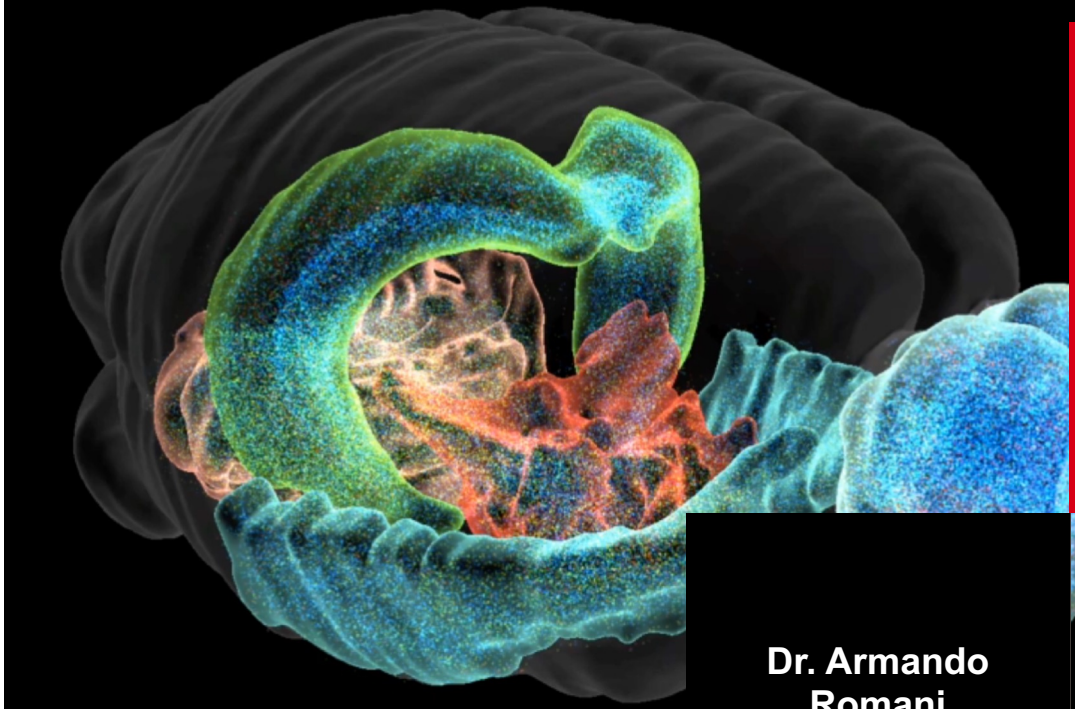


Computational neuroscience: biophysics - Lecture 2

Dr. Armando Romani

EPFL, 2024

- Blue Brain Project

Neuroinformatics: Data and Metadata

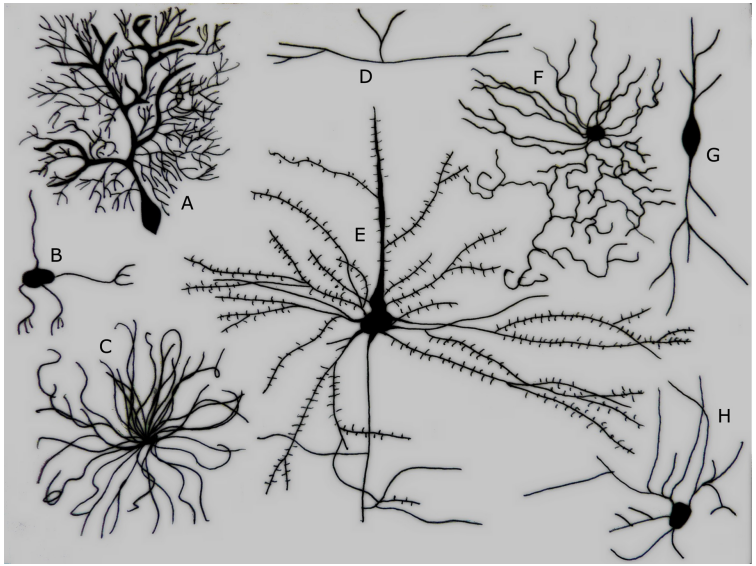
Lecture Overview

- Scope
- Approaches / Resources

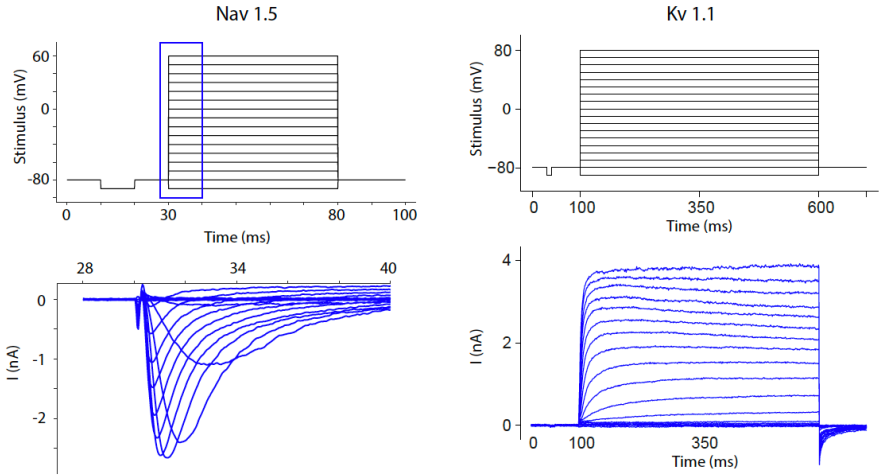
Lecture Overview

- **Scope**
- Approaches / Resources

Neuroscience Data

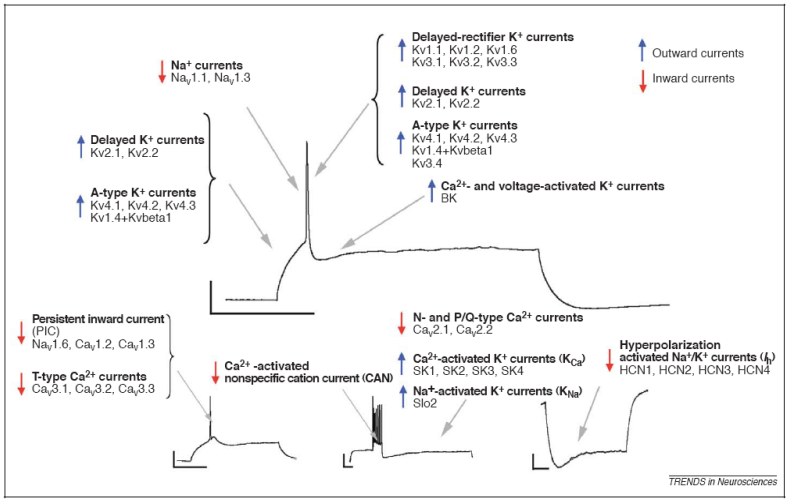


Morphology reconstructions – lecture 3



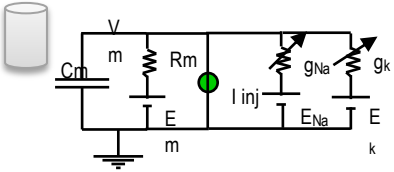
Single channel recordings – lecture 4

Neuroscience Data



Single cell recordings – lecture 4

Simplified Model: Single Compartment and ion channel formalism



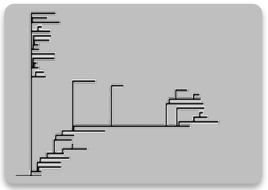
$$\frac{C_m dV_m}{dt} = \frac{E_m - V_m}{R_m} + I_{channels}$$

$$\frac{dm}{dt} = \alpha_m(V_m)(1-m) - \beta_m(V_m)m$$

$$\frac{dh}{dt} = \alpha_h(V_m)(1-h) - \beta_h(V_m)h$$

$$I_{channel} = m^n h g_{channel}(V_m - E_{channel})$$

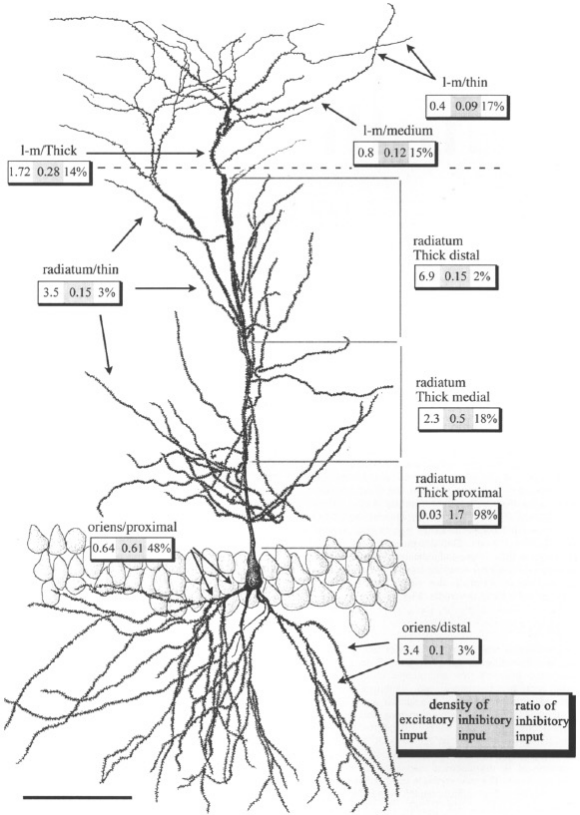
Cellular Model: Cable and ion channel formalism



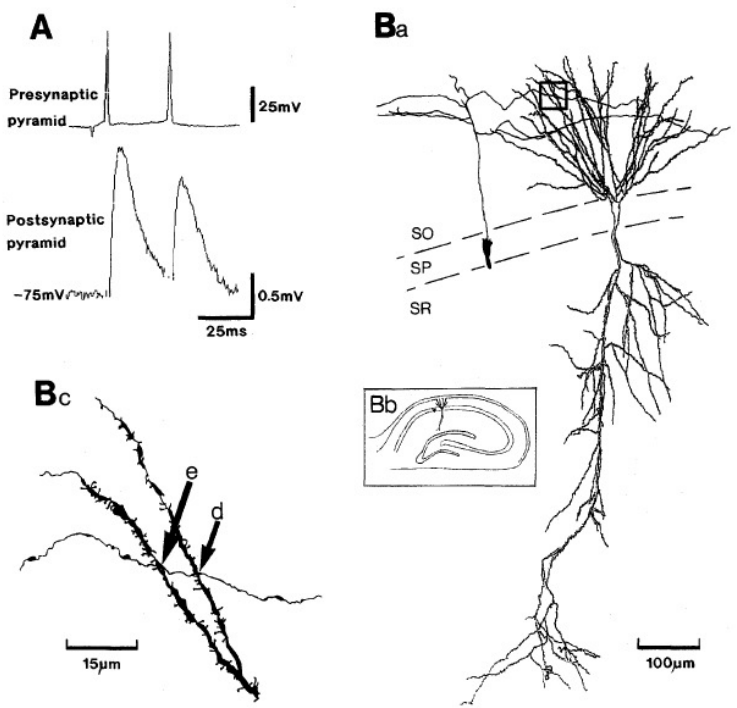
$$\frac{C_m dV_m}{dt} = \frac{E_m - V_m}{R_m} + I_{channels} + \frac{2(V_{m_{i+1}} - V_{m_i})}{R_{a_{i+1}} + R_a} + \frac{2(V_{m_{i-1}} - V_{m_i})}{R_{a_{i-1}} + R_a}$$

Computational models – lecture 5 and 6

Neuroscience Data

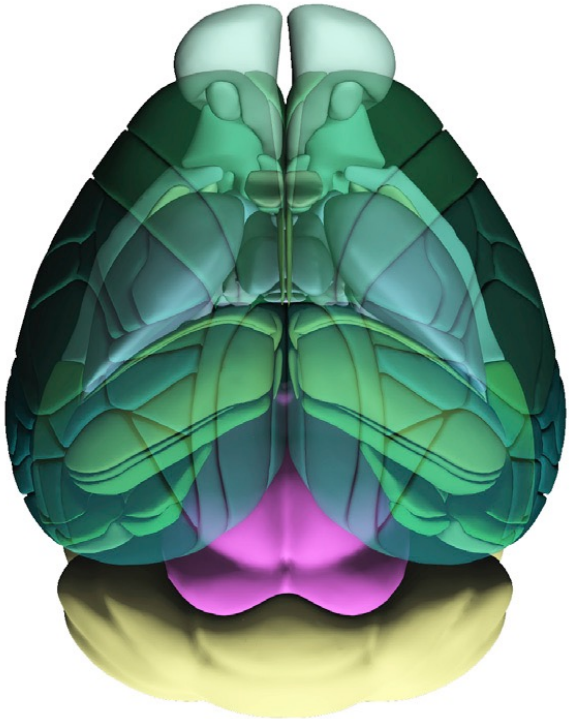


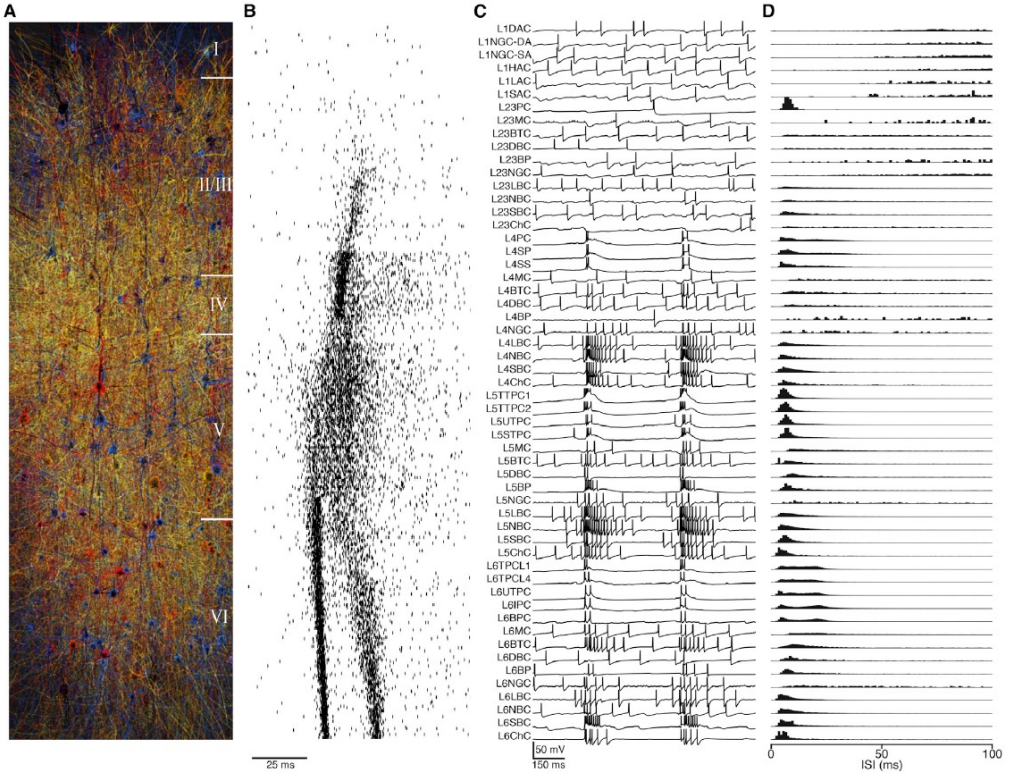
Number of synapses – lecture 7



Pair recordings – lectures 7 and 8

Neuroscience Data

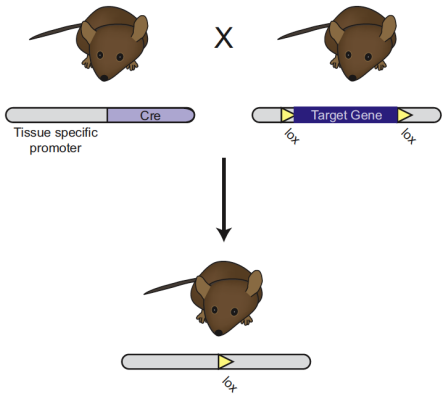


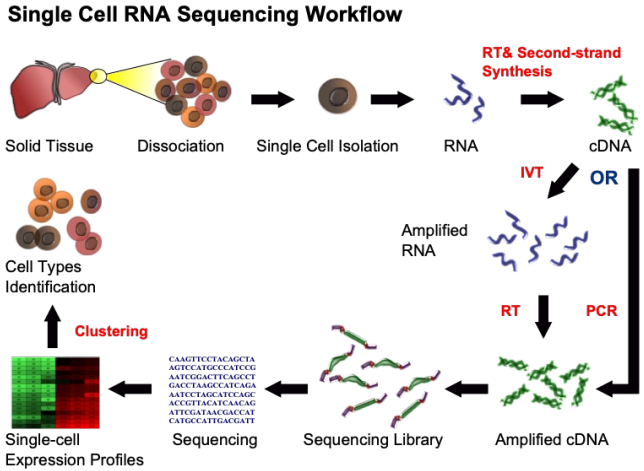


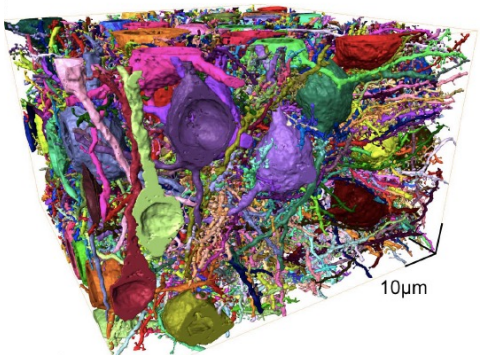
Atlas – lecture 9

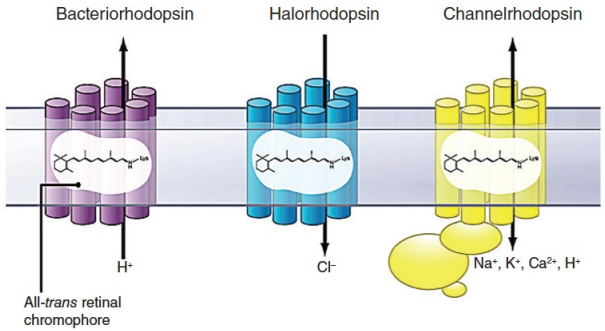
Network activity – lecture 10, 11, and 12

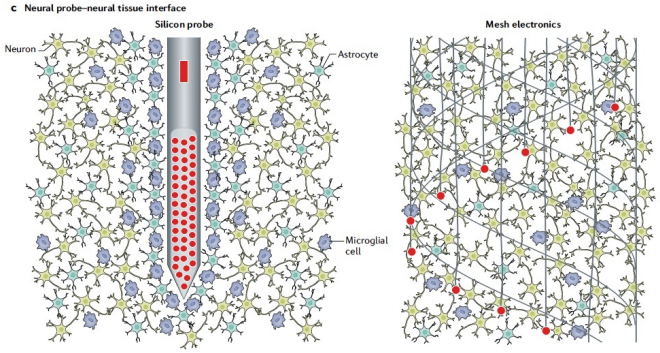
Neuroscience Data

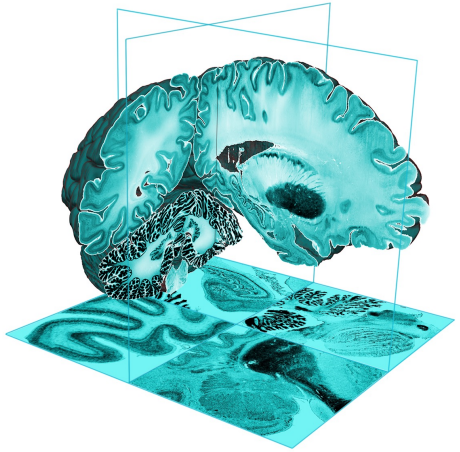












…and much more! – lecture 14

Neuroscience Data

- Experimental measure (raw or processed)
- Information extracted from literature (raw or processed)
- Model artifact

Neuroscience Data

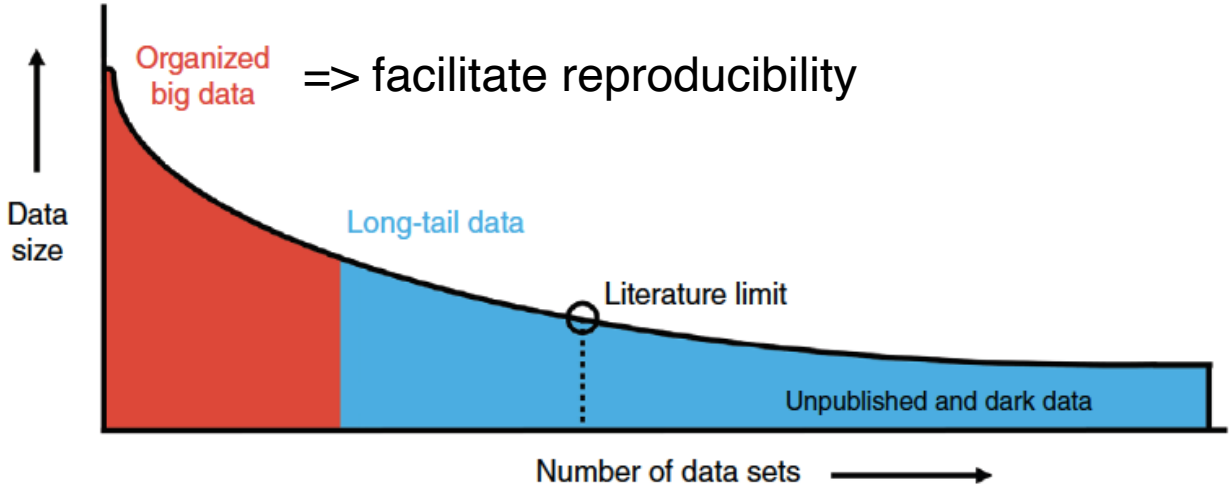


Figure 1 Schematic illustration of long-tail data. Studies that have plotted data set size against the number of data sources reliably uncover a skewed distribution. Well-organized big science efforts featuring homogenous, well-organized data represent only a small proportion of the total data collected by scientists. A very large proportion of scientific data falls in the long-tail of the distribution, with numerous small independent research efforts yielding a rich variety of specialty research data sets. The extreme right portion of the long tail includes data that are unpublished; such as siloed databases, null findings, laboratory notes, animal care records, etc. These dark data hold a potential wealth of knowledge but are often inaccessible to the outside world.

Ferguson et al., 2014

Problem

Reproducibility / reusability by the same person over time, by different people in the same lab or in other labs

hippo_planes_for_morph-collage.txt

-0.02014812020230947,0.6864996194426534,-0.7268509652999063,-1604.1923270574614 0.02014812020230947,-0.6864996194426534,0.7268509652999063,1654.1923270574614
-0.019954736969346867,0.7276227597727851,-0.685687193939861,-1685.5191394721055 0.019954736969346867,-0.7276227597727851,0.685687193939861,1735.5191394721055
-0.020011572342902184,0.7157871516529508,-0.6980317259988408,-1662.2457659248844 0.020011572342902184,-0.7157871516529508,0.6980317259988408,1712.2457659248844
-0.02006873300917579,0.7036735147228083,-0.7102399810156117,-1638.3429607097721 0.02006873300917579,-0.7036735147228083,0.7102399810156117,1688.3429607097721

Based on memory, filename, and content:

- Pairs of planes
- A,B,C,D coefficients of the planes
- 101 equally-spaced planes

But…

- How was it produced?
- Who made it?
- The file is not archived
- Multiple version?

Lecture Overview

- Scope
- **Approaches / Resources**

Data Life Cycle

Griffin et al., 2018

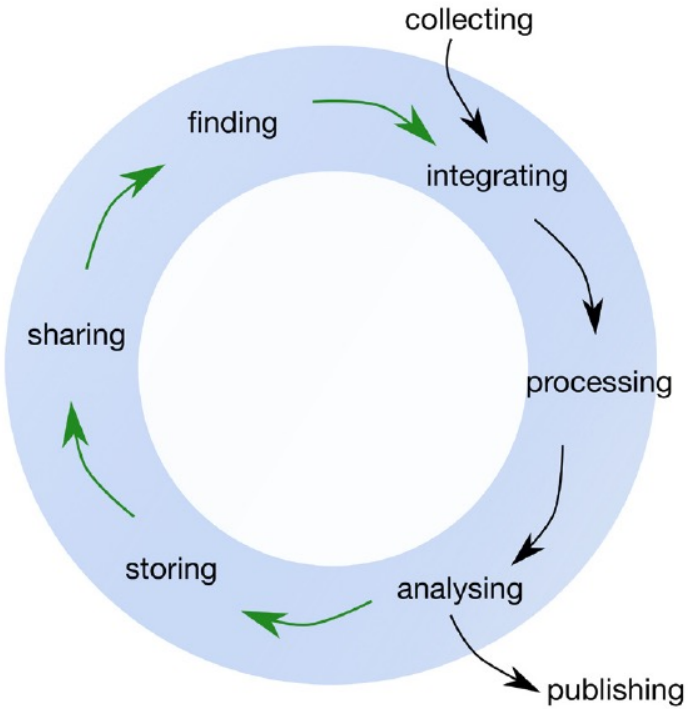
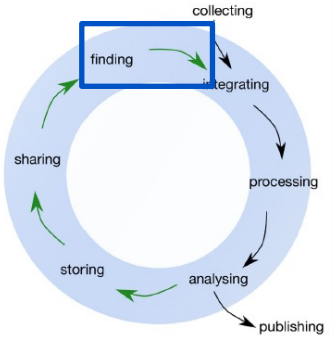


Figure 1. The Data Life Cycle framework for bioscience, biomedical and bioinformatics data that is discussed throughout this article. Black arrows indicate the 'traditional', linear view of research data; the green arrows show the steps necessary for data reusability. This framework is likely to be a simplified representation of any given research project, and in practice there would be numerous 'feedback loops' and revisiting of previous stages. In addition, the publishing stage can occur at several points in the data life cycle.

EPFL

Finding

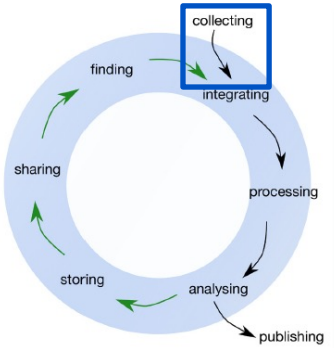


- Poor findability: supplementary material to articles, on personal or institutional websites, or in non-discipline-specific repositories
- High findability: public data repository
- Poor quality of the user-submitted data in public repositories
- Public datasets often require extra curation before reuse (e.g. NeuroMorpho.org)
- The results of extra curation may not find their way back into the repositories
- Repositories are often not easily searched by generic web search engines
- Registries, which form a secondary layer linking multiple, primary repositories, may offer a more convenient way to search across multiple repositories

EPFL

Legend: challenge, solution

Collecting

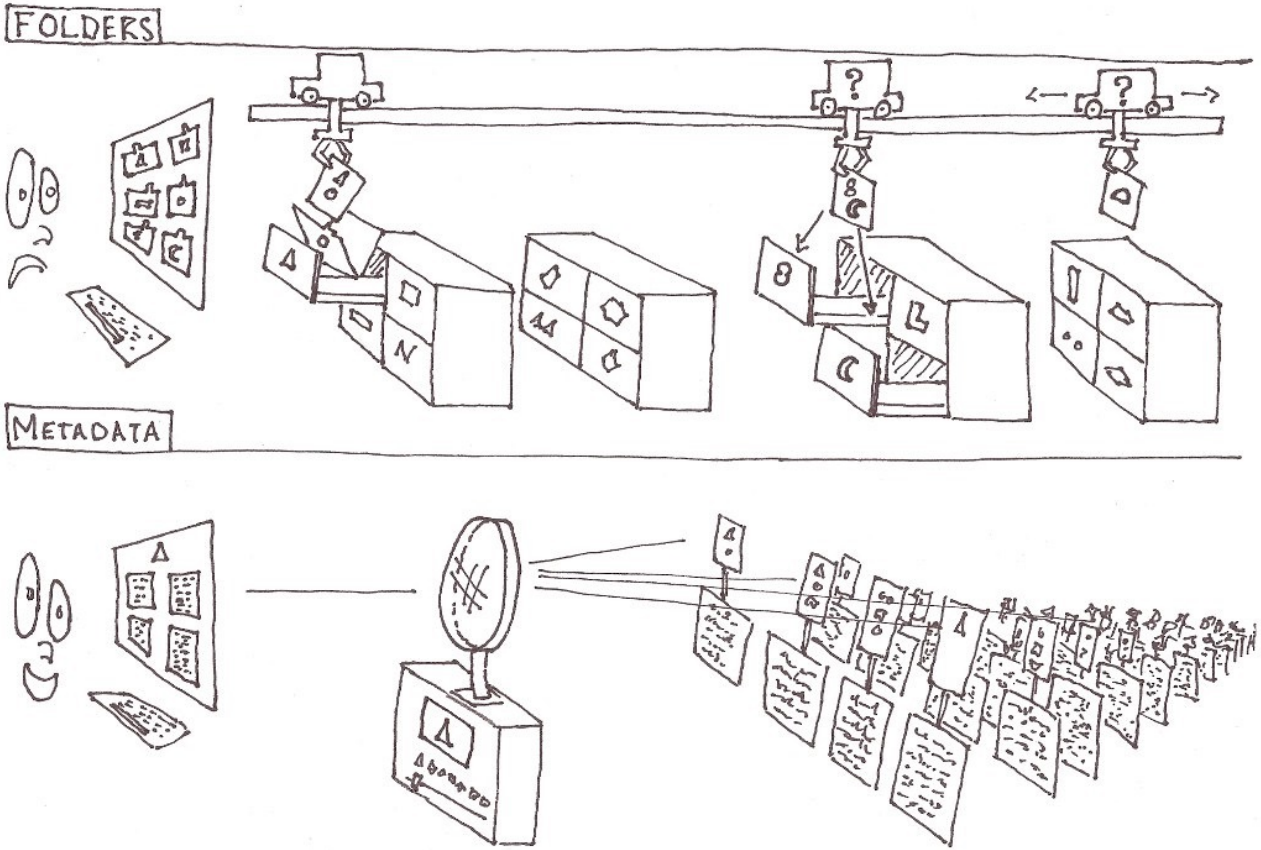


Search for specific data

- Metadata
- Provenance
- Ontology
- Standard vocabulary

Legend: challenge, solution

Folders Vs Metadata



Metadata: data about data

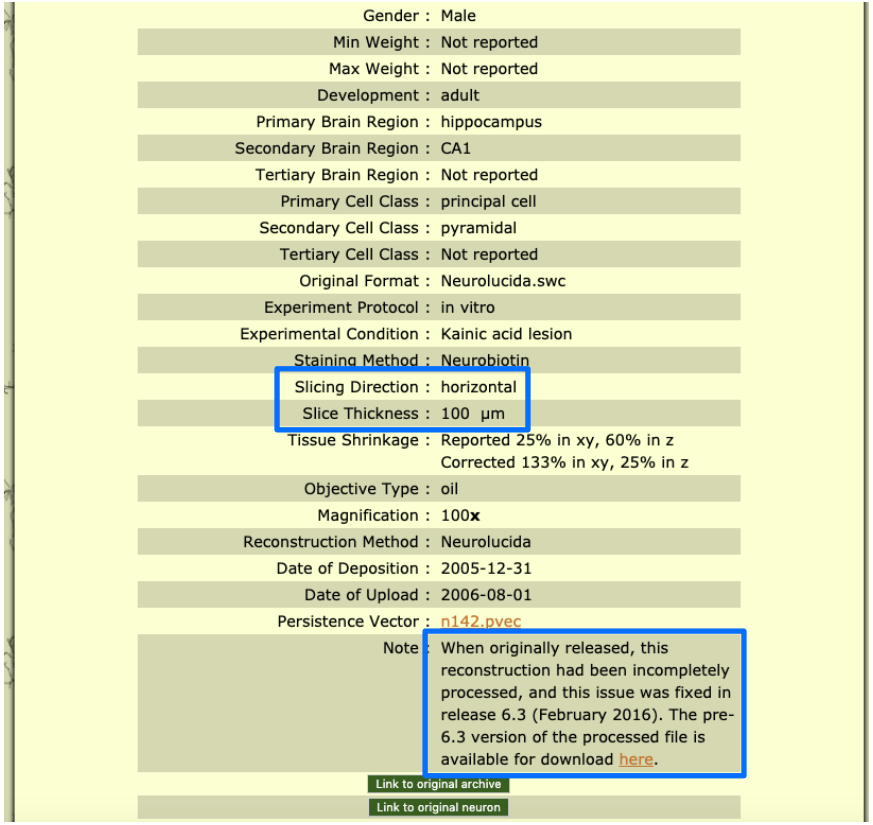
Provenance

Metadata about the origin and transformations of the data

- Who produces the data?
- Is this derived data? – If so, from which dataset?
- What processes / analyses were used to produce / transform the data?
- Which software and version?

Provenance


NeuroMorpho.Org
Version 7.9 - Released: 12/13/2019 - Content: 121544 neurons
HOME
BROWSE
SEARCH
LITERATURE COVERAGE
TERMS OF USE
HELP
0
Morphology File (Standardized)
Morphology File (Original)
Log File (Standardized)
Log File (Original)
Get above files zipped
3D Neuron Viewer - Java, legacy
3D Neuron Viewer - WebGL, novel
Animation
Details about selected neuron
NeuroMorpho.Org ID : NMO_00132
Neuron Name : n142
Archive Name : Turner
Species Name : rat
Strain : Fischer 344
Structural Domains : Dendrites, Soma, Axon
Physical Integrity : Dendrites Complete, Axon Incomplete
Morphological Attributes : Diameter, 3D, Angles
Min Age : Not reported
Max Age : Not reported


Gender : Male
Min Weight : Not reported
Max Weight : Not reported
Development : adult
Primary Brain Region : hippocampus
Secondary Brain Region : CA1
Tertiary Brain Region : Not reported
Primary Cell Class : principal cell
Secondary Cell Class : pyramidal
Tertiary Cell Class : Not reported
Original Format : Neurolucida.swc
Experiment Protocol : in vitro
Experimental Condition : Kainic acid lesion
Staining Method : Neurobiotin
Slicing Direction : horizontal
Slice Thickness : 100 µm
Tissue Shrinkage : Reported 25% in xy, 60% in z
Corrected 133% in xy, 25% in z
Objective Type : oil
Magnification : 100x
Reconstruction Method : Neurolucida
Date of Deposition : 2005-12-31
Date of Upload : 2006-08-01
Persistence Vector : n142.pvec
Note : When originally released, this reconstruction had been incompletely processed, and this issue was fixed in release 6.3 (February 2016). The pre-6.3 version of the processed file is available for download here.
Link to original archive
Link to original neuron

Metadata

- Metadata is what makes the data searchable and useful
- Metadata is not always sufficient to a given scope
- The scope changes as a function of the project, but also of the time
- Describing a piece of data for complete reproducibility is challenging and requires much effort
- Attempts have been made to request a minimum set of information to the data producer

Minimal Information for Neuroscience Datasets (MINDS)

- **Subject**: age, sex, species, strain (when applicable),
- **Methods**, protocols, equipment and parameters used in experiments
- **Classification** data category, data format, cell type
- **Brain location** using brain atlas parcellation or spatial coordinates/transform,
- **Contributors** and their **affiliations**,
- Access to **the data** and the format via persistent identifier URL (e.g. DOI)

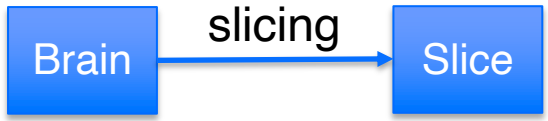
Hill, 2016

Metadata

If metadata is well structured, uses consistent element names and contains element values with specific descriptions from agreed upon vocabularies, it enables machine readability, aggregation, integration and tracking across datasets […]. One key approach in best practice metadata collection is to use ***controlled vocabularies*** *built from* ***ontology*** *terms.* (Griffin et al., 2018)

Ontology

- Ontology is a way of organizing and defining concepts and the relationships between them
- It provides machine-interpretable representations of some aspect of biological reality
- Sourcing metadata element values from ontologies ensures that the terms used in metadata are consistent and clearly defined
- define provenance rules



Standard vocabulary

CA1 Trilaminar

Name		Supertype
	CA1 (i)0113p-SUB_111 Trilaminar	421:Schaffer Collateral-Related:SCR RO-Targeting-III

Synonym(s)
CA1 double projecting cell
CA1 double projection cell
CA1 double-projection cell
CA1 double-projection neuron
CA1 hippocampo-subicular projection cell
CA1 horizontal trilaminar cell
CA1 horizontal trilaminar interneuron
CA1 horizontal trilaminar neuron
CA1 nonpyramidal projection neuron
CA1 oriens retrohippocampal projection cell
CA1 oriens-retrohippocampal cell
CA1 oriens-retrohippocampal nonpyramidal neuron
CA1 oriens-retrohippocampal projection neuron
CA1 septal-projecting GABAergic cell
CA1 TL cell
CA1 trilaminar cell
CA1 trilaminar interneuron

EPFL

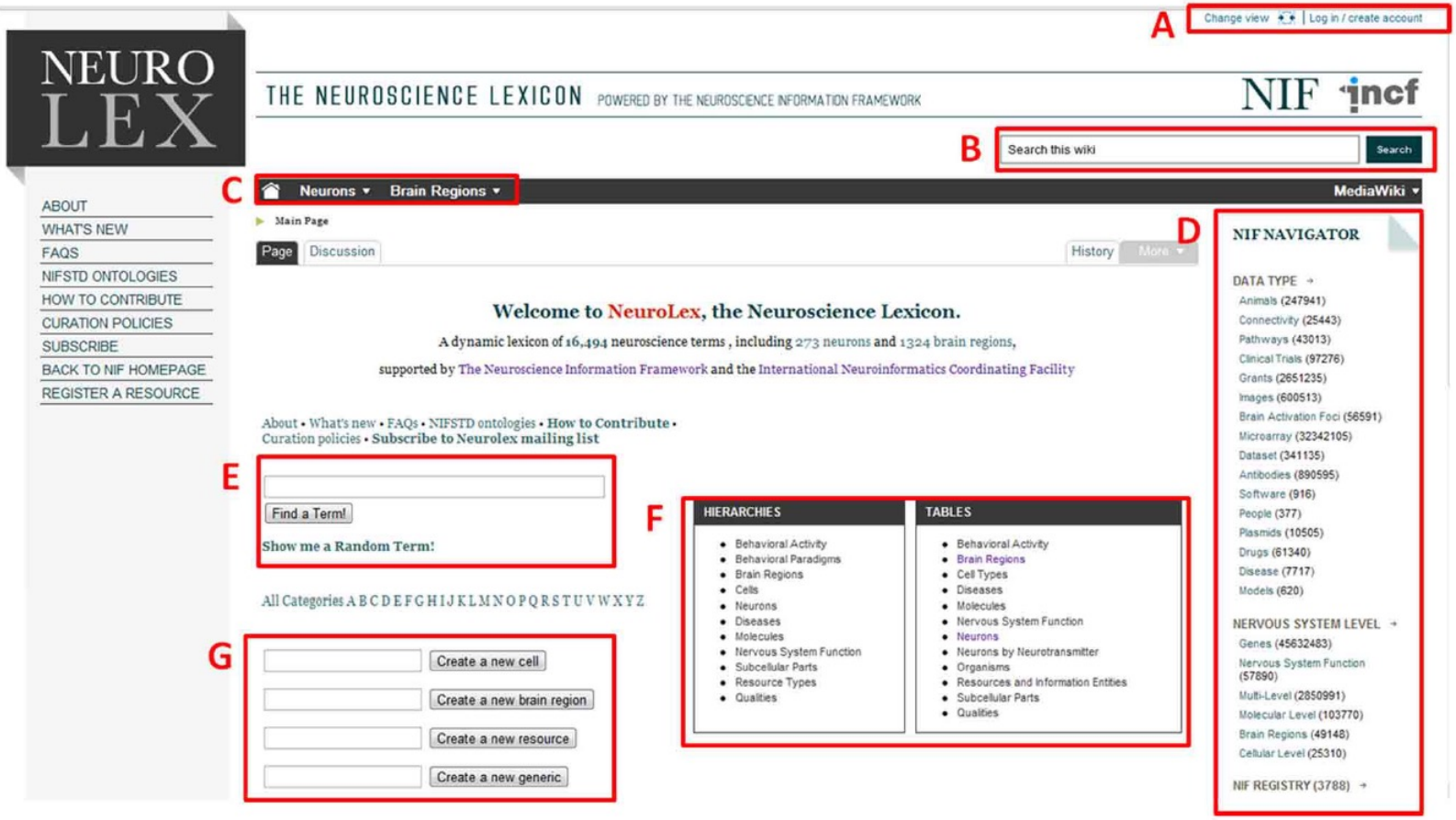
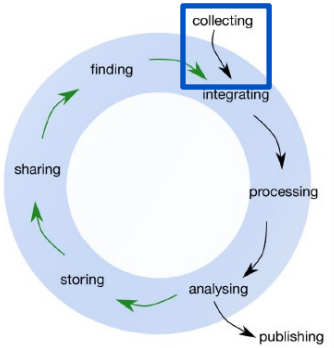


FIGURE 1 | Landing page for NeuroLex.org. Several features are highlighted. **(A)** Login/user management controls. **(B)** Global site search bar. **(C)** Quick navigation to neuron or brain region information. **(D)** NIF Navigator, connecting the Neuroscience Information Framework B's federated resources to each NeuroLex page. **(E)** Global site search bar. **(F)** Quick navigation to hierarchies or tables containing detailed information about diverse entities in Neuroscience. **(G)** Quick creation forms for cells, brain regions, resources, and generic page contents.

Standardization

- Standardization does not apply only to vocabulary
- Reproducibility also requires standard formats and
- Standard tools

Collect data: manual vs. automatic approaches



Manual approaches

- Expert-based
- They tend to be of higher quality, but they produce small datasets

Automatic approaches

- Text mining, automatic classification, curation…
- They can introduce thresholds to keep a certain level of quality. However, they tend to incorporate data of different quality, but they can digest most of the available data

EPFL

Manual approaches: Bezaire and Soltesz, 2013

HIPPOCAMPUS 23:751–785 (2013)

Quantitative Assessment of CA1 Local Circuits: Knowledge Base for Interneuron-Pyramidal Cell Connectivity

Marianne J. Bezaire,* and Ivan Soltesz

Bezaire and Soltesz, 2013

Method

- Combine selected pieces of information to estimate the numbers of the main cell types in the hippocampus CA1
- Use of assumptions and generalizations to fill the gaps (see part 2 of the lecture)

TABLE 4.

Estimated Number of Each Type of Interneuron

			Layer			
Interneuron type	Fraction (%)	Total	SO	SP	SR	SLM
Neurogliaform family	**32.2**	**12,390**	**980**	**5,410**	**3,030**	**2,970**
Ivy	22.9	8,810	980	5,410	2,420	0
Neurogliaform	9.3	3,580	0	0	610	2,970
SOM expressing	**9.3**	**3,580**	**3,580**	**0**	**0**	**0**
O-LM	4.3	1,640	1,640	0	0	0
Double projection	2.0	760	760	0	0	0
CB- septal proj.	0.5	190	190	0	0	0
Oriens-retrohipp.	1.7	640	640	0	0	0
Other SOM+ cells[a]	0.9	350	350	0	0	0
PV expressing	**23.9**	**9,210**	**2,200**	**6,460**	**550**	**0**
PV+ Basket	14.4	5,530	1,320	3,880	330	0
Bistratified	5.7	2,210	530	1,550	130	0
Axo-axonic	3.8	1,470	350	1,030	90	0
CCK expressing	**13.9**	**5,370**	**1,140**	**1,070**	**1,960**	**1,200**
CCK+ Basket	9.4	3,600	780	940	1,170	710
ADI	1.0	390	0	0	390	0
SCA	1.0	400	0	0	400	0
PPA	1.3	490	0	0	0	490
CCK Misc.	1.3	490	360	130	0	0
Interneuron-specific	**19.4**	**7,470**	**780**	**3,190**	**1,450**	**2,050**
IS I	11.0	4,250	780	1,800	780	890
IS II	5.1	1,970	0	480	450	1,040
IS III	3.2	1,250	0	910	220	120
Other interneurons	**1.2**	**480**				
Total interneurons	**100**	**38,500**	**8,680**	**16,130**	**6,990**	**6,220**

Automatic approaches: Attili et al. 2022

Received: 21 July 2021 | Revised: 25 January 2022 | Accepted: 28 February 2022

DOI: 10.1111/ejn.15639

RESEARCH REPORT

EJN European Journal of Neuroscience FENS WILEY

Quantification of neuron types in the rodent hippocampal formation by data mining and numerical optimization

Sarojini M. Attili[1] | Keivan Moradi[1] | Diek W. Wheeler[2] | Giorgio A. Ascoli[1,2]

EPFL

WELCOME TO THE HIPPOCAMPOME PORTAL

v1.12 - Released: 05/05/2022
527,802 Pieces of Knowledge (PoK) & 46,004 Pieces of Evidence (PoE)

The Hippocampome is a curated knowledge base of the circuitry of the hippocampus of normal adult, or adolescent, rodents at the mesoscopic level of neuronal types. Knowledge concerning dentate gyrus, CA3, CA2, CA1, subiculum, and entorhinal cortex is distilled from published evidence and is continuously updated as new information becomes available. Each reported neuronal property is documented with a pointer to, and excerpt from, relevant published evidence, such as citation quotes or illustrations.

The goal of the Hippocampome is dense coverage of available data characterizing neuronal types. The Hippocampome is a public and free resource for the neuroscience community, and the knowledge is presented for user-friendly browsing and searching and for machine-readable downloading.

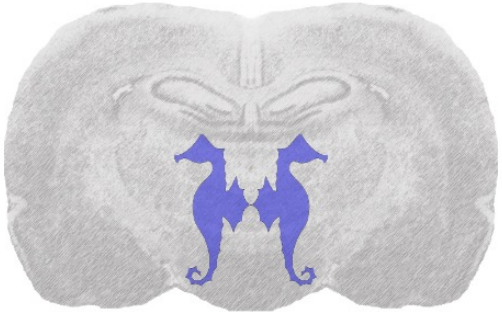
If you have feedback on either functionality or content, or if you would like to be informed when the first official version is released, please fill out the feedback form or email us at Hippocampome.org@gmail.com.

The release of v1.0 on 09/23/2015 includes 3,697 PoK and 13,888 PoE. Reference: **Wheeler et al., 2015**. *Hippocampome.org: A knowledge base of neuron types in the rodent hippocampus*. eLife 2015;4:e09960.

The release of v1.1 on 10/17/2016 includes 3,697 PoK, 13,888 PoE, and the Neuron Term Portal, which allows one to view definitions for terms and phrases used on this website. Reference: **Hamilton et al., 2016** *Name-calling in the hippocampus (and beyond): coming to terms with neuron types and properties*. Brain Informatics 2017 Mar;4(1):1-12; doi:10.1007/s40708-016-0053-3.

The release of v1.2 on 11/06/2016 includes 3,936 PoK, 14,399 PoE, and a clickable Connectivity Matrix, an interactive Connectivity Navigator, and the ability to Search by Connectivity. Reference: **Rees et al., 2016** *Graph theoretic and motif analyses of the hippocampal neuron type potential connectome*. eNeuro Nov 2016, ENEURO.0205-16.2016; DOI: 10.1523/ENEURO.0205-16.2016.

The release of v1.3 on 06/30/2017 includes 10,822 PoK, 21,285 PoE, and a downloadable list of Allen Brain Atlas (ABA) predictions of marker expressions and a utility for viewing the effects of thresholds on ABA marker expression predictions. Reference: **Hamilton et al., 2017** *Molecular fingerprinting of principal neurons in the rodent hippocampus: a neuroinformatics approach.* Journal of Pharmaceutical and Biomedical Analysis 2017 Sep 10;144:269-278; doi: 10.1016/j.jpba.2017.03.062.



The release of v1.7 on 10/08/2019 includes Izhikevich models for most of the neuron types, downloadable parameter and CARLSim4-simulation files, and the ability to perform simulations of the firing patterns. Reference: **Venkadesh, et al., 2019** *Simple models of quantitative firing phenotypes in hippocampal neurons: comprehensive coverage of intrinsic diversity.* PLOS Computational Biology 2019 Oct 28;15(10):e1007462; doi: 10.1371/journal.pcbi.1007462.

The release of v1.8 on 12/28/2020 includes an additional 5,152 PoK, 6,055 PoE, and new browsable matrices for neurite lengths, somatic path distances, numbers of potential synapses, numbers of contacts, and connection probabilities. Reference: **Tecuatl, et al., 2020** *Comprehensive estimates of potential synaptic connections in local circuits of the rodent hippocampal formation by axonal-dendritic overlap.* Journal of Neuroscience 2021 Feb 24;41(8):1665-1683; doi: 10.1523/JNEUROSCI.1193-20.2020.

The release of v1.9 on 02/27/2021 includes an additional 91 PoK, 768 PoE, and a new browsable matrix for in vivo recordings. Reference: **Sanchez-Aguilera, et al., 2021** *An update to Hippocampome.org by integrating single-cell phenotypes with circuit function in vivo.* PLoS Biology 2021 May 6;19(5):e3001213; doi: 10.1371/journal.pbio.3001213.

The release of v1.10 on 08/03/2021 includes the Cognome, a literature review and knowledge base of spiking neural circuit and network simulations of the hippocampal formation. Reference: **Sutton and Ascoli, 2021** *Spiking neural networks*

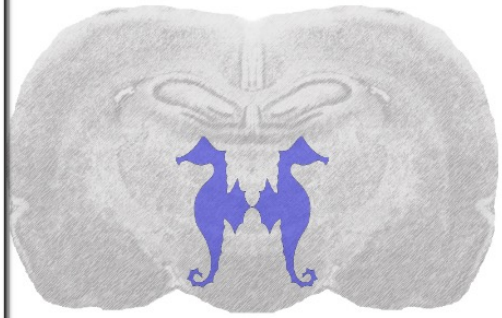
WELCOME TO THE HIPPOCAMPOME P

v1.12 - Released: 05/05/2022
527,802 Pieces of Knowledge (PoK) & 46,004 Pieces o

The Hippocampome is a curated knowledge base of th
hippocampus of normal adult, or adolescent, rodents a
level of neuronal types. Knowledge concerning dentate
CA2, CA1, subiculum, and entorhinal cortex is distilled
evidence and is continuously updated as new informat
available. Each reported neuronal property is documer
pointer to, and excerpt from, relevant published evide
citation quotes or illustrations.

The goal of the Hippocampome is dense coverage of a
characterizing neuronal types. The Hippocampome is a
resource for the neuroscience community, and the kno
presented for user-friendly browsing and searching an
readable downloading.

If you have feedback on either functionality or content
like to be informed when the first official version is rele
out the feedback form or email us at Hippocampome.org@gmail.com.



The release of v1.0 on 09/23/2015 includes 3,697 PoK and 13,888 PoE. Reference: **Wheeler et al., 2015**. *Hippocampome.org: A knowledge base of neuron types in the rodent hippocampus*. eLife 2015;4:e09960.

The release of v1.1 on 10/17/2016 includes 3,697 PoK, 13,888 PoE, and the Neuron Term Portal, which allows one to view definitions for terms and phrases used on this website. Reference: **Hamilton et al., 2016** *Name-calling in the hippocampus (and beyond): coming to terms with neuron types and properties*. Brain Informatics 2017 Mar;4(1):1-12; doi:10.1007/s40708-016-0053-3.

The release of v1.2 on 11/06/2016 includes 3,936 PoK, 14,399 PoE, and a clickable Connectivity Matrix, an interactive Connectivity Navigator, and the ability to Search by Connectivity. Reference: **Rees et al., 2016** *Graph theoretic and motif analyses of the hippocampal neuron type potential connectome*. eNeuro Nov 2016, ENEURO.0205-16.2016; DOI: 10.1523/ENEURO.0205-16.2016.

The release of v1.3 on 06/30/2017 includes 10,822 PoK, 21,285 PoE, and a downloadable list of Allen Brain Atlas (ABA) predictions of marker expressions and a utility for viewing the effects of thresholds on ABA marker expression predictions. Reference: **Hamilton et al., 2017** *Molecular fingerprinting of principal neurons in the rodent hippocampus: a neuroinformatics approach.* Journal of Pharmaceutical and Biomedical Analysis 2017 Sep 10;144:269-278; doi: 03.062.

The release of v1.7 on 10/08/2019 includes Izhikevich models for most of the neuron types, downloadable parameter and CARLSim4-simulation files, and the ability to perform simulations of the firing patterns. Reference: **Venkadesh, et al., 2019** *Simple models of quantitative firing phenotypes in hippocampal neurons: comprehensive coverage of intrinsic diversity.* PLOS Computational Biology 2019 Oct 28;15(10):e1007462; doi: 10.1371/journal.pcbi.1007462.

The release of v1.8 on 12/28/2020 includes an additional 5,152 PoK, 6,055 PoE, and new browsable matrices for neurite lengths, somatic path distances, numbers of potential synapses, numbers of contacts, and connection probabilities. Reference: **Tecuatl, et al., 2020** *Comprehensive estimates of potential synaptic connections in local circuits of the rodent hippocampal formation by axonal-dendritic overlap.* Journal of Neuroscience 2021 Feb 24;41(8):1665-1683; doi: 10.1523/JNEUROSCI.1193-20.2020.

The release of v1.9 on 02/27/2021 includes an additional 91 PoK, 768 PoE, and a new browsable matrix for in vivo recordings. Reference: **Sanchez-Aguilera, et al., 2021** *An update to Hippocampome.org by integrating single-cell phenotypes with circuit function in vivo.* PLoS Biology 2021 May 6;19(5):e3001213; doi: 10.1371/journal.pbio.3001213.

The release of v1.10 on 08/03/2021 includes the Cognome, a literature review and knowledge base of spiking neural circuit and network simulations of the hippocampal formation. Reference: **Sutton and Ascoli, 2021** *Spiking neural networks*

Browse neuron type census matrix

	Neuron Type	Rats	Mice
DG(18)	Granule	962200	394502
	Hilar Ectopic Granule	90	37
	Semilunar Granule	163633	67090
	Mossy	25573	10485
	Mossy MOLDEN	5206	2134
	AIPRIM	8559	3509
	DG Axo-axonic	235	96
	DG Basket	1612	661
	DG Basket CCK+	128	52
	HICAP	833	342
	HIPP	6006	2462
	HIPROM	10	4
	MOCAP	9067	3717
	MOLAX	5421	2223
	MOPP	377	155
	DG Neurogliaform	599	246
	Outer Molecular Layer	10	4
	Total Molecular Layer	7989	3275
CA3(25)	CA3 Pyramidal	183844	75376
	CA3c Pyramidal	44123	18090

Legend: +/green: Excitatory -/red: Inhibitory

Attili et al., 2022: method

- we assigned a variable for each neuron type
- in order to account for information regarding neuronal groups in certain layers that did not correspond to any Hippocampome.org neuron type, we added nine ancillary variables
- divisive scaling factor of 0.41 to convert mouse data to rat
- Numerical densities were converted into neuronal numbers
- Relationships are converted in equations

EPFL

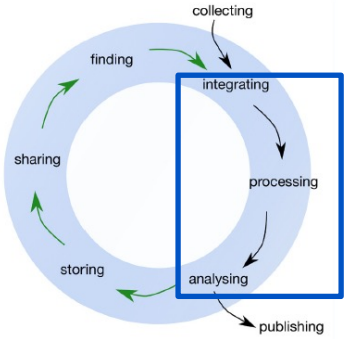
Equation	$x_{69}+x_{72}+x_{75}+x_{79}+x_{81}+x_{84}+x_{86}+x_{87}+x_{91}+x_{94}+x_{96}+x_{97}+x_{100}+x_{114}+x_{131}=324{,}000$	$\frac{x_6}{x_8}=\frac{31}{30}$	$x_7+x_8\le 3{,}680$ $x_7+x_8+x_{17}+x_{21}+x_{22}\ge 3{,}680$
Normalized form	$(x_{69}+x_{72}+x_{75}+x_{79}+x_{81}+x_{84}+x_{86}+x_{87}+x_{91}+x_{94}+x_{96}+x_{97}+x_{100}+x_{114}+x_{131})/324{,}000-1=0$	$\frac{30x_6}{31x_8}-1=0$	$\frac{x_7+x_8}{3{,}680}-1\le 0$ $\frac{x_7+x_8+x_{17}+x_{21}+x_{22}}{3{,}680}-1\ge 0$
Least squares form	$((x_{69}+x_{72}+x_{75}+x_{79}+x_{81}+x_{84}+x_{86}+x_{87}+x_{91}+x_{94}+x_{96}+x_{97}+x_{100}+x_{114}+x_{131})/324{,}000-1)^2=0$	$\left(\frac{30x_6}{31x_8}-1\right)^2=0$	$max\left(0,\frac{x_7+x_8}{3{,}680}-1\right)^2=0$ $max\left(0,1-\frac{x_7+x_8+x_{17}+x_{21}+x_{22}}{3680}\right)^2=0$
With weights	$r=10\times((x_{69}+x_{72}+x_{75}+x_{79}+x_{81}+x_{84}+x_{86}+x_{87}+x_{91}+x_{94}+x_{96}+x_{97}+x_{100}+x_{114}+x_{131})/324{,}000-1)^2$	$r=1\left(\frac{30x_6}{31x_8}-1\right)^2$	$r=1\times max\left(0,\frac{x_7+x_8}{3680}-1\right)^2$ $r=1\times max(0,1-(x_7+x_8+x_{17}+x_{21}+x_{22})/3680)^2$

$$r=\sum_i^n w_i\left(\left(z_i^{type}-z_i^{data}\right)/z_i^{data}\right)^2,$$

overall residual to be minimized by numerical optimization

Attili et al., 2022

Integrating, processing, and analysing



- Data are often heterogeneous
- Convert data into the same format and granularity
- Uniform metadata (Interoperability)
- Recording and reporting how data is processed and analysed
- Full reproducibility requires access to the software, software versions, workflow, dependencies and operating system used as well as the data and software code itself
- Hosting source code in a repository where it receives a unique identifier and is under version control (e.g. GitHub)

Legend: challenge, solution

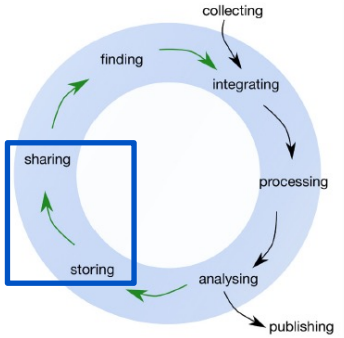
Publishing

- Journals increasingly recommend or require deposition of raw data in a public repository
- The vast majority of scientific journals require inclusion of processing and analysis methods in ‘sufficient detail for reproduction’
- ‘data journals’: scientific journals that publish papers describing datasets. This gives authors a vehicle to accrue citations for data production alone, which is typically not well recognized under the traditional publishing model
- Open Access Publishing
- Publications have unique ID (PubMed ID, DOI) as well as researchers (ORCID)

EPFL

Legend: challenge, solution

Storing, sharing



- Public repositories
- Local data storage solutions during the processing and analysis stages
- Some domains require large storage resources
- Human data require privacy regulations

Legend: challenge, solution

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

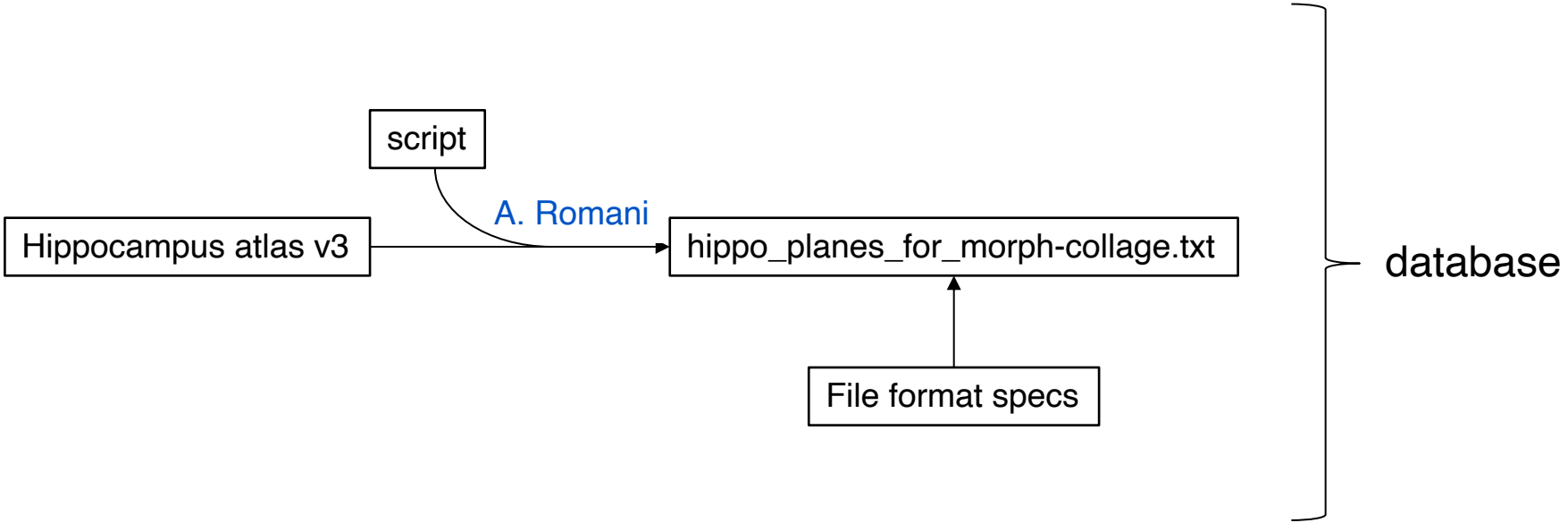
To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

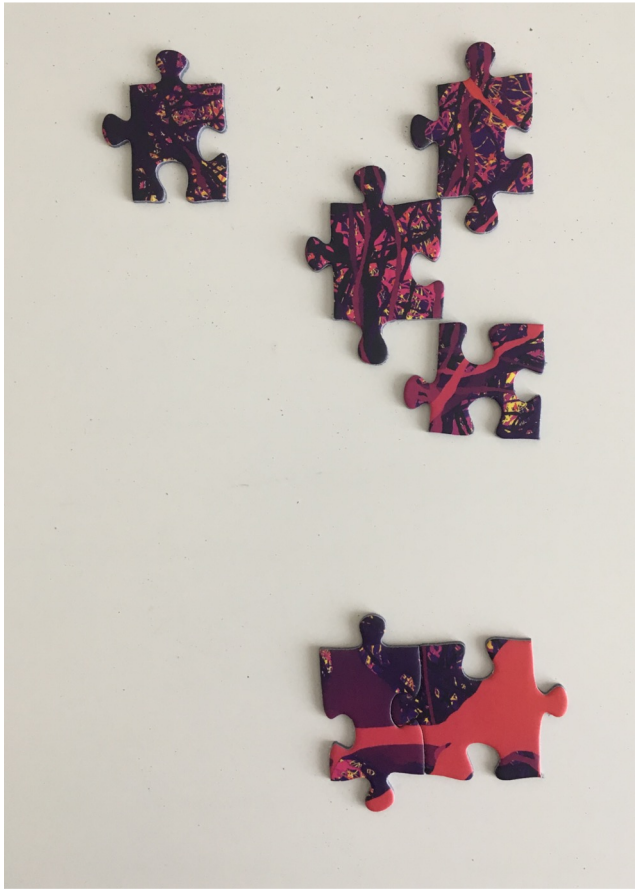
Wilkinson et al., 2016

It's time to fix my problem...



What you would like

The reality



How to overcome sparsity and reproducibility of data?

Examples of some issues with existing data

Cell densities

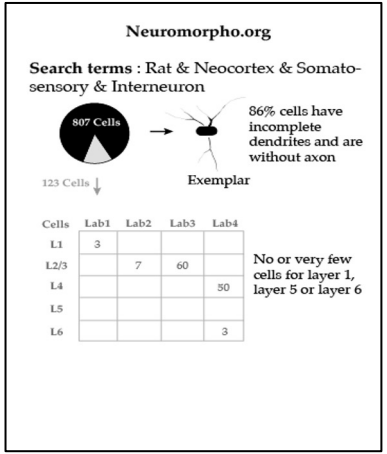
Reference	Species (region, age)	Microcircuit cell count	Microcircuit cell density/mm³
Knox, 1982	Rat (V1, adult)	~15,000	~40,000
Peters et al. 1987	Rat (V1, adult)	~15,000	~75,000
Beaulieu 1992	Rat (mPfC, adult)	~5,000	~36,400
DeFelipe et al. 2002	Rat (S1, adult)	~8,000	~5,000
Garcia et al. 2006	Rat (S1, p14)	~7,000	~52,000
Meyer et al. 2010	Rat (S1, p27)	~20,000	~82,000

~4x variability

~2.5x variability

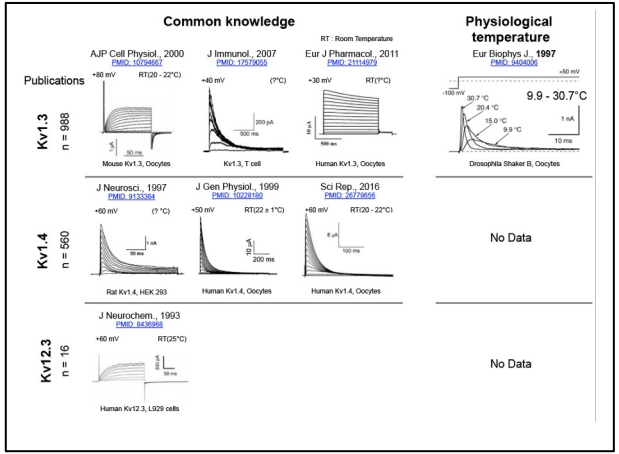
Numbers in literature can **vary** up to **4 fold**

Morphologies



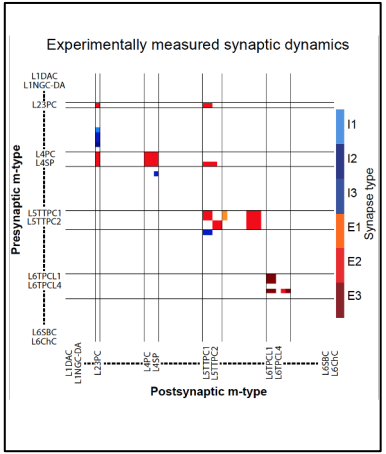
Usually data is **incomplete or missing**

Ion channel kinetics



Data often is **inconsistent** and raw data largely not publicly **available**

Synaptic physiology



Very few pathways have been characterized. Data can be very **sparse**

What you have learnt

- Data is not enough, it is important to have metadata.
- Rich metadata and standardizations enable reproducibility.
- MINDS.
- Ontology, provenance, standard vocabulary…
- FAIR principles.
- Lifecycle of the data.
- Data sparseness.