Computational methods for data discovery, harmonization and integration

Pang, Chao

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)