# Jarvis: A Multimodal Visualization Tool for Bioinformatic Data

Mark Hutchens[1], Nikhil Krishnaswamy[1],
Brent Cochran[2], and James Pustejovsky[1]
{mhutchens,nkrishna,jamesp}@brandeis.edu
brent.cochran@tufts.edu

[1]Brandeis University, Waltham, MA USA
[2]Tufts University School of Medicine, Boston, MA USA

**Abstract.** In this paper we present Jarvis, a multimodal explorer and navigation system for biocuration data, from both curated sources and text-derived datasets. This system harnesses voice and haptic control for a bioinformatic research context, specifically manipulation of data visualizations such as heatmaps and word clouds showing related terms in the dataset. We combine external speech systems with Clustergrammer [1] for the generation of bioinformatic queries, the BoB interface [2] for answering queries in that domain, and the VoxML framework [12] for manipulating the results and semantic grounding. We deploy the resulting system to iOS on an iPad for use by researchers over a test dataset of gene expression in tumor samples. The intent is to integrate multimodal control (here voice and haptics), so as to facilitate interaction with and analysis of data, taking advantages of using both modalities.

**Keywords:** multimodal · bioinformatics · haptics · visualization

## 1 Motivations

Due to the size of current curated biological datasets, such as protein-protein interaction networks, navigating and exploring the data in such collections can be a challenge. Information visualization is a valuable technique to navigate and cogently understand it, and the visualization should have the ability to smoothly manipulate large quantities of data in a variety of ways, including interactions between different visual techniques [16, 22].

At the same time, the naming schemes and underlying ontologies used in bioinformatics datasets, such as those for genes/proteins, discourage pure voice-based interactions for understanding queries (for instance, "angiotensin-converting enzyme 2" is usually referred to by its acronym, ACE2, which sounds like the phrase "ace two," a term difficult to ground semantically directly from a domain-independent speech model). Additional modalities, e.g., haptics indicating regions in which these terms appear, simplifies grounding these entities to the data, hence multimodal methods for manipulating the data can be helpful. Voice

commands can include demonstratives such as "this" and "that", while a haptic interface can specify intended targets.

Our system, named Jarvis, combines speech and haptic controls to encourage more robust and flexible question and answer interactions over biocuration data. More generally, this platform enables a user to navigate and explore any dataset using multimodal queries, with more traditional language input as well through pointing, swiping, tapping, and other haptic gestures. As we demonstrate, the modality of expression for a query or part of a query varies considerably, depending on the content and context. That is, sometimes it is more appropriate (and easier) to simply point to a region on a heatmap rather than try to describe it in language; in other contexts, the exact term is needed, where there would be no recourse to a gesture. By allowing the user to interact through both modalities of language and haptics (individually and together), we hope to enhance the navigability and potential for discovery of results returned through complex queries.

## 2 Prior Work

With the advent of tablets and smartphones, haptic interfaces have become a common method of interacting with such devices. As speech recognition technology has improved, voice commands have also increased in prevalence, particularly in contexts where touching the device is prohibited or ill-advised, such as driving. However, despite the co-occurrence of these two modal interfaces on the same device, they rarely overlap in the same use case. Some notable exceptions are in work to integrate conversational with situational context in request fulfillment, where users (for example) can ground to locations on maps using the touchscreen interface and make requests that are localized within those locations [8]. The underlying data, regardless of its domain or format, and its visualization provides the user with background knowledge when making a relevant request, and integrating multiple interactive modalities should also allow computer systems to take advantage of such background knowledge [9, 20].

While there has been considerable work on the analysis of both biological and biomedical literature (e.g., [10, 13]), there has been less research devoted to natural language dialogue-based interactions over biocuration datasets. Of particular note in this area is the work presented in [4, 6, 14], where computational agents with deep knowledge of biological processes are queried through natural language expressions. More recently, [23] extends this to allow for dialogue-based interactions over biological knowledge bases (`http://pathwaymap.indra.bio`).

Visualization of dense data for bioinformatics has been accomplished through clustergrams [19], and dimensionality reductions have been performed through manifold approximation [15] and principal component analysis [22, 3]. Other visualization tools include Cytoscape [21] and ProViz [7]. We bring together the integration of multiple modalities, the background data derived from bioinformatic literature, and the aforementioned visualization techniques in the Jarvis interface.

## 3 Architecture

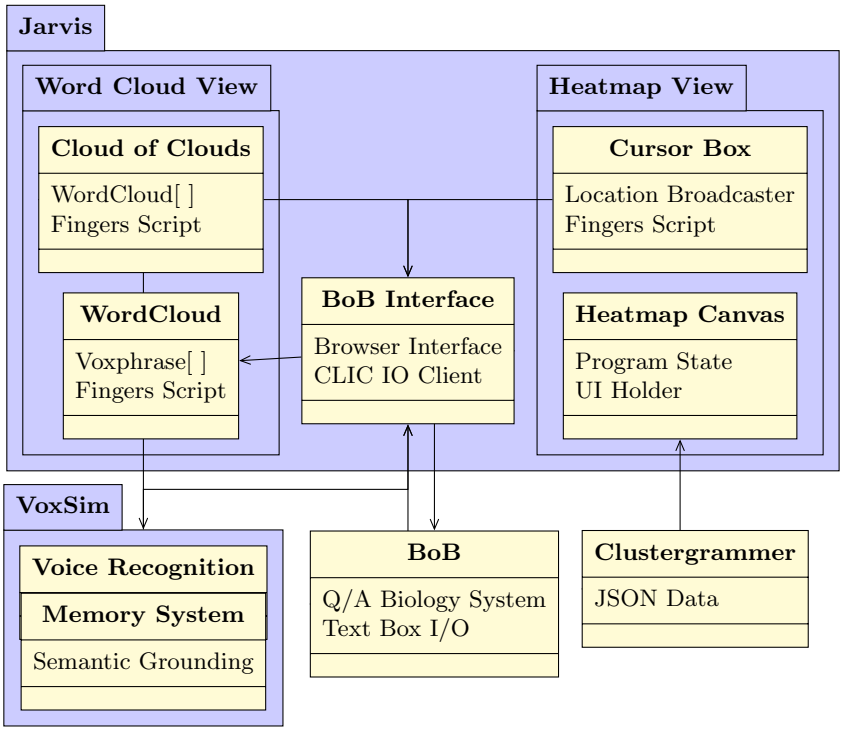


**Fig. 1.** The main components of Jarvis

Jarvis uses the Unity game engine [5] to render graphics and process I/O. The rendered environment is developed on top of the VoxSim platform [12], a Unity-based semantic event simulator that facilitates the manipulation of the visualized objects. This allows for movement of the objects by voice command. It also provides semantic grounding to know what words like "this" may refer to in context of previous inputs and in each modality.

VoxSim is built on the modeling language VoxML [18], which encodes the semantics of *voxemes* or visual instantiations of lexical items. This allows the visualized objects to be manipulated in 3D space based on concrete properties like concavity or symmetry, or abstract properties like location in 3D space or graspability. Jarvis exploits the abstract properties of voxemes to render elements from the underlying dataset as manipulable objects to facilitate data exploration.

Objects implemented as voxeme-derived instances[1] in VoxSim include both words displayed to the user and the groupings of them in clouds. Objects are made interactable by touch on the iPad with scripts from the publicly-available Unity asset *Fingers*, which parallels native iOS gestures. The Unity object has access to both the properties of the voxeme class and the gestures accessible through Fingers.

The data for heatmaps is generated through Clustergrammer [1], which groups hierarchically-clustered heatmaps from gene expression data and saves them to JSON files. These are visualized on the Heatmap Canvas. The heatmap visualization algorithm is also adapted from Clustergrammer.

Through Unity APIs, Jarvis can be configured to consume recognized speech from a variety of services, including Google SR, IBM Watson, or custom speech models. The VoxSim platform also facilitates external parsers through TCP sockets and REST connections.

Natural language processing and question answering is done via a connection to the BoB biocuration system [2]. An example of an exchange with BoB is given in Table 1. Jarvis uses BoB's associated CLIC IO Client (also [2]) to format requests relating to larger datasets such that BoB may parse them, and Jarvis may read the results. It attaches lists of data to phrases passed into BoB and receives the desired results.

Fig. 1 shows the architecture and components of Jarvis.

| | |
|---|---|
| USER: | Create the gene set. [*This also passes a JSON structure containing all of 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 1.** A typical exchange with BoB

## 4 Haptic Control

The user can interact with Jarvis via a combination of voice and haptic control. Table 2 shows the gestures available in the Jarvis interface and what their associated function. Use of gestures depends on the visualized context and the technique currently being used (e.g., heatmap or word cloud). Some gestures may mean different things in different context and not all gestures have a use in all contexts.

[1] "Voxphrases" in Fig. 1 are voxemes representing words as manipulable 3D objects. These are always billboarded to display facing the camera.

In Fig. 2 we see a selector box and the heatmap we wish to zoom on. The heatmap shown represents data on gene expression over tumor samples, that is run through Clustergrammer [1] to group the presence of proteins along rows and source tissue samples along columns.

| **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 2.** Gestures available in Jarvis

The colors in the heatmap are salient to relationships between proteins and gene expression, so analyzing the redder area in the lower left that displays higher associated activity is likely to be useful. These few dozen proteins are best grabbed by selecting a region, which we do by dragging the selection box.

Haptic commands through touch on the iPad are used to drag and resize the selector box, and saying "Zoom in here" is used to zoom in. This simultaneously grounds the selection for transmission to BoB. In the absence of voice commands, such as in the event of speech recognition failure, a corresponding text input box is provided to facilitate the same natural language functionality.

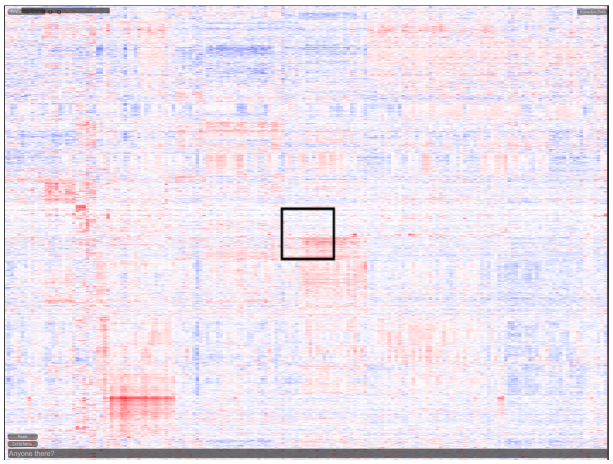

**Fig. 2.** A heatmap of proteins vs. tissue samples with user interface overlaid

## 5 Voice Interaction

As mentioned above, speech-to-text is handled by packages connected to the VoxSim platform [12], such as IBM Watson or Google Speech Recognition. When the user asks a question like "Which of these are transcription factors?" the system reads the currently selected genes and passes the list and question to BoB. The resultant list is passed back to Jarvis for visualization.

The two lists are then visualized as word clouds, with each word in the cloud individually interactable to encourage item-to-item juxtaposition. The full list of genes selected initially corresponds to one word cloud, while the subset of transcription factors can be separated. Each word's position in the clouds is determined by factors in the underlying data, such as frequency of occurrence or similarity to a data point represented by another selected word.


**Fig. 3.** A cloud of proteins for manipulation

In Figure 3 we see a cloud of protein names that BoB returned. These proteins can be organized in the cloud based on a number of attributes, and the cloud itself responds to haptic commands as well. A subset of the cloud, e.g. which proteins are transcription factors, may be pulled out and manipulated independently.

When a word is selected, the cloud can be reordered to correspond to relationships with that word. In future, the 3-dimensional display would be useful for groupings related to up to three different criteria simultaneously.

## 6 Multimodal Integration

Certain modalities are better suited to grounding certain information than others. For example, in human-to-human conversation, deictic gesture—i.e., pointing—grounds naturally to locations, while language may be better at grounding concept labels or attribute descriptions. In a tablet environment, such as that on

which Jarvis is deployed, projective gesture maps neatly to haptics and touch feedback, and the gesture set listed in Table 2.

The nature of bioinformatic data, particularly protein and gene names, poses a problem for purely speech-based systems. In the parlance used in biology, gene names which may be initialisms or collections of characters that are not easily pronounced, such as "MAPK" nonetheless have conventionalized pronunciations in common use (e.g., "map-kay"). These pronunciations are not likely to be covered by the phonetic model of a speech recognition system; a user saying "map-kay" is likely to result is a transcription of "Map K[ay]" (or as might be the case on a smartphone, an inferred request for directions to the house of someone named "Kay").

Therefore, navigating this domain via speech alone is exceedingly difficult. Providing text input may alleviate the speech recognition issues but is time consuming and still does not solve the problem if the user does not know quite how to phrase their request. In addition, the typical ways in which large bioinformatic datasets are typically presented (heatmaps, graphs, word clouds) do not necessarily lend themselves well to being explored purely using language. These kinds of data presentations are inherently visual as well.

Including another modality helps with this. Using haptics to indicate regions or entities of interest allows the user to ground their actions and requests to specific entities using easily recognizable demonstratives ("this," "that," "these," etc.), obviating the need to try to pronounce or spell out the entity references in the data. This makes navigation easy and more tractable for a large dataset.

In principle, it would be possible to navigate and interact with the data using only haptic gesture, by linking defined actions to other haptic gestures (e.g., double or triple tap, two-fingered pan, long press, etc.—see Table 2), but the limit in the number of gestures allowed limits the available vocabulary. Therefore, the functionality that is provided by speech and language input is also crucial. The mixture of haptics and language allows for less discursive language to ask the same question due to entities being focused using haptics and passed along using language.

## 7 Evaluation

Because Jarvis is under active development, evaluation of its capabilities is still in the early stages. Nonetheless, since the goal of Jarvis in this particular use case is well-defined—to enable biologists to accomplish novel discoveries in large datasets—we can at least evaluate the usability of the system in accomplishing this task. In addition, we can evaluate its interactive capabilities for accomplishing data exploration over large datasets of arbitrary provenance (see Sec. 8); strictly biological data is not a requirement for useful multimodal exploration, it is simply an illustrative use case.

### 7.1 Interactive Usability

We propose a method for evaluating the multimodal capabilities of the interaction using simple metrics and object and event semantics, one agnostic to the precise modalities in use in the interaction [11], making it ideal to assess particular areas where the Jarvis system needs improvement.

When defining a multimodal interaction, it is necessary to specify the vocabulary expressible in each modality. The system discussed in [11], also built on VoxSim, uses speech and projective gesture recognition in a Blocks World environment. To evaluate Jarvis, we swap out the projective gesture recognition (pointing, pushing, grasping, etc.) for the available haptic gestures (see Table 2), and instead of specific objects in the interaction, we may instead use delineated regions of the data in the view presented if the presentation allows (e.g., heatmap).

A robust multimodal evaluation scheme should be able to be applied to a human-computer interaction on a system and return a result that is representative of the system's coverage of the total possible interactions within the system's domain (e.g., exploration of bioinformatic data using a variety of visualization techniques).

Previous multimodal evaluation using this schema presented assumed that a user must be truly naive, having very little to no knowledge of exactly what the system understands. However, evaluating in this manner defeats the purpose of a system like Jarvis, whose assumed users must be domain experts. Therefore, evaluating the system's coverage should represent how easy it is for the domain expert, without much prior knowledge of the interface, to use that interface to accomplish their task.

An interaction consists of "moves" taken by each participant, which are logged live. These are timestamped and coded by participant and modality ($S$ for speech, $G$ for gesture, $A$ for action; these are listed in the subscripts in the sample log in Table 3).

Usability metrics of the system can be conditioned on a particular modality. For instance, user studies in aggregate might find that speech input is a point of difficulty, with users struggling to figure out appropriate phrasings, where haptic input provides for an easier semantic grounding of entities (some of these possibilities are shown in the sample log in Table 3). Conversely, they may discover that haptic input is less precise than necessary, requiring the ability to select finer-grained regions than currently allowed.

The sample log in Table 3 is based on the BoB dialogue from Table 1. We model the USER and two "agents," BoB and JARVIS. This is to separate the biocuration backend functionality provided by BoB from the grounding interface provided by Jarvis. BoB normally delivers output through text, and so the speech output attributed to BoB here is actually delivered through Jarvis via text-to-speech. Therefore Jarvis and BoB are perceived as one from the perspective of the user, but BoB is nevertheless given attribution to focus on the semantic content. This allows us to evaluate difficulties due to the BoB output versus

difficulties due to the Jarvis interface, giving Jarvis a way to evaluate both itself and its backend.

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

**Table 3.** Sample interaction log.

In the sample interaction, we see Jarvis making moves that facilitate moves by the user (e.g., moves 0-1, Jarvis presents the data in heatmap form, facilitating the user selecting a region by haptic panning). The data presentation in this fashion followed by a user action over it can be assumed to mean that the user understands how to use this type of data presentation, along with the haptic gesture. This can be further quantified by looking at the length of the delay between when the data is shown to when the user selects an area. 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 this type of data presentation.

We also see places where the dialogue breaks down or fails to proceed smoothly. At move 3, the user says something that BoB does not understand, and so BoB gives the user a reason why. The user responds with a new instruction that grounds the demonstrative "these" to a particular set, that allows BoB to complete the request. Since BoB prompts for a correction that later incorporates information gathered through the Jarvis interface, we can see that it needed more information at move 4, which is alleviated through use of an additional modality provided by Jarvis. This then becomes a way of validating intent detection within a dialogue systems (cf. [24]).

### 7.2 Fidelity of Data Transfer

Successful discovery using a multi-component system like Jarvis, particularly one intended for use with arbitrary backends, depends on the fidelity of data transferred between the subsystems. Therefore, we propose the following methods to evaluate fidelity of particular subcomponents:

*Speech Recognition.* One of the primary difficulties that arises in a multimodal system is poor recognition of one modality hamstringing the interaction in other modalities, often by forcing the interaction into a bad state from which it cannot recover. To assess quality of speech recognition specifically and figure out where it needs improvement, we can have users execute a scripted dialogue that provides a known ground truth, and then measure the accuracy of the recognized input compared to that reference using standard metrics, e.g., BLEU score [17].

*Semantic Grounding.* Correct semantic grounding is essential to ensure that the correct information extracted from the underlying data is passed to BoB along with the query. To assess this through the interaction, we look at "blocks" bounded by moves that negate a prior move and redirect the interaction to new focus objects or actions (e.g., move 5 in Table 3). We assume that within a block, information is being correctly grounded, thus enabling the user to make satisfactory requests, so the longer a block proceeds without redirection or correction, the better the grounding mechanism is performing. Thus, we can also assess grounding via speech vs. grounding via haptics.

*Visualization Accuracy.* Numerous aspects of the visualization can be assessed for accuracy. BoB will return gene subsets represented as word clouds, and the visualization must be sure to accurately represent all elements in the gene subset. Each voxeme object representing a gene in the subset has an underlying textual representation (to which speech and language input is grounded), so this can simply be compared to the string representing the same gene returned by BoB. A 100% match between the two sets means 100% coverage of Jarvis over the BoB data.

Region selection can also be assessed for consistency, by calculating the overlap when a user attempts to select the same region multiple times and calculating how the selections, determined from the selection box's size and location (see Fig. 2) overlap. This can be assessed using haptics with selection via mouse as a baseline to see how much precision is gained or lost from the use of haptics.

Region selection can also be correlated to selection in the underlying data by selecting the same indicated region multiple times and calculating the variance over the data subsets that are extracted by selecting that region.

All these component-specific evaluations can be combined with the general time-based usability evaluation (Sec. 7.1) to determine the errors in components that led to difficulties in the overall interaction, e.g., does a particular misrecognized word regularly lead to a command BoB fails to understand or does an inaccurate presentation of the data make it difficult to make a discovery based on the previous query? Therefore, different levels of functionality each have distinct kinds of evaluation criteria measuring performance and usability. Evaluations using domains experts are planned and ongoing.

## 8 Future Work

Using the results of our planned evaluation, we anticipate being able to fine-tune speech recognition language models to suit the biocuration domain and BoB's capability. However, the particular domain for a multimodal Jarvis interface is determined by the data source and the backend inference and query engine. Hence, while bioinformatic data is our current test case, the Jarvis interface is not limited solely to the display of biological data. For future developments, we would like to implement the following additions to enable robust interaction and exploitation of arbitrary datasets:

- 3D visualizations of the heatmap view to add an additional dimension. This will be useful for representing time sequence data. In the biocuration use case, for example, this could represent single-cell or cell-cycle proliferation data. Such a visualization would be requested by a swipe to the left or right from the heatmap view, effectively rotating the 3D heatmap display, to view the temporal dimension side-on.
- More extended commands for data manipulation. For example, the Fingers package can recognize double and triple taps, which could be used on words for highlighting different kinds of relationships between them, or rearranging locations in clouds to show which genes regulate which proteins, etc.
- More in-depth gene-to-gene relationships in the word cloud visualization. For instance, grouping genes by known gene clusters, or genes that encode similar proteins or share generalized functions. For other types of data, such as protein-protein interaction, these relationship could also be proteins that interact with other proteins or classes or proteins in similar ways (e.g., "TNF-inhibitors," or proteins that collectively inhibit the expression of members of the tumor necrosis factor superfamily).
- Better integration between Fingers and VoxSim. There is a memory stack in VoxSim that tracks which objects the user has interacted with, and any interaction through haptics should be guaranteed to be tracked.
- The ability to "save state" in a visualization and revisit it (for example using the "swipe" gesture to swap views). This would allow users to conduct multiple searches that may lead them on long paths through the data, and then subsequently return "home" and compare their results to make higher-level discoveries.

## 9 Conclusion

In this paper, we have introduced a multimodal tool for visualizing and exploring bioinformatic datasets. This tool, Jarvis, combines speech and haptic control with a robust biocuration dialogue system in iOS on an iPad. These features encourage smooth interactions over complex data in this domain.

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. For large, complex datasets such as those containing biological entities and relations, multiple grounding techniques should not only allow users to have a less cumbersome user experience, but also allow them to switch modalities to achieve greater (or less) precision when desired, and allow one modality's strength to mitigate weaknesses of the other(s) when interacting with the specifics of a dataset. This also allows the same underlying interface to be used with multiple datasets potentially in multiple domains.

**Acknowledgements** 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. We would like to thank everyone at the Boston office of Smart Information Flow Technologies, particularly Laurel Bobrow, Robert Bobrow, Mark Burstein, David McDonald, and Matthew McLure; and Benjamin Gyori and John Bachman at Harvard Medical School. All remaining errors are, of course, those of the authors alone.

## References

[1] Nicolas F. Fernandez et al. "Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data". In: (2017). URL: https://doi.org/10.1038/sdata.2017.151.

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

[3] Neil R. Clark and Avi Ma'ayan. "Introduction to Statistical Methods for Analyzing Large Data Sets: Gene-Set Enrichment Analysis". In: *Science Signaling* 4.190 (2011), 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.

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

[5] Will Goldstone. *Unity Game Development Essentials*. Packt Publishing Ltd, 2009.

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

[7] Peter Jehl et al. "ProViz—a web-based visualization tool to investigate the functional and evolutionary features of protein sequences". In: *Nucleic Acids Research* 44.W1 (Apr. 2016), 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.

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

[9] Michael Johnston. "Multimodal integration for interactive conversational systems". In: *The Handbook of Multimodal-Multisensor Interfaces: Language Processing, Software, Commercialization, and Emerging Directions-Volume 3*. 2019, pp. 21–76.

[10] Jin-Dong Kim. *Biomedical natural language processing*. 2017.

[11] Nikhil Krishnaswamy and James Pustejovsky. "An evaluation framework for multimodal interaction". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[12] Nikhil Krishnaswamy and James Pustejovsky. "VoxSim: A Visual Platform for Modeling Motion Language". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 54–58. URL: https://www.aclweb.org/anthology/C16-2012.

[13] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

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

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

[16] Kay L O'Halloran et al. "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 (2018), pp. 11–30.

[17] Kishore Papineni et al. "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. 2002, pp. 311–318.

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

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

[20] Ethan Selfridge and Michael Johnston. "Interact: Tightly-coupling Multimodal Dialog with an Interactive Virtual Assistant". In: *Proceedings of the*

*2015 ACM on International Conference on Multimodal Interaction*. 2015, pp. 381–382.

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

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

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

[24] Zhao Yan et al. "Building task-oriented dialogue systems for online shopping". In: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.