A Researcher’s Perspective on the GDPR 2018
Legal Constraints on Data-driven SSH Research

Kristoffer L Nielbo
knielbo@sdu.dk
knielbo.github.io

datakube|Department of History|University of Southern Denmark
<table>
<thead>
<tr>
<th>Time</th>
<th>Section</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>Data, OD &amp; initial concerns</td>
<td>deluge, obligations, culture</td>
</tr>
<tr>
<td>0.10</td>
<td>GDPR and scientific research</td>
<td>safeguards, exemptions, pseudonymization</td>
</tr>
<tr>
<td>0.20</td>
<td>Discussion</td>
<td>...</td>
</tr>
</tbody>
</table>
– the data deluge is transforming knowledge discovery and understanding in every domain of human inquiry
– knowledge discovery depends critically on advanced computing capabilities
  a large part of these data are unstructured and fundamentally cultural
– we have limited experience with computational approaches and lack common standards for managing first- and third-party data
– with the advent of machine learning data combined with rich domain knowledge are our most valuable assets

– ‘generic’ benchmarking data sets are being replaced by domain relevant datasets

– probabilistic notions of ‘fairness’ (i.e., parity definitions) are challenged by impossibility results

GDPR grants users the right to a logical explanation of how algorithms use their data
DATA-RELATED CHALLENGES

data obligations
– personal data protection policies (e.g., GDPR) are basically a good thing
– intellectual property rights (e.g., copyright) can also be a good thing
– BUT both can be a hindrance to research

cultural issues
– develop an organisational culture of analytics
– computational and data literacy in SSH
– data silos at private and public data providers
OPEN DATA

we spend $\frac{1}{3}$ of the total global research budget on publishing and communicating results that 99% of people cannot access (DataCite|FORCE2017)

OD/ODbL
- share, create and adapt data irrespective of intellectual property rights, making data a public good
- improvement of human welfare >> ethics of privacy

make data-driven research more like open software

OD::GDPR
- GDPR is about returning ownership of data to the individual through explicit consent
- most controversial is delete and provide
- if data are kept in accordance with GDPR’s safeguards, data re-use in research can become easier
General Data Protection Regulation | GDPR

summary
- protect privacy rights of the data subject
- harmonizing (R)DM practices within EU (Digital Single Market)
- exemptions for scientific, historical, health and statistical research
- data re-use and sensitive categories (∼ override rights) → implement safeguards

- making data protection a daily practice
- processing of personal sensitive data for research shall be of public interest
- provide adequate technical and organisational measures ensuring data protection
WHOSE DATA IS IT ANYWAY?

I DEMAND THE RIGHT TO BE FORGOTTEN!

SORRY, WHO ARE YOU AGAIN?

THE RIGHT TO BE FORGOTTEN

MOST PEOPLE DON'T CARE ABOUT THE DATA ON THEM, THAT'S OUT THERE.

WHATEVER.

IF PEOPLE TAKE THEMSELVES OUT OF A DATA SET, IT IS NO LONGER VALID.

THERE ISN'T ONE BIG BUTTON YOU CAN PUSH.

YOUR PERSONAL DATA IS ONLY PERSONAL DATA IF SOMEONE USES IT.

PEOPLE NEED TO BE MORE AWARE OF THEIR RIGHTS.

PEOPLE THINK THERE'S NOTHING THEY CAN DO ANYWAY.

Casually uploading his DNA profile...

It's a headache for future historians

Data on you is sold to others...

2017: NY DATA

Casually uploading his DNA profile...

Casually uploading his DNA profile...

Casually uploading his DNA profile...

Casually uploading his DNA profile...

Casually uploading his DNA profile...
GDPR | scientific research

safeguards
– “conditions for exemptions”
– data protection by design
– technical and organizational measures in accordance with data minimisation
– pseudonymisation - personal data that is protected by coding and encryption
– Data Protection Officer (DPO) is mandatory (when large scale \(^1\) data collection is involved)

issues
– data re-use (\(\sim\)secondary use) should not be incompatible with initial collection (exemption for research)
– Big Data are not specified (and opposed to data minimization)
– cloud computing seems problematic
– data mining lacks apriori hypotheses, which complicates consent

\(^1\)N = 5000+ or 12+ months (v2014)
Pseudonymization

definition
– data that cannot be attributed to data subject without additional data
– iow. removal of direct identifiers

anonymization
– not GDPR relevant
– removal of direct and indirect identifiers

why
– facilitate data re-use
– safeguard for scientific, historical and statistical purposes
– essential “by-design” feature
– used to meet data security requirements
– delete and provide is not necessary if data subject is not identifiable
SoMe example

-data
  - web-based and predominately unstructured data
  - majority are publicly available
  - often available through API

-principles
  - consent/opt-in policy
  - parent consent
  - right to be forgotten (delete and provide)

-strategies
  - ‘walled garden’ of Terms and Conditions
  - but ‘silence, pre-ticked boxes or inactivity’ is inadequate to confer consent
  - exempt from notice requirement, if data are from publicly available sources
  - utilize anonymous data sources
– or “why you don’t have to worry about GDPR’s approach to cloud computing”
– affordable and scalable compute + RDM for research projects @ SDU
THANK YOU

knielbo@sdu.dk

& THANKS TO
Culture Analytics © Institute of Pure and Applied Mathematics (UCLA)
especially Tina Eliassi-Rad