Implications of the Manifold Hypothesis

Tom Cunningham

2025-10-15

In short: recent developments in AI can be thought of as computers getting access to the latent space that underlies reality.

Summary

In short. Reality has a low-dimensional structure. The signals we send and receive are high-dimensional (images, audio, text), but they are *clustered* such that they can be represented almost without loss in a low-dimensional space.

We have only recently taught computers to perform the mapping between low- and high-dimensional representations, and the economic effects of AI are a consequence of this.

The manifold hypothesis. Bengio et al. (2012) define the manifold hypothesis: *"real-world data presented in high dimensional spaces are expected to concentrate in the vicinity of a manifold of much lower dimensionality."*

This can be thought of as a nonlinear version of principal components analysis. Computer scientists say that neural nets work well because they're a good fit for the nature of real-world manifolds (Bengio says neural nets encode a prior that the manifold is smooth, sparse, and hierarchical).

signal (high dimensional)	state (low dimensional)
image	objects, angle, lighting
audio	text, speaker, tone, volume
text	meaning, style, phrasing

Statistical implications of the manifold hypothesis.

1. **Signals are redundant:** if you remove a pixel from an image or a word from a text you can predict what the missing point was with high confidence.
2. **Signals are compressible:** you can reconstruct a signal with high accuracy using a low-dimensional representation.
3. **Unsupervised learning is useful:** Passive observation of the world helps you learn the manifold, and so makes you much better at subsequent supervised learning. Likewise learning to predict one label helps you predict other labels (transfer learning).
4. **Shallow algorithms fail at high dimensional tasks:** Non-hierarchical learning methods (regression, decision trees) work well if you feed them the low-dimensional state, but work badly if you feed them the raw high-dimensional signal.

Application to human abilities. It is useful to think of the human brain as extracting a low dimensional state from a high dimensional signal (perceptual input, words, etc.).

Most of our perceptual ability is clearly unconscious: we are able to make very subtle inferences from huge amounts of sense data but we have limited conscious introspection into that judgment, e.g. judging how far away a tree is, judging how old a person is from their face, judging someone's identify from their voice. Psychologists will often say that the pre-conscious brain is doing some kind of Bayesian inference before presenting the results to the conscious brain.[1]

[1] A classic reference is Pylyshyn (1999) on modularity of perception.

An important observation about human capabilities is that they are *asymmetric*: it's trivial to recognize whether an object has some property (a joke is funny, a picture is beautiful) but it's far harder to create an object that has that property. This type of asymmetry can arise from a computational asymmetry (P vs NP), but I think the cause is different, it's because our konwledge is tacit.[2]

[2] I wrote a blog post about evidence for tacit knowledge, & a paper formalizing a tacit knowledge explanation of this asymmetry, & trying to apply it to economic decision-making.

Application to computer history. Computers have been able to beat humans at all sorts of computational tasks since the 1940s, including low-dimensional statistical inference (e.g. linear regression).

Only recently they've become able to match human ability in making inference about high-dimensional objects like text, audio, images. This is because it takes a lot of data to learn the mapping between low-dimensional state and high-dimensional signal. Notably computers can do the mapping both ways: they can recognize whether a picture contains a cat (like a human), and they can also create a picture that contains a cat (much harder for humans).

Application to recommenders. Historically recommenders which used the *content* of items were not very effective, e.g. recommending music to someone based on their preferences over tempo, or recommending web pages based on raw text matches. As a consequence the most successful recommender and classifier algorithms relied on engagement instead of content: e.g. pagerank, collaborative filtering.

However neural nets are now powerful enough to extract the latent semantic content of an item, so we can directly model someone's judgment of an item as a function of its content, instead of proxying their judgment with other peoples' judgment of that item. Implications: (1) we can now recommend content even when we have no engagement data (solving the "cold start" problem); (2) we're no longer constrained by what content already exists, we can now synthesize new content to maximize some function, e.g. synthesize advertisements to maximize click-through rate. (See a case study on the history of content classification here).

Application to intellectual property. Suppose that humans can recognize the properties of an object (whether a joke is funny, painting is pretty, etc.), but they can't create an object with a given set of properties. Then society will be characterized by *copying*: people will repeat the same jokes, reproduce the same paintings. In this world it's efficient to have protection of intellectual property to incentivize discovery of new objects.

However now suppose we teach a computer to recognize the properties of an object (which means learning the manifold), but it can also do the reverse, i.e. it can easily create a funny joke, or paint a pretty painting. In equilibrium there will now be much less imitation, and the efficient intellectual property regime will change. We shouldn't allow the first person (or algorithm) who is able to see the entire latent landscape to claim ownership of everything that they discover.

Application to Communication. I wrote a note last year with some prediction about how LLMs will affect communications, which I think is consistent with this manifold hypothesis.

1. For properties where human judgment is the ground truth (e.g. whether something is grammatical, is hate speech, is pornographic), then AI classifiers will achieve perfect accuracy, & this favors defense.

2. For properties which refer to some outside fact (whether a statement is true, whether an image depicts a real event, whether a painting is a forgery), then AI synthesis will degrade the ordinary human ability to make inferences from the content of the item, & so we will have to rely relatively more on signals of provenance.

Application to LLMs. We can think of pre-training an LLM as discovering the low-dimensional structure of text, and that low-dimensional structure includes all the knowledge expressed in the training text. Once you have learned how to transform a piece of high-dimensional text into a low-dimensional semantic representation then suddenly a lot of intractable problems become tractable. E.g. you can train a model on a few pairs of (question,answer) and get a pretty useful chatbot (AKA ChatGPT). This would be a wildly intractable problem without the low-dimensional representation.

There's another interesting observation about chatbots: they're mostly trained to give an answer that the user prefers. But why would you ask someone a question if you already know which answer is best? This is explicable either (1) if we train a chatbot to maximize the preferences of an expert, not the average person; (2) if it's easier for humans to recognize a good answer than to create one, either because of tacit knowledge or a computational constraint.

Application to wages. Here is a very stylized model which incorporates the manifold hypothesis. Suppose that comparative advantage across people is entirely due to their private knowledge: painters know how to paint, programmers know how to program, etc.. It's hard to justify this assumption in a purely rational model because you could just write down the information and sell it, but it makes more sense if knowledge is tacit and there's learning-by-doing.

Suppose now that LLMs can observe everyone's actions and extract their tacit knowledge. Intuitively, you now have an assistant in your pocket who can answer any question that you'd normally go to a domain expert for, and so the rents to expertise will collapse.

This can be formalized as a pure trade model, where each agent has a vector of productivities, and there's some global equilibrium price vector. The LLM makes private knowledge public, and so raises everyone's productivity at each task to the level of the world expert. This has the following implications:

1. Wages of the highest-paid fall, but aggregate output increases ("leveling up")

2. Exchange will fall (and so GDP may fall) because you can do most things yourself – e.g. you wouldn't call a doctor because you can diganose yourself.

3. The incentives to acquire new knowledge decline (insofar as an LLM can extract your knowledge by watching you work).

However it's notable that LLMs can beat most people on most knowledge-based questions, and yet they still have relatively minor productivity effects, so there's something missing from their set of capabilities, & that's still somewhat of an open question.

Types of Problems

General setup: you're given a question q Q, choose an answer a A, and get payoff y(q,a).
My claim is the *shape* of the function y(.,.) will determine everything.

There are two useful subtypes.

1. *Low-dimensional question* (landscape navigation). Suppose we face the same problem over and over but there are many possible distinct answers, then so there's an explore-exploit problem. More precisely, suppose $y(q, a)|q$ is a rugged function of a.
2. *Low-dimensional answer* (question-answering). Suppose we constantly get different questions but the answer is just a binary of scalar. This is like a classic supervised learning problem.

type	problem	question	answer	notes
questions	What digit is in this image (MNIST)?			
	What is the sum of X and Y?			
	What was last transaction by account XXX?			Many real-world record-keeping problems like this
	What chess move to play here?			
	Color pixel in Mandelbrot	low	low	
landscape				

Claims about capacities of human & computer brains:

- Classic computers are bad when there's high manifold curvature.
- Classic computers are good when there's

Alternative Setup

[TODO: add a diagram showing generating process, breadth and depth of generating process]

Two characteristics of prediction problems.

1. *Manifold dimensionality.* – high dimensionality if it's irreducibly complex, e.g. a telephone book full of numbers.
2. *Manifold curvature.* – high curvature if it's complicated to map the surface dimensions to the output – low curvature if it's all linear.

	low manifold dimension	high manifold dimension
low manifold curvature	- add two numbers	- telephone number from name - business database - directions from address
high manifold curvature	- encrypt data	- classify an image
	- play best chess move	- understand language

	low manifold dimension	high manifold dimension
	- prove theorem true/false - find optimum point - color a pixel in mandelbrot set - predict position of a star	- predict weather (chaotic system)

The learning rate depends on both dimensionality and curvature. A nonparametric estimator will be slow if either (A) there's high intrinsic dimensionality; or (B) the manifold is highly curved.

However if you *know* the shape of the manifold already, then you're only limited by the intrinsic dimensionality.

Computers are good at problems with low curvature. They can learn sets of facts and follow logical rules very well. They're extremely good at problems with *low curvature*.

Humans have progressively found lower-dimensional representations of many problems. E.g. Newton, Copernicus, Mendeleev, etc., all found much simpler latent representations of many problems.

They have discovered the curvature of the manifold, so apparently high-dimensional problems become low-dimensional.

The manifold hypothesis: many problems have low-dimensional representations. Problems which *appear* to have high dimension actually have low dimension.

Implication: it's valuable to do unsupervised or self-supervised pre-training, to figure out the manifold, and then you can do supervised learning on top of that.

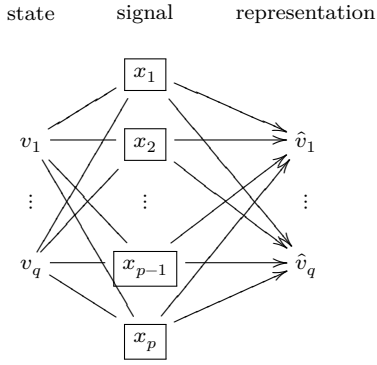
Economics literature on returns to experience.

- Deming (2024): more complex occupations have steeper returns to experience.
- Nedelkoska (2025) "[we] find that employees receive a positive wage premium to the complexity of their job and that workers in highly complex occupations acquire twice as much skills throughout life compared to less complex occupations."

Formal Setup [SKETCH]

[THESE ARE ALL ROUGH WORKING NOTES – DON'T TAKE TOO SERIOUSLY]

See `2023-10-26-models-of-ai-and-the-world.qmd`



Q = dimension of state	
P = dimension of signal	
N = number of observations	
$x_n \in \mathbb{R}^P$	observation (image, text, sound)
$v_n \in \mathbb{R}^Q$	state (objects, meaning, words)
$P \gg Q$	(signals have higher dimensionality than state)
$f : R^Q \to \mathbb{R}^P$	1:1 mapping between signal and state

signal (q-dimensional)	state (p-dime
image	objects,
text	meaning, styl
audio	speaker

Observations:

The space of signals will be sparse. Because $P > Q$ and the mapping is 1:1 most realizations of $x \in X$ would never be observed. E.g. most configurations of pixels are static, most configurations of words are gibberish: they don't have any interpretation at all (no v) and so you would almost never encounter them.[3]

[3] More realistically, you assume they are generated by some other process without the usual latent state involved.

Signals will be redundant. The state v over-determines the signal x, so if you remove some of the information from the signal then you can reconstruct them with very high confidence.

Technically the state-space has high dimensionality, but effectively it has low dimensionality. In some cases we think signals are *under-determined* by the state rather than over-determined, i.e. that $q > p$, e.g. (1) images are 2D projections of a 3D world, so the signal seems lower-dimensional than the state; (2) a given sentence is often ambiguous between many possible meanings, and the disambiguatin is done by context, implying the set of sentences is lower-dimensional than the set of meanings.

In principle the state has very high dimensionality, higher than the signal, but in practice we have such strong priors about the state that its effective dimensionality is lower than the signal space. A full representation of the state requires a dimensionality $\bar{q} \gg p$, but in practice the state has such strong regularities that the majority of the variation requires a much lower-dimensional representation q, so we have:

$$\underbrace{\bar{q}}_{\substack{\text{dimensionality} \\ \text{of world}}} \gg \underbrace{p}_{\substack{\text{dimensionality} \\ \text{of signal}}} \gg \underbrace{q}_{\substack{\text{dimensionality} \\ \text{of state}}}$$

Signals are highly compressible. This model implies that you can compress signals from a P-dimensional object down to a Q-dimensional object.

Computers can learn the state but it takes a lot of data. We can model computers as slowly learning how to transform between v and x but it requires an enormous amount of training data. We talk about the computer's problem as a PCA problem below.

Extension: humans find it easier to decode than encode. Humans have an asymmetry about some things: they tend to be better at decoding than encoding, e.g. we can recognize whether a picture looks like Rishi Sunak but we can't draw a picture that looks like him. We can model this as humans being composed of two agents, conscious and pre-conscious: the pre-conscious brain has private information which it uses to calculate $\hat{v} = E[v|x]$, and the conscious brain only observe the posteriors from the pre-conscious brain (nice analogy: a person with a sniffer dog).

Problem: No Closed-Form Solution

We want a model where you observe a matrix of p features for n cases, which are generated from some lower-dimensional representation. There are two problems:

1. Doing the inversion – you infer the low-dimensional state for each case. This is straight-forward (I think).
2. Learning the inversion – finding an optimal low-dimensional decomposition. It's not clear to me whether we can get analytic solutions. Udell says *"low rank approximation problems are not convex, and in general cannot be solved globally and efficiently."*

Model 1: PCA

$$\underbrace{\begin{bmatrix} h_1^1 \dots h_p^1 \\ \ddots \\ h_1^n \dots h_p^n \end{bmatrix}}_{\substack{\text{observed dataset} \\ n \text{ cases and } p \text{ features}}} = \underbrace{\begin{bmatrix} w_1^1 \dots w_q^1 \\ \ddots \\ w_1^n \dots w_p^n \end{bmatrix}}_{\substack{p \text{ loadings} \\ \text{on } q \text{ factors}}} \underbrace{\begin{bmatrix} x_1^1 \dots x_q^1 \\ \ddots \\ x_1^n \dots x_q^n \end{bmatrix}}_{\substack{n \text{ latent vectors} \\ \text{on } q \text{ factors}}}$$

Bengio et al. write it as:

$$h = W^T x + b$$

and give a probabilistic interpration ("PCA can be given a natural probabilistic interpretation (Roweis, 1997; Tipping and Bishop, 1999) as factor analysis:")

$$p(h) = N(h; 0, \sigma_h^2 I)$$
$$p(x|h) = N(x; Wh + \mu_x, \sigma_x^2 I)$$

Tipping and Bishop (1999) write:

$$t = Wx + u$$

and they show that if you assume Normal distribution of x and u then you can find a weight matrix W which maximizes likelihood. They do *not* assume a prior over the weights themselves (p614).

PCA with a single factor (gymnastics)

If there's a single factor plus noise then we can write everything like this:

$$\underbrace{\begin{bmatrix} h_1^1 \dots h_P^1 \\ \ddots \\ h_1^N \dots h_P^N \end{bmatrix}}_{\substack{\text{observed dataset} \\ N \text{ cases and } P \text{ features}}} = \underbrace{\begin{bmatrix} x^1 \\ \vdots \\ x^N \end{bmatrix}}_{\substack{\text{latent value} \\ \text{for each case}}} \underbrace{\begin{bmatrix} w_1 & \dots & w_P \end{bmatrix}}_{\substack{\text{weight for} \\ \text{each feature}}} + \begin{bmatrix} \varepsilon_1^1 \dots \varepsilon_P^1 \\ \ddots \\ \varepsilon_1^N \dots \varepsilon_P^N \end{bmatrix}$$

We can also write it out like this:

$$\underbrace{h_{n,p}}_{\substack{\text{feature } p \text{ of} \\ \text{case } n}} = \underbrace{x_n}_{\substack{\text{avg score} \\ \text{of this case}}} \times \underbrace{w_p}_{\substack{\text{avg score} \\ \text{of this feature}}} + \underbrace{e_{n,p}}_{\text{noise}}$$

Answer when $P = 1$. Suppose we have a single feature and there's no noise. So there's many contestants and just one judge. Then we have this:

$$\underbrace{h_n}_{\substack{\text{observed score} \\ \text{of each case}}} = \underbrace{x_n}_{\substack{\text{true} \\ \text{value}}} + \underbrace{w}_{\substack{\text{common} \\ \text{noise}}}$$

Assume we have mean-zero Gaussian priors over all RHS variables, then we can write:

$$\begin{aligned}
E[x_n|h_n] &= \frac{\sigma_x^2}{\sigma_x^2 + \sigma_w^2} h_n \\
E[w|h_n] &= \frac{\sigma_w^2}{\sigma_x^2 + \sigma_w^2} h_n \\
E[w|h] &= \frac{\sigma_w^2}{\sigma_w^2 + \sigma_x^2/N} \frac{1}{N} \sum_{m=1}^{N} h_m \\
E[x_n|h] &= h_n - E[w|h] \\
&= h_n - \frac{\sigma_w^2}{\sigma_w^2 + \sigma_x^2/N} \bar{h}
\end{aligned}$$

Implications:

- when $N = 1$ then $\hat{x}_n = \frac{\sigma_x^2}{\sigma_w^2 + \sigma_x^2} h_n$
- when $N = 2$ then $\hat{x}_1 = \frac{\sigma_w^2/2 + \sigma_x^2/2}{\sigma_w^2 + \sigma_x^2/2} h_1 - \frac{\sigma_x^2/2}{\sigma_w^2 + \sigma_x^2/2} h_2$
- when $N \to \infty$ then $E[x_n|h] \simeq h_n - \bar{h}$.

Old answer (but I think it's wrong). Suppose we have priors over the RHS variables that are mean-zero with known variances. Then *I think* our estimate will be as follows, but I need to confirm:

$$\begin{aligned}
\hat{x}_n = E[x_n|h] &= \frac{n\sigma_a^2}{n\sigma_a^2 + \sigma_e^2} \left(\frac{1}{P} \sum_{p=1}^{P} h_{n,p} - \frac{1}{NP} \sum_{m=1}^{N} \sum_{p=1}^{P} h_{m,p} \right) \\
&= \frac{n\sigma_a^2}{n\sigma_a^2 + \sigma_e^2} \left(\bar{h}_{n\cdot} - \bar{h}_{\cdot\cdot} \right)
\end{aligned}$$

- Suppose there's no noise: then we learn relative row values and relative column values exactly, but we're missing overall calibration. Our posterior is just the relative row value.

Variance of the posteriors [UNFINISHED]. We now derive the variance of the posteriors. We'll start by assuming no noise:

$$\begin{aligned}
\hat{x}_n &= \frac{1}{P} \sum_{p=1}^{P} h_{n,p} - \frac{1}{NP} \sum_{m=1}^{N} \sum_{p=1}^{P} h_{m,p} \\
&= \\
V[\hat{x}_n - x_n] &= \sigma_N + \frac{1}{P^2} \sigma_P^2
\end{aligned}$$

(For derivation see `2024-07-20` note.)

ChatGPT prompt.

> I want to find an analytically tractable expression for a model with dimensionality reduction. Suppose we have H=x'w+e, so we have $h_{n,p} = x_n + w_p + e_{n,p}$

Suppose we observe the cells of this 2x2 matrix, which is formed by adding rows and columns:

$$\begin{bmatrix} x_1 + y_1 & x_2 + y_1 \\ x_2 + y_1 & x_2 + y_2 \end{bmatrix}$$

We have mean-zero Gaussian priors over x and y, can you write an expression for the posterior

Model 2: Questions and Answers

This is the model I used in my *imitation* note. We observe a set of n questions and answers. Each question is a vector of p attributes, and the answer is the weighted sum of those attributes, but we need to infer the weights.

$$\underbrace{\begin{bmatrix} a^1 \\ \vdots \\ a^n \end{bmatrix}}_{\text{answers}} = \underbrace{\begin{bmatrix} q_1^1 + ... \, q_p^1 \\ \vdots \\ q_1^n + ... \, q_p^n \end{bmatrix}}_{\substack{\text{multi-attribute} \\ \text{questions}}} \underbrace{\begin{bmatrix} w^1 \\ \vdots \\ w^p \end{bmatrix}}_{\text{weights}}$$

$$\underset{\substack{n \times 1 \\ \text{observed}}}{\underbrace{a}} = \underset{\substack{n \times p \\ \text{observed}}}{\underbrace{Q}} \cdot \underset{\substack{p \times 1 \\ \text{unobserved}}}{\underbrace{w}}$$

Then we have a simple expression for the posterior:

$$\hat{w} = E[w|Q, a] = Q'(QQ')^{-1}a$$

Misc

- Another model: multidimensional signal, multidimensional state, and a mapping matrix:

$$\underset{\substack{P \times 1 \\ \text{signal}}}{\underbrace{x_n}} = \underset{\substack{P \times Q \\ \text{mapping}}}{\underbrace{A}} \cdot \underset{\substack{Q \times 1 \\ \text{state}}}{\underbrace{v_n}} + \underset{\substack{P \times 1 \\ \text{noise}}}{\underbrace{e_n}}$$

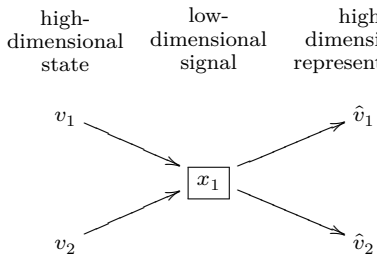
- Related: solving a least-squares model when over-determined or under-determined - tweet, Wikipedia

Applications

Summary:

In many fields we model problems as $n > p$ and $q > p$.

- Statistics: $n > p$
- Economics: $n = \infty$, $q > p$.
- Signal extraction: $n = \infty$, $q > p$.
- Perception: (…).



Application to Communication

Application to internal properties. An internal property of a signal is one that depends entirely on the content of the signal itself, not on anything outside the signal: e.g. whether a text is hate speech, whether a photo contains nudity, whether a song is catchy.

1. **Baseline: human judgment.** ordinarily humans can immediately judge $\hat{v}(x)$ and using $\hat{v}$ tell whether it has a given internal property.
2. **Computer with small m:** When m is small the computer learns only a very crude approximation of A. In practice we give the computer labelled data, $\hat{v}_1$, and we give the computer a very small set of features x_1, x_2, and the computer runs a simple regression, $\hat{v}_1 \sim x_1 + x_2$.
3. **Evasion of computer:** If the human knows the computer's algorithm, $\hat{v}_1(x_1, x_2)$, then it's trival to get around it: even if the human doesn't know $x \to v$ perfectly, still they can just fiddle with x_1 and x_2, e.g. mispelling the trigger words, or changing the colour of an image.
4. **Computer with large m:** Now the computer learns A perfectly, they have human-level performance, and it's impossible to evade it.

Application to external properties. An external property depends on something outside the signal: e.g. whether a photo depicts something that actually happened, whether a poem was written by Shakespeare. The content of the signal can be informative but it's not definitive.

1. **Baseline: signals all created by the world.** Consider recorded media (photo, audio), and suppose that they are only created by events that actually happened. Thus encoding, $v \to x$, is done by the laws of physics, and decoding is done by the human brain.
2. **Manipulation by humans.** Suppose now a human wants to modify or synthesize a signal. It's hard! Humans can automatically convert $x \to \hat{v}$, but that is done by a pre-conscious part of the brain, so they don't know how to tweak x to change $\hat{v}$. They could make random changes and see what happens to $\hat{v}$ but this is extremely slow. In addition, because the x-space is sparse, the receiver would recover a $\hat{v}$ which has extremely low probability, and they would infer that the signal has been tampered with (in practice: the photo would have weird inconsistencies, or the audio would be clipped).
3. **Computer with small m.** A computer with a small m might be able to do crappy classification ($x \to v$), but it wouldn't be very robust: e.g. it would learn that images with yellow backgrounds are typically of camels. You could reverse the algorithm to produce an x which maximizes $\hat{v}$ but it would be very ugly & obviously fake.
4. **Computer with large m.** Now suppose the computer can perfect convert $x \leftrightarrows v$: they can synthesize an arbitrary image, & it's impossible to discriminate from a real image. Suppose receivers are naive, they think that if they observe x, then $\hat{v}(x)$ really happened. Then strategic senders can arbitrarily manipulate their beliefs, e.g. creating photos of politicians doing scandalous things.
5. **Equilibrium.**
    1. If all senders are strategic then I think you get a babbling equilibrium: you now learn nothing from x.
    2. If some senders are strategic then there will be some non-zero persuasive power of media. In the medium run you'd expect more entry by strategic senders until the returns to creating media go to zero: you could write equilibrium with the share of fakes pinned down by the intersection of two curves: (1) creation of fake media as a function of credence; (2) credence in media as a function of prevalence of fakes.

3. Platforms might pay some cost c to check the veracity of some media, when the probability of being fake exceeds a threshold. This would put a ceiling on the influence of fake media.

Application to Intellectual Property

(see `AI and intellectual property`, I think it can be put in this framework)

Applications to Economics

Most models of inference in economics have the agent receives a signal that is not fully revealing of the state because it has lower cardinality than the state (e.g. a binary signal), or because there is noise (then you can think of the noise as part of the state, and so the state is higher-dimensional than the signal). E.g.:

- Estimating the productivity of an employee from their education.
- Estimating demand conditions from sales.
- Estimating the competence of a politician from economic conditions.
- Estimating the quality of a product from peer usage of that product.

Yet in practical situations information sets often *do not* seem to be discrete or univariate, instead they're enormously rich: we have the newspaper, we have millions of datapoints about employees, we have the internet. Let's try to reinterpret these situations:

- Estimating the productivity of an employee from meeting them and watching them work.
- Estimating demand conditions from reading the newspaper and trying to infer the state of the world.
- Estimating the competence of a politician from their speeches, their mannerisms, how journalists and other politicians talk about them.
- Estimating the quality of a product from the label.

Application to Recommender Algorithms

1. **We want to predict a person's response to an item.** The item could be a post, a song, a video, an advertisement, a product. The person's response could be clicking, purchasing, upvoting, or labelling as toxic.
2. **Model: low-dimensional semantics.** I will state a simple model of human judgment and then try to describe some of the facts we observe about recommendation. Assume that the content of each item (pixels, characters, etc.) can be represented in a low-dimensional space, the "semantics." Each person's response is fully determined by the semantics, though different people have different weights.
3. **World without computers.** There are many practical decisions people make about others' judgments: will people like this concerto? how many copies is this book likely to sell? is this painting obscene? Then we can use a mixture of our own judgments and simple statistics about others judgments – e.g. what are the best-selling books, what are the most-beloved paintings.
4. **Very small computers.** Suppose our computers can digitize data about preference but not the full content of the items themselves. Then we can do collaborative filtering to predict preference, irrespective of the content.

5. **Small computers.** Suppose we now have access to the full digital representation of each piece of content. Somewhat surprisingly this information is not very useful: the relationship between surface-level features and preference is pretty noisy. Shallow features are somewhat informative: e.g. the tempo of a song, the proportions of a painting, can be predictive of preference, but these features don't have much predictive value relative to collaborative filtering.
6. **Large computers.** Now suppose we have computers that are *large* such that they can extract the semantics of each item with high accuracy.
    - *Implication:* the content discovery problem (AKA cold start) is solved.
    - *Implication:* we will start synthesizing content. (However a classifier that has high accuracy on its training set could perform badly off-distribution, and thus do badly in synthesizing content).

Suppose our raw data looks like this, where we have observations $j \in 1 \dots J$:

x^j	high dimensional input
$i^j \in N$	identity of rater
$R^j \in R$	rating

Search engines are similar: History of search engines:

1. Text match between query and document.
2. Pagerank for the quality of a document.
3. Predict click-through rate and satisfaction rate.
4. Semantic match.

Application to Statistics

Applied to statistics. The canonical statistical inference problem is where the number of observations is larger than the number of features ($n > p$), e.g. linear regression is only well-defined when $n > p$.

Application to Chatbots

If you train on sufficiently many held-out words then you eventually learn the mapping from x to v.

Now you can post-train on question/answer pairs (SFT) or on human preferences (RLHF), and you'll get very accurate very quickly.

Related Literature

Tipping and Bishop (1999) "Probabilistic Principal Component Analysis" They show that "the maximum likelihood estimators for the isotropic error model … do correspond to principal component analysis."

Factor analysis differs from PCA in having a separate noise term:

$$t = Wx + \mu + \epsilon$$

"The model parameters may be deremined by maximum likelihood, although … there is no closed form analytic solution for W and Ψ