Supplemental Material

Dimensionality Reduction in the Wild: Gaps and Guidance

M. Sedlmair, M. Brehmer, S. Ingram, and T. Munzner

University of British Columbia, Canada

Abstract

We provide the following supplemental material along with the paper submission "Dimensionality Reduction in the Wild: Gaps and Guidance":

- Appendix A: Full list of interviews and associated usage examples: domain, date, location, and duration of interviews, who interviewed them, publications, documents, & artifacts solicited by interviewees
- Appendix B: Interview foci & questions
- Appendix C: Data analysis examples

Appendix A: Full list of interviews

Example	Domain	Date	Dur.	Loc.	Interviewers	Pubs[1]	Docs
FishPop	fisheries sciences	2010-09-17	2h	remote	MS,SI	[6, 18]	m
MoCap	machine learning	2011-11-23	1h	UBC	MS,MB	[2, 31]	s
Music	media informatics	2011-01-22	0.5h	phone	MS	[4]	m,m,t[8]
Concept[2]	life sciences	2010-11-18	1h	phone	MS,SI,TM		v,e
NPAlgo	machine learning	2010-12-01	1h	phone	MS,TM	[19, 20, 21, 22]	
SeqAln	bioinformatics	2010-04-20	1.5h	phone	MS,SI,TM	[5, 9, 10] [17, 14, 23] [32, 29]	e
GamMdl	machine learning	2010-12-07	2h	UBC	MS	[33]	m
		2011-04-07	0.5h	UBC	MS		
Search	search engine opt.	2011-01-26	1.5h	phone	MS,TM		m,e
ProstCan	bioinformatics	2011-04-04	1.5h	remote	MS	[28]	m
		2011-04-29	4h	remote	MS,SI,TM		
EpiGen	bioinformatics	2010-11-16	2h	remote	MS,HY,TMÖ	[24]	s,s
		2010-12-21	2h	remote	MS,HY,TMÖ		
StrucGen	bioinformatics	2011-04-20	1h	phone	MS,SI,TM	[16]	v
FlockSim	mathematics	2011-03-18	1.5h	SFU	MS,SB	[7, 11, 12]	t[1]
		2011-04-05	1h	SFU	MS,SB,TMÖ	[13, 15]	
CompBio	comp. bio.	2010-10-13	1h	remote	MS,SI		
ChemRel	comp. chem.	2010-10-07	1h	phone	MS,SI,TM		
MedImg	comp. vision	2011-04-05	1.5h	SFU	MS,SB	[3, 25, 26]	m
		2011-04-15	1.5h	SFU	MS,SB	[27, 30]	
TxtDocs	journalism	2012-03-05	3h	UBC	MS,MB,TM		m,w
BoatAct	marine/ocean sci.	2011-01-20	1.5h	phone	MS,SI		m,d,e,v
Polymers[2]	structural chem.	2010-11-18	1h	phone	MS,SI,TM		e
(excluded)	comp. vision	2010-10-06	1h	SFU	MS,HY,TMÖ		

Table 1: Full list of interviews and corresponding usage examples.

The first five entries at the top of Table 1 correspond to the usage examples described in Section 5 of the paper. The single entry at the bottom of the table shows the interview that was excluded from further analysis.

Interviewers: Michael Sedlmair (MS), Matthew Brehmer (MB), Stephen Ingram (SI), and Tamara Munzner (TM); Hamidreza Younesy (HY), Steven Bergner (SB), and Torsten Möller (TMÖ) are with the School of Computing Science at Simon Fraser University (SFU), Burnaby BC, Canada.

Additional documents from interviewees (Docs): unpublished manuscript (m), data (d), presentation slides (s), thesis (t), visualization screenshots (v), additional email correspondence with us (e), web site / blog (w).

Note[1]: Articles published by interviewees and/or referred to us by interviewees.

Note[2]: CONCEPT & POLYMERS were interviewed together.

Appendix B: Interview foci & questions

A: Data (+ Data Analysis)

- How does your data look like?
- One dataset, more datasets?
- What are the major problems, challenges in the data analysis?
- Which information in the data is important for you / what do you read from the data?
- What else do you want to read from the data

B: Task (+ Goals)

- What are you doing?
- What are you working on?
- What are your Goals?
- What is the ultimate goal?
- What data analysis tasks are involved in your work?
- How important is data analysis in your daily work?
- What other tasks apart from data analysis?
- Collaboration or alone?
- What are the questions/hypotheses you try to answer by analyzing your data?

C: Current Practices (Tools), Problems and Challenges

- What are the current tools you use for data analysis?
- What Visualizations are you currently using?
- How is your procedure in analyzing the data with these tools (hypotheses)?
- Good things/ bad things about these tools?
- What are you missing with these tools (perfect analysis tool)

D: Dim Reduction

- Do you use DimRed in your work?
- If not yet, why do you think it is important for you?
- What are your expectations?

E: Patterns of Interest

- Clusters
- Outliers
- Correlation between dimensions (between axis, should be rare after Dim-Red)
- Finding Meaningful LowDim Axes

Appendix C: Data analysis examples

In this section we provide examples from process of data collection and analysis, conducted in the spirit of grounded theory, as described in Section 3 of the paper. Names are obscured to preserve anonymity.

First, interview transcripts (e.g. Figure 1) and notes (e.g. Figure 2), along with publications and documents of our interviewees (e.g. Figure 3) were open-coded to identify concepts.

In Figures 4 and 5, the concepts identified in interview notes and documents, along with their properties and dimensions, were organized into axial codes in our interviewee summaries.

Figures 6, 7, 8, and 9 represent an iterative process of selective coding, the focus on conceptual relationships between axial codes. This led to the development of our descriptive taxonomy, described in Section 4 of the paper.

Figure 10 represents the result of categorizing usage examples as happy, struggle, or fail.

Figure 11 reflects the practices of memoing and theorizing, as prescribed by the methodology of grounded theory, occurring throughout the data analysis process.

sequences, then additional information like fields that can superimpose. ultimate challenge is to learn how the structure relates to the genomes.

me: functional

functional information. all the information you measure or compute about experimental result. 3D structur eitself, various ways that can be represented in less atomistic forms. properties taht can compute or associate with surface.

binding cavity or cleft, for which small molecule or another part of the proetin could be substrate or site of information. those are things where once we have a 3D structure we can create a quant model. 3D structure becomes a quant model.

then question of creating biological hypotheses - given that kind of shape, we want to look at how many observations do we have of that shape or similar shapes that have epxerimentally exhibited a particular biological function.

maybe protein undergoes or facilities bio function of a particular type. if you want to start exploring and eo associations between function and shape, fo rinstance, or properties that may help to explain function that are mapped onto shape, like elctrostatic potential, then that's the kinds of simple view of a particular problem.

to give practical example, very large project protein structure initiative set out to solve 3D structures at a genomic scale. the way the human genome project tried to sequence the human genome. in a relatively uncharacteristic way for structural biology, they solved structures where did not know bi9ochemical function of them. normally it's hypothesis driven, where people try to understand function by looking at 3D structure dat.a but in this case idea was to solve structures that cover lots of sequence space, then to back and analyze late.r so have lots of structures where don't know the function.

computational task to find something you can see or compare about structure or one of its reduced forms where might have a clue about bio context. by looking at thingss wh

functional stuff. is it just out of GO ontology? or do you have your own?

GO describes biochemical functions. normally this particular protein is involved in a pathway that performs something related to cell

Figure 1: An example of an interview transcript (usage example StrucGen).

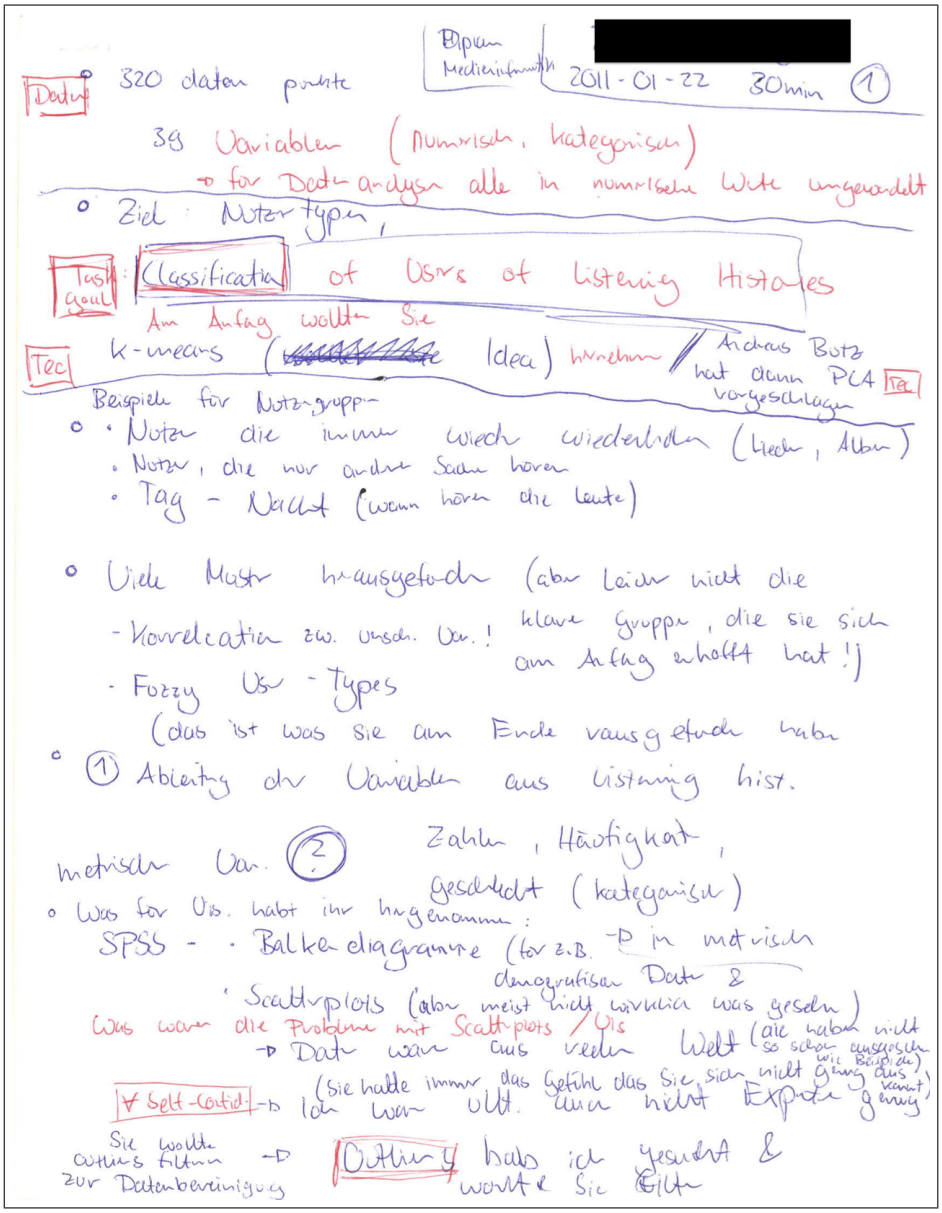
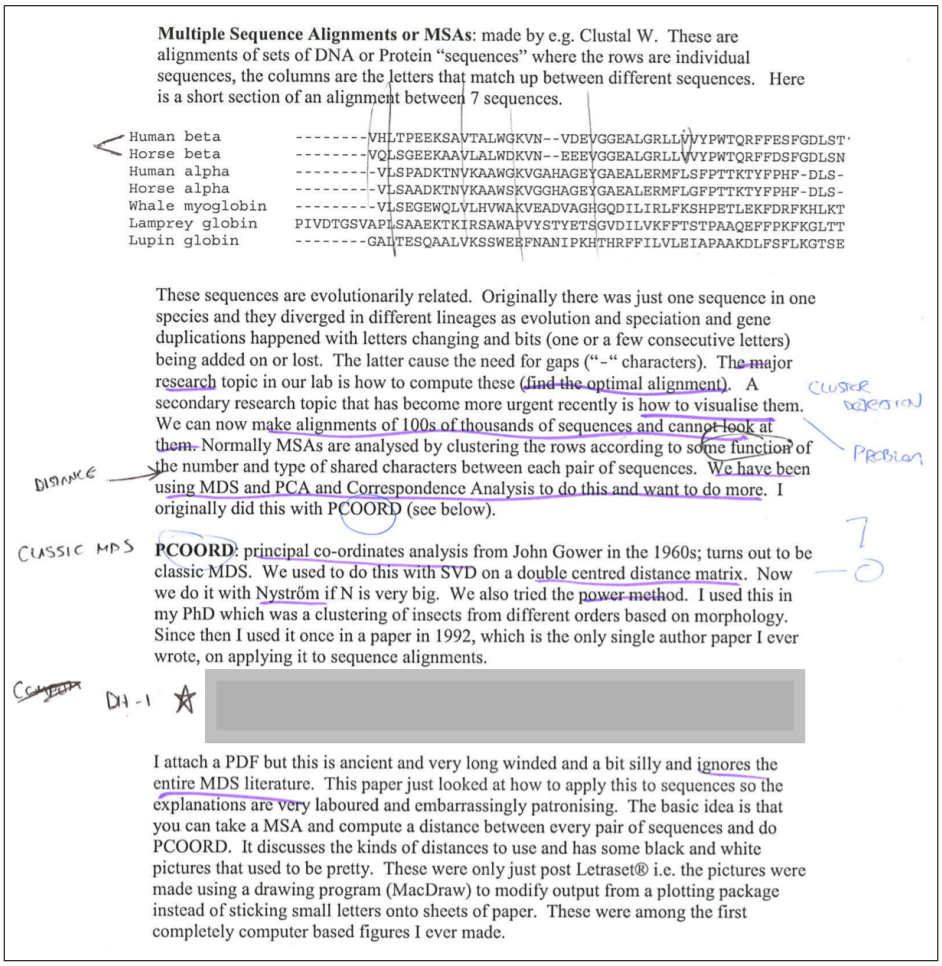


Figure 2: Raw interview notes (blue ink) for usage example Music, with post-hoc open-coding (red ink) to identify concepts.



Multiple Sequence Alignments or MSAs: made by e.g. Clustal W. These are alignments of sets of DNA or Protein "sequences" where the rows are individual sequences, the columns are the letters that match up between different sequences. Here is a short section of an alignment between 7 sequences.

```
Human beta        --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
Horse beta        --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
Human alpha       ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
Horse alpha       ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
Whale myoglobin   ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
Lamprey globin    PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
Lupin globin      --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
```

These sequences are evolutionarily related. Originally there was just one sequence in one species and they diverged in different lineages as evolution and speciation and gene duplications happened with letters changing and bits (one or a few consecutive letters) being added on or lost. The latter cause the need for gaps ("-" characters). The major research topic in our lab is how to compute these (find the optimal alignment). A secondary research topic that has become more urgent recently is how to visualise them. We can now make alignments of 100s of thousands of sequences and cannot look at them. Normally MSAs are analysed by clustering the rows according to some function of the number and type of shared characters between each pair of sequences. We have been using MDS and PCA and Correspondence Analysis to do this and want to do more. I originally did this with PCOORD (see below).

PCOORD: principal co-ordinates analysis from John Gower in the 1960s; turns out to be classic MDS. We used to do this with SVD on a double centred distance matrix. Now we do it with Nyström if N is very big. We also tried the power method. I used this in my PhD which was a clustering of insects from different orders based on morphology. Since then I used it once in a paper in 1992, which is the only single author paper I ever wrote, on applying it to sequence alignments.

I attach a PDF but this is ancient and very long winded and a bit silly and ignores the entire MDS literature. This paper just looked at how to apply this to sequences so the explanations are very laboured and embarrassingly patronising. The basic idea is that you can take a MSA and compute a distance between every pair of sequences and do PCOORD. It discusses the kinds of distances to use and has some black and white pictures that used to be pretty. These were only just post Letraset® i.e. the pictures were made using a drawing program (MacDraw) to modify output from a plotting package instead of sticking small letters onto sheets of paper. These were among the first completely computer based figures I ever made.

Figure 3: Document sent to us by an interviewee (usage example SeqAln), with post-hoc open-coding to identify concepts.

Summary: Interview, [redacted] 2010-09-17

------- personal ----------

Biologist with statistics background.

Domain: Biology - Fisheries.
Type of user: Data analyst for fisheries data.

Work place: DOF.
Overall User Goal: Give recommendations about which harvest control rule to use. Publish papers.

Experience with DR techniques:
- No

---------- *project* ----------

Short project description:
Comparison of mathematical models simulating the behavior of fish populations. All models take a set of parameters such as carrying capacity and productivity. Each model is run with a variety of parameter combinations and Carrie checks if the fish population has died out for this model/parameterization.

Data:
- Models tested with different parameterization
- simulation data (input and output)
- (n/dim unknown so far)

Tasks:
- Evaluating different harvest control rules (HCRs)
- Finding best HCR (i.e. Most robust against a variety of assumptions about fish populations)

Pattern of interest:
- correlations

Goals/Metrics:
- Level of extirpation

Current analysis techniques (Visualization, Stats, DR):
- R
- Matrices of line plots

Current problems / challenges:

Figure 4: Early version for an interviewee summary (usage example FishPop), italicized headings the result of axial coding: organizing concepts, properties, and dimensions.

Summary: [redacted]

Based on:

- 1.5h Skype interview (11-01-26 MS + TM Notes)
- CHI '11 submission
- Email communication

Type of user	Researcher, wants to better understand search browser intent
User goal	Conduct research, publish papers, build better search UI/support for targeted ads
DR User	Non-expert, some background (alumni UBC InfoVis group), recruited machine learning expert to help with classification
Experience with DR techniques	Used correspondence analysis to view structure of query tasks. Exposed to DimStiller, tried to use it but it didn't work out: couldn't understand results.

Brief research description:

A better understanding of search browser query intent is needed to improve search browser user interfaces and support targeted advertising that better matches queries. They seek to build, refine, and validate a search query intent taxonomy, and then classify queries with their validated taxonomy.

Several hundred [redacted]users were recruited to maintain a diary of their search query tasks over the course of several months. Queries were annotated by these users. Search metrics and topical content related to these queries were also collected. The researchers constructed a taxonomy based on this dataset, based on two intrinsic dimensions and 12 query intent clusters. This taxonomy will be used to classify future search queries based on intent.

Use case Instance:

Search query logs decomposed into individual search tasks from potentially hundreds-thousands of [redacted]users need to be mapped to a low-dimensional space. The researchers seek to validate the intrinsic dimensionality of their taxonomy and its search query intent clusters with this data.

Data
- N = 1290 - 1463 search tasks (from 36 users); subsequent field study with 300 people (N much higher) (small by [redacted]standards)
- D = many: (dirty data (sparse, incomplete), 12-13 diary/questionnaire fields, search metrics (clicks, refinement events, abandonment, query length), topical content, mix of categorical and numerical dimensions

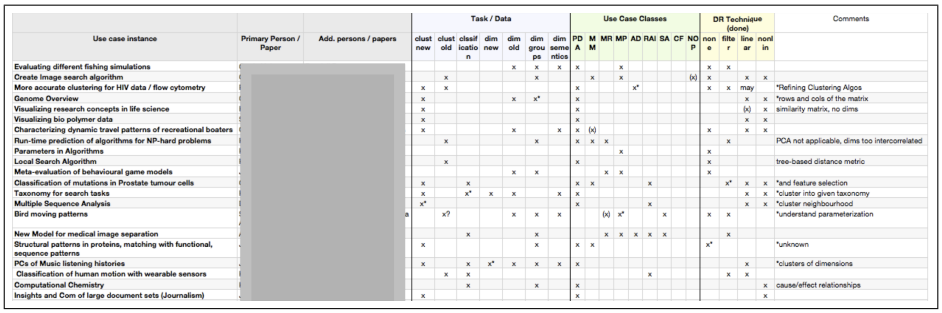
Figure 5: Later version of an interviewee summary (usage example SEARCH), bold headings the result of axial coding: organizing concepts, properties, and dimensions.

Short description of current project	Domain	Experience with DR	Type of user	Data	N	Dim	Task	Pattern of interest	Metrics	Current Analysis Techniques	Current Problems / Challenges
Evaluating different fishing simulation models in terms of harvest control rule	Biology - Fisheries	No	Analyst	Models tested with different parameterization	?	5 to 10	- Evaluating HCR - Finding best HCR	- Correlations between dimensions - Correlations in general	- Insight: Level of extirpation	- R - Line Graphs	- Missing Overview - Tedious interaction
Helping SpinPro Designers to reuse their images.	Vision	Yes	Researcher / Tool designer	Image database (design galleries)	?	25	- Developing new search algorithms - Debugging algorithms	- Clusters / Groupings	- Insight: What is the right algorithm - Insight: Model improvement/ optimization	- Matlab	
?? No explicit examples	Computational chemistry	Yes	Analyst	e.g. compounds with features represented in fingerprints	?	?	- Analyzing data for somebody else - Communicating findings	?	- Finding cause/effect relationships - Communicating cause/effect relationships	- MDS, correlation matrices - Pipeline Plot / Knime - Spotfire / Tableau (rare)	- Sparse data - Missing guidance - Is information really true? - How to trace back data analysis to build better compounds...
Finding better clustering methods for biologists	Biology - Cancer / HIV	Yes	Analysts / Algo designers	Cancer / HIV / FlowCap	20.000	5 to 20	- Comparing clustering algorithms (differences/uniques) - Compare automatic clustering vs manuel results - Developing automatic clustering techniques	- Clusters / Classifications	- Better Analysis - Faster clustering - More accurate clustering	- R - C for speed - Heatmaps - Scatterplots - By biologists: FlowJo tool	- Scalability Communication with biologists - Misslabels (dirty data) - PCA removes important information - PCA hard to interpret for biologists - Interactive clustering with visualization would be great
Recreational boating activity	coastal and ocean resources	little / no	researcher	survey data (ordinal and ratio) from recreational boaters	543 survey respondents	39 survey questions (mix or ordinal (Likert scale) and ratio (i.e. boat length)	- find meaningful clusters of types of recreational boaters - characterize boaters	clusters correlations b/w dims factors that characterize boaters (meaningful new dims)	- insight into clusters	- PCA - MDS - SPLOMs - k-means clustering - focal point clustering	- missing guidance - stability of clusters - the "blob" - dirty data
Epidemics. Providing a better overview over genomes.	Biology - Genomes	Little	Researcher / Tool designer	Heatmaps of genome parts	10.000 - 100.000	160 (8x20)	- Understanding problem - Implementing new tool	- Clusters	- Insight: Getting Overview		- The manual sorting of the rows in the heatmap
Visualizing research concepts in life science	Life science	Little	Researcher / Tool designer	Research concepts	?	20. 000	- Implementing a VisTool	- Clusters	- Insight into Clusters	- co-occurrence matrix of research concepts - force-directed layout: 2d scatterplots	- Scalability - Blob - Are clusters real
Visualizing bio polymer data	Biology - polymer bio.	Yes	Researcher / Tool designer	feature vectors about about polymer molecules	10.000 - 100.000	up to 1.000	- Implementing a VisTool	- Clusters	- Insights	- MDS - force-directed layouts - correlation matrices - mdsi, c, mavis	- Are my clusters real - See clusters in higher D
Different projects with simulating run time of algorithms	- Computational Intelligence -	Yes	Researcher / Algo designer	Features about algos (eg parameters), run time, sampled	100.000 - millions	100-200	- Generation of knowledge about algorithm parameterization, etc. - Come up with new algos, models, etc	- correlation between dimensions - clusters of algorithms (UC3)	- Insight	- Scatterplots and co - mathematical calculations	- usual DR techniques do not work because of the dependencies between his variables (run-time!)
Meta-evaluation of behavioural game models	Behavioural game theory	No	Researcher / Algo designer	Sampling models with different parameters	100.000 - millions	5 to 8	- Getting insights into the problem	- Clusters of dims		- mathematical calculations - Line Graphs - Basically computational meth - Line Graphs, Scatterplots	- Data sampling / generating a data set that includes all maximas
Taxonomy for search tasks	Web search machines	Yes	Researcher	data from a big survey where people kept a diary of their search tasks	300 Subjects	lots of (aspects that people used to describe their tasks)	- Come up with (lowd) axes for the taxonomy - understanding how tasks fit in predefined buckets	- low dimensional axes - clusters in this low dimensional space	- Correct Taxonomy: can other people put their task task in correctly		- what are dims/what are
Classification of mutations in Prostate tumour cells	bioinformatics	Yes	Researcher / Algo designer	genomic and clinical data from cancer patients	600 subjects	1.4 M (genomic exons), 20 clinical variables	- classify potential cancer patients based on their genomic data into categories: aggressive, non-aggressive, none - communicating the results to physicians	- clusters - clusters at each stage of the analysis pipeline	- comparisons against physicians' noisy ground truth data - more accurate classification	- R - filtering, feature selection, PCA, MDS - scatterplots used at end of the pipeline	- is the analysis pipeline valid? trust - batch effect: noisy, perhaps mostly indifferentiated data - need insight at interim stages of the analysis pipeline - noisy ground truth - dimensions highly intercorrelated - runtime complexity
Multiple Sequence Analysis	bioinformatics	Yes	Researcher / Tool designer	protein sequences, each 10-30K letters long	- 100K sequences (use case 1,2) - 5 - 50 sequences (use case 3)	- UC 1,2: dimensions not meaningful (N^2 distance matrix) - UC 3: several hundred to 50K	- UC1: verify alignment of 100K sequences, locate clusters of sequences, drill down - UC2: use DR avoid computing large distance matrix, as scaffolding to sequence alignment algorithm - UC3: analysis of a virus strain where D >> N	- previously unknown clusters - cluster neighbourhood - clusters that facilitate sequence alignment (UC2)	- UC1: cluster detection, hypothesis generation, cluster insight - UC2: scalable cluster alignment	- many DR techniques and custom algorithms (particularly for UC2): MDS, PCA, CA, PCOORD, mBED, FastMap, SeedMap, BGA, MADE4, CIA - phylogenetic trees, reticulate networks	- trees do not scale past 1K - communication of results in publications (no consensus in the research community) - UC3: D >> N
Models of biological aggregate (flock) behaviour	biological mathematics	unknown	Researcher	2D spatial + temporal biological aggregate motion produced by a model and ground truth data from recorded aggregate motion	# of samples = # of aggregates (exact N unknown) (not individuals within an aggregate)	14 model parameters	- given a set of model parameters, compare the predicted aggregate movement with recorded ground truth data - reduce / refine the number of model parameters, perform sensitivity analysis - generate and validate hypotheses	- clusters - original parameters contributing to the model - groups of dimensions	- comparison with ground truth (recorded biological aggregate motion) - model stability, computation speed	- Matlab to filter model parameters - linear and weakly linear analytical methods, numerical simulations, bifurcation analysis, - Paraglide, multidimensional visualization tool	- stability of parameterization: how quickly is the model changing (need vis for sensitivity analysis) - computational processing time - sparse statistics - visualizations difficult to interpret
Models of medical image segmentation parameters	computer graphics	Yes	Researcher / Algo designer	2D / 3D / 4D medical images (SPECT, dynamic PET), model with input and output parameters, ground truth data (manual image segmentations)	50 - 250 samples, very noisy, less if working in 4D (10-15 samples)	6 to 8 input parameters, 6 to 10 output parameters	- model parameterization, refining, reducing # parameters - sensitivity analysis - human guidance and verification of parameter choice and settings - comparison against ground truth	- clusters as means (areas of parameter stability - shapes of clusters - groups of dimensions - cluster neighbourhood	- comparisons against physicians' ground truth data - more accurate classification - model stability, computation speed	- Paraglide, multidimensional visualization tool - feature selection / manual DR	- needs visualization support - overfitting the model - stability of the model - doesn't trust DR algorithms / understand meaning of new dimensions: results in blob problem
Constructing a structural genomics database for proteins	structural genomics / bioinformatics	unknown	Researcher / Tool designer	protein data base including: - 3D structural data - sequence data - functional data (gene ontology)	range from 70K to 15M protein models	huge space, 3 groups of dimensions	- model building - visualization for communication - hypothesis validation and generation: cluster structure space, correlate with known clusters in functional and sequence space	- clusters as means and ends - groups of dimensions	- smallest number of degrees of freedom needed in structural models - insight into clusters	- dendrograms - custom visualization tools	- need for intuitive visualization - combining categorical and quantitative data, establishing distance metrics
Classifying music listener user types on last.fm	media informatics	Little	researcher / analyst	listening histories and demographic data from Last.fm	310 users (sampled from 5000, sparse data)	49-48 dimensions in 7 categories, mix of categorical and numerical	- hypothesis validation: validate existence of hypothesized user groups	- clusters - semantics of reduced	- separable user groups - validated hypotheses	- k-means clustering - PCA	- no clear user groups after performing initial PCA (blob

Figure 6: Early selective coding: focusing on conceptual relationships across summaries.

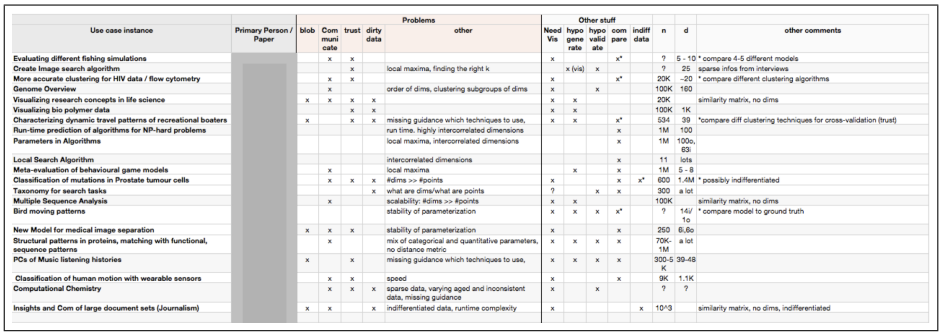
Task Taxonomy										Other stuff						Problems				Data Taxonomy							Notes	
Pure Data Analysis				Make/ Refine Model			Reduce for Algo Input	Curve Fitting	Manifold Finding	Hypotheses		... xxx		Need Vis		The blob (false negative?)	Communicate to others	Trust the results	other	Type	Scale		Struct. of dims		Smooth Manifold	Indifferentiated		done
Clusters		Dims																										
mean	end	mean	end	Generate	Reduce	Paramerterize				Generation	Validation	Explore	Compare	Yes	No						points	dims	semantics of dims	groups of dims				
X				X		X				X	X							X	local maxima, finding the right k	Images - Feature vectors	?	25		X			no case study: not interesting enough?	x
	X									X	X						X		use PCA, but removes too much important information for them	HIV data	20k	~20					no case study: not interesting enough?	x
no			X			X				X	X		X				X	X		sim fishery data	?	05 to 10	X				no case study: vismon	x
unclear										no	X			X			X	X	sparse, varying aged and inconsistent data	different	?	?		X			no case study: very top level interview (not enough info)	x
	X									X		X		X		X	X	X		research concepts	20k	(similarity matrix)					possible case study: descriptive	x
	X									X				X				X	are my clusters real?	feature vectors of polymer molecules	10k-100k	up to 1k					possible case study: descriptive	x
	X									X		X		X			X		clustering of a subgroup of dimensions (rows of heatmaps), order of dims is interesting	genomes	10-100k	160					possible case study: descriptive	x
X		X		X	X	X	X			X	X		X						PCA, MDS, etc do not work for grouped data	sampled in/outs from algos	100k-1M	100-200		X			no case study: no DR	x
		X				X				X			X				X		local maxima	sampled in/outs from game models	100k-1M	5 - 8		X			no case study: no DR	x
	X		X							X						X		X	missing values in the data, clustering is unstable/changes	traffic data from survey	543	39 vars					possible case study: descriptive	x
X				X			X			X			X	X			X	X		Classification of mutations in Prostate tumour cells	~600	1.4M	no	x		maybe	possible case study: predictive	x
	X		X								X	X	X	?	?				what are dims/what are points	diary survey about search tasks	300	a lot	x				possible case study: descriptive	x
X	X						X			X		X		X			X		scalability, #dims >> #points	multiple sequence analysis	100K	(similarity matrix)					possible case-study: predictive	x
X		X		X		X	X					X	X	X		X	X	X	Stability of parameterization, speed	medical image segmentation	250 (s)	6i/6o		X			no case-study - paraglide	x
		X			X	X				X	X		X	X					Stability of parameterization, speed	simulated animal group patterns	?	14i/pattern out	X	X			no case-study - paraglide	x
X	X		X	X	X					X	X		X	X			X		mix of categorical and quantitative parameters, no distance metric	structural genomics / protein structures	70K- 15M	a lot		X			no case study: not enough info	x
	X		X							X	X	X	X	X		X		X		last.fm listener data	310 - 5K	39 - 48	X	X			possible case study: descriptive	x
X							X						X	X			X	X	speed,	classifying human motion	9120	1170					possible case-study: predictive	x
	X									X	X	X	X	X		X	X	X	indifferentiated data, word order not preserved by bag-of-words model	Investigating the Iraqi war logs, wikileaks info dumps	10^3	(similarity matrix)				X	no case study - MoDisco	

Figure 7: Continued selective coding: focusing on conceptual relationships.



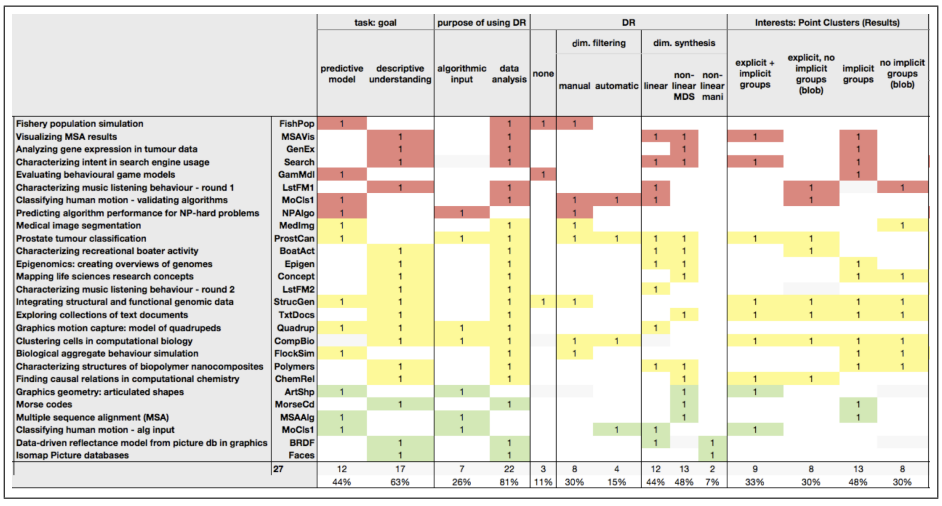
Use case instance	Primary Person / Paper	Add. persons / papers	Task / Data							Use Case Classes								DR Technique (done)				Comments
			clust new	clust old	clssif icatio n	dim new	dim old	dim grou ps	dim seme ntics	PD A	M M	MR	MP	AD	RAI	SA	CF	NO P	non e	filte r	line ar	nonl in	
Evaluating different fishing simulations							x	x	x	x			x						x	x			
Create Image search algorithm				x				x			x		x					(x)	x		x	x	
More accurate clustering for HIV data / flow cytometry			x	x						x				x*					x	x	may		*Refining Clustering Algos
Genome Overview			x				x	x*		x										x		x	*rows and cols of the matrix
Visualizing research concepts in life science			x							x											(x)	x	similarity matrix, no dims
Visualizing bio polymer data			x							x											x	x	
Characterizing dynamic travel patterns of recreational boaters			x				x		x	x	(x)								x		x	x	
Run-time prediction of algorithms for NP-hard problems				x				x		x	x	x								x			PCA not applicable, dims too intercorrelated
Parameters in Algorithms													x						x				
Local Search Algorithm				x						x									x				tree-based distance metric
Meta-evaluation of behavioural game models							x	x			x	x							x				
Classification of mutations in Prostate tumour cells			x		x					x	x				x					x*	x	x	*and feature selection
Taxonomy for search tasks			x		x*	x	x		x	x												x	x	*cluster into given taxonomy
Multiple Sequence Analysis			x*							x					x							x	x	*cluster neighbourhood
Bird moving patterns		a		x?			x	x	x			(x)	x*			x			x	x		x		*understand parameterization
New Model for medical image separation					x			x				x	x	x	x	x				x			
Structural patterns in proteins, matching with functional, sequence patterns			x					x		x	x							x*					*unknown
PCs of Music listening histories			x		x	x*	x	x	x	x											x		*clusters of dimensions
Classification of human motion with wearable sensors				x	x										x				x	x			
Computational Chemistry					x			x		x												x	cause/effect relationships
Insights and Com of large document sets (Journalism)			x							x												x	

Figure 8: Continued selective coding: focusing on conceptual relationships.



Use case instance	Primary Person / Paper	Problems					Other stuff							
		blob	Com muni cate	trust	dirty data	other	Need Vis	hypo gene rate	hypo valid ate	com pare	indiff data	n	d	other comments
Evaluating different fishing simulations			x	x			x			x*		?	5 - 10	* compare 4-5 different models
Create Image search algorithm				x		local maxima, finding the right k		x (vis)	x			?	25	sparse infos from interviews
More accurate clustering for HIV data / flow cytometry			x	x			x			x*		20K	~20	* compare different clustering algorithms
Genome Overview			x			order of dims, clustering subgroups of dims	x		x			100K	160	
Visualizing research concepts in life science		x	x	x	x		x	x				20K		similarity matrix, no dims
Visualizing bio polymer data				x	x		x	x				100K	1K	
Characterizing dynamic travel patterns of recreational boaters		x		x	x	missing guidance which techniques to use,	x	x		x*		534	39	*compare diff clustering techniques for cross-validation (trust)
Run-time prediction of algorithms for NP-hard problems						run time. highly intercorrelated dimensions				x		1M	100	
Parameters in Algorithms						local maxima, intercorrelated dimensions				x		1M	100o, 6k	
Local Search Algorithm						intercorrelated dimensions				x		11	lots	
Meta-evaluation of behavioural game models			x			local maxima		x		x		1M	5 - 8	
Classification of mutations in Prostate tumour cells			x	x	x	#dims >> #points	x			x	x*	600	1.4M	* possibly indifferentiated
Taxonomy for search tasks					x	what are dims/what are points	?		x	x		300	a lot	
Multiple Sequence Analysis			x			scalability: #dims >> #points	x	x				100K		similarity matrix, no dims
Bird moving patterns						stability of parameterization	x	x	x	x*		?	14/ 1o	* compare model to ground truth
New Model for medical image separation		x	x	x		stability of parameterization	x			x		250	6i,6o	
Structural patterns in proteins, matching with functional, sequence patterns			x			mix of categorical and quantitative parameters, no distance metric	x	x	x	x		70K-1M	a lot	
PCs of Music listening histories		x		x		missing guidance which techniques to use,	x	x	x	x		300-5K	39-48	
Classification of human motion with wearable sensors			x	x		speed	x			x		9K	1.1K	
Computational Chemistry			x	x	x	sparse data, varying aged and inconsistent data, missing guidance	x		x			?	?	
Insights and Com of large document sets (Journalism)		x	x		x	indifferentiated data, runtime complexity	x				x	10^3		similarity matrix, no dims, indifferentiated

Figure 9: Continued selective coding: focusing on conceptual relationships.



		task: goal		purpose of using DR		DR						Interests: Point Clusters (Results)			
		predictive model	descriptive understanding	algorithmic input	data analysis	none	dim. filtering: manual	dim. filtering: automatic	dim. synthesis: linear	dim. synthesis: non-linear MDS	dim. synthesis: non-linear mani	explicit + implicit groups	explicit, no implicit groups (blob)	implicit groups	no implicit groups (blob)
Fishery population simulation	FishPop	1			1	1	1								
Visualizing MSA results	MSAVis		1		1				1	1		1		1	
Analyzing gene expression in tumour data	GenEx		1		1					1				1	
Characterizing intent in search engine usage	Search		1		1				1	1		1		1	
Evaluating behavioural game models	GamMdl	1			1	1								1	
Characterizing music listening behaviour - round 1	LstFM1		1		1				1				1		1
Classifying human motion - validating algorithms	MoCls1	1			1		1	1	1				1		
Predicting algorithm performance for NP-hard problems	NPAlgo	1		1			1								
Medical image segmentation	MedImg	1			1		1								1
Prostate tumour classification	ProstCan	1		1	1		1	1	1	1		1	1		
Characterizing recreational boater activity	BoatAct		1		1				1	1			1		
Epigenomics: creating overviews of genomes	Epigen		1		1				1	1				1	
Mapping life sciences research concepts	Concept		1		1					1				1	1
Characterizing music listening behaviour - round 2	LstFM2		1		1				1						
Integrating structural and functional genomic data	StrucGen	1	1		1	1	1					1	1	1	1
Exploring collections of text documents	TxtDocs		1		1					1		1	1	1	1
Graphics motion capture: model of quadrupeds	Quadrup	1	1	1	1				1						
Clustering cells in computational biology	CompBio		1	1	1		1	1				1	1	1	1
Biological aggregate behaviour simulation	FlockSim	1			1		1							1	1
Characterizing structures of biopolymer nanocomposites	Polymers		1		1				1	1				1	1
Finding causal relations in computational chemistry	ChemRel		1		1					1		1	1		
Graphics geometry: articulated shapes	ArtShp	1		1						1		1			
Morse codes	MorseCd		1		1					1				1	
Multiple sequence alignment (MSA)	MSAAlg	1		1						1				1	
Classifying human motion - alg input	MoCls1	1		1				1	1			1			
Data-driven reflectance model from picture db in graphics	BRDF		1		1				1		1				
Isomap Picture databases	Faces		1		1						1				
	27	12	17	7	22	3	8	4	12	13	2	9	8	13	8
		44%	63%	26%	81%	11%	30%	15%	44%	48%	7%	33%	30%	48%	30%

Figure 10: Classifying usage scenarios: happy, struggle, fail

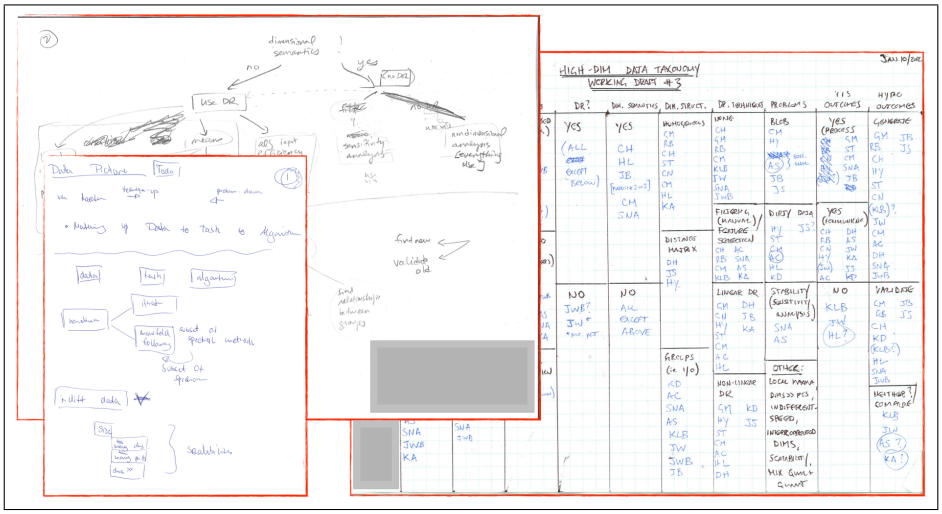


Figure 11: Examples of memos and ongoing theorizing occurring during the course of data analysis.

References

[1] S. N. Abdolyousefi. Equilibria of a Nonlocal Model for Biological Aggregations: Linear Stability and Bifurcation Studies. Master's thesis, Simon Fraser University, 2011.

[2] K. Altun, B. Barshan, and O. Tunçel. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43(10):3605–3620, 2010.

[3] S. Andrews, G. Hamarneh, and A. Saad. Fast random walker with priors using precomputation for interactive medical image segmentation. In *Proc. Intl. Conf. Medical Image Computing and Computer-Assisted Intervention: Part III (MICCAI)*, pages 9–16. Springer-Verlag, 2010.

[4] D. Baur, J. Büttgen, and A. Butz. Listening factors: A large-scale principal components analysis of long-term music listening histories. In *Proc. Conf. Human Factors in Computing Systems (CHI)*. ACM, in press.

[5] G. Blackshields, F. Sievers, W. Shi, A. Wilm, and D. G. Higgins. Sequence embedding for fast construction of guide trees for multiple sequence alignment. *Algorithms for Molecular Biology*, 5(21):1–11, 2010.

[6] M. Booshehrian, T. Möller, R. M. Peterman, and T. Munzner. Vismon: Facilitating analysis of trade-offs, uncertainty, and sensitivity in fisheries management decision making. *Computer Graphics Forum (Proc. EuroVis)*, in press.

[7] J. Buhl, D. J. T. Sumpter, I. D. Couzin, J. J. Hale, E. Despland, E. R. Miller, and J. Simpson, S. From disorder to order in marching locusts. *Science*, 312(5778):1402–1406, 2006.

[8] J. Büttgen. What's in a History? A Large-Scale Statistical Analysis of Last.FM Data. Master's thesis, Ludwig-Maximilians Universitat Munchen, 2010.

[9] A. C. Culhane, G. Perriere, and D. G. Higgins. Cross-platform comparison and visualization of gene expression data using co-inertia analysis. *BMC Bioinformatics*, 4(59), 2003.

[10] A. C. Culhane, J. Thioulouse, G. Perriere, and D. G. Higgins. MADE4: an R package for multivariate analysis of gene expression data. *Bioinformatics*, 21(11):2789–2790, 2005.

[11] R. Eftimie. *Modeling Group Formation and Activity Patterns in Self-Organizing Communities of Organisms*. PhD thesis, University of Alberta, 2008.

[12] R. Eftimie. Hyperbolic and kinetic models for self-organized biological aggregations and movement: a brief review. *Journ. Mathematical Biology*, 2011.

[13] R. Eftimie, G. de Vries, M. A. Lewis, and F. Lutscher. Modeling group formation and activity patterns in self-organizing collectives of individuals. *Bulletin of Mathematical Biology*, 69(5):1537–1565, 2007.

[14] A. Fagan, A. C. Culhane, and D. G. Higgins. A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics*, 7(13):2162–2171, 2007.

[15] R. C. Fetecau and R. Eftimie. An investigation of a nonlocal hyperbolic model for self-organization of biological groups. *Journ. Mathematical Biology*, 61(4):545–579, 2010.

[16] M. J. Gabanyi, P. D. Adams, K. Arnold, L. Bordoli, L. G. Carter, J. Flippen-Andersen, L. Gifford, J. Haas, A. Kouranov, W. a. McLaughlin, D. I. Micallef, W. Minor, R. Shah, T. Schwede, Y.-P. Tao, J. D. Westbrook, M. Zimmerman, and H. M. Berman. The structural biology knowledgebase: a portal to protein structures, sequences, functions, and methods. *Journ. Structural and Functional Genomics*, 12(2):45–54, 2011.

[17] D. G. Higgins. Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets. *Computer Applications in the Biosciences (CABIOS)*, 8(1):15–22, 1992.

[18] C. Holt and M. Bradford. Evaluating benchmarks of population status for Pacific salmon. *N. American Journ. Fisheries Management*, 31(2):363–378, 2011.

[19] F. Hutter, T. Bartz-Beielstein, H. Hoos, K. Leyton-Brown, and K. Murphy. Sequential Model-Based Parameter Optimization: an Experimental Investigation of Automated and Interactive Approaches. In T. Bartz-Beielstein, M. Chiarandini, L. Paquete, and M. Preuss, editors, *Experimental Methods for the Analysis of Optimization Algorithms*, pages 363–414. Springer-Verlag, 2010.

[20] F. Hutter, H. Hoos, and K. Leyton-Brown. Tradeoffs in the empirical evaluation of competing algorithm designs. *Annals of Mathematics and Artificial Intelligence*, 60:65–89, 2010.

[21] F. Hutter, H. Hoos, K. Leyton-Brown, and K. Murphy. Time-bounded sequential parameter optimization. In C. Blum and R. Battiti, editors, *Learning and Intelligent Optimization*, volume 6073 of *Lecture Notes in Computer Science*, pages 281–298. Springer, 2010.

[22] K. Leyton-Brown, E. Nudelman, and Y. Shoham. Empirical hardness models: Methodology and a case study on combinatorial auctions. *Journ. ACM*, 56(4):22:1–22:52, 2009.

[23] S. F. Madden, S. B. Carpenter, I. B. Jeffery, H. Bjorkbacka, K. A. Fitzgerald, L. A. O'Neill, and D. G. Higgins. Detecting microRNA activity from gene expression data. *BMC Bioinformatics*, 11(257), 2010.

[24] C. B. Nielsen, S. D. Jackman, I. Birol, and S. J. M. Jones. ABySS-Explorer: visualizing genome sequence assemblies. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis)*, 15(6):881–888, 2009.

[25] A. Saad, G. Hamarneh, and T. Möller. Exploration and visualization of segmentation uncertainty using shape and appearance prior information. *IEEE Trans. Visualization and Computer Graphics (Proc. Vis)*, 16(6):1366–1375, 2010.

[26] A. Saad, G. Hamarneh, T. MÃűller, and B. Smith. Kinetic modeling based probabilistic segmentation for molecular images. In D. Metaxas, L. Axel, G. Fichtinger, and G. SzÃľkely, editors, *Medical Image Computing and Computer-Assisted Intervention âĂŞ MICCAI 2008*, volume 5241 of *Lecture Notes in Computer Science*, pages 244–252. Springer, 2008.

[27] A. Saad, T. Möller, and G. Hamarneh. ProbExplorer: Uncertainty-guided Exploration and Editing of Probabilistic Medical Image Segmentation. *Computer Graphics Forum*, 29(3):1113–1122, 2010.

[28] Y. Saeys, I. n. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[29] W. Shi, F. Lei, C. Zhu, F. Sievers, and D. G. Higgins. A complete analysis of HA and NA genes of influenza A viruses. *PloS ONE*, 5(12):e14454, 2010.

[30] T. Torsney-Weir, A. Saad, T. Möller, B. Weber, H. C. Hege, J. M. Verbavatz, and S. Bergner. Tuner: principled parameter finding for image segmentation algorithms using visual response surface exploration. *IEEE Trans. Visualization and Computer Graphics (Proc. Vis)*, 17(12):1892–1901, 2011.

[31] O. Tunçel, K. Altun, and B. Barshan. Classifying human leg motions with iniaxial piezoelectric gyroscopes. *Sensors*, 9(11):8508–8546, 2009.

[32] I. M. Wallace and D. G. Higgins. Supervised multivariate analysis of sequence groups to identify specificity determining residues. *BMC Bioinformatics*, 8(135), 2007.

[33] J. Wright and K. Leyton-Brown. Beyond equilibrium: Predicting human behavior in normal-form games. In *Proc. AAAI Conf. Artificial Intelligence (AAAI)*, pages 901–907, 2010.