THE UNIVERSITY *of* EDINBURGH
School of Philosophy, Psychology and Language Sciences

Using Jupyter Notebook

A python taster course for social sciences research

Welcome!

About us...

Jenny Culbertson

Andres Karjus

Matt Spike

Alexander Martin

Welcome!

About you...

1. **What University are you from?**
2. **What is your research area?**
3. **How many of you...**

   Collect your own data?

   Have taken a stats class?

   Know a programming language? Which one?

Agenda

1. **Introduction**
   Basic notebook “syntax”
2. **Getting into some data**
   Data analysis 1: a classic psychology experiment
3. **Break (11:30)**
4. **Getting into some data**
   Data analysis 2: data from oxford admissions
5. **Improving your workflow**
   Why use Jupyter Notebook?

Introduction

66 Project GA Experiment Progress Cont'd

Continued from Page 59

DAY 10

0uM GA Cubes: Explants are darker green than GA treatment cubes. Explants are ~2-4.5cm long. Stems are ~~1 MEDIUM~~ 0.5-1.5cm long with leaves ~1-4.5cm long. White spots on leaves only present on ~1/4 of the leaves → a decrease from Day 8 (p. 59)

1 uM GA Cubes: Explants much paler green than 0uM cubes, some yellowing starting to occur. Explants are ~3-9cm long with stems ranging between ~0cm (an explant has no stem) and ~6cm long and leaves between ~1-4.5cm long. All leaves are curling inward.

3.5uM GA Cubes: Explants are a lighter green than 0uM GA cubes and all leaves are curling inward. Explants are ~3.5-9.5cm long with stems ranging between ~1.5-5.5cm long and leaves between ~1-4cm. Some explants are still tipped over.

7uM GA Cubes: Explants are darker than the ~~E MEDIUM~~ 1uM and 3.5uM GA cubes but lighter green than the 0uM GA cubes and all leaves are curled inward. Explants are ~3-8.5cm long with stems ranging between ~1.5-5cm and leaves between ~1-5cm.

MBP 18 SEP14

Continued on Page 70

Signed 18 SEP14 Date

Read and Understood By

Signed Date

67 Project Tissue Culture

Continued from Page

Plant Material

Nodal segments with axial buds were cut from a vegetative CW plant as outlined by Lata et al. (2009).

Mother Plant Label	38
# of Cuttings Made	18
Approx. Length	~15
Axillary Buds Present	Yes
Avg # of Leaves Intact	1-3

Notes: Side branch from 38 fell off. So cut off as many explants from that branch

Initials: DHL Date: DHL 19 SEP 14

~~DHL 19 SEP14~~

Inoculation

Disinfected explants were inoculated on MS Shooting Medium under aseptic conditions as outlined by Lata et al. (2009).

MS Shooting Media Batch #	KGP17Sep14
# of Explants per Cube	25
# of Cubes	7
Total # of Explants Inoculated	175
Cube Label(s): MJC19Sept14-1	KGP19Sept14 (1-6)

Notes:

Initials: MBP Date: 19 Sept 14

KGP19Sep14

MBP

Continued on Page 68

Signed Date

Read and Understood By

Signed Date

What's a Notebook?

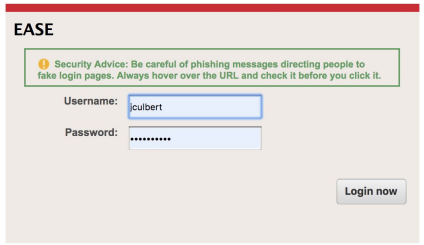
Great things about a notebook:

1. You can mix text, tables, drawings, calculations *all in one place*
2. You can *document* things
   - your design, procedure, participants
   - why you made specific decisions
   - what you've not done yet
3. They can help you quickly transition to a poster, talk, manuscript

In short: they are good for your workflow

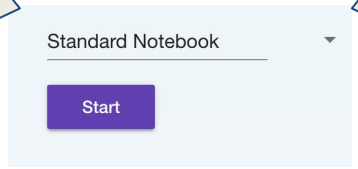
Time to find your notebook!

1 https://ease.ed.ac.uk



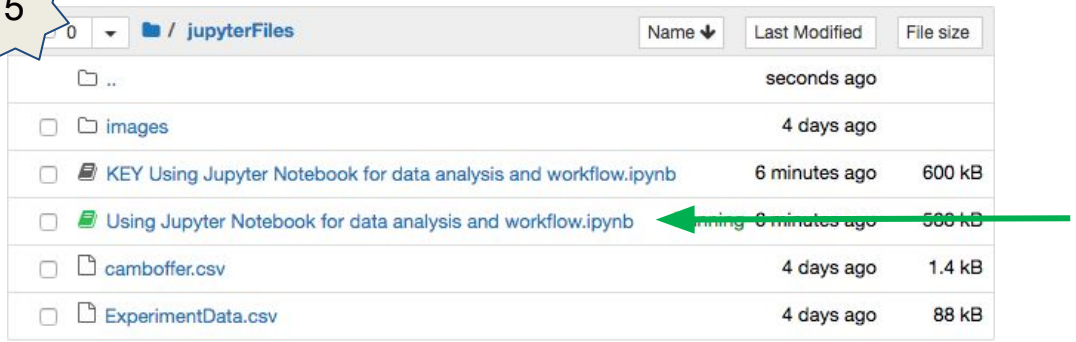
2 https://noteable.edina.ac.uk/login

3



4

5



Dataset 1: The Stroop Task

Background

The Stroop Task is one of the best known psychological experiments. It is named after John Ridley Stroop. The "Stroop effect" describes the difficulty people have in naming the ink color of a color word if there is a mismatch between ink color and word. For example, the word GREEN printed in red ink.

It is easier to measure key presses than the time it takes to name something aloud; so, there "manual" Stroop tasks involve pressing keys.

Stroop demo

https://www.psytoolkit.org/experiment-library/stroop.html

The data

There are 6 types of **trials**

text	letterColour	condition
red	red	congruent
red	green	incongruent
blue	blue	congruent
blue	red	incongruent
green	green	congruent
green	blue	incongruent

The data

Each participant does 6 repetitions of these 6 types of trials...so how many trials each?

text	letterColour	condition
red	red	congruent
red	green	incongruent
blue	blue	congruent
blue	red	incongruent
green	green	congruent
green	blue	incongruent

RED

The data

participant	text	letterColour	condition	RT
1	red	red	congruent	0.92

BLUE

The data

participant	text	letterColour	condition	RT
1	red	red	congruent	0.92
1	blue	blue	congruent	1.01

GREEN

The data

participant	text	letterColour	condition	RT
1	red	red	congruent	0.92
1	blue	blue	congruent	1.01
1	green	red	incongruent	1.22
...				

GREEN

The data

participant	text	letterColour	condition	RT
1	red	red	congruent	0.92
1	blue	blue	congruent	1.01
1	green	red	incongruent	1.22
...				
2	green	green	congruent	0.86

RED

The data

participant	text	letterColour	condition	RT
1	red	red	congruent	0.92
1	blue	blue	congruent	1.01
1	green	red	incongruent	1.22
...				
2	green	green	congruent	0.86
2	red	blue	incongruent	1.18

BLUE

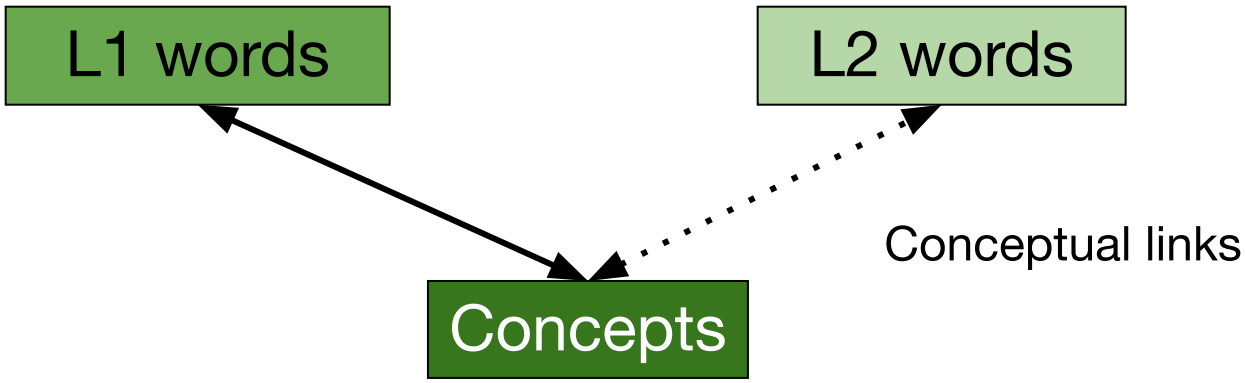
The data

participant	text	letterColour	condition	RT
1	red	red	congruent	0.92
1	blue	blue	congruent	1.01
1	green	red	incongruent	1.22
...				
2	green	green	congruent	0.86
2	red	blue	incongruent	1.18
2	blue	red	incongruent	1.14
...				

Dataset 1: The Stroop Task

Do L2 speakers show the stroop effect?

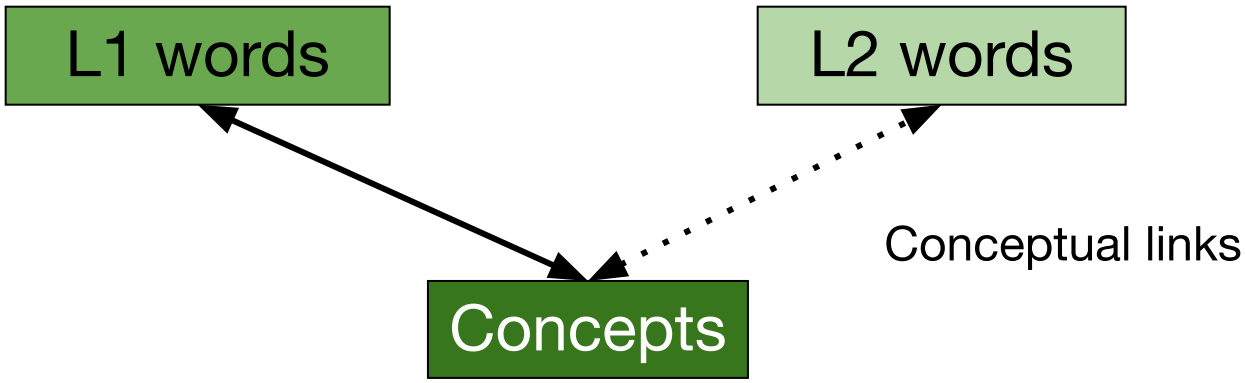
Some theories of L2 knowledge argue that the links between words and concept are stronger for the L1 than the L2.



Dataset 1: The Stroop Task

Do L2 speakers show the stroop effect?

Prediction: stronger interference for L1 than L2 speakers



ROJO

The data

participant	text	letterColour	condition	RT	group
1	red	red	congruent	0.92	L1
1	blue	blue	congruent	1.01	L1
1	green	red	incongruent	1.22	L1
...					
2	green	green	congruent	0.86	L1
2	red	blue	incongruent	1.18	L1
2	blue	red	incongruent	1.14	L1
...					

The data

participant	text	letterColour	condition	RT	group
1	red	red	congruent	0.92	L1
...					
2	green	green	congruent	0.86	L1
...					
3	green	blue	incongruent	0.86	L2
...					
4	blue	blue	congruent	0.94	L2
...					

Data analysis

Improving your workflow: Don’t forget what you’ve done!

Example of a (real-life) structure

- import stuff
- load in all data and concatenate files together

Entrée [2]:
```
cheminD = '../data/3nouns/'
filesD = [i for i in os.listdir(cheminD) if re.compile('.*-coded\.csv').match(i)]
sujs = set([i.split('-')[0] for i in filesD])

#for i in sujs: print i
```

Production

Entrée [3]:
```
df = pd.concat([pd.read_csv(cheminD+'{}-coded.csv'.format(s)) for s in sujs], ignore_index=True)
```

- check for participants not meeting a certain criterion

Example of a (real-life) structure

Check for bad subjects

Entrée [7]:
```
singleMods = df[df.trialType=='1mod2'].groupby('sujet', as_index=False).homo.mean()

badParticipants = list(singleMods[singleMods.homo<0.85].sujet.unique())
extraWeirdos = [
    'ENG4003', # did prenominal 2mod orders
]
badParticipants += extraWeirdos
print badParticipants
#df[df.sujet.isin(badParticipants)].groupby('cond').sujet.nunique()

df = df[~df.sujet.isin(badParticipants)]

df['condC'] = df.cond.map(lambda x: 0.5 if x=='dems' else -0.5)

df.groupby('cond').sujet.nunique()

# must exclude 3002 for not having enough data in 2mod because only gave 1mod responses (now done auto)

df.to_csv('../data/3nouns/cooked-for-anaR.csv', index=None)
```

```
['ENG2005', 'ENG2017', 'ENG2023', 'ENG3001', 'ENG3002', 'ENG3014', 'ENG4006', 'ENG4007', 'ENG4019', 'ENG4003']
```

Entrée [8]:
```
df.groupby('cond').sujet.nunique()
```

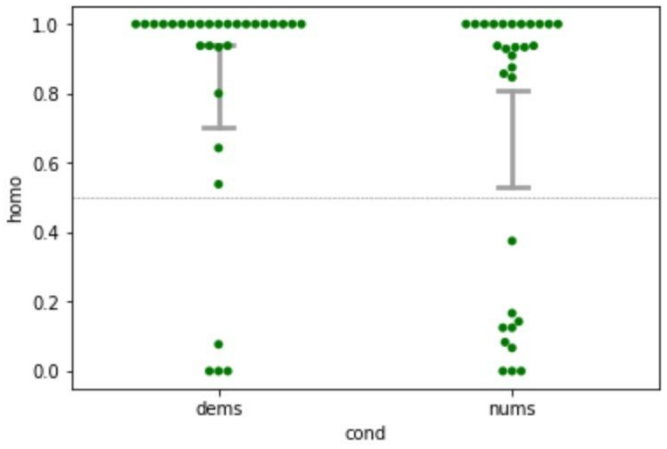
Out[8]:
```
cond
dems    30
nums    30
Name: sujet, dtype: int64
```

Example of a (real-life) structure

Entrée [9]:
```
aggregators = {'homo':mean}

gp = df[df.trialType=='2mod'].groupby(['sujet', 'cond'], as_index=False).agg(aggregators)
```

Entrée [10]:
```
fig, ax = subplots()

order = ['dems', 'nums']

sns.barplot(x='cond', y='homo', data=gp, ax=ax, alpha=0, capsize=0.1, errcolor='darkgrey', order=order)
sns.swarmplot(x='cond', y='homo', data=gp, ax=ax, color='green', order=order)

ax.set(ylim=(-0.05,1.05))
ax.plot(range(-1,4), [0.5]*5, color='grey', linestyle='--', linewidth=0.5)
```

Out[10]:
```
[<matplotlib.lines.Line2D at 0x1a109798d0>]
```



Example of a (real-life) structure

Entrée [11]:

```
%%R

library('lme4')
setwd('/Users/Alexander/Documents/Science/U20/POS/NPO/lab/data/')

d <- read.csv('./3nouns/cooked-for-anaR.csv')
test <- subset(d, d$trialType=='2mod') # only select two modifier trials
```

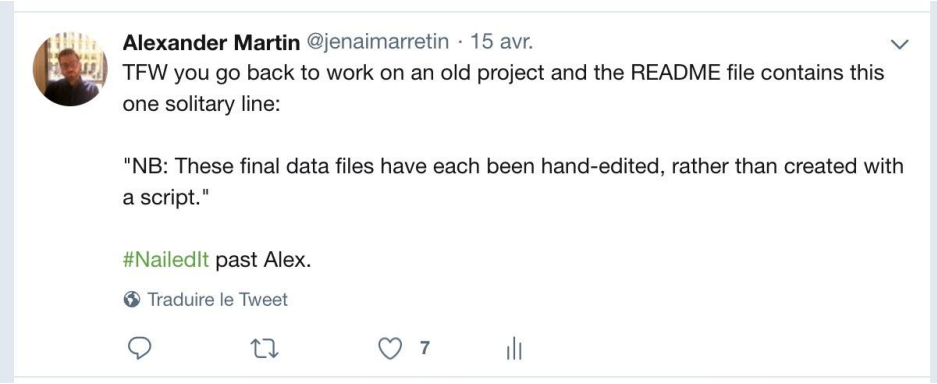
Entrée [12]:

```
%%R

m1 <- glmer(homo ~ condC + (1|sujet), data=test, family=binomial)
m0 <- glmer(homo ~ 1 + (1|sujet), data=test, family=binomial)

anova(m1, m0)
```

```
Data: test
Models:
m0: homo ~ 1 + (1 | sujet)
m1: homo ~ condC + (1 | sujet)
   Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
m0  2 410.23 419.8 -203.11   406.23
m1  3 407.24 421.6 -200.62   401.24 4.9934      1    0.02544 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Benefits

1. Overstretched memory load relief
   - How did I exclude participants again?
   - What did my models look like *exactly*?
   - It's all in one place!
2. Easily shareable with collaborators
   - Ok, it's a bit ugly for public sharing, but you can make a nicer one for OSF or github...your collaborators can get in on it right away.
   - Readable by people who don't know how to code!
3. Makes your workflow **explicit**, **accessible**, and **reproducible** (including by five-years-from-now you)
   - Automating everything can avoid tweets like this:



Resources

1. Berkeley Initiative for Transparency in the Social Sciences
   - Plenty of resources for all kinds of transparency practices for people with all kinds of data
2. Open Science Framework
   - Great place for pre-registrations where you lay out your data collection and analysis plans *before* you do them
3. Github
   - Resource for tracking the history of your updates on all file types

Extra fun stuff!!