Overview of Categories

They are colored similarly in the labeling interface.

Layout Categories

Light Colors / for blocks of contents

Textual Blocks

- Title
- Author
- Abstract
- Keywords
- Section
- Paragraph
- List
- Bibliography

Inserted Blocks

- Equation
- Algorithm
- Figure
- Table
- Caption

Top/bottom

- Header
- Footer
- Footnote

Other

- Not-this-paper

Semantic Categories

Dark Colors / Individual Tokens

Author Details

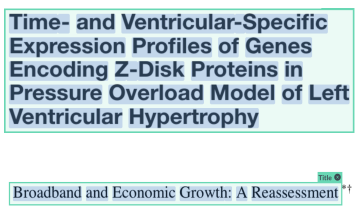
- Author/Name
- Author/Affilation
- Author/Email

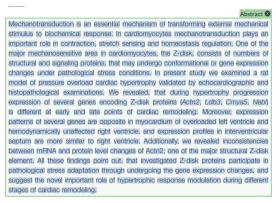
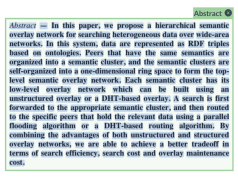
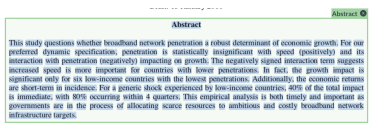
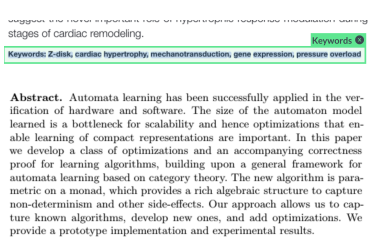
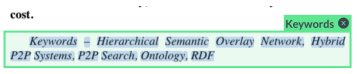
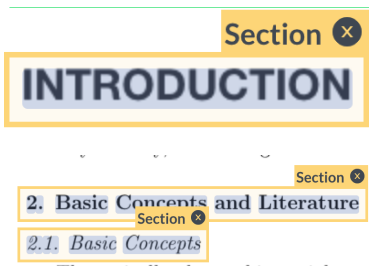
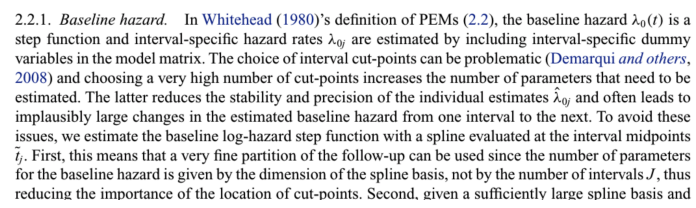
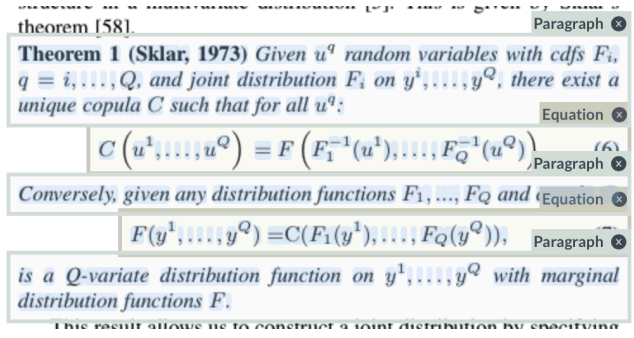
Paper Details

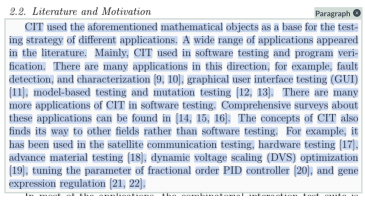
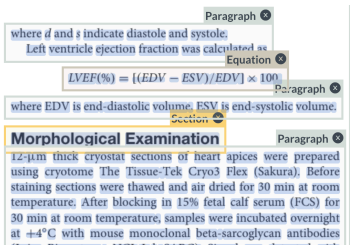
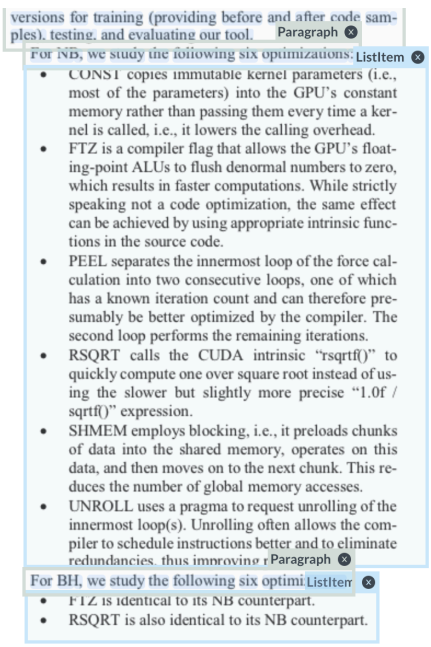
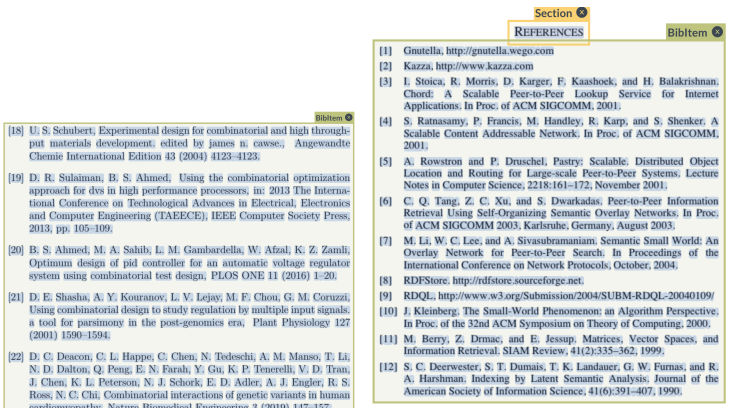
- Paper/Venue

Detailed Explanation

Textual Blocks

Category	Examples	Definition
Title		The title of a paper. - There could be multiple lines for a title. - Please do not include the tailing asterisk if possible (2nd example).
Author		The author *region* of a paper, including author names, affiliations, and emails

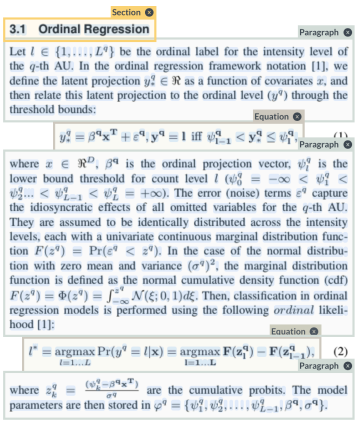
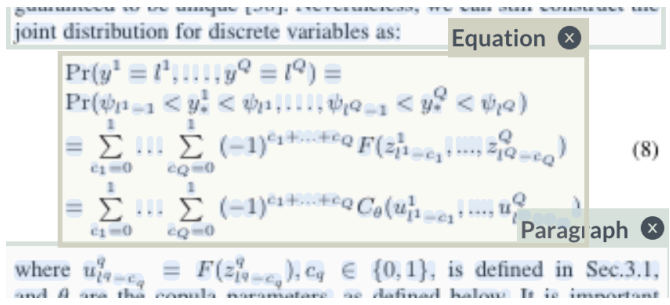
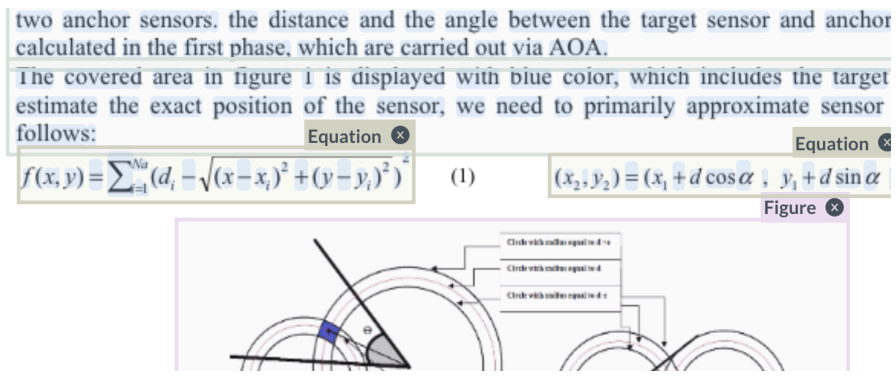
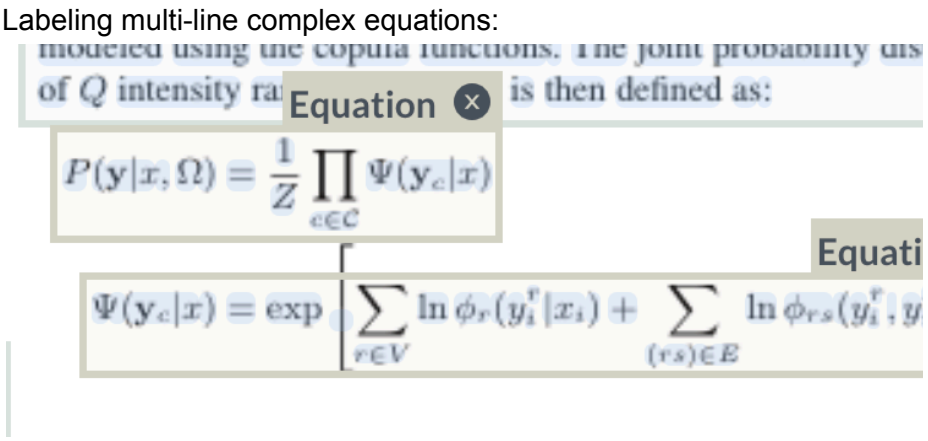
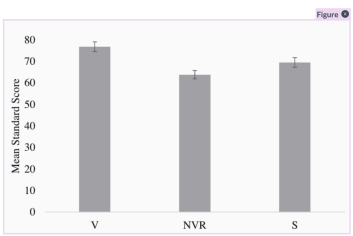
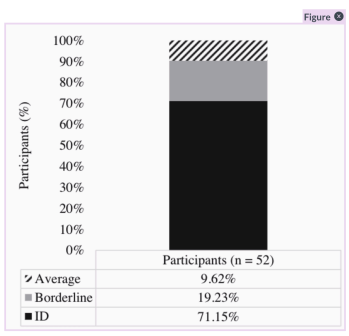
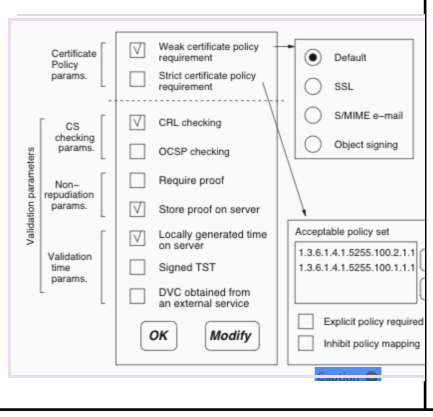
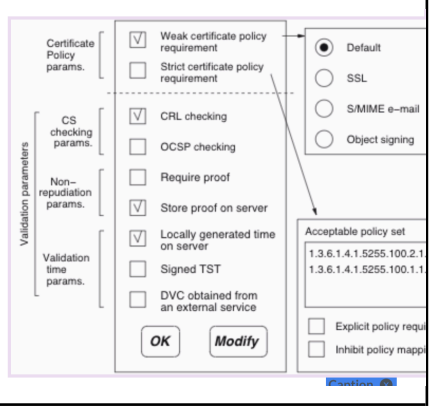
Abstract	▼ Case 1  ▼ Case 2a  ▼ Case 2b 	The abstract text region for a paper. - The abstract could be implicit - just a paragraph right after the title, but feels like a summary of the paper, e.g., case 1. - Please include the abstract title if it exists. 1. It could be “inline”, (case 2a) 2. Or separated from the text (case 2b) - Please separate *keywords* from the abstract region. - There could be additional text next to the abstract, e.g., paper categories, copyrights. We do not need to create labels for them -> See [Objects do not need labeling]().
Keyword	 	The *Keywords* word and paper keywords, usually appear after the abstract. Keywords might not always occur, as shown in the 3rd example.
Section		The section headers, including the section number when it appears. For now, we label different levels of section headers using the same category. And for consecutive section headers, we label them separately. For cases when the section header is incorporated in a paragraph, we label the header together with the paragraph as a paragraph category, and do not create a separate `header` block for the header. ▼  This also applies to cases where there’s a bolded “Theorem” , “Equation”, “Definition” in the beginning of a paragraph. Please do NOT create “section header” annotations for them, but just treat them as regular paragraph text. ▼ 

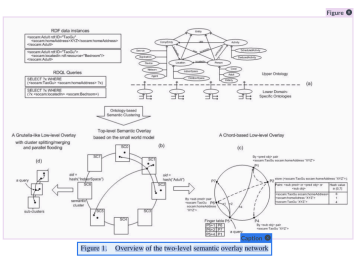
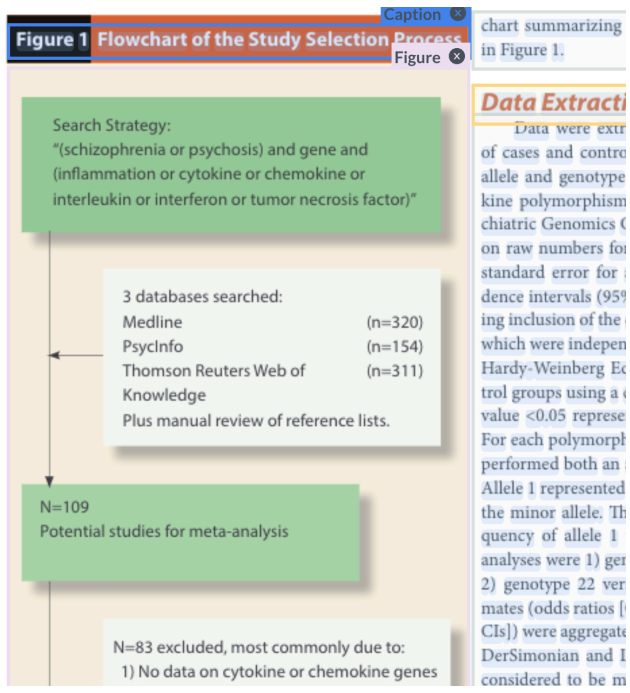
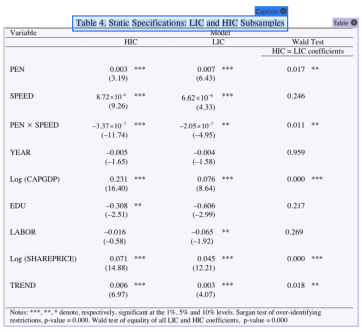
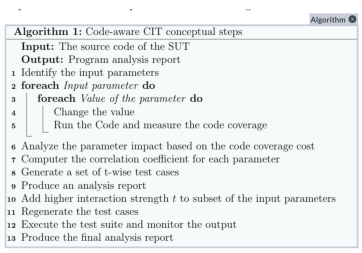
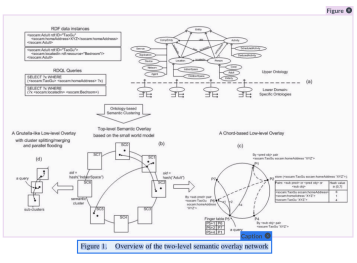
Paragraph	▼ Case 1  ▼ Case 2 	A paragraph of paper's body text. Sometimes a paragraph might be interrupted by other "inserted" contents, and we need to reflect that in the labeling (case 2)
List		The whole list region Sometimes a paragraph might be inserted in between items, and we need to reflect that in the labeling (shown left)
Bibliography		The whole bibliography region on a page * Similarly, we just label the whole region to speed up annotation. For multi column papers, please create one box for each column. Please do not include the reference headers → they should be section headers.
Not-this-paper		Some text from other papers. This usually happens for papers taken from journals, see the left example (highlighted in dashed red rectangle). How to annotate: 1. Firstly please label the text according to regular labeling schemas 2. After finishing labeling the text, draw a huge "not-this-paper" block including all the text that does not belong to the current paper.

Inserted Blocks

Usually the inserted blocks do not have good token representations (the token bounding box sizes may be different from the actual "visual" size of the objects). So the free-form annotation function might be frequently used in this case.

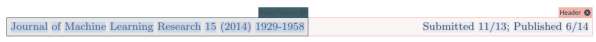
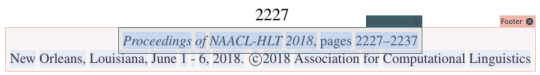
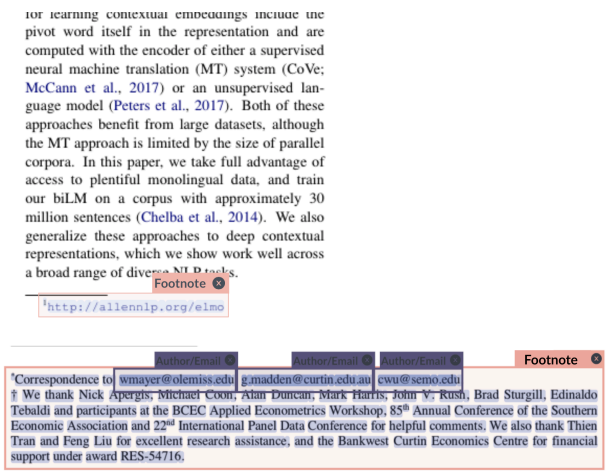
Category	Examples	Definition

Equation	 3.1 Ordinal Regression Let $l \in \{1, \ldots, L^q\}$ be the ordinal label for the intensity level of the q-th AU. In the ordinal regression framework notation [1], we define the latent projection $y_*^q \in \Re$ as a function of covariates x, and then relate this latent projection to the ordinal level (y^q) through the threshold bounds: $y_*^q = \beta^q \mathbf{x}^T + \varepsilon^q, \mathbf{y}^q = \mathbf{l}$ iff $\psi_{l-1}^q < \mathbf{y}_*^q \leq \psi_l^q$, (1) where $x \in \Re^D$, β^q is the ordinal projection vector, ψ_l^q is the lower bound threshold for count level l ($\psi_0^q = -\infty < \psi_1^q < \psi_2^q \ldots < \psi_{L-1}^q < \psi_L^q = +\infty$). The error (noise) terms ε^q capture the idiosyncratic effects of all omitted variables for the q-th AU. They are assumed to be identically distributed across the intensity levels, each with a univariate continuous marginal distribution function $F(z^q) = \Pr(\varepsilon^q < z^q)$. In the case of the normal distribution with zero mean and variance $(\sigma^q)^2$, the marginal distribution function is defined as the normal cumulative density function (cdf) $F(z^q) = \Phi(z^q) = \int_{-\infty}^{z^q} \mathcal{N}(\xi; 0, 1) d\xi$. Then, classification in ordinal regression models is performed using the following *ordinal* likelihood [1]: $l^* = \underset{l=1,\ldots,L}{\operatorname{argmax}} \Pr(y^q = l \mid \mathbf{x}) = \underset{l=1,\ldots,L}{\operatorname{argmax}} \mathbf{F}(\mathbf{z}_l^q) - \mathbf{F}(\mathbf{z}_{l-1}^q)$, (2) where $z_k^q = \frac{(\psi_k^q - \beta^q \mathbf{x}^T)}{\sigma^q}$ are the cumulative probits. The model parameters are then stored in $\varphi^q = \{\psi_1^q, \psi_2^q, \ldots, \psi_{L-1}^q, \beta^q, \sigma^q\}$.	The display equation inside a paper. And the equations within the text (inline equations) are considered as texts. Please need NOT to label the equation numbers. (You may see screenshots with the equation number labeled. That's from a previous version.) For math heavy papers, the equation and text structure could be complicated.  $\Pr(y^1 = l^1, \ldots, y^Q = l^Q) =$ $\Pr(\psi_{l^1-1} < y_*^1 < \psi_{l^1}, \ldots, \psi_{l^Q-1} < y_*^Q < \psi_{l^Q})$ $= \sum_{c_1=0}^{1} \cdots \sum_{c_Q=0}^{1} (-1)^{c_1+\ldots+c_Q} F(z_{l^1-c_1}^1, \ldots, z_{l^Q-c_Q}^Q)$ $= \sum_{c_1=0}^{1} \cdots \sum_{c_Q=0}^{1} (-1)^{c_1+\ldots+c_Q} C_\theta(u_{l^1-c_1}^1, \ldots, u_{l^Q-c_Q}^Q)$ joint distribution for discrete variables as: Equation $\Pr(y^1 = l^1, \ldots, y^Q = l^Q) =$ $\Pr(\psi_{l^1-1} < y_*^1 < \psi_{l^1}, \ldots, \psi_{l^Q-1} < y_*^Q < \psi_{l^Q})$ $= \sum_{c_1=0}^{1} \cdots \sum_{c_Q=0}^{1} (-1)^{c_1+\ldots+c_Q} F(z_{l^1-c_1}^1, \ldots, z_{l^Q-c_Q}^Q)$ (8) $= \sum_{c_1=0}^{1} \cdots \sum_{c_Q=0}^{1} (-1)^{c_1+\ldots+c_Q} C_\theta(u_{l^1-c_1}^1, \ldots, u_{l^Q-c_Q}^Q)$ Paragraph where $u_{l^q-c_q}^q = F(z_{l^q-c_q}^q)$, $c_q \in \{0, 1\}$, is defined in Sec.3.1,  two anchor sensors. the distance and the angle between the target sensor and anchor calculated in the first phase, which are carried out via AOA. The covered area in figure 1 is displayed with blue color, which includes the target estimate the exact position of the sensor, we need to primarily approximate sensor follows: Equation $f(x,y) = \sum_{i=1}^{Na} (d_i - \sqrt{(x-x_i)^2 + (y-y_i)^2})^2$ (1) Equation $(x_2, y_2) = (x_1 + d\cos\alpha\ ,\ y_1 + d\sin\alpha$ Figure Labeling multi-line complex equations:  modeled using the copula functions. The joint probability dis of Q intensity ra Equation is then defined as: $P(\mathbf{y}\|x, \Omega) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \Psi(\mathbf{y}_c\|x)$ Equati $\Psi(\mathbf{y}_c\|x) = \exp\left[\sum_{r \in V} \ln \phi_r(y_i^r\|x_i) + \sum_{(rs) \in E} \ln \phi_{rs}(y_i^r, y\right.$
Figure	▼ Case 1  ▼ Case 2  	A figure/chart/image in the paper, excluding the captions. For figures without a clear bounding box, we draw a **minimum bounding box** that includes the full figure region (case 1). Not so good  Better (less white space between the annotated bounding box and the figure boundary) 

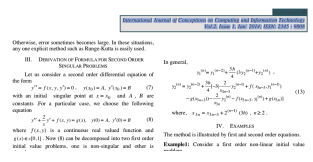
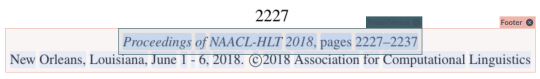
	 Figure 1: Overview of the two-level semantic overlay network	Sometimes a figure might have some table-like structures (case 2). We do not need to label all images, see [Objects do not need labeling]().
		Sometimes a figure could be combined with the caption - please label them separately. → Similarly for tables 
Table	 Table 4. Static Specifications: LIC and HIC Subsamples Notes: ***, **, * denote, respectively, significant at the 1%, 5% and 10% levels. Sargan test of over-identifying restrictions, p-value = 0.000. Wald test of equality of all LIC and HIC coefficients, p-value = 0.000	A table region inside a paper. For now, we include the table notes inside the labeled table region. (And we label the caption separately.) Sometimes we will have large table/figures that's being rotated:
Algorithm	 **Algorithm 1:** Code-aware CIT conceptual steps **Input:** The source code of the SUT **Output:** Program analysis report Identify the input parameters **foreach** *Input parameter* **do** **foreach** *Value of the parameter* **do** Change the value Run the Code and measure the code coverage Analyze the parameter impact based on the code coverage cost Computer the correlation coefficient for each parameter Generate a set of t-wise test cases Produce an analysis report Add higher interaction strength t to subset of the input parameters Regenerate the test cases Execute the test suite and monitor the output Produce the final analysis report	Algorithm blocks usually have unique visual properties (different from code blocks) - it starts (or ends) with a header indicating it's describing an algorithm. Then it has a list like structure listing steps and equations for an algorithm. Sometimes it can also appear as "pseudo code" blocks.
Caption	 Figure 1: Overview of the two-level semantic overlay network	Usually started with Figure x or Table x.

Top/Bottom

Category	Examples	Definition

Header		Refers to the text at the top of a page, sometimes it may appear above a line separator. Sometimes headers might contain important information like paper venues.
Footer		Refers to the text at the bottom of a page, sometimes it may appear below a line separator. Sometimes footers might contain important information like paper venues.
Footnote		The biggest difference between a footer and a footnote is that, footnote points to somewhere in the original text. Thus, footnote usually begins with a number or some characters (e.g., *) for linking. Note: 1. When labeling, please also include the the index 2. It might also contain important information like author emails, etc.

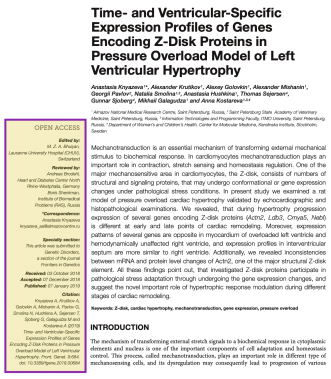
Semantic Regions

Category	Screenshots	Definition
Author/Name		Author names usually appear in the beginning of a paper . Please also include the superscripts or reference number (usually at the end for each author) when possible. And for some IEEE papers, they have author details at the end of the paper - please just neglect them and don't create annotations for them.
Author/Affiliation		Author affiliations
Author/Email		Author emails. Sometimes might appear within a footer/header.
Paper/Venue		The venues for a paper. Note: 1. Usually the journal name / conference are something we care about. But please also include the volume number / DOI when it's possible. 2. But sometimes a paper venue information might also appear directly next to some texts indicating copyright, time, or location. We do not need the them, see the example below:  (The 2nd line contains only location, time, and copyright.)

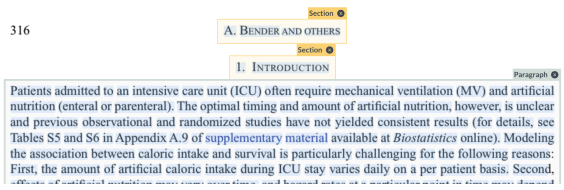
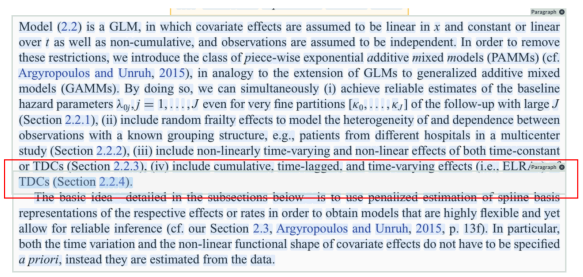
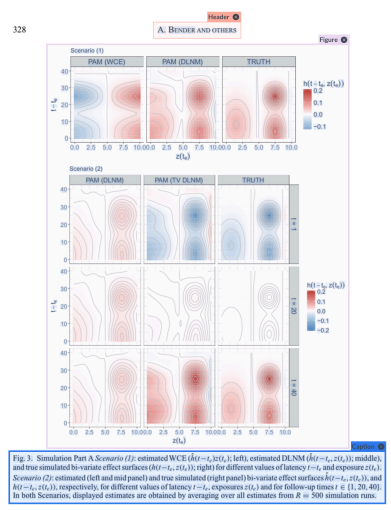
Texts that don't need annotation

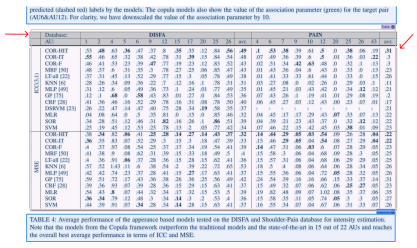
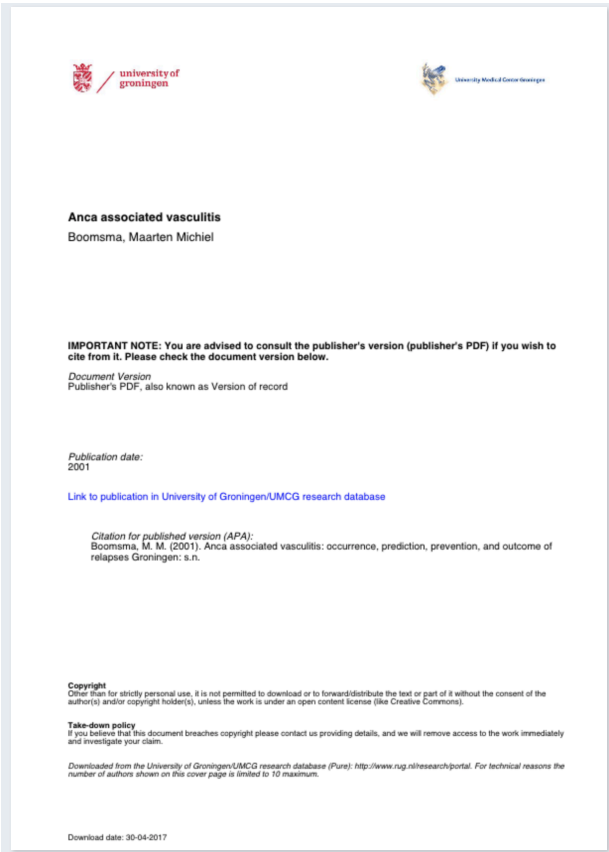
For objects not listed in the previous sections, we usually don't need to assign tags for them. Specifically, we don't need to label:

Category	Examples	Definition

Irrelevant Text	 **Time- and Ventricular-Specific Expression Profiles of Genes Encoding Z-Disk Proteins in Pressure Overload Model of Left Ventricular Hypertrophy** *Anastasia Knyazeva[1*], Alexander Krutikov[2], Alexey Golovkin[1], Alexander Mishanin[1], Georgii Pavlov[1], Natalia Smolina[1], Anastasia Hushkina[1], Thomas Sejersen[3], Gunnar Sjoberg[3], Mikhail Galagudza[1] and Anna Kostareva[1,4]* [1] Almazov National Medical Research Centre, Saint Petersburg, Russia, [2] Saint Petersburg State Academy of Veterinary Medicine, Saint Petersburg, Russia, [3] Information Technologies and Programming Faculty, ITMO University, Saint Petersburg, Russia, [4] Department of Women's and Children's Health, Center for Molecular Medicine, Karolinska Institute, Stockholm, Sweden OPEN ACCESS *Edited by:* M. J. A. Wisser, Lausanne University Hospital (CHUV), Switzerland *Reviewed by:* Andreas Brodehl, Heart and Diabetes Center North Rhine-Westphalia, Germany; Boris Sheremet, Institute of Biomedical Problems (IMG), Russia **Correspondence:* Anastasia Knyazeva knyazeva_ae@almazovcentre.ru *Specialty section:* This article was submitted to Genetic Disorders, a section of the journal Frontiers in Genetics *Received:* 09 October 2018 *Accepted:* 07 December 2018 *Published:* 07 January 2019 *Citation:* Knyazeva A, Krutikov A, Golovkin A, Mishanin A, Pavlov G, Smolina N, Hushkina A, Sejersen T, Sjoberg G, Galagudza M and Kostareva A (2019) Time- and Ventricular-Specific Expression Profiles of Genes Encoding Z-Disk Proteins in Pressure Overload Model of Left Ventricular Hypertrophy. Front. Genet. 9:664. doi: 10.3389/fgene.2018.00664 Mechanotransduction is an essential mechanism of transforming external mechanical stimulus to biochemical response. In cardiomyocytes mechanotransduction plays an important role in contraction, stretch sensing and homeostasis regulation. One of the major mechanosensitive area in cardiomyocytes, the Z-disk, consists of numbers of structural and signaling proteins, that may undergo conformational or gene expression changes under pathological stress conditions. In present study we examined a rat model of pressure overload cardiac hypertrophy validated by echocardiographic and histopathological examinations. We revealed, that during hypertrophy progression expression of several genes encoding Z-disk proteins (*Actn2*, *Ldb3*, *Cmya5*, *Neb6*) is different at early and late points of cardiac remodeling. Moreover, expression patterns of several genes are opposite in myocardium of overloaded left ventricle and hemodynamically unaffected right ventricle, and expression profiles in interventricular septum are more similar to right ventricle. Additionally, we revealed inconsistencies between mRNA and protein level changes of Actn2, one of the major structural Z-disk element. All these findings point out, that investigated Z-disk proteins participate in pathological stress adaptation through undergoing the gene expression changes, and suggest the novel important role of hypertrophic response modulation during different stages of cardiac remodeling. **Keywords: Z-disk, cardiac hypertrophy, mechanotransduction, gene expression, pressure overload** **INTRODUCTION** The mechanism of transforming external stretch signals to a biochemical response is cytoplasmic elements and nucleus is one of the important components of cell adaptation and homeostasis control. This process, called mechanotransduction, plays an important role in different type of mechanosensing cells, and its dysregulation may consequently lead to progression of various	Some random text might appear in the paper, including, but not limited to: 1. Page Numbers 2. Paper copyrights 3. Some additional review or funding information just next to the paper abstract or titles 4. Text in other languages 5. Sometimes a dedicated front page might appear, indicating some extra information about the paper. 6. Some random text that does not fit in the existing categories, e.g., the example below and time-dependent covariate effects in survival analysis. *Statistics in Medicine* **33**, 5318–5337. XIAO, Y., ABRAHAMOWICZ, M., MOODIE, E. E. M., WEBER, R. AND YOUNG, J. (2014). Flexible marginal structural models for estimating the cumulative effect of a time-dependent treatment on the hazard: Reassessing the cardiovascular risks of didanosine treatment in the Swiss HIV cohort study. *Journal of the American Statistical Association* **109**, 455–464. [*Received February 6, 2017; revised January 15, 2018; accepted for publication January 16, 2018*] Title **Prevalence of reduced bone mineral density in patients with anti-neutrophil cytoplasmic antibodies associated vasculitis and the role of immunosuppressive therapy; a cross- sectional study** Author Maarten M. Boomsma [1], Coen A. Stegeman [2], Andrea B. Kramer [3], Marjan Karsijns [3], D. Albertus Piers [4], Jan Willem Cohen Tervaert [5] Author/Affiliation Department of Internal Medicine, Division of [1] Clinical Immunology, and [2] Nephrology, University Hospital Groningen; [3] Medical Faculty, University of Groningen; Department of [4] Nuclear Medicine, University Hospital Groningen; Department of Internal Medicine, [5] Division of Clinical Immunology, University Hospital Maastricht, The Netherlands. *Submitted for publication*

Common Issues

ID	Example	Issue and Fix
1	 316 Section A. BENDER AND OTHERS Section 1. INTRODUCTION Paragraph Patients admitted to an intensive care unit (ICU) often require mechanical ventilation (MV) and artificial nutrition (enteral or parenteral). The optimal timing and amount of artificial nutrition, however, is unclear and previous observational and randomized studies have not yielded consistent results (for details, see Tables S5 and S6 in Appendix A.9 of supplementary material available at *Biostatistics* online). Modeling the association between caloric intake and survival is particularly challenging for the following reasons: First, the amount of artificial caloric intake during ICU stay varies daily on a per patient basis. Second,	**Issue: Confusing page header looks like a section header** Sometimes the header could be really confusing - it might look like a section header. For this case, the header (A. BENDER AND OTHERS) is confused as a section header. **Fix:** Follow the top to bottom order when labeling a page. Pay extra attention near the top and bottom of the page.
2	 Paragraph Model (2.2) is a GLM, in which covariate effects are assumed to be linear in x and constant or linear over t as well as non-cumulative, and observations are assumed to be independent. In order to remove these restrictions, we introduce the class of *p*iece-wise exponential *a*dditive *m*ixed *m*odels (PAMMs) (cf. Argyropoulos and Unruh, 2015), in analogy to the extension of GLMs to generalized additive mixed models (GAMMs). By doing so, we can simultaneously (i) achieve reliable estimates of the baseline hazard parameters $\lambda_{0j}, j = 1, \ldots, J$ even for very fine partitions $[\kappa_0, \ldots, \kappa_J]$ of the follow-up with large J (Section 2.2.1), (ii) include random frailty effects to model the heterogeneity of and dependence between observations with a known grouping structure, e.g., patients from different hospitals in a multicenter study (Section 2.2.2), (iii) include non-linearly time-varying and non-linear effects of both time-constant or TDCs (Section 2.2.3), (iv) include cumulative, time-lagged, and time-varying effects (i.e., ELR/Paragraph TDCs (Section 2.2.4). ~~The basic idea – detailed in the subsections below – is to use penalized estimation of spline basis~~ representations of the respective effects or rates in order to obtain models that are highly flexible and yet allow for reliable inference (cf. our Section 2.3, Argyropoulos and Unruh, 2015, p. 13f). In particular, both the time variation and the non-linear functional shape of covariate effects do not have to be specified *a priori*, instead they are estimated from the data.	**Issue: Overlapping between continuous paragraph blocks** It's really easy to create incorrect overlapping boxes for paragraph regions, as shown in this case (the two paragraph boxes overlap in the highlighted red region). **Fix:** When labeling continuous blocks, please be cautious when creating boxes near the boundaries.
3	 328 Header A. BENDER AND OTHERS Figure Scenario (1) PAM (WCE) PAM (DLNM) TRUTH Scenario (2) PAM (DLNM) PAM (TV DLNM) TRUTH Caption Fig. 3. Simulation Part A *Scenario (1):* estimated WCE ($\hat{h}(t-s, z(t_s); t$ left), estimated DLNM ($\hat{h}(t-t_s, z(t_s))$; middle), and true simulated bi-variate effect surfaces ($h(t-t_s, z(t_s))$; right) for different values of latency $t-t_s$ and exposure $z(t_s)$. *Scenario (2):* estimated (left and mid panel) and true simulated (right panel) bi-variate effect surfaces $\hat{h}(t-t_s, z(t_s))$, and $h(t-t_s, z(t_s))$, respectively, for different values of latency $t-t_s$, exposures $z(t_s)$ and for follow-up times $t \in \{1, 20, 40\}$. In both Scenarios, displayed estimates are obtained by averaging over all estimates from $R = 500$ simulation runs.	**Issue: how to label composite figures (a huge figure with many subplots)?** **Fix:** For a composite within the same page, label a huge figure block

4	 predicted (dashed red) labels by the models. The copula models also show the value of the association parameter (green) for the target pair (AU6&AU12). For clarity, we have downscaled the value of the association parameter by 10. Table 4: Average performance of the apperance based models tested on the DISFA and Shoulder-Pain database for intensity estimation. Note that the models from the Copula framework outperform the traditional models and the state-of-the-art in 15 out of 22 AUs and reaches the overall best average performance in terms of ICC and MSE.	**Issue: Inaccurate table boundary** As shown in the example, the labeled boundary does not perfectly delineate the table boundary. **Fix:** Please use the free-form annotation function to label tables and figures
5	 **Anca associated vasculitis** Boomsma, Maarten Michiel **IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.** *Document Version* Publisher's PDF, also known as Version of record *Publication date:* 2001 Link to publication in University of Groningen/UMCG research database *Citation for published version (APA):* Boomsma, M. M. (2001). Anca associated vasculitis: occurrence, prediction, prevention, and outcome of relapses Groningen: s.n. **Copyright** Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons). **Take-down policy** If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim. *Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.* Download date: 30-04-2017	**Issue: A cover page with no-specific information listed above** **Fix:** Label nothing