Jarvis: A Multimodal Visualization Tool for Bioinformatic Data

Mark Hutchens[1], *Nikhil Krishnaswamy*[1],
Brent Cochran[2] and James Pustejovsky[1]
[1]Brandeis University, [2]Tufts University School of Medicine

HCI International 2020
July 24, 2020

Motivations

- Curated biological datasets can be challenging to navigate and explore;
  - e.g., size of protein-protein interaction networks.
- Information visualization is a valuable navigation and understanding tool
  - provided the visualization has the ability to manipulate large quantities of data;
  - including interactions between different visualization techniques (Tao et al., 2004; O'Halloran et al., 2018).

Motivations

- Bioinformatic naming schemes and ontologies discourage speech-based interactions;
  - e.g., difficulty pronouncing or resolving entities;
  - "angiotensin-converting enzyme 2" → **ACE2** ("ace-two")
- Other modalities may be better at grounding certain types of information:
  - e.g., deictic gesture to locations.

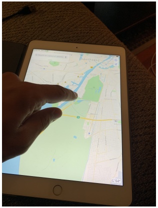


Figure: "We should be *here*."

Motivations

- Multimodal methods for manipulating data can be helpful, e.g.,
  - Voice commands with demonstratives such as “this” and “that”;
  - Haptic interface can specify intended targets.
- **Jarvis** combines speech and haptic controls to encourage robust and flexible Q+A interactions;
  - Test use case: Biocuration data;
  - More generally, data exploration with multimodal queries, e.g.,
    - language input, pointing, swiping, tapping, etc.
- We hope to enhance the navigability and potential for discovery of results returned through complex queries.

Prior Work

- Tablets/smartphones made haptic interfaces commonplace;
- Improved speech recognition did the same for voice commands.
- Integrative crossmodal interfaces (e.g., localizing to maps) (Johnston, 2009; Selfridge and Johnston, 2015; Johnston, 2019).
- Analysis of both biological and biomedical literature (e.g., Kim, 2017; Lee et al., 2020).
- Language interaction over biocuration datasets (Friedman et al., 2017; Gyori et al., 2017; McDonald et al., 2016; Todorov et al., 2019 - `http://pathwaymap.indra.bio`).

Prior Work

- Visualization of dense data for bioinformatics via clustergrams (Schonlau, 2002).
- Dimensionality reduction through manifold approximation and principal component analysis (Tao et al., 2004; Clark and Ma'ayan, 2011; McInnes and Healy, 2018).
- Other visualization tools e.g., Cytoscape, ProViz (Shannon et al., 2003; Jehl et al., 2016).
- Jarvis brings together multiple modalities, background data from bioinformatic literature, and the aforementioned visualization techniques.

Architecture

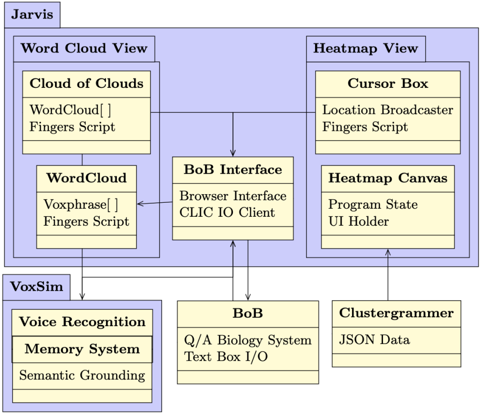
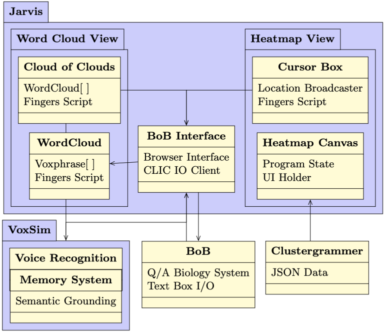


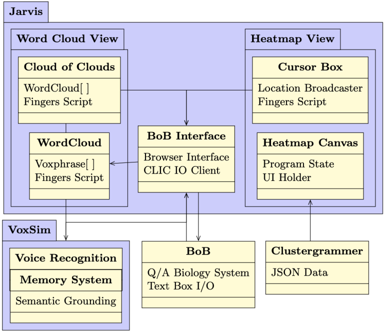
Figure: The main components of Jarvis

Architecture



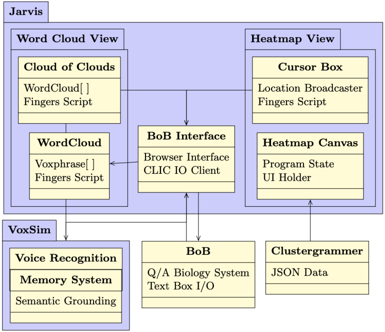
- **VoxSim**: semantic visualization engine that facilitates manipulating visualized objects.
- **BoB**: dialogue system for biological Q+A.
- **Clustergrammer**: input data about e.g., gene-protein interactions in the form of clustergrams.
- **Heatmap Canvas**: visualizes the heatmap from Clustergrammer data.

Architecture



- User can zoom on a desired location.
- Data within the bounds is saved and sent to BoB.
- **Voice Recognition** recognizes the query and semantically grounds entities in it to entities in the world/heatmap data.
- BoB processes query and returns subset of the genes it receives.

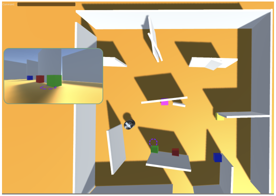
Architecture



- **Word Cloud View** receives both the original set and subset (Cloud 1: visualized full set; Cloud 2: visualized subset).
- Individual proteins can be manipulated in the Word Cloud View.
- Subsequent questions can be asked about those proteins.

VoxSim

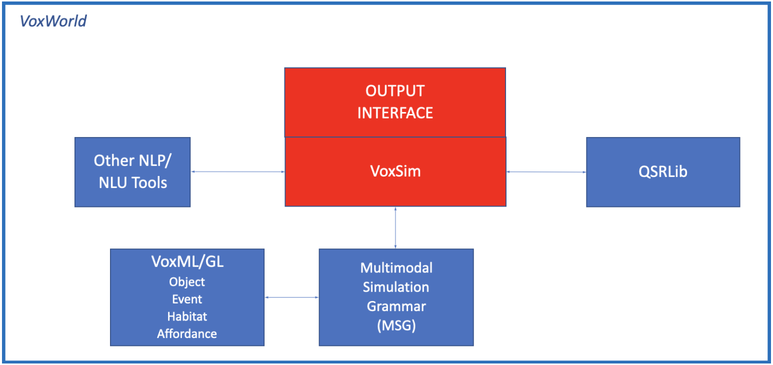
- Unity game engine-based event simulator used to develop intelligent agent behaviors;
  - Typical use-case: interactive embodied agents.



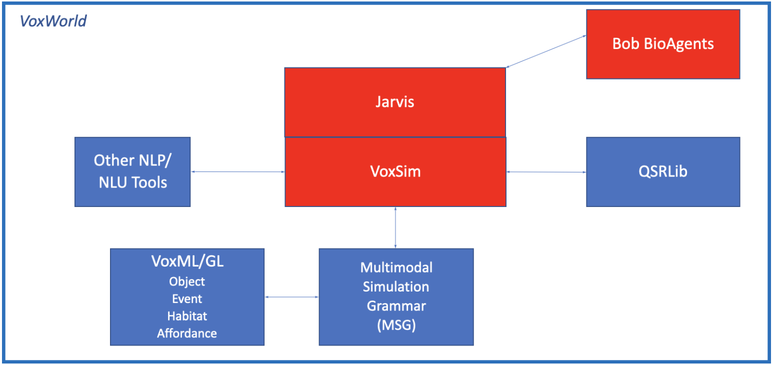
VoxSim

- VoxSim is built on the modeling language VoxML (Pustejovsky and Krishnaswamy, 2016);
  - VoxML encodes the semantics of visualized lexical items (*voxemes*);
  - Allows visualized objects to be manipulated in 2D and 3D space.
- Jarvis exploits the abstract properties of voxemes to render elements from the underlying dataset as manipulable objects to facilitate data exploration.

VoxSim


VoxWorld
OUTPUT INTERFACE
VoxSim
Other NLP/ NLU Tools
QSRLib
VoxML/GL
Object
Event
Habitat
Affordance
Multimodal Simulation Grammar (MSG)

VoxSim


VoxWorld
Bob BioAgents
Jarvis
Other NLP/ NLU Tools
VoxSim
QSRLib
VoxML/GL
Object
Event
Habitat
Affordance
Multimodal Simulation Grammar (MSG)

Usage

- Objects are made interactable through VoxSim and the *Fingers* Unity asset, which parallels native iOS gestures.
- Speech recognition via Google SR.
  - VoxSim supports arbitrary 3rd-party endpoints.
- Heatmap data via Clustergrammer;
  - Groups hierarchically-clustered heatmaps from gene expression data and saves them to JSON.
  - Heatmap visualization algorithm is also adapted from Clustergrammer.

Language Input

- Natural language processing and question answering through BoB biocuration system (Burstein et al., 2019).

USER:	Create the gene set. [*Passes a JSON structure containing all the genes selected from the visual heatmap interface.*]
BoB:	Okay.
BoB:	I created the gene-set selection with 7 items.
BoB:	What would you like to do next?
USER:	Which of these are transcription factors?
BoB:	Of those 7 genes, PLAGL1 is a transcription factor.

Table: A typical exchange with BoB

Haptic Control

- Use of gestures depends on the visualized context and technique (e.g., heatmap or word cloud).
- Some gestures may mean different things in different contexts.
- The gesture *inhabits* the visual context:
  - $H_{[1]}$ = in heatmap; $H_{[2]}$ = in word cloud.

Gesture	**Heatmap Interpretation**	**Word Cloud Interpretation**
Tap	—	Semantically ground word
Swipe	Swap View	Swap View
Pan	Move selection box	—
Scale	Resize selection box	Zoom camera
Rotate	—	Rotate Word Cloud
Long Press	—	Semantically ground cloud

Table: Gestures available in Jarvis

Multimodal Integration

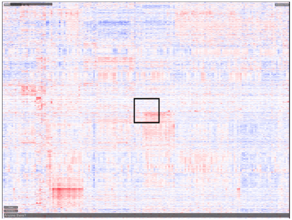


Figure: A heatmap of gene expression vs. tissue samples with UI overlaid

- This heatmap represents gene expression over tumor samples (rows: genes; columns: source tissue).
- Haptics on tablet drag and resize the selector box;
- Voice (e.g., "zoom in here") zooms in, grounds the selected area to the data.

Multimodal Integration

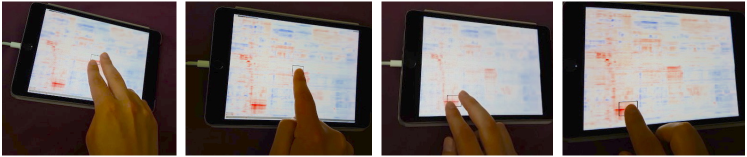


Figure: Interacting with Jarvis interface

- With a query, e.g., "Which of these are transcription factors?" the system reads the currently selected genes and passes the list and query to BoB;
- The resultant subset is passed back to Jarvis for visualization.

Multimodal Integration

Figure: Clouds of protein names for manipulation

- The two lists are then visualized as word clouds;
    - Each word in the cloud is individually interactable.
- Each word's position in the clouds is determined by factors in the underlying data, such as frequency of occurrence or similarity to another data point.

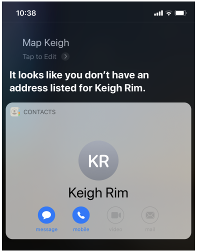
Multimodal Integration

Figure: Clouds of protein names for manipulation

- The clouds respond to haptic input as well;
  - A subset of the cloud, e.g. which proteins are transcription factors, may be pulled out and manipulated independently.
- The cloud can be reordered to correspond to relationships with a selected word.

Multimodal Integration

- The nature of bioinformatic data, particularly protein and gene names, poses a problem for purely speech-based systems;
- Gene names may be initialisms or difficult to pronounce;
  - e.g., “MAPK” has a conventionalized pronunciation (“map-kay”).
  - Say this to a smart phone:



Multimodal Integration

- Navigating bio domain via speech alone is difficult.
- Providing text input helps but is time-consuming and doesn't solve the problem if user doesn't know how to phrase requests.
- Large bioinformatic datasets are typically presented visually (heatmaps, graphs, word clouds);
  - Not necessarily easy to interact with these linguistically.
- Haptics to indicate regions or entities of interest allows simpler language ("this," "that," "these," etc.);
  - Obviates the need to pronounce or spell out entity references.
- Makes navigation more tractable for a large dataset.
- Allows for less discursive language to ask the same question.

Evaluation

- Late-breaking paper: evaluation still planned.
- Goal of Jarvis in this use case is well-defined: to enable biologists to accomplish novel discoveries in large datasets;
  - We can evaluate the usability of the system in accomplishing this task.
- We can also evaluate interactive capabilities for data exploration over arbitrary large datasets;
- Strictly biological data is not a requirement for useful multimodal exploration, it is simply an illustrative use case.

Interactive Usability

- We propose a multimodal evaluation method based on Krishnaswamy and Pustejovsky, 2018.
- An interaction consists of timestamped "moves" coded by participant and modality (S for speech, G for gesture, A for action.

1	JARVIS_A	CREATE_HEATMAP(data[])	0.000000
2	USER_G	PAN_TO (<.14674;.24371>)	1.145281
3	USER_S	"Which of these are transcription factors?"	2.452981
4	BoB_S	"I am having trouble, possibly because I don't know what 'these' refers to."	5.803915
5	BoB_s	"I don't know what genes you mean."	7.818170
6	USER_S	"Create the gene set."	8.642095
7	BoB_S	"Okay."	10.041973
8	BoB_S	"I created the gene-set selection with 7 items."	12.803915
9	BoB_S	"What would you like to do next?"	14.500183
10	USER_S	"Which of these are transcription factors?"	15.661427
11	JARVIS_A	CREATE_WORDCLOUD(geneset[])	18.891054
12	JARVIS_A	CREATE_WORDCLOUD(subset[])	18.891054
13	BoB_S	"Of those 7 genes, PLAGL1 is a transcription factor."	18.891054

Table: Sample interaction log.

Interactive Usability

- Moves 0-1: Jarvis presents heatmap, facilitating the user selecting a region by haptic panning;
- We can quantify time between data presentation and user interacting with it.
    - A long delay may signal confusion on the part of the user, e.g., in how to use the system and/or what to do with data presentation.
- Dialogue breaks down at move 3: user says something that BoB does not understand, so BoB gives the user a reason why.
- User responds with new instruction that grounds the demonstrative “these” to a particular set.
- Jarvis facilitated repair at move 6.

Fidelity of Data Transfer

- **Speech Recognition**: Have users execute a scripted dialogue that provides a known ground truth, then compare the recognized input to that reference using, e.g., BLEU score (Papineni et al., 2002).
- **Semantic Grounding**: Look at "blocks" in the log bounded by moves that negate a prior move and redirect the interaction to new focus objects or actions.
    - Information is being correctly grounded within a block.
    - Longer blocks means better continuous grounding performance.
    - Can assess grounding through language and grounding through haptics.

Fidelity of Data Transfer

- **Visualization Accuracy**:
  - Match list of genes passed to BoB to list of visualized gene names.
  - Have user select same region multiple times through mouse or haptics and calculating overlap between successive selections.
  - Correlate region selection to underlying data by calculating overlap in retrieved data over multiple selections.

Conclusion

- Jarvis, combines speech and haptic control with a robust biocuration dialogue system;
- These features encourage smooth interactions over complex data.
- The underlying mechanism that enables the integration of the two distinct modalities is the transformation of data into a manipulable object.
- This allows domain specialists to navigate through the data using multiple grounding techniques.
- The same underlying interface is adaptable to multiple datasets potentially in multiple domains.

Future Work

- 3D visualizations of the heatmap view add an additional dimension
  - Useful for representing time sequence data, e.g., single-cell or cell-cycle proliferation data.
- More extended commands for data manipulation.
  - e.g., double- or triple-taps to highlight relationships or rearrange elements.
- More in-depth gene-to-gene relationships in the word cloud visualization.
  - e.g., grouping genes by known clusters, similar protein encoding, or generalized functions.
- Better integration between Fingers and VoxSim.
  - Use the VoxSim contextual memory to track objects the user has previously interacted with.
- "Save state" in a visualization and revisit it (e.g., the "swipe" gesture to swap views).

Acknowledgments

- Laurel Bobrow, Robert Bobrow, Mark Burstein, David McDonald, and Matthew McLure (Smart Information Flow Technologies)
- Benjamin Gyori and John A. Bachman (Harvard Medical School)
- This work is supported in part by US Defense Advanced Research Projects Agency (DARPA), Contract W911NF-15-C-0238; and DTRA grant DTRA-16-1-0002; Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Acknowledgments

Thank you!

References I

Burstein, Mark et al. (2019). "Using Multiple Contexts to Interpret Collaborative Task Dialogs". In: *Advanced In Cognitive Systems*.

Clark, Neil R. and Avi Ma'ayan (2011). "Introduction to Statistical Methods for Analyzing Large Data Sets: Gene-Set Enrichment Analysis". In: *Science Signaling* 4.190, tr4–tr4. ISSN: 1945-0877. DOI: 10.1126/scisignal.2001966. eprint: https://stke.sciencemag.org/content/4/190/tr4.full.pdf. URL: https://stke.sciencemag.org/content/4/190/tr4.

Friedman, Scott et al. (2017). "Learning by reading: Extending and localizing against a model". In: *Advances in Cognitive Systems* 5, pp. 77–96.

References II

Gyori, Benjamin M et al. (2017). "From word models to executable models of signaling networks using automated assembly". In: *Molecular systems biology* 13.11.

Jehl, Peter et al. (Apr. 2016). "ProViz,Äîa web-based visualization tool to investigate the functional and evolutionary features of protein sequences". In: *Nucleic Acids Research* 44.W1, W11–W15. ISSN: 0305-1048. DOI: 10.1093/nar/gkw265. eprint: https://academic.oup.com/nar/article-pdf/44/W1/W11/18787253/gkw265.pdf. URL: https://doi.org/10.1093/nar/gkw265.

Johnston, Michael (2009). "Building multimodal applications with EMMA". In: *Proceedings of the 2009 international conference on Multimodal interfaces*, pp. 47–54.

References III

- Johnston, Michael (2019). "Multimodal integration for interactive conversational systems". In: *The Handbook of Multimodal-Multisensor Interfaces: Language Processing, Software, Commercialization, and Emerging Directions-Volume 3*, pp. 21–76.
- Kim, Jin-Dong (2017). *Biomedical natural language processing*.
- Krishnaswamy, Nikhil and James Pustejovsky (2018). "An evaluation framework for multimodal interaction". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Lee, Jinhyuk et al. (2020). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4, pp. 1234–1240.

References IV

McDonald, David et al. (2016). "Extending biology models with deep NLP over scientific articles". In: *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

McInnes, Leland and John Healy (2018). "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *ArXiv* abs/1802.03426.

O'Halloran, Kay L et al. (2018). "A digital mixed methods research design: Integrating multimodal analysis with data mining and information visualization for big data analytics". In: *Journal of Mixed Methods Research* 12.1, pp. 11–30.

References V

Papineni, Kishore et al. (2002). "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pp. 311–318.

Pustejovsky, James and Nikhil Krishnaswamy (2016). "VoxML: A Visualization Modeling Language". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4606–4613.

Schonlau, Matthias (2002). "The clustergram: A graph for visualizing hierarchical and nonhierarchical cluster analyses". In: *The Stata Journal* 2.4, pp. 391–402.

References VI

Selfridge, Ethan and Michael Johnston (2015). "Interact: Tightly-coupling Multimodal Dialog with an Interactive Virtual Assistant". In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 381–382.

Shannon, P et al. (Nov. 2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks". In: *Genome Research* 13.11, pp. 2498–2504. DOI: 10.1101/gr.1239303.

Tao, Ying et al. (2004). "Information visualization techniques in bioinformatics during the postgenomic era". In: *Drug Discovery Today: BIOSILICO* 2.6, pp. 237–245.

References VII

Todorov, Petar V et al. (2019). "INDRA-IPM: interactive pathway modeling using natural language with automated assembly". In: *Bioinformatics* 35.21, pp. 4501–4503.