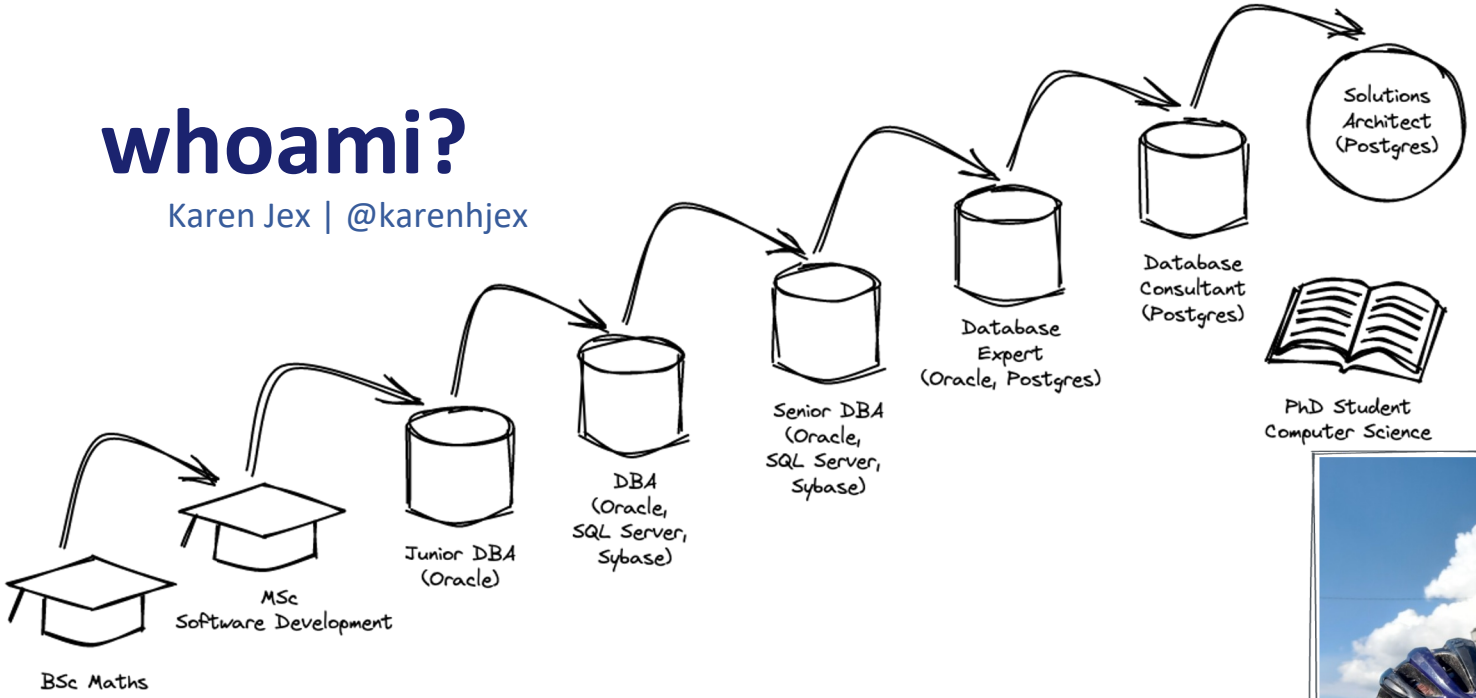
Unlocking Open Data using an Open Source Database

Karen Jex | Senior Solutions Architect @ Crunchy Data

PG Conf EU, Berlin | October 2022

whoami?

Karen Jex | @karenhjex



Agenda

- Choice of PhD
- Introduction to Open Data
- Challenges of Table Extraction
- Research Approach
- Existing Methods
- Genetic Algorithm Example
- Research Impact
- Plans for the Future

Agenda

- **Choice of PhD**
- Introduction to Open Data
- Challenges of Table Extraction
- Research Approach
- Existing Methods
- Genetic Algorithm Example
- Research Impact
- Plans for the Future

Possibilities
Possibility
Possibility
Possibility
Possibility
Possibility

Choosing the PhD

“Unlocking Open Data through Wrapper Generation”

- findaphd.com
- Department of Computer Science, University of Manchester
- Professor Norman Paton
- Part time & Remote

Agenda

- Choice of PhD
- **Introduction to Open Data**
- Challenges of Table Extraction
- Research Approach
- Existing Methods
- Genetic Algorithm Example
- Research Impact
- Plans for the Future

Open Data Definition

https://en.wikipedia.org/wiki/Open_data

Open data is data that is openly accessible, exploitable, editable and shared by anyone for any purpose, even commercially. Open data is licensed under an open license … The growth of the open data movement is paralleled by a rise in intellectual property rights.

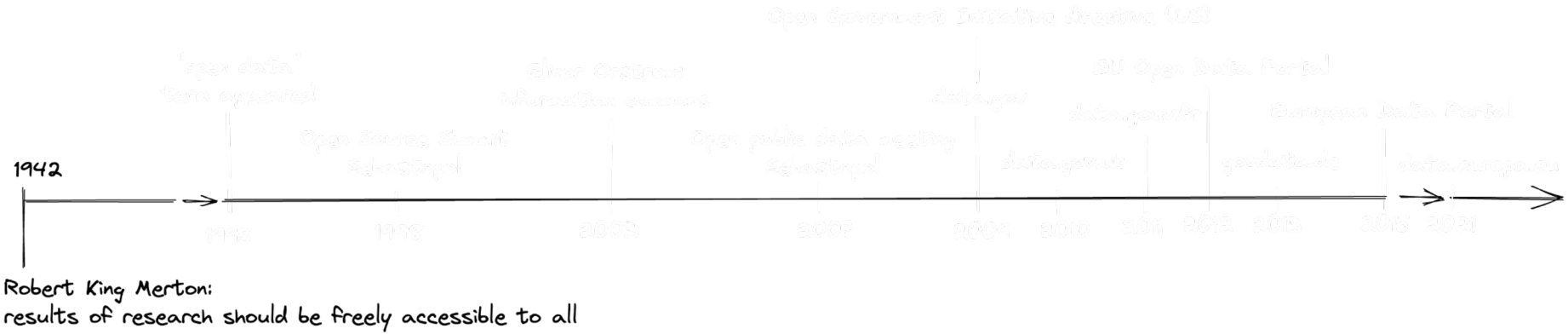
The philosophy behind open data has been long established … but the term "open data" itself is recent, gaining popularity with the rise of the Internet and World Wide Web and, especially, with the launch of open-data government initiatives such as Data.gov, Data.gov.uk and Data.gov.in …

One of the most important forms of open data is open government data (OGD), which is a form of open data created by ruling government institutions. Open government data's importance is born from it being a part of citizens' everyday lives, down to the most routine/mundane tasks that are seemingly far removed from government.

The abbreviation FAIR/O data is sometimes used to indicate that the dataset or database in question complies with the principles of FAIR data and also carries an explicit data-capable open license.

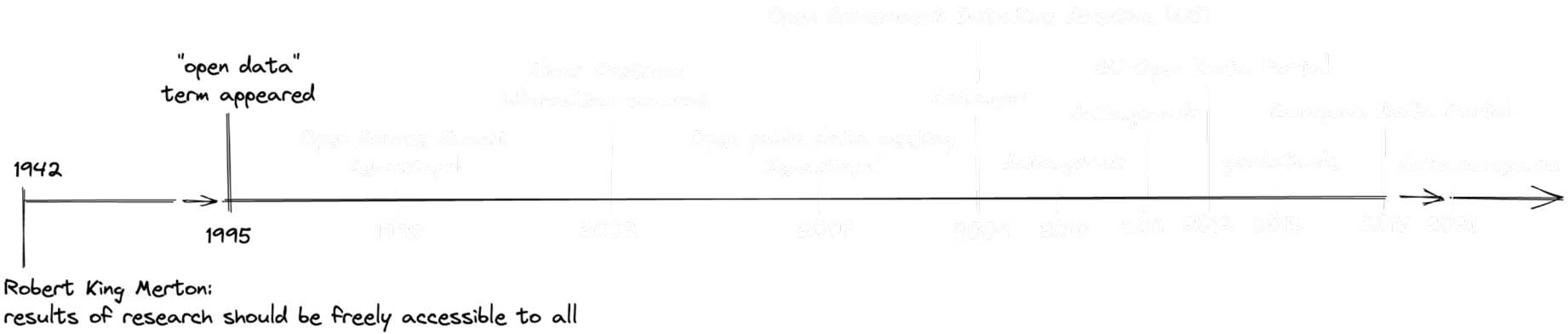
Open Data History

http://parisinnovationreview.com/articles-en/a-brief-history-of-open-data



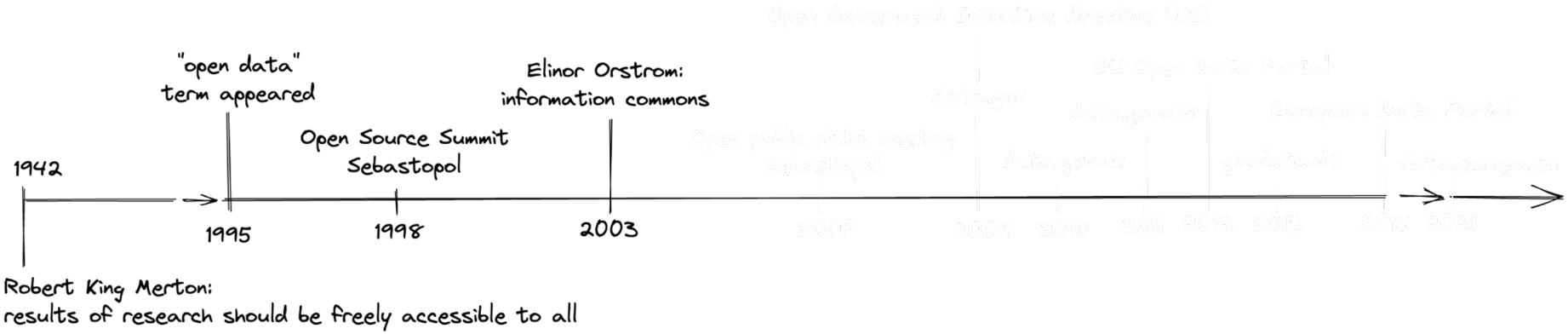
Open Data History

http://parisinnovationreview.com/articles-en/a-brief-history-of-open-data



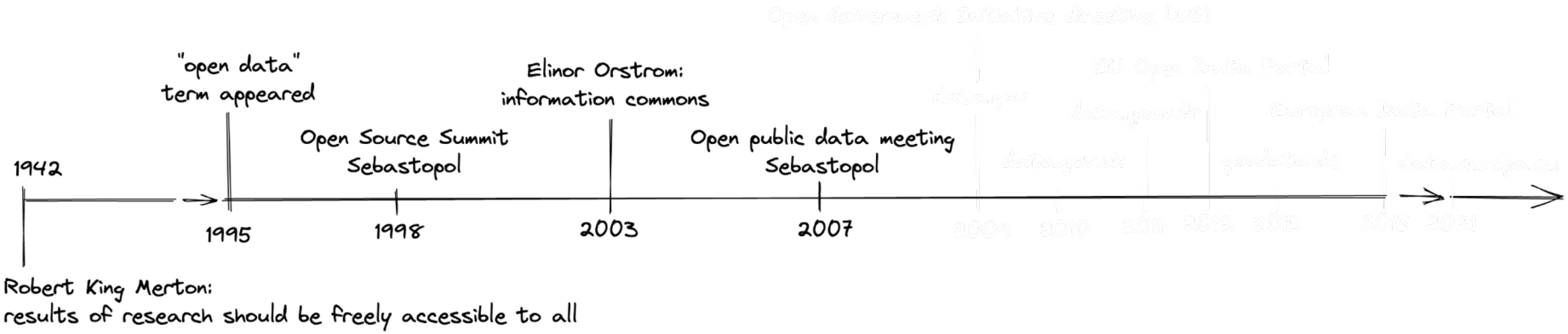
Open Data History

http://parisinnovationreview.com/articles-en/a-brief-history-of-open-data



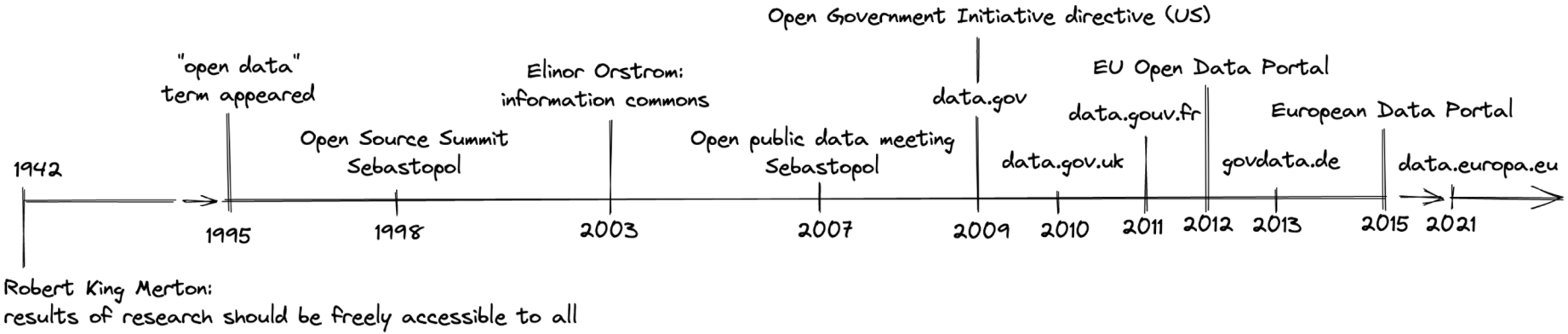
Open Data History

http://parisinnovationreview.com/articles-en/a-brief-history-of-open-data

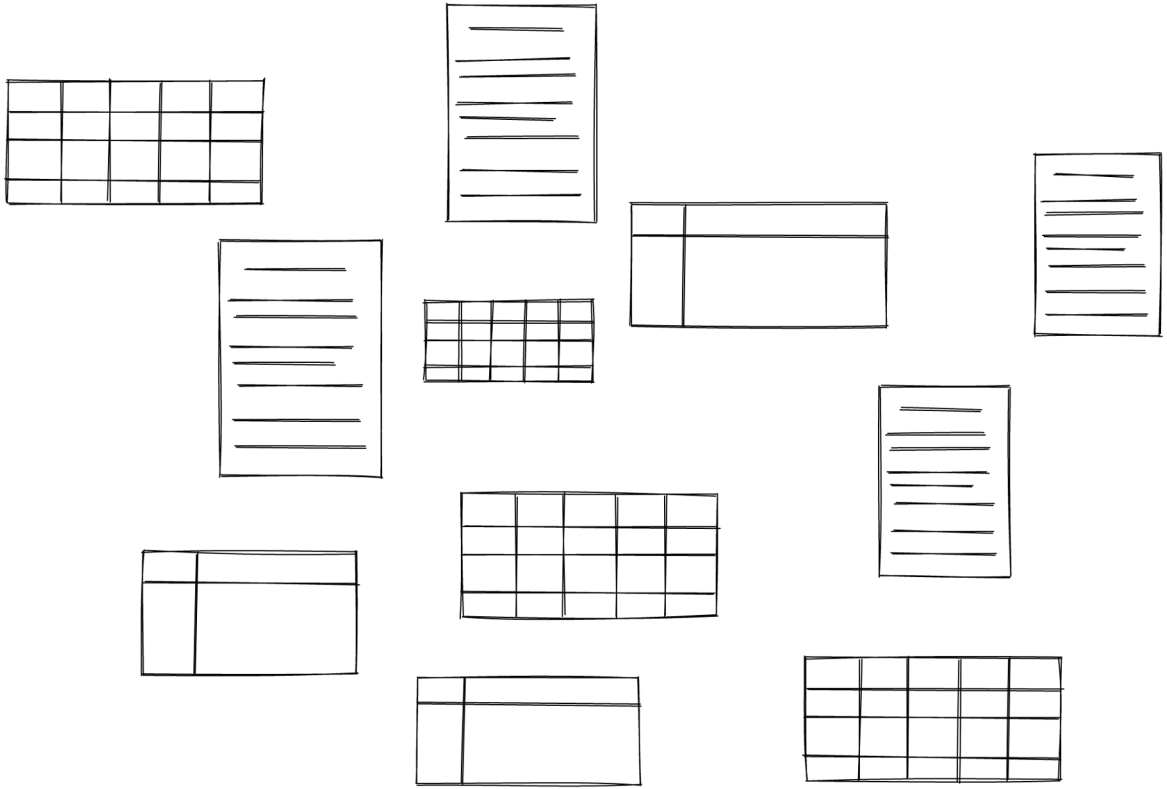


Open Data History

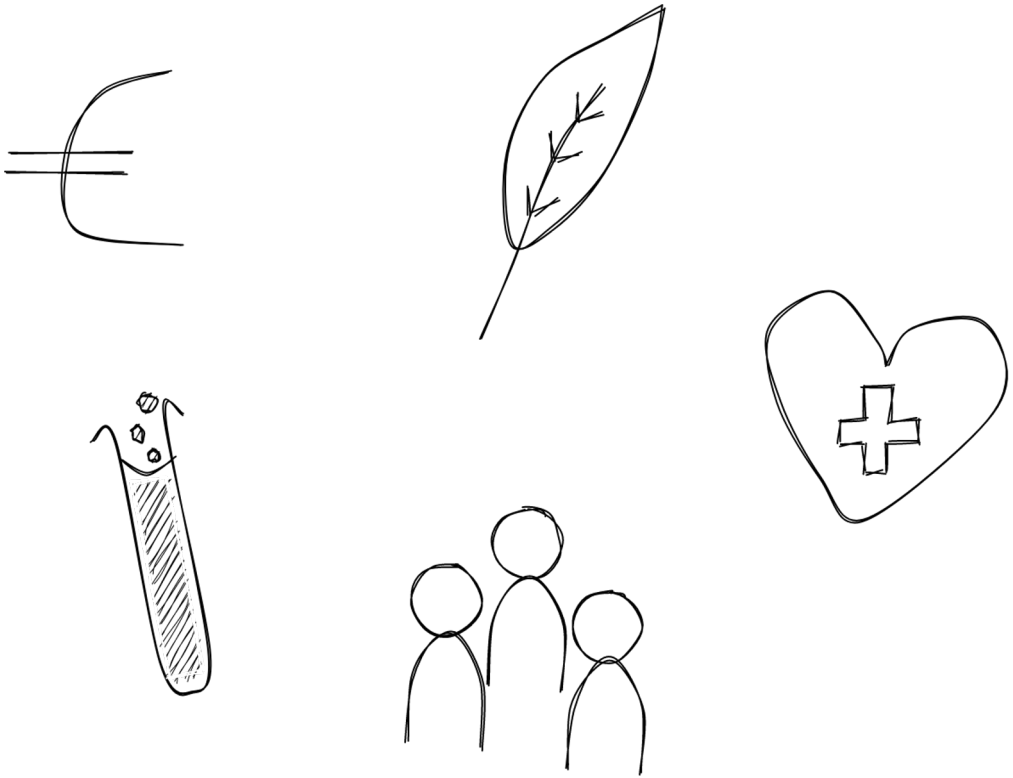
http://parisinnovationreview.com/articles-en/a-brief-history-of-open-data



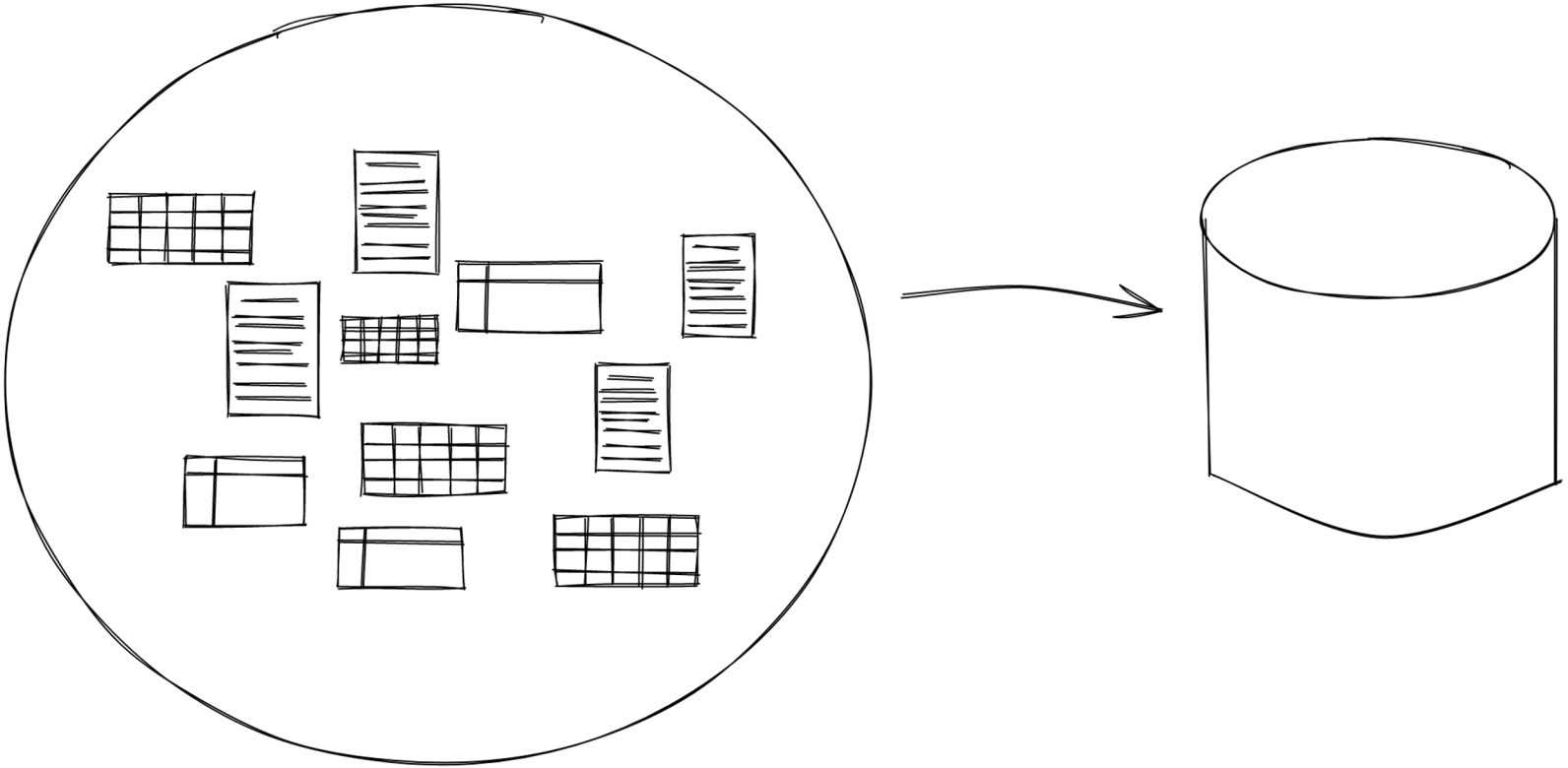
Unlocking Open Data through Wrapper Generation



Unlocking Open Data through Wrapper Generation



Unlocking Open Data through Wrapper Generation



Open Data Sources

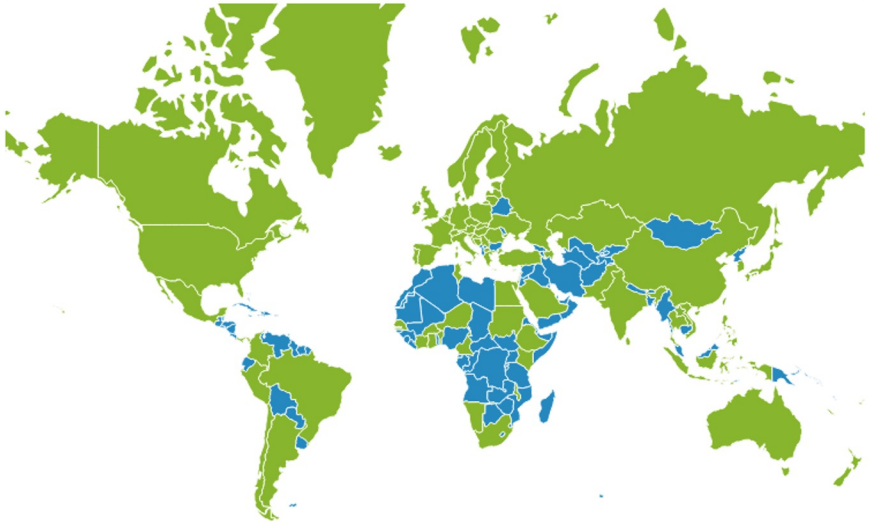
- World Bank Open Data
- WHO (World Health Organization)
- Google Public Data Explorer
- Registry of Open Data on AWS (RODA)
- European Union Open Data Portal
- U.S. Census Bureau
- Data.gov
- UNICEF Dataset
- Data.gov.uk
- Open Data Network
- NASA Planetary Data System
- NASA Earth Data
- FiveThirtyEight
- Kaggle

Registry of Research Data Repositories

https://www.re3data.org/

Browse by country

Graphical Text



Browse by subject

Graphical Text

click to zoom into subjects or to select a bottommost subject in the hierachy as filter for the re3data search page
shift + click on a top subject to select it as filter

Open Data Formats

- ZIP
- CSV
- GEOJSON
- PDF
- HTML
- XLS
- JSON
- SHP
- KML
- TXT

Agenda

- Choice of PhD
- Introduction to Open Data
- **Challenges of Table Extraction**
- Research Approach
- Existing Methods
- Genetic Algorithm Example
- Research Impact
- Plans for the Future

Challenges Presented by CSV Files

"While current approaches for automatic table discovery focus on spreadsheet and HTML files, CSV files pose a different challenge due to their lack of rich embedded meta-data such as cell formatting, table styling, and formulas."

Christodoulakis, C., Munson, E. B., Gabel, M., Brown, A. D., & Miller, R. J. (2020). **Pytheas: Pattern based table discovery in CSV files**. *Proceedings of the VLDB Endowment*, *13*(11), 2075–2089.

RFC4180 https://www.rfc-editor.org/rfc/rfc4180

Common Format and MIME Type for Comma-Separated Values (CSV) Files

Optional header as first line

Same number of fields (separated by a comma) in each line

Fields optionally enclosed in double quotes

```
field_name,field_name,field_name CRLF
aaa,bbb,ccc CRLF
"aaa","bbb","cc""c" CRLF
"aaa","b CRLF
bb","ccc" CRLF
zzz,yyy,xxx
```

One record per line, delimited by line break

Double quote inside field escaped by double quote

Last line may or may not have line break

Fields containing line breaks, double quotes, commas enclosed in double-quotes

RFC4180: What it doesn’t specify

- Date format

dd-mon-yyyy yyyy-mm-dd mm/dd/yyyy dd mm yyyy

- Encoding

UTF-8 UTF-16 UTF32 ANSI ASCII

Open Data Example: Simple CSV

https://www.kaggle.com/datasets/johnharshith/world-happiness-report-2021-worldwide-mortality

	A	B	C	D	E	F	G	H	I	J	K
1	Country name	Regional indicator	Ladder score	Standard error of ladder score	upperwhisker	lowerwhisker	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity
2	Finland	Western Europe	7.842	0.032	7.904	7.78	10.775	0.954	72	0.949	-0.098
3	Denmark	Western Europe	7.62	0.035	7.687	7.552	10.933	0.954	72.7	0.946	0.03
4	Switzerland	Western Europe	7.571	0.036	7.643	7.5	11.117	0.942	74.4	0.919	0.025
5	Iceland	Western Europe	7.554	0.059	7.67	7.438	10.878	0.983	73	0.955	0.16
6	Netherlands	Western Europe	7.464	0.027	7.518	7.41	10.932	0.942	72.4	0.913	0.175
7	Norway	Western Europe	7.392	0.035	7.462	7.323	11.053	0.954	73.3	0.96	0.093
8	Sweden	Western Europe	7.363	0.036	7.433	7.293	10.867	0.934	72.7	0.945	0.086
9	Luxembourg	Western Europe	7.324	0.037	7.396	7.252	11.647	0.908	72.6	0.907	-0.034
10	New Zealand	North America and ANZ	7.277	0.04	7.355	7.198	10.643	0.948	73.4	0.929	0.134
11	Austria	Western Europe	7.268	0.036	7.337	7.198	10.906	0.934	73.3	0.908	0.042
12	Australia	North America and ANZ	7.183	0.041	7.265	7.102	10.796	0.94	73.9	0.914	0.159
13	Israel	Middle East and North Africa	7.157	0.034	7.224	7.09	10.575	0.939	73.503	0.8	0.031
14	Germany	Western Europe	7.155	0.04	7.232	7.077	10.873	0.903	72.5	0.875	0.011
15	Canada	North America and ANZ	7.103	0.042	7.185	7.021	10.776	0.926	73.8	0.915	0.089
16	Ireland	Western Europe	7.085	0.04	7.164	7.006	11.342	0.947	72.4	0.879	0.077
17	Costa Rica	Latin America and Caribbean	7.069	0.056	7.179	6.96	9.88	0.891	71.4	0.934	-0.126
18	United Kingdom	Western Europe	7.064	0.038	7.138	6.99	10.707	0.934	72.5	0.859	0.233
19	Czech Republic	Central and Eastern Europe	6.965	0.049	7.062	6.868	10.556	0.947	70.807	0.858	-0.208
20	United States	North America and ANZ	6.951	0.049	7.047	6.856	11.023	0.92	68.2	0.837	0.098
21	Belgium	Western Europe	6.834	0.034	6.901	6.767	10.823	0.906	72.199	0.783	-0.153
22	France	Western Europe	6.69	0.037	6.762	6.618	10.704	0.942	74	0.822	-0.147
23	Bahrain	Middle East and North Africa	6.647	0.068	6.779	6.514	10.669	0.862	69.495	0.925	0.089
24	Malta	Western Europe	6.602	0.044	6.688	6.516	10.674	0.931	72.2	0.927	0.133
25	Taiwan Province of China	East Asia	6.584	0.038	6.659	6.51	10.871	0.898	69.6	0.784	-0.07
26	United Arab Emirates	Middle East and North Africa	6.561	0.039	6.637	6.484	11.085	0.844	67.333	0.932	0.074
27	Saudi Arabia	Middle East and North Africa	6.494	0.056	6.604	6.384	10.743	0.891	66.603	0.877	-0.149
28	Spain	Western Europe	6.491	0.042	6.574	6.408	10.571	0.932	74.7	0.761	-0.081

world-happiness-report-2021

Open Data Example: non-Trivial Spreadsheet

https://data.gov.uk/dataset/1f9e8832-d668-43ff-9511-d01f8ce0af22/handwashing-consumer-tracker

Header

	A	B	C	D	E
1	Home				
2					
3	20-066068-02 - 15th - 18th January 2021				
4	PUBLIC USE ONLY				
5	Handwashing				
6	Adults aged 16-75 in England, Wales and Northern Ireland				
7	Q2. Thinking about when you are at home at the moment, how often do you wash your				
8	All Adults aged 16-75 in England, Wales and Northern Ireland				
9					

Home

20-066068-02 - 15th - 18th January 2021

PUBLIC USE ONLY

Handwashing

Adults aged 16-75 in England, Wales and Northern Ireland

Q2. Thinking about when you are at home at the moment, how often do you wash your hands in each of the following ways? Select only one per statement. - Summary

All Adults aged 16-75 in England, Wales and Northern Ireland

	Q2. Thinking about when you are at home at the moment, how often do you wash your hands in each of the following ways? Select only one per statement.			
	With water only	soap/handwash and warm water	With soap/handwash and cold water	With hand sanitising wipes or gel
	(A)	(B)	(C)	(D)
Unweighted base	2062	2062	2062	2062
Weighted base	2062	2062	2062	2062
Never	1172 57% BCD	156 8%	731 35% B	784 38% BC
Sometimes	575 28% B	341 17%	698 34% AB	816 40% ABC
Most of the time	162 8%	1044 51% ACD	449 22% AD	298 14% A
Always	7 * 	458 22% ACD	47 2% AD	25 1% A
Prefer not to say	146 7% B	63 3%	137 7% B	139 7% B
NET: Never	1172 57% BCD	156 8%	731 35% B	784 38% BC
NET: Sometimes/most of the time	737 36%	1385 67% ACD	1147 56% A	1114 54% A
NET: Always	7 * 	458 22% ACD	47 2% AD	25 1% A

irements of the international quality standard for market research, ISO 20252 and with the Ipsos MORI Terms and Conditions.

Overlap formulae used

ColumnProportions(5%): A/B/C/D Minimum Base: 30(**) Small Base: 100(*)

ColumnMeans(5%): A/B/C/D Minimum Base: 30(**) Small Base: 100(*)

Open Data Example: non-Trivial Spreadsheet

https://data.gov.uk/dataset/1f9e8832-d668-43ff-9511-d01f8ce0af22/handwashing-consumer-tracker

Headings

Q2. Thinking about when you are at home at the moment, how often do you wash your hands in each of the following ways? Select only one per statement.

With water only	With soap/handwash and warm water	With soap/handwash and cold water	With hand sanitising wipes or gel
(A)	(B)	(C)	(D)

Home

20-066068-02 - 15th - 18th January 2021
PUBLIC USE ONLY
Handwashing
Adults aged 16-75 in England, Wales and Northern Ireland

Q2. Thinking about when you are at home at the moment, how often do you wash your hands in each of the following ways? Select only one per statement. - Summary

All Adults aged 16-75 in England, Wales and Northern Ireland

	Q2. Thinking about when you are at home at the moment, how often do you wash your hands in each of the following ways? Select only one per statement.			
	With water only (A)	soap/handwash and warm water (B)	With soap/handwash and cold water (C)	With hand sanitising wipes or gel (D)
Unweighted base	2062	2062	2062	2062
Weighted base	2062	2062	2062	2062
Never	1172 57% BCD	156 8%	731 35% B	784 38% BC
Sometimes	575 28% B	341 17%	698 34% AB	816 40% ABC
Most of the time	162 8%	1044 51% ACD	449 22% AD	298 14% A
Always	7 * 	458 22% ACD	47 2% AD	25 1% A
Prefer not to say	146 7% B	63 3%	137 7% B	139 7% B
NET: Never	1172 57% BCD	156 8%	731 35% B	784 38% BC
NET: Sometimes/most of the time	737 36%	1385 67% ACD	1147 56% A	1114 54% A
NET: Always	7 * 	458 22% ACD	47 2% AD	25 1% A

irements of the international quality standard for market research, ISO 20252 and with the Ipsos MORI Terms and Conditions.

Overlap formulae used

ColumnProportions(5%): A/B/C/D Minimum Base: 30(**) Small Base: 100(*)

ColumnMeans(5%): A/B/C/D Minimum Base: 30(**) Small Base: 100(*)

Open Data Example: non-Trivial Spreadsheet

https://data.gov.uk/dataset/1f9e8832-d668-43ff-9511-d01f8ce0af22/handwashing-consumer-tracker

Summary Data

13	Unweighted base	2062	2062	2062	2062
14					
15	Weighted base	2062	2062	2062	2062
16					

Home

20-066068-02 - 15th - 18th January 2021

PUBLIC USE ONLY

Handwashing

Adults aged 16-75 in England, Wales and Northern Ireland

Q2. Thinking about when you are at home at the moment, how often do you wash your hands in each of the following ways? Select only one per statement. - Summary

All Adults aged 16-75 in England, Wales and Northern Ireland

	Q2. Thinking about when you are at home at the moment, how often do you wash your hands in each of the following ways? Select only one per statement.			
	With water only (A)	soap/handwash and warm water (B)	With soap/handwash and cold water (C)	With hand sanitising wipes or gel (D)
Unweighted base	2062	2062	2062	2062
Weighted base	2062	2062	2062	2062
Never	1172 57% BCD	156 8%	731 35% B	784 38% BC
Sometimes	575 28% B	341 17%	698 34% AB	816 40% ABC
Most of the time	162 8%	1044 51% ACD	449 22% AD	298 14% A
Always	7 *	458 22% ACD	47 2% AD	25 1% A
Prefer not to say	146 7% B	63 3%	137 7% B	139 7% B
NET: Never	1172 57% BCD	156 8%	731 35% B	784 38% BC
NET: Sometimes/most of the time	737 36%	1385 67% ACD	1147 56% A	1114 54% A
NET: Always	7 *	458 22% ACD	47 2% AD	25 1% A

irements of the international quality standard for market research, ISO 20252 and with the Ipsos MORI Terms and Conditions.

Overlap formulae used

ColumnProportions(5%): A/B/C/D Minimum Base: 30(**) Small Base: 100(*)

ColumnMeans(5%): A/B/C/D Minimum Base: 30(**) Small Base: 100(*)

Open Data Example: non-Trivial Spreadsheet

https://data.gov.uk/dataset/1f9e8832-d668-43ff-9511-d01f8ce0af22/handwashing-consumer-tracker

Data (with calculated percentages)

17	Never	1172	156	731	784
18		57%	8%	35%	38%
19		BCD		B	BC
20	Sometimes	575	341	698	816
21		28%	17%	34%	40%
22		B		AB	ABC
23	Most of the time	162	1044	449	298
24		8%	51%	22%	14%
25			ACD	AD	A
26	Always	7	458	47	25
27		*	22%	2%	1%
28			ACD	AD	A
29	Prefer not to say	146	63	137	139
30		7%	3%	7%	7%
31		B		B	B

Home

20-066068-02 - 15th - 18th January 2021
PUBLIC USE ONLY
Handwashing
Adults aged 16-75 in England, Wales and Northern Ireland

Q2. Thinking about when you are at home at the moment, how often do you wash your hands in each of the following ways? Select only one per statement. - Summary

All Adults aged 16-75 in England, Wales and Northern Ireland

	Q2. Thinking about when you are at home at the moment, how often do you wash your hands in each of the following ways? Select only one per statement.			
	With water only (A)	soap/handwash and warm water (B)	With soap/handwash and cold water (C)	With hand sanitising wipes or gel (D)
Unweighted base	2062	2062	2062	2062
Weighted base	2062	2062	2062	2062
Never	1172 57% BCD	156 8%	731 35% B	784 38% BC
Sometimes	575 28% B	341 17%	698 34% AB	816 40% ABC
Most of the time	162 8%	1044 51% ACD	449 22% AD	298 14% A
Always	7 *	458 22% ACD	47 2% AD	25 1% A
Prefer not to say	146 7% B	63 3%	137 7% B	139 7% B
NET: Never	1172 57% BCD	156 8%	731 35% B	784 38% BC
NET: Sometimes/most of the time	737 36%	1385 67% ACD	1147 56% A	1114 54% A
NET: Always	7 *	458 22% ACD	47 2% AD	25 1% A

irements of the international quality standard for market research, ISO 20252 and with the Ipsos MORI Terms and Conditions.

Overlap formulae used

ColumnProportions(5%): A/B/C/D Minimum Base: 30(**) Small Base: 100(*)

ColumnMeans(5%): A/B/C/D Minimum Base: 30(**) Small Base: 100(*)

Open Data Example: non-Trivial Spreadsheet

https://data.gov.uk/dataset/1f9e8832-d668-43ff-9511-d01f8ce0af22/handwashing-consumer-tracker

35	NET: Never	1172	156	731	784
36		57%	8%	35%	38%
37		BCD		B	BC
38	NET: Sometimes/most of the time	737	1385	1147	1114
39		36%	67%	56%	54%
40			ACD	A	A
41	NET: Always	7	458	47	25
42		*	22%	2%	1%
43			ACD	AD	A

Aggregated Data

Home

20-066068-02 - 15th - 18th January 2021

PUBLIC USE ONLY

Handwashing

Adults aged 16-75 in England, Wales and Northern Ireland

Q2. Thinking about when you are at home at the moment, how often do you wash your hands in each of the following ways? Select only one per statement. - Summary

All Adults aged 16-75 in England, Wales and Northern Ireland

	Q2. Thinking about when you are at home at the moment, how often do you wash your hands in each of the following ways? Select only one per statement.			
	With water only (A)	soap/handwash and warm water (B)	With soap/handwash and cold water (C)	With hand sanitising wipes or gel (D)
	2062	2062	2062	2062
	2062	2062	2062	2062
	1172 57% BCD	156 8%	731 35% B	784 38% BC
	575 28% B	341 17%	698 34% AB	816 40% ABC
	162 8%	1044 51% ACD	449 22% AD	298 14% A
	7 * ACD	458 22% ACD	47 2% AD	25 1% A
	146 7% B	63 3%	137 7% B	139 7% B

NET: Never	1172 57% BCD	156 8%	731 35% B	784 38% BC
NET: Sometimes/most of the time	737 36%	1385 67% ACD	1147 56% A	1114 54% A
NET: Always	7 *	458 22% ACD	47 2% AD	25 1% A

irements of the international quality standard for market research, ISO 20252 and with the Ipsos MORI Terms and Conditions.

Overlap formulae used

ColumnProportions(5%): A/B/C/D Minimum Base: 30(**) Small Base: 100(*)

ColumnMeans(5%): A/B/C/D Minimum Base: 30(**) Small Base: 100(*)

Open Data Example: non-Trivial Spreadsheet

https://data.gov.uk/dataset/1f9e8832-d668-43ff-9511-d01f8ce0af22/handwashing-consumer-tracker

44

45 **irements of the international quality standard for market research, ISO 20252 and with the Ipsos MOR**

46 *Overlap formulae used*

47 *ColumnProportions (5%): A/B/C/D Minimum Base: 30(**) Small Base: 100(*)*

48 *ColumnMeans (5%): A/B/C/D Minimum Base: 30(**) Small Base: 100(*)*

Footer

Home

20-066068-02 - 15th - 18th January 2021
PUBLIC USE ONLY
Handwashing
Adults aged 16-75 in England, Wales and Northern Ireland

Q2. Thinking about when you are at home at the moment, how often do you wash your hands in each of the following ways? Select only one per statement. - Summary

All Adults aged 16-75 in England, Wales and Northern Ireland

	Q2. Thinking about when you are at home at the moment, how often do you wash your hands in each of the following ways? Select only one per statement.			
	With water only	soap/handwash and warm water	With soap/handwash and cold water	With hand sanitising wipes or gel
	(A)	(B)	(C)	(D)
Unweighted base	2062	2062	2062	2062
Weighted base	2062	2062	2062	2062
Never	1172 57% BCD	156 8%	731 35% B	784 38% BC
Sometimes	575 28% B	341 17%	698 34% AB	816 40% ABC
Most of the time	162 8%	1044 51% ACD	449 22% AD	298 14% A
Always	7 *	458 22% ACD	47 2% AD	25 1% A
Prefer not to say	146 7% B	63 3%	137 7% B	139 7% B
NET: Never	1172 57% BCD	156 8%	731 35% B	784 38% BC
NET: Sometimes/most of the time	737 36%	1385 67% ACD	1147 56% A	1114 54% A
NET: Always	7 *	458 22% ACD	47 2% AD	25 1% A

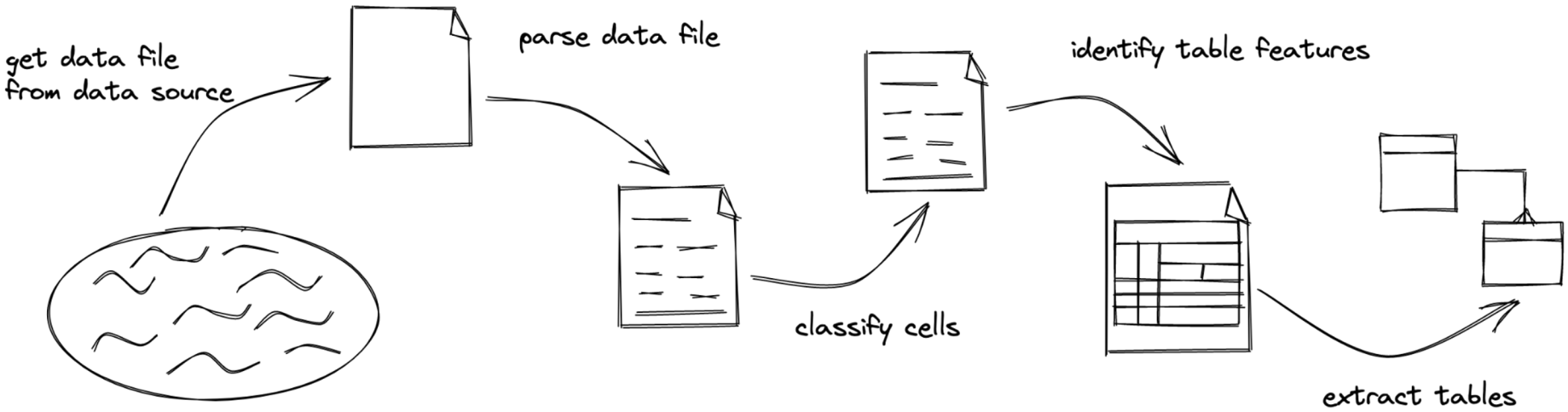
irements of the international quality standard for market research, ISO 20252 and with the Ipsos MORI Terms and Conditions.

Overlap formulae used

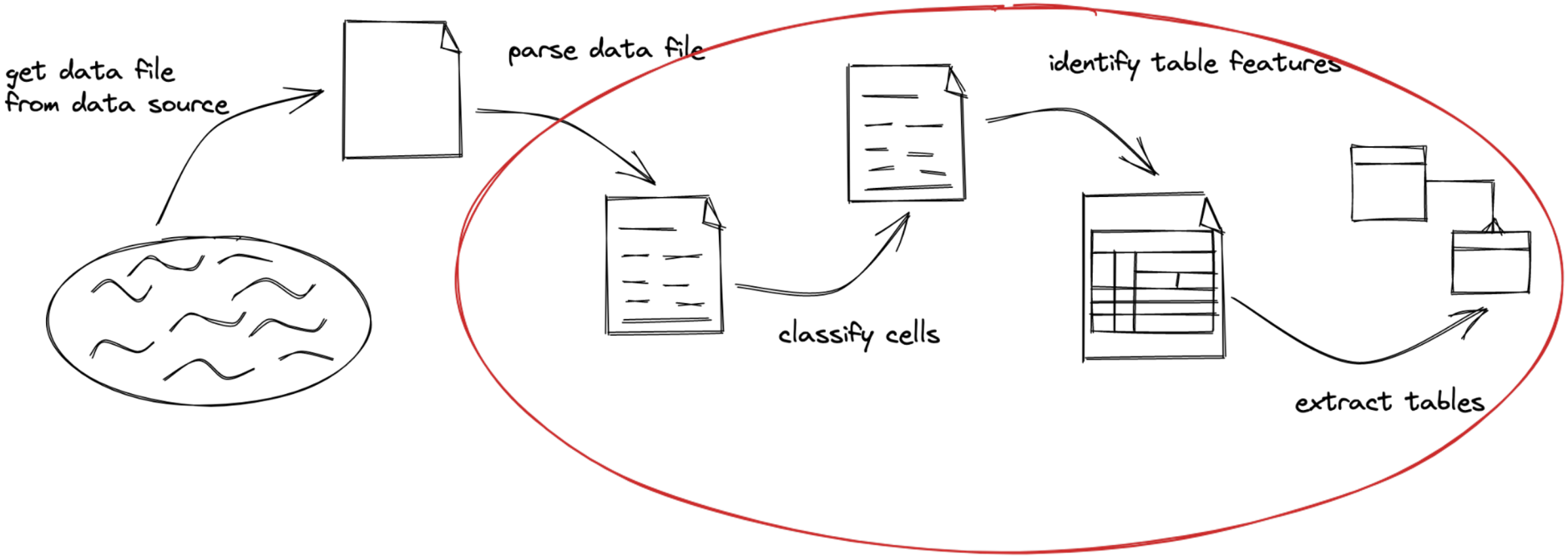
ColumnProportions (5%): A/B/C/D Minimum Base: 30(**) Small Base: 100(*)

ColumnMeans (5%): A/B/C/D Minimum Base: 30(**) Small Base: 100(*)

Steps to Table Extraction



Steps to Table Extraction



Agenda

- Choice of PhD
- Introduction to Open Data
- Challenges of Table Extraction
- **Research Approach**
- Existing Methods
- Genetic Algorithm Example
- Research Impact
- Plans for the Future

Approach

- Identify and classify existing work
- Benchmark existing methods
- Identify gaps in current work
- Decide on direction of my research

Agenda

- Choice of PhD
- Introduction to Open Data
- Challenges of Table Extraction
- Research Approach
- **Existing Methods**
- Genetic Algorithm Example
- Research Impact
- Plans for the Future

Related Work: Methods

- Manual
- Automated
- Semi-automatic
- Interactive

- Speculative
- Genetic Algorithm
- Rule-based
- Pattern-based
- Random forest

Related Work: Languages

- Python
- Java
- C++
- R
- Apache Spark
- …

Related work: papers

- **Structure Detection in Verbose CSV Files.**
  Lan Jiang, Gerardo Vitagliano, Felix Naumann | EDBT 2021: 193-204
- **Multi-hypothesis CSV parsing**
  T Döhmen, H Mühleisen, P Boncz| ACM Intl Conf Proceeding 2017
- **A Genetic-Based Search for Adaptive Table Recognition in Spreadsheets.**
  Elvis Koci, Maik Thiele, Oscar Romero, Wolfgang Lehner | ICDAR 2019: 1274-1279
- **Pytheas: Pattern-based Table Discovery in CSV Files.**
  Christina Christodoulakis, Eric Munson, Moshe Gabel, Angela Demke Brown, Renée J. Miller
  Proc. VLDB Endow. 13(11): 2075-2089 (2020)

Existing Tools: Academic

- Hypoparsr
- Strudel
- Pytheas
- DeExcelarator
- xtable_finder
- TabbyXL
- InfoGather+
- Muppets
- Senbazuru
- XLIndy

Existing Tools: Commercial

- ETL
- Spreadsheet
- COPY

Agenda

- Choice of PhD
- Introduction to Open Data
- Challenges of Table Extraction
- Research Approach
- Existing Methods
- **Genetic Algorithm Example**
- Research Impact
- Plans for the Future

Evolutionary Algorithms

Evolutionary Algorithm (EA)

… a generic population-based metaheuristic optimization algorithm. An EA uses mechanisms inspired by biological evolution, such as reproduction, mutation, recombination, and selection. Candidate solutions to the optimization problem play the role of individuals in a population, and the fitness function determines the quality of the solutions … Evolution of the population then takes place after the repeated application of the above operators.

https://en.wikipedia.org/wiki/Evolutionary_algorithm

Genetic Algorithm (GA)

- Specific case of Evolutionary Algorithm
- Developed by John Holland in 1960s/70s
- Inspired by natural selection
- Uses mutation, crossover & selection on a population
- Used to solve optimization problems

GA Pseudocode

Sean Luke, Essentials of Metaheuristics Second Edition, p.37

```
popsize ← desired population size                    ▷ This is basically λ. Make it even.

P ← {}
for popsize times do
    P ← P ∪ {new random individual}
Best ← □
repeat
    for each individual P_i ∈ P do
        AssessFitness(P_i)
        if Best = □ or Fitness(P_i) > Fitness(Best) then
            Best ← P_i
    Q ← {}                                           ▷ Here's where we begin to deviate from (μ, λ)
    for popsize/2 times do
        Parent P_a ← SelectWithReplacement(P)
        Parent P_b ← SelectWithReplacement(P)
        Children C_a, C_b ← Crossover(Copy(P_a), Copy(P_b))
        Q ← Q ∪ {Mutate(C_a), Mutate(C_b)}
    P ← Q                                            ▷ End of deviation
until Best is the ideal solution or we have run out of time
return Best
```

define # individuals in population

create initial population by generating *popsize* individuals

assess each individual and keep track of the best one

create a new generation using Crossover and Mutation

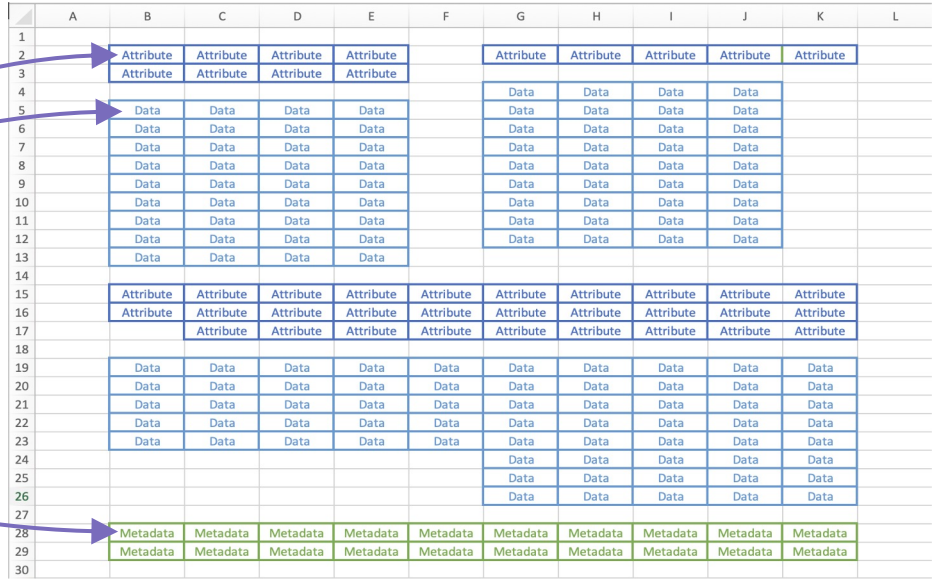
create new generations until end condition met

Return best individual found

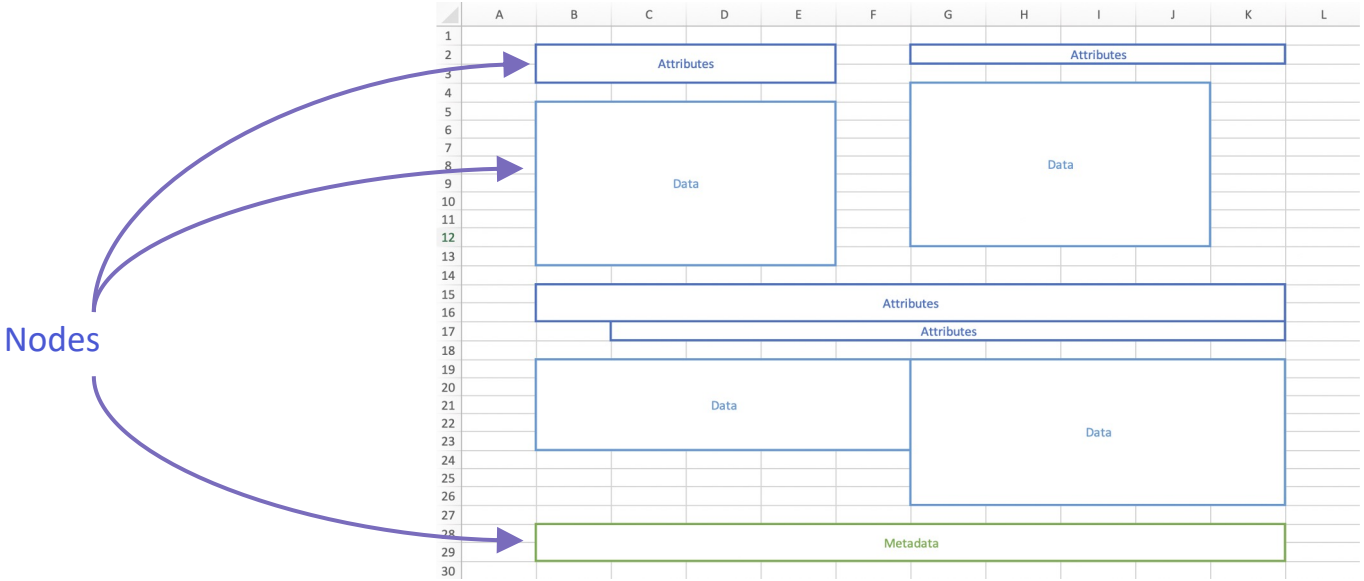
Applying GA to Table Extraction

Input: Annotated Spreadsheet

Labels

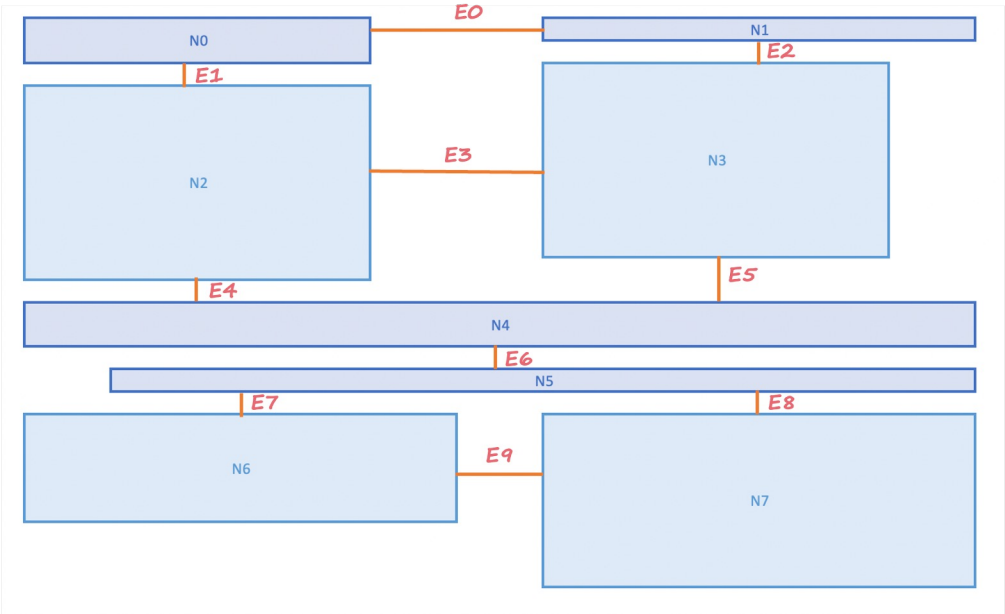


Applying GA to Table Extraction



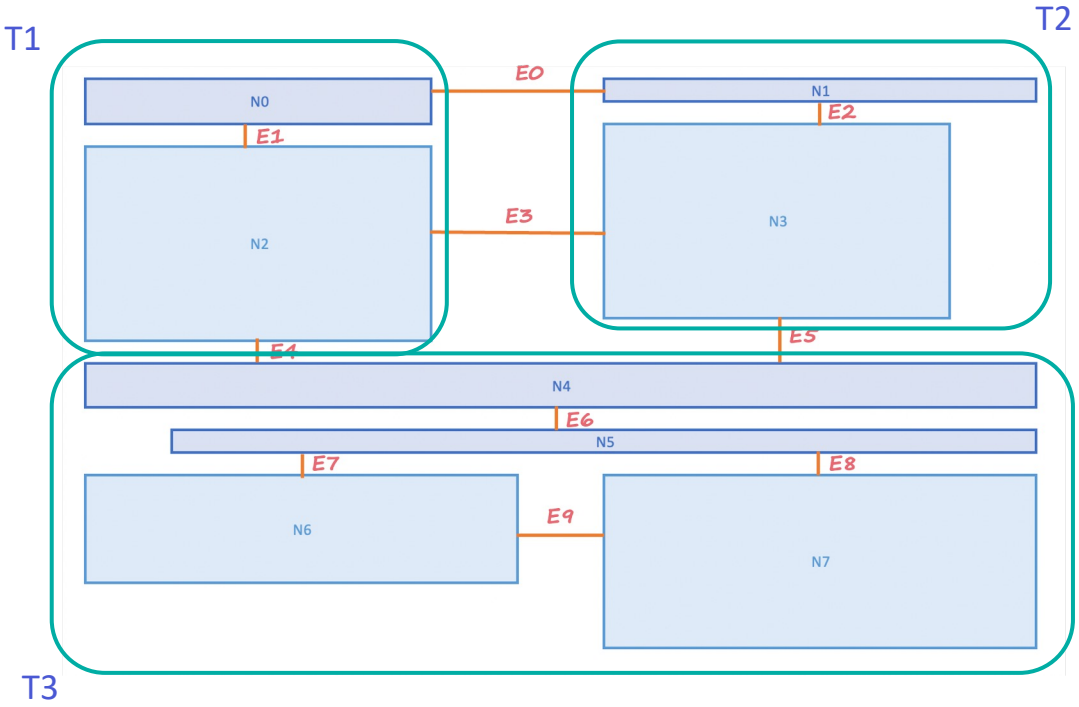
Applying GA to Table Extraction

Nodes and Edges



Applying GA to Table Extraction

Tables



Applying GA to Table Extraction

Expressing table extraction as a problem to be solved by GA

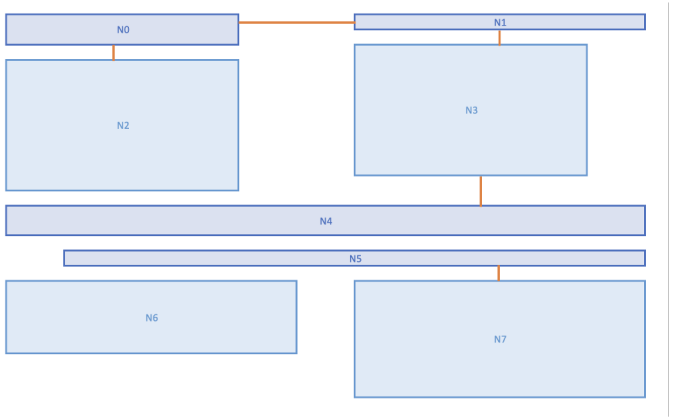
	E0	**E1**	**E2**	E3	E4	E5	**E6**	**E7**	**E8**	**E9**
Ideal Solution	0	**1**	**1**	0	0	0	**1**	**1**	**1**	**1**

(0,1,1,0,0,0,1,1,1,1)

Applying GA to Table Extraction

Expressing table extraction as a problem to be solved by GA

	E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	
Ideal Solution	0	1	1	0	0	0	1	1	1	1	(0,1,1,0,0,0,1,1,1,1)
Solution 2	1	1	1	0	0	1	0	0	1	0	(1,1,1,0,0,1,0,0,1,0)



Applying GA to Table Extraction

Additional Steps Required

- Determine list of rules to calculate the best solution (ie the one that is the most “table like”)
  - How well aligned the header and data nodes are
  - How much space there is between nodes
  - If the header appears above the data as expected
- Train the solution to assign weightings to the rules

Applying GA to Table Extraction

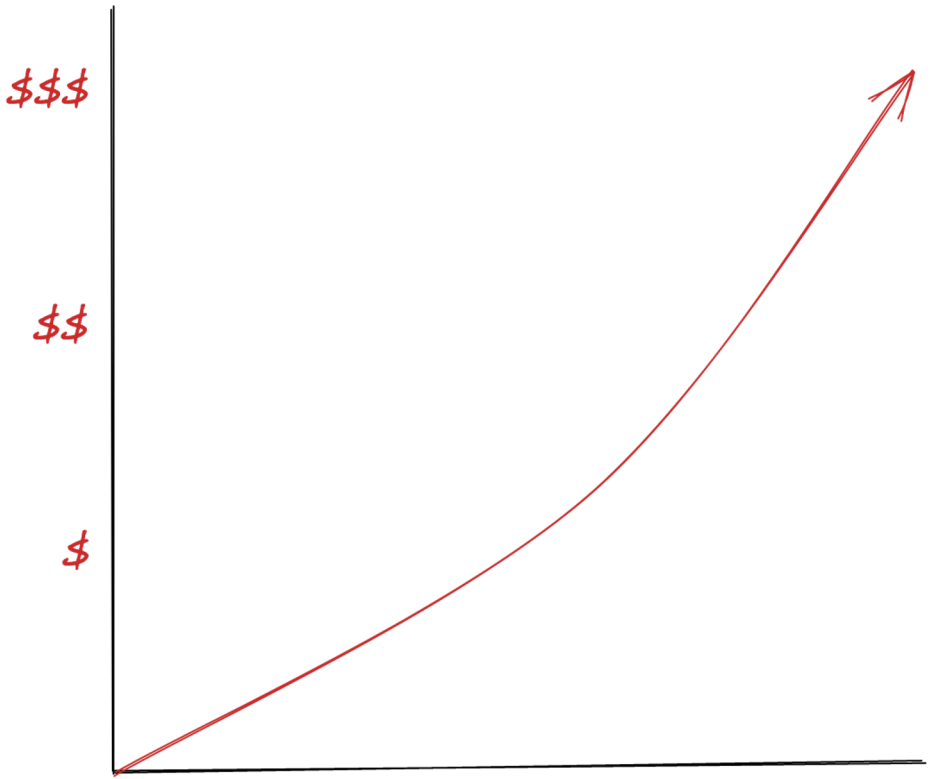
Advantages and Disadvantages of the Approach

- Allows for multiple tables
  Vertically and/or horizontally stacked
- Requires annotated spreadsheet as input
  Annotation is usually a manual task

Agenda

- Choice of PhD
- Introduction to Open Data
- Challenges of Table Extraction
- Research Approach
- Existing Methods
- Genetic Algorithm Example
- **Research Impact**
- Plans for the Future

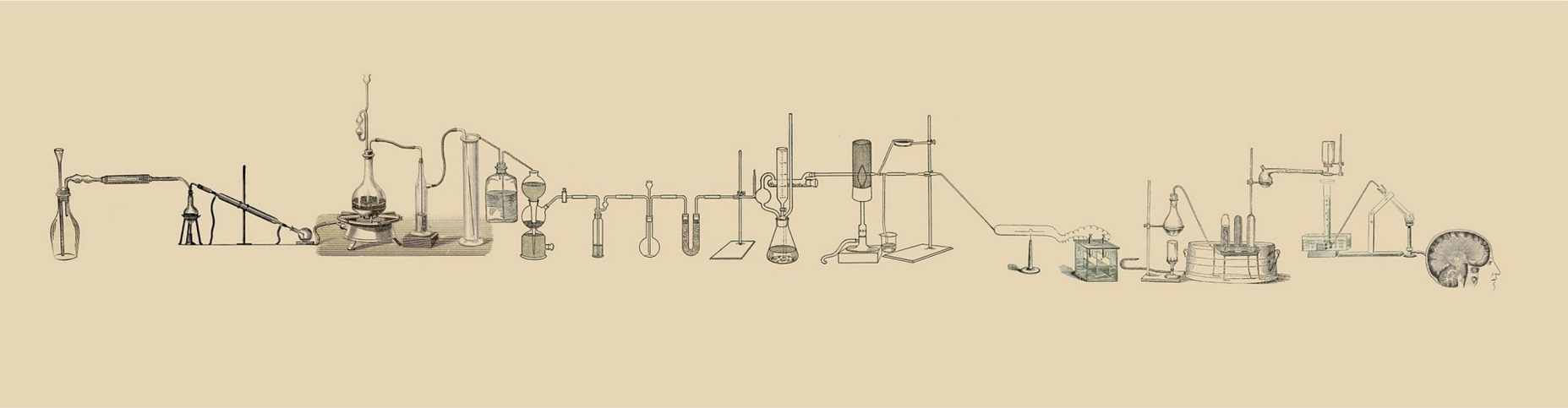
Research Impact



Research Impact

- Need to get value from this data
- This is challenging

Research Impact



Research Impact

- Existing commercial tools are of limited value
  - Respond to one aspect of the problem
  - Expect structured files in a specific format
  - Require detailed knowledge of the data

Agenda

- Choice of PhD
- Introduction to Open Data
- Challenges of Table Extraction
- Research Approach
- Existing Methods
- Genetic Algorithm Example
- Research Impact
- **Plans for the Future**

Future Plans

- Next year: Experimental evaluation of existing methods
- Next 5 years: Identify and develop new methods
- Longer term:
  - Fully automated table extraction from Open Data
  - PostgreSQL extension?

Conclusions

- Complex problem
- Many different existing approaches to parts of the problem
- Difficult to identify one complete solution
- Huge potential impact

Thank You!

Karen Jex | @karenhjex | karen.jex@crunchydata.com
Slides: https://karenjex.blogspot.com

Image acknowledgements

- Possibilities: Image by Gerd Altmann from Pixabay
- Research Impact (lab): Image by holdentrils from Pixabay