The growing challenges of big data in the agricultural and ecological sciences

chris.rawlings@rothamsted.ac.uk

Head of Computational and Systems Biology

BBSRC bioscience for the future

ROTHAMSTED RESEARCH

Food Security

‘Demand for food is projected to increase by 50% by 2030 and double by 2050’

Rothamsted Research

- Rothamsted is an independent scientific research institute
- Longest running agricultural research institute in the world (est. 1843)
- Delivering knowledge, innovation and new practices to increase crop productivity and quality
- Develop environmentally sustainable solutions for agriculture

Main funder

BBSRC
bioscience for the future

Overview

- From Genotype to Phenotype data
- Environmental data
- Modelling and simulation
- How does this all change how biology is done and what biologists need to do?

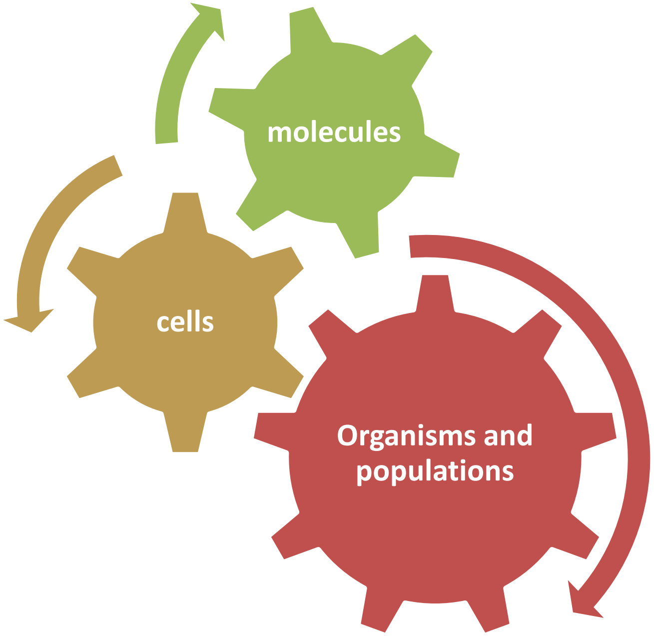
Data Rich Interactions in Agri-Ecology

- Next generation sequencing
  - Genomes of host organisms
    - Large and complex
  - Genomes of pest and pathogen organisms
- Managing / integrating multi-omics datasets
  - Variety of data resources
  - Transcriptomics and metabolomics most important
  - Importance of model organisms
  - Range of data types

ELIXIR

- Measuring phenotypes and traits –
  - Until now - low throughput
  - Now moving to high throughput
  - Range of automation and imaging technologies
- Measuring the environmental factors
  - affecting agricultural production
  - Increasing sustainability of agriculture

AnaEE



ISBE

Modelling interactions and dynamics of biological systems at all scales

Genotype + Environment = Phenotype

- Each individual has same genes but is defined by variations
  - Each individual organism has a **genotype**
  - Described by the set of genes and associated polymorphisms (SNPs, rearrangements)
- Each individual organism has:
  - **Phenotype(s)**
  - Observable characteristics – shape, colour
  - Behaviours
- Phenotype is influenced by **environment**
  - Particularly in the case of plants - sessile

Measuring phenotypes and traits

- Measuring phenotypes and traits –
  - Until now - low throughput
  - Now moving to high throughput
  - Range of automation and imaging technologies
- Measuring the environmental factors
  - affecting agricultural production
  - Increasing sustainability of agriculture

Crops - Phenotype

- Agronomic
  - Crop yield
  - Total biomass
- Plant Morphology
  - Architectural - macro-measurements
    - E.g. stem number, height, thickness, root structure
    - Light Interception – canopy
  - Non-destructive and destructive methods
- Biophysical and Biochemical
  - Photosynthesis
  - Tissue composition e.g.
    - Lipid content
    - Sugars – bioethanol
- **Environmental factors**
  - Weather
  - Soil moisture
  - Treatments (nutrients, pest management etc).

LiCOR Fluorometer

Delta T Sunscan

Sentek Diviner

Relating genotype to phenotype

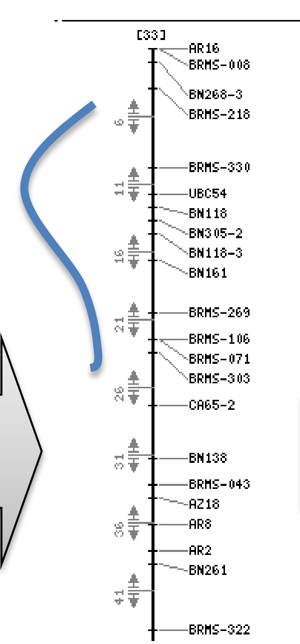
ROTHAMSTED RESEARCH

Genetics

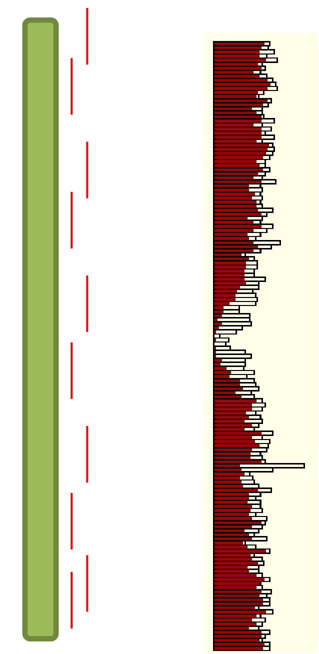
- Studies of disease families
- Population studies (association genetics)
- Crop breeding populations
- Originally genetic markers
- Now genotyping by sequencing

Phenotypic measurements

QTL **Linkage maps**



Genomic sequence
S.viminalis chromosome



BBSRC bioscience for the future

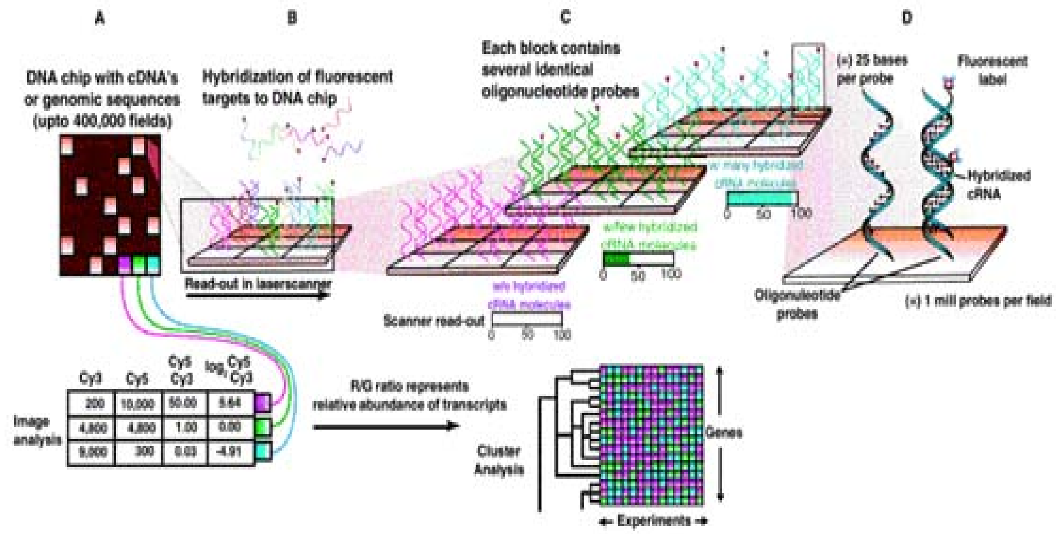
Relating genotype to phenotype

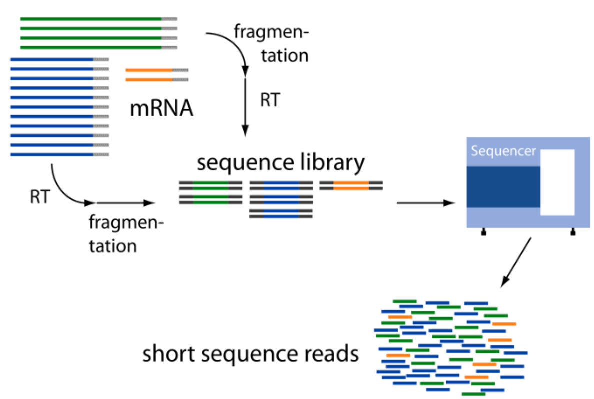
ROTHAMSTED RESEARCH

Transcriptomics

- Every gene transcript represented on a microarray
- Measure level of expression across genome in different cell types
- Now done by RNA-Seq – another use of Next Generation Sequencing

• Variations of these and combinations of these studies



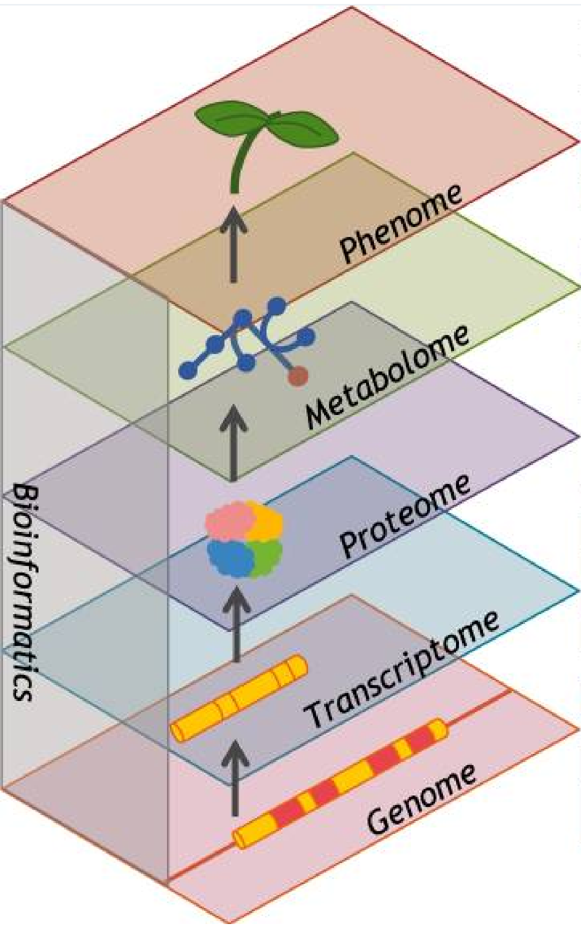


GeneChips

BBSRC bioscience for the future

Range of Plant and Crop ‘Omics Data

ROTHAMSTED RESEARCH



Omics resources	Arabidopsis	Rice	Soybean
Integrated database	TAIR [1]	Gramene [2]	SoyBase [3]
Mutant lines	FOX line [4], Ac/Ds tag line [5], T-DNA tag line [6], TILLING[7]	Ac/Ds tag line [8], T-DNA tag line [9], Tos17 mutant panel [10]	TILLING [11], Ac/Ds tag line [12]
Natural variation	NASC [13], ABRC [14]	Oryzabase [15], IRRI [16]	Legume base [17]
Metabolic map	AraCyc [18]	RiceCyc [19]	(PlantCyc [20], MedicCyc [21])
Metabolite profile	PRIMe [22], GMD@CSB.DB [23]	NA	NA
Proteome / modificome profile	PPDB [24], PhosPhAt [25]	Rice proteome database [26]	Soybean proteome database [27], Proteomics of oilseeds [28]
Interaction	AtPIN [29], AtPID [30]	NA	NA
Subcellular localization	SUBA [31], NASC proteome database [32]	Rice proteome database [26]	Soybean proteome database [27]
Full-length cDNA clones, ESTs	RAFL clones [33], RARGE [34]	KOME [35]	Soybean full-length cDNA database[36]
Microarray, GeneChip	AtGenExpress [37], Genevestigator [38]	RICEATLAS [39], Genevestigator [38]	Genevestigator [38], SGMD [40], SoyXpress [41]
Noncoding RNA	Arabidopsis MPSS [42]	Rice MPSS [43]	NA
Chip-cip / Chip seq data	SIGnAL [44]	NA	NA
Genome sequence, gene annotation	TAIR [1]	RAP-DB [45], TIGR/MSU rice [46]	SoyBase [3], Phytozome [47]
Molecular markers, variation data	TAIR[1], Nordborg lab. [48]	Gramene[2], OryzaSNP [49]	Soymap [50]
Re-sequencing	Arabidopsis 1001 [51]	NA	NA
Focused gene family database (eg. Transcription factor)	RARTF [52], AGRIS [53], DATF [54]	DRTF [55], GRASSIUS [56]	SoybeanTFDB [57], LegumeTFDB [58]

Source: Mochida et.al. “Genomics and bioinformatics resources for crop improvement.” (2010)

BBSRC bioscience for the future

New - Phenomics – HT Phenotyping

- Automated measurements
- Exploit high resolution, multi-spectral cameras
- Image processing, computer vision techniques
- Non-invasive
- Measurements e.g.
  - Plant growth and development
  - Photosynthetic activity
  - Stress measurements

National Plant Phenomics Centre

ROTHAMSTED RESEARCH

A new facility, unique in the UK, will employ a combination of robotics, automated image analysis and genomics to understand how plant growth is genetically controlled

http://www.aber.ac.uk/en/media/Example-of-Research---National-Plant-Phenomics-Centre.pdf

PRIFYSGOL ABERYSTWYTH UNIVERSITY

BBSRC bioscience for the future

Australian Plant Phenomics Facility
Field Based Technologies

ROTHAMSTED RESEARCH

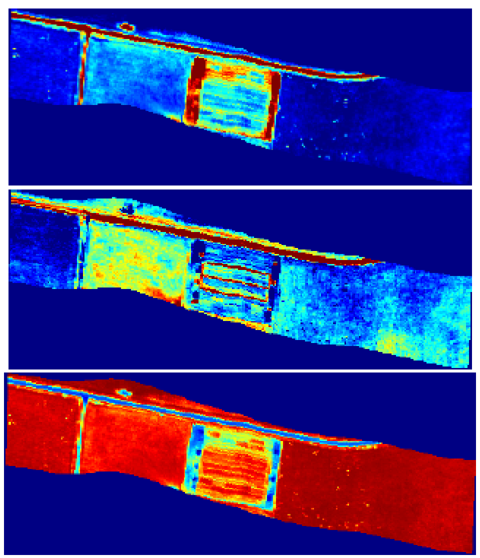
Phenomobile

http://www.plantphenomics.org.au/

http://www.plantphenomics.org/hrppc/capabilities/fieldmodule

BBSRC bioscience for the future

Remote Sensing – Hyperspectral Imaging

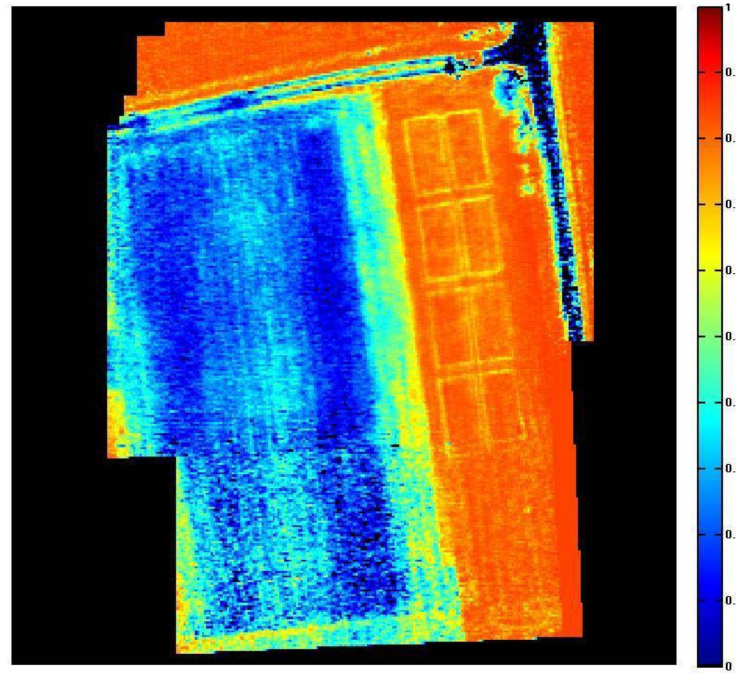


PC1

PC2

NDVI

Nitrogen, herbicides and disease

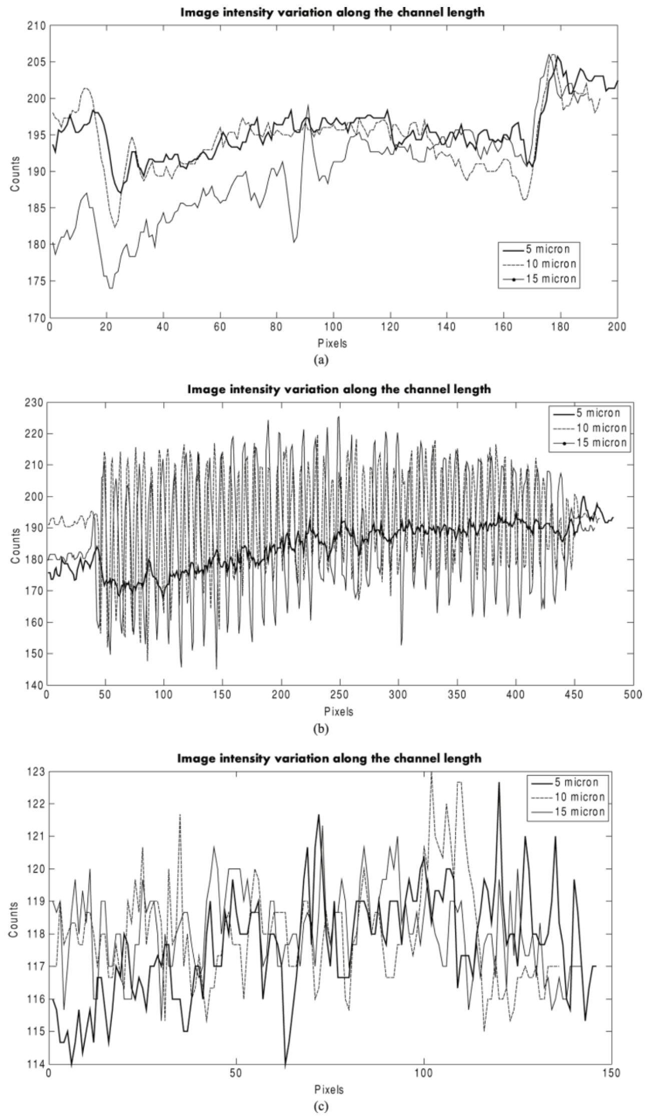


NDVI mapping from UAV (Normalized Difference Vegetation Index)

http://cadair.aber.ac.uk/dspace/handle/2160/2940

PRIFYSGOL ABERYSTWYTH UNIVERSITY

New Trend - Quantitative data from images



Merged Three-Channel Image

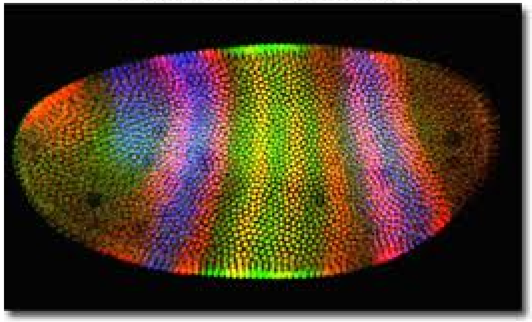
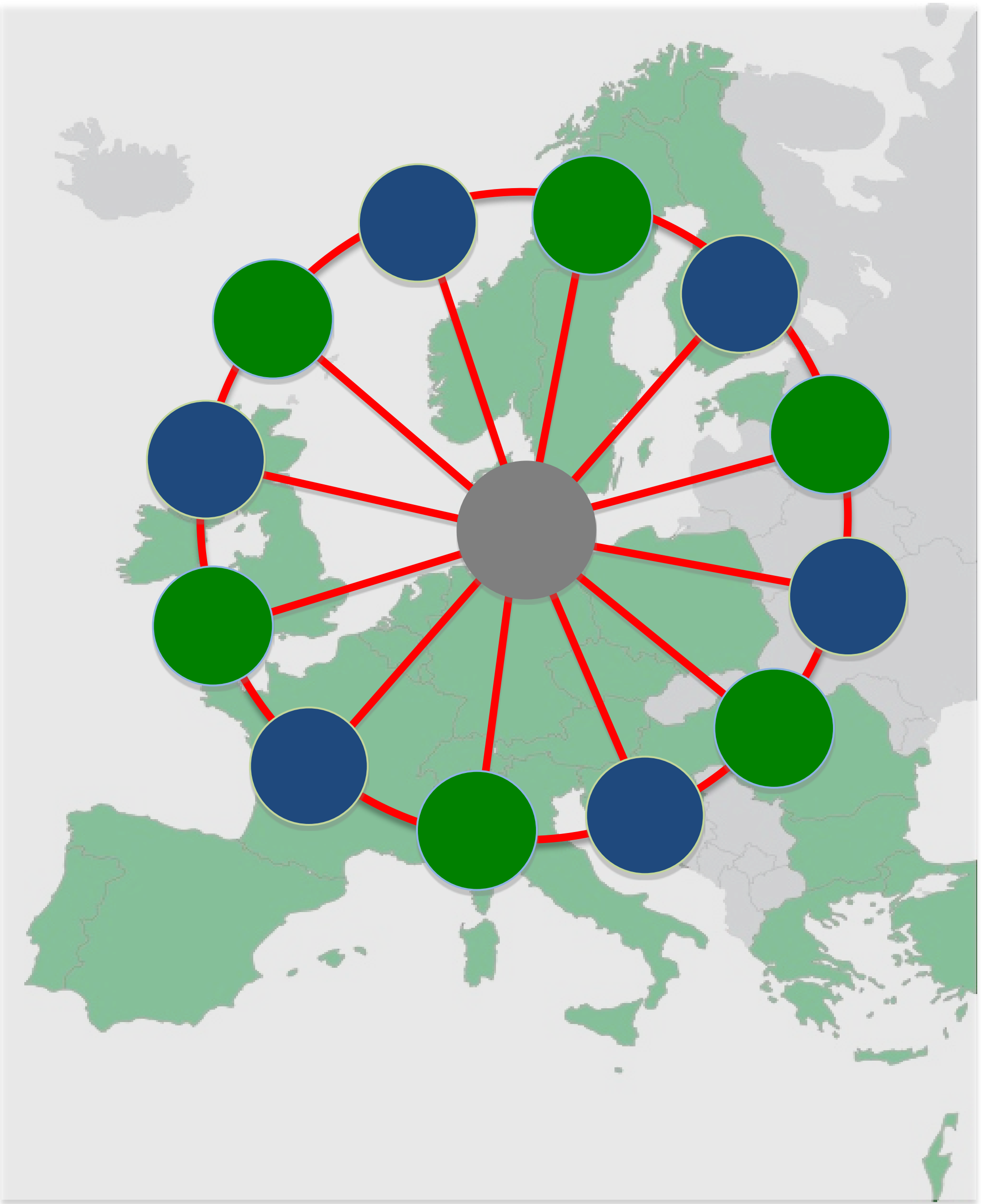


Figure 1

Distributed research infrastructure model for Euro-BioImaging



Euro-BioImaging HUB

coordination & support
access, data, training,
European infrastructure management

FLAGSHIP TECHNOLOGY NODES

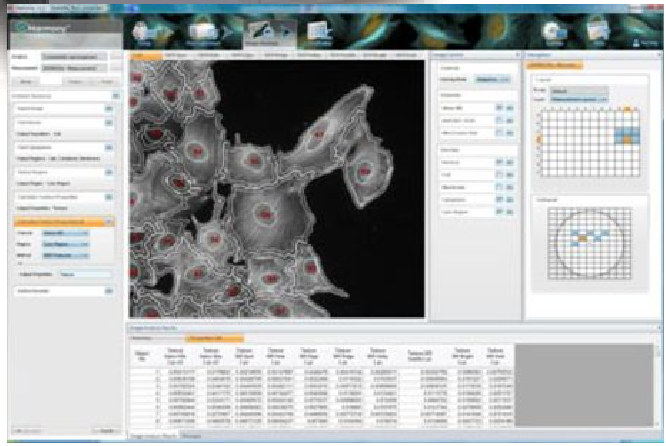
offer an innovative technology
at European leading level

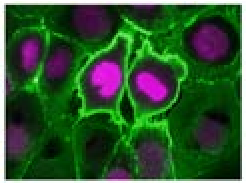
MULTIMODAL TECHNOLOGY NODES

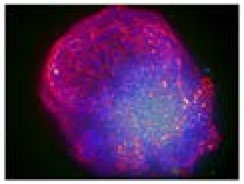
provide excellence by integration of
multiple imaging technologies at one site

High Content Screening

- Multi well robotics
- Live cell imaging
  - Advanced image management and analysis software
- Dynamic and spatially resolved quantitative data
- Multispectral Confocal microscopy
- e.g. Perkin Elmer Opera System
  - Developed for pharmaceutical industry to screen cell cultures for new drugs
  - Applications in plant science emerging
- 100,000 image sets per day







Summing up so far

- Crops for the future will be based on advanced research and breeding using molecular methods
- The genome sequencing and other ‘omics technologies are creating a data deluge
- Distributed in centres around globe – particularly so for of agricultural species
- Next data tsunami coming from image based technologies used in high throughput phenotyping
- Being used at all levels of biological and geographical scale
- Obvious challenges inherent in data integration, analysis and visualisation

Agricultural Interactions with the Environment

Delivering sustainable intensification

Sustainable Intensification

- The demands on land for food production are increasing
- New management methods needed to increase intensity of agriculture
- Challenge is to do this without significant harm to environment
  - Sustainable intensification
  - Intelligent management of agricultural ecosystems
- Be ready for climate change

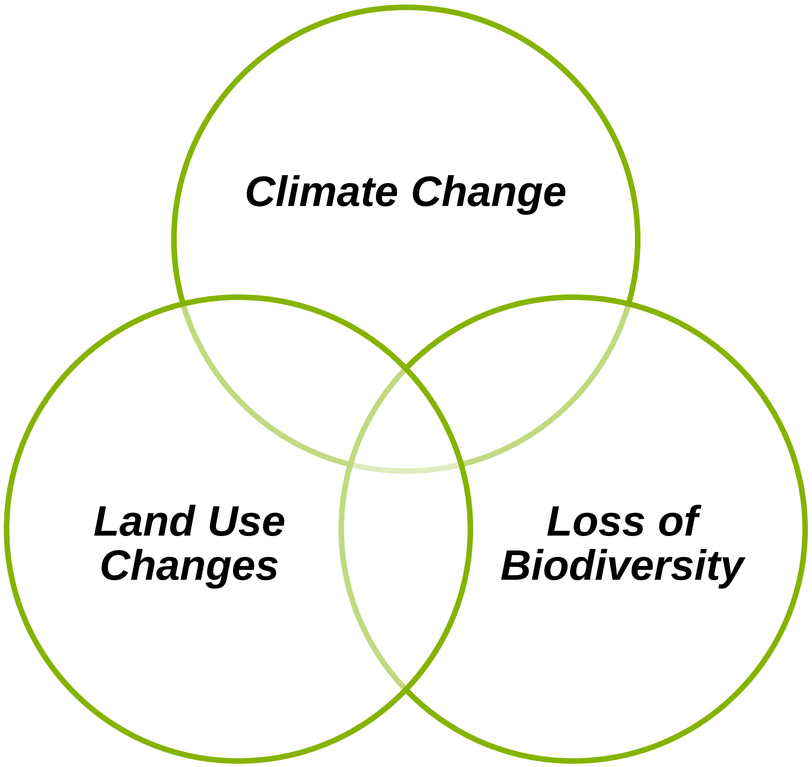
AnaEE

A European infrastructure for analysis and experimentation on ecosystems

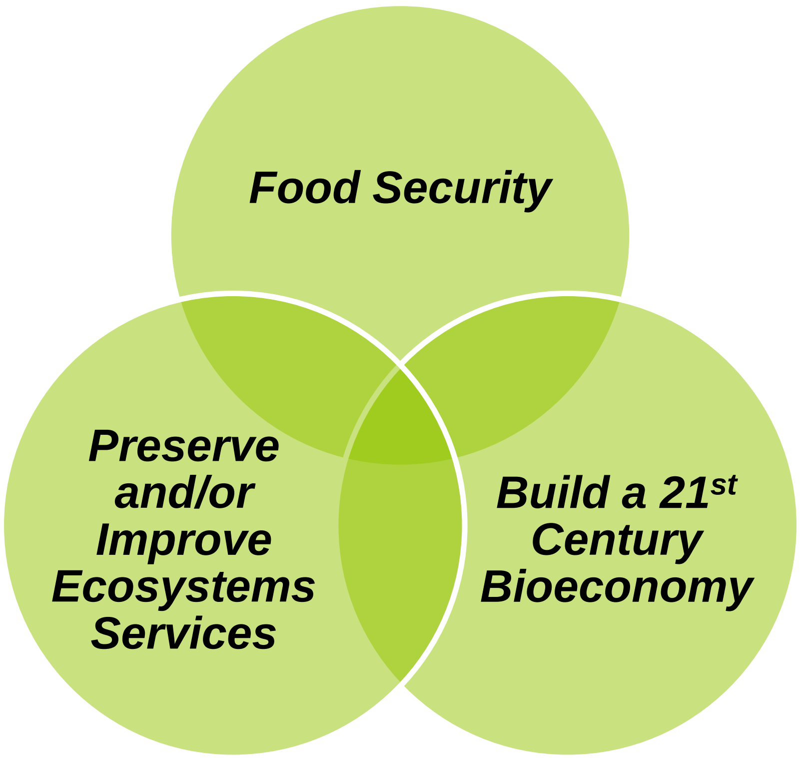
http://www.anaee.com/

Challenges facing Europe and the world

Global Changes



Need to Adapt



Our strategic objectives

Foster **capacity building** in ecosystem science by providing **state of the art facilities** and **structuring the research community**

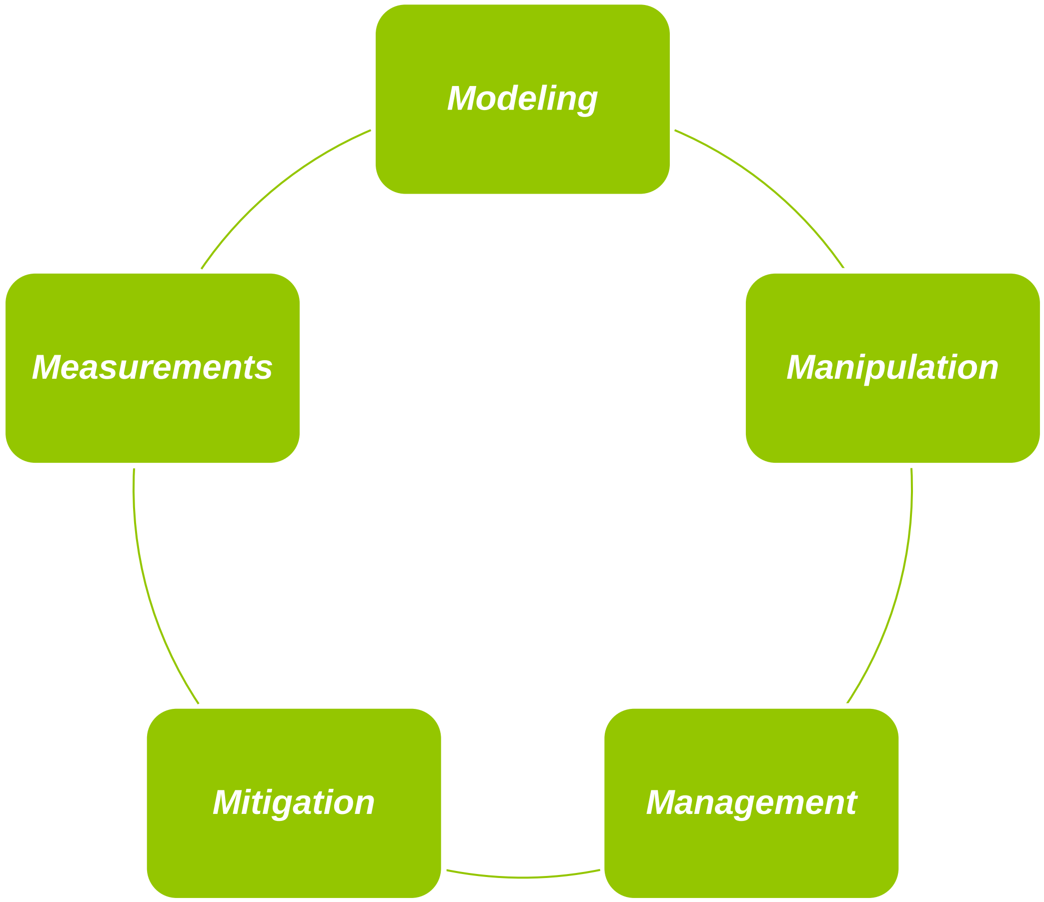
High-quality scientific data and services to assess impacts and risks associated with environmental changes

Help develop policies and engineer techniques that will allow buffering of and/or adaptation to these changes

Contribute to developing a **21st century bioeconomy**

Analysis and Experimentation on Ecosystems

Our approach



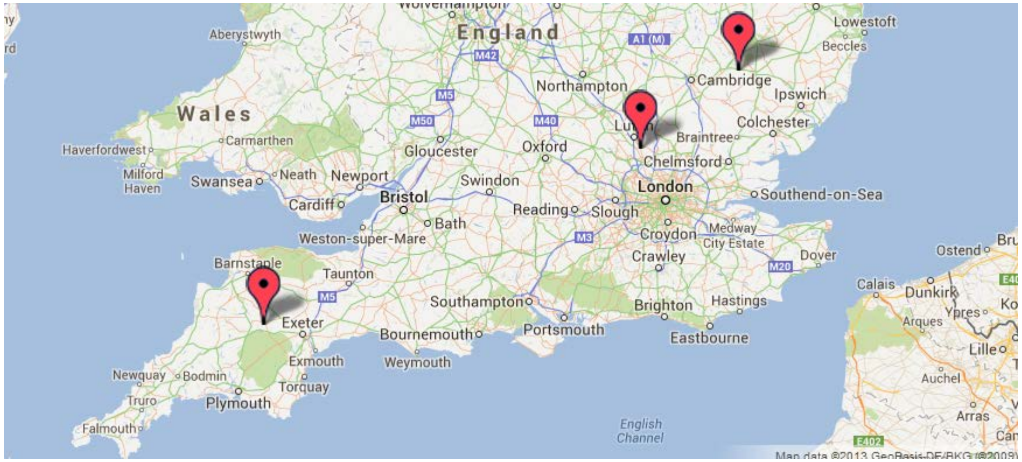
Our Model

- A world-class **distributed experimental infrastructure** for **enabling ecosystem research**
- A **coordinated set** of **experimental platforms across Europe** to **analyse**, **test** and **forecast** the response of ecosystems to environmental and land use changes.
- ANAEE will be the key instrument for carrying out terrestrial ecosystem research within the **European Research Area**.
- Scale: **Europe** including **full range of Europe's ecosystems and climate zones**
- Wide range of **environmental and societal implications** for policymakers

Rothamsted Research
where knowledge grows

Rothamsted North Wyke Farm Platform

Example AnaEE Site

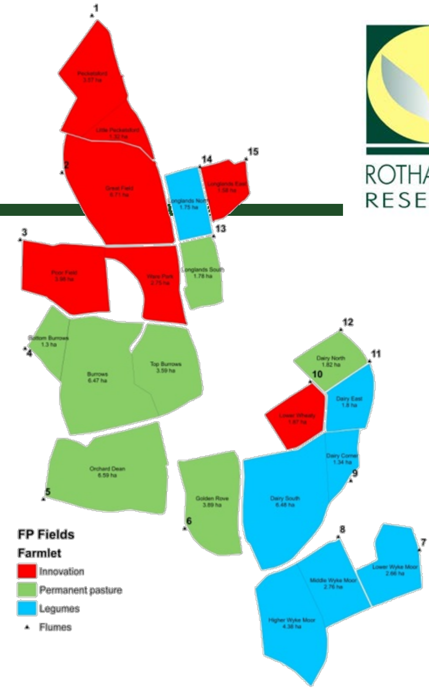


BBSRC
bioscience for the future

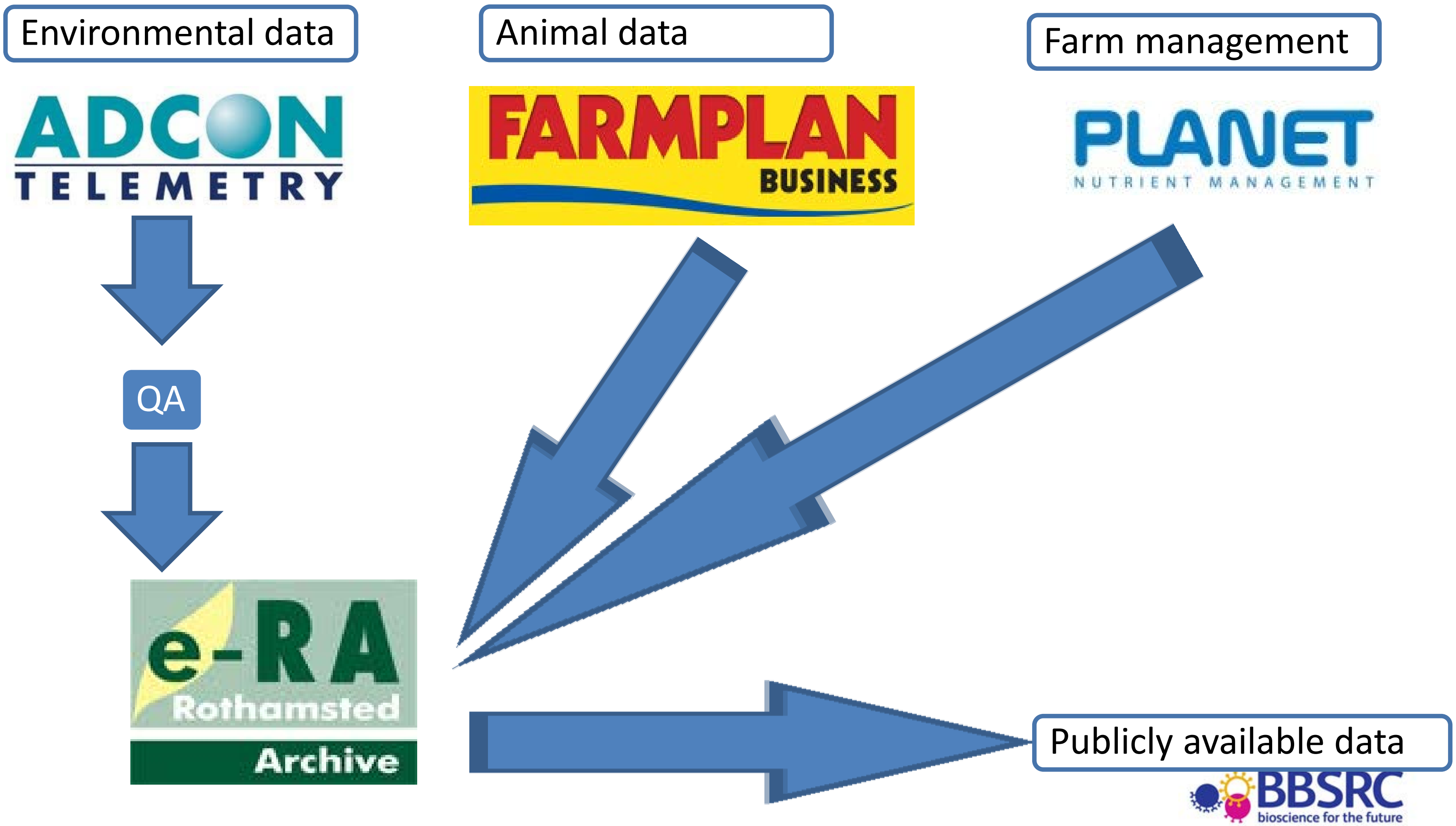
ROTHAMSTED RESEARCH

Main features

- Three “farmlets”
- Highly instrumented
  - Most instrumented farm in Europe(?)
  - Realtime data capture from 15 monitoring stations
- Known topology and hydrology
- Long term experiments – just starting
  - Baseline data 2 years
- Integration with remote sensing data
  - Satellite, hyperspectral imaging

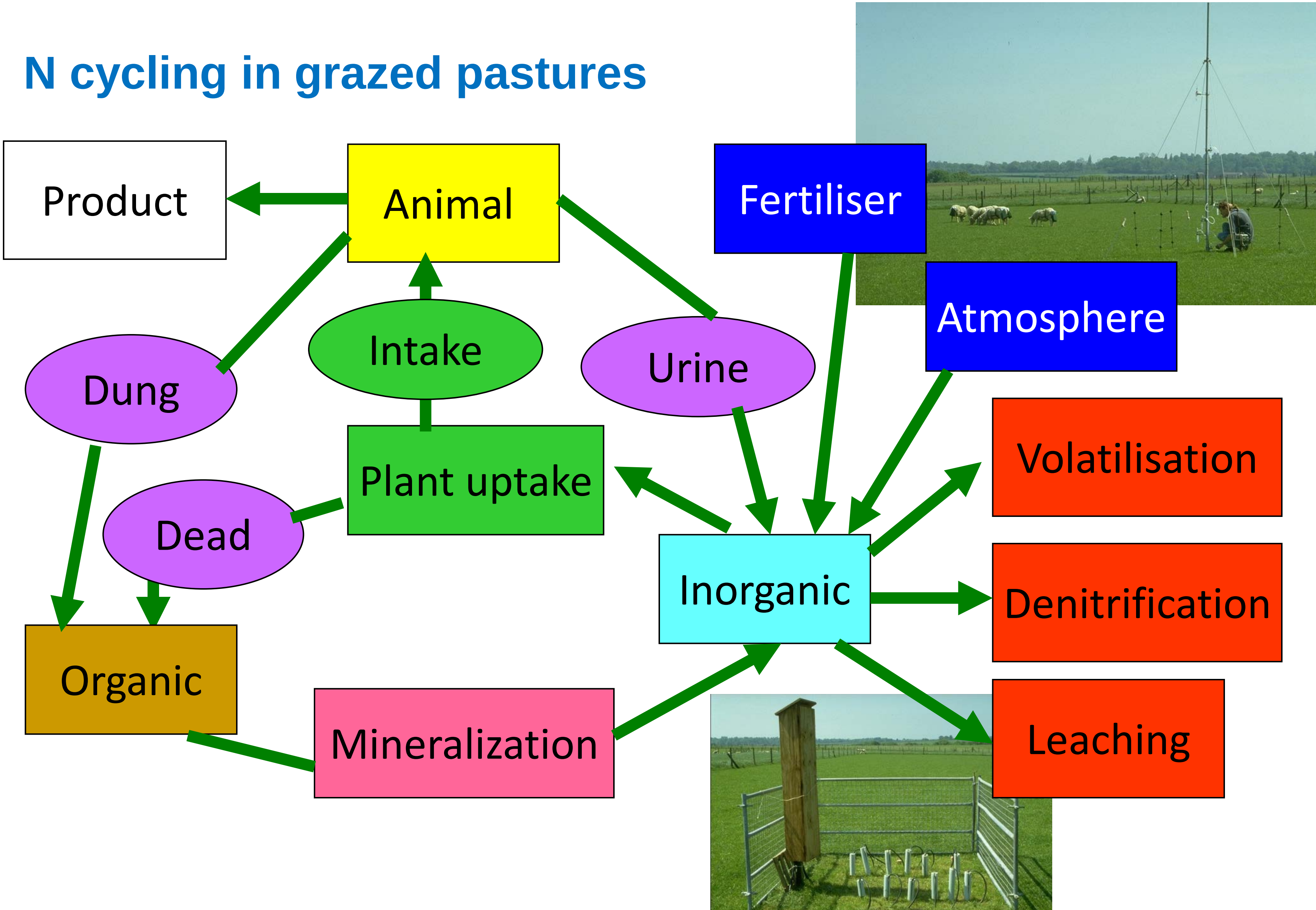


Databases



Environmental data
Animal data
Farm management
ADCON TELEMETRY
FARMPLAN BUSINESS
PLANET NUTRIENT MANAGEMENT
QA
e-RA Rothamsted Archive
Publicly available data
BBSRC bioscience for the future

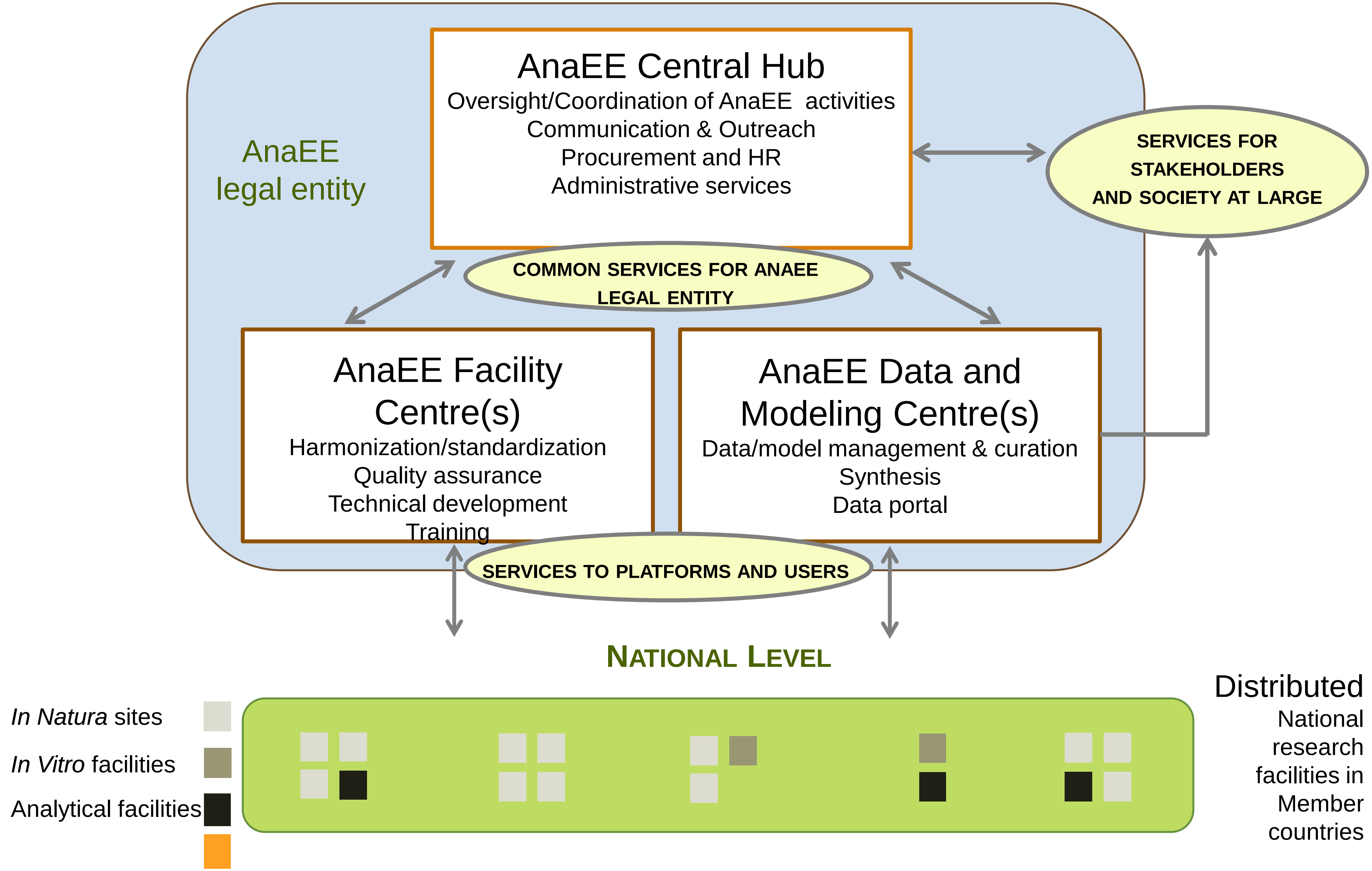
N cycling in grazed pastures



PARSONS et al. (1991) Uptake, cycling and fate of nitrogen in grass-clover swards continuously grazed by sheep. *J. Agric. Sci., Camb.* 116, 47-61.

JARVIS et al. (1991) Micrometeorological studies of ammonia emission from sheep grazed swards. *J. Agric. Sci., Camb.* 117, 101-109.

Possible AnaEE Service structure



Characteristics of A-E Data

- Very wide range of studies
  - Different objectives, different data
- Widely Different Scales
  - Spatial : lysimeter - catchment
  - Temporal : realtime – annual
- Different Types of Environment
  - Controlled (mesocosm), Managed, Natural,
- Different geographical locations

Challenges - Integration/Interoperabilty

- Many competing standards
- Metadata not been considered in many situations
  - Metadata standards poorly developed

Modelling and Simulation

Systems Biology

Two approaches

- Study all components of a biological system/organism
  - Integrative approach
  - Understand interactions
  - Link interactions to phenotypes to identify emergent properties of the network
- Systems Engineering appriach
  - Build computer and mathematical models
  - Generally associated with biochemical pathway modelling
  - Create simulations
  - Well developed examples in single cell organisms
    - Yeast, Bacteria
- Biology should become more like physics and amenable to engineering approaches
  - Physicists conduct experiments to test models
  - **Predict phenotype from genotype**

ISBE Infrastructure for Systems Biology Europe

http://isbe.eu/

- An infrastructure for the integration and synthesis of systems biology across the Data Generation, Integration and Stewardship Centres identified by the project.
- The distributed, interconnected infrastructure using
  - best practices,
  - standards,
  - technical infrastructure,
  - software
  - capacity for data
  - model management and distribution.
- To propose and promote a framework and best practices for model and data management for Systems Biology in Europe.
- To collaborate with standardisation activities for model and data management

Modelling and Simulation Equally Important in Agro-ecology

- Agro-ecology uses many mathematical models
  - Long tradition
  - dynamics of ecosystems – function and structure
  - Crop productivity and crop management decision support
  - predicting impacts of climate change
- Models can be linked to
  - other models
  - datasets used for calibration and validation
- Predictive models
  - Contribution to policy development in agriculture
- Contribution to climate modelling
  - Soil carbon models (Rothamsted) now part of climate change prediction models (Hadley Centre)

How has big data changed things?

- **Process of life science research is changing**
  - Data mining the public data resources will save time and money
  - *In silico* research before *in vitro/in vivo*
  - Many biologists spend less time in the lab, more on the computer
  - Outsourced data generation – research service companies not just for commercial research e.g. next gen sequencing
    - Almost no research time in the lab
  - Competitive edge can come from the software and *in silico* methods providing more focussed research
  - Some biologists only work on the computer
    - Bioinformatics, Computational Biologists

How is big data changing things?

- **Creates new infrastructure requirements for life sciences**
  - Data centres with relevant expertise
    - European Bioinformatics Institute (EU)
    - National Center for Biotechnology Information (USA)
  - International collaborations
    - ELIXIR

http://www.elixir-europe.org/

- **Internet connectivity**
  - Terabyte data movements (next gen sequencing)
  - Remote working with data centres
  - Data integration – semantic web, federation
  - Collaboration environments

- **High performance computing (PRACE)**
  - Modelling simulation (ISBE)
  - Data analysis and visualisation
  - Different requirements from physical sciences..

How has big data changed things?

- **Need for cross-disciplinary working**
  - Computer Science and Electronic engineering
  - Mathematics and Computer science
- **Development of new informatics-led sub-disciplines and careers**
  - Bioinformatics
  - Computational Biologists
  - Cheminformatics
  - Data managers and data scientists
- **Life science informatics as research and service**
  - In commercial and academic sectors

Conclusions

- Biology is a big-data discipline, Ecology is becoming one
- Applications are everywhere and the demands are growing
- Major drivers of growth include:
  - Next generation sequencing
  - Imaging of all levels of biological scale
  - Remote (and not so remote) sensing – UAVs…
- Driving demand in compute, storage, networking
- Major challenges in
  - Development of standards
  - Data integration
  - Visualisation
- Shortages of computational capacity and skills across biology

THE END

ROTHAMSTED RESEARCH

BBSRC
bioscience for the future