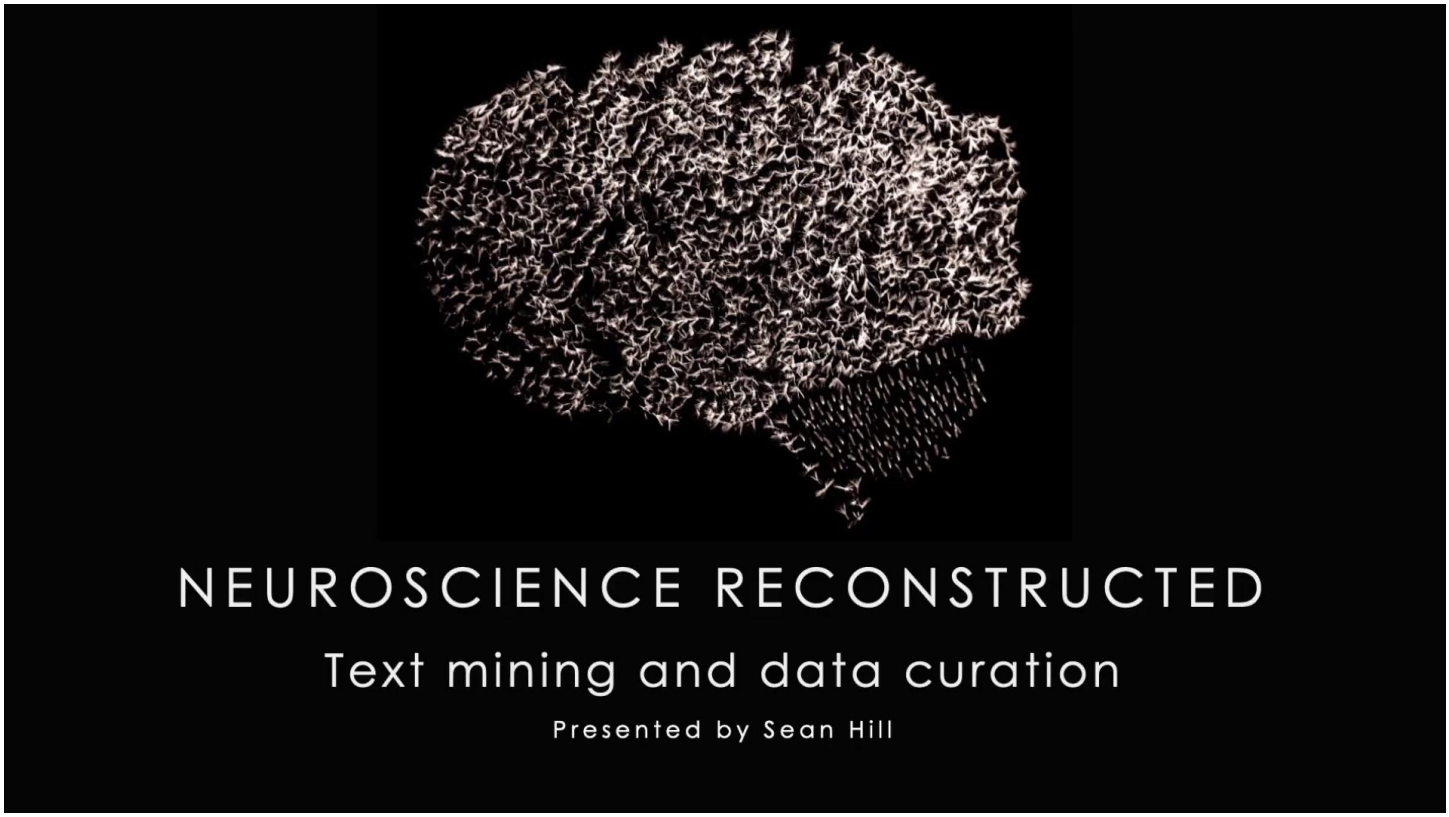
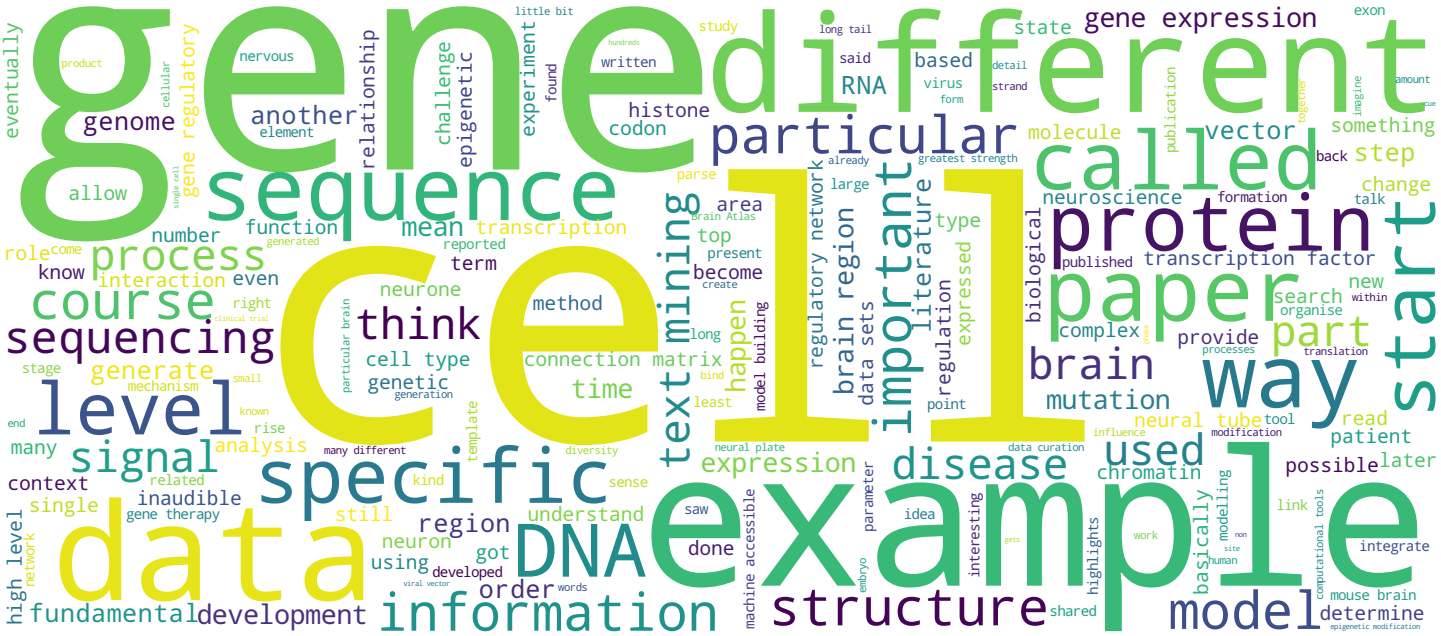




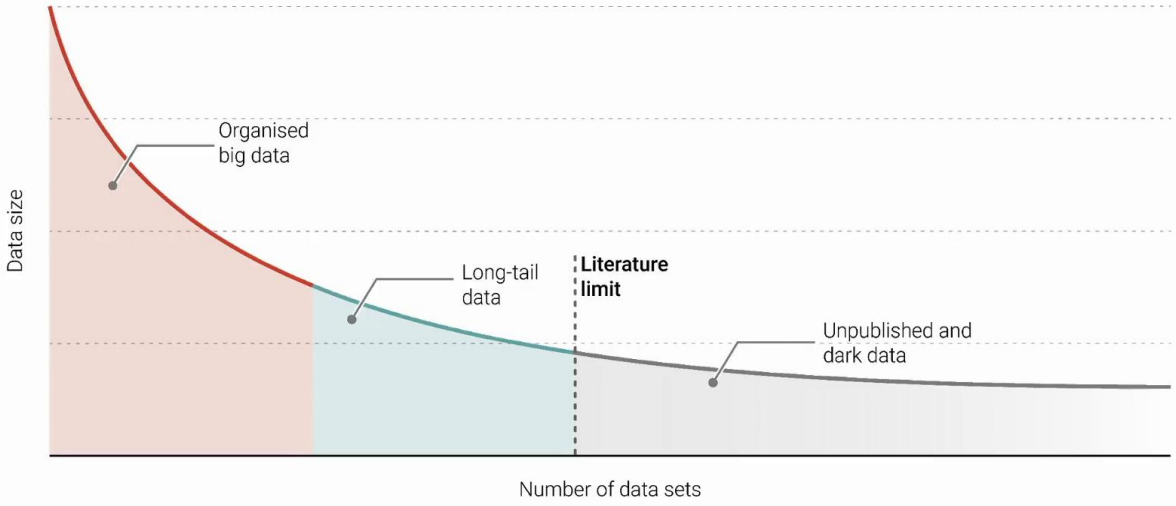
Search MOOC

Video

EPFL

Where to find data?

- In **organized big data sets** in databanks
  - Machine-accessible
  - Computational tools needed for data interpretation
- In **printed documents or unpublished** (long-tail data)
  - Not structured in a machine-accessible way
  - O(100k) publications per year in neuroscience



Adapted from Ferguson AR, Nielson JL, Cragin MH, Bandrowski AE & Martone ME. Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nature Neuroscience 17, 1442–1447 (2014)*. doi:10.1038/nn.3838

In organising big data sets in data banks, we've got to make the data machine accessible, and we need the computational tools for data interpretation. But in printed documents or unpublished documents, data is just written. It's put into tables, into a PDF, or it's written in words. It's not structured in a machine accessible way that you could use in an analysis or a model building pipeline. There are about 100,000 publications every year in neuroscience. A lot of the key data is actually locked up in these online publications. At the same time, if we think of the entirety of data, so there are some large data sets that are that are very large, well-organised, but this long tail of data in this publication really highlights the challenge of most data is actually, in this long tail, and that includes the data that is published just through figures and descriptions in papers. But there's also this unpublished dark data where somebody may say something qualitative in a paper. There's underlying data to that, but it's never really published or shared. This is a major challenge in neuroscience. Happily, there are a number of initiatives from funders to encourage, and in many cases now, start to require data to be shared and reused outside of the context of just a paper.

Notes

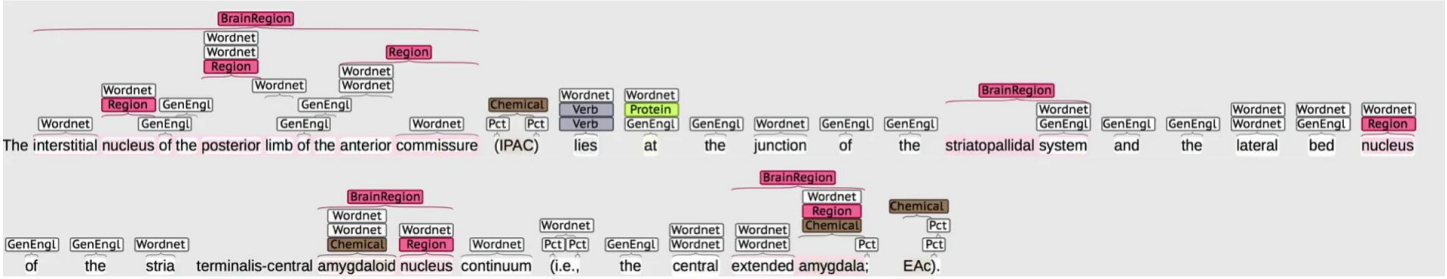
Summary

0m 05s

Text mining

Text mining tools can help find valuable data "hidden" in publications.

- Make the data accessible for search, analysis and the integration of such data into computational models.
- Need for manual review and verification
- No standardization, no centralization of the data



Renaud Richardet

But for those papers, since there's a huge history of literature and there are many more papers still being published, text mining tools can provide a way to find potentially, to find data hidden in those publications and make that data accessible for search and analysis that can be used for computational models. It can help accelerate the review of literature. I think that's really where text mining has its greatest strength. When you've got to look at what is known about a particular brain area, for example, you can do a much faster survey, high-level survey by using text mining. Of course, one of the challenges is that there's really no standardization in the way that the data is actually described in the text. When we're doing text mining, part of what you have to do is you have to actually define, for example, in this case, named entity recognition to go in and parse the sentence into its components, recognising which terms, for example, are proteins, what are the relationships to other types of entities called maybe brain regions, and parse out of all of these sentences meaning, essentially, what's the relationship? Is there a positive statement being said about in this brain region?

Notes

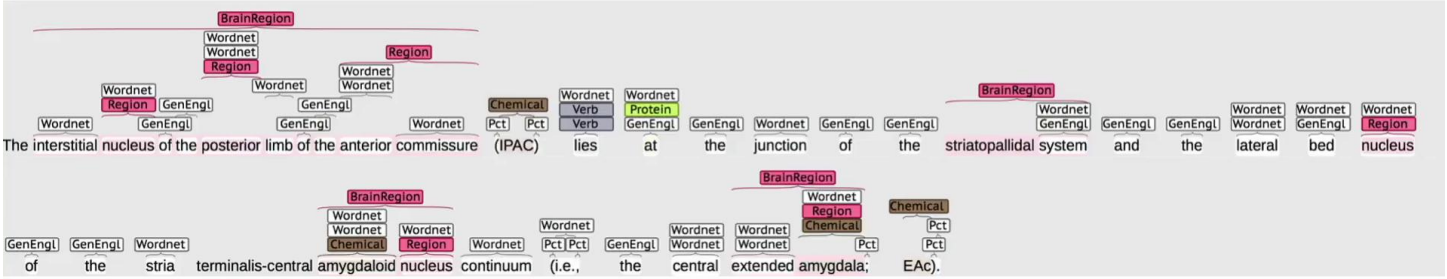
Summary

1m 51s

Text mining

Text mining tools can help find valuable data "hidden" in publications.

- Make the data accessible for search, analysis and the integration of such data into computational models.
- Need for manual review and verification
- No standardization, no centralization of the data



Renaud Richardet

This protein is expressed. When we do that, you can start to actually find, at least broadly, useful information from the literature. I think that, again, the greatest strength of this is in reviewing the literature rather than finding exact data just because of the variability in how scientists express themselves.

Notes

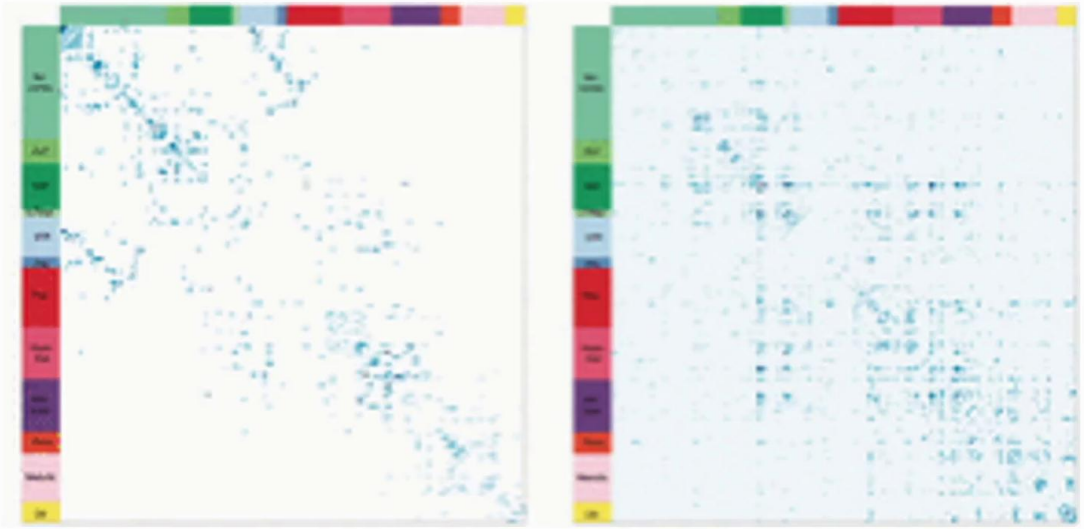
Summary

3m 27s

Mining brain connectivity

Connection matrix from the Allen Brain Atlas

Connection matrix extracted from the literature



Richardet R, Chappelier J-C, Telefont M,Hill S. Large-scale extraction of brain connectivity from the neuroscientific literature, *Bioinformatics*, Volume 31, Issue 10, 15 May 2015, Pages 1640–1647. doi: 10.1093/bioinformatics/btv025

But we did do this for statements of connectivity between brain regions. This paper actually highlights we built a connection matrix by passing out brain connectivity statements from hundreds of thousands of papers or many thousands of papers. By comparing the connection matrix that was generated from the text mining process, so that's on the right, to an actual connection matrix from the Allen Mouse Brain Atlas. We found that we could actually see some similarities in the overall connectivity that's reported. At least, there was a high-level validation that we're finding similar type of information from largescale literature text mining.

Notes

Summary

3m 55s

Data curation

Data curation is the **organization, integration and annotation of data** collected from various sources. Computational tools are useful to:

- Extract key parameters from the literature
- Enable precise annotation of the data
- Provide machine readable description of parameters, units and linked to existing brain concepts (ontologies)
- Integrate with modeling process
- Track provenance from source document, position and computable definition

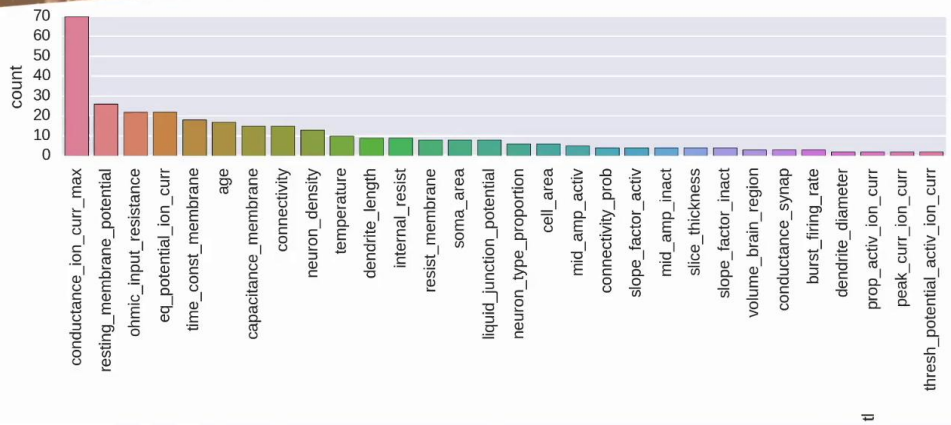
Another way to find useful data and to to organise data is actually through data curation. Data curation is the organization, integration, and annotation of data from various different sources. There are useful computational tools to help extract key parameters from the literature, to enable precise annotation of the data, and to turn, through a human oriented curation process, to provide machine readable descriptions of the parameters, the units, and links to existing brain concepts or ontologies. This is something that we've done in order to integrate with modelling processes. If you want to integrate particular parameters into your modelling, it's very useful to have a clear annotation as to where the parameter values came from in the literature. Of course, keeping track of where did that come from? Which specific statement, in which paper? In order to give proper attribution, and also to be able to go back and check and revalidate that the parameter actually means what you think it means.

Notes

Summary

4m 41s

Parameter annotation



Cell	Regional part of brain	Values	Unit	Species
Thalamus interneuron small	Lateral geniculate body	48400	1/mm**3	Wistar Rat
Thalamus relay cell	Lateral geniculate body	246800	1/mm**3	Wistar Rat
Thalamus interneuron small	Ventral posteromedial nucleus	13300	1/mm**3	Wistar Rat
Thalamus relay cell	Ventral posterolateral nucleus	150300	1/mm**3	Wistar Rat
Thalamus relay cell	Ventral posteromedial nucleus	304300	1/mm**3	Wistar Rat
Thalamic reticular nucleus cell - GABAergic	Thalamic reticular nucleus	157140	1/mm**3	Wistar Rat
Thalamic reticular nucleus cell - non-GABAergic	Thalamic reticular nucleus	29820	1/mm**3	Wistar Rat
Thalamus interneuron small	Ventral posterolateral nucleus	5900	1/mm**3	Wistar Rat
Thalamus relay cell	Ventral posterior nucleus	51011	1/mm**3	Rat

O'Reilly C, Lavarone E. and Hill, SL. A Framework for Collaborative Curation of Neuroscientific Literature. *Front. Neuroinform.* 2017;16(6):e9356. doi:10.3389/fninf.2017.00027; under a CC BY 4.0 license

In this work with Christian O'Reily, we developed a framework for collaborative curation of neuroscientific literature where you could actually annotate different properties, for example, peak conductances, resting membrane potentials, connectivity, dendroid length, and so on, from many different papers, and organise that in a way that then could be useful to search and use in model building. Here are some examples where we identified information about the thalamus, thalamus cells in particular brain regions of the thalamus and specific values with their units along with the specific species that's reported in the paper. This gives us a way of building up another database of useful properties linked to the papers, to the original publications.

Notes

Summary

5m 59s