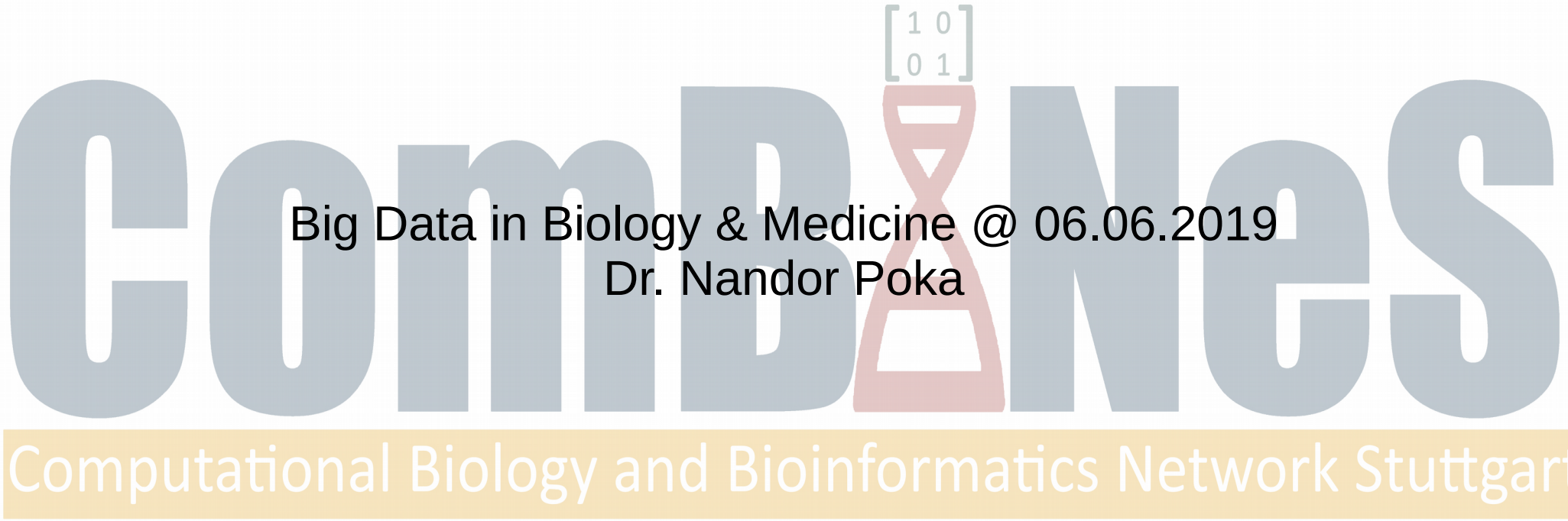
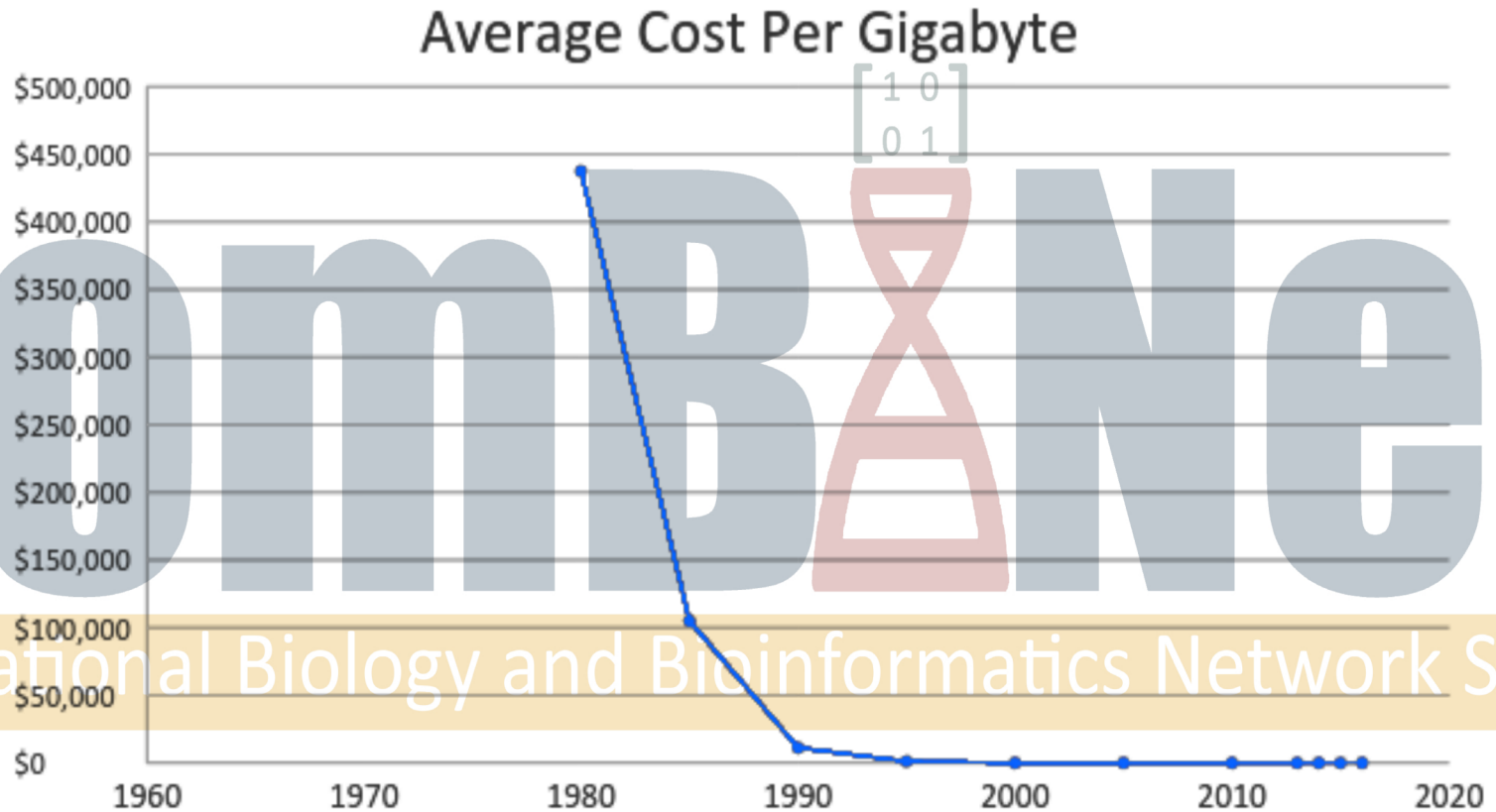
ComBINeS

Computational Biology and
Bioinformatics Network Stuttgart

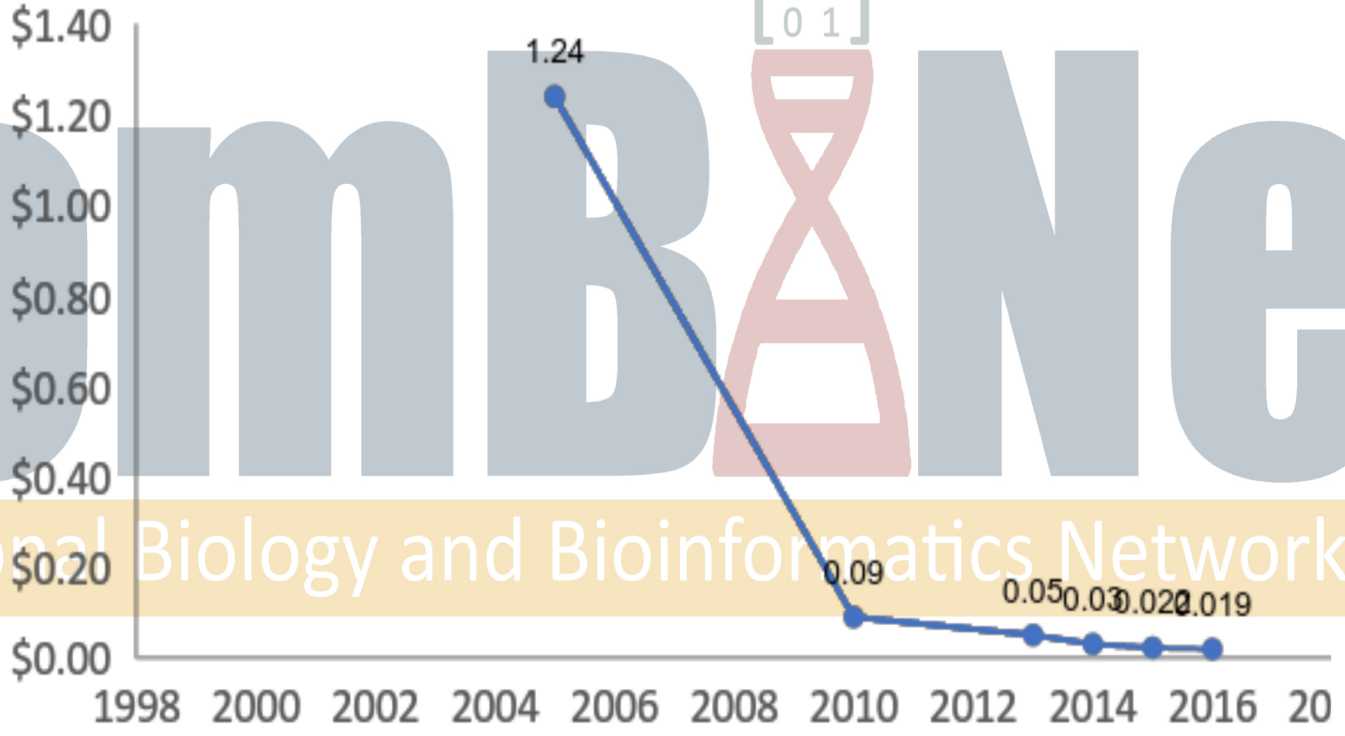


Big Data in Biology & Medicine @ 06.06.2019
Dr. Nandor Poka

Big Data

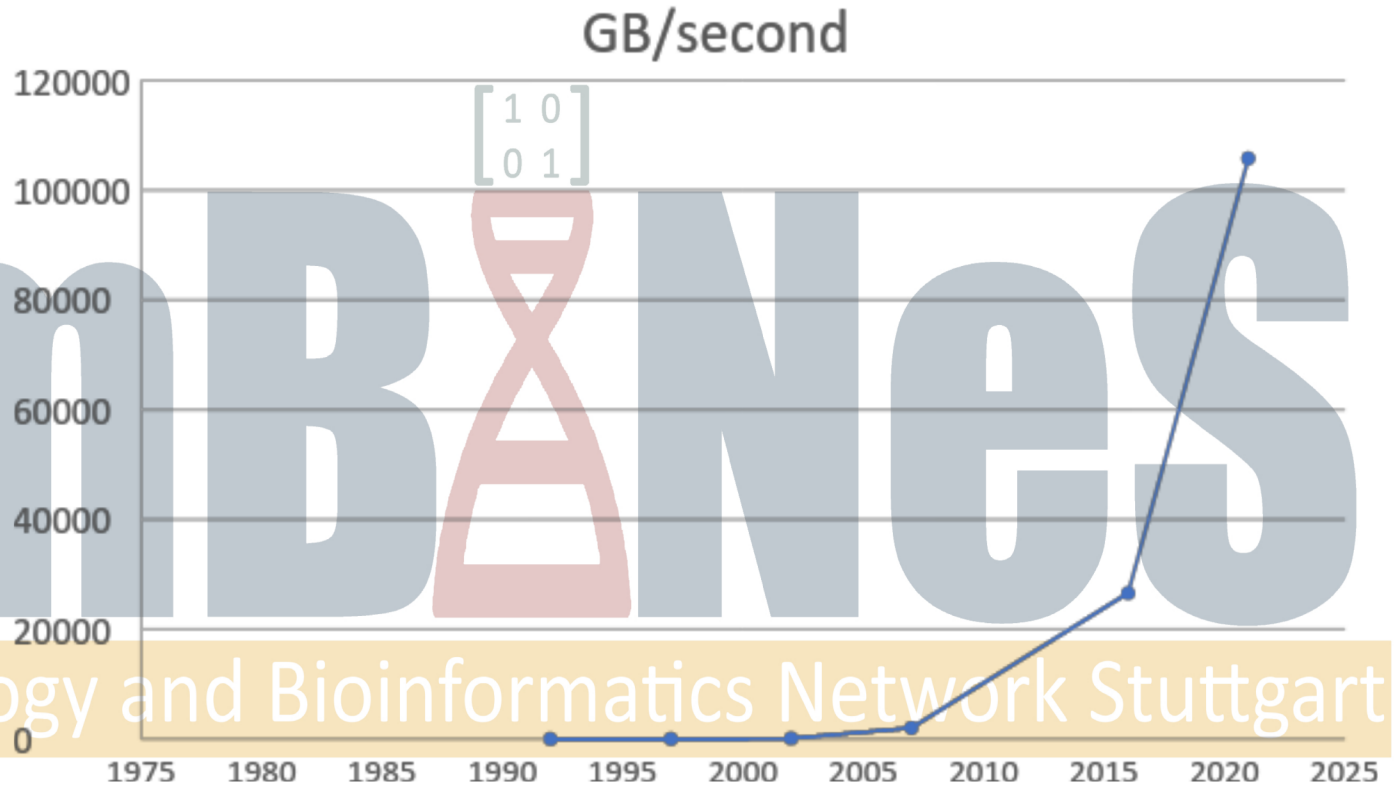


Data source: http://www.statisticbrain.com/average-cost-of-hard-drive-storage/



Data source: http://www.statisticbrain.com/average-cost-of-hard-drive-storage/

Year	Global Internet Traffic
1992	100 GB per day
1997	100 GB per hour
2002	100 GB per second
2007	2,000 GB per second
2016	26,600 GB per second
2021	105,800 GB per second



But what is Big Data?

- Like old data, just a lot more… too much to deal with traditional methods
- System logs of various kinds
- Financial transactions
- Online purchases
- Etc.

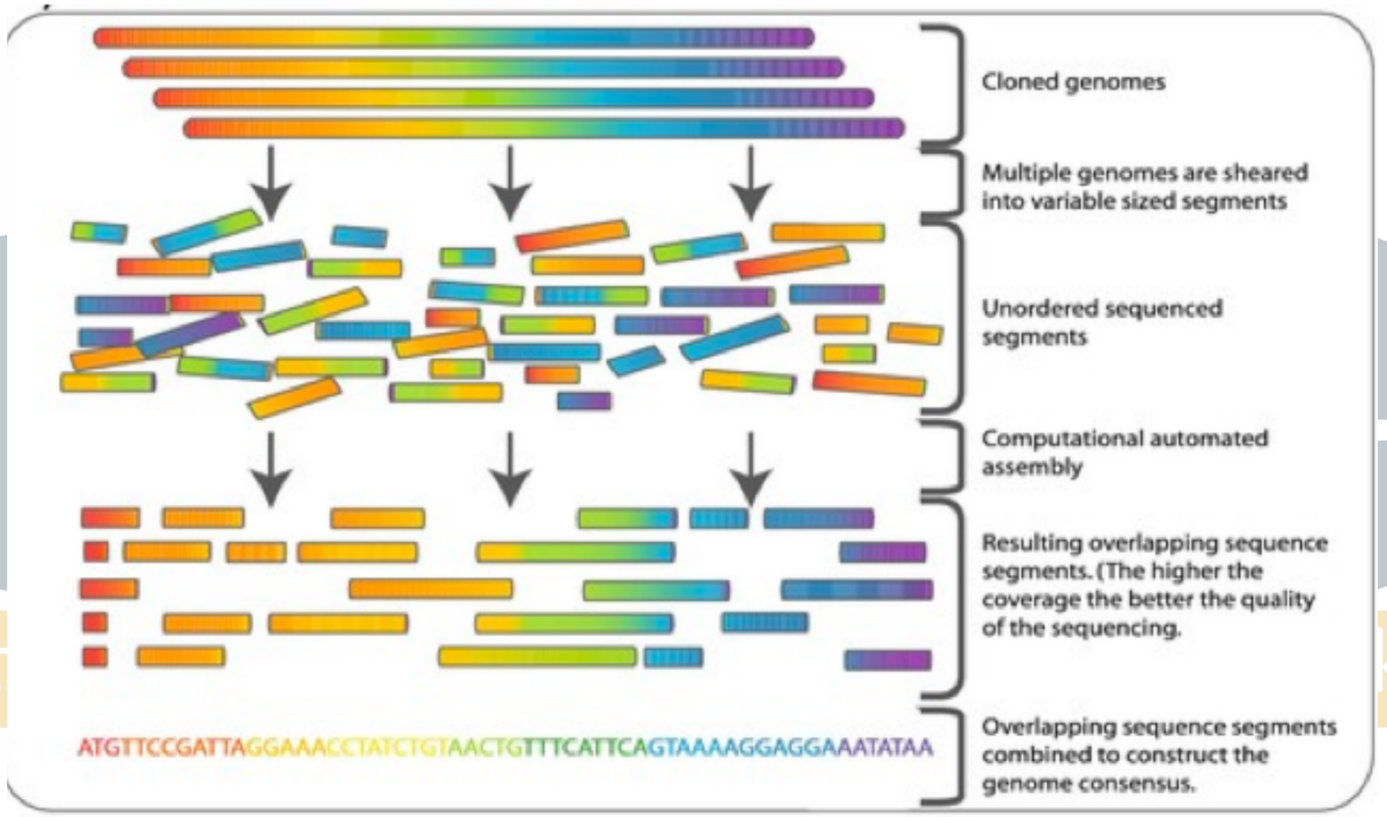
“New age” Big Data

- Social media – likes, posts, contacts
- User location data – navigation searches, phone location data
- Community knowledge – recommendations, product ratings
- Personal (non-professional) health data – fitness trackers

What sources of data we have in Biology and Medicine?

- Sequencing data
- Genomics / proteomics (and other “omics”) data
- Gene / protein interaction networks
- Medical images
- Electronic personal medical data
- Wearable fitness data

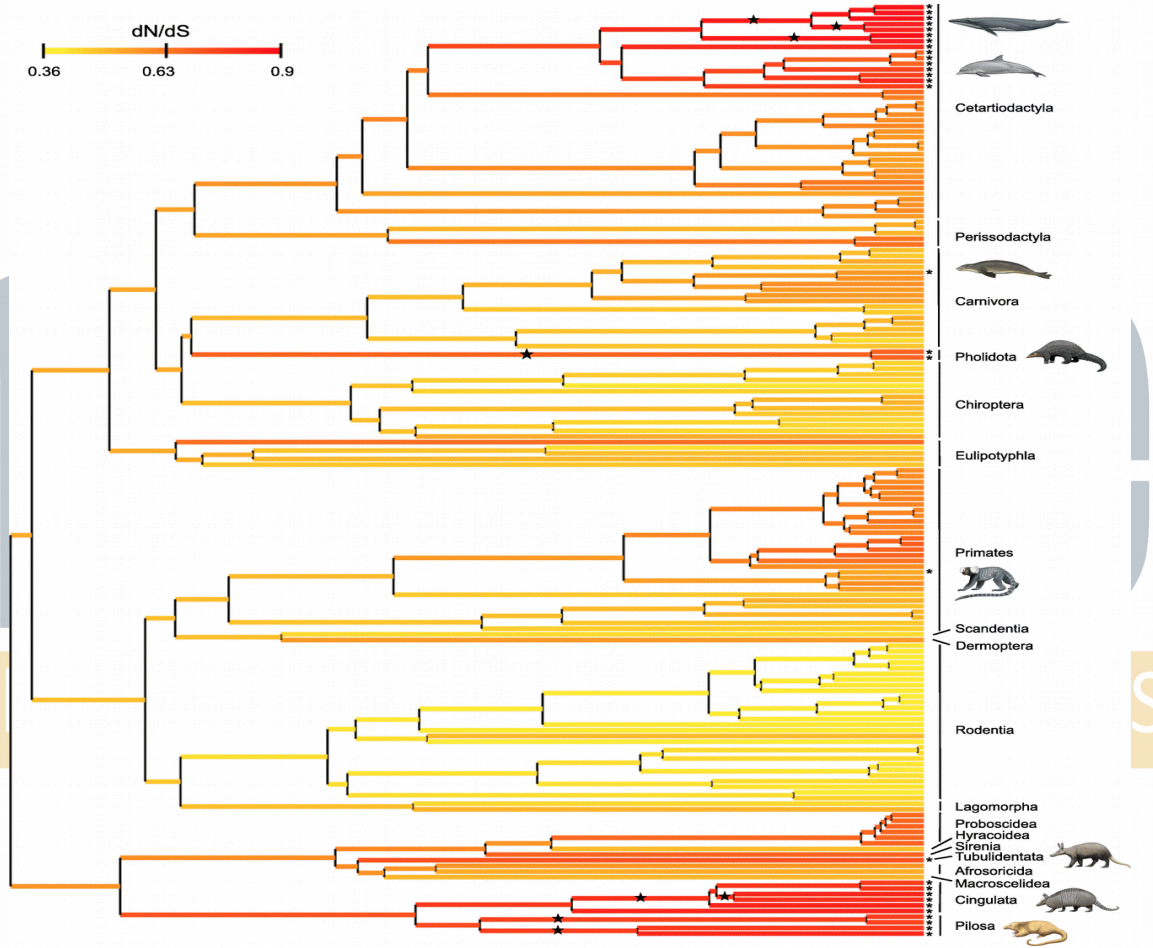
Sequencing



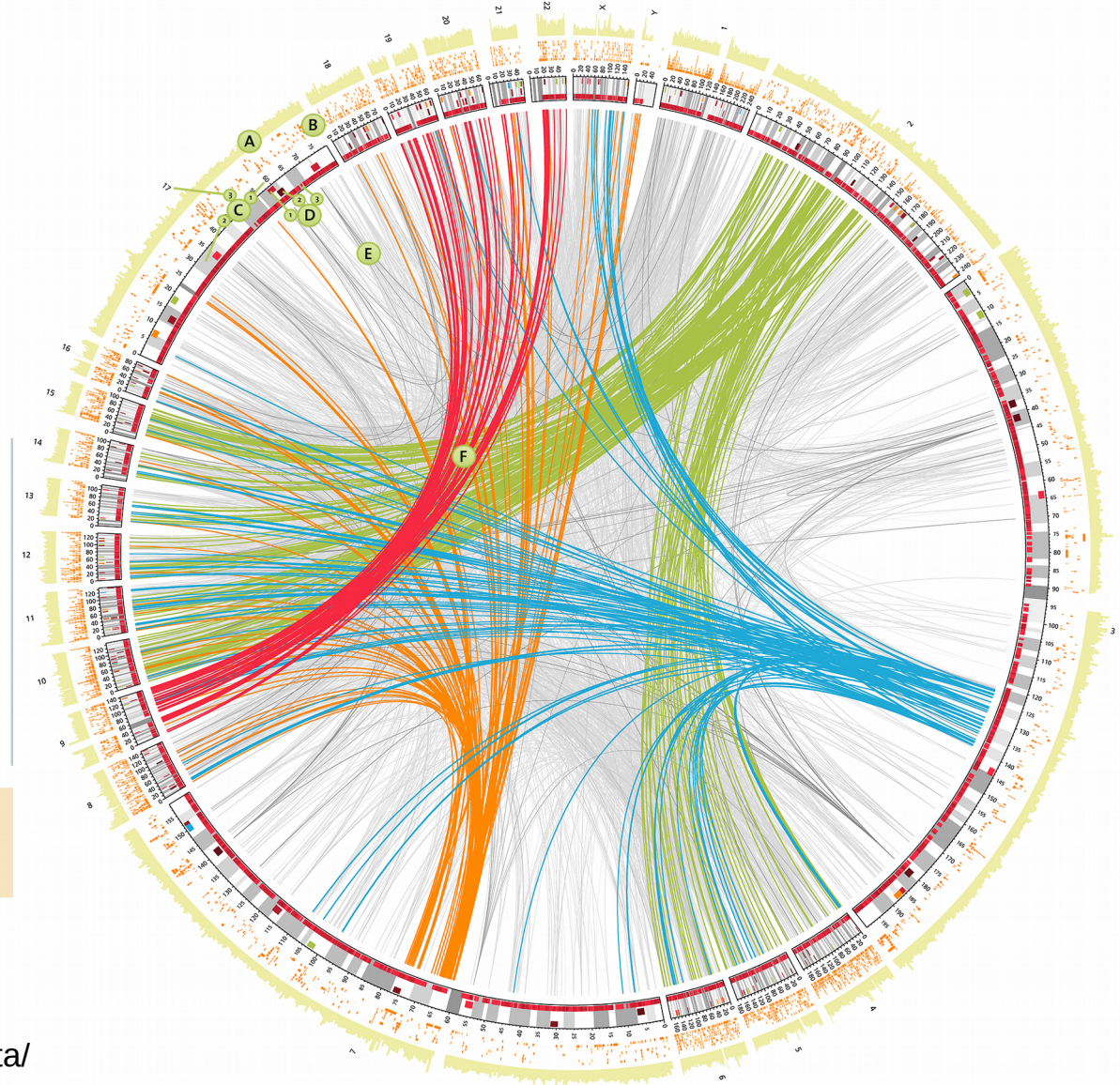
Source:

Computational Biology Methods and Their Application to the Comparative Genomics of Endocellular Symbiotic Bacteria of Insects. Biol. Procedures Online (2009). Commins, J., Toft, C., Fares, M. A.

Genomics and other “omics”



Source:
https://doi.org/10.1186/s12862-019-1359-6

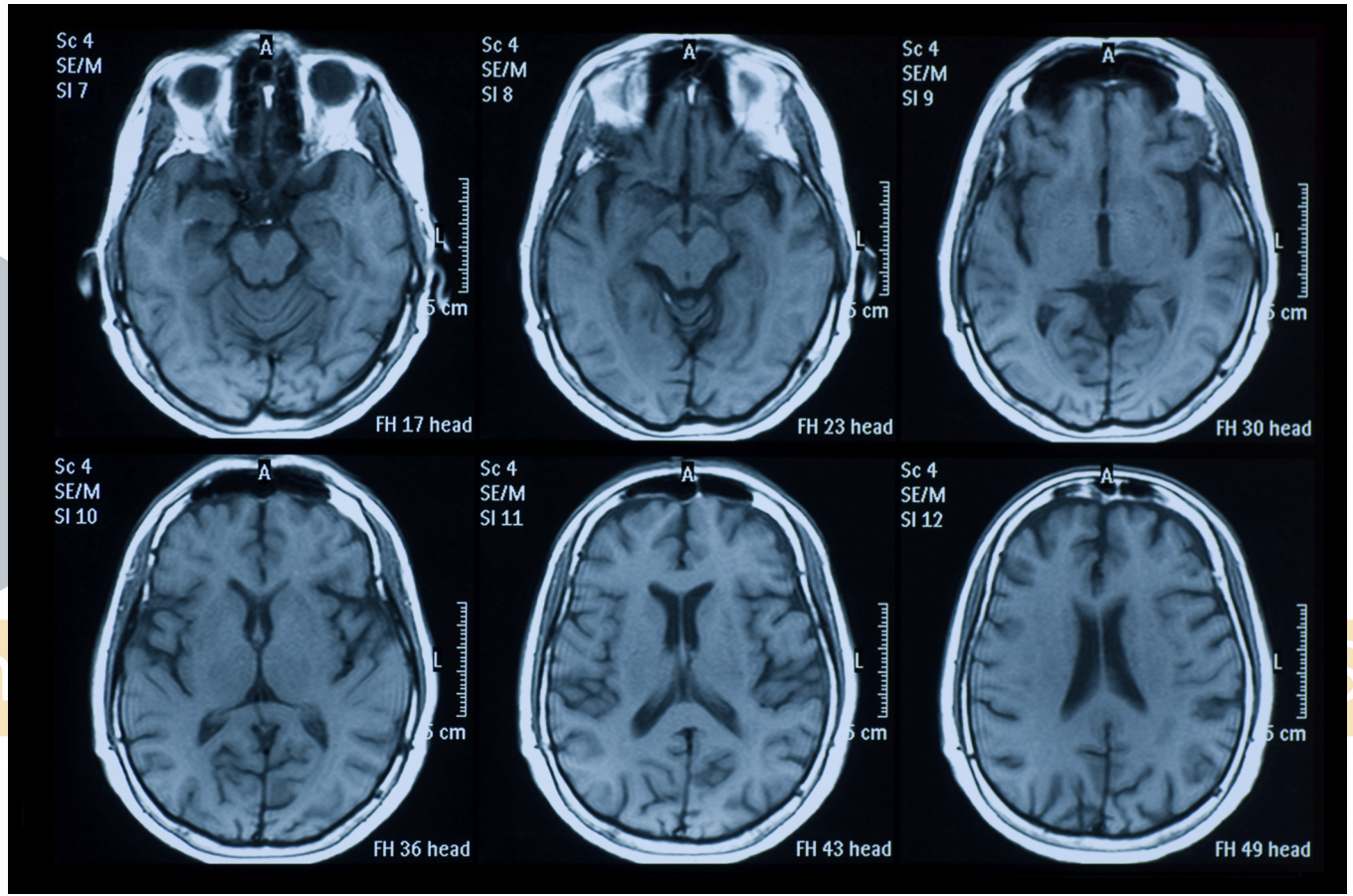


Computational

k Stuttgart

http://circos.ca/intro/genomic_data/

Medical imaging

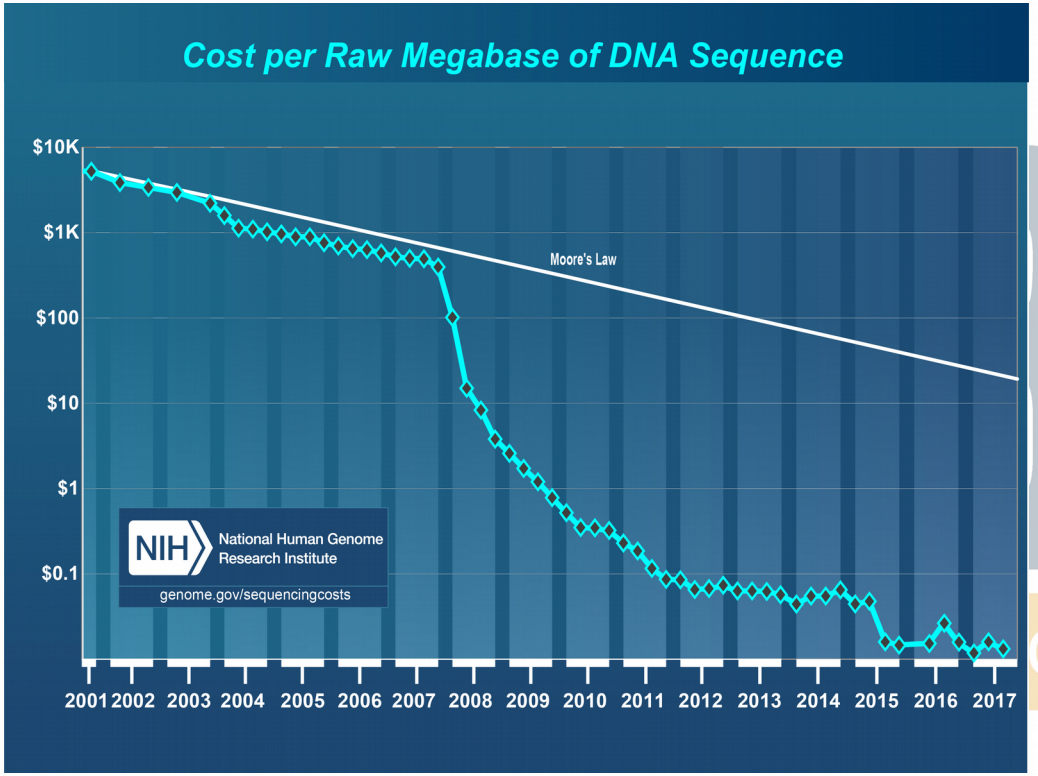


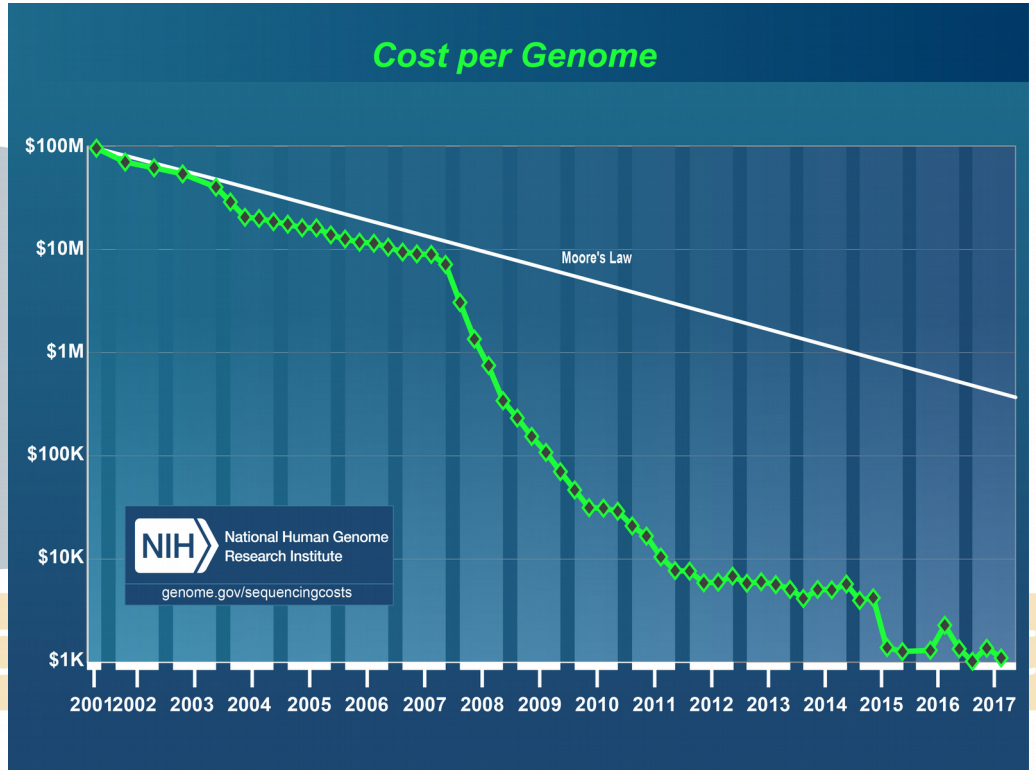
http://time.com/2860630/mri-scans-can-detect-early-onset-of-parkinsons-study-finds/

What type of data?

- Sequences → plain text, often compressed
- “Omics” data → heterogenous, but typically includes sequence data (often with added metadata / markup), tables / databases, graphs, 3D molecule models
- Medical images → high-res often multi-layered images / videos
- Medical records → can be anything
- Personal fitness data → varies, but somewhat unified → eg. Google Fit data format / types

How is this Big Data?





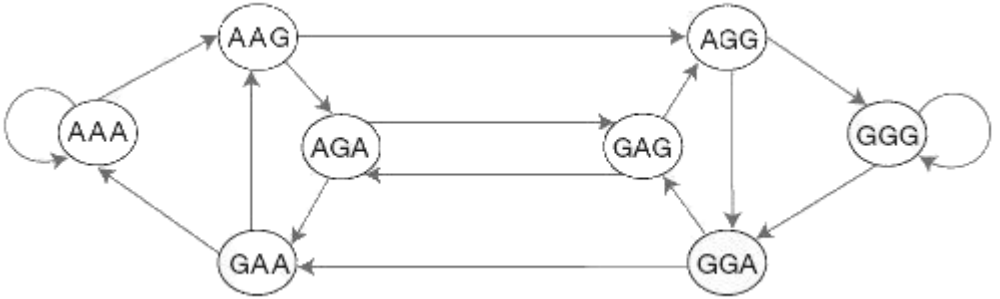
https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

- The human diploid genome is approx 6.6Gb (GigaBase) → in plain ASCII it's 6.6GB (GigaByte)
- Typical sequencing is done with 100 base long reads and 40X coverage → ~2.64*10^9 reads → 264GB raw sequence
- ~ 2x more with quality data → 0,5TB+
- Some organisms have genome size of 100-200Gb+ genome sizes


```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCA
+
!''*((((***+))%%%++)(%%%%).1
```

For assembly

- de Bruijn graph with 66 Mill. Vertices
- Min. same amount of directed edges
-



“Omics” data

- Sample from an gene expression experiment data
- 47191 rows by 405 colums = ~19,1M data points

ID_REF	1NL.AVG_Signal	1NL.Detection Pval	1TU.AVG_Signal	1TU.Detection Pval	2NL.AVG_Signal
ILMN_1762337	61.1	0.24156	59.8	0.31558	64.3
ILMN_2055271	63.3	0.16364	68.8	0.07662	67.4
ILMN_1736007	54.7	0.51818	61.6	0.24545	55.5

Dataset published: Wozniak MB, Le Calvez-Kelm F, Abedi-Ardekani B, Byrnes G et al. Integrative genome-wide gene expression profiling of clear cell renal cell carcinoma in Czech Republic and in the United States. PLoS One 2013;8(3):e57886. PMID: 23526956 Direct download: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40435

Dataset further used in this paper to make devise interaction network https://doi.org/10.1016/j.gdata.2017.10.006

Medical data

- MRI scans → ~3Mpx per image → ~1.7MB as lossless PNG
- 20 images per patient → 24MB / Scan
- In Germany, in 2015, 136.2 MRI scans / 1.000 resident
- (82.175M / 1000) * 136.2 = ~11.2M Scans → ~1,14PB (Petabyte) of data.

What can be achieved bio / medicine side?

- “full stack” organism analysis
- Rapid pathogen detection – in plants / animals as well
- Personalized medicine – Cas9 / CRISPR → “on-site” gene repairing
- Rapid drug development

Challenges

- Data privacy
- Cost – manual labour, ingredient and equipment cost
- Quantity vs. quality – data is abundant, but heterogeneous
- Visualization
- Speed(up)
- Free / commercial software

What can be done on IT side?

- Better, faster software / algorithms – cluster / parallel computing
- AI for applied cases – detecting micro fractures on X-rays, detecting tumors on MRI
- Better equipments, data acquirement
- Applied Data analysis / predictions:
  - Live predictions – epilepsy, heart attacks
  - Early onset detection of neurodegenerative diseases (eg. Parkinsons)

Get your share of the data!

- https://www.ebi.ac.uk/services
- https://www.uniprot.org/
- https://www.ncbi.nlm.nih.gov/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5506686/

Questions?