Crowdsourcing Historical Tabular Data – 1961 Census of England and Wales

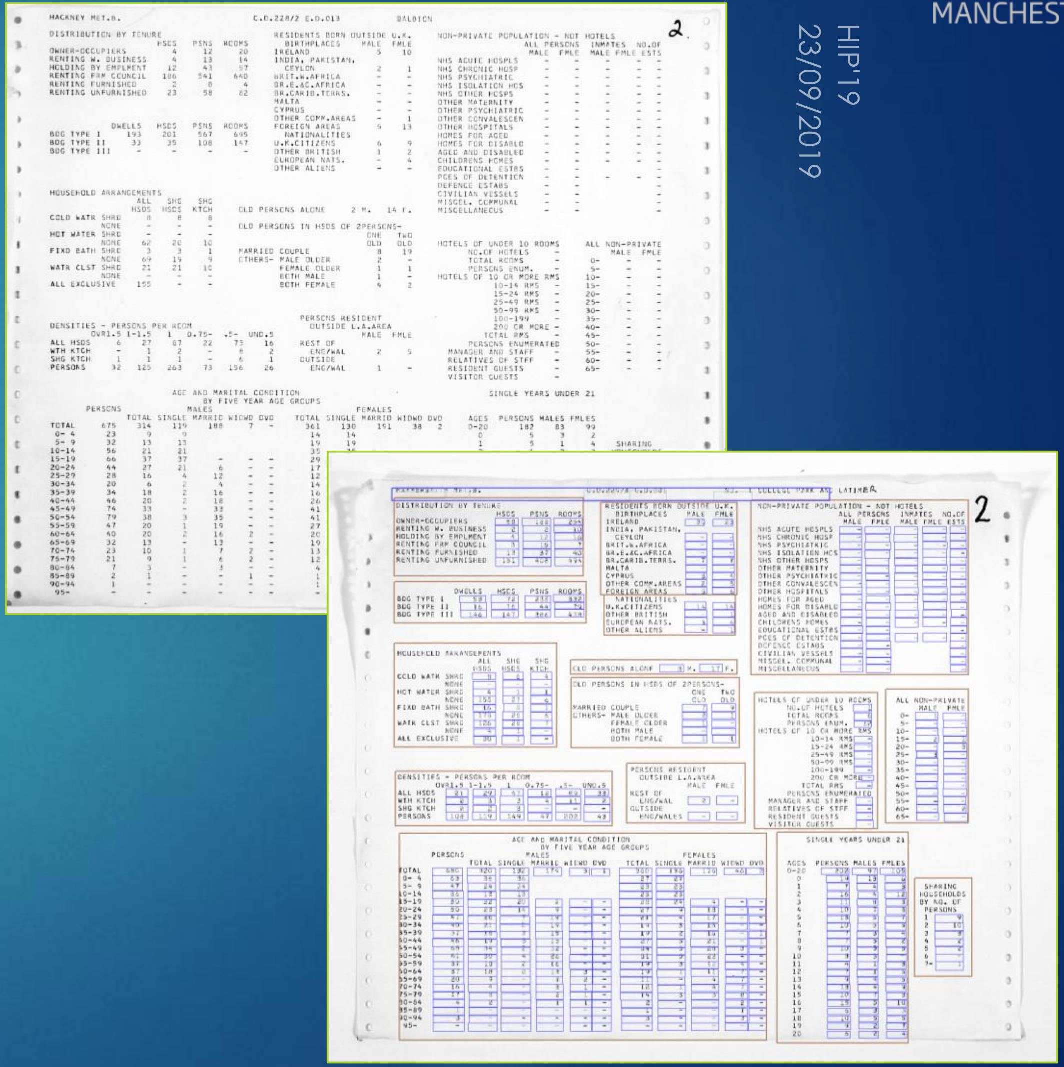
CHRISTIAN CLAUSNER, JUSTIN HAYES AND APOSTOLOS ANTONACOPOULOS

PATTERN RECOGNITION AND IMAGE ANALYSIS RESEARCH LAB, UK

HIP'19, SYDNEY, AUSTRALIA

The 1961 Census Digitisation Project

- Millions of data items trapped in 100,000+ pages (tables)
- Main part of project in 2018/2019
  - For Office for National Statistics
- Automated processing pipeline
  - About 98% correct results
  - Requires post-correction
- Two other publications, this one is an experience paper concentrating on the crowdsourcing aspects



HACKNEY MET.B. C.D.228/2 E.D.013 DALSTON 2.

DISTRIBUTION BY TENURE

	HSDS	PSNS	ROOMS
OWNER-OCCUPIERS	4	12	20
RENTING W. BUSINESS	4	13	14
HOLDING BY EMPLMENT	12	43	57
RENTING FRM COUNCIL	106	541	640
RENTING FURNISHED	2	8	4
RENTING UNFURNISHED	23	58	82

	DWELLS	HSDS	PSNS	ROOMS
BDG TYPE I	193	201	567	695
BDG TYPE II	33	35	108	147
BDG TYPE III	-	-	-	-

RESIDENTS BORN OUTSIDE U.K.

BIRTHPLACES	MALE	FMLE
IRELAND	5	10
INDIA, PAKISTAN, CEYLON	2	1
BRIT.W.AFRICA	-	-
BR.E.&C.AFRICA	-	-
BR.CARIB.TERRS.	-	-
MALTA	-	-
CYPRUS	-	-
OTHER COMM.AREAS	-	1
FOREIGN AREAS	5	13
NATIONALITIES		
U.K.CITIZENS	6	9
OTHER BRITISH	1	2
EUROPEAN NATS.	-	4
OTHER ALIENS	-	-

NON-PRIVATE POPULATION - NOT HOTELS

	ALL PERSONS		INMATES		NO.OF
	MALE	FMLE	MALE	FMLE	ESTS
NHS ACUTE HOSPLS	-	-	-	-	-
NHS CHRONIC HOSP	-	-	-	-	-
NHS PSYCHIATRIC	-	-	-	-	-
NHS ISOLATION HOS	-	-	-	-	-
NHS OTHER HOSPS	-	-	-	-	-
OTHER MATERNITY	-	-	-	-	-
OTHER PSYCHIATRIC	-	-	-	-	-
OTHER CONVALESCEN	-	-	-	-	-
OTHER HOSPITALS	-	-	-	-	-
HOMES FOR AGED	-	-	-	-	-
HOMES FOR DISABLD	-	-	-	-	-
AGED AND DISABLED	-	-	-	-	-
CHILDRENS HOMES	-	-	-	-	-
EDUCATIONAL ESTBS	-	-			-
PCES OF DETENTION	-	-	-	-	-
DEFENCE ESTABS	-	-			-
CIVILIAN VESSELS	-	-			-
MISCEL. COMMUNAL	-	-			-
MISCELLANEOUS	-	-			-

HOUSEHOLD ARRANGEMENTS

		ALL HSDS	SHG HSDS	SHG KTCH
COLD WATR	SHRD	8	8	8
	NONE	-	-	-
HOT WATER	SHRD	-	-	-
	NONE	62	20	10
FIXD BATH	SHRD	3	3	1
	NONE	69	19	9
WATR CLST	SHRD	21	21	10
	NONE	-	-	-
ALL EXCLUSIVE		155	-	-

OLD PERSONS ALONE 2 M. 14 F.

OLD PERSONS IN HSDS OF 2PERSONS-

		ONE OLD	TWO OLD
MARRIED COUPLE		8	19
OTHERS-	MALE OLDER	2	-
	FEMALE OLDER	1	1
	BOTH MALE	1	-
	BOTH FEMALE	4	2

HOTELS OF UNDER 10 ROOMS	
NO.OF HOTELS	-
TOTAL ROOMS	-
PERSONS ENUM.	-
HOTELS OF 10 OR MORE RMS	
10-14 RMS	-
15-24 RMS	-
25-49 RMS	-
50-99 RMS	-
100-199	-
200 OR MORE	-
TOTAL RMS	-
PERSONS ENUMERATED	
MANAGER AND STAFF	-
RELATIVES OF STFF	-
RESIDENT GUESTS	-
VISITOR GUESTS	-

ALL NON-PRIVATE

	MALE	FMLE
0-	-	-
5-	-	-
10-	-	-
15-	-	-
20-	-	-
25-	-	-
30-	-	-
35-	-	-
40-	-	-
45-	-	-
50-	-	-
55-	-	-
60-	-	-
65-	-	-

DENSITIES - PERSONS PER ROOM

	OVR1.5	1-1.5	1	0.75-	.5-	UND.5
ALL HSDS	6	27	87	22	73	16
WTH KTCH	-	1	2	-	8	2
SHG KTCH	1	1	1	-	6	1
PERSONS	32	125	263	73	156	26

PERSONS RESIDENT OUTSIDE L.A.AREA

	MALE	FMLE
REST OF ENG/WAL	2	5
OUTSIDE ENG/WAL	1	-

AGE AND MARITAL CONDITION BY FIVE YEAR AGE GROUPS

	PERSONS	MALES					FEMALES				
		TOTAL	SINGLE	MARRID	WIDWD	DVD	TOTAL	SINGLE	MARRID	WIDWD	DVD
TOTAL	675	314	119	188	7	-	361	130	151	38	2
0- 4	23	9	9				14	14			
5- 9	32	13	13				19	19			
10-14	56	21	21				35				
15-19	66	37	37	-	-	-	29				
20-24	44	27	21	6	-	-	17				
25-29	28	16	4	12	-	-	12				
30-34	20	6	2	4	-	-	14				
35-39	34	18	2	16	-	-	16				
40-44	46	20	2	18	-	-	26				
45-49	74	33	-	33	-	-	41				
50-54	79	38	3	35	-	-	41				
55-59	47	20	1	19	-	-	27				
60-64	40	20	2	16	2	-	20				
65-69	32	13	-	13	-	-	19				
70-74	23	10	1	7	2	-	13				
75-79	21	9	1	6	2	-	12				
80-84	7	3	-	3	-	-	4				
85-89	2	1	-	-	1	-	1				
90-94	1	-	-	-	-	-	1				
95-	-	-	-	-	-	-	-				

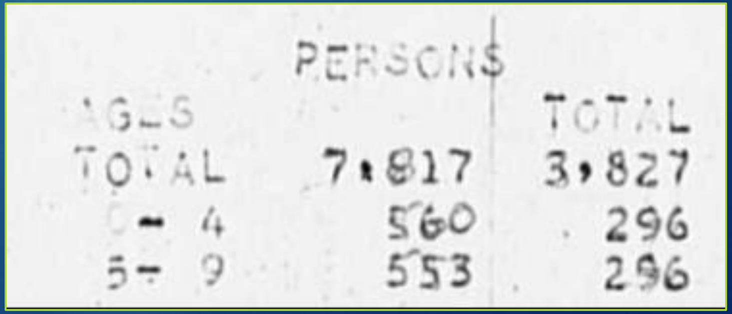
SINGLE YEARS UNDER 21

AGES	PERSONS	MALES	FMLES
0-20	182	83	99
0	5	3	2
1	5	1	4

SHARING

Challenges

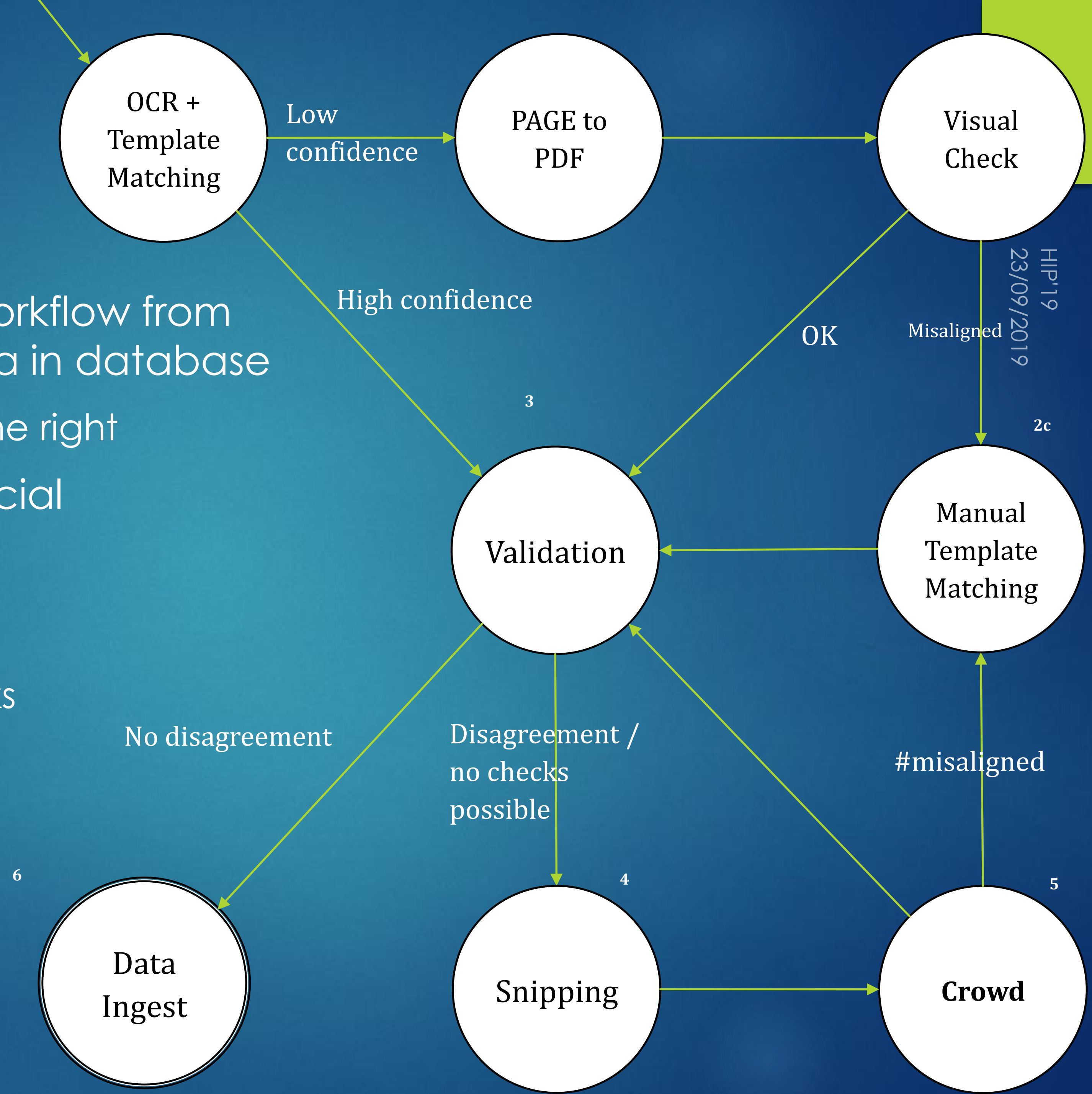
- Inconsistent scan quality (illumination, warping, skew, scaling, placement)
- Faint print, handwritten corrections
- Microfilm scratches and general degradation
- Missing parts, printing errors
- Unorganised data (pages not in any particular order)
- Dense tables, sometimes with no separation between columns



	PERSONS	
AGES		TOTAL
TOTAL	7,817	3,827
0- 4	560	296
5- 9	553	296

Workflow

- Complete digitisation workflow from image to structured data in database
  - Simplified workflow in the right
- Validation of data is crucial
- Identify errors by
  - Visual checks
  - Automated crosschecks
- Manual intervention
  - In part in-house
  - Mostly by crowd



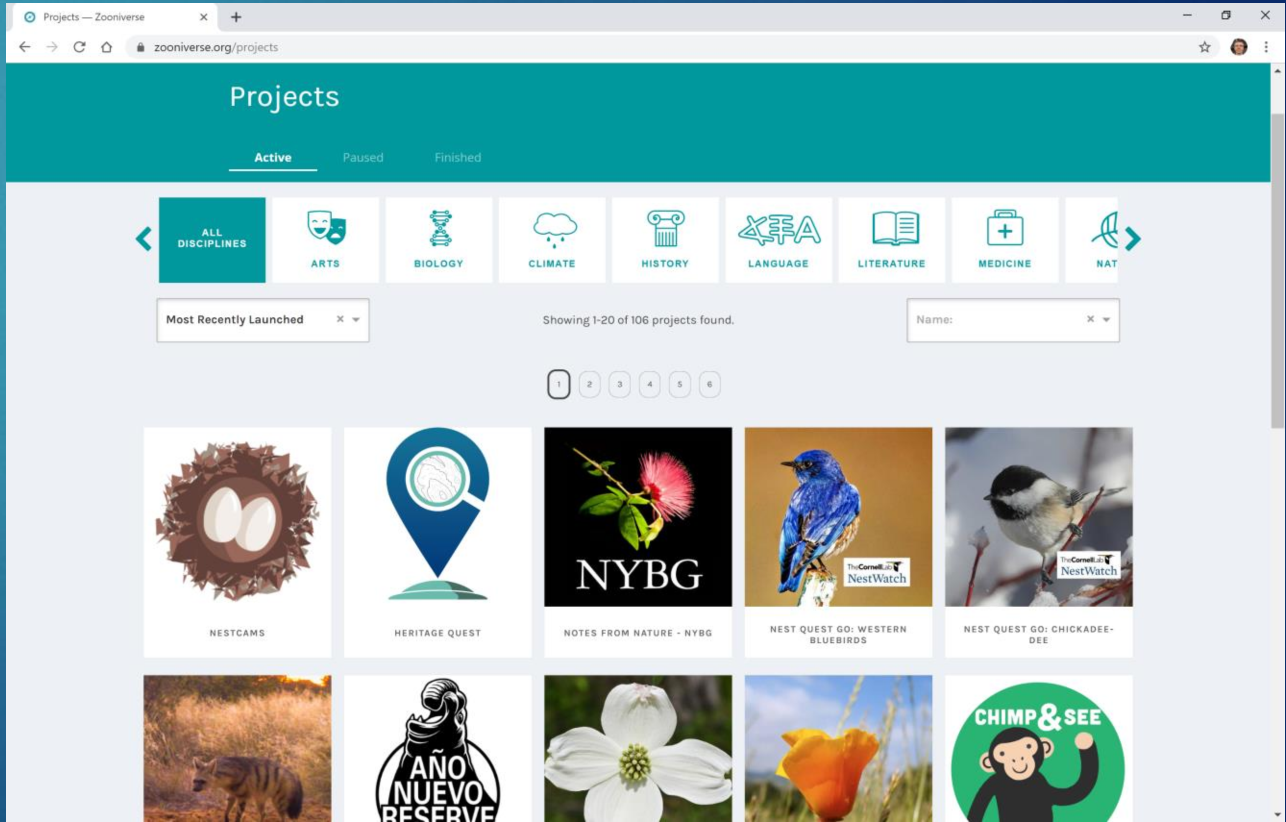
HIP'19 23/09/2019

Zooniverse

- We used Zooniverse for crowdsourcing
- Public platform (also open source)
- Big base of volunteers
- Free for projects that benefit the public good
- Easy to use
- Good support

HIP'19
23/09/2019

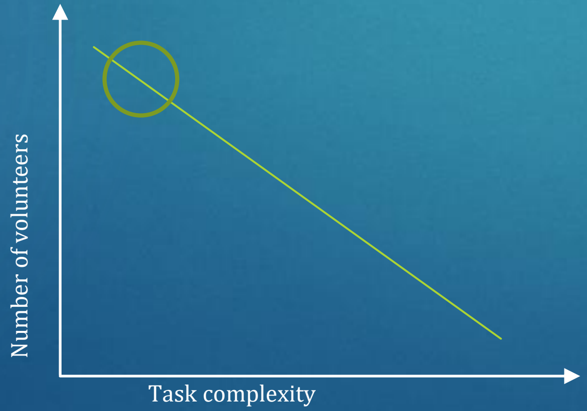
https://www.zooniverse.org/

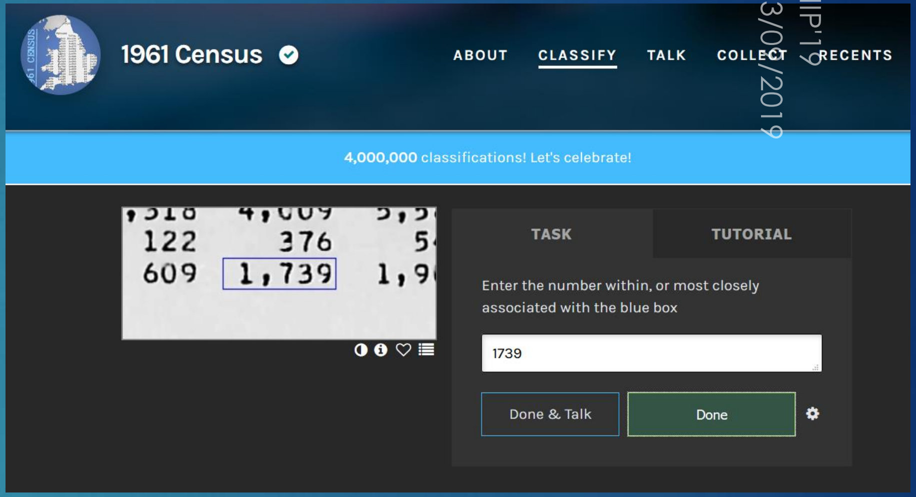


Showing 1-20 of 106 projects found.

Micro Tasks

- Task for volunteers as simple as possible
- "Enter text for highlighted table cell"
  - We don't even show the OCR result
- Problematic or unclear cases can be tagged (Talk section with hashtags)

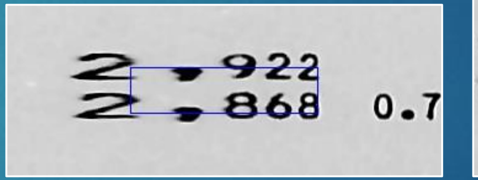


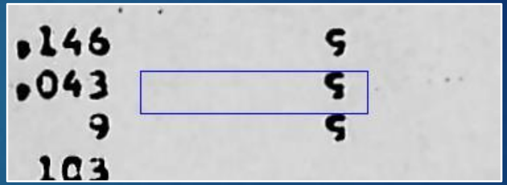

1961 Census

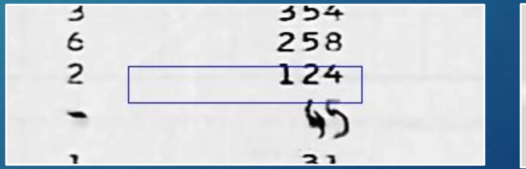
4,000,000 classifications! Let's celebrate!

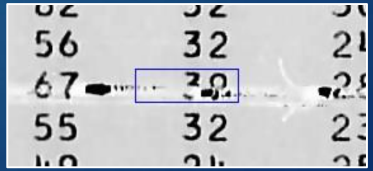
Enter the number within, or most closely associated with the blue box

1739


2 922
2 868
0.7

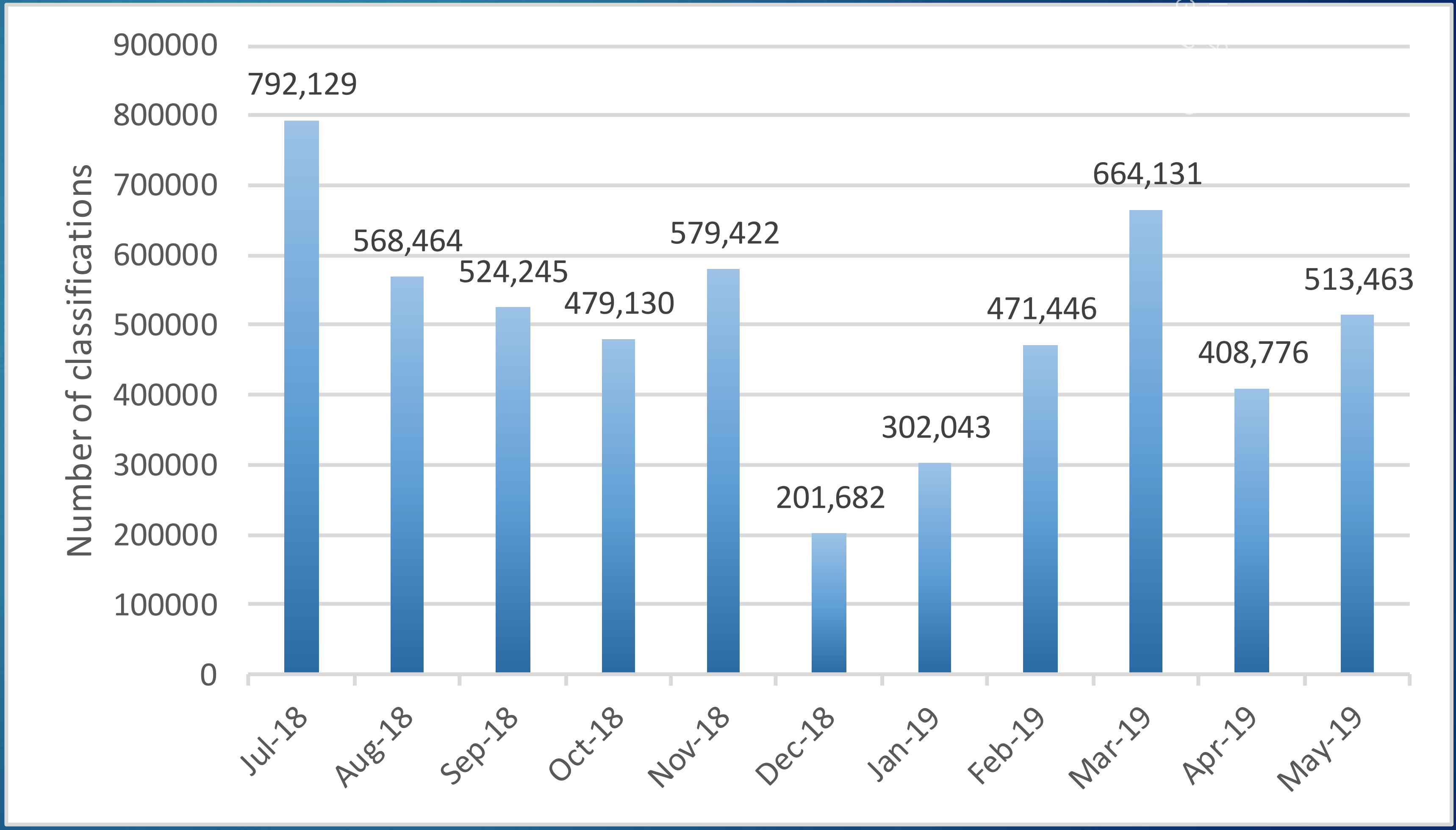

146
043
9
103

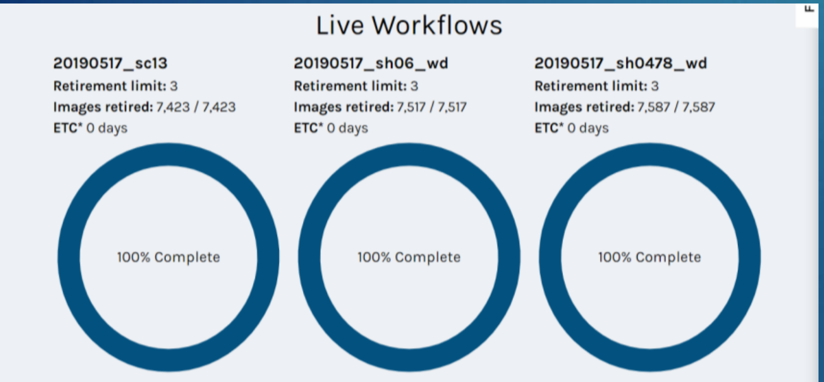



56 32
67 30
55 32

Census Zooniverse Project

- One of the most active projects in the time period
- No promotion
- Difficult to provide enough data



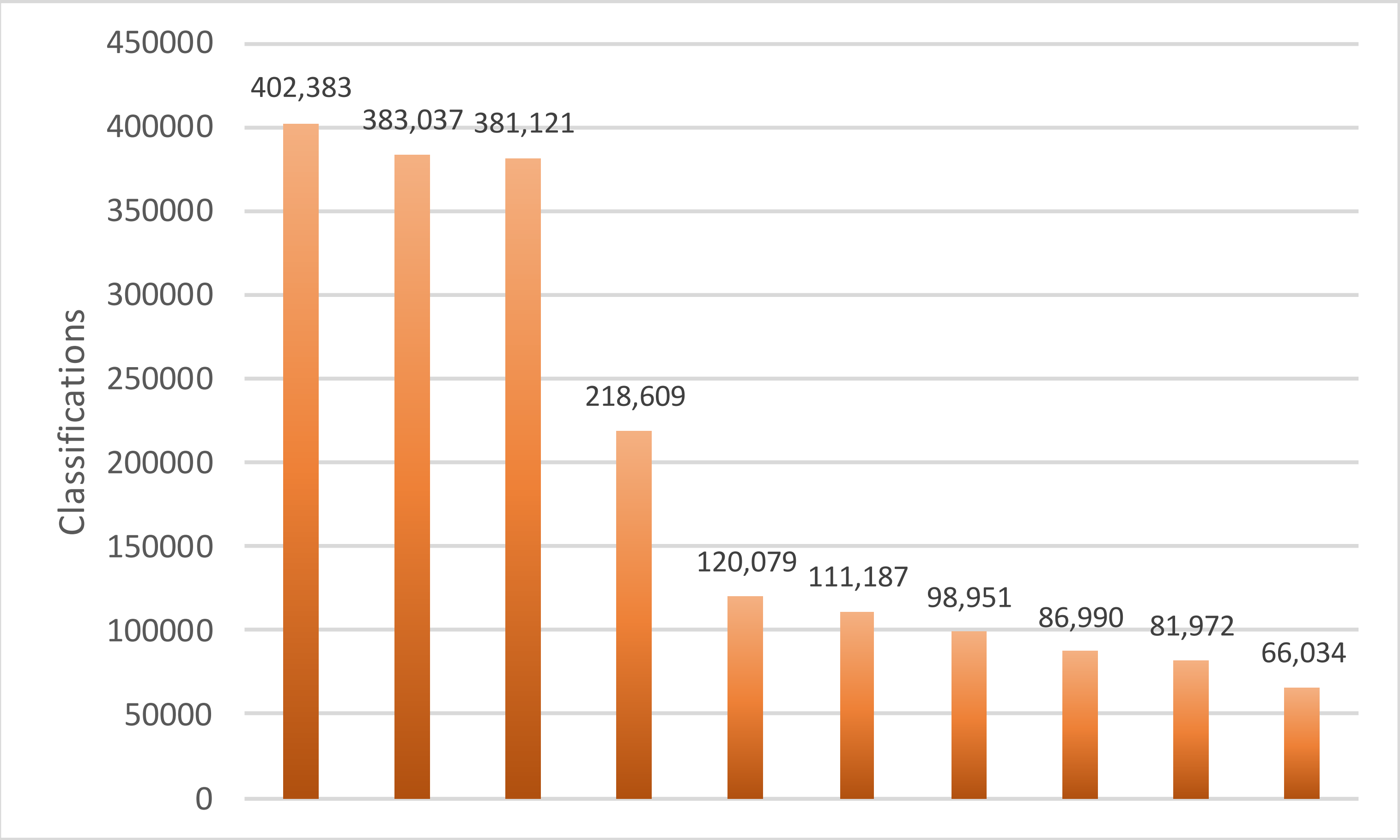

Live Workflows

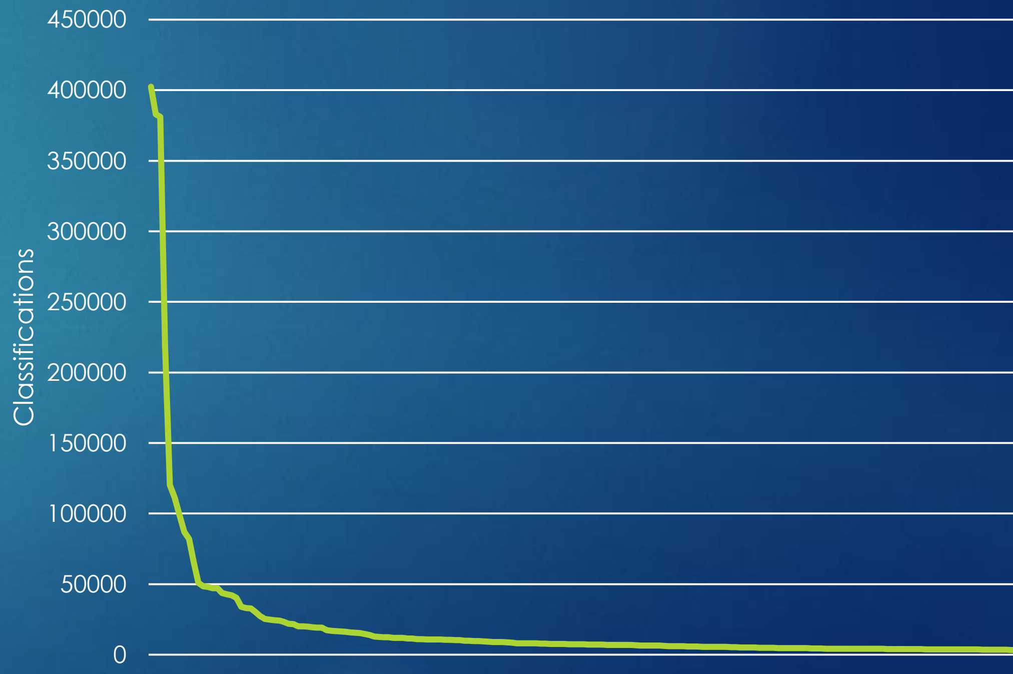
20190517_sc13
Retirement limit: 3
Images retired: 7,423 / 7,423
ETC* 0 days
100% Complete

20190517_sh06_wd
Retirement limit: 3
Images retired: 7,517 / 7,517
ETC* 0 days
100% Complete

20190517_sh0478_wd
Retirement limit: 3
Images retired: 7,587 / 7,587
ETC* 0 days
100% Complete

User Activity





University of Salford MANCHESTER

Great Participation

- Large user base with auto-promotion of new/active/stagnant projects on Zooniverse
- High interest in historical projects (and UK)
- Micro-tasking (mindfulness?)
- User engagement
- Consistency in data provision
- Power users (special attention)

Discussion

- Crowdsourcing was very successful for the Census 1961 project
- Accuracies
  - OCR about 98%
  - Cell recognition in total about 95%
  - Correctness after crowdsourcing about 99.5%
    - Rest corrected by expert

Problems

- Malicious users
  - Needs vigilance from our side
  - Can be blocked from Zooniverse side
- Bugs in the Zooniverse platform
  - We had a nasty one where text entered by users was incomplete
  - Fast fix
- Problems with data upload at busy times
  - Need to work around it

Conclusion

- Worth it
- Over 5 million corrections in a few months
- Volunteers liked it (even demanded more data)
- Possibly more to come in near future

Questions?

- zooniverse.org/projects/dataliberation/1961-census
- primaresearch.org/publications