

# PREREQUISITES FOR DATA ANALYTICS AND AI

ÁKOS BERNARD JÓZWIAK



## OUTLINE

# OUTLINE

- What do we need to have before actually analysing data?
- Focusing on data (in the food safety domain)
  - Missing data
  - Missing standards, catalogues
  - Importance of ontologies
  - Machine readable data
- What can we do?



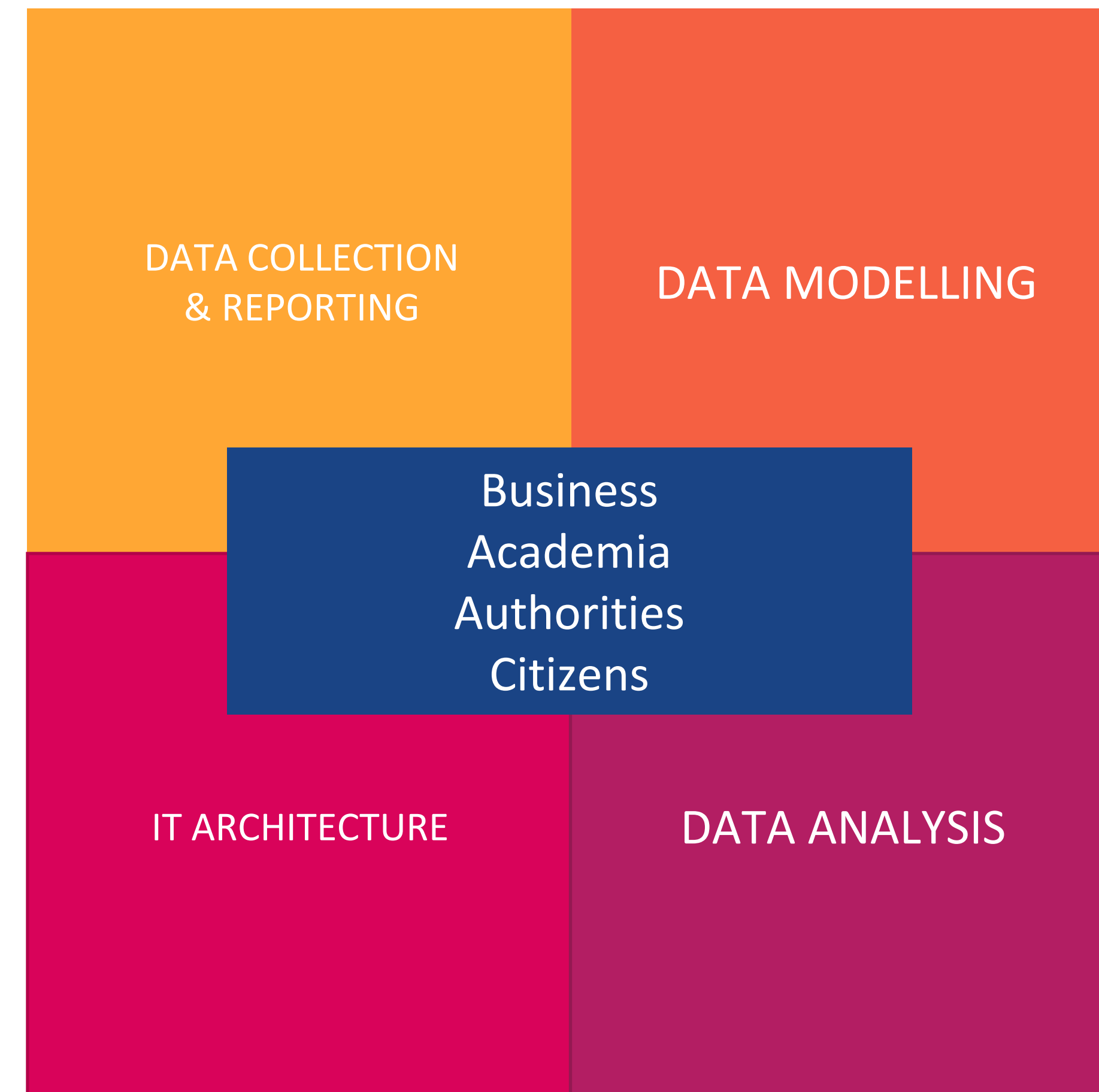
PREREQUISITES

# WHAT DO WE NEED?

Data



People



## PREREQUISITES

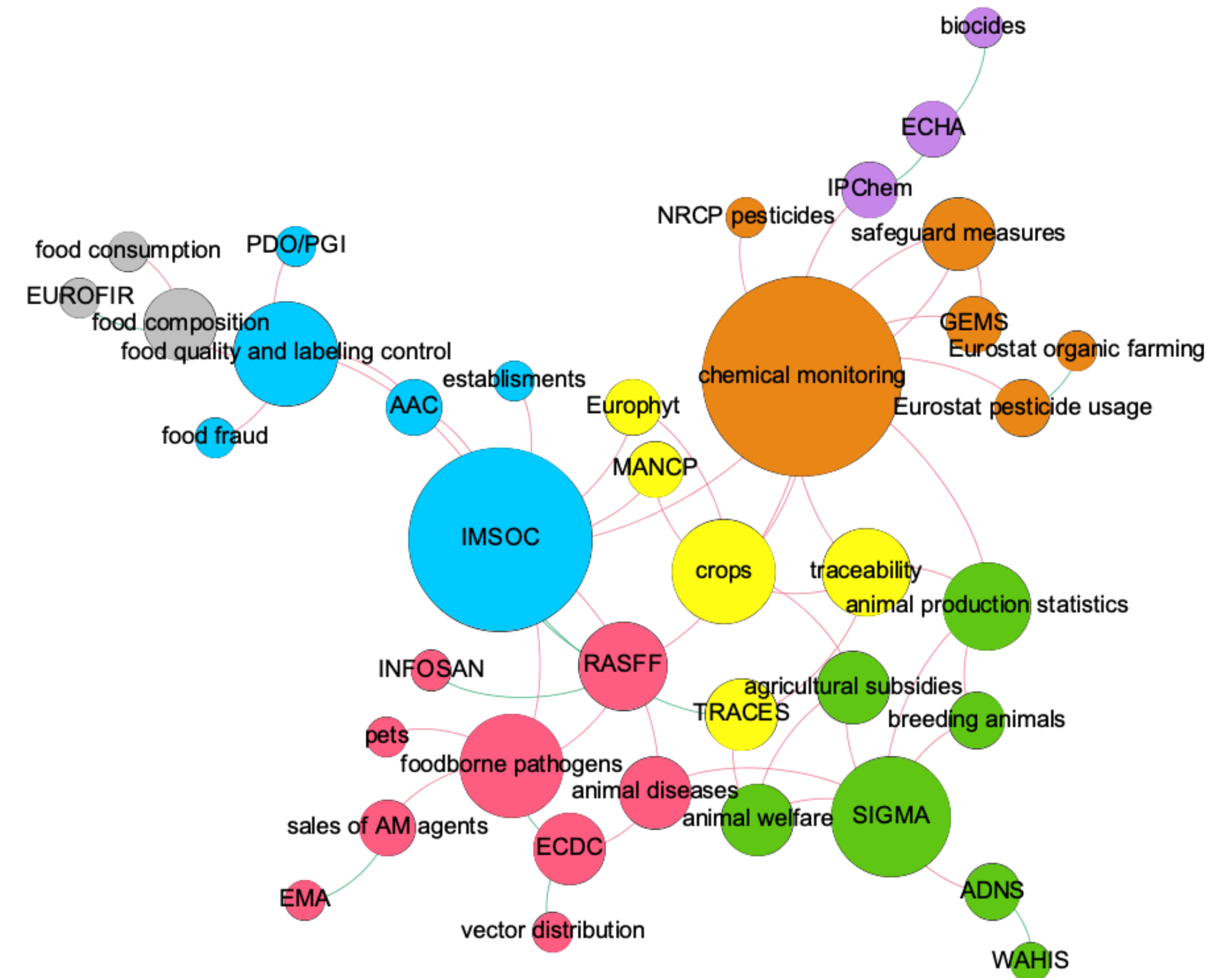
# DATA

- Quality (Completeness, Validity, Uniqueness, Timeliness, Consistency, Accuracy)
- Quantity?
- Granularity?
- Representativity?
- Interoperability?
- ...

## EFSA ADVISORY GROUP ON DATA

# EFSA ADVISORY GROUP ON DATA

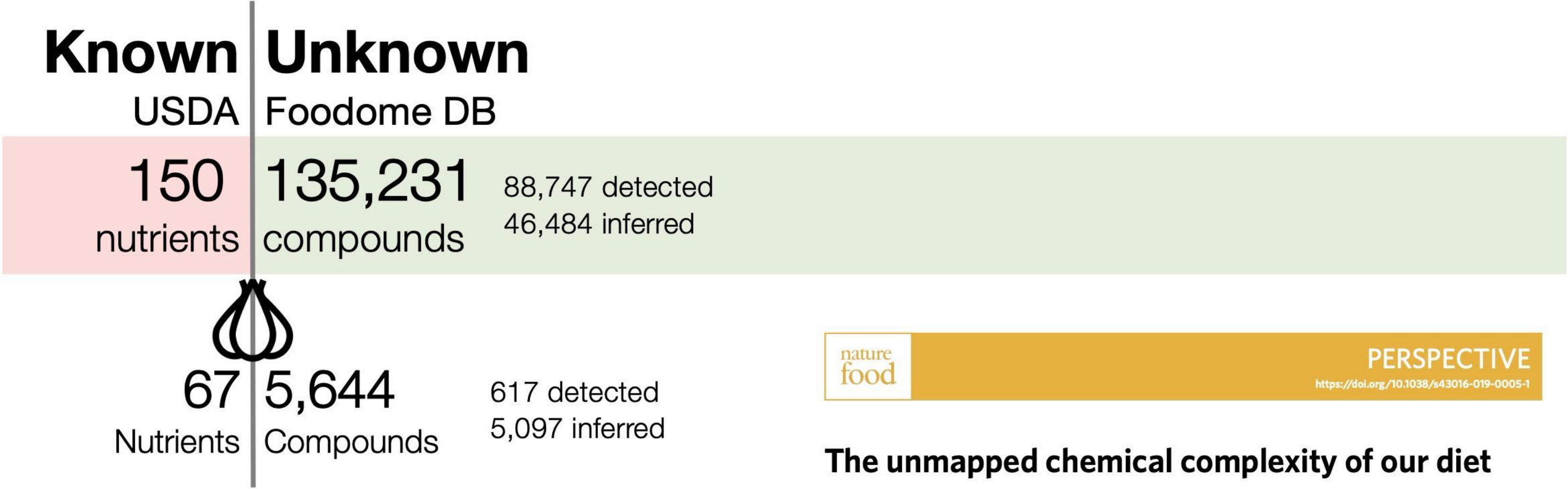
- Act as a governance body providing recommendations
  - <https://doi.org/10.2903/sp.efsa.2020.EN-1901>
- Act as a Think Tank providing input on project ideas
- Act as a channel providing access to knowledge, expertise, competencies and staff in Member States
- Provide strategic input on and oversight of alignment of EFSA's data roadmap



EXAMPLES

# MISSING DATA

- Molecular level food composition data



**PERSPECTIVE**  
<https://doi.org/10.1038/s43016-019-0005-1>


## The unmapped chemical complexity of our diet

Albert-László Barabási<sup>1,2,3\*</sup>, Giulia Menichetti<sup>1</sup> and Joseph Loscalzo<sup>2</sup>

Our understanding of how diet affects health is limited to 150 key nutritional components that are tracked and catalogued by the United States Department of Agriculture and other national databases. Although this knowledge has been transformative for health sciences, helping unveil the role of calories, sugar, fat, vitamins and other nutritional factors in the emergence of common diseases, these nutritional components represent only a small fraction of the more than 26,000 distinct, definable biochemicals present in our food—many of which have documented effects on health but remain unquantified in any systematic fashion across different individual foods. Using new advances such as machine learning, a high-resolution library of these biochemicals could enable the systematic study of the full biochemical spectrum of our diets, opening new avenues for understanding the composition of what we eat, and how it affects health and disease.

## EXAMPLES

# REPRESENTATIVITY PROBLEMS

- Sampling strategies:
  - objective (i.e., random)
  - selective (i.e., risk-based)
  - suspect
  - (convenient)

*statistically limited interpretability /  
biased results*
- Challenges:
  - Central level random sampling plan, executed on a risk basis locally: what strategy is reported then?
  - Many questionable reporting practices, inconsistencies
  - E.g.: Veterinary drug residues sampling programmes

## EXAMPLES

# MISSING / MISALIGNED STANDARDS

- Can we link RASFF data with EFSA contaminants and consumption data?
- Not yet (although EFSA and COM are working on it)

RASFF (own) catalogues on food categories and hazards

≠

EFSA catalogues on food categories (FoodEx2) and hazards (PARAM)

## EXAMPLES

# IMPORTANCE OF ONTOLOGIES

- Ontology: a generalized, semantic data model
- Research projects aiming for utilising data for better food systems safety: connecting various (open source) data with the help on ontologies and common identifiers
- Need for standardised, interoperable ontologies
  - Food classification: FoodON is the one used by the research community, not FoodEx2. Is it fine for EFSA, COM, MS authorities?
  - Inter-agency exchange of chemical contaminants data: which ontology to choose?
  - No common international ontology of animal diseases
  - ...

## EXAMPLES

# MACHINE READABLE DATA?

- Building data lakes for research and/or control purposes
  - Need for interoperable, connected ontologies
  - Easy to access data (Repositories, direct database access, API, ...)
  - FAIR (Findable, Accessible, Interoperable, Reusable)
- Do we have that?

**Foodome project:** connecting different layers

## 1. Food Composition Data

USDA (it gathers several composition databases)  
FooDB  
Scientific Literature  
Online Grocery Data

## 2. Food Classification Data/Ontology

USDA ChooseMyPlate  
Foodon  
NOVA

## 3. Population Consumption and Health Statistics (MACRO)

NHANES  
Nurses' Health Study  
UK Biobank  
FAOSTAT

## 4. Bioinformatics, Cheminformatics, and Metabolomics Databases (MICRO)

STITCH  
ChEMBL  
CTD  
PubChem  
MetaboLights

## PREREQUISITES

# PEOPLE

- Creation and development of (big) databases is not only an IT problem
- The ability of analysis and evaluation of *input data* and *results*: high-level knowledge of food chain science is needed enabling interpretation and validation
- Data literacy
  - Basic statistics is in the food safety risk assessment curricula
  - But data science is not (or very rare)
  - Future (or current) risk assessors need data generation, retrieval, manipulation and analysis knowledge

## OUTLOOK

# WHAT CAN WE DO?

- Investing in data generation
- Building ontologies
- Sharing tools, standards, data, models...
- Publishing, open data standards
  - e.g., FSKX
- Education
- Change management

## OUTLOOK

# ALSO: DATA CULTURE

1. data literacy should be encouraged to spread
2. strong analytical skills: separate from technical skills, it means the ability to ask a good question that you can answer with data
3. statistical/technical skills: not everybody needs to be a data scientist, but organizations do need the power of statistics
4. data visualization
5. willingness to learn: willing to make mistakes and learn from them
6. mentoring
7. data storytelling: make complex data more accessible



# THANK YOU FOR YOUR ATTENTION

## CONTACT

Ákos Józwiak

Research Director | Digital Food Institute, University of Veterinary Medicine Budapest

Head of Food and Nutrition Science | Syreon Research Institute

[akos.jozwiak@syreon.eu](mailto:akos.jozwiak@syreon.eu)

LinkedIn: [akosbernardjozwiak](https://www.linkedin.com/company/akosbernardjozwiak)

<https://syreon.eu>

<https://dfi.univet.hu/en/>