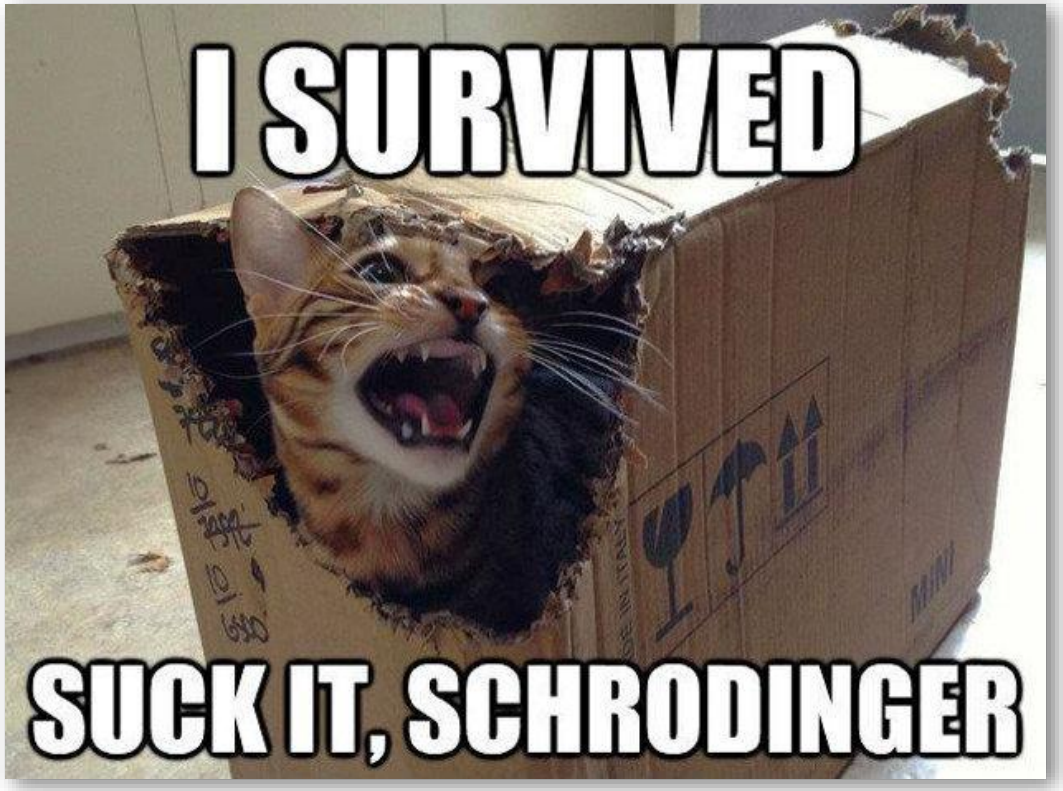
CS 498 Probability & Statistics

Lecture 01



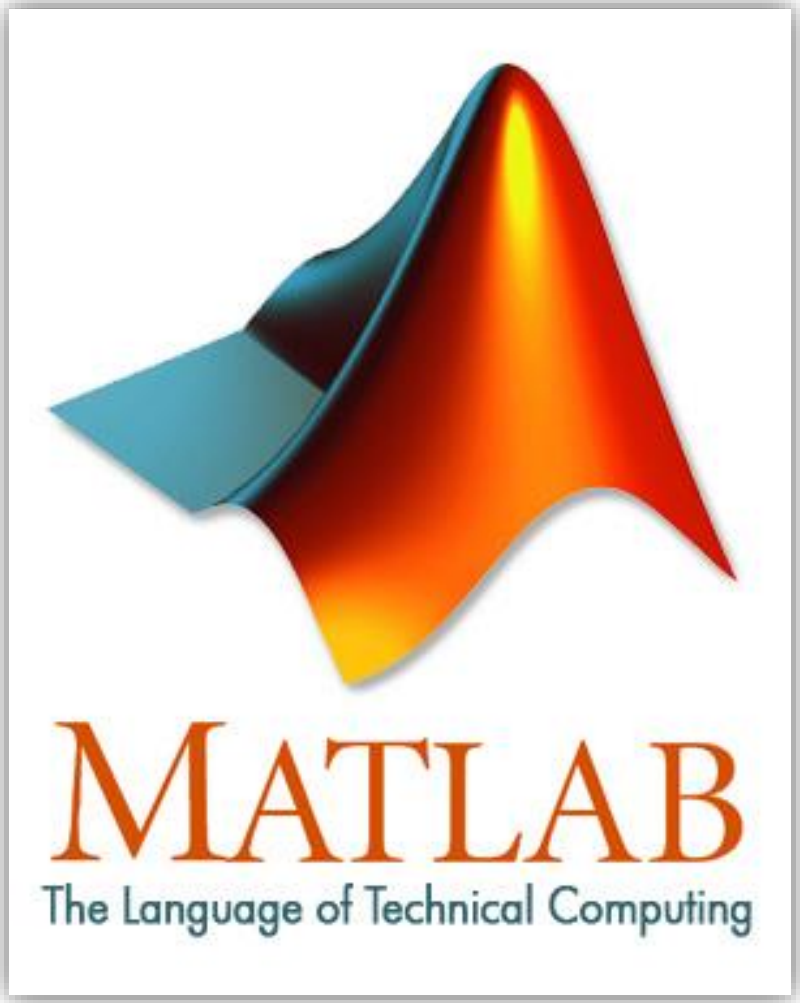
Course logistics

- Class schedule
  - MWF 11:00-11:50 am
  - 1214 Siebel Center.
- Office hours
  - TBD
- Evaluation
  - Homework, midterm, final

- Instructor
  - David Forsyth
  - Email: daf@illinois.edu
  - SC3310 (best way to reach)
- TA: Zicheng Liao
  - Email: liao17@illinois.edu

http://luthuli.cs.uiuc.edu/~daf/courses/Probcourse/Probcourse-2013/498-home.html

Where to start?



About matlab

- "The language of technical computing"
  - The language of **MATRIX**
- Easy interface (C-like), simple syntax, and well-documented.
- Interpreted rather than compiled
- Cross-platform
- Cross-language (matlab <==> C/C++)
- Free student license!

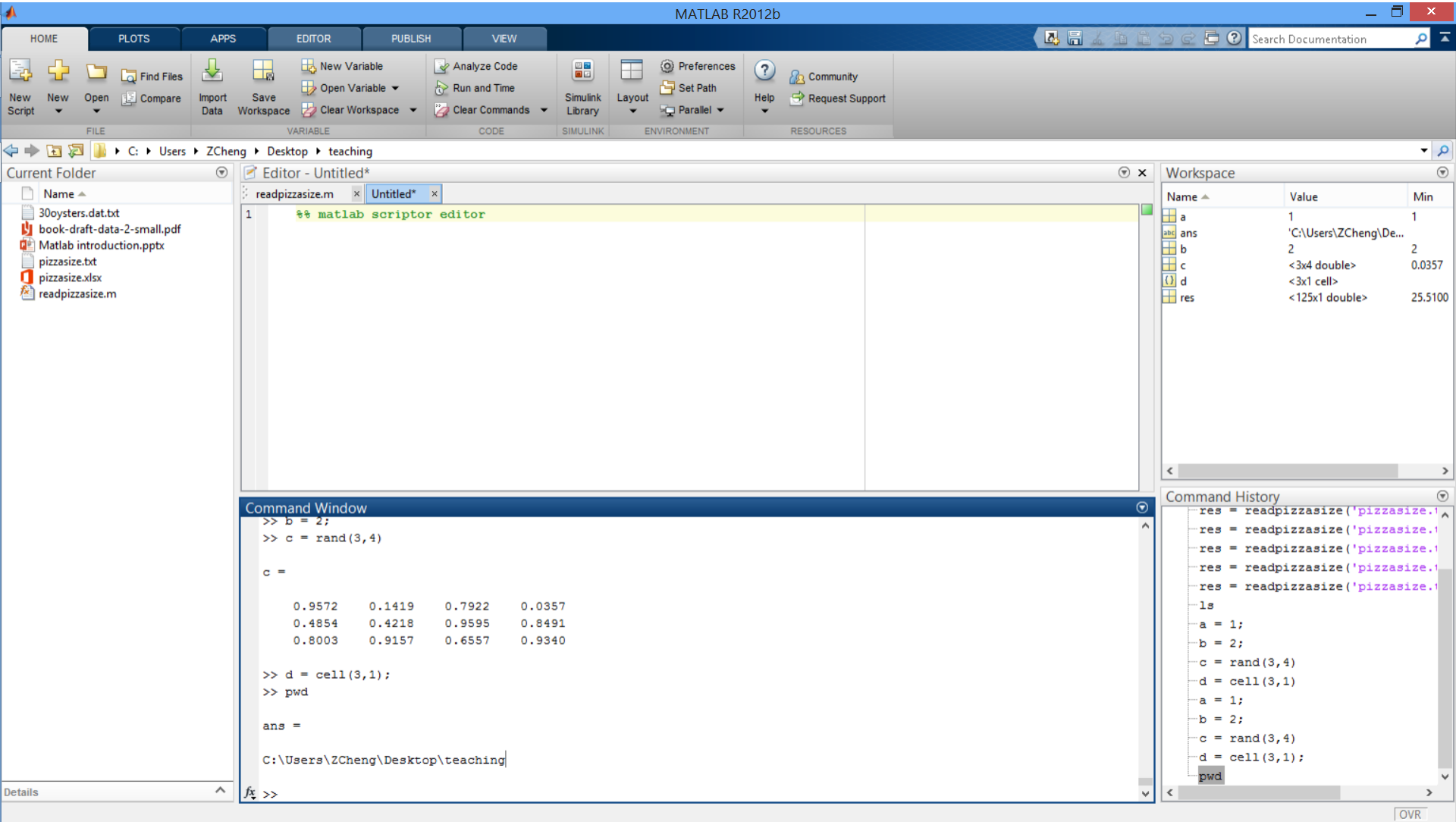
Install matlab

- Half way to success
  - Step 1: go to http://webstore.illinois.edu/home/
  - Step 2: follow instructions.. And you're all set

Install matlab

- Half way to success
  - Step 1: go to http://webstore.illinois.edu/home/
  - Step 2: follow instructions.. And you're most likely to run into sorts of problem
    - Check out a matlab package (I am using R2012b)
    - **License.dat**, **installation** file, and installation **key**.
    - Strictly follow http://dl.webstore.illinois.edu/docs/ii/matlabconc.htm
    - Tricky part: connect to the license manager on server
      - Physically connected to campus network
      - VPN (http://dl.webstore.illinois.edu/docs/ii/vpn.htm)
        - » Follow instructions.. And you're all set.

You've got here!



Create a scalar variable, vector, matrix

```
>> a = 1; b = 2;        %create variable a=1, b=2

>> c = [1 0 1]          %create a row vector
c =
    1    0    1

>> c = [1, 0, 1]        %comma is equivalent to ' '
c =
    1    0    1

>> c = [1; 0; 1]        %create a column vector with semicolons
c =
    1
    0
    1

>> c = [1; 0 1]         %rows must match in dimension
Error using 'vertcat'
CAT arguments dimensions are not consistent.
```

Create a scalar variable, vector, matrix

```
>> d = [1 -2 0; 0 1 2]                %create a 2x3 matrix
d =
    1   -2    0
    0    1    2

>> e = zeros(3,3)                     %create a 3x3  zero matrix
e =
    0    0    0
    0    0    0
    0    0    0

>> f = ones(3,3)                      %create a 3x3  matrix with all 1
f =
    1    1    1
    1    1    1
    1    1    1

>> g = rand(2)      %create a 2x2  matrix with random values
g =
   0.6557   0.8491
   0.0357   0.9340
```

Indexing

```
>> a = [1 2 3 4 5 6 7 8 9 10];
>> a = 1:10              %quick way to create a sequence
a =
     1    2    3    4    5    6    7    8    9    10

>> a(3)     %retrieve the 3rd elm, 1-based indexing, C is 0-based
ans =
     3

>> a(end)       %retrieve the last element
ans =
     10

>> a(2:6)       %retrieve a sub-sequence
ans =
     2    3    4    5    6

>> a(:)'        %colon retrieves the whole vector
ans =
     1    2    3    4    5    6    7    8    9    10
```

Indexing

```
>> a = rand(3,3)
a =  <c1>       <c2>       <c3>
<r1>0.6555   0.0318   0.0971
<r2>0.1712   0.2769   0.8235
<r3>0.7060   0.0462   0.6948

>> a(2,3)              %retrieve element at row 2 column 3
ans =
      0.8235

>> a(8)                %column-major indexing; C is row-major
ans =
      0.8235

>> a(1,:)              %retrieve the whole first row
ans =
      0.6555   0.0318   0.0971

>> a(2:3,1:2)          %retrieve a sub-matrix
ans =
      0.1712   0.2769
      0.7060   0.0462
```

Basic operators: + - * /

```
>> a + b    % a = 1, b = 2
ans =
      3
>> c - a    % vector - scalar
ans =
      0   -1    0
>> a - c    % scalar – vector, a = 1, c = [1   0   1]
ans =
      0    1    0

>> c * b    % =c * b, vector-scalar multiplication; commutative
ans =
      2    0    2

>> c / b            % vector divided by scalar, c = [1 0 1], b = 2
ans =
    0.5000       0    0.5000

>> b / c            % scalar divided by vector
Error using  /
Matrix dimensions must agree.
```

Basic operators: + - * / .* ./

```
>> c + d              % vector plus vector c = [1  0  1], d = [2  2  -1]
ans =
      3     2     0

>> c + 1:5            % a 1x3 vector plus a 1x5 vector
Error using  +
Matrix dimensions must agree.

>> e = [2; 2; -1]  % a 3x1 column vector
e =
      2
      2
     -1
>> c + e              % a row vector plus a column vector
Error using  +
Matrix dimensions must agree.
>> e'                 % transpose of e
ans =
      2     2    -1
>> c + e'
ans =
      3     2     0
```

Basic operators: + - * / .* ./

```
>> c*d              %c = [1  0  1], d = [2  2  -1]
Error using  *
Inner matrix dimensions must agree.

>>c*d'             %dot product
ans =
      1

>> c.*d            %element-wise operation,  [1*2  0*2  1*(-1)]
ans =
      2    0    -1

>> c./d
ans =
    0.5    0   -1.0

>> e = 1:5;
>> c./e
Error using  ./
Matrix dimensions must agree.
```

Basic operators: + - * / .* ./

```
%---implement dot product of two vectors---%
>> c = [1 0 1]; d = [2   2  - 1];

%C-style impl: always try to avoid for loops if possible
>> ans = 0;
>> for i = 1:length(c)
        ans = ans + c(i)*d(i);
   end
>> disp(ans);
       1

%matlab way of doing it
>> c*d';                 %already shown
>> dot(c, d);            %matlab built-in function
>> sum(c.*d)             %your way of doing it explicitly
ans =
       1
```

Basic operators: ^ .^

```
>> 5^2                  % 5 to the power of 2
ans =
      25

>> d^2                  % d = [2   2  - 1]
Error using  ^
Inputs must be a scalar and a square matrix.

>> d.^2
ans =
      4     4     1

>> 2.^d                 % scalar .^ vector
ans =
      4.0000    4.0000    0.5000

>> d.^c                 % vector .^ vector, c = [1 0 1]
ans =
      2     1    -1
```

Logical subscripting

- Logical operators: &(and), |(or), xor, a>b, etc

```
>> if (2 > 3) || (1&1)
        disp('true');
   else
        disp('false');
   end
true

>> a = 1:4;              %a=[1  2  3  4]

>> res = a>2
res =
   0    0    1    1     %logical type
```

Logical subscripting

- Logical operators: &(and), |(or), xor, a>b, etc

```
>> a = 1:4;                %a=[1  2  3  4]

>> res = a>2
res =
    0    0    1    1      %logical type

%--- continue from here---%
>> a(a>2)                  %=a(logical([0 0 1 1])), not a([0 0 1 1])
ans =
    3    4

%%
>> a = randn(1, 10000);  % 10000 samples from normal distribution
>> sum(a<1 & a>-1)/10000     %guess an answer..
ans =
       0.6732                    %1-sigma of normal distribution
```

Concatenate

```
>> a = [1 2];

>> a = [a 3]                    %concatenate a scalar
a =
   1    2    3

>> a = [a [3 2 1]]              %concatenate a with a vector
a =
   1    2    3    3    2    1

>> b = 1:6;  a = [a; b]         %concatenate in the vertical dim
a =
   1    2    3    3    2    1
   1    2    3    4    5    6

>> a = [a; 1:7]                 %dimension must match
Error using vertcat
CAT arguments dimensions are not consistent.
```

Delete

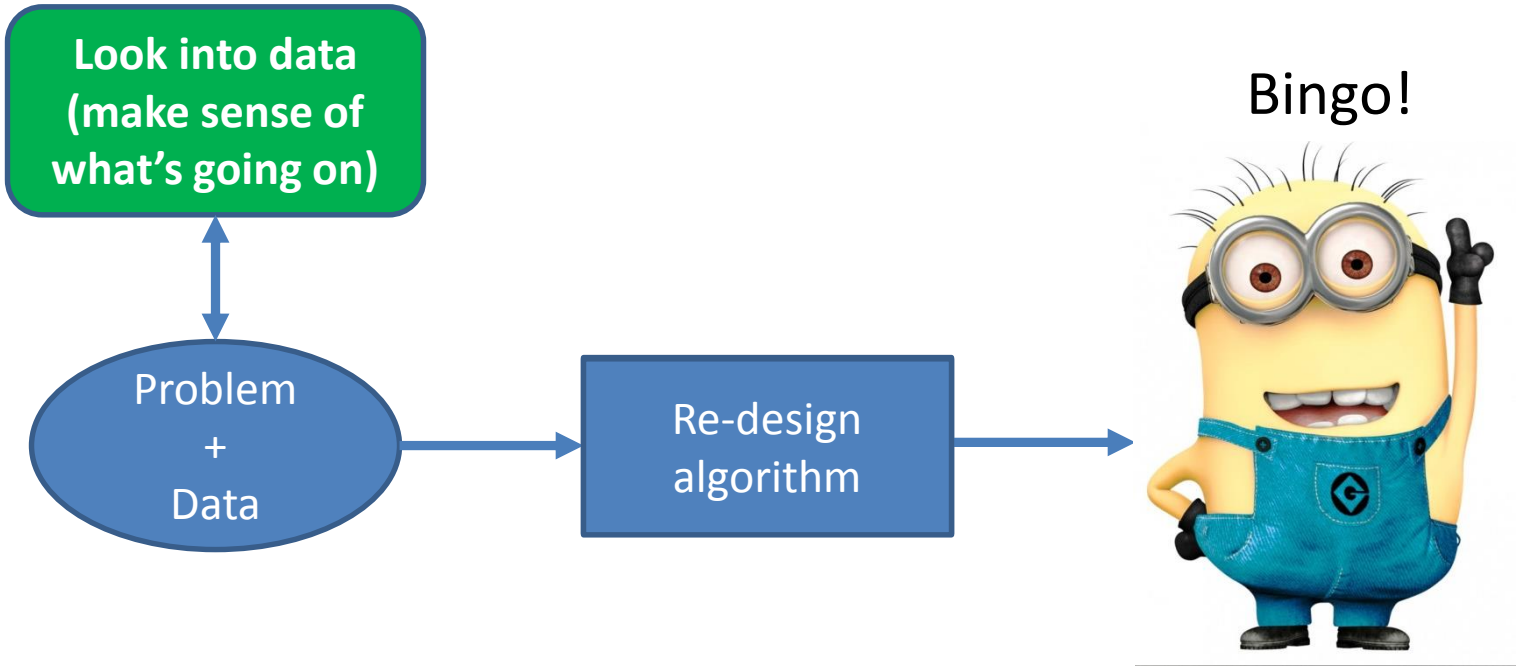
```
>> a
a =
   1   2   3   3   2   1
   1   2   3   4   5   6

>> a(1,:) = []                    %delete the first row
a =
   1   2   3   4   5   6          %matrix size changed

>> a(2:3) = []                    %delete two elements in a vector
a =
   1   4   5   6

>> a(2) = []                      %delete one more element
a =
   1   5   6
```

Online resources

- A quick tutorial
  - http://web.eecs.umich.edu/~aey/eecs451/matlab.pdf
- Get started with matlab
  - http://www.mathworks.com/help/pdf_doc/matlab/getstart.pdf
- Matlab online document **(everything is here!)**
  - http://www.mathworks.com/help/matlab/
  - >> doc func_name
  - >> doc; search with key words

First tools for looking at Data

- It’s all about data
- “what’s going on here?”
- Descriptive statistics



Datasets

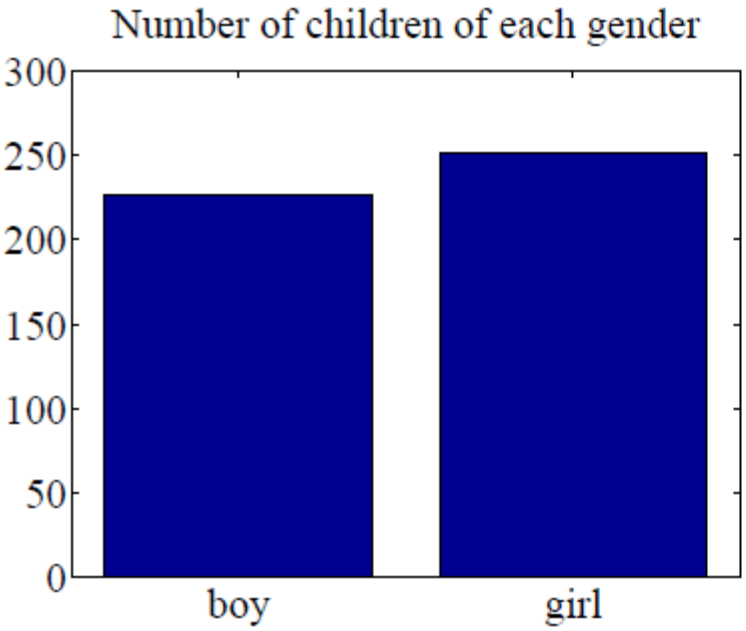
- School dataset

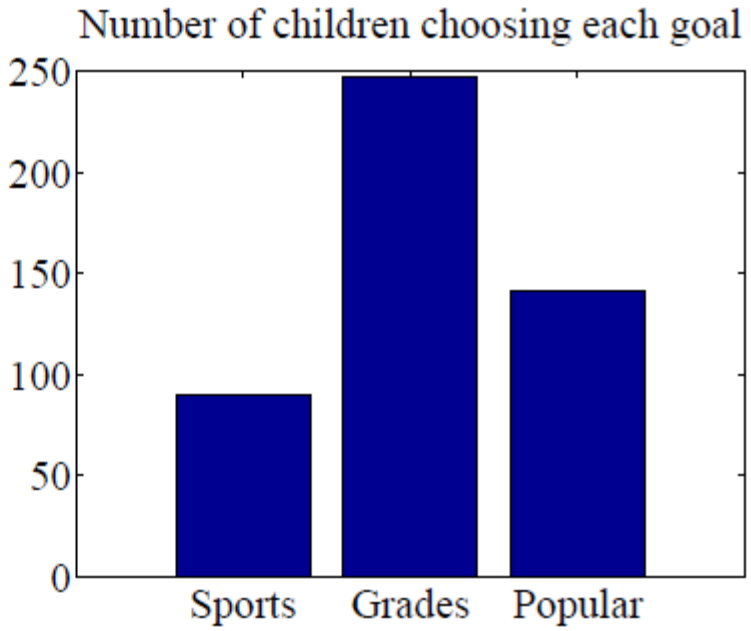
Gender	Goal	Gender	Goal
boy	Sports	girl	Sports
boy	Popular	girl	Grades
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	girl	Grades
girl	Popular	girl	Sports
girl	Grades	girl	Popular
girl	Sports	girl	Grades
girl	Sports	girl	Sports

http://lib.stat.cmu.edu/DASL/Datafiles/PopularKids.html

Bar charts

- Count of categorical data





matlab\plotschooldata.m

(Walk through the whole process)

Datasets

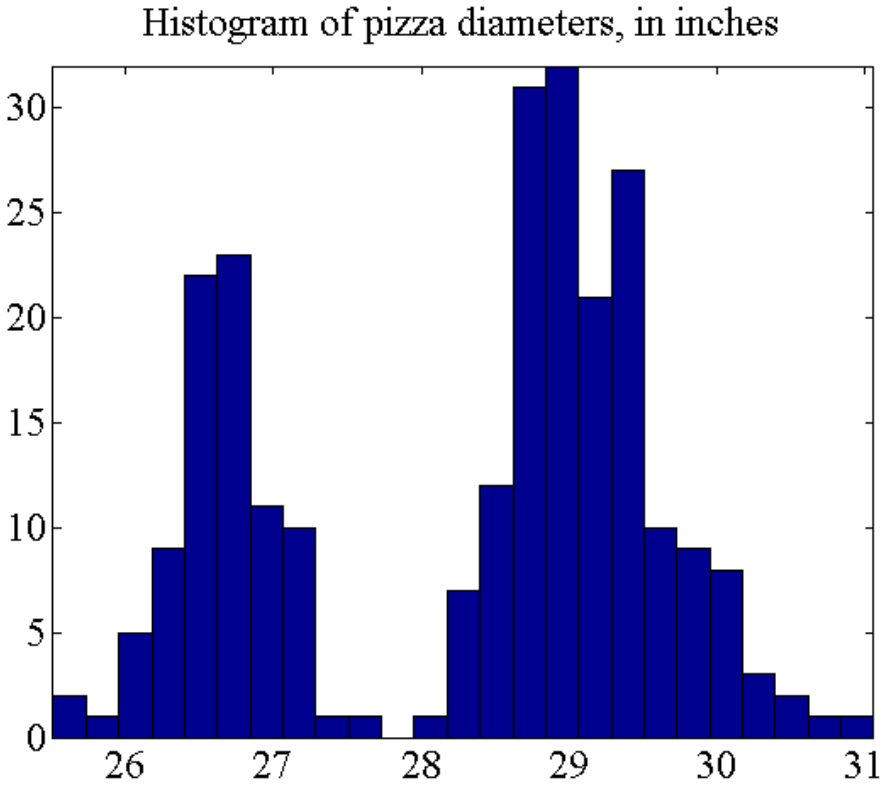
- Pizza size dataset

Store	Diameter
Dominos	29.4
Dominos	29.63
Dominos	27.06
EagleBoys	30.05
EagleBoys	29.47
EagleBoys	30
EagleBoys	30.01
EagleBoys	28.66
EagleBoys	30.44
Dominos	26.38
Dominos	28.86
Dominos	26.6

http://www.amstat.org/publications/jse/jse_data_archive.htm

Histogram

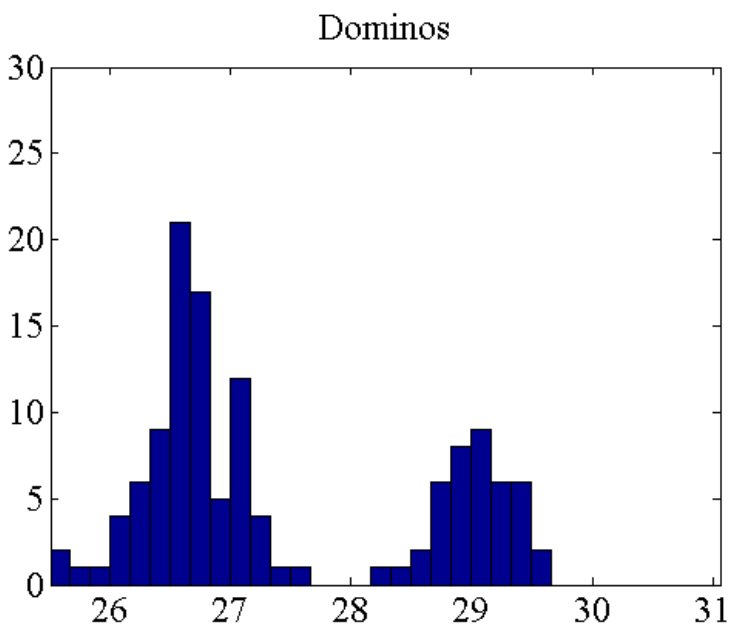
- Count of continuous data in even (or uneven) intervals

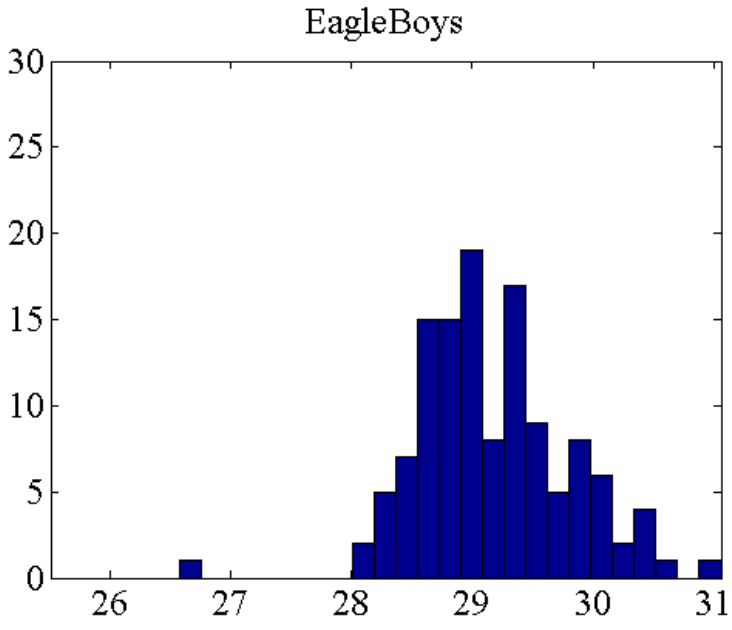


matlab\plotpizzasize.m

Class-conditional histogram

- Histogram of a certain class





matlab\plotpizzasize_condhist.m

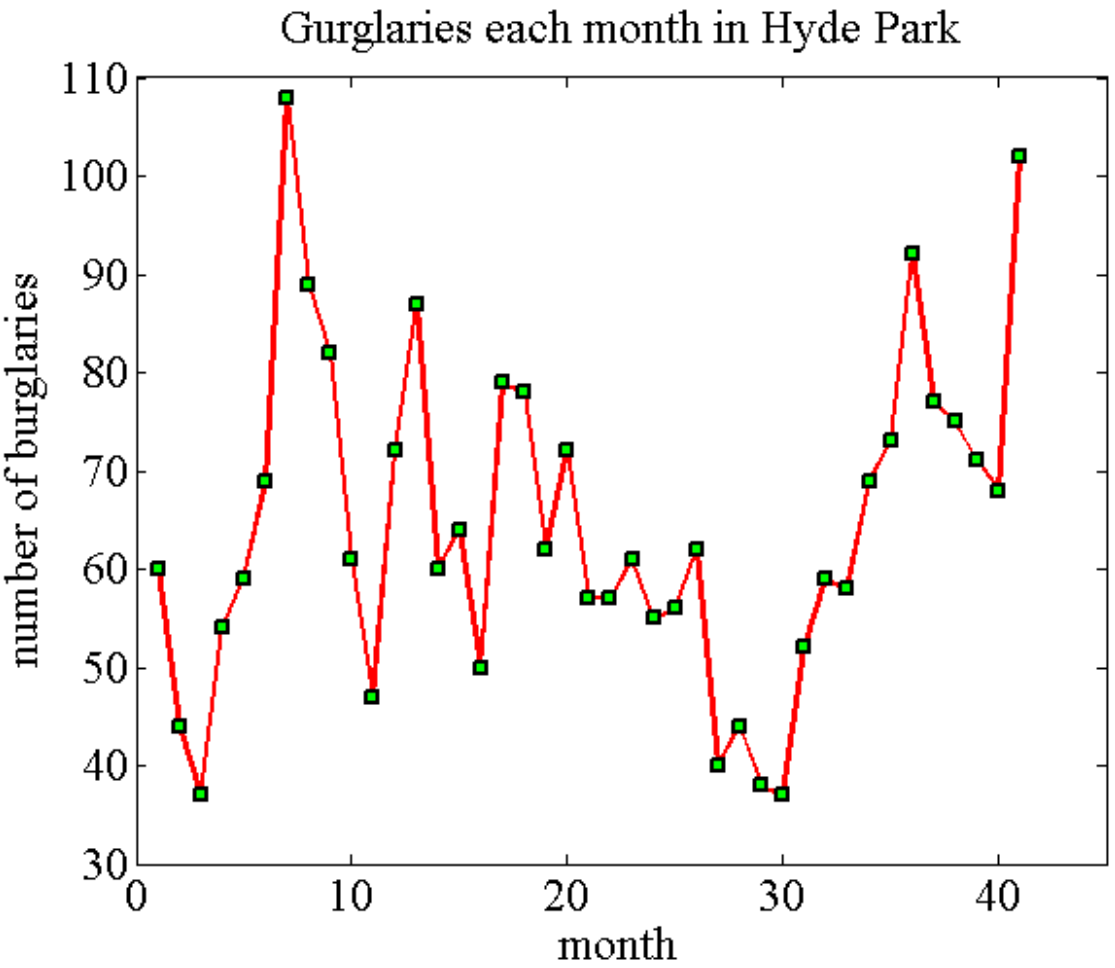
Series data

month	number of burglary
1	60
2	44
3	37
4	54
5	59
6	69
7	108
8	89
9	82
10	61
11	47
12	72
13	87
14	60
15	64
16	50
17	79
18	78
19	62
20	72

Number of burglaries each month in Hyde Park

http://lib.stat.cmu.edu/DASL/Datafiles/timeseriesdat.html

Plot series data



matlab\plotburglary.m

Summarizing 1D data

- Mean
- Standard deviation
- Variance
- Median
- Percentile
- Interquartile range

Index	net worth
1	100, 360
2	109, 770
3	96, 860
4	97, 860
5	108, 930
6	124, 330
7	101, 300
8	112, 710
9	106, 740
10	120, 170

Net worth of people you meet in a bar

Mean

– Mean: $mean(\{x\}) = \frac{1}{N}\sum_{i=1}^{N} x_i$

```
>>  a = [1 2 3 5 6];
>> mean(a)
ma =
       3.4

>> a = [1 2 3; 4 5 6];
>> mean(a)            %by default, take mean per-column
ans =
      2.5000    3.5000    4.5000

>>mean(a, 2)         %take mean in the 2nd dimension (row)
ans =
      2
      5
```

Median

– Median: The data half way along the sorted data points

```
>>  a = [1 2 3 5 6];
>> median(a)
ma =
       3

>> a = [a 6];            %a = [1  2  3  5  6  6]
>> median(a)             %take the mean of the two middle points
ans =
       4

>> median([1 2 2 2 2 2 5 10 15 100])     %biased measure
ans =
       2
```

Std. and variance

– Standard deviation: $std(\{x\}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - mean(\{x\})\right)^2}$

– Va

```
>> a = [1 2 3 5 6];
>> std(a)
ans =
        2.0736          %not exactly by the formula
>> std(a,1)             %based on the above formula
ans =
   1.8547
>> var(a,1)             %variance
ans =
        3.4400
>> std(a,1)^2           %variance = std^2
ans =
        3.4400
>> mean((a-mean(a)).*(a-mean(a)))      %what var(a) does
ans =
        3.4400
```

Percentile and interquartile range

- Percentile: The k-th percentile is the value such that $k\%$ of the data x is less than or equal to.
- Interquartile range: $\mathrm{iqr}(\{x\}) = prctile(x, 75) - prctile(x, 25)$

```
>>  a = rand(10000,1);
>>  prctile(a, 20)      %20th-percentile of 0-1 random samples
ans =
       0.1991           %as expected

>>  prctile(a, 80)      %80th-percentile: ~0.8
ans =
      0.7978

>>  iqr(a)              %interquartile range of a: ~0.5
ans =
      0.4984

>>  prctile(a, 75) - prctile(a, 25)          %sanity check
ans =
      0.4984
```

Summarizing 1D data

```
>> networths = [100360, 109770, 96860,
97860, 108930, 124330, 101300,…
112710,106740, 120170];

>> m = mean(networths)
m =

    107903

>> sd = std(networths)
sd =

  9.2654e+03

>> v = var(networths)
v =

  8.5848e+07
```

Index	net worth
1	100, 360
2	109, 770
3	96, 860
4	97, 860
5	108, 930
6	124, 330
7	101, 300
8	112, 710
9	106, 740
10	120, 170

Net worth of people you meet in a bar

Summarizing 1D data

```
>> bnetworths = [networths, 1e9];

>> bm = mean(bnetworths)

bm =
       9.1007e+07

>> bsd = std(bnetworths)

bsd =
       3.0148e+08

>> bv = var(bnetworths)

bv =
      9.0889e+16
```

Index	net worth
1	100, 360
2	109, 770
3	96, 860
4	97, 860
5	108, 930
6	124, 330
7	101, 300
8	112, 710
9	106, 740
10	120, 170
11	**1e9**

A billionaire comes in

Sensitive to outliers!

Summarizing 1D data

```
>> md = median(networths)
md =

       107835

>> bmd = median(bnetworths)
bmd =

       108930
```

Index	net worth
1	100, 360
2	109, 770
3	96, 860
4	97, 860
5	108, 930
6	124, 330
7	101, 300
8	112, 710
9	106, 740
10	120, 170
11	**1e9**

Networths with a billionaire

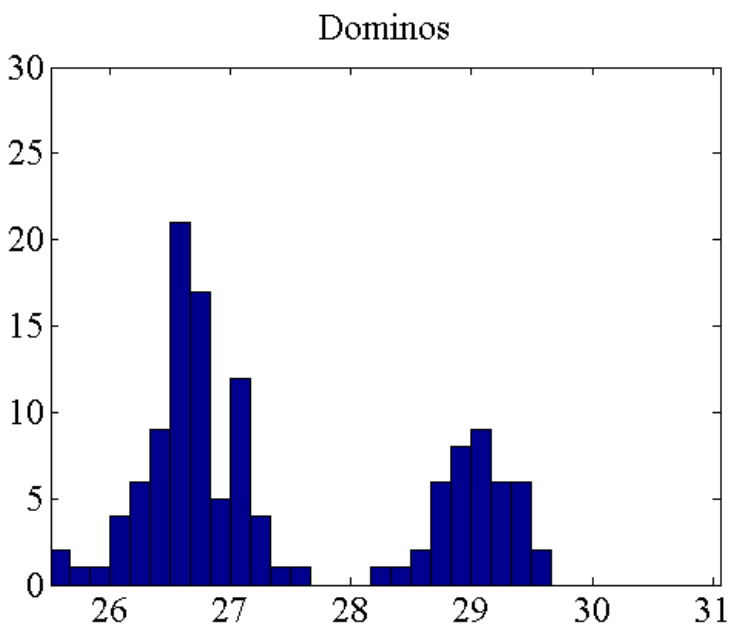
Summarizing 1D data

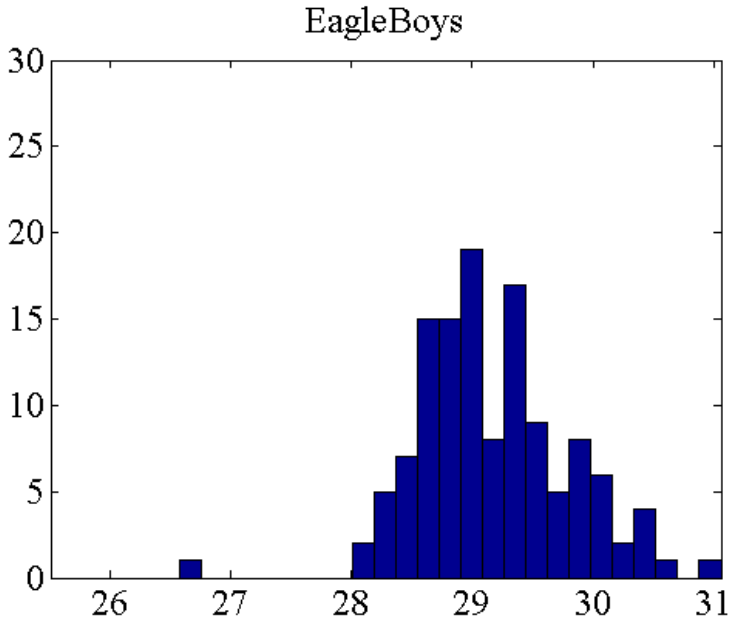
```
>> pcts = prctile(networths, [25 50 75])

pcts =
       100360     107835     112710

>> bpcts = prctile(bnetworths, [25 50 75])

bpcts =
       100595     108930     118305

>> interqtl = iqr(networths)

interqtl =   12350

>> binterqtl = iqr(bnetworths)

binterqtl =  17710
```

Index	net worth
1	100, 360
2	109, 770
3	96, 860
4	97, 860
5	108, 930
6	124, 330
7	101, 300
8	112, 710
9	106, 740
10	120, 170
11	**1e9**

Networths with a billionaire

The pizzasize puzzle

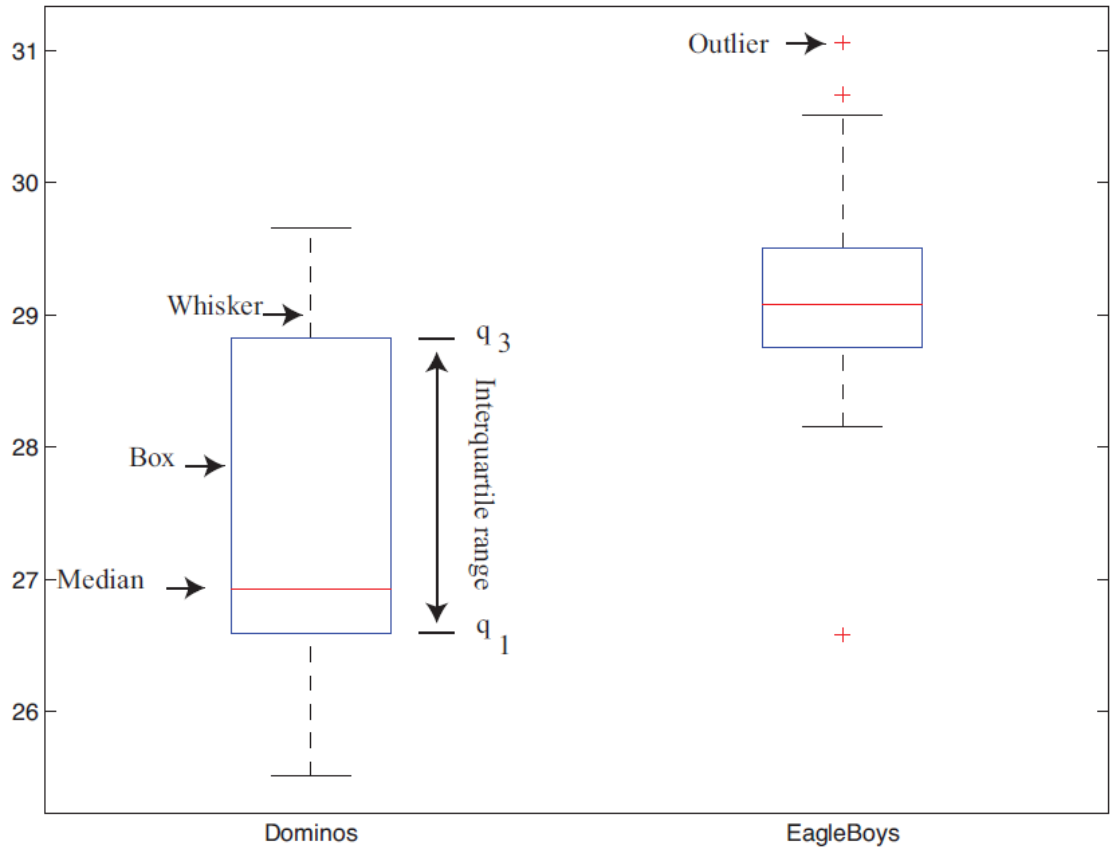




- Understand what's going on
- Look at other labels: type of crust and type of topping
- Cannot compare many histogram together
  - Need a more compact plot

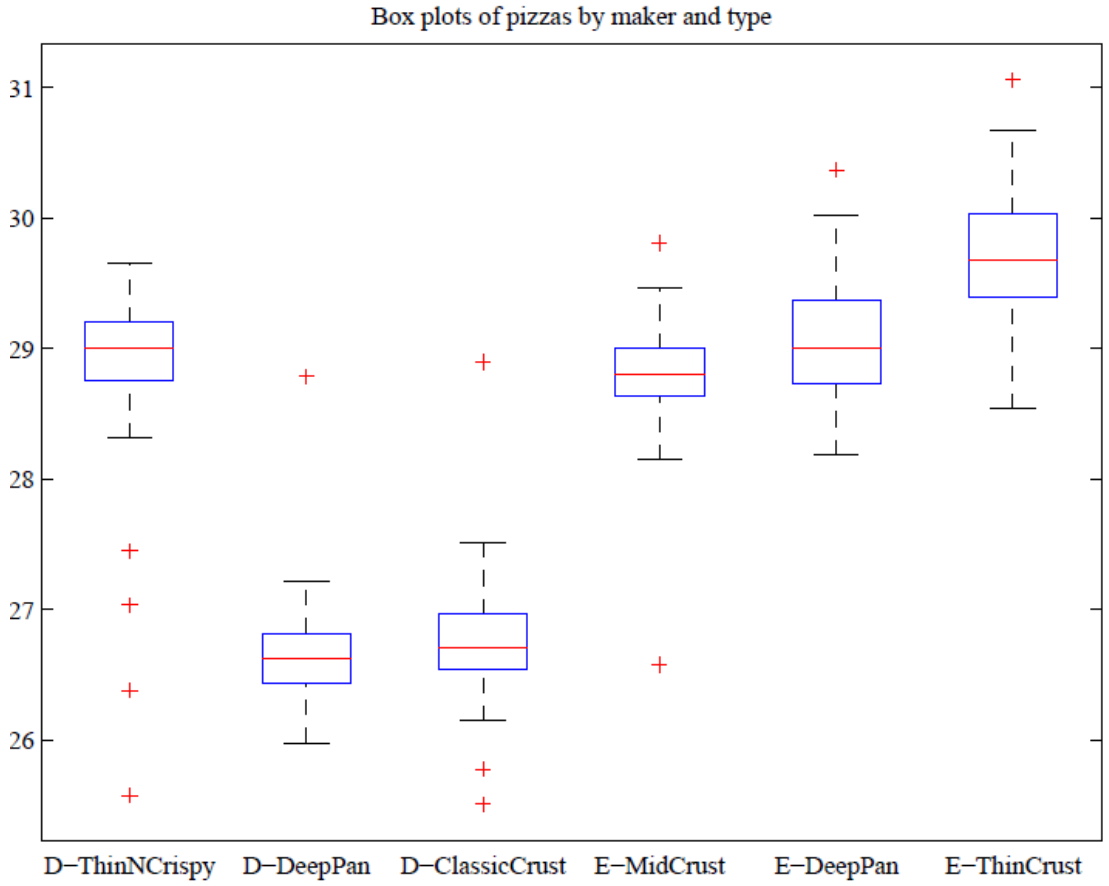
boxplot

- More compact way of summarizing data than histogram



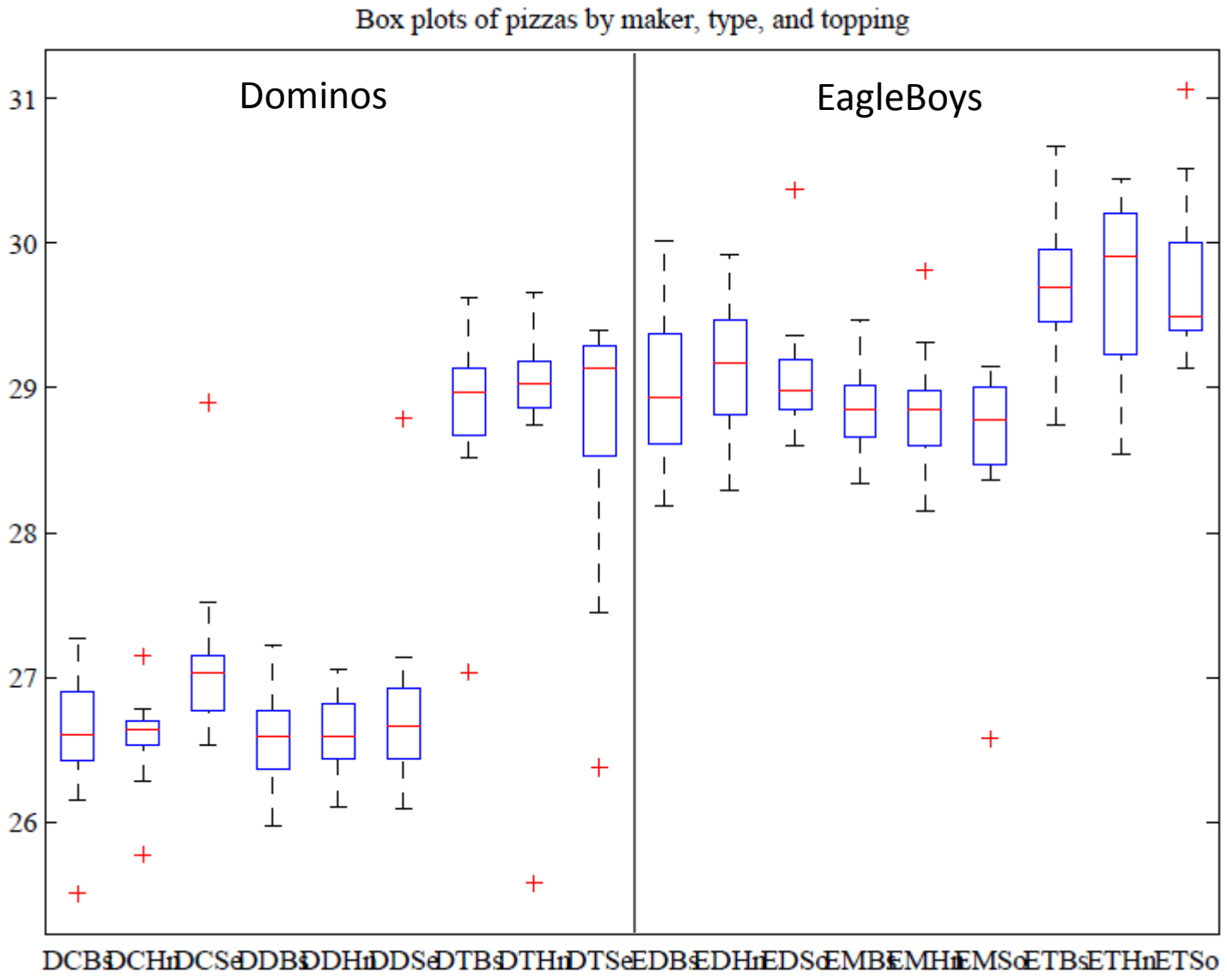

```
>> boxplot([dsizes esizes], 'whisker', 1.5);
```

Boxplot with type of crust



- EagleBoys has tighter control over size
- Dominos ThinNCrispy is unusual
  - shrinking during baking
  - control portion by weight
  - mistakes by chef (?)

Boxplot with crust and topping



Wrap-up

- “A matlab start is half way to success”

- It’s all about data.
  - Plot data with bar chart, histogram, series plot and box plot.

  - Summarize 1D data with mean, std, variance, median, percentile and interquartile range.