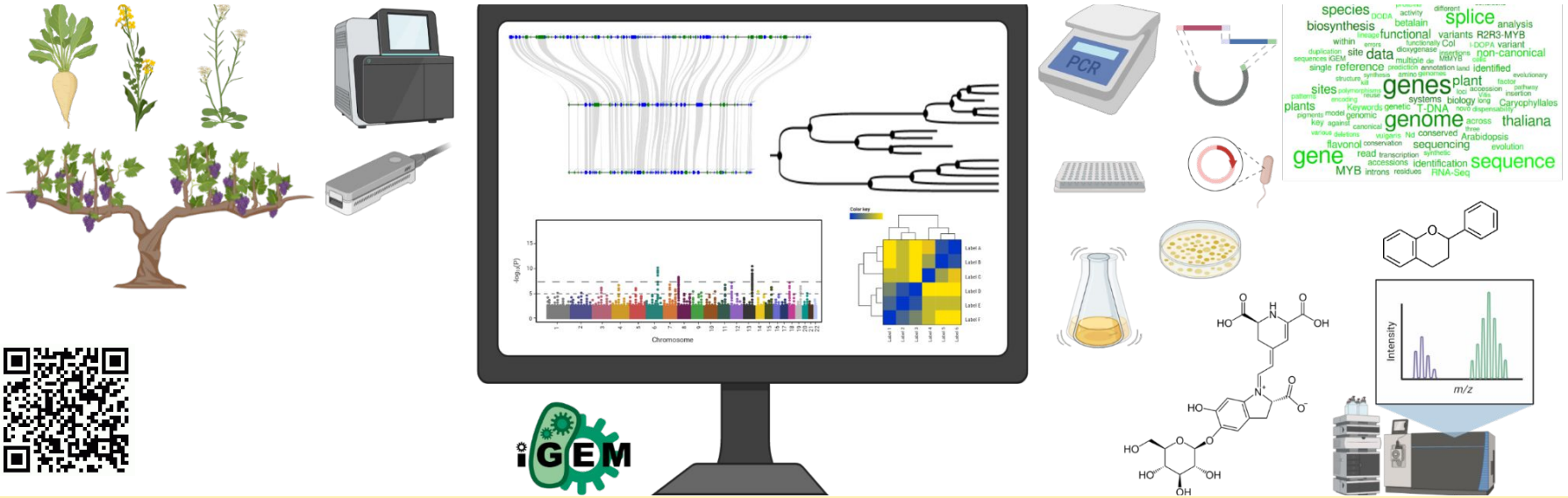
Technische Universität Braunschweig



GE32/MM12 - Introduction to Data

Prof. Dr. Boas Pucker (Plant Biotechnology and Bioinformatics)

Availability of slides

- All materials are freely available (CC BY) - after the lectures:
  - StudIP: **GE32/MM12**
  - GitHub: https://github.com/bpucker/teaching

- Questions: Feel free to ask at any time

- Feedback, comments, or questions: b.pucker[a]tu-braunschweig.de

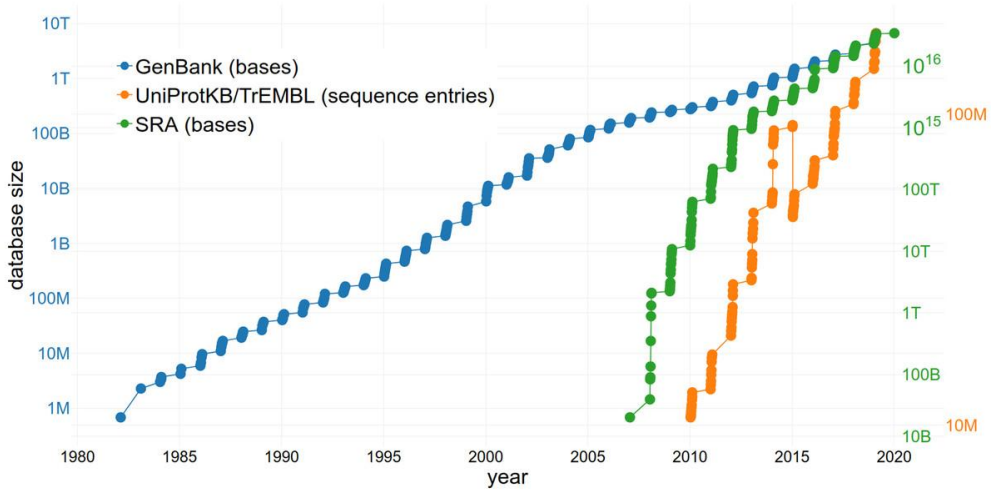
What is Data Literacy?

What is Data Literacy?

- ‘Competency to handle data properly’
- ‘Read, understand, create, communicate data as information’ (wikipedia)

Why is Data Literacy important?

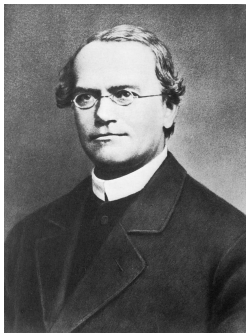
- We live in a world of data: important for science, business, and society
- Objective: data informed decision making
- Sizes and complexity of data sets are increasing (‘BigData’)
- Data is accumulating over time



Sielemann et al., 2020: 10.7717/peerj.9954

More definitions

- Data: collection of observations
  - 2020: 312 red flowers; 95 white flowers
  - 2021: 298 red flowers; 98 white flowers



- Knowledge: accumulation of facts and data
  - Ratio of red flowers to white flowers is 3:1

- Insights: grasping the underlying nature of knowledge; understanding general concepts
  - Flower color has a genetic basis
  - Red flower allele dominant over white flower allele

What collections/databases do you know?

Data management

- Larger data sets require more efficient data management
- Data management plans required in project proposals
- Many services and organizations emerging (NFDI)

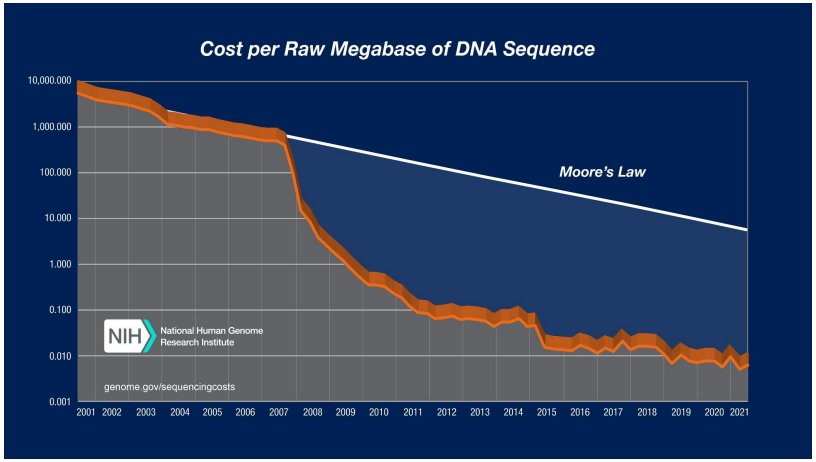
Data protection and security

- Data protection is becoming a huge issue in the EU
- Avoid any personal data in your research data sets
- Data security is gaining relevance
- German universities are frequently attacked:
  - Gießen (2019): offline for weeks + 1.7 million direct costs
  - HHU Düsseldorf (2020): ransom attack on clinic
  - Bochum, Dresden, Freiburg, Berlin

https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en

Growth of databases

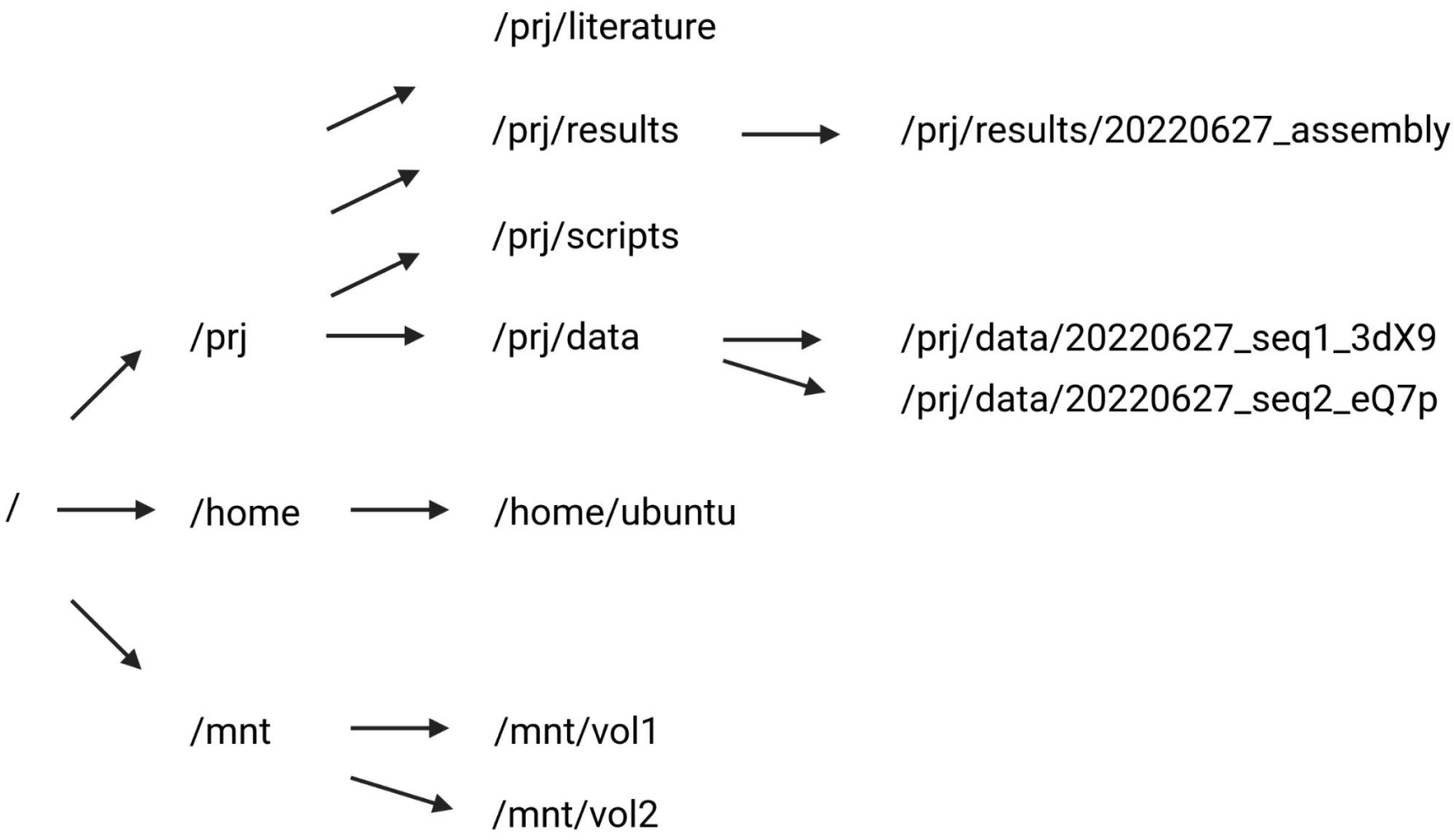
- Data generation costs are dropping
  - Rapid sequencing technology development
  - Resolution of pictures is increasing
  - Robotics supports data acquisition
- Data storage capacities are increasing
- Potential of data reuse is recognized



How to structure your data?

- Document every step (e.g. in a README)
  - Origin of data sets; versions of tools, parameters of analyses
- Keep raw data sets separated from scripts and results
- Sort data by project
- Structure analysis related data sets/results by date

Example: Linux file system



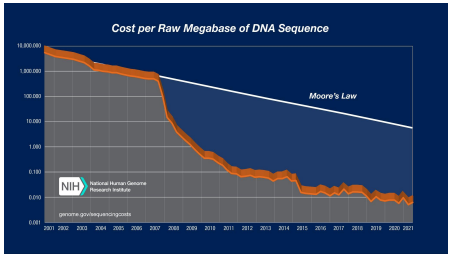
Naming files

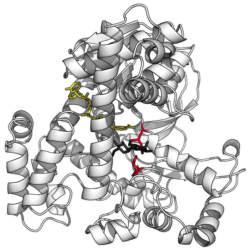
- File names should be informative
- Never use ‘new‘ or ‘final’; use version numbers instead
- Use date as file name prefix (e.g. 2022-06-27 or 20220627)
- Never use spaces in file or folder names (underscore or minus as replacement)
- Include your initials as suffix in collaborative projects

What types of (plant) data do you know?

Types of data in plant sciences

- Sequence data
- Enzyme information
- Scientific publications
- Geographic positions
- Transcriptomics/Proteomics/Metabolomics
- Pictures/Videos
- Phenotyping data





https://commons.wikimedia.org/wiki/File:Glucosidase_enzyme.png

https://commons.wikimedia.org/wiki/File:Azimutalprojektion-schief_kl-cropped.png (CY BY-SA)

Documentation

- Document as much as possible
- Others must be able to repeat your experiments/analyses
- Document dates of data acquisition and processing
- Document all steps of data processing
- Document versions and parameters of applied tools

Electronic Laboratory Notebooks (ELN)

- Links between data sets and documentation
- Avoids repetitive documentation (copy&paste); version control
- Automatic search
- No issues with handwriting
- Quick and regular backups possible
- Accessible from everywhere
- Collaborative

Technical solutions for electronic lab notebooks

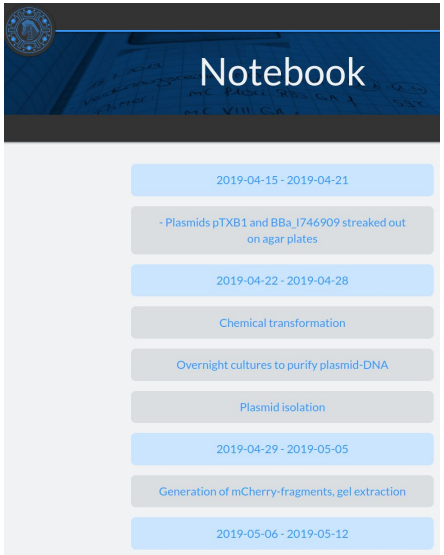
- Simple wiki page

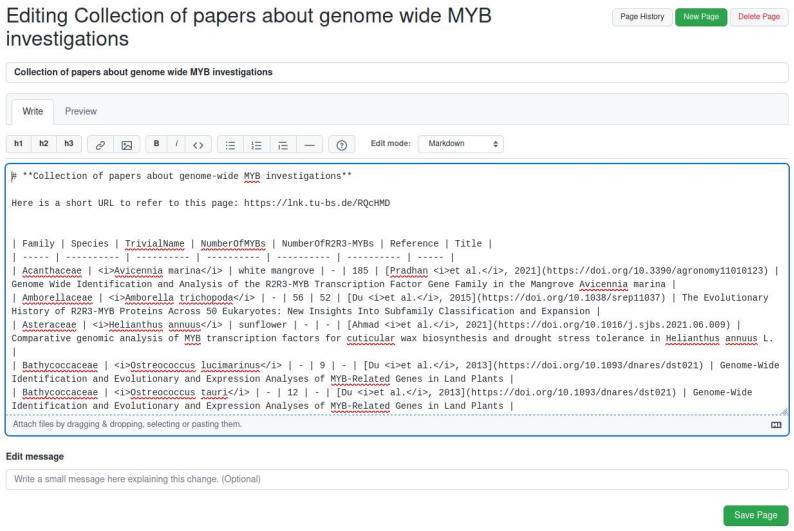
- Dedicated systems developed by research institutions
  - GABI-Kat LIMS
  - Chemotion
  - e!DAL
  - gitlab

- Commercial offers
  - Benchling
  - Dotmatic's
  - Signals Notebook (Perkin Elmer)
  - LabArchives

Example: wiki

- Electronic lab notebook for collaborative learning
- iGEM wiki to document team projects
- Github wiki to document tools

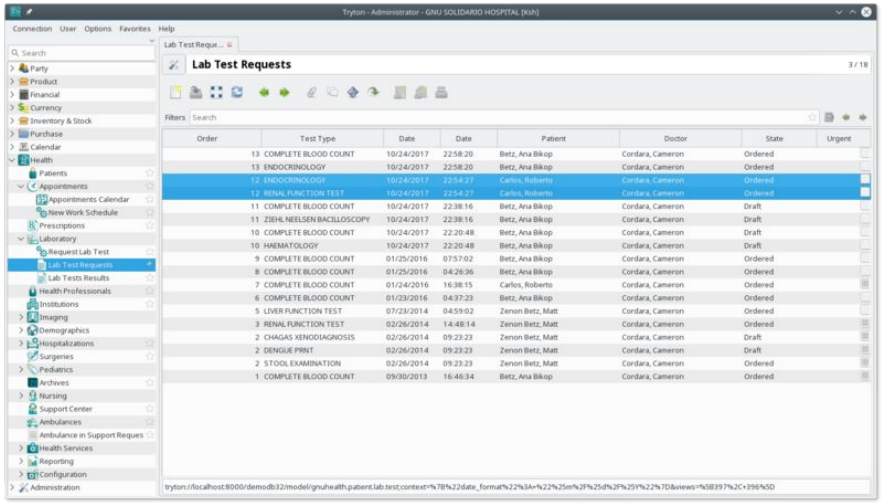




https://2019.igem.org/Team:Bielefeld-CeBiTec/Notebook
https://github.com/bpucker/MybMonday/wiki

Laboratory Information Management System (LIMS)

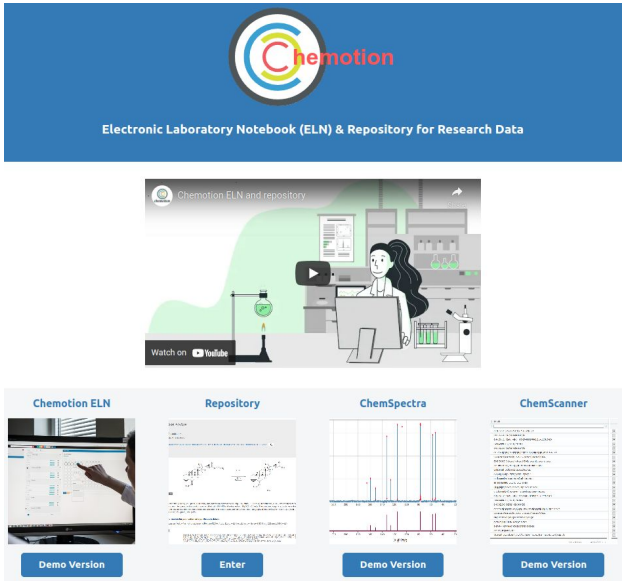
- Workflows are represented and samples are tracked
- Information are linked between analysis
- Different levels of permissions/access
- Examples:
  - Handle a collection of T-DNA insertion lines (GABI-Kat)
  - Manage all samples submitted for sequencing
  - Manage all oligonucleotide orders



https://commons.wikimedia.org/wiki/File:Gnulims_lab_requests.png

Example: Chemotion

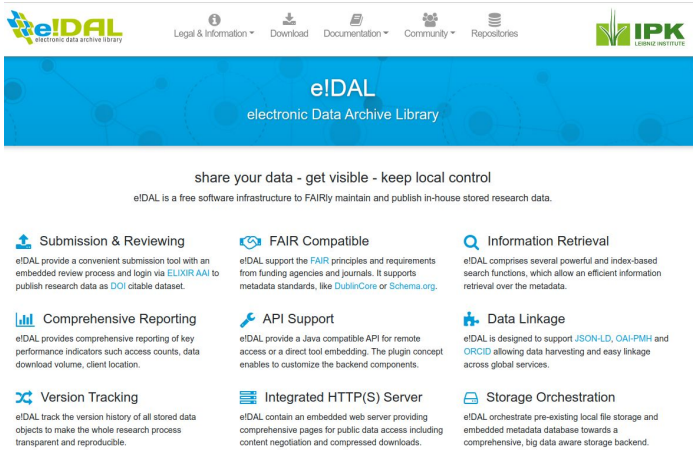
- Research data management tool for chemists
- Electronic Laboratory Notebook (ELN)
- Repository for research data (easy transfer from ELN)
- DOIs are assigned to datasets and protocols



https://www.chemotion.net/chemotionsaurus/index.html; © Chemotion - KIT

Example: e!DAL

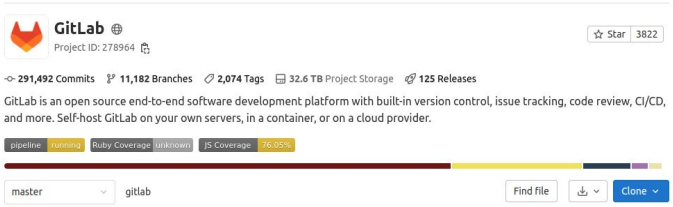
- Maintained at IPK; supported by de.NBI
- DOIs are assigned to submitted data sets
- Reporting about access statistics
- Version tracking
- FAIR compatible and data linkage



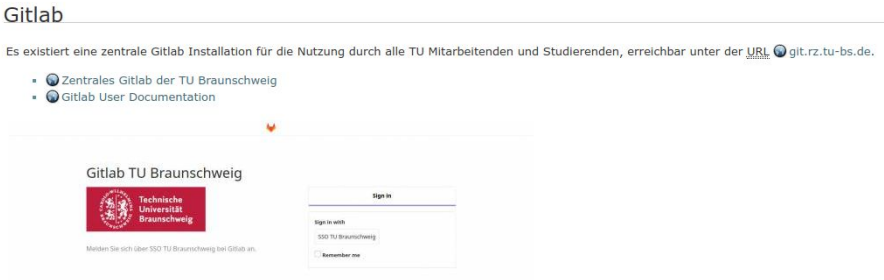
https://edal.ipk-gatersleben.de/

Example: gitlab

- Free version control solution
- Can be used for software development, but also suitable as ELN
- GITZ offers a central service at TUBS: https://doku.rz.tu-bs.de/doku.php?id=server:gitlab
- Non-commercial alternative to github



https://gitlab.com/gitlab-org/gitlab



https://doku.rz.tu-bs.de/doku.php?id=server:gitlab

Example: Benchling

- Cloud-based platform for biotechnology
- Templates for documentations
- Convenient to use
- Compatible with various platforms
- Suitable for certified processes

Lab 4.0: digital lab for higher efficiency

- Connection of lab and analysis processes
- Data from different instruments are synchronized through a cloud
  - Values and pictures are directly inserted into the lab book
- Automation of processes (robotics)
  - Pipetting robot for large scale sample processing
  - Automatic phenotyping facilities
- Samples are labeled with barcodes
  - Barcodes and scanning avoid human errors
  - Can also prevent fraud

Summary: data is everywhere

- Size of data sets is growing
- More sophisticated data collection methods
- Databases enable dissemination/reuse
- Electronic documentation
- Digitalization makes processes more efficient

Linux introduction (1)

Connection to virtual machine (VM):

```
$ ssh -i /path/to/private_key ubuntu@123.133.7.49 -p 1234
(base) ubuntu@agilezuse-10552:~$
```

'$' is used to indicate that following text needs to go into terminal

'#' indicates comment (should not be transferred into terminal)

Frequent issues:

1) Path to private key file not correct
2) Private key file too public

Linux introduction (2)

- Moving through the folder structure:
  - $ cd /full/path/to/folder #change into specific folder
  - $ cd subfolder #change into subfolder
  - $ cd .. #change into parent directory

- Checking content of a folder:
  - $ ls #shows content of current folder
  - $ ls -lh #shows more details
  - -l triggers display of additional details
  - -h human readable
  - -a show also hidden files

Linux introduction (3)

- Python scripts:
  - $ python3 contig_stats3.py --inputfile assembly.fasta
  - $ python3 contig_stats3.py --inputfile assembly.fasta --min_contig_len 1000
  - Different parts of command are separated by space
  - Spaces in filenames are not tolerated

- Other tools:
  - toolname parameter1 parameter2 parameter3

Time for questions!

Questions

1. What is Data Literacy?
2. Why is Data Literacy important?
3. What are important considerations when naming files?
4. What types of data exist?
5. What needs to be documented?
6. What are advantages of an ELN?
7. What is LIMS?
8. What characterizes lab 4.0?