Introduction to Weapons of Math Destruction

Big Data, Inequality, and the Age of the Algorithm

Cathy O'Neil

Intro & Chapter 1

The Algorithmic Promise

- **The Goal:** Replace biased, inconsistent human decision-making with ”cold, hard numbers.”
- **The Theory:** Everyone is judged by the same rules; bias is eliminated.
- **The Reality:** Many models are opaque, unregulated, and reinforce existing discrimination.

”The era of the algorithm... should lead to greater fairness... but the opposite is true.”

The Three Pillars of a WMD

To be a "Weapon of Math Destruction," a model must have these three traits:

1. **Opacity:** The model is a "black box." The people affected don't know how it works or how to challenge it.
2. **Scale:** It affects massive numbers of people (e.g., all teachers in a city or all applicants for a loan).
3. **Damage:** It has a negative impact on people's lives, often creating a "vicious cycle" or "downward spiral."

Case Study: D.C. Teacher Evaluations

The Intention: Fire ”bad” teachers and reward ”good” ones using objective data.

- **The Model:** Used ”value-added modeling” based on standardized test scores.
- **The Flaw:** The model couldn’t account for external factors (poverty, erasure/cheating in previous years).
- **The Result:** Excellent teachers like Sarah Wysocki were fired without any clear explanation or ”feedback loop” to correct the error.

The Danger of "Proxies"

When models lack direct data, they use **proxies** (stand-ins).

What they want to measure	**The Proxy used**
Good Teacher	

The Danger of ”Proxies”

When models lack direct data, they use **proxies** (stand-ins).

What they want to measure	**The Proxy used**
Good Teacher	Standardized Test Scores
Responsible Employee	

The Danger of ”Proxies”

When models lack direct data, they use **proxies** (stand-ins).

What they want to measure	**The Proxy used**
Good Teacher	Standardized Test Scores
Responsible Employee	Credit Scores
Future Criminality	

The Danger of ”Proxies”

When models lack direct data, they use **proxies** (stand-ins).

What they want to measure	**The Proxy used**
Good Teacher	Standardized Test Scores
Responsible Employee	Credit Scores
Future Criminality	Zip Code / Social Circle

The Danger of ”Proxies”

When models lack direct data, they use **proxies** (stand-ins).

What they want to measure	**The Proxy used**
Good Teacher	Standardized Test Scores
Responsible Employee	Credit Scores
Future Criminality	Zip Code / Social Circle

Problem: Proxies often correlate with race and class rather than merit.

The Counter-Example: Baseball Models

Why is "Moneyball" *not* a WMD?

- **Transparency:** Stats (home runs, RBIs) are public and understood by everyone.
- **Relevant Data:** Uses actual performance (hits, strikes) rather than "proxies" like zip codes.
- **Feedback Loops:** Models are updated daily. If a prediction is wrong, the model is "tweaked" immediately.

Discussion Questions

1. Why does O'Neil say that "the privileged are processed by people, the masses by machines"?
2. Can a model be "fair" if it is 95% accurate but ruins the lives of the 5% it gets wrong?
3. How can we create "feedback loops" for algorithms used in hiring or education?