

University of Groningen

## Computational methods for data discovery, harmonization and integration

Pang, Chao

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*

Publisher's PDF, also known as Version of record

*Publication date:*

2018

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*

Pang, C. (2018). *Computational methods for data discovery, harmonization and integration: Using lexical and semantic matching with an application to biobanking phenotypes*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen.

### Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

### Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

## Propositions

1. The fact that we use human language when capturing scientific data inevitably introduces heterogeneity.
2. To realize the promise of personalized medicine we need to bridge heterogeneity and enable large scale integrated analysis .... but
3. Manually harmonizing biobank data to enable integrated analysis is (too) complex and time-consuming (bioshare consortium).
4. Full automation of data harmonization not yet possible because computational representation of knowledge is incomplete .... however
5. Semi-automatic systems allow users to more efficiently harmonize data and generate high quality training data for machine learning approaches.
6. Machine learning promises the ultimate solution to enable full automation for the harmonization challenges.
7. Healthcare data needs to be coded using standard vocabularies or ontologies to unleash its values.
8. Implementation of the FAIR principles is essential to enable discovery and reuse of scientific knowledge and data as a basis for reproducible science.
9. The difference between a data scientist and a data engineer is the understanding of the domain knowledge.
10. "If we want to harmonize data, we need to harmonize people first."  
(BioSHaRE consortium)
11. "A shared beer always tastes better" (Oscar Wagner)