Research Methods for Data Science
DIT875, 2018-2019

Main lecturer: John Hughes

Course responsible: Graham Kemp (kemp@chalmers.se)

Learning outcomes

Knowledge and understanding

- extract and summarize the current knowledge about a specific topic in data science from original articles
- clearly describe the scientific or technical problems treated within a specific topic in data science
- identify the essential points of an article

Learning outcomes

Competence and skills

- retreive information that is required to understand a topic not treated in the primary sources
- write well organized and well formulated text with proper scientific argumentation
- explain and communicate a topic to readers that are not necessarily experts in the domain
- plan a research project, such as a master's thesis, based on problem analysis and with a clearly shaped goal, and predict its feasibility

Learning outcomes

Judgement and approach

- review scientific sources critically
- analyze and evaluate the reasons for the choice of a solution method
- identify possible ethical and societal consequences of a method, design or system
- evaluate possible decisions, based on general ethical values
- apply ethical principles in scientific writing, including proper citation and use of statistical statements

Course content

The following topics are covered in the course:

- technical writing in data science, being practiced on a topic of free choice and on a research proposal
- structuring a scientific text
- communicating a topic to different audiences
- theories on ethics, with examples from data science
- identification and analysis of ethical and societal issues in data science
- ethics and good practice in research and publishing

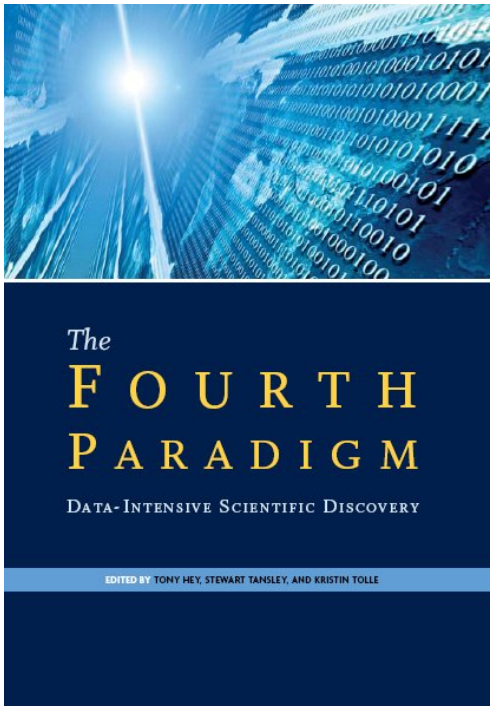
Assessment

- The course is examined by a written thesis proposal, normally carried out in pairs, and individual written assignments.

Some previous Master's projects

- Constructing a Context-aware Recommender System with Web Sessions (3Bits Consulting AB)
- Machine Learning for On-line Advertising Using Contextual Information (Admeta)
- The Identification of Target Proteins from Patents - Mining of biological entities from a full-text patent database (AstraZeneca)
- Browser Fingerprinting (Burt)
- Learning to rank, a supervised approach for ranking of documents (Findwise)
- Entity Disambiguation in Anonymized Graphs Using Graph Kernels (Recorded Future)
- Using Classification Algorithms for Smart Suggestions in Accounting Systems (SpeedLedger)
- Cluster User Music Sessions (Spotify)
- Extracting Data from NoSQL Databases - A Step towards Interactive Visual Analysis of NoSQL Data (TIBCO Software)
- Pattern Recognition in a Distributed Message Passing (Volvo Technology)

The Fourth Paradigm



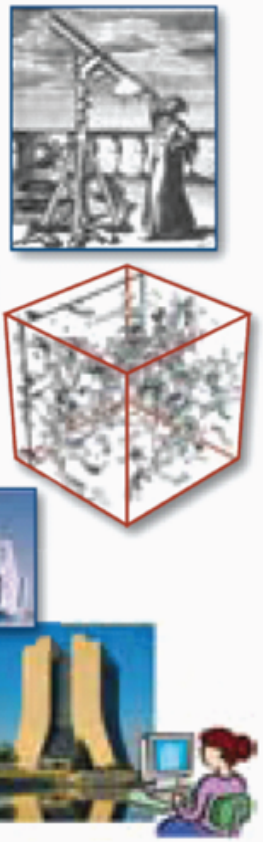
Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.

http://research.microsoft.com/en-us/collaboration/fourthparadigm/default.aspx

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments
    or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files
    using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - \mathrm{K}\frac{c^2}{a^2}$$



From “Jim Gray on eScience: a transformed scientific method”, introduction to “The Fourth Paradigm: Data-Intensive Scientific Discovery” edited by Tony Hey, Stewart Tansley, and Kristin Tolle

European Bioinformatics Institute

- www.ebi.ac.uk
- “We develop databases, tools and software that make it possible to align, verify and visualise the diverse data produced in publicly funded research, and make that information freely available to all. ”
- EMBL-EBI data centres can store over 120 Petabytes (120,000 Terabytes) of data.
- Every weekday, well over 27 million requests are made to EMBL-EBI websites.

Source: https://www.ebi.ac.uk/about/our-impact

Research Data Lifecycle

https://www.ukdataservice.ac.uk/manage-data/lifecycle

Re-using data

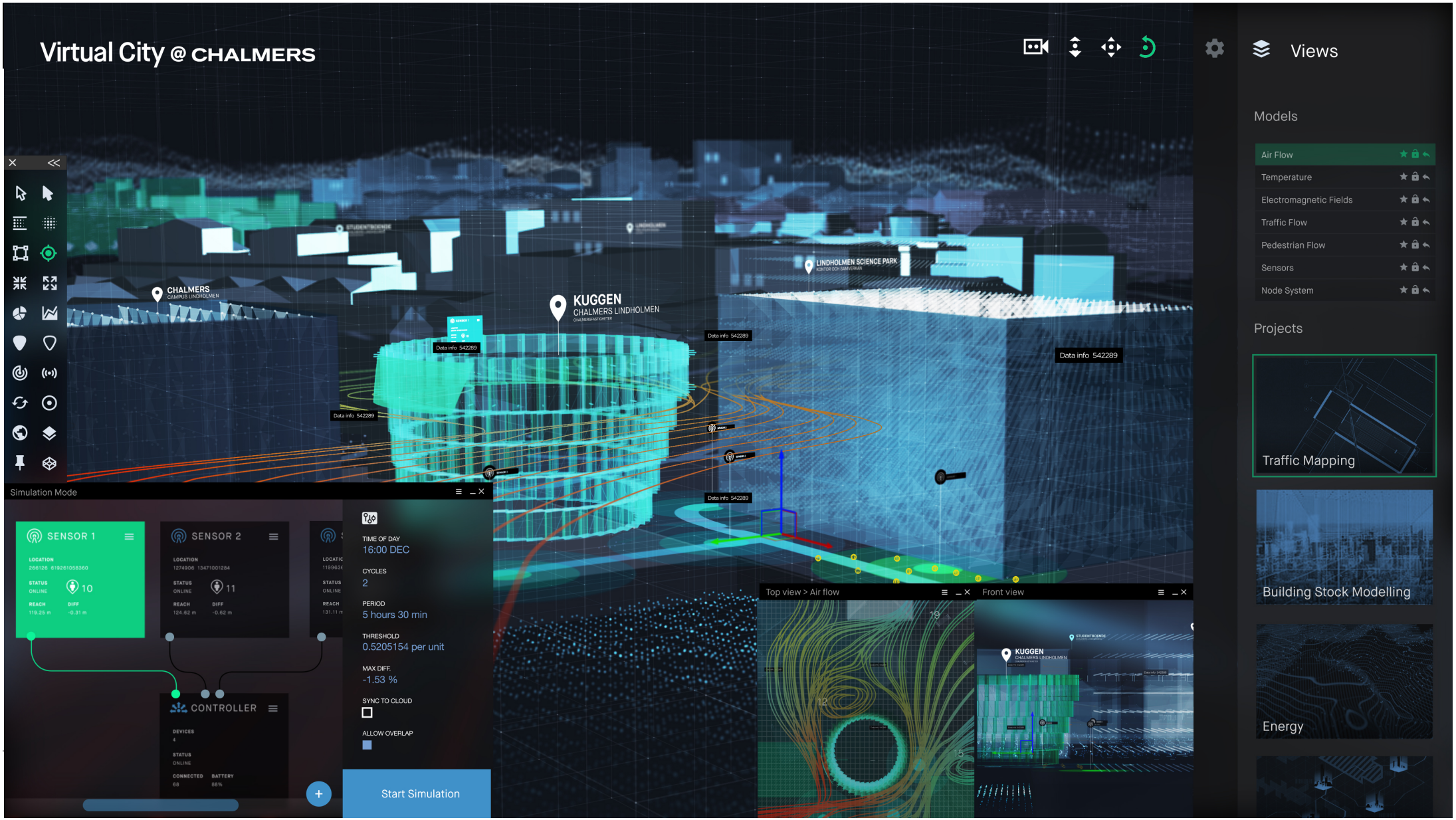
- Added value if we can combine two or more data sets

- The following slides are from a presentation that was given at a workshop on Virtual Cities.
- The aim of that presentation was to highlight some of the challenges that might be faced when combining data sets.
- Think ahead, and plan for future re-use of data.

FINANCIAL TIMES SURVEY, 2001-11-27

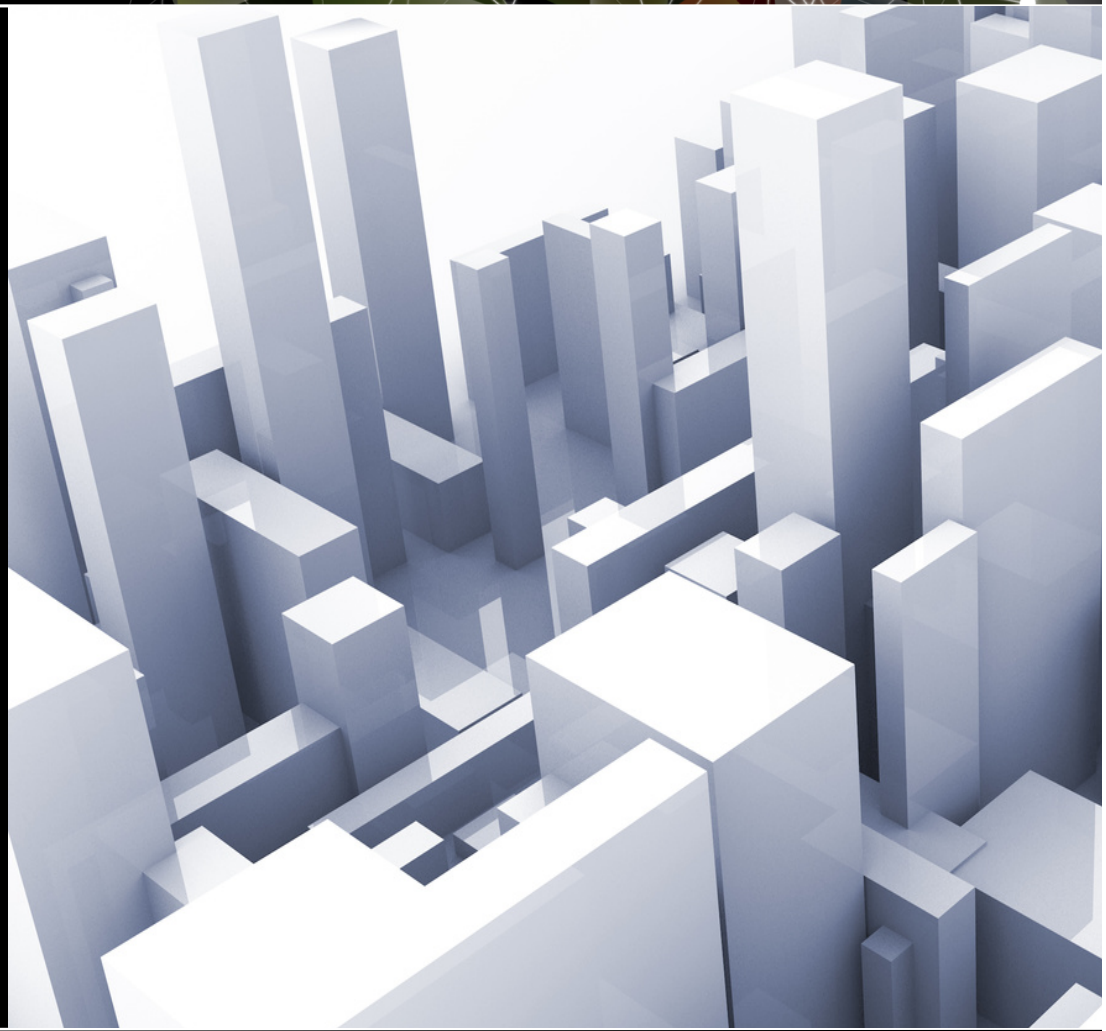
"Data integration and management is an area with less glamour than high-performance computing …

… but, probably, more practical relevance for the biotech industry. Researchers need to organise and integrate information about genes and proteins from many different sources, in many formats and file types, so that they can uncover patterns and associations."


Virtual City @ CHALMERS
Views
Models
Air Flow
Temperature
Electromagnetic Fields
Traffic Flow
Pedestrian Flow
Sensors
Node System
Projects
Traffic Mapping
Building Stock Modelling
Energy
CHALMERS
CAMPUS LINDHOLMEN
KUGGEN
CHALMERS LINDHOLMEN
LINDHOLMEN SCIENCE PARK
Data info 542289
Simulation Mode
SENSOR 1
SENSOR 2
CONTROLLER
TIME OF DAY
16:00 DEC
CYCLES
2
PERIOD
5 hours 30 min
THRESHOLD
0.5205154 per unit
MAX DIFF.
-1.53 %
SYNC TO CLOUD
ALLOW OVERLAP
Start Simulation
Top view > Air flow
Front view

A VIRTUAL LABORATORY

- Simulate the Smart City
- Optimize the Smart City
- Analyze the Smart City
- Plan the Smart City
- Realize the Smart City



CHALMERS
UNIVERSITY OF TECHNOLOGY

Data Integration Challenges in Biology

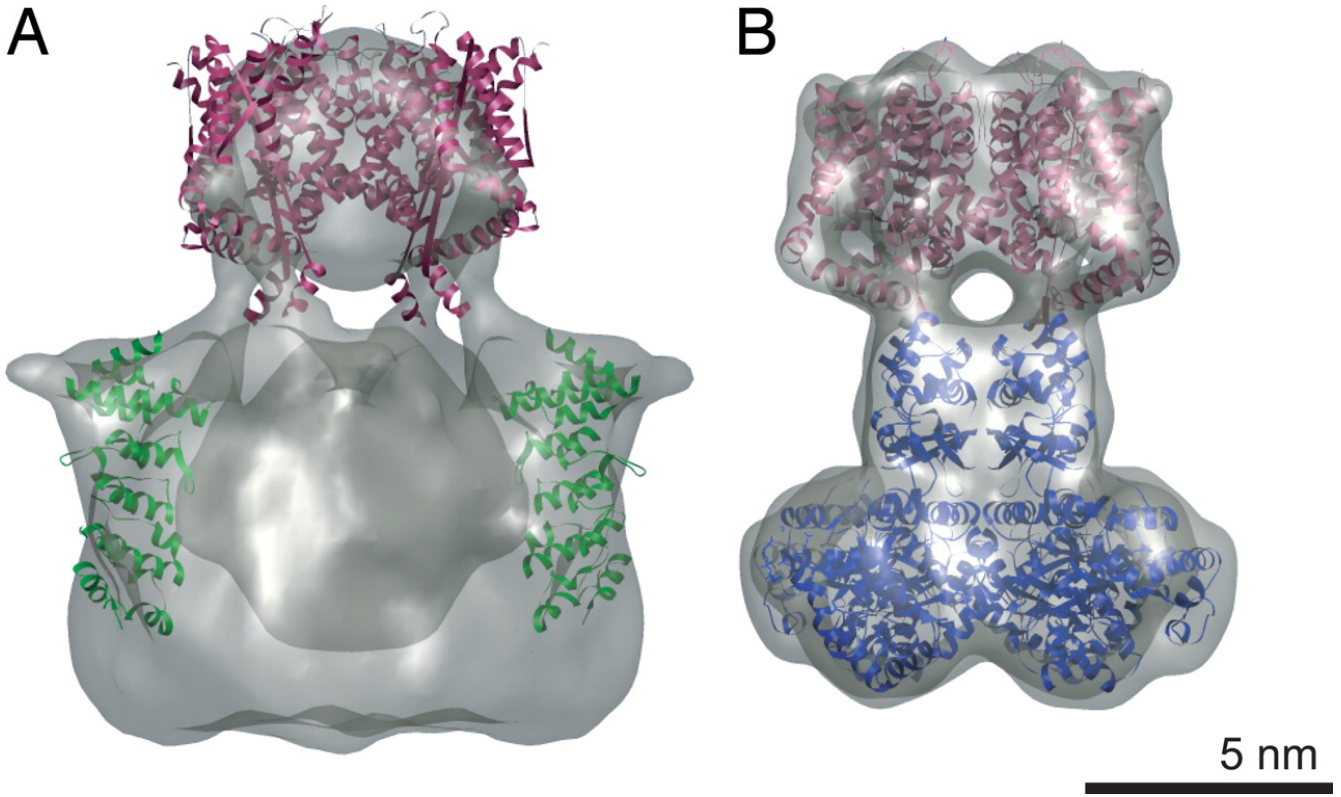
Graham Kemp
Computer Science and Engineering

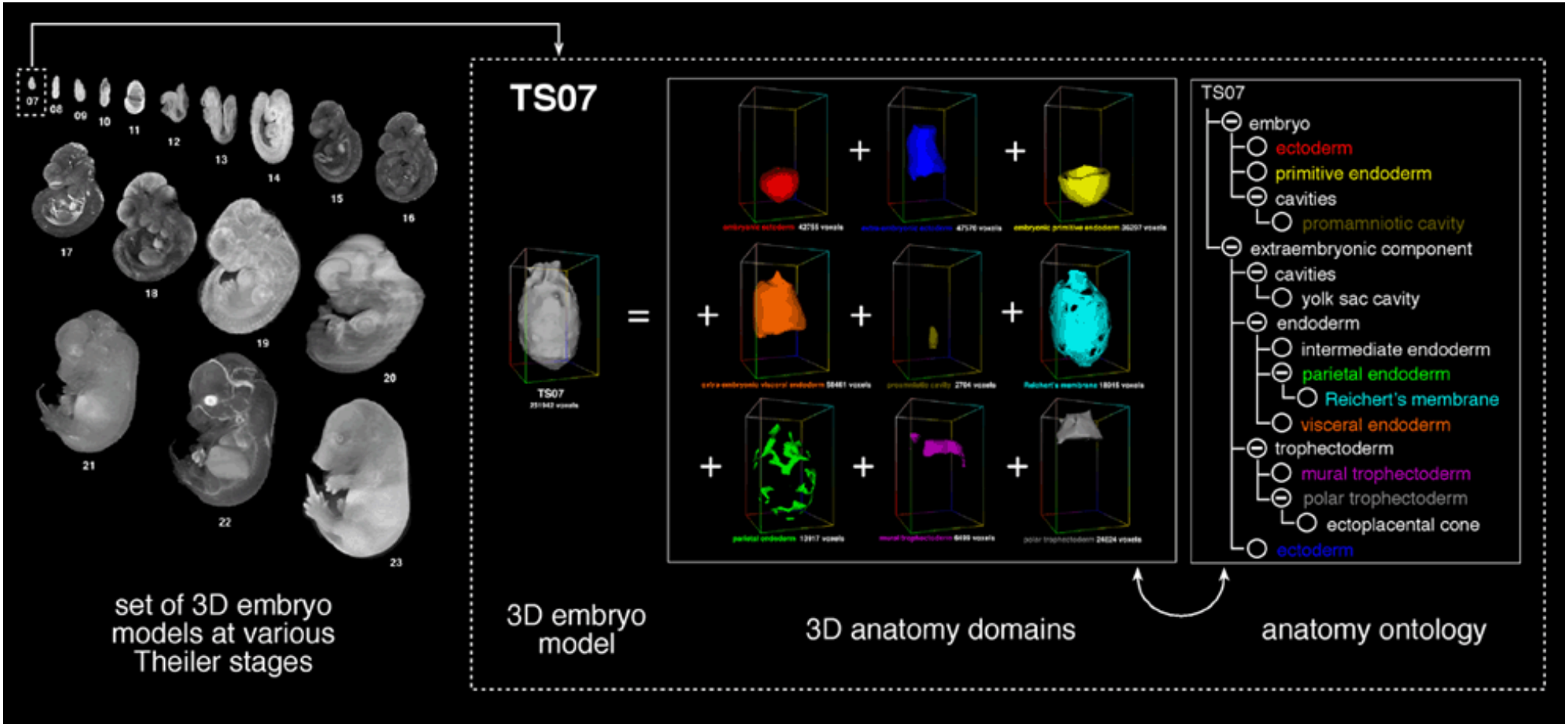
WHAT IS BIOLOGY?

- Ecosystem
- Community
- Population
- Organism
- Organ System
- Organ
- Tissue
- Cell
- Molecule

HIGH AND LOW RESOLUTION DATA

Moiseenkova-Bell, V.Y., Stanciu, L.A., Serysheva, I.I., Tobe, B.J. and Wensel, T.G., 2008. Structure of TRPV1 channel revealed by electron cryomicroscopy. *Proceedings of the National Academy of Sciences*, *105*(21), pp.7451-7455.



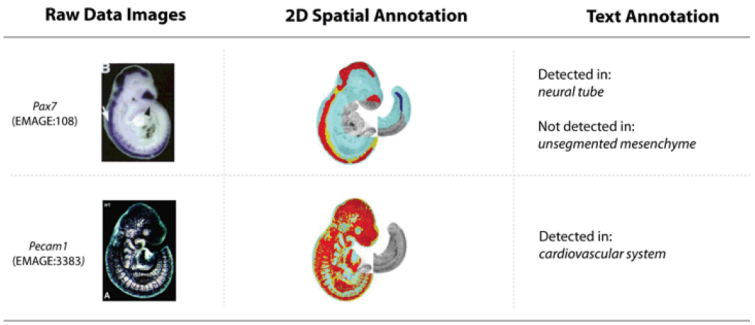


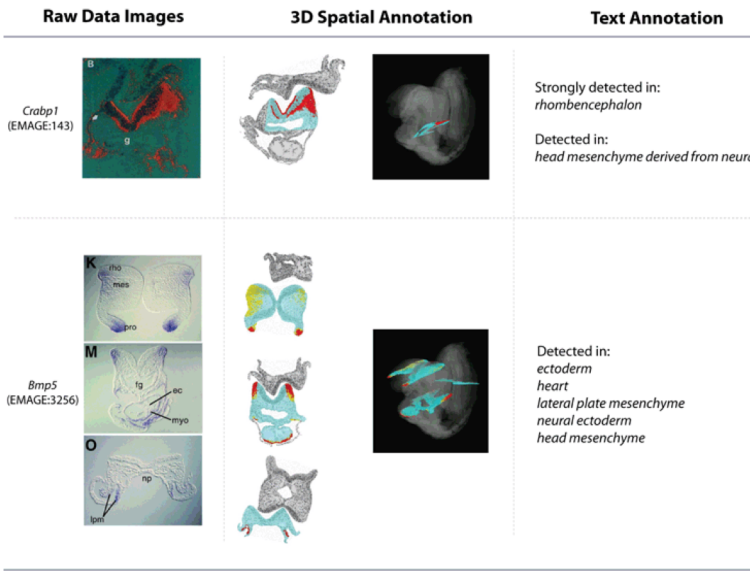
https://www.emouseatlas.org/emage/about/what_is_emage.php

EMAGE DATA ANNOTATION METHODS

- Spatial annotation of 2D and 3D models
- 2D section data mapped onto 3D virtual embryo model

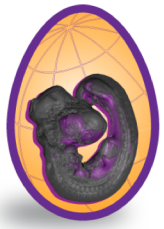
https://www.emouseatlas.org/emage/about/what_is_emage.php





COMPARING DIFFERENT SPECIES

emouseatlas



echickatlas

DATA INCOMPATIBILITY

1. ***Scale-difference***. For example, in one database four values (cold, cool, warm, hot) are used to classify climates of cities, while in another database the average temperatures in Fahrenheit may be recorded.
2. ***Level of Abstraction***. For example, in one database "labor cost" and "material cost" may be recorded separately, while in another they are combined into "total cost." Another example is recording an employee's "average salary" instead of his or her "salary history" for the previous five years.
3. ***Inconsistency Among Copies of the Same Information***. Certain information about an entity may appear in several databases, and the values may be different due to timing, errors, obsolescence, etc.

Smith, J.M., Bernstein, P.A., Dayal, U., Goodman, N., Landers, T., Lin, K.W. and Wong, E., 1981. Multibase: integrating heterogeneous distributed database systems. In *Proceedings of the National Computer Conference, 1981* (pp. 487-499). ACM.

DO WE HAVE ANY OF THESE CHALLENGES?

- Combining representations across many scales?
- Combining data with different resolutions?
- Modelling changes and growth over time?
- Analysing data to understand systems?
- Mapping data to 3D reference models?
- Finding similarities and differences in different data sets?
- Bringing in new data from outside our initial area of interest?
- Dealing with variety and complexity of data?