# The future of data(base) education: Is the "cow book" dead?

Zachary G. Ives
University of Pennsylvania
zives@cis.upenn.edu

Rachel Pottinger
University of British Columbia
rap@cs.ubc.ca

Arun Kumar
University of California, San Diego
arunkk@eng.ucsd.edu

Johannes Gehrke
Microsoft
johannes@microsoft.com

Jana Giceva
Technical University of Munich
jana.giceva@in.tum.de

## ABSTRACT

This panel encourages a debate over the future of database education and its relationship to Data Science: Are Computer Science (CS) and Data Science (DS) different disciplines about to split, and how does that effect how we teach our field? Is there a "data" course that belongs in CS that all of our students should take? Who is the traditional database course, e.g. based on the "cow book", relevant to? What traditional topics should we *not* be teaching in our core data course(s) and which ones should be added? What do we teach the student who has one elective for data science? How does our community position itself for leadership in CS given the popularity of DS?

**PVLDB Reference Format:**
Zachary G. Ives, Rachel Pottinger, Arun Kumar, Johannes Gehrke, and Jana Giceva. The future of data(base) education: Is the "cow book" dead?. PVLDB, 14(12): 3239 - 3240, 2021.
doi:10.14778/3476311.3476394

## 1 PANEL OVERVIEW

The academic database field is at an inflection point. With the advent of data science as its own discipline distinct from computer science, as well as the incredible burst of enthusiasm around machine learning (ML), artificial intelligence (AI), and big data — many prospective students have a perception of our field as an increasingly niche area. A key question is whether the academics among the VLDB community need to adapt, in order to better teach a new generation of students for their future careers, but also to attract top graduate students into the field.

The panelists and audience will debate how to approach database education, in light of the "fork" between computer science and data science, and argue why or whether there is a central "data course" or database course that every student should take. Among other questions they will consider is how the first course in databases, as well as the database *systems* course, needs to be refocused in the data science era; whether there is a *different* course for data scientists; and finally, what key ideas our field should be known for by every student.

## 2 PANELISTS

**Rachel Pottinger**, Professor of Computer Science at the University of British Columbia (UBC).
*Research:* Rachel works on data understanding, exploration, integration, and metadata.
*Rationale:* As former Associate Department Head of the Undergraduate Program at UBC, Rachel helped lead the creation of a data science minor.

**Arun Kumar**, Assistant Professor of Computer Science and Engineering (CSE) and Halicioglu Data Science Institute (HDSI) at the University of California, San Diego (UCSD).
*Research:* Arun's main research interests are in data management and systems for ML/AI analytics, focusing on issues of usability, scalability and resource efficiency.
*Rationale:* Arun is a co-author of *Data Management in ML Systems* [1], the first research-based textbook on systems for ML. He created and teaches courses on database system implementation, as well as systems for analytics and ML, aimed at CS and DS students. He helped shape HDSI's data systems course series, CSE's MS specialization in DS, and a new joint CSE-HDSI online MS in DS.

**Johannes Gehrke**, Technical Fellow, Microsoft.
*Research:* Johannes works on database systems, distributed systems, and machine learning.
*Rationale:* Johannes is a co-author of the widely used textbook, *Database Management Systems* [4] (a.k.a "the cow book"). He has experience from academia (Cornell ), startup (FAST Search and Transfer), large-company product group (Office 365 and Teams), and research lab (MSR); he has hired many data scientists and engineers into his teams.

**Jana Giceva**, Assistant Professor of Computer Science at the Technical University of Munich (TUM) and core member of the Munich Data Science Institute (MDSI).
*Research:*. Jana works on designing cloud-native data systems for modern hardware, paying special interest on the interactions with computer systems, HW/SW co-design and compilers.
*Rationale:*. Jana has taught courses on data management, computer architecture and cloud computing. She also led the computer system's curriculum (re-)design while at Imperial College London.

## 3 PANEL CHAIR

**Zachary Ives** is Professor and chair of Computer and Information Science at the University of Pennsylvania and a co-founder of

Proceedings of the VLDB Endowment, Vol. 14, No. 12 ISSN 2150-8097.
doi:10.14778/3476311.3476394

Blackfynn, Inc. Zack's research interests include data integration, distributed query processing, data provenance and authoritativeness, and applied machine learning. He has received an ICDE 2013 Ten-Year Most Influential Paper Award as well as the 2017 SWSA Ten-Year Award at the International Semantic Web Conference. He has been an Associate Editor for the Proceedings of the VLDB Endowment, a Program Co-Chair for the ACM SIGMOD conference, and will be the General Chair for SIGMOD 2022.

Zack is a co-author of the textbook *Principles of Data Integration* [2], a winner of the Lindback Award for Distinguished Teaching, and teaches several courses in distributed systems, databases, and data science. He helped define the curriculum for Penn's Networked and Social Systems Engineering program, one of the first undergraduate degrees at the intersection of computer science, network science, and data science, as well as the Penn's MSE in Data Science. He also co-developed the curriculum of the OpenDS4All project [3], a set of open-source materials for teaching data science.

## REFERENCES

[1] Matthias Boehm, Arun Kumar, and Jun Yang. 2019. *Data Management in Machine Learning Systems.* Morgan & Claypool, USA.

[2] AnHai Doan, Alon Halevy, and Zachary G. Ives. 2012. *Principles of Data Integration.* Morgan Kauffmann.

[3] The Linux Foundation. 2019. OpenDS4All. https://github.com/odpi/OpenDS4All

[4] Raghu Ramakrishnan and Johannes Gehrke. 2000. *Database Management Systems* (2nd ed.). McGraw-Hill, Inc., USA.