.csv`
  - `okcupid/okcupid_data_sign.csv`
- Word Sense Disambiguation in Russian (4 words)
  - look into `russian_wsd/`: 16 files (4 different settings per word)
- Slovene Regional Language Variants
  - `slovene-dialects.csv`
- Pragmatic conditionals
  - `trainif.csv`

Running on a **real** test set

With a simple modification you can apply to the `run_experiment.py` script, you can eventually evaluate your model on completely unseen data: a 15% of the whole dataset that gets held out while you develop.

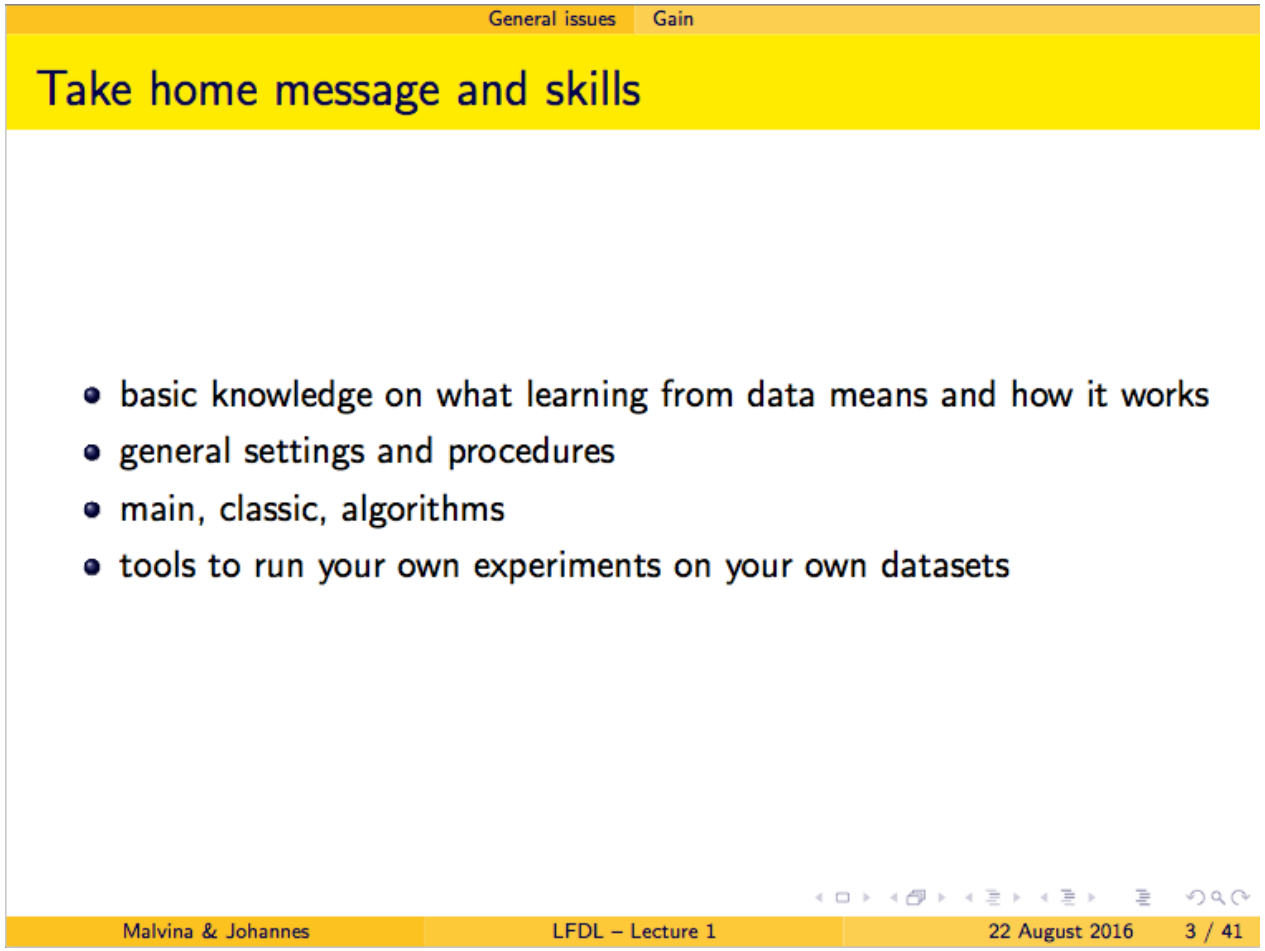
Change:

```
#print('\nResults on the test set:')
#evaluate_classifier(clf, test_X, test_y, args)
```

to

```
print('\nResults on the test set:')
evaluate_classifier(clf, test_X, test_y, args)
```

Bye bye



Malvina: m.nissim@rug.nl
Johannes: j.bjerva@rug.nl

repo: github.com/bjerva/esslli-learning-from-data-students