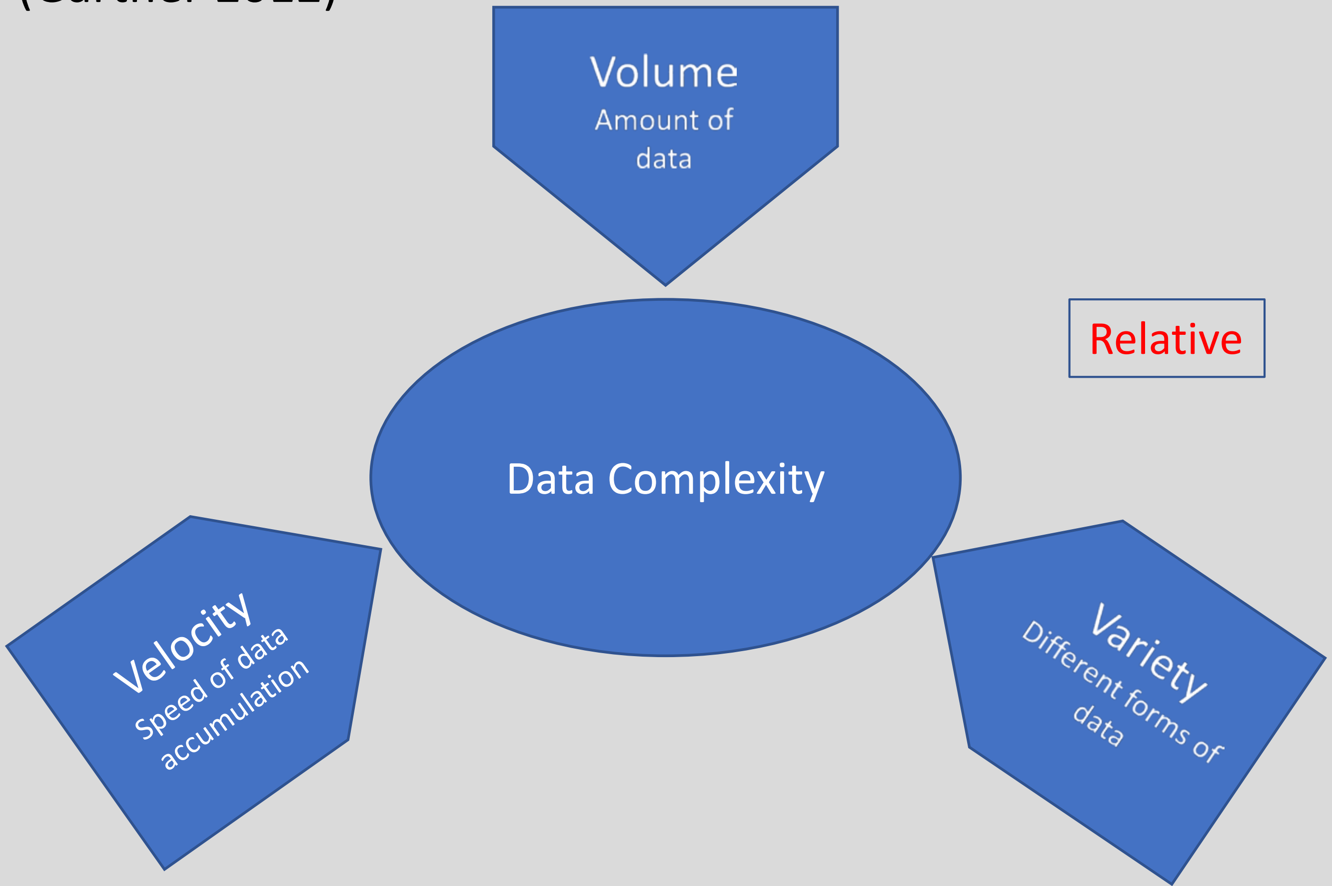
Future use of Big Data in animal and plant breeding

Peer Berg

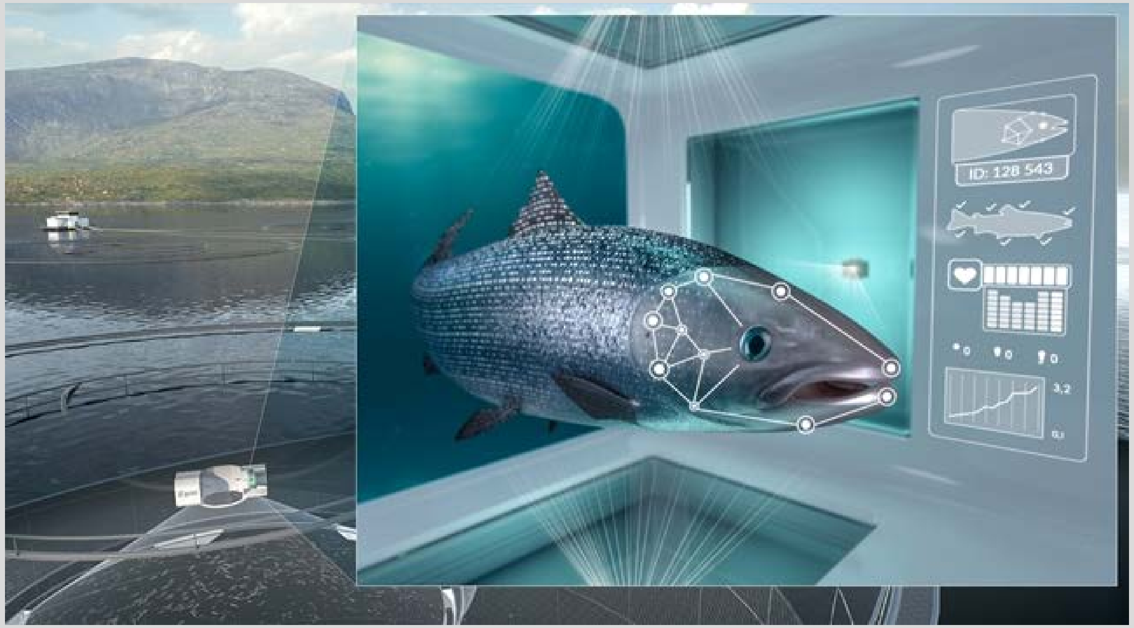
What is Big Data?

- Big data is high-volume, high-velocity and/or high variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation. (Gartner 2012)
- The 3 Vs.



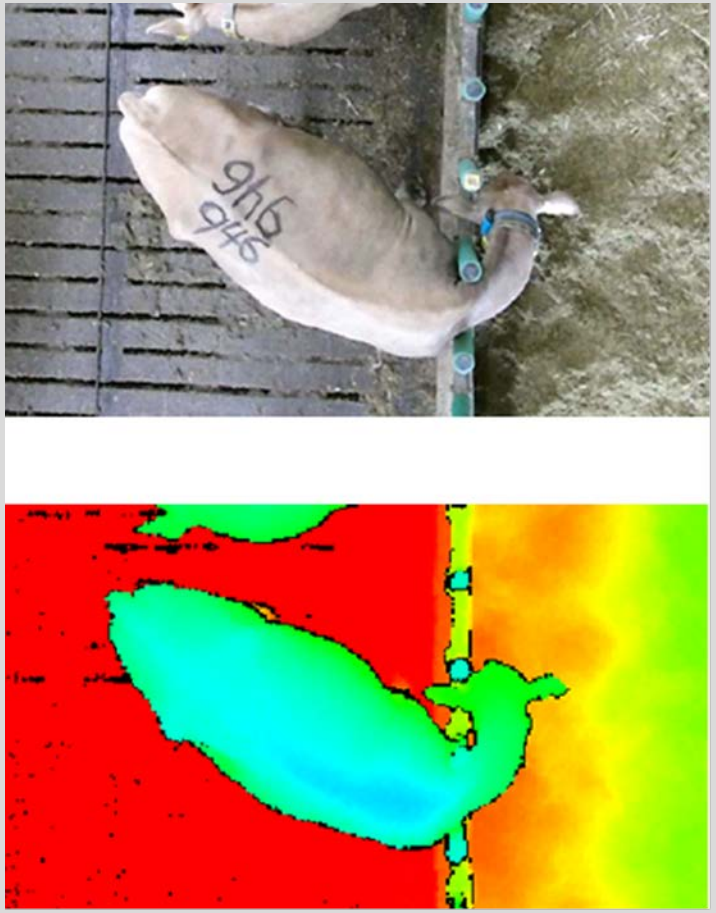
Examples

Pig Atlas, Topigs-Norsvin

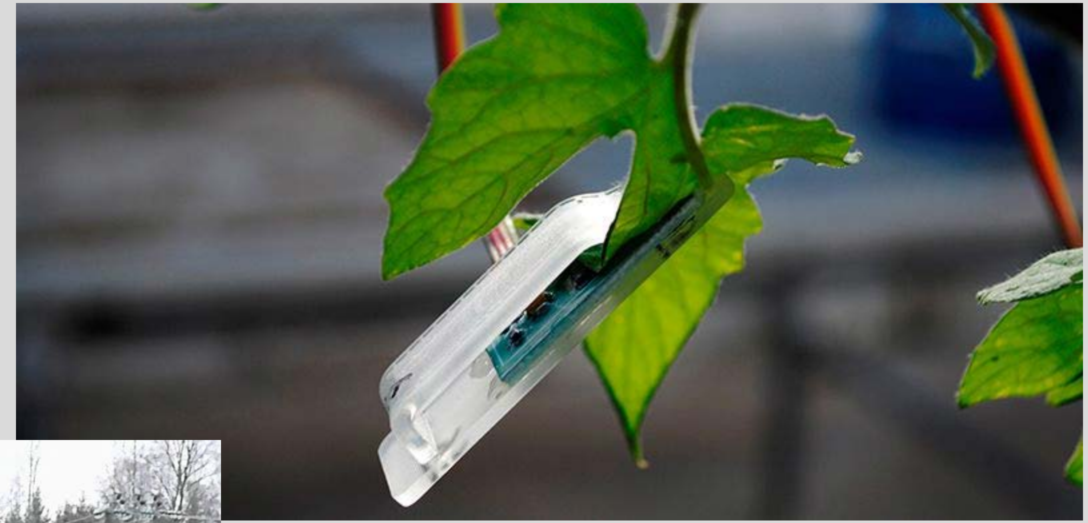


iFarm, www.cermaq.com

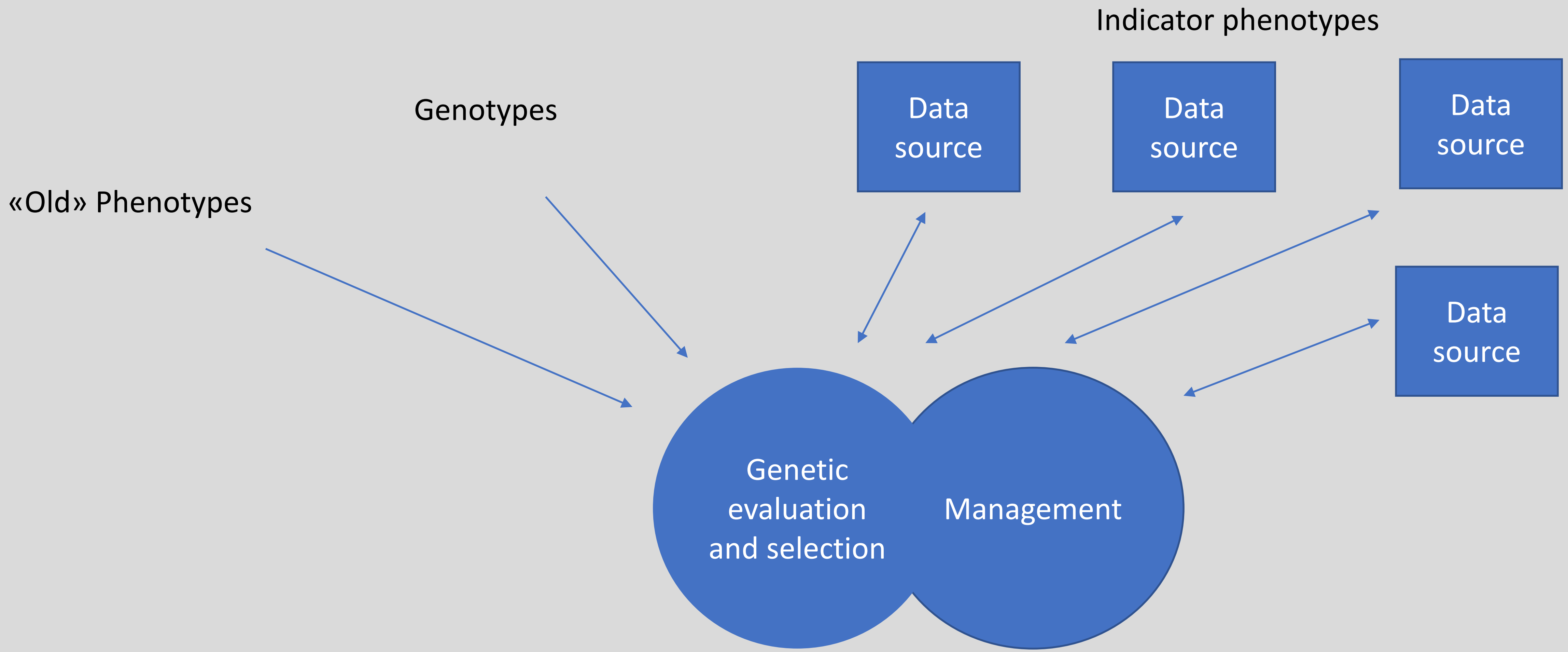
Examples



Feed intake, Viking Genetics



Individual plant measurements, Amin Afzal


Indicator phenotypes
Genotypes
«Old» Phenotypes
Data source
Data source
Data source
Data source
Genetic evaluation and selection
Management

Big Data in breeding

Precision animal breeding

.. the aim should be to meet certain goals:

- to improve the precision with which breeding outcomes can be predicted;
- to avoid the introduction and advance of characteristics deleterious to well-being; and
- to manage genetic resources and diversity between and within populations as set out in the Convention on Biological Diversity.

Flint & Woolliams 2008

Potentials

Increase accuracy
- More information
- Better modelling

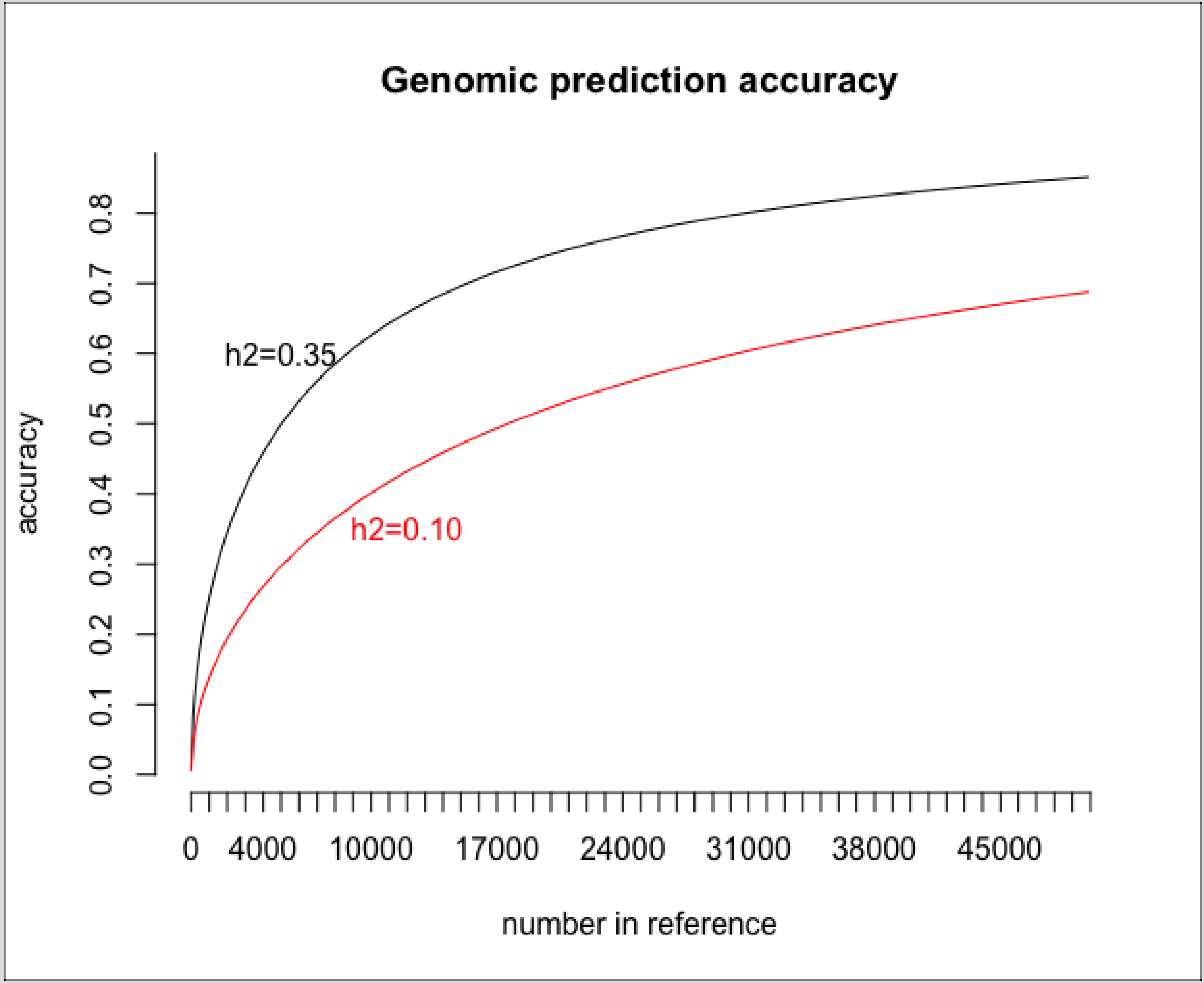
Increasing i:
- Multi-stage to single stage selection

Increase genetic variance
- Traits ignored
- Better modelling

$$\Delta G = \frac{i \cdot r_{IA} \cdot \sigma_A}{L}$$

Reducing L:
- Small options in livestock
- Some options in plants

GS



Genomic prediction accuracy
accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
h2=0.35
h2=0.10
0
4000
10000
17000
24000
31000
38000
45000
number in reference

Accuracy

Direct response

$$\Delta G = \frac{i \cdot r_{IA} \cdot \sigma_A}{L}$$

Traits difficult to measure in large numbers

Indirect response

$$\Delta G_{X|Y} = \frac{i \cdot r_{IA} \cdot r_g \cdot \sigma_{Ax}}{L}$$

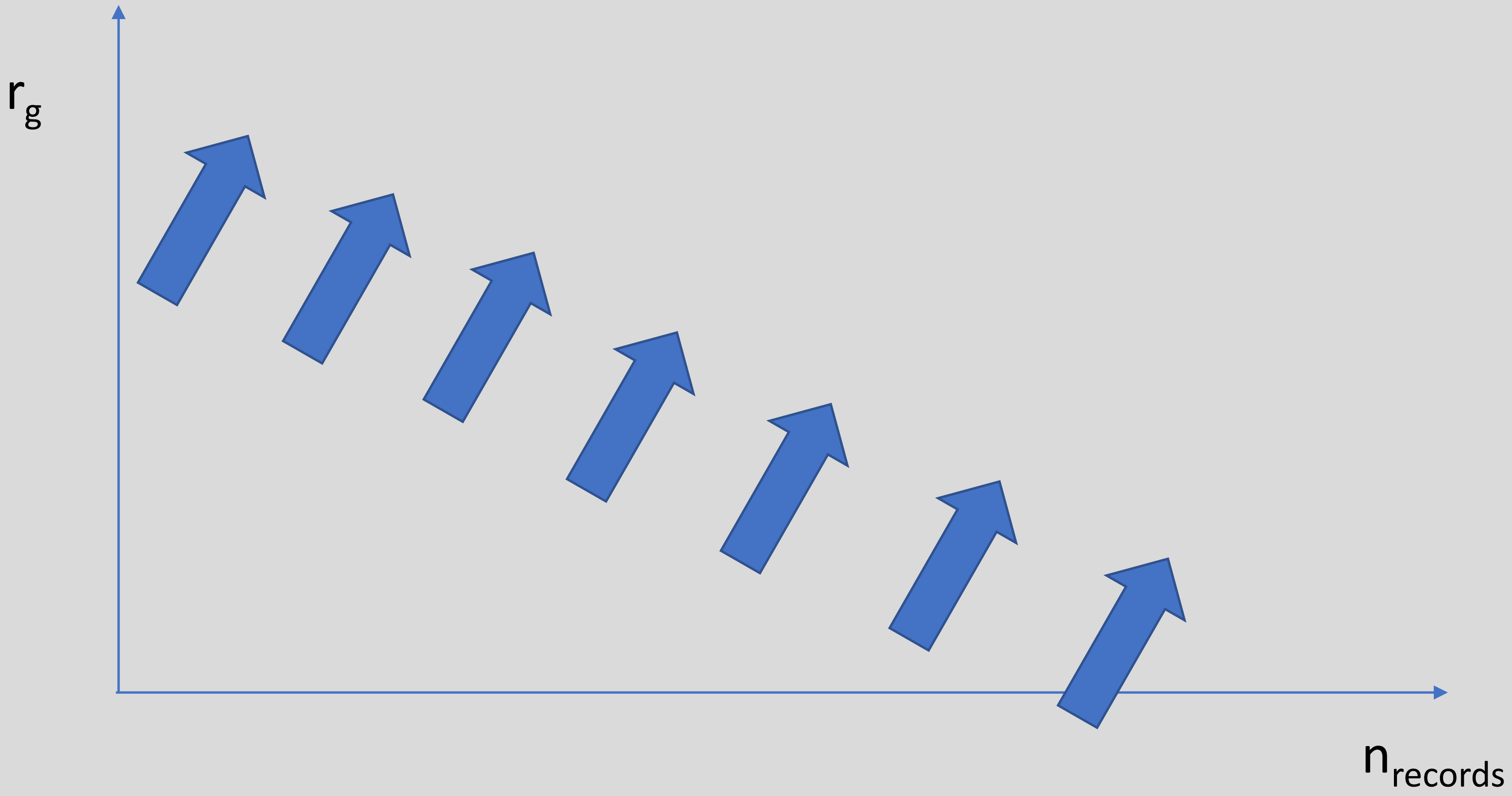
Surveilance of traits not selected for

Better modelling

- Reduce environmental variance
  - Disease resistance
    - Modelling risk
    - Resistance vs. Tolerance
- Environmental variance -> genetic variance
  - Indirect genetic effects
    - Indirect effect models
      - Simplistic
    - Measurering interactions
      - Increased total genetic variance
      - More accurate indirect effects

Not infected
Or
Resistance/Tolerant

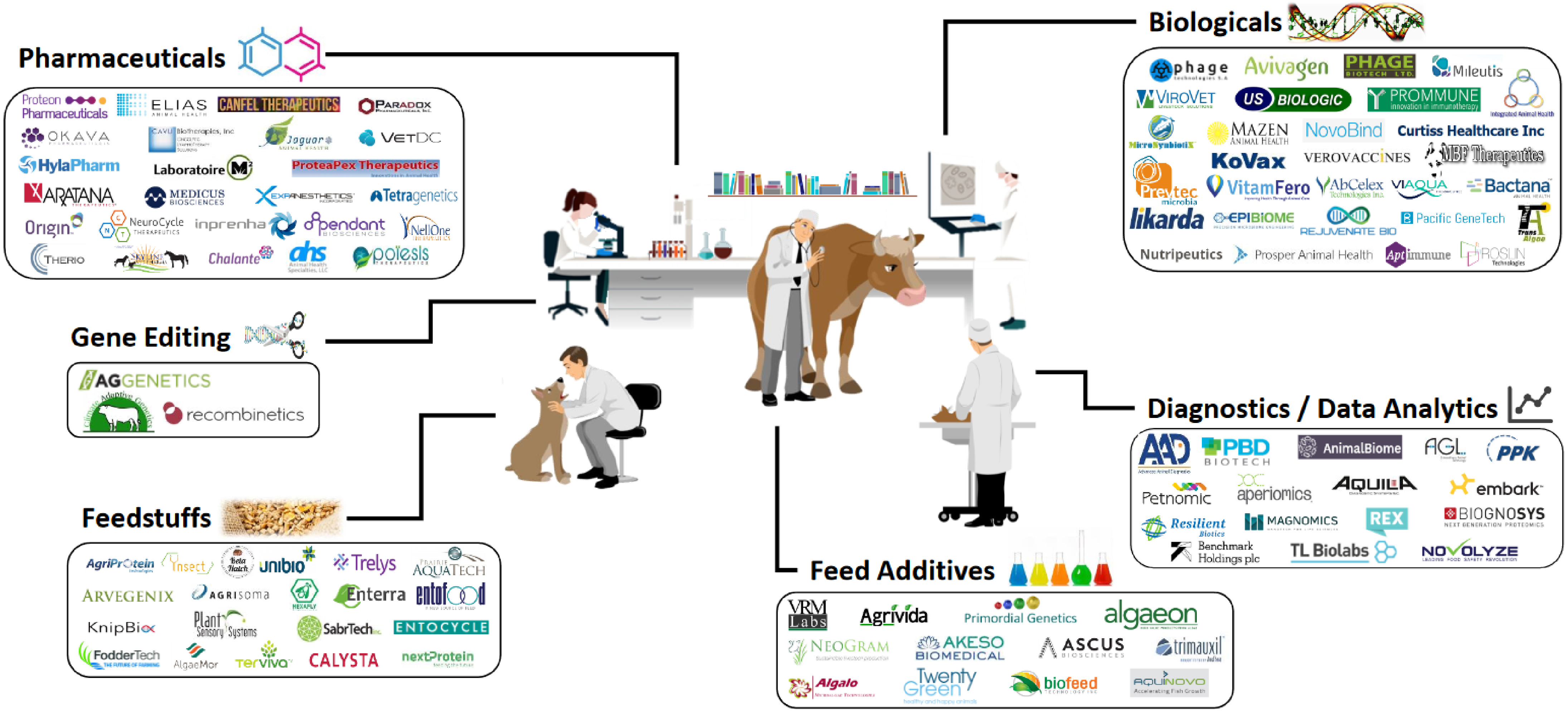
Infected
Susceptiple


r_g
$n_{records}$

Whats happening?

Animal Biotech Landscape 2017

TechAccel
Technology Acceleration Partners

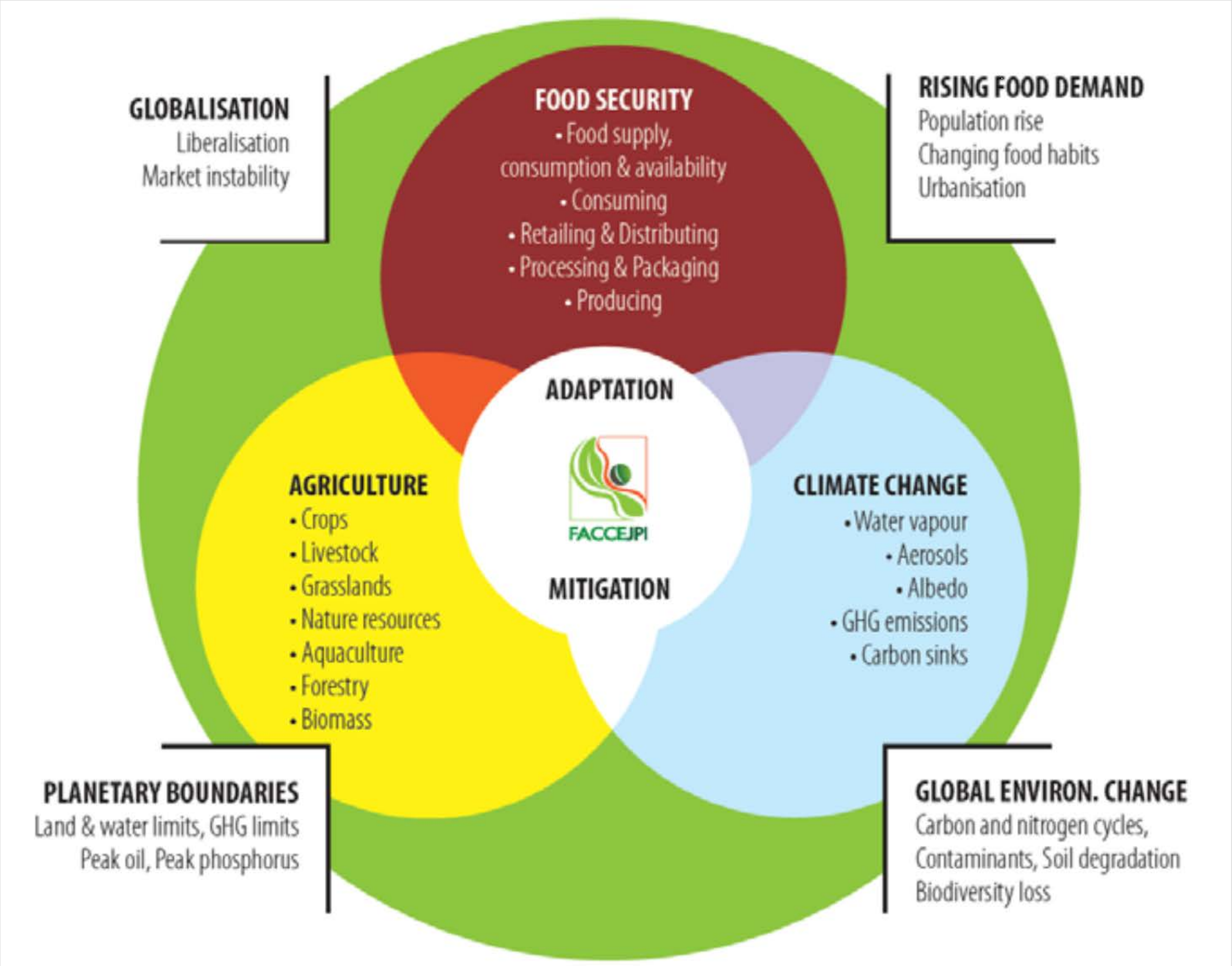


ow.ly/PhWF30ixuST

FACCE-JPI

- provide coherence in research programming across Europe to meet the societal challenge of *jointly* ensuring
  - food security,
  - adaptation to climate change impacts and
  - mitigation of greenhouse gas emissions.

23 countries + NZ



FACCE-JPI Workshops 2017

- **Workshop on Technologies**
  - **Fostering the adoption of existing (and emerging) technologies for primary production in the context of climate change that are on the edge of being mature but not yet widely adopted**
- **Workshop on Big Data**
  - **To explore research needs and research gaps**
  - **To identify potential application and integration (reassembling) of relevant new and existing data**
  - **To maximise impact in FACCE - JPI projects through use of existing data**
  - **To identify infrastructures and tools to be used by FACCE - JPI at joint action level**

New Technologies

- New breeding technologies / gene editing
- Automated phenotyping
- Veterinary / health
- Precision Livestock Farming
- Next generation feed
- Housing and manure management
- wrt.
  - Research, Networks, Infrastructure, Integration of farmers, Integration of consumers and society, Innovation and integration of industry

Integration

Required to deposit data in Research data repositories

Experiences

- Big Data in Agriculture
  - DuPont Sciences Symposia Series, Edinburgh, May 2018 - https://www.bigdataag2018.org/
  - Start-up companies
    - Comparative and predictive solutions based on publicly available data and in-house data (supplied by customers)
    - «Data Silos»
- Mimiro
  - Daughter company of Tine and Felleskjøpet
  - Storing or linking all (dairy) relevant data
  - Commercial focus on selling applications and access to data.

Big Data - agriculture

- Many “data owners”
- Diverse sources

- Challenges
  - Variety - Integration
  - Volume of data
  - Velocity

- **Public**
  - **Open Access**
- **Commercial**
  - **Business models**
    - **Reference/calibration**
    - **Knowledge exchange**
    - **Pay for data**
- **Farmers**
  - **Motivation**
    - **Decision making**
    - **Bench marking**
    - **Incentives (subsidies, right to produce, deregulation)**

Data sources

Research data → Research repository

Commercial data → Ownership / Motivation to share / Business models

The risk: Data silos

PLOS | BIOLOGY

PERSPECTIVE

Domesticated Animal Biobanking: Land of Opportunity

Linn F. Groeneveld[1]*, Sigbjørn Gregusson[2], Bernt Guldbrandtsen[3], Sipke J. Hiemstra[4], Kristian Hveem[5], Juha Kantanen[6,7], Hannes Lohi[8], Lina Stroemstedt[9], Peer Berg[1]

Interbull
Eurogenomics

Many initiatives

- **F**indable
- **A**vailable
- **I**nteroperable
- **R**eusable

- Research Data Alliance
- Agricultural Data Interest Group
- On-farm data sharing WG
- European Open Science Cloud
- GODAN (600+ partners)
- ELIXIR – Norwegian node
- CGIAR Platform for Data in Agriculture
- FAIRDOM
- …

Technologies

- HADOOP
  - Distributed file storage and analysis
- Machine learning
  - Data reduction
  - To assign priors to SNP effects (Perez-Enciso 2017)
  - Lack of training data
    - Lack of records
    - Lack of clearly defined phenotype
      - Disease resistance
      - Robustness

NEWS · 01 NOVEMBER 2018

Machine learning spots natural selection at work in human genome

Scientists are using artificial intelligence to identify genetic sequences molded by evolutionary pressures.

Nature **563**, 167 (2018)

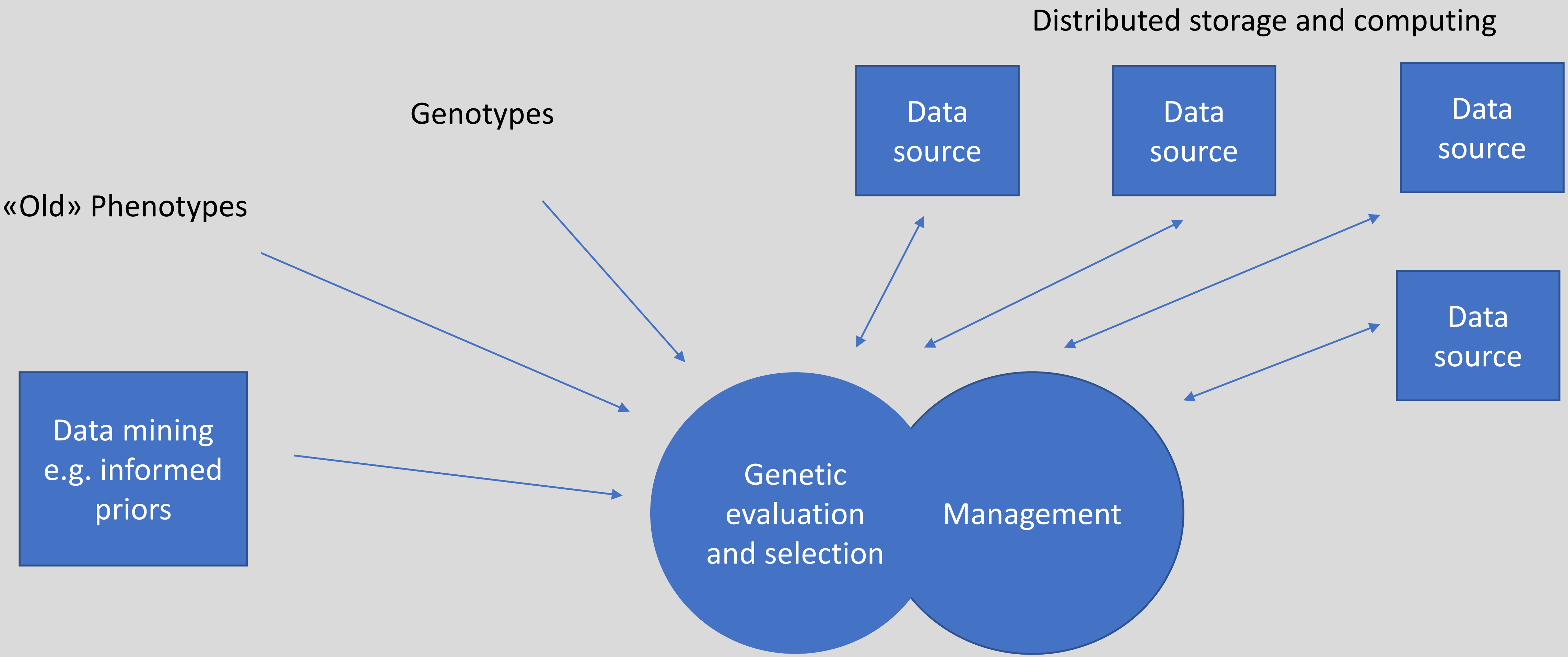
Deep Learning (DeepSweep) training on simulated data → 20.000 potential nucleotides under selection → Lactase gene is marked

Table 4
State of the art of Big Data applications in Smart Farming and key issues.

Stages of the data chain	State of the art	Key issues
Data capture	Sensors, Open data, data captured by UAVs (Faulkner and Cebul, 2014) Biometric sensing, Genotype information (Cole et al., 2012) Reciprocal data (Van 't Spijker, 2014)	Availability, quality, formats (Tien, 2013)
Data storage	Cloud-based platform, Hadoop Distributed File System (HDFS), hybrid storage systems, cloud-based data warehouse (Zong et al., 2014)	Quick and safe access to data, costs (Zong et al., 2014)
Data transfer	Wireless, cloud-based platform (Karim et al., 2014; Zhu et al., 2012), Linked Open Data (Ritaban et al., 2014)	Safety, agreements on responsibilities and liabilities (Haire, 2014)
Data transformation	Machine learning algorithms, normalize, visualize, anonymize (Ishii, 2014; Van Rijmenam, 2015)	Heterogeneity of data sources, automation of data cleansing and preparation (Li et al., 2014)
Data analytics	Yield models, Planting instructions, Benchmarking, Decision ontologies, Cognitive computing (Van Rijmenam, 2015)	Semantic heterogeneity, real-time analytics, scalability (Li et al., 2014; Semantic Community, 2015)
Data marketing	Data visualization (Van 't Spijker, 2014)	Ownership, privacy, new business models (Orts and Spigonardo, 2014)

Wolfert et al. 2017

Vision


Distributed storage and computing
Genotypes
«Old» Phenotypes
Data source
Data source
Data source
Data source
Data mining e.g. informed priors
Genetic evaluation and selection
Management

Discussion points

- Vision for use of Big Data
  - Resource focus vs. Application focus
    - Requirements
    - Applications
- Hypothesis vs. data-mining driven R&D?
- Sharing of research data
  - Who owns the data we analyse?
- Who owns commercial data?
  - Farmer, technology provider, organisation
- How to get access to data?
  - Research/routine application