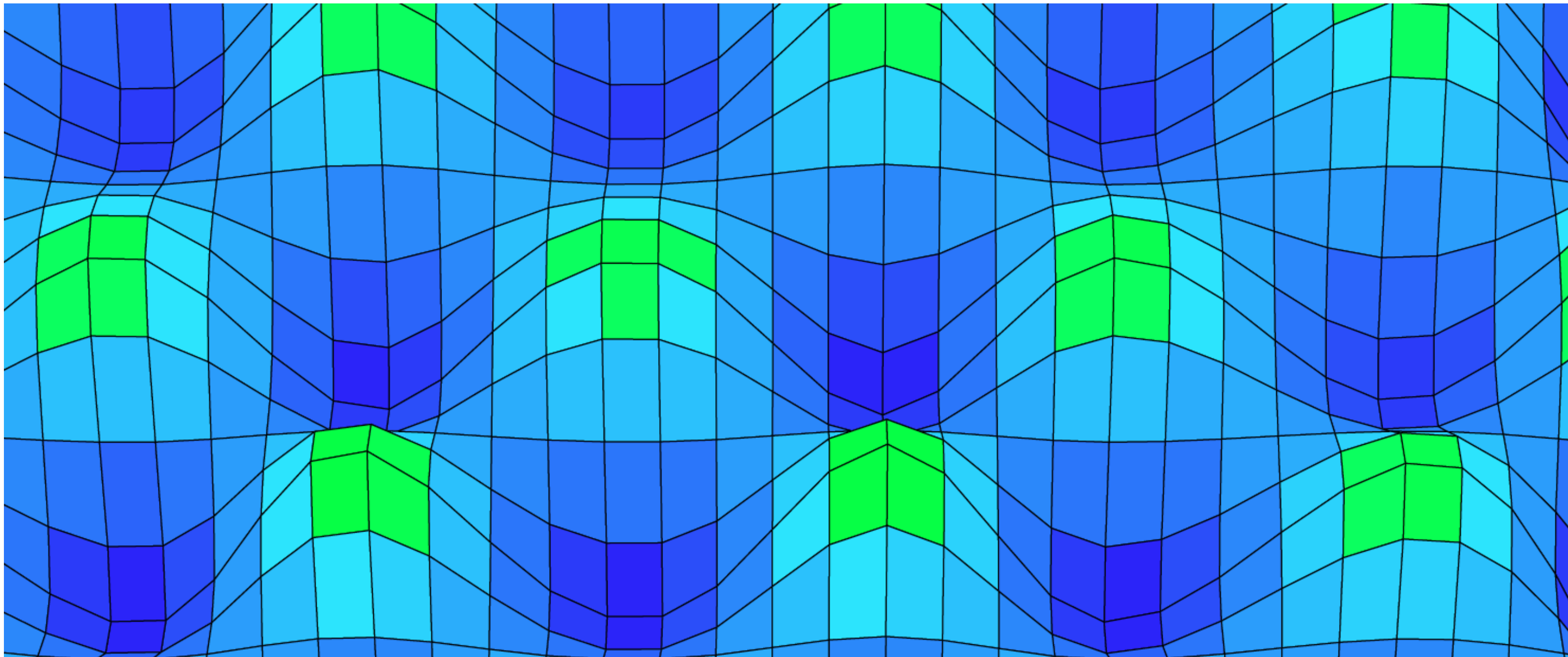
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät
Seminar für Sprachwissenschaft



Machine learning in linguistics

Peter Hendrix

Machine learning

"Machine learning explores the study and construction of algorithms that can learn from and make predictions on data"

http://en.wikipedia.org/wiki/Machine_learning

What’s cooking?

- Kaggle: “What’s cooking?”
- Text classification
- Predict cuisine based on ingredients

What's cooking?

```
# Load data
load("data/cooking.rda")
nrow(data)
[1] 39774
#
# List classes
sort(unique(data$cuisine))
 [1] "brazilian"    "british"      "cajun_creole"
 [4] "chinese"      "filipino"     "french"
 [7] "greek"        "indian"       "irish"
[10] "italian"      "jamaican"     "japanese"
[13] "korean"       "mexican"      "moroccan"
[16] "russian"      "southern_us"  "spanish"
[19] "thai"         "vietnamese"
```

What’s cooking?

```
# Look at first recipe
data$ingredients[[1]]
[1] "romaine lettuce"       "black olives"
[3] "grape tomatoes"        "garlic"
[5] "pepper"                "purple onion"
[7] "seasoning"             "garbanzo beans"
[9] "feta cheese crumbles"
#
# Which cuisine?
```

What’s cooking?

```
# Look at first recipe
data$ingredients[[1]]
[1] "romaine lettuce"       "black olives"
[3] "grape tomatoes"        "garlic"
[5] "pepper"                "purple onion"
[7] "seasoning"             "garbanzo beans"
[9] "feta cheese crumbles"
#
# Which cuisine?
data$cuisine[1]
[1] "greek"
```

What's cooking?

```
# Look at another recipe
data$ingredients[[9]]
 [1] "olive oil"                   "purple onion"
 [3] "fresh pineapple"             "pork"
 [5] "poblano peppers"             "corn tortillas"
 [7] "cheddar cheese"              "ground black pepper"
 [9] "salt"                        "iceberg lettuce"
[11] "lime"                        "jalapeno chilies"
[13] "chopped cilantro fresh"
#
# Which cuisine?
```

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

What's cooking?

```
# Look at another recipe
data$ingredients[[9]]
 [1] "olive oil"                 "purple onion"
 [3] "fresh pineapple"           "pork"
 [5] "poblano peppers"           "corn tortillas"
 [7] "cheddar cheese"            "ground black pepper"
 [9] "salt"                      "iceberg lettuce"
[11] "lime"                      "jalapeno chilies"
[13] "chopped cilantro fresh"
#
# Which cuisine?
data$cuisine[9]
[1] "mexican"
```

What's cooking?

```
# Look at a third recipe
data$ingredients[[5971]]
 [1] "fish sauce"    "napa cabbage" "scallions"
 [4] "fresh ginger" "garlic"       "chili flakes"
 [7] "chili powder" "salt"         "water"
[10] "daikon"       "pears"
#
# Which cuisine?
```


EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

What's cooking?

```
# Look at a third recipe
data$ingredients[[5971]]
 [1] "fish sauce"    "napa cabbage" "scallions"
 [4] "fresh ginger" "garlic"        "chili flakes"
 [7] "chili powder" "salt"          "water"
[10] "daikon"       "pears"
#
# Which cuisine?
data$cuisine[5971]
[1] "korean"
```


What’s cooking?

- Basic preprocessing:
    - stemming
    - bag of words
    - remove sparse terms ($n < 10$)

What's cooking?

- Training:
  - stratified sampling from labeled data
    - training set (n = 29,774)
    - validation set (n = 10,000)
  - tune model parameters using validation set performance
- Fit different classification algorithms

What's cooking?

algorithm	performance	time
gradient boosting (`xgboost`)	80.5%	<10 mins
deep learning (`h2o`)	80.4%	18 hours
neural net (`h2o`)	79.7%	1.5 hours
multinomial regression (`glmnet`)	77.7%	3.5 hours
support vector machine (`e1071`)	77.6%	1.5 hours
random forest (`randomForest`)	75.6%	20 mins
discrimination learning (`ndl`)	75.1%	<10 mins
partial least squares (`caret:pls`)	74.8%	1 hour
discriminant analysis (`MASS`)	74.6%	<10 mins
rule-based (`C50`)	71.5%	45 mins
decision tree (`C50`)	69.5%	30 mins
k nearest neighbhors (`class`)	65.9%	13 hours
naive Bayes (`klaR`)	62.5%	20 mins

What's cooking?

- Improve performance:
  - additional preprocessing
  - proper cross-validation
  - ensembling and/or stacking

Machine learning

"All models are wrong, but some are useful"

George Box

Machine learning

- Which models are useful?
- Statisticians favourite answer: “it depends”
- General criteria:
    - performance
    - computational efficiency
    - interpretability
    - plausibility