Blue Brain Project

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

NLP for neuroscience

Neuroinformatics,
Blue Brain Project
EPFL, Switzerland

Renaud Richardet, PhD

SGAICO November 2015, s-i.ch

Agenda

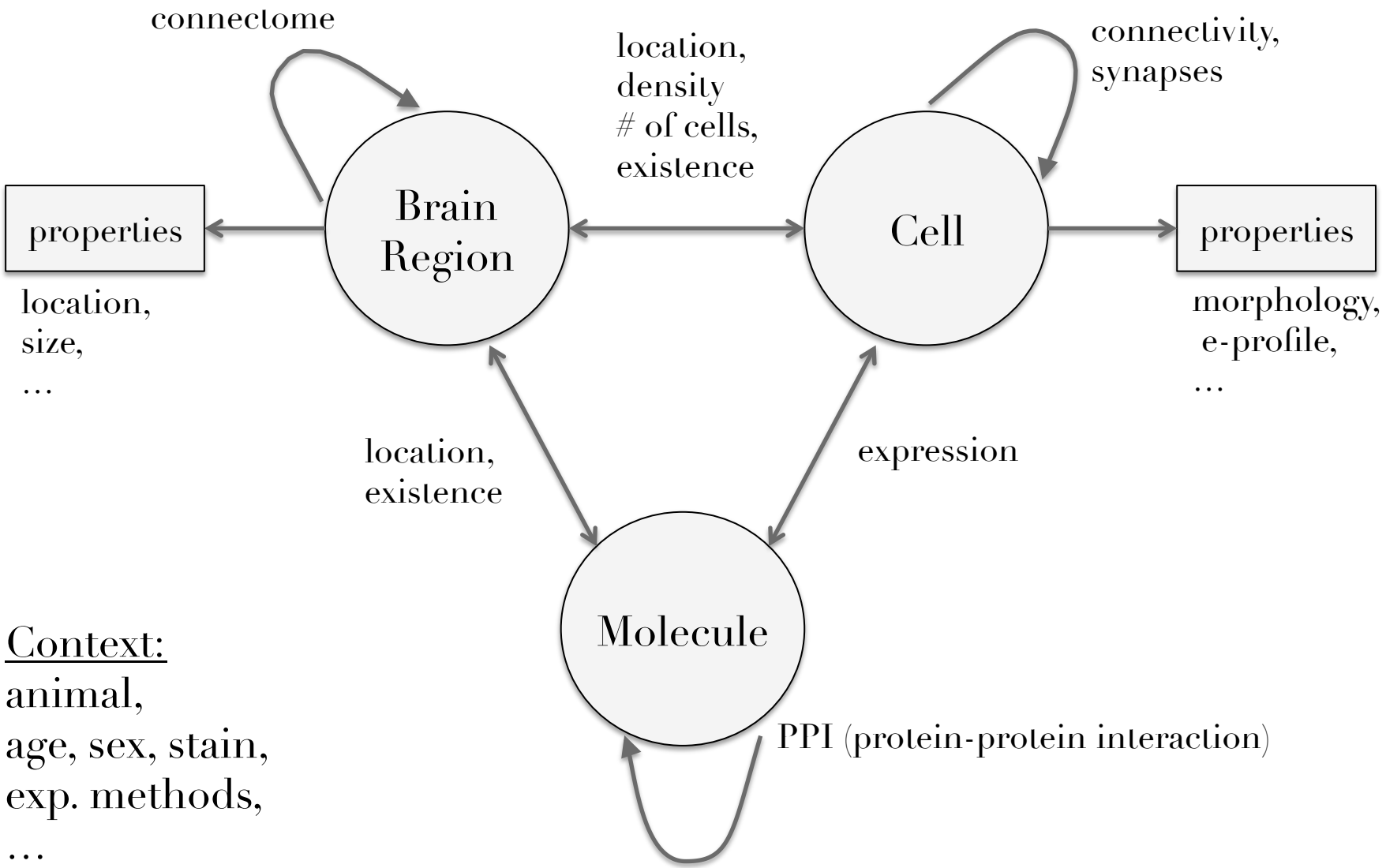
- Introduction: NLP for neuroscience
- braiNER: extracting brain region connectivity from scientific articles
- Agile text mining: neuroNER
- Topic modelling
- Synthesis

Introduction: NLP for Neuroscience

- NLP in the neuroscientific domain, e.g.:
  - identify named entities (proteins/gene, species, methods, …)
  - extract events (protein-protein interactions, …)
  - create a knowledge base of brain region connections
  - curation, integration in NS models

- Why important?
  - Highly valuable knowledge in text form within papers
  - 1 new paper each minute on PubMed on average

Model of Neuroscientific Entities & Relationships



ML-based NERs for Biomedical NLP

- Sentences, tokens, POS (OpenNLP, MaxEnt)
- Abbreviations (Second String, HMM)
- Species (Linnaeus, HMM)
- Chemicals (OSCAR4)
- Genes and proteins (BANNER, Gimli, CRF)
- *Brain regions (braiNER, CRF)*

Requirements for successful experiments

- realistic expectations from domain experts (precision and recall)
- strong commitments of both neuroscience researcher and NLP researcher regarding collaboration and communication
- sufficient amounts of accessible raw textual data
- means to evaluate a tasks' performance (or willingness to create evaluation data)
- availability of NLP models like NERs (or possibility to create them)

Agenda

- Introduction: NLP for neuroscience
- braiNER: extracting brain region connectivity from scientific articles
- Agile text mining: neuroNER
- Topic modelling
- Synthesis

brainNER

(thesis p.64)

Bioinformatics, 2015, 1–8
doi: 10.1093/bioinformatics/btv025
Advance Access Publication Date: 20 January 2015
Original paper

OXFORD

Data and text mining

Large-scale extraction of brain connectivity from the neuroscientific literature

Renaud Richardet[1,*], Jean-Cédric Chappelier[2], Martin Telefont[1] and Sean Hill[1]

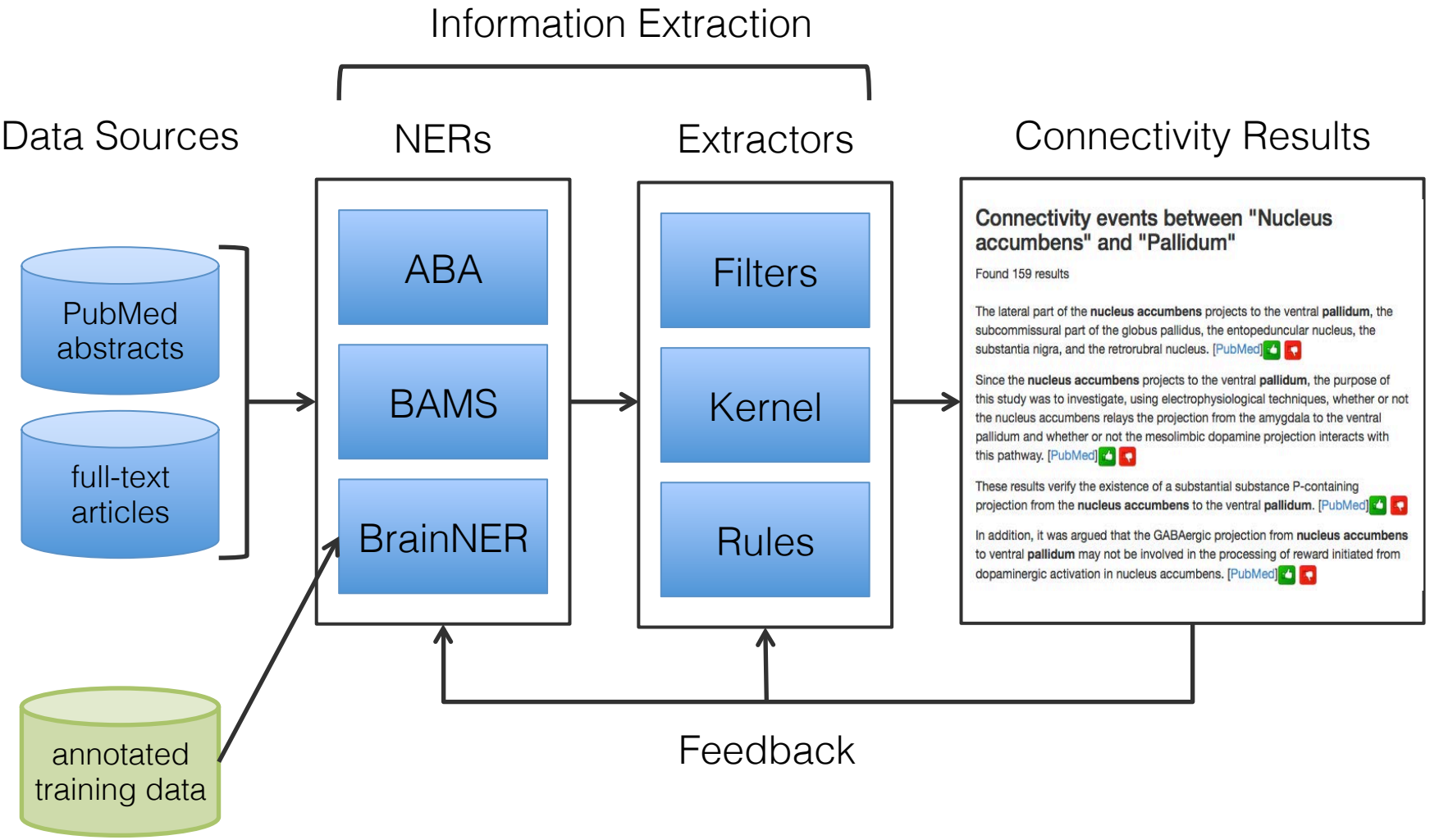
[1]Blue Brain Project, Brain Mind Institute and [2]School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

- scope: metascale brain connectivity
- goal: accelerate literature review, by providing a centralized repository of connectivity data, mined from the neuroscientific literature
- large scale (8B words)

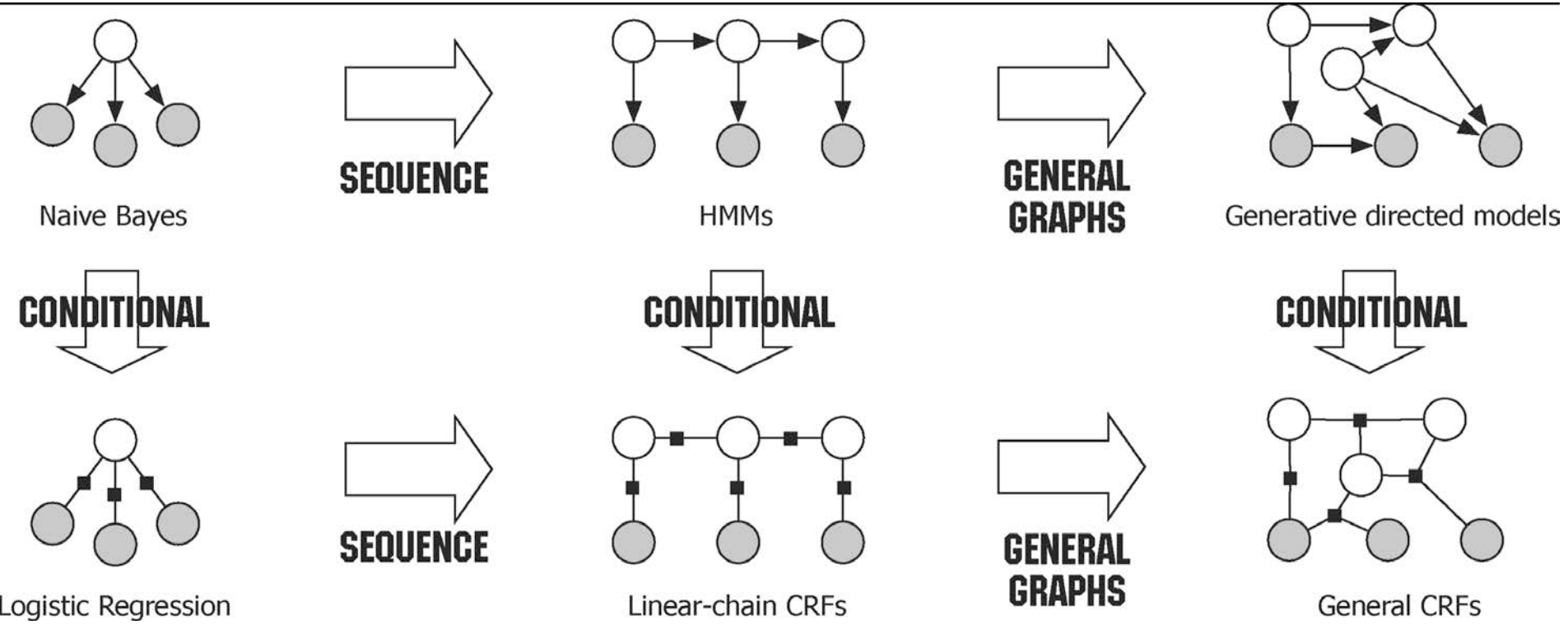
Examples of connectivity statements between brain regions

	Sample Sentence	Connectivity Statement, Comment
1	The nucleus accumbens (AC) receives projections from both the substantia nigra (SN) and the ventral tegmental area (VTA) (Dworkin, 1988).	(SN, VTA) → AC
2	Substantial numbers of tyrosine hydroxylase-immunoreactive cells in the dorsal raphe nucleus (DR) were found to project to the nucleus accumbens (AC) (Stratford and Wirtshafter, 1990).	DR → AC
3	The dentate gyrus (DG) is, of course, not only an input link between the entorhinal cortex (Ent) and the hippocampus proper (CAs), but also a major site of projection from the hippocampus (CA), as are the amygdala (Amg), entorhinal cortex (Ent), and septum (Spt) (Izquierdo and Medina 1997).	CAs → DG → Ent, (CA, Amg, Ent, Spt) → DG Complex, long range relationships
4	This latter nucleus (N?), which projects to the striatum (CP), receives inputs from motor cortex (MO) as well as the basal ganglia (BG), and is situated to integrate these and then provide feedback to the basal ganglia (BG) (Strutz 1987).	MO → N? → CP, BG ↔ N? Anaphora: ”latter nucleus (N?)” was defined in previous sentence
5	In this review, we summarize a classic injury model, lesioning of the perforant path, which removes the main extrahippocampal input to the dentate gyrus (Perederiy and Westbrook 2013).	Injury model, not normal conditions
6	The most commonly proposed mechanism is that the periaqueductal gray of the midbrain (PAG) or the cerebral cortex (Cx) have descending influences to the spinal cord (SpC) to modulate pain transmission at the spinal cord (SpC) level (Andersen 1986).	PAG → SpC, Cx → SpC ”proposed” implies an hypothesis, not a finding

Text Mining: Methods



Models

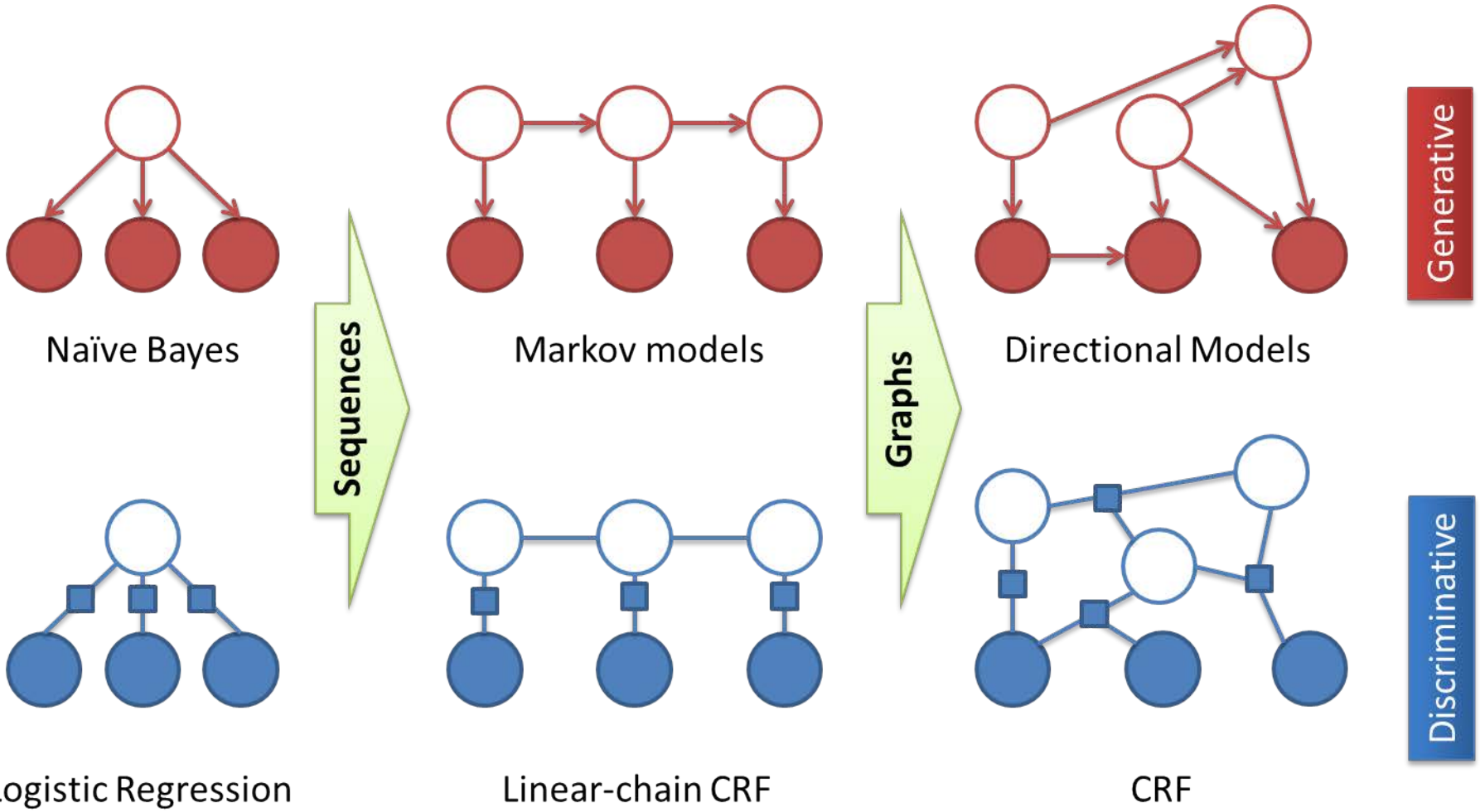


Sutton & McCallum, 2010

Features for brain region NER

PartOfSpeech	DictionaryMatcher	IsSingleChar	NumbersThenLetters
Lemma	Prefix{2,3,4}	IsSingleDigit	OnlyLettersThenNumbers
Lowercase	Suffix{2,3,4}	IsDoubleDigit	OnlyNumbersThenLetters
Capitalization	StartsWith{+,-}	HasDash	Preceding(1)+Following(1){ DictionaryMatcher, HasDash, HasQuote, HasSlash, *IsRealNumber,* StartsWith{+,-}, *IsPunctuation* }
HasGreekLetter	EndWith{%}	HasQuote	
HasNumeric	IsRealNumber	HasSlash	Preceding(3)+Following(3) {PartOfSpeech,Lemma}
IsRomanLetter	IsPunctuation	LettersThenNumbers	

Models



Adapted from C. Sutton, A. McCallum, "An Introduction to Conditional Random Fields", ArXiv, November 2010

Evaluation

Table 3. Performance comparison of brain region NER models against the WhiteText corpus (partially matching spans)

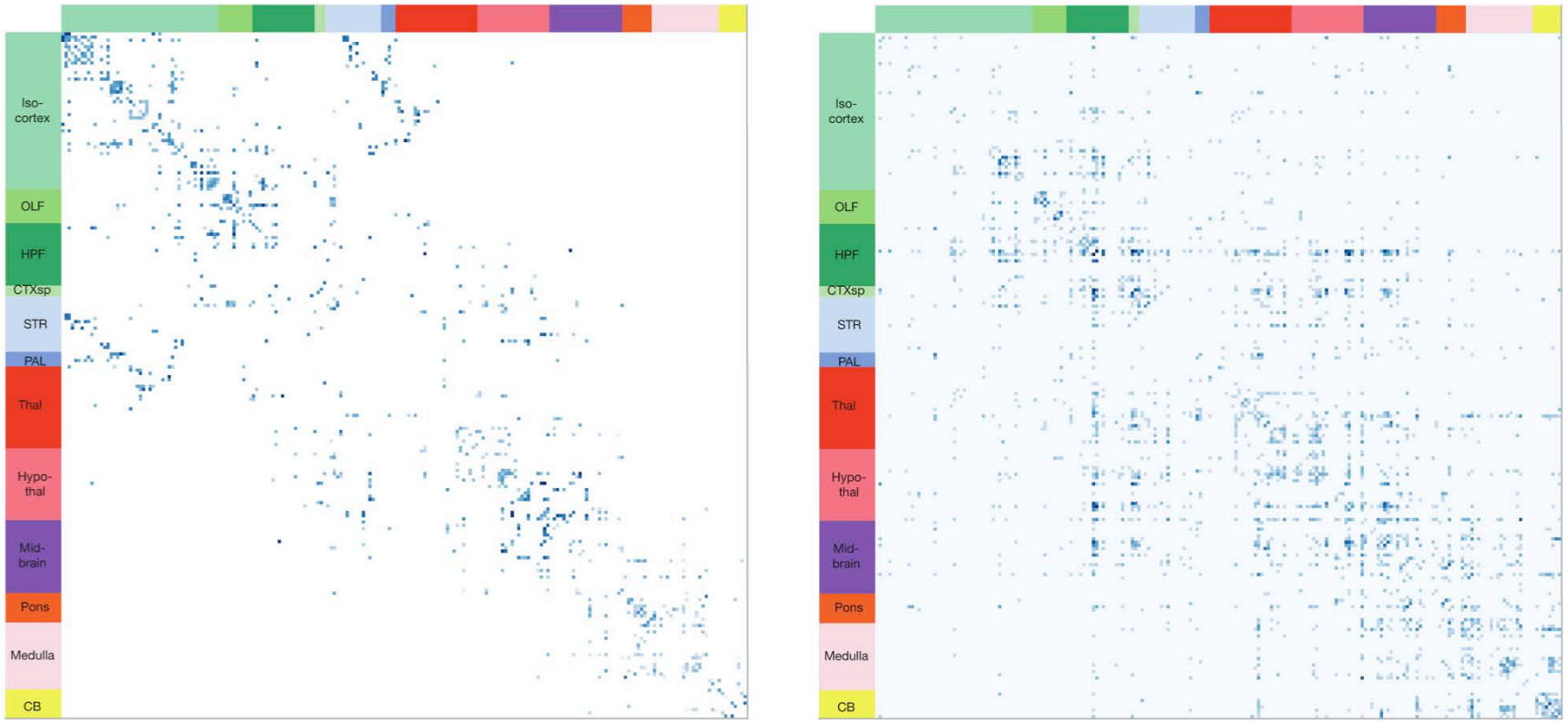
Model	Exact comparison			Lenient comparison		
	Precision	Recall	*F* score	Precision	Recall	*F* score
ABA lexicon	58.4%	11.1%	18.6%	89.9%	16.9%	28.5%
ABA-SYN lexicon	58.4%	21.9%	31.9%	**92.1%**	34.2%	49.9%
BAMS lexicon	61.1%	11.0%	18.6%	90.7%	16.2%	27.5%
BAMS-SYN lexicon	61.3%	17.5%	27.2%	89.8%	25.5%	39.7%
WhiteText (French *et al.*, 2009)	81.3%	76.1%	78.6%	91.6%	**85.7%**	**88.6%**
BraiNER-W (features from WhiteText)	83.6% (3.3)	76.4% (4.6)	79.8% (3.9)	87.1% (3.6)	77.8% (7.4)	82.1% (5.8)
BraiNER (with additional features)	**84.6%** (1.3)	**78.8%** (1.2)	**81.6%** (0.9)	88.4% (1.0)	81.0% (1.8)	84.6% (1.3)

For machine learning-based NERs [French *et al.* (2009) and BraiNER], average values over 8-fold cross validation with splits at document level and 5 repetitions, including standard deviation in parenthesis where appropriate.

Table 4. Evaluation of extraction models against the WhiteText corpus

Extractor	Prec.	Recall	*F* score
All co-occurrences (all permutations)	9%	**100%**	16%
Filter sentence > 500 characters	10%	93%	18%
Filter sentence with > 7 brain regions	11%	80%	19%
Keep if contain trigger words	15%	53%	23%
Keep nearest neighbor co-occurrence	28%	51%	36%
All filters (FILTERS)	45%	31%	37%
Shallow linguistic kernel (KERNEL)	60%	68%	**64%**
Ruta rules (RULES)	**72%**	12%	21%
FILTERS and KERNEL	66%	**19%**	**29%**
FILTERS and RULES	80%	7%	13%
KERNEL and RULES	81%	10%	18%
FILTERS and KERNEL and RULES	**82%**	7%	12%
(FILTERS or KERNEL) and RULES	80%	11%	19%

Comparison of the inter-region connectivity matrices



- Left: connection matrix from Allen Brain Atlas (ipsilateral, symmetrized), see Figure 4a of [Oh 2014].
- Right: connection matrix from the results extracted from the literature (LIT)
- LIT exhibits structural similarity with ABA (precision between ABA and LIT are significantly closer than from random matrices).
- Precision evaluated at 78% against in-vivo connectivity data from ABA.

Results

Table 5. Statistics of the corpora used, extracted brain regions and connections using all three extractors (`FILTERS` or `KERNEL` or `RULES`)

Corpus	Corpus statistics		Brain regions			Connectivity statements		
	Documents	Words	ABA	BAMS	BraiNER	ABA	BAMS	BraiNER
All PubMed abstracts	13 293 649	2.1×10^9	1 705 549	1 918 561	1 992 747	41 965	50 331	188 994
Full-text neuroscience articles	630 216	6.1×10^9	2 327 586	2 514 523	2 751 952	62 095	72 602	279 100

- 13.2 million PubMed abstracts and 630,216 full-text publications related to neuroscience
- Over 4 million (ABA) and 4 million (BAMS) brain region mentions
- Over 100,000 (ABA) and 120,000 (BAMS) potential brain region connections

frontiers in Neuroanatomy

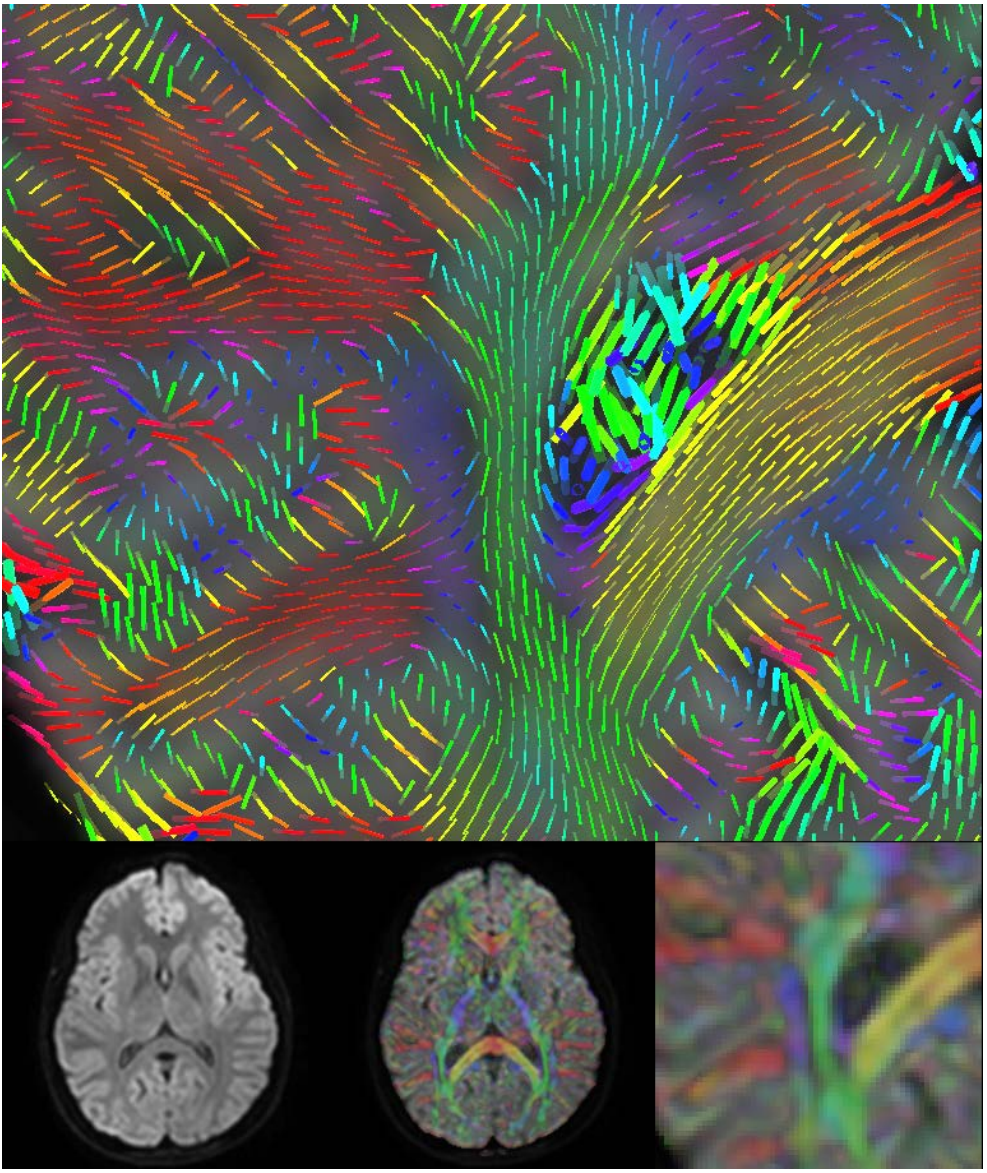
TECHNOLOGY REPORT ARTICLE

Front. Neuroanat., 27 May 2015 | http://dx.doi.org/10.3389/fnana.2015.00066

Automatic target validation based on neuroscientific literature mining for tractography

Xavier Vasques[1,2,3†], Renaud Richardet[1†], Sean L. Hill[1], David Slater[4,5], Jean-Cedric Chappelier[6], Etienne Pralong[5], Jocelyne Bloch[5], Bogdan Draganski[4,5] and Laura Cif[4,5,7*]

DTI



- DTI (diffusion tensor imaging) requires definition of potential brain region targets
- We compared manual literature review and text mining to find targets

Methods

- Manual literature review (LIT): Dr. Laura Cif (CHUV)
- Text mining (TM): Renaud Richardet (HBP)
- Diffusion tensor imaging (DTI): Dr. Xavier Vasques (HBP)
- 3 brain regions
  - internal globus pallidus (GPi) ← well known
  - subthalamic nucleus (STN) ← well known
  - nucleus accumbens (NAcc) ← "unknown"

Connectivity:

Literature (LIT)

Globus Pallidus internus	
Afferents	**Efferents**
Subthalamic nucleus	Thalamus
Substantia nigra pars compacta	Lateral habenula
Ventral tegmental area	Substantia nigra
Neostriatum	Pedunculopontine nucleus
	Cerebral cortex (rat)
	Neostriatum
Subthalamic nucleus	
Afferents	**Efferents**
Primary motor cortex	Globus Pallidus internus
Supplementary motor area	Globus Pallidus externus
Frontal eye field	Substantia nigra pars compacta
Somatosensory cortex	Substantia nigra pars reticulata
Anterior cingulate	Ventral thalamic nuclei ipsilaterally
Globus Pallidus externus	Parafascicularis thalamic nucleus contralaterally (rat)
Substantia nigra pars compacta	Substantia innominata
Ventral tegmental area	Ventral pallidum
Dorsal raphe nucleus	Pedunculopontine nucleus
Pedunculopontine nucleus	Ipsilateral cortex (rat)
Centro-median/parafascicularis complex	Neostriatum (rat)
	Spinal cord (rat)
Nucleus Accumbens	
Afferents	**Efferents**
Orbitofrontal cortex	Ventral pallidum
Anterior cingulate	Substantia nigra pars compacta
Subgenual cortex	Substantia nigra pars reticulata
Pregenual cortex	Ventral tegmental area
Hippocampus	Hippocampus
Parahippocampal cortex	Caudate
Amygdala	Putamen
Substantia nigra pars compacta	Medio-dorsal thalamus
Ventral tegmental area	Cingulate gyrus
	Substantia innominata (rat)
	Lateral preoptic area (rat)
	Lateral hypothalamic area (rat)

Table 1. *Summary of the manual literature review*

Text mining (TM)

Nucleus Accumbens							
ABA			Brainer				
Region	Score	Region	Score	Region	Score	Region	Score
Ventral tegmental area	454	ventral tegmental area	238	anterior olfactory nucleus	2	dorsal prefrontal cortex	1
Caudoputamen	412	striatum	95	ventral pallidus	2	hypothalamic arcuate nucleus	1
Cerebral cortex	295	prefrontal cortex	68	ventromedialmesencephalic tegmentum	2	perirhinal	1
Striatum-like amygdalar nuclei	175	amygdala	54	anterior striatum	2	orbital prefrontal cortex	1
Hippocampal region	122	medial prefrontal cortex	52	dorsomedial striatum	2	nucleus of the lateral olfactory tract	1
Ammon's horn	93	hippocampus	47	accessory olfactory bulb	2	frontal and cingulate cortices	1
Hippocampal formation	70	hippocampal	41	bed nucleus of the stria terminalis	2	rostral neostriatum	1
Pallidum	61	basolateral amygdala	40	paraventricular nucleus	2	basolateral amygdala–nucleus accumbens	1
Midbrain	53	caudate-putamen	39	anterior limbic cortex	2	limbic cortex	1
Subiculum	38	cortical	35	thalamic paraventricular nucleus	2	caudate/putamen	1
Thalamus	28	mesolimbic	31	tegmenti pedunculopontinus pars compacta	2	islands of calleja	1
Hypothalamus	28	hippocampal formation	29	mesencephalic ventromedial tegmental	2	medial extended amygdala	1
Periaqueductal gray	23	ventral pallidum	26	ventral pallidal area	2	vomeronasal nerve	1
Olfactory tubercle	22	ventral striatum	20	arcuate nucleus	2	ventral subiculum of the hippocampus	1
Basolateral amygdalar nucleus	19	caudate putamen	16	medial substantia nigra pars reticulata	2	lateral and basolateral amygdaloid nuclei	1
fimbria	18	thalamus	14	striatal caudate putamen	2	substantia nigra reticulata	1
Nucleus raphe pontis	18	neostriatum	13	substantia innominata	2	dorsal frontal cortex	1
Entorhinal area	18	septum	13	lateral septum	2	cea	1
Dorsal nucleus raphe	13	caudate nucleus	13	ventral striatal	2	subcortical regions	1
Globus pallidus, external segment	12	mesencephalic	13	anterolateral hypothalamus	2	supramammillary nucleus	1
medial forebrain bundle	11	amygdaloid	12	vta	2	ventromedial nucleus of the hypothalamus	1
Paraventricular nucleus of the thalamus	11	limbic	12	rostral substantia innominata	2	medial preoptic area	1
Lateral preoptic area	9	dorsal raphe nucleus	11	caudoputamen	2	anterior ventral tegmental area	1
Nucleus of the solitary tract	8	paraventricular of the thalamus	11	basolateral amygdaloid nucleus	2	median raphe	1
stria terminalis	8	corpus striatum	11	lateral hypothalamic area	2	olfactory tubercule	1
Substantia innominata	7	forebrain	10	ventral tegmentum	2	fos-positive nuclei	1
Locus ceruleus	6	neocortical	9	limbic forebrain	2	caudal amygdala	1
Orbital area	6	caudate	9	locus coeruleus	2	basal striatal	1
Interpeduncular nucleus	5	substantia nigra	9	anterior cingulate	2	dorsal medial prefrontal cortex	1
internal capsule	4	periaqueductal gray	9	nucleus accumbens septi	2	dorsal raphé	1
Infralimbic area	4	frontal cortex	8	hypothalamus	2	lateral thalamus	1
Ventral posteromedial nucleus of the thalamus	4	subicular	8	parafascicular nuclei	2	neostriatal	1
Substantia nigra, compact part	4	anterior cingulate cortex	7	spinal	2	mesolimbic dopaminergic system	1
Prelimbic area	4	thalamic	7	medial caudate-putamen	2	dorsal septum	1
Lateral hypothalamic area	3	striatal	7	medial substantia nigra	2	piriform cortex	1
Nucleus raphe magnus	3	median raphe nucleus	7	thalamic nuclei	2	accumbens nucleus	1
Median eminence	3	dorsal striatum	7	mesolimbic dopaminergic pathway	2	central amygdaloid region	1
Cerebellum	2	ventral subiculum	7	paraventricular nucleus of the hypothalamus	2	anterior commissure	1
Medial preoptic area	2	basal ganglia	6	neocortical fields	2	dorsal and ventral striatum	1
Pedunculopontine nucleus	2	ventral tegmental	6	dorsomedial medulla	1	medial pallium	1
Paraventricular hypothalamic nucleus	2	mesoaccumbens	6	brainstem	1	ta	1
fasciculus retroflexus	2	fimbria	6	periaqueductal gray matter	1	nucleus basalis	1
Arcuate hypothalamic nucleus	2	a10	6	midline thalamic	1	dorsal subiculum	1
Midbrain reticular nucleus, retrorubral area	2	subiculum	6	striatum proper	1	orbital frontal cortex	1
Isocortex	2	paraventricular nucleus of the thalamus	5	bed nucleus of stria terminalis	1	arcuate nucleus of hypothalamus	1
Dorsomedial nucleus of the hypothalamus	2	ventral hippocampus	5	nucleus of the solitary tract	1	ca3 region	1
Brain stem	2	lateral preoptic area	5	dorsolateral frontal cortex	1	diagonal band of broca	1
Bed nuclei of the stria terminalis	2	limbic structures	5	medial ventral striatum	1	telencephalic	1
Substantia nigra, reticular part	2	olfactory tubercle	5	hypothalamic	1	septal area	1
Parafascicular nucleus	2	medial frontal cortex	5	midline thalamic nuclei	1	subventricular zone	1
Main olfactory bulb	2	ventral mesencephalic tegmentum	5	excitatory tegmental area	1	dorsal caudate-putamen	1
Inferior colliculus	1	lateral hypothalamus	5	ventral part of the striatum	1	mt	1
Globus pallidus, internal segment	1	midbrain	5	ventromedial mesencephalon	1	nucleus raphe dorsalis	1
Basic cell groups and regions	1	accumbens	4	ventrolateral striatum	1	cerebral cortex	1
Supramammillary nucleus	1	entorhinal cortex	4	raphe nucleus	1	posterior hypothalamus	1
Parataenial nucleus	1	subpallidal areas	4	raphe	1	paraventricular thalamic	1
Cerebellar nuclei	1	orbitofrontal cortex	4	dorsolateral striatum	1	cerebrospinal fluid	1
Intermediodorsal nucleus of the thalamus	1	ventral mesencephalon	4	ventral forebrain	1	prefrontal cortical	1
Taenia tecta	1	prefrontal	4	periformical lateral hypothalamus	1	g.j.	1
Diagonal band nucleus	1	cortex	4	amygdala central nucleus	1	mesencephalic tegmentum	1
Agranular insular area	1	nucleus tractus solitarii	4	dorsal raphe	1	infralimbic cortex	1
Interbrain	1	basal forebrain	4	medial septum	1	forebrain areas	1
Superior central nucleus raphe, medial part	1	amygdaloid nuclei	4	intra-nucleus accumbens	1	pedunculopontine tegmental nucleus	1
Lateral amygdalar nucleus	1	dentate gyrus	3	dentate	1	spinal cord	1
Anterior olfactory nucleus	1	ventral pallidal	3	si	1	midline thalamus	1
Ventromedial hypothalamic nucleus	1	cortical regions	3	fornix	1	medial amygdala	1
Lateral septal nucleus	1	putamen	3	limbic striatum	1	hippocampal system	1
Posterior amygdalar nucleus	1	medial striatum	3	vl	1	median cervical nucleus	1
hypoglossal nerve	1	telencephalon	3	subcortical structure	1	stria terminalis	1
Suprachiasmatic nucleus	1	basolateral nucleus	3	septal region	1	ct	1
Dorsal peduncular area	1	basolateral nucleus of the amygdala	3	mesolimbic dopaminergic	1	ventral tegmental area of tsai	1
Retrohippocampal region	1	globus pallidus	3	substantia nigra zona compacta	1	cerebral spinal fluid	1
Primary somatosensory area	1	smith ad	3	extrapyramidal and limbic systems	1	lateral septal nucleus	1
Dorsal motor nucleus of the vagus nerve	1	limbic areas	3	olfactory bulb	1	cpu	1
Mammillary body	1	caudatoputamen	3	olfactory areas	1	periamygdaloid cortex	1
Medulla	1	bla	3	ca1 hippocampal	1	basal value	1
Parabrachial nucleus	1	prelimbic cortex	3	pfc	1	prefrontal cortical regions	1
ansa peduncularis	1	neocortex	2	medial raphe nucleus	1	insula	1
Field CA1	1	ventral hippocampal	2	limbic circuit	1	entorhinal	1
Perirhinal area	1	substantia nigra pars compacta	2	locus ceruleus	1	area 24	1
Posterior hypothalamic nucleus	1	ventromedial striatum	2	ip	1	vomeronasal	1
vomeronasal nerve	1	limbic system	2	zona incerta	1	perirhinal cortex	1
Piriform area	1	cingulate	2	preoptic area	1	posterior striatum	1
Subthalamic nucleus	1	dorsomedial prefrontal cortex	2	bilateral amygdala	1	retrohippocampal region	1
Medial habenula	1	nucleus raphe magnus	2	m 248	1	lpo	1
						dura mater	1
						forebrain regions	1
						raphé nuclei	1

Table 4. Brain regions for which connections have been found in the literature for the nucleus accumbens based on ABA and braiNER lexicons.

Brain Regions Co-occurrences Matrix

	Cerebral cortex	Caudoputamen	Cerebellum	Ammon's horn	Brain stem	Thalamus	Striatum-like amygdalar nuclei	Hypothalamus	Midbrain	Medulla	Hippocampal region	Nucleus accumbens	Periaqueductal gray	Nucleus raphe pontis	Entorhinal area	Dentate gyrus	Paracentral nucleus	Somatomotor areas	Locus ceruleus	Ventral tegmental area	Main olfactory bulb	Isocortex	Visual areas	Globus pallidus, external segment	Pons	Nucleus of the solitary tract	Hippocampal formation	Inferior colliculus	Subiculum	Inferior olivary complex
Cerebral cortex		1720	709	1169	520	1896	1026	341	221	368	247	447	106	132	155	42	49	634	205	236	201	184	461	74	84	59	176	42	57	37
Caudoputamen	1720		160	296	84	388	350	90	481	26	86	637	10	137	19	8	7	110	16	126	40	135	18	664	15	2	19	6	53	
Cerebellum	709	160		152	550	348	28	92	107	140	69	4	8	35	1	8	56	263	129	13	46	40	28	14	106	11	6	19	1	329
Ammon's horn	1169	296	152		75	112	753	130	89	9	703	147	10	145	640	812	12	11	91	29	40	367	17	8	6	3	96	3	216	2
Brain stem	520	84	550	75		419	203	519	201	211	52	6	128	308	7	5	8	53	111	28	22	80	35	8	99	147	23	79	3	41
Thalamus	1896	388	348	112	419		293	260	249	44	27	52	54	20	19	2	11	198	25	26	9	171	76	70	44	10	25	60	7	32
Striatum-like amygdalar nuclei	1026	350	28	753	203	293		614	88	32	136	302	137	57	110	39	10	12	55	74	67	54	63	16	11	180	94	13	78	3
Hypothalamus	341	90	92	130	519	260	614		233	187	53	54	194	85	8	9	2	9	101	52	30	19	5	13	30	90	21	12	12	1
Midbrain	221	481	107	89	201	249	88	233		163	41	68	403	233	1	5	5	5	37	190	17	25	9	10	286	7	6	93	2	35
Medulla	368	26	140	9	211	44	32	187	163		2	1	190	109		1	4	6	102		42	9	1		304	265	1	3		47
Hippocampal region	247	86	69	703	52	27	136	53	41	2		219	4	58	232	423	5	7	45	12	15	73	3	1	4		96		91	
Nucleus accumbens	447	637	4	147	6	52	302	54	68	1	219		30	24	19	1		1	11	655	8	2		52		10	84	1	51	
Periaqueductal gray	106	10	8	10	128	54	137	194	403	190	4	30		109			1	3	42	12		4		1	13	15	2	13	1	
Nucleus raphe pontis	132	137	35	145	308	20	57	85	233	109	58	24	109		10	28	3	1	261	47	30	28	18	4	24	14	42	3	2	12
Entorhinal area	155	19	1	640	7	19	110	8	1		232	19		10		481	1	1	5	2	37	36			2		216		161	
Dentate gyrus	42	8	8	812	5	2	39	9	5	1	423	1		28	481		2		22	1	11	5					113	1	34	
Paracentral nucleus	49	7	56	12	8	11	10	2	5	4	5		1	3	1	2		1	3		16	3	5		2		3	2	1	2
Somatomotor areas	634	110	263	11	53	198	12	9	5	6	7	1	3	1	1		1				1	14	37	14	4					9
Locus ceruleus	205	16	129	91	111	25	55	101	37	102	45	11	42	261	5	22	3			48	72	24	13	3	55	28	18	1		6
Ventral tegmental area	236	126	13	29	28	26	74	52	190		12	655	12	47	2	1			48		2	3	5	2	3		11		4	3
Main olfactory bulb	201	40	46	40	22	9	67	30	17	42	15	8		30	37	11	16	1	72	2		17	3	3	3	14	8	2	1	
Isocortex	184	135	40	367	80	171	54	19	25	9	73	2	4	28	36	5	3	14	24	3	17		25	3	2		51	10	9	3
Visual areas	461	18	28	17	35	76	63	5	9	1	3			18			5	37	13	5	3	25		1	18		2	4	1	3
Globus pallidus, external segment	74	664	14	8	8	70	16	13	10		1	52	1	4				14	3	2	3	3	1		2		1	5		1
Pons	84	15	106	6	99	44	11	30	286	304	4		13	24	2		2	4	55	3	3	2	18	2		13		16		3
Nucleus of the solitary tract	59	2	11	3	147	10	180	90	7	265		10	15	14					28		14				13		1			4
Hippocampal formation	176	19	6	96	23	25	94	21	6	1	96	84	2	42	216	113	3		18	11	8	51	2	1		1			54	3

link

Regions connected to: Nucleus accumbens

The table below lists brain regions for which *connectivity events* have been found in the literature. A connectivity event is a statement from a scientific article indicating that two brain regions are connected. The *score* represents the number of connectivity events, normalized by the confidence that each event has been extracted correctly (precision).

Click on a region to view the individual connectivity events:.

Region	Score
Ventral tegmental area	454
Caudoputamen	412
Cerebral cortex	295
Striatum-like amygdalar nuclei	175
Hippocampal region	122
Ammon's horn	93
Hippocampal formation	70
Pallidum	61
Midbrain	53
Subiculum	38
Thalamus	28

search link

Connectivity events between "Nucleus accumbens" and "Ventral tegmental area"

Found 655 results

INTRODUCTION The **nucleus accumbens** receives a large dopaminergic projection from the **ventral tegmental area** (AIO region; Ungerstedt, 1971). [PubMed]

BDNF is responsible for normal expression of dopamine D3 receptor in **nucleus accumbens** where it receives mesolimbic projections from **ventral tegmental area**, involving in reward system (Guillin et al., 2001). [PubMed]

Similarly, the **nucleus accumbens** projects to the **ventral tegmental area** and to a dorsomedial portion of the substantia nigra, (Swanson and Cowan, 1975; Conrad and Pfaff, 1976; Nauta et al., 1978; Phillipson, 1978, 1979), and DARPP-32-containing neurons of the nucleus accumbens are a probable source of labeled terminals in those nuclei. [PubMed]

Dopamine is a neurotransmitter in the projection from the **ventral tegmental area** to the **nucleus accumbens**, while GABA is contained in the projections from the nucleus accumbens to the ventral pallidum and from the ventral pallidum back to the ventral tegmental area. [PubMed]

These findings provide additional evidence that dopaminergic (A10) neurons projecting from the **ventral tegmental area** to the **nucleus accumbens**, contribute to locomotor activity. [PubMed]

The dopamine terminals were destroyed by injection of 6-OHDA into the **ventral tegmental area** which projects to the **nucleus accumbens**. [PubMed]

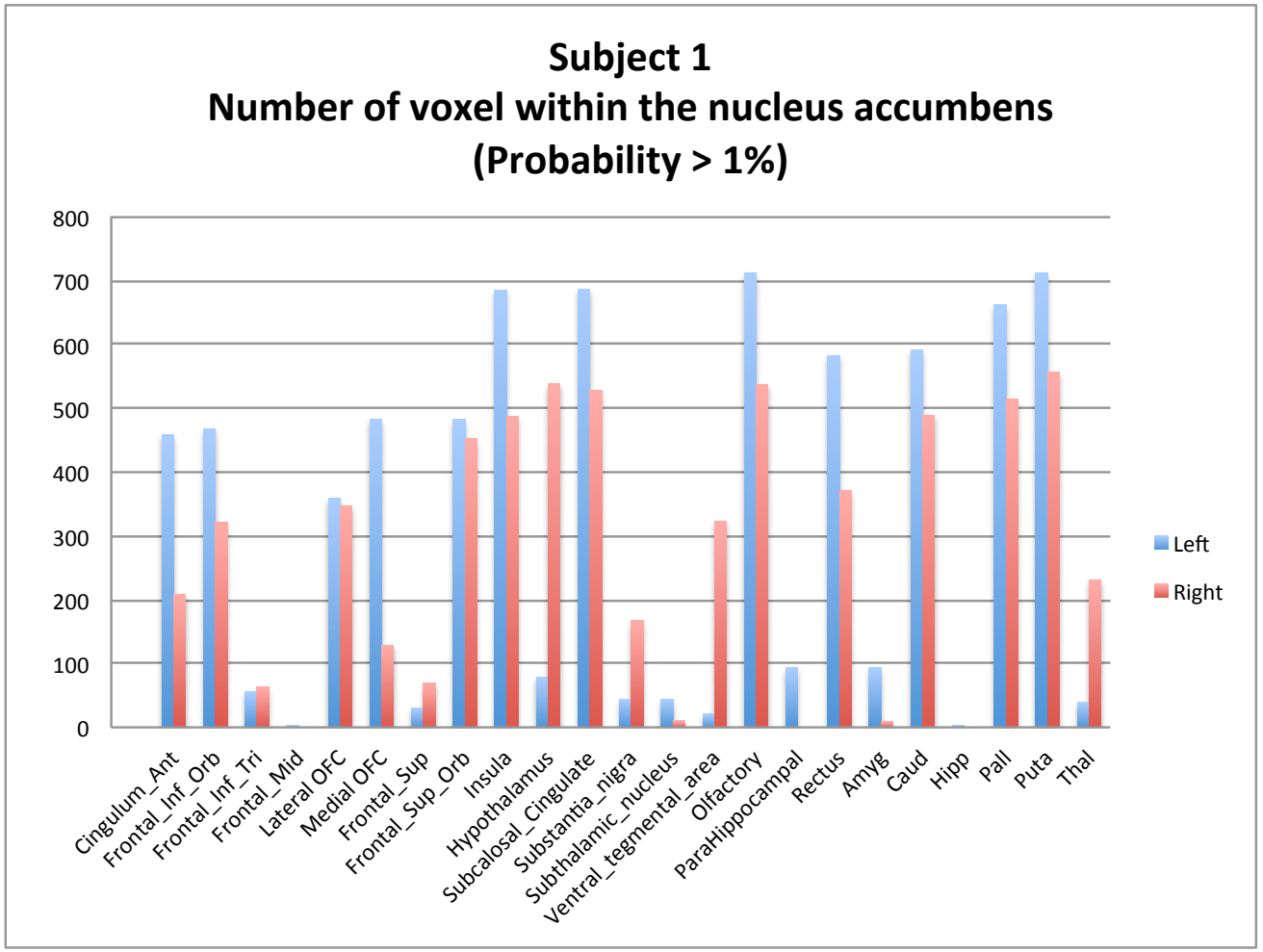
In contrast, many studies have pointed to the mesolimbic dopaminergic pathway projecting from the **ventral tegmental area** to the **nucleus accumbens** as a critical site for the initiation of psychological dependence on opioids. [PubMed]

Evidence suggests that inhibition of tonic pain is mediated by activation of mesolimbic dopamine neurons, arising from the cell bodies of the **ventral tegmental area** and projecting to the **nucleus accumbens**. [PubMed]

link

DTI

Figure 3 shows the strength connectivity within the NAcc calculated in percentage of NAcc voxels connected to the targets.



Results

	found by LIT	proposed by TM	missed by TM	recall
Gpi	10	32	0	1,00
STN	23	31	1	0,96
Nacc	21	85	0	1,00
Overall	54	148	1	0,98

Table 5. The overall efficiency of the TM against LIT.

- TM retrieved most of the targets found by LIT
- LIT took ~ 1 week
- TM took ~ 2h

Agenda

- Introduction: NLP for neuroscience
- braiNER: extracting brain region connectivity from scientific articles
- Agile text mining: neuroNER
- Topic modelling
- Synthesis

Domain expertise vs machine learning

"In data science, domain expertise is more important than machine learning skill."

Debate at O'Reilly's Stata data science conference, 2012

Domain expertise vs machine learning

"In data science, domain expertise is more important than machine learning skill."

Debate at O'Reilly's Stata data science conference, 2012

Domain expertise vs machine learning

"In data science, domain expertise is more important than machine learning skill."

Debate at O'Reilly's Stata data science conference, 2012

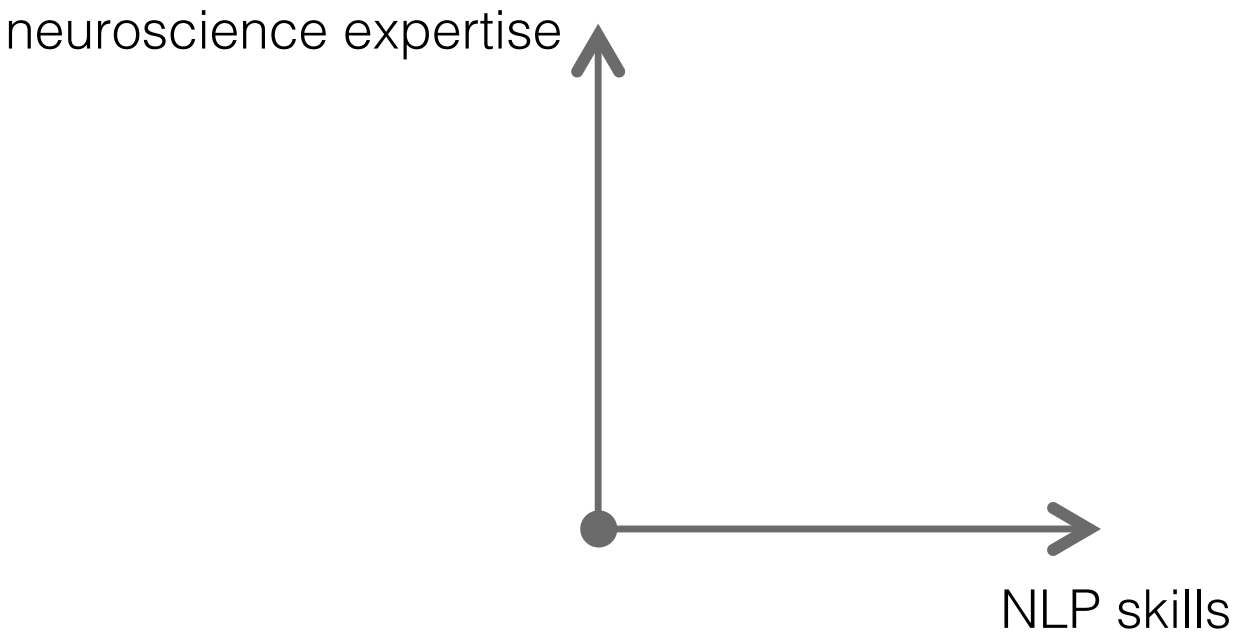
neuroscience expertise

NLP skills

Domain expertise vs machine learning

"In data science, domain expertise is more important than machine learning skill."

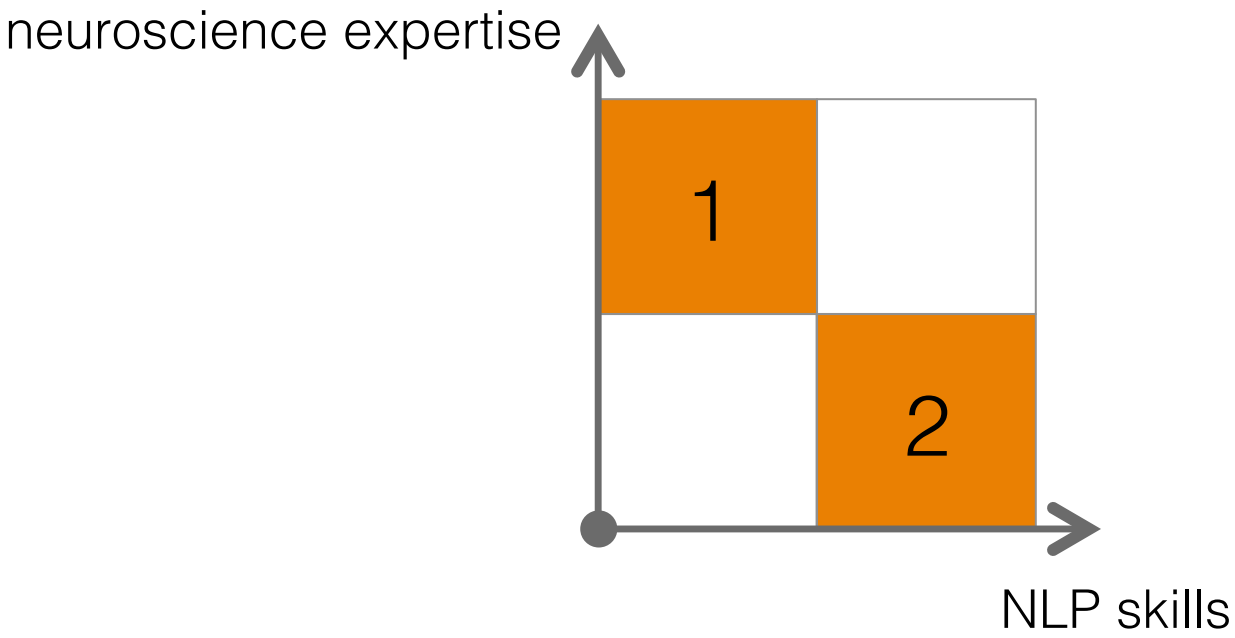
Debate at O'Reilly's Stata data science conference, 2012



Domain expertise vs machine learning

"In data science, domain expertise is more important than machine learning skill."

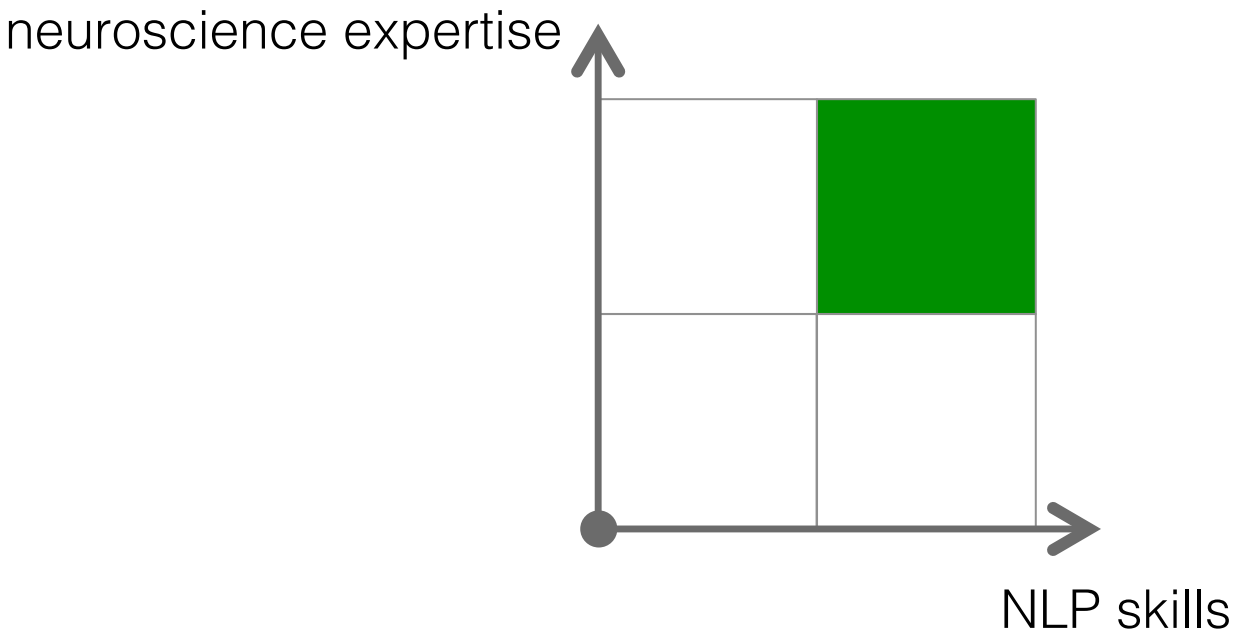
Debate at O'Reilly's Stata data science conference, 2012



Domain expertise vs machine learning

"In data science, domain expertise is more important than machine learning skill."

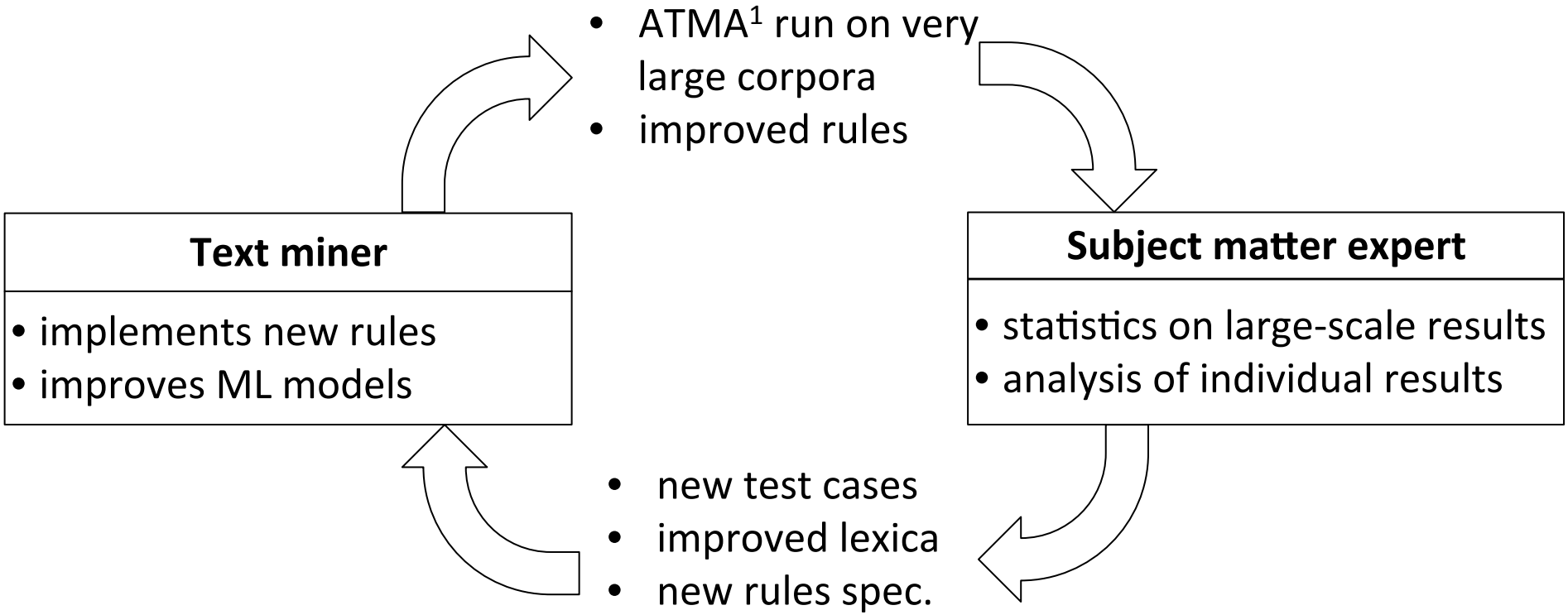
Debate at O'Reilly's Stata data science conference, 2012



Agile text mining

- custom text mining applications
- facilitate collaboration between:
  - subject matter experts
  - text miners
- short iteration cycles
- accessible results
- functional tests
- scalable

Agile text mining: Lifecycle

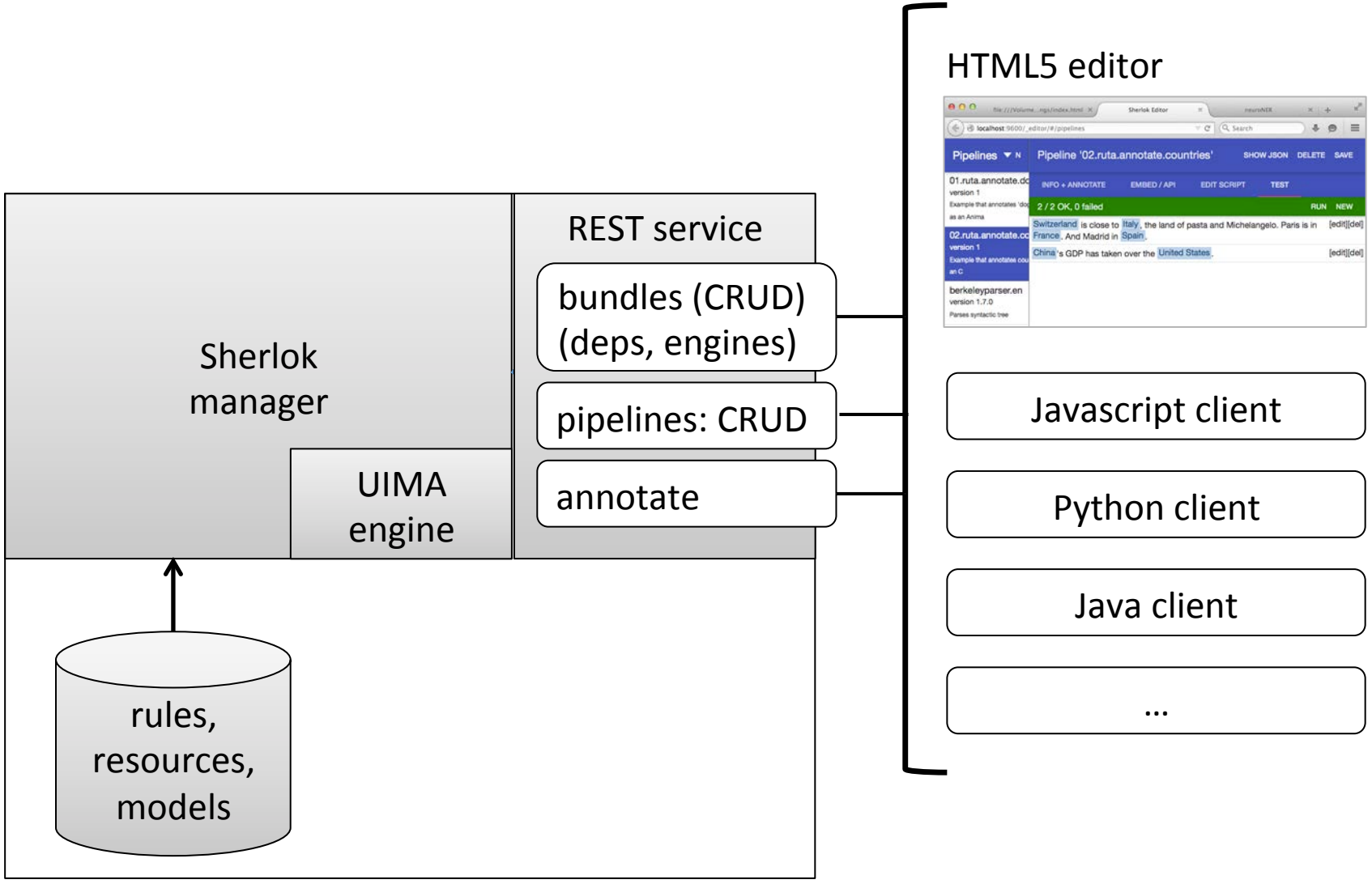


[1] ATMA: agile text mining application

Sherlok

- supports the lightweight development of *agile text mining* applications
- facilitates the collaboration between domain experts and text miners
- functional testing
- scripting language (Ruta)
- reusing existing NLP models
- integrate remote resources (e.g. Github.com)
- simple integration (JSON REST service)
- scalable

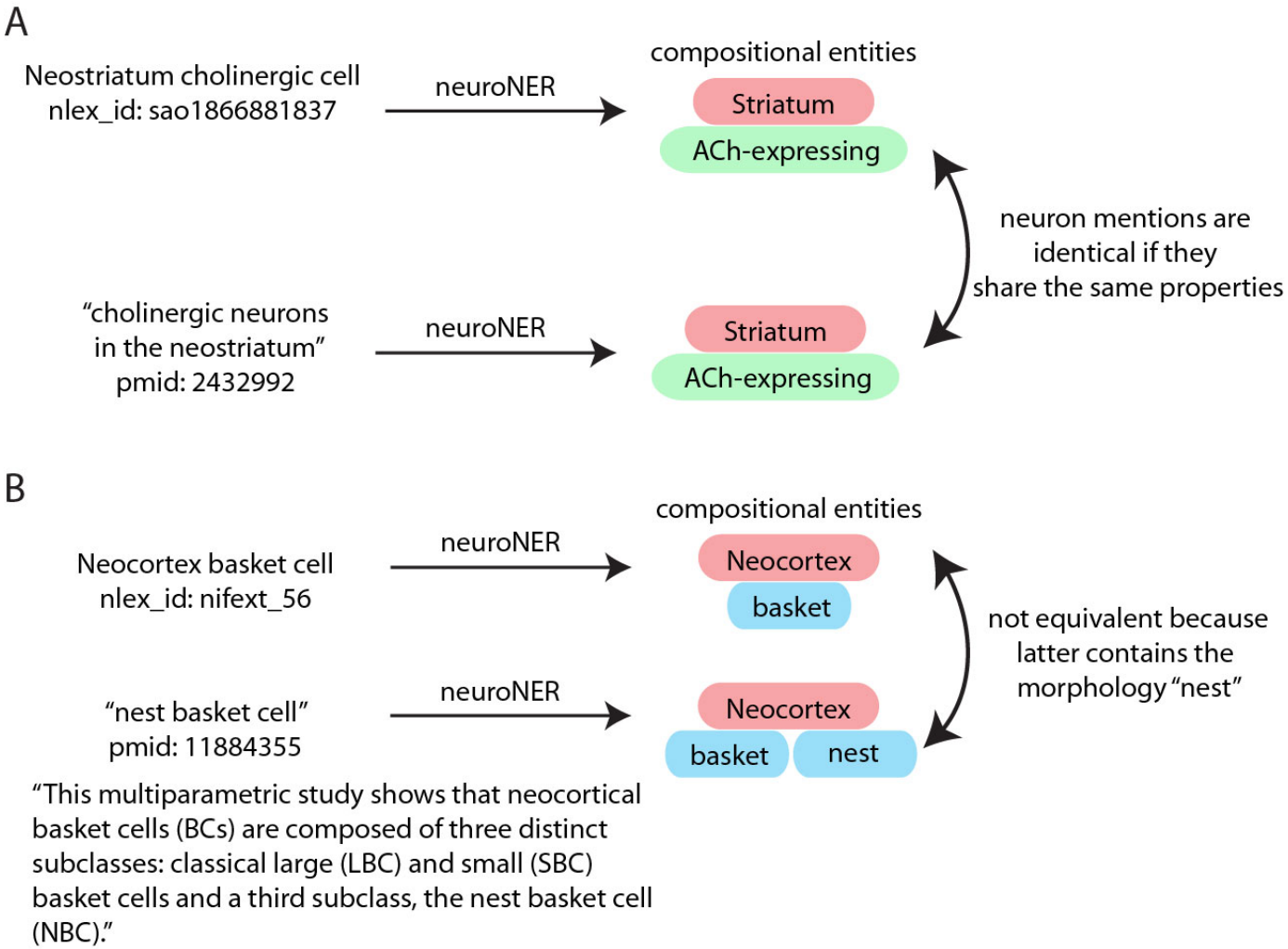
Sherlok Architecture: NLP as a service



neuroNER

- Neuroscientists regularly disagree on how neuron types should be described for example, the Petilla convention
- As experimental methodology improvements (e.g. single-cell RNA sequencing) promise to completely redefine neuron classification schemes, I believe that it is essential to first summarize decades worth of neuron study
- neuroNER: an automated method for identifying and normalizing textual mentions of neuron types and subtypes
- bottom-up approach
- neuron identity as compositional [Hamilton+ 2012]
- applied at large scale (Sherlok)
- results viewable and searchable through an easy-to-use web-based interface

Compositional Approach



Building neuroNER: a named entity recognizer for neuron

Stage 1: defining domains and enumerating termlists used to describe neuron function

Category	Examples	Extracted properties
brainregion	CA1, pedunculus cerebri, cortical	525,934
function	olfactory, primary motor, presynaptic	308,652
morphology	bipolar, candelabrum, tufted, Purkinje	167,513
species	rat, mouse, human, bovine	147,259
size	large, medium, narrow, giant	63,983
neurotransmitter	dopaminergic, GABAergic, 5HT	52,834
developmental stage	foetal, embryonic, post natal day 2	52,492
orientation	horizontal, descending, upper	10,717
protein and genes	calbindin, mGluR1, Cck, NPY-expressing	9,287
layer	layer 4a, L2/3	4,583
electrophysiology	depolarized, burst, fast-spiking	2,940

Stage 2: building natural language processing rules for identifying neuron mentions and conjoined adjectives

In the cingulate cortex (CG), the five populations represented three major classes of interneurons2, 9 (parvalbumin-positive fast-spiking basket cells, somatostatin-positive regular-spiking bipolar and multipolar cells, and cholecystokinin-positive irregular-spiking bipolar and multipolar cells) and two major classes of projection neurons (thick-tufted layer 5 nonadapting pyramidal neurons and layer 6 adapting corticothalamic neurons; Fig. 1).

E12 induction gave rise to cortical GABA neurons expressing parvalbumin (PV), somatostatin (SST), but not vasoactive intestinal peptide (VIP) (Figure 2J-L).
to cortical GABA neurons expressing parvalbumin, somatostatin, but not vasoactive intestinal peptide

Dlx genes continue to express in subsets of GABAergic neurons in embryonic, postnatal, and mature brains,

In mature cortex (P21), both cohorts settled in deep cortical layers despite their different migration routes, with a larger fraction of Dlx5-CreER labeled neurons situated deeper in layer 6 than Dlx1-CreER labeled neurons

BA n.N.BrainRegionProp [9]
BA n.N.Developmental [5]
BA n.N.Electrophysiology [4]
BA n.N.Function [4]
BA n.N.Layer [3]
BA n.N.Missing [6]
BA n.N.Morphology [8]
BA n.N.Neuron [14]
BA n.N.NeuronWithProperties [34]
BA n.N.NeurotransmitterProp [3]
BA n.N.ProteinName [10]
BA n.N.ProteinProp [4]
BA n.N.ProteinTrigger [6]

Stage 3: extensive iteration and testing

Ruta rules

```
PACKAGE neuroner;

Document{-> RETAINTYPE(BREAK, SPACE)}; // BREAK so that two neurons on two different lines do not collide
// SPACE needed to match lexical resources

DECLARE Annotation NeuronProperty(STRING name, STRING ontologyId); // base class for properties

DECLARE NeuronTrigger;  // acts as a trigger
Document{-> MARKFAST(NeuronTrigger, 'bluima/neuroner/neuron_triggers.txt', true, 3)};

// DEVELOPMENTAL STAGES
DECLARE NeuronProperty Developmental;
Document{-> MARKTABLE(Developmental, 1, 'bluima/neuroner/hbp_developmental_ontology.csv', true, 2, "", 2, "ontologyId" = 2)};

// NEUROTRANSMITTER
DECLARE NeuronProperty NeurotransmitterProp(STRING name, STRING ontologyId);
Document{-> MARKTABLE(NeurotransmitterProp, 1, 'bluima/neuroner/hbp_neurotransmitter_ontology.csv', true, 2, "", 2, "ontologyId" = 2)};

// LAYER
DECLARE NeuronProperty Layer;
Document{-> MARKTABLE(Layer, 1, 'bluima/neuroner/hbp_layer_ontology.csv', true, 2, "", 2, "ontologyId" = 2)};

// MORPHOLOGY
DECLARE NeuronProperty Morphology;
Document{-> MARKTABLE(Morphology, 1, 'bluima/neuroner/hbp_morphology_ontology.csv', true, 2, "", 2, "ontologyId" = 2)};

// SPECIES
WORDTABLE speciesWt = 'bluima/neuroner/ncbi_species_top1000.csv';
DECLARE NeuronProperty Species;
Document{-> MARKTABLE(Species, 1, speciesWt, true, 2, "", 0, "ontologyId" = 2)};

// ORIENTATION
DECLARE NeuronProperty Orientation;
"(?i)inverted|horizontal|descending|upper|lower" -> Orientation;

// SIZE
DECLARE NeuronProperty Size;
"(?i)large|medium|small|narrow|giant" -> Size;
```

Ruta rules (2)

```
84  //
85  // AGGREGATE NEURONS ////////////////////////////////////////////////////////////////////////////////////////
86
87  // aggregate multiple triggers together
88  (NeuronTrigger "and"? NeuronTrigger){-> SHIFT(NeuronTrigger)};
89
90  DECLARE Neuron;                // matches the whole span of a neuron definition
91  DECLARE PreNeuron, PostNeuron; // context before and after a neuron trigger
92
93  // gather NeuronProperty occurring before and after Neuron into Pre and PostNeuron
94  NeuronProperty+{-> MARK(PreNeuron, 1, 1)} NeuronTrigger;
95  (NeuronProperty+ ("-" | COMMA | "and" | (COMMA "and") | (COMMA "or")))* {-> MARK(PreNeuron, 1, 2)} NeuronProperty+ NeuronTrigger;
96  (("and" | "in" | "of" | "with") "the"? NeuronProperty+)+ {-> MARK(PostNeuron)};
97  NeuronTrigger NeuronProperty+ {-> MARK(PostNeuron, 2, 2)};
98
99  // aggregate Pre and PostNeurons into Neuron, remove them
100 PreNeuron NeuronTrigger PostNeuron {-> MARK(Neuron, 1, 3)};
101 NeuronTrigger PostNeuron {-> MARK(Neuron, 1, 2)};
102 PreNeuron NeuronTrigger {-> MARK(Neuron, 1, 2)};
103 //PreNeuron{-> DEL};
104 //PostNeuron{-> DEL};
105
106 NeuronTrigger{NOT(REGEXP("[Cc]ells?")) -> MARK(Neuron, 1, 1)}; // remove single isolated Neuron, unless "cell"
107
108 Neuron{CONTAINS(BREAK) -> DEL}; // remove neuron that have line breaks
109 // TODO check above with pdf (new lines?)
110
111 (Neuron{-> UNMARK(Neuron)}){PARTOFNEQ(Neuron)}; // only keep longest Neuron
112
```

Elasticsearch front-end

Neuron Brainregion Function Morphology Neurotransmitter Species Developmental Size Protein Layer

Immunostaining of planar neurons was light, comparable to that of excitatory neurons (pyramidal neurons in the DCN), whereas immunostaining of radiate neurons was dark, comparable to that of glycinergic neurons (cartwheel cells in the dorsal cochlear nucleus and principal cells in the medial nucleus of the trapezoid body). (PubMed)

A linear quantitative analysis of layer IV basket cell connectivity data suggests that on average basket cells (1) comprise 25-35% of all GABAergic neurons in layer IV (3552-4736 cells mm(-3)), (2) account for 30-41% of all putative inhibitory dendritic synapses of layer IV spiny stellate cells (145-195 synapses cell(-1)) and a similar proportion of layer IV basket cells (25-37%, 71-107 synapses cell(-1)), and (3) provide each layer IV spiny cell with 13-45 axons and each layer IV basket cell with 6-29 axons. (PubMed)

The list of neurochemically identified distinct cell types can be given as follows: five types GABA-containing cell types with secondary markers and at least one without; two glycinergic cell types and one interplexiform cell where glycine colocalizes with somatostatin; one dopaminergic amacrine cell and also a variant of this with interplexiform morphology; two types of serotoninergic cells; three NADPHdiaphorase-positive cells, one substance P-positive cell type without identified second marker; one CCK-positive cell type without identified second marker and the calbindin positive cells (at least one but potentially more types). (PubMed)

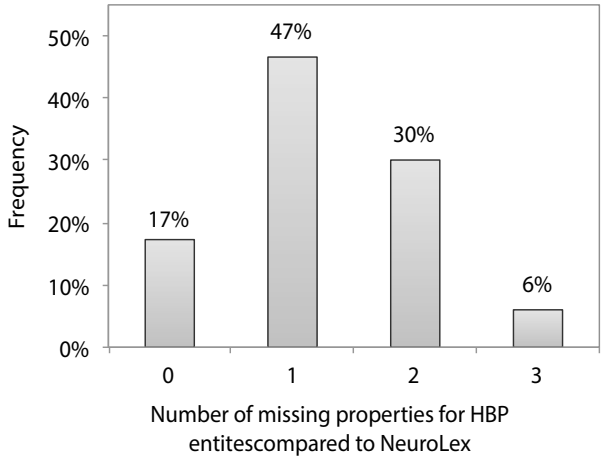
Here we document cooperative interactions between cerebral-buccal interneuron 2 and cerebral-buccal interneuron 12, characterize synaptic input to cerebral-buccal interneuron 2 and cerebral-buccal interneuron 12 from buccal peripheral nerve 2,3, describe a synaptic connection between cerebral-buccal interneuron 1 and buccal neuron B34, further characterize connections made by cerebral-buccal interneurons 2 and -12 with B34 and B61/62, and describe a novel, inhibitory connection made by cerebral-buccal interneuron 2 with a buccal neuron. (PubMed)

Fast-spiking nonpyramidal neurons, including chandelier cells, basket cells, neurogliaform cells, double bouquet cells, net basket cells, bitufted cells, and regular-spiking pyramidal neurons all respond to stimulation of multiple whiskers on the contralateral face. (PubMed)

Evaluation

An evaluation corpus was created to evaluate neuroNER's precision and recall. 200 sentences were manually evaluated, resulting in a **82% recall and 98% precision**.

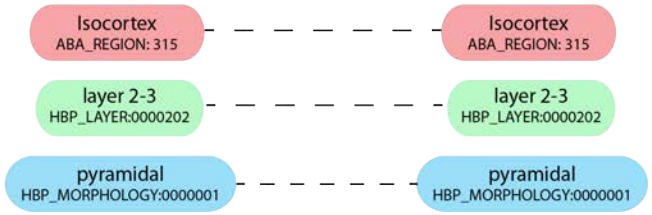
In a second evaluation, we used the neuroNER to cross-compare NeuroLex and BBP neuron lists. 17% of BBP entities can be fully normalized to NeuroLex, whereas 36% are missing more than 1 property. For most of these cases, the missing property is a layer term not present in NeuroLex.



Correspondence btw HPB and NeuroLEX

A

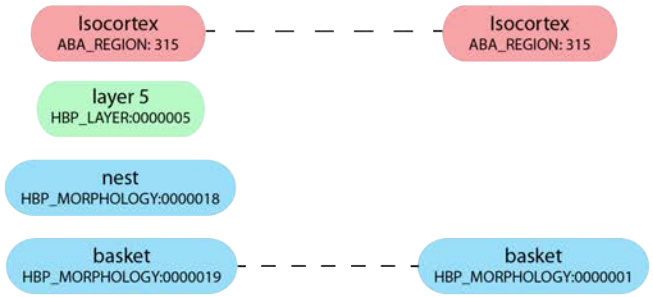
Layer II/III Pyramidal Cell
HBP_CELL:0000028



Neocortex pyramidal layer 2-3 cell
NeuroLEX_CELL: nifext_49

B

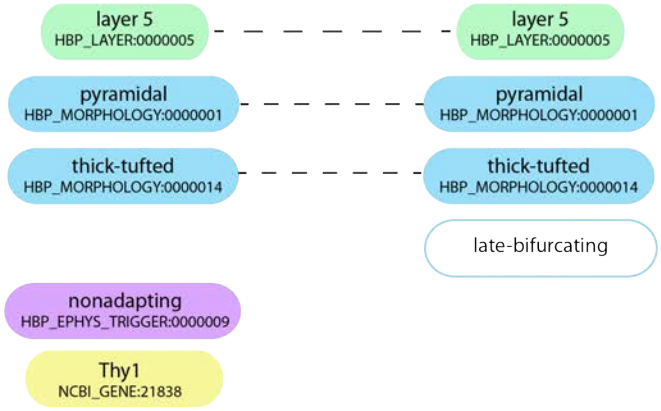
Layer V Nest Basket Cell
HBP_CELL:0000131



Neocortex basket cell
NeuroLEX_CELL: nifext_56

C

"thick-tufted layer 5 nonadapting pyramidal neurons expressing Thy1"
Sugino et al 2006,
PMID: 16369481



Layer V Late bifurcating Thick Tufted Pyramidal Cell
HBP_CELL:0000037

Agenda

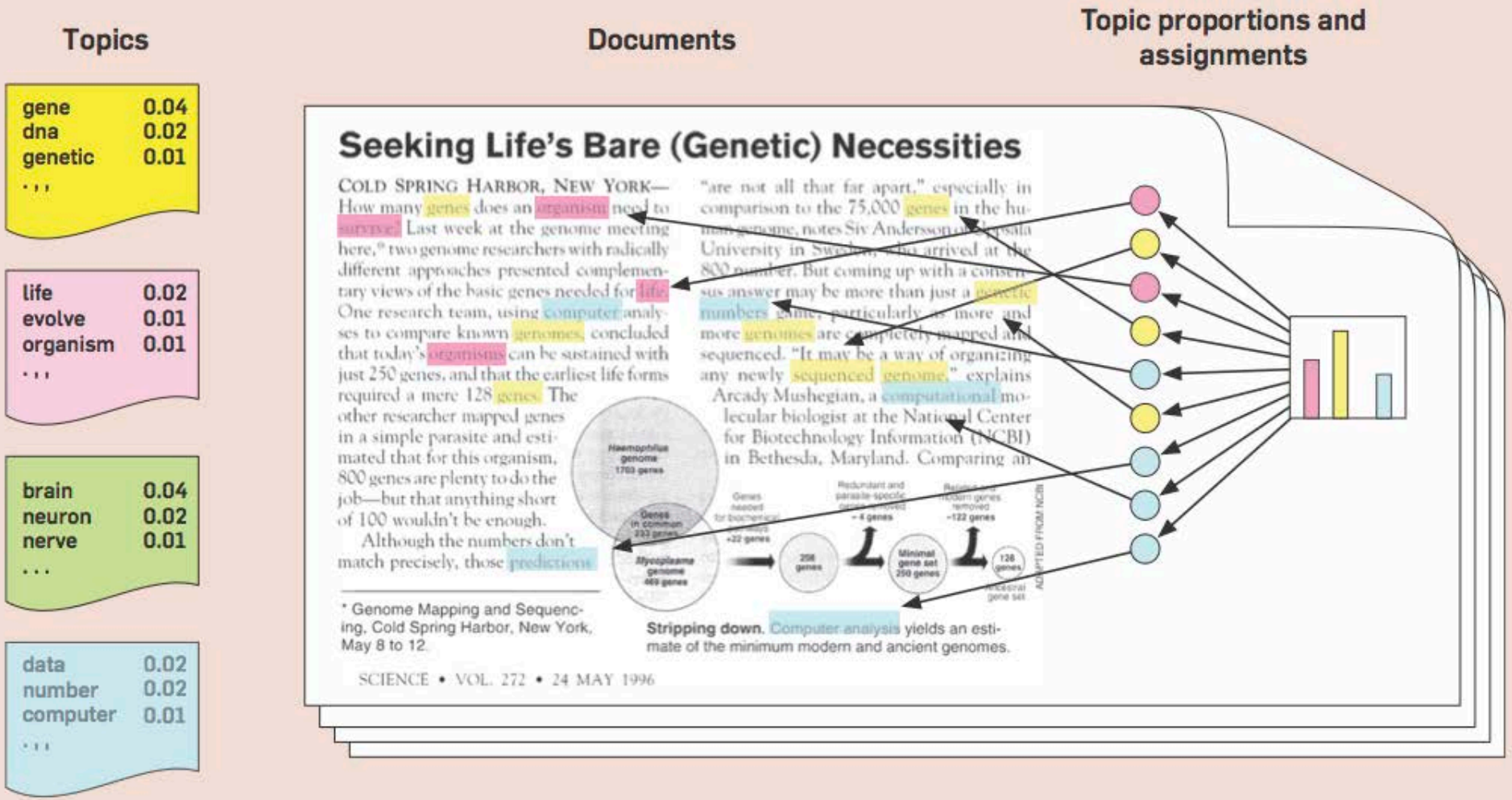
- Introduction: NLP for neuroscience
- braiNER: extracting brain region connectivity from scientific articles
- Agile text mining: neuroNER
- Topic modelling
- Synthesis

Topic Models

topic 1	topic 2	topic 3	topic 4	topic 7	topic 8	topic 9	topic 12	topic 13	topic 21
venous	apoptosis	liver	activity	brain	treatment	current	electron	pregnancy	system
vein	cell	hepatic	enzyme	cerebral	patient	channel	microscopy	women	data
portal	death	hepatocytes	activities	cortex	management	potential	surface	fetal	computer
thrombosis	p53	hepatitis	phosphatase	seizures	disease	membrane	layer	maternal	information
pulmonary	toxin	cirrhosis	enzymes	cortical	therapy	mv	membrane	pregnant	image
artery	apoptotic	bile	alkaline	epilepsy	diagnosis	k+	cells	delivery	device
veins	bcl-2	kidney	acid	temporal	risk	conductance	scanning	birth	developed
arterial	dna	chronic	found	regions	care	cells	structure	infants	analysis
catheter	induced	biliary	specific	frontal	clinical	action	junctions	gestation	program
shunt	fas	serum	rat	lesions	important	voltage	observed	placental	devices

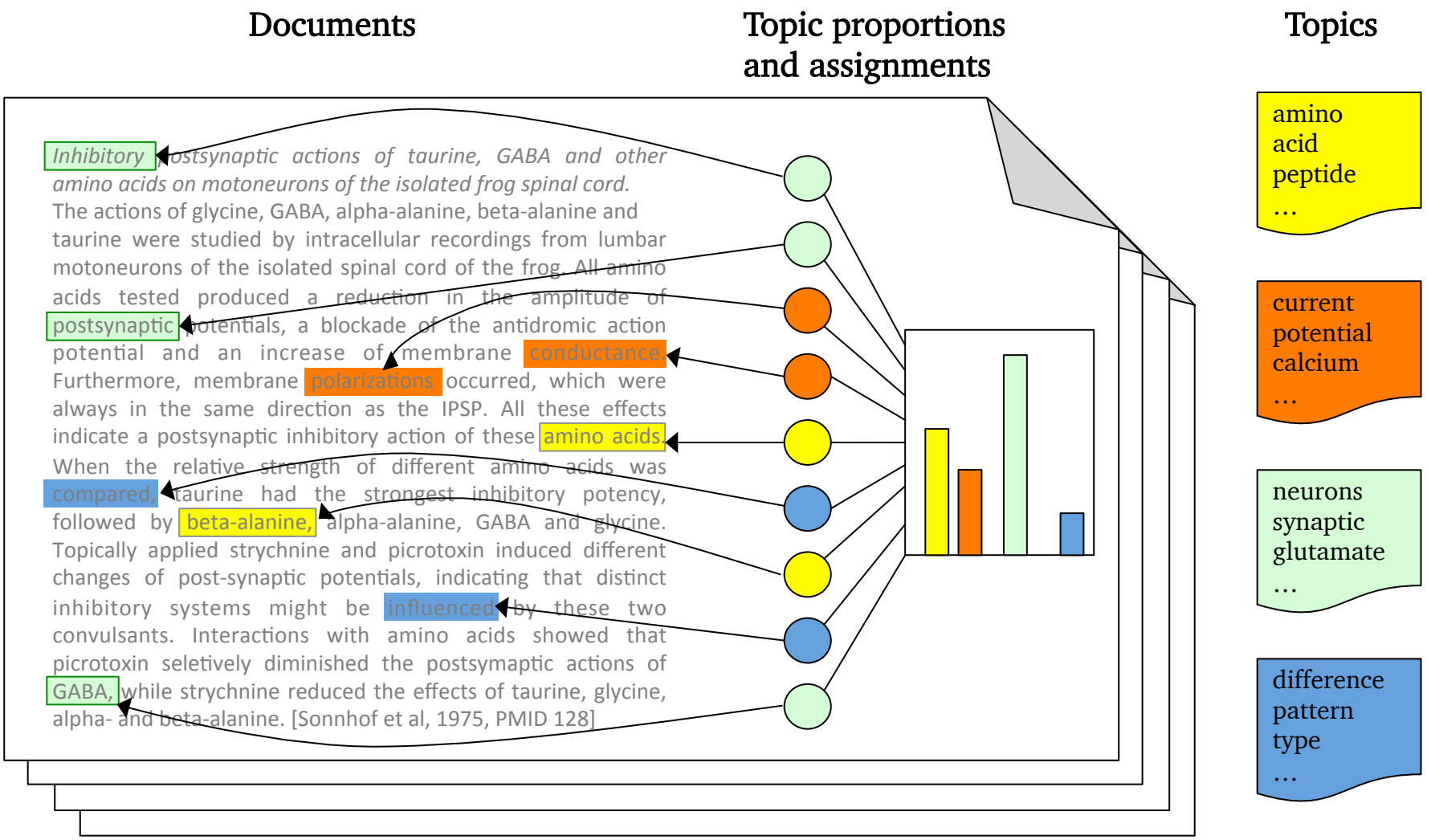
Selected topics from PubMed DCA model

Topic Models



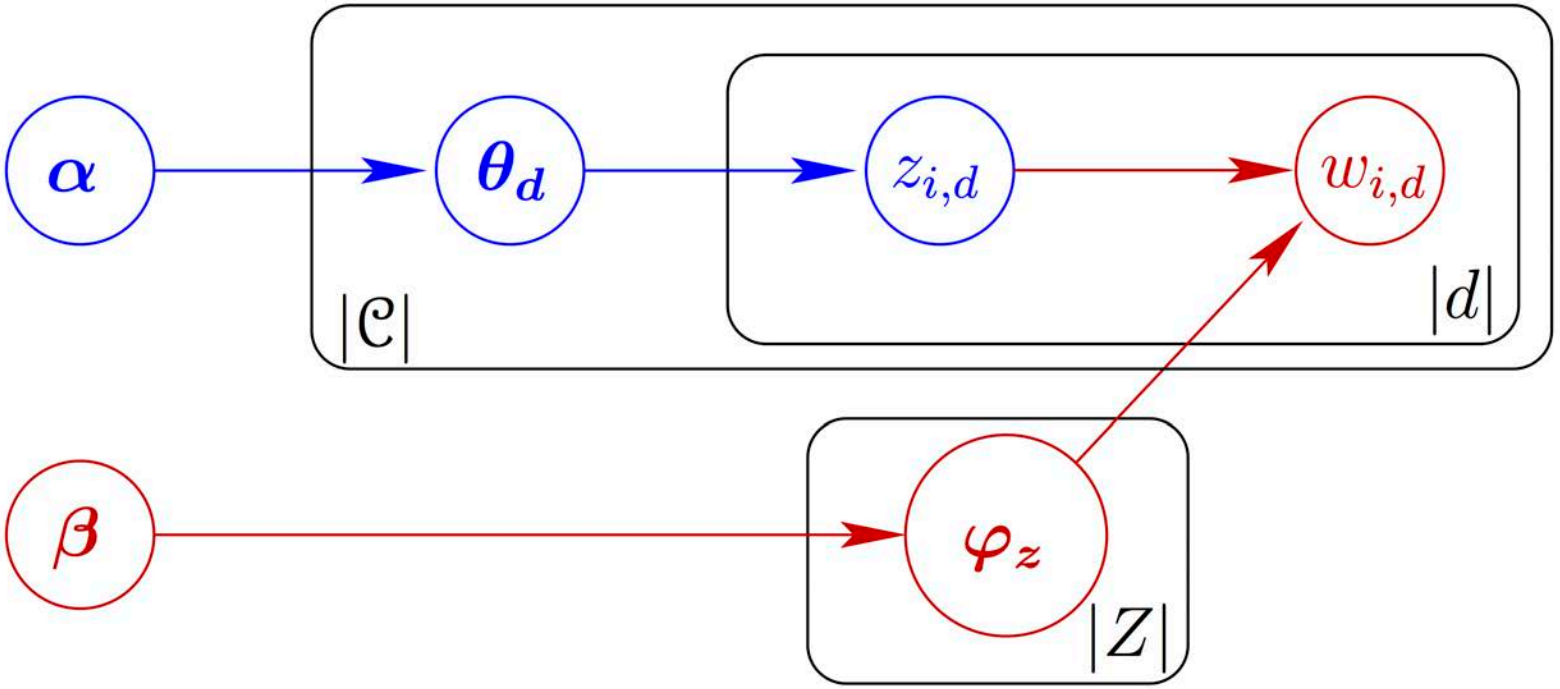
D. Blei, *Probabilistic topic models*. Communications of the ACM, 55(4):77–84, 2012.

Topic Models



D. Blei, *Probabilistic topic models*. Communications of the ACM, 55(4):77–84, 2012.

Topic Models



(c) LDA

Chappelier, 2011 in *Modèles statistiques pour l'accès à l'information textuelle*

Topic Models: Example PmId 128

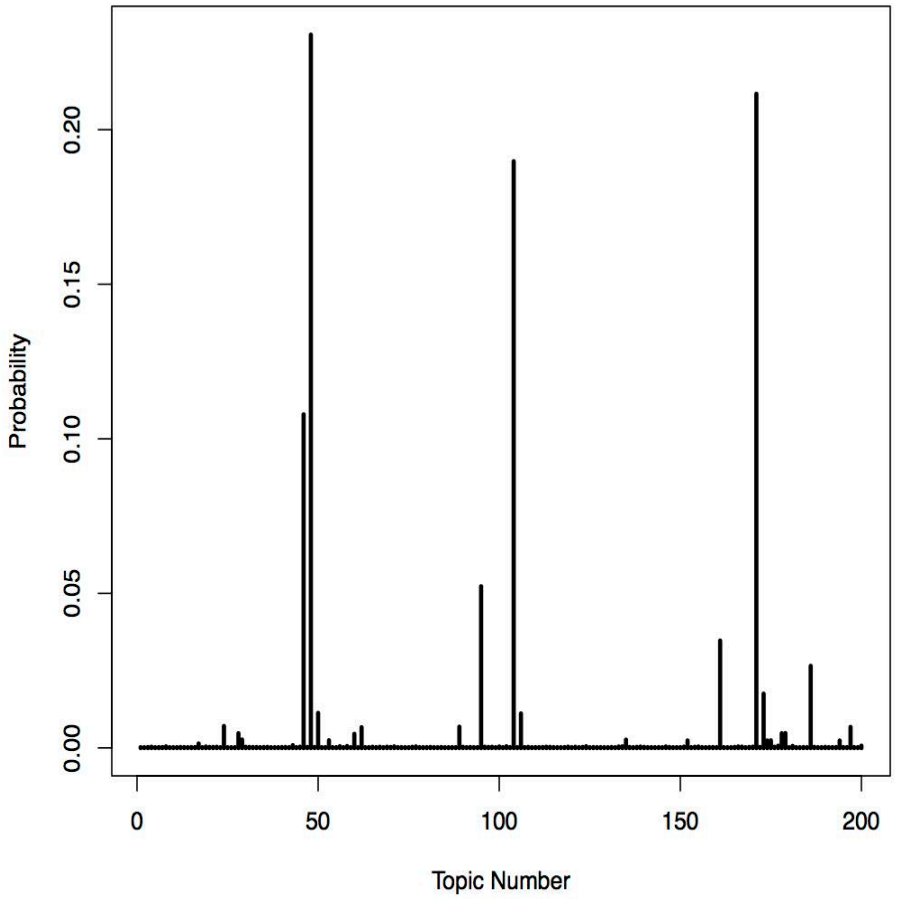
Inhibitory postsynaptic actions of taurine, GABA and other amino acids on motoneurons of the isolated frog spinal cord

http://www.ncbi.nlm.nih.gov/pubmed/128

The actions of glycine, GABA, alpha-alanine, beta-alanine and taurine were studied by intracellular recordings from lumbar motoneurons of the isolated spinal cord of the frog. All amino acids tested produced a reduction in the amplitude of postsynaptic potentials, a blockade of the antidromic action potential and an increase of membrane conductance. Furthermore, membrane polarizations occurred, which were always in the same direction as the IPSP. All these effects indicate a postsynaptic inhibitory action of these amino acids. When the relative strength of different amino acids was compared, taurine had the strongest inhibitory potency, followed by beta-alanine, alpha-alanine, GABA and glycine. Topically applied strychnine and picrotoxin induced different changes of post-synaptic potentials, indicating that distinct inhibitory systems might be influenced by these two convulsants. Interactions with amino acids showed that picrotoxin seletively diminished the postsymaptic actions of GABA, while strychnine reduced the effects of taurine, glycine, alpha- and beta-alanine. But differences in the susceptibility of these amino acid actions to strychnine could be detected: the action of taurine was more sensitively blocked by strychnine compared with glycine, alpha- and beta-alanine. With regard to these results the importance of taurine and GABA as transmitters of postsynaptic inhibition on motoneurons in the spinal cord of the frog is discussed.

$action_{171}$ $glycin_{200}$ $GABA_{200}$ $taurin_{200}$ $studi_{200}$ $intracellular_{161}$ $record_{106}$ $lumbar_{104}$ $motoneuron_{104}$ $isol_{200}$ $spinal_{104}$ $cord_{104}$ $frog_{46}$ $amino_{48}$ $acid_{48}$ $test_{89}$ $produc_{200}$ $reduct_{200}$ $amplitud_{200}$ $postsynapt_{200}$ $potenti_{200}$ $blockad_{200}$ $antidrom_{200}$ $action_{171}$ $potenti_{200}$ $increas_{200}$ $membran_{200}$ $conduct_{104}$ $Furthermor_{200}$ $membran_{200}$ $polar_{48}$ $occur_{104}$ $alway_{200}$ $direct_{171}$ $IPSP_{200}$ $effect_{171}$ $indic_{200}$ $postsynapt_{200}$ $inhibitori_{200}$ $action_{171}$ $amino_{48}$ $acid_{48}$ $relat_{200}$ $strength_{46}$ $differ_{95}$ $amino_{48}$ $acid_{48}$ $compar_{200}$ $taurin_{200}$ $strongest_{171}$ $inhibitori_{200}$ $potenc_{200}$ $follow_{104}$ $GABA_{200}$ $glycin_{200}$ $Topic_{173}$ $appli_{200}$ $strychnin_{200}$ $picrotoxin_{46}$ $induc_{200}$ $differ_{95}$ $chang_{200}$ $potenti_{200}$ $indic_{200}$ $distinct_{95}$ $inhibitori_{200}$ $system_{104}$ $influenc_{200}$ $convuls_{200}$ $Interact_{171}$ $amino_{48}$ $acid_{48}$ $show_{171}$ $picrotoxin_{46}$ $selet_{200}$ $diminish_{171}$ $postsymapt_{200}$ $action_{171}$ $GABA_{200}$ $strychnin_{200}$ $reduc_{200}$ $effect_{171}$ $taurin_{200}$ $glycin_{200}$ $differ_{95}$ $suscept_{200}$ $amino_{48}$ $acid_{48}$ $action_{171}$ $strychnin_{200}$ $detect_{89}$ $action_{171}$ $taurin_{200}$ $sensit_{200}$ $block_{171}$ $strychnin_{200}$ $compar_{200}$ $glycin_{200}$ $regard_{95}$ $result_{171}$ $import_{161}$ $taurin_{200}$ $GABA_{200}$ $transmitt_{200}$ $postsynapt_{200}$ $inhibit_{171}$ $motoneuron_{104}$ $spinal_{104}$ $cord_{104}$ $frog_{46}$ $discuss_{50}$

Topic Models: Example PmId 128



topic 45	topic 47	topic 103	topic 170
receptor	amino	nerve	effect
glutamate	acid	spinal	inhibition
release	peptide	cord	activity
synaptic	acids	sensory	inhibitory
acid	residues	peripheral	action
neurons	sequence	motor	inhibited
acetylcholine	protein	sympathetic	vitro
cholinergic	residue	axon	results
slices	synthetic	dorsal	treatment
excitatory	lysine	fiber	mechanism
hippocampal	fragment	cervical	vivo

Marc Zimmermann, 2013

Topic Models: Evaluation of Libraries

Name	Author	Version	Model(s)	Parameter Estimation	Likelihood Estimation	Language	Deployment	Reference
Mallet	McCallum et al	2.0.7 (last change in Repo regarding topic models: Jan 2012, release with parallel topic models: 2008)	LDA (+ hyper-parameter optimization.)	Distributed Gibbs Sampling	Left-to-right particle sampler	Java	Multi-threaded, single-machine	[9]
DCA	Wray Buntine	0.202 (released August 2009, first release February 2009)	LDA, Gamma Poisson Models, Pachinko Allocation (experimental) + hyper-parameter optimization.	Gibbs Sampling	Left-to-right sequential sampler	C	Multi-threaded, single-machine	[2], [1]
Online LDA	Matthew Hoffman	September 2010	LDA (no hyper-parameter opt.)	Online VB	NA	Python	Single machine, online	[7]
Vowpal Wabbit	John Langford and Matthew Hoffman	Last relevant change: August 2012. First version: Jan 2012	LDA (no hyper-parameter opt.)	Online VB	NA	C++	Single-machine, no parallelization of LDA	[6], [7]
PLDA	Yi Wang et al.	Last changes: Jul 2011. First version 2008.	LDA (no hyper-parameter optimization)	Parallel Gibbs Sampling (MPI implementation of asynchronous LDA with optimizations)	NA	C++	Parallel deployment (MPI)	[14]
Mr. LDA	Ke Zhai et al.	Last changes: August 2012, Development start: Jan 2012	LDA	Distributed VB	Lower-bound for likelihood given	Java	MapReduce (Hadoop)	[15]
Hadoop LDA	Xiance Si	Last changes: March 2012, Dev start: April 2010	LDA (no hyper opt) Parallel Gibbs Sampling	Gibbs Sampling	NA	Java	MapReduce (Hadoop)	[11]

Table 3.1.: Overview over implementations considered.

Marc Zimmermann, 2013

Topic Models: Evaluation of Libraries

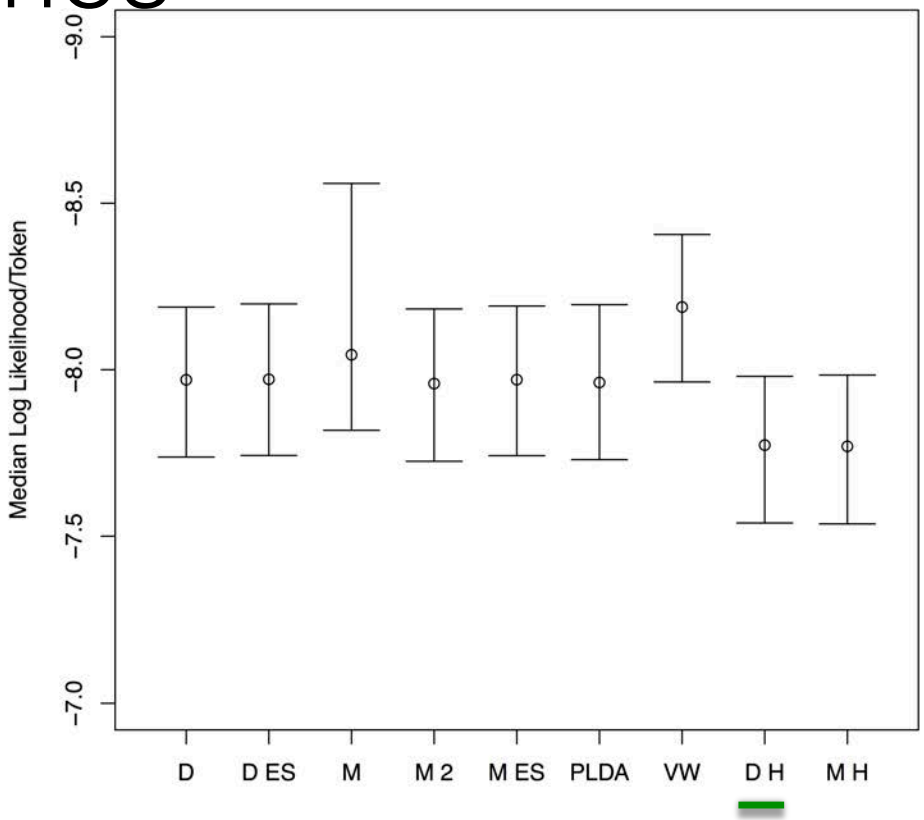


Figure 3.3.: Comparisons of held-out likelihoods on the 20 Newsgroups corpus. The likelihood was estimated using 10-fold cross-validation. The vertical bar represent a 95% confidence interval for the median. The letters *D* and *M* stand for DCA and Mallet respectively. *ES* stands for *early stopping* and *H* for hyper-parameter optimization. Experiments for Mallet have been run several times since the first time it showed some unusual behavior ("M"). Only the result of one additional run is shown here ("M 2").

Marc Zimmermann, 2013

Topic Models: Evaluation of Libraries

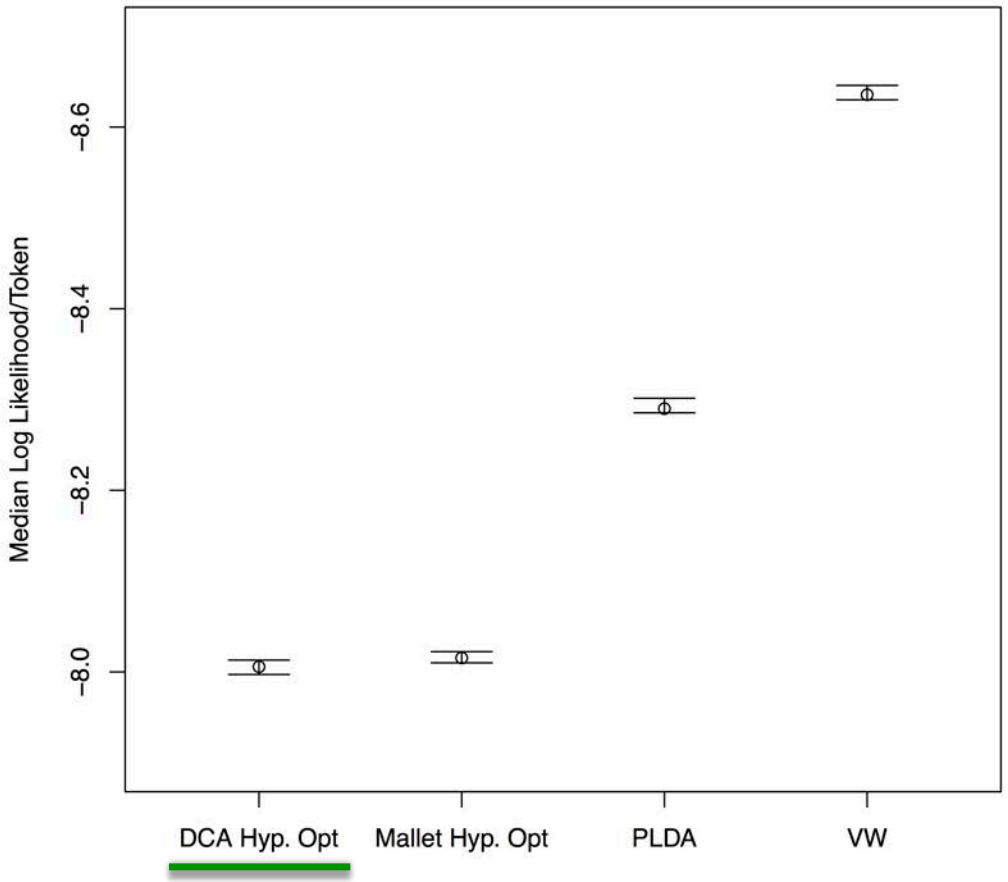


Figure 3.4.: Comparisons of held-out likelihoods on a subset of the PubMed Abstracts corpus. The likelihood was estimated using 10-fold cross-validation. The vertical bar represent a 95% confidence interval for the median.

Marc Zimmermann, 2013

Training for PubMed Model

- using DCA-0202
- 200 topics
- hyperparameters optimization
- trained on 21’034’484 PubMed abstracts
- word minimum frequency: 100
- stopword list (524 common English words)
- vocabulary size: 107’941 words
- iterations (Gibbs sampling): 1000 (~1d, 8 threads)

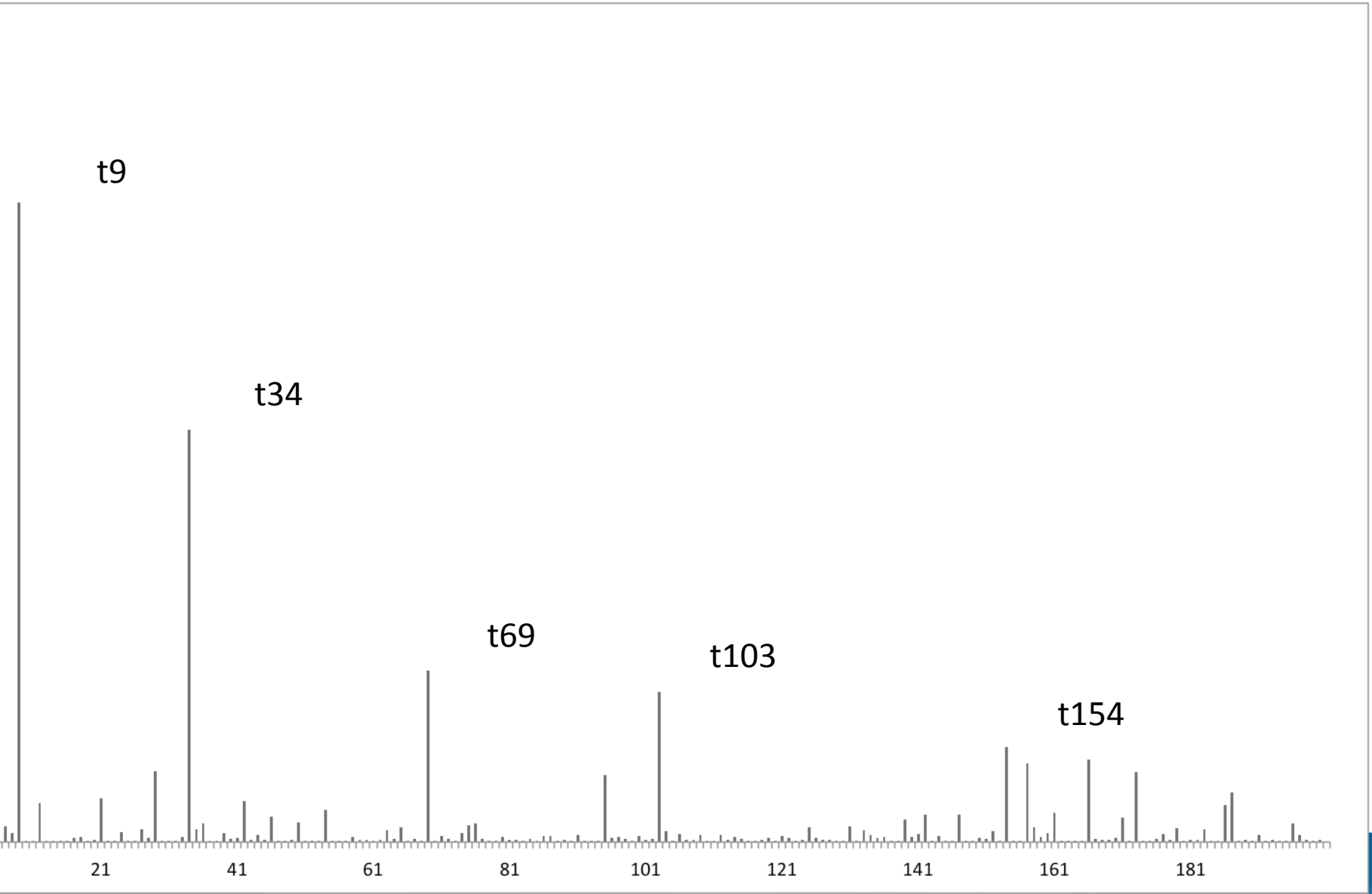
Evaluation on Corpus of Collected PDFs from BBP

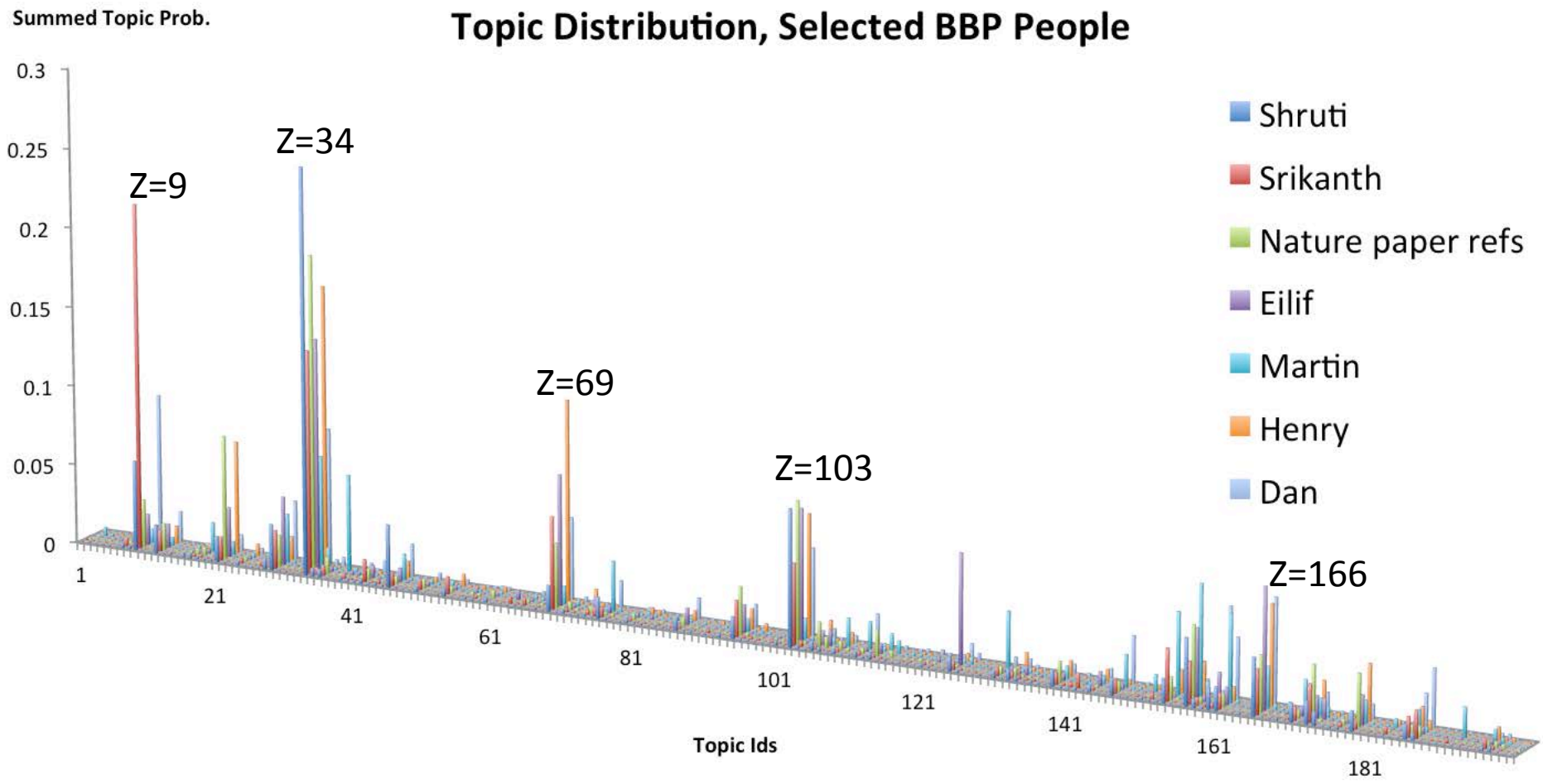
- PubMed id extracted with ***Pdf2PmId*** tool (indexes all abstracts and performs 4-gram text search)

Dataset	# Pdfs	# PmIds Resolved
Eilif	291	208
Martin's zotero	1493	1031
Shruti	839	544
Srikanth	147	101
Dan	363	284
Nature paper refs (zotero)	275	81
Henry's papers*	N.A.	72

*) Henry's paper's PubMed ids resolved from PubMed search

Topic Distribution: Srikanth



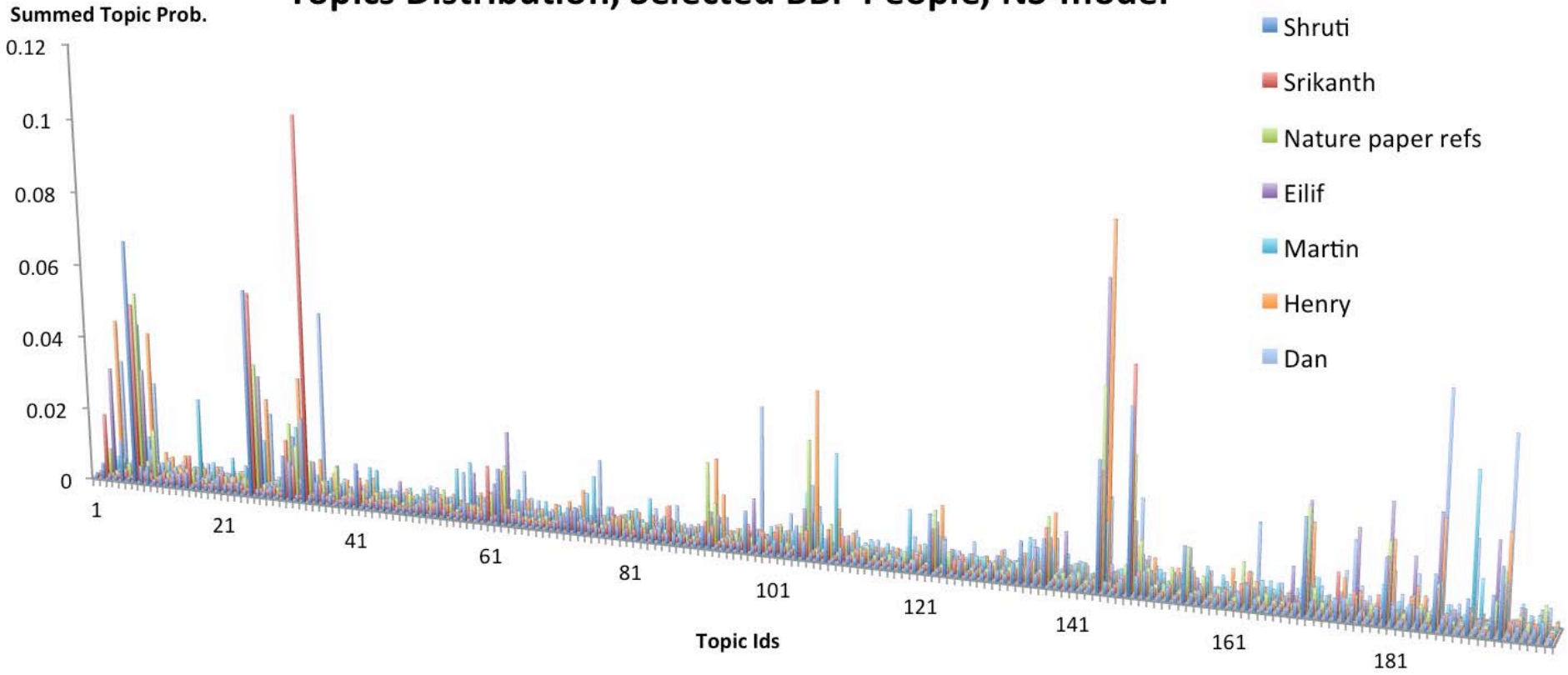


topic 9		topic 34		topic 69		topic 103		topic 166	
current	k+	neurons	glial	model	parameters	stimulation	evoked	system	functional
channel	conductance	cell	astrocytes	data	results	activity	recorded	process	factors
potential	cells	neuronal	layer	analysis	approach	response	motor	mechanisms	state
membrane	action	immunoreac.	synaptic	method	experimental	neurons	stimuli	conditions	function
mv	voltage	brain	olfactory	based	time	electrical	reflex	development	result

Training for PubMed-NS Model

- using DCA-0202
- 200 topics
- hyperparameters optimization
- training on 1'030'546 NS PubMed abstracts
  - get all journals from 1678 pdfs (from BBP people)
  - manual filtering --> 84 journals
  - fetching all ids --> 1'773'358 PubMed Ids
- word minimum frequency: 100
- stopword list (524 common English words)
- vocabulary size: 28'756 words
- using annotation disambiguation
- iterations (Gibbs sampling): 1000 (~12h, 3 threads)

Topics Distribution, Selected BBP People, NS-model



Rank:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Shruti	7	25	148	144	170	33	194	180	122	62
Srikanth	33	148	25	7	144	2	180	30	60	137
Nature	144	7	148	25	105	170	91	180	30	137
Eilif	144	25	7	180	2	186	170	62	193	176
Martin	6	190	108	15	30	144	104	118	56	74
Henry	144	2	105	7	30	186	25	194	170	91
Dan	186	33	194	97	2	7	148	144	190	25

topic 2	**topic 6**	**topic 7**	**topic 25**	**topic 30**	**topic 33**	**topic 105**	**topic 144**	**topic 148**	**topic 170**	**topic 180**	**topic 186**
model	hormone	inhibitory	dendritic	understanding	current	method	network	potential	cortical	activity	synaptic
experimental	estrogen	inhibition	layer	recent	channel	data	neural	action	area	firing	plasticity
data	steroid	excitatory	dendrite	mechanism	potential	algorithm	information	membrane	cortex	unit	long-term
predict	estradiol	synaptic	spine	development	inactivation	approach	input	slice	visual	discharge	potentiation
base	testosterone	transmission	axon	molecular	block	propose	circuit	recording	subcortical	spike	change
prediction	progesterone	interneuron	cell	review	conductance	set	functional	depolarization	neocortex	burst	induction
simulation	level	gabaergic	neuron	provide	potassium	image	processing	amplitude	region	spontaneous	mechanism

Agenda

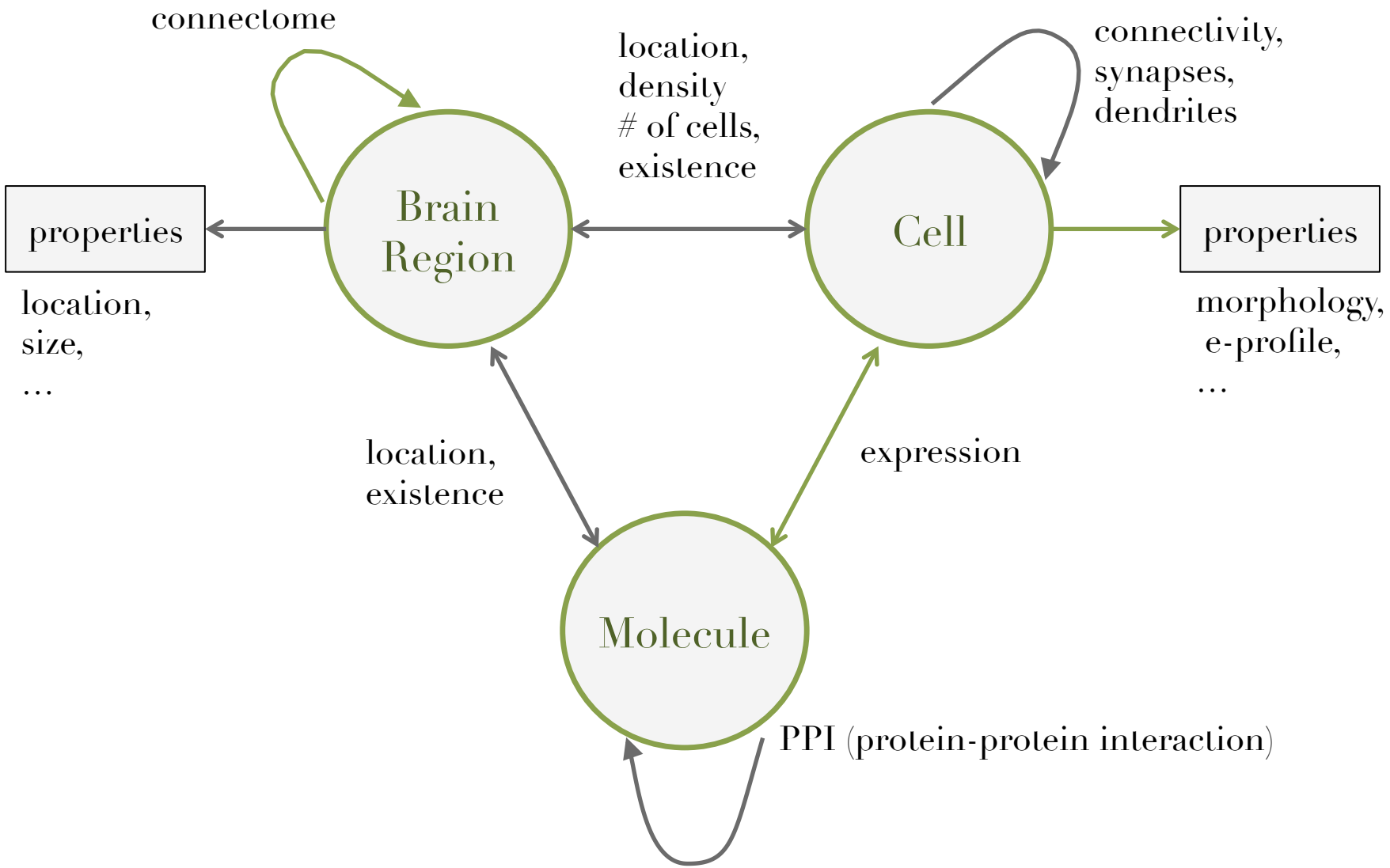
- Introduction: NLP for neuroscience
- braiNER: extracting brain region connectivity from scientific articles
- Agile text mining: neuroNER
- Topic modelling
- Synthesis

Synthesis

Contributions:

- agile text mining
- large scale neuroscience corpora
- bluima, Sherlok
- connectome (braiNER)
- neuroNER

Model of Neuroscientific Entities & Relationships



Future directions

- neuroinformatics platform / OpenMinTed
  - register textual resources
  - continuous integration of new resources
  - register annotations & relationships
  - reuse existing tools & data spaces
  - search throughout annotations and relations

- integrate provenance tracking into neuroscientific model parameters

Thanks

Prof. Henry Markram
Dr. Jean-Cédric Chappelier (EPFL, IC)
Prof. Sean Hill
Dr. Catherine Zwahlen
Dr. Martin Telefont
Dr. Laura Cif (CHUV)
Dr. Xavier Vasques
Dr. Shreejoy Tripathy (UBC, Canada)

Thanks

NI & BBP Team:

Dace Stiebrina
Daniel Keller
Eilif Muller
Emily Clark
Felix Schürmann
Iurii Katkov
James King
Jean-Denis Courcol
Jeff Muller
Julian Shillcock
Katrien Van Look
Marc-Oliver Gewaltig
Martin Ouellet
Michael Reimann
Mohameth Sy
Rafael Nogueira
Ranjan Rajnish (LNMC)
Samuel Kerrien
Srikanth Ramaswamy
Tsolmongerel Papilloud
Vincent Delattre (LNMC)
Werner Van Geit

EPFL Students:

Joëlle Portmann	(Feb - Jun 2012)
Marc Zimmermann	(Sept - Dec 2012)
Orianne Rollier	(Feb - June 2013)
Luca La Spada	(Feb - June 2013)
Samuel Kimoto	(July - Aug 2013)
Philemon Favrod	(Sept - Dec 2013)
Erick Cobos	(Feb - July 2014)
Luca La Spada	(Sept 2014 - March 2015)
Marco Antognini	(Feb 2015 – July 2015)