Advanced Data Management (CSCI 490/680)

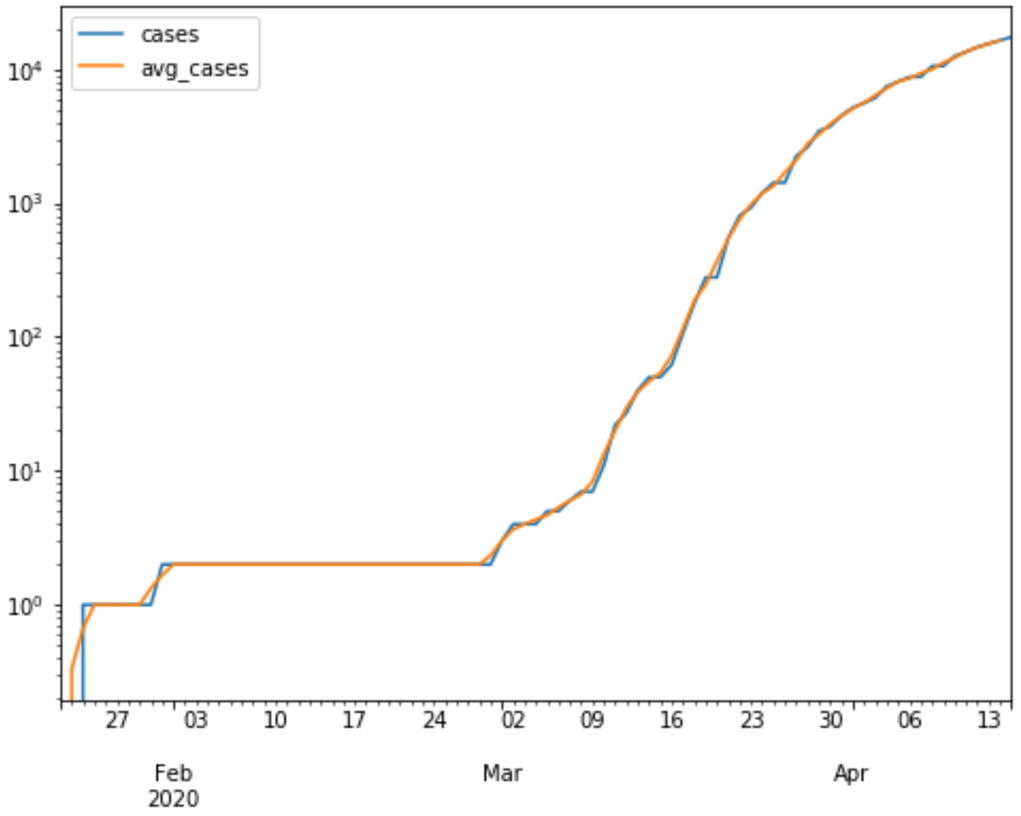
Machine Learning and Databases

Dr. David Koop

Reading Quiz

- Before continuing this lecture, go to Blackboard and complete the reading quiz on today's reading

Assignment 5

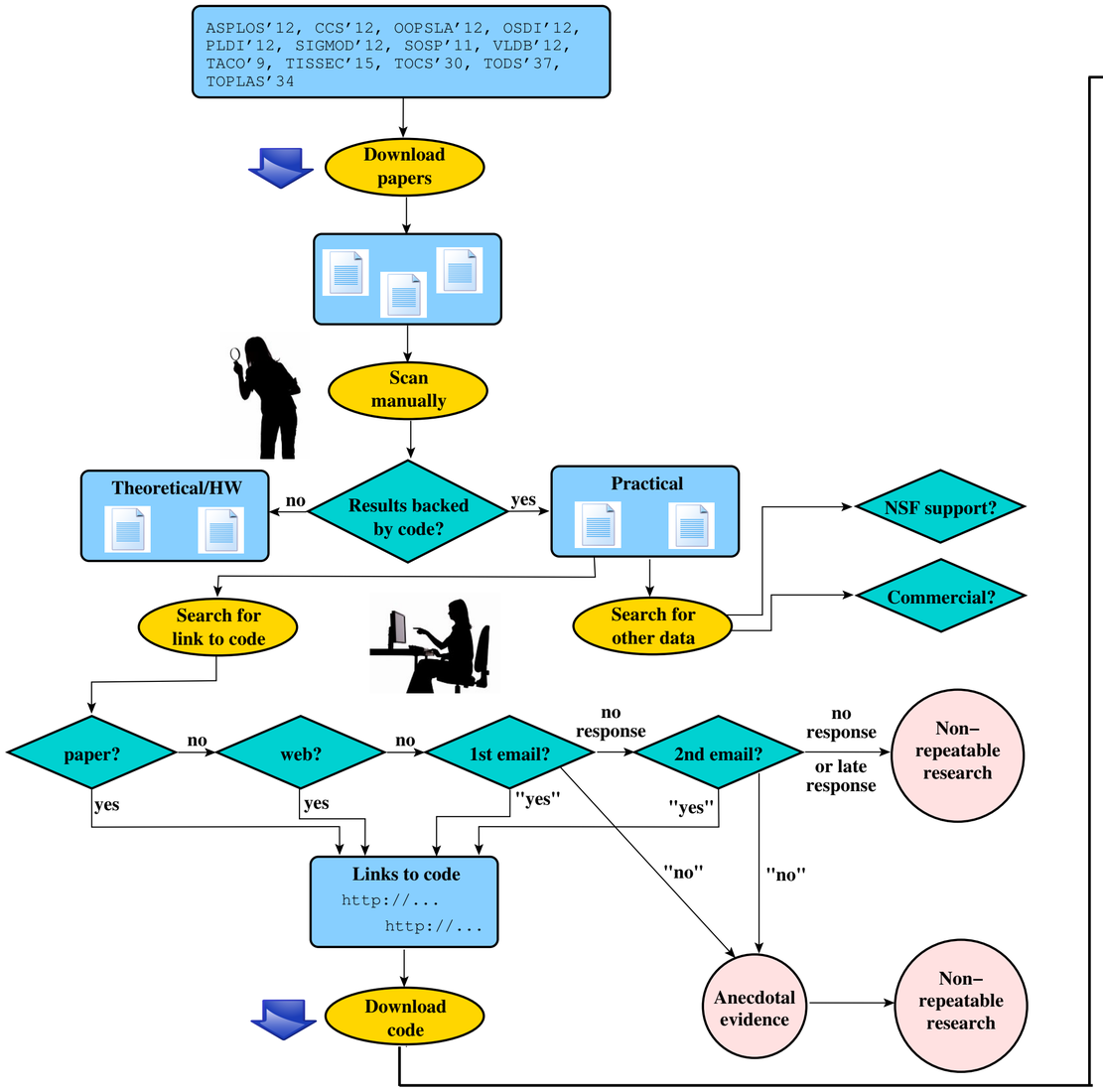


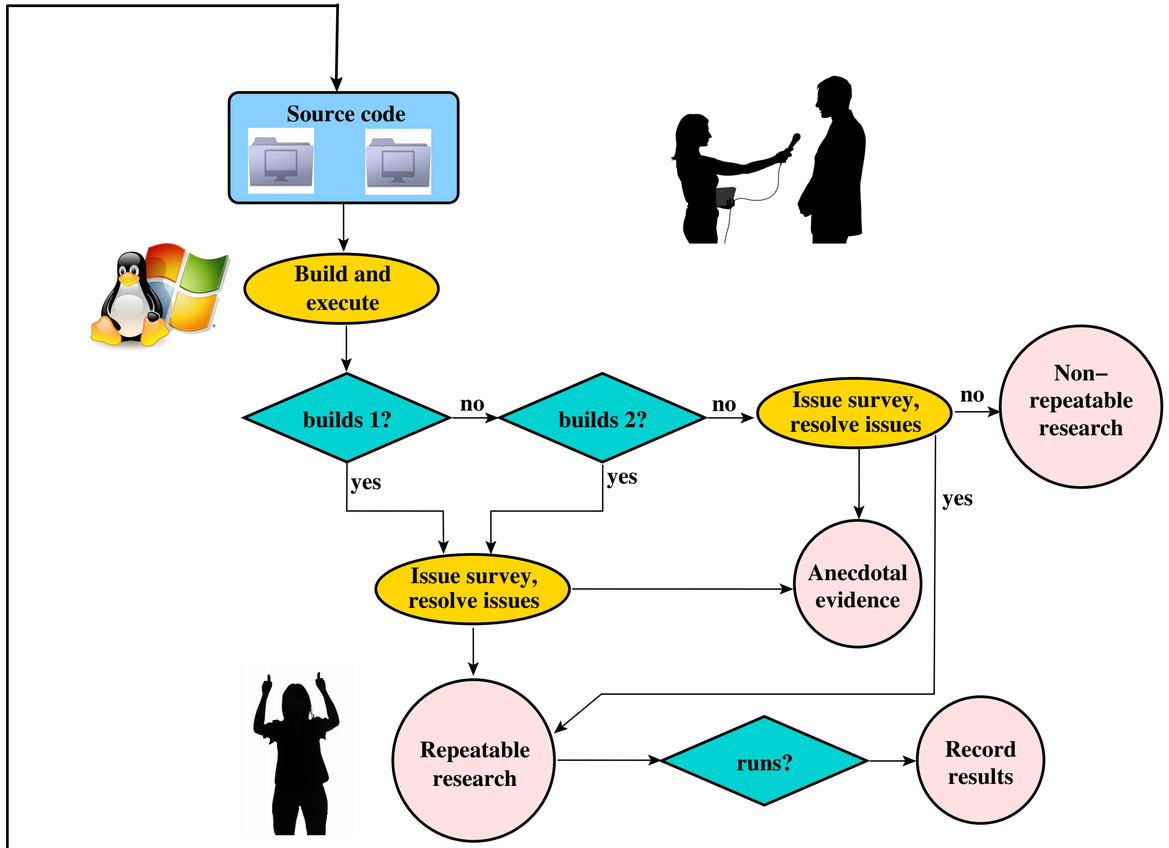
- Due Thursday
- Questions?
- Note about `%-m` strftime conversion: use `%#m` on Windows

Final Exam and Review

- Final Exam
  - Tuesday, May 5 from 4-5:50pm
  - Online
  - Similar format to Test 2
  - Comprehensive but with more focus on last few weeks of class
- Review
  - Thursday, April 30
  - Submit questions via email or discussion

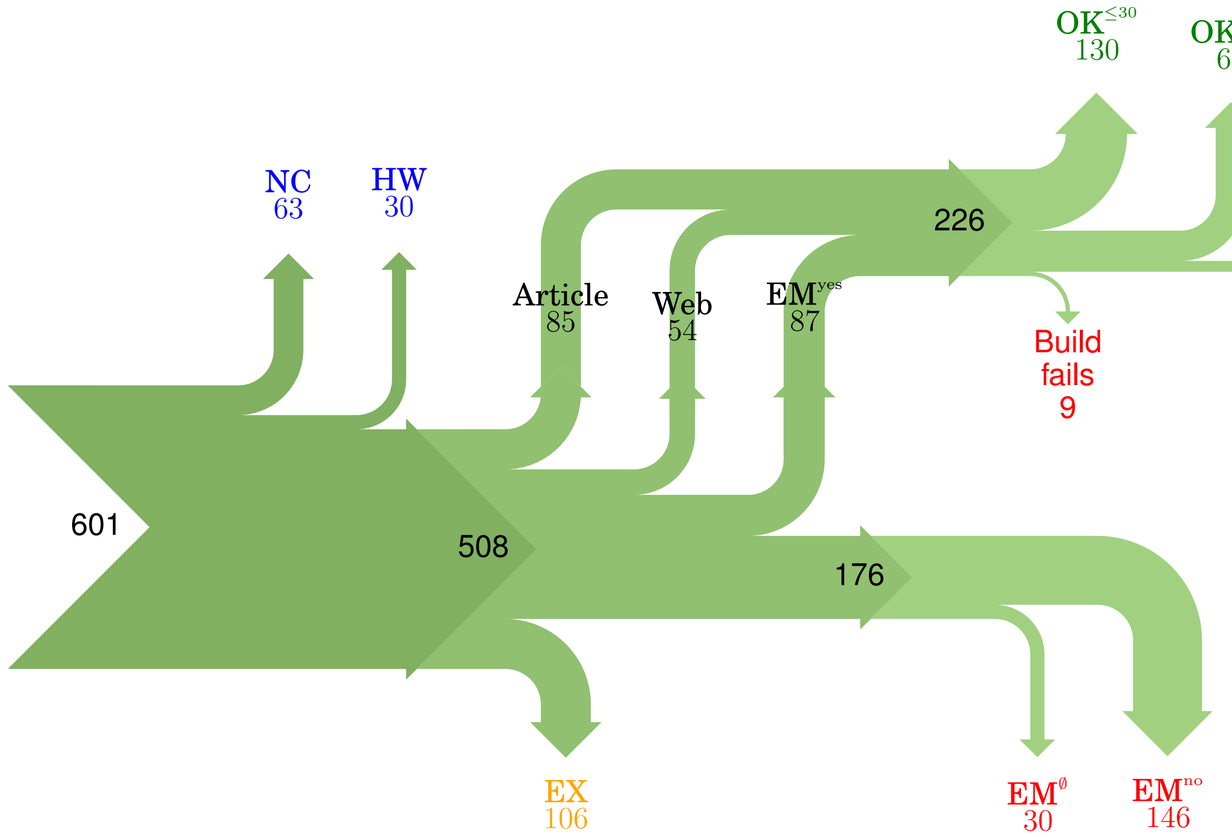
Checking Computational Results in Systems





[Collberg and Proebsting, 2015]

Repeatability Results



Notation	Number of papers ...
HW	excluded due to replication requiring special hardware
NC	excluded due to results not being backed by code
EX	excluded due to overlapping author lists
BC	where the results are backed by code
Article	where code was found in the paper itself
Web	where code was found through a Web search
EM^{yes}	where the author provides code after receiving an email message
EM^{no}	where the author responds to an email message saying code cannot be provided
$EM^{\emptyset}$	where the author does not respond to email requests within two months
$OK^{\leq 30}$	where code is available and we succeed in building the system in ≤30 minutes
$OK^{>30}$	where code is available and we succeed in building the system in >30 minutes
OK^{Auth}	where code is available and we fail to build, and the author says the code builds with reasonable effort
Fails	where code is available and we fail to build, and the author says the code may have problems building

Figure 11: Study result. Blue numbers represent papers that were excluded from consideration, green numbers papers that are weakly repeatable, red numbers papers that are non-weakly repeatable, and orange numbers represent papers that were excluded (due to our restriction of sending at most one email to each author).

[Collberg and Proebsting, 2015]

Excuses for not sharing

- Versioning
- Available Soon
- No Intention to Share
- Personnel Issues
- Lost Code
- Academic Tradeoffs
- Industrial Lab Tradeoffs
- Obsolete HW/SW
- Controlled Usage
- Privacy/Security
- Design Issues

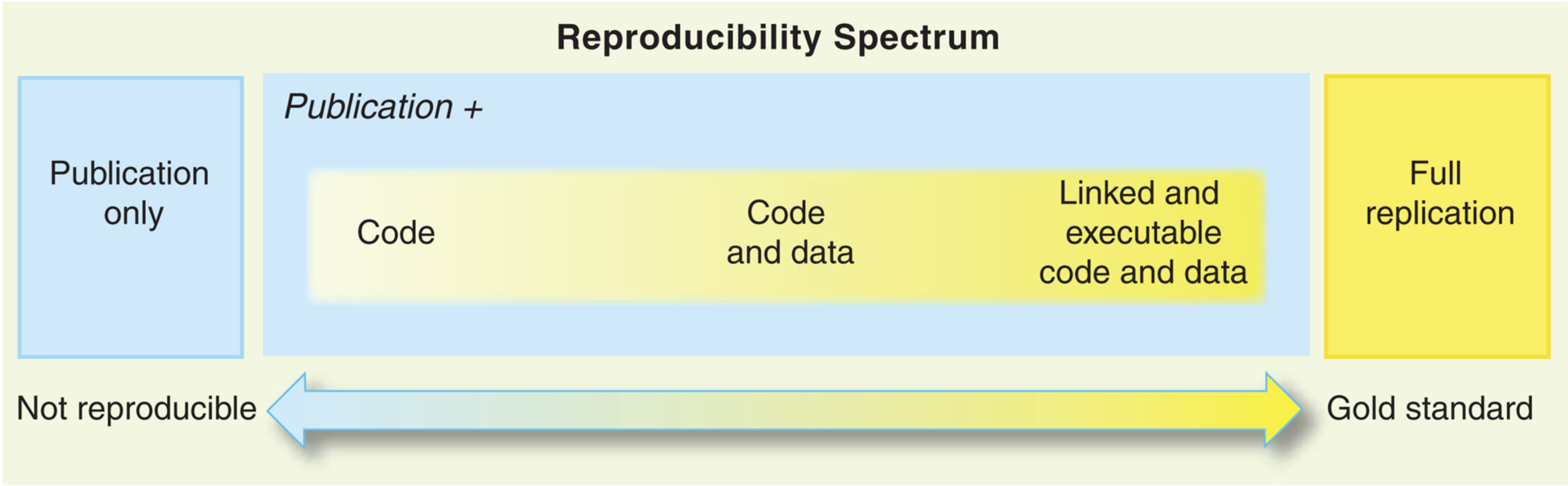
[Collberg and Proebsting, 2015]

Reproducible Research

- Science is verified by replicating work independently
- Replication Issues:
  - Requires many resources to replicate (Sloan Digital Sky Survey)
  - Requires significant computing power (Climate Model Simulation)
  - Requires too much time or very specific circumstances (Environment Epidemiology)
- Reproducibility
  - Replication of the analysis based on the collected data (not replicating the data collection itself)
  - Better if we have the actual code or available executables

[R. D. Peng]

Reproducibility Spectrum



[R. D. Peng]

10 Rules for Reproducible Computational Research

- Rule 1: For Every Result, Keep Track of How It Was Produced
- Rule 2: Avoid Manual Data Manipulation Steps
- Rule 3: Archive the Exact Versions of All External Programs Used
- Rule 4: Version Control All Custom Scripts
- Rule 5: Record All Intermediate Results, When Possible in Standardized Formats

[Sandve et al., 2013]

10 Rules for Reproducible Computational Research

- Rule 6: For Analyses That Include Randomness, Note Underlying Random Seeds
- Rule 7: Always Store Raw Data behind Plots
- Rule 8: Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
- Rule 9: Connect Textual Statements to Underlying Results
- Rule 10: Provide Public Access to Scripts, Runs, and Results

[Sandve et al., 2013]

(Database) Reproducibility Research Topics

- Design and Management of Experiment Repositories
- Querying and Searching Experiments
- Mining Experiments

[J. Freire et al.]

Notebook Reproducibility

- Use notebooks from Github (~1 million)
  - Unambiguous cell order? 81.99%
- Study notebook dependencies
  - Dependencies Available? 13.72%
  - Dependencies Install? 5.03%
- Study notebook executability
  - Execute: 24.11% of unambiguous cell order
  - Matched results: 4.03%

[Pimentel et al., 2019]

Dataflow Notebooks

In [a0a358]:
```
raw_df = pd.read_csv("fifa17-top20-women.txt",sep="-",header=None)
```

raw_df:

	0	1	2
0	Caroline Seger	Sweden	85
1	Wendie Renard	France	85
2	Steph Houghton	England	85
...	...	...	...

In [aaa3c6]:
```
column_names = {0: "Name", 1: "Country", 2: "Rating"}
```

column_names:
```
{0: 'Name', 1: 'Country', 2: 'Rating'}
```

In [a249ea]:
```
named_df = raw_df.rename(columns=column_names)
```

named_df:

	Name	Country	Rating
0	Caroline Seger	Sweden	85
1	Wendie Renard	France	85
2	Steph Houghton	England	85
...	...	...	...

In [aab079]:
```
named_df.groupby("Country").size().sort_values(ascending=False)
```

Out[aab079]:
```
Country
USA          6
Canada       3
Brazil       3
...        ...
```

[D. Koop et al.]

Dataflow Notebooks

- Persistent Identifiers

```
In [a0a358]: raw_df = pd.read_csv("fifa17-top20-women.txt",sep="-",header=None)
```

raw_df:

	0	1	2
0	Caroline Seger	Sweden	85
1	Wendie Renard	France	85
2	Steph Houghton	England	85
...	...	...	...

```
In [aaa3c6]: column_names = {0: "Name", 1: "Country", 2: "Rating"}
```

```
column_names: {0: 'Name', 1: 'Country', 2: 'Rating'}
```

```
In [a249ea]: named_df = raw_df.rename(columns=column_names)
```

named_df:

	Name	Country	Rating
0	Caroline Seger	Sweden	85
1	Wendie Renard	France	85
2	Steph Houghton	England	85
...	...	...	...

```
In [aab079]: named_df.groupby("Country").size().sort_values(ascending=False)
```

```
Out[aab079]: Country
             USA        6
             Canada     3
             Brazil     3
             ...      ...
```

[D. Koop et al.]

Dataflow Notebooks

- Persistent Identifiers
- Named Outputs

```
In [a0a358]: raw_df = pd.read_csv("fifa17-top20-women.txt",sep="-",header=None)
```

raw_df:

	0	1	2
0	Caroline Seger	Sweden	85
1	Wendie Renard	France	85
2	Steph Houghton	England	85
...	...	...	...

```
In [aaa3c6]: column_names = {0: "Name", 1: "Country", 2: "Rating"}
```

column_names:

```
{0: 'Name', 1: 'Country', 2: 'Rating'}
```

```
In [a249ea]: named_df = raw_df.rename(columns=column_names)
```

named_df:

	Name	Country	Rating
0	Caroline Seger	Sweden	85
1	Wendie Renard	France	85
2	Steph Houghton	England	85
...	...	...	...

```
In [aab079]: named_df.groupby("Country").size().sort_values(ascending=False)
```

```
Out[aab079]: Country
             USA          6
             Canada       3
             Brazil       3
             ...        ...
```

[D. Koop et al.]

Dataflow Notebooks

- Persistent Identifiers
- Named Outputs
- Unnamed Outputs

```
In [a0a358]: raw_df = pd.read_csv("fifa17-top20-women.txt",sep="-",header=None)
```

raw_df:

	0	1	2
0	Caroline Seger	Sweden	85
1	Wendie Renard	France	85
2	Steph Houghton	England	85
...	...	...	...

```
In [aaa3c6]: column_names = {0: "Name", 1: "Country", 2: "Rating"}
```

```
column_names: {0: 'Name', 1: 'Country', 2: 'Rating'}
```

```
In [a249ea]: named_df = raw_df.rename(columns=column_names)
```

named_df:

	Name	Country	Rating
0	Caroline Seger	Sweden	85
1	Wendie Renard	France	85
2	Steph Houghton	England	85
...	...	...	...

```
In [aab079]: named_df.groupby("Country").size().sort_values(ascending=False)
```

```
Out[aab079]: Country
             USA        6
             Canada     3
             Brazil     3
             ...      ...
```

[D. Koop et al.]

Dataflow Notebooks

- Persistent Identifiers
- Named Outputs
- Unnamed Outputs
- Connection by Variable Reference

```
In [a0a358]: raw_df = pd.read_csv("fifa17-top20-women.txt",sep="-",header=None)
```

raw_df:

	0	1	2
0	Caroline Seger	Sweden	85
1	Wendie Renard	France	85
2	Steph Houghton	England	85
...	...	...	...

```
In [aaa3c6]: column_names = {0: "Name", 1: "Country", 2: "Rating"}
```

```
column_names: {0: 'Name', 1: 'Country', 2: 'Rating'}
```

```
In [a249ea]: named_df = raw_df.rename(columns=column_names)
```

named_df:

	Name	Country	Rating
0	Caroline Seger	Sweden	85
1	Wendie Renard	France	85
2	Steph Houghton	England	85
...	...	...	...

```
In [aab079]: named_df.groupby("Country").size().sort_values(ascending=False)
```

```
Out[aab079]: Country
             USA        6
             Canada     3
             Brazil     3
             ...      ...
```

[D. Koop et al.]

Improving Databases

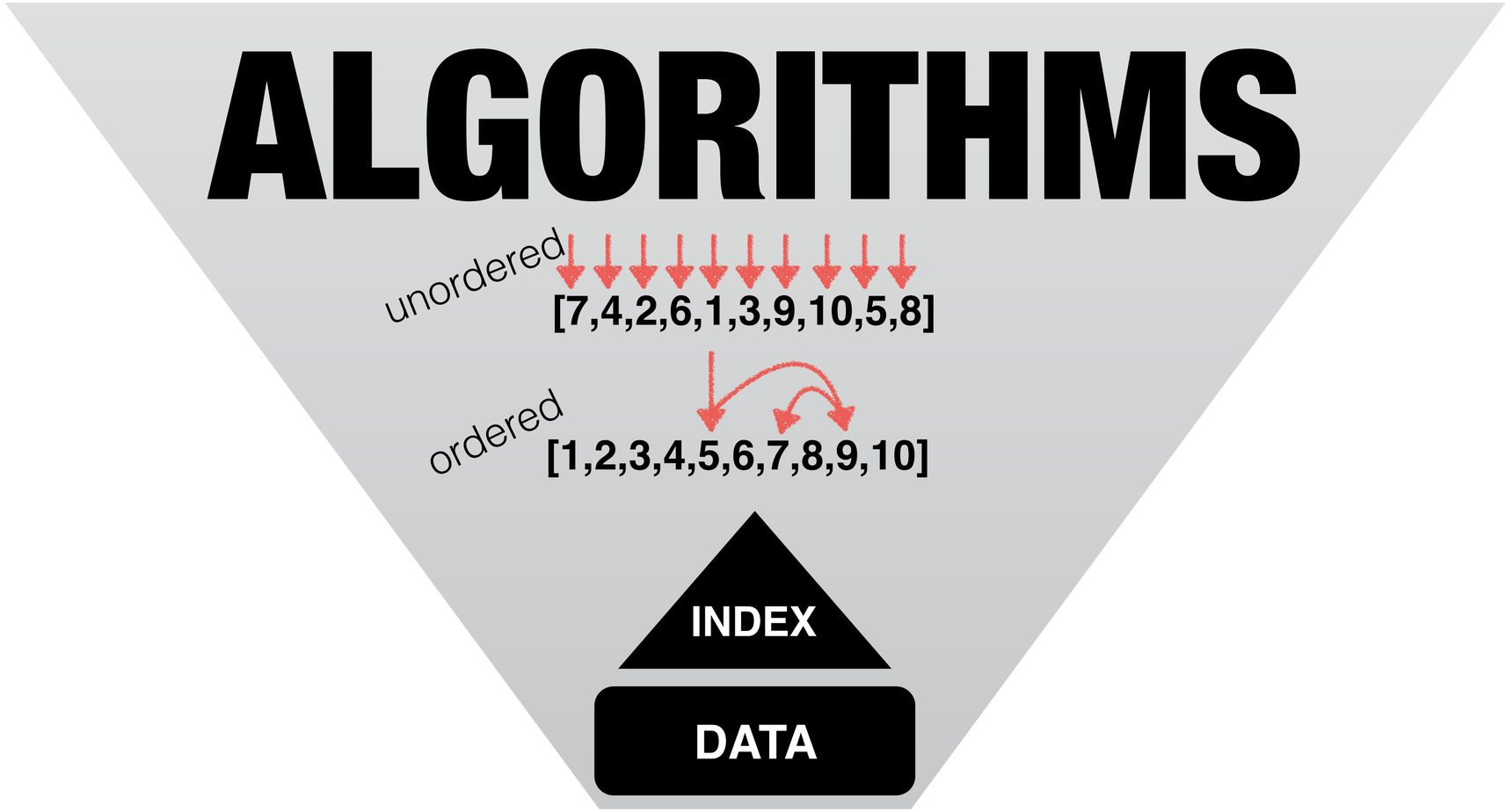
LEARNED AND **SELF-DESIGNING** DATA STRUCTURES

DASlab
@ Harvard SEAS

MIT DSAIL
Data Systems and AI Lab

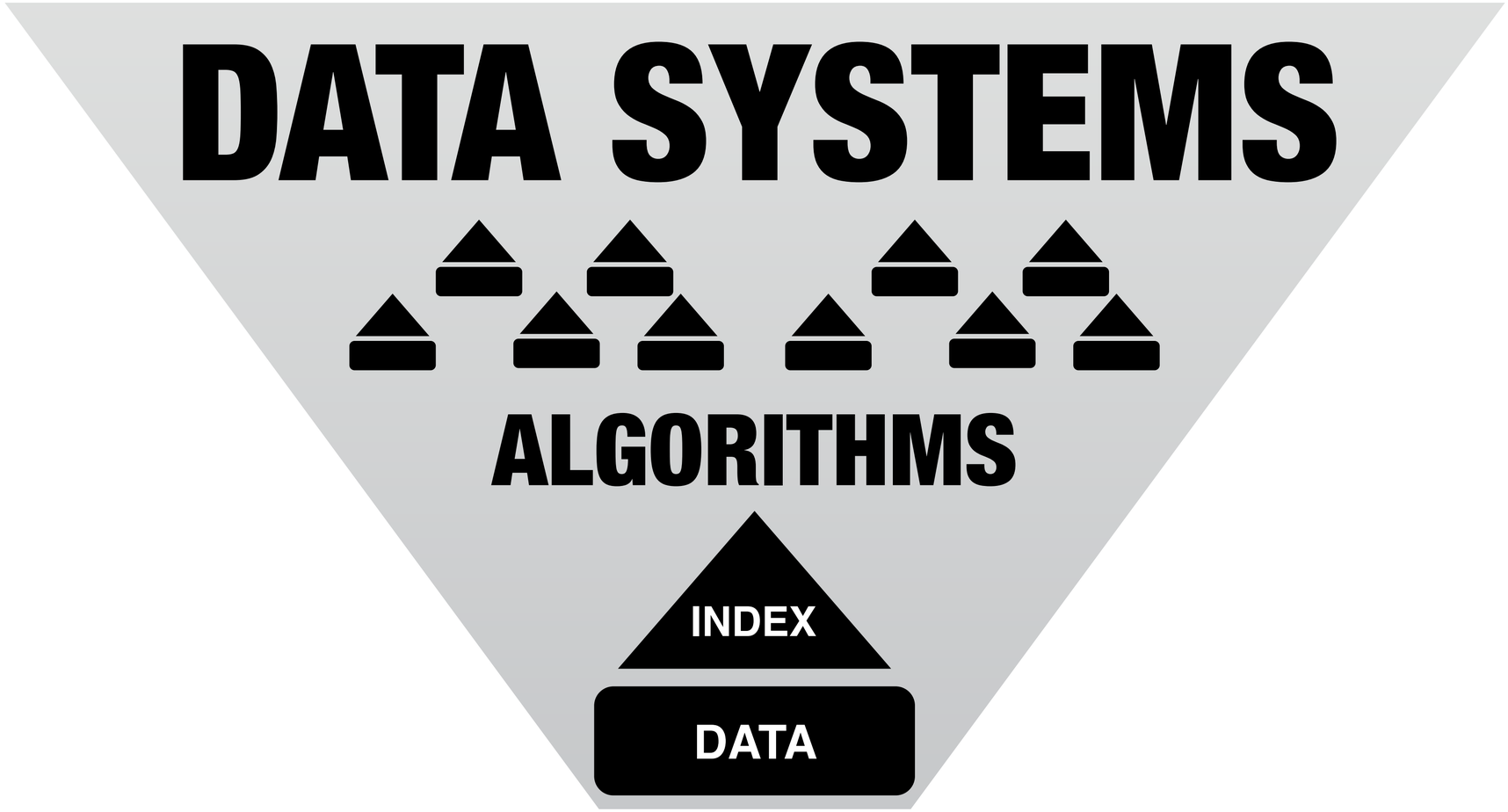
Stratos Idreos & Tim Kraska

Algorithms rely on the order of data



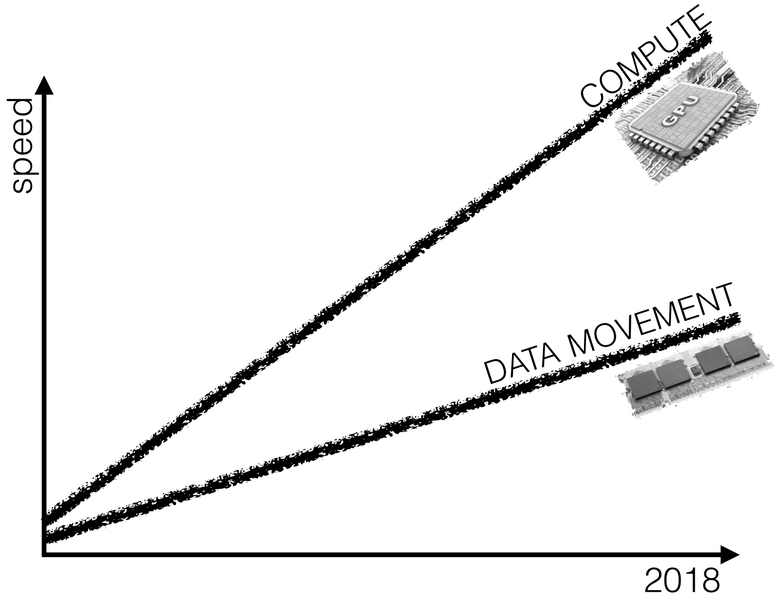
[S. Idreos, 2019]

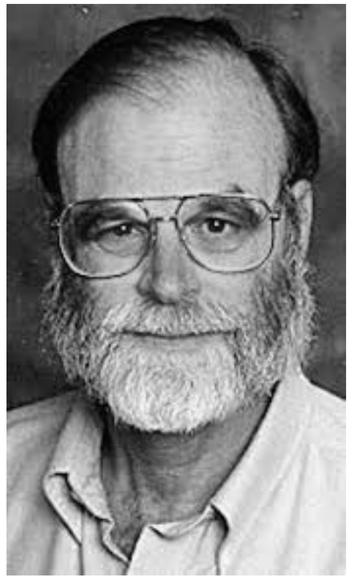
Data systems rely on algorithms



[S. Idreos, 2019]

Data structures define performance





register = this room

caches = this city

memory = nearby city

disk = Pluto

Jim Gray, Turing Award 1998

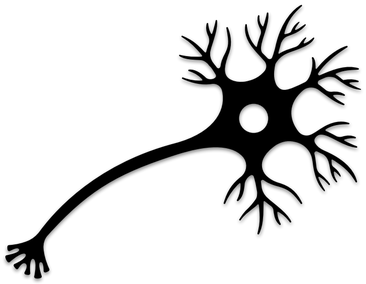
[S. Idreos, 2019]

How do I make my **data system** run x times as fast?  (sql,nosql,bigdata, …)

How do I minimize my **bill** in the **cloud**?

How do I extend the **lifetime** of my hardware? 

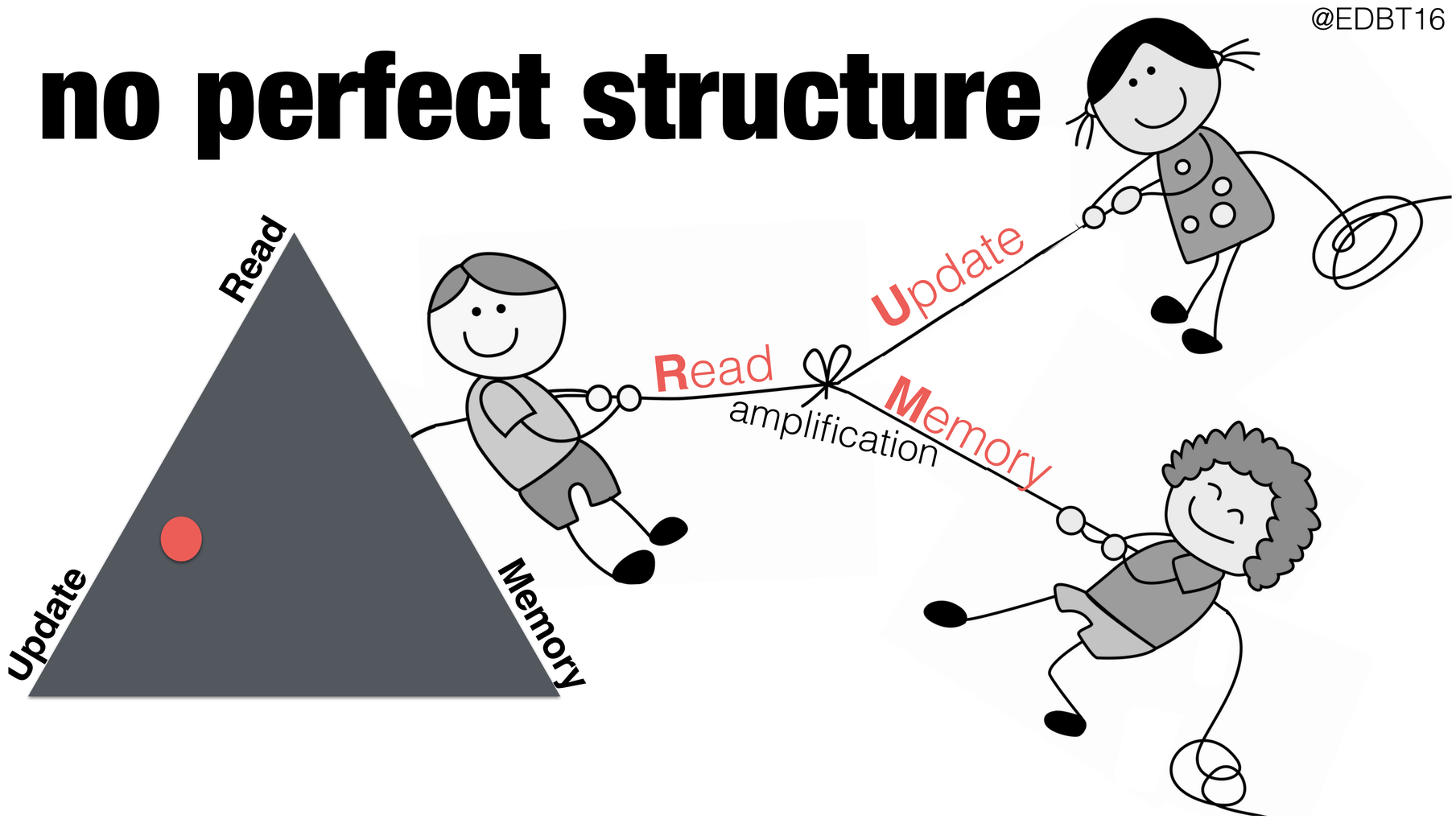
How to accelerate **statistics** computation for data science/ML? 



How do I train my **neural network** x times faster?

[S. Idreos, 2019]

Tradeoffs in each structure

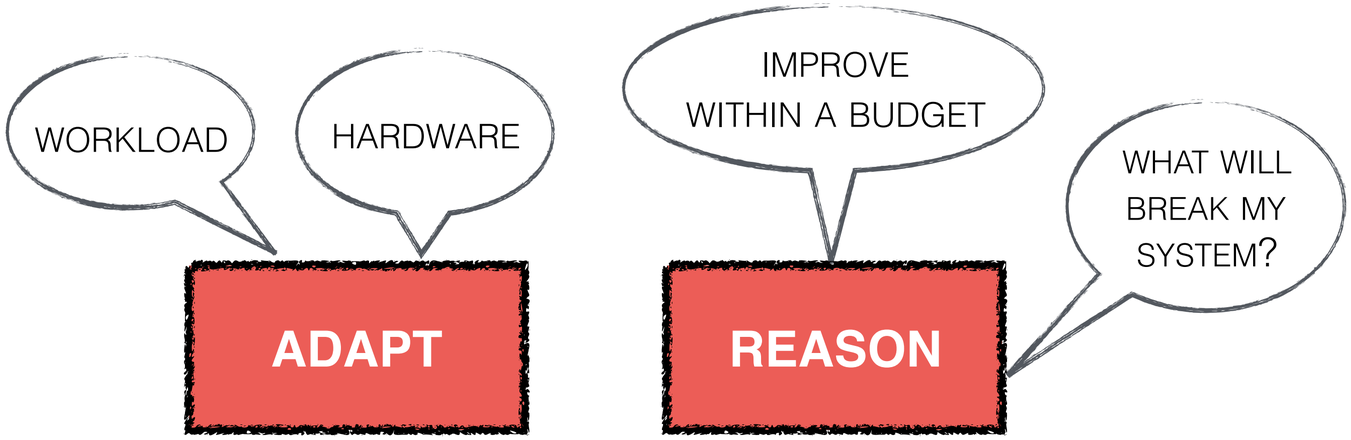


[S. Idreos, 2019]

New Applications Demand Change

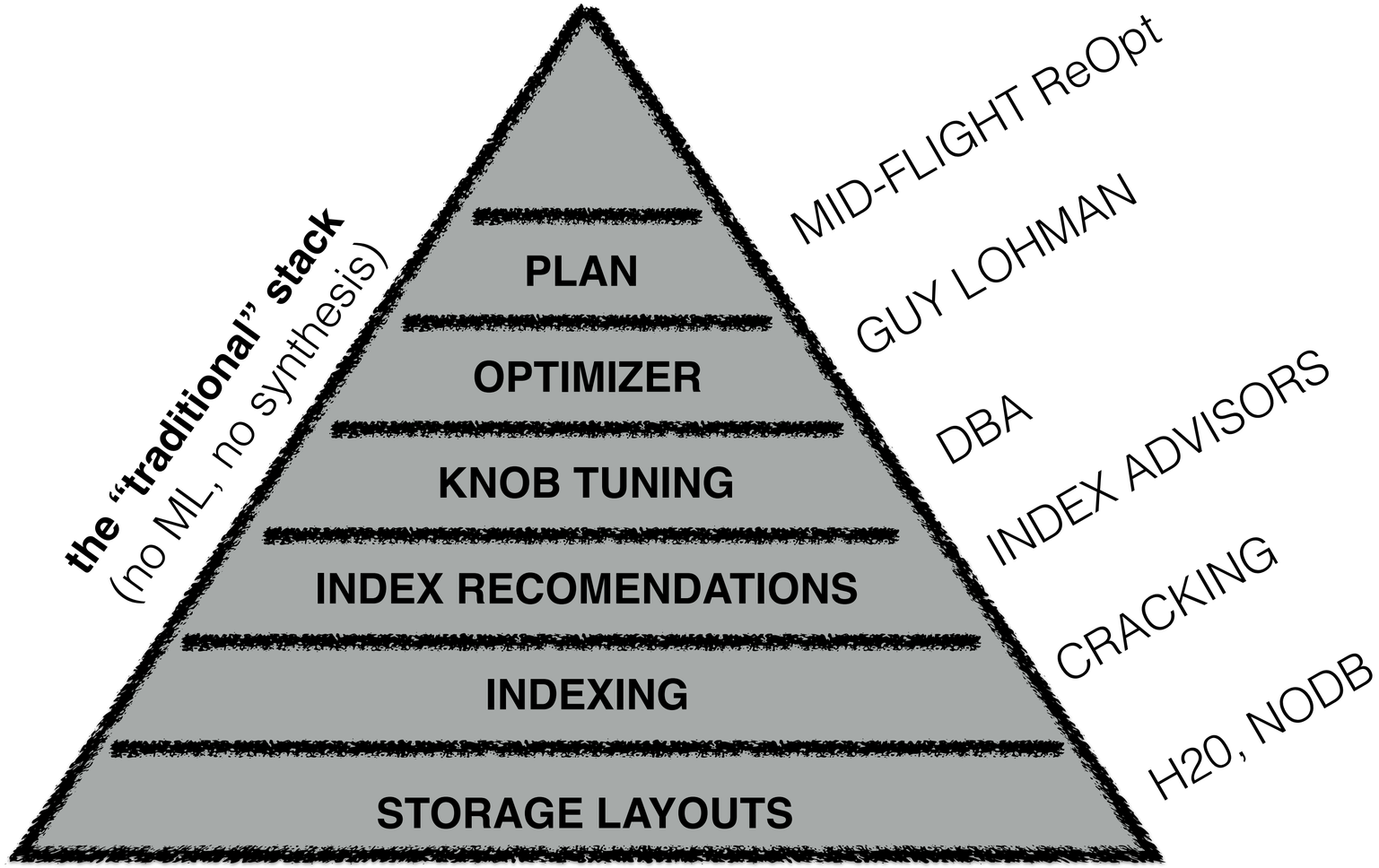
NEW APPLICATIONS

existing systems need to change too



[S. Idreos, 2019]

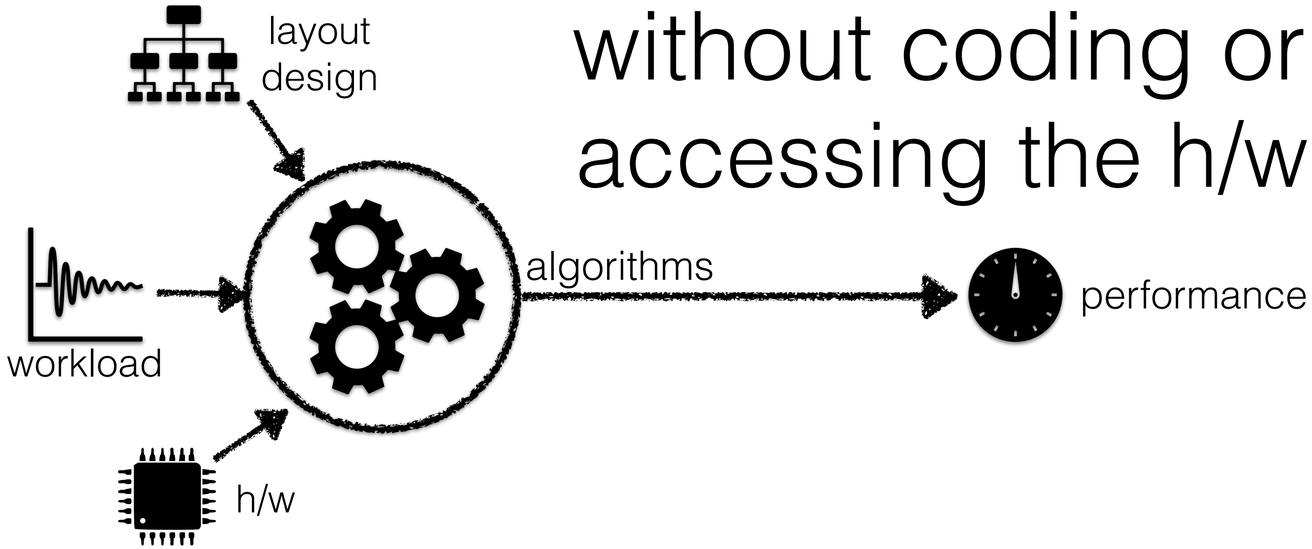
"Traditional" Database Research



[S. Idreos, 2019]

Self-designing systems

Data Calculator



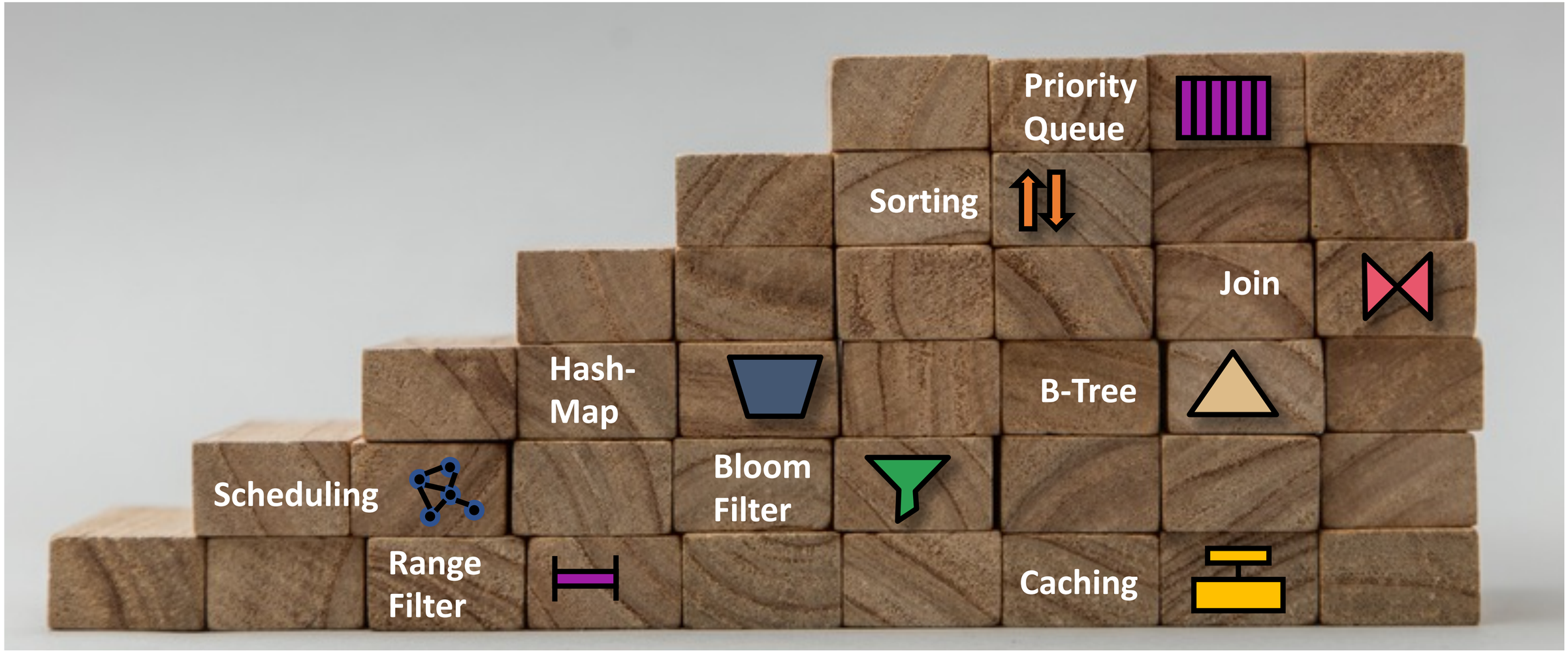
DASlab @ Harvard SEAS

[S. Idreos, 2019]

SageDB: a learned database system

T. Kraska, M. Alizadeh, A. Beutel, E. H. Chi, J. Ding, A. Kristo, G. Leclerc, S. Madden, H. Mao, and V. Nathan

Learned Data Structures and Algorithms



Discussion

- Is this the future?
- What about comparison baselines?
- Lots of work being done in this area

Reminders

- Assignment 5 Due Thursday
- Final Exam Review Thursday (send questions!)
- Final Exam on Tuesday, May 5 from 4-5:50pm