Intro and regressions

11

Probability and Statistics

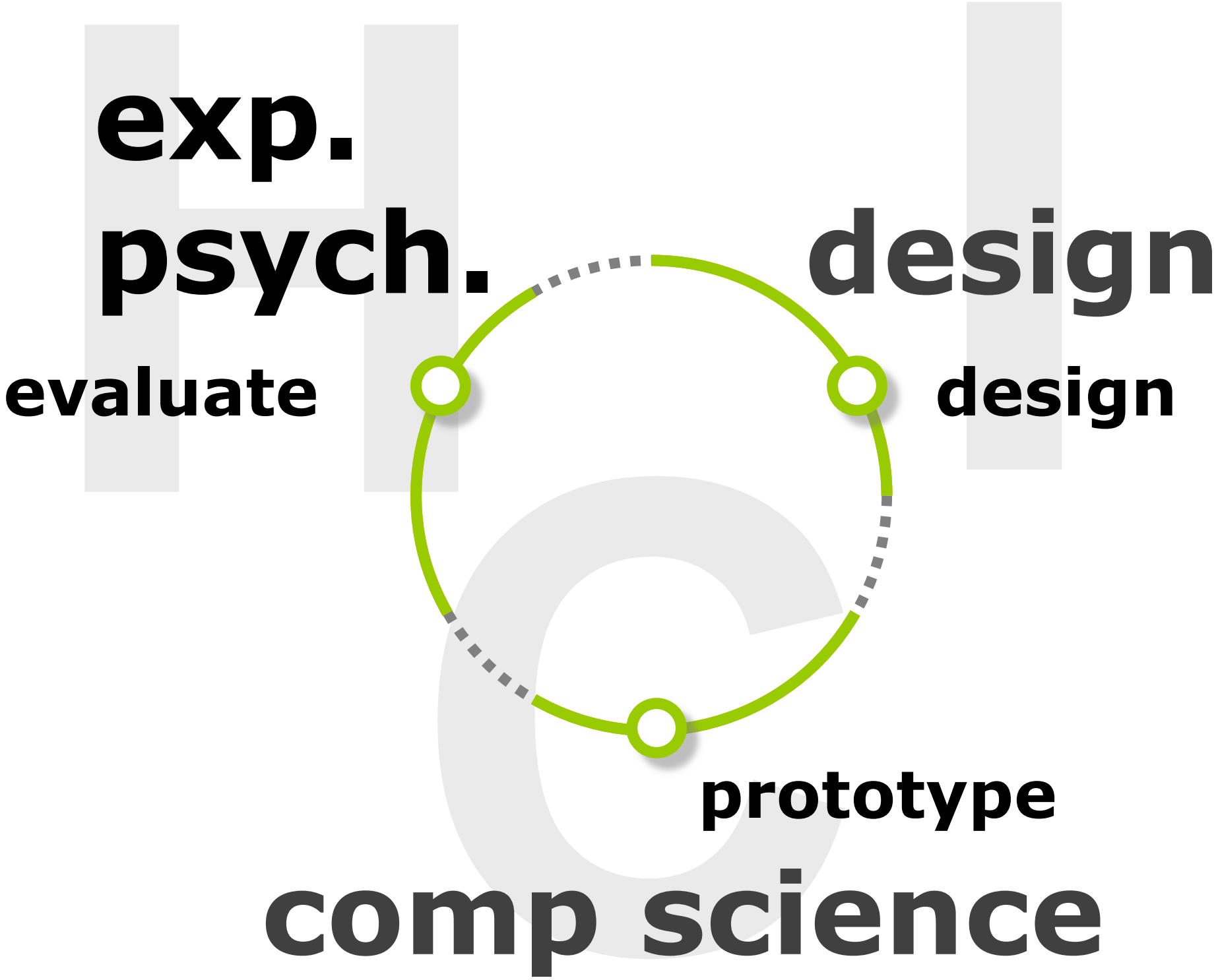
COMS10011

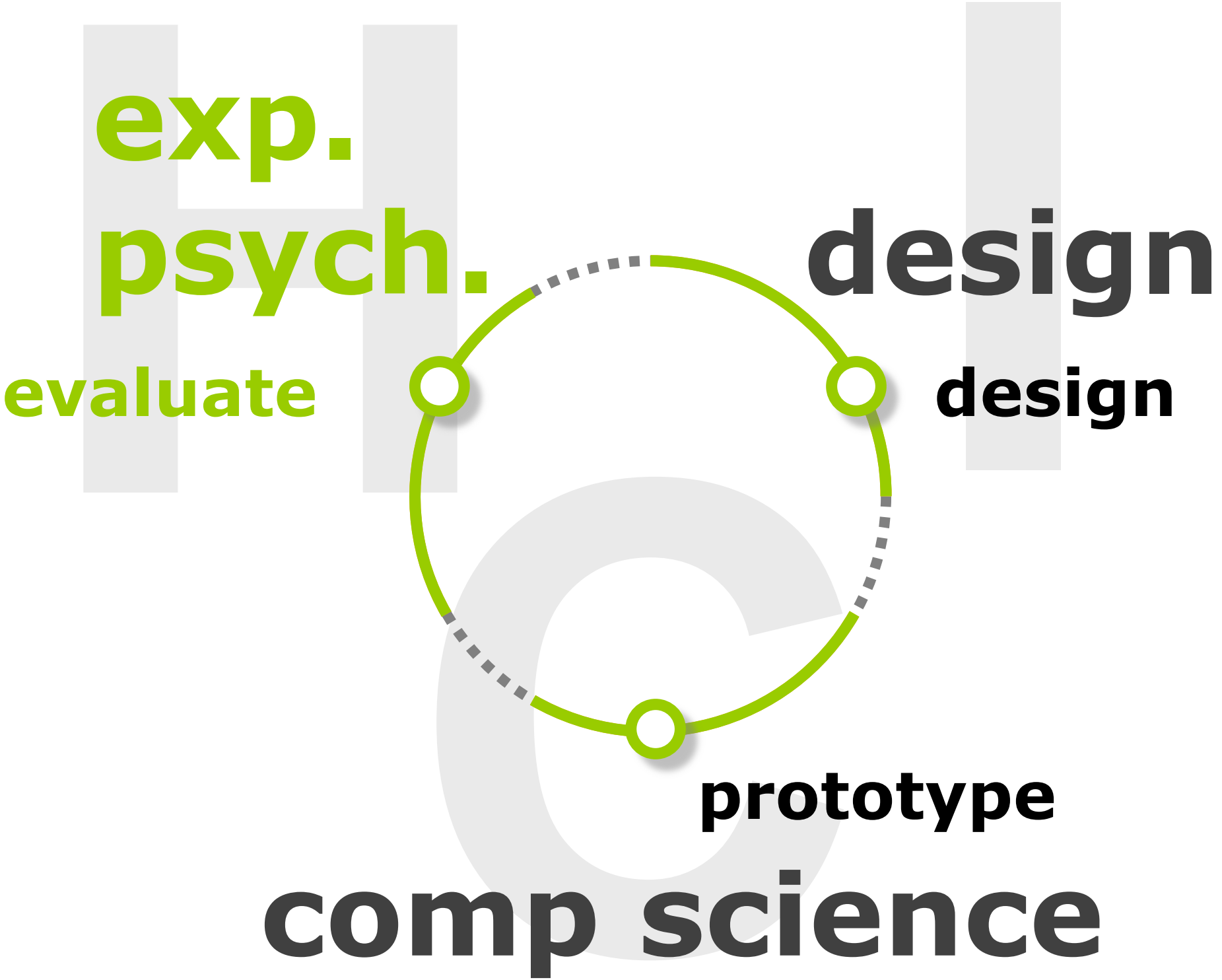
Dr. Anne Roudaut
csxar@bristol.ac.uk

who am i?

Human Computer Interaction (HCI)::

a multidisciplinary field of study focusing on the design of computer technology and, in particular, the interaction between humans (the users) and computers


exp.
psych.
design
evaluate
design
prototype
comp science


exp.
psych.
design
evaluate
design
prototype
comp science

experimental psychology::

the branch of psychology concerned with the scientific investigation of the responses of individuals to stimuli in controlled situations

e.g. bandwagon effect (one of our many cognitive biases) phenomenon whereby the rate of uptake of beliefs, ideas, fads and trends increases the more that they have already been adopted by others


#WYFFT
abc

what is the link with statistic?

well, like in many fields related to computer science, statistics is the main tool to **evaluate, demonstrate, predict or analyse data**

let’s start with an example

imagine you are designing a graphical interface for a new application on a laptop

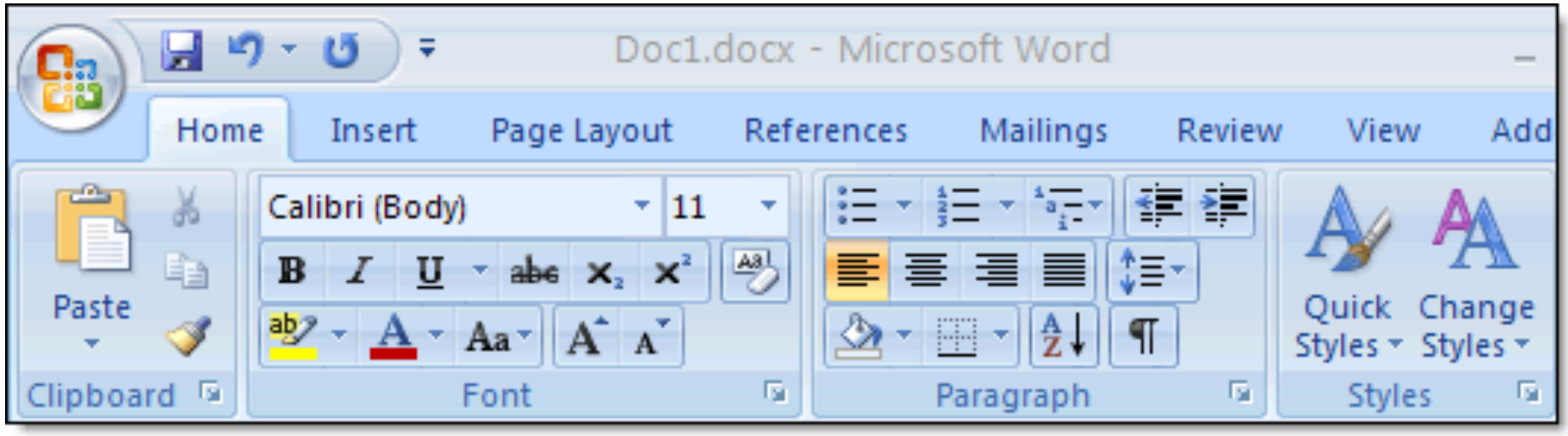
how big should the buttons/icons be?

Fitts' law ::

the time required to **acquire a target** of size *w* at distance *d* can be described as $T = a + b \log (1 + d/w)$

$T = a + b$ IndexDifficulty

(depends on input device)



e.g. one reason why we have ribbons in Word now

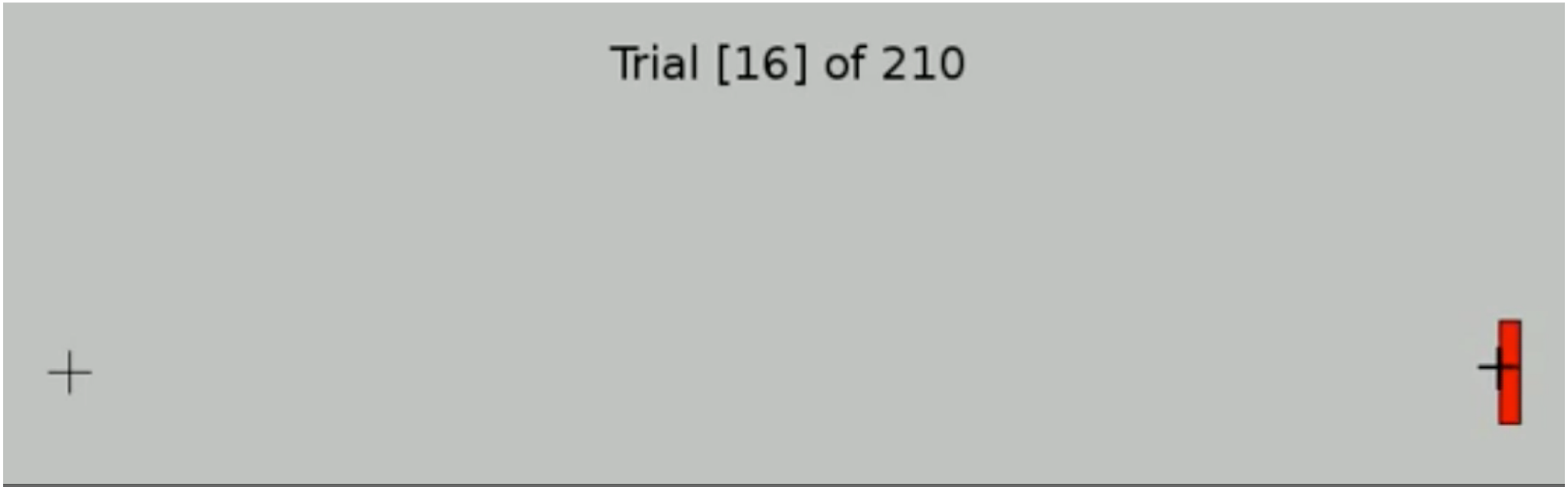

smaller glasses = harder and further = harder

Fitts' law ::

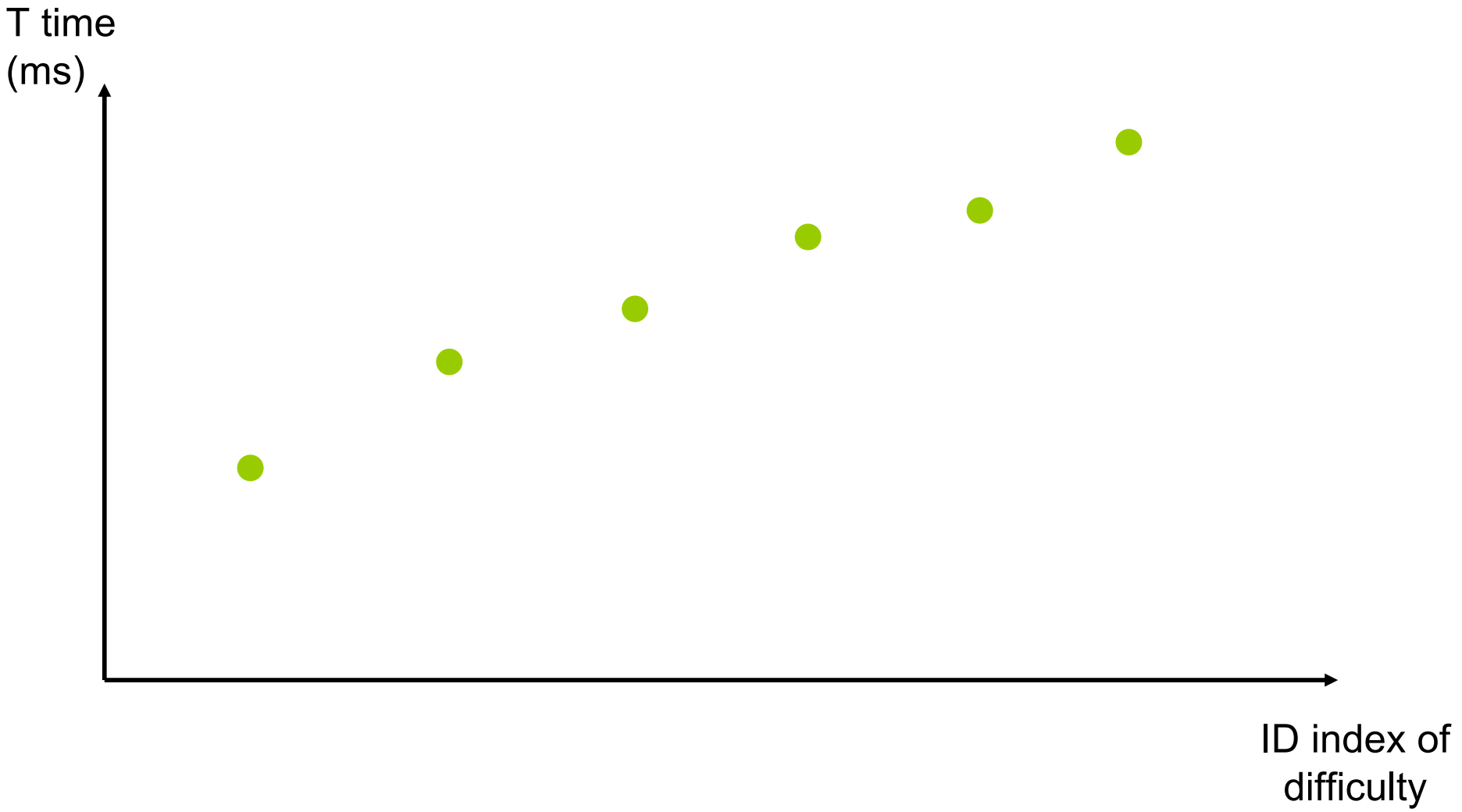
the time required to **acquire a target**

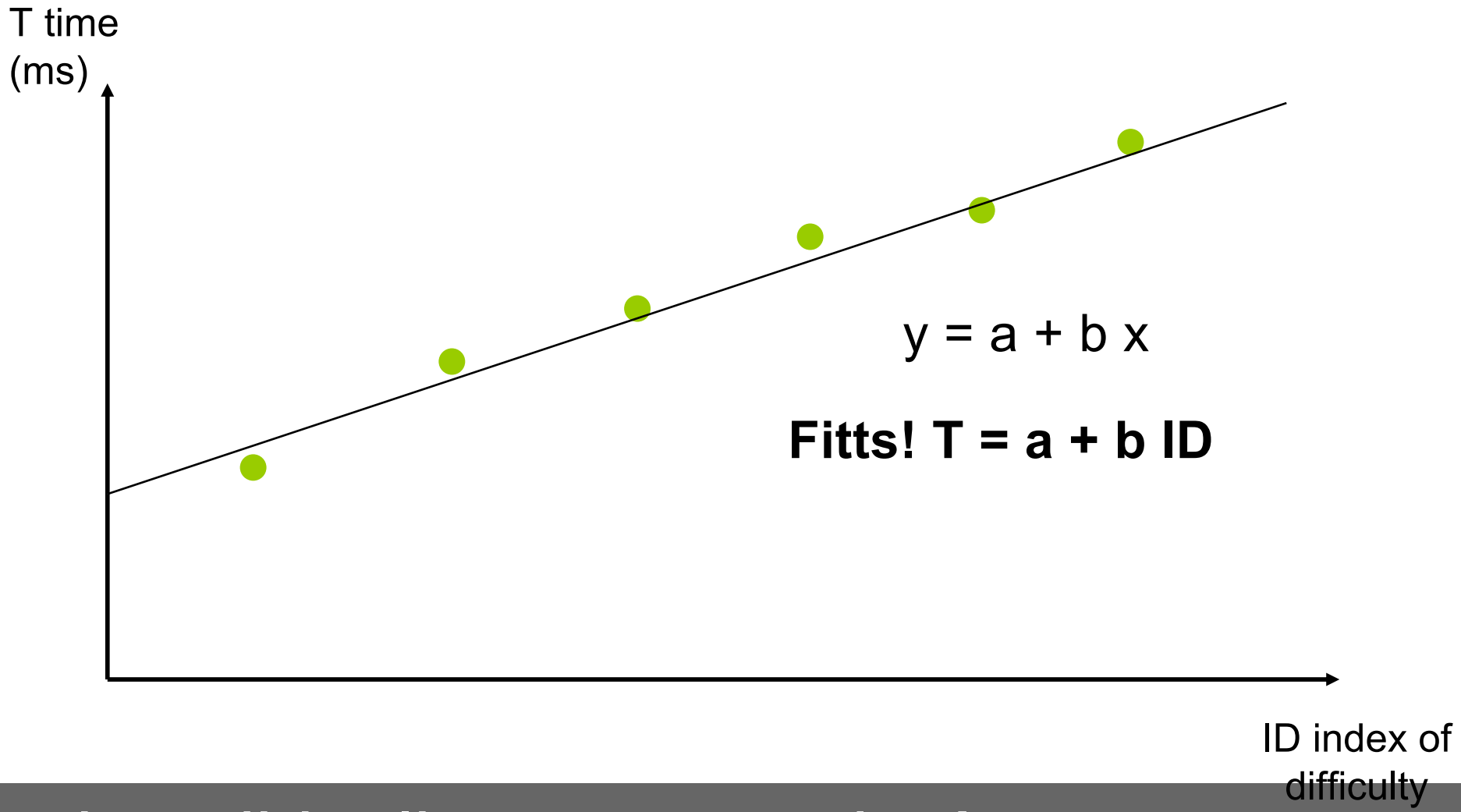
$$T = a + b\,ID$$

but where does this equation come from?



let’s run an experiment and ask one participant to click on targets of IDs


T time
(ms)
ID index of
difficulty



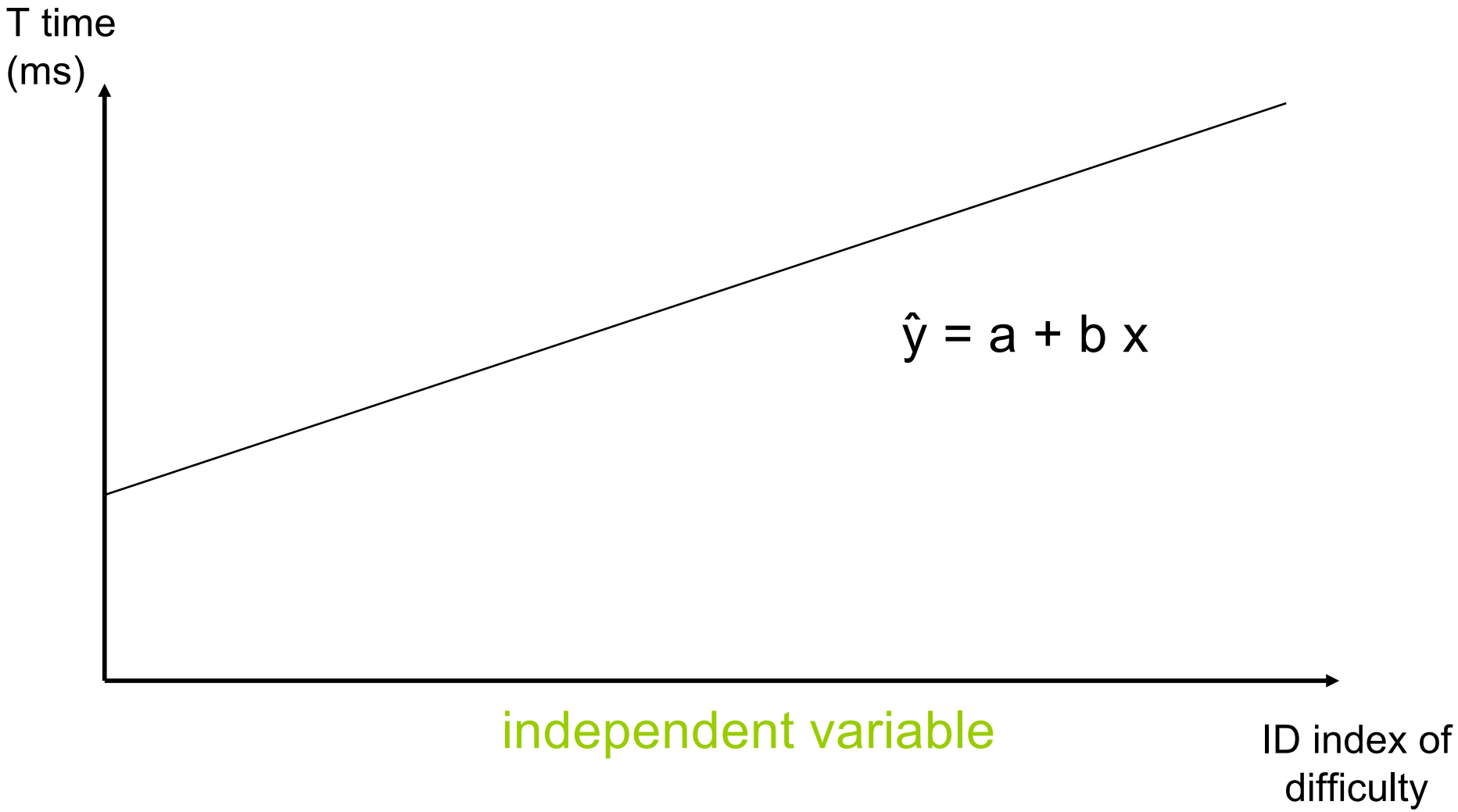
we just did a linear regression!

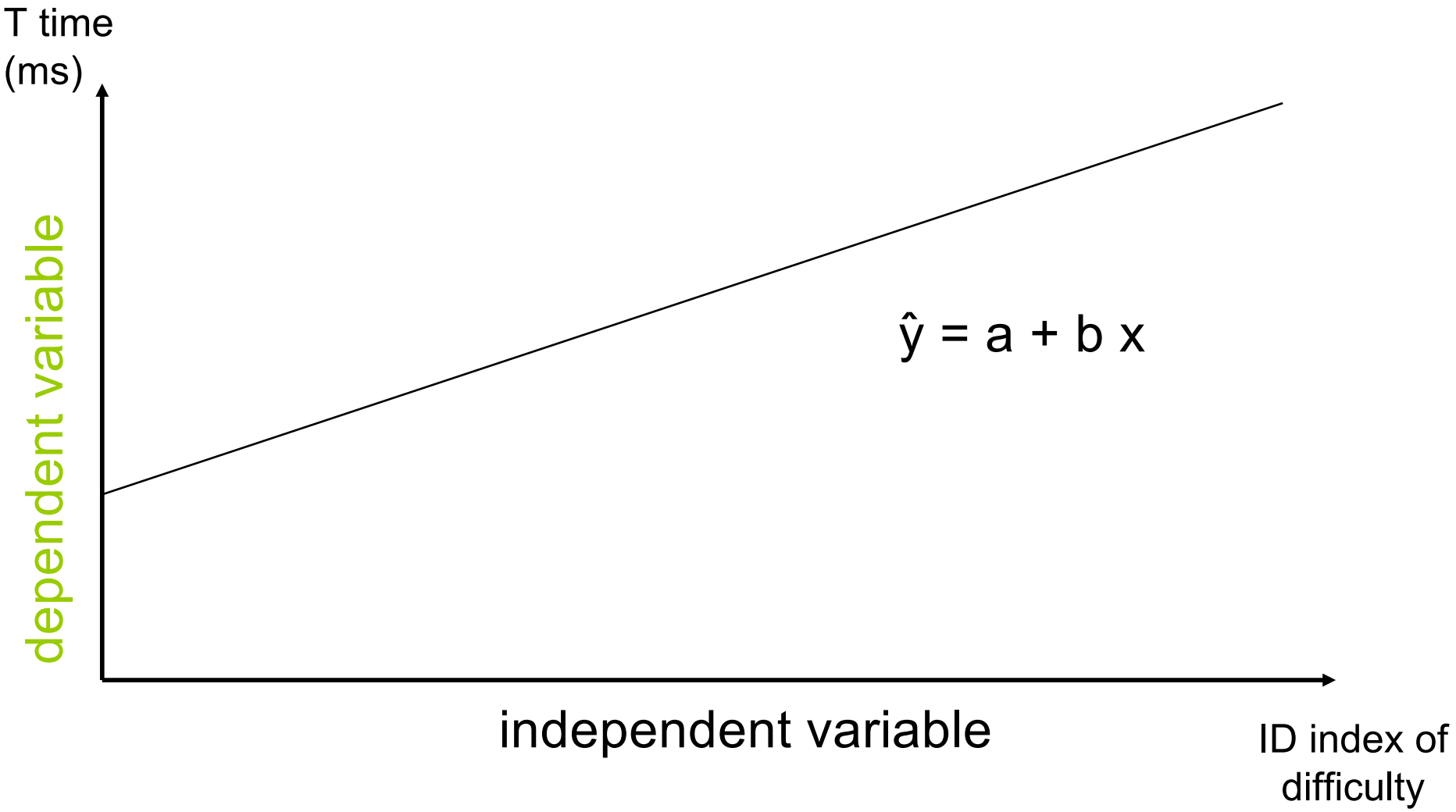
regression ::

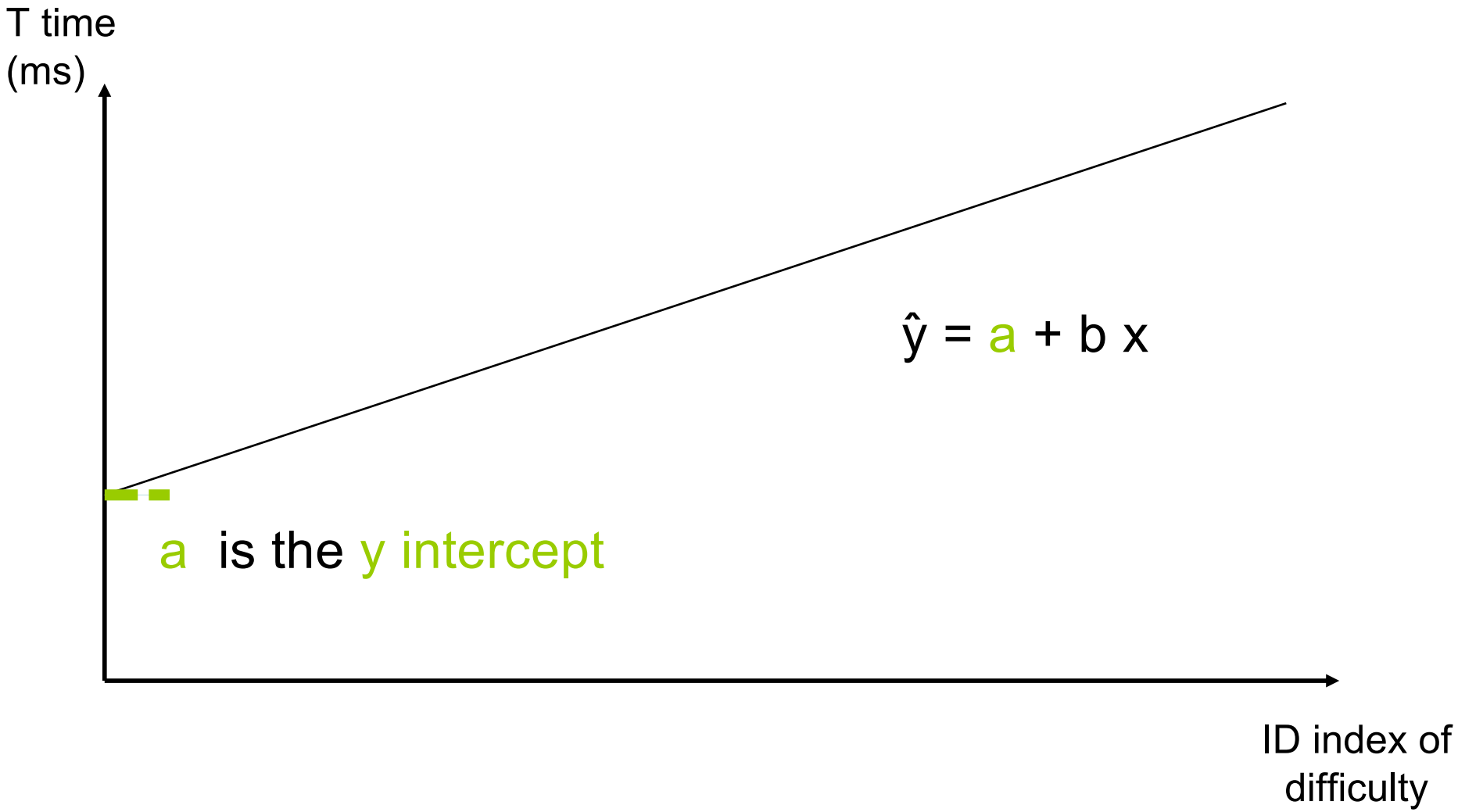
a technique for determining the statistical relationship between two or more variables where a change in a **dependent variable** is associated with, and depends on, a change in one or more **independent variables**

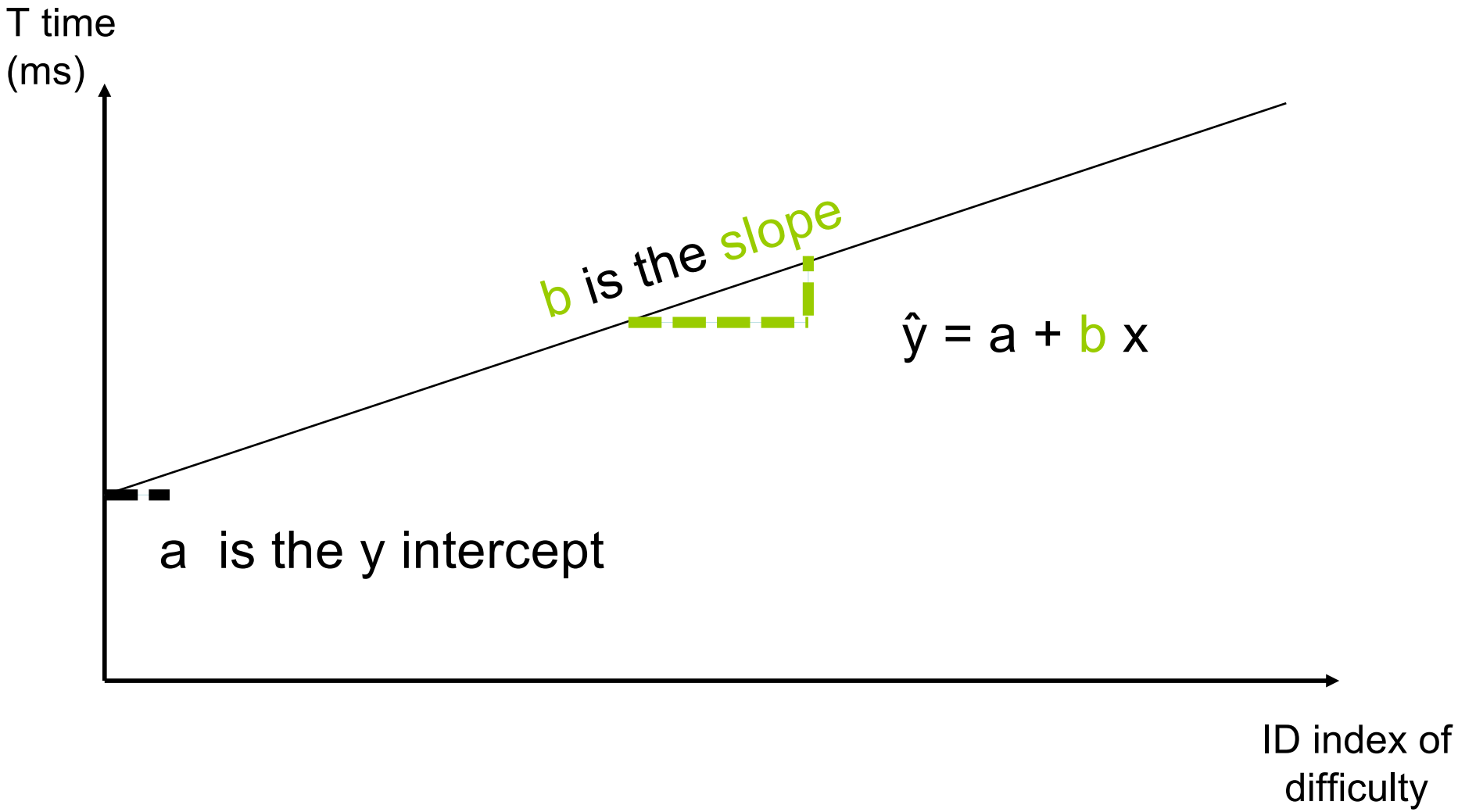
arguably the most basic technique for **machine learning**

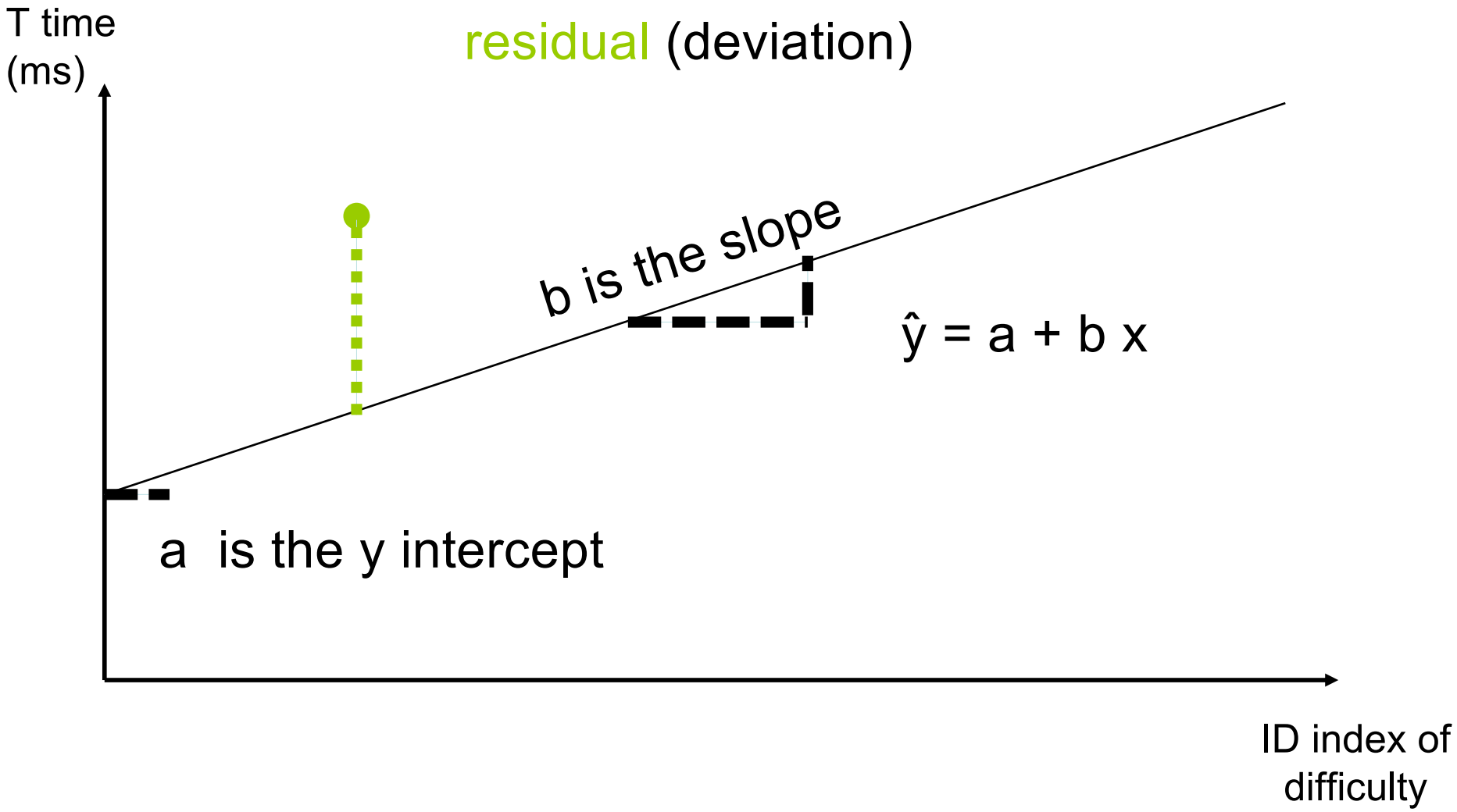
quick terminology of regressions


T time
(ms)
ŷ = a + b x
independent variable
ID index of
difficulty

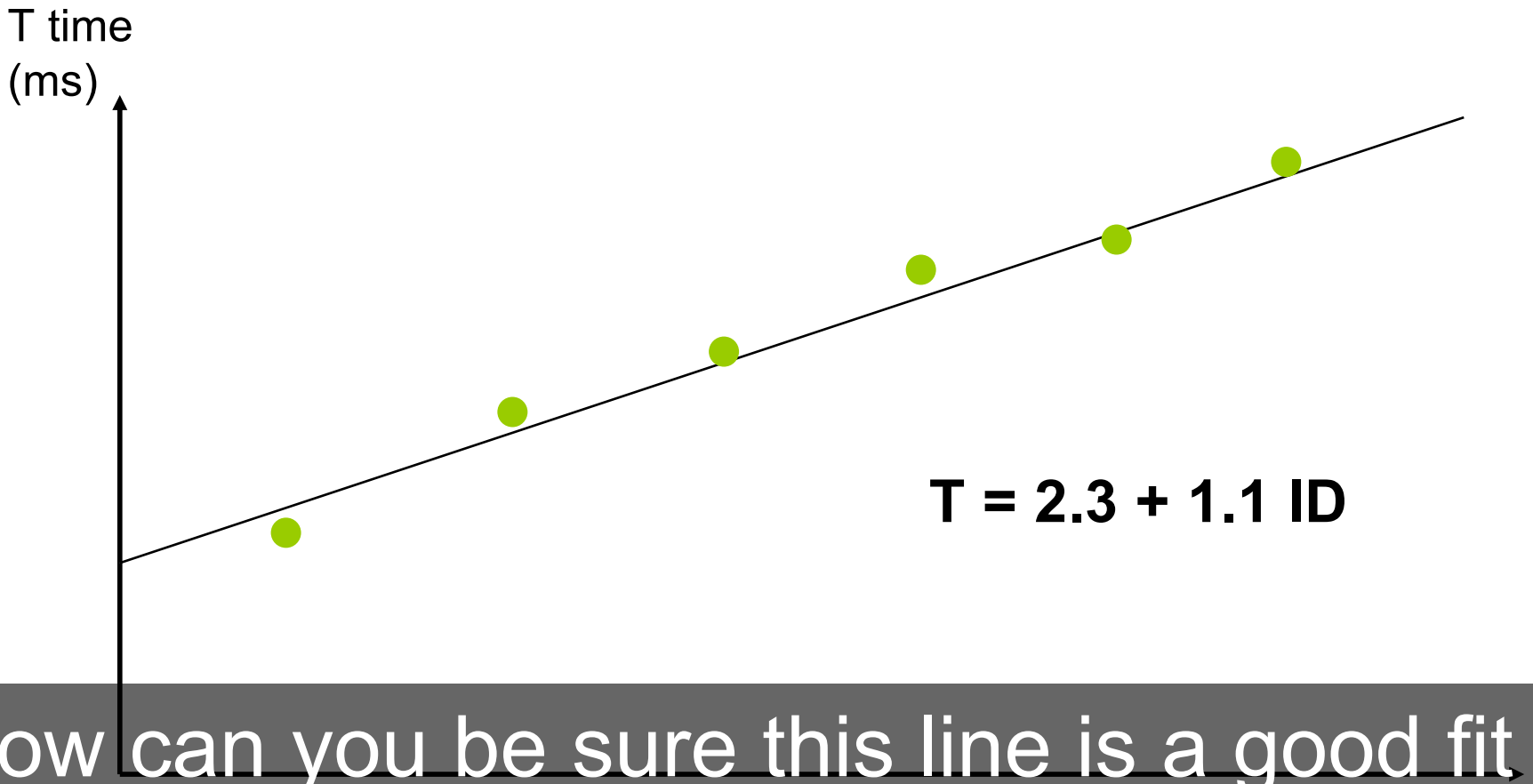

T time
(ms)
dependent variable
ŷ = a + b x
independent variable
ID index of
difficulty


T time
(ms)
ŷ = a + b x
a is the y intercept
ID index of
difficulty

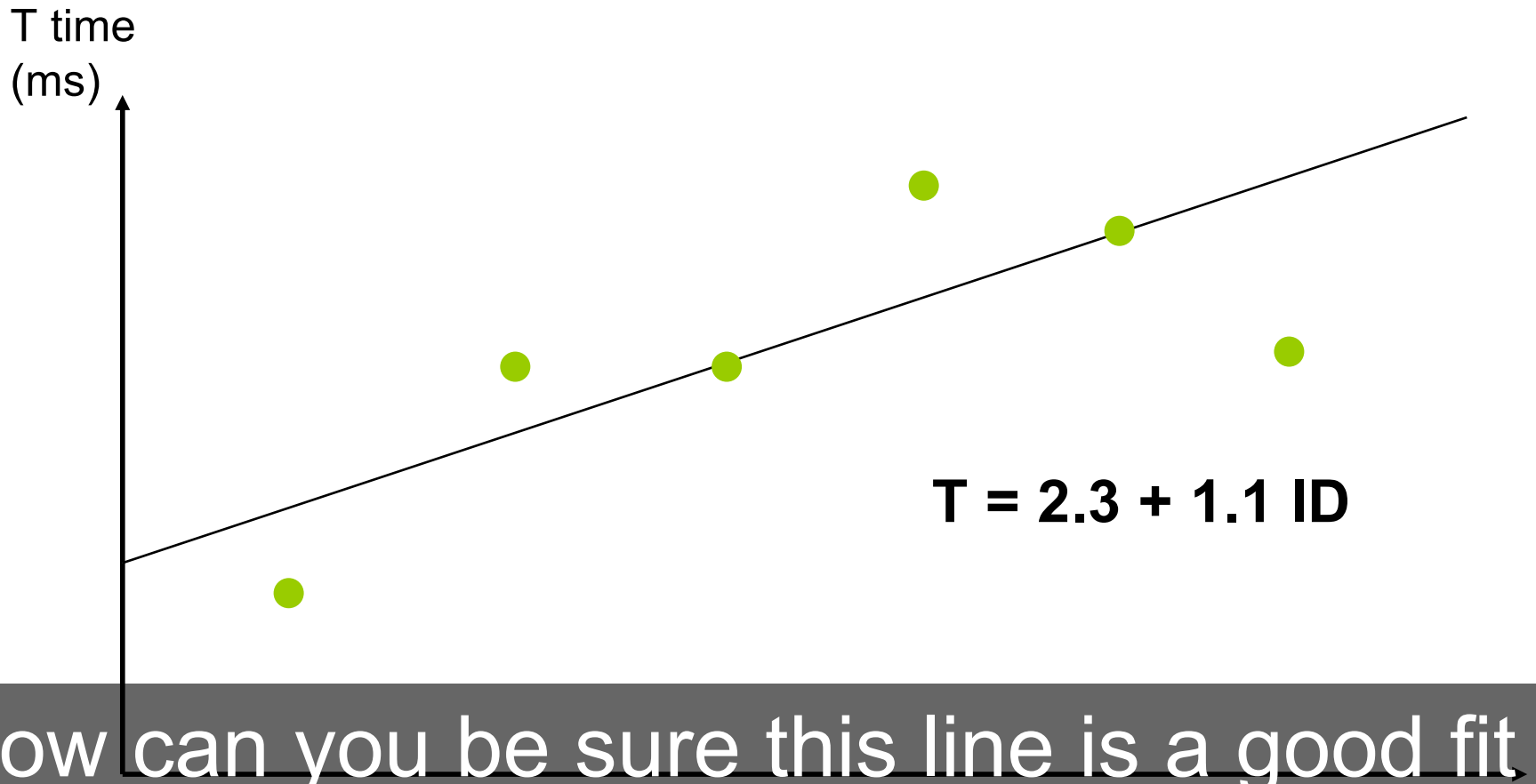

T time
(ms)
b is the slope
ŷ = a + b x
a is the y intercept
ID index of
difficulty


T time
(ms)
residual (deviation)
b is the slope
ŷ = a + b x
a is the y intercept
ID index of
difficulty

goodness of fit



how can you be sure this line is a good fit to our data?



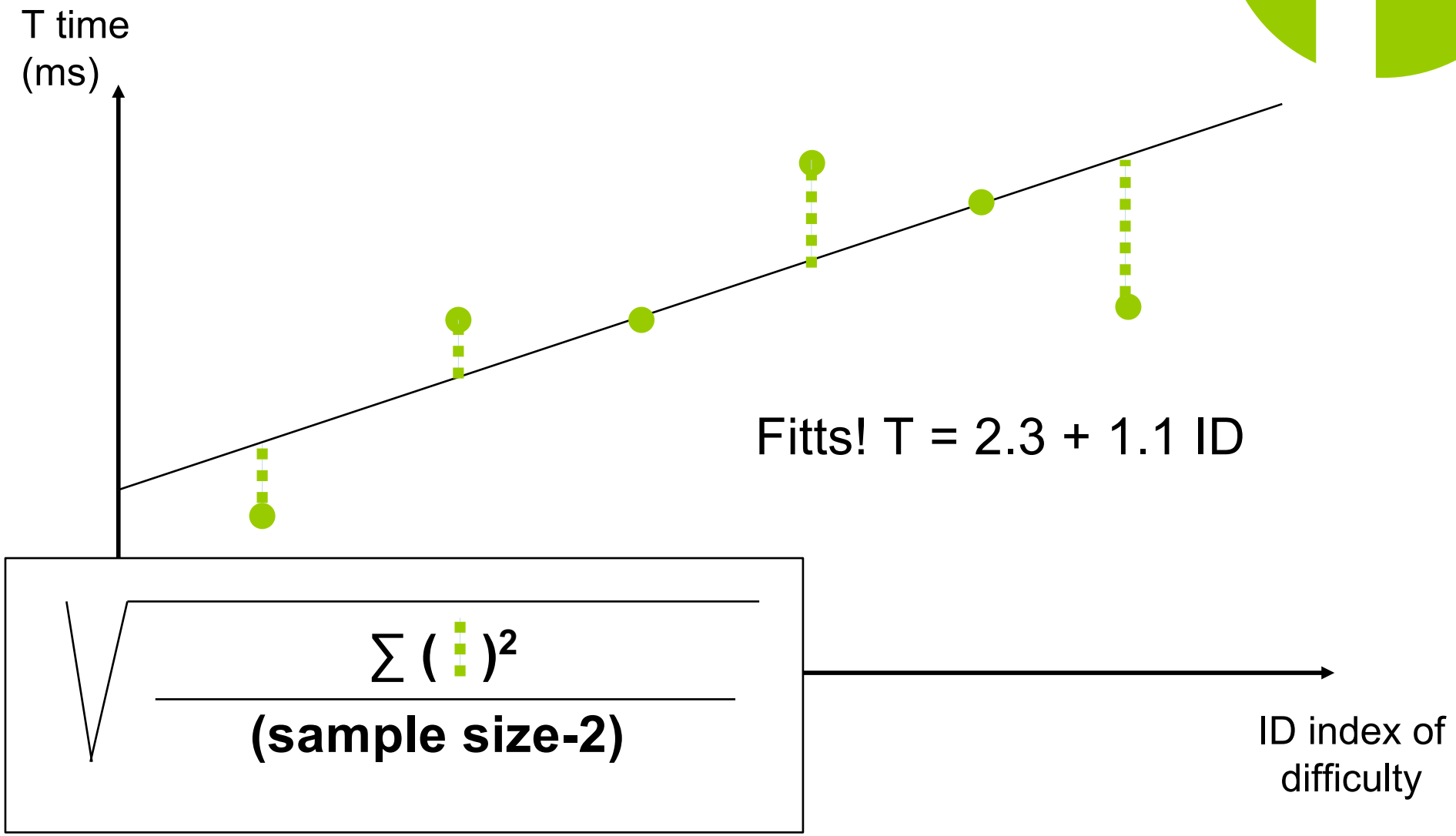
how can you be sure this line is a good fit to our data? what about now?

<1min brainstorming with your neighbor>

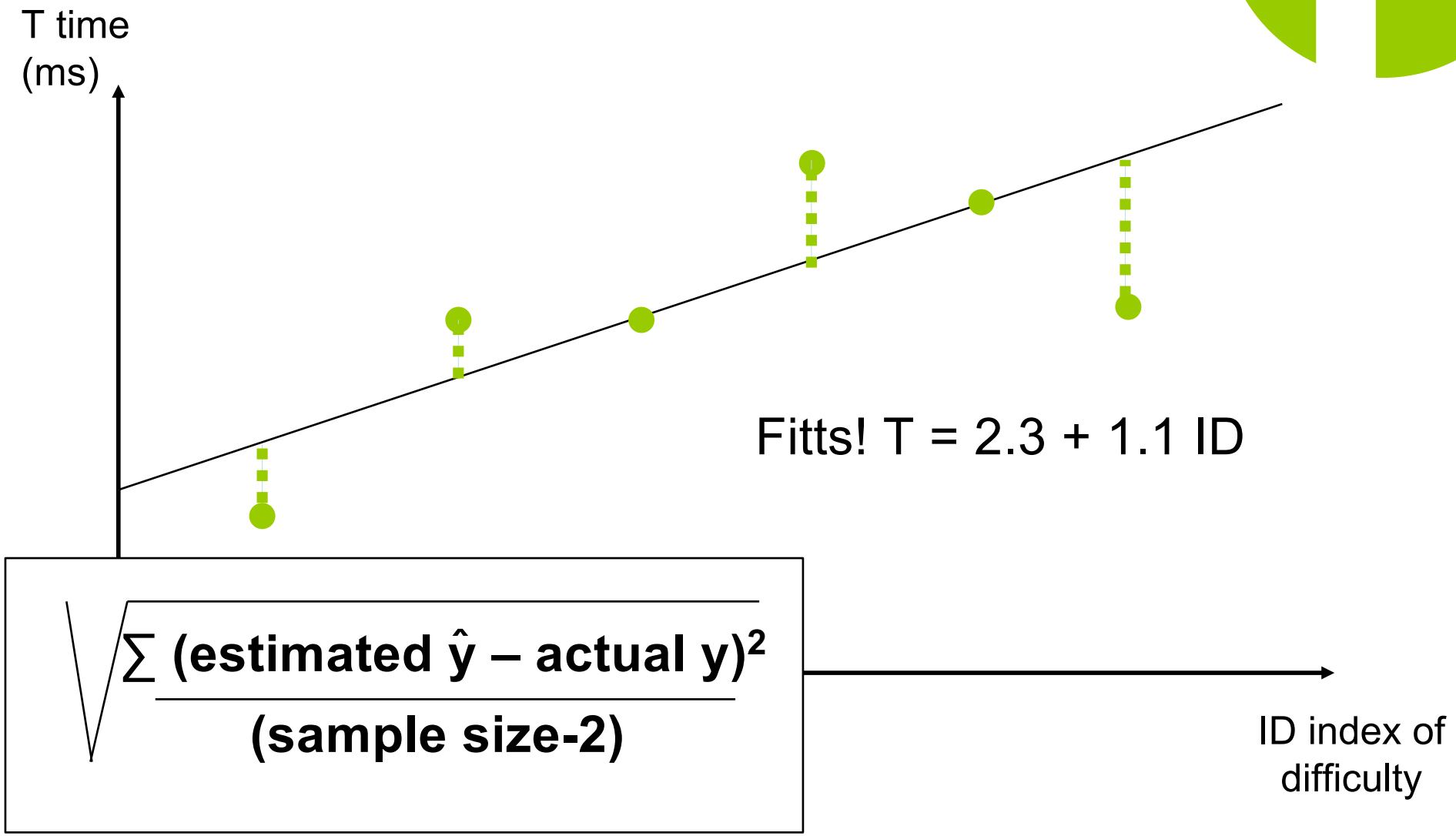
we can compute the **goodness of fit** with several methods

e.g. standard error of the estimate
or R squared

standard error of the estimate

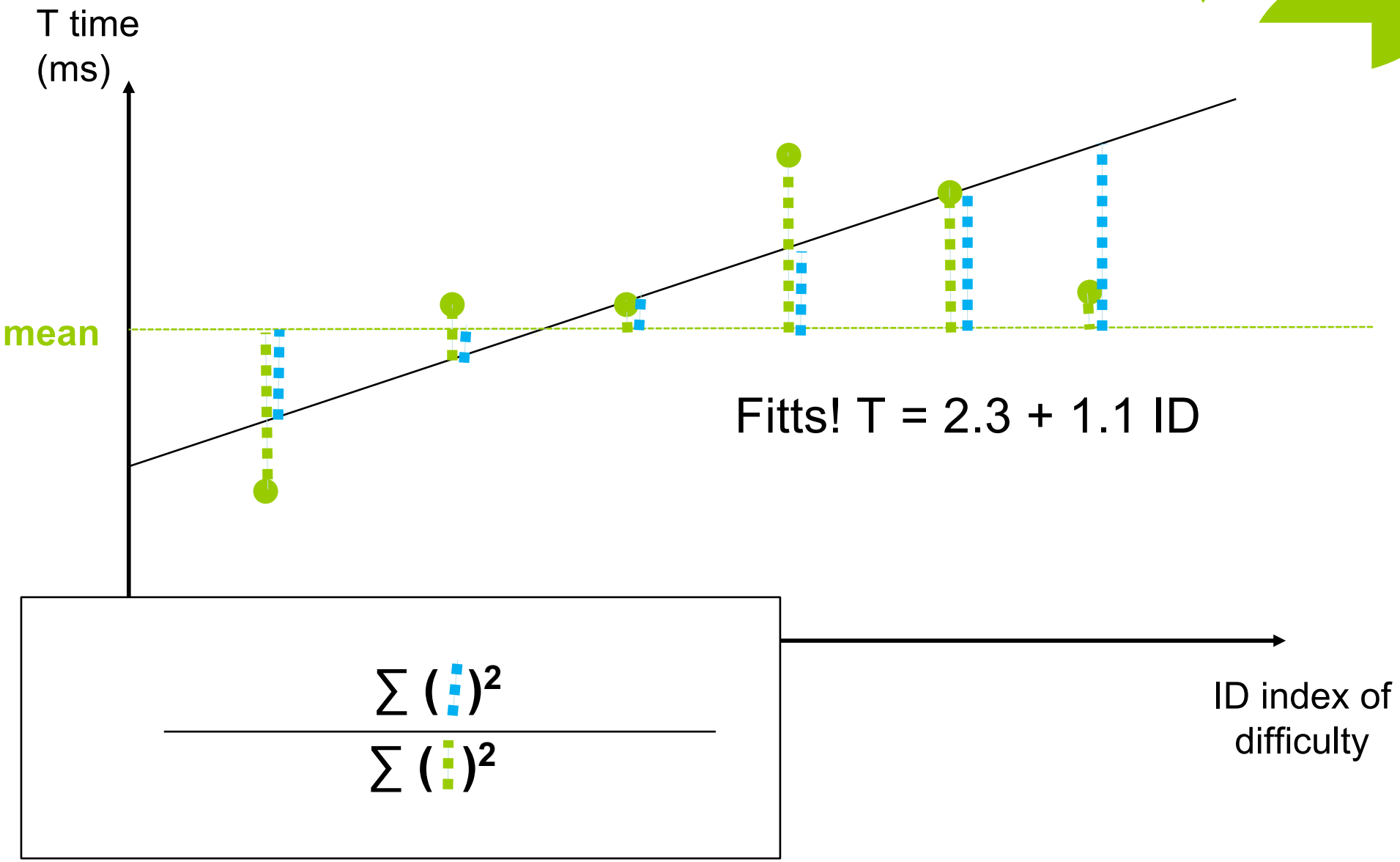


standard error of the estimate



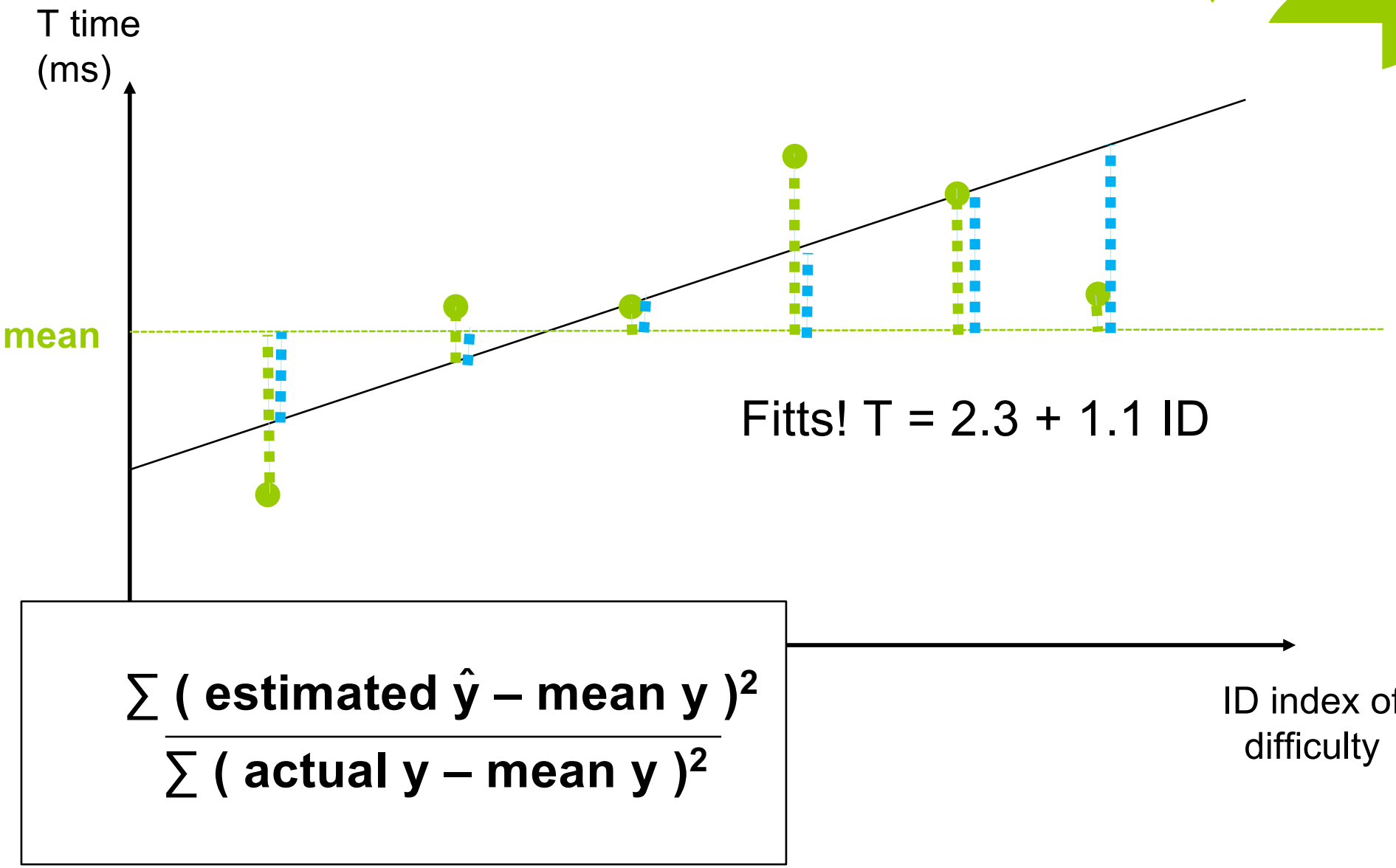
S gives a standard error in the metric of the data (the less the better)

R squared



$$\frac{\sum (\text{dashed})^2}{\sum (\text{dashed})^2}$$

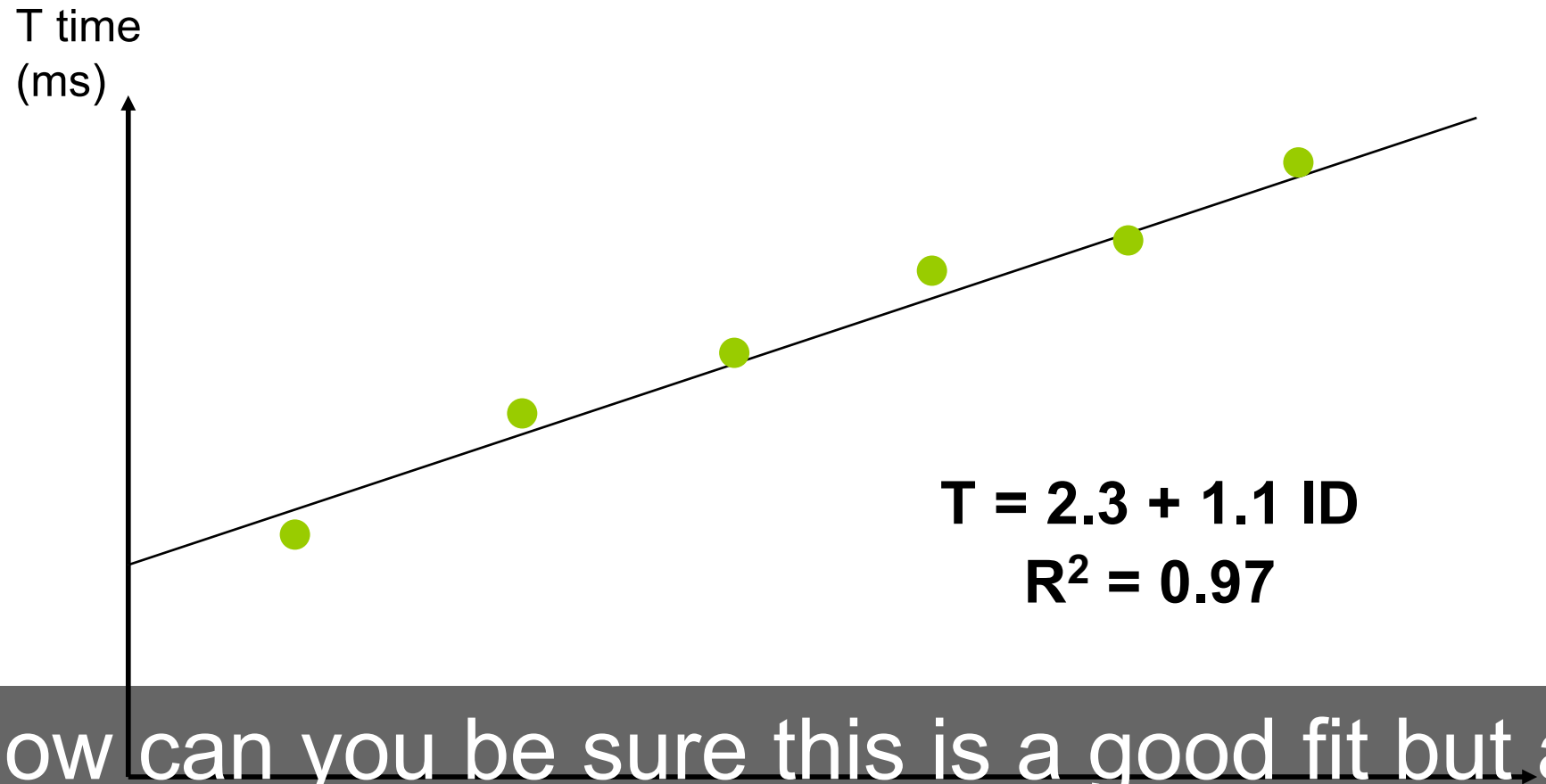
R squared



$$\frac{\sum (\text{estimated } \hat{y} - \text{mean } y)^2}{\sum (\text{actual } y - \text{mean } y)^2}$$

R^2 gives a percentage and 100% means perfect fit (>70% is better)

we become sure at X % that the equation fits the data



how can you be sure this is a good fit but also a good representation of the human ability?

to **gain additional confidence** we **repeat**

we gain trust in a model if it fits the data with **little error** when

1. it is verified with a **lot of data**
2. it holds across very **different people,**
3. it is verified in **independent studies...**

The information capacity of the human motor system in controlling the amplitude of movement.

PM Fitts - Journal of experimental psychology, 1954 - psycnet.apa.org

Reports of 3 experiments testing the hypothesis that the average duration of responses is directly proportional to the minimum average amount of information per response. The results show that the rate of performance is approximately constant over a wide range of movement amplitude and tolerance limits. This supports the thesis that" the performance capacity of the human motor system plus its associated visual and proprioceptive feedback mechanisms, when measured in information units, is relatively constant over a considerable …

☆ 99 Cited by 7707 Related articles All 18 versions Web of Science: 3367

Fitts's original paper is probably the most cited in HCI, studies done and redone many times

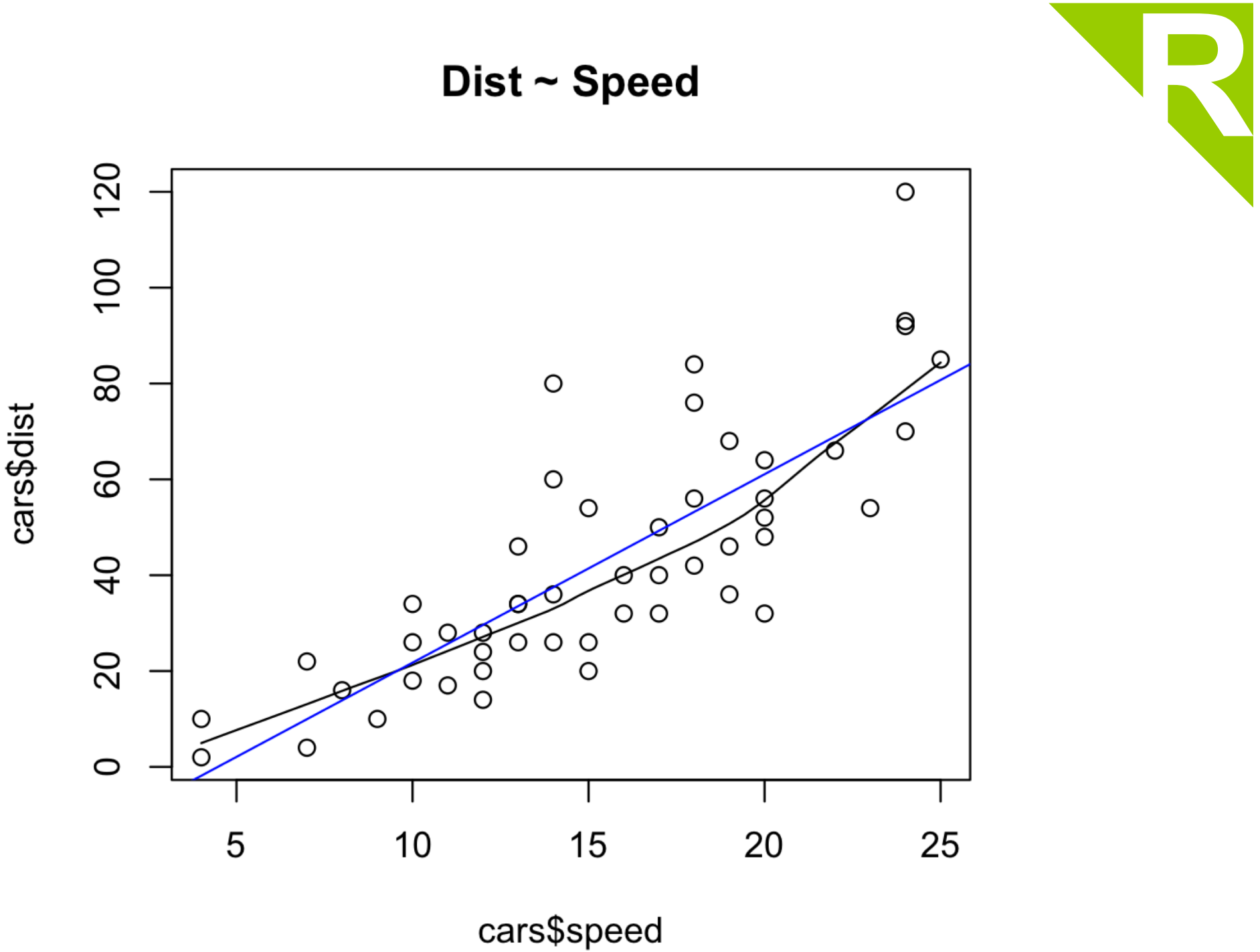
practically

```
vpn-user-244-044:~ nenisea$ R

R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> print ("hello world!")
[1] "hello world!"
>
```

we will be using R and I will try to give you as much as possible of examples

in your terminal

```
head(cars) # cars is a table that already comes with R and
contain 50 observations of speed and distance in two rows

scatter.smooth(x=cars$speed, y=cars$dist, main="Dist ~
Speed")

linearMod <- lm(dist ~ speed, data=cars) # build linear
regression model

abline(linearMod, col="blue") # draw the regression line

summary(linearMod) # goodness of fit
```


Dist ~ Speed
cars$dist
0
20
40
60
80
100
120
5
10
15
20
25
cars$speed

```
Call:
lm(formula = dist ~ speed, data = cars)


Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

dist = -17.6 + 2.9 * speed

```
Call:
lm(formula = dist ~ speed, data = cars)


Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

also note this

<0.01

<0

(will explain next week)

other applications

predictive analysis

predicting the number of items a consumer will purchase
e.g. check your amazon/ebay history!

predicting the number of shopper who will pass in front of a particular public billboard and use the data for advertisement bidding

predicting the number of claim in a given time period (used by insurance companies)

optimize processes

understand the impact of each machine on a production line on the quality of a product

understand the relationship between wait times of callers and number of complaints in a call centre

a retail store manager wanting to extend shopping hours to increase sales, but regression indicates that increase in revenue not sufficient to support rise in operating expenses

here a curve = polynomial fitting

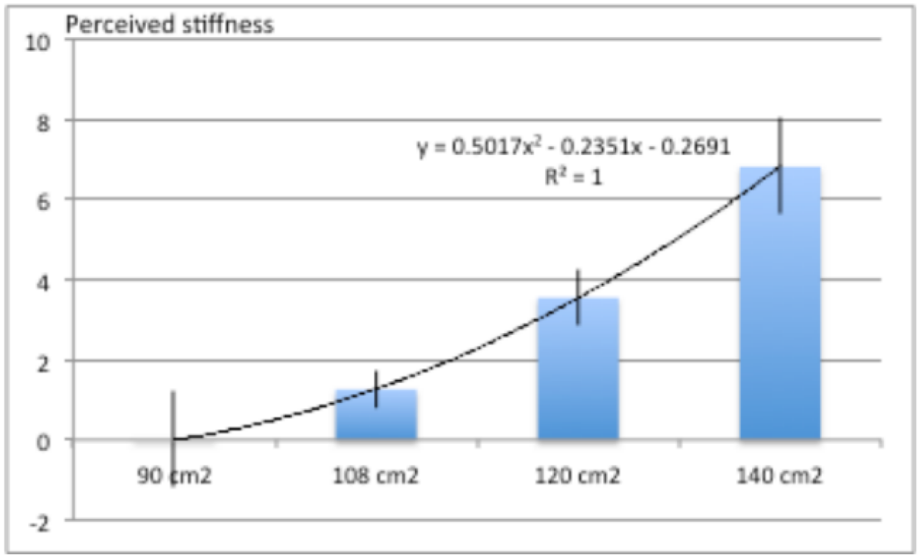


Figure 11. Bradley-Terry-Luce model output as well as a polynomial regression.

Our results are illustrated in Figure 11. We observed a clear distinction between the perceived stiffness of the 4 patches, the size of the patch increasing the perceived stiffness. In particular A is the least restrictive, followed by B, then C and then D is the most restrictive. We found that each paired comparison was significant ($p<0.0125$). This thus allows us to compare the different patches and conclude that D is the most efficient patch. We also performed a polynomial regression on our data and found a very accurate fit: $y=0.5017x2-0.2351x-0.2691$ ($R^2=1$). This suggests a quadratic correlation between the area of the patch and the perceived stiffness, which allows us to imagine bigger patches in order to restrict movements of the knee, which would require more stiffness. Of course further investigations need to be done to confirm this.

betwe
the ai
jam th

Prelin

The g
lab. A
the co
potent

First c
One p
hands
sugge
sugge
player
the u
"defro
where
player
for ter
mover
our id

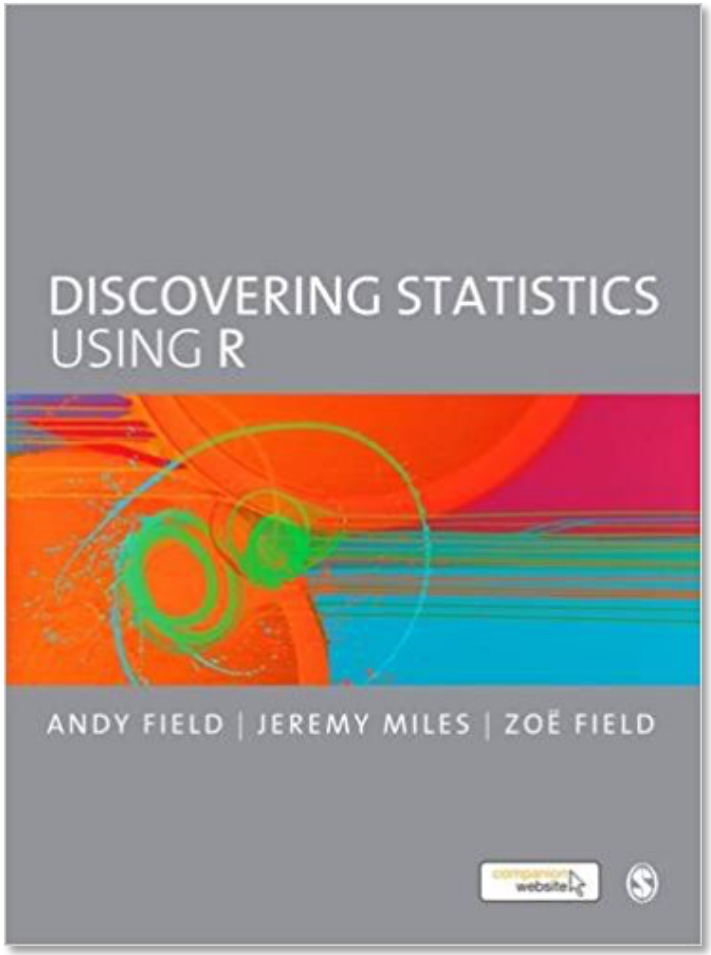
We al
becor
that th
in two
impro
particl
imple

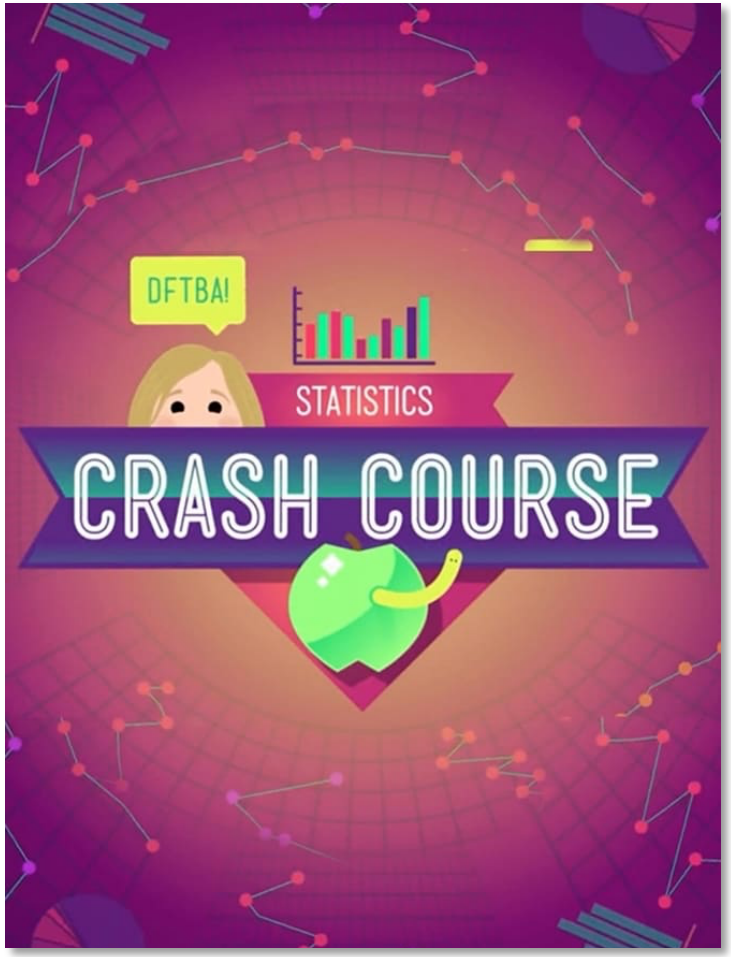
Patch

polynomial regression model

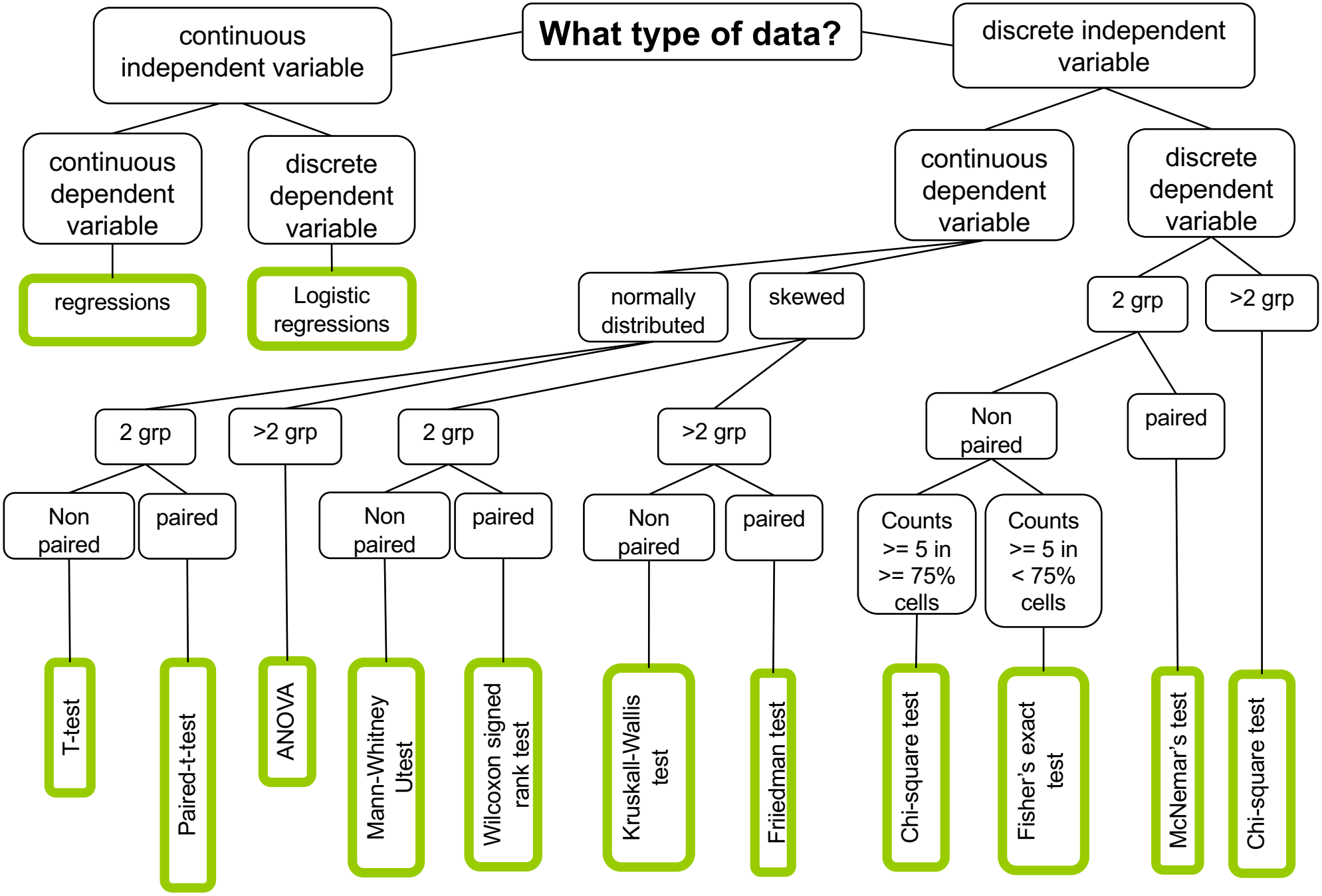
```
Mod2 <- lm(dist~poly(speed,2,raw=TRUE), data=cars)

Mod3 <- lm(dist~poly(speed,3,raw=TRUE), data=cars)

Mod4 <- lm(dist~poly(speed,4,raw=TRUE), data=cars)
```

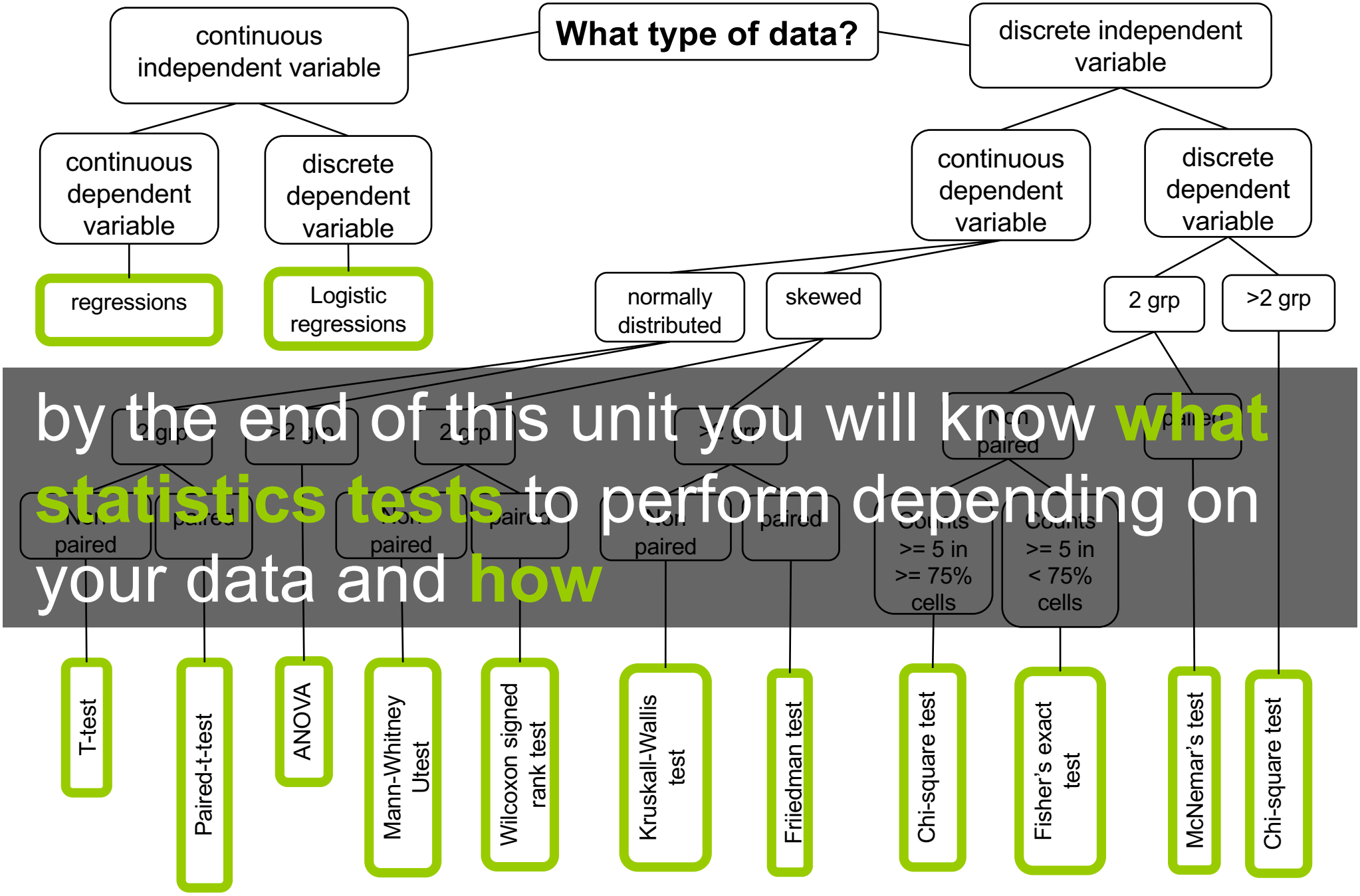
the next half of this unit

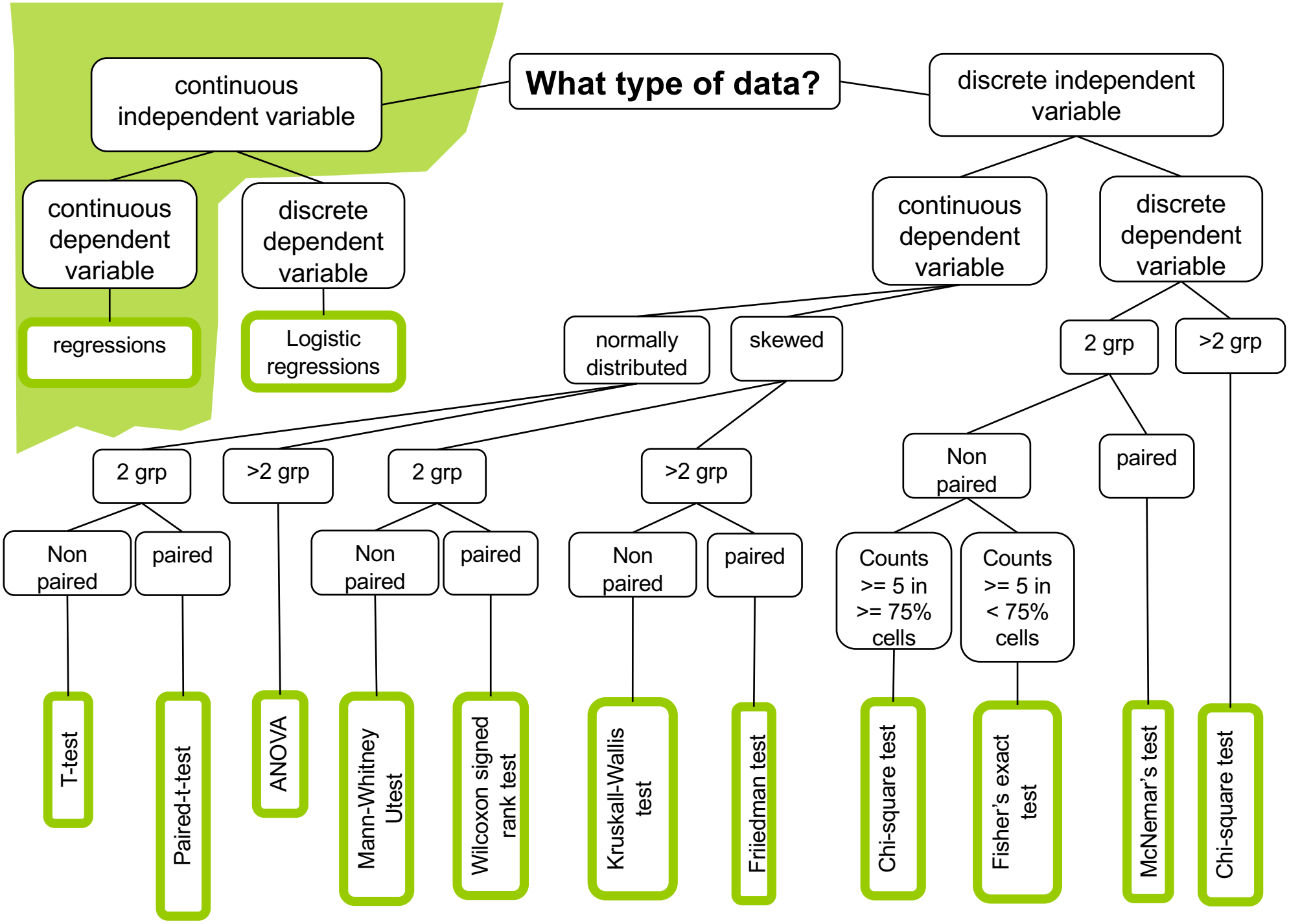


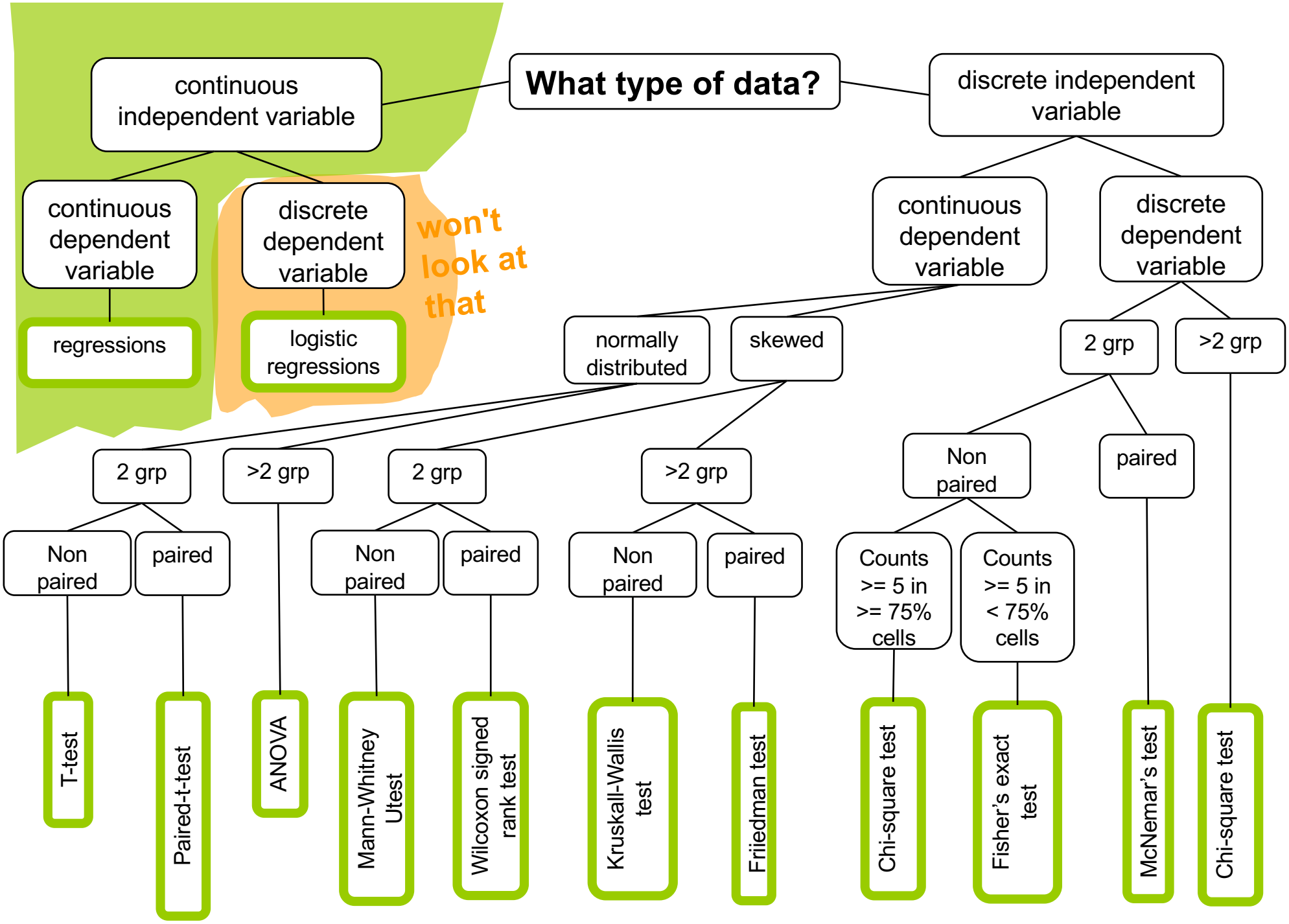


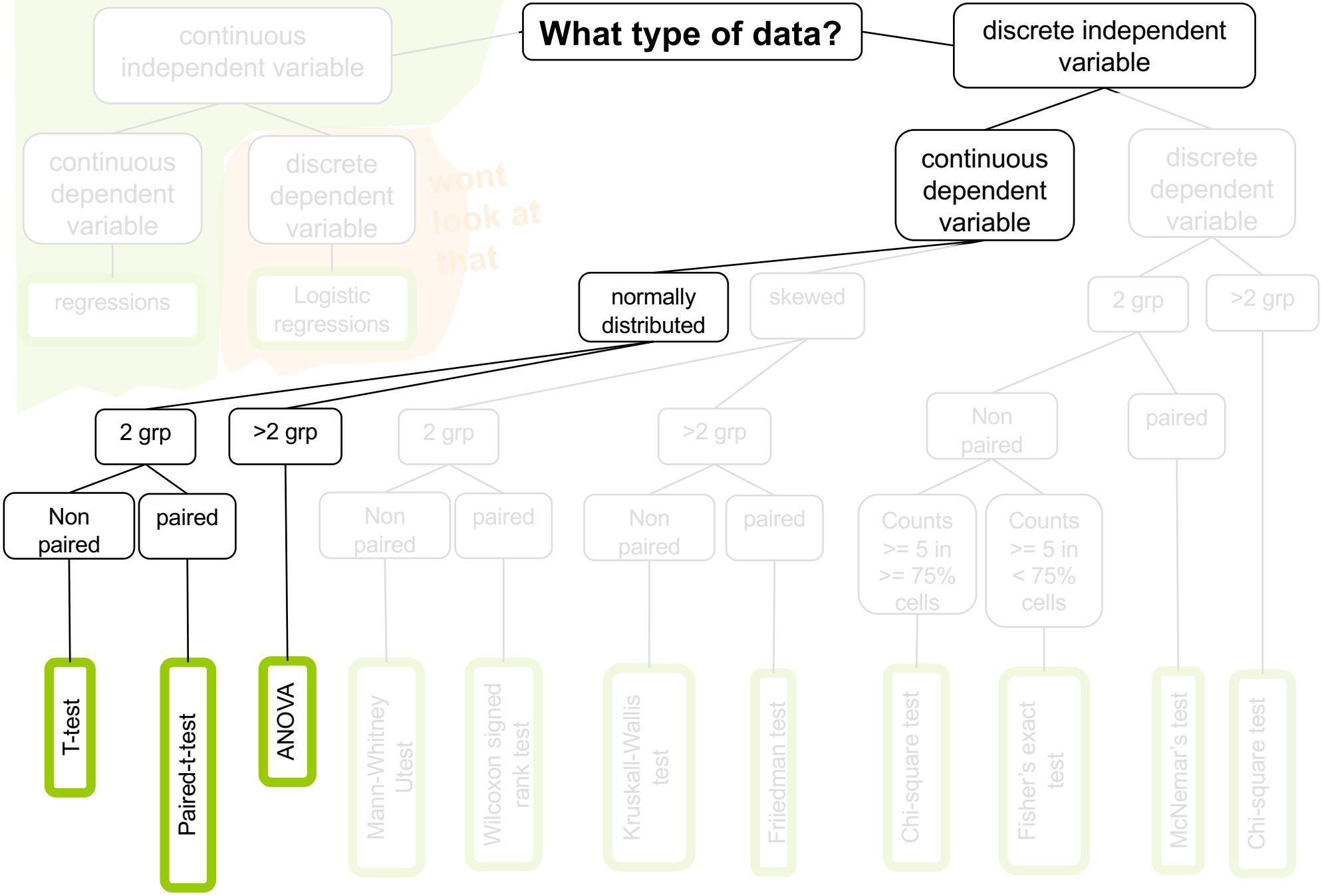
the **text book** I am using and a suggestion of **YouTube video channel**


What type of data?
continuous independent variable
discrete independent variable
continuous dependent variable
discrete dependent variable
regressions
Logistic regressions
continuous dependent variable
discrete dependent variable
normally distributed
skewed
2 grp
>2 grp
2 grp
>2 grp
Non paired
paired
Non paired
paired
Non paired
paired
Non paired
paired
T-test
Paired-t-test
ANOVA
Mann-Whitney Utest
Wilcoxon signed rank test
Kruskall-Wallis test
Friedman test
Counts >= 5 in >= 75% cells
Counts >= 5 in < 75% cells
Chi-square test
Fisher's exact test
McNemar's test
Chi-square test


What type of data?
continuous independent variable
discrete independent variable
continuous dependent variable
discrete dependent variable
continuous dependent variable
discrete dependent variable
regressions
Logistic regressions
normally distributed
skewed
2 grp
>2 grp
by the end of this unit you will know what statistics tests to perform depending on your data and how
paired
Counts >= 5 in >= 75% cells
Counts >= 5 in < 75% cells
T-test
Paired-t-test
ANOVA
Mann-Whitney Utest
Wilcoxon signed rank test
Kruskall-Wallis test
Friedman test
Chi-square test
Fisher's exact test
McNemar's test
Chi-square test


What type of data?
continuous independent variable
discrete independent variable
continuous dependent variable
discrete dependent variable
regressions
Logistic regressions
continuous dependent variable
discrete dependent variable
normally distributed
skewed
2 grp
>2 grp
2 grp
>2 grp
Non paired
paired
Non paired
paired
Non paired
paired
Non paired
paired
T-test
Paired-t-test
ANOVA
Mann-Whitney Utest
Wilcoxon signed rank test
Kruskall-Wallis test
Friedman test
2 grp
>2 grp
Non paired
paired
Counts >= 5 in >= 75% cells
Counts >= 5 in < 75% cells
Chi-square test
Fisher's exact test
McNemar's test
Chi-square test


What type of data?
continuous independent variable
discrete independent variable
continuous dependent variable
discrete dependent variable
won't look at that
regressions
logistic regressions
continuous dependent variable
discrete dependent variable
normally distributed
skewed
2 grp
>2 grp
2 grp
>2 grp
Non paired
paired
Non paired
paired
Non paired
paired
Non paired
paired
Counts >= 5 in >= 75% cells
Counts >= 5 in < 75% cells
T-test
Paired-t-test
ANOVA
Mann-Whitney Utest
Wilcoxon signed rank test
Kruskall-Wallis test
Friedman test
Chi-square test
Fisher's exact test
McNemar's test
Chi-square test


What type of data?
discrete independent variable
continuous dependent variable
discrete dependent variable
normally distributed
skewed
2 grp
>2 grp
Non paired
paired
T-test
Paired-t-test
ANOVA
continuous independent variable
continuous dependent variable
discrete dependent variable
regressions
Logistic regressions
wont look at that
2 grp
>2 grp
Non paired
paired
Mann-Whitney Utest
Wilcoxon signed rank test
Non paired
paired
Kruskall-Wallis test
Friedman test
Non paired
paired
Counts >= 5 in >= 75% cells
Counts >= 5 in < 75% cells
Chi-square test
Fisher's exact test
2 grp
>2 grp
McNemar's test
Chi-square test

R plots

https://www.harding.edu/fmccown/r/

good videos

https://www.youtube.com/watch?v=WWqE7YHR4Jc

https://www.youtube.com/playlist?list=PLF596A4043DBEAE9C

end