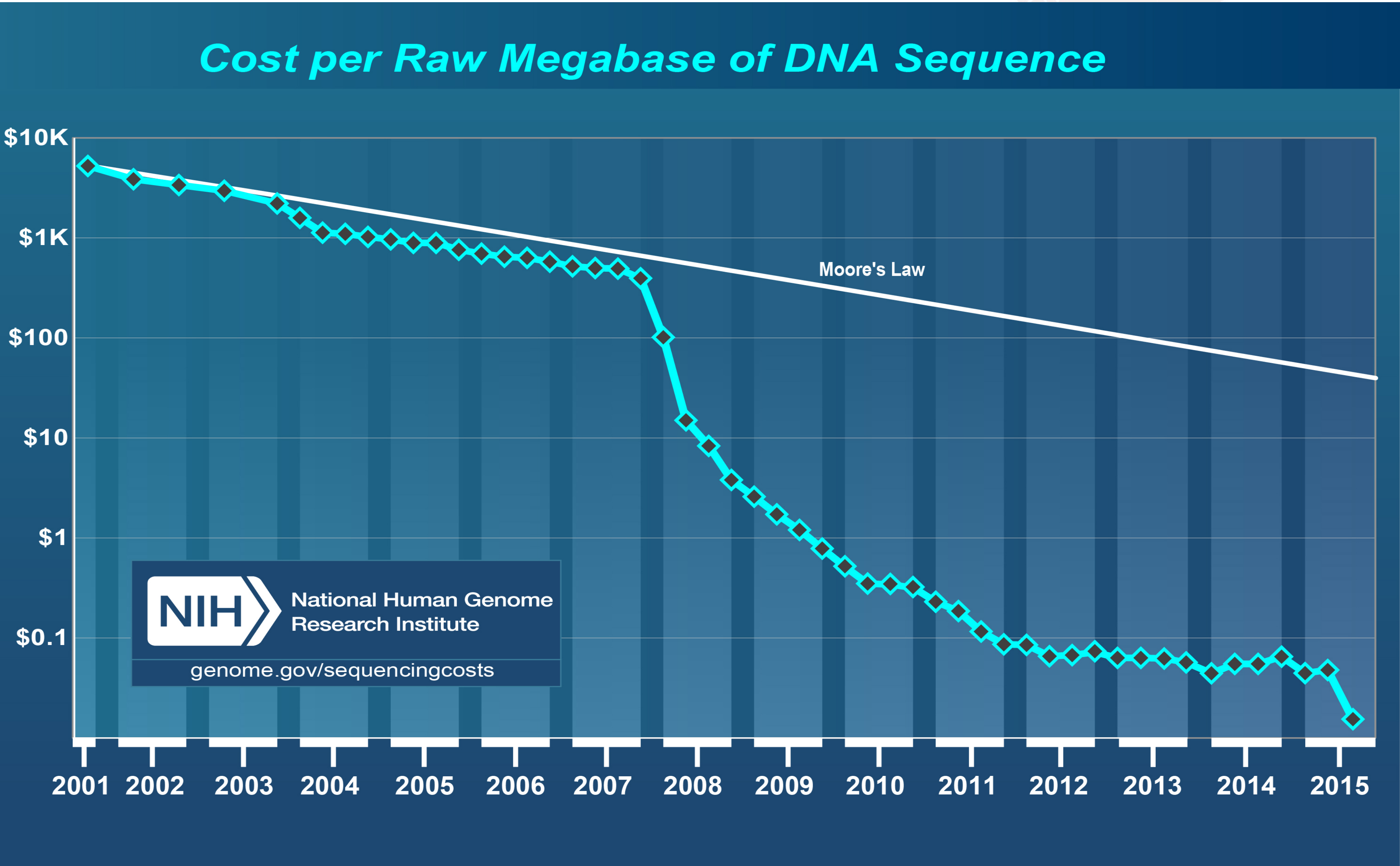
Biomedical Computing Center at KI

Population Scale human whole-genome sequencing

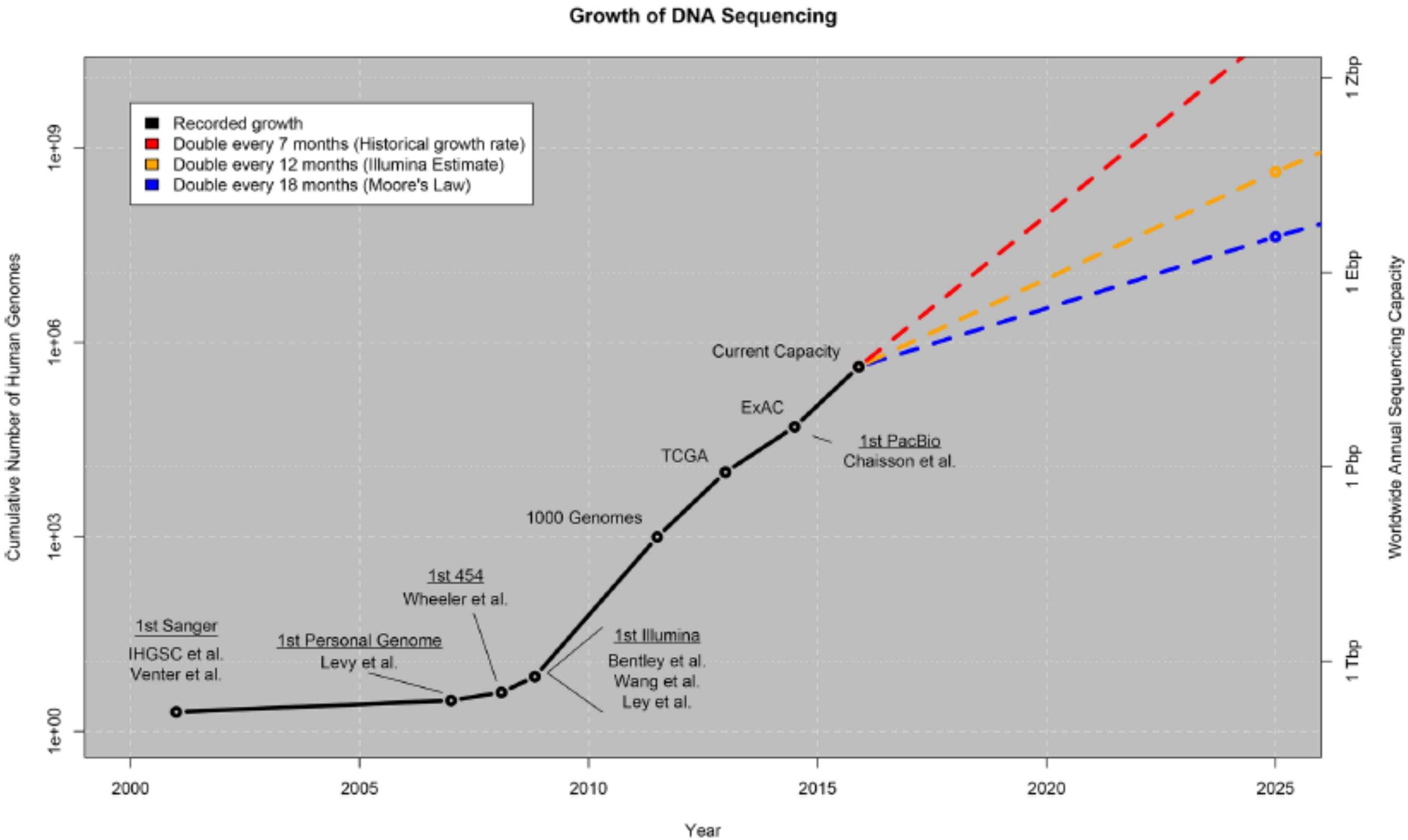
- HiSeq X Ten System
  - ≈18,000 humane genomes/year with 30X coverage (< $1000 per genome[*])
  - That means ≈3.6 PB of row data/ year[#]

[*]http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-hiseq-x-ten.pdf
[#]https://medium.com/precision-medicine/how-big-is-the-human-genome-e90caa3409b0

Sequencing capacity is out scaling Moore's Law.



“Big Data: Astronomical or Genomical?”



PLoS Biol. 2015 Jul 7;13(7):e1002195

Big Biomedical Data

- **Volume**
  - How ‘big’ exactly the data has to be to qualify as ‘big data’ is not specified. It is estimated that 90 per cent of the data in the world today has been created in the last two years,.

- **Velocity:** Every millisecond counts
  - Data need handling based on speed;
  - The speed of decision making – the time taken from data input to decision output – is a critical factor in the big data discussion. Emerging technologies are capable of processing vast volumes of data in real or near real time

- **Variety**
  - Heterogeneous Data - comes in many forms
  - Source data can be diverse, and it may require time and effort to shape it into a form fit for processing and analysis. The capacity of a system to analyze a variety of source data is crucial as it can yield insights not achievable by consulting one type of data in isolation. Too diverse in structure that no single data model captures all elements of the data

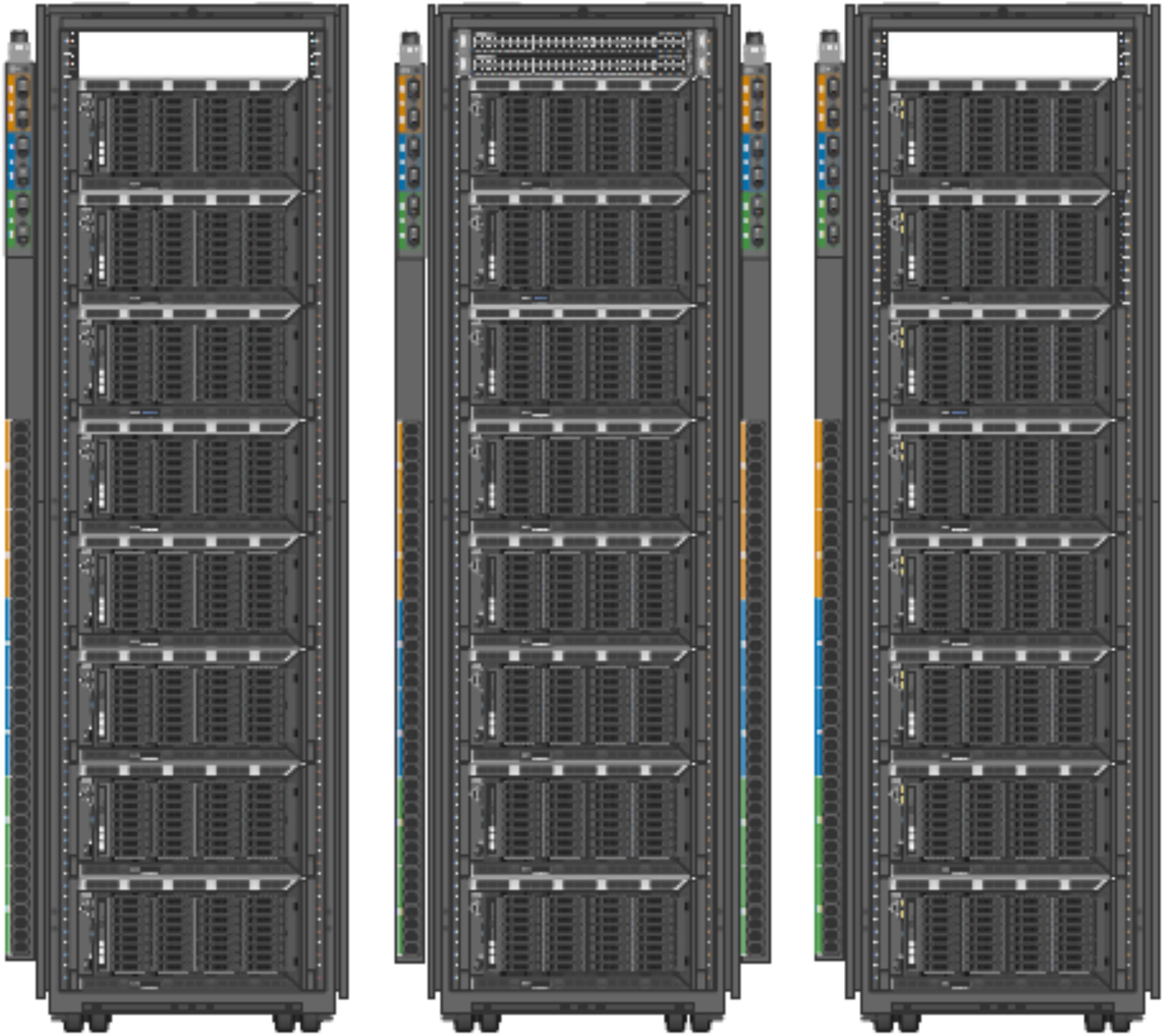
The Big Data Problem

- Data growing faster than computation speeds
- Data centers face bottlenecks:
  - Stalling CPU speeds
  - Storage capacity
- Solution is to distribute data over cluster of machines
- **Data Center is based on Hardware and Software stack that provides possibility to scale up to several thousands of machines.**

Data Center - Software Stack

- Data Center is based on **Apache Hadoop** and **Apache Spark computing framework**.
- Apache Hadoop and Spark offer possibility to store and process vast amounts of diverse data in parallel using inexpensive and **open-source approaches**.
- **The scalable and flexible nature of Hadoop-based systems addresses issues of future expansion and can be easily adapted to any new addition from the biomedical field.**

NIASC Data Center - Hardware Stack



- 24 computer nodes
- 672 CPU cores,
- Storage - 2 Petabytes
- RAM - 6 Terabytes


LAS AMERICAS
BEACH SPIRIT

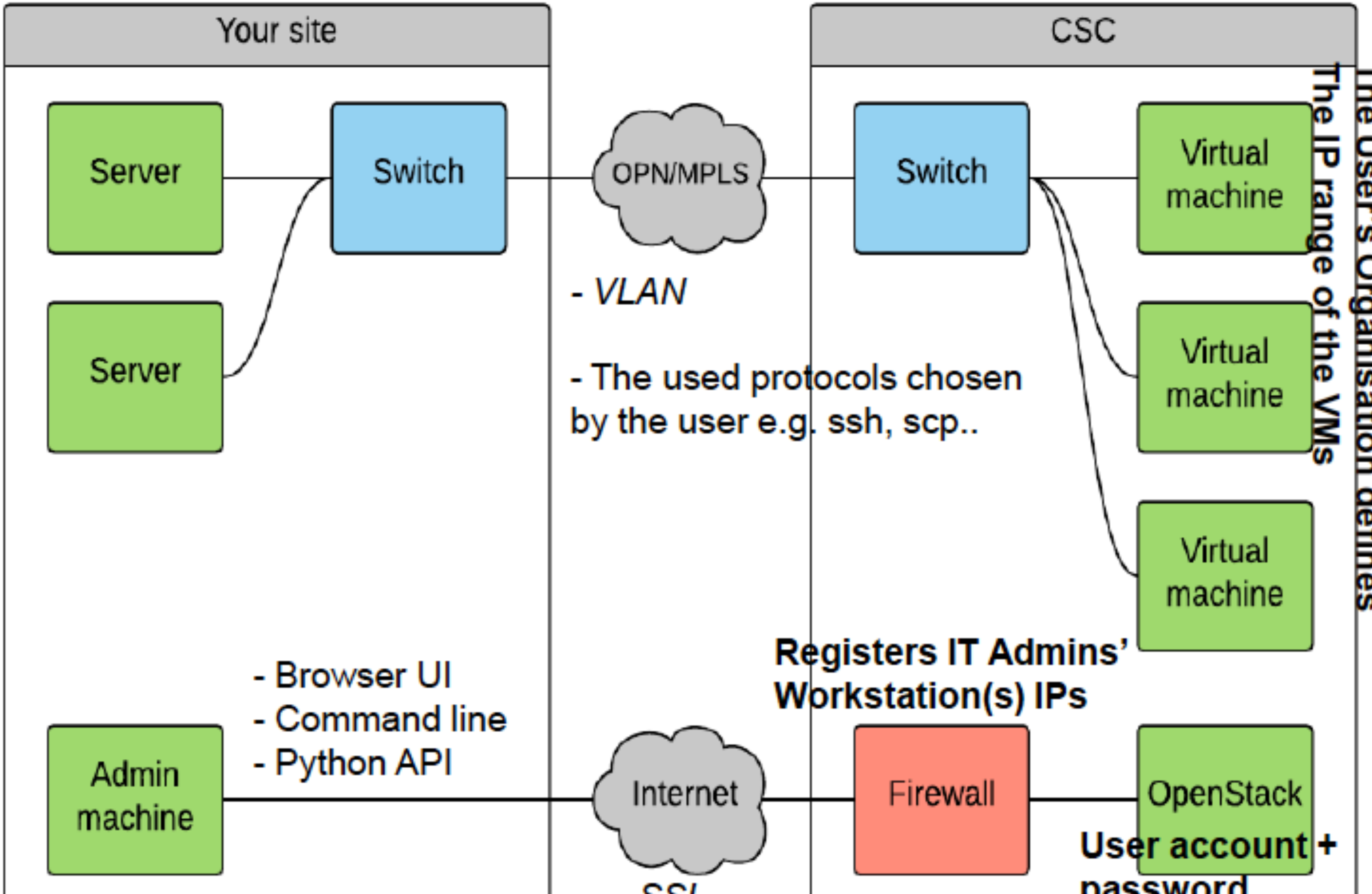
Hops

[Hadoop For Humans]

www.hops.io

@hopshadoop

Integration to ePouta cloud



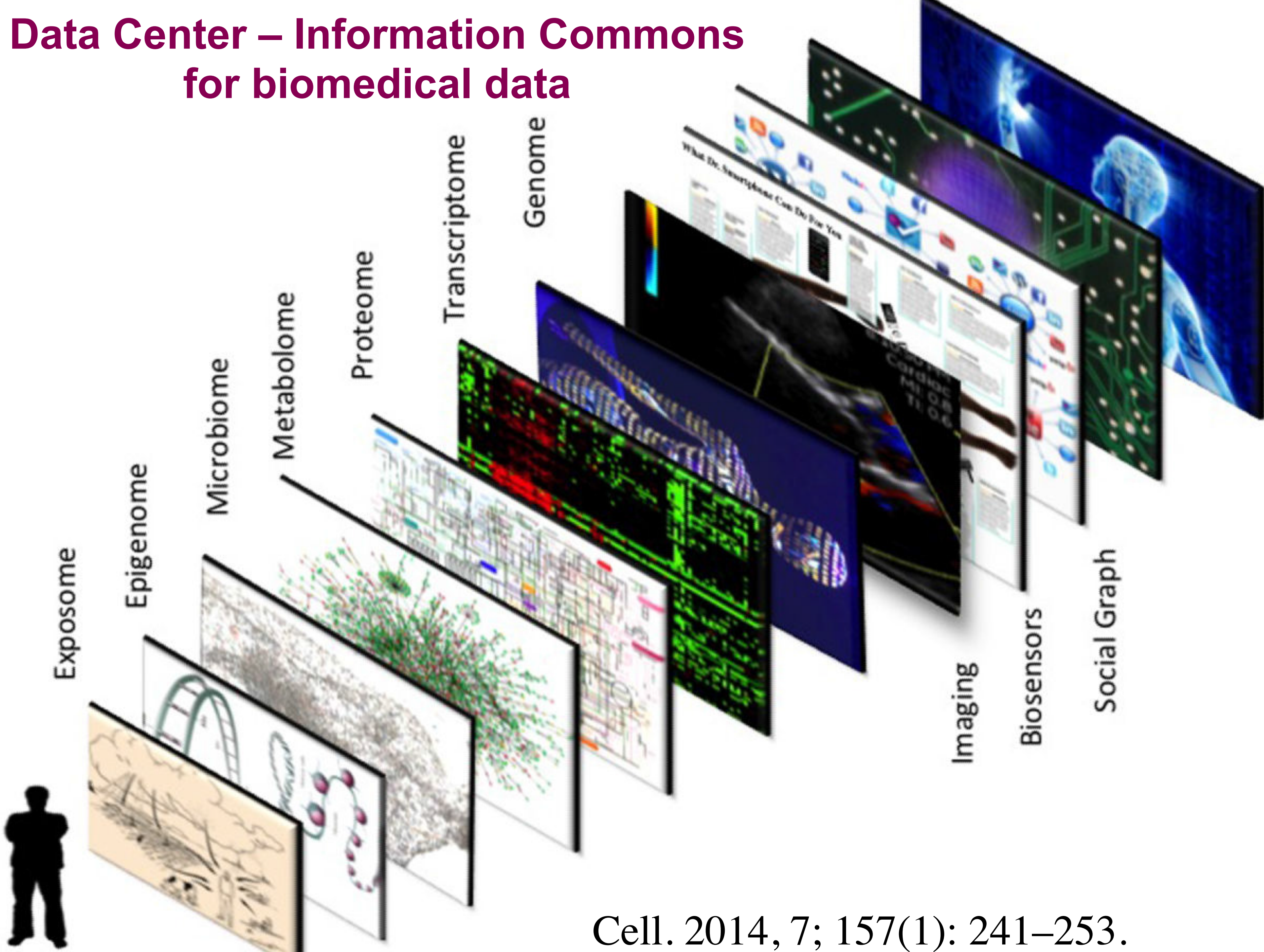
Development of Open Source software for parallel analysis of metagenomic data

- NIASC coordinating office and ICM Warsaw, are developing a bioinformatics pipeline to analyze viral metagenomics datasets, generated by NGS platforms.

- Workflow will give a possibility to utilize computing resources of the Polish Grid Infrastructure (http://www.plgrid.pl).

- Later, the workflow will also have a graphical user interface.
- NCBI BLASTn source code was modified to speed up its performance. The modified blastn algorithm is now 10 times faster on the PLGRID infrastructure.

Data Center – Information Commons for biomedical data



Cell. 2014, 7; 157(1): 241–253.