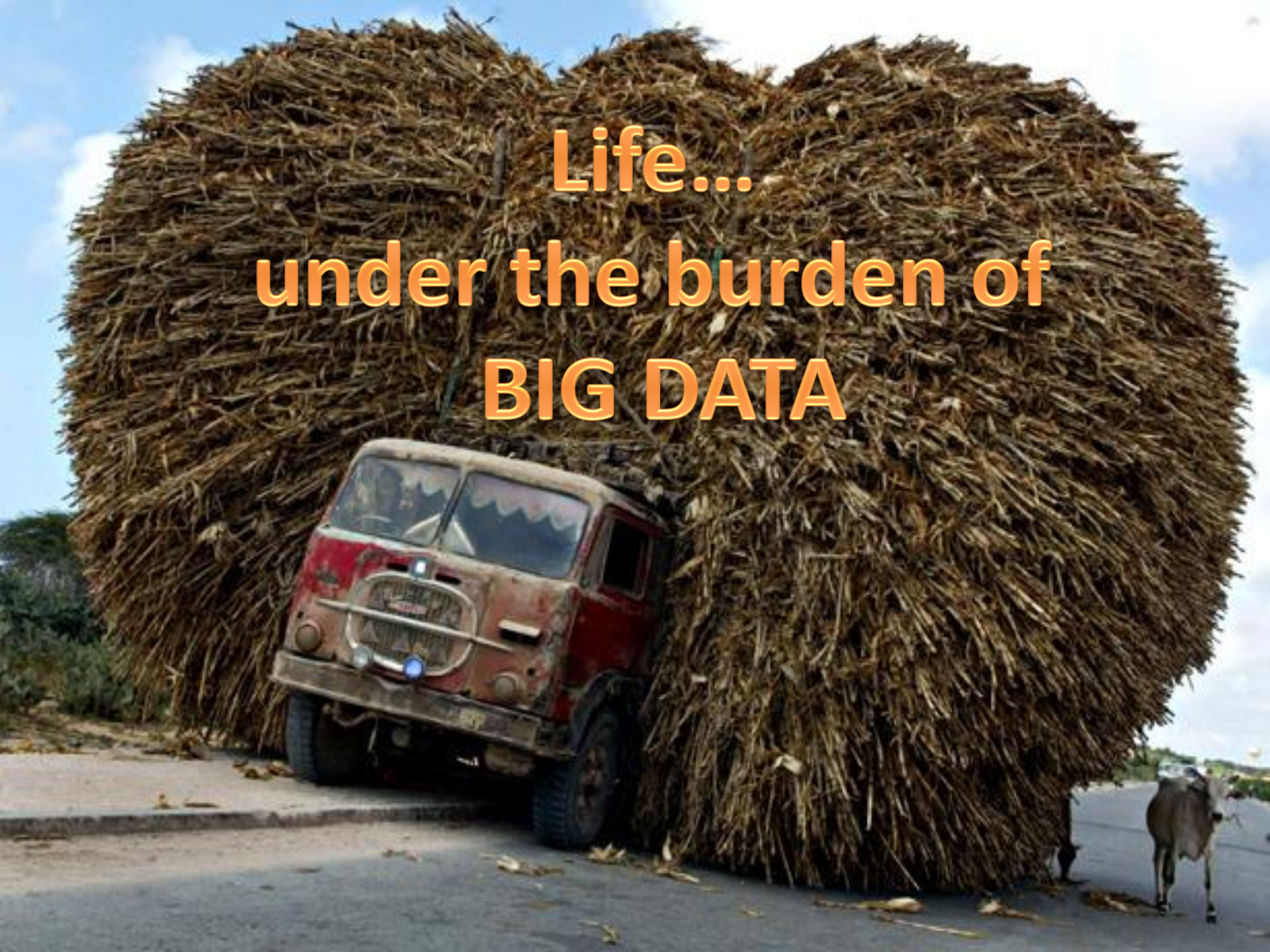
BT3041:
Analysis and Interpretation of Biological Data

Instructor: V. Srinivasa Chakravarthy

Slot: A

Room:

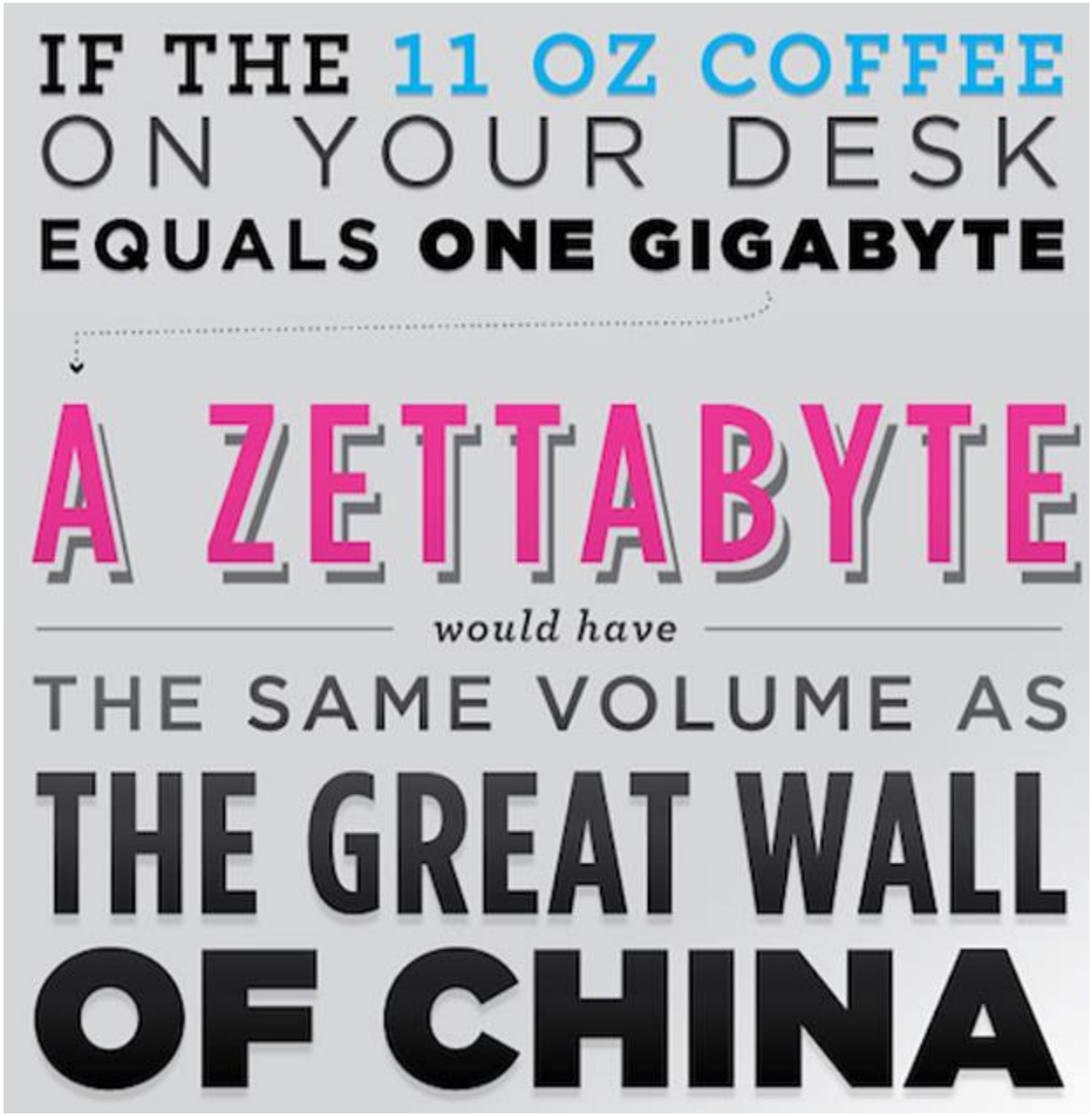

Life…
under the burden of
BIG DATA

Unidades de Medidas de Almacenamiento

Medida	Simbología	Equivalencia	Equivalente en Bytes
byte	b	8 bits	1 byte
kilobyte	Kb	1024 bytes	1 024 bytes
megabyte	MB	1024 KB	1 048 576 bytes
gigabyte	GB	1024 MB	1 073 741 824 bytes
terabyte	TB	1024 GB	1 099 511 627 776 bytes
Petabyte	PB	1024 TB	1 125 899 906 842 624 bytes
Exabyte	EB	1024 PB	1 152 921 504 606 846 976 bytes
Zetabyte	ZB	1024 EB	1 180 591 620 717 411 303 424 bytes
Yottabyte	YB	1024 ZB	1 208 925 819 614 629 174 706 176 bytes
Brontobyte	BB	1024 YB	1 237 940 039 285 380 274 899 124 224 bytes
Geopbyte	GB	1024 BB	1 267 650 600 228 229 401 496 703 205 376 bytes

www.tiposdecomputadora.wordpress.com

1 Exabyte = 36,000 hours of HD video


IF THE 11 OZ COFFEE
ON YOUR DESK
EQUALS ONE GIGABYTE
A ZETTABYTE
would have
THE SAME VOLUME AS
THE GREAT WALL
OF CHINA

Cisco expects Internet traffic to hit 2 **zettabytes** annually by 2019.
the Internet is currently experiencing a 23 percent compound annual growth rate in traffic.

http://www.fierceenterprisecommunications.com/story/internet-traffic-will-hit-1-zettabyte-2016/2015-05-27

Telecom

- Nearly 5 billion mobile-phone subscriptions worldwide
- Over 3 billion people accessing the internet.[1]

- The world's effective capacity to exchange information through telecommunication networks:
  - 281 petabytes in 1986,
  - 471 petabytes in 1993,
  - 2.2 exabytes in 2000,
  - 65 exabytes in 2007[8]
  - predicted to reach 667 exabytes annually by 2014.(Wiki)

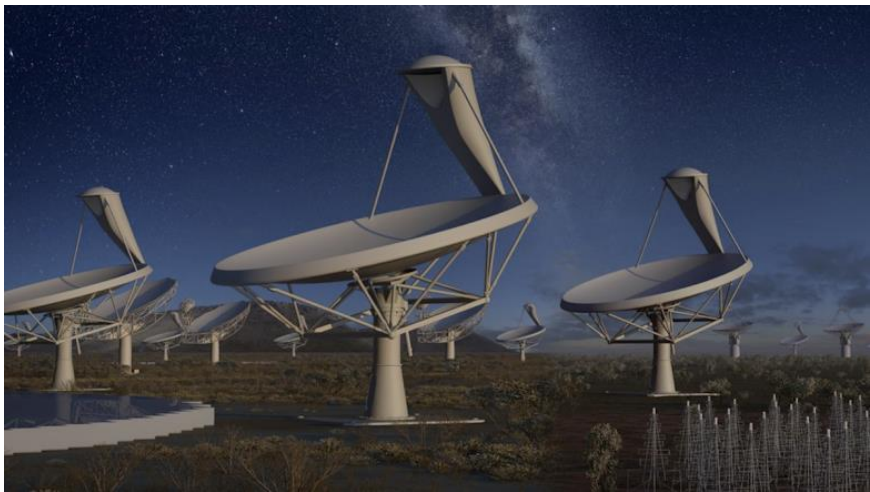
Video Surveillance

- Up to 5.9 million closed-circuit television cameras in UK
- Including 750,000 in "sensitive locations" such as schools, hospitals and care homes.
- 1 camera for every 11 people
- A few GB/camera/day

http://www.telegraph.co.uk/technology/10172298/One-surveillance-camera-for-every-11-people-in-Britain-says-CCTV-survey.html

Space Exploration

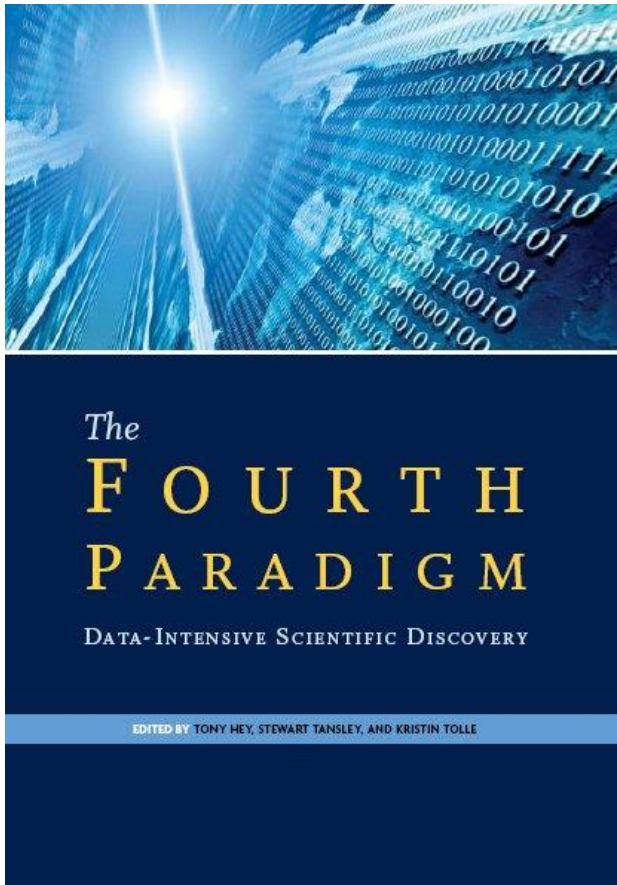
- Square Kilometer Array (SKA) project
- Radio telescopes spread over 1 sq km area
- “sensitive enough to detect airport radar on a planet 50 light years away”
- Generates 750 terabytes every SECOND!



http://venturebeat.com/2014/10/05/how-big-data-is-fueling-a-new-age-in-space-exploration/

DATA in modern world

- Data as the fourth pillar of science
- The first 3 pillars are:
  - Theory
  - Experiment
  - Computation



Jobs!

- Needed by 2018, in US alone:
  - 140,000 to 190,000 big data analysts
  - 1.5 million managers who understand big data

file:///D:/BACKUPD/courses/biol_data/intro/Big_data_McKinsey_Company.htm#sthash.2NWbgp5G.dpuf

Big Data Jobs in India

- Number of analytics jobs in India had doubled since last year (2016)
- 50,000 jobs available in Big Data in 2017
- India currently contributes to 12 per cent of worldwide analytics and data science job openings, making it the largest analytics hub in the world, outside the US.
- Amazon, Citi, HCL, Goldman Sachs and IBM stand out to be the leading organisations with most number of analytics openings

https://www.ndtv.com/india-news/big-data-jobs-in-high-demand-amazon-hcl-ibm-among-top-recruiters-study-1714709


Big data in Biology

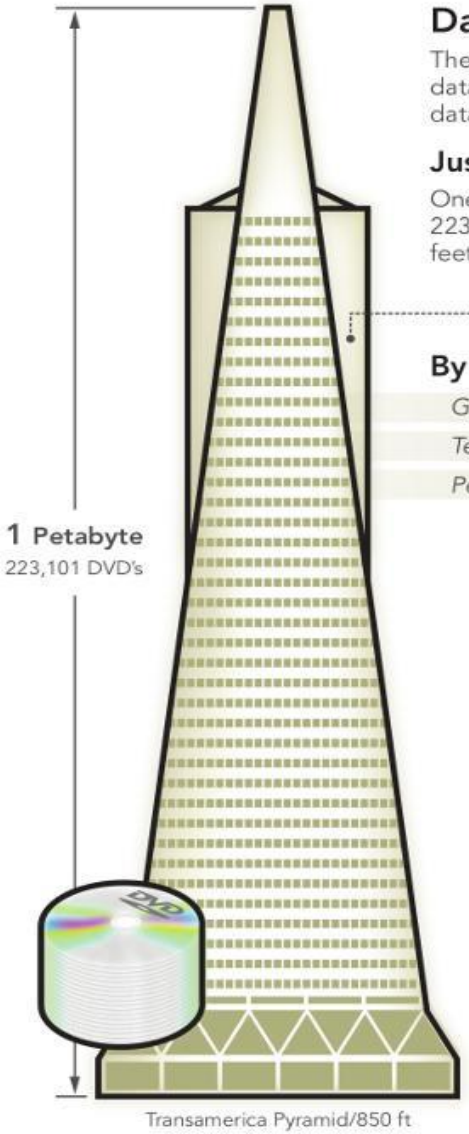
SLAC's Linac Coherent Light Source (LCLS) X-ray laser

Data Extremes at LCLS

The LCLS data team manages about 10 petabytes of data – including user-generated data and tape copies of raw data – which is about three times more than the total data library for movie-streaming and rental company Netflix.

Just how Big is a Petabyte?

One petabyte amounts to more than enough data to fill 223,000 DVDs – if stacked, those DVDs would measure 878 feet tall, or the equivalency of the Transamerica Pyramid.



Transamerica Pyramid/850 ft

By the Numbers:

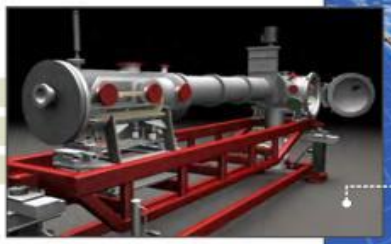
Gigabyte 1,000,000,000 bytes

Terabyte 1,000,000,000,000 bytes

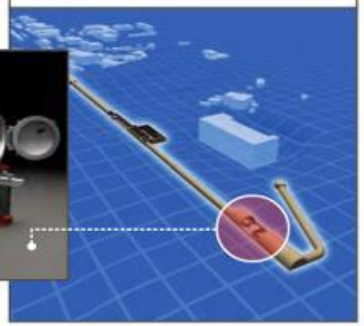
Petabyte 1,000,000,000,000,000 bytes

Why all that Data?

A crystallography experiment at LCLS-CXI can generate **2.5 million** images per day on average for a single 12-hour shift.



CXI Instrument



LCLS/CXI/Far Hall/Hutch 5

Interesting Facts:

The movie Avatar is reported to have taken over **1 petabyte** of local storage at Weta Digital for the rendering of the 3D CGI effects.

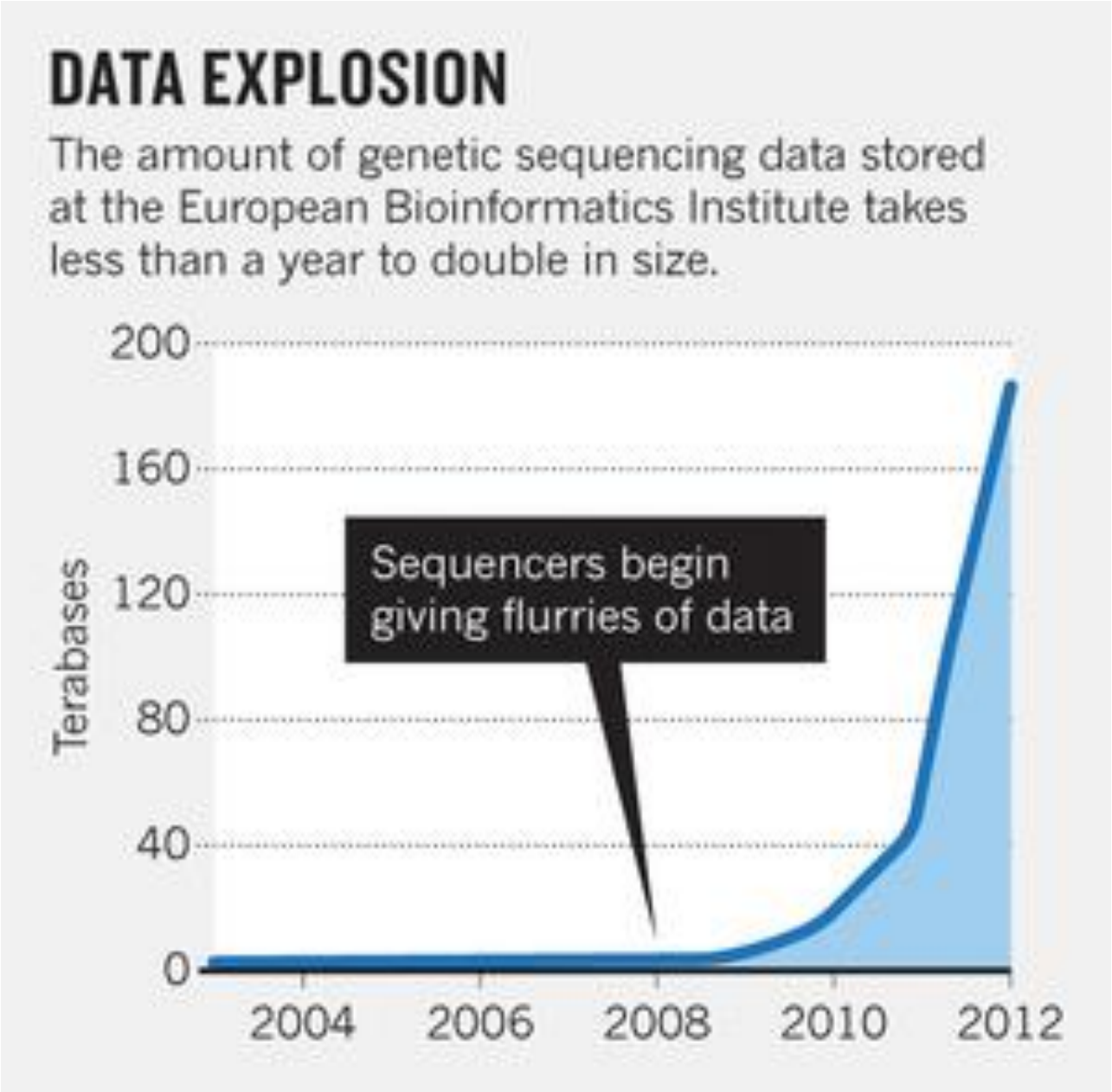
It is estimated that the human brain's ability to store memories is equivalent to about **2.5 petabytes** of binary data.

CERN amassed about **200 petabytes** of data from the more than 800 trillion collisions looking for the Higgs boson.

SLAC ENERGY

A Big Data place

- The European Bioinformatics Institute (EBI) in Hinxton, UK,
  - part of the European Molecular Biology Laboratory
  - one of the world's largest biology-data repositories,
  - currently stores **20 petabytes** of data and back-ups
  - Data about genes, proteins and small molecules.



(Marx, Nature, 2013)

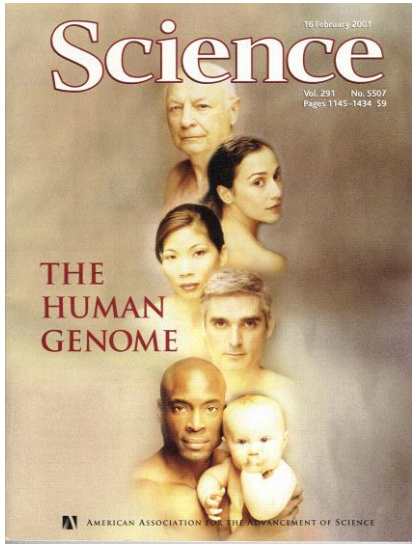
…Where is it all coming from?

Another Big Data place

- Beijing Genomics Institute (BG) in Shenzen, China
- “The Sequence Factory”
- 157 genome sequencing instruments working 24X7
- Samples from people, plants, animals and microbes.
- Each day, it generates 6 terabytes of genomic data.
- Every instrument can decode one human genome per week (used to take months or years and many staff).
- (Storage for 1 human genome = 200 GB)

THE ‘OM’ICS

HUMAN GENOMICS



Human Genome Project

- Aim
  - Identify sequence of bases on all 23 human chromosomes (3 billion bases/3Gb)
  - Identify genes within those sequences (~30 000 genes)
  - Locate the position of the genes on the chromosomes
- $6 bn, 1000 scientists, 50 countries, 10+ years!
- Human genome can now be sequenced in a few days on the ‘next-generation sequencing’ (NGS) machines
- Full genome data being collected from disease conditions
  - the combined cancer genome and normal genome from a single patient constitutes about 1 terabyte (10^{12} bytes)
  - a million genomes would generate an exabyte (10^{18} bytes). ”

(Courtesy: Karthik Raman)

Types of Genomics

- Disease genomics
  - Millions of patients per disease
- DNA profiling
  - Family lineages, parenting, forensics etc
- Comparative genomics
- Plant genomics
- Bacterial genomics
- Viral genomics

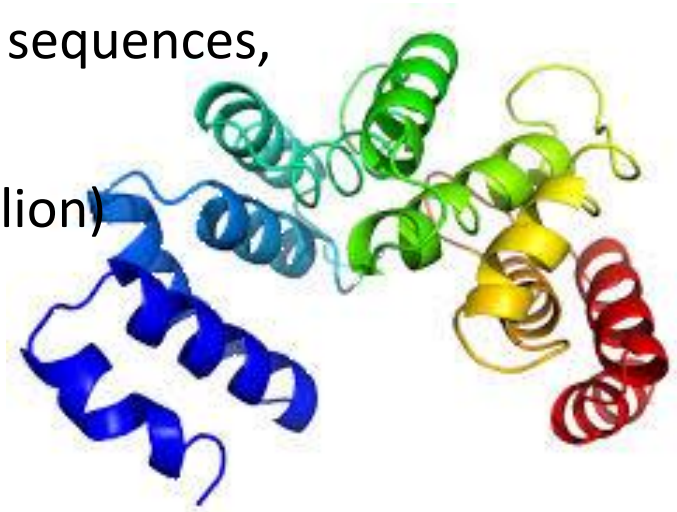
Transcriptome

- The set of all RNA molecules in a given cell, population of cells or an organism
- A single gene may produce many different types of mRNA molecules, so a transcriptome is much more complex than the genome that encodes it.

Proteins

- Peptide: a chain of amino acids (AAs)
- Assuming an average size of 200 AAs, number of possible proteins is 20^{200} > # protons in the universe
- Assume:
  - there are 10^7–10^8 species on Earth and
  - 10^3–10^5 genes/species,
  - ➔ there are 10^{10}–10^{13} unique protein sequences,
  - <<possible sequence space,
  - >> known protein number (about 1 billion)

https://www.ncbi.nlm.nih.gov/genbank/statistics/



Single Protein Study

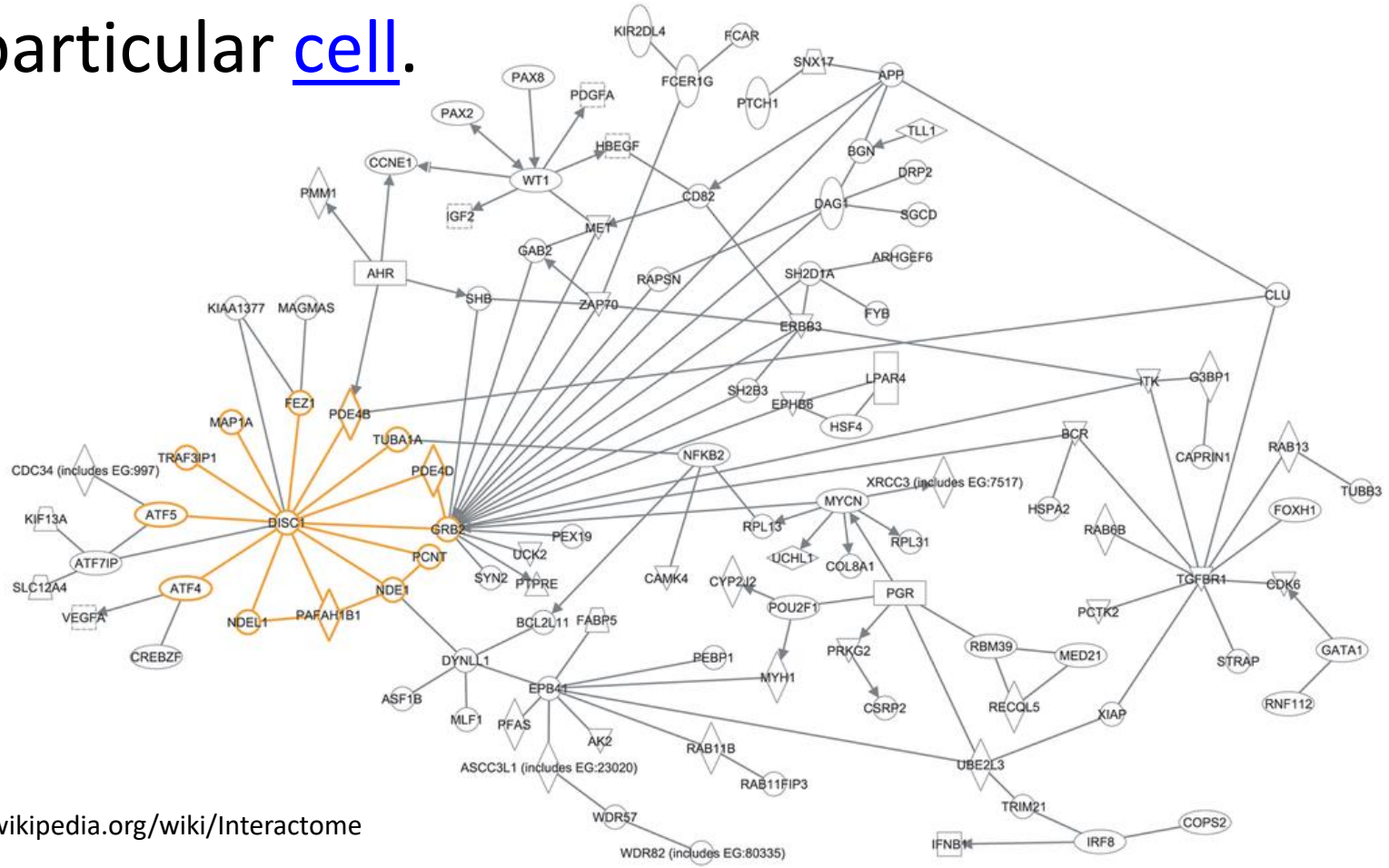
- Structure prediction
  - Secondary structure
  - Tertiary structure
  - Quaternary Structure
- Can be quite complex
  - P53 – tumor suppression gene assoc protein
  - P53 mutation database exists
  - 60,000 publications on p53 alone!!!

Proteomics

- The full complement of proteins expressed in a cell, organ or an organism

Interactome

- is the whole set of molecular interactions in a particular cell.

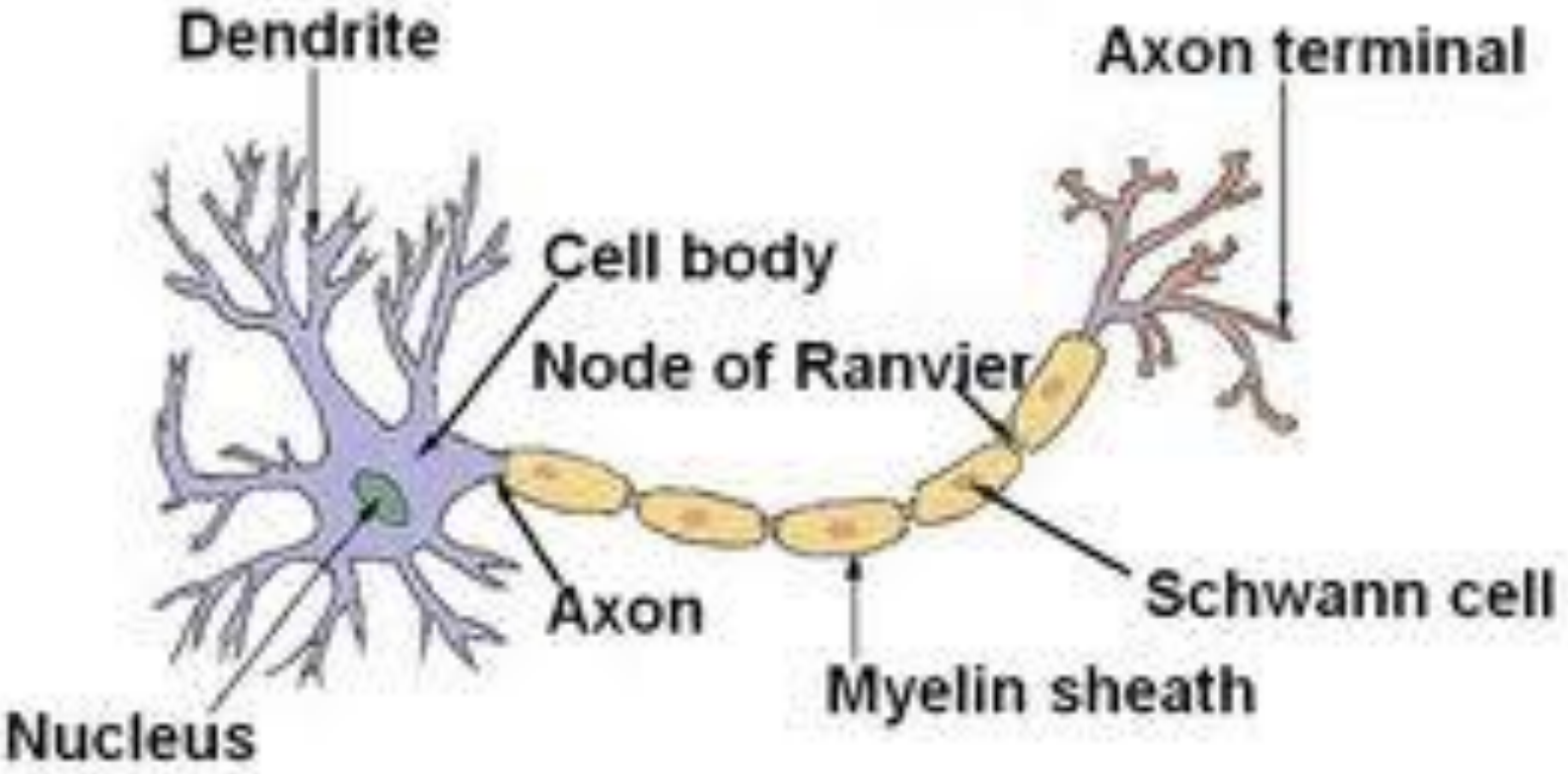


https://en.wikipedia.org/wiki/Interactome

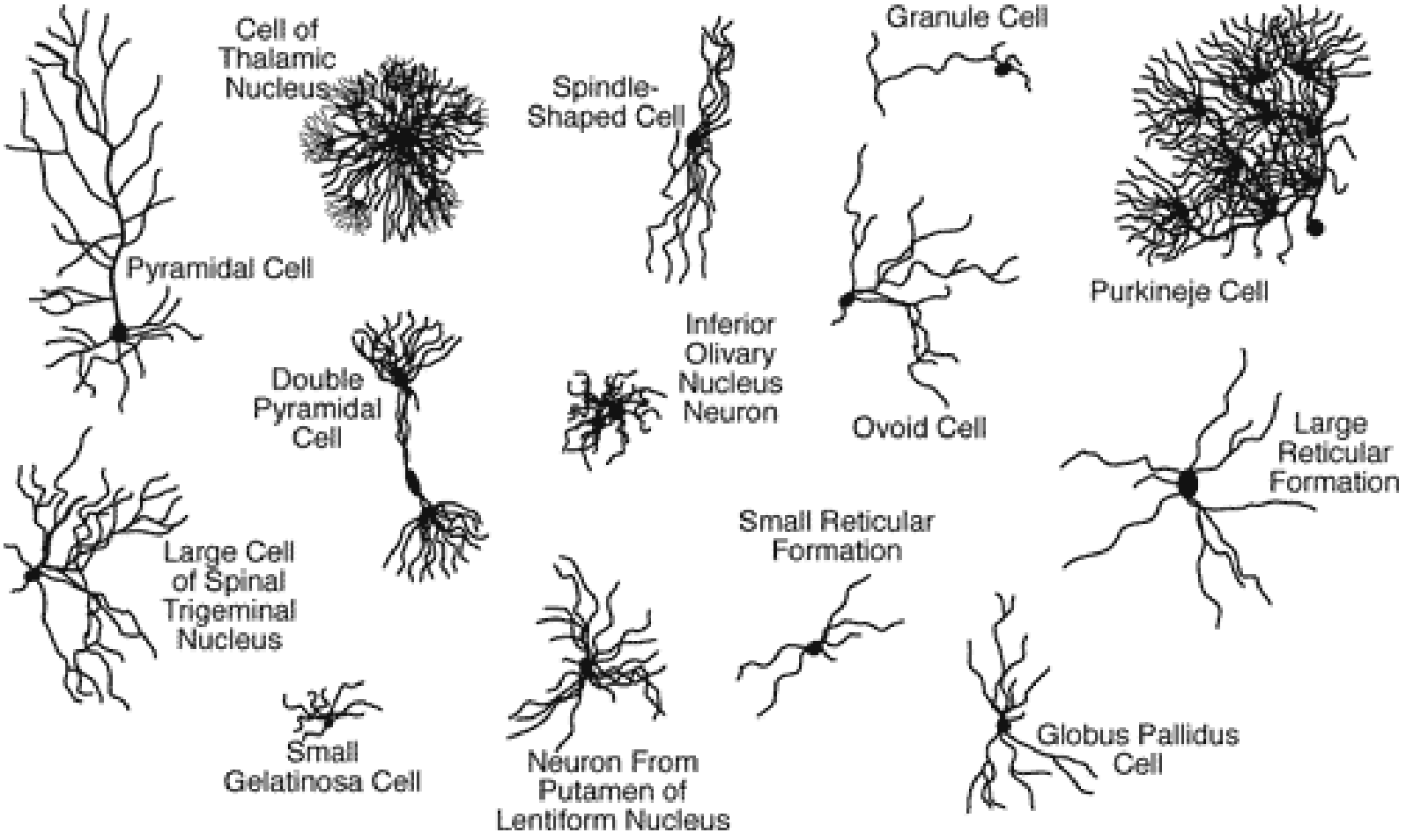
NOW…

LET’S GET TO THE CELL LEVEL

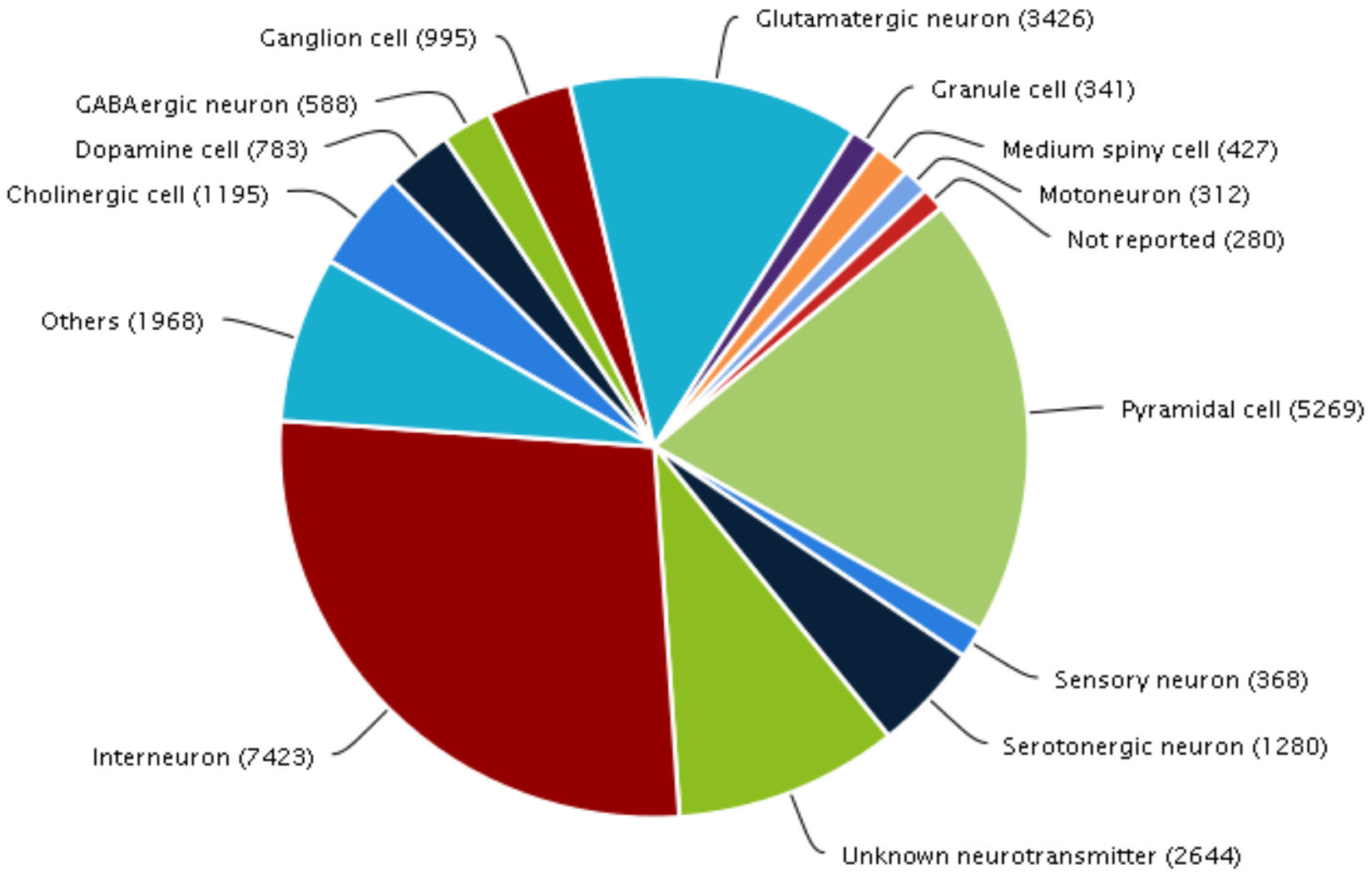
Structure of a Typical Neuron



Neurons come in different Shapes

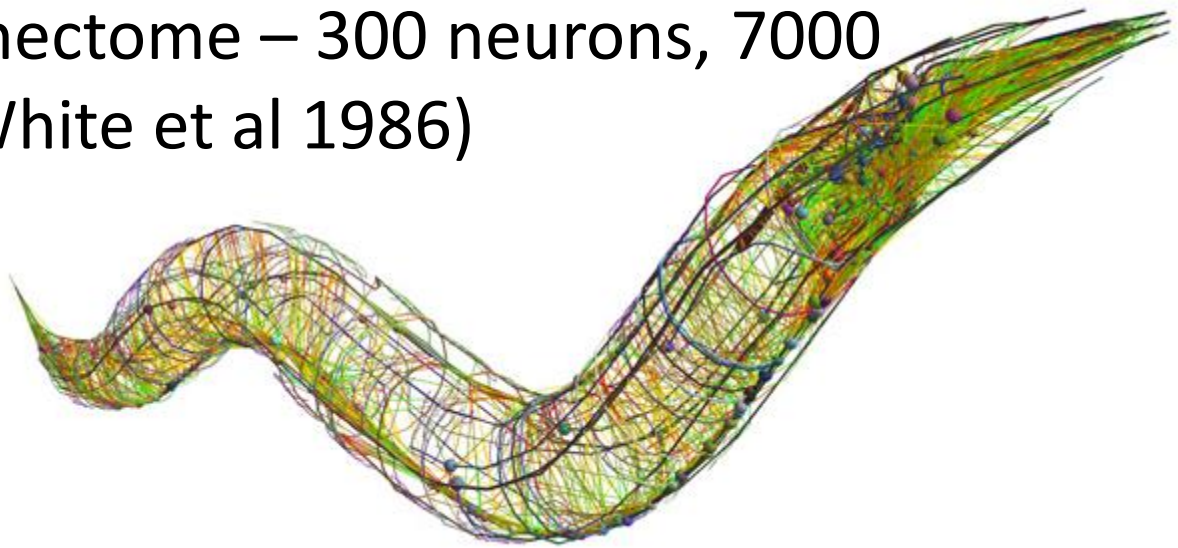


Neuromorph

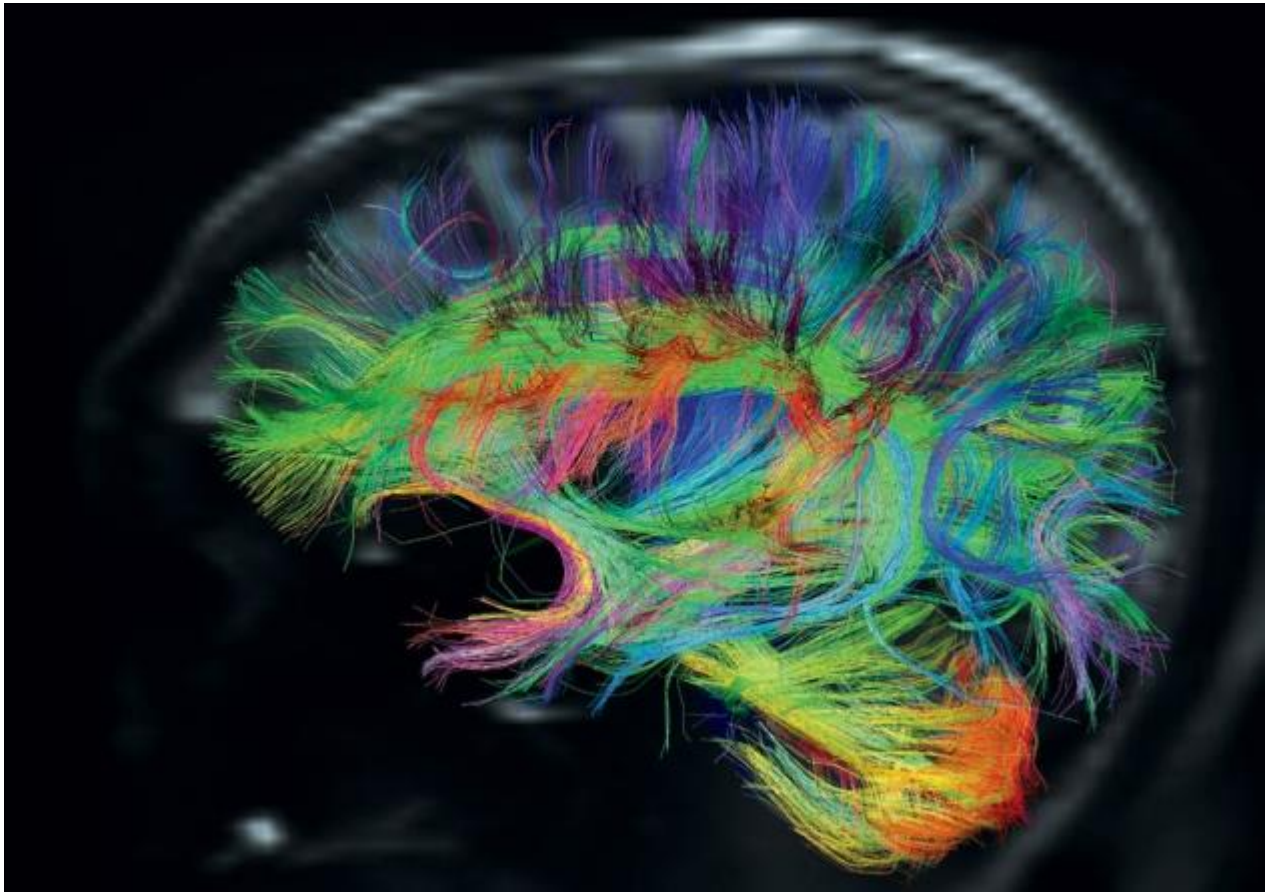

Glutamatergic neuron (3426)
Granule cell (341)
Medium spiny cell (427)
Motoneuron (312)
Not reported (280)
Pyramidal cell (5269)
Sensory neuron (368)
Serotonergic neuron (1280)
Unknown neurotransmitter (2644)
Interneuron (7423)
Others (1968)
Cholinergic cell (1195)
Dopamine cell (783)
GABAergic neuron (588)
Ganglion cell (995)

The Connectome Project

- To find out the complete wiring diagram of the brain
- Human brain
  - Has 100 billion neurons
  - Each neuron has about 1k-10k connections
- Feasible for smaller organisms
  - C. Elegans connectome – 300 neurons, 7000 connections (White et al 1986)



Human Connectome Project



- 1200 individuals

- fMRI + dMRI +
+ MRI + MEG

-Washington U +
U. Minnesota

http://www.humanconnectome.org/

Connectome Data Sizes

HCP Data Sizes (per Subject)		
Session	**Format**	**.zip File Size**
Structural	Unprocessed	70.99 MB
	Preprocessed	1.19 GB
Resting State fMRI	Unprocessed	2 GB
(each of 2 sessions)	Preprocessed	3.24 GB
Task fMRI (avg per Task)	Unprocessed	490 MB
	Preprocessed	771 MB
(all 7 Tasks)	Unprocessed	3.43 GB
	Preprocessed	5.4 GB
Diffusion	Unprocessed	2.18 GB
	Preprocessed	2.81 GB
Group-Average on Unrelated 20	Additionally Processed	289 MB
Total (per Subject)	Unprocessed	9.81 GB
	Preprocessed	15.77 GB
	Both	25.58 GB
Total (5 Subjects)	Unprocessed	62.16 GB
	Preprocessed	78.83 GB
	Both	141 GB
Total (20 Subjects)	Unprocessed	247.34 GB
	Preprocessed	315.05 GB
	Both	562.39 GB
Total (68 Subjects)	Unprocessed	815.4 GB
	Preprocessed	1.058 TB
	Both	**1.873 TB**

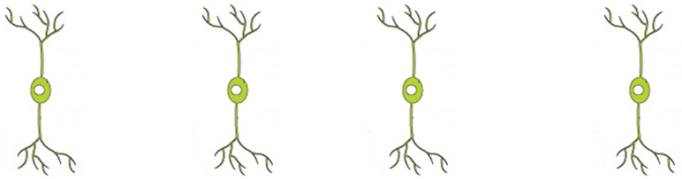
For 68 subjects:
1.8 Terabytes!!

The Hierarchy

Tissue/organ

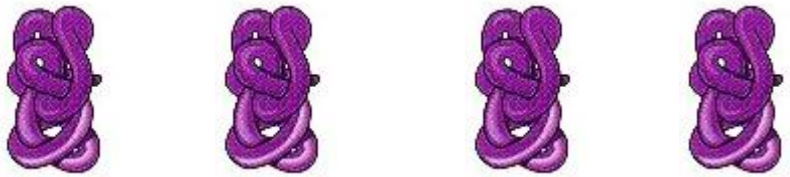
Large scale networks

Cell (e.g. neuron)



Microcircuits

Proteome



Metabolome/
Interactome

Transcriptome

Regulatory
Networks

Genome

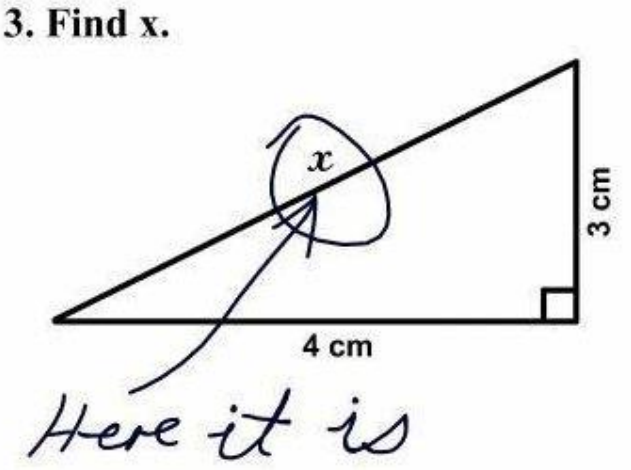
ANALYZE THAT!

Questions?

- How to represent a DATA OBJECT?
- How to compare objects?
  - Same type or different?
  - How different?
- How to group/cluster objects based on similarity?
- How to assign objects to classes?
- How to compare groups of objects?
  - Are two groups of objects really different?

Course Structure

- Mathematical Preliminaries
  - Vectors, vector spaces
  - Eigenvalues and eigenvectors
  - Derivatives in higher dimensions
  - Linear Least Squares problem
  - Optimization
    - Lagrange multipliers
  - Probability and Bayes theorem



Unsupervised Learning methods

- Clustering
  - K-means
  - Hierarchical clustering
  - Scale-based clustering
  - Fuzzy clustering
  - Graph based clustering
  - Self-organizing map
- Dimensionality reduction
  - PCA and ICA

Classification

- Prototype-based classification
  - K Nearest-neighbor classifier
  - Learning Vector Classification

Classification

- Discriminant-based classification
  - Linear Discriminant Analysis
  - Neural Networks –
    - Multilayer perceptron
    - Radial Basis Function Network
    - Deep neural networks
  - Support Vector Machines
  - Bayesian Classifier

Text Books

- Introduction to Data Mining – Tan/Steinbach/Kumar
- Neural Networks: A classroom approach – Satish Kumar
- Analysis of Biological Data – Whitlock/Schluter

Grading

- Quiz I – 20
- Quiz 2 – 20
- Assignments – 20
- Endsem – 40

- Grading policy – RG!!!


May the
DATA
be with you!