

# Data mining in the times of the GDPR

Jesus M. Gonzalez-Barahona

Universidad Rey Juan Carlos  
@jgbarah <http://github.com/jgbarah/presentations>

3 de septiembre de 2019

Data mining &  
GDPR

Jesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

# Should we worry?

# Mining software repositories

*Should we worry  
when we research  
based on data  
extracted from  
software development repositories?*

Let's analyze from two (interrelated)  
points of view:

- legal requirements
- ethical requirements

Data mining &  
GDPR

Jesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

While the EU's ethics review process is primarily concerned with ethics issues, your project must demonstrate compliance with the GDPR. However, the fact that your research is legally permissible does not necessarily mean that it will be deemed *ethical*.

Crucially, if your research proposal involves the processing of any personal data, whatever method is used, you – and all of your partners, collaborators and service providers – must, if called upon, be able to demonstrate compliance with both legal and ethical requirements. Such requests could come from data subjects, funding agencies or data protection supervisory authorities.

H2020 document on Ethics and Data Protection, by EC

# Applicable law

In the European Union (and elsewhere):

- **General Data Protection Regulation (GDPR)**  
Affects all of EU, and rights of EU citizens
- Specific law in member states
- Recommendations by national data protection agencies

Similar law in other jurisdictions:

- **California Consumer Privacy Act**

### Data mining & GDPR

Jesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

It is highly likely that if your project involves any data about identifiable persons, even if they are not directly participating in the research, you are processing personal data and must comply with EU and national law. Only data that have been fully and irreversibly anonymised are exempt from these requirements. Importantly, while **pseudonymisation** can provide individual data subjects with a degree of protection and anonymity, pseudonymised data still fall within the scope of personal data because it is possible to re-identify the data subject (see below).

Even if your project is using only **anonymised data**, the origin or acquisition of the data may still raise significant ethics issues.

## H2020 document on Ethics and Data Protection, by EC

# “Higher risks” related to GDPR in research

Types of personal data	<ul style="list-style-type: none"> <li>* racial or ethnic origin</li> <li>* political opinions, religious or philosophical beliefs</li> <li>* genetic, biometric or health data</li> <li>* sex life or sexual orientation</li> <li>* trade union membership</li> </ul>
Data subjects	<ul style="list-style-type: none"> <li>* children</li> <li>* vulnerable people</li> <li>* people who have not given their explicit consent to participate in the project</li> </ul>
Scale or complexity of data processing	<ul style="list-style-type: none"> <li>* large-scale processing of personal data</li> <li>* systematic monitoring of a publicly accessible area on a large scale</li> <li>* involvement of multiple datasets and/or service providers, or the combination and analysis of different datasets (i.e. big data)</li> </ul>
Data-collection or processing techniques	<ul style="list-style-type: none"> <li>* privacy-invasive methods or technologies (e.g. the covert observation, surveillance, tracking or deception of individuals)</li> <li>* using camera systems to monitor behaviour or record sensitive information</li> <li>* data mining (including data collected from social media networks), ‘web crawling’ or social network analysis</li> <li>* profiling individuals or groups (particularly behavioural or psychological profiling)</li> <li>* using artificial intelligence to analyse personal data</li> <li>* using automated decision-making that has a significant impact on the data subject(s)</li> </ul>
Involvement of non-EU countries	<ul style="list-style-type: none"> <li>* transfer of personal data to non-EU countries</li> <li>* collection of personal data outside the EU</li> </ul>

H2020 document on Ethics and Data Protection, by EC



## Data mining &amp; GDPR

Jesus M.

Gonzalez-Barahona

Should we worry?

What should we do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

Types of personal data	<ul style="list-style-type: none"> <li>* racial or ethnic origin</li> <li>* political opinions, religious or philosophical beliefs</li> <li>* genetic, biometric or health data</li> <li>* sex life or sexual orientation</li> <li>* trade union membership</li> </ul>
Data subjects	<ul style="list-style-type: none"> <li>* children</li> <li>* vulnerable people</li> <li>* people who have not given their explicit consent to participate in the project</li> </ul>
Scale or complexity of data processing	<ul style="list-style-type: none"> <li>* large-scale processing of personal data</li> <li>* systematic monitoring of a publicly accessible area on a large scale</li> <li>* involvement of multiple datasets and/or service providers, or the combination and analysis of different datasets (i.e. big data)</li> </ul>
Data-collection or processing techniques	<ul style="list-style-type: none"> <li>* privacy-invasive methods or technologies (e.g. the covert observation, surveillance, tracking or deception of individuals)</li> <li>* using camera systems to monitor behaviour or record sensitive information</li> <li>* data mining (including data collected from social media networks), 'web crawling' or social network analysis</li> <li>* profiling individuals or groups (particularly behavioural or psychological profiling)</li> <li>* using artificial intelligence to analyse personal data</li> <li>* using automated decision-making that has a significant impact on the data subject(s)</li> </ul>
Involvement of non-EU countries	<ul style="list-style-type: none"> <li>* transfer of personal data to non-EU countries</li> <li>* collection of personal data outside the EU</li> </ul>

# Children & vulnerable people

Maybe we don't know...

...but they may be in our dataset

- they are subject to special protection
- even when we usually cannot tell who they are...
- ...others could

This situation is very difficult to deal with

# Privacy invasive methods

- Example: who is working off-hours
- Methodology: tracking individual activity in all available data sources
- Risk: tagging specific people

You can learn working hours, days off, vacation...

# Profiling individuals or groups

- Example: activities by newcomers
- Methodology: tracking individual activity in all available data sources
- Risk: tagging specific people

You can show specific activity of persons

# Using AI to analyze personal data

- Example: find out experts
- Methodology: analyze activity to find out experts in some languages, using AI
- Risk: singling out specific persons

# Large-scale

- Usually, several data sources
- The more data, the better
- If we can combine datasets, we do

The better the research, the riskier

# Data mining from social media

- Our data source are social media
- Of course we mine data from them

Data mining is the core of our business

# No explicit consent

- We collect data from services...
- ...which didn't get explicit consent for our cases
- Even if they got, they don't guarantee that for us
- In summary: usually, no explicit consent

Can we avoid this case?



# Transfer of data across EU border

- Collecting data from non EU data sources
- Sharing data with non-EU researchers

Can we avoid these scenarios?

Data mining &  
GDPR

Jesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

# What should we do?

# Detailed analysis

Ethics issues raised by our methodology:

- data collection and processing operations
- ethics issues that these raise
- mitigation of these issues in practice.

Submission to the Research Ethics Committee.

# Detailed analysis

- Mandatory for EC-funded research proposals
- Important: involve the DPO  
(Data Protection Officer)
- Maybe: requirement to conduct a DPIA  
(Data Protection Impact Assessment)

# Mitigation

Anonymization, pseudonymization

...but even in this case, ethics issues:

- origin of the data
- potential misuse of methodology or findings
- potential for deanonymization

Data mining &  
GDPR

Jesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

# Some definitions

# Data processing

*(2) Data processing [includes] any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.*

GDPR, Article 4.

# Actors subject to GDPR

- Data controller: determines the purposes and means of the processing of personal data
- Data processor: processes personal data on behalf of the controller



## DPIA

## Data Protection Impact Assessment:

*Process designed to assess the data-protection impacts [...] and, [...] to ensure that remedial actions are taken as necessary to correct, avoid or minimise the potential negative impacts on the data subjects.*

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

# DPIA likely required

A company systematically monitoring its employees' activities, including the monitoring of the employees' work station, internet activity, <i>etc.</i>	<ul style="list-style-type: none"> <li>- Systematic monitoring</li> <li>- Data concerning vulnerable data subjects</li> </ul>
The gathering of public social media data for generating profiles.	<ul style="list-style-type: none"> <li>- Evaluation or scoring</li> <li>- Data processed on a large scale</li> <li>- Matching or combining of datasets</li> <li>- Sensitive data or data of a highly personal nature</li> </ul>

H2020 document on Ethics and Data Protection,  
by European Commission

Data mining &  
GDPR

Jesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

# A case study

# The problem

You have a collection of  
all commits fulfilling some properties  
from a large set of public repositories.

You analyze number of committers per time  
period.

How can you publish the dataset  
in a reproduction package?

# The problem

Let's assume we have  
a legitimate basis  
for the data processing

(is research a legitimate basis?)

Data mining &  
GDPRJesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

```
commit 491f9205c36fdf54b4bbb7f25ba83b6cb99874b9
Author: Jesus M. Gonzalez-Barahona <jgb@gsync.es>
Date:   Mon May 13 13:33:04 2019 +0200
```

Add notice about GNOME Extension needed  
for indications.

Data mining &  
GDPRJesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

```
commit 491f9205c36fdf54b4bbb7f25ba83b6cb99874b9
Author: Jesus M. Gonzalez-Barahona <jgb@gsyc.es>
Date:   Mon May 13 13:33:04 2019 +0200
```

Add notice about GNOME Extension needed  
for indications.

# Personal data?

Data mining &  
GDPRJesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

# Personal data

*(1) 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;*

## GDPR, Art. 4



# Personal data

*Personal data include data such as internet protocol (IP) addresses (unique identifiers that can be used to identify the owner of devices connected to the internet) and data from 'smart meters' monitoring energy usage by addresses linked to identifiable persons.*

H2020 document on Ethics and Data Protection,  
by European Commission

Certainly, it includes names & email identifiers.

# Personal data

Jesus M. Gonzalez-Barahona <jgb@gsyc.es>

# Personal data

But... wait!!!!

Our data is public data!!

# Open source data

## [Box 4] Using 'open source' data

The fact that some data are publicly available does not mean that there are no limits to their use.

On the contrary, **if you take 'open source' personal data about identifiable persons and create new records or files/profiles, you are processing personal data about them** and must have a lawful/legitimate basis for doing so.

**You must ensure that the data processing is fair to the data subject and that their fundamental rights are respected.**

H2020 document on Ethics and Data Protection,  
by European Commission

# Open source data

If your research project uses **data from social media networks** and you do not intend to seek the data subjects' explicit consent to the use of their data, you must assess whether those persons actually intended to make their information public (e.g. in the light of the privacy settings or limited audience to which the data were made available).

It is not enough that the data be accessible; they must have been made public to the extent that the data subjects do not have any **reasonable expectation of privacy**. **You must also ensure that your intended use of the data complies with any terms and conditions published by the data controller.**

If you are in any doubt as to what you can and cannot do with this kind of data, you should seek advice from your DPO or a suitably qualified expert and include their opinion in your proposal.

H2020 document on Ethics and Data Protection,  
by European Commission

# How to fix the problem

- Anonymize: Strip all personal data  
(but still... more on this later)
- Pseudonymize personal data  
(the dataset will be much richer, but still...)

# Pseudonymizing

*(5) 'pseudonymisation' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;*

## GDPR, Art. 4

# Pseudonymization

## [Box 2] Pseudonymisation and anonymisation: understanding the difference

**Pseudonymisation** entails substituting personally identifiable information (such as an individual's name) with a unique identifier that is not connected to their real-world identity, using techniques such as coding or hashing. However, if it is possible to re-identify the individual data subjects by reversing the pseudonymisation process, data protection obligations still apply. They cease to apply only when the data are fully and irreversibly anonymised.

**Anonymisation** involves techniques that can be used to convert personal data into anonymised data. Anonymisation is increasingly challenging because of the potential for re-identification.

**Re-identification** is the process of turning pseudonymised or anonymised data back into personal data by means of data matching or similar techniques.

H2020 document on Ethics and Data Protection,  
by European Commission



# Pseudonymizing 1

```
echo -n "<jgb@gsyc.es>" | sha256sum  
3fdffb4a435cc3a5bab7d96b3cc2cefea90ca879f7fba0341
```

```
commit 491f9205c36fdf54b4bbb7f25ba83b6cb99874b9  
Author: 3fdffb4a435cc3a5bab7d96b3cc2cefea90ca879f7fba0341  
Date: Mon May 13 13:33:04 2019 +0200
```

Add notice about GNOME Extension needed...

# Pseudonymizing 1

Not good enough:

Data returned for an Email Append

Input (Encrypted Email)	Recovered Email	First Name	Last Name	Address	City	State	Zip	Phone
cbf05329de4e57e4cba09471448ddb98	joe.smith@gmail.com							
5469703a9c26d5e8be4e46bef4596e2f836088c0	commonme@yahoo.com	Don	Johnson	478 19TH PL W	Redmond	WA	98052	4255557892

In addition to reversing hashed email addresses, Datafinder also provides personal information including name, address and phone number associated with an email address.

Four cents to deanonymize: Companies reverse hashed  
email addresses, by Gunes Acar

# Pseudonymizing 1

For the curious:

- Estimated: 5 billion email addresses (2018)
- Amazon EC2: 450 billion hashes/sec.
- Lists of (targeted) email lists for sale
- Data breaches leaking billions of addresses

There is a whole business ecosystems around  
email addresses

Four cents to deanonymize: Companies reverse hashed

## Pseudonymizing 2

- Hash “name + address”

Better, but still subject to attack if you harvested addresses

- Salt the hash, use different algorithm
- Non-hash functions (eg, sequential code)
- Encryption instead of hash

Better, but you need to disclose details if you want others to merge with your dataset

# Pseudonymizing 3

A possibility emerges:

*“Encryption / coding*

*Datasets: public*

*Key / coding table:*

*only to researchers asking for it*

(variant: reference them in the paper)

Is this good enough for “legitimate use”?

# Pseudonymizing 3

commit 491f9205c36fdf54b4bbb7f25ba83b6cb99874b9

Author: 4334345

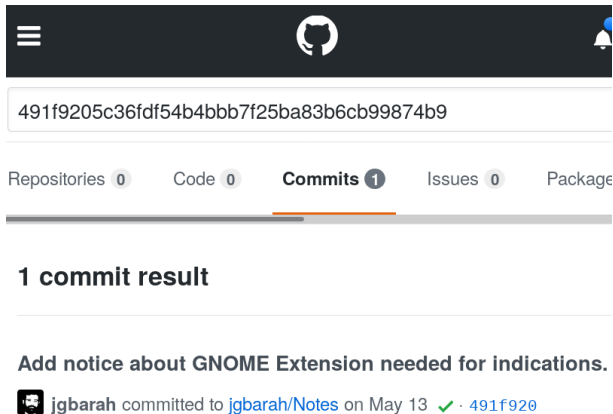
Date: Mon May 13 13:33:04 2019 +0200

Add notice about GNOME Extension needed...

Separate table:

4334345, Jesus M. Gonzalez-Barahona <jgb@gsyc.es>

# We still have a problem




The screenshot shows the GitHub interface for a repository named 'jgbarah/Notes'. The repository has 0 repositories, 0 code files, 1 commit, 0 issues, and 0 packages. The commit history shows a single commit by 'jgbarah' on May 13, 2019, with the message 'Add notice about GNOME Extension needed for indications.' and a green checkmark indicating it is the latest commit. The commit hash is 491f920.

491f9205c36fdf54b4bbb7f25ba83b6cb99874b9

Repositories 0 Code 0 **Commits 1** Issues 0 Package

## 1 commit result

Add notice about GNOME Extension needed for indications.

 jgbarah committed to [jgbarah/Notes](#) on May 13 ✓ · [491f920](#)

# We still have a problem

```
curl https://archive.softwareheritage.org/api/1/1/491f9205c36fdf54b4bbb7f25ba83b6cb99874b9/
```

```
{"author":  
  {"name": "Jesus M. Gonzalez-Barahona",  
    "fullname": "Jesus M. Gonzalez-Barahona <jgb@g",  
    "email": "jgb@gsyc.es"}  
}  
...  
}
```



# What can we do?

Pseudonymize the hash too:

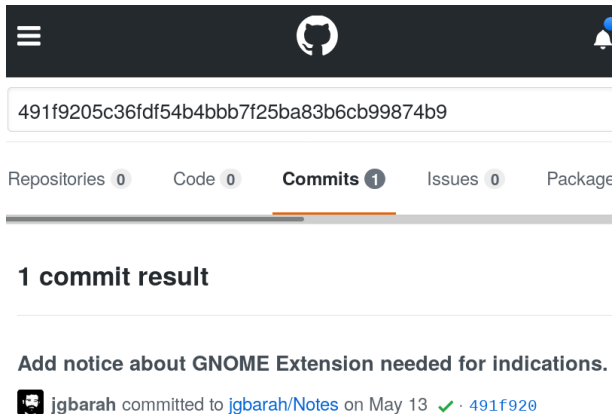
```
commit 6777888876876
```

```
Author: 4334345
```

```
Date: Mon May 13 13:33:04 2019 +0200
```

Add notice about GNOME Extension needed...

# We still have a problem (2)




The screenshot shows the GitHub interface for a commit. At the top is a dark navigation bar with a menu icon, the GitHub logo, and a notification bell. Below this is a search bar containing the commit hash 491f9205c36fdf54b4bbb7f25ba83b6cb99874b9. A horizontal bar below the search bar displays statistics: Repositories 0, Code 0, Commits 1 (highlighted with an orange bar), Issues 0, and Package 0. The main content area is titled "1 commit result" and contains the text "Add notice about GNOME Extension needed for indications." followed by a commit entry: a user avatar, the username "jgbarah", the text "committed to jgbarah/Notes on May 13", a green checkmark, and the commit hash "491f920".

491f9205c36fdf54b4bbb7f25ba83b6cb99874b9

Repositories 0 Code 0 **Commits 1** Issues 0 Package 0

## 1 commit result

Add notice about GNOME Extension needed for indications.

 jgbarah committed to [jgbarah/Notes](#) on May 13 ✓ · [491f920](#)

## We still have a problem (2)

- Commit comment can be used to deanonymize author.
- Date can be used to deanonymize author.

`commit 6777888876876`

`Author: 4334345`

`Date: 5353453453`

`Comment: 4343434334`

...and separate coding tables

# Why we need tables

- For reproduction: commits per time period
- for reuse: identity merging
- for reuse: relationship between message and time of the day
- for reuse: link to issues
- ...

# When data is public...

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

This is a problem for anyone  
having access to the data  
that allows deanonymization

But also for anyone letting others deanonymize

Big trouble if it is publicly available data

## Data mining & GDPR

Jesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

# By design

# GDPR approach

## Data protection by design (DPbD):

*Data controllers are required to implement appropriate technical and organisational measures to give effect to the core data-protection principles of GDPR.*

GDPR, Articles 5 and 25

# DPbD in research

## Data protection by design:

- Anonymization / pseudonymization
- data minimization
- cryptography (hashing, encrypting)
- data protection focused service providers & storage
- procedures for exercising fundamental rights (access, consent)



Data mining &  
GDPR

Jesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

# Important details

# Timing of anonymization

- Collection (no personal data is processed)  
Example: web form, no browser tracking  
Example: anonymous dataset from 3rd party
- Later than collection:  
Raw data is not anonymized  
(needs special protection)

# Data minimization

*(1) Data processing must be lawful, fair and transparent. It should involve only data that are necessary and proportionate to achieve the specific task or purpose for which they were collected*

GDPR, Article 5

# Data minimization

- Collect minimal personal data
- Anonymize and pseudonymize
- Store data securely
- Dispose data when no longer needed
- Limit access to data

# Informed consent

*Explain to participants what your research is about, what their participation in your project will entail and any risks that may be involved. When they have fully understood, see and obtain their express permission.*

GDPR, Article 4 (11), Article 7

# Managing consent

You need to document and archive consent

You need to be able of producing evidence

Consent management applications:

- ethically robust, secure
- model consent processes
- manage, document, evidence

Data mining &  
GDPR

Jesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

# Long term

# A call for action

Maybe we researchers need to work with developers to learn what is legitimate use for them?

Maybe developers should specify what is legitimate use for them, in their repositories?



Data mining &  
GDPR

Jesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

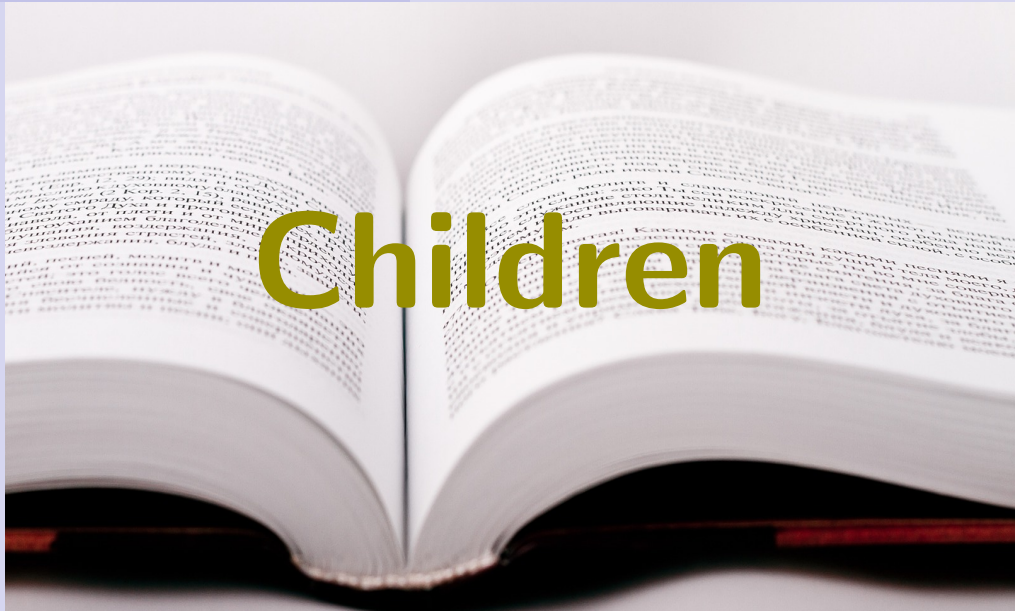
Important details

Long term

Children

To probe further

# Children



Data mining &  
GDPRJesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

*If your research project involves collecting data from children, you must follow the “EC Guidance note on informed consent”, in particular the provisions on obtaining the consent of a parent/legal representative and, where appropriate, the assent of the child. [...] it is imperative that any information you address to a child is in age-appropriate and plain language that they can easily understand.*

H2020 document on Ethics and Data Protection,  
European Commission

Data mining &  
GDPRJesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

*The GDPR establishes special safeguards for children in relation to “information society services”, a broad term covering all internet service providers, including social media platforms. These include a requirement for verified parental consent in respect of information society services offered directly to children aged under 16. Individual Member States may provide for this threshold to be lowered to 13.*

H2020 document on Ethics and Data Protection,  
European Commission

Data mining &  
GDPR

Jesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

# To probe further

# References

- **GDPR portal** by EC  
(includes full text of GDPR)
- **H2020 document on Ethics and Data Protection**, European Commission
- **GDPR and Research: An Overview for Researchers**, UK Research and Innovation

Data mining &  
GDPR

Jesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further

# Credits

## Data mining & GDPR

Jesus M.  
Gonzalez-Barahona

Should we worry?

What should we  
do?

Some definitions

A case study

By design

Important details

Long term

Children

To probe further



©2019 Jesus M. Gonzalez-Barahona.

Some rights reserved. This document is distributed under the terms  
of the Creative Commons License "Attribution-ShareAlike 4.0",  
available in

<http://creativecommons.org/licenses/by-sa/4.0/>

This document (including source) is available from  
<https://github.com/jgbarah/presentaciones>