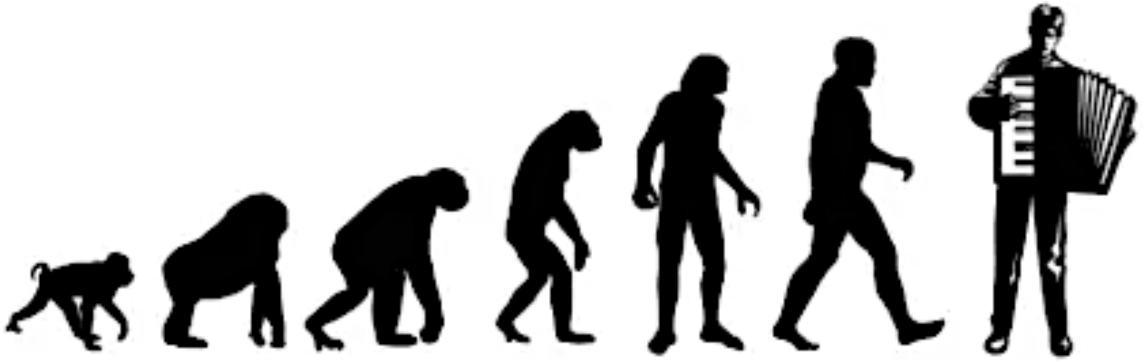
DIFFERENT THINKING with EXAMPLES

# CLIMBING OUT OF THE BOX



**MARK HERELD**
Senior Experimental Systems Engineer
Mathematics and Computer Science Division
Argonne National Laboratory
and the University of Chicago

Monday, October 21, 2019
Leiden, Netherlands

Biodiversity Informatics 101 Workshop
Biodiversity Next 2019

# HERE'S WHERE I AM COMING FROM…

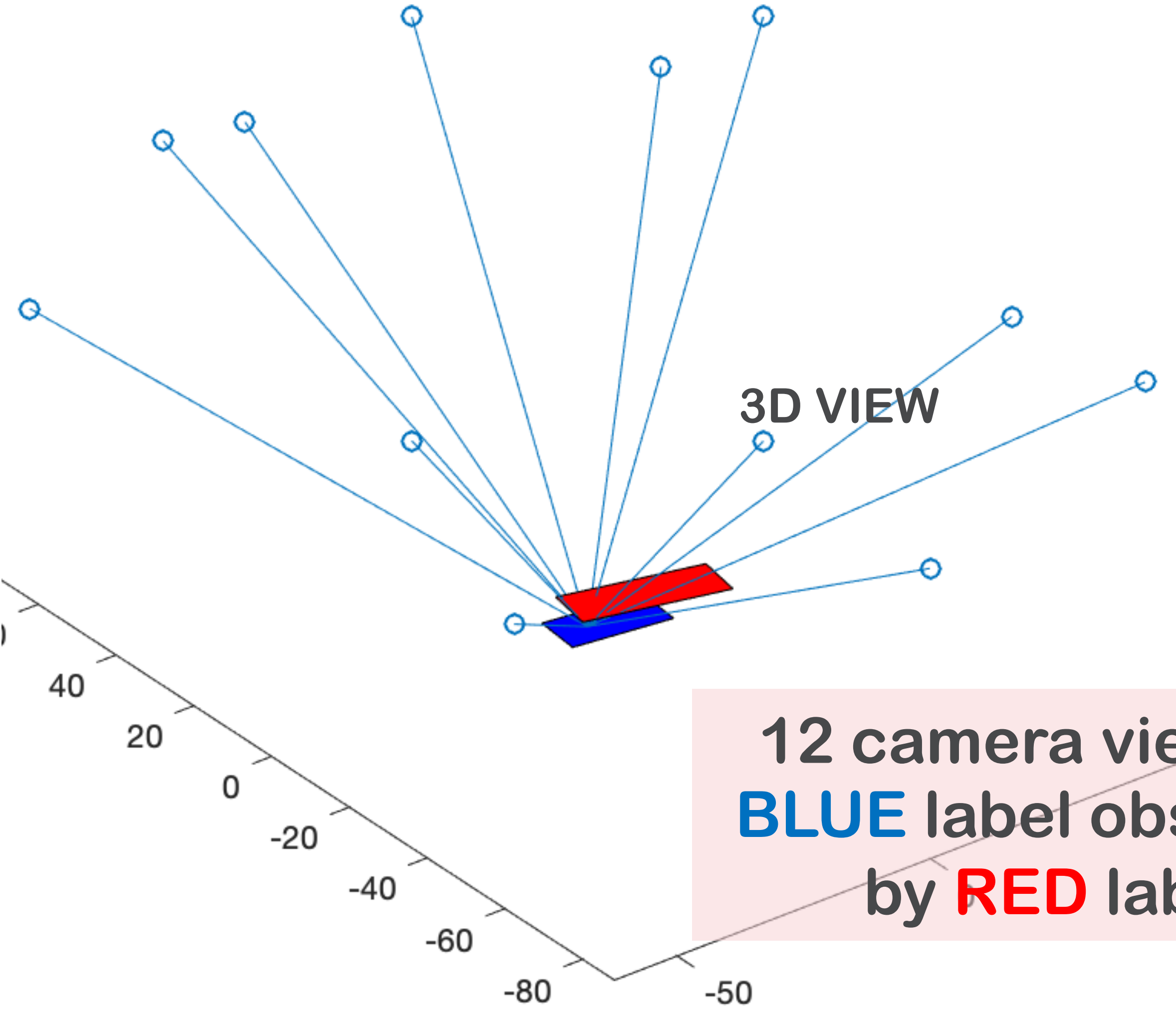
- **digitizing pinned insects fast enough** to complete the task in a matter of a few years rather than many decades or more than a century
- why: to enable **new biodiversity science** that would be otherwise impossible without this historically unique large-scale dataset
- oh, and… I'm a **physicist**, **computer scientist**, and **engineer**

- **look for bottlenecks** to solving the problem(s): technological, social, operational
- **challenge assumptions**: what are the real underlying requirements, what can be postponed or ignored, why has it been done this way until now?
- remember: **many solutions exist**, remain flexible, find one

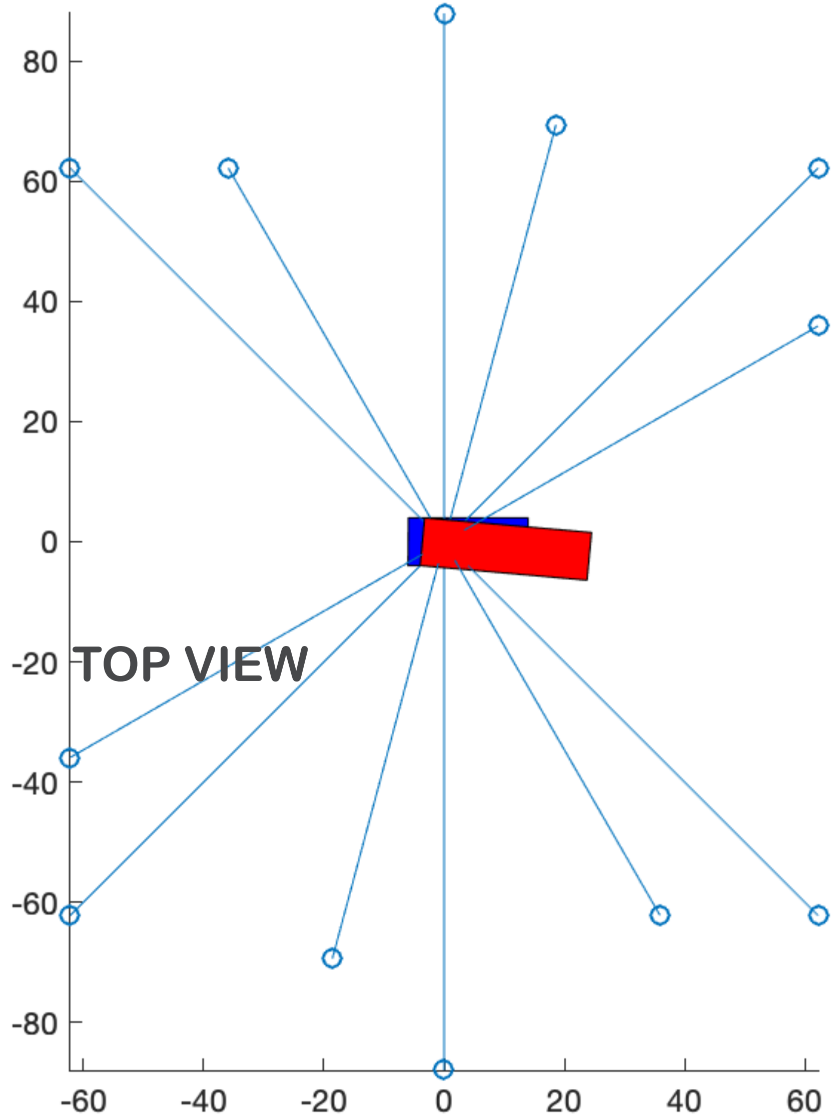
## DIGITIZING THE INFORMATION ON LABELS

# DO WE HAVE TO REMOVE THE LABELS FROM THE PINS?

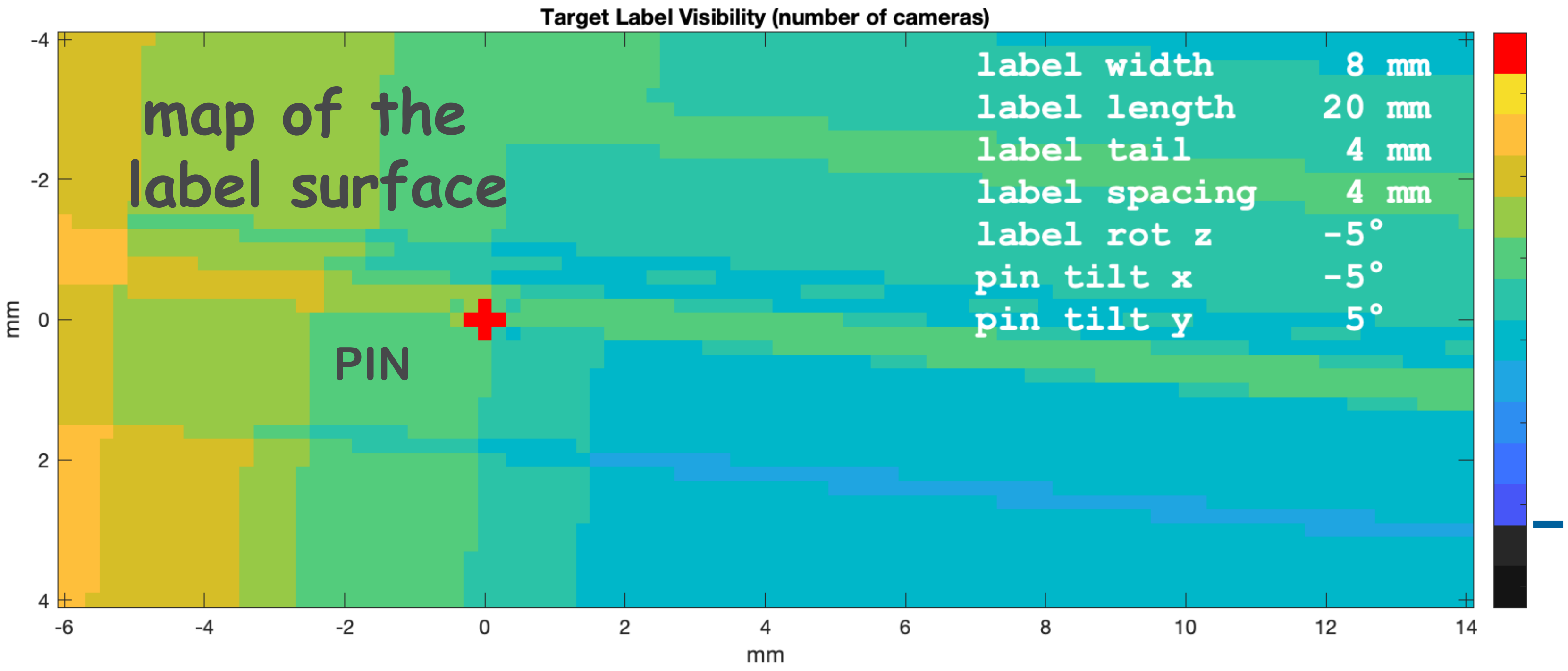
# MULTI-VIEW VISIBILITY OF PINNED LABELS



12 camera views of BLUE label obscured by RED label



# MULTI-VIEW VISIBILITY OF PINNED LABELS



| | |
|---|---|
| label width | 8 mm |
| label length | 20 mm |
| label tail | 4 mm |
| label spacing | 4 mm |
| label rot z | -5° |
| pin tilt x | -5° |
| pin tilt y | 5° |

**12** Seen by ALL cameras

**8**

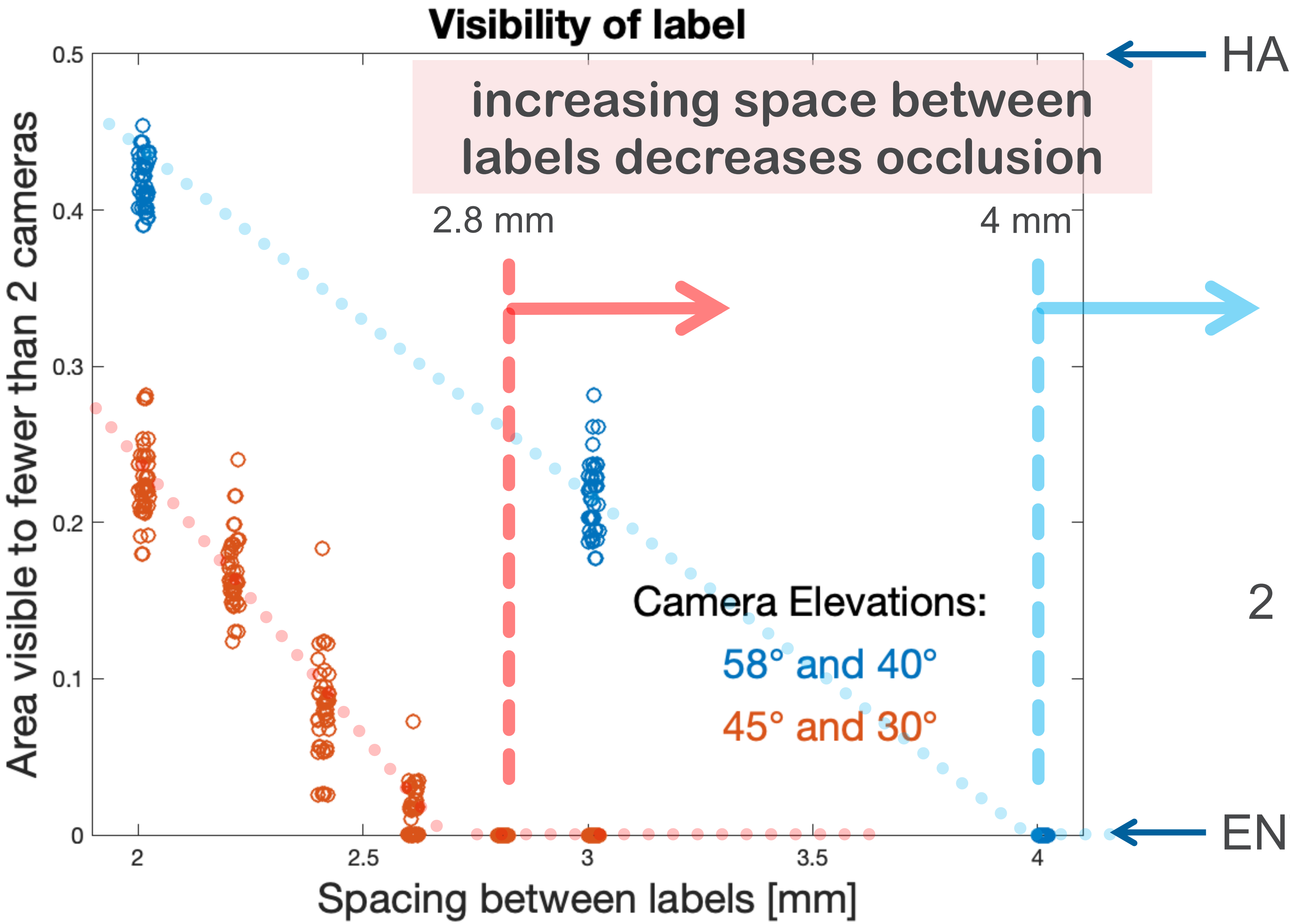
**4**

2 or MORE

**0** Seen by NO cameras

**heat map of number of cameras**
**with clear view of BLUE label**

# MULTI-VIEW VISIBILITY OF PINNED LABELS



increasing space between labels decreases occlusion

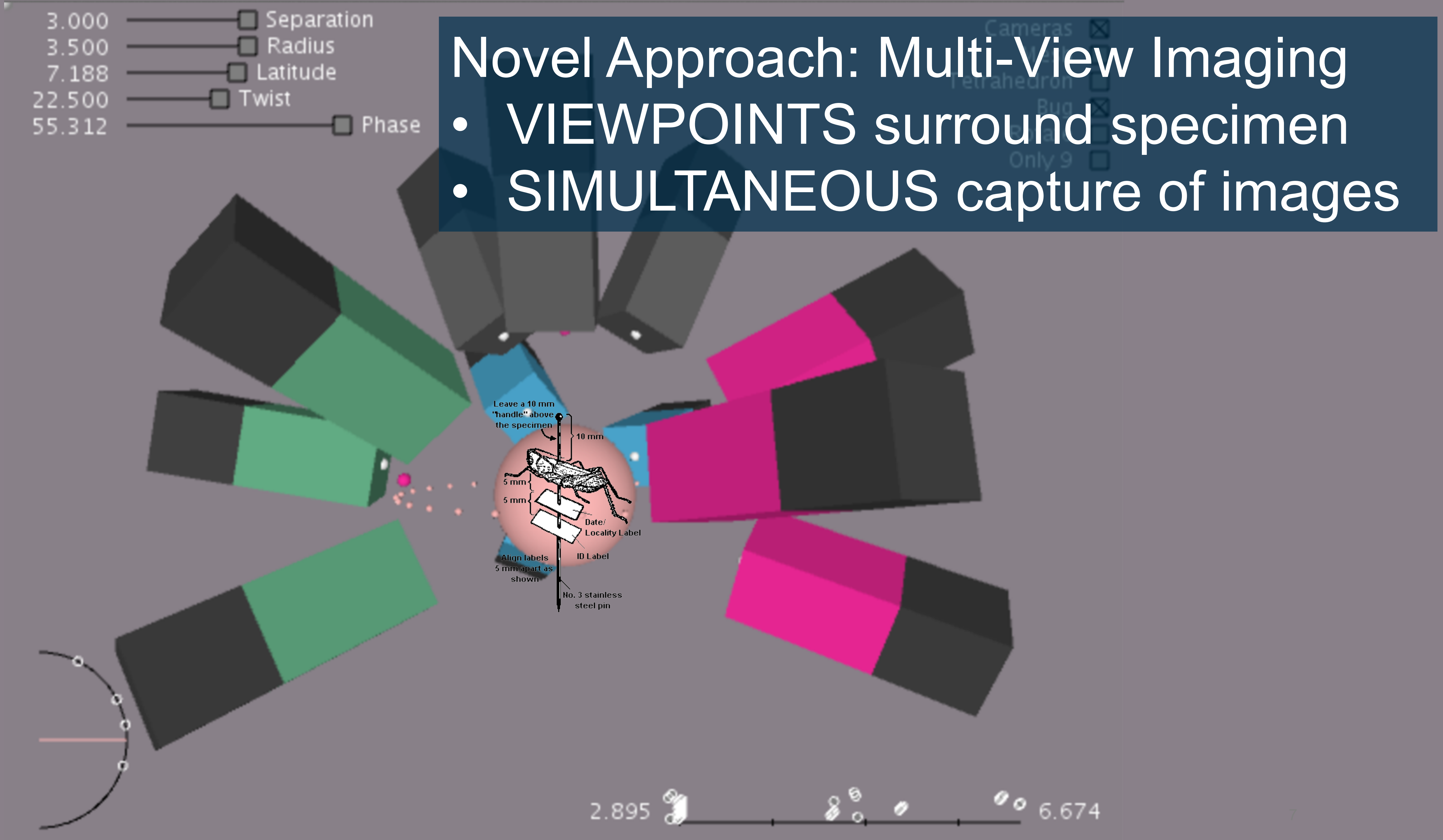
HALF of label VISIBLE

“VISIBLE” = 2 or MORE cameras

ENTIRE label VISIBLE

## Novel Approach: Multi-View Imaging

- VIEWPOINTS surround specimen
- SIMULTANEOUS capture of images



3.000 Separation
3.500 Radius
7.188 Latitude
22.500 Twist
55.312 Phase

Cameras ☒
[illegible]
Tetrahedron □
Bug ☒
[illegible]
Only 9 □

2.895 6.674

Leave a 10 mm "handle" above the specimen
10 mm
5 mm
5 mm
Date/ Locality Label
ID Label
Align labels 5 mm apart as shown
No. 3 stainless steel pin

# DIGITIZING A COLLECTION WITH MILLIONS OF SPECIMENS

CAN WE DO IT *FAST* ENOUGH TO FINISH BEFORE WE DIE?

# WORK FASTER: THINK “*PARALLEL*”

## THINK ABOUT STEPS

1st ● pull specimen from drawer and place on work area

2nd ● add unique bar code to the pin

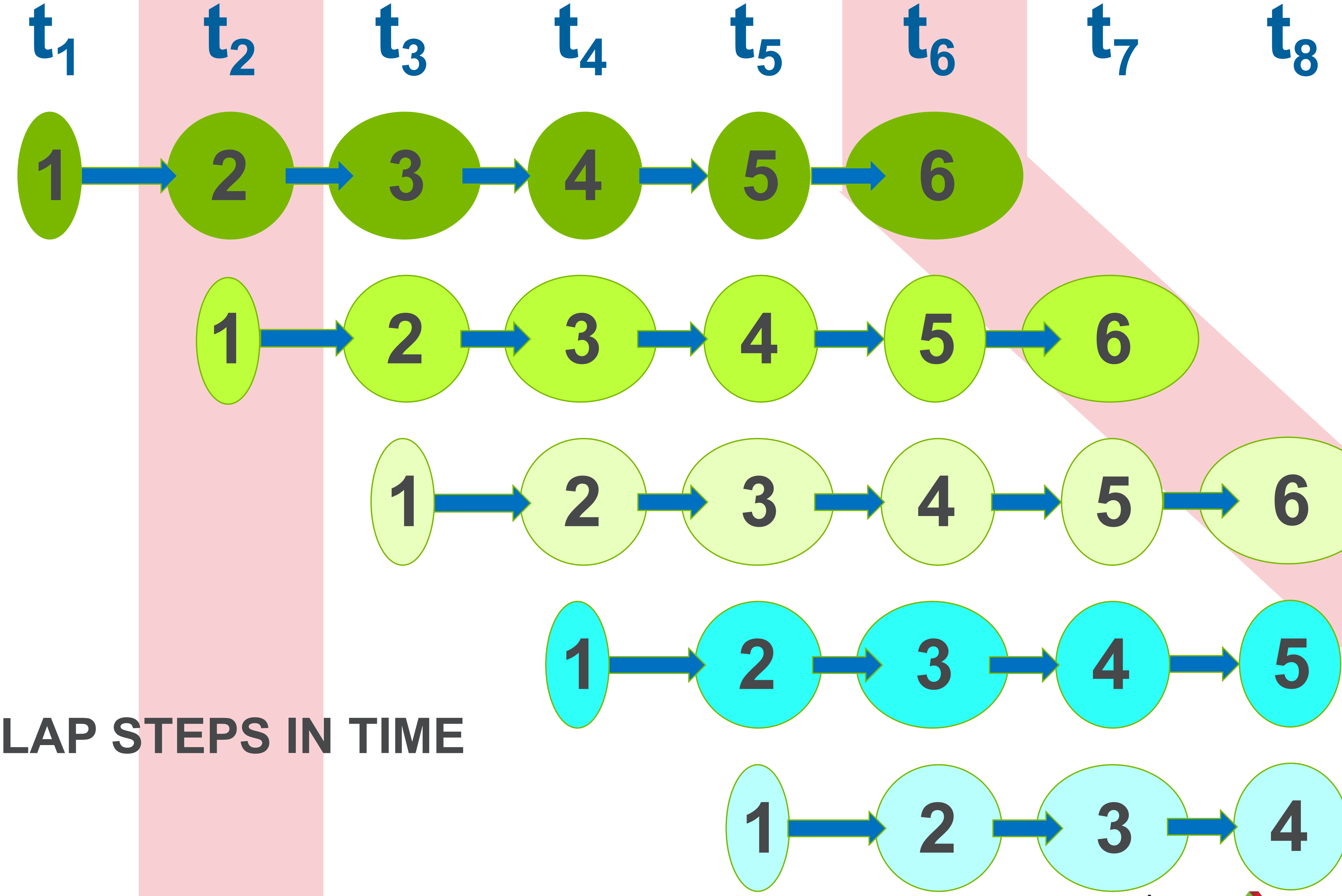
3rd ● space labels evenly along pin with special tool

4th ● place specimen in camera focal zone

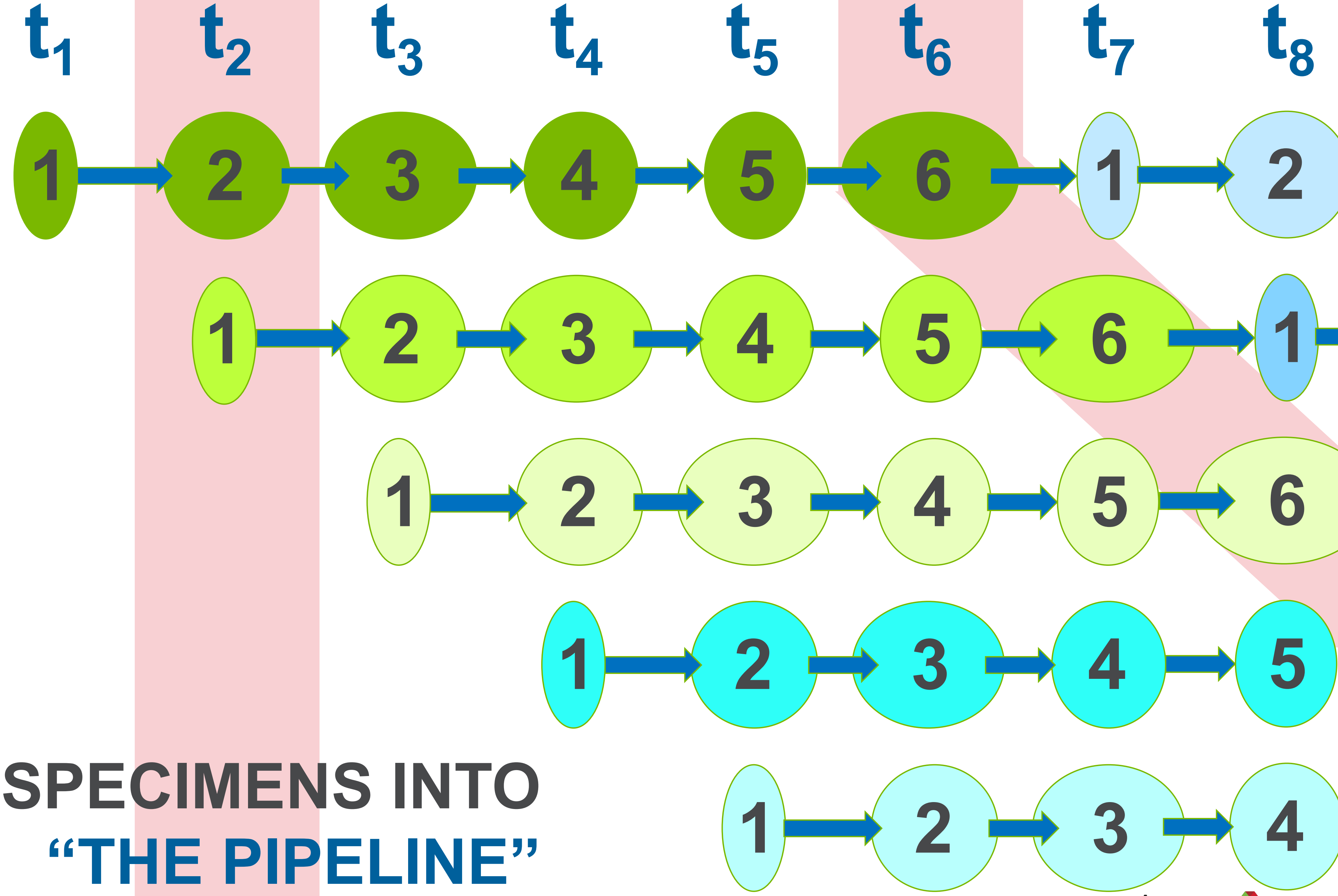
5th ● capture image of specimen

| | $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | |
|---|---|---|---|---|---|---|
| 1st | 1 | | | | | pull specimen from drawer and place on work area |
| 2nd | | 2 | | | | add unique bar code to the pin |
| 3rd | | | 3 | | | space labels evenly along pin with spe |
| 4th | | | | 4 | | place specimen in camera foc |
| 5th | | | | | 5 | capture image of spec |

THINK ABOUT TIME



# OVERLAP STEPS IN TIME



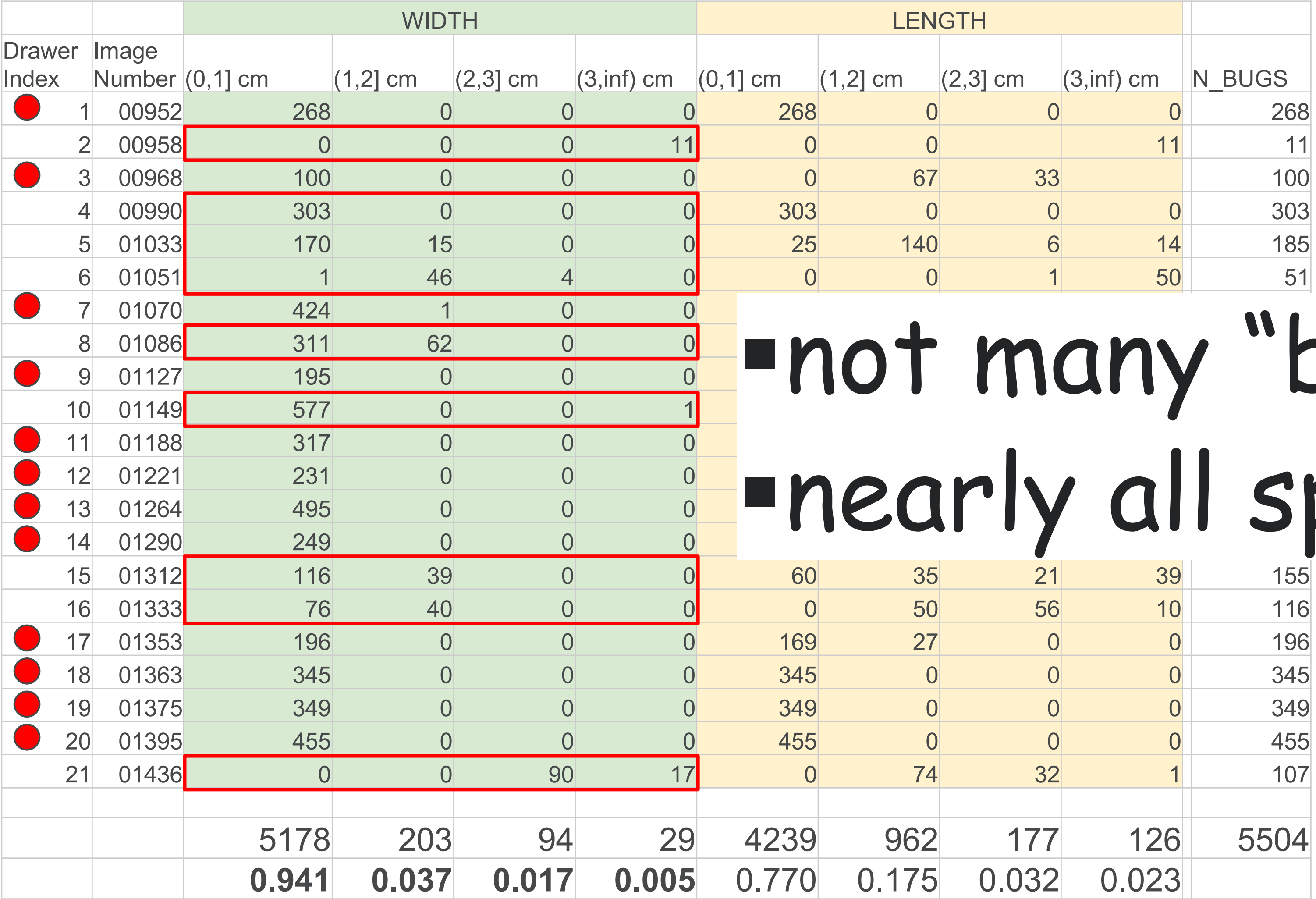
## INJECT SPECIMENS INTO "THE PIPELINE"

# BUT: BUTTERFLIES WILL BLOCK THE LABELS FROM VIEW

WILL THE "BUTTERFLIES" DEAL OUR QUEST A CRUSHING BLOW?

# TABULATE BUG SIZES

## random 21 drawers out of the 15 thousand in the collection



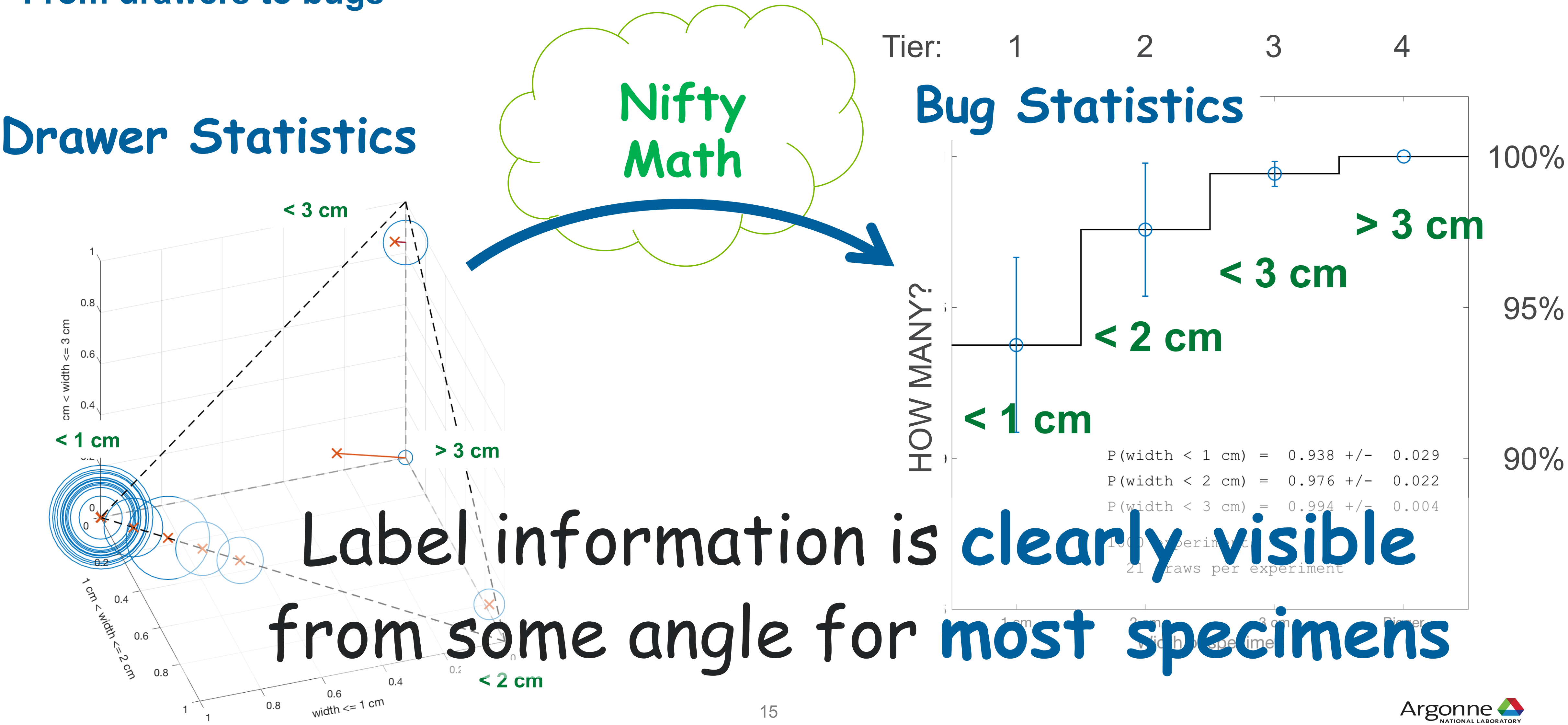
| | | WIDTH | | | | LENGTH | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Drawer Index | Image Number | (0,1] cm | (1,2] cm | (2,3] cm | (3,inf) cm | (0,1] cm | (1,2] cm | (2,3] cm | (3,inf) cm | N_BUGS |
| ● 1 | 00952 | 268 | 0 | 0 | 0 | 268 | 0 | 0 | 0 | 268 |
| 2 | 00958 | 0 | 0 | 0 | 11 | 0 | 0 | | 11 | 11 |
| ● 3 | 00968 | 100 | 0 | 0 | 0 | 0 | 67 | 33 | | 100 |
| 4 | 00990 | 303 | 0 | 0 | 0 | 303 | 0 | 0 | 0 | 303 |
| 5 | 01033 | 170 | 15 | 0 | 0 | 25 | 140 | 6 | 14 | 185 |
| 6 | 01051 | 1 | 46 | 4 | 0 | 0 | 0 | 1 | 50 | 51 |
| ● 7 | 01070 | 424 | 1 | 0 | 0 | | | | | |
| 8 | 01086 | 311 | 62 | 0 | 0 | | | | | |
| ● 9 | 01127 | 195 | 0 | 0 | 0 | | | | | |
| 10 | 01149 | 577 | 0 | 0 | 1 | | | | | |
| ● 11 | 01188 | 317 | 0 | 0 | 0 | | | | | |
| ● 12 | 01221 | 231 | 0 | 0 | 0 | | | | | |
| ● 13 | 01264 | 495 | 0 | 0 | 0 | | | | | |
| ● 14 | 01290 | 249 | 0 | 0 | 0 | | | | | |
| 15 | 01312 | 116 | 39 | 0 | 0 | 60 | 35 | 21 | 39 | 155 |
| 16 | 01333 | 76 | 40 | 0 | 0 | 0 | 50 | 56 | 10 | 116 |
| ● 17 | 01353 | 196 | 0 | 0 | 0 | 169 | 27 | 0 | 0 | 196 |
| ● 18 | 01363 | 345 | 0 | 0 | 0 | 345 | 0 | 0 | 0 | 345 |
| ● 19 | 01375 | 349 | 0 | 0 | 0 | 349 | 0 | 0 | 0 | 349 |
| ● 20 | 01395 | 455 | 0 | 0 | 0 | 455 | 0 | 0 | 0 | 455 |
| 21 | 01436 | 0 | 0 | 90 | 17 | 0 | 74 | 32 | 1 | 107 |
| | | | | | | | | | | |
| | | 5178 | 203 | 94 | 29 | 4239 | 962 | 177 | 126 | 5504 |
| | | **0.941** | **0.037** | **0.017** | **0.005** | 0.770 | 0.175 | 0.032 | 0.023 | |

- not many “butterflies”
- nearly all specimens < 2 cm

# HOW MANY ARE SMALLER THAN 2 CM?

**98%**

## From drawers to bugs



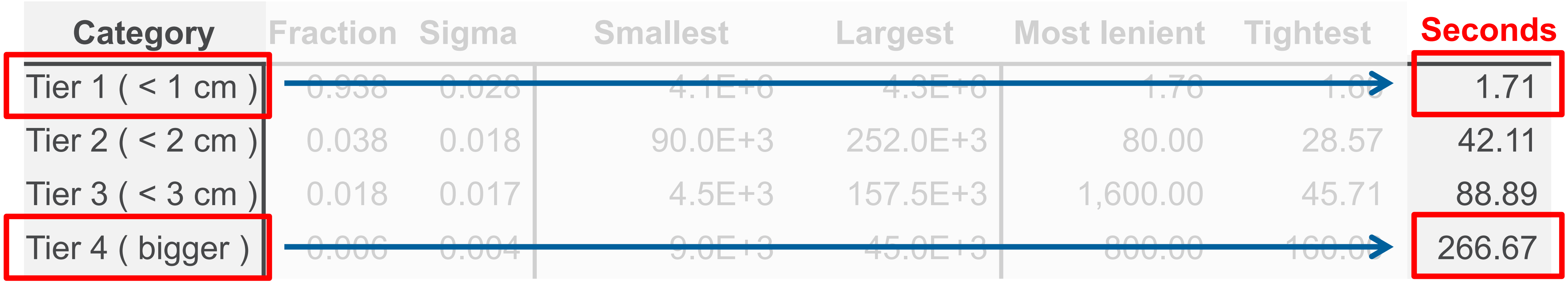
```
P(width < 1 cm) =  0.938 +/-  0.029
P(width < 2 cm) =  0.976 +/-  0.022
P(width < 3 cm) =  0.994 +/-  0.004

1000 experiments
21 draws per experiment
```

Label information is **clearly visible** from some angle for **most specimens**

# HOW MUCH TIME CAN WE SPEND ON EACH SPECIMEN?

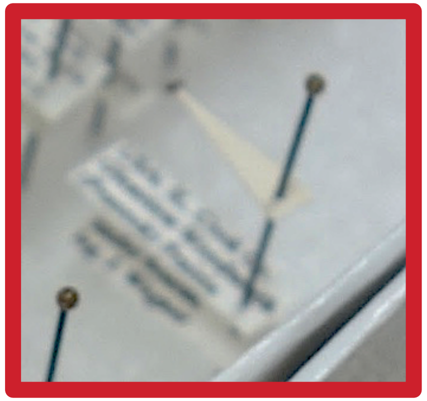
**EQUATION: (Time Budget / Number of Specimens)**

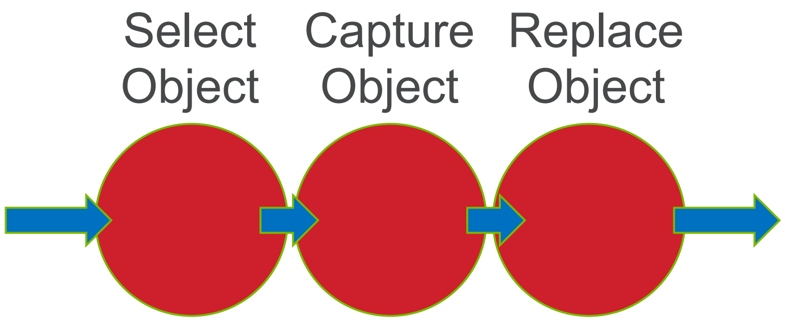


| Category | Fraction | Sigma | Smallest | Largest | Most lenient | Tightest | Seconds |
|---|---|---|---|---|---|---|---|
| Tier 1 ( < 1 cm ) | 0.938 | 0.028 | 4.1E+6 | 4.3E+6 | 1.76 | 1.66 | 1.71 |
| Tier 2 ( < 2 cm ) | 0.038 | 0.018 | 90.0E+3 | 252.0E+3 | 80.00 | 28.57 | 42.11 |
| Tier 3 ( < 3 cm ) | 0.018 | 0.017 | 4.5E+3 | 157.5E+3 | 1,600.00 | 45.71 | 88.89 |
| Tier 4 ( bigger ) | 0.006 | 0.004 | 9.0E+3 | 45.0E+3 | 800.00 | 160.00 | 266.67 |

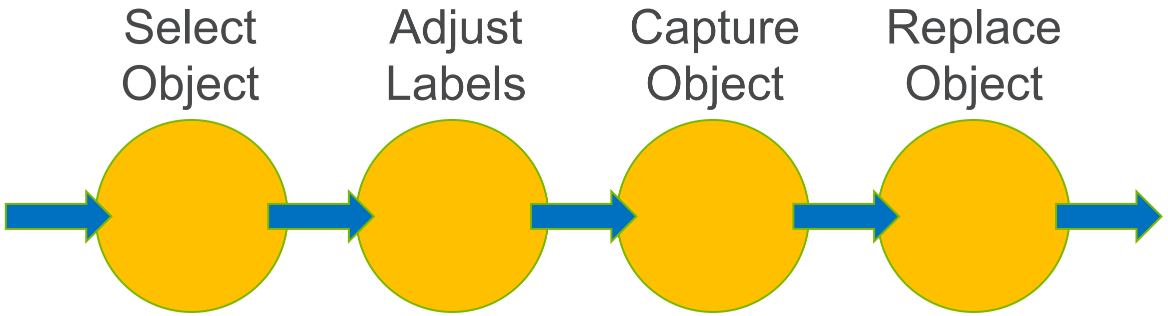
## Label information is **clearly visible** from some angle for **most specimens**

# DIFFERENT PIPELINES FOR EACH LEVEL OF DIFFICULTY

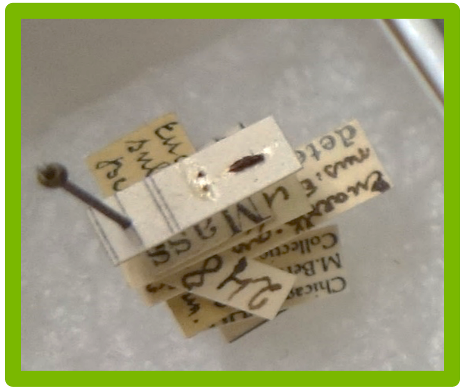


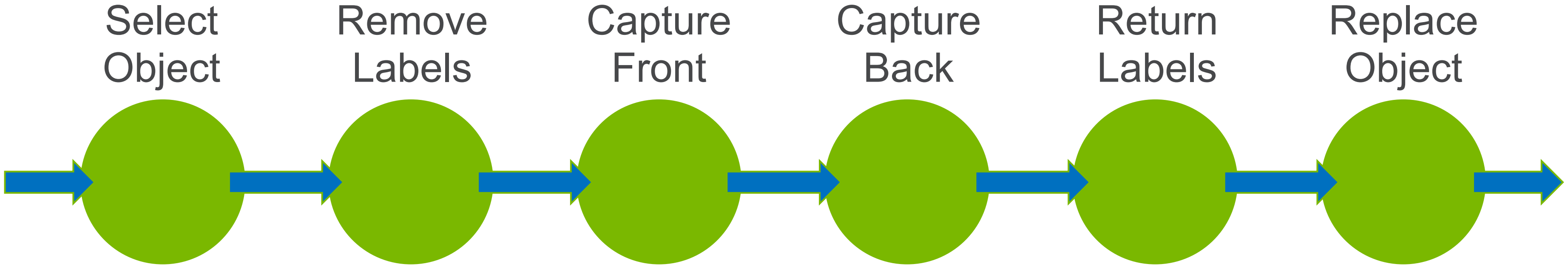


Select Object → Capture Object → Replace Object

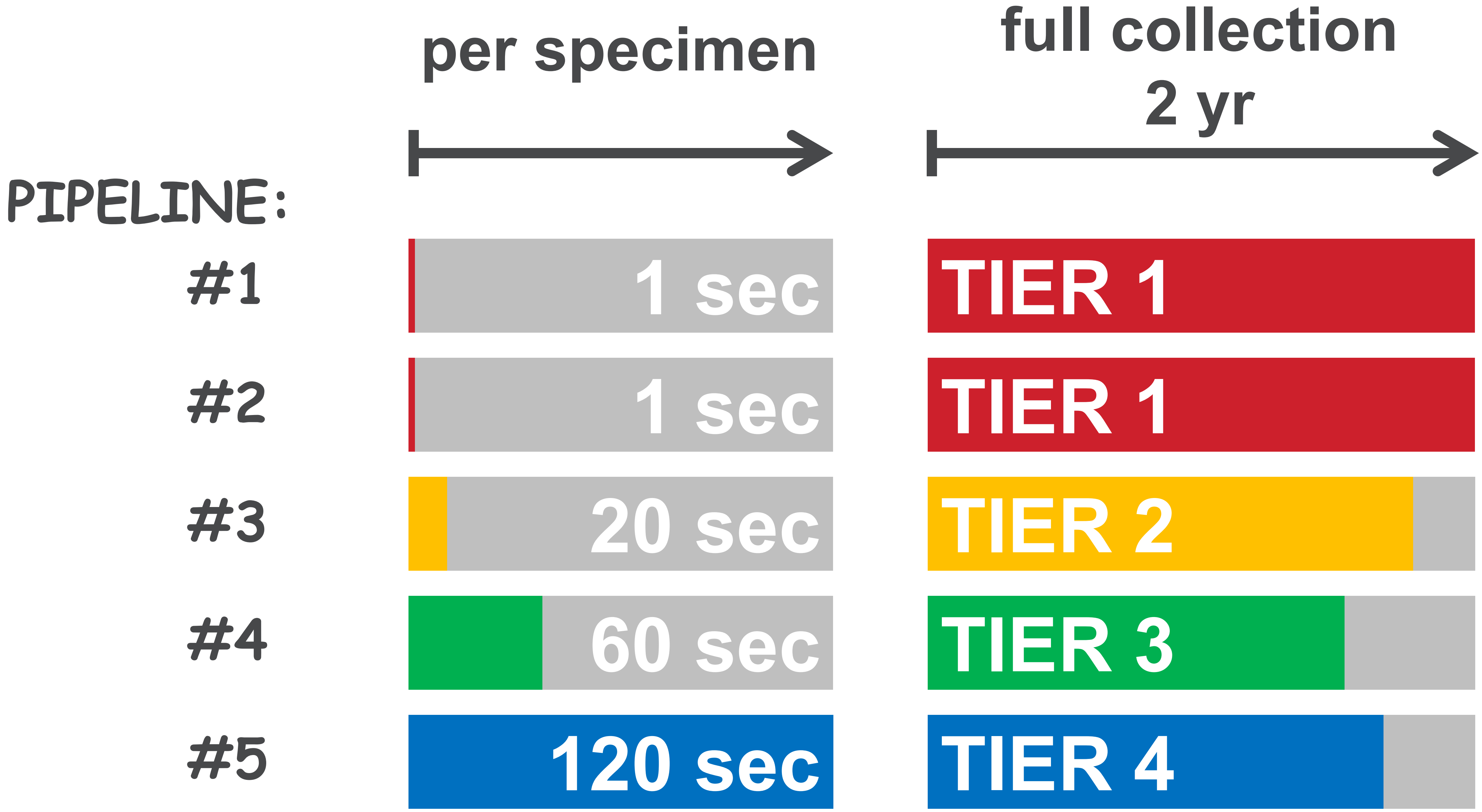


Select Object → Adjust Labels → Capture Object → Replace Object





Select Object → Remove Labels → Capture Front → Capture Back → Return Labels → Replace Object



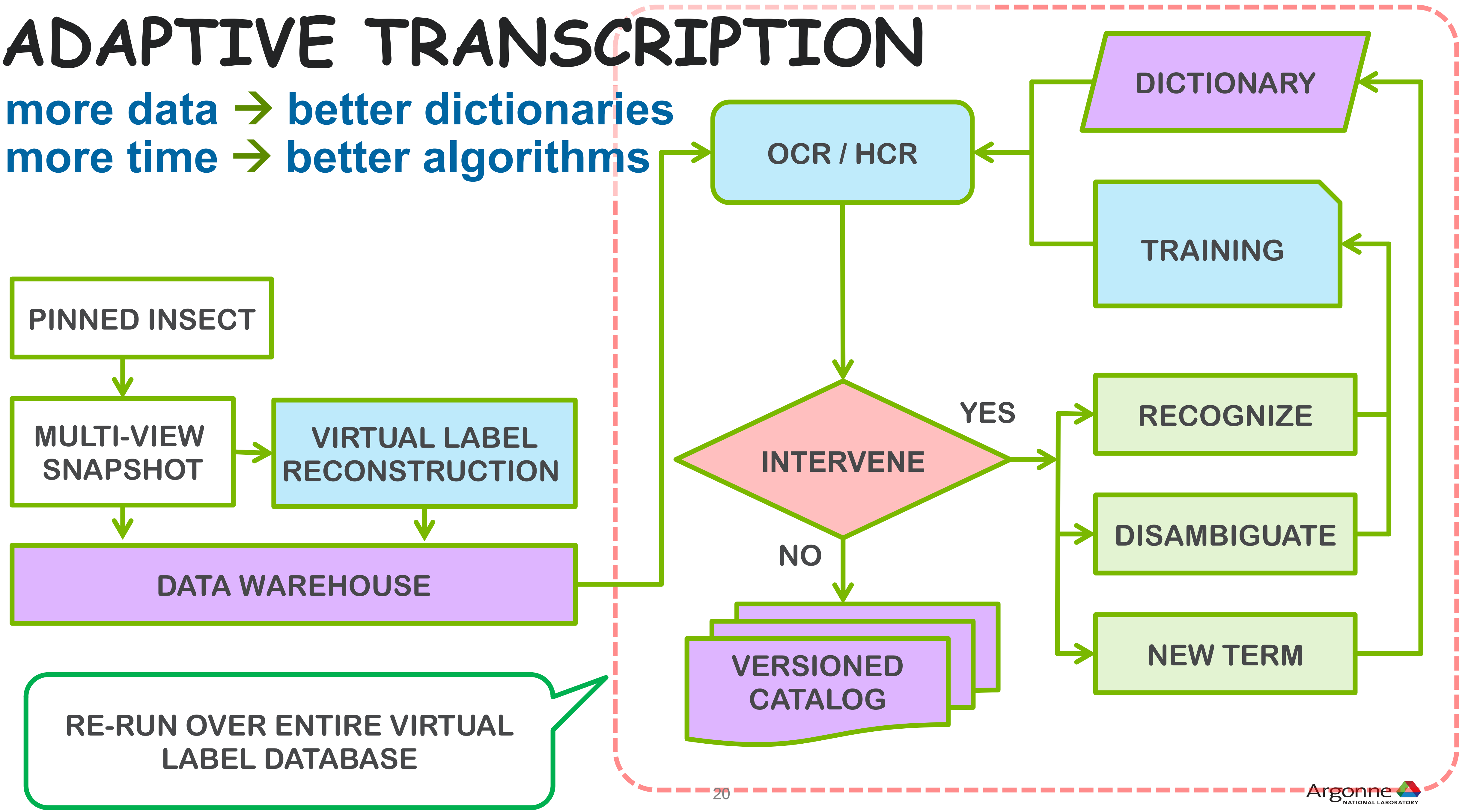
| PIPELINE: | per specimen | full collection 2 yr |
|---|---|---|
| #1 | 1 sec | TIER 1 |
| #2 | 1 sec | TIER 1 |
| #3 | 20 sec | TIER 2 |
| #4 | 60 sec | TIER 3 |
| #5 | 120 sec | TIER 4 |

# TRANSCRIBING FROM IMAGE TO DATABASE

DO WE HAVE TO SOLVE THIS PROBLEM RIGHT NOW?

# ADAPTIVE TRANSCRIPTION

more data → better dictionaries
more time → better algorithms



PINNED INSECT

MULTI-VIEW SNAPSHOT

VIRTUAL LABEL RECONSTRUCTION

DATA WAREHOUSE

OCR / HCR

DICTIONARY

TRAINING

INTERVENE

YES

NO

RECOGNIZE

DISAMBIGUATE

NEW TERM

VERSIONED CATALOG

RE-RUN OVER ENTIRE VIRTUAL LABEL DATABASE

# THE MESSAGE

- **look for bottlenecks** to solving the problem(s): technological, social, operational

- **challenge assumptions**: what are the real underlying requirements, what can be postponed or ignored, why has it been done this way until now?

- remember: many solutions exist, remain flexible, **your solution is out there**

# ADDITIONAL INFO

- Come to my talk!: “LightningBug ONE” in *SI67 Digitization Next* symposium, Wed 13:30 – 15:00

- Project Links:
  - http://lightningbug.tech
  - https://silo18.github.io/LightningBugONE/

- Hereld M, Ferrier N (2019) LightningBug ONE: An Experiment in High-Throughtput Digitization of Pinned Insects. Biodiversity Information Science and Standards 3: e37228. https://biss.pensoft.net/article/37228/
- Agarwal N, Ferrier N, Hereld M (2018) Towards Automated Transcription of Label Text from Pinned Insect Collections. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* https://doi.org/10.1109/wacv.2018.00027
- Hereld M, Ferrier N, Agarwal N, Sierwald P (2017) Designing a High-Throughput Pipeline for Digitizing Pinned Insects. *2017 IEEE 13th International Conference on e-Science (e-Science)* https://doi.org/10.1109/escience.2017.88