

It's a Noisy World Out There ☹️!



Linda Rising

www.lindarising.org

linda@lindarising.org

<https://www.linkedin.com/in/lindarising/>



5-minute break

- Questions? Use the index cards 😊!
- We all have a conscious attention limit of about 40-45 minutes.
- This session is 75 minutes, so, we will take a **5-minute break** after 35 minutes.
- Stand, if you are able and want to; stretch; talk; take a nap; go outside; whatever 😊!



How do we make decisions?

Our mostly informal approach goes something like:

Gather, examine data

Do some simple analysis

Check our intuition or feelings (System 1)

Reach a decision

Called “clinical judgment” -- extensively studied by psychologists.



Results?

- Do we compare with actual results?
- If we do, we rationalize or explain any discrepancies ☹!
- Research shows correlation is ~55%, barely better than chance.
- Save time by simply flipping a coin!



Paul Meehl -- 1954

Meehl (University of Minnesota) reviewed 20 studies comparing clinical judgment with mechanical prediction (models, algorithms).

Mechanical prediction generally superior to human judgment.

Reactions to Meehl's findings: shock, disbelief, contempt, and dismissal.

Meehl: evidence for the advantage of the mechanical approach was “massive and consistent.”



Follow-on review -- 2000

A 2000 review confirmed that mechanical rules outperform clinical judgment.

Results understate advantages of mechanical prediction, which is **FASTER** and **CHEAPER** than clinical judgment.

Humans had an unfair advantage in many studies -- they had **PRIVATE INFORMATION** not supplied to the model.

Results were clear -- simple models beat humans.



Why don't we do something about it? We don't like it! ☹️!

Mechanical models are controversial. If a decision has a big impact on our lives, we prefer that a human make it.

BUT what if a mechanical model makes better decisions?

The idea that mechanical models make better decisions than humans in many fields is an old one.

The passing of time has not been kind to humans ☹️!



Oregon Research Institute - 1968

- Researchers created a simple algorithm measuring likelihood that an ulcer was malignant
- Doctors listed 7 factors, equally weighted.
- Then, doctors judged probability of cancer in 96 stomach ulcers.
- Doctors shown each ulcer twice, randomly mixing duplicates so doctors wouldn't notice.



Initial results

- Simple first algorithm assumed to be starting point.
- Surely, algorithm should be more complex to capture the subtleties of the doctors' thinking.
- Results were unsettling. Researchers described results as “generally terrifying.”
- Simple first algorithm was very good at predicting the doctors' diagnoses.




The doctors might want to believe that their thought processes were complicated, but the simple model captured them very well.

This doesn't mean the doctors' thinking was simple, just that it could be captured by a simple model.

More surprisingly, doctors' diagnoses were all over the place. They didn't agree with each other. When presented with duplicates of the same ulcer, every doctor contradicted him/herself, giving different diagnoses.

If you wanted to know if you had cancer, you were better off using the simple algorithm than asking a doctor to study the X-ray.

The simple algorithm outperformed not just the group of doctors. It outperformed the single best doctor.



The surprising success of equal-weighting schemes has an important practical implication: it is possible to develop useful algorithms without prior statistical research. Simple equally weighted formulas based on existing statistics or common sense are often good predictors of significant outcomes.

Daniel Kahneman, *Thinking, Fast and Slow*



What about algorithms?

- Human judgment can be replaced with algorithms.
- People have competed against algorithms in several hundred contests of accuracy over the past 60 years, from predicting the life expectancy of cancer patients to predicting the success of graduate students.
- Algorithms were more accurate than human professionals in about half the studies, & essentially tied with the humans in the others.
- Ties are victories for the algorithms, which are more cost-effective.



Is this AI?

- **Well...not really.**
- **We are talking about simple, transparent, easily understood formulas based on human input.**
- **We are not talking about complex, black box, generative programs that produce unintelligible, unknowable decision-making.**
- **That's a BIG difference.**



Noise

- That experts (of any kind, in any domain) cannot even agree with themselves is “noise” – the random variability in results.
- Given the same data twice, we make different decisions.
- Algorithms win, at least partly, because they don’t do this. The same inputs generate the same outputs every time.
- They don’t get distracted. They don’t get bored. They don’t get mad. They don’t get annoyed. They don’t have off days.



Simple algorithms can work

- Algorithms can be developed with a small number of cases or use commonsense.
- Select a few (6 to 8) variables related to the outcome.
- Assign variables equal weight in the prediction formula, setting their sign in the obvious direction (+ for assets, - for liabilities).
- Algorithms that weight variables equally & don't rely on data have been successful in personnel selection, election forecasting, predictions about football games, & other applications.

Virginia Apgar, M.D. (1909-1974)

Apgar Scoring System

Indicator		0 Points	1 Point	2 Points
A	Activity (muscle tone)	Absent	Flexed arms and legs	Active
P	Pulse	Absent	Below 100 bpm	Over 100 bpm
G	Grimace (reflex irritability)	Floppy	Minimal response to stimulation	Prompt response to stimulation
A	Appearance (skin color)	Blue; pale	Pink body, Blue extremities	Pink
R	Respiration	Absent	Slow and irregular	Vigorous cry



Ask the right question

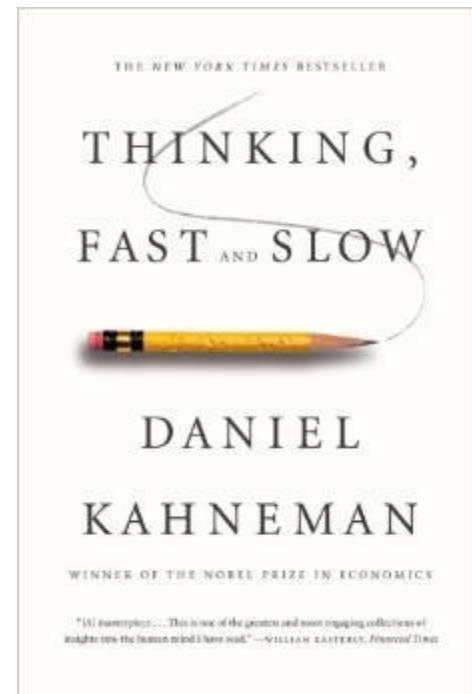
- **BUT aren't those algorithms biased?? Wrong question. Instead, ask:**
- **“How do the bias & performance of algorithms compare with the status quo?”**
- **Rather than asking whether algorithms are flawed, ask how their flaws compare with those of humans.**
- **Research on algorithmic decision-making dates back several decades and reach a similar conclusion:**
- **Algorithms are less biased & more accurate than the humans they replace.**

Bias vs Noise

Bias – predictable error that always has the same effect on thinking and decision-making

Noise – variable error that can move thinking and decision-making in any direction

Search on YouTube: Linda Rising Thinking Fast and Slow



Accurate, Noisy, Biased, Noisy & Biased



A. ACCURATE



B. NOISY



C. BIASED



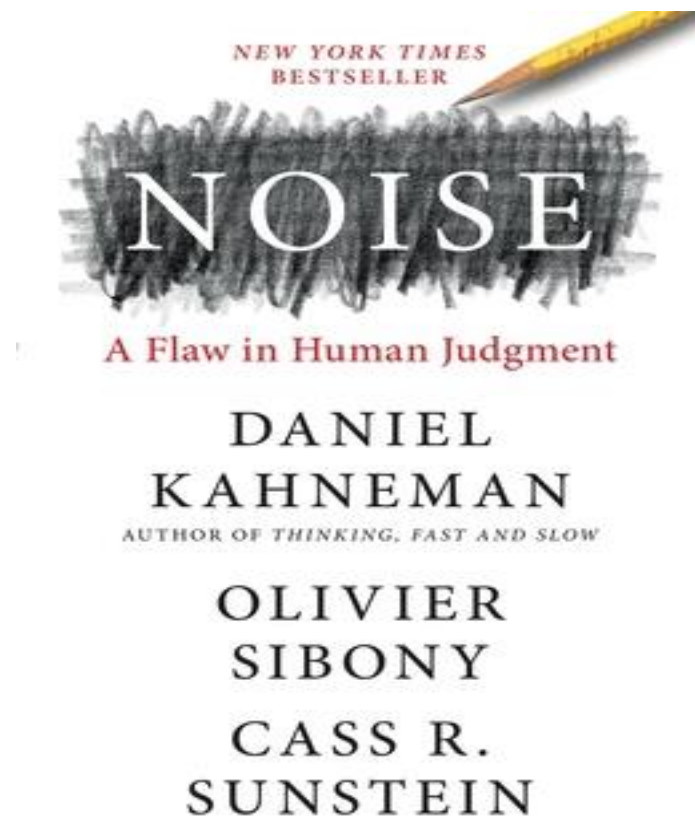
D. NOISY AND BIASED

Bias AND Noise



“Wherever there is judgment, there is noise and more of it than we think.”

Everyone is concerned about bias, but no one worries about noise. Kahneman, Sibony, and Sunstein want to change that.





How bad is it?

- Experts often contradict prior judgments when given the same data at different times.
- Software developers asked on different days produced estimates that differed by 71%, on average.
- Fingerprint examiners expressed different opinions when shown the same prints weeks apart.
- Diagnostic accuracy of melanomas was 64% -- misdiagnosed melanomas in 1 of every 3 lesions.
- Reviews of 426 patients in state hospitals found only 50% agreement on type of patient's mental illness.



Sources of Noise

- We have different personalities, beliefs, passions, emotions.
- We are different from ourselves, depending on whether we're cranky, what time of day it is, whether it's nice out, if our favorite sports team just won/lost.
- Study found judges in Louisiana gave harsher sentences to kids, especially black kids, the week after LSU college football team lost a game.
- Study of judicial decisions in France & US found judges more lenient on defendants' birthdays.



What to do?

- **Step #1. Be aware. Not enough but a good start.**
- **Track decision accuracy.**
- **It takes effort to create even a simple algorithm so don't go overboard.**
- **Don't eliminate humans and human intuition 😊!**
- **Always allow people to adjust results a little. Kahneman calls this “disciplined intuition.” Allowing wiggle room improves overall quality of decisions.**



Premortem

- Research shows premortems more reliably reduce noise & bias than other techniques.
- Ask the group to imagine a decision was made, and the outcome was a disaster.
- Each group member writes why the result was bad, then group reviews the reasons, creates action plans.
- Premortems rely on “prospective hindsight” that research shows generates more scenarios than projecting into the future.
- Gary Klein, “Performing a Project Premortem,” Harvard Business Review, September 2007.



Noise audit

- Noise can be measured without knowing the correct response.
- If targets were removed, you know nothing about accuracy, but you would know something was wrong with the scattered shots of B & D. Wherever the bull's-eye was, they did not come close to hitting it.
- Run an experiment where a few realistic cases are evaluated independently by several professionals. Scattering of judgments can be seen without knowing the correct answer.
- Experiments are called noise audits.



Noise audit example

- In an investment firm, senior leaders designed a case for analysts (who should apply identical methods).
- On average, between any two analysts there was a 44% difference in evaluations.
- Leaders' preliminary estimates were ~10%.
- The degree of variability in judgment between people is always greater than expected.



Successful noise audits

- **Noise audits only work if leaders will accept unpleasant results & act on them.**
- **Buy-in is easier if leaders view the study as their creation, so cases should be compiled by respected team members & cover typical problems.**
- **All team members should participate in the audit.**
- **The organizational unit (team, department, company) must own the process.**



Hiring & Interviews

- Interviews are a minefield of noise & biases.
- No complex judgment task has been the focus of so much field research.
- Correlation between the ratings of 2 interviewers after interviewing the same candidate is 62 to 65%.
- Variability largely the result of noise.
- Most organizations expect variability & require several interviewers and aggregate results, typically, in a meeting, which has its own problems.



Structured Interview

- List the attributes, qualities or skills expected and how they should be rated.
- Knowledge of programming language X: 1 – just learning, 2 – real, live project experience, 3 – several years of several projects.
- Ask several interviewers to meet each candidate, make **independent** judgments, fill out a detailed evaluation form.
- Facilitator collects and tallies evaluations before discussion in a meeting.



Google's hiring process

- **Several people separately interview each candidate.**
- **Interviewers have guidelines for judging candidates on specific criteria.**
- **Candidates are graded on a predetermined scale for each criterion.**
- **Hiring is gathering data, not getting a vague, overall impression based on a short conversation.**
- **After data has been collected and tallied, hiring committee meets & decides who to bring on board.**



Estimates are bad ☹️!

- Estimates are created, but rarely reviewed after project is completed.
- 5 teams of developers asked to estimate effort for a project completed in the past. Original estimate 1240 hours. Project took 2400 hours.
- 5 estimates: 1100, 1339, 1500, 1550, 2251
- One estimate close to actual but 4 were optimistic (as was original estimate).



Software estimation

- Estimations of future effort based on past effort.
- If you do new things, you'll never accurately estimate time & effort to build the software.
- Two approaches to estimations: ask an “expert” or use a model. Second tries to capture the expertise of the first -- both are noisy.
- The real question isn't, “How long will this take,” but “What can you build for this much money?”
- The answer is...



Build iteratively

Forget the models. Instead:

- 1) Prioritize requirements. Identify the core functionality.**
- 2) Build framework & core functionality.**
- 3) If there's time & money, continue.**
- 4) If there's still time & money, continue.**
- 5) If there's still time & money, ...**



Inside vs Outside Views

Inside View focuses on case at hand -- the plan & obstacles to completion, constructs scenarios of future progress, extrapolates current trends.

Outside View ignores details of case at hand, avoids detailed forecasting, focuses on data from a class of cases similar in relevant aspects to current project – reference class.

Daniel Kahneman (b 1934)



- Beware the 'inside view,' The McKinsey Quarterly, November 1, 2011.



The Outside View

Start with the outside view -- a baseline.

Adjust the baseline with information about the current project. How is the current project unique & different from the baseline? Is it less complex? Is it being built on a different platform? Is the team more experienced?

This is especially applicable if you have never done anything like this before.



Sources of Group Noise

- Who arrives first, last, early, late
- Who speaks first, last
- Who speaks with confidence
- Who speaks the loudest and most frequently
- Who is wearing black (or other significant color)
- Who is seated next to whom
- Who smiles, frowns or gestures at the right moment
- Who is the youngest, oldest, most recently hired, veteran employee...



The “Wisdom of Crowds”?

- There are “wise crowds” whose average (mean) judgment is close to the correct answer, but there are also crowds that follow tyrants, fuel market bubbles, believe in magic or hold a shared illusion.
- Small differences can lead one group to a firm “yes” and an essentially identical group to an emphatic “no.”



Music Download Experiment

- **14,000 participants randomly assigned – web site for unknown bands and songs**
- **(1) control group could see only names of bands & songs (could hear and download any they liked)**
- **(2) experimental group, further sub-divided into 8 “worlds” – could not only see bands & songs but also how many downloads by others in their “world”**



Music Download Results

- **Because groups were similar you would expect that good songs would always win and bad songs would always lose.**
- **Group rankings were wildly disparate. The worst songs (as established by the control group) never ended at the top. The best never ended at the bottom, but otherwise, almost anything could happen.**



How does this apply in organizations?

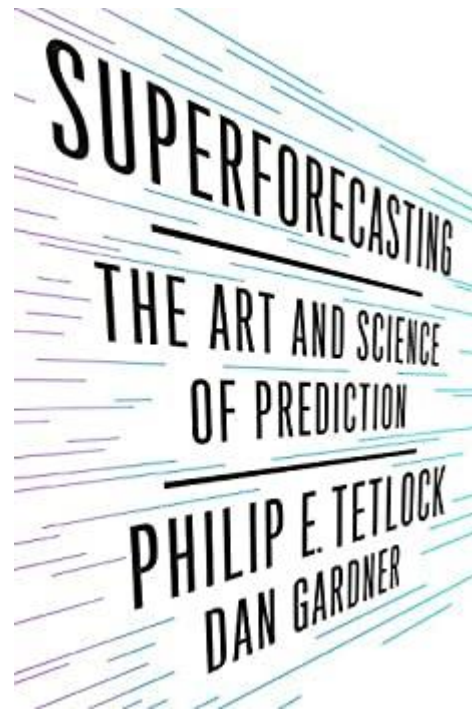
Suppose a small group of 10 is deciding whether to adopt a new initiative. If one or two advocates speak first, they could shift the group in their preferred direction. The same is true if skeptics speak first. Similar groups might make different judgments because of who spoke first (the equivalent of early downloads). The popularity of songs has a close analogue in professional judgments. Groups are noisy ☹! The solution? Keep judgments as independent as possible for as long as possible.



Reducing noise reduces bias

- The U.S. intelligence community sponsored a forecasting tournament so scientists could measure prediction accuracy.
- Results showed 1-in-50 were “superforecasters,” consistently making better predictions than other participants.
- Difference was **~50% from reducing noise**, 25% from reducing bias, 25% from increasing information.
- Noise & bias are independent sources of error, so reducing either improves forecasts.

Superforecasting - <https://goodjudgment.com>





Better decisions? Reduce noise. Use “decision hygiene.”

- **Be aware.**
- **Do a noise audit.**
- **Consider simple algorithms.**
- **Use structured interviews.**
- **In general, keep judgments as independent as possible for as long as possible.**
- **Use premortems.**
- **Adopt an outside view.**
- **Keep learning ☺!**
- **Thanks for listening ☺!**