Data: The Big Picture

Database and SQL: Course Notes | Brendan Shea, Ph.D. (Brendan.Shea@rctc.edu)

This chapter will look at some "big picture" issues involving data and databases. First, we'll think about the role that data plays in an organization's larger goals and how we can design databases to best meet this goal. Then, we'll briefly discuss the sorts of (unique) ethical issues that confront data ethics.

1 Contents

2 The Life of a Database

In this lesson, we'll take a high-level (or "bird's eye point of view") look at how database design, implementation, and maintenance work.

2.1 Information Systems

Databases are fundamentally embedded within larger **information systems,** and database design must consider this. Information systems consist of (a) people, (b) hardware, (c) application programs (software), (d) databases, and (e) the policies, procedures, and practices that enable these components to work together. The information system exists to turn raw "data" into knowledge, and database design and modeling is a central (arguably the *most important)* part of this.

2.2 SOFTWARE DEVELOPMENT LIFE CYCLE

The Software Development Life Cycle (SDLC) provides a conceptual tool for understanding how information systems are developed and maintained. The SDLC is an "abstract" model that describes what *tasks* must be completed instead of how to achieve them. There are a variety of more specific modeling (from ER modeling to "Agile" to "Rapid Application Development") that give details on how to accomplish these tasks. The SDLC consists of the following stages:

Planning involves a **needs assessment** of an organization's data needs, determining the **scope** of the proposed system, and conducting a **feasibility study** to determine constraints regarding cost/personnel/equipment/etc.. By the end of this process, you should have a general idea of what users *need* from the information system and what resources are available to create, modify, or replace the system.

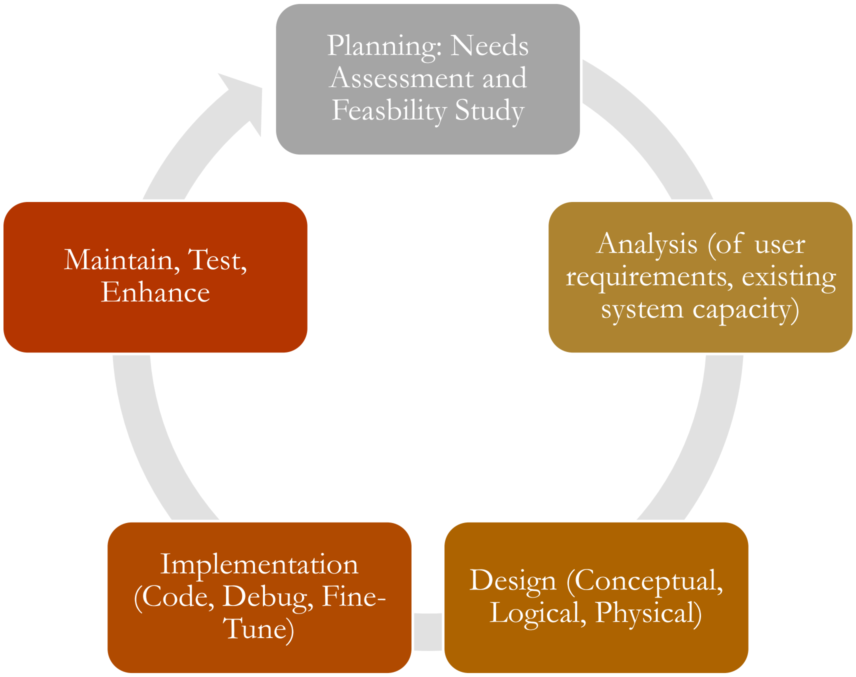
Analysis involves talking to end-users in more detail to determine their needs and creating "conceptual" maps of different users' views of data (e.g., using multiple E-R diagrams or object diagrams). This may involve additional feasibility considerations (i.e., determining how detailed of a model we can/should implement).

Detailed system **design** involves the creation of precise technical specifications (whether these are for software, new employee training, or hardware capabilities, etc.) that will eventually need to be implemented. A comprehensive conceptual model is completed, and this is mapped to a logical model (relational, object, etc.) with appropriate entities, attributes, and relationships. By the end of this, we should have a list of testable "outcome goals" against which our implementation can be measured. For databases, this is a database schema.

Implementation involves the physical "creation" of the system. This might include writing and debugging code, installing and testing hardware, or training employees. For databases, this will often involve the selection of a DBMS, and the SQL code needed to create tables, load data, and so on.

Maintenance and Testing involve a continual effort to (a) find and correct errors, (b) adapt to changes in the organization's environment (such as the availability of new technology or an increase in the business size), and (c) improve and perfect operations.

IMPORTANT: The SDLC is an "iterative process" in which discoveries we make at a given stage might require us to "go back" to the decisions made at the "previous" stage. It is also a cycle that will be repeated, as old systems must be updated/replaced to meet new business challenges or opportunities.


Planning: Needs Assessment and Feasbility Study
Analysis (of user requirements, existing system capacity)
Design (Conceptual, Logical, Physical)
Implementation (Code, Debug, Fine-Tune)
Maintain, Test, Enhance

2.3 Levels of Database Design

Database Design can itself be broken into several different levels

Conceptual Model Design begins by taking the (often vague and ambiguous) descriptions of data given by end-users, managers, application programmers, etc. and turning these into a formal model of the type an ER diagram gives. This generally involves the following:

1. The formulation of business rules, as simple, declarative statements describing the data model.
2. The creation of a graphical ER, UML, or Object model that displays the entities, attributes, and relations. We must also identify primary keys and foreign keys. Finally, we must normalize the model, which will often involve the creation of additional entities.
3. The internal and external **validation** of this model, by checking it for (a) internal consistency/coherence and (b) external "agreement" with the end-user needs that this model is meant to capture.
4. At the end, we want a model that is "as simple as possible, but no simpler." That is, we want all and only the data that our end-users need. More entities does not mean more awesome!

Logical Model Design is the process of "mapping" a conceptual model to a more specific model that will be implemented in the previous step. For example, we might choose a relational data model, a NoSQL/JSON model, or (rarely) something else. For a relational model, this means we must create

1. Create tables along with their attributes and constraints. We begin by creating a table for each strong entity, subtypes of these entities, weak entities, and relations.
2. After completing this process, again normalize the database and do internal/external validation.

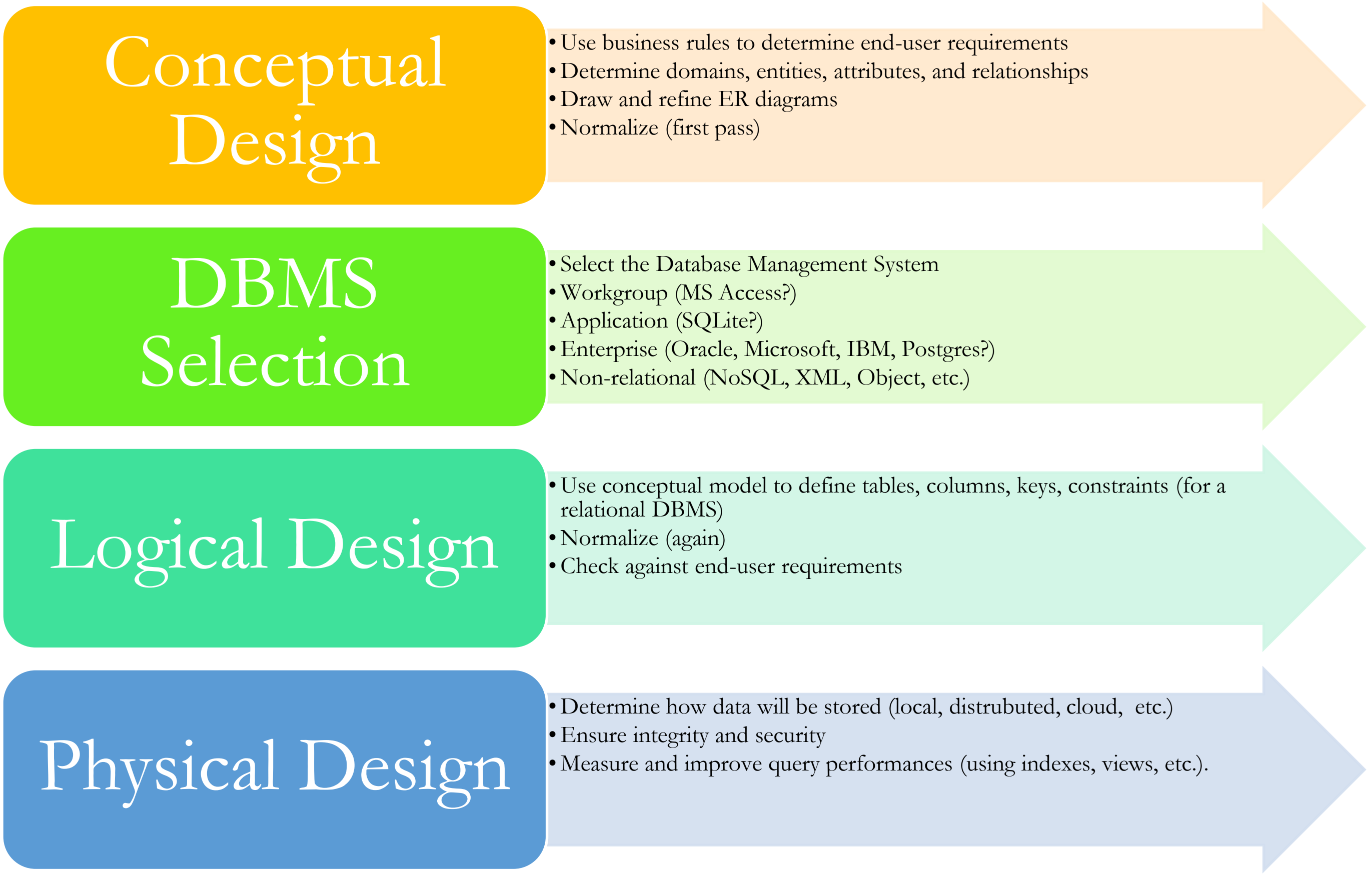
Physical Model Design takes the logical model and "maps" it to a particular physical model. We choose a particular DBMS software (Postgrs, Oracle, SQLite, etc.) Here, we write the SQL (or, MongoDB, etc.) code that will create the tables. We also develop database **views** for different groups of users, **triggers/functions** to perform database updates, choose and create **indexes** to maximize performance, and so on. The database administrator decides on **user groups** and levels of access. While we might have various goals here, two of the most important include (a) optimizing performance and (b) ensuring data security.

The physical model may involve decisions about storage and location. For example, will the database be hosted on a single mainframe, **sharded** between many computers, or on the "cloud"? These may have for the other choices we make during this design stage.

Database Testing and Evaluation occur immediately after implementation but must also be performed throughout the life of the database. These tests using include answering at least the following two questions:

1. **Is User Data Secure?** This involves answering questions such (1) do users have appropriate types of passwords? (2) do users have the correct level of **access rights** (to read or change just the data they need)? (3) are **audit trails** (records that track which users accessed the database, and what they did) in place, and (4) is appropriate **data encryption** used (that is, we should "encode" data on the disk so that only authorized users with the appropriate "key" can read/access it).
    a. Depending on the nature of the system, we may also need to ensure the physical security of the drives that store the data.
2. **Does the Database Perform Well?** Database administrators spend a lot of time thinking about *which* sorts of queries end-users are running on the database and *how* these queries perform. Query performance can be optimized by judicious creation of indexes (as well as the type of index), trigger functions, (materialized) views, database page size, how database transactions **lock** the records they are working with (which prevents other uses from changing the data), and other factors.

"Top Down" vs "Bottom Up" vs "A Bit of Both". One might pursue two general strategies in designing database models (and, really, in creating basically any scientific/mathematical model). A **top-down** approach begins by specifying the abstract types of *entities* that you want to have in your data model (for example, "Customers" and "Products") and then proceeds to make decisions about the attributes and relations of these entities. A **bottom-up** approach begins by focusing on the *attributes* (for example, "customer_name" and "customer_id") and the way they relate to one another, only then proceeding to "cluster" relevant attributes into entities. Most real-world modeling is, by necessity, going to involve a bit of each. For example, ER modeling (with its focus on "entities" and "relations") is a top-down process. In contrast, database normalization (which focuses almost exclusively on the relationships between attributes) is a bottom-up process. Both play a role in sound design.



Conceptual Design

- Use business rules to determine end-user requirements
- Determine domains, entities, attributes, and relationships
- Draw and refine ER diagrams
- Normalize (first pass)

DBMS Selection

- Select the Database Management System
- Workgroup (MS Access?)
- Application (SQLite?)
- Enterprise (Oracle, Microsoft, IBM, Postgres?)
- Non-relational (NoSQL, XML, Object, etc.)

Logical Design

- Use conceptual model to define tables, columns, keys, constraints (for a relational DBMS)
- Normalize (again)
- Check against end-user requirements

Physical Design

- Determine how data will be stored (local, distrubuted, cloud, etc.)
- Ensure integrity and security
- Measure and improve query performances (using indexes, views, etc.).

3 Data Ethics

In general, **ethics** is the study of right and wrong behavior. While most of us learn basic ethical principles during our childhood (like “don’t hit your siblings” or “it’s wrong to lie”), many professions have specific ethical rules and guidelines associated that go beyond this “fourth-grade morality.” Such varieties of **professional ethics** have historically included the following:

- **Biomedical ethics** deals with ethical issues in medicine and biological research. These include issues related to patient “autonomy” and “privacy” treatment of patients at the end of life, the guidelines for ethical research, and the fair distribution of limited health care resources.
- **Military ethics** concerns the requirements for waging a “just” war, including the decision to go to war in the first place and the conditions that individual soldiers must abide by in the fighting of this war.
- **Engineering ethics** concerns the sorts of conflicts that engineers, architects, and designers/builders can encounter when considering the impact of their inventions (for example, how should we wait “consumer safety” vs. “cost” vs. “environmental impact” vs. “could this be used as a weapon”?). I

There are three main reasons that professional ethics is essential for these professions. First, doctors, soldiers, and engineers have a great deal of *power* over the well-being of other people. So, we want to make sure this is used correctly. Second, all of these professions (unfortunately) have long histories of abuse, where doctors, soldiers, and engineers who “didn’t follow the rules” have harmed people. Finally, these professions are notoriously difficult to regulate "from the outside" since they require expert knowledge that voters/politicians/managers lack.

In recent years, there has been a concerted effort (by schools, employers, governments, and concerned groups of professionals) to formulate some principles of **data ethics, computer ethics,** and/or **ethics of artificial intelligence (AI).** This reflects the fact that, in today's society, data professionals have begun to resemble the professions just mentioned. They have lots of power, there are well-known abuses, and the technical issues make it difficult for outsiders to know what is happening, or how to respond.

3.1 FOUR PRINCIPLES OF DATA ETHICS

While data ethics is still relatively young and hasn't yet "coalesced" on any set of rules, many scholars have proposed basing it on the widely adopted **Four-Principles** approach to biomedical ethics. These principles are as follows:

The principle of **respect for autonomy** requires that we respect and support the ability of people to make their *own* decisions about how to live their life. This includes issues such as:

- Under what conditions is it OK to collect and store data on people without their **consent**?
- How can we ensure that those "affected" by the use of databases (by governments, corporations, etc.). be able to access this data in a way that makes it **understandable** to them?

The principle of **nonmaleficence** requires that we avoid (physically or psychologically) *harming* people. This includes such as issues as:

- How can we ensure that customer data is **secure** from threats posed by hackers, foreign governments, and others?
- How can we ensure that data collected for one (morally OK) purpose isn't "reused" for some other, more nefarious purpose?

The principle of **beneficence** requires that we give aid to people. It requires that we think about "big-picture" issues of cost-benefit and the **consequentialist** (or **utilitarian)** approach to ethics (how can we make sure that we are "making the world a better place, overall?").

- How can we (as both societies and individuals) ensure that "Big Data" leads to a healthier, happier society? (For example, has social media left us better off? Worse off? A bit of each?)
- To what extent do highly compensated data professionals have an obligation to "give back"? (For example, by contributing to open-source projects, serving as tutors, etc.).

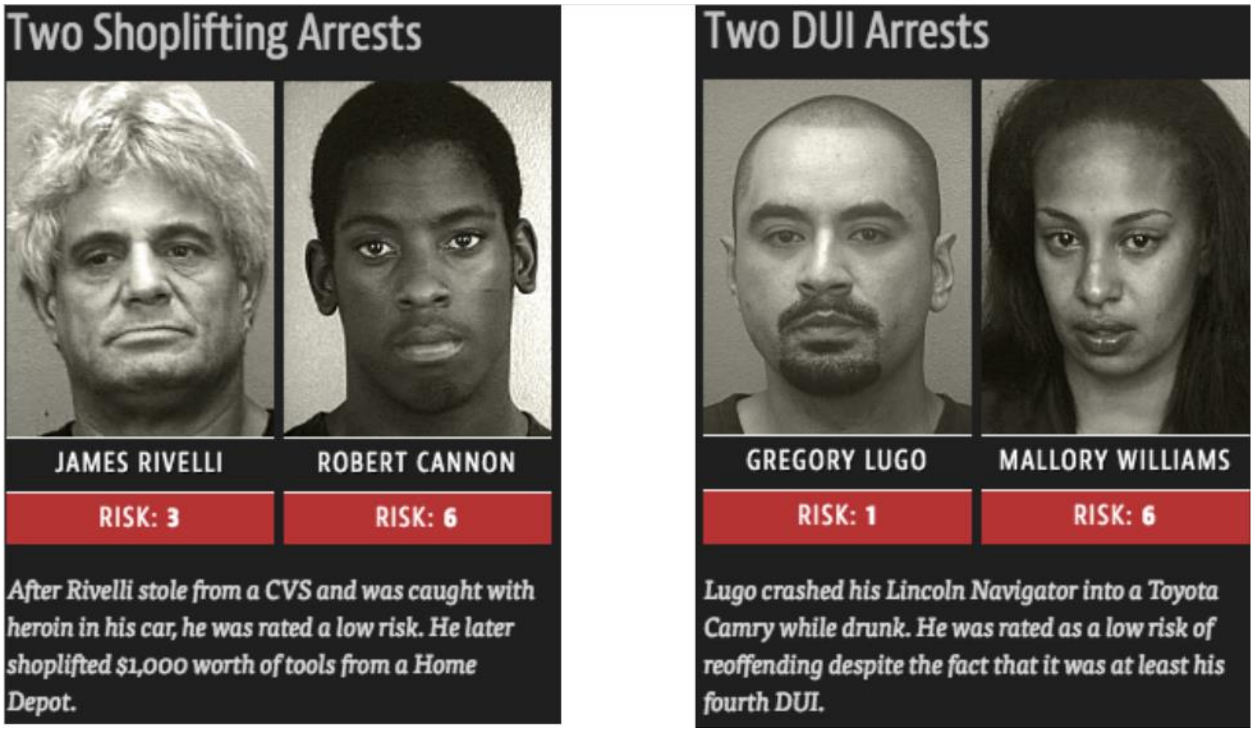
The principle of **justice** requires that we distribute both good things (money! food! jobs!) and bad things (punishment) *fairly,* and that people "get what they deserve."

- How can we ensure that AI algorithms—which are now widely used for employment, criminal sentencing, and many other things—don't "learn" to discriminate against the same groups that existing society discriminates against?
- How can we ensure that the *benefits* of data are distributed *fairly?* For example, most people don't want to lead in a world where just a small handful of people reap the "benefits", while everyone else pays the "costs."

We'll look at a few case studies related to these principles, with a focus on issues of "Big Data" and "Artificial Intelligence."

3.2 Case Study: Crime and punishment

Over the last decade, AI (such as Northpointe's "COMPAS") has been increasingly used to provide "risk assessments" for judges, parole boards, and others involved in criminal sentencing. This AI is "trained" using large databases of people who have formerly been tried and/or convicted. The risk assessments purport to identify the probability that an individual will commit crimes *in the future*. This, in turn, affects decisions about how long the individual will be sentenced, whether they will receive parole, etc. However, recent research (Propublica 2016) has suggested that COMPAS is biased against black defendants—it *overestimates* how likely they are to reoffend, even as it *underestimates* the probability that otherwise similar white defendants will reoffend.



After Rivelli stole from a CVS and was caught with heroin in his car, he was rated a low risk. He later shoplifted $1,000 worth of tools from a Home Depot.

Lugo crashed his Lincoln Navigator into a Toyota Camry while drunk. He was rated as a low risk of reoffending despite the fact that it was at least his fourth DUI.

Figure 1 Source: PruPublica.

Software systems based on Big Data/AI can be biased for several reasons, including:

1. **Biased data.** The "machine learning" algorithms used for many tasks need to be "trained" using data. Biased training data -> biased output.
2. **Societal bias.** Training data might reflect social biases based on gender, race, age, religion, etc. The algorithm will then "learn" to internalize these biases.
3. **Algorithmic bias.** Bias may result from problems with the algorithm itself, perhaps as a result of lack of diversity on the development team, or failure to test it adequately.
4. **User bias.** Users may apply AI in selective ways that reflect their own biases.

Other areas in which the risk of AI bias and discrimination might arise include:

- Interview and hiring
- Loan approval
- Addiction risk
- Policing and security

AI systems can "learn" to be biased based on age, gender, nationality, religion, race, etc. *even when the underlying algorithms don't include these factors explicitly.*

AI Bias: Open Problems. Testing for bias is difficult, because

1. the ways in which an AI might be biased are numerous,
2. an initially unbiased AI can "become" biased based on program updates or new training data, and
3. programs based on machine-learning, deep-learning, or neural network paradigms can appear as **"black boxes"** even to their developers. A "black box" is a program whose process for producing outputs we don't understand.

Voluntary standards for AI bias detection and mitigation are being developed, but these are still a work in progress.

3.3 Case Study: The Surveillance State

In recent years, China has used AI and Big Data methods to create a burgeoning security state, where "The Great Firewall" restricts residents' access to the internet, cameras (connected to facial recognition-AI) track many public spaces, and social media is monitored by AI to to detect "problem behavior." Finally, there are plans for a robust "social credit system" whereby citizens are evaluated wholistically. Many of these measures are (so far) relatively inefficient and/or in limited use. However, their more widespread use in Xinjiang to track/control the Uyghur people suggests the threat isn't wholly theoretical. China has begun to export this technology to other non-democratic regimes.

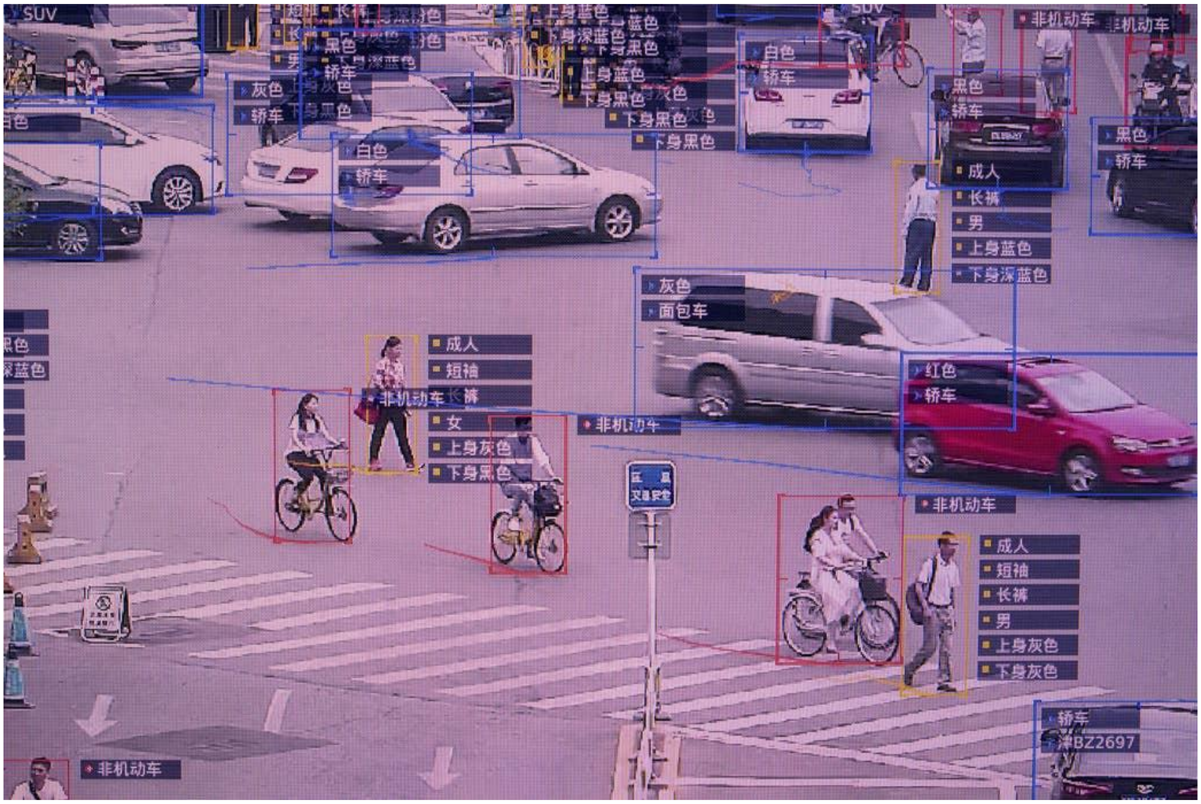


Figure 2 A 2018 demonstration of SenseTime's AI-based capacities. Source: Bloomberg/Getty

Issue: Surveillance and manipulation:

- AI makes gathering and analyzing data about individuals much easier. This, in turn, can allow for greater control/manipulation. Scenarios include:
- Aspiring totalitarians could use AI to monitor/control their populations in ways previously unimagined.
- Social media sites can use AI to target users individually, with the goal of maximizing time spent on the site, regardless of the cost to the user or society.
- Advertisers and special-interest groups can use AI to harvest user data and target ads (as in the Facebook-Cambridge Analytica scandal).
- AI makes it easier to build and market and "addictive" technologies.

Ethical issues raised by AI surveillance and manipulation include:

- To what extent should/can individuals have control over how their "data" is gathered and used by AIs? This is an issue of "autonomy."
- How can we prevent the use of AIs to harm/control people? Existing models of nuclear arms treaties and genetic engineering standards are promising but imperfect.
- Medical researchers, credit agencies, police forces, intelligence agencies and many other private/governmental bodies have legitimate need to gather and analyze data using AI. This needs to be weighed against the above risks.

3.4 CASE STUDY: SLAUGHTERBOTS

Numerous militaries have pursued development of AI-based drones that can navigate their way to a target, "recognize" the target, and (perhaps) kill/destroy the target without the need for human direction. In summer 2021, there were reports that Israel deployed the first such drone swarm in a military operation in Gaza. Stuart Russell (a leading AI researcher) has argued that slaughterbots are a potential "weapon of mass destruction", and one that would be much more difficult to control than either nuclear or chemical weapons. He warns that rogue states, corrupt government officials, or others could easily use these to kill large numbers of people.

Issue: Autonomous weapons and AI accountability. In short- to mid-term, we may see AI weapons of the following sort:

- Cheap, mass-producible drones of the type just described. These could easily overwhelm enemy defenses or enable the killing of enemy targets without harm to the surrounding infrastructure.
- Autonomous tanks, troop carriers, fighter jets, auto-targeting rifles, anti-missile systems, etc.
- Algorithms for attacking enemy infrastructure such as electrical grids, cell phone networks, or medical systems.

Ethical Issues: AI Weapons. If states have a legitimate right to self-defense, they have a prima facie right to research/manufacture weapons needed for this defense. However, they need to consider:

1. The ease with which AI weapons could be captured/used by bad actors.
2. How AI weapons will be monitored and held "accountable" for errors (such as inappropriately targeting civilians).
3. The potential that their own development of AI weapons may lead to an "arms race," with the resulting bad consequences.

4. The “Campaign to Stop Killer Robots”, among others, has argued that the most dangerous types of AI weapons ought to be banned altogether.

Robot-human interaction. There is a general need to set standards for human-robot interactions :

- Autonomous vehicles must balance the objective to drive their occupant(s) quickly and safely with the need to minimize harm to pedestrians and others.
- Industrial robots must balance productivity vs. risks to worker safety and/or product safety
- AI software used in areas such as psychiatry or medicine (either by itself or as part of a human-AI “team”) must minimize harm to patients, while delivering cost-effective care.

3.5 Review Questions

1. How do we ensure that people's data is used ethically?
2. What are four proposed principles of data ethics? Can you explain what they “mean” in your own words?
3. How can we ensure that data is used beneficially?
4. How can data be used to create a fairer, more just society?
5. What are the risks of data breaches?
6. How can we protect people's data from being mishandled?
7. What are the consequences of unethical data practices?
8. How can data be used to exploit people?
9. How can we make sure that data is used for good?
10. What are some of the implications of new data technologies?

4 Reading: Optimizing Schools (from Princeton University Case Study)[1]

In 2012, Minerva High School, a public school in Pittsburgh, PA, with nearly 3,000 students and 180 classroom teachers, reached a depressing milestone. That year, the student dropout rate hit its highest point

[1] This work is licensed under a Creative Commons Attribution 4.0 International License.

ever at nine percent. This meant that nearly one out of every ten students who entered the school as a freshman left without graduating at all. And only 55 percent of Minerva students were

able to graduate on time, compared to the state average of 76.4 percent. The school board responded to these dismal statistics by demanding that the school principal, Mr. Vulcani, address this growing problem, or risk the loss of funding – even threatening a possible shutdown. Mr. Vulcani feared for the school's future, but he was at a loss. Students were habitually disengaged, and teachers had been demotivated by frequently shifting incentives, the imposition of new pedagogical approaches and increasingly negative assessments.

Mr. Vulcani met with school board members who suggested he put the vast and varied datasets the school had already collected about its students' behavior to use. In addition to the typical performance, disciplinary and attendance records, the school had given students scannable ID cards that registered when they were used to open the library doors, confirm attendance in class, purchase snacks or lunches and the like. Even the school's Wi-Fi network tracked students' internet use and monitored their movements throughout the campus with an impressive degree of accuracy, using their mobile phones as proxies. These measures produced large quantities of data. The school board members suggested that developments in data science and machine learning could be applied to this information in order to shed light on the causes of what appeared to be an irreversible trend towards high dropout rates. Understanding what causes students to drop out might suggest appropriate interventions and could inform the creation of new incentive structures for teachers and students.

Mr. Vulcani took these suggestions to heart and contracted a local data science company, Hephaestats, that promised insights into business processes through novel approaches using artificial intelligence. Together they formulated these specific goals:

1. To identify predictors of student disengagement as an indicator for dropping out, and to apply machine learning tools to these predictors in order to flag at-risk students;

2. To equip teachers with fine-grained information to allocate resources and assist at-risk students by suggesting specific interventions, such as talking to the student directly, adjusting their workload and schedule, contacting parents or meeting with the student's counselor;

3. To be transparent in the way the system is used.

Mr. Vulcani and the school board agreed to provide Hephaestats with their existing databases, spanning several years, and gave them access to new data as it was collected. Students and parents were not notified of this agreement, nor were they given the opportunity to opt out. Knowing from previous experience how difficult it is to get parents and administrators to come to a consensus on any new initiative, and given the urgency of the situation, Mr. Vulcani believed this was for the best. Besides, he argued, this decision was supported by the school board and fell within his general mandate to promote positive educational outcomes for all.

> *Discussion Question #1:*
>
> How should decisions to adopt AI technologies be made? In this case, it came from above – a suggestion from the school board, implemented by the principal. Who are the other relevant stakeholders? Should they have been involved in the decision to utilize Hephaestats? To what extent? How does the decision not to include an opt-out option affect the legitimacy of the school's actions?

Upon receipt of the student data, Hephaestats began with a broad policy of data analysis looking at a large number of predictors, ranging from various student demographics (e.g. race, ethnicity, gender, mobility, address, home life) to academic factors (e.g. grades, GPA, test results, history of disciplinary action, attendance) to teacher statistics (e.g. certifications, degrees, percent of students failing per class, years of teaching). Hephaestats then harvested new data for a full academic year, allowing its machines to correlate the data of students that previously dropped out with information about current students in order to recognize patterns. Ultimately, Hephaestats was able to produce its own synthetic data by generating inferences that would not have been possible without inputting the school's original data into its algorithms.

> *Discussion Question #2:*
>
> Did the school violate the privacy of its students by sharing their data with Hephaestats? If so, is this breach justifiable, and on what grounds? Would you feel differently if you were the parent of an atrisk student versus the parent of a valedictorian?

Using all this data, Hephaestats was able to identify eight key indicators that, in combination, predicted whether a student would drop out with 92 percent accuracy. The reasons ranged from predictable administrative issues (e.g. an overly ambitious or unsuitable combination of chosen courses, course scheduling) to external factors (e.g. balancing a job with school, domestic responsibilities), as well as previously unconsidered factors (e.g. poor nutritional options in the school cafeteria). Hephaestats then supplied teachers with profiles of at-risk students that helped them better understand why an individual student might be struggling and suggested targeted approaches for helping him or her. These treatment protocols included things like tutoring, modifying assignments, talking with the student's guardians and, in certain extreme cases, contacting local authorities to address possible problems at home.

Some teachers readily followed the recommendations made by Hephaestats, and there was an immediate boost in student engagement. At the same time, the administration used the information provided by Hephaestats to adjust certain aspects of the campus environment in order to nudge students towards better behavior. For example, students were encouraged to change some of their courses to make their schedules more manageable. Some of the most popular sugary foods were also removed from the cafeteria, and all teachers were instructed to provide more attention to struggling students, rather than the students who were already expected to do well.

By the end of the 2016-17 academic year, Minerva High School appeared to have made an impressive turnaround. The four-year graduation rate had risen from 55 percent to 85 percent – a praiseworthy increase, according to the superintendent, that was distinctly higher than the district's average. The dropout rate similarly improved, from nine percent to only five percent. Nearly all Minerva students now left school with their diplomas, and a higher percentage than ever before were being accepted to four-year colleges. In an interview with a local news reporter, Mr. Vulcani praised the work done by Hephaestats, which he argued was instrumental to this improvement in student outcomes.

> *Discussion Question #3:*
>
> How might we define a successful outcome for Minerva High School? What was the school hoping to optimize for by contracting Hephaestats? To what extent is this end legitimate? In order to achieve this end, did Hephaestats ask the right questions and use the right proxies?
>
> *Discussion Question #4:*
>
> Graduation statistics did, indeed, improve after Hephaestats came on the scene. But correlation doesn't necessarily imply causation. What are some alternative explanations for the improvement in dropout rates?

The seeming success of Hephaestats' approach was overshadowed to some extent by concerns raised by students and their parents when they were finally told about Hephaestats' involvement. Not only did they learn that Hephaestats had been using student data to make its recommendations, but they were informed that the school's data collection and use policies were going to continue into the foreseeable future in order to maintain the system's accuracy. Effective as Hephaestats had been in improving educational outcomes, this decision, delivered from above, made many students and their parents uncomfortable. Emboldened by their critiques, some teachers and administrators also began to publicly voice criticisms of and opposition to the new system.

4.1 ETHICAL OBJECTION #1: PRIVACY

Critics claimed that the data provided by the school to train Hephaestats' algorithms amounted to a fishing expedition, whereby vast amounts of data were provided without regard for its sensitivity. Surely, they argued, the school was breaching privacy laws or at least some expected norms by handing over such a broad range of data to a commercial entity. And even if the goals were laudable, it didn't seem right to collect and use student data without first getting consent from students or parents. If the community allowed such a blatant violation of privacy in this instance, they worried it would set a bad precedent.

4.2 ETHICAL OBJECTION #2: DEHUMANIZATION OF STUDENTS AND FACULTY

Many students didn't like the idea of being treated as research subjects – even if it was for their own good. Several reported feeling like subjects to be optimized in a living lab, where every action could be scrutinized by an invisible, all-powerful computer. Critics acknowledged that the new AI approach to enhancing educational outcomes might be beneficial for those students who could be targeted for intervention, but that overall, the system seemed to be primarily geared towards improving the school's ranking. This sidelined the school's subtler aims, such as providing a safe space for students to fail, the opportunity to experiment with interests and ideas, and a nurturing environment in which students could grow to be thoughtful, self-sufficient adults. Teachers argued vehemently that their training, experience and intuitions were being overridden by the new AI system.

4.3 ETHICAL OBJECTION #3: TRANSPARENCY

In general, students, parents and teachers felt they had been forced to trust a process for evaluating risk and identifying solutions that they could not scrutinize themselves. If they were to be beholden to such an authoritative, powerful and opaque system, then they wished to be involved in the design process as participants, rather than serving merely as its data subjects. They also insisted on a procedure for appealing disputed results and requested the source code of Hephaestats' systems. To emphasize the urgency of their request, they threatened to file a formal Freedom of Information Act (FOIA) request.

In addition to these ethical considerations, the school was criticized for over-enthusiasm about using artificial intelligence as a way to modernize education. Several teachers argued that the machines were just telling administrators the same things teachers had been saying for years. Working on the frontlines, teachers often know why individual students are failing and what to do about it. They just rarely have sufficient institutional support to make the necessary changes. Thus, these teachers attributed much of the improvement in graduation statistics after the introduction of Hephaestats merely to the "Consultant Effect."

One of the school's math teachers had a slightly different take, but similarly questioned the "wisdom" of AI systems like Hephaestats. She wrote an op-ed in the local newspaper, describing her concern that the new system was simply rehashing statistical methods, rather than using a new form of intelligence. The teacher explained that all the flaws, limits and biases that are well-known in statistics were being swept under the rug by rebranding the system as artificial intelligence, thereby "blackboxing" the processes. She argued that blind faith in such an unchecked system of statistical governance may lead to the kinds of long-term problems that still plague the field of statistics.

> *Discussion Question #5:*
>
> The rhetorical decision to call a technology "AI" imbues it with a certain mystique. But quantitative and statistical methods, such as those used in many AI systems (including Hephaestats) inherently involve generalizations. While there is value in using statistics to understand social problems and make predictions, the methodologies may not be useful on an individual basis. What is the danger of calling a system that deals in particulars "AI"? What are the advantages?

Representatives from Minerva High School and Hephaestats met with concerned students, parents and teachers to respond to their worries about the new system. While the school admitted to handing over vast amounts of data, they considered the privacy concerns unfounded. First, the databases were all pseudonymized and no record was being kept of the link between data points and students' identities in the raw dataset. Second, even if the data had been identifiable, they argued that they did not need to seek consent from students or their parents because the data was collected for the legitimate purpose of improving educational outcomes.

For its part, Hephaestats resisted calls to release its proprietary algorithms. Representatives from the company argued that the source code would be meaningless to the public. The original code was developed as a means to enable Hephaestats' algorithms to learn from patterns and correlations in the existing data, but the code had long since evolved into complex neural networks with millions or even billions of nodes. It would be impossible for Hephaestats (or any other machine learning experts) to adequately interpret the current functioning of such a system. However, Hephaestats did provide some tools that would allow students to understand which data points were used to recommend a particular course of action. Given the state of the art of such technology, this was the best they could feasibly do.

Hephaestats' representatives agreed with the math teacher that their system was largely based on statistical methods. However, they noted that the system makes use of some recent developments in the field of machine learning, which is both a subfield of artificial intelligence and predictive statistics. Since the term "artificial intelligence" is more commonly used, they decided to use that labeling, while making sure to correct their processes and calculations for the known limitations in the field. They did not intend to mislead.

4.4 Reflection & Discussion Questions

4.4.1 Privacy

When designing a system of AI governance, some trade-offs are inevitable. For example, individual privacy considerations often must be balanced against the desire to achieve legitimate social ends. The extent to which specific values are embedded in systems reflects the priorities and preferences of the systems' designers. And the extent to which users accept and utilize these systems is likewise reflects users' priorities. In the Minerva case, the school board and Mr. Vulcani decided that some infringement of students' privacy was a reasonable price to pay for a lower dropout rate.

- How should decisions about the appropriate balance between privacy and improving educational outcomes be made? Who should be involved in this process and to what extent?
- How does the issue of privacy change in the school setting? Would the appropriate balance between privacy and social welfare be different in a private school? In an office setting? Think of this question using both ethical and constitutional law frameworks.

4.4.2 Autonomy

Autonomy is an individual's ability to make decisions for herself and act upon them.

Hephaestats, like many other AI systems, may compromise this value. For teachers, the use of Hephaestats' system could mean that their priorities and judgements are overridden. The issue of autonomy is still more complex for students being targeted for interventions. Students who are still minors are not thought to have the same degree of autonomy as adults, either in theory or law. Treating minors as less autonomous can be good in terms of protecting more vulnerable members of society, but also raises some difficult questions about paternalism.

- Should Hephaestats provide students with their risk profiles? Should students have a right of appeal? Should they be able to opt out of being assessed? Would it be possible to include them in decisions regarding the design and deployment of Hephaestats, and if so, how?
- Hephaestats offered several options to address the student dropout rate at Minerva High School, but they mostly emphasized a "nudging" model. This meant that students typically did not receive explicit mandates or directives; rather, they were nudged along in certain directions through changes to the incentive structures (e.g. by making sugary foods less available). Nudging represents a softer form of influence, but it is influence nonetheless. AI systems designed to "nudge" are justified on the basis that they adjust choice architecture to help people make the right decisions. But the right decisions for whom? Who decides? Are nudges a better way of producing desired outcomes than more explicit exertions of power? Why or why not?

4.4.3 Consequentialism

Some people argue that certain actions are impermissible regardless of what good outcomes they might bring about; others believe that the ends may justify the means. The Minerva administrators and their partner, Hephaestats, had both good ends and, they argued, appropriate means. But in complex AI systems, it may be quite challenging to even keep track of the various means in use. If all these means must be evaluated independently of the ends they're used to bring about, it may be very difficult to evaluate the permissibility of different actions. Furthermore, when AI is deployed to solve real world problems, each step of the implementation must be tracked as well. Considering the difficulty of assessing each of these steps in their entirety, school officials and Hephaestats preferred to focus on their noble end of reducing the student dropout rate.

- Even if nearly everyone felt the school dropout rate was a problem, not all stakeholders agreed with Hephaestats about the appropriateness of their means, namely, their use of student data without consent to produce un-auditable results. These dissenters might argue that the way Hephaestats went about reducing the dropout rate undermined its ultimate success in achieving this "noble" end. What would you say?
- If we accept that all significant stakeholders ought to have a voice in determining the values they want their communities to promote, does it follow that they should be involved in decision-making about the means of achieving those ends as well? How would schools go about including them?

4.4.4 Rhetoric

The use of language is very important, especially in framing and describing new, developing technologies. Hephaestats chose to label and promote their IT solution as "artificial intelligence," but they also could have labeled their approach as a matter of social science or statistics, instead.

- Was Hephaestats right to call its technology "AI"? If not, how should they have formulated their description of the system to the school board, students, parents and teachers?

- What are the implications of calling something "AI"? What kinds of political and social intentions does that decision reflect? Does the "AI" labeling evoke a kind of responsibility beyond what is typically expected from an IT system? If so, what could be done to mitigate these concerns.