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

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

data/sube | Department of History | University of Southern Denmark



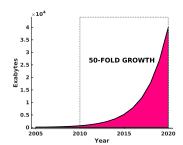


PROGRAM

0.00	Data, OD & initial concerns	deluge, challenges and copyright
0.10	GDPR and scientific research	safeguards, exemptions, pseudonymization
0.20	Discussion	





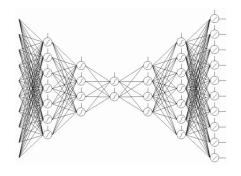


- the data deluge is transforming knowledge discovery and understanding in every domain of human inquiry
- knowledge discovery depends critically on advanced computing capabilities
 - a large part of these data are unstructured and fundamentally cultural
- we have limited experience with computational 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 controllers

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 & (to some extend) data protection policies
- 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 could 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 (\sim override rights) \rightarrow 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







GDPR scientific research

content

- broad definition with a health science bias (health, genomic, and biometric data)

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)

exemption 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



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
- right to be forgotten (delete and provide)
- parent consent

strategies

- 'walled garden' of Terms and Conditions
- 'Silence, pre-ticked boxes or inactivity' is inadequate to confer consent
- exempt from notice requirement, if data are from publicly available sources



 $^{{\}sf -}$ or why you don't have to worry about GDPR's old-fashioned approach to cloud computing



⁻ cheap, scalable compute and data management for research projects @ SDU

THANK YOU

knielbo@sdu.dk knielbo.github.io

& THANKS TO

Culture Analytics @ Institute of Pure and Applied Mathematics (UCLA)

Tina Eliassi-Rad

