HARMONY

Harmony

Open source NLP/AI tool for psychologists and social and health sciences
Data discovery and harmonisation

harmonydata.ac.uk
github.com/harmonydata/

Thomas Wood
Fast Data Science

Bettina Moltrecht, PhD
Centre for Longitudinal Studies
UCL

UCL
UKRI Economic and Social Research Council

MEET THE HARMONY TEAM

Rachel Gomes
University College London

Richard Thomas
UK-LLC

Eoin McElroy
Ulster University

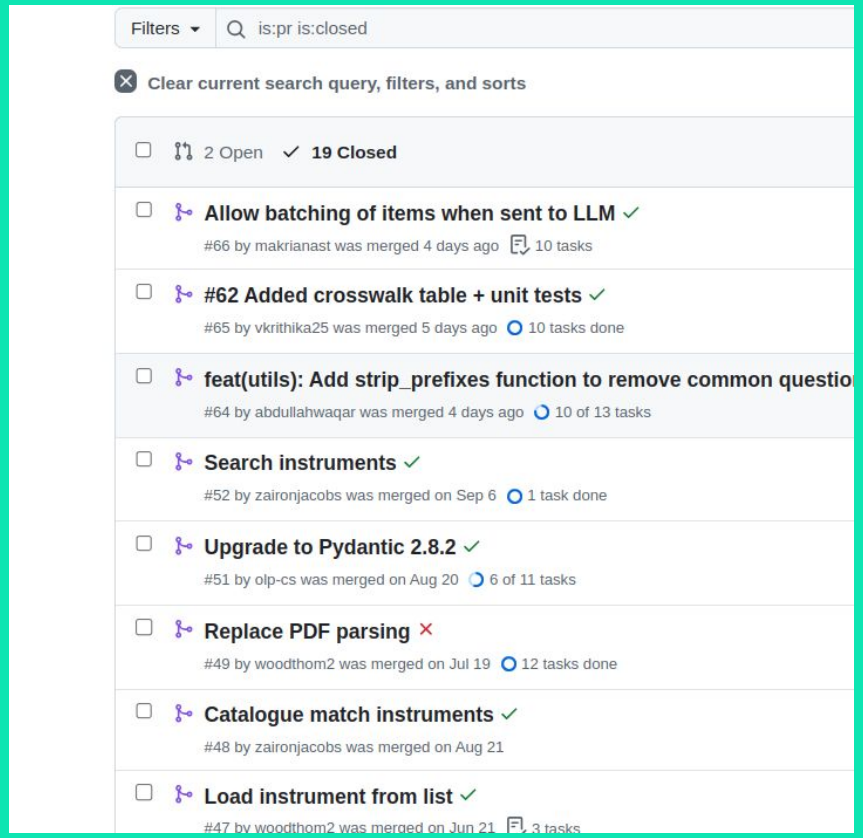
Bettina Moltrecht
University College London

Louise Arseneault
Kings College London

Thomas Wood
Fast Data Science

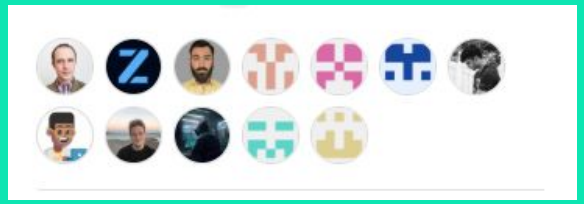
Mauricio Hoffmann
Universidade Federal de Santa Maria

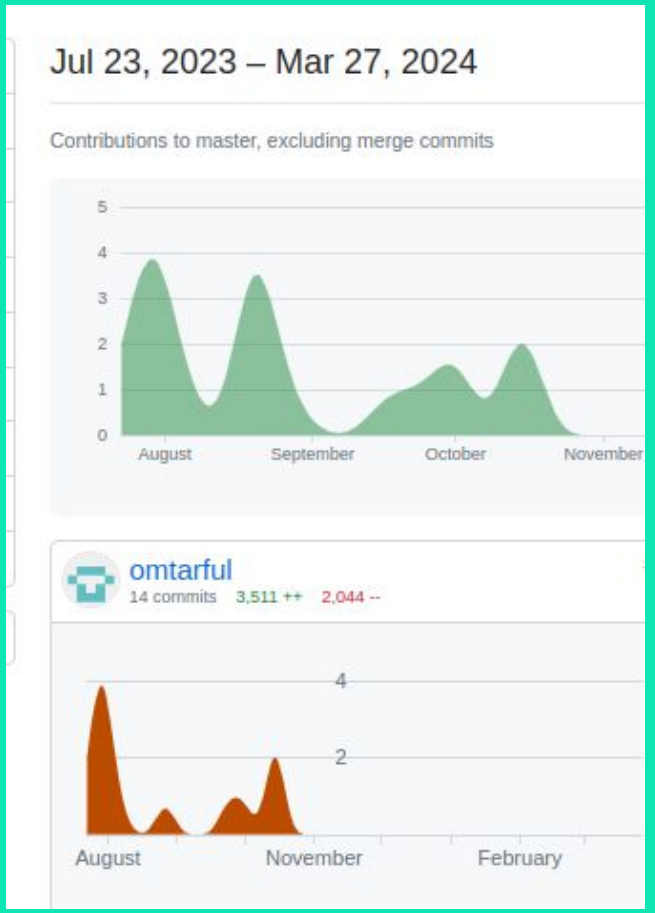
CONTRIBUTORS



...plus people around the world who have made 24 pull requests to Github!

vkrithika25
makrianast
shahid-0
EveWCheng
0x48piraj
nikhildevre
omtarful
Gabe-ferr
abdullahwaqar
deannavarley





Partners

Ulster University

UCL

Fast Data Science

wellcome

UKRI Economic and Social Research Council

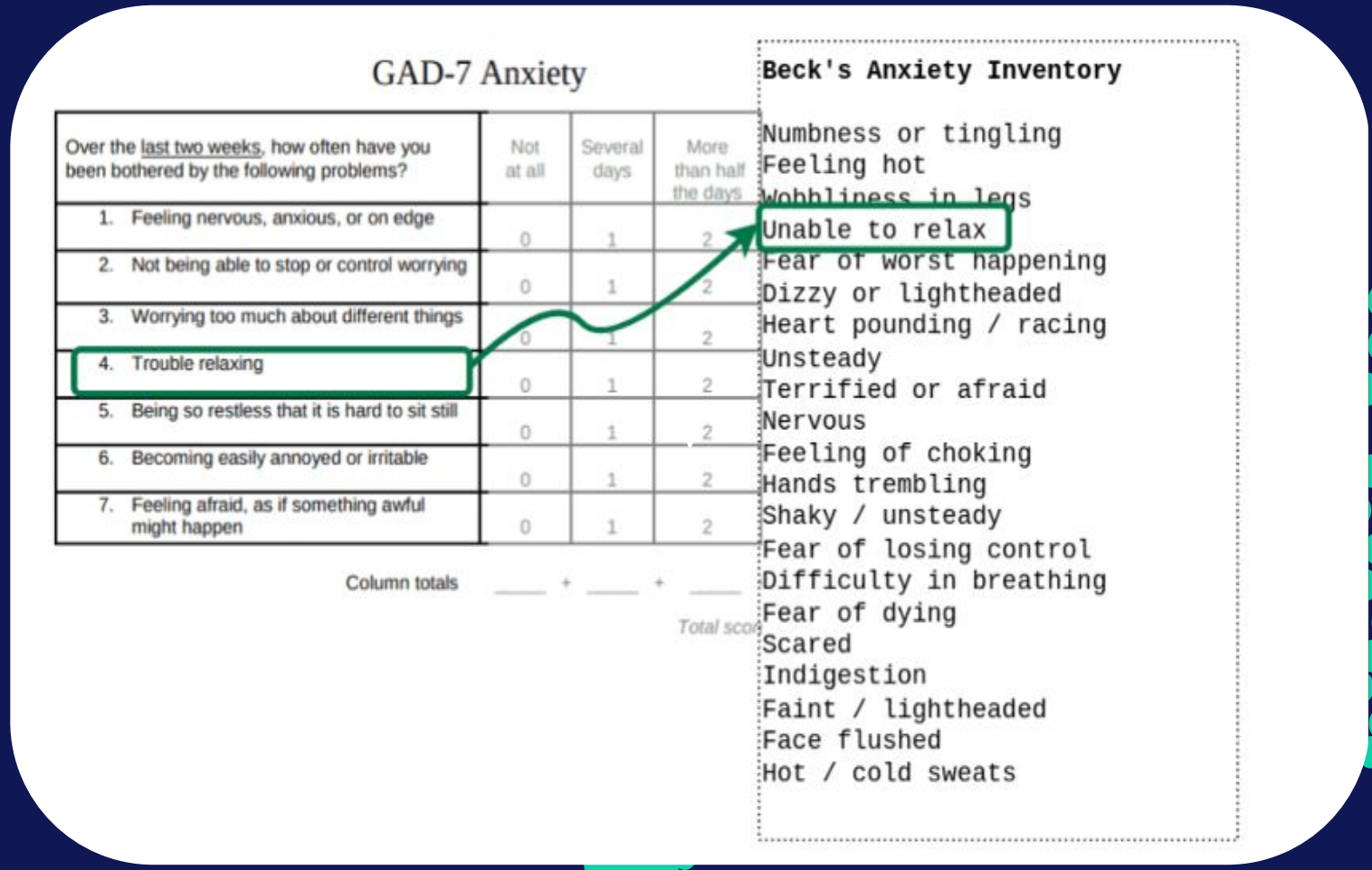
UK Longitudinal Linkage Collaboration

The challenge…

Combining studies…

Retrospective harmonisation

We have to match different survey items, sometimes in different languages



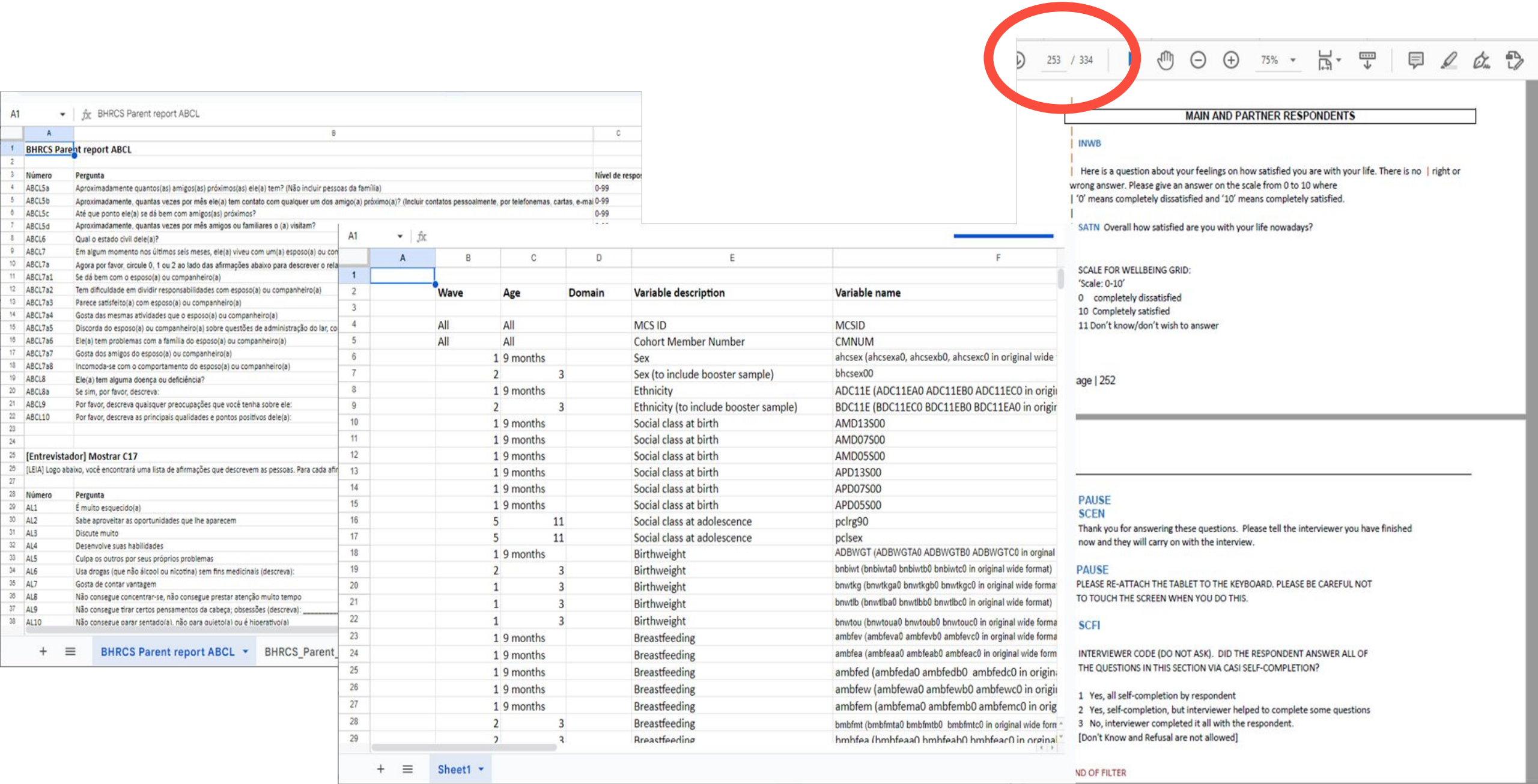
An example...

Research using data from the UK and Brazil

UK: Millennium cohort study
19,000 children

Brazilian High Risk Cohort Study for Childhood Mental Health Conditions (BHRCS).
2,511 young people

Surveys were in different languages, using different instruments



BHRCS Parent report ABCL

Número	Pergunta	Nível de respo
ABCL5a	Aproximadamente quantos(as) amigos(as) próximos(as) ele(a) tem? (Não incluir pessoas da família)	0-99
ABCL5b	Aproximadamente, quantas vezes por mês ele(a) tem contato com qualquer um dos amigo(a) próximo(a)? (Incluir contatos pessoalmente, por telefonemas, cartas, e-mai	0-99
ABCL5c	Até que ponto ele(a) se dá bem com amigos(as) próximos?	0-99
ABCL5d	Aproximadamente, quantas vezes por mês amigos ou familiares o (a) visitam?	
ABCL6	Qual o estado civil dele(a)?	
ABCL7	Em algum momento nos últimos seis meses, ele(a) viveu com um(a) esposo(a) ou con	
ABCL7a	Agora por favor, circule 0, 1 ou 2 ao lado das afirmações abaixo para descrever o rela	
ABCL7a1	Se dá bem com o esposo(a) ou companheiro(a)	
ABCL7a2	Tem dificuldade em dividir responsabilidades com esposo(a) ou companheiro(a)	
ABCL7a3	Parece satisfeito(a) com esposo(a) ou companheiro(a)	
ABCL7a4	Gosta das mesmas atividades que o esposo(a) ou companheiro(a)	
ABCL7a5	Discorda do esposo(a) ou companheiro(a) sobre questões de administração do lar, co	
ABCL7a6	Ele(a) tem problemas com a família do esposo(a) ou companheiro(a)	
ABCL7a7	Gosta dos amigos do esposo(a) ou companheiro(a)	
ABCL7a8	Incomoda-se com o comportamento do esposo(a) ou companheiro(a)	
ABCL8	Ele(a) tem alguma doença ou deficiência?	
ABCL8a	Se sim, por favor, descreva:	
ABCL9	Por favor, descreva quaisquer preocupações que você tenha sobre ele:	
ABCL10	Por favor, descreva as principais qualidades e pontos positivos dele(a):	
	[Entrevistador] Mostrar C17	
	[LEIA] Logo abaixo, você encontrará uma lista de afirmações que descrevem as pessoas. Para cada afir	
Número	Pergunta	
AL1	É muito esquecido(a)	
AL2	Sabe aproveitar as oportunidades que lhe aparecem	
AL3	Discute muito	
AL4	Desenvolve suas habilidades	
AL5	Culpa os outros por seus próprios problemas	
AL6	Usa drogas (que não álcool ou nicotina) sem fins medicinais (descreva):	
AL7	Gosta de contar vantagem	
AL8	Não consegue concentrar-se, não consegue prestar atenção muito tempo	
AL9	Não consegue tirar certos pensamentos da cabeça; obsessões (descreva):	
AL10	Não consegue parar sentado(a), não para quieto(a) ou é hiperativo(a)	

BHRCS Parent report ABCL | BHRCS_Parent_

	Wave	Age	Domain	Variable description	Variable name
	All	All		MCS ID	MCSID
	All	All		Cohort Member Number	CMNUM
	1	9 months		Sex	ahcsex (ahcsexa0, ahcsexb0, ahcsexc0 in original wide
	2	3		Sex (to include booster sample)	bhcsex00
	1	9 months		Ethnicity	ADC11E (ADC11EA0 ADC11EB0 ADC11EC0 in origi
	2	3		Ethnicity (to include booster sample)	BDC11E (BDC11EC0 BDC11EB0 BDC11EA0 in origir
	1	9 months		Social class at birth	AMD13S00
	1	9 months		Social class at birth	AMD07S00
	1	9 months		Social class at birth	AMD05S00
	1	9 months		Social class at birth	APD13S00
	1	9 months		Social class at birth	APD07S00
	1	9 months		Social class at birth	APD05S00
	5	11		Social class at adolescence	pclrg90
	5	11		Social class at adolescence	pclsex
	1	9 months		Birthweight	ADBWGT (ADBWGTA0 ADBWGTB0 ADBWGTC0 in orginal
	2	3		Birthweight	bnbiwt (bnbiwta0 bnbiwtb0 bnbiwtc0 in original wide format)
	1	3		Birthweight	bnwtkg (bnwtkga0 bnwtkgb0 bnwtkgc0 in original wide forma
	1	3		Birthweight	bnwtlb (bnwtlba0 bnwtlbb0 bnwtlbc0 in original wide format)
	1	3		Birthweight	bnwtou (bnwtoua0 bnwtoub0 bnwtouc0 in original wide forma
	1	9 months		Breastfeeding	ambfev (ambfeva0 ambfevb0 ambfevc0 in orginal wide forma
	1	9 months		Breastfeeding	ambfea (ambfeaa0 ambfeab0 ambfeac0 in original wide form
	1	9 months		Breastfeeding	ambfed (ambfeda0 ambfedb0 ambfedc0 in origin
	1	9 months		Breastfeeding	ambfew (ambfewa0 ambfewb0 ambfewc0 in origi
	1	9 months		Breastfeeding	ambfem (ambfema0 ambfemb0 ambfemc0 in orig
	2	3		Breastfeeding	bmbfmt (bmbfmta0 bmbfmtb0 bmbfmtc0 in original wide form
	2	3		Breastfeeding	bmbfea (bmbfeaa0 bmbfeab0 bmbfeac0 in orginal

Sheet1

253 / 334 75%

MAIN AND PARTNER RESPONDENTS

INWB

Here is a question about your feelings on how satisfied you are with your life. There is no right or wrong answer. Please give an answer on the scale from 0 to 10 where '0' means completely dissatisfied and '10' means completely satisfied.

SATN Overall how satisfied are you with your life nowadays?

SCALE FOR WELLBEING GRID:
'Scale: 0-10'
0 completely dissatisfied
10 Completely satisfied
11 Don't know/don't wish to answer

age | 252

PAUSE
SCEN
Thank you for answering these questions. Please tell the interviewer you have finished now and they will carry on with the interview.

PAUSE
PLEASE RE-ATTACH THE TABLET TO THE KEYBOARD. PLEASE BE CAREFUL NOT TO TOUCH THE SCREEN WHEN YOU DO THIS.

SCFI

INTERVIEWER CODE (DO NOT ASK). DID THE RESPONDENT ANSWER ALL OF THE QUESTIONS IN THIS SECTION VIA CASI SELF-COMPLETION?

1 Yes, all self-completion by respondent
2 Yes, self-completion, but interviewer helped to complete some questions
3 No, interviewer completed it all with the respondent.
[Don't Know and Refusal are not allowed]

ND OF FILTER

Matching items Brazil_ UK .XLSX

File Edit View Insert Format Data Tools Help

Share

Menus 90% $ % .0 .00 123 Calibri 11

A1 Cohort

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Cohort	Variable name	wave	Question	Question in Portugu	Response opti	Recoding	Comments	Name (harmonized)	Final coding	Recoding/Comments		Response option	Question	wave	Variable name
7		PARENT-CHILD RELATIONSHIP (in contact: yes=0/no=1; quality: 3= not close, 2=somewhat, 1= very close 0= extremely close)														
10	BHRCS	mrelation (reg22)	dom_s01_reg_w0	How does/did the child and the biological mother get along when they are/were together?	Como a criança e mãe biológica se dão ou se davam quando estão/estavam juntas?	1, Very well|2, Well|3, More or less|4, bad|5, Very bad|88, Does not apply|99, Does not know	Brazil 4|5 = 3 Brazil 3= 2 Brazil 2= 1 Brazil 1= 0		mcrelation	3= not close, 2=somewhat close, 1= very close 0= extremely close	In MCS this is mostly mothers, but can also be father/other caregiver, so we either combine Brazil or do only	UK 4 = 0 UK 3= 1 UK 2= 2 UK 1= 3	1 Not very close 2 Fairly close 3 Very close 4 Extremely close 5 Dont know/Dont wish to answer	Overall, how close would you say you are to CM? (7)Overall, how close would you say you are to your father?	5, 6,7	EPSCHC00 FPSCHC00 (parent_cm interview) FCRLQM00 (cm interview) FCRLQF00 (cm interview) GCRLQF00 (cm interview) GPSCHC00parent_cm interview)
13		Marital status														
14	BHRCS	reg31	dom_s01_reg_w0w2	What is the marital status of the biological mother?	A mãe biológica da criança está no momento: casada ou morando junto com o pai biológico, casada ou morando junto com companheiro,	88, Refuses|99, Does not know|1, Married or living together with the biological	Brazil 1= 1 Brazil 2= 2 Brazil 3|4= 3 Brazil 5= 4		maritalstatus	1= Married or living with biological parent 2= Separated and living with someone else 3= Separated /Divorced		I think the recoding needs changed as we do not have Separated and living with someone else easily avaailable in MCS	Refused | -9 Dont know | -8 Not applicable | -1 Legally separated | 1 Married, 1st and only marriage | 2 Remarried, 2nd	Current legal marital status	5 , 6, 7	EPFCIN00 (parent interview) FPFCIN00 (parent interview) GPFCIN00 (parent interview)
15		PARENT DEATH (yes=1, no=0)														
16	BHRCS	reg36	dom_s01_reg_w0w2	Is the child's biological father known?	O pai biológico da criança é desconhecido, conhecido ou falecido?	1, Unknown|2, Known|3, Deceased	Brazil 3 = 1, Brazil 1|2 = 0		fatherstatus	0= not dead 1= dead	Coding options reduced as of wave 6	3=1 "dead" all other =0	1 Resident full-time in household | 2 Resident part-time in household | 3 Deceased | 4 Non-resident, in contact | 5 Non-resident, not in contact |6 Non resident, contact	Natural father status	5,6,7	ENATF00 Family derived FDNATF00 (family derived) GDNATF00 (family derived)

1 exposure var | control variables | 2 harmonised items


... months later ...

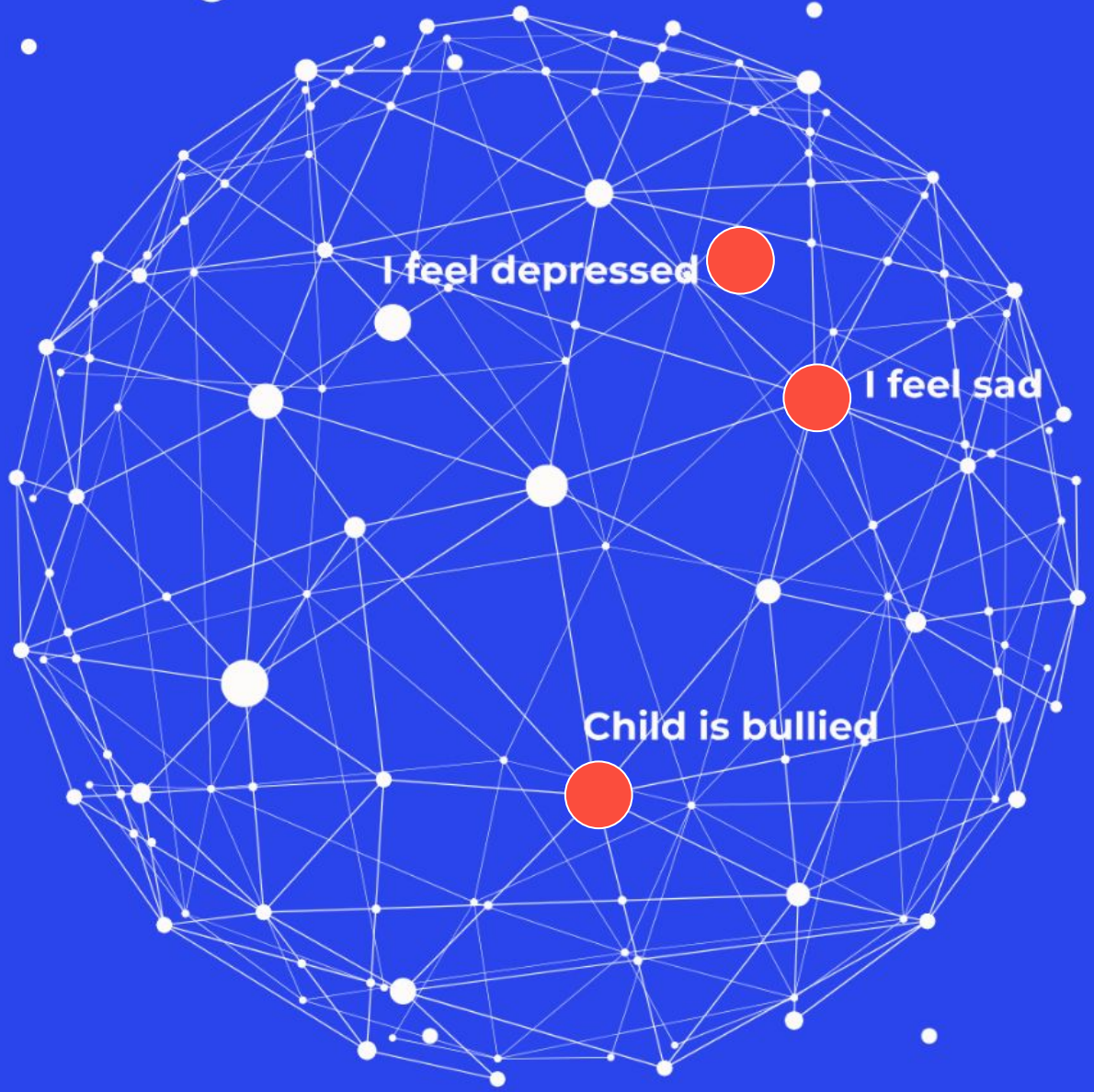
HARMONY

https://harmonydata.ac.uk/app

How does Harmony work?

1. Process the PDF to get the question items out
2. Convert all questions to sentence embeddings
3. Calculate the cosine similarity

https://fastdatascience.com/natural-language-processing/semantic-similarity-with-sentence-embeddings/



Sentences have a related meaning

Vectors point in a similar direction

Cosine score is high – close to 1

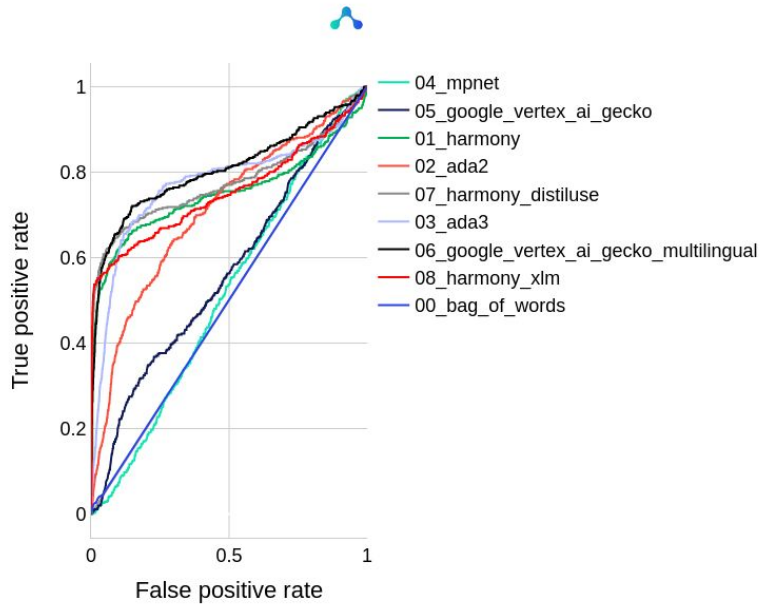
Sentences have unrelated meaning

Vectors are orthogonal

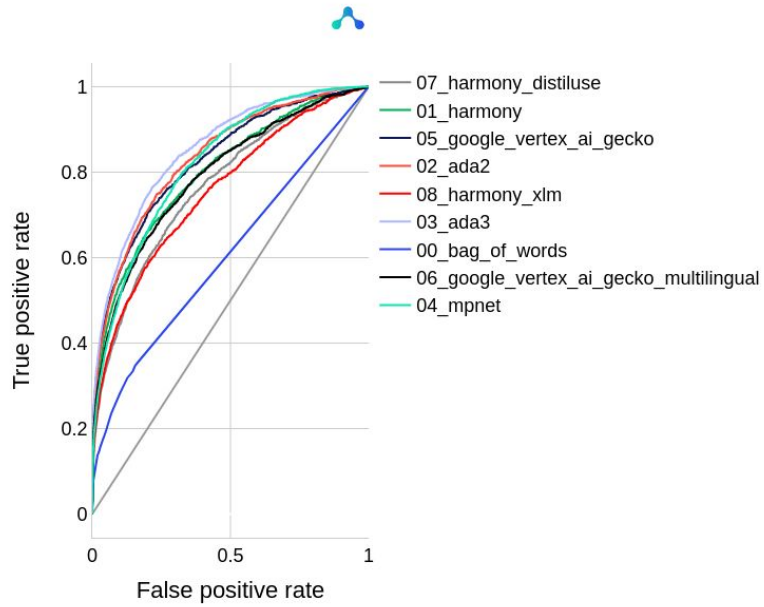
Cosine score is low – closer to 0

Evaluating Harmony

ROC on GAD 7 multilingual dataset



ROC on McElroy et al childhood dataset

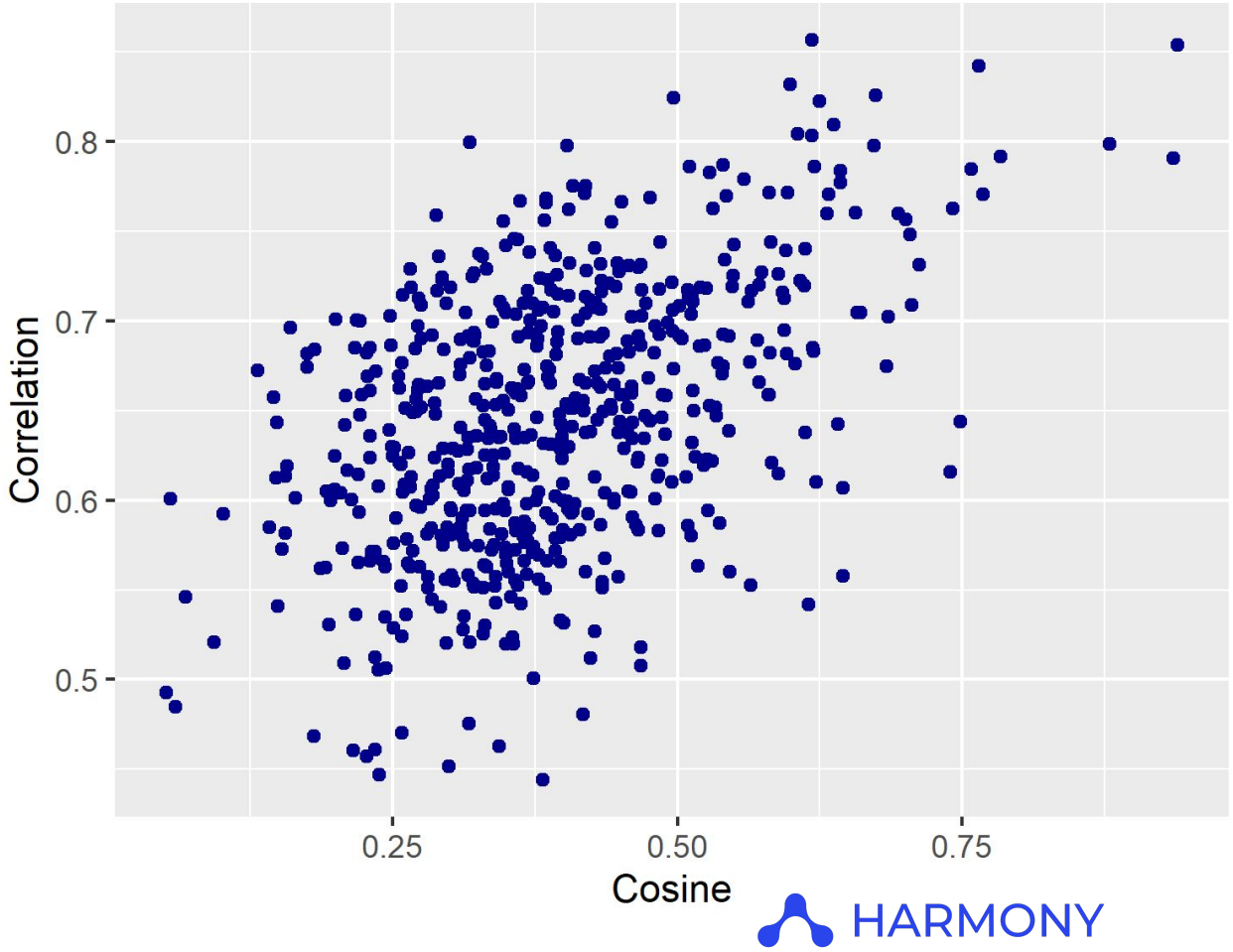


Real correlations?

	A	B	C
1	Questionnaire	Item number	Content
2	IDQ		
3	IDQ		
4	IDQ		
5	IDQ		
6	IDQ		
7	IDQ		
8	IDQ		
9	IDQ		
10	IDQ		
11	IAQ		
12	IAQ		
13	IAQ		
14	IAQ		
15	IAQ		
16	IAQ		
17	IAQ		
18	IAQ		

	A	B	C	D	E
1	Supplementary File 2. Correlaiton and cosine coefficients for item pairs				
2		from	to	spearman	cosine
3	1	1	10	0.719538559	0.61149627
4	2	1	11	0.719244021	0.445720732
5	3	1	12	0.731182941	0.711875081
6	4	1	13	0.665979411	0.571581244
7	5	1	14	0.703580795	0.511955619
8	6	1	15	0.709961188	0.297138691
9	7	1	16	0.691486691	0.50184983
10	8	1	17	0.608636306	0.259960353
11	9	1	18	0.738485659	0.37026453
12	10	1	19	0.798613012	0.879561961
13	11	1	2	0.824578169	0.496798843
14	12	1	20	0.573991369	0.572757006
15	13	1	21	0.628075543	0.655307412
16	14	1	22	0.625625193	0.354875147
17	15	1	23	0.718334673	0.681931853
18	16	1	24	0.683265023	0.334293664
19	17	1	25	0.587003714	0.53709048

Real correlations?



MCElroy et al,

Using natural language processing to facilitate the harmonization of mental health questionnaires: a validation study using real-world data

Harmony Discovery

https://harmonydiscovery.fastdatascience.com/

We are allowing researchers to discover datasets and studies by searching in a vector index

A giant vector index of UK studies across disciplines

- Currently using Elasticsearch (experimenting with Pinecone)

Designs: https://www.figma.com/...

Open source

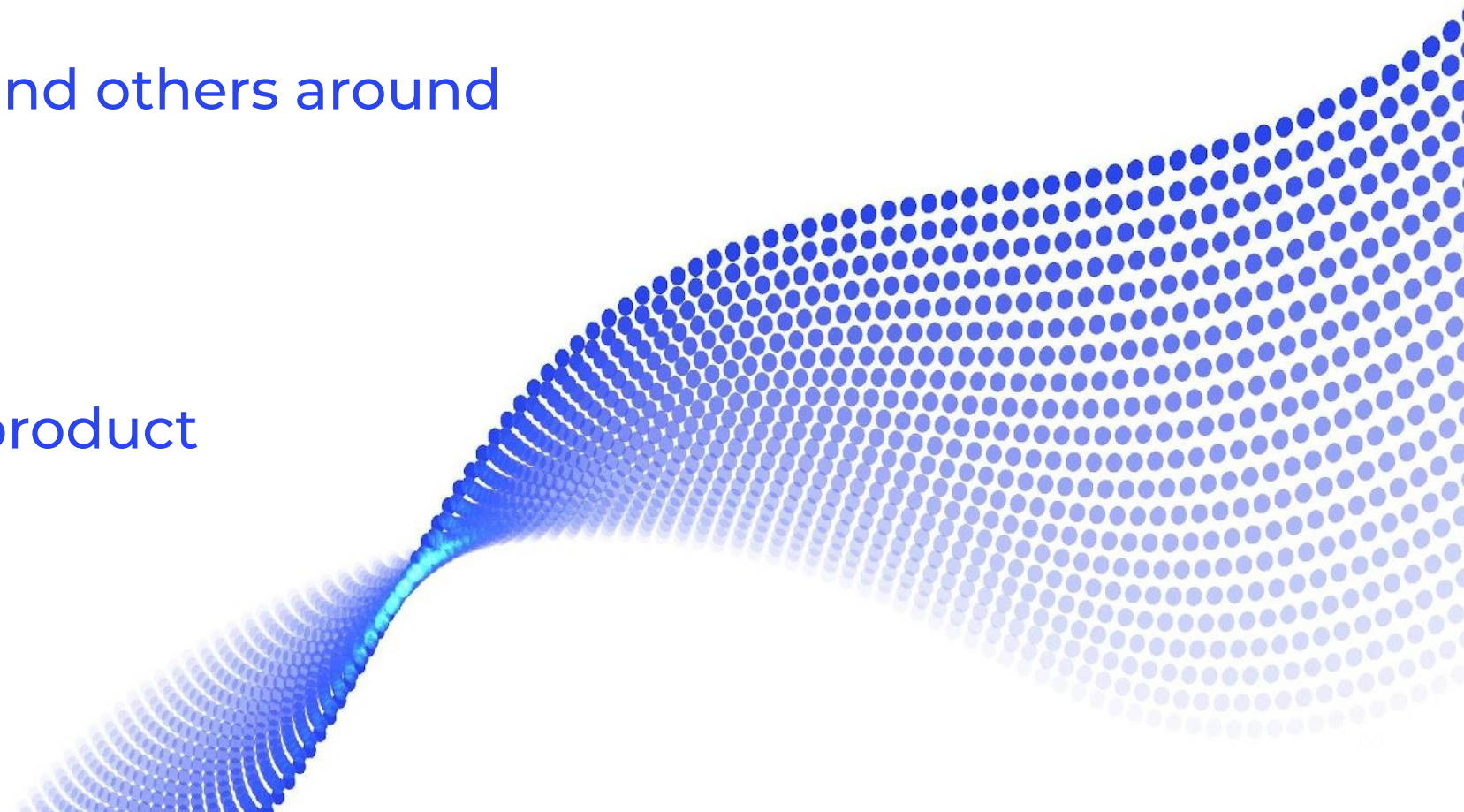
https://github.com/harmonydata

Free for researchers and others around the world

MIT License

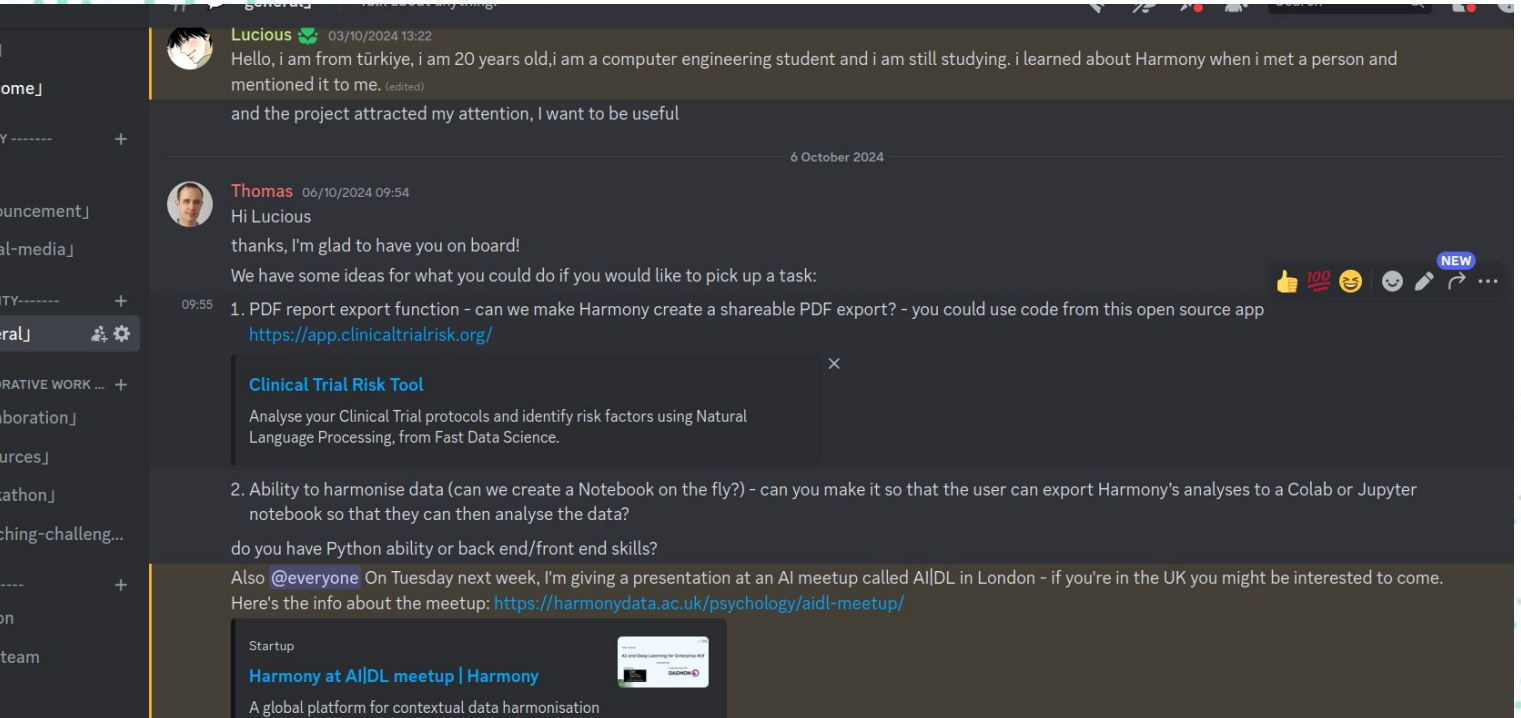
It’s not a monetised product

(online) hackathons



Discord

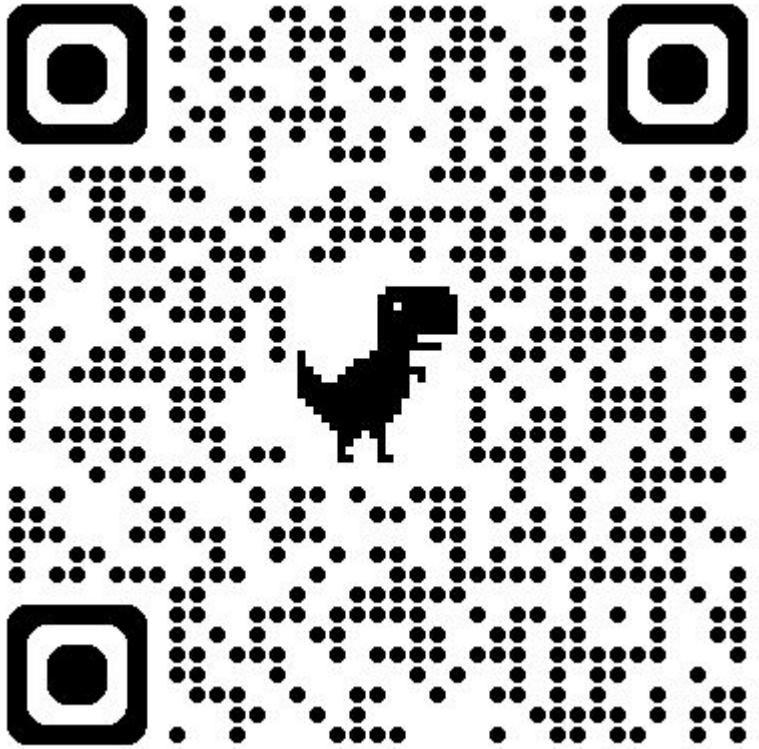
discord.gg/harmonydata



Open source tasks

We are always looking for collaborators.
Ideas include

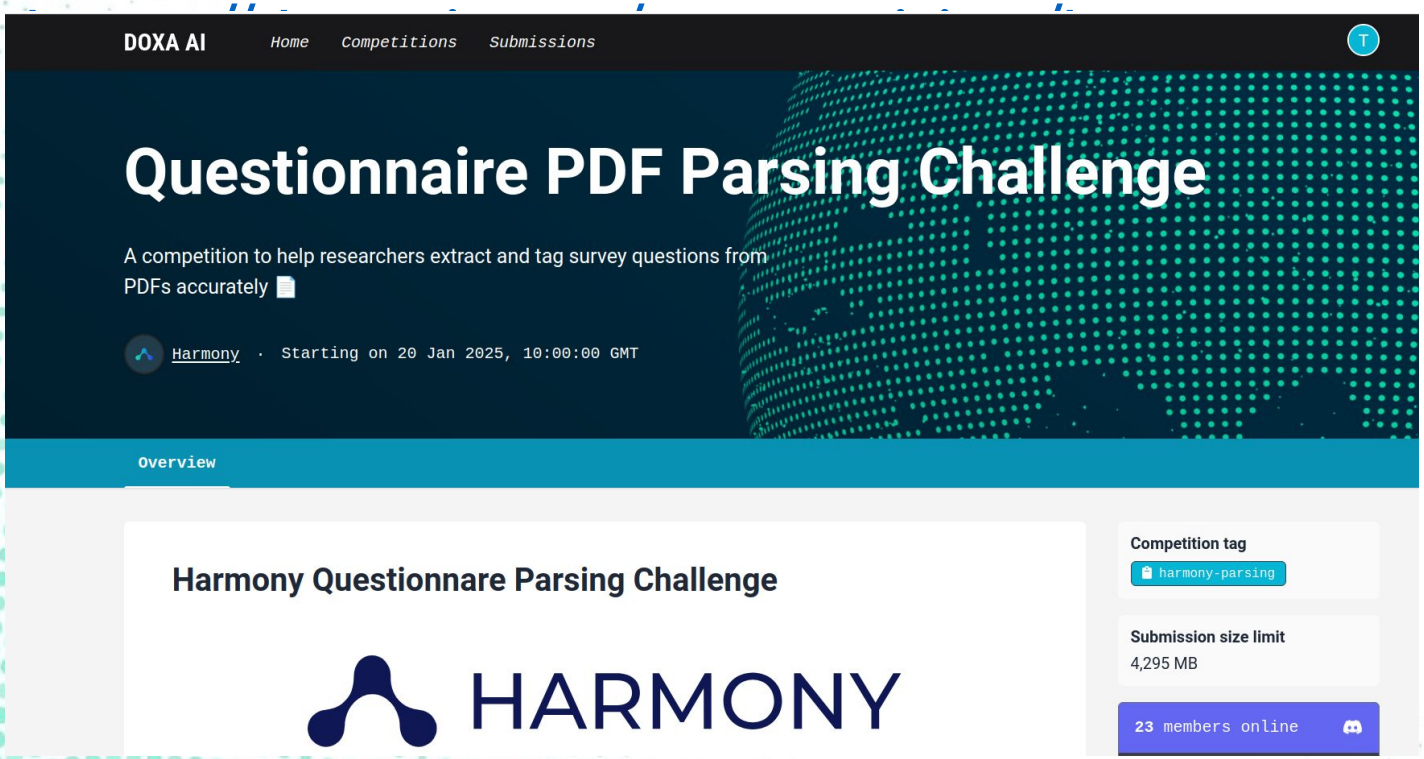
- Create browser plugin
- Integration with more platforms
- PDF report export function (can we make Harmony create a shareable PDF export?)
- Ability to harmonise data (can we create a Notebook on the fly?)



Search

#	Participant	Submitted	MAE	R^2
1	Alex @ayetsu	11 hours ago	20.683	0.14331
2	Daksh @Daksh	1 day ago	20.906	0.16552
3	Akshat @akshat	9 hours ago	21.198	0.08753
4	Jeremy @jezz	1 month ago	21.261	0.11211
5	Gabriele Monti @montig	2 weeks ago	21.629	0.04189
6	idle @Idlehunter	2 weeks ago	21.770	0.01900
7	Jake @jakedorman64	1 month ago	21.797	0.09679
8	Louis de Wardt @louisdewardt	3 weeks ago	21.797	0.09679
9	Denis @denis_chaykovskiy	2 weeks ago	21.797	0.09679
10	Rafi Ahmed Patel @rafa	6 days ago	21.797	0.09679
11	Thomas @thomas	1 month ago	24.666	-0.62098

Train PDF parser



HARMONY

Thank you for listening!

harmonydata.ac.uk

discord.gg/harmonydata

linkedin.com/company/harmonydata

github.com/harmonydata/harmony

@harmony_data

@harmony_data

b.moltrecht@ucl.ac.uk thomas@fastdatascience.com

HARMONY

Harmonise questionnaire items - with Harmony

Harmony is a tool for retrospective harmonisation of questionnaire items.

If you want to compare items from different surveys, such as GAD-7 and PHQ-9, Harmony can identify which questions match.

FAQs - Privacy policy

B

Upload or drag and drop any **pdf, docx or xlsx** file here

Choose from existing instruments

SCARED English, GAD-7 English, CES_D English

- SCARED English
- GAD-7 English
- CES_D English

HARMONISE

Options

Match Threshold

Search

Show within-instrument matches

SHARE

EXPORT

B

Found 8 matches

SCARED English - Q28	People tell me that I worry too much	81	GAD-7 English - Q3.0 Worrying too much about different things
GAD-7 English - Q7.0	Feeling afraid, as if something awful might happen	78	CES_D English - Q10 I felt fearful.
SCARED English - Q35	I worry about how well I do things	76	GAD-7 English - Q3.0 Worrying too much about different things
SCARED English - Q15	When I get frightened, I feel like things are not real	74	GAD-7 English - Q7.0 Feeling afraid, as if something awful might happen
SCARED English - Q6	When I get frightened, I feel like passing out	72	GAD-7 English - Q7.0 Feeling afraid, as if something awful might happen
SCARED English - Q9	People tell me that I look nervous	71	GAD-7 English - Q1.0 Feeling nervous, anxious, or on edge
SCARED English - Q12			GAD-7 English - Q7.0

Options

Match Threshold

Search

fam

Show within-instrument matches

SHARE

EXPORT

B

Found 113 matches

SCARED English - Q29 I don't like to be away from my family	46	CES_D English - Q8 I felt that I could not shake off the blues even with help from my family or friends.
SCARED English - Q29 I don't like to be away from my family	39	CES_D English - Q20 I could not get "going."
SCARED English - Q29 I don't like to be away from my family	38	CES_D English - Q19 I felt that people dislike me.
SCARED English - Q29 I don't like to be away from my family	37	CES_D English - Q1 I was bothered by things that usually don't bother me.
GAD-7 English - Q2.0 Not being able to stop or control worrying	36	CES_D English - Q8 I felt that I could not shake off the blues even with help from my family or friends.
SCARED English - Q29 My child doesn't like to be away from his/her family	34	CES_D English - Q8 I felt that I could not shake off the blues even with help from my family or friends.
SCARED English - Q31		CES_D English - Q8