# CSC 343
# Introduction to Databases

csc343, Introduction to Databases

UNIVERSITY OF
TORONTO

# Our first hour or so

- Some key concepts
- Examples to motivate the course
- Admin info

# Databases and DBMSs

- Databases are everywhere, often behind the scenes.

- DBMS (Database Management System):
  "A powerful tool for creating and managing large amounts of data efficiently and allowing it to persist over long periods of time, safely." [Ullman and Widom, FCDB]

- Database: a collection of data managed by a DBMS.

# Data models

Every DBMS is based on some data model: a notation for describing data, including

- the structure of the data
- constraints on the content of the data
- operations on the data

Some specific data models:

- relational data model
- semistructured data model
- unstructured data — (key, value) pairs
  - value could be anything: a full document, a video, etc.
- graph data model

# The Relational Data Model

- Main concept is a “relation.”
  - Based on the concept of relations in math.
  - Can think of as tables of rows and columns.

Teams

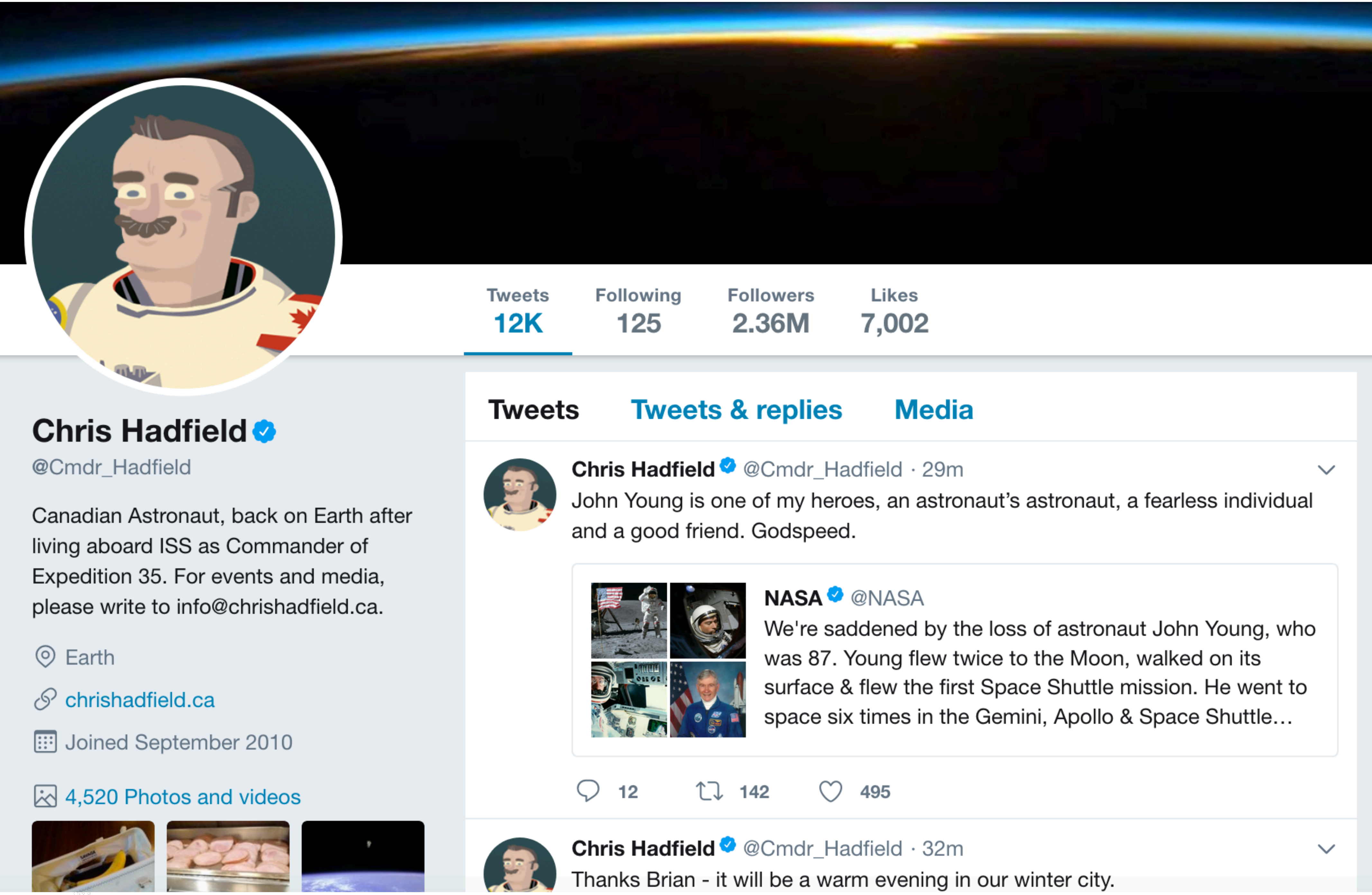
| Name | Home Field | Coach |
|---|---|---|
| Rangers | Runnymede CI | Tarvo Sinervo |
| Ducks | Humber Public | Tracy Zheng |
| Choppers | High Park | Ammar Jalali |

Games

| Home team | Away team | Home goals | Away goals |
|---|---|---|---|
| Rangers | Ducks | 3 | 0 |
| Ducks | Choppers | 1 | 1 |
| Rangers | Choppers | 4 | 2 |
| Choppers | Ducks | 0 | 5 |

# Example …

- A dataset scraped from Twitter
- Defining a schema that expresses its structure
- Creating an instance that contains the data
- Writing some queries on the data



Tweets 12K | Following 125 | Followers 2.36M | Likes 7,002

**Chris Hadfield**

@Cmdr_Hadfield

Canadian Astronaut, back on Earth after living aboard ISS as Commander of Expedition 35. For events and media, please write to info@chrishadfield.ca.

Earth

chrishadfield.ca

Joined September 2010

4,520 Photos and videos

Tweets | Tweets & replies | Media

**Chris Hadfield** @Cmdr_Hadfield · 29m

John Young is one of my heroes, an astronaut's astronaut, a fearless individual and a good friend. Godspeed.

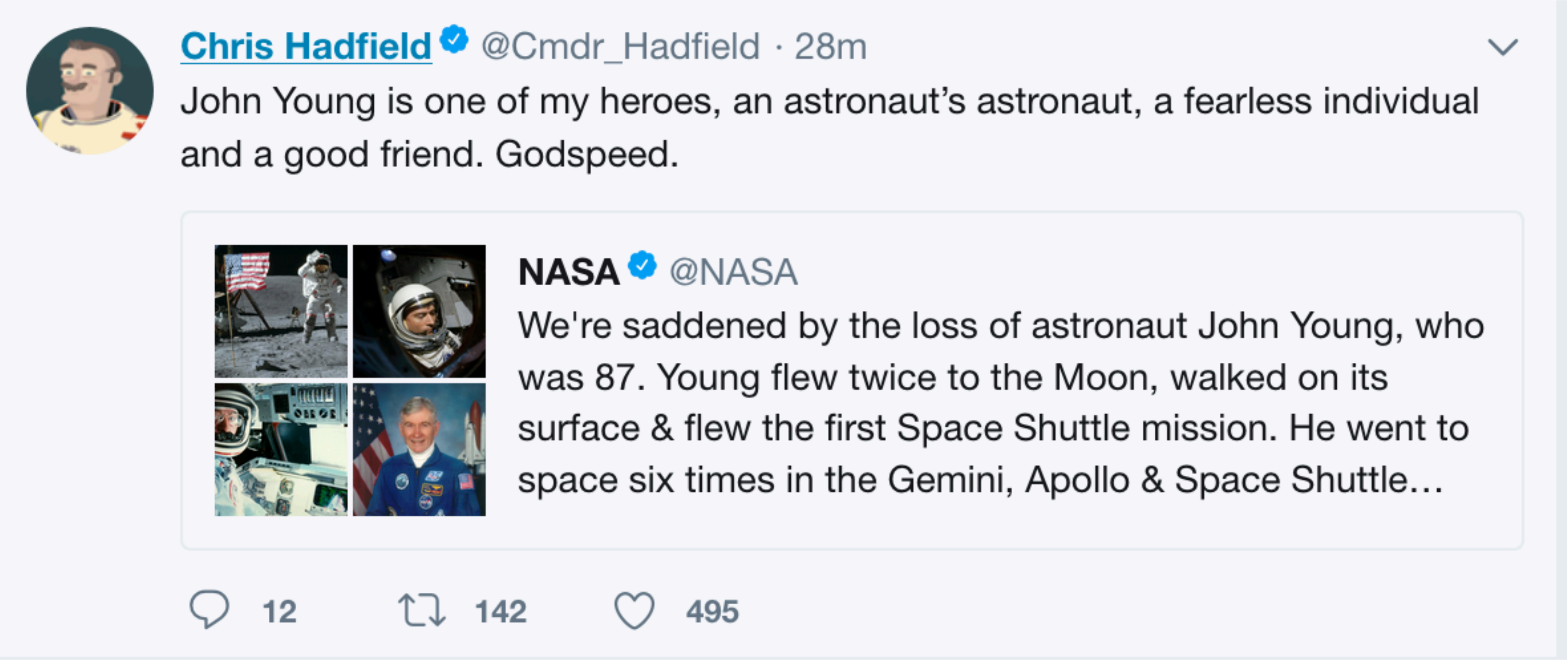
> **NASA** @NASA
>
> We're saddened by the loss of astronaut John Young, who was 87. Young flew twice to the Moon, walked on its surface & flew the first Space Shuttle mission. He went to space six times in the Gemini, Apollo & Space Shuttle…

12 · 142 · 495

**Chris Hadfield** @Cmdr_Hadfield · 32m

Thanks Brian - it will be a warm evening in our winter city.

https://twitter.com/cmdr_hadfield?lang=en



Chris Hadfield @Cmdr_Hadfield · 28m

John Young is one of my heroes, an astronaut's astronaut, a fearless individual and a good friend. Godspeed.

NASA @NASA

We're saddened by the loss of astronaut John Young, who was 87. Young flew twice to the Moon, walked on its surface & flew the first Space Shuttle mission. He went to space six times in the Gemini, Apollo & Space Shuttle…

12 142 495

## How do you describe:

- a tweet
- sender of tweet
- who sender follows
- retweets

# Example Partial Twitter Schema

```
CREATE TABLE Profile (
        ID VARCHAR(50),
        name VARCHAR(50),
        location VARCHAR(50),
        url VARCHAR(150),
        bio VARCHAR(500),
        PRIMARY KEY (ID)
);

CREATE TABLE Follows (
        a VARCHAR(50),
        b VARCHAR(50),
        PRIMARY KEY(a, b),
        FOREIGN KEY (a) REFERENCES Profile(ID)
);

CREATE TABLE Tweets (
       ID INTEGER,
       userid VARCHAR(50),
       content VARCHAR(140),
       PRIMARY KEY (ID)
);
```

# What a DBMS provides

- Ability to specify the logical structure of the data
  - explicitly
  - and have it enforced

- Ability to query or modify the data.

- Good performance under heavy loads (huge data, many queries).

- Durability of the data.

- Concurrent access by multiple users/processes.

# Overall architecture of a DBMS

- The DBMS sits between the data and the users or between the data and an application program

- Within the DBMS are layers of software for:
  - parsing “queries”
  - implementing the fundamental operations
  - optimizing queries
  - maintaining indices on the data
  - accessing the files that store the data and indices
  - management of buffers
  - management of disk space

# How  to find all Vic students with over 80 in csc207?

Students:

| Student # | Name | College |
|---|---|---|
| 1234 | Fred Flintstone | Vic |
| 2345 | Wilman Flintstone | UC |
| 3456 | Betty Rubble | Vic |
| etc. | etc. | etc. |

Grades:

| Student # | Course | Grade |
|---|---|---|
| 9876 | aps105 | 78 |
| 3456 | csc207 | 85 |
| 2345 | csc207 | 92 |
| etc. | etc. | etc. |

# What this course is about

- csc443 is about implementation of the DBMS itself

- csc343 is about *using* DBMSs:
  - defining schemas and instances
  - writing queries
  - connecting to code written in a general-purpose language
  - rigorous underlying principles

# Why study databases?

- Interesting concepts and techniques.

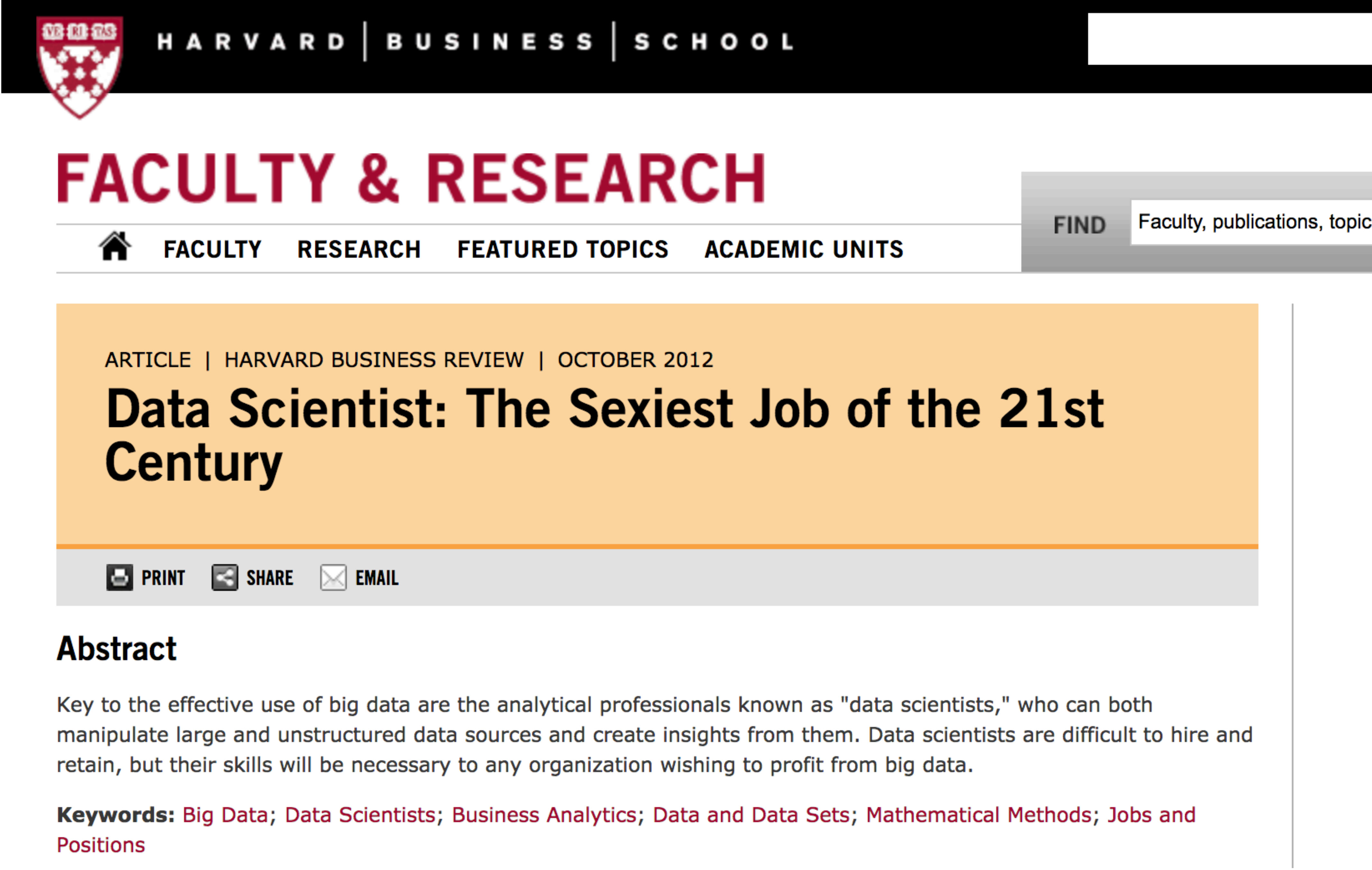
- Spans computer science, including OS, languages, theory,  AI, multimedia, logic.

- Databases have become increasingly important
  - shift from a focus on computation to information
  - data increases in volume and diversity.

- Jobs: In demand and well paid.

- Research: Many open problems.

- Foundation of Data Science



ARTICLE | HARVARD BUSINESS REVIEW | OCTOBER 2012

# Data Scientist: The Sexiest Job of the 21st Century

## Abstract

Key to the effective use of big data are the analytical professionals known as "data scientists," who can both manipulate large and unstructured data sources and create insights from them. Data scientists are difficult to hire and retain, but their skills will be necessary to any organization wishing to profit from big data.

**Keywords:** Big Data; Data Scientists; Business Analytics; Data and Data Sets; Mathematical Methods; Jobs and Positions