Bioinformatics on Azure

Insights to Biodiversity with the Microsoft Data Platform

About Us

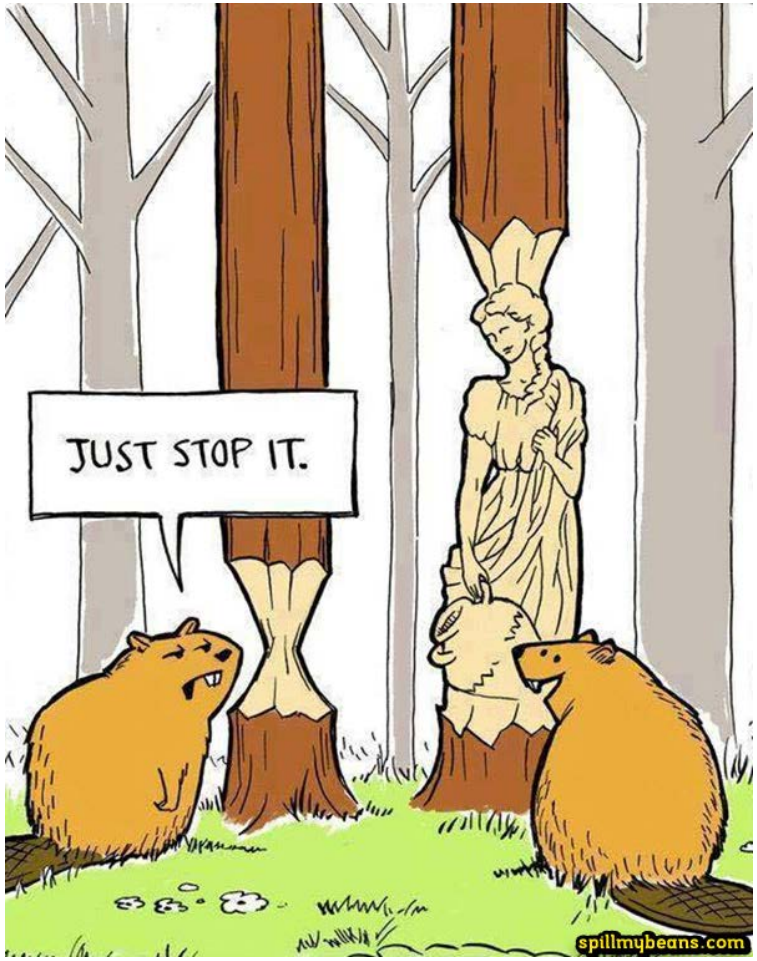
Tillmann Eitelberg

- CEO oh22information services GmbH
- PASS Regional Mentor Germany
- Vice-president PASS Germany
- Chapter Leader Cologne/Bonn, Germany
- Microsoft Data Platform MVP
- www.ssis-components.net
- www.sqlpodcast.de

Ameli Kirse

- Master of Science in Organismic, Evolutionary and Palaeobiology
- Working with Next Generation Sequencing Data
- Focused on molecular biodiversity research
- Currently PhD position on HTP sequencing and analysis

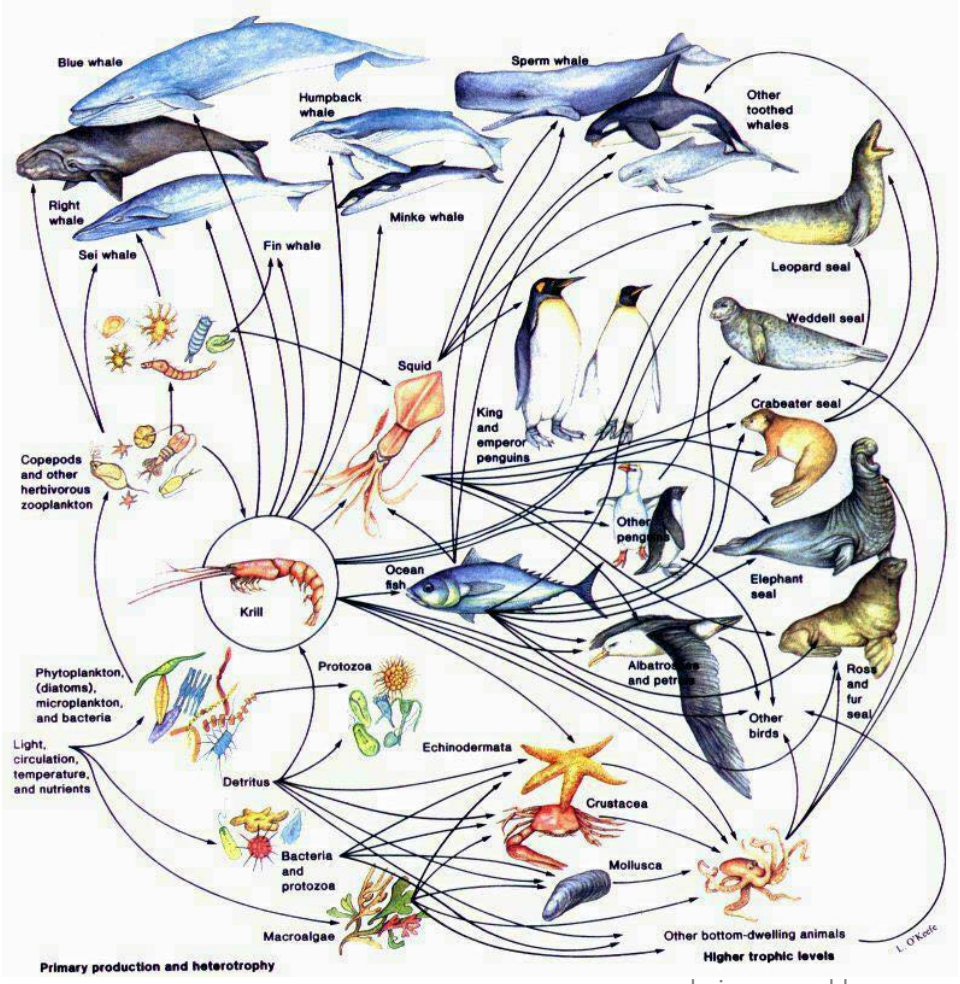
Biodiversity is important



spillmybeans.com

Ecosystem engineers are actively creating, maintaining and modifying habitats

Biodiversity is important



weedscience.weebly.com

Changes in food webs can alter community structures with unforeseeably consequences

Biodiversity is important – ecosystem services

Biodiversity is important

Ecosystem services

Humans benefit from a high biodiversity

→ conservation sites

Biodiversity is important

Osmoderma eremita
(hermit beetle)

Stuttgart 21
(Underground train station)

Examples for active biodiversity protection

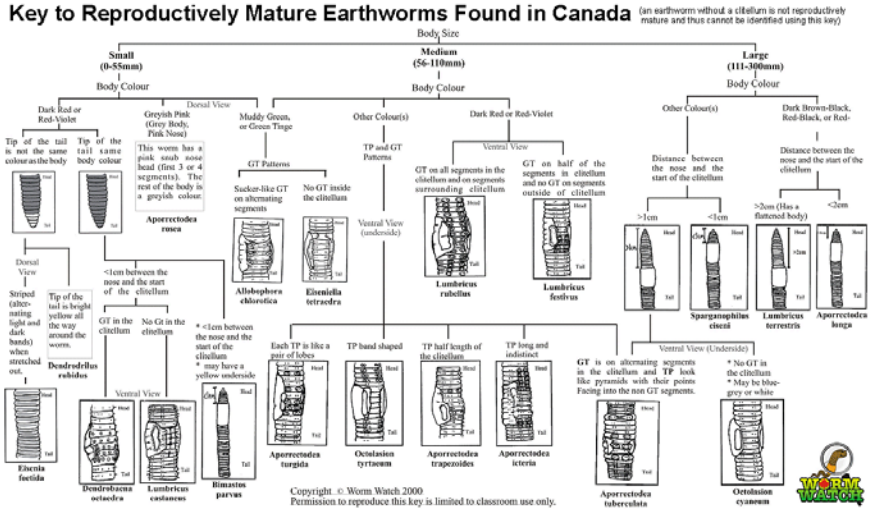
World Youth Day 2005

Epidalea calamita vs. Pope

Species identification yesterday

1. Collection of species (traps, canopy fogging, etc.)

2. Identification on the basis of expertise knowledge and Identification keys

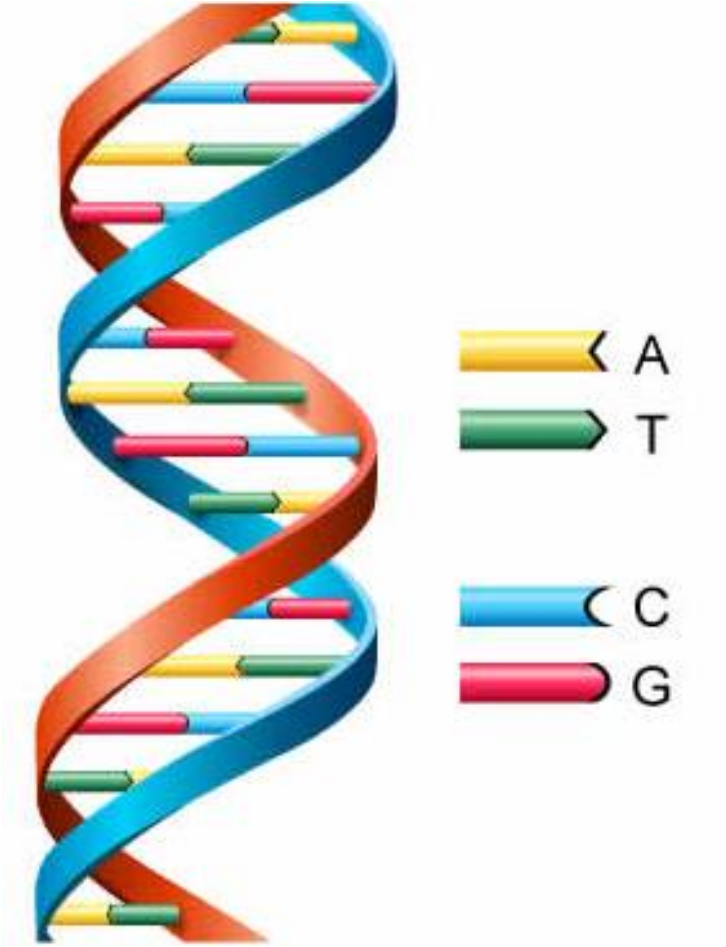


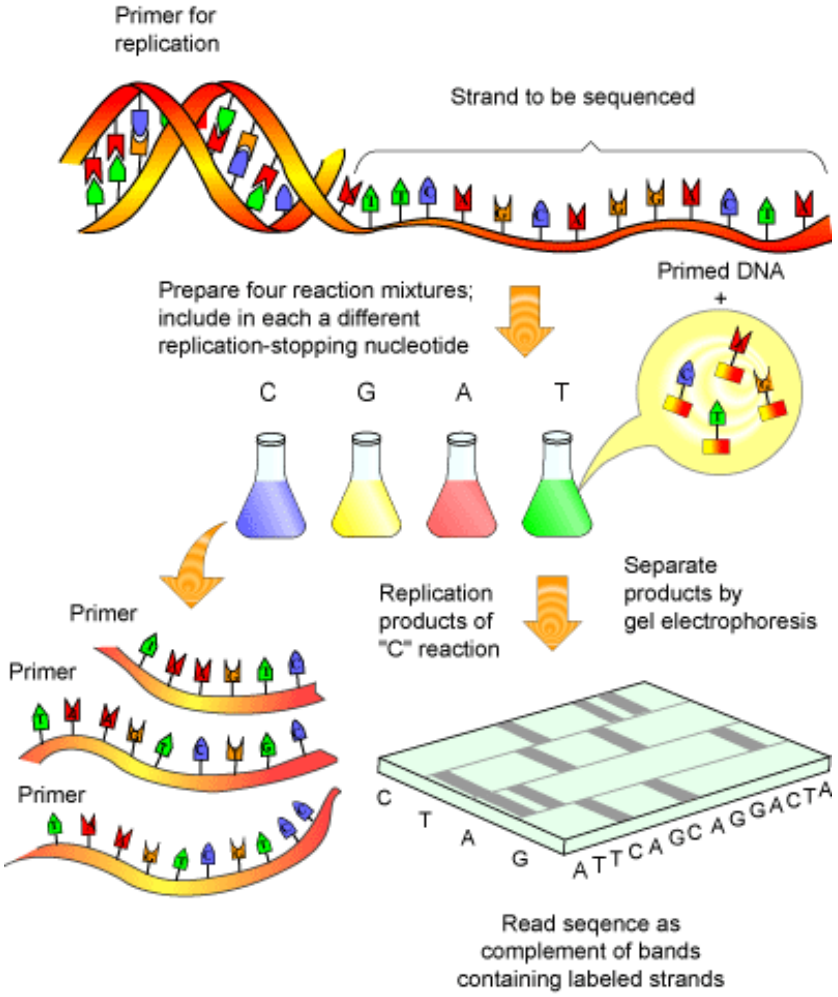
Species identification yesterday

Cryptic species

Phenotypic plasticity

Molecular Biology





Barcoding

Sample collection

Metadata

Photo documentation

DNA extraction

PCR

Sequencing

Barcoding

Time consuming collection of species

Everything found?

Uncultivable bacteria

Next Generation Sequencing

Illumina Miseq

- Bridge Amplification
- Up to 15M reads per run
- 15GB
- Paired-end reads (300bp)

Illumina MiSeq - NGS

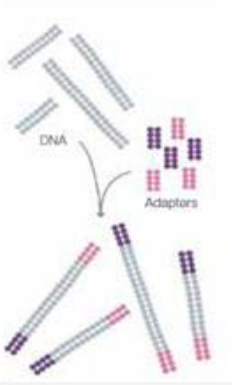


Figure 1

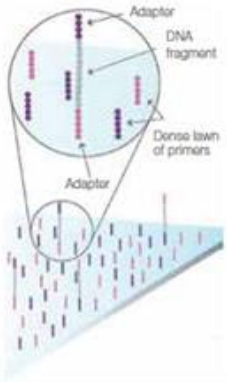


Figure 2

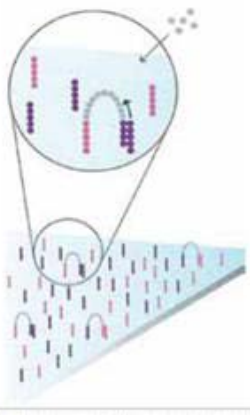


Figure 3

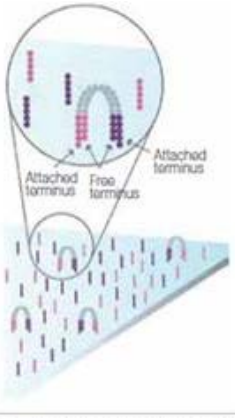


Figure 4

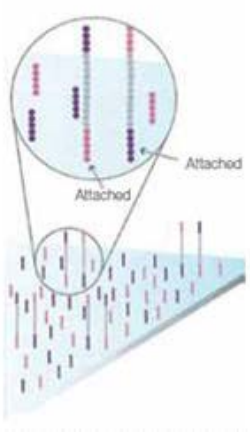


Figure 5

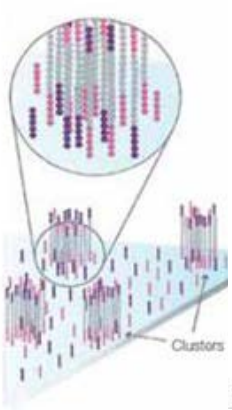


Figure 6

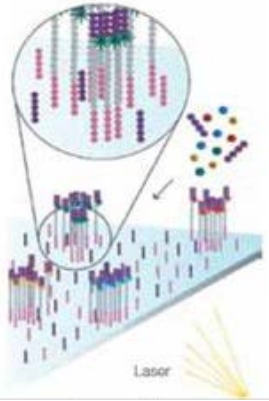


Figure 7

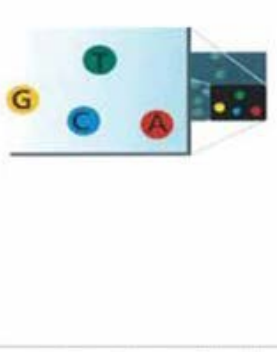


Figure 8

http://www.biorigami.com

Malaise Trap

Environmental Metabarcoding

Every organism is leaving its DNA unconsciously behind

blogs.scientificamerican.com

DNA Extraction

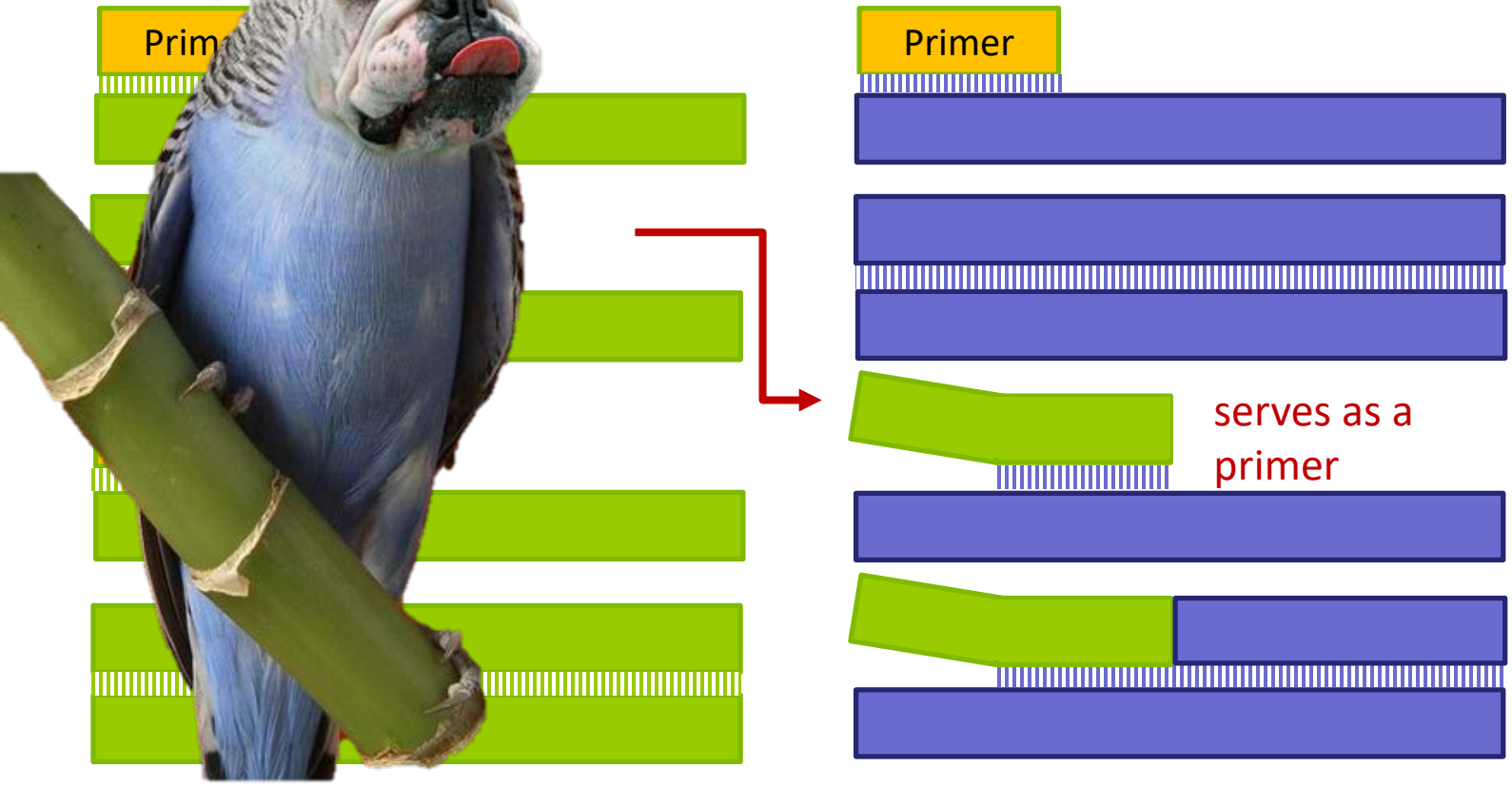
Analyzed via bioinformatic tools

Typical Analysis workflow

- Depending on NGS approach used
  - Illumina: merge paired reads
- Demultiplexing
- Quality filtering
- Clustering (OTU detection)
- Chimera removal
- Comparisson to database / Alignment
- Identification of species

Chimeras

Incomplete extension leading to a foreign DNA strand, which is copi[illegible]letion in the upcoming PCR cycles

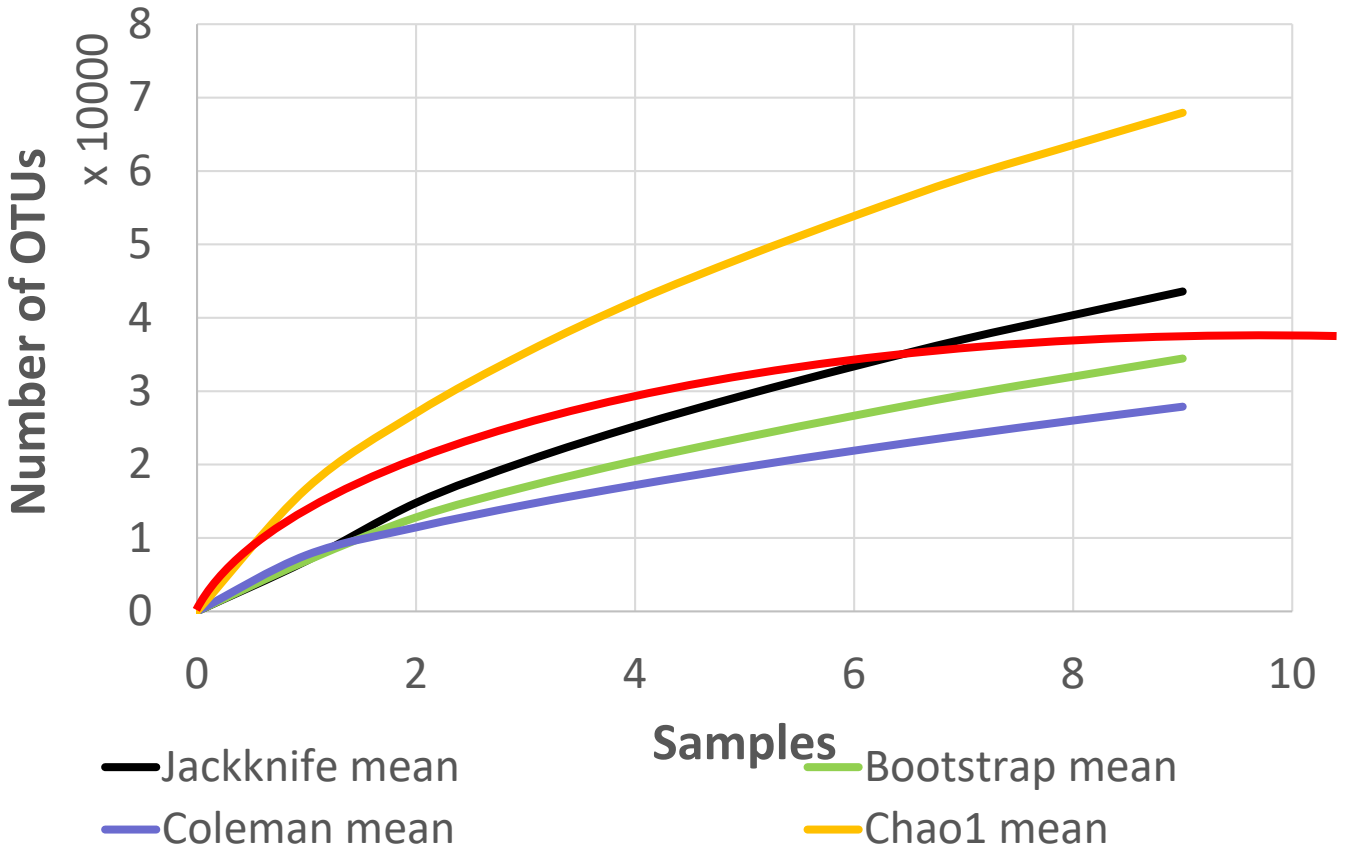


Typical Analysis workflow

- Depending on NGS approach used
  - Illumina: merge paired reads
- Demultiplexing
- Quality filtering
- Clustering (OTU detection)
- Chimera removal
- Comparisson to database / Alignment
- Identification of species

Rarefaction Curve

How do we know if we have taken enough samples?



Pipelines: QIIME

Interface

- Comprises several programs and algorithms
- Written in Python

Pros

1) Broad range of implemented programs + algorithms
2) Several scripts can be run on multiple processors
3) Biom files and R-Packages

Cons

1) Installation
2) Error messages are leaving room for interpretations
3) Analysis steps are often time consuming
4) No information if program is still running or how long it will take
5) Each step is generating a temporary file

Pipelines: QIIME

- Primarily based on python
- Can be used on Mac via MacQIIME
- For VirtualBox as well as for Amazon EC2 exist preconfigured images
- Original statement from QIIME website:

  QIIME has a lot of dependencies and can (but doesn't have to) be very challenging to install.

Pipelines: Mothur

Program

- Several programs and algorithms are implemented
- Written in C++

Cons

1) Expects the user to perform the whole analysis
2) Each step is generating a temporary file

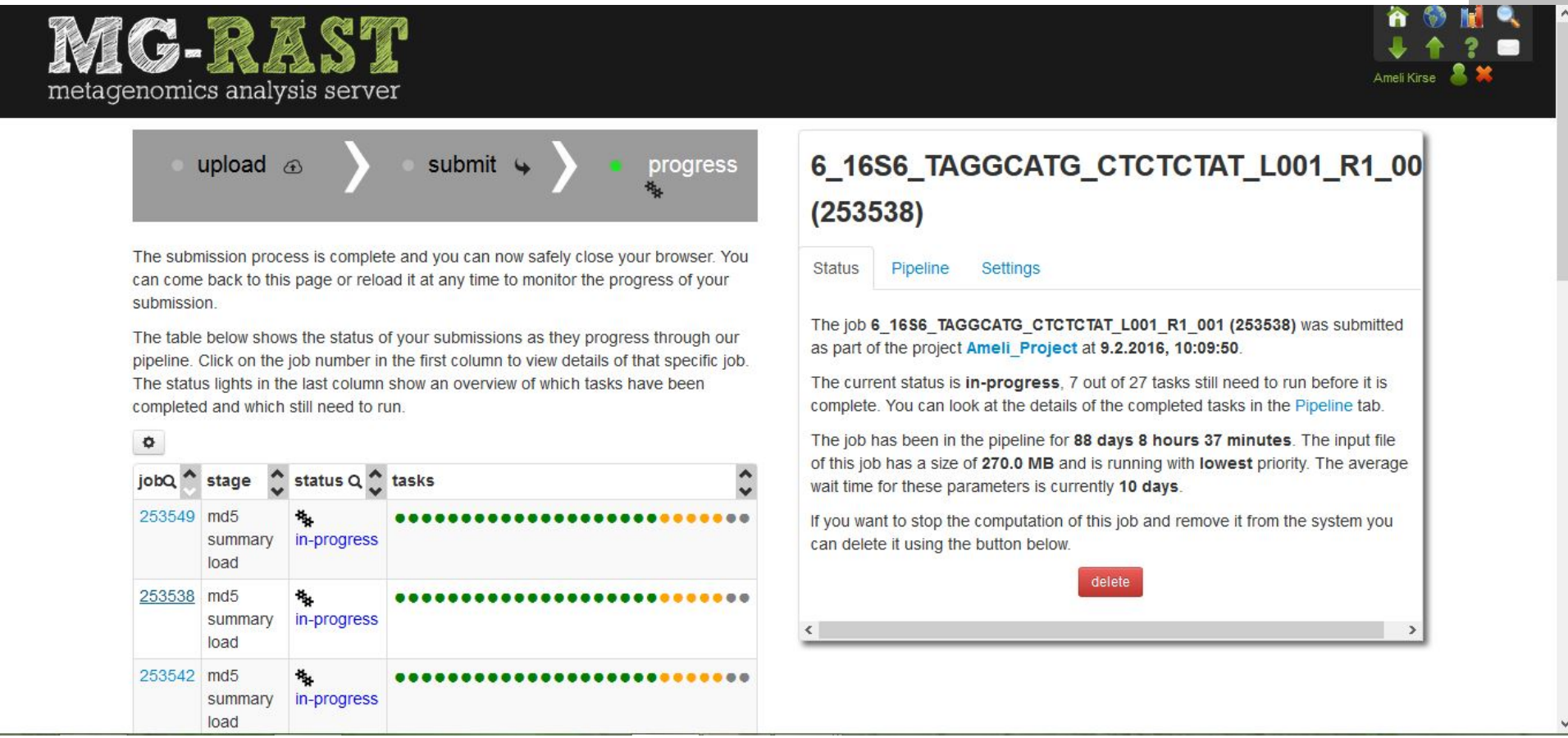
Pros

1) Broad range of implemented algorithms and programs
2) Comparatively short runtime
3) Several statistical and graphical tools are available
4) Very active online community

Pipelines: Mothur

- Available for Linux, Mac and Windows
- Based on C++
- Installation is just decompressing the provided ZIP file
- Active development / release cycles
- Source Code hosted on GitHub

Pipeline: MG-Rast



Pipeline: MG-Rast

- RESTful API
- Uses JSON as its data format
- Hosted on AWS
- MG-Rast is OpenSource and hosted on GitHub
- Local Installation is possible but not supported (We have not been funded to create a readily installable version)

Pipelines - Cons

- Time consuming!
  - Waiting for free cores
- Complex environment which is mostly configured by an administrator
- Information about runtime are lacking
- Limited data storage
- Command line
- Basic knowledge about programming languages is required
- Questionable results

Further analysis

- All pipelines provide tools for the analysis of the dataset
  (alpha-diversity, beta-diversity etc.)
- Biologist love extravagant/individualized plots
- R → Very powerful (ggplot2, vegan, BiodiversityR)
- EstimateS, STAMP, etc.

EstimateS
Statistical Estimation of Species Richness and Shared Species from Samples

Why Big Data – Metasystematics databases

silva
high quality ribosomal RNA databases

GREENGENES
The 16S rRNA Gene Database and Tools

➔ Some online tools are available

silva ngs

500MB

rdpipeline

2000 Sequences

Current Data Storage Requirements

0.5 PB

Twitter's storage needs today are estimated at 0.5 petabytes per year

1 EB

YouTube currently requires from 100 Petabytes to 1 Exabyte for storage

Current Data Storage Requirements

100 PB

The National Astronomical Observatory of Japan devotes ~100 petabytes to storage

100 PB

For genomics more than 100 petabytes of storage are currently used by only 20 of the largest institutions

1 EB

Square Kilometre Array (SKA) project is expected to lead to a storage demand of 1 exabyte per year

Future Data Storage Requirements

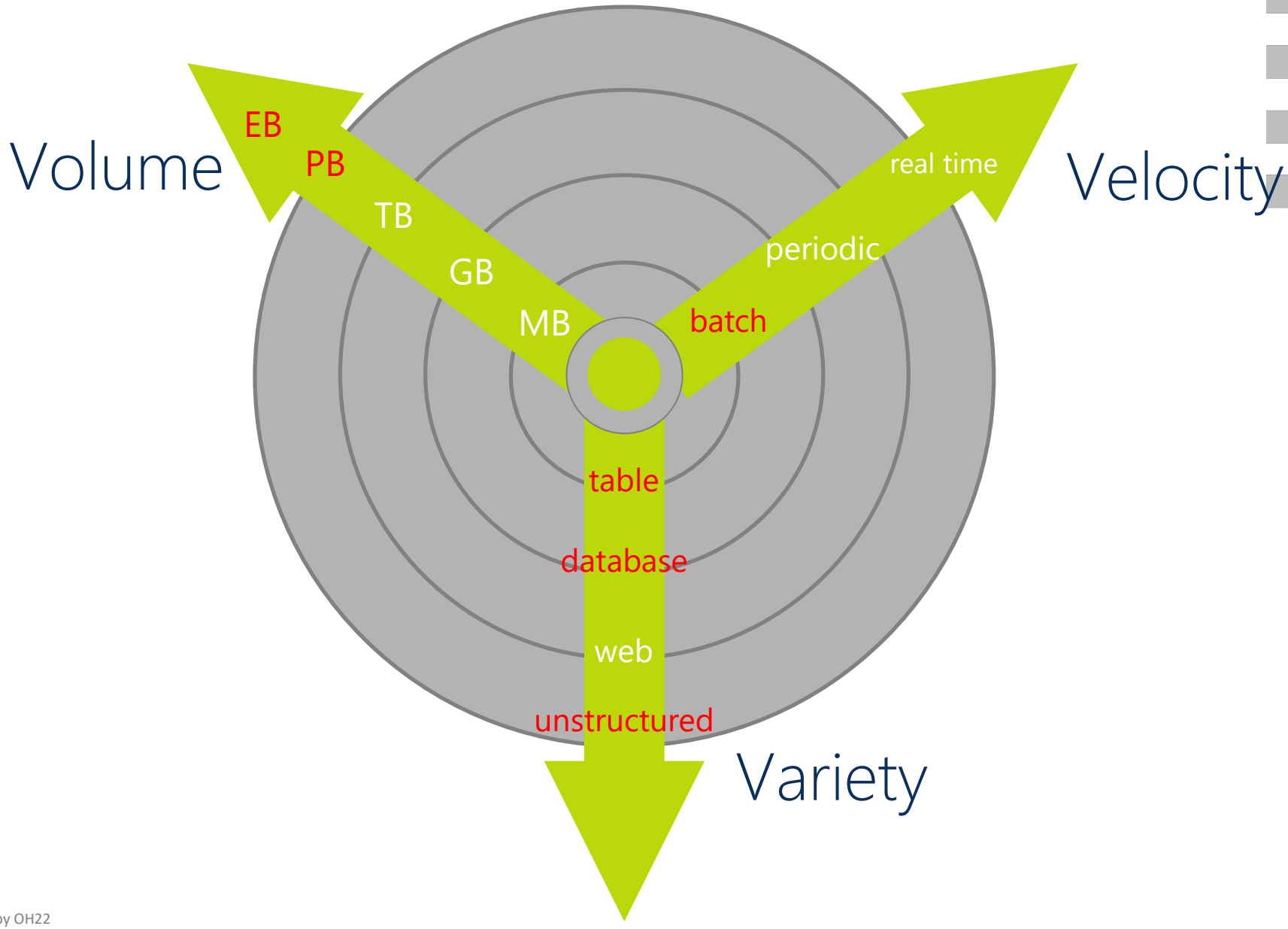
2 EB

YouTube require between 1 and 2 Exabyte's additional storage per year by 2025

40 EB

40 Exabyte's of storage capacity will be needed by 2025 just for the human genomes.

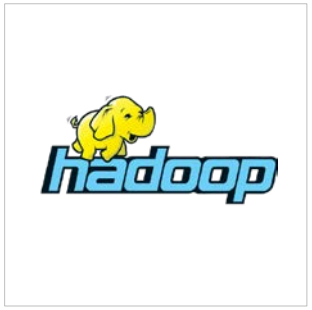
Big Data - **3 Vs of Big Data**

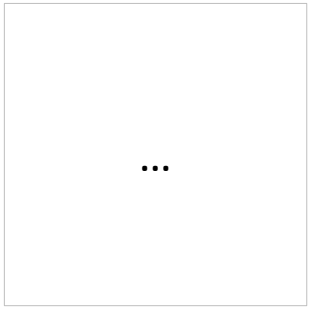


Microsoft Azure “Pipeline”

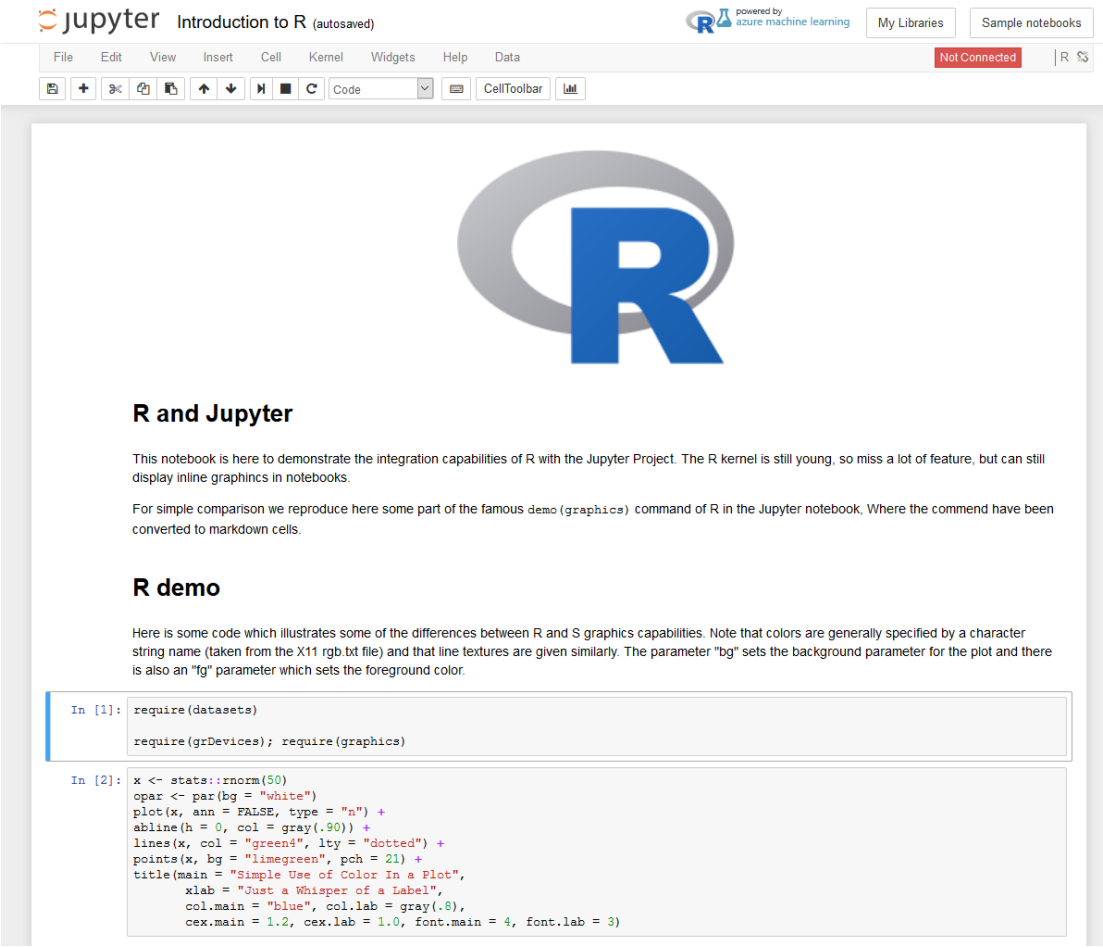
- Use of new techniques that are optimized for large data sets
- Independent Administration
- Virtually no limitation of space and / or Performance
- Optimize Workflows
- Improved presentation of results
- Simple interactive use of the results
- Using established bioinformatics algorithms and libraries

Data Analysis on Azure





Azure Notebooks (Jupyter)



http://notebooks.azure.com

Microsoft Azure – Using R

- With Microsoft Power BI and SQL Server 2016 we can now use R
- This provides more opportunities for the processing of data with existing R scripts and libraries
- Also this offers the possibility to visualize results with Power BI and SSRS

Microsoft Azure - Using R

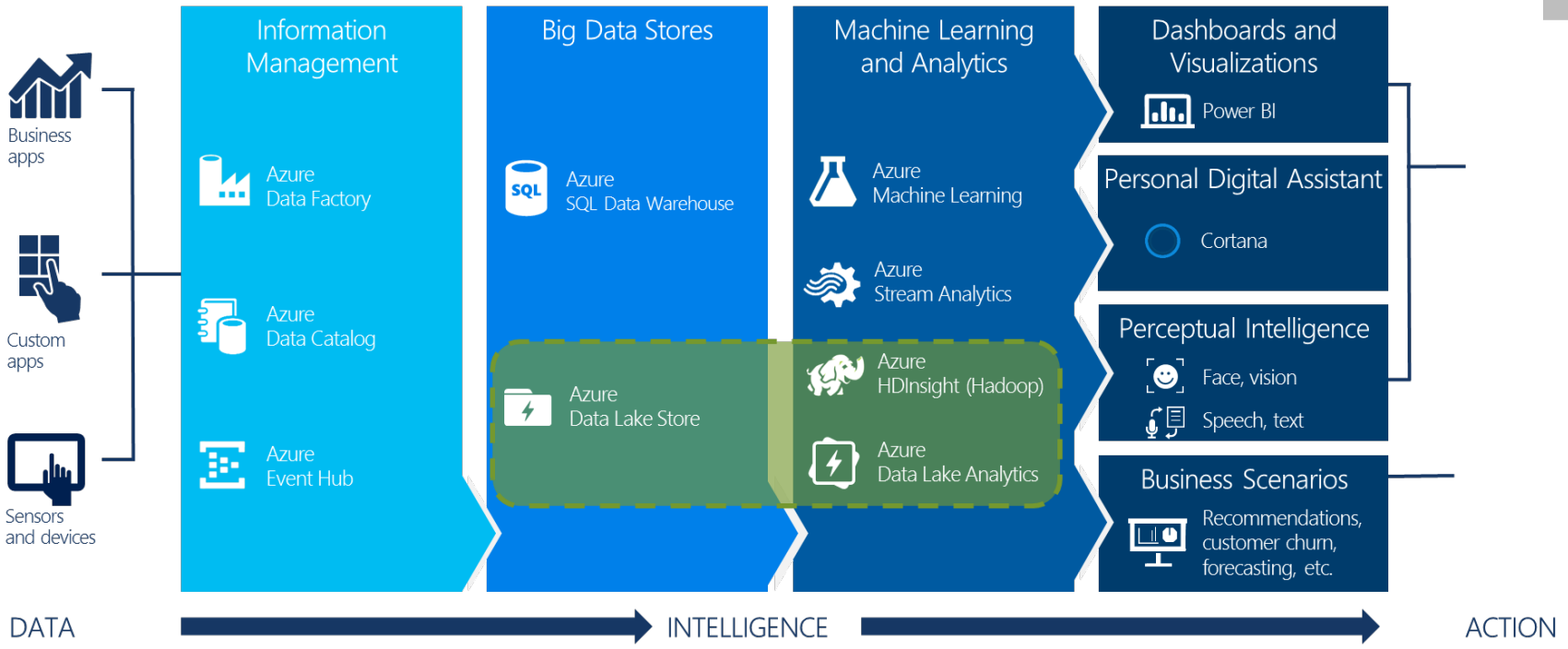
- Easily create a machine with 20 cores and 140 GB Memory
- Windows with R Server or SQL Server R Services
- It’s also possible to use Hadoop + Spark + R Server
- German Cloud provides better data security (from German point of view)

Microsoft Azure - Using R

```
> require("vegan")
Lade nötiges Paket: vegan
Lade nötiges Paket: permute
Lade nötiges Paket: lattice
This is vegan 2.3-2
Warning message:
Paket 'permute' wurde unter R Version 3.2.5 erstellt
> library(vegan)
> AandC <- read.delim("C:/Projects/Ami/AndC/AandC.txt", row.names=1)
> View(AandC)
>
> tran_NMDS <- metaMDS(AandC)
Square root transformation
Error: cannot allocate vector of size 2.2 Gb
1: In as.matrix.dist(dist) :
  Reached total allocation of 15969Mb: see help(memory.size)
2: In as.matrix.dist(dist) :
  Reached total allocation of 15969Mb: see help(memory.size)
3: In as.matrix.dist(dist) :
  Reached total allocation of 15969Mb: see help(memory.size)
4: In as.matrix.dist(dist) :
  Reached total allocation of 15969Mb: see help(memory.size)
>
```

- Partial multithreading
  ➔ Microsoft R Open provides optional multi-threaded math libraries

Azure Data Lake
as part of Cortana Intelligence Suite



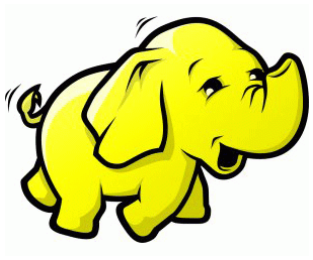
Azure Data Lake

A Platform for data scientists to store and process data of any size and shape.

Batch, real-time, and interactive analytics made easy

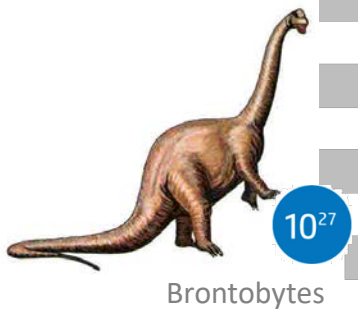
Azure Data Lake Store

- HDFS - Apache Hadoop Distributed File System
- Integrated in Azure Data Lake Analytics and HDInsight
- Future Integrations in
  - Hortonworks, Cloudera, MapR
  - Spark, Storm, Flume, Sqoop, Kafka
  - …

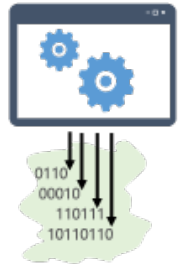


Azure Data Lake Store

- No fixed limits on file size
- Unstructured and structured data in their native format
- Massive throughput to increase analytic performance
- High durability, availability, and reliability
- Azure Active Directory access control

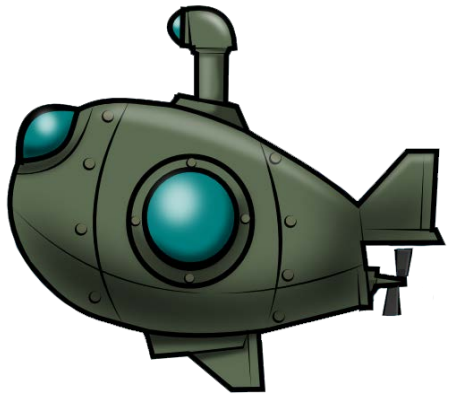


Brontobytes



Azure Data Lake Analytics

- Using U-SQL for analysis
- Process unstructured and structured data
- Extend U-SQL with C#/.NET
  - C# expressions
  - User-defined operators (UDO)
  - User-defined function (UDF)
  - User-defined aggregates (UDAGG)
- Declarative nature of SQL and custom imperative power of C#



.NET BIO

- Bioinformatics library for .NET from Microsoft Research
- includes parsers for common bioinformatics file formats
- algorithms for manipulating DNA, RNA, and protein sequences
- connectors to biological web services
- Open Source Project from Outercurve
- Hosted on GitHub
- Apache 2 License

OUTERCURVE FOUNDATION

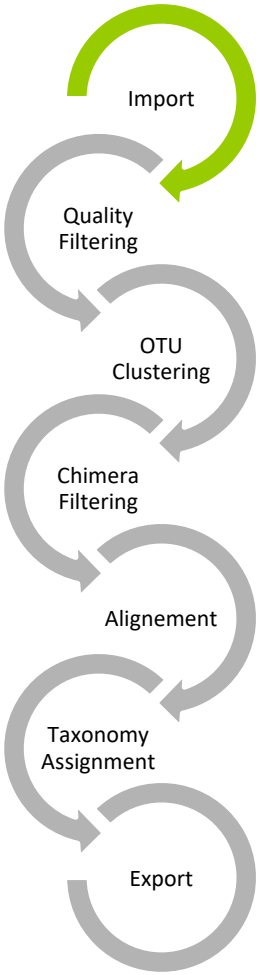
Microsoft Research

Power BI

Power BI

- Comprehensive toolset for reporting
  - Power BI Service
  - Power BI Desktop
  - Power BI Mobile Apps
  - Power BI Developer
  - Power BI Embedded
- A variety out-of-the-box charts
- Additional charts can be generated via R or D3.js

Microsoft Azure “Pipeline"

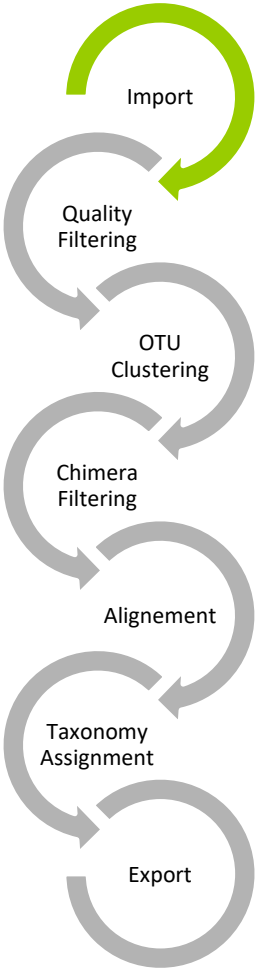


- Import Sequences generated by Illumina Miseq
- FASTQ is a text based format
  - One record goes over 4 lines
  - Record includes sequences as well as quality information's

```
@MSQ-M01442:174:000000000-AEW84:1:1101:10036:8807 1:N:0:CTCTCTACTAGATCGC
TAGCGTATATTAAAGTTGTTGAGGTTAAAAAGCTCGTAGTTGAACCTTGGGCCTGGCTGGCCGGTCCGCCTCA …
+
EFGGGFFEGGGGGDGGGGGGGGGGGGGGGFEFGGFCEGGEGFFGGGGG<FGGGDGGGGGGGGGGGGGGGCCC
```

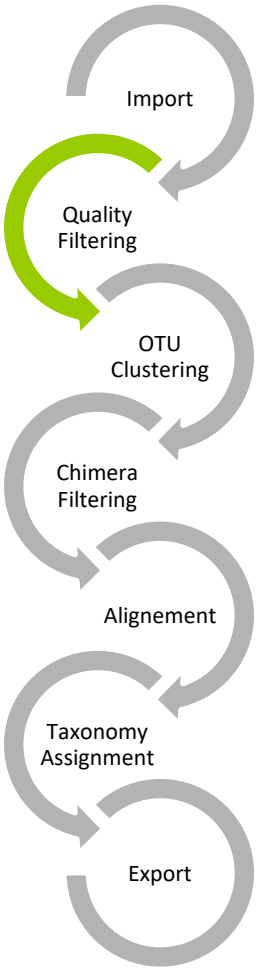
- We also import the SILVA database to avoid using web services

Microsoft Azure “Pipeline"



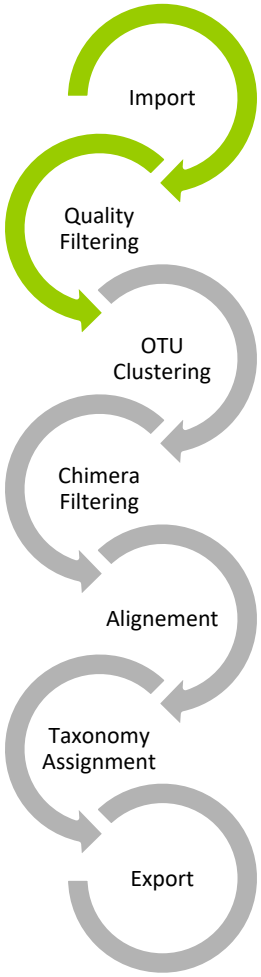
- Custom Extractor for U-SQL
  - C# + dotnetbio Library
- Preparations:
  - Libraries require strong names for Azure Data Lake
  - Some dotnetbio Code has to be adjusted due to some limitations of the original library
    (memory limitations)

Microsoft Azure “Pipeline"



- Filter data which doesn’t reach our quality criteria
  - Remove Primer
  - Remove Sequences shorter than 100 nucleotides
  - Remove non reverse complemented sequences in dataset
    - ACGTACCGAT
    - TGCATGGCTA
  - Ilumnia Quality Score
    - Overall < 19%
    - Per Nucleotid < 15%

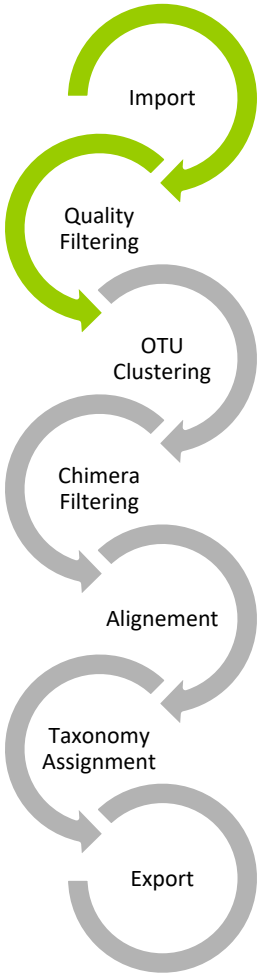
Microsoft Azure “Pipeline"



DEMO

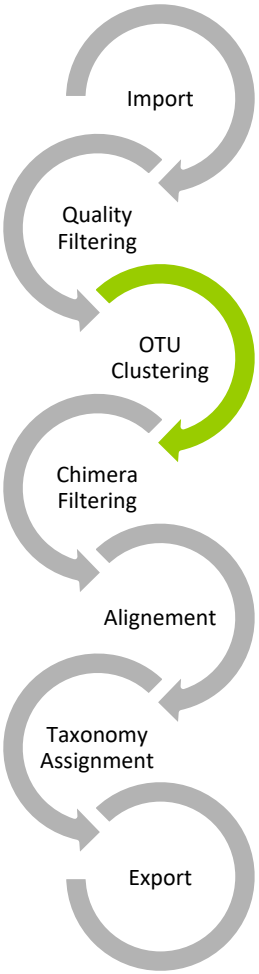
```
public override IEnumerable<IRow> Extract(IUnstructuredReader input, IUpdatableRow output)
{
   if (input.Length > 0)
   {
      var fqParser = new FastQParser
      {
         FormatType = FastQFormatType.Illumina_v1_8,
         Alphabet = Alphabets.RNA
      };

      IList<QualitativeSequence> seqsOriginal = new FastQParser()
         .Parse(input.BaseStream)
         .Cast<QualitativeSequence>()
         .ToList();

      foreach (var item in seqsOriginal)
      {
         SequenceToRow(item, output);
         yield return output.AsReadOnly();
      }
   }
   else
   {
      throw new Exception("input file not found");
   }
}
```

Microsoft Azure “Pipeline"



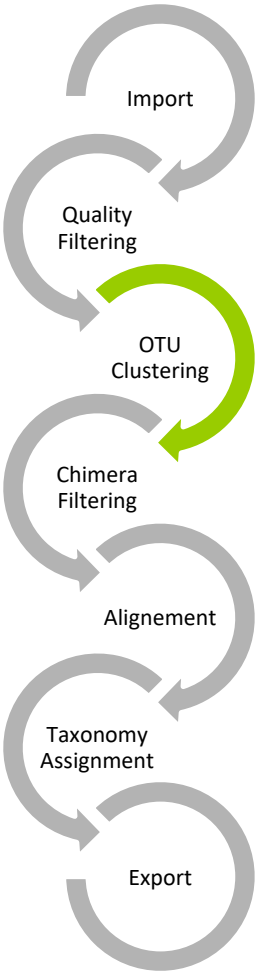

```
DECLARE @sequencePath = @"\SampleData\ERR1273621.fastq";

@e = EXTRACT SequenceId string,
             Sequence string,
             Counter int,
             ReversedSequence string,
             ComplementedSequence string,
             ReverseComplementedSequence string,
             IndexOfNonGap long,
             LastIndexOfNonGap long,
             AlphabetName string,
             HasAmbiguity bool,
             HasGaps bool,
             HasTerminations bool,
             IsComplementSupported bool,
             LowNucleotides int,
             QualityScore double
FROM @sequencePath
USING new oh22is.Analytics.Formats.FastQExtractor();
```

Microsoft Azure “Pipeline"

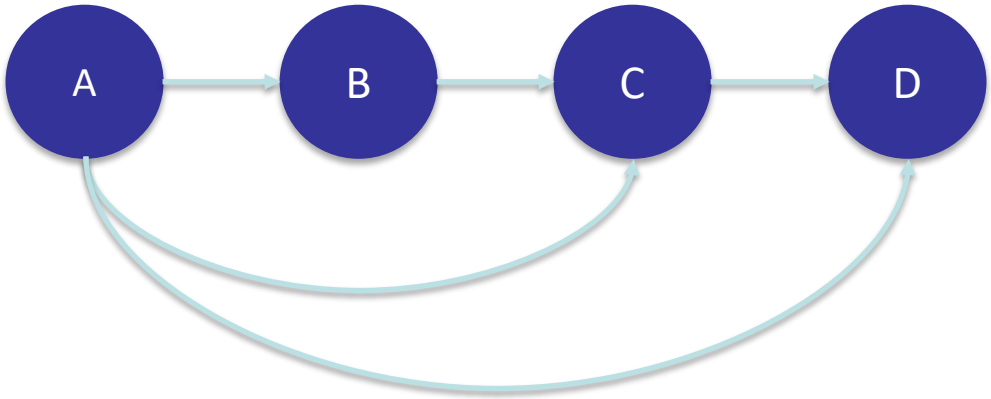


- Cluster sequences into OTU
  (Operational Taxonomic Unit)
- Sequences with a similarity score higher than 97% belong to the same OTU
- MUMmer or NUCmer (NUCleotide MUMmer) Algortihm is used by dotnetbio
- Current disadvantage of our pipeline: many common algorithms like SortMeRNA, Mothur, TRIE, UCLUST, USEARCH, BLAST, USEARCH61, SUMACLUST, SWARM, CD-Hit can not be used

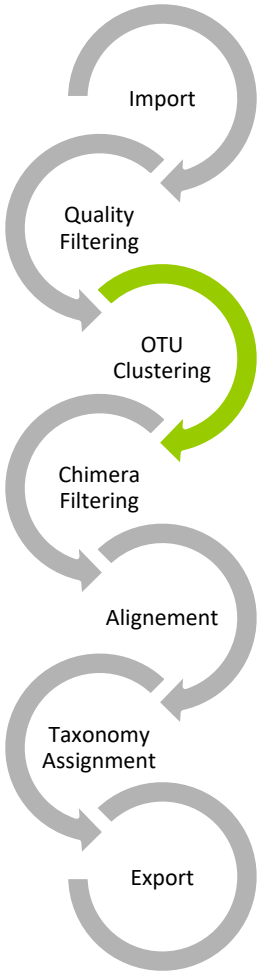
Microsoft Azure “Pipeline"



- Building OTU with own algorithm
- Implementing algorithm like Levenshtein and/or Hirschberg in C#
- Building the transitive closure for matched sequences

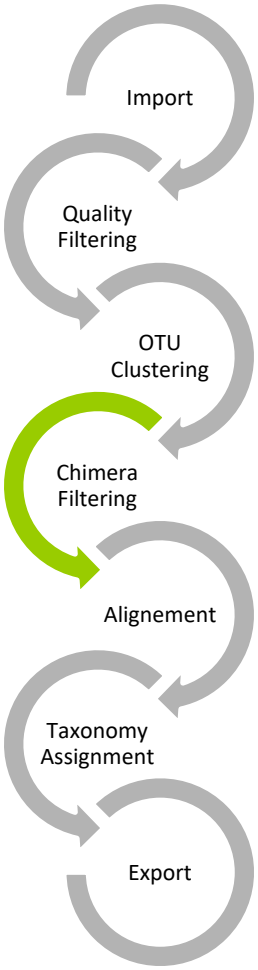


Microsoft Azure “Pipeline"



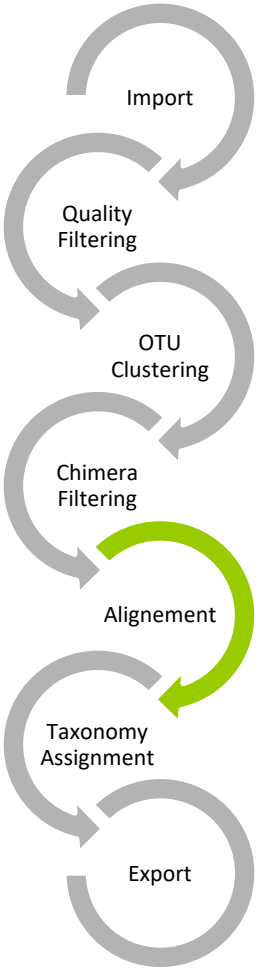
- Currently U-SQL cannot directly process recursive queries such as with CTE
- To achieve similar results, a Table-Valued Function can be created
  - Disadvantage: It's no real recursion, the iteration is fixed
- A further possibility is the development of a recursive reducer

Microsoft Azure “Pipeline"



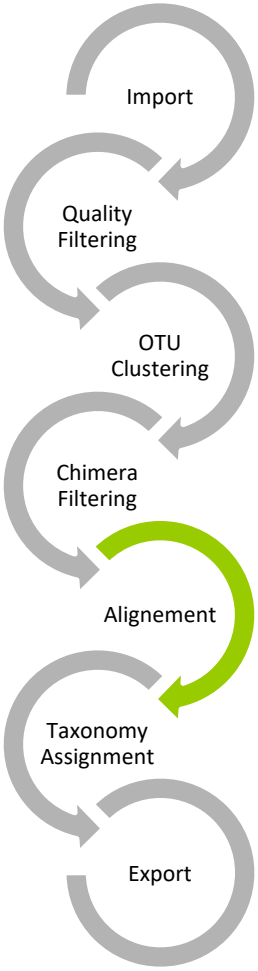
- Chimeras are a product of the unwanted combination of two or more sequences
- dotnetbio offers a comprehensive range of functions for the Chimera detection

Microsoft Azure “Pipeline"



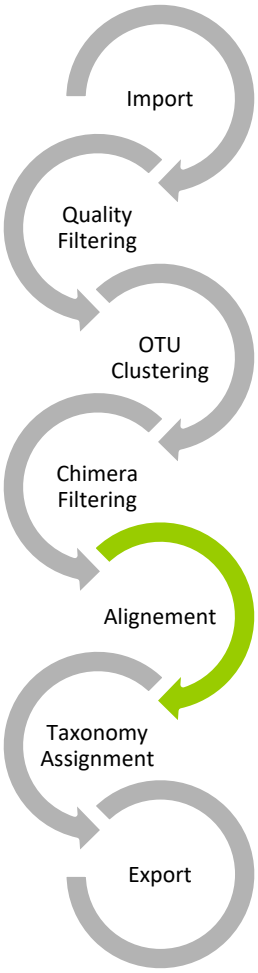
- Alignment or mapping is the process of searching data in reference databases like BLAST, SILVA or GreenGenes
- SILVA is currently the most comprehensive RNA database
  - 4 million 16S/18S
  - 400.000 23S/28S

Microsoft Azure “Pipeline"



- From the perspective of a (relational) database developer, the system is very strange
- Data is stored in text files
  - Sequences are stored in ARB files
  - Taxonomy data is stored in CSV files
  - Keys between both tables look like *AB002522.1.1416*
- To search within the database the Needleman-Wunsch algorithm is used

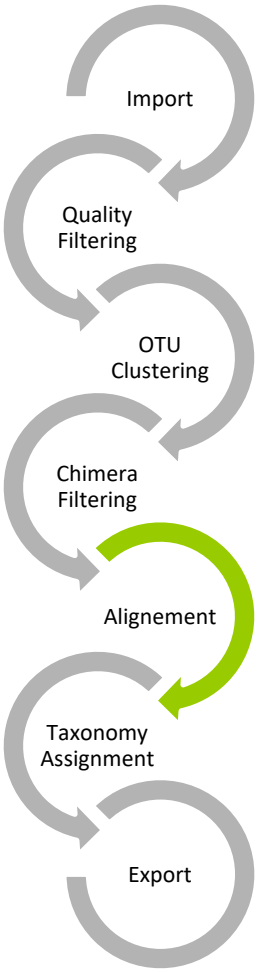
Microsoft Azure “Pipeline"



- Needleman-Wunsch algorithm is used to compare biological sequences
- the similarity between two sequences is calculated
  - ACTCCTTAA
  - A-TCC--AA

	-	A	C	T	C	C	T	T	A	A
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7
T	-2	0	0	1	0	-1	-2	-3	-4	-5
C	-3	-1	1	0	2	1	0	-1	-2	-3
C	-4	-2	0	0	1	3	2	1	0	-1
A	-5	-3	-1	-1	0	2	2	1	2	1
A	-6	-4	-2	-2	-1	1	1	1	2	3

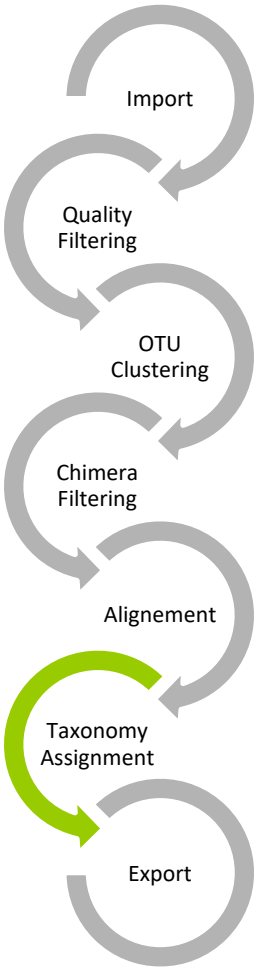
Microsoft Azure “Pipeline"



DEMO

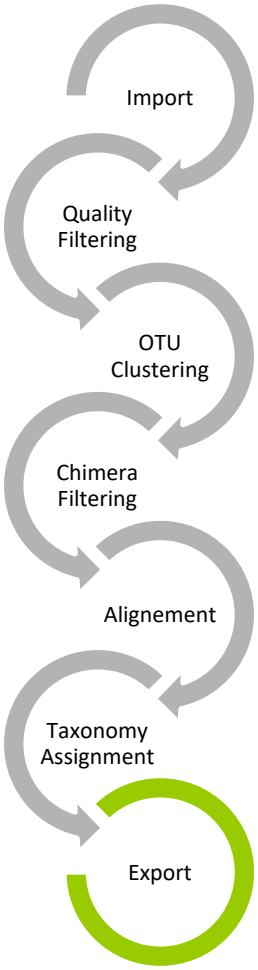
```
public static int AlignmentScore(string sequenceA, string sequenceB,
                                 int gapPenalty, int matchScore, int mismatchScore)
{
            var align = Align(sequenceA, sequenceB, gapPenalty,
                              matchScore, mismatchScore)[1];
            var c = align.Count(x => x == '-');
            return Convert.ToInt32(100 / align.Length * c);
}
```

Microsoft Azure “Pipeline"



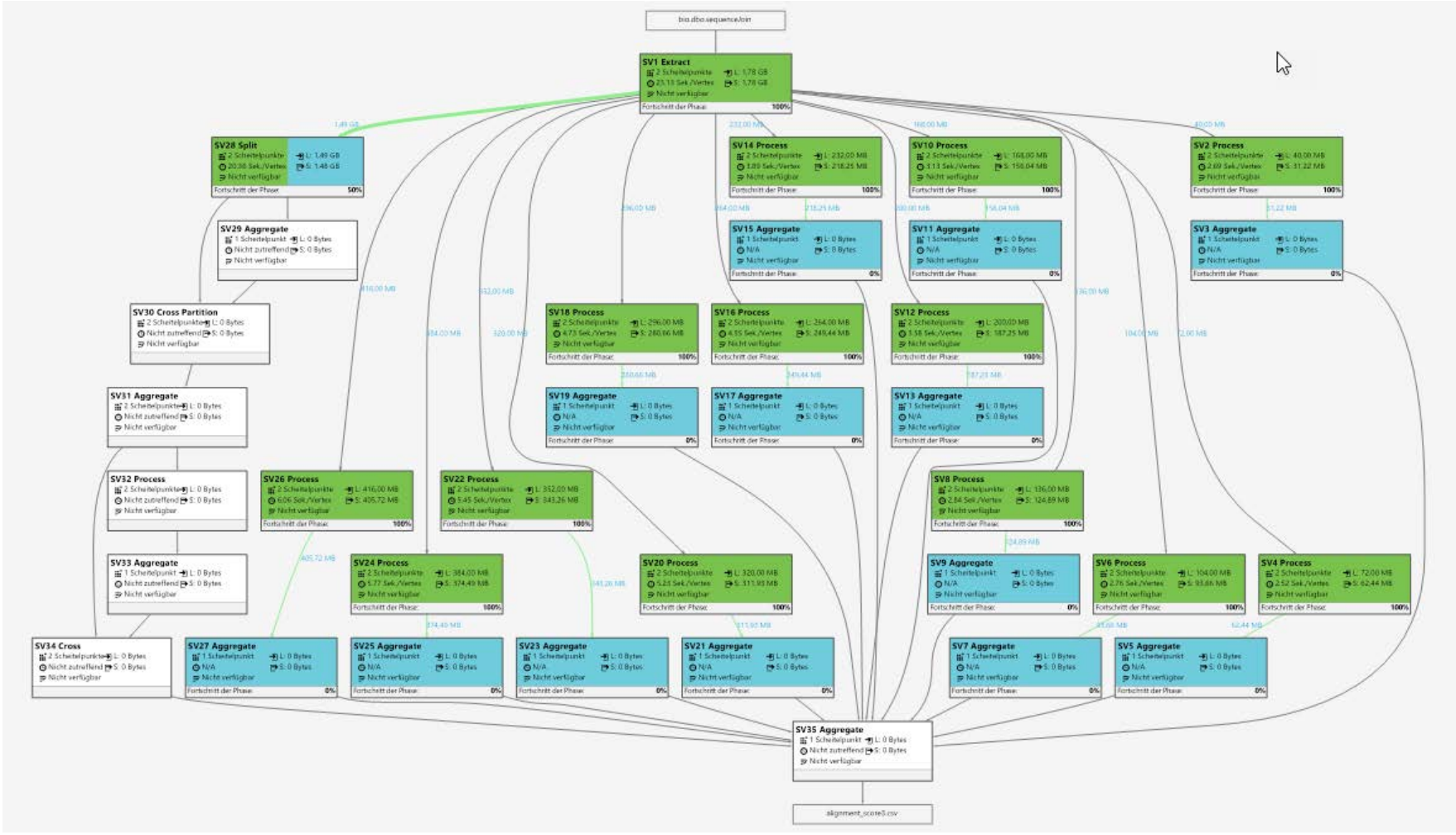
- Taxonomy Assignment is usually one step in the described workflow
- Size doesn‘t really matter – simply join the data
- With U-SQL we already joined the data together in the alignment process

Microsoft Azure “Pipeline"

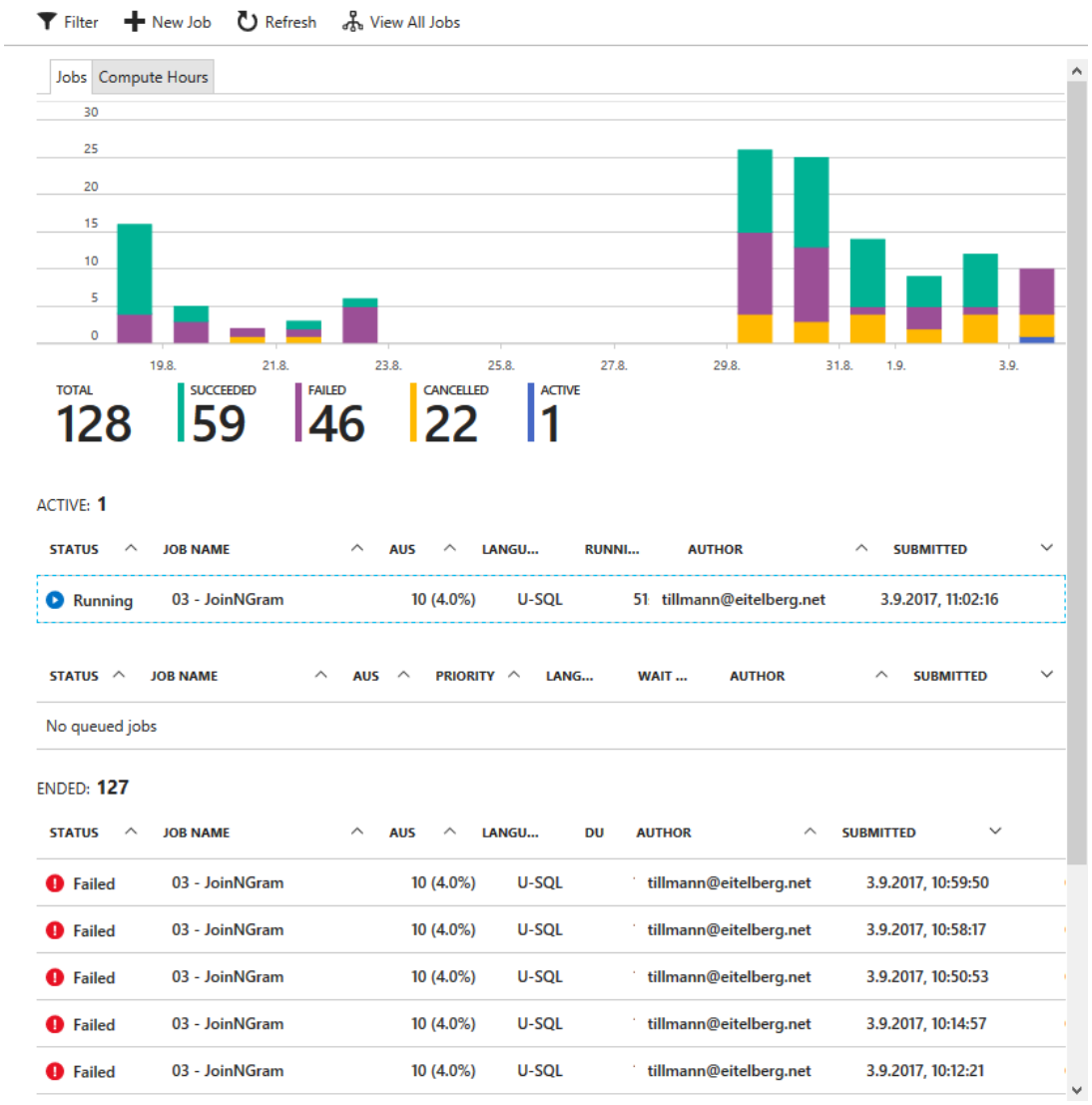


- Data can be exported to an BlobStorage or used directly from Azure Data Lake
- After exported the data, it can easily consumed by applications like Power BI
- Microsoft Power BI is also supported as a target
- Use Polybase to consume data with SQL Server 2016

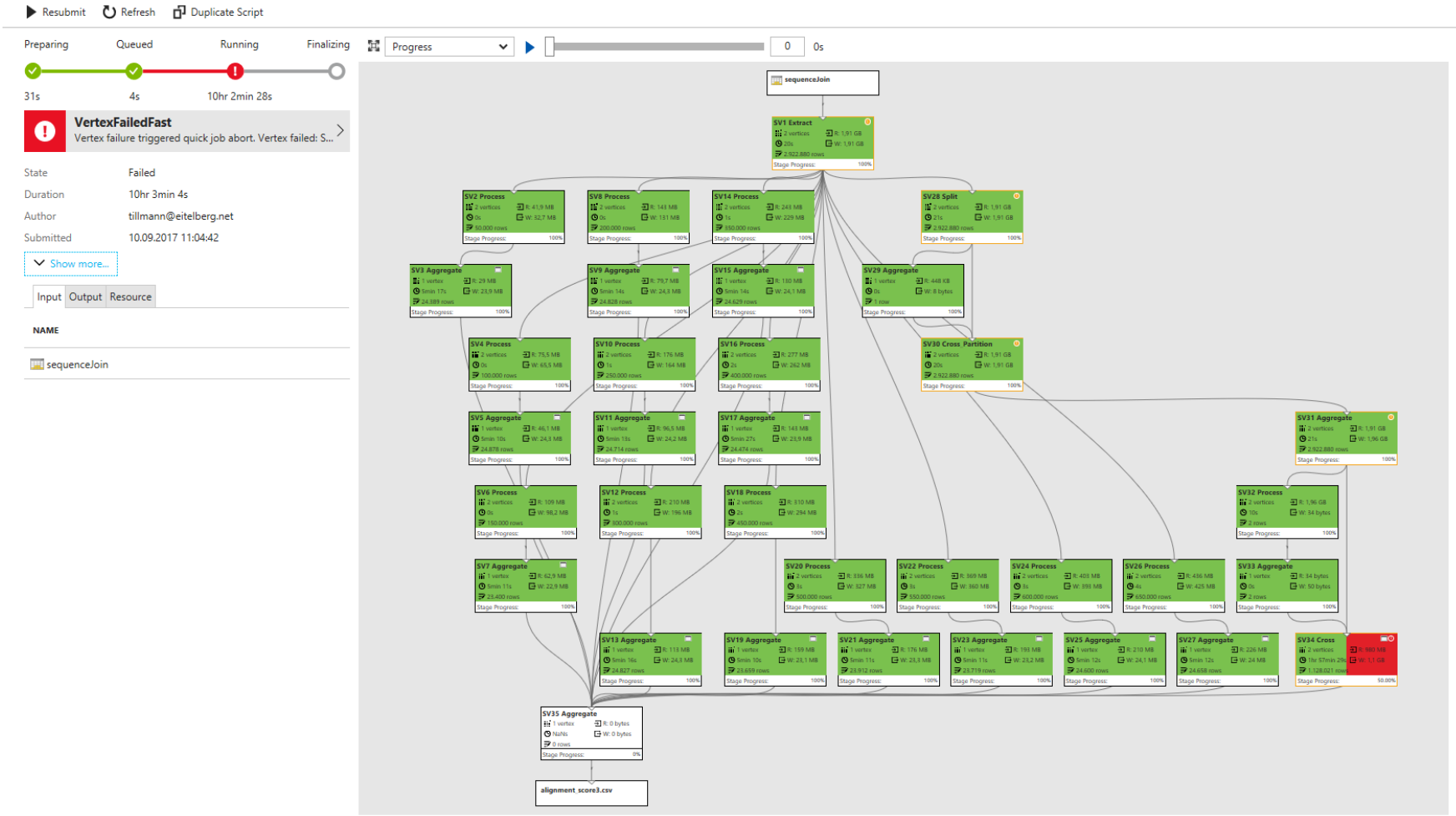
Microsoft Azure “Pipeline"



Microsoft Azure “Pipeline"



Microsoft Azure “Pipeline"



Microsoft Azure “Pipeline"

- Visualize data with Microsoft Power BI
- It’s possible to access data directly from ADLS
- Create Reports and Dashboards
- Ask questions about your data

Microsoft Azure “Pipeline"

DEMO

Power BI

Further Use Cases