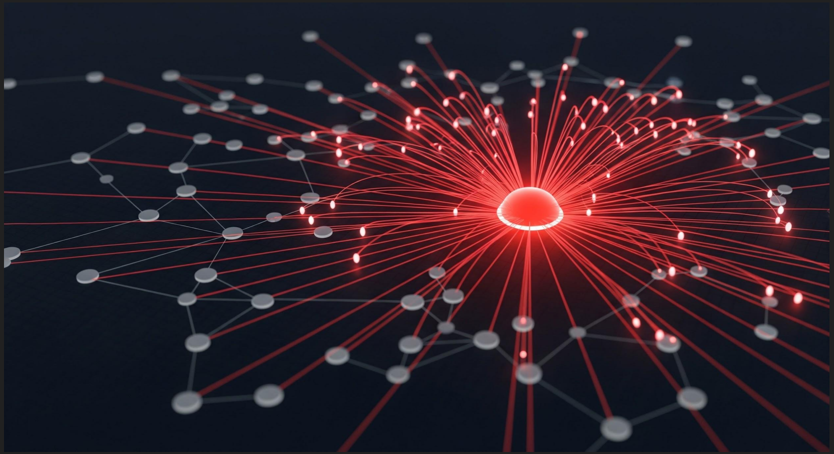
Probabilistic Machine Learning Against Disinformation

Uncertainty, Robustness, and Trust

Prof. Benjamin Guedj
Professor of Machine Learning and Foundational Artificial Intelligence, Centre for Artificial Intelligence, Department of Computer Science, UCL
Research director, Inria
Turing Fellow, The Alan Turing Institute

International Crime Sciences Conference 2025

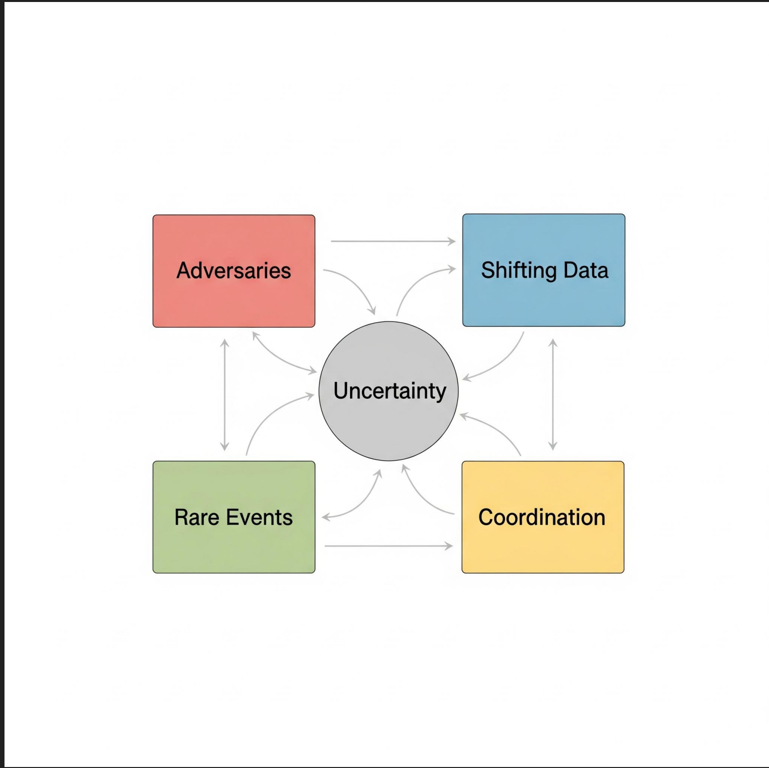
The Disinformation Problem



- AI now allows *cheap, scalable, highly convincing* synthetic content
- Disinformation spreads faster than verification
- Tactics evolve constantly → yesterday's data ≠ today's threats
- Consequences span **crime**, **security**, and **societal trust**

Why Disinformation Is Hard

- **Adaptive adversaries**: attackers change style, tone, and patterns
- **Fast-evolving data**: models trained last month already lag
- **Low signal-to-noise**: harmful content is rare but impactful
- **Coordinated behaviour**: networks of bots amplify narratives
- **High uncertainty**: ground truth is ambiguous or delayed



Why Probabilistic Machine Learning?

- Traditional ML gives **a single answer**
- Probabilistic ML gives **a distribution over answers**
- It quantifies **how uncertain** the model is
- Uncertainty spikes when data **shift** or inputs are **manipulated**
- Helps decide **when not to trust** a prediction

Uncertainty Quantification

- Models should **know what they don’t know**
- Useful for detecting **ambiguous**, **manipulated**, or **out-of-distribution** content
- Uncertainty is a **signal**, not a weakness
- Enables safer decisions: *flag*, *abstain*, *prioritise for human review*

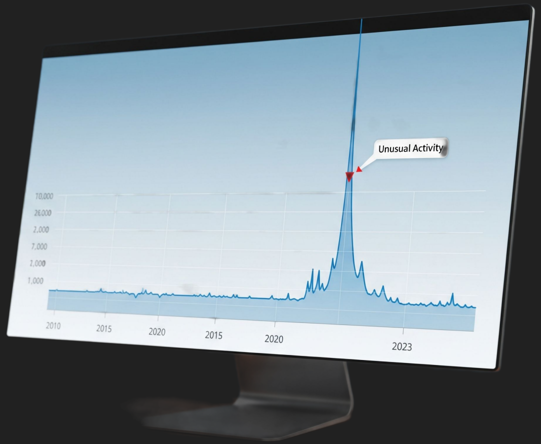
Distribution Shift: Why It Breaks Models

- Attackers constantly change style, structure, and timing
- Models trained on “yesterday’s data” quickly become obsolete
- Theory helps: PAC-Bayes links **training error**, **model complexity**, and **future performance**
- Key idea: we need models that **generalise under uncertainty**, not just fit past data
- This is a key area in theoretical machine learning

$$\mathbb{E}_Q[L] \leq \hat{L} + \sqrt{\frac{\mathrm{KL}(Q\|P) + \log(1/\delta)}{2n}}$$

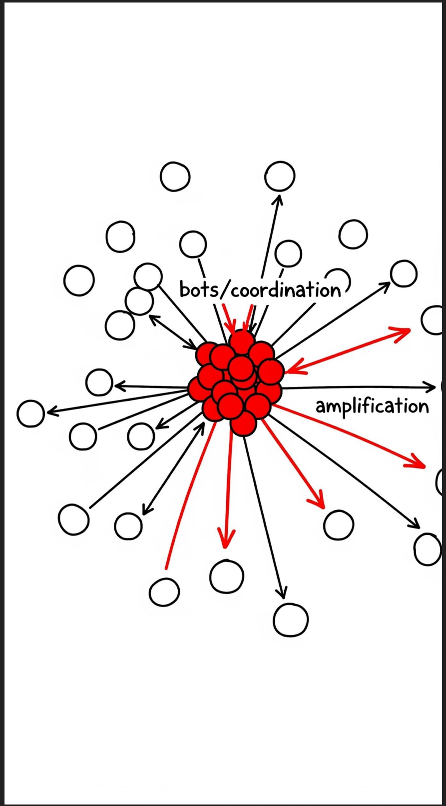
Anomaly Detection

- Disinformation often leaves **subtle statistical fingerprints**
- Detect rare patterns in:
  - writing style
  - metadata (timing, device, IP clusters)
  - propagation dynamics
- Useful for spotting **synthetic**, **manipulated**, or **out-of-distribution** content
- Works even when attackers try to mimic normal behaviour



Network Modelling

- Disinformation spreads through **relationships**, not isolated messages
- Graph models reveal **coordinated clusters** and amplification patterns
- Key signals of manipulation:
  - synchronous posting
  - repeated resharing within tight communities
  - identical message templates across accounts
- Early detection often comes from **network structure**, not content

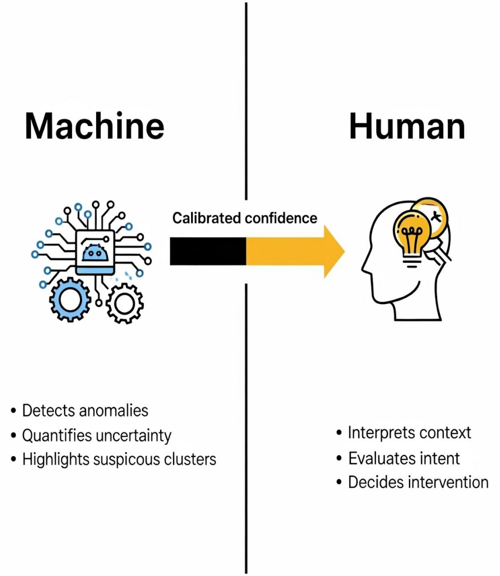


Counterfactual Reasoning

- Understand *how* content could influence different groups
- Explore “what if?” scenarios:
    - What if the message were phrased differently?
    - What if it targeted another demographic?
    - What if it spread earlier/later?
- Helps estimate **potential harm**, not just detect anomalies
- Supports smarter **intervention strategies** (timing, prioritisation)

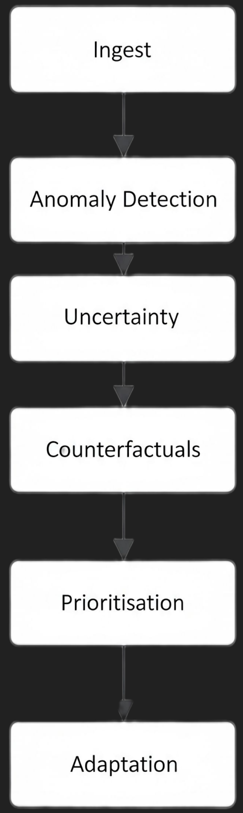
Human + Machine: Calibration & Prioritisation

- Disinformation requires **context and judgement**
- Probabilistic models guide humans by signalling:
  - *high uncertainty*
  - *high potential harm*
  - *unusual activity patterns*
- Calibrated models ensure confidence that what we get is close to reality
- Goal: **prioritise analyst attention**, not replace it



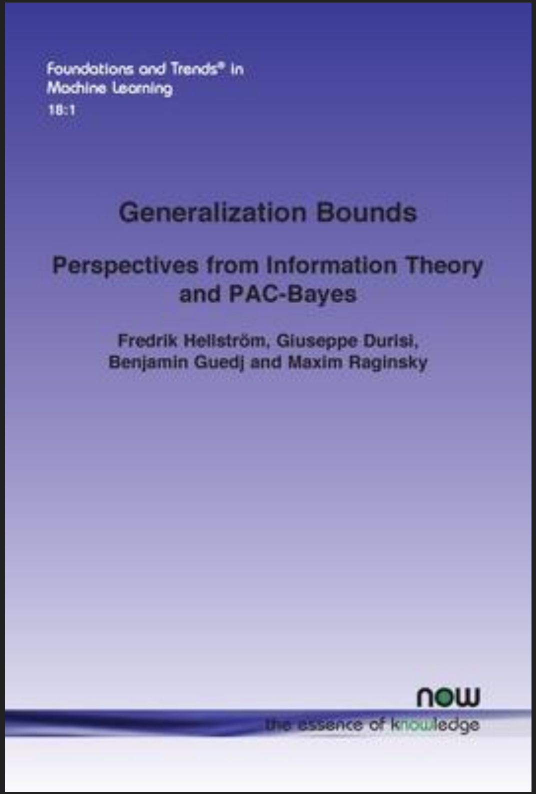
A Practical Probabilistic Pipeline

- **1. Ingest**
  Collect content + metadata + network signals
- **2. Detect anomalies**
  Spot rare or unusual patterns (text, propagation, timing)
- **3. Quantify uncertainty**
  Identify cases where the model is unsure
- **4. Assess impact (counterfactuals)**
  Estimate who could be influenced and how
- **5. Prioritise**
  Send the highest-risk items to analysts
- **6. Learn & adapt**
  Update models as attackers evolve



Takeaways

- Disinformation is **adaptive**, **fast**, and **adversarial**
- Traditional ML fails under **shift**, **manipulation**, and **coordination**
- Probabilistic ML provides:
    - **uncertainty quantification**
    - **robustness under drift**
    - **anomaly & network detection**
    - **risk-aware prioritisation**
- Goal: support **trust**, **safety**, and **effective investigation**



https://bguedj.github.io

b.guedj@ucl.ac.uk