INF5830 – 2017 FALL
NATURAL LANGUAGE PROCESSING

Jan Tore Lønning & Andrei Kutuzov

Today

- Part 1: Course overview
  - What?
  - How?

- Part 2: ”Looking at data”:
  - Descriptive statistics

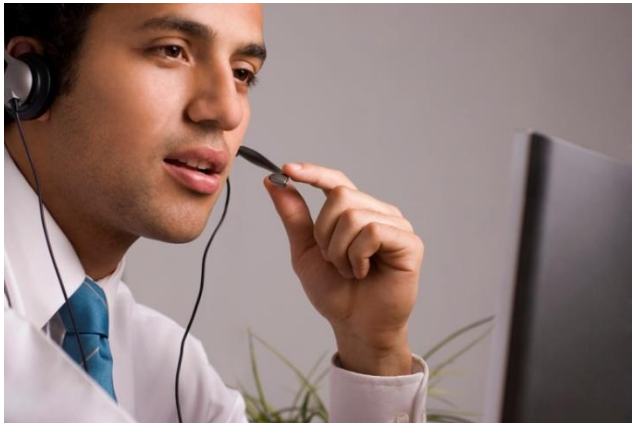
What?

Name game

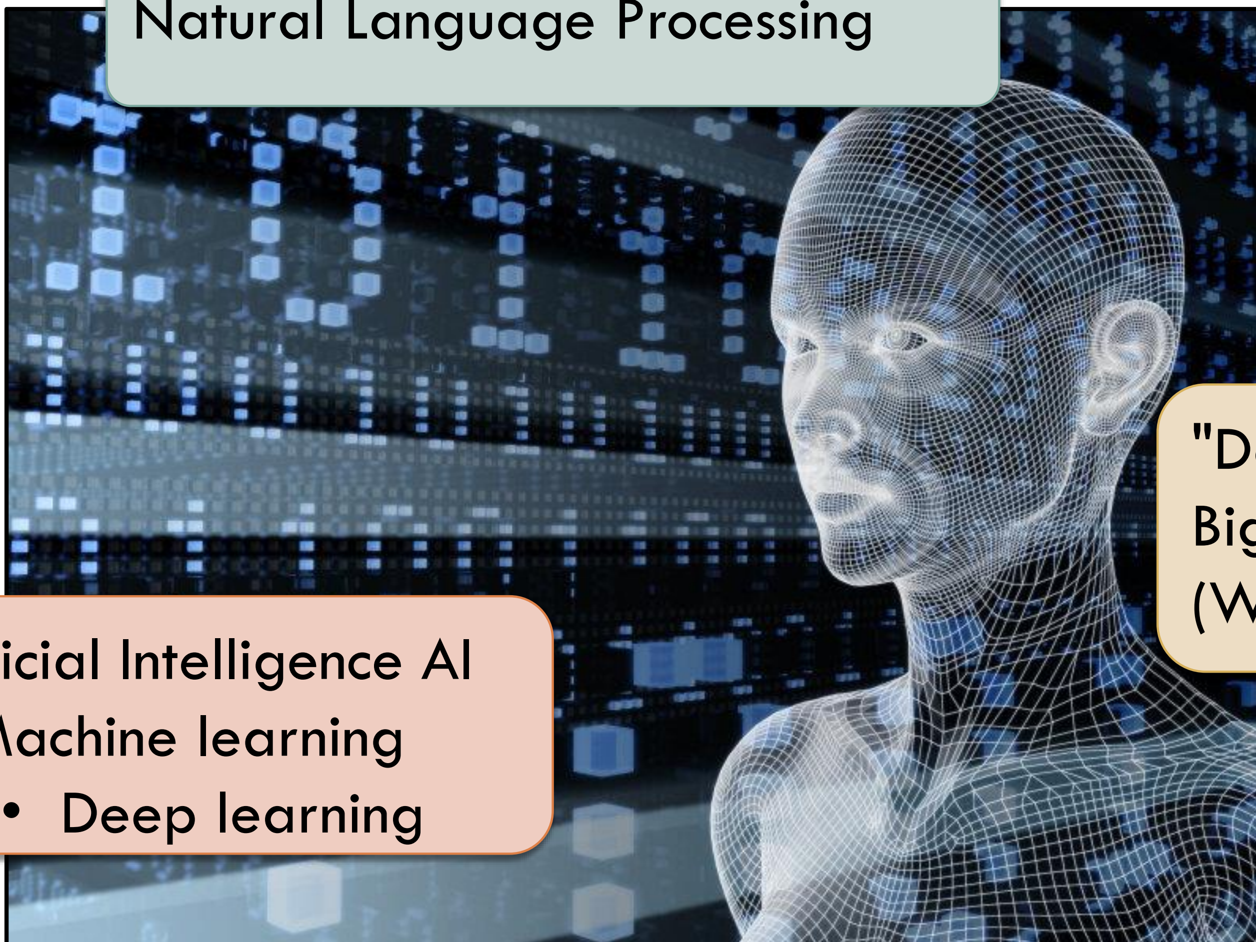
- Computational Linguistics
  - Traditional name, stresses interdisciplinarity
- Natural Language Processing
  - Computer science/AI/NLP
  - ”Natural language” a CS term
- Language Technology
  - Newer term
  - Stresses applicability
  - LT today is not SciFi (AI), but part of everyday app(lication)s
- The terms are more or less interchangeable

NLP applications - examples

- Translation (Google translate)
- Dialogue systems:
  - Personal devices
    - (Apple's Siri, Amazon Alexa, …)
  - Phone services
    - Directory, banking, tickets, etc.
- Text analyses, web data, "data science":
  - Personalization
  - Sentiment analyses
  - Intelligence



Megatrends



Natural Language Processing

"Data science"
Big data
(WWW)

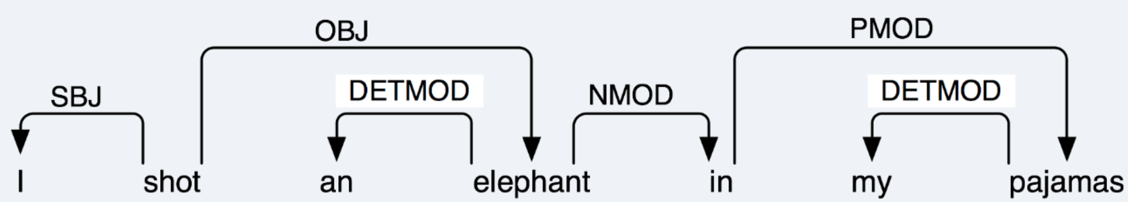
Arificial Intelligence AI
- Machine learning
  - Deep learning

What

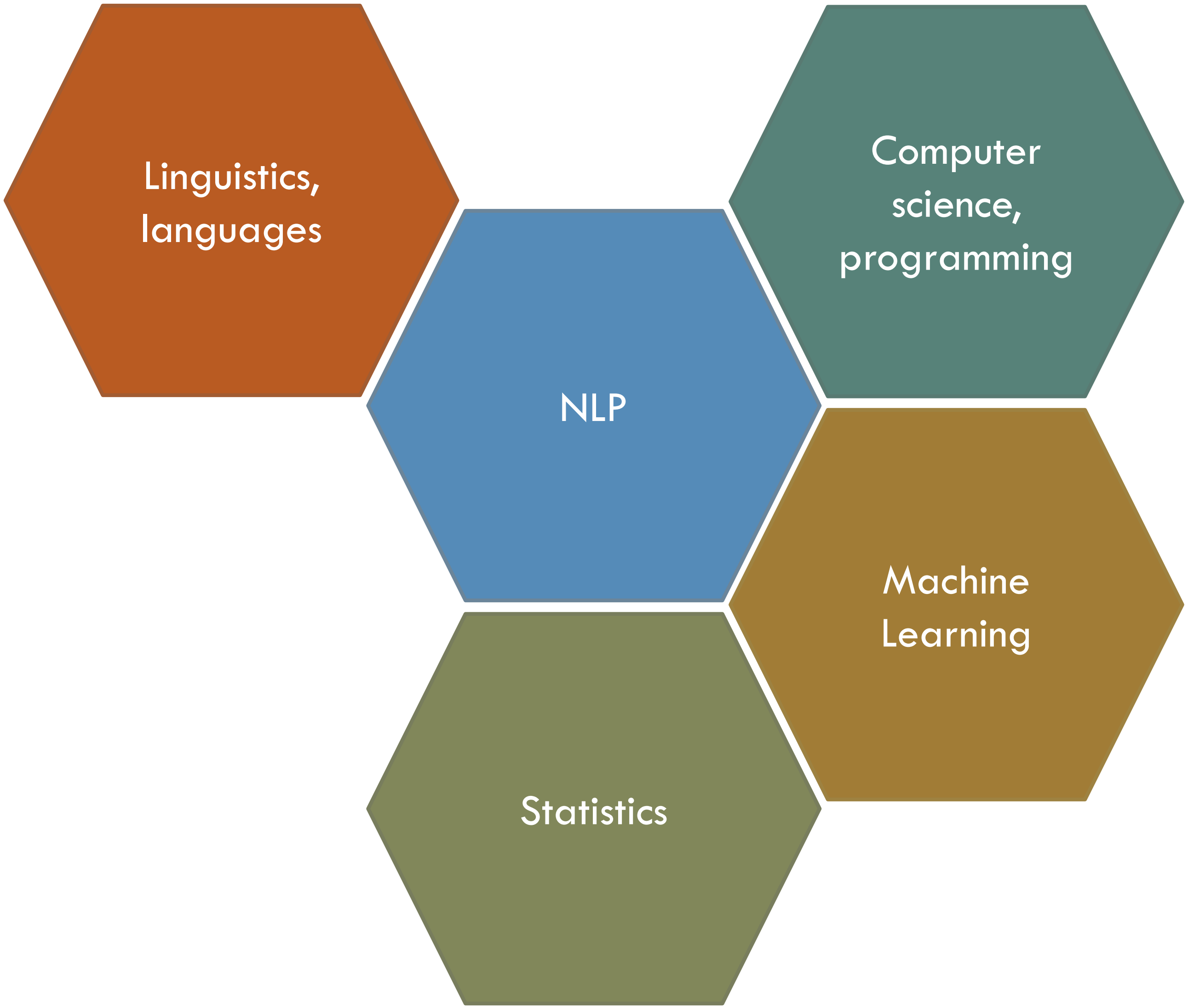
http://www.uio.no/studier/emner/matnat/ifi/INF5830/

- Follow steps in bottom-up data-driven text systems
- Learn to set-up and carry out experiments in NLP:
  - Machine learning
  - Evaluation
  - Applications of simple statistics
- Dependency parsing
- Role labeling
- … and more

Some steps when processing text

<table>
<tr><td>Split into sentences</td><td>Obama says he didn't fear for 'democracy' when running against McCain, Romney.</td></tr>
<tr><td>Tokenize (normalize)</td><td>| Obama | says | he | did| not | fear | for | ‘ | democracy | ‘ | when | running | against | McCain | , | Romney | .</td></tr>
<tr><td>Tag</td><td>Obama_N says_V he_PN did_V not_ADV fear_V …</td></tr>
<tr><td>Lemmatize</td><td>Says_V → say_V, did_V → do_V, running_V → run_V …</td></tr>
<tr><td>Parsing (dependency)</td><td>
</td></tr>
<tr><td>Coreference resolution</td><td>Obama says he did not …..</td></tr>
<tr><td>Semantic relation detect.</td><td>Fear(Obama, Democracy) Run_against(Obama, McCain),..</td></tr>
<tr><td>Negation detection</td><td>… did not fear … → Not(Fear(Obama, Democracy))</td></tr>
</table>

NLP is based on



Why statistics and probability in NLP?

1. "Choose the best"
(=the most probable given the available information)

- *bank* (Eng.) can translate to b.o. *bank* or *bredd* in No.
  - Which should we choose?
  - What if we know the context is "*river bank*"?
- *bank* can be Verb or Noun,
  - which tag should we choose?
  - What if the context is *they bank the money* ?
- A sentence may be ambiguous:
  - What is the most probable parse of the sentence?

Use of probabilities and statistics, ctd.:

2. In constructing models from examples (ML):
   - What is the best model given these examples?

3. Evaluation:
   - Model1 is performing slightly better than model 2 (78.4 vs. 73.2), can we conclude that model 1 is better?
   - How large test corpus do we need?

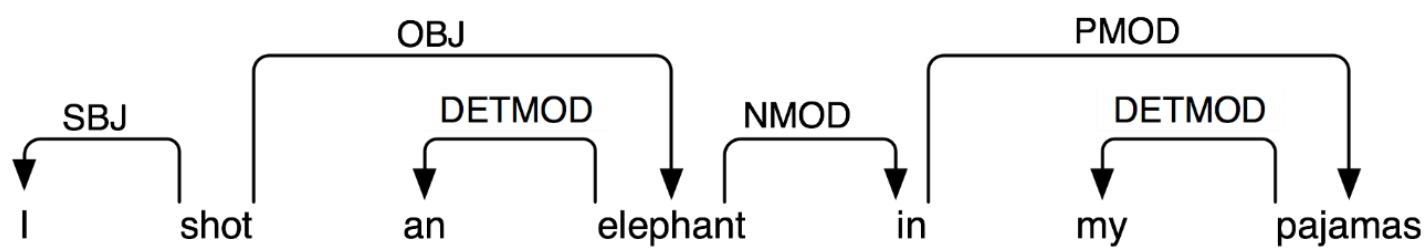
Machine learning in NLP

- "Machine learning" is the term for systems that improves by training
- Plays a mayor part in modern NLP
- For example, machine translation systems that are trained on earlier translated texts

Machine learning (ML) in INF5830

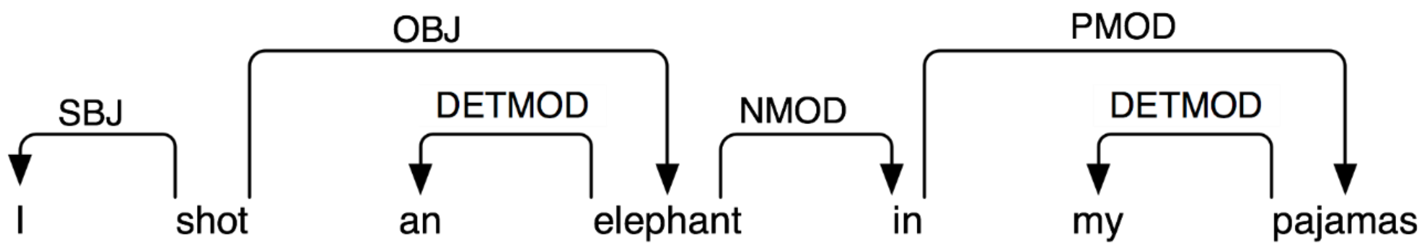
- Consider several NLP tasks that use ML
- Learn how to carry out experiments and evaluate them
- More in-depth on some ML-methods:
  - Naive Bayes,
  - Decision trees
  - Maximum entropy

Data-driven dependency parsing



- The parser is learned from data (machine learning)
- Increasing interest in dependency-based approaches to syntactic parsing in recent years:
  - new methods emerging
  - applied to a wide range of languages
  - CoNLL shared tasks (2006, 2007)

Data-driven dependency parsing



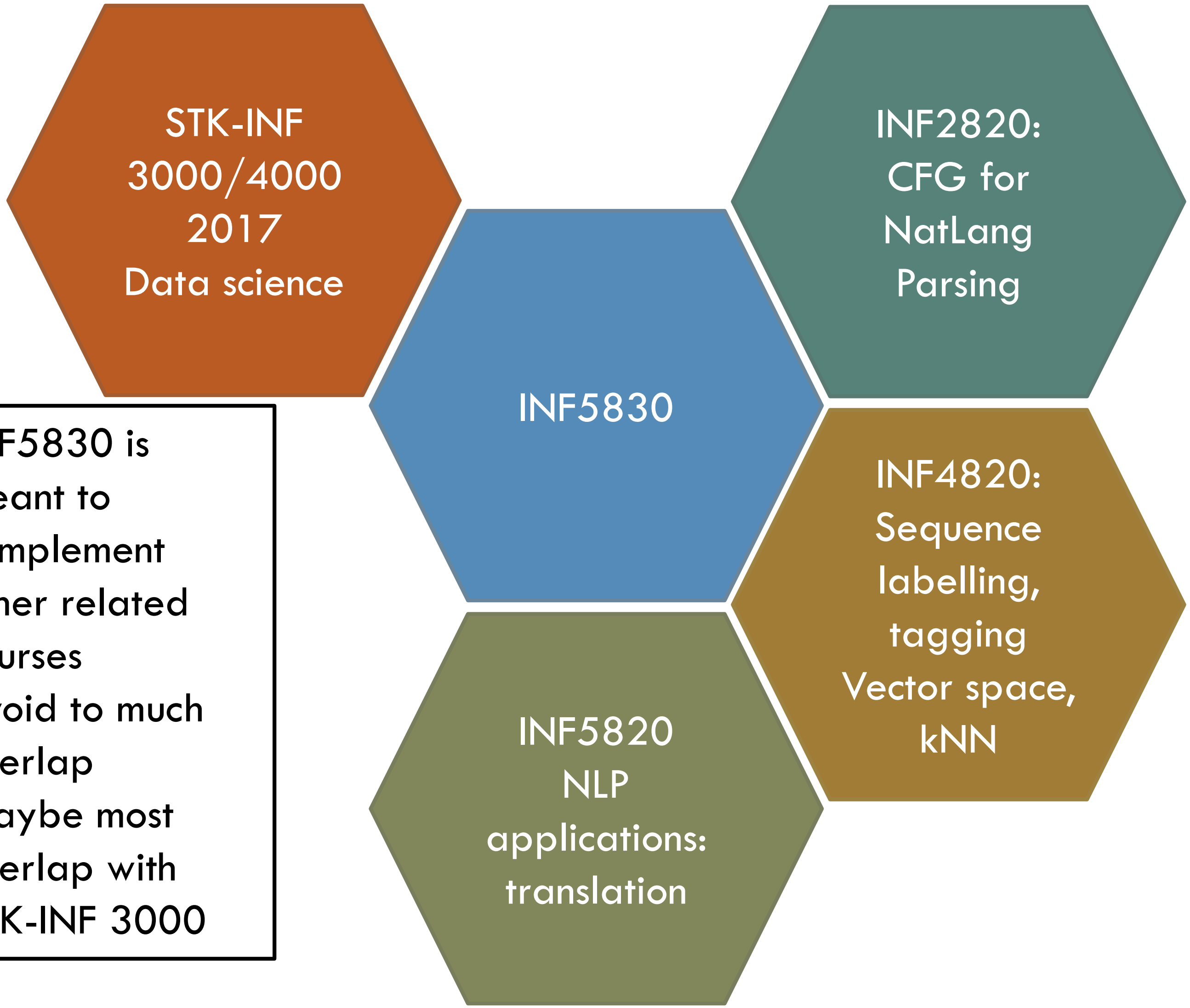
- Parsing provides "scaffolding" for semantic analysis
- Useful for down-stream applications:
  - opinion mining
  - information extraction
  - syntax-informed statistical machine translation
  - etc…

Semantic role labeling

- Semantic argument classification
  - CoNLL08, 09 shared tasks: syntactic and semantic parsing of English (2008) and other languages (2009)
  - dependency representations for semantic role labeling
- Syllabus: linguistics “classics” and research articles

What?

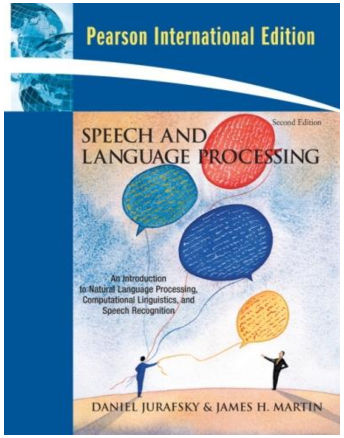
Related courses

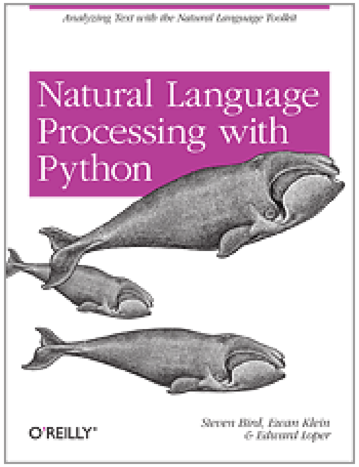


- INF5830 is meant to complement other related courses
- Avoid to much overlap
- Maybe most overlap with STK-INF 3000

Syllabus (online)

- Lectures: Presentations put on the web
- Parts of books:
  - Jurafsky and Martin,
    Speech and Language Processing
    - 3. ed (in progress), chapters online
      - *(Your advised to own 2. ed.)*
  - S. Bird, E. Klein and E. Loper:
    Natural Language Processing with Python
    - (Available online)
- Articles/web-pages/distributed material
- A book in statistics may be useful, e.g.
  - Sarah Boslaugh:*Statistics in a Nutshell*
  - www.openintro.org/stat/textbook.php





Computational "Work Bench"

- Python, general programming language:
  - Well-suited for text
  - Readable, structured code
- Packages, extensions
  - NLTK:
    - Python toolbox for NLP
    - Emphasis on training
  - NumPy
  - SciPy (stats)
  - Matplotlib
  - scikit-learn (machine learning)
- Programs for Dep. parsing

Bundle of Python-packages

- Free
- Widely used for "data science" and machine learning
- Working in the same environment as you program (contrast to R)
- Packages for deep learning uses Python (e.g. TensorFlow)

From last year's evaluation

"Even though a lot of the students taking the course this semester did not have sufficient background knowledge in statistics (myself included), I would have preferred if more time was spent on the core subject matter rather than spending time on simple statistics. The resources available (Khan Academy etc.) for learning the statistics required for this course are good enough that it should be sufficient to tell the students what we would need to learn on our own and some resources for learning it and leave it up to the students to learn this on their own if they need to. Before we start having obligatory assignments in every subject at the same time a month or so into the semester, we do have the spare time to do this on our own."

Tutorials

- We will follow this advise, with a twist
- Since your background varies, we will give some extra tutorials on subject that some of you may know and other's do not know
- Mondays 14.15-16
- First, Monday 28 Aug, Probabilities

Schedule

- Lectures: Tuesday 10.15-12
- Lab sessions: Monday 12.15-14 (not all weeks)
- Extra tutorials, Some Mondays 14.15-16 (first month or so)
- It is not 2 different groups!
  - The schedule it misleading:
- Written exam
  - December 20 at 2:30 PM (4 hours).
  - For they who fail: Exam spring 2018
    - Requires approved obligs this semester

Looking at data

Data

- Advise in "data science", machine learning and data-driven NLP:
  Start by taking a look at your data
  - (But tuck away your test data first)
- General form:
  - A set of objects
  - To each object some associated properties
    - Called variables in statistics
    - Features in machine learning
    - (Attributes in OO-programming)

Example data set: email spam

	spam	chars	lines breaks	'dollar' occurs. numbers	'winner' occurs?	format	number
1	no	21,705	551	0	no	html	small
2	no	7,011	183	0	no	html	big
3	yes	631	28	0	no	text	none
4	no	2,454	61	0	no	text	small
5	no	41,623	1088	9	no	html	small
…							
50	no	15,829	242	0	no	html	small

From OpenIntro Statistics
Creative Commons license

There are more variables (properties) in the data set

Example data set: email spam

	spam	chars	lines breaks	'dollar' occurs. numbers	'winner' occurs?	format	number
1	no	21,705	551	0	no	html	small
2	no	7,011	183	0	no	html	big
3	yes	631	28	0	no	text	none
4	no	2,454	61	0	no	text	small
5	no	41,623	1088	9	no	html	small
…							
50	no	15,829	242	0	no	html	small

50 individuals (objects, items, …)
lines

7 variables

4 categorical variables

3 numeric variables

Some words of warning

- This is how data sets often are presented in texts on
  - Statistics
  - Machine learning
- But we know that there is a lot of work before this
  1. Preprocessing text
  2. Selecting properties (variables)
  3. Extracting the properties

Text as a data set

	token	POS
1	He	PRON
2	looked	VERB
3	at	ADP
4	the	DET
5	lined	VERB
6	face	NOUN
7	with	ADP
8	vague	ADJ
9	interest	NOUN
10	.	.
11	He	PRON
12	smiled	VERB
13	.	.

- Two "variables":
  - Token type
  - POS (part of speech)

Types of (statistical) variables

Categoric		**Numeric**	
Binary (useful to separate out fo ML)		Discrete (counting)	Continuous (measuring)

Categorical variables

- Categorical:
  - Person: Name
  - Word: Part of Speech (POS)
    - {Verb, Noun, Adj, …}
  - Noun: Gender
    - {Mask, Fem, Neut}
- Binary/Boolean:
  - Email: spam?
  - Person: 18 ys. or older?
  - Sequence of word: Grammatical English sentence?

Numeric varaiables

- Discrete
  - Person: Years of age, Weight in kilos, Height in centimeters
  - Sentence: Number of words
  - Word: length
  - Text: number of occurrences of *great,* (42)
- Continuous
  - Person: Height with decimals
  - Program execution: Time
  - Occurrences of a word in a text: Relative frequency (18.666…%)

Frequencies of categorical variables

Frequencies

- Given a set of objects O
  - Which each has a variable which takes values from a set V
- To each v in V, we can define
  - The absolute frequency of v in O:
    - the number of elements x in O such that x.f = v
      - (requires O finite)
  - The relative frequency of v in O:
    - The absolute frequency/the number of elements in O

Universal POS tagset (NLTK)

Tag	Meaning	English Examples
ADJ	adjective	*new, good, high, special, big, local*
ADP	adposition	*on, of, at, with, by, into, under*
ADV	adverb	*really, already, still, early, now*
CONJ	conjunction	*and, or, but, if, while, although*
DET	determiner, article	*the, a, some, most, every, no, which*
NOUN	noun	*year, home, costs, time, Africa*
NUM	numeral	*twenty-four, fourth, 1991, 14:24*
PRT	particle	*at, on, out, over per, that, up, with*
PRON	pronoun	*he, their, her, its, my, I, us*
VERB	verb	*is, say, told, given, playing, would*
.	punctuation marks	*. , ; !*
X	other	*ersatz, esprit, dunno, gr8, univeristy*

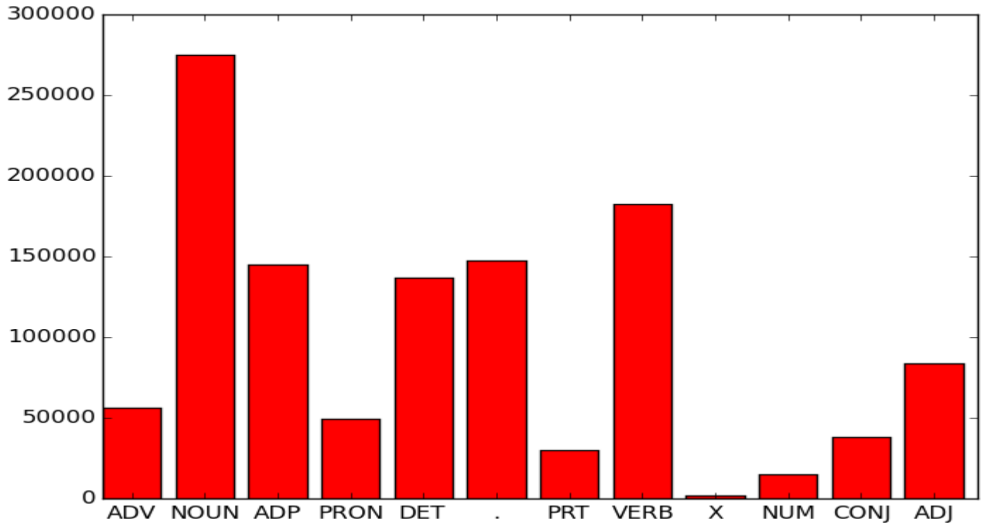
Distribution of universal POS in Brown

- Brown corpus:
  - ca1.1 mill. words
- For each word occurrence:
  - feature: simplified tag
  - 12 different tags
- Frequency(absolute)
  - for each of the 12 values:
  - the number of occurrences in Brown
- Frequency (relative)
  - the relative number
    - Same graph pattern
    - Different scale

Cat	Freq
ADV	56 239
NOUN	275 244
ADP	144 766
NUM	14 874
DET	137 019
.	147 565
PRT	29 829
VERB	182 750
X	1 700
CONJ	38 151
PRON	49 334
ADJ	83 721

(Numbers from 2015)

Distribution of universal POS in Brown



Cat	Freq
ADV	56 239
NOUN	275 244
ADP	144 766
NUM	14 874
DET	137 019
.	147 565
PRT	29 829
VERB	182 750
X	1 700
CONJ	38 151
PRON	49 334
ADJ	83 721

To better understand our data we may use graphics.
For frequency distributions, the bar chart is the most useful

Frequencies

- Frequencies can be defined for all types of value sets V (binary, categoric, numeric) as long as there are only finitely many observations or V is countable,
- But doesn't make much sense for continuous values or for numeric data with very varied values:
  - The frequencies are 0 or 1 for many (all) values

More than one feature

Example NLTK, sec. 2.1

	can	could	may	might	must	will
news	93	86	66	38	50	389
religion	82	59	78	12	54	71
hobbies	268	58	131	22	83	264
science_fiction	16	49	4	12	8	16
romance	74	193	11	51	45	43
humor	16	30	8	8	9	13

- Example of a contingency table
- Observations, O, all occurrences of the five modals in Brown
- For each observations, two parameters
  - f1, which modal, V1 = {can, could, may, might, must, will}
  - f2, genre, V2={news, religion, hobbies, sci-fi, romance, humor}

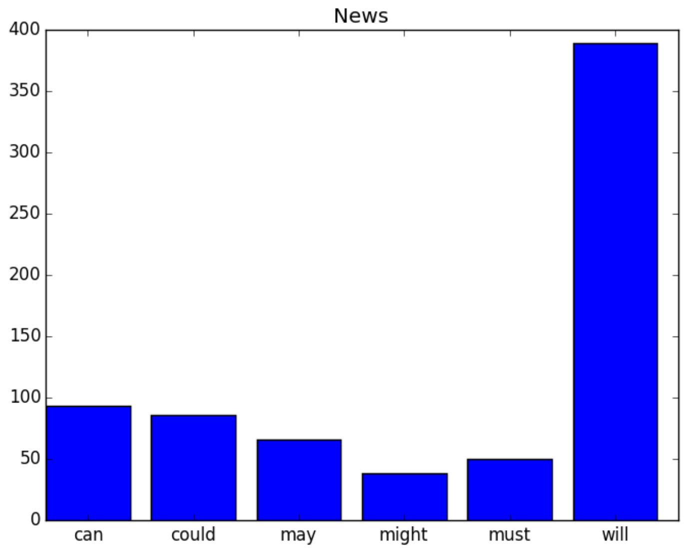
Example NLTK, sec. 2.1

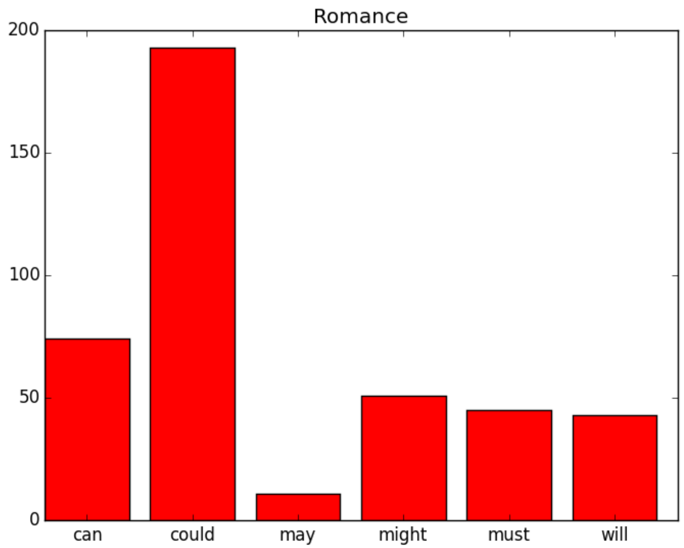
	can	could	may	might	must	will
news	93	86	66	38	50	389
religion	82	59	78	12	54	71
hobbies	268	58	131	22	83	264
science_fiction	16	49	4	12	8	16
romance	74	193	11	51	45	43
humor	16	30	8	8	9	13

- Each row and each column is a frequency distribution
- We can make a chart for each row and inspect the differences

Example NLTK, sec. 2.1

	can	could	may	might	must	will
news	93	86	66	38	50	389
religion	82	59	78	12	54	71
hobbies	268	58	131	22	83	264
science_fiction	16	49	4	12	8	16
romance	74	193	11	51	45	43
humor	16	30	8	8	9	13

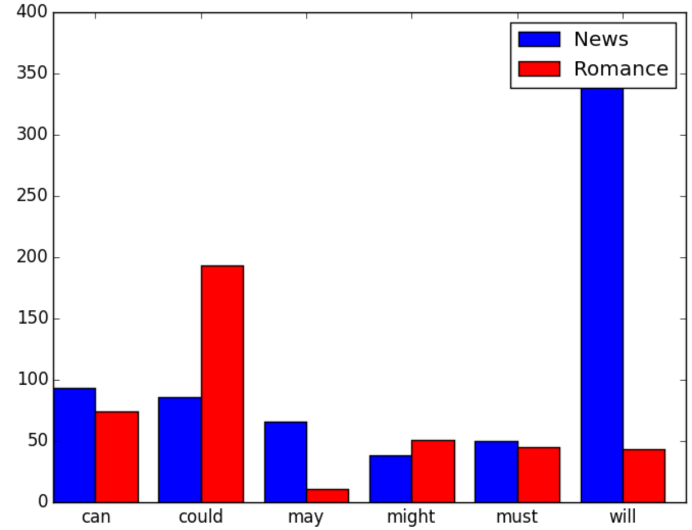




Example NLTK, sec. 2.1

	can	could	may	might	must	will
news	93	86	66	38	50	389
religion	82	59	78	12	54	71
hobbies	268	58	131	22	83	264
science_fiction	16	49	4	12	8	16
romance	74	193	11	51	45	43
humor	16	30	8	8	9	13

- Or one may combine several frequency distributions into one chart in some way

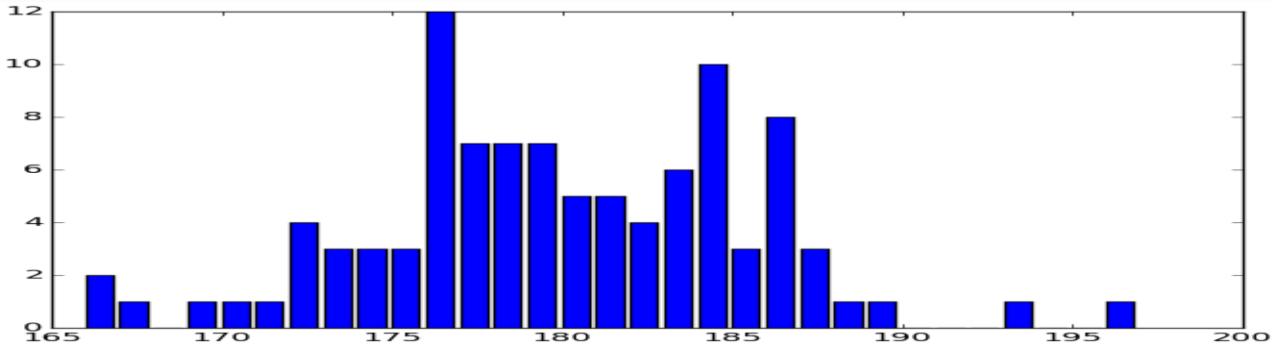


Numerical data

Numeric values

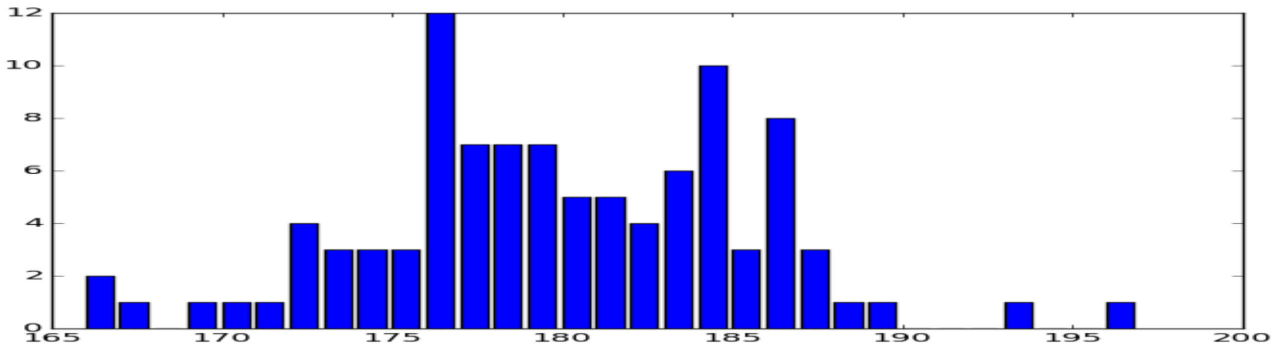
173 172 173 183 177 177 186 180 178 187 179 181 184 172 180 180 171 176 186 175 176 181 176 177 178 176 174 186 172 175 186 183 185 184 176 179 175 193 181 178 177 183 196 187 184 179 182 184 181 176 185 180 176 176 176 167 178 182 176 186 179 176 166 186 169 186 183 178 186 184 179 177 174 176 184 174 177 178 173 182 182 184 185 172 179 179 189 178 170 183 166 188 187 184 184 177 181 180 183 184

Ex 1



- When we have a set of objects with a numeric feature, we may ask more questions:
  - Max?
    - 196
  - Min?
    - 166
  - Middle, average?

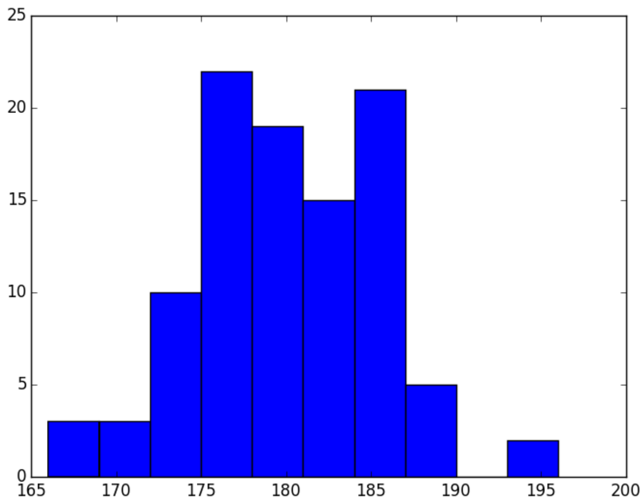
Mean, median, mode



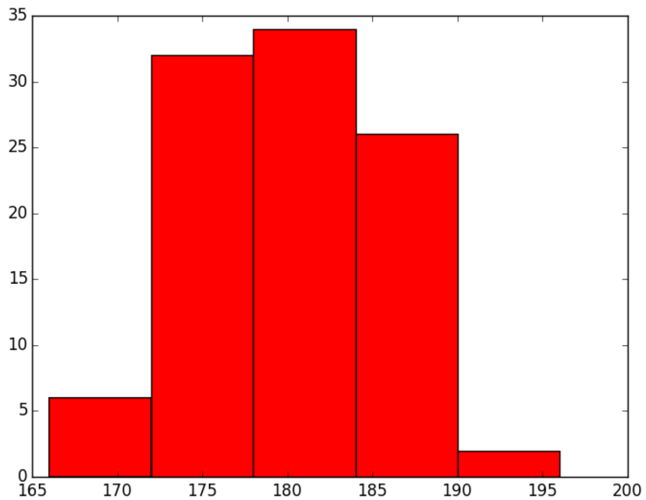
- 3 ways to define "middle", "average"
  - Median: equally many above and below, in the example: 179
    - Formally, if the objects are ordered $x_1, x_2, \ldots, x_n$, then the median is $x_{(n/2)}$ if n is even and $(x_{(n-1)/2} + x_{(n+1)/2})/2$ if it is odd.
  - Mean: ex: 179.54
    - $\bar{x} = (x_1 + x_2 + \cdots + x_n)/n = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Mode, the most frequent one, ex: 176

Observe:
Mean and median may be different, e.g.
- Sentence length
- Income

Histogram for numeric dates



Ex 1: 10 bins

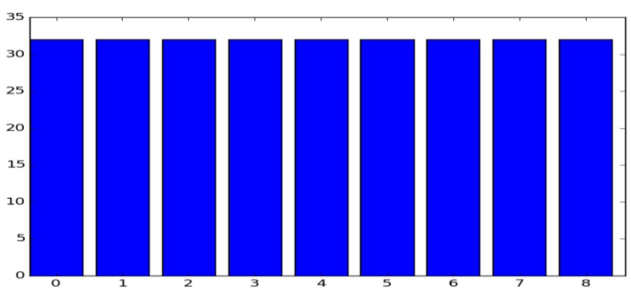


Ex 1: 5 bins

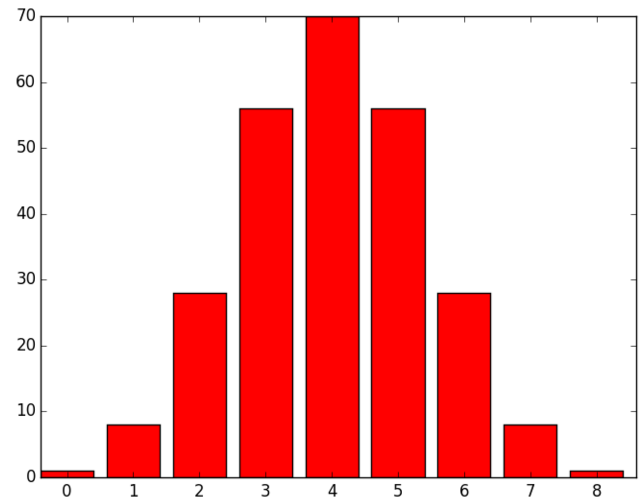
- Split the set of values into n equally sized intervals
- For each interval, ask how many individuals take a value in the interval
- Over the interval, draw a rectangle with height proportional to this frequency
- The y-axis may be tagged with (the shape remains the same)
  - Absolute frequencies
  - Relative frequencies, or
  - Densities (= absolute frequencies/elements in the interval)

Dispersion

- Median or mean does not say everything
- Nor does max, mean or range (=max-min)
- Example:
  - Two sets
  - The same median=mean=4, min:0, max:8



Ex 2: Uniform



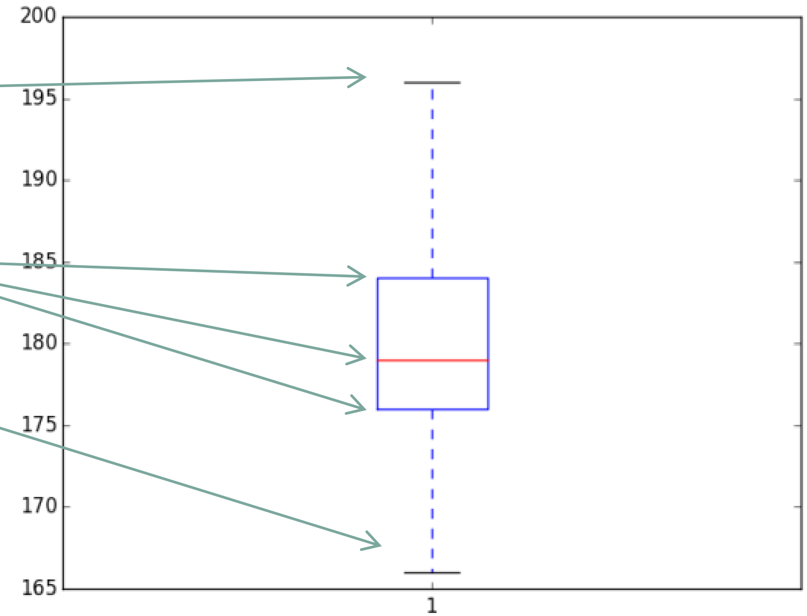
Ex 3: Binomial

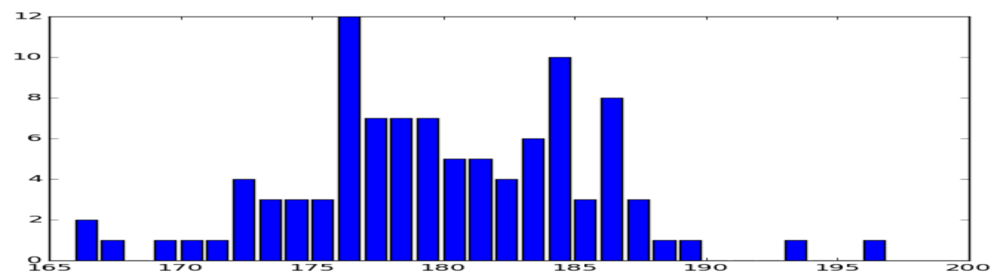
Median, quartile, percentile

- The *n*-percentile *p*:
  - *n* percent of the objects are below *p*
  - (100–*n*) percent are above *p*
  - (where $0<n<100$)
- Median is the 50-percentile
- Quartiles are the 25-, 50-, 75-percentiles
  - Split the objects into 4 equally big bins
  - Example 1: 176, 179, 184
  - Example 2: 2, 4, 6; Example 3: 3, 4, 5

Boxplot

- Example 1:
  - Max 196
  - Quartiles:
  - 176, 179, 184
  - Min 166
- Also good for continuous data
- (The exact definition may var when "outlayers")



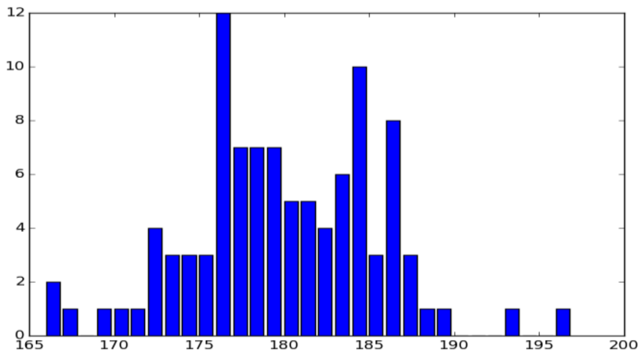


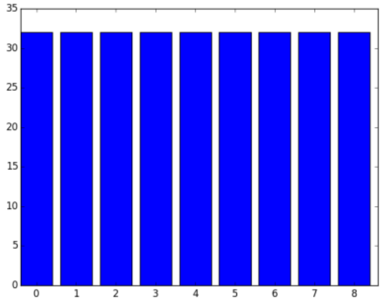
Variance

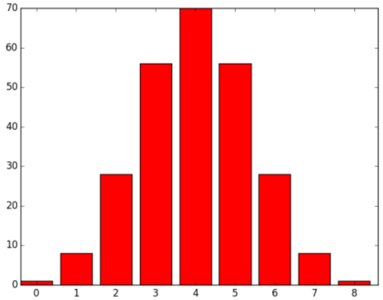
- Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Variance: $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Idea:
  - Measure how far each point is from the mean
  - Take the average
  - Square – otherwise the average would be 0
- Standard deviation: square root of the variance
  - "Correct dimension and magnitude"

Beware:
For some purposes we will later on divide by (n-1) instead of n.
We return to that!

The examples



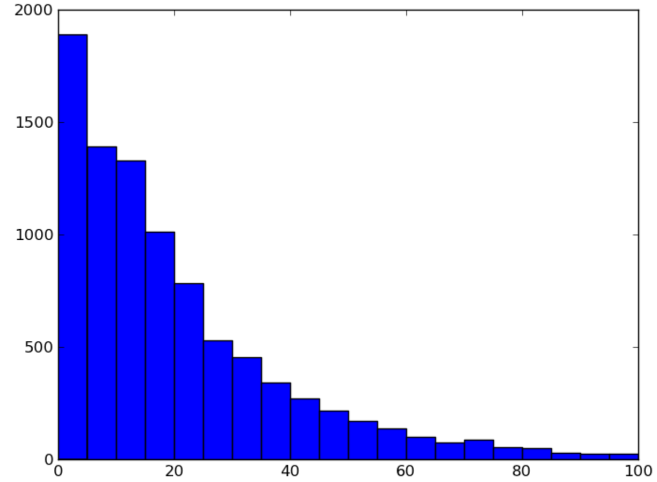


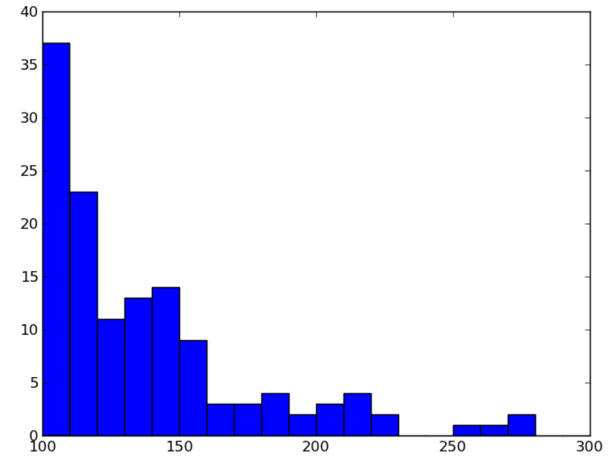


EX	Min	25%	Median	75%	Max	Mean	Vari.	s.d
1	166	176	179	184	196	179.54	30.33	5.5
2	0	2	4	6	8	4	6.67	2.58
3	0	3	4	5	8	4	2.0	1.414

Example: sentence length

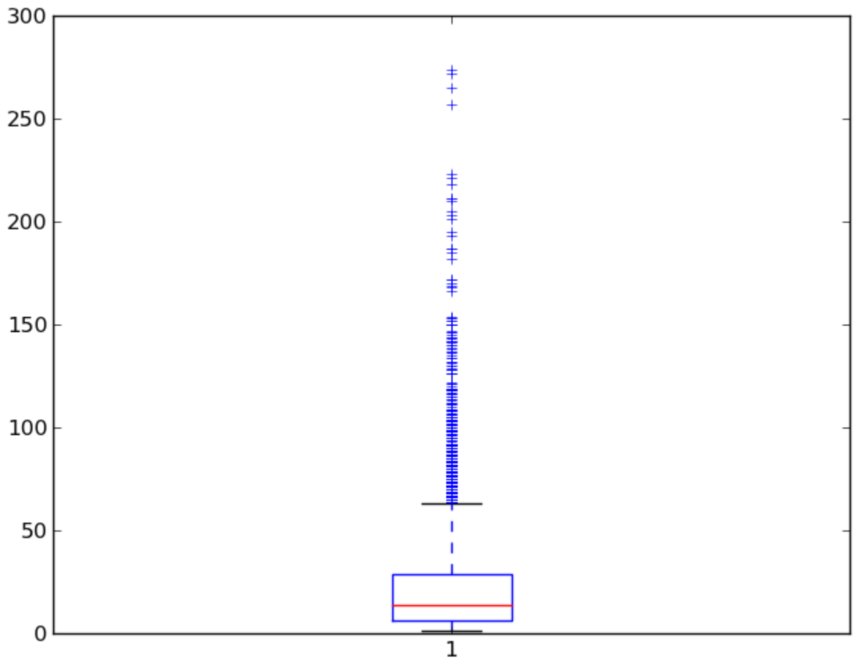
- NLTK: austen-emma.txt
- Number of sentences: 9111
- Length:
  - Min: 1
  - Max: 274
  - Mean: 21.3
  - Median: 14
  - Q1-Q2-Q3: 6-14-29
  - Std.dev.: 23.86





Example: sentence length

- NLTK: austen-emma.txt
- Number of sentences: 9111
- Length:
  - Min: 1
  - Max: 274
  - Mean: 21.3
  - Median: 14
  - Q1-Q2-Q3: 6-14-29
  - Std.dev.: 23.86



Boxplot with outliers

Take home

- Statistical variables:
  - Categoric
  - Numeric
    - Discrete
    - Continuos
- Frequencies
- Median
  - Quartiles, percentiles
- Mean
  - Variance
  - Standard deviation
- Tables
- Bar chart
- Histogram
- Boxplot