Agenda: Quick review of Chapters 1 – 7, followed by discussion of Chapters 8 – 9 (Risks and Control). My comments are in brackets.

1. **NOW**

Data is being collection in a big way **NOW**. Examples: Flu prediction (analyzing search terms faster than medical reports), airline fare changes, astronomy, human genome.

2. **MORE**

From *sampling* to *all*. Allows drilling down into detail and unanticipated uses.

3. **MESSY**

Increased volume means less accuracy. "More trumps better." Most analyses don’t require great accuracy. Causes of messiness: natural errors, disparate sources, different formatting, transformations (e.g., extracting movie opinions from Twitter messages). Example: Yahoo table-of-contents or keywording vs. Google page-rank searching.

4. **CORRELATION**

*What* not *why*. Example: Amazon and Netflix recommendations; Pop-Tart sales before hurricanes. [Eat when cold? Fast to prepare? Long shelf life if not eaten? No dieting when death is imminent? *Doesn’t matter.*] Valid hypothesis not required, and often not even attempted.

5. **DATAFICATION**

Putting information into quantifiable format. Examples: Captain’s logs, treating books as digitized text instead of images, location as coordinates, social media interactions, physical-activity sensors.

6. **VALUE**

All data is valuable, but keeping it relies on cheap acquisition and storage. Useful for immediate and future unanticipated purposes. Example of reuse: search terms. Large collectors (e.g., mobile phone providers) can sell data to analyzers. Value may diminish over time (e.g., Netflix movie watching). Digital trail (e.g., mistyped search terms, mistakes on quizzes) called *digital exhaust*. Increasing availability of open data (e.g., data.gov).

7. **IMPLICATIONS**

Three types of big-data companies: Holders, skills, creative users ("big-data mindset"). Baseball after *Moneyball* shows that a competitive edge disappears.

8. **RISKS**

*Privacy*, *Propensity* (predicting and punishing bad behavior), *Dictatorship* of Data (overuse for decision making; repression).
A. Privacy

Not all data contains personal information (e.g., oil-refineries, manhole covers). Can determine individual from large, anonymized data sets. Secondary uses may not have been authorized.

Specific authorization prevents good uses (e.g., flu trends, fires in buildings); blanket authorization provides no protection. Opting-out still reveals information (blurred houses indicate possible burglary targets).

_Individual notice and consent, opting out, and anonymization_ have lost their effectiveness because of bigger data and smarter analysis.

Businesses are only out to make money, but governments can do harm.

B. Propensity

Predicting future behavior, and punishing without any bad actions. *(Minority Report* film.) Does deploying police to areas of predicted crime interfere with civil rights? Assigning special counseling to at-risk children? Should airport screening be based on profiling? What happens to presumption of innocence?

C. Dictatorship of Data

Over-reliance on numbers: McNamara and the Vietnam War body count.

Underlying data can be of poor quality and biased. Misanalyzed. Not capture what we think it does (e.g., standardized test to measure school quality; SAT and GPA for job candidates).

9. CONTROL

Authors advocate shift from individual consent at time of collection to holding data users accountable. [But should we leave assessment of harm to the user, who has other priorities?]

Regulators would establish different time frames for reuse depending on risk and societies’ values. Right to keep data longer coupled with responsibility for its use and requirement to destroy it. [Do you trust the regulators? Will they be starved for funds by corporate-controlled lawmakers?]

Deliberately blurring data helps privacy (e.g., from 18 people who rented film to <50 who rented film).

Individuals responsible for _behavior_, not _propensities_. Safeguards: open data, open algorithms, means to disprove predictions. [Bad example: No Fly List.]

Algorithmist: Ensure best practices, monitor usage, review accuracy and validity. [Who watches them? What authority do they have. Internal ombudspeople have no power.]

10. NEXT

Authors: Need to establish new principles by which we govern ourselves. [Fat chance.] Big data needed for dealing with climate change, improving health care, economic development, conflict prevention.

Big data doesn’t replace invention.

[My conclusion: Volume of data collected and sophistication of analysis will continue to increase; regulation won’t keep up.]