



Introduction to Bioinformatics

Molecular and Genomics Informatics Core

By Alexander Lemenze, PhD

Our Journey Today

MEETING AGENDA

Computer Resources

Basics computer components

Operating systems

Local and Cloud resources

Languages and basic access

What is a language

What languages are commonly used

How can I view my data (Genome Browsers)

Databases and Community

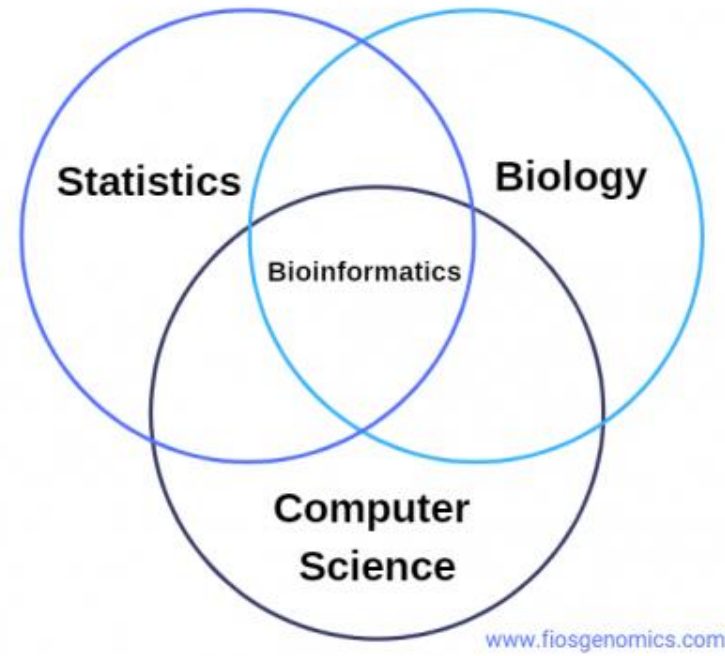
Soap box on reproducibility

Common databases for use

Community help

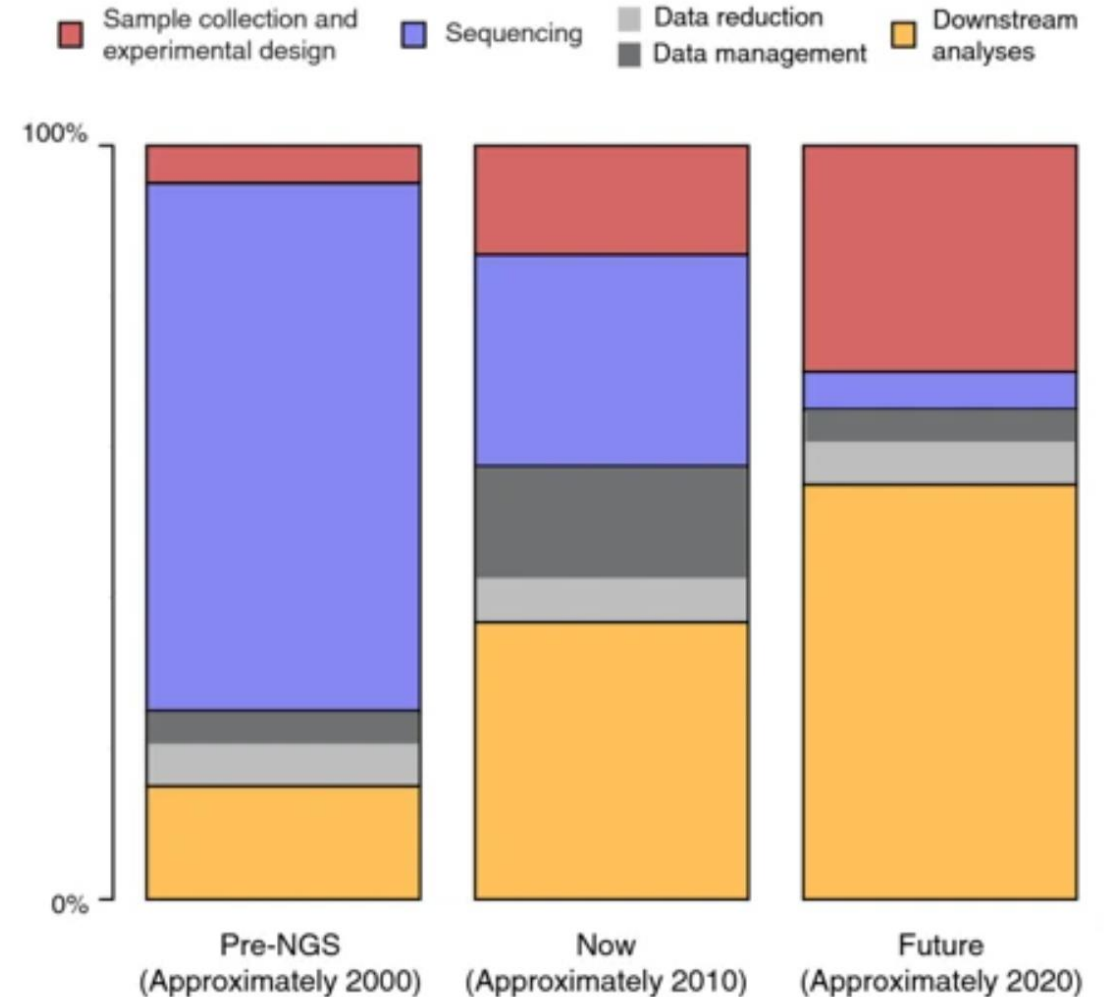
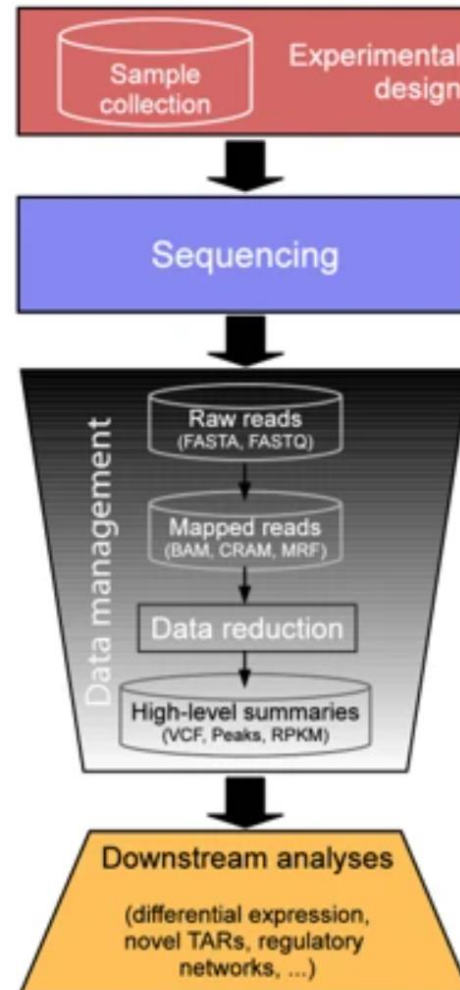
What is Bioinformatics

- Per google:
 - “the science of collecting and analyzing complex biological data such as genetic codes.”
- Types of “bioinformatics”
 - Biostatistician
 - Computational biologist
 - Algorithm development
 - Pipeliners

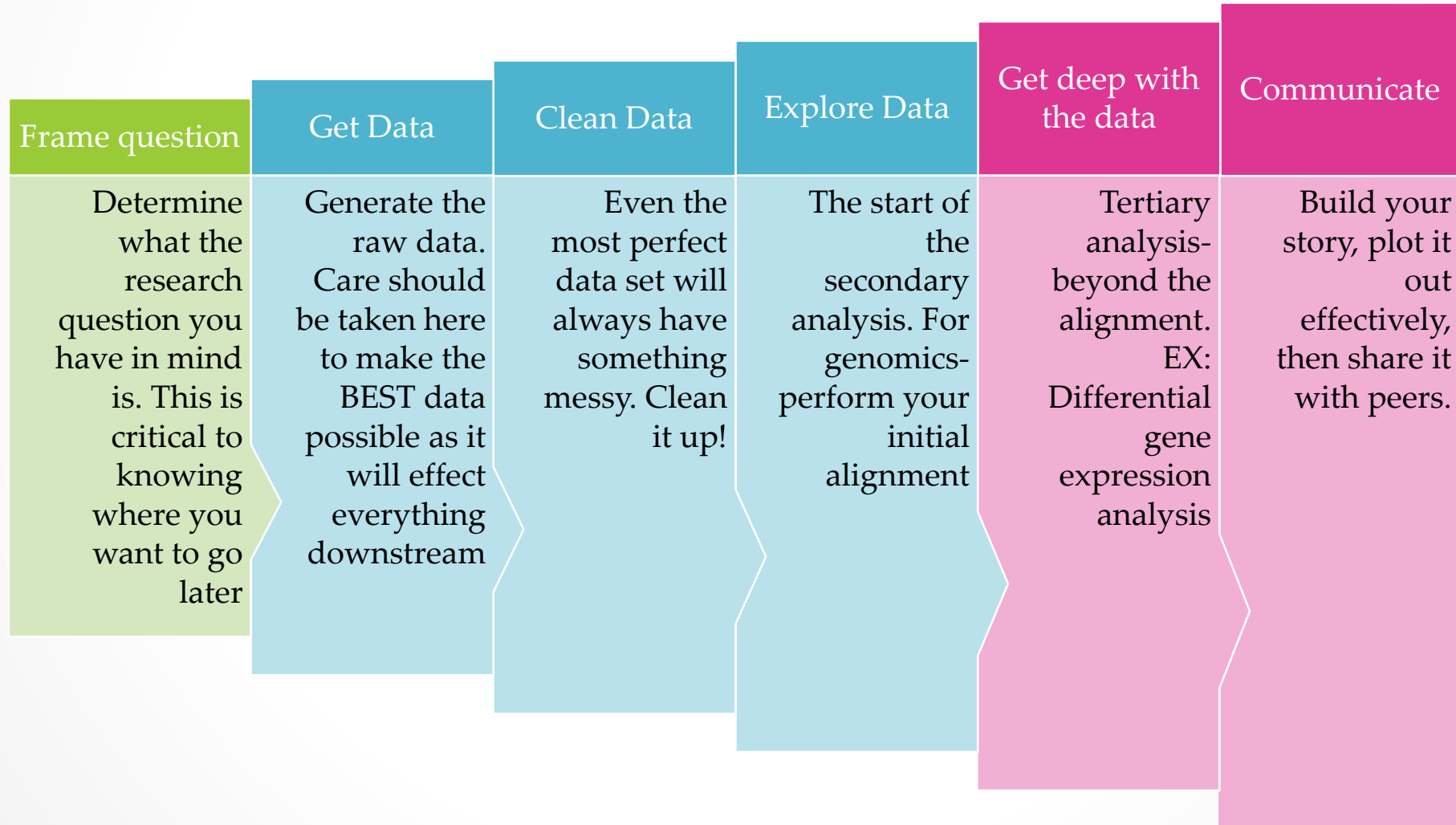


Why Bioinformatics

- Sequencing data is now more cost effective than ever.
- DATA DELUGGGGEEE!!



Stages of data science



An abstract graphic in the top right corner of the slide. It consists of a complex network of small, dark blue circular nodes connected by thin, light blue lines. The nodes are arranged in a way that suggests a hierarchical or interconnected structure, with some nodes having many connections and others having fewer. The overall shape of the network is roughly triangular, pointing towards the top right corner.

Computer Resources

The basics:

CPU

- Central Processing Unit
 - “Brains” of the computer
 - Terms associated:
 - Cores
 - Threads
 - Processors

Minimum:
24 cores

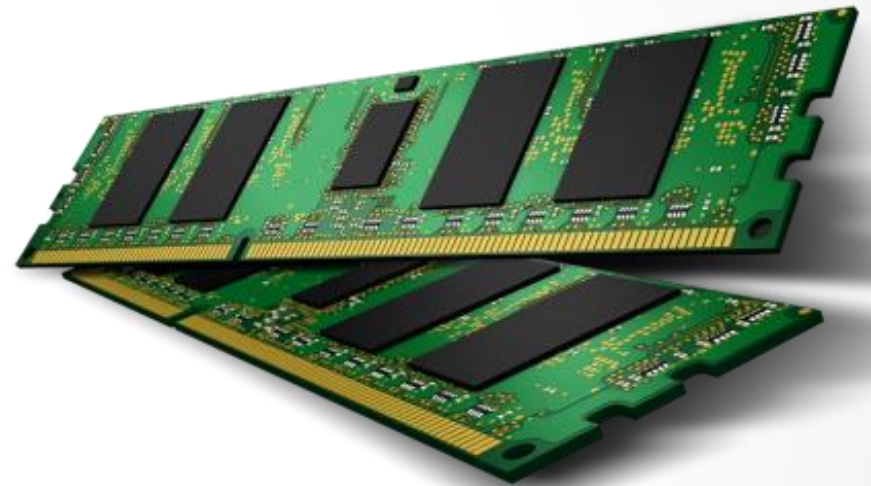


The basics:

RAM

- Random-access memory
 - Thinking capacity of computer

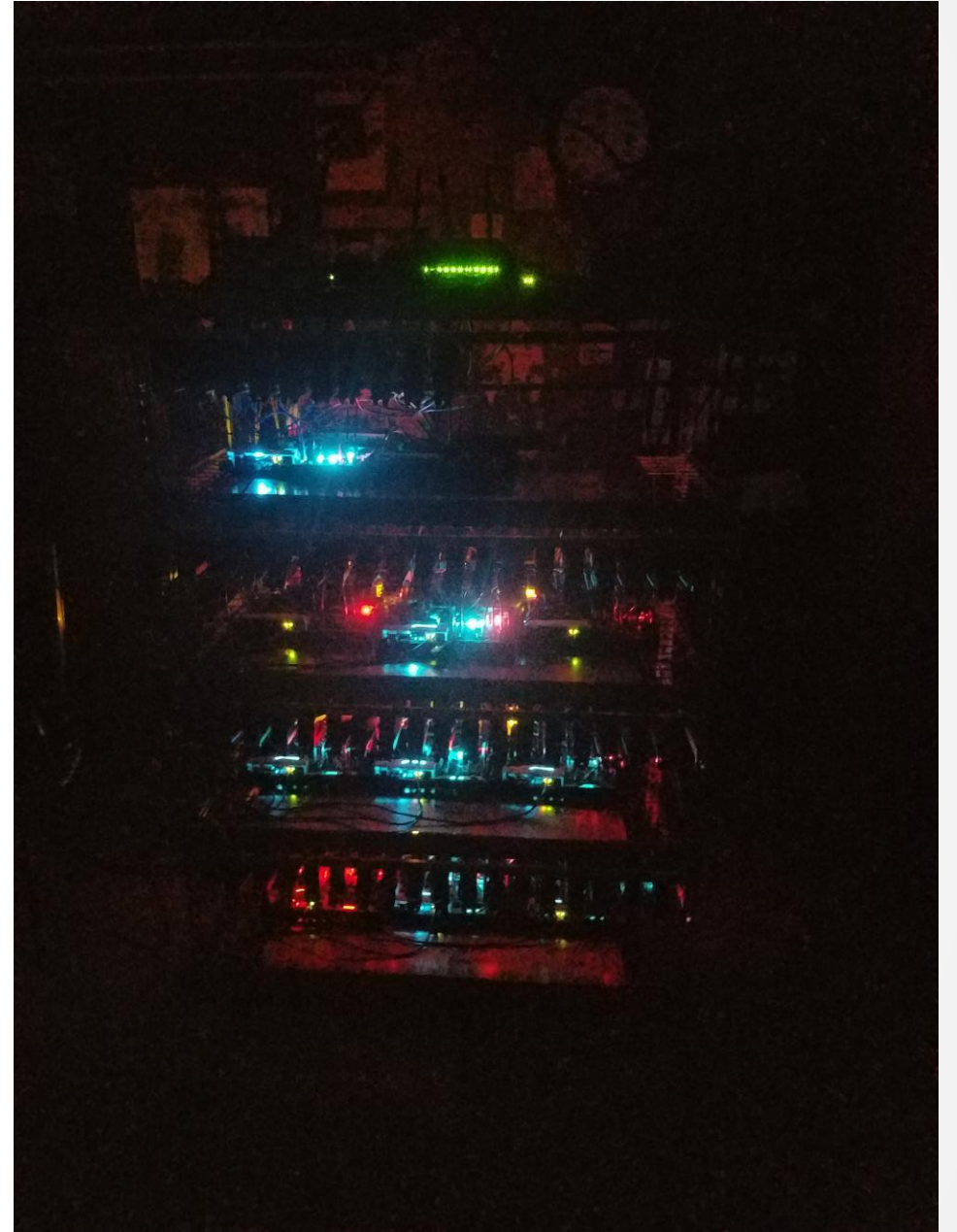
Minimum:
64 Gb



The basics:

GPU

- Graphical Processing Unit
 - Usually for visual effects (hence the graphical)
 - Big for gaming and cryptocurrency
 - Can run algorithms more effectively in some cases
 - -> Process acceleration



The basics:

Hard drive

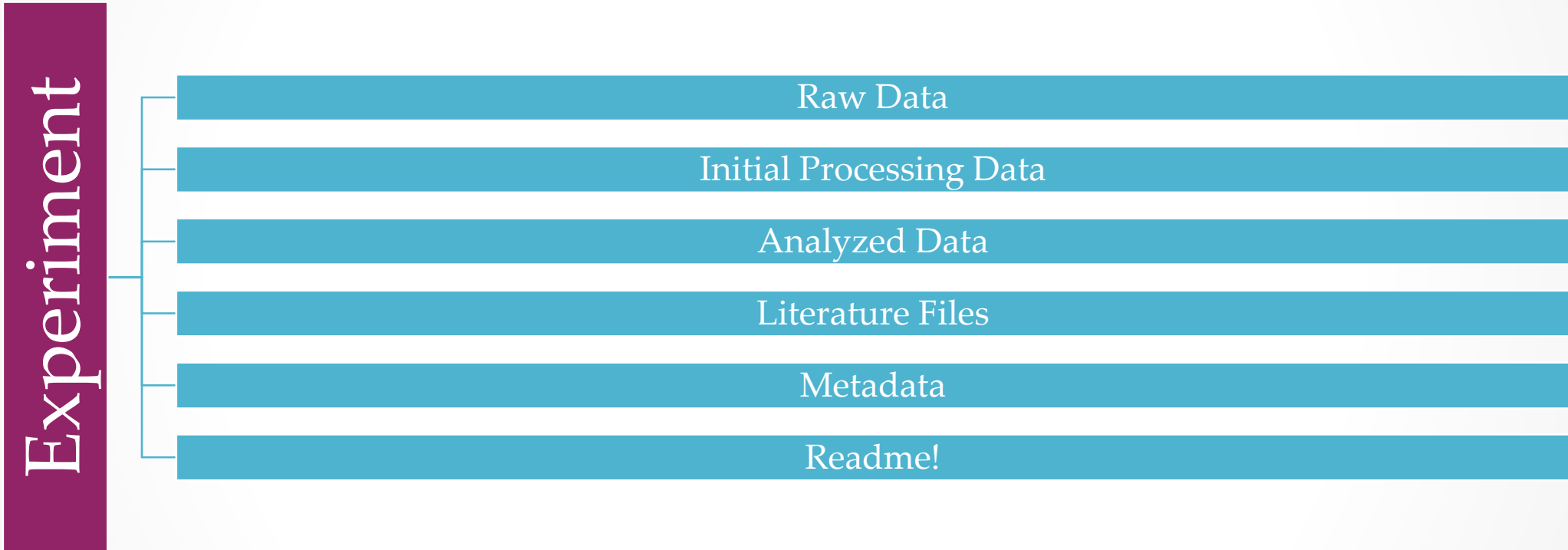
- Hard drive
 - Storage space!
- Estimated raw data per run:
 - Illumina:
 - ~20-200Gb/run
 - ONT:
 - ~2 Tb/run
 - Rich raw data

Minimum for a local machine:
10 Tb



Soap Box on Data management

Example Structure



Operating Systems

- **Windows**

- Infrequent for bioinformatics
- Many tools incompatible
- Drastically improving with virtual machines and linux integration

- **Mac**

- Unix based like linux, more often used for bioinformatics and visuals.

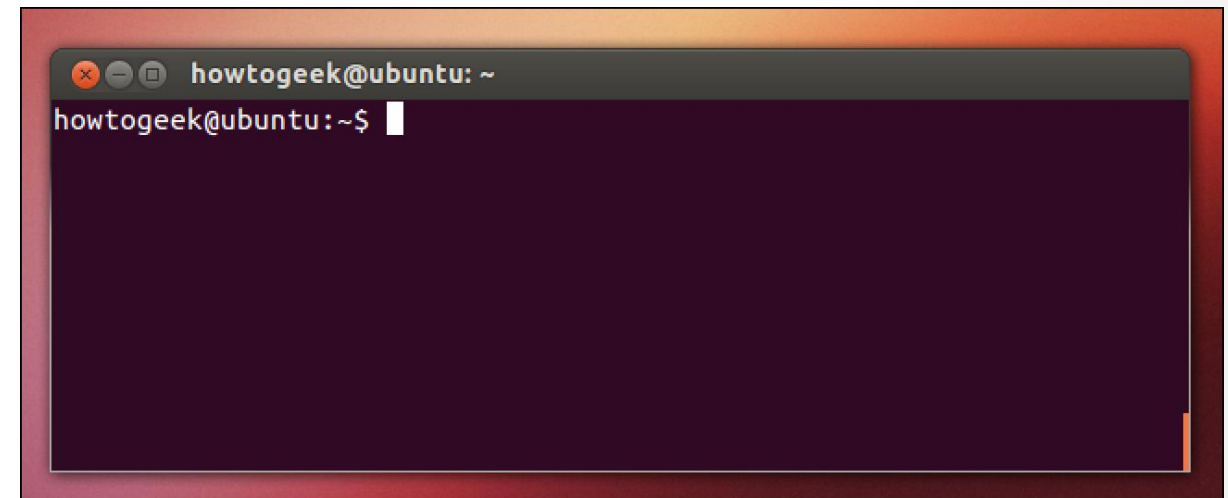
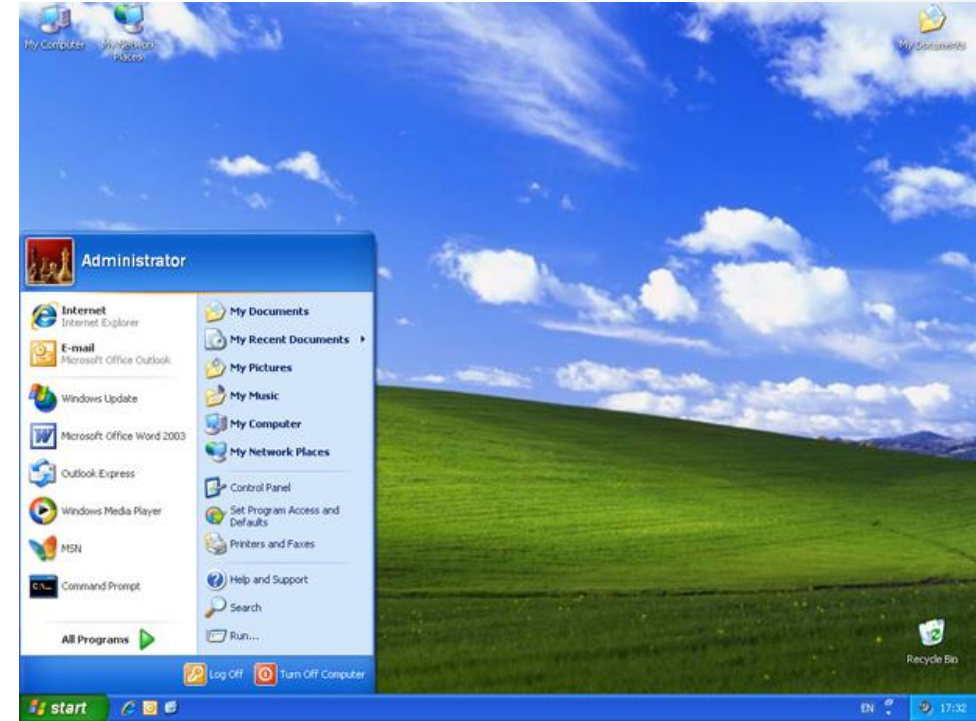
- **Linux**

- Flavors of Ubuntu, Centos, Redhat, etc etc
- Unix based. Most common platforms for bioinformatics.



GUI vs CLI

- Graphical User Interface
 - Pretty
 - Easier to use
 - Slower
 - Limited to presets
- Command Line Interface
 - Not so pretty
 - Much faster
 - Flexible

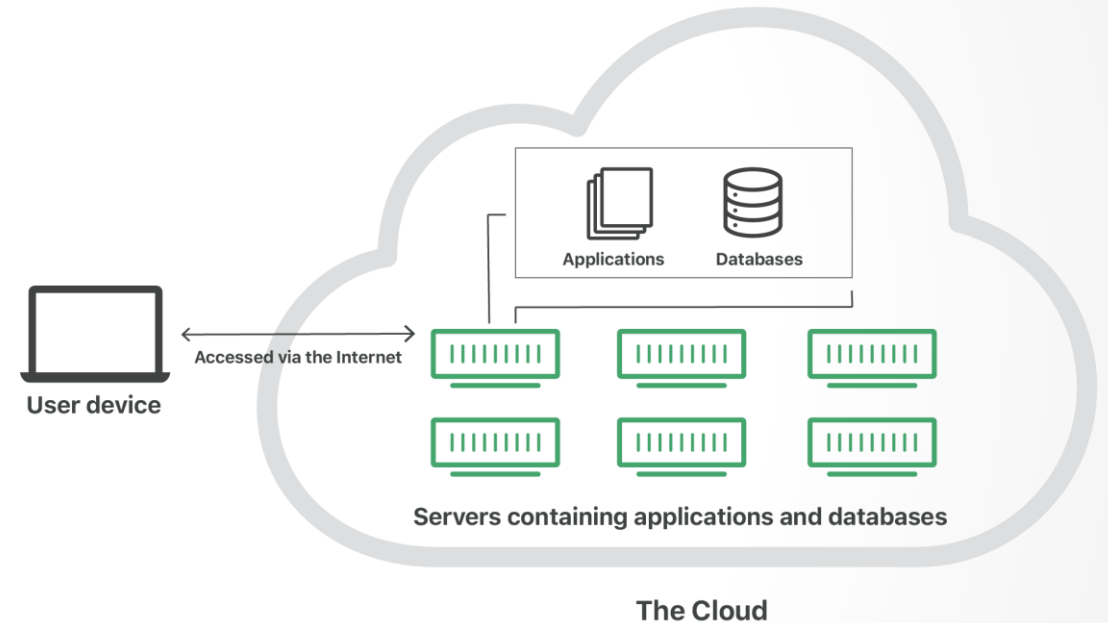


Local Resources

- Workstations
 - Easy to maintain personal set ups
 - Laptop portability
 - Expensive to have full bore station for each member
 - Usually defaults to a mid range station per person
- High Performance Compute Cluster
 - Intensive computer power
 - Scalable to needs
 - Shared amongst investigators to reduce downtime
 - Amarel @ Rutgers OARC

What is the Cloud

- Decentralized and distributed compute and storage that are accessed through the internet



<https://www.cloudflare.com/learning/cloud/what-is-the-cloud/>

Cloud Resources

- Amazon Web Services
 - AWS
- Google Cloud Platform
 - GCP
- Microsoft Azure Cloud
 - Azure



Google Cloud Platform



An abstract graphic in the top right corner of the slide. It consists of a complex network of thin, light blue lines connecting numerous small, dark blue circular nodes. The nodes are distributed across the upper right portion of the slide, with some clusters being denser than others, creating a web-like or molecular structure.

Languages and Basic tools

What is a programming language

- Computers speak in 1's and 0's
- Allows for logic to be constructed
 - Functions
 - Targets/objects
 - Data structures



Basic Language

Unix/Bash

- Standard operations
- First line scripting



```
BASH(1)                                     General Commands Manual                                     BASH(1)
NAME
    bash - GNU Bourne-Again SHell

SYNOPSIS
    bash [options] [command_string | file]

COPYRIGHT
    Bash is Copyright (C) 1989-2013 by the Free Software Foundation, Inc.

DESCRIPTION
    Bash is an sh-compatible command language interpreter that executes commands read from the standard input or from a file. Bash also incorporates useful features from the Korn and C shells (ksh and csh).

    Bash is intended to be a conformant implementation of the Shell and Utilities portion of the IEEE POSIX specification (IEEE Standard 1003.1). Bash can be configured to be POSIX-conformant by default.

OPTIONS
    All of the single-character shell options documented in the description of the set builtin command can be used as options when the shell is invoked. In addition, bash interprets the following options when it is invoked:

    -c      If the -c option is present, then commands are read from the first non-option argument command_string. If there are arguments after the command_string, they are assigned to the positional parameters, starting with $0.
    -i      If the -i option is present, the shell is interactive.
    -l      Make bash act as if it had been invoked as a login shell (see INVOCATION below).
    -r      If the -r option is present, the shell becomes restricted (see RESTRICTED SHELL below).

Manual page bash(1) line 1 (press h for help or q to quit)
```

Basic Language

Python

- Higher level programming language
- General purpose and focused on readability
- Most common uses- full programs and data wrangling



Basic Language

R

- Another higher level language
- Primarily for statistical and graphical operations
- Most common uses then- maths and plotting



Basic Tools

Viewing your data

- Genome Browsers
 - Interactive Genome Browser (IGV) by the Broad
 - University of California Santa Cruz (UCSC) Genome Browser.



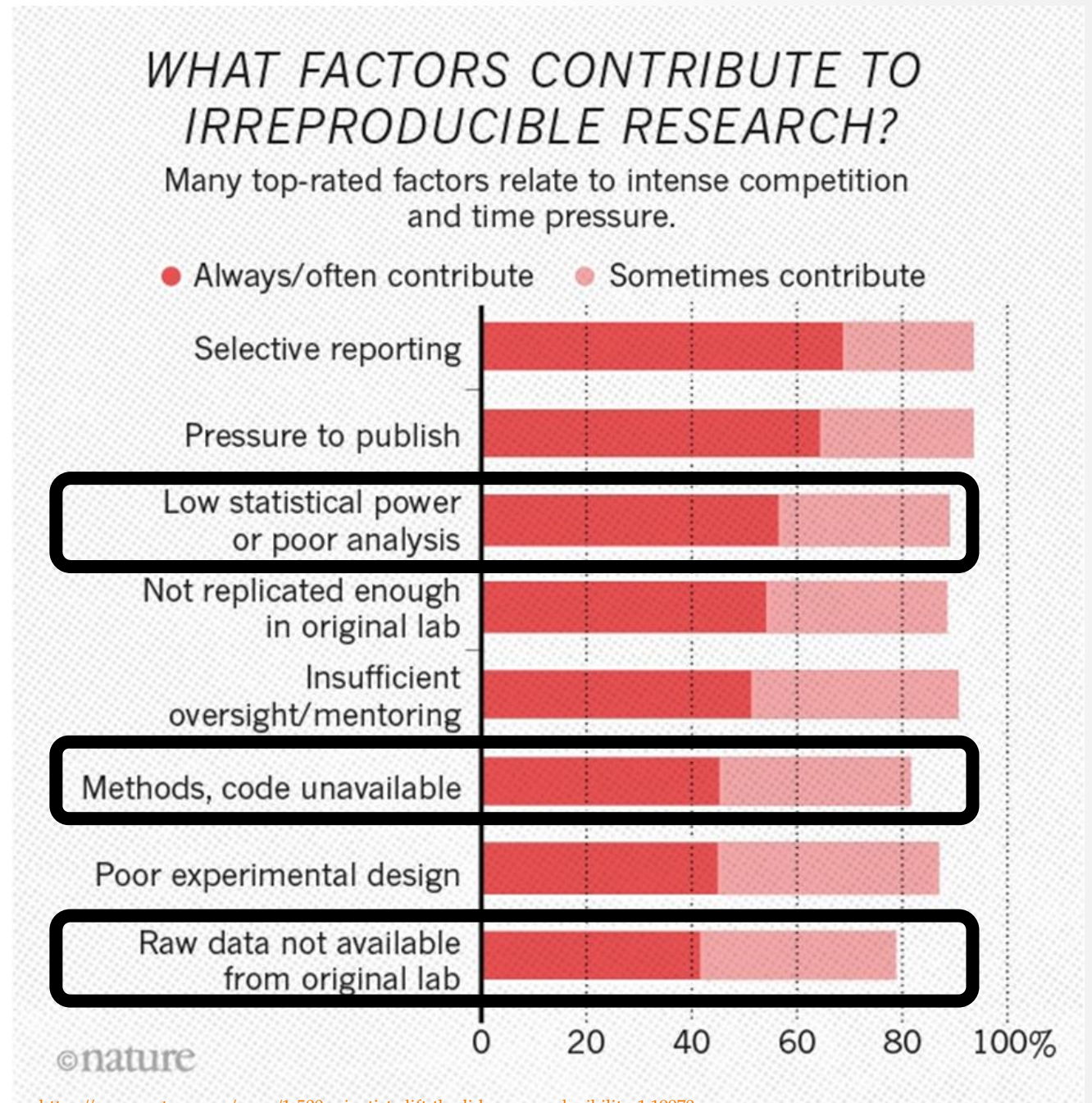
Databases and Community



Reproducibility

- Critically important for science as a whole

Bioinformatics plays into several major facets



Common databases

NCBI

- National Center for Biotechnology Information
 - Hosts end point data
 - Like Pubmed!
 - And a variety of other data sets
 - Genome references
 - BLAST



Common databases

GEO

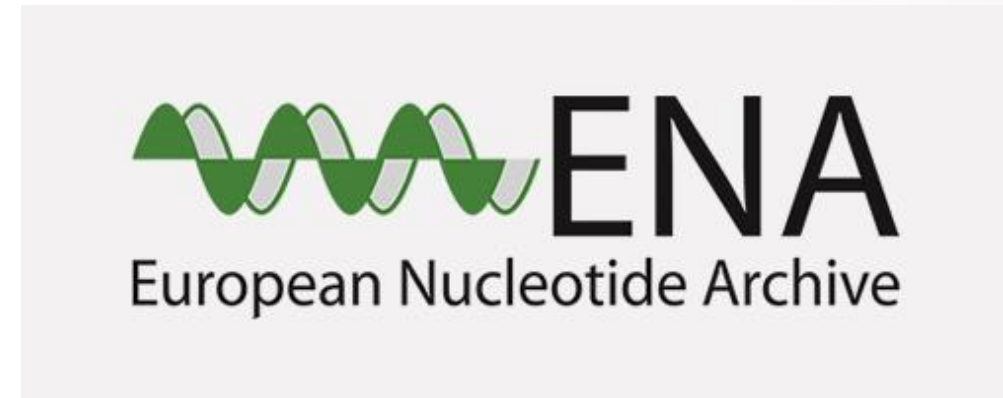
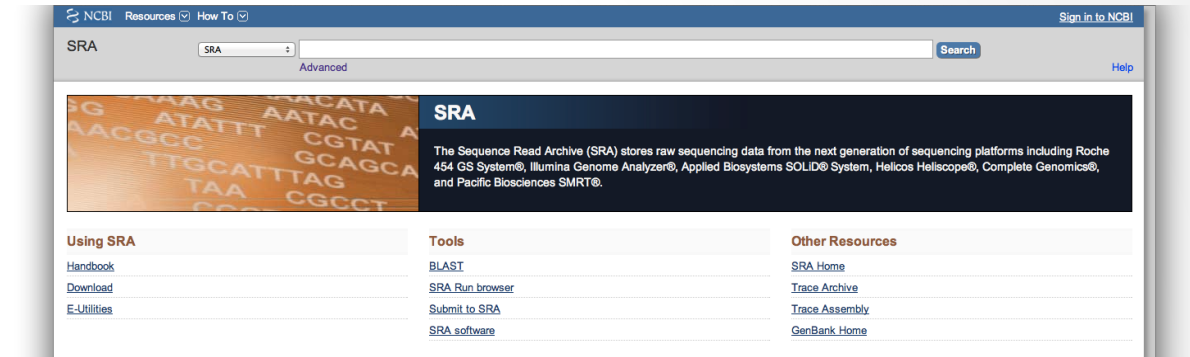
- Gene Expression Omnibus
 - Hosted by NCBI
 - Repository for experiments of expression data
 - “Easily” queriable
 - Currently >3.6 million individual samples



Common databases

SRA/ENA

- Short Reads Archive
 - NCBI hosted
- European Nucleotide Archive
 - European Bioinformatics Institute
 - Also hosts ENSEMBL



Community

Reach out for help

- Biostars
- Stackoverflow
- Google!



DOCTORS:



PROGRAMMERS:





THANK YOU