

# The changing nature of big data for a cooperative herd improvement organisation

ICAR 2021

Bevin Harris

# a story from the late 1990s

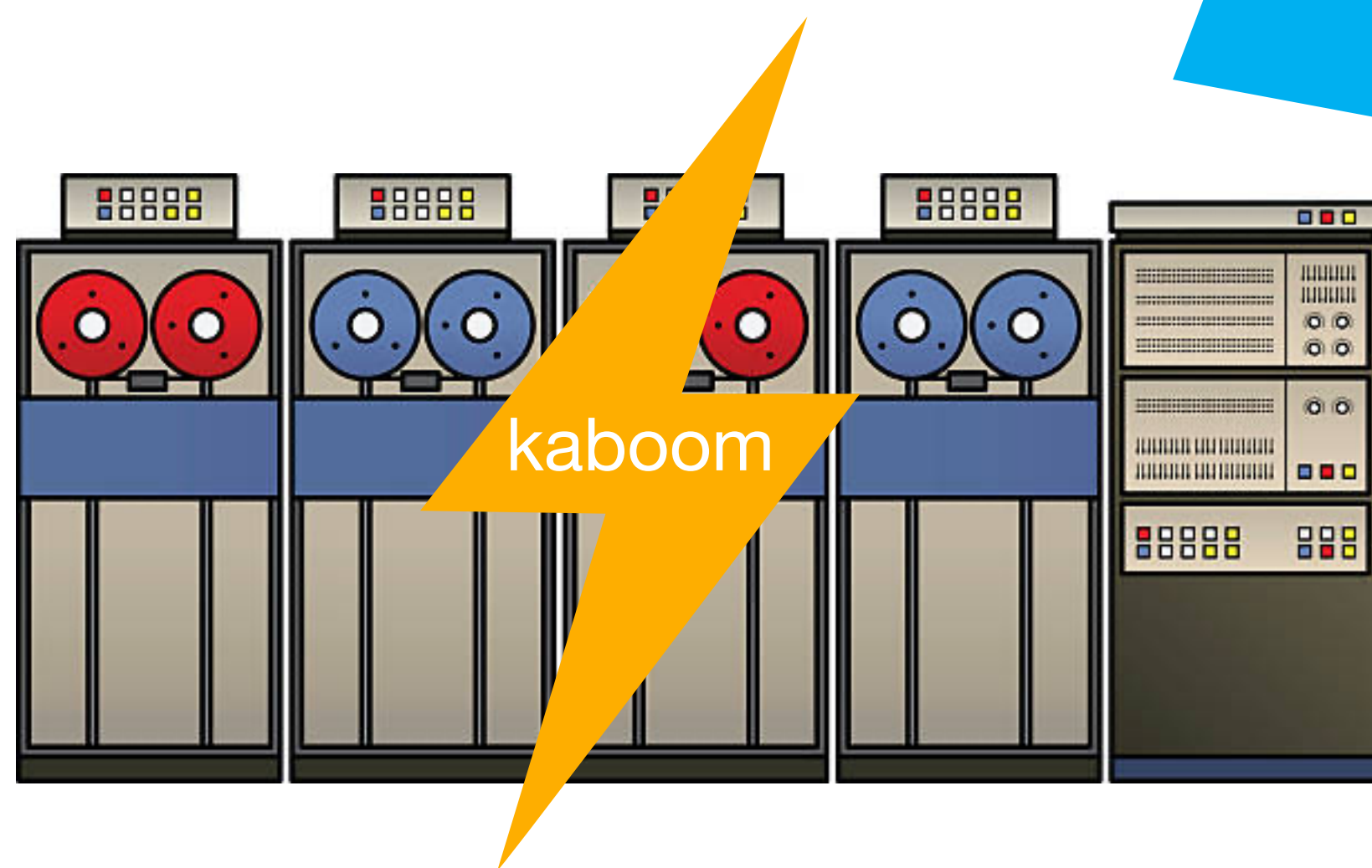
what did big data looked like in the dark ages



400 herds/day  
120,000 records/day



Data transferred to mainframe  
computer

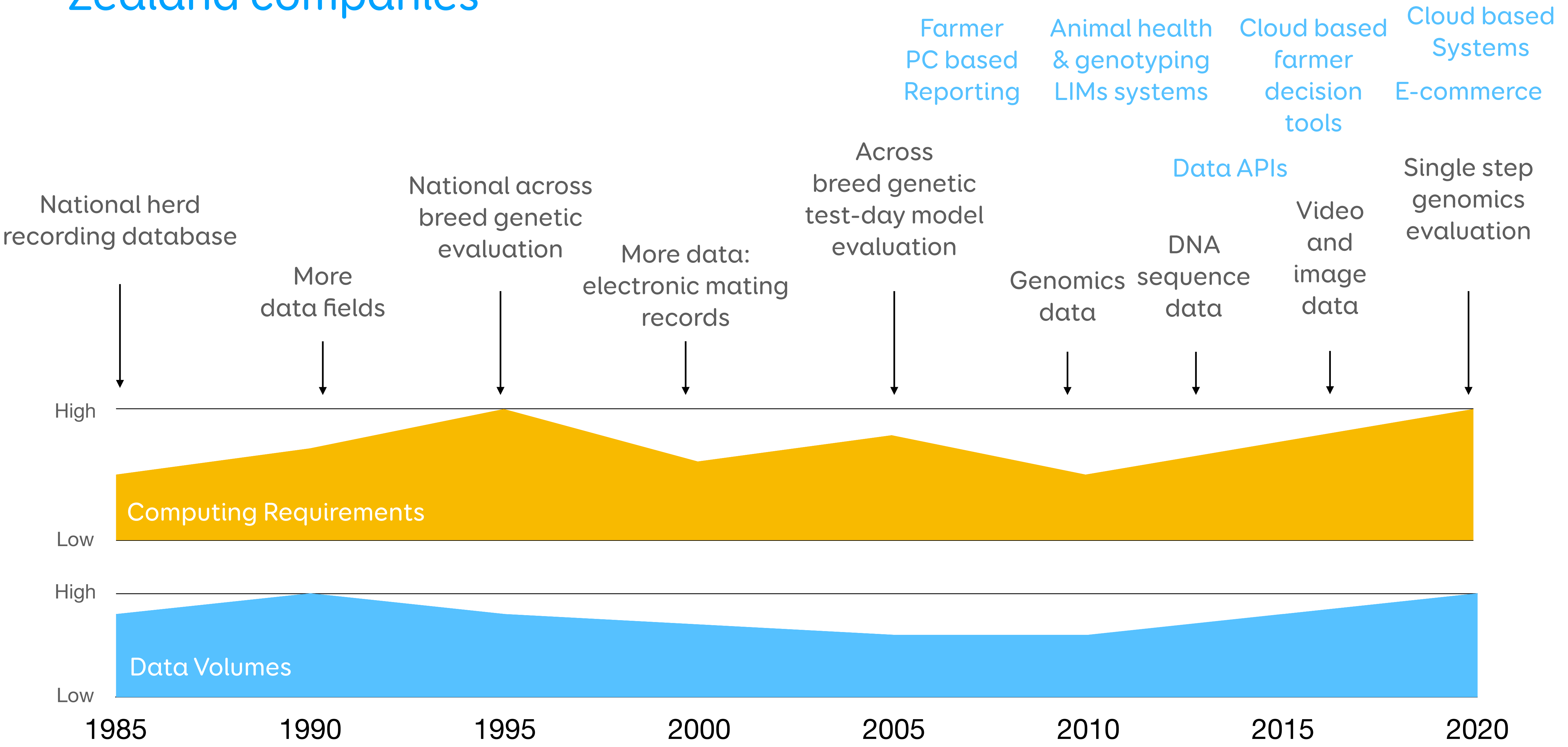


8 hours processing  
time



how times  
have changed

# A history of data and computing volumes for LIC relative other New Zealand companies



what is big  
data in 2021?

# The 5 Vs of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: *volume*, *variety* and *velocity*. Over time, other Vs have been added to descriptions of big data:

| VOLUME                                                                              | VARIETY                                                                               | VELOCITY                                                                              | VERACITY                                                                              | VALUE                                                                                 |
|-------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| The amount of data from myriad sources.                                             | The types of data: structured, semi-structured, unstructured.                         | The speed at which big data is generated.                                             | The degree to which big data can be trusted.                                          | The business value of the data collected.                                             |
|  |  |  |  |  |

# Examples of big data at LIC

From mature to embryonic data sources

**Mature:** DNA Sequence and SNP data

**Mature:** Image and video

**In-between:** Mid Infra Red spectra

**Embryonic:** Milk-omics

# DNA Sequence and SNP data

Dairy cattle breeding and genomics

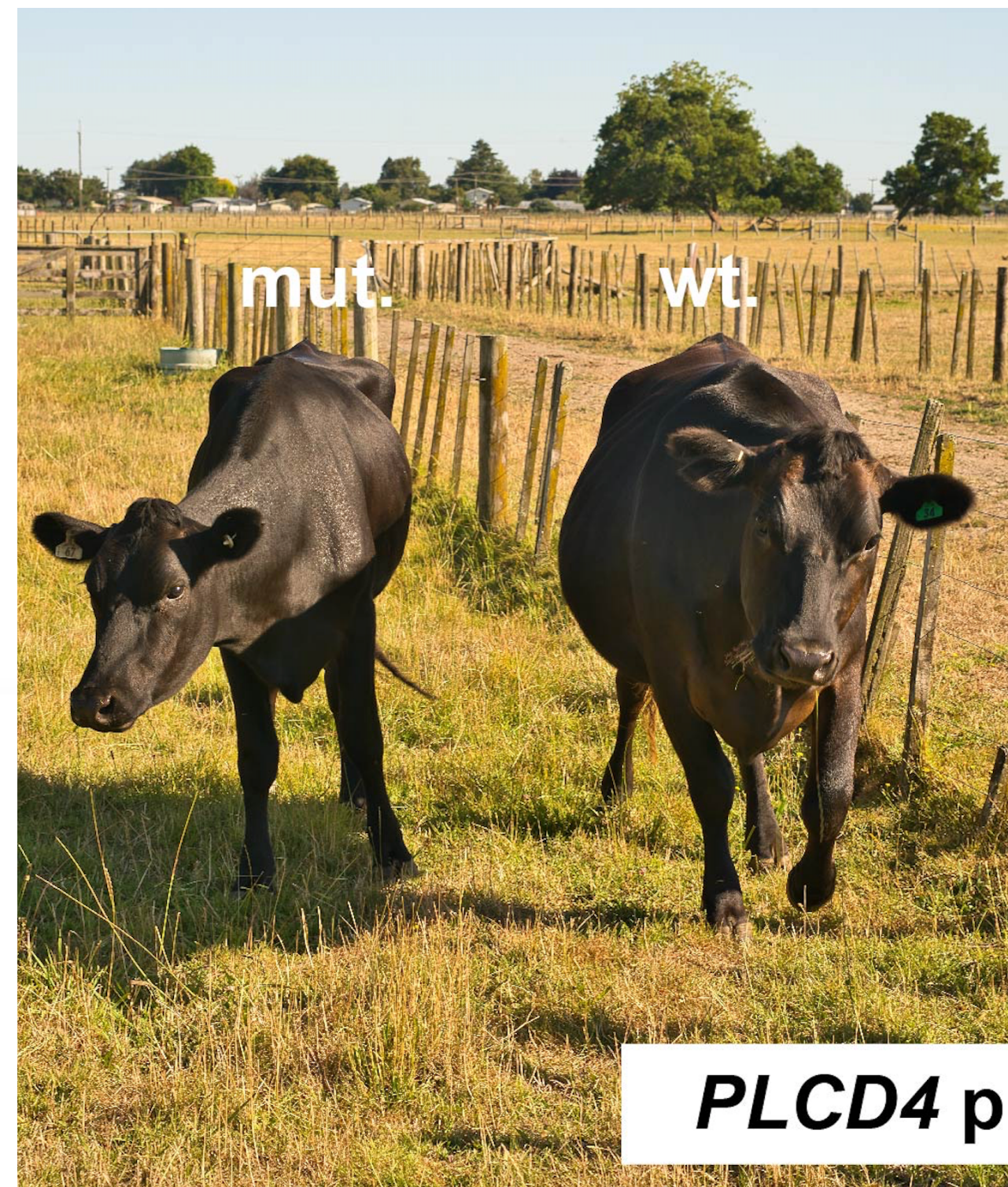
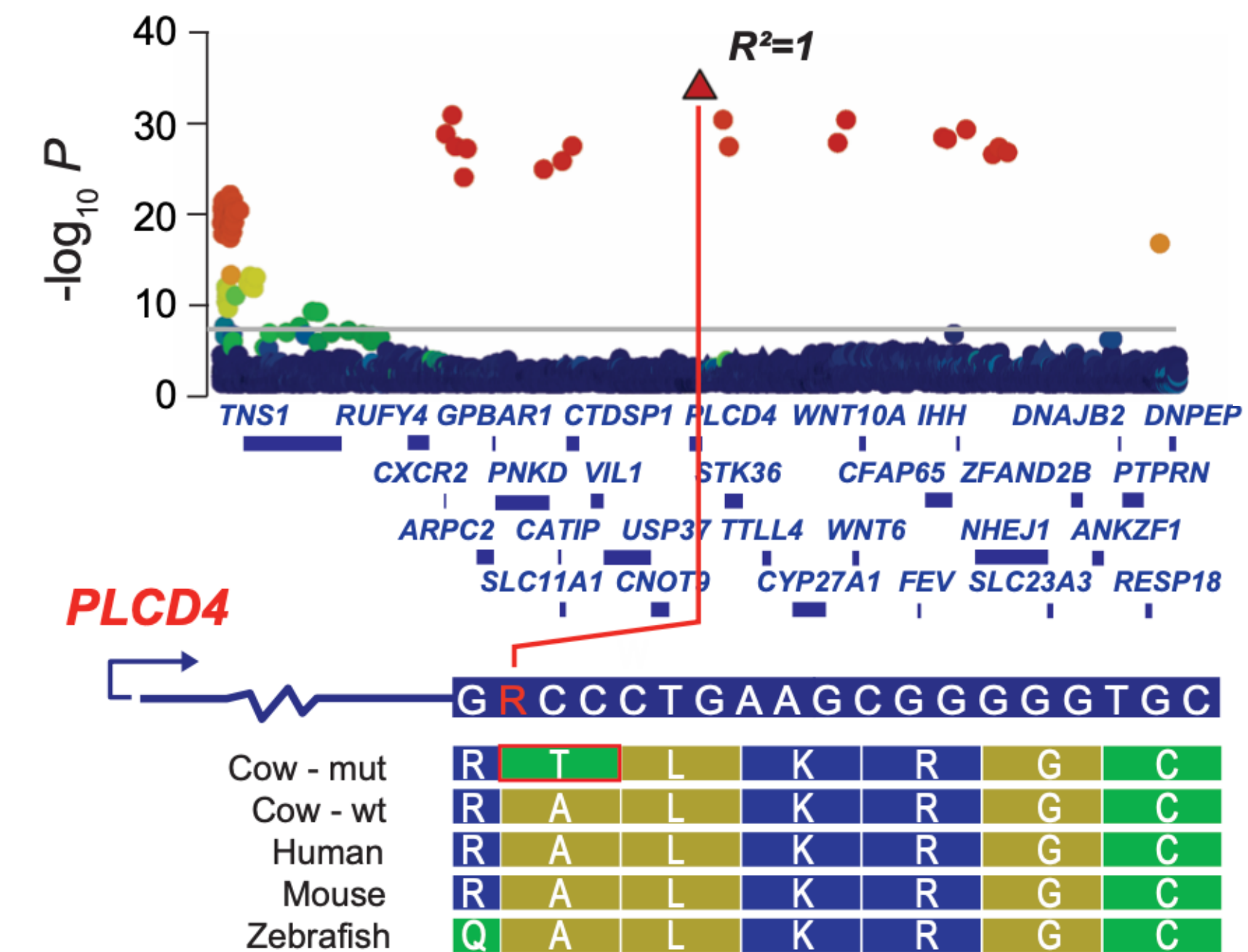
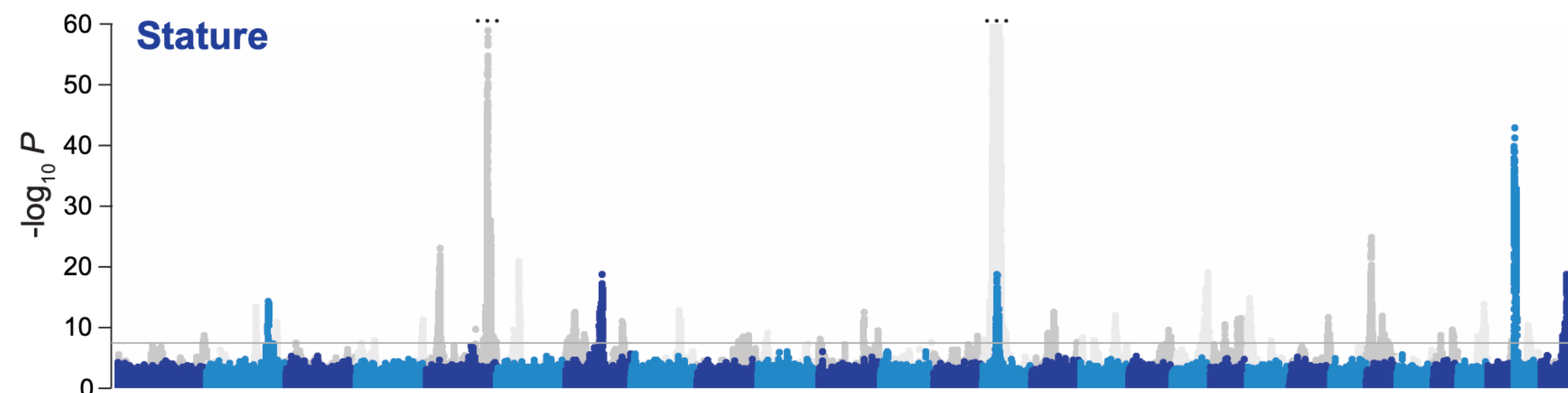
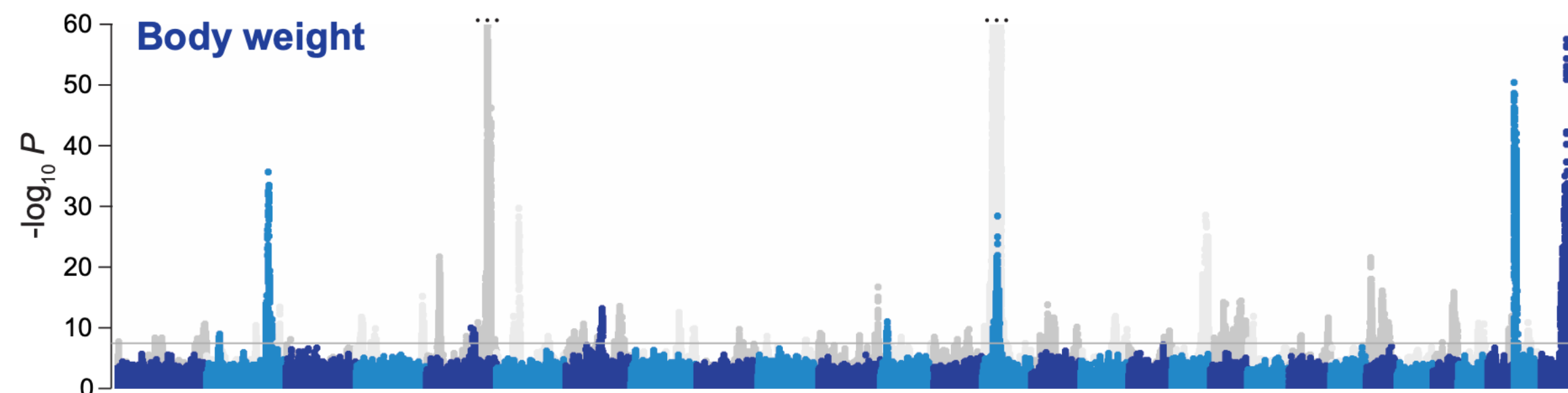
**Volume:** 18 Tb/year Sequence and of compressed data

**Variety:** Semi-structured: Ids, SNP calls, SNP manifests etc.,

**Velocity:** Sequence: large lumps (1TB) and SNP: 20,000 samples/week

**Veracity:** High quality, easy to detect data from samples with errors,  
requires a quality assurance pipeline

**Value:** Breeding programs, genomic evaluation, gene discovery and  
parentage assignment



**PLCD4 p.Ala326Thr**

# Image and video

On-farm automation

**Volume:** 1 Tbs/week

**Variety:** Unstructured: Image and video data in varying formats

**Velocity:** Continuous

**Veracity:** Variable quality requires sophisticated image quality checking pipeline

**Value:** New products for farmers to remove labour extensive tasks



# Milk Infra Red Spectra

Generating more value from a herd test sample

**Volume:** 11 million milk samples annually: 100Gb data/day: 20Tb/year

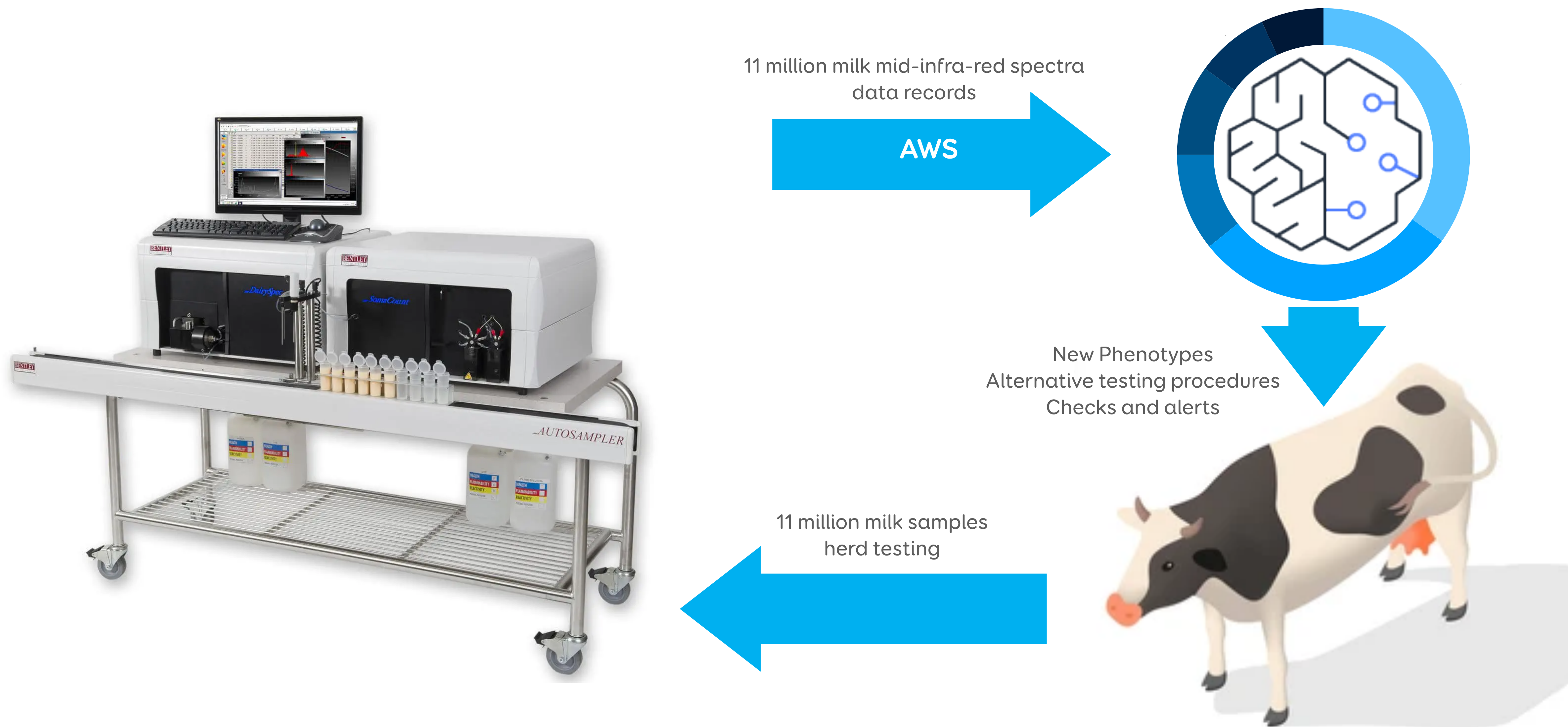
**Variety:** Semi-structured: spectra data: phenotype data:  
genomic data: cow descriptive data

**Velocity:** Daily

**Veracity:** High quality, requires data cleansing pipe to detect samples with errors

**Value:** New diagnostics from milk samples

# Extracting increased value from herd-test samples



# Milk-omics

More animal information from herd test samples – Milk microbiome

**Volume:** DNA Sequence: Multiple genomes: Variable coverage

**Variety:** Unstructured: Raw DNA sequence data: Multiple genomes

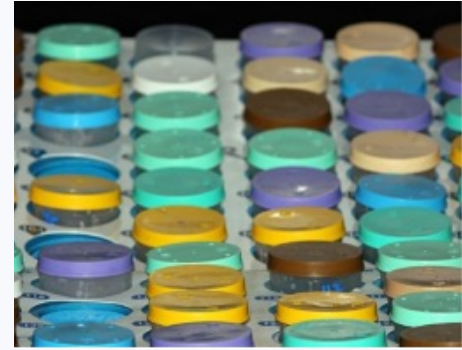
**Velocity:** 6Tb per week

**Veracity:** High quality easy to detect data from samples with errors

**Value:** Adding value to herd testing product

# Detection of all species simultaneously!

**Herd test samples**



**Bulk milk samples**



**Faecal samples**



**Effluent samples**



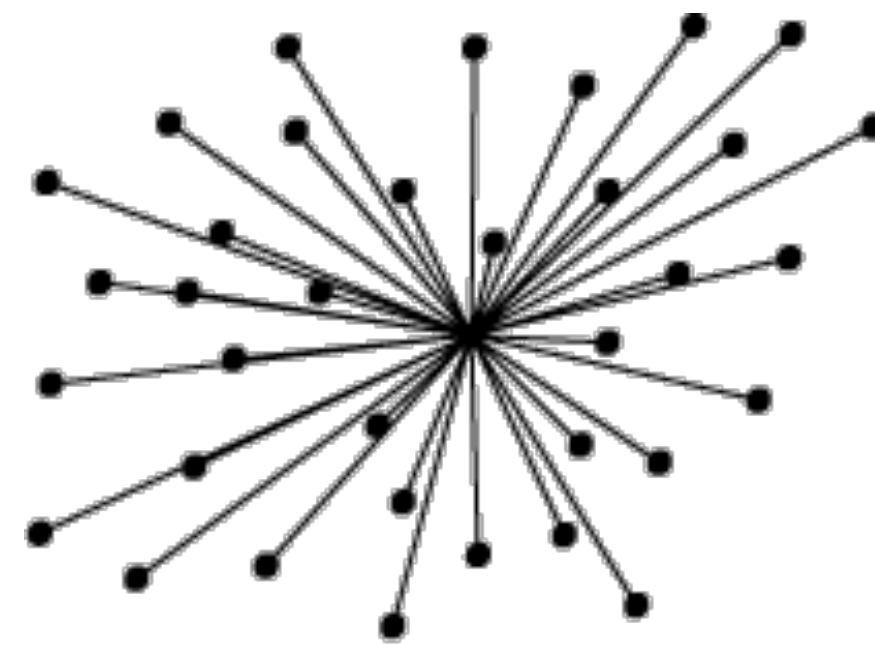
*Early illness diagnostic test development through milk and faeces-based DNA-sequencing*

# Implications to our business

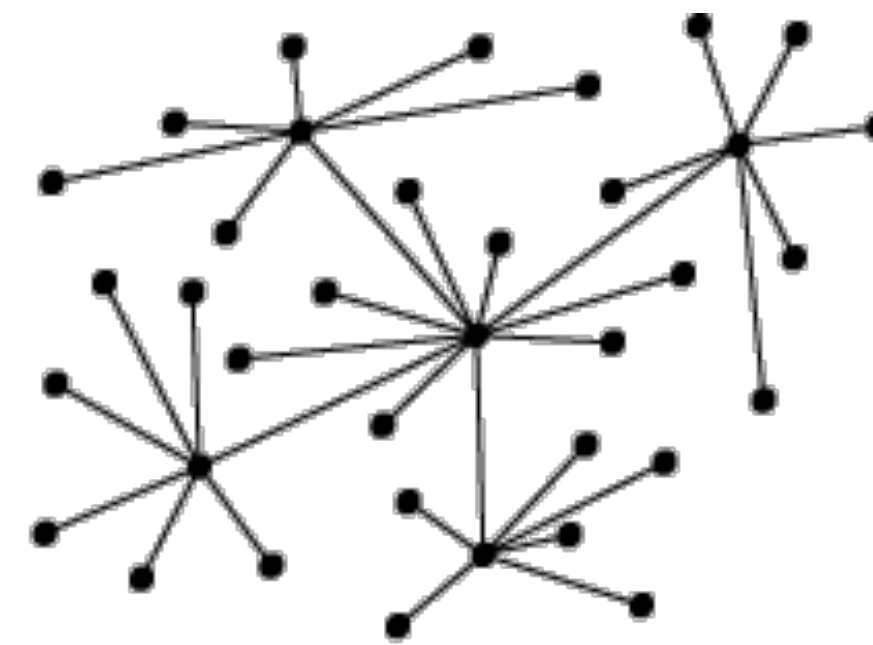
The future of data within LIC

# Key data process changes

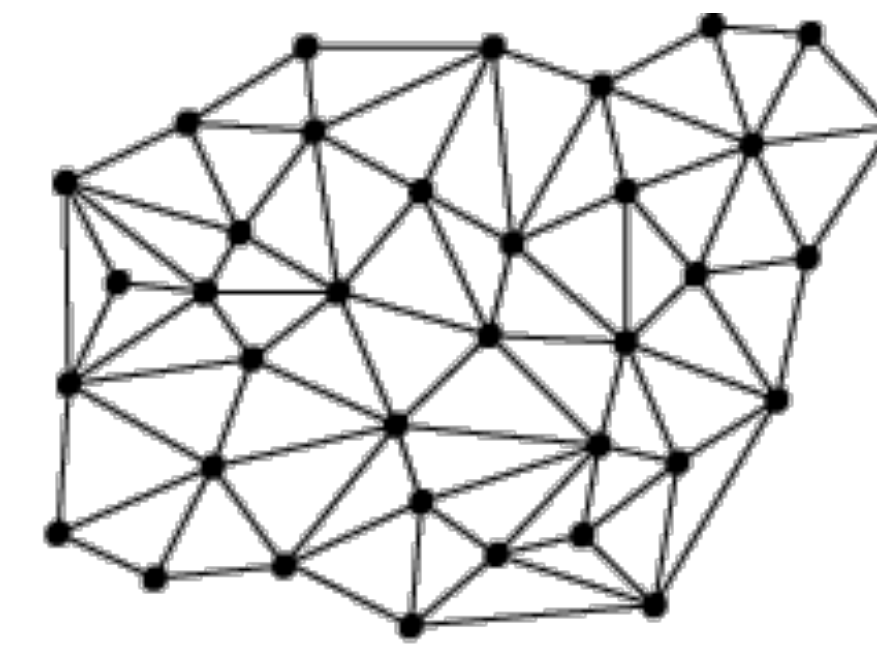
Data Lakes to Data Meshes



centralised



decentralised



distributed

Simplicity

Consistency

Buy over build

Automate - quality + exception  
reporting

Limit interactions with legacy systems

# Key data process changes

## Automation of data:

- No humans in the process
- Alerts and monitoring of data flows
- 100% cloud based
- Cheap cold storage for raw data

## Data in right place and storage format

- AWS S3
- Parquet (column based formats)
- Amazon SageMaker Data Wrangler
- Amazon EMR

# Key skill base changes

## Data processing and automation

- Data engineers/wrangers
- Data architects
- Bioinformaticians

## Science

- Machine learning and artificial intelligence
- Image/Video analysis specialists
- Statisticians



Thank you for your attention