2022/05/31 HUG2022

BD5: an open data format for representing quantitative biological dynamics data

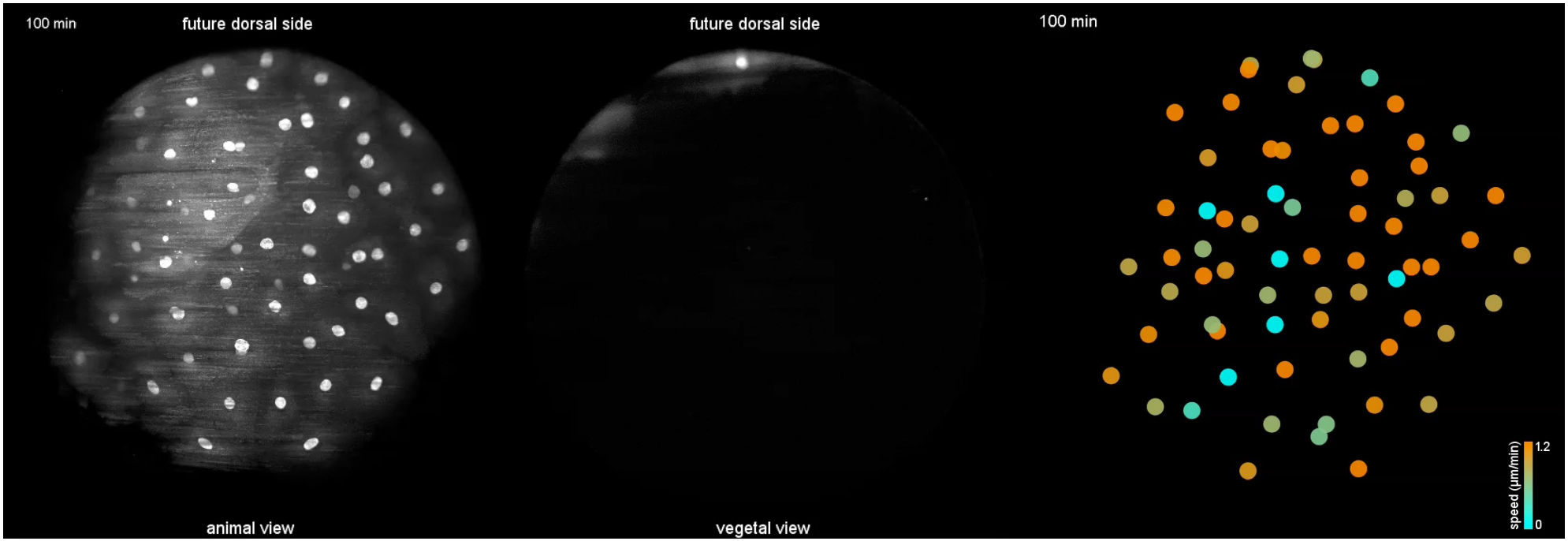
Koji Kyoda[1], Kenneth H.L. Ho[2], Hiroya Itoga[1], Yukako Tohsato[3], Shuichi Onami[1]
[1]RIKEN BDR, [2]Francis Crick Institute, [3]Ritsumeikan University

Bioimage informatics

- Live-cell imaging can capture spatiotemporal dynamics of biological phenomena.
- Using image analysis, (x, y, z, t, c) data can be obtained from microscopy images.

image data

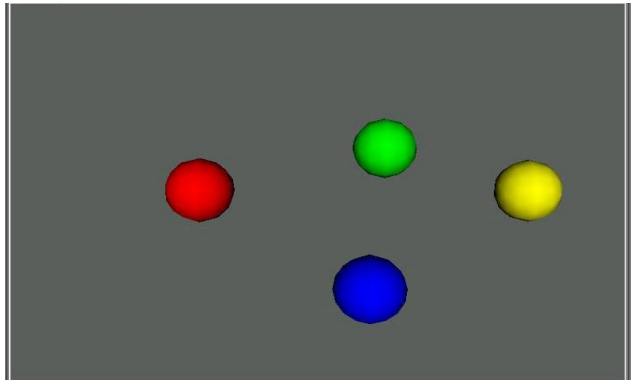
quantitative data



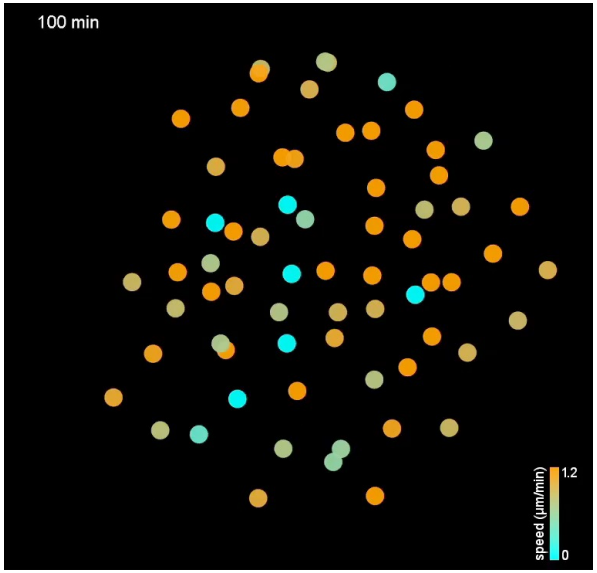
(Keller et al. 2008)

Quantitative biological dynamics data

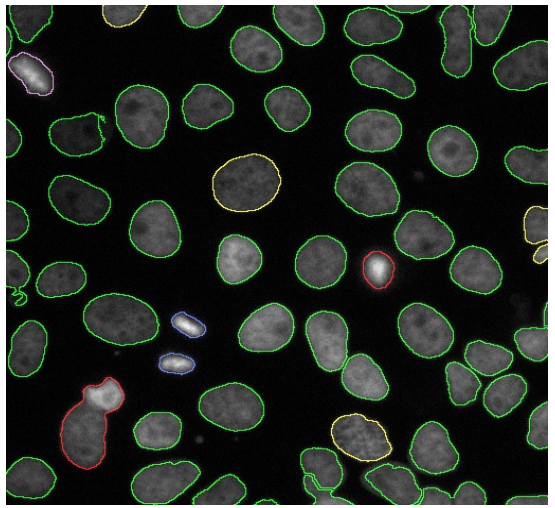
C. elegans (Bao et al., 2006)



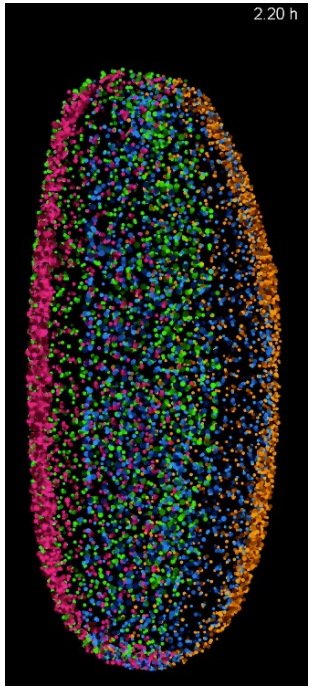
zebrafish (Keller et al., 2008)



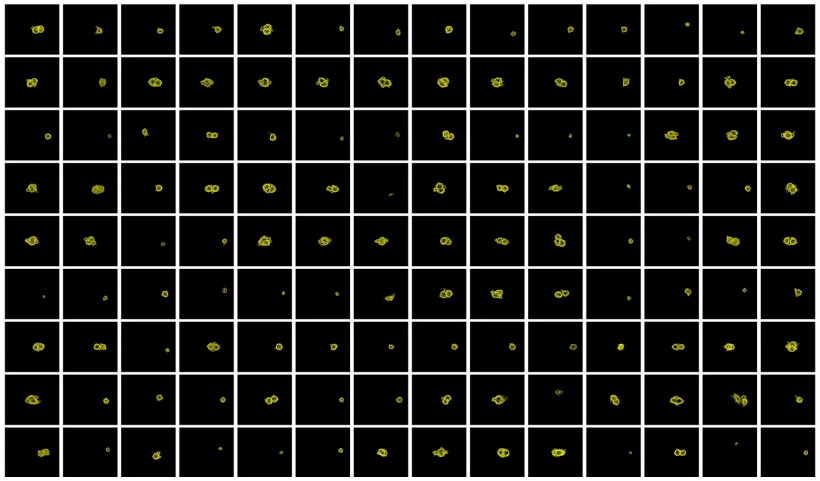
HeLa cell (Held et al., 2010)



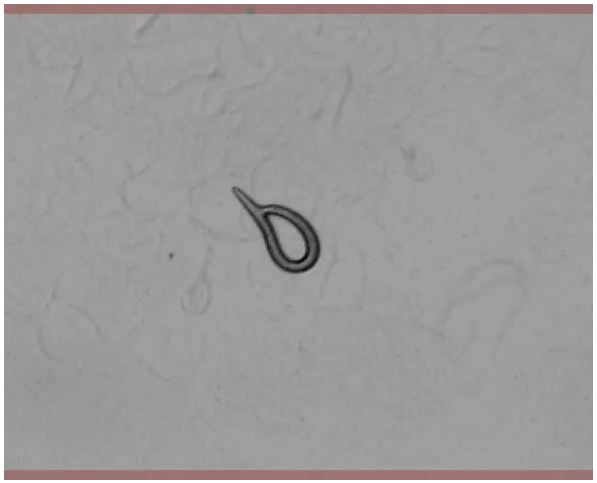
D. melanogaster (Keller et al., 2010)



C. elegans (Kyoda et al., 2013)



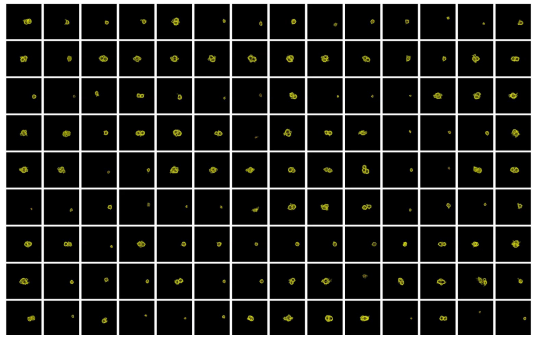
C. elegans (Yemini et al., 2013)



Problem

- Research groups used different data formats.
- It is often difficult to reuse their data because of:
  - intricate data structure
  - the lack of detailed explanations

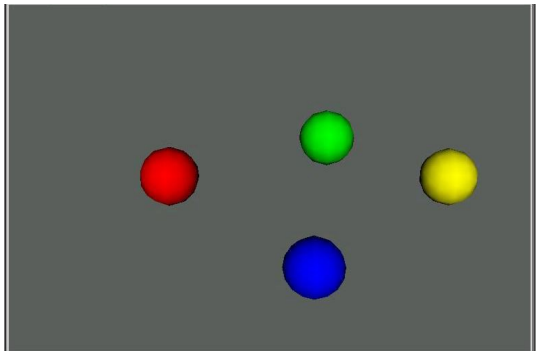
(Kyoda et al. 2013)



0000, 1000, 51.8232, 37.7117, 27.1820, 1, P0, 17
477, 342, 23, 211111111210111212112…….
486, 321, 24, 212112222100121222102…….
……
1000, 2000, 33.3146, 32.1893, 25.2702, 2, P0, 17
320, 257, 23, 21111211111211111112…….

Text file

(Bao et al. 2006)

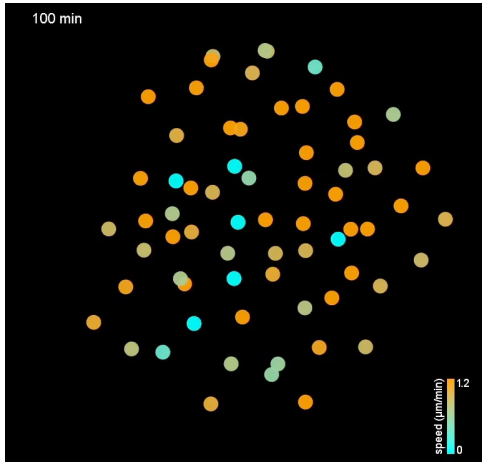


t001-nuclei
t002-nuclei
…..

1, 1, -1, 4, -1, 380, 366, 16.1, 80, EMS, 21281015,
2, 1, -1, 2, -1, 387, 153, 16.6, 86, ABp, 2836266,
3, 1, -1, 3, -1, 189, 251, 17.2, 88, ABa, 2850348,
4, 1, -1, 5, -1, 562, 269, 18.1, 80, P2, 2168825

Separated text files

(Keller et al. 2008)



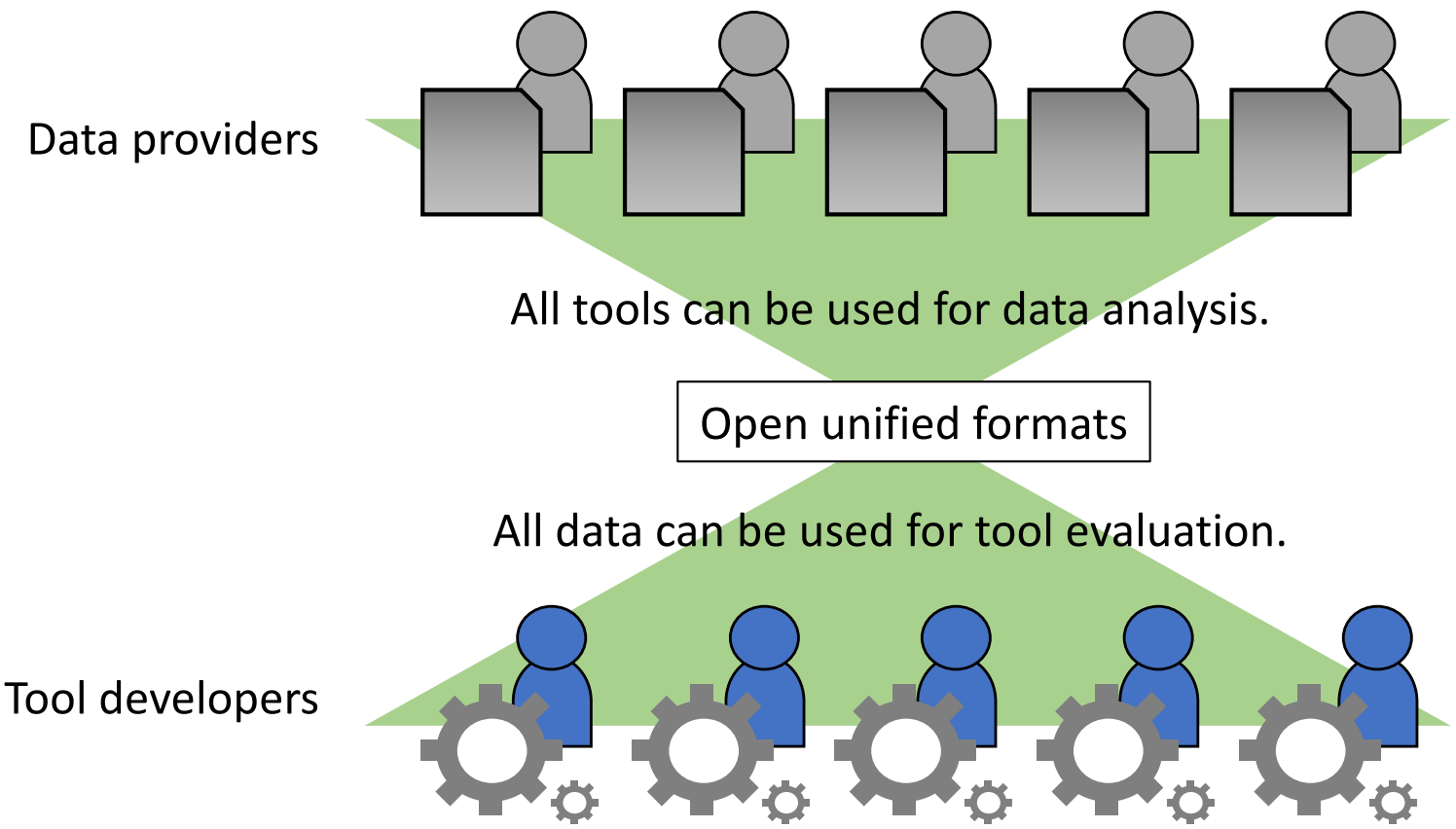
<68x17 double>
<69x17 double>
…..

611.2904, 563.0863, 80.3444, 13.8991, ….
485.5862, 546.2255, 91.1758, 14.0556, ….
562.8646, 459.1969, 95.8143, 14.2670, ….
……

Matlab file

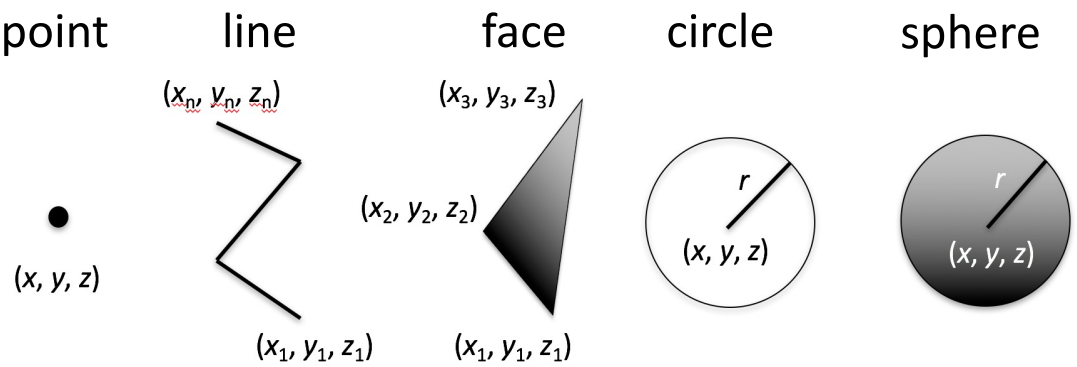
Open unified data formats

- Allowing
  - Data analysis and comparison
  - Tool development and its evaluation



BDML: Biological Dynamics Markup Language

- An open unified format for representing quantitative data of biological dynamics

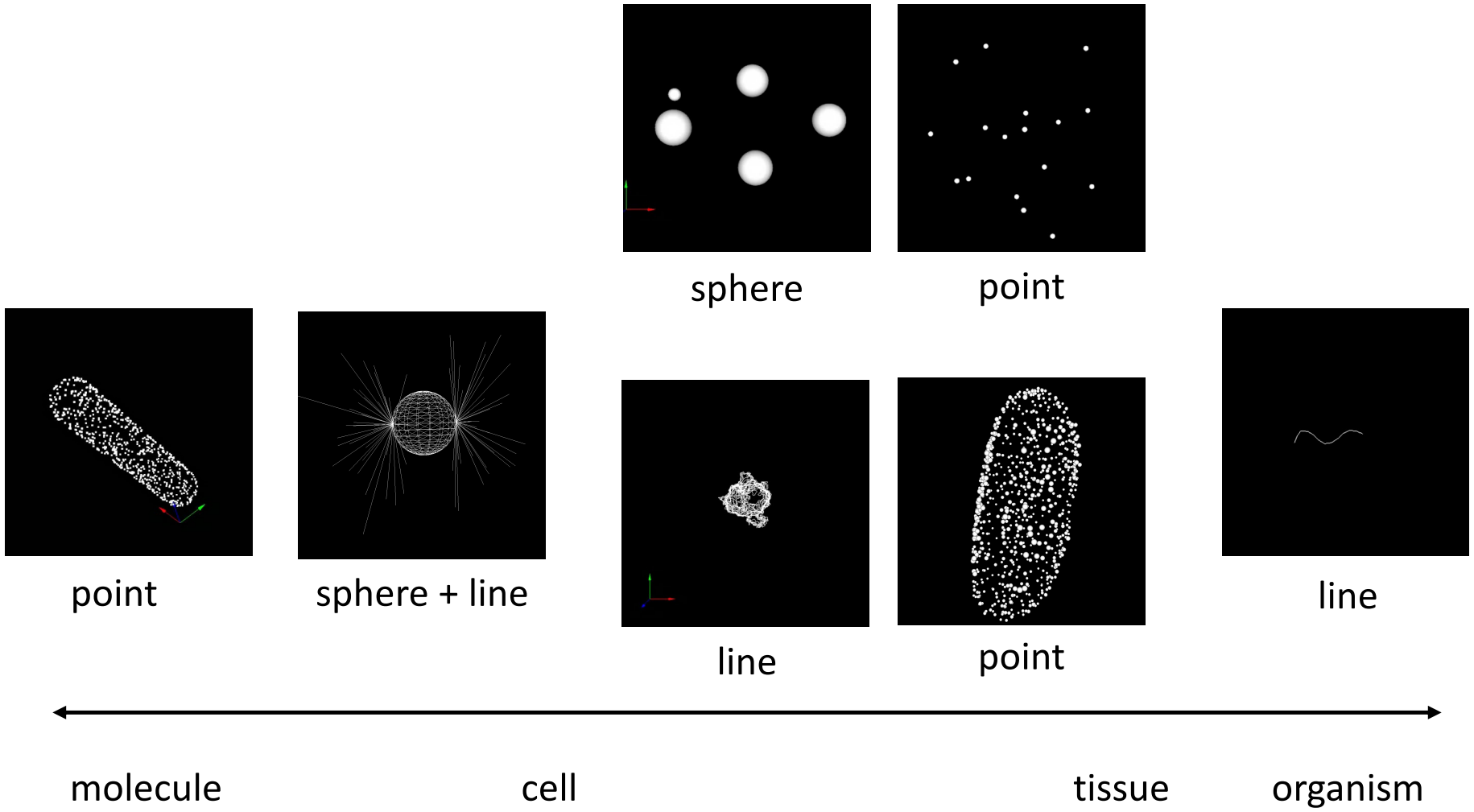



```
<scaleUnit>
  <tScale>20</tScale>
  <tUnit>second</tUnit>
</scaleUnit>
<component>
  <componentID>100</componentID>
  <time>1</time>
  <measurement>
     <point><xyz><x>10.32</x><y>30.42</y><z>18.32</z></xyz></point>
  </measurement>
</component>
<component>
  <componentID>101</componentID>
  <time>2</time>
  <prevID>100</prevID>
  <measurement>
     <point><xyz><x>9.57</x><y>32.05</y><z>14.91</z></xyz></point>
  </measurement>
</component>
```

Kyoda et al. (2015) Bioinformatics

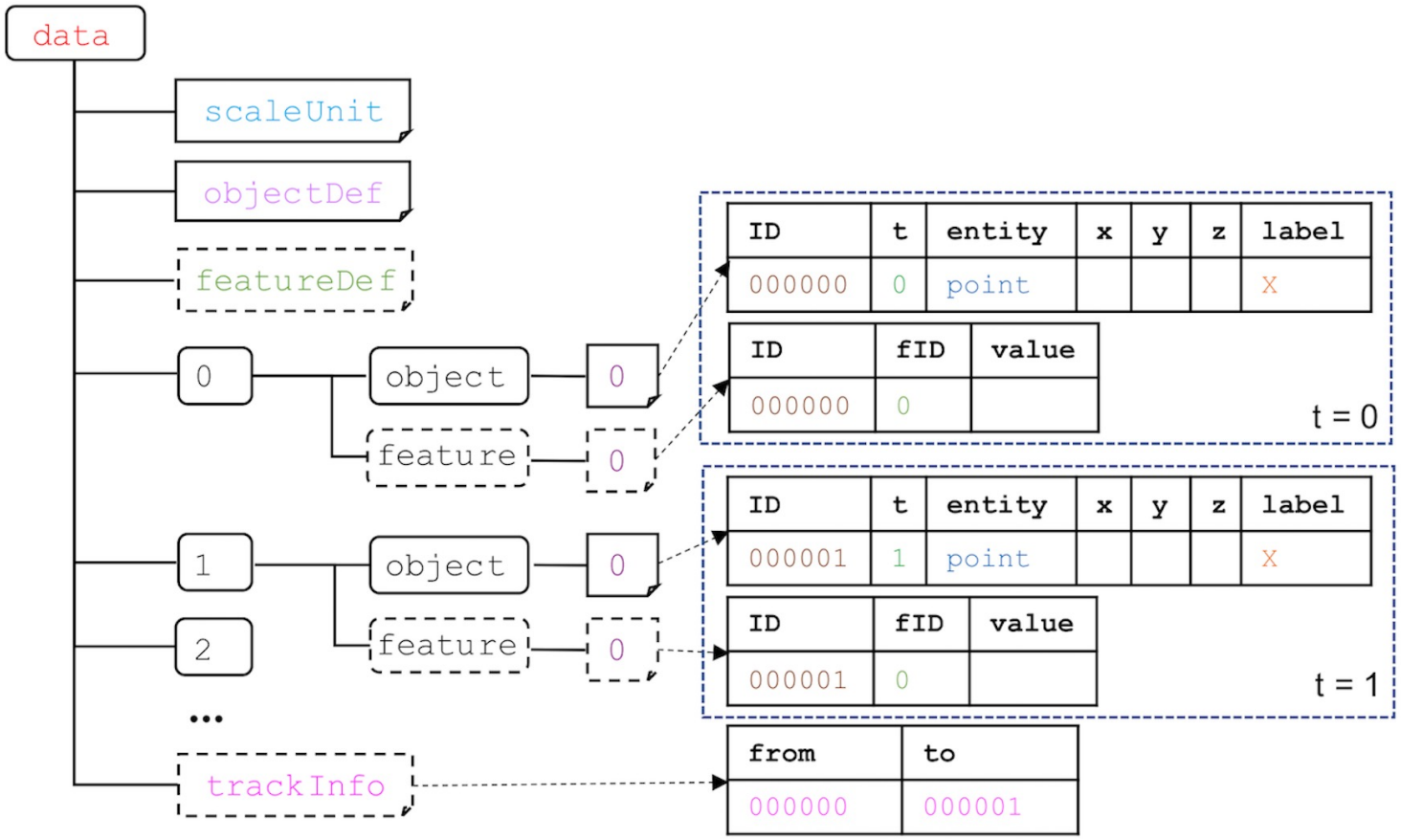
Biological dynamics described in BDML

- Data ranging from molecules to organisms



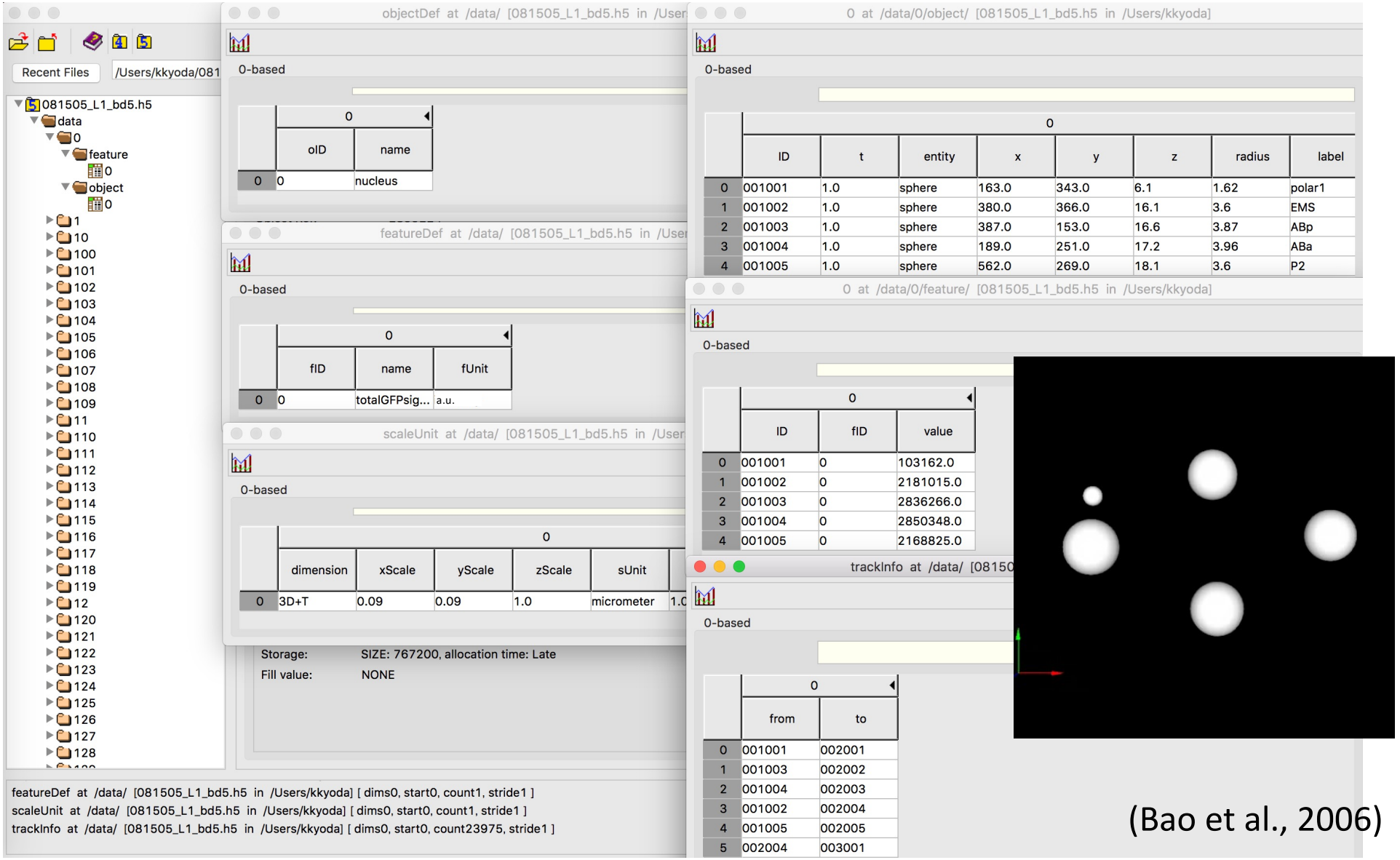
BD5

- HDF5-based data format for representing quantitative biological dynamics data



Kyoda et al. (2020) PLoS One

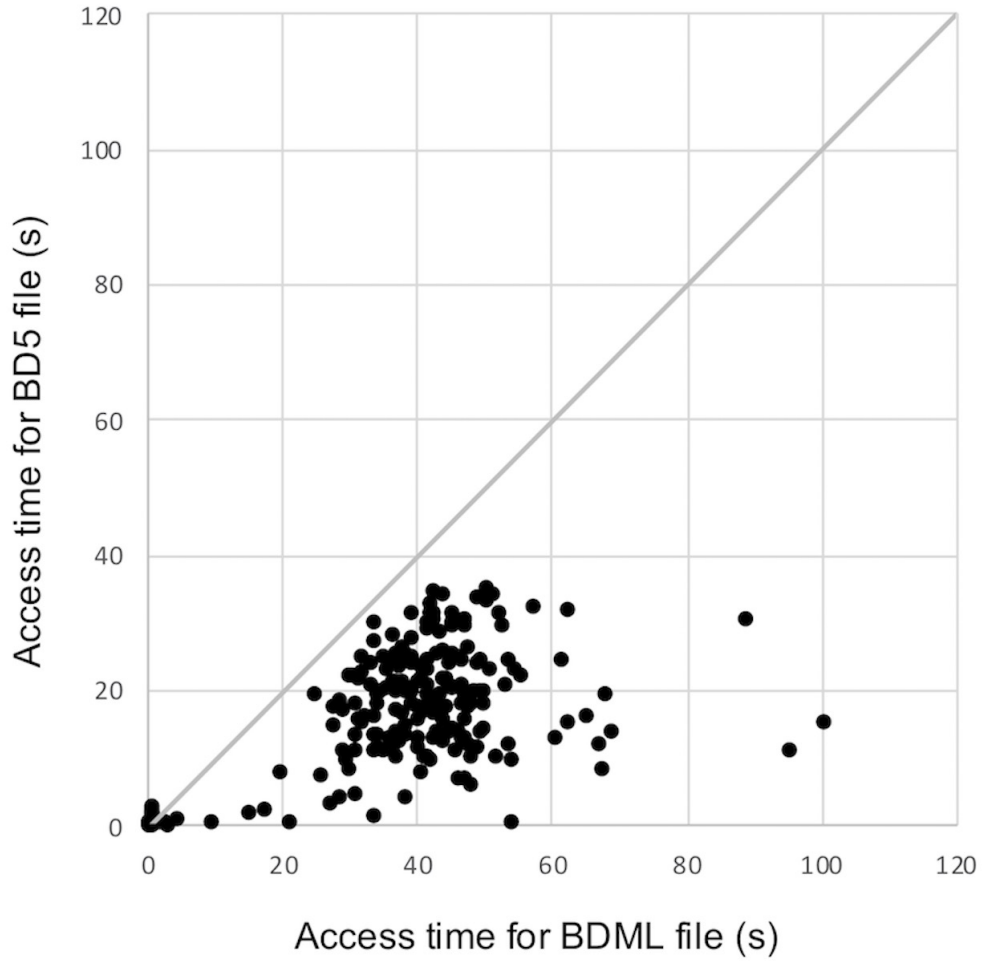
Example



(Bao et al., 2006)

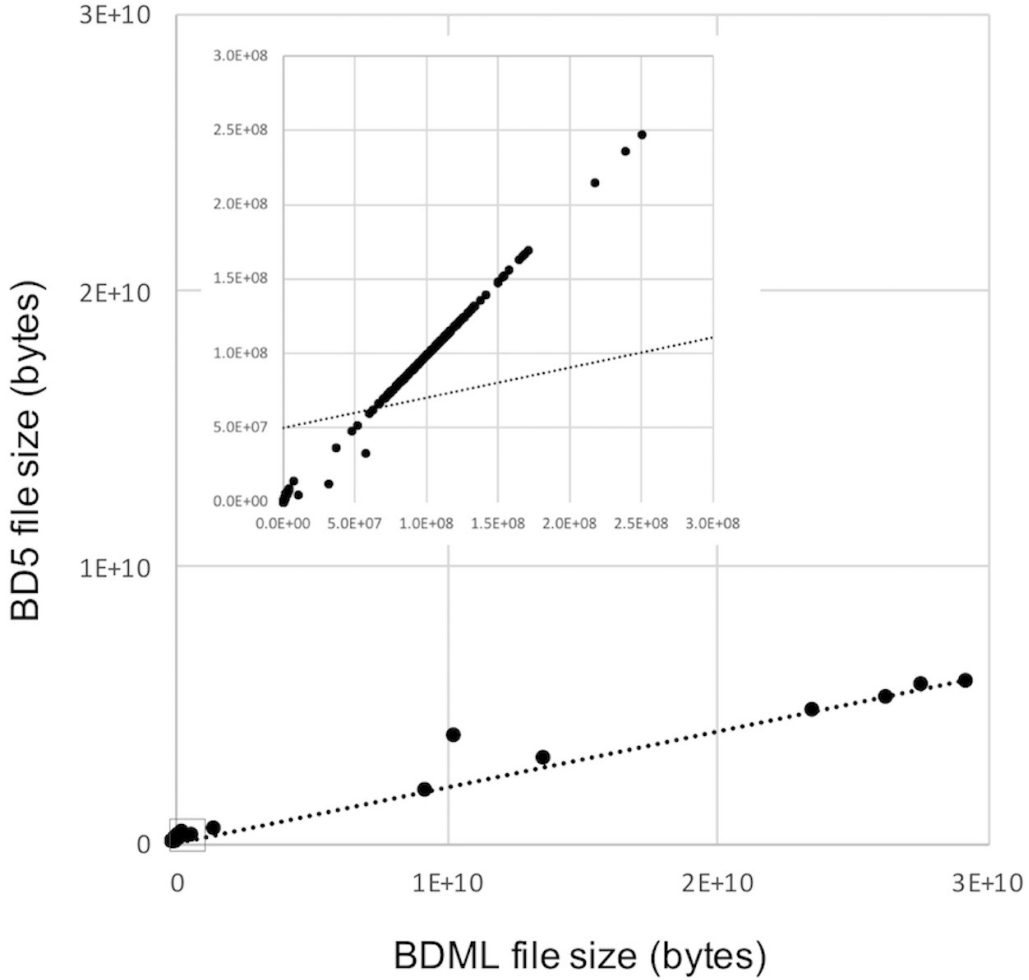
Fast data access

- Compared with BDML, BD5 enables fast access to quantitative data owing to random access to the HDF5-based file.



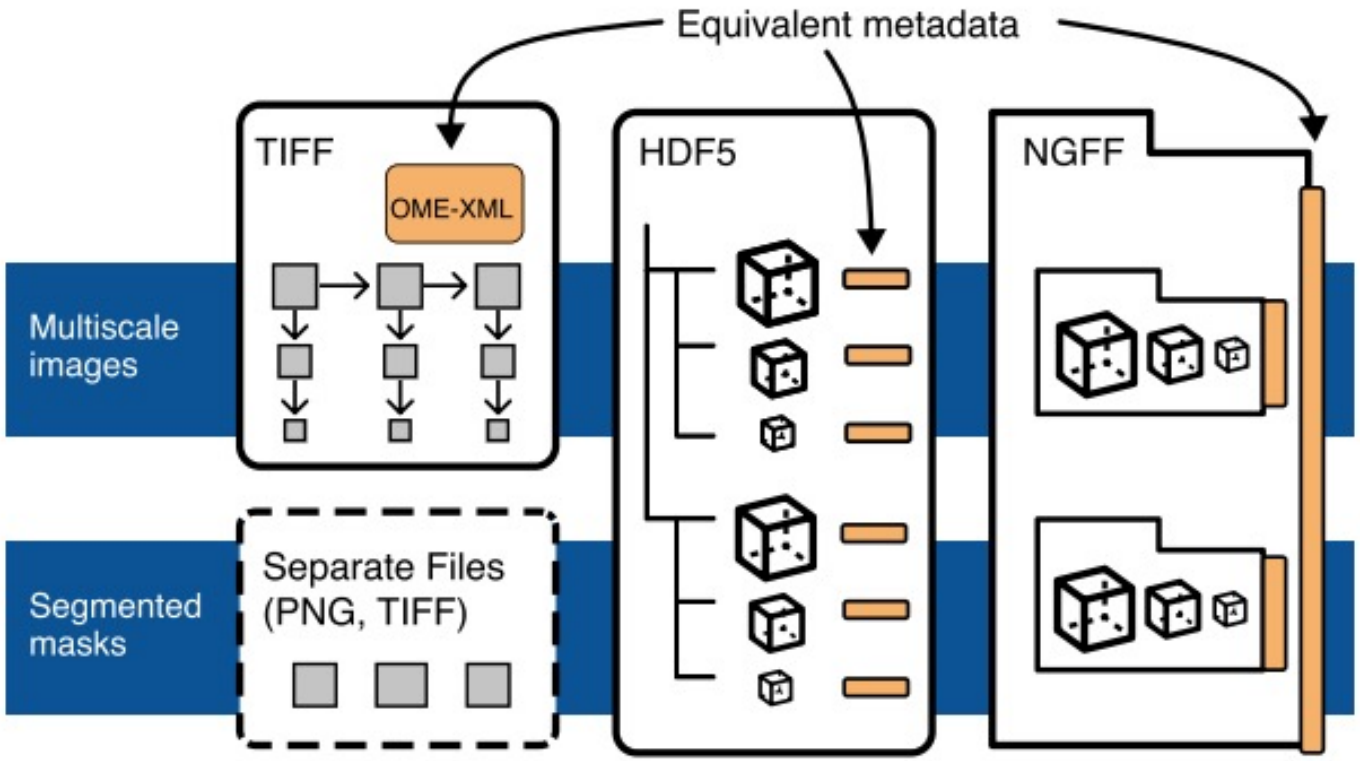
File size reduction

- BD5 enables fast transfer of large quantitative data because the file size is dramatically reduced.



Bioimaging data format

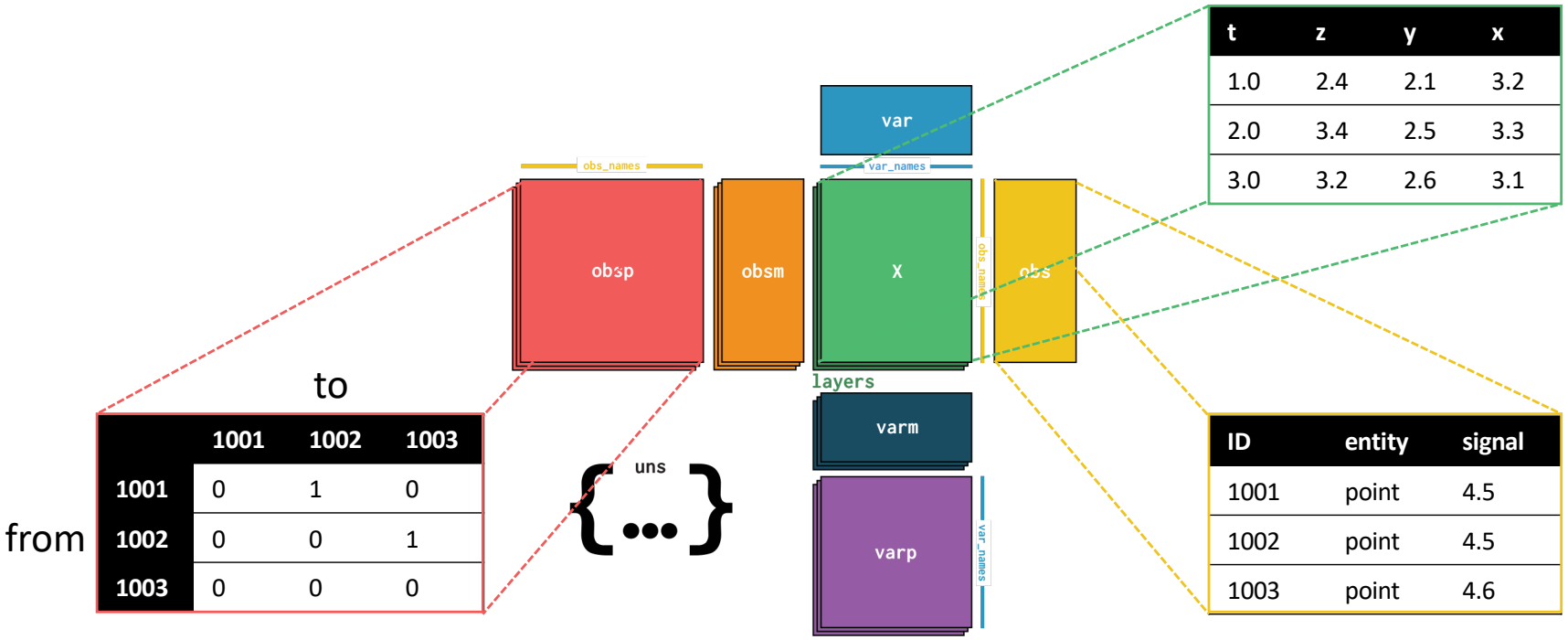
- Next generation file format for bioimaging data
  - ome-zarr is a zarr-based format for storing bioimaging data.



(Moore et al., Nat. Methods, 2021)

BD-zarr

- Dynamics data is stored in AnnData (https://anndata.readthedocs.io)
  - Store coordinates information of biological objects in ***X*** array
  - Store features information as separate ***obs*** array
  - Store tracking information as separate ***obsp*** array

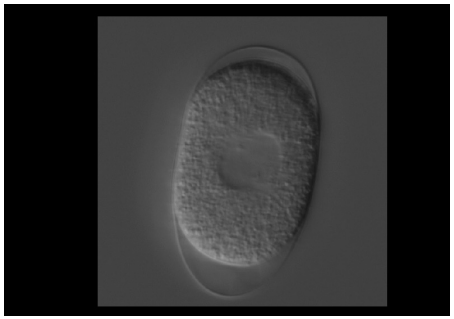


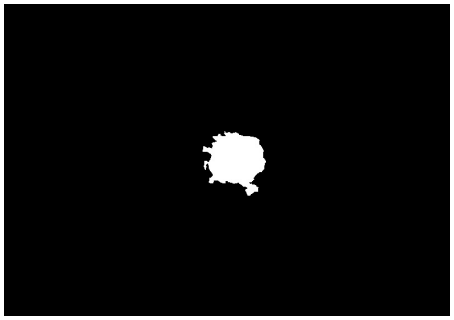
Example

- Early worm embryogenesis data (Kyoda et al., 2020)

```
wt-N2-081015-01
   |
   |--- 0
   |     |
   |     |--t
   |        |--c
   |           |--z
   |              |--y
   |                 |--x
   |
   |--- labels
   |      |
   |      |--0
   |         |
   |         |--t
   |            |--...
   |
   |--- dyn
          |
          |-- X
          |
          |-- obs
          |
          |-- obsp
```

image data





Pixel-based ROI data

Dynamics data

X position data

	t	z	y	x
0	1.0	39.922649	109.316254	307.414764
1	2.0	39.498207	113.885712	309.511322
2	3.0	39.999203	111.549751	315.685699
3	4.0	40.121613	112.833496	321.046295
4	5.0	42.206738	115.653198	332.908386

obs feature data

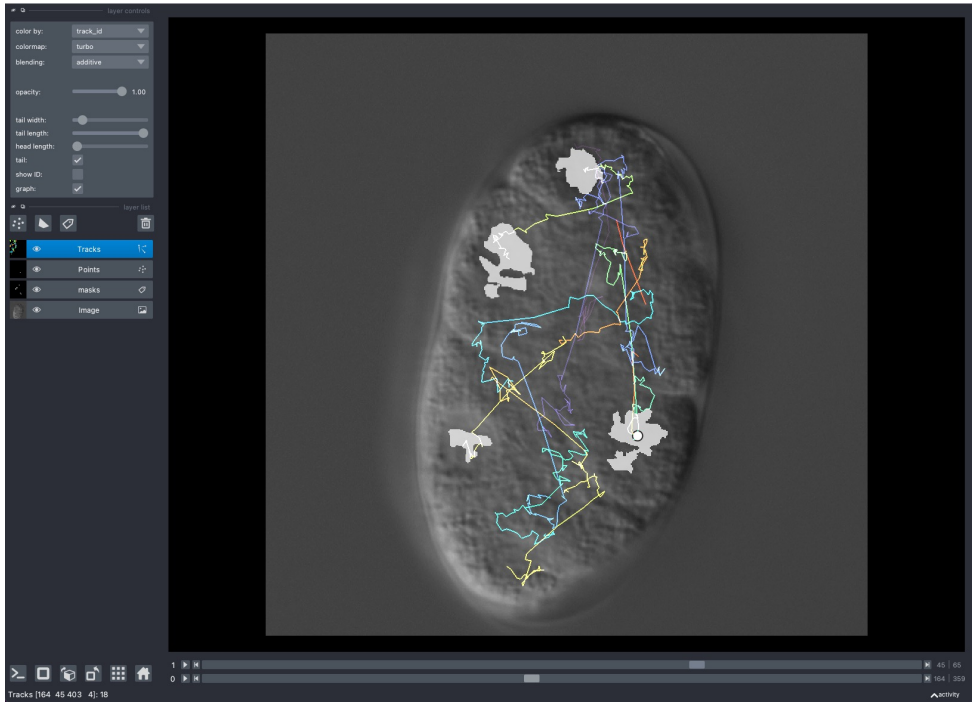
	id	entity	name	sphericity	volume
0	1000	line	P0	0.837884	76238.3
1	2000	line	P0	0.811657	88108.6
2	3000	line	P0	0.854491	110026.0
3	4000	line	P0	0.771704	107802.0
4	5000	line	P0	0.815484	138333.0

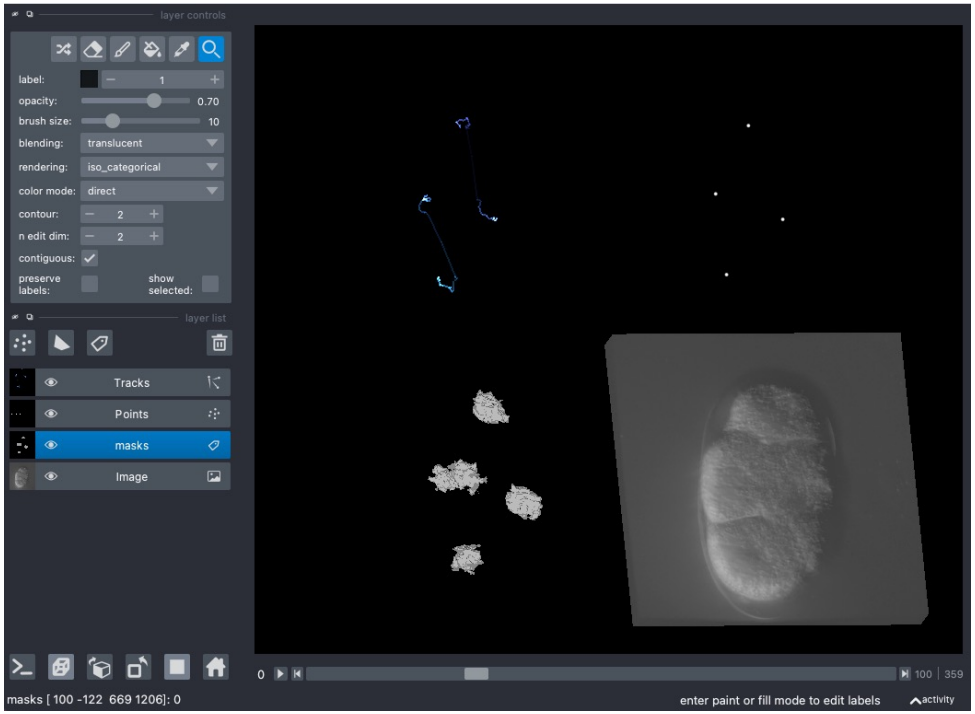
obsp tracking data

```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]], dtype=int8)
```

Visualization of BD-zarr data

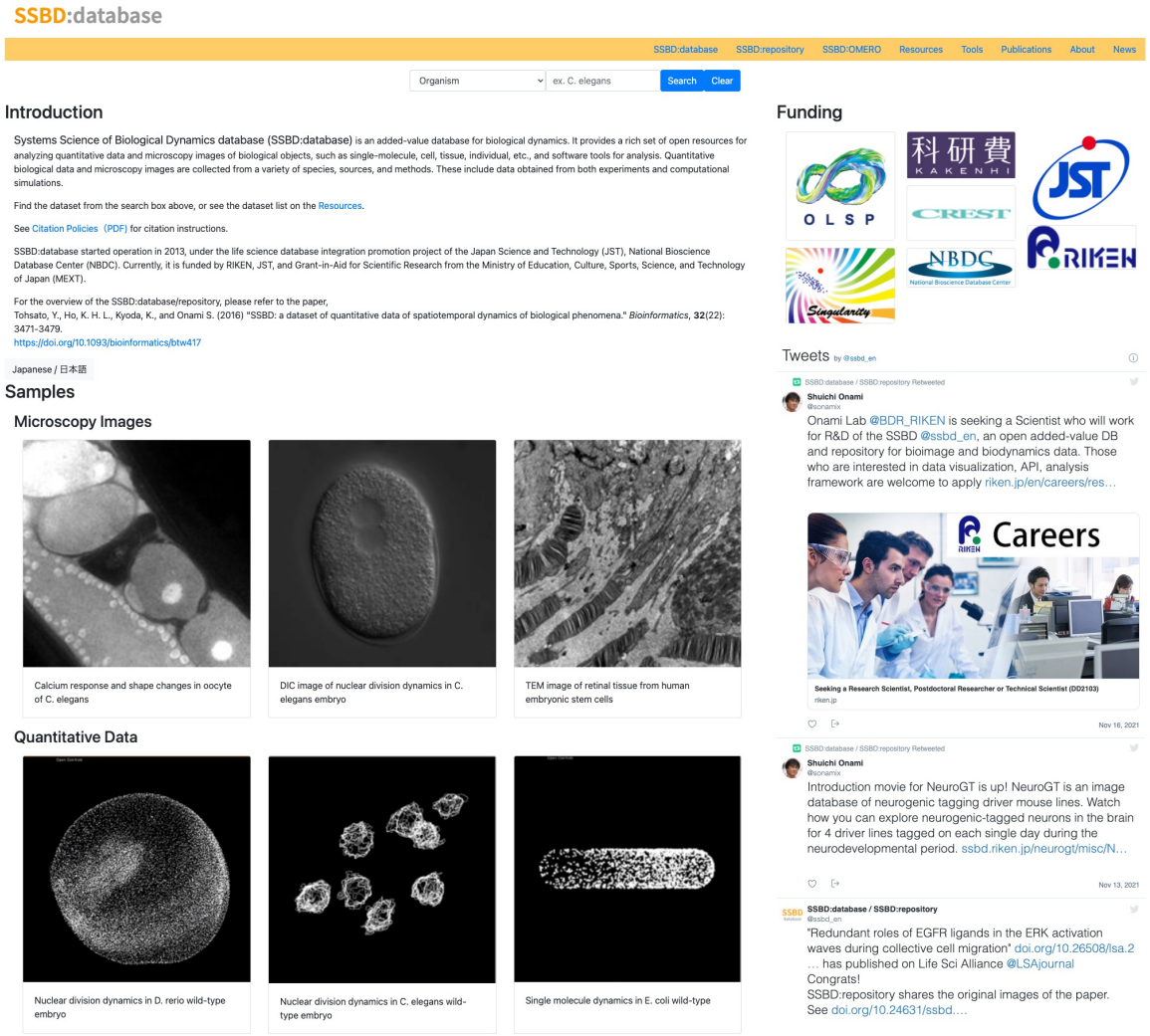
- with napari image viewer (https://napari.org/)





Data sharing of bioimaging data

- SSBD:database (https://ssbd.riken.jp) stores and shares quantitative data and image data of biological dynamics with rich meta data.



Summary

- We have developed BDML/BD5 based on XML/HDF5 for representing quantitative data of biological dynamics.
- Compared with BDML, the BD5 format has two advantages:
  - faster access and retrieval of quantitative data
  - Smaller file size, faster transfer of files in large datasets
- Following the current development in the bioimaging community, we are working on developing a Zarr-based format that are functionally compatible with BD5, HDF5-based format.

Acknowledgement

- RIKEN OLSP (Funding)

- Norio KOBAYASHI (RIKEN R-IH)
- Hideyuki Jitsumoto (RIKEN R-IH)
- Information Systems Division

- Bioimaging Community
  - Josh Moore (University of Dundee)
  - Kevin Yamauchi (ETH Zürich)
- Special acknowledgement to EMBL-EBI's Embassy Cloud and the BioImage Archive for providing valuable S3 storage and support for this project.