# STREAMING MESHES EXTRACTION AND VISUALIZATION FROM EXTREME-RESOLUTION BRAIN IMAGES

by

Giulia Clementi

A dissertation submitted to the faculty of
Roma Tre University
in partial fulfillment of the requirements for the

Laurea Magistrale in Ingegneria Informatica

Department of Engineering

Roma Tre University

October 2017

# ABSTRACT

This is the abstract.

*To my family:*

*my mom, that dedicated me her sweetest artworks,*

*my uncle Gianfranco, grandmother Tullia, grandfather Domenico,*

*to the Piras-Biggio branch, that are always too much proud of me,*

*to Manuel: he standed my mood after too much hours of study, - too less if you hear my mom - he's still by my side.*

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# NOTATION AND SYMBOLS

| | |
|---|---|
| $\mathring{A}$ | Angstrom |
| $[\partial_d]$ | Boundary operator |
| $[\delta^d]$ | Coboundary operator |
| $\mathcal{F}$ | Scalar field |

# ACRONYMS

| | |
|---|---|
| $1D$ | One dimensional |
| $2D$ | Two dimensional |
| $3D$ | Three dimensional |
| $AAV$ | Adeno-Associated Virus |
| $AAV9$ | Adeno-Associated Virus serotype 9 |
| $BRC$ | Binary Row Compressed |
| $B-rep$ | Boundary Representation |
| $Cap$ | Capsid protein |
| $CEDMAV$ | Center for Extreme Data Management, Analysys and Visualization |
| $CNS$ | Central Neurvous System |
| $COO$ | Coordinate representation |
| $CSC$ | Compressed Sparse Colums |
| $CSR$ | Compressed Sparse Row |
| $CVD$ | Computer Visual Design |
| $DNA$ | DeoxyriboNucleic Acid |
| $EM$ | Electron microscope |
| $fMRI$ | Functional Magnetic Resonance Imaging (fMRI) |
| $fs$ | femtosecond |
| $GFP$ | Green Fluorescent Protein |
| $HBI$ | 4-(p-hydroxybenzylidene)imidazolidin-5-one |
| $HZ$ | Hierarchical Z |
| $INFN$ | Istituto Nazionale di Fisica Nucleare |
| $I/O$ | Input/Output |
| $LAR$ | Linear Algebraic Representation |
| $LLNL$ | Lawrence Livermore National Laboratory |
| $LSST$ | Large Synoptic Survey Telescope |
| $mm$ | millimeters |
| $MRI$ | Magnetic Resonance Imaging |
| $NIH$ | National Institute of Health |
| $nm$ | nanometers |
| $PB$ | Petabyte |
| $PDB$ | Protein Data Bank |
| $pyplasm$ | Programming LAnguage for Solid Modeling |
| $Rep$ | Recombinase protein |
| $RNA$ | RiboNucleic Acid |
| $ROI$ | Region of Interest |
| $SC$ | SuperComputing |
| $SC$ | Structural Component |
| $SCI$ | Scientific Computing and Imaging |
| $SEM$ | Scanning Electron Microscope |
| $SpMV$ | Sparse Matrix-Vector |
| $ssDNA$ | single-stranded DeoxyriboNucleic Acid |
| $TB$ | Terabyte |

# ACKNOWLEDGEMENTS

I am grateful to Prof. Paoluzzi and Prof. Pascucci for have believed and invested in me to make this thesis possible. Special acknowledgements goes to all the team at Scientific Computing and Imaging (SCI) Institute led by Prof. Pascucci and to the staff, for their warm welcome in Salt Lake City, for having made me feel immediately part of the group and for having taught me the real sense of research, dialogue and collaboration. Thanks to Danilo Salvati, for his patient help and his generousity and generally speaking to the others at the Computational Visual Design (CVD) Lab at Roma Tre University.

# PREFACE

This is the preface.

# PART I

# PROBLEM STATEMENT AND STATUS QUAESTIONIS

# CHAPTER 1

# BIG DATA AND DATA-DRIVEN SCIENTIFIC DISCOVERY

This chapter is based on the transcript of a colloquium given by Professor Valerio Pascucci at the Department of Mathematics and Physics of the Roma Tre University on 28th November 2016. It will be described how the Big Data field caused a revolution in the paradigm by which scientists and engineers conduct scientific research. Furthermore, it will be discussed the main features of the infrastructures and techniques that need to be implemented to analyse and understand the Big Data models.

## 1.1 Data deluge and role of data management, analysis and visualisation

Dealing with the big data is one of the greatest challenges of our time. Astrophysics don't actually look anymore through telescopes to observe the sky. Instead, they are "looking" through large-scale, complex instruments which relay data to datacenters, and only then they look at the information on their computers [5]. The most productive facilities for astronomy in the world adopt a brute-force approach: they go around and survey the sky collecting as much data as possible. In northern Chile, alongside the existing Gemini South and Southern Astrophysical Research Telescopes, the Large Synoptic Survey Telescope (LSST) is currently under construction, and will photograph the entire available sky every few nights. The camera is expected to gather 15 TB a day of uncompressed data, over 200000 pictures per year and 100 PB of data in 10 years. Managing and data mining the output of the telescope is expected to be the most technically complex part of the project. Initial computer requirements are estimated at 100 teraflops of computing power and 15 PB of storage, rising as the project collects data. As highlighted by this introductory example, one of the key success factors in the scientific research now and in the future will lie in the ability to develop and exploit new tools in computer science to support the entire research

cycle, from data management to analysis and visualisation.

## 1.2 Data-driven scientific discovery

The way we carry out science an engineering, has undergone a dramatic change, due to the rise of the big data era.

### 1.2.1 Evolution of the scientific method: the four paradigms

The evolutionary path of the scientific method has been described by Jim Gray by the transition through four research paradigms:

#### 1.2.1.1 Empirical science

Empirical science is based on the scientific method, defined sistematically for the first time by Galileo Galilei (1564-1642) and consisting in the use of empirical evidence - systematic observation, measurement, and experiment - in order to formulate and test the initial hypotheses.

> All knowledge about reality begins with experience and terminates in it. Conclusions obtained by purely rational processes are, so far as Reality is concerned, entirely empty. It was because he recognized this, and especially because he impressed it upon the scientific world that Galileo became the father of modern physics and in fact of the whole of modern natural science.
> *Albert Einstein [2]*

#### 1.2.1.2 Theoretical science

Theoretical science employs mathematical models, physical laws and abstractions of physical objects and systems to rationalise, explain and predict natural phenomena. The recourse to the physical-mathematical formalism has become more and more accentuated in modern physics in XX century. The aim is to understand the infinite big and small phenomena, that can't be easily replicated and observed in laboratories. In this context, theoretical science and deductive reasoning are often the most feasible survey methods to knowledge discovery, in opposition to the scientific method.

#### 1.2.1.3 Computational science

Developed in the military field starting from the World War II, computational science implements numerical analysis to solve problems for which a quantitative theory already

exists, involving a wide range of different disciplines, such as physics, applied mathematics, and computer science. Computer simulations reproduce the behaviour of the system using the mathematical model. The simulations are a key tool for information that can not be tested empirically and are used to explore, gain new insights and estimate the performances of systems which are too complex for analytical solution.

#### 1.2.1.4 Data driven investigation

The simulations have carried us through much of the last half of the last millennium. At this point, we attended the data deluge, the explosion of the data from the experimental sciences. In the new model, data driven investigation, the data are captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline. The techniques and technologies for such data-intensive science are so different that it's worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration. [3]

### 1.2.2 "The end of theory: the data deluge makes the scientific method obsolete"[1]

This is the title of a provocative and revolutionary piece, written by Chris Anderson in 2008, in which he calls into question the old paradigm, in which "data, without model, are just noise" and physics had to drift into theoretical speculation, because of the impossibility to conduct experiments that would falsify the hypothesis. He wonders if today the mathematical model is needed at all, since it's only an approximation of the truth. The alternative is to employ the data and analyse them through computing clusters and statistical algorithms to find patterns where science cannot, without starting from initial hypothesis. Chris Anderson describes the World Wide Web as a huge, shared knowledge basis, in which petabytes of data are stored in the cloud, without the possibility and the need to keep the data locally any more, because of the too low capacity of hard disks.

## 1.3 Examples of Big Data applications

The new paradigm, the data driven investigation, is applied today to every field of scientific research. The result is the rise of a new concept of the scientific research itself, in

which computer science is the basic tool to explore and analyse complex datasets at ever finer scale. To grow in this kind of environment, an ever wider range of collaborations between institutions and researchers from all over the world is necessary, as well as the fusion of different disciplines to form a unique branch of research. The following are only few of the possible applications and are expected to have a huge impact on society in the next years:

- Climate data: The analysis of several terabytes of historic climate data allows us not only to extract information from the data, run simulations and visualise in an effective way the results, but also to find trends - like for instance global warming and climate change - and project in the future, to understand what will happen in the next years. In fig.1 is shown an animated visualisation produced by averaging results from 15 of the most advanced climate models in the world. Simulation of historical and projected climate data shows a strong warming trend in the lower atmosphere, underneath a cooling layer in the upper atmosphere. The model results were contributed by participants in the World Climate Research Programme's Climate Model Intercomparison Project (CMIP) and were archived by the U.S. Departement of Energy's Program for Climate Model Diagnosis and Intercomparison (PCMDI) at Lawrence Livermore National Laboratory. The data was analysed using the Climate Data Analysis Tools developed by PCMDI. The visualisation was produced with the integrated `ViSUS` tools developed by VACET.

- Materials science: Materials science is a syncretic discipline hybridising metallurgy, ceramics, solid-state physics and chemistry in addiction to engineering. It's aimed to discovery and design new materials. Many of the most pressing scientific problems humans currently face are due to the limits of the materials that are available. Thus, breakthroughs in materials science are likely to affect the future of technology significantly.

- Computational fluid dynamics: CFD is aimed to solve and analyse problems that involve fluid flows and to simulate the interaction of liquids and gases with surfaces defined by boundary conditions. Ongoing research yields software that improves the accuracy and speed of complex simulation scenarios such as turbulent flows. For

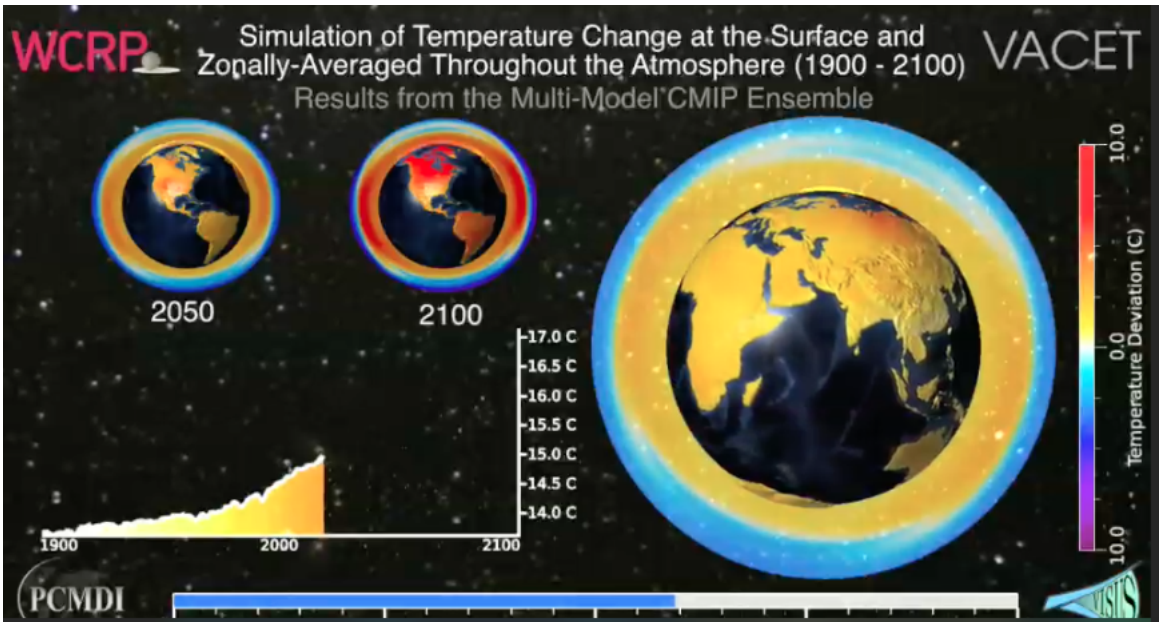


**Figure 1.1**. Simulation showing the climate change from historical and projected climate data.

instance, the Rayleigh-Taylor instability, is an instability of an interface between two fluids of different densities which occurs when the lighter fluid is pushing the heavier fluid. Fig.2 shows a scientific visualisation of an extremely large simulation of a Rayleigh-Taylor instability problem. The simulation has been generated by the `VisIt` Visualisation System, developed at the Lawrence Livermore National Laboratory.

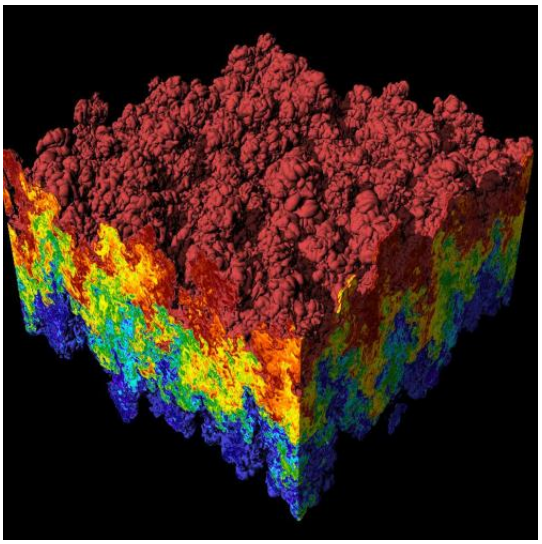

**Figure 1.2**. Simulation of the Rayleight-Taylor instability problem (LLNL).

- Energy research: this application is obviously expected to have a strong impact in the future. Several disciplines, such as chemistry and engineering, are involved. It's possible to simulate the generation of energy from different sources. In Fig.3 is shown a combustion simulation, displayed on the POWDER (Parallel, Out-of-core Wall for Display of Extreme Renders), a shared SCI Institute resource (University of Utah).

**Figure 1.3**. The authoress and a combustion simulation (SCI Institute, University of Utah).

- Study of the universe and its origins: this kind of study allows the simulation of past and future events in the universe, like for instance the Andromeda-Milky Way collision, involving the two largest galaxies in the Local Group and predicted to occur in about 4 billion years.

- Personalised medicine: Personalised medicine is aimed to select optimal therapies based on the context of a patient's genetic content or other molecular or cellular analysis. This branch of research encompasses all sorts of personalisation forms, starting from pharmacogenomics, by means of which we'll soon be able to synthesize drugs that are specific for the individual patient.

- Neuroscience: This field of research allows the multi-scale exploration of the brain. It's a fascinating and extreme challenging discipline: it will suffice to consider that at the moment we have no detailed circuit diagrams of the brains of humans or any other mammals. I'll talk extensively about the mapping of the brain in the second chapter, since it's one of the main topics of this thesis.

## 1.4 Scientific infrastructures

At this point of the treatment, it's clear that today the data are a key aspect in conducting scientific analysis and visualisation and a major driver of any kind of scientific investigation. This section focuses on the infrastructure that has to be built generally speaking to deal with a Big Data model both in terms of the physical infrastructure and the software architecture, providing also a list of the main activities, techniques and requirements.

### 1.4.1 High Performance Computing (HPC)

The scale of events being simulated by computer simulations has far exceeded anything imaginable using traditional mathematical modelling, thanks to the development of High Performance Computing. The graph in fig. 4 shows a clear trend, an unprecedented evolution of HPC resources in the last years. The employment of HPC offers a new challenge to computer scientists: "running efficiently Big Data computations is a Big Data problem"[4]. There's a growing complexity in managing massive logs, complex memory hierarchies, multi-dimensional data, complex I/O pathways, network connections, power consumption and so on. In this sense performance analysis plays a central role in HPC.

### 1.4.2 Storing the data: the cloud

Cloud computing is another emerging paradigm. It offers obvious advantages, however while it's currently in use in search engines and in the hosting of Web sites, the revolution in scientific computing has still to happen. In many cases, the nodes need to be tightly integrated with a very low latency. In yet other cases, very high I/O bandwidth is required. Certainly, more specialised data clouds are bound to emerge soon.

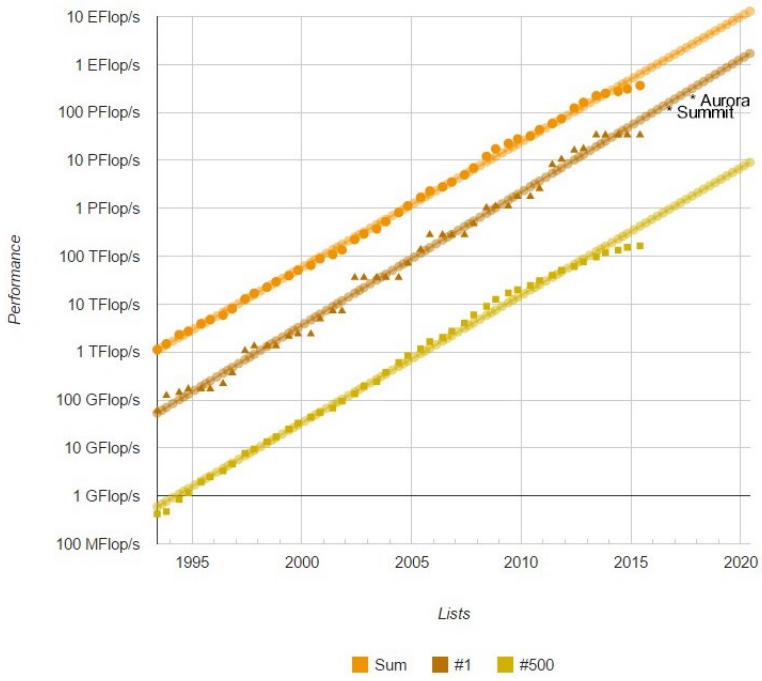


**Figure 1.4**. Evolution of the HPC resources.

### 1.4.3 Web server

The idea is to set up a web server, to support a client-server model in addition to the traditional standalone application. In this way, researchers from all over the world will be able to stream immediately, resample the data, comparing simulations, look at statistical variation and so on. A great challenge in this sense is to scale in the number of connections that can be established and maintained robustly.

## 1.5 Activities, requirements and techniques

### 1.5.1 Critical activities in knowledge discovery

- HPC
- Scientific computing
- Machine learning
- Data analysis
- Data mining
- Data exploration

- Data visualisation, often of multivariate data
- Uncertainty quantification: this area is intended to estimate the accuracy of the prediction. The system should also be able to explore a high dimensional space of possible configurations.
- Verification and validation: these is the most difficult part in terms of violating the scientific method, since it could be impossible to experiment empirically the different alternatives.

### 1.5.2 Requirements

- Interactivity;
- Scalability in the number of cores;
- I/O layer: this is a key point, since traditional systems would be stalled by I/O cost in Big Data applications;

### 1.5.3 Techniques

- Analysis on-the-fly versus offline precomputation of the outputs: both the approaches are common, depending on the particular application;
- Multi-resolution representations of massive data and visualisations in streaming-mode versus information-preserving abstraction: The multi-resolution approach is essential to avoid the stall of the system and offer immediately a visualisation to the user. However, the limits of this approach have been highlighted and the risk after several steps is losing too much information. One of the aims of this thesis is to rethink the multiscale representation of massive data models as it's insufficient to deal with Big Data, since data analysis results often don't represent well important trends. New data abstractions are needed: the magic behind algebraic topology and a pure mathematical approach is that it's possible to extract features and express a very complex concept in few bytes without losing information. Topology is only one of the broad spectrum of complementary techniques that we'll have to set up in the next years.

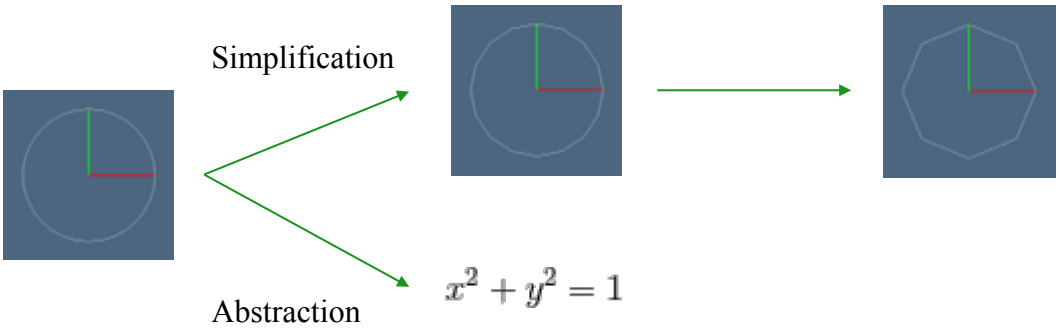


**Figure 1.5**. Difference between simplification and abstraction.

## 1.6 Summary and conclusions

The big data are a great opportunity to achieve new scientific discoveries and engineering innovations. The computer science community is today a central player in the development of modern science. The scientific discovery is the result of a wide range of interdisciplinary and intercontinental collaborations. Data generation, processing and exploration have driven the new scientific revolution. They are the "telescope" of modern science and engineering. [4]

## 1.7 References

[1] C. Anderson, *The end of theory: The data deluge makes the scientific method obsolete*, The fourth paradigm Data-Intensive scientific discovery, (2009).

[2] A. Einstein, *On the method of theoretical physics*, Philosophy of Science, 1 (April 1934), pp. 163–169.

[3] T. H. G. Bell and A. Szalay, *Beyond the data deluge*, Science, 323 (2009), pp. 1297–1298.

[4] V. Pascucci, *Transcript of the colloquium given at roma tre university*, (2016).

[5] K. T. Tony Hey, Stewart Tansley, *The fourth paradigm Data-Intensive scientific discovery*, Microsoft Research, Redmond,Washington, 2009.

# CHAPTER 2

# GENERATING THE MAP OF THE CEREBRAL CIRCUITRY

The study of the brain is a fascinating and extreme challenging discipline: it will suffice to consider that at the moment we have no detailed circuit diagrams of the brains of humans or any other mammals.[5] How the brain works and how its function follows from its structure, is still a mystery. Several disciplines deal with the understanding of the brain: psychology, neurology, psychiatry, cognitive science and many others. The field of *neuroscience* encompasses all this different approaches.

Historically, the oldest method of studying the brain is anatomical, and until the middle of the 20th century, much of the progress in neuroscience came from the development of better *cell stains*[1] and better microscopes. *Neuroanatomy* studies the large-scale structure of the brain as well as the microscopic structure of neurons and their components, especially synapses.

As will be discussed, today we're able to map the human brain at macro level into regions, that can be roughly associated with specific types of activities. We know that this kind of knowledge does not suffice, because many regions of the brain are able to participate in completing complex tasks.

On the other hand, the mapping of the brain at micro level poses enormous technical difficulties. However, the discovery of cell-type specific tracers and new methods for

---

[1]The most part of the animal tissue is colorless or transparent, since made for the most part of water without pigments. The result is that animal tissue without further techniques, is almost invisible under the microscope. Therefore, since the birth of scientific histology, a set of substances has been collected, able to color the cells or their various parts to make them immediately distinguishable. Then, cell staining is an auxiliary technique used in microscopy to enhance contrast in the microscopic image. Staining can be done *"in vivo"*, dyeing living tissues, *"in vitro"*, colouring cells or structures that have been removed from their biological context, or *"ex vivo"*, in which many cells continue to live and metabolize until they are fixed. *Fixation* is an another critical step in the preparation of histological sections, by which biological tissues are preserved from decay, preventing autolysis or putrefaction. Fixation terminates any ongoing biochemical reactions, and may also increase the mechanical strength or stability of the treated tissues.

high-resolution imaging and 3D reconstruction of large datasets have led a renaissance in neuroanatomy. [1] The development of *immunostaining* [2] techniques has allowed investigation of neurons that express[3] specific sets of genes. Moreover, *functional neuroanatomy* uses medical imaging techniques to correlate variations in human brain structure with differences in cognition or behavior. The ultimate goal is the generation of the *human connectome*, a complete circuit diagram of the brain.

> The connectome will significantly increase our understanding of how functional brain states emerge from their underlying structural substrate, and will provide new mechanistic insights into how brain function is affected if this structural substrate is disrupted. *Sporns et al.*[7]

To sum up, this area of research is rich in terms of new discoveries, approaches and challenges. This chapter presents an overview of the state of the art in the field of connectomics, including a brief description of the main histological and imaging techniques, issues, limits and a list of relevant research projects in this area.

## 2.1 Connectome

As explained in the introduction, the connectome is a complete circuit diagram of the brain. Since the birth of the term, the relation of duality between "connectome" and "genome" has been highlighted:

> It is clear that, like the genome, which is much more than just a juxtaposition of genes, the set of all neuronal connections in the brain is much more than the sum of their individual components. The genome is an entity itself, as it is from the subtle gene interaction that life emerges. In a similar manner, one could consider the brain connectome, set of all neuronal connections, as one single entity, thus emphasizing the fact that the huge brain neuronal communication capacity and computational power critically relies on this subtle and incredibly complex connectivity architecture. *Patric Hagmann [4]*

So connectome determines literally who we are, evidently supporting a great number of variable dynamic states, depending on current sensory inputs, global brain state, learn-

---

[2]The contrast for imaging can come from *immunostaining*, based on the principle of conjugation antigen-antibody to detect a specific protein in a sample. In addition, in immunostaining, the antibodies are labeled with fluorescent tags that are the key to final imaging result.

[3]*Expression* is the process by which information from a gene is used in the synthesis of a functional gene product, often proteins, but also functional RNA. This process is used by all known life - eukaryotes, including multicellular organisms, prokaryotes, bacteria and archaea, and viruses - to generate the macromolecular machinery for life. The steps by which this happens include the transcription, RNA splicing, translation, and post-translational modification of a protein.

ing, development and experience.

Connectomics may range in scale from a detailed map of the full set of neurons and synapses within part or all of the nervous system of an organism to a macro scale description of the functional and structural connectivity between all cortical areas and subcortical structures, depending on the levels of resolution in brain imaging. Since this kind of exploration encompasses the study of the brain at different length scales, from the entire brain to few nanometers, so, of course, we could not hope for a single imaging technique that covers that all, so for that purpose we need to distinguish different scales and a variety of bioimaging tools and technologies that has been developed.[10] These scales can be categorized as *nanoscale*, *microscale*, *mesoscale* and *macroscale*.

### 2.1.1 Nanoscale and microscale

At the microscopic level, the human brain comprises billions of neurons, each connected to other neurons by up to several thousand synaptic connections. The aim of the following subsection is to quantify the amount of data to handle. This is a crucial point: if we don't know the scale of the problem, we'll go on looking for the needle in the haystack, the nuclei and the dendrites within an unmanageable amount of neural data.

#### 2.1.1.1 Big data application

A good part of what we need to know about the problem can be found within the article *"Discovering the Wiring Diagram of the Brain"* of J. W. Lichtman et al. (Harvard University). Their laboratory is giving extremely important contributions to the discovery of the connectome. The first considerations concern the features and the amount of the connections generally speaking:

> To get a sense of the scale of the problem, consider the cerebral cortex of the human brain, which contains more than 160 trillion synaptic connections. These connections originate from billions of neurons. Each neuron receives synaptic connections from hundreds or even thousands of different neurons, and each sends information via synapses to a similar number of target neurons. This enormous fan-in and fan-out can occur because each neuron is geometrically complicated, possessing many receptive processes (dendrites) and one highly branched outflow process (an axon) that can extend over relatively long distances. [...] The staggering numbers and complex cellular shapes are not the only daunting aspects of the problem. The circuits that connect nerve cells are nanoscopic in scale. The density of synapses in the cerebral cortex

> is approximately 300 million per cubic millimeter. [...]

Here they describe some introductory concepts about the imaging techniques they use for their analysis. As previously said, every tool is specific for a length scale. *Functional Magnetic Resonance Imaging* (fMRI) and *diffusion magnetic resonance imaging* are non-invasive imaging technologies, giving important results in living subjects at macroscale. fMRI is also used in the resting state and during tasks to study the functions of the connectome circuits. On the other side electron microscopy (EM) (see 2.1) [4] is needed when looking for data at microscale. The downside is that this kind of analysis currently requires post-mortem tissue. This ideas will be further deepened in the course of this chapter.

> Functional magnetic resonance imaging (fMRI) has provided glimpses into the macroscopic 3-D workings of the brain. However, the finest resolution of fMRI is approximately 1 cubic millimeter per voxel, the same cubic millimeter that can contain 300 million synapses. Thus there is a huge amount of circuitry in even the most finely resolved functional images of the human brain. Moreover, the size of these synapses falls below the diffraction-limited resolution of traditional optical imaging technologies. [...] Presently, the gold standard for analyzing synaptic connections is to use electron microscopy (EM), whose nanometer (nm) resolution is more than sufficient to ascertain the finest details of neural connections. But to map circuits, one must overcome a technical hurdle: EM typically images very thin sections (tens of nanometers in thickness), so reconstructing a volume requires a serial reconstruction, whereby the image information from contiguous slices of the same volume is recomposed into a volumetric dataset. There are several ways to generate such volumetric data, but all of these have the potential to generate astonishingly large digital image data libraries.

The good news is that we currently have microscopy techniques that are able to image at the resolution sufficient to identify every single neuron, dendrite and axon. However, we need to roughly calculate the amount of memory, necessary to save the images of the human brain at microscale resolution. Lichtman and the others start from a volume of 1 cubic mm of brain, the resolution must be such to allow the reconstruction by EM of all the synaptic circuitry:

> Unambiguously resolving all the axonal and dendritic branches would require sectioning at probably no more than 30 nm. Thus the 1 mm depth would

---

[4] An *electron microscope* uses a beam of electrons instead of light, exploiting the wave-particle duality of electrons. It consists in an electron emission source, electromagnetic lenses and an electron detector. A very thin sample is positioned along the electron beam. The electron beam is produced, accelerated and then focused on the sample by the lenses. The beam passes through the sample which modifies it and imprints the image. The beam is then magnified by other lenses and detected for example using fluorescence.

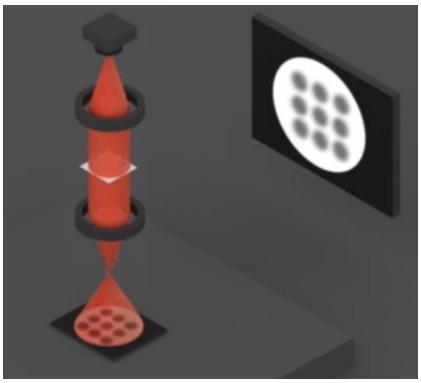

**Figure 2.1**. Electron microscope schema

> require 33,000 images. Each image should have at least 10 nm lateral resolution to discern all the vesicles (the source of the neurotransmitters) and synapse types. A square-millimeter image at 5 nm resolution is an image that has roughly $4 \times 10^{10}$ pixels, or 10 to 20 gigapixels. So the image data in 1 cubic mm will be in the range of 1 petabyte. The human brain contains nearly 1 million cubic mm of neural tissue.

After this analysis of the problem, the conclusion is that the discovery of the connectome at this resolution is a Big Data application, pushing the boundaries of our current memory availability and computing resources.

### 2.1.2 Mesoscale

In this case we refer to a spatial resolution of hundreds of micrometers. This kind of connectome attempts to capture anatomically and functionally distinct neuronal populations, formed by local circuits (e.g. cortical columns) that link hundreds or thousands of individual neurons. This scale still presents very ambitious technical challenges.

### 2.1.3 Macroscale

A connectome at the macroscale - millimeter resolution - attempts to capture large brain systems that can be parcellated[5] into anatomically distinct modules - areas, parcels or nodes - each having a distinct pattern of connectivity.

The initial explorations in macroscale human connectomics were done using either equally sized regions or anatomical regions with unclear relationship to the underlying

[5] *Parcellation* refers to the division of the brain into functionally distinct parcels, brain regions with distinct architectures, connectivity, functions, and topography.

functional organization of the brain (e.g. gyral and sulcal-based regions). Connectomic databases at the mesoscale and macroscale may be significantly more compact than those at cellular resolution, but they require effective strategies for accurate anatomical or functional parcellation of the neural volume into network nodes.

## 2.2 From histological samples to 3D interactive visualizations

This section describes in detail the process applied at the University of Utah by the laboratory of Prof. Angelucci to obtain a stack of images from tissue samples. The images, in this case at microscale, are the raw data from which starts the work of the team led by Prof. Pascucci at the SCI Institute. It will also be explained a different approach and results, useful to project in comparison, the one developed at Harvard by Lichtman and Reid.

### 2.2.1 CLARITY tissue clearing

CLARITY is a technique for preparation of histologic samples, developed at the Stanford University School of Medicine. It allows detailed views of the cerebral structure and connections, while maintaining the organ completely intact. The technique consists of removing completely the lipid substrate in the brain by replacing it with a gel made of various components, as acrylamide, bis-acrylamide and formaldehyde. The result is making the tissue completely transparent in respect of the light, while the lipid reticulum is naturally opaque. The cerebral structures are completely visible and, at the same time, their integrity is preserved. This kind of potential to drill down had been reached before only through biopsy. Previously, similar attemps had failed, due to the interference between this kind of preparation and the fluorescent tagging, aimed to highlight cells and anatomical structures. The degree of visibility in the structure allows the mapping of the brain at molecular scale, offering the amazing opportunity to appreciate the finer biological tissue structures, leaving unaltered the stuctures at macroscale. It differs from the traditional techniques, like microtomy, in which the brain is dissected into thin slices.

> Clarity is the transformation of intact biological tissue into an hybrid form in which specific components are replaced with exogenous elements that provide new accessibility or functionality. [2]

In terms of brain imaging, obviously, the ability for CLARITY imaging to reveal specific

structures in such unobstructed detail has led to promising avenues of future applications in local circuit wiring and connectome generation. For instance, thanks to CLARITY, a peculiar pattern has been discovered, where neurons connected back to themselves and their neighbours, which has been observed in animals to be brought back to autism-like behaviours.

CLARITY can be used with little or no modifications to clear many other organs such as lever, pancreas, spleen, testes, and ovaries and other species such as zebrafish. While bone requires a simple decalcification step, similarly, plant tissue requires an enzymatic degradation of the cell wall. On the other side, the main disadvantage of the technique is the length of time it takes to create and image a sample.

To summarise:

> CLARITY is powerful. It will enable researchers to study neurological diseases and disorders, focusing on diseased or damaged structures without losing a global perspective. That's something we've never before been able to do in three dimensions. *Francis Collins*

#### 2.2.1.1 Procedure

The process of applying CLARITY imaging begins with a postmortem tissue sample. A series of chemical treatments must be applied to achieve transparency, in which the lipid content of the sample is removed, while almost all of the original proteins and nucleic acids are left in place. The purpose of this is to make the tissue transparent and thus amenable to detailed microscopic investigation of its constituent functional parts, predominantly proteins and nucleic acids. To accomplish this, the preexisting protein structure has to be placed in a transparent scaffolding which preserves it, while the lipid components are removed. This 'scaffolding' is made up of hydrogel monomers such as acrylamide. The addition of molecules like formaldehyde can facilitate attachment of the scaffolding to the proteins and nucleic acids that are to be preserved, and the addition of heat is necessary to establish the actual linkages between the cellular components and the acrylamide. Once this step is complete, the protein and nucleic acid components of the target tissue's cells are held firmly in place, while the lipid components remain detached. Lipids are then removed over 1-2 weeks of passive diffusion in detergent, or accelerated by electrophoretic methods to only hours to days. The large majority of non-lipid molecules, such as proteins

and DNA, remain unaffected by this procedure, thanks to the acrylamide gel and chemical properties of the molecules involved. The tissue expands during this process, but as needed can be restored to its initial dimensions with a final step of incubation in refractive index matching solution. By this stage in the process, the sample has been fully prepared for imaging.

### 2.2.2 Fluorescent imaging techniques

Fluorescent imaging techniques are revolutionising biology and medicine. They're based on the use of fluorescent proteins, that allow the tracking and visualization of molecular processes in living cells as well as whole organisms. In this sense, the strategy that has been proved to be successful is the use of light from living organisms. Light given off by living organisms is referred to as *bioluminescence*. This is proper of the 90% of the organisms under the ocean. By far, the most exiting and recently useful bioluminescent organism is the *Jellyfish Aequorea Victoria*, which not only is bioluminescent, but has a fluorescent protein associated with it, called Green Fluorescent Protein (GFP). The bioluminescent protein in the jellyfish is the *aequorin*, giving off blue light that is absorbed by the GFP. In response to this absorption the GFP gives off its own green light. [6]

The characteristic of fluorescence has become an extremely important new tool for biologists, in particular because the molecule responsible for fluorescence in the jellyfish, the GFP, has been shown to be isolatable and capable of being inserted into an exogenous organism. This incredible discovery led to giving the nobel prize in medicine in 2008 to Shimomura, who was the first to isolate the GFP, Chalfie, who was the first actually inserting the GFP in an exogenous organism, and Tsien, for the development of GFP. [6]

#### 2.2.2.1 Fluorescent tagging

*Fluorescent tagging* or *labeling* can serve similar purposes in respect to cell staining. The main advantages are that this technique is compatible with live tissue and doesn't require

---

[6] *Fluorescence* is the property of a substance to convert light from one color to another. The *fluorophore* is the name of the fluorescent compound. The incoming light hits the fluorophore and in response, not only it reflects the incoming light, but it can also give off light of a different color (green, in fig. 2.2 [3] ). Looking at the electron orbital diagram, the electrons are circulating around the nucleus of a molecule that's part of the fluorophore and what happens is that when blue light comes is, it excites the electrons and causes a jump to a different orbit. The electrons will stay in this state until at some point they drop down to the original orbit and in response to the shift a light is given off, in this case green fluorescent light.

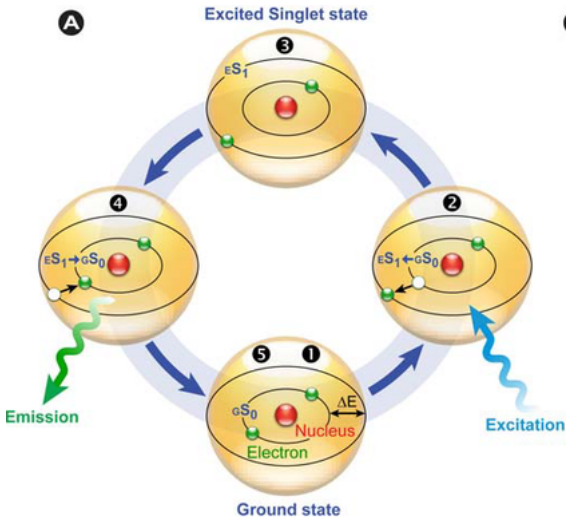


**Figure 2.2**. Principle of fluorescence.

necessarily fixation. The fluorescent tag is the fluorophore, the molecule that is attached chemically or biologically to aid in the labeling and detection of the biomolecule of interest. The fluorophore selectively binds to a specific region or functional group on the target molecule. GFP is one of the most common tags.

#### 2.2.2.2 Fluorescent labels: insight into the GFP

Of the various methods of labeling biomolecules, fluorescent labels are advantageous in that they are highly sensitive even at low concentration and non-destructive to the target molecule folding and function. As previously said, GFP (2.3, source Protein Data Bank (PDB)[7]) is a naturally occurring fluorescent protein from the Jellyfish Aequorea Victoria that is widely used to tag proteins of interest. GFP emits a photon in the green region of the light spectrum when excited by the absorption of light.

The primary structure[8] consists of 238 amino acids. The secondary structure[9] is consisting of eleven $\beta$-strands and two $\alpha$-helixes. The domain[10] is a $\beta$-barrel, a nearly perfect cylinder,

[7]The *Protein Data Bank* is an open database for the 3D structural data of large biological molecules, such as proteins and nucleic acids.

[8]*Protein primary structure* is the linear sequence of amino acids in a peptide or protein.

[9]The *secondary structure* is made of regularly repeating local structures stabilized by hydrogen bonds.

[10]A *protein domain* is a conserved part of a given protein sequence and tertiary structure that can evolve, function, and exist independently of the rest of the protein chain. Each domain forms a compact three-

42$\mathring{A}$[11] long and 24$\mathring{A}$ in diameter, creating what is referred to as a "beta-can" formation, which is unique to the GFP-like family. The first $\alpha$-helix is at the base of the barrel, the second is positioned along the central axis, inside the barrel structure, and contains the covalently bonded chromophore HBI (see fig. 2.4), responsible of the fluorescent emission. The barrel gives the HBI its stability and allows it to be able to be stitched together with proteins of interest. Five shorter alpha helices form caps on the ends of the structure.

Researchers have mutated the original fluorophore by changing the wavelength of light absorbed to include other colors of fluorescence. These variants are produced by the genetic engineering of the GFP gene. The *brainbows* (2.9 and 2.10) are now introduced as a very interesting technique exploiting 90 types of fluorescent proteins to trace individual neurons in complex dense network of the brain.

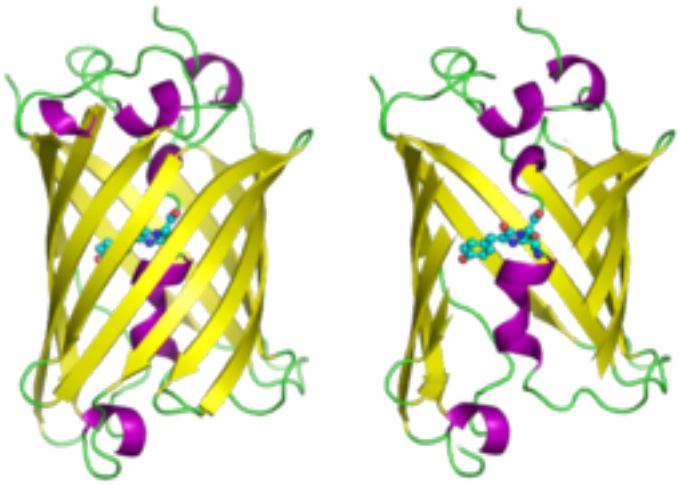

**Figure 2.3**. Green Fluorescent Protein (GFP)

dimensional structure and often can be independently stable and folded.

[11] 1 *angstrom* is equal to 0.1nm

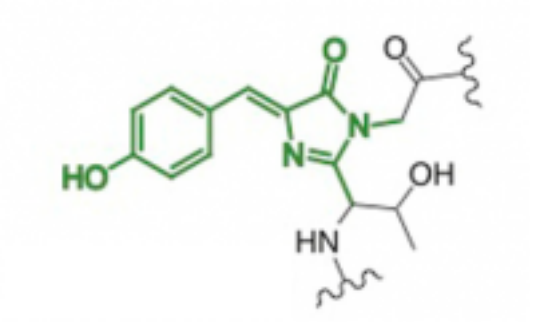


**Figure 2.4**. Chromophore of GFP

#### 2.2.2.3 Adeno-associated virus

A viral vector[12] can be used to express the GFP within the cell or organism of interest. The one used by the neurobiologists at the University of Utah is the Adeno-Associated Virus serotype 9.

Adeno-associated virus (AAV) is a small virus (25nm), belonging to the genus Dependoparvovirus, which infects humans and some other primate species. AAV is not currently known to cause disease and can infect both dividing and quiescent cells.

AAV is *replication-defective*, also called *helper dependent virus*: this means that an infection cannot occur except in the presence of a suitable helper virus. In other words it's dependent on the assistance of a helper virus in order to replicate, such as Adenovirus or Herpesvirus. The adeno-associated virus (AAV) was first identified as a dependovirus in the 1960s in the laboratories at National Institutes Health (NIH). The AAV genome is built of *single-stranded* deoxyribonucleic acid (ssDNA), either positive or negative-sensed, which is about 4.7 kilobase long. The genome comprises two genes: *Recombinase protein (Rep)* and *Capsid protein (Cap)*. The former encodes the proteins required for the AAV life cycle, including integration and replication, the latter contains proteins, which interact

[12] *Viral vectors* are aimed to deliver genetic material into cells. This process can be performed in vivo or in vitro. Viruses have evolved specialized molecular mechanisms to efficiently transport their genomes inside the cells they infect. All *viruses* bind to their hosts and introduce their genetic material into the host cell as part of their replication cycle. This genetic material contains basic 'instructions' of how to produce more copies of these viruses, hacking the body's normal production machinery to serve the needs of the virus. The host cell will carry out these instructions and produce additional copies of the virus, leading to more and more cells becoming infected. There are two main types of virus infection: lytic and lysogenic (2.5). Shortly after inserting its DNA, viruses of the lytic cycle quickly produce more viruses, burst from the cell and infect more cells. Lysogenic viruses integrate their DNA into the DNA of the host cell and may live in the body for many years before responding to a trigger. The virus reproduces as the cell does and does not inflict bodily harm until it is triggered. The trigger releases the DNA from that of the host and employs it to create new viruses.

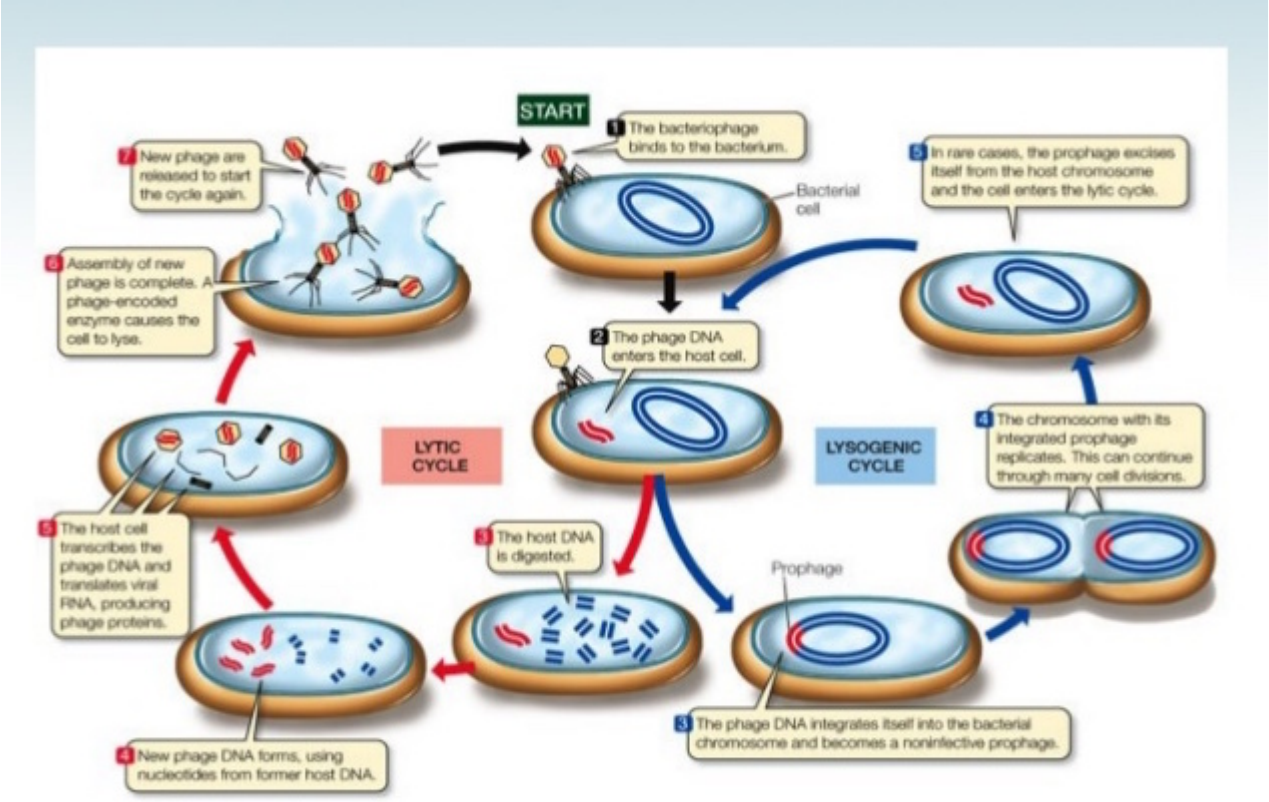


**Figure 2.5**. Lytic and lysogenic cycle.

together to form the icosahedral capsid[13]. AAV normally don't integrate into the genome of the host cell. However, in absence of the helper virus, AAV can install a latency state and integrate into the genome of the host cell in a specific site, thanks to the action of the recombinase protein Rep. The AAV remains quiescent until the presence of helper virus causes a super-infection. In case of co-infection of the host cell by the helper virus the replication process of the AAV starts through lytic cycle and virions[14] production.

Several serotypes of AAV have been isolated from various tissue samples. As of 2006 there have been 11 AAV serotypes described, the 11th in 2004. All of the known serotypes

[13]The *capsid* is the protein shell of the virus, protecting it from the external environment. Some viruses are *enveloped*, meaning that the capsid is coated with a lipid membrane.

[14]While not inside an infected cell or in the process of infecting a cell, viruses exist in the form of independent particles, also known as *virions*, consisting of two or three parts: (I) the genetic material made from either DNA or RNA, long molecules that carry genetic information; (II) a protein coat, called the capsid, which surrounds and protects the genetic material; and in some cases (III) an envelope of lipids that surrounds the protein coat when they are outside a cell.

can infect cells from multiple diverse tissue types. Tissue specificity is determined by the capsid serotype.

#### 2.2.2.4 Adeno-associated virus serotype 9

The AAV9 expresses[15] fluorescent proteins within the infected neurons. It's able to go through the blood-brain barriers and label the central nervous system (CNS). The virus has been genetically modified in the following way: some genes of the Cap and Rep have been removed, in order to avoid lysis and create space for an insert of exogenous DNA. The exogenous DNA eliminates the dipendency from the helper virus.

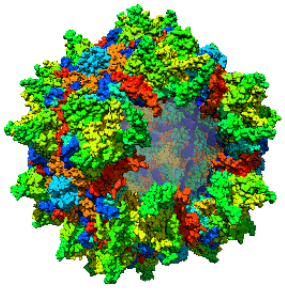

**Figure 2.6**. Adeno-Associated Virus serotype 9 (AAV9)

[15] see note 3

### 2.2.3 Microscopy

*Fluorescent microscope*[16] standard *confocal*2.8 [17], *two-photon*, or *light-sheet* imaging methods are all suitable to detect the fluorescence emitted down to the scale of protein localization, thus resulting in the final highly detailed and three-dimensional images.

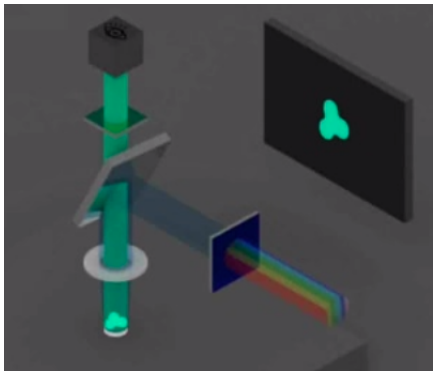

**Figure 2.7**. Fluorescent microscope schema

#### 2.2.3.1 Two-photon excitation microscopy

Two-photon excitation microscopy is a variant of confocal microscopy. The base idea the same: the light excites the sample that, in response, emits fluorescent light. The main issue in fluorescence microscopy is that fluorescence is emitted along the entire illuminated cone and not just at focus. This problem is solved in confocal microscopy through the use of a pinhole, able to select only the light coming from the target point in the sample.

[16] A fluorescence microscope is an optical microscope using fluorescent samples. It consists of magnifying lenses, a dichroic mirror and filters. The first filter selects the light which will excite the fluorophores contained in the sample, in 2.7 (from http://toutestquantique.fr/en/) blue. This blue light illuminates a large portion of the sample. When the fluorophores in the sample are illuminated with the proper wavelenght they emit a fluorescent light of another wavelenght. The dichroic mirror and the second filter select only this fluorescent light emitted by the sample. Therefore, the microscope detects the image of only the fluorescent part of the sample. Fluorescent microscopes allow to detect the presence and localization of a very small amount of molecules in biological samples.

[17] The confocal microscope is a special type of fluorescence microscope. Two pinholes are positioned at confocal positions. The light beam is focused by the first pinhole on only a small part of the sample. If fluorophores are present there, they emit light which is then filtered by the dichroic mirror and the filter. The second pinhole, positioned in the focal plane, selects only the light coming from the target point in the sample. The objective thus collects only this light. The surface of the sample is then scanned by moving either the sample or the light beam. This allows to reconstruct a 2D image at a given height. One can then move vertically to obtain images at different heights. A volume image of the sample can be reconstructed using proper software. Confocal microscopes allow to detect fluorescent molecules in 3D with a good spatial resolution.

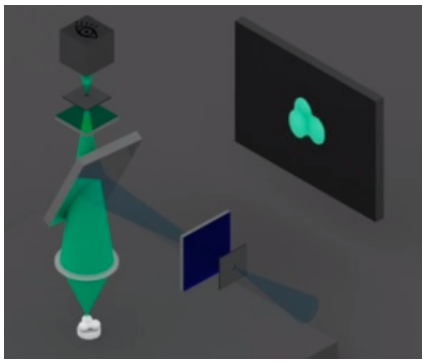

**Figure 2.8**. Confocal microscope schema

In confocal microscopy, the main issue in the imaging depth is the limited ability of the light to penetrate into the tissue and the phenomena of absorption and scattering, that prevent the light from being focused. It has been proved that imaging in the near-infrared minimises both absorption and scattering.

The idea behind two-photons microscopy is that two photons reach the electrons almost at the same time (they arrive within less then 1fs[18]): this is a fundamental requirement. Each photon carries approximately half the energy necessary to excite the molecule. So, for each excitation, two photons of infrared light are absorbed instead of one photon of blue light. This photons take an electron to the excited state.

The real strength of two-photon microscopy is that there is no out-of-focus light: the focus is the only point to emit light. For this reason, a two-photon microscope can achieve the same resolution and z-sectioning capability of a confocal microscope without the pinhole. At this point, in order to obtain the image, it suffices to assign to the focus - a pixel in the grid that forms the image - the global emission intensity coming from the sample when exciting the focus, without considering the paths of the light. So the image is created scanning point-by-point the sample and recording the intensity at each spot. To summarise, while the excitation part is almost identical in respect to confocal microscopy, the detection part is much simpler in two-photon microscopy, since the pinhole isn't needed and to generate the image it suffices to record the global emission intensity at each point.

Moreover, two-photons optics allows multiple channel to be collected simultaneously, since the same wavelength can excite multiple dyes. An important property is that the

[18] femtosecond

emission intensity depends on the square of excitation intensity. As a consequence, a high excitation, a high peak of power is needed, preferably against a normal average power. The solution resides in the use of a pulsed laser. A two-photon microscope is an expensive technology, since this kind of laser costs between 100 and 200 thousands dollars. To conclude, two-photon microscopy is ideal when working with slides of thickness 200nm-several mm. It can be applied to both living animals and fixed samples often treated with clearing techniques. The results are excellent images at cellular detail and high-resolution. [9]

### 2.2.4 Microscale imaging of circuits in clarity-treated primate visual cortex

- On the living neural tissue an injection of Adeno-Associated Virus serotype 9 is done. The virus starts to replicate within the cells and expresses a fluorescent protein. The aim of this step is to label the neurons and the blood vessels.

- Beginning from the postmortem tissue, CLARITY clearing is applied, to replace the lipid substrate with a gel and make the sample transparent.

- The tissue is sectioned through a *microtome*[19].

- Imaging is done through a two-photon microscope.

- The extreme-resolution 3D data are stored into a stack of images. Every image is a pixel grid.

- The work of neurobiologists is complete. The data are given to computer scientists to perform management, analysis and visualization.

### 2.2.5 Nanoscale imaging of brain circuits through Electron Microscopy

- To track individual neurons and to disambiguate wires traveling over long distances, the labeling is done by means of 90 different fluorescent colors which are random combinations of red, green and blue.

---

[19] *Microtomy* is a method for the preparation of thin sections for various kinds of materials. The microtome is the tool used to cut this thin slices, known as sections.

**Table 2.1**. From tissue sample to 3D image: example application[1]

| | |
|---|---|
| Genus | Macaque |
| Site | Projection neurons between visual cortical areas V1 and V2 |
| Diameter injection of AAV in V2 | 1mm |
| Volume of labeled axons in V1 | 60mm$^3$ |
| Z-resolution | 1$\mu$m |

| Volume size | 5mm$^3$ | 60mm$^3$ |
|---|---|---|
| Data generated | 130GB | 1.6TB |
| Time needed | 96hours | 19days |

- Achieving the finest resolution through Electron Microscopy requires imaging very thin slices of neural tissue. The *ultramicrotome*[20] developed at Harvard is called *Automatic Tape-Collecting Lathe Ultramicrotome* (ATLUM) [5] and is able to collect sections of lateral resolutions 5 nm or better. The thousands of sections are placed on a long carbon-coated tape.

- Imaging is done through *Scanning Electron Microscope* (SEM)[21].

- The images must be stitched together into very large images and stacked into volumes.

[20] An ultramicrotome is a microtome allowing the preparation of extremely thin sections.

[21] A scanning electron microscope consists in an electron source, electromagnetic lenses and an electron detector. It uses an electron beam instead of light, based on wave-particle duality. The electron beam is accelerated and focused on a sample using the lenses. The sample emits secondary electrons which are then detected. The number of detected electrons depends on the variations of the sample's surface. By scanning the beam and detecting the variation of the number of emitted electrons, one can reconstitute the surface topography. The electron beam can also ionize the atoms and make them emit X-rays. The ray's energy depends on the elementary composition of the sample. By scanning once again the beam and detecting the X-rays energy, one can deduce the chemical nature of the material and its spatial variation. Other types of interactions between the beam and the surface allow to perform various complementary analysis. The SEM thus allows to obtain a magnified image of the surface of thick samples and to analyze their composition.

- The final result is multicolor labeling. The obtained maps, the brainbows (2.9 [22]), are visualised through a streaming web application. Fig. 2.10 [5] is obtained tracking the neurons through the stack of slices and following each neuron's complex branching structure. The result are the shown treelike structures. An insight about neurotracking will be provided in the next section.

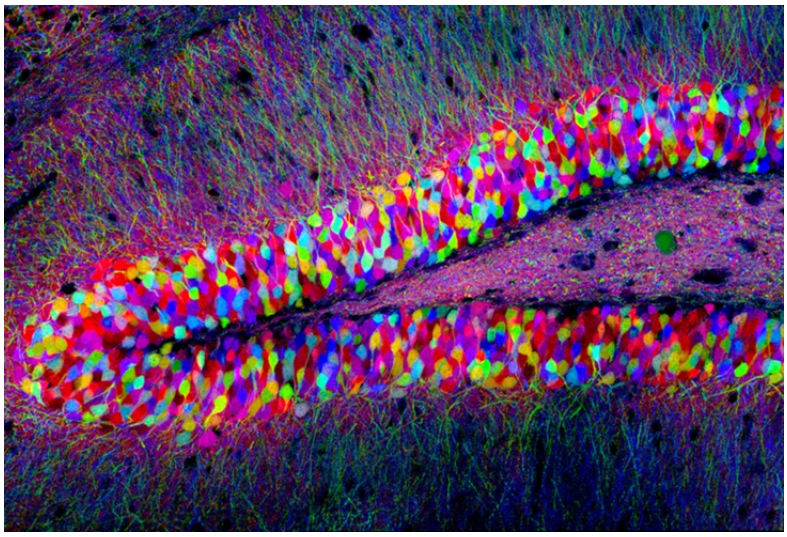

**Figure 2.9**. Brainbow, Harward University (1)

## 2.3 Applications

### 2.3.1 Anatomical and functional mapping

As previously said, the functions of the connectome circuits are being studied through functional MRI in the resting state and during tasks. To understand how neural structures result in specific functional behavior such as consciousness, it is necessary to build theories that relate functions to anatomical connectivity. The bond between structural and functional connectivity is not straightforward. Computational models of whole-brain network dynamics are valuable tools to investigate the role of the anatomical network in shaping functional connectivity. Computational models can also be used to predict the dynamic effect of lesions in the connectome.

[22] http://braintour.harvard.edu/archives/portfolio-items/brainbow

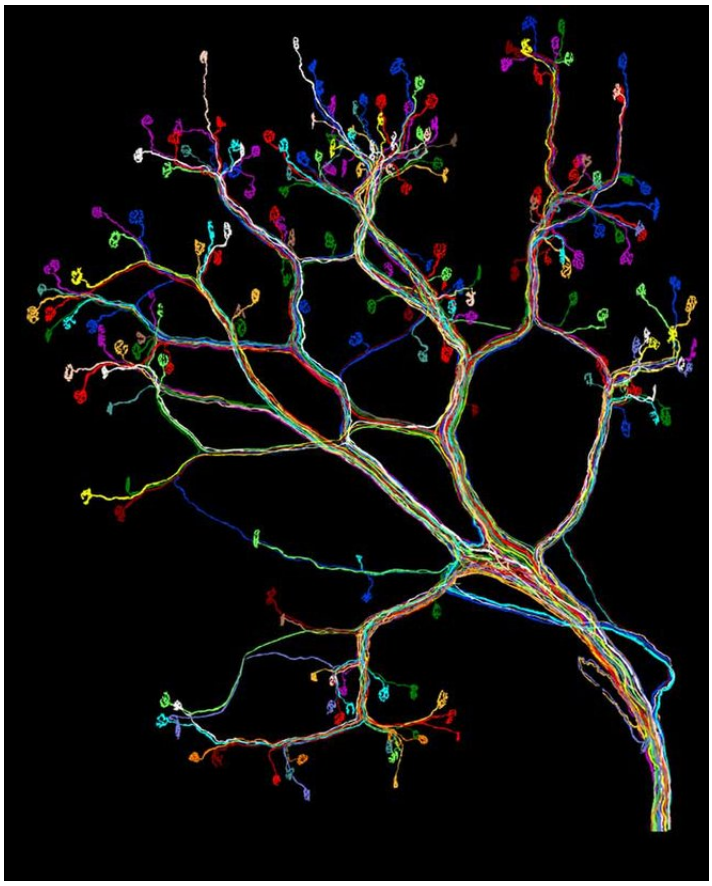

**Figure 2.10**. Brainbow, Harward University (2)

### 2.3.2 Neuro-tracker: automatic tracing of the connections

*Digital reconstruction* or tracing of 3D neuron structures is a critical activity in discovering the wiring and functions of the brain. It's challenging, especially when a 3D microscopic image has low *signal-to-noise ratio* (SNR) and fragmented neuron segments due to the intrinsic punctuated neurite structures (e.g. *synaptic boutons*) or imperfections in sample preparation. Notably such datasets are common for the nervous systems of different animals. Most of the existing methods have used various *structural components* (SC), e.g. 3D spheres, ellipsoids, cylinders, lines segments or irregular compartments, to model a neuron's morphology. The strategy is usually to build-up the reconstruction by incrementally adding more and more such SCs into the morphological modeling of a neuron. The present thesis suggests a similar approach based on the topology boundary and co-boundary operators. In fig. 2.11[8] is shown an exaple of neuron tracing and reconstruction from images through various SC examples. In (a) there's 3D reconstruction

of a fruit fly[23] neuron. The entire morphology model can typically be decomposed as individual segments, shown in different colors (b), which are connected at the branching points. Typically, each segment can be traced/reconstructed separately (c). In (d) is illustrated the modeling of image voxel information using a series of spherical SCs. The edge of image region (bright voxels) best matches to the aggregation of SCs. In (e) other types of SCs such as ellipsoids and cylinders, are used in locally matching the image content and thus growing the reconstruction. (f) shows a maximum intensity project of a 3D stack of a fruit fly lamina neuron. The neurites are highly punctuated, have high contrast in image intensity and appear to be broken. In (g) is presented a stained pyramidal neuron of a mouse brain region, where axonal varicosities make it hard to grow a reconstruction using local searching based on SCs.

### 2.3.3 Study of the brain as a network or graph

The ideal representation of the complex wiring biagram of the brain is in form of a large graph and the maps at different scales can be joint into a single hierarchical visualization of the neural organization of a given species that ranges from single neurons to populations of neurons to cortical areas. The connectome can be studied as a network by means of network science and graph theory. In case of a micro-scale connectome, the nodes of the graph are the neurons, and the edges correspond to the synapses between those neurons. For the macro-scale connectome, the nodes correspond to the regions of interest (ROIs), while the edges of the graph are derived from the axons interconnecting those areas. Thus connectomes are sometimes referred to as *brain graphs*, as they are indeed graphs in a mathematical sense which describe the connections in the brain or, in a broader sense, the whole nervous system. *Statistical graph theory* is an emerging discipline which is developing sophisticated pattern recognition and inference tools to parse these brain

[23]The Drosophila connectome, once completed, will be a complete list of the roughly 135000 neurons in the brain of the fruit fly *Drosophila melanogaster*, along with all of the synapses between these neurons. As of 2013, the Drosophila connectome is a work in progress. The choice of the fruit fly has important underlying reasons: on the one hand, researchers prefer an organism small enough that the connectome can be obtained in a reasonable amount of time. On the other hand, since one of the main uses of the connectome is to relate structure and behavior, an animal with a large behavioral repertoire is desirable. It's also very helpful to use an animal with a large existing community of experimentalists, and many available genetic tools. Surprisingly the fruit fly exhibits hundreds of different behaviors that have been qualitatively and quantitatively studied over the years and tens of thousands of genetic variants are available.

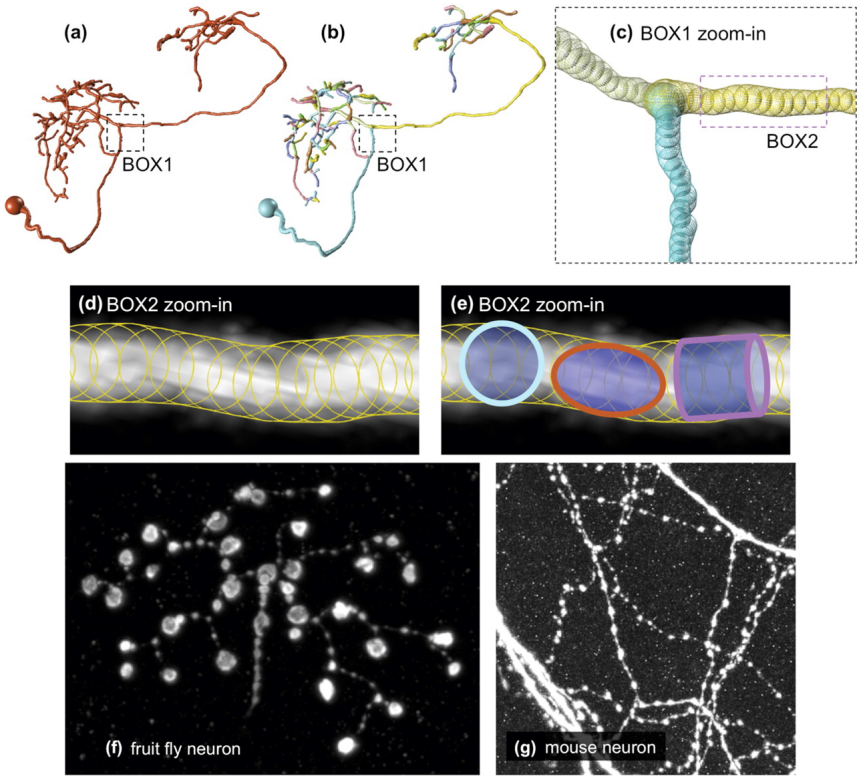


**Figure 2.11**. Automatic tracing of neurons.

graphs.

## 2.4 Limits of the approach

To summarize, these are the main challenges in the field of human connectomics:

- Large size and complexity of the primate brain, compared to mouse brain [1];
- Current non-invasive imaging techniques can't capture the brain's activity on a neuron-by-neuron level. Mapping the connectome at the cellular level in vertebrates currently requires post-mortem microscopic analysis of limited portions of brain tissue.
- State of the art in tools to annotate the data;
- Long images acquisition times. [24]
- Amount of the data;

[24] To address this issues, several groups are building high-throughput serial electron microscopes.

- Quality of the images: low signal-to-noise ratio (SNR) and fragmented neuron segments [8];

- State of the art in theory and algorithms for the analysis of the huge brain-graphs. In this field, statistical graph theory is an emerging discipline which is developing sophisticated pattern recognition and inference tools to parse these brain-graphs.

## 2.5 Summary and conclusions

Connectomics is the research field aimed to generate a hierarchical mapping of the brain at different scales, from the full set of neurons and synapses to a macro scale description of the functional and structural connectivity between cortical areas and subcortical structures. This is a multidisciplinary Big-Data application, that pushes the boundaries of our technologies and computational resources. Although we're still not able to generate a complete map of the human brain, several research groups from the most important academic institutions in the world are working on this problem. A strong collaboration between neurobiologists and computer scientists has been set up, in which neurobiologists obtain 3D images from neural tissue through complex histological and imaging techniques. Computer scientists work to manage, analyse and visualise the results. This thesis investigates two of the main activities that are needed to deal with the data provided by neurobiologists: the model extraction and the simplification.

## 2.6 References

[1] A. G. S. M. V. P. A. A. C. CHRISTENSEN, F. FEDERER, *Large scale imaging and 3d visualization of long-range circuits in clarity-treated primate visual cortex*.

[2] J. K. S. Y. K. S. A. A. S. D. T. J. M. J. J. Z. K. A. M. J. D. A. K. P. S. B. H. R. C. G. L. G. V. D. K. CHUNG, K.; WALLACE, *Structural and molecular interrogation of intact biological systems*, Nature, 497 (2013), pp. 332–337.

[3] G. DRUMMEN, *Fluorescent probes and fluorescence (microscopy) techniques - illuminating biological and biomedical research*, 17 (2012), pp. 14067–90.

[4] P. HAGMANN, *From diffusion MRI to brain connectomics*, thesis, EPFL, Lusanne, 2005.

[5] H. P. M. F. C. JEFF W. LICHTMAN, R. CLAY REID, *Discovering the wiring diagram of the brain*, (2009).

[6] J. LIPPINCOTT-SCHWARTZ, *Intracellular fluorescent imaging: An introduction*.

[7] R. K. O. Sporns, G. Tononi, *The human connectome: A structural description of the human brain*, PLoS Computational Biology.

[8] H. Peng, F. Long, and G. Myers, *Automatic 3d neuron tracing using all-path pruning*, Bioinformatics, 27 (2011), pp. i239–i247.

[9] K. Thorn, *Two photon microscopy*.

[10] X. Zhuang, *Super-resolution fluorescence microscopy*.