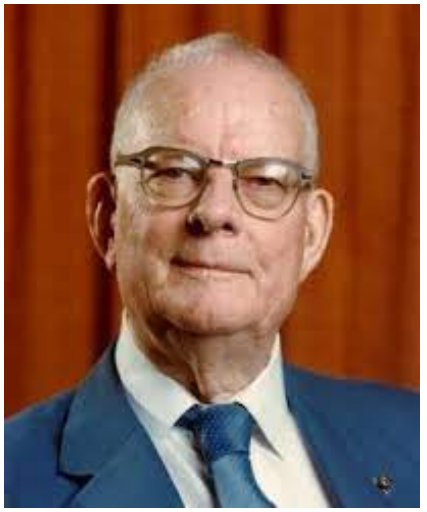
Statistics in the Knowledge Economy

David Banks
Duke University

1. Introduction

William Edwards Deming had a new vision for industry. He understood that poor quality costs money, and that corporate management was not magically brilliant.



He used experimental design, process control, and sampling theory to change the world with statistics. His work made the world richer.

He proved the power of statistical thinking.

I think the field of statistics stands at the threshold of a new vision for industry, and I hope that many of us shall step forward to embrace the change.

In my career, I've seen 2.5 theoretical breakthroughs:

- Brad Efron's bootstrap
- Gelfand and Smith's Markov chain Monte Carlo
- phase transitions in large p, small n regression.

I don't expect to see another theoretical advance of comparable magnitude.

But I do hope we shall see a statistical surge in **data engineering**.

If one uses statistics to infer the existence of the Higgs boson or the presence of an extra-solar planet, one is doing data science.

But if one is taking a lot of data and using smart algorithms and statistical perspectives and optimization to enable Uber to position its drivers 1% better, then one is doing data engineering.

At this level of detail, there are no general theorems. Every application requires a bespoke solution.

Data engineering is creating new businesses and new services. It is essential to Google maps, ride-sharing services, recommender systems, TikTok, Vivino, YouTube, insurance, and everything Amazon. It is essential to the information economy.

Information technology is the future of industry. And it is a target rich environment for statisticians.

Some key new challenges include:

- **Computational advertising.**
- **Autonomous vehicles.**
- **Large language models.**
- Optimal control of manufacturing processes.
- Finance industries.

There is much more, and our PhD students are flocking to those industries and application areas.

2. Computational Advertising

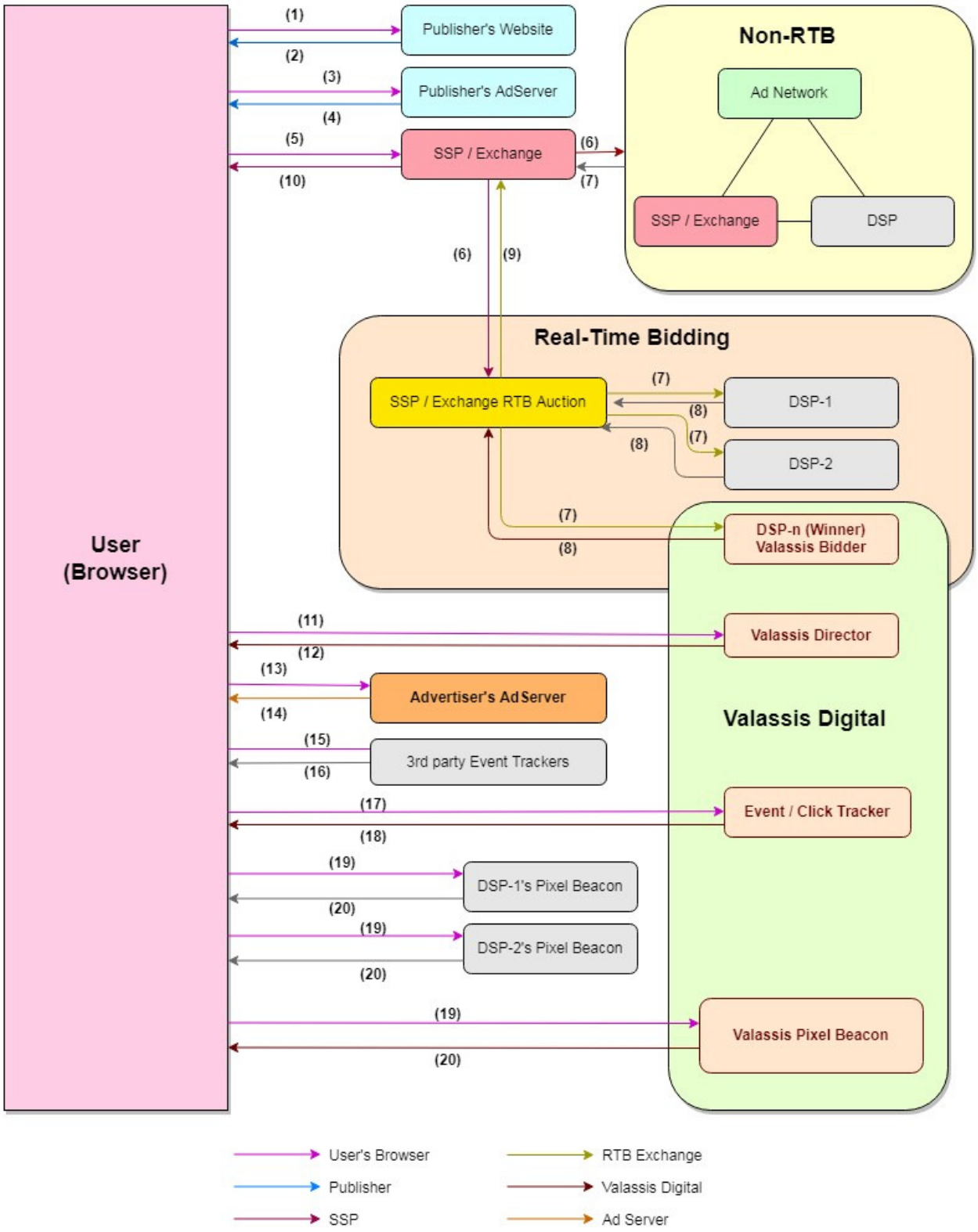
Computational advertising is an emerging field.

- It is a $309 billion industry.
- The ecology of ad buy and auctions is complex, and companies with better information will make more profit.
- It uses Deming's tools of experimental design, process control and sampling, as well as engaging with many other areas of statistics (e.g., causal inference, predictive inference, spatio-temporal modeling, networks).
- It poses new and important research problems, especially in connection with recommender systems, which have wider applications.
- Statisticians can make huge contributions.

2.1 Real-Time Bidding

Computational advertising (CA) is a young, fast-moving field. The online advertising market is projected to be valued at $982 billion by 2025. It is the dominant revenue stream for many major IT companies.

One component of CA is on-line ad clicks. When you type "pizza" into their browser, it triggers a virtual auction that lasts a few milliseconds. Domino's and Papa John's and Pizza Hut bid for your eyeballs. The highest "qualified" bids are displayed, with the highest bidder getting the top position. But the process is actually much more complex.


User
(Browser)
Publisher's Website
Publisher's AdServer
SSP / Exchange
Non-RTB
Ad Network
SSP / Exchange
DSP
Real-Time Bidding
SSP / Exchange RTB Auction
DSP-1
DSP-2
DSP-n (Winner)
Valassis Bidder
Valassis Director
Valassis Digital
Advertiser's AdServer
3rd party Event Trackers
Event / Click Tracker
DSP-1's Pixel Beacon
DSP-2's Pixel Beacon
Valassis Pixel Beacon
(1)
(2)
(3)
(4)
(5)
(6)
(7)
(8)
(9)
(10)
(11)
(12)
(13)
(14)
(15)
(16)
(17)
(18)
(19)
(20)
User's Browser
Publisher
SSP
RTB Exchange
Valassis Digital
Ad Server

1. A person's browser contacts the publisher's website (e.g., CNN.com).
2. The publisher's website sends back content including placements that will need to be fulfilled though an AdServer.
3. The browser contacts the publisher's AdServer to fulfill placements that will not go up for auction. If the user has an in-app or a browser ad blocker this interchange will not happen.
4. The publisher's AdServer sends back predefined ad content.
5. For other placements the browser will contact an Exchange with placement information and an indication on whether the placement should go out for bid, or if another private (guaranteed) deal has been setup for the placement.

6. If the placement is marked for real-time bidding (RTB), an auction is set up by the Exchange.

7. Once the auction is initiated the demand side platforms (DSPs) are simultaneously contacted to participate in the auction. This all runs in parallel.

8. If a DSP decides to bid its bid offer and a director tag (for tracking) are returned to the auction.

9. The winning bid from the auction and information on the bidder, the DSP, is sent to the Exchange

10. The Exchange returns the DSP wrapped ad tag that contains everything needed to track the bid, impression, and director log joins. It's essentially a link to the director so one can obtain user information and pass back the actual creative ad tag.

11. The browser contacts the DSP's director to obtain the AdServer tag, which is a link to the advertising agency for (one of) the company's that won the bid.

12. The browser receives the AdServer tag.

13. The browser then uses the AdServer tag to contact the advertiser's AdServer requesting the impression.

14. The advertiser's AdServer returns all of the impression assets and creative content back to the browser.

15. If the campaign/line item is using a third party for tracking it is contacted with the impression and user information. This service is now almost always used.

16. The third party tracker sends back a 1×1 pixel as a verification or handshake token.

17. The ad tag (sent in #10) contains the javascript that the browser/app will load when the impression renders. These signals are passed back ONLY AFTER the creative content is loaded which is why these actions/behaviors are so far down the chain.

18. The DSP sends back a 1×1 pixel as a verification/handshake.

19. If the DSPs are tracking pixel fires this information is sent from the browser.

Even this is an oversimplification. The economic ecology of ecommerce is evolving. For example, it used to be that nearly all auctions were second-price; now they are nearly all first-price auctions.

CA touches on many aspects of statistics. Important topics include design of experiments, causal inference, recommender systems, predictive inference, and time series modeling.

But CA can also engage with text and sentiment analysis, dynamic network analysis, probabilistic ad contracting, spatio-temporal processes, censored data analysis, and many other statistical fields.

This is a good opportunity for academic research, since most companies that work in this space are not interested in proving theorems, but rather eking out an extra half percent of profit or putting out urgent fires. And corporate data scientists are generally not encouraged to publish.

2.2 Recommender Systems

Recommender systems are the workhorses of CA. They help Demand Side Platforms decide which ads to bid upon for display. They make movie, book, music, and dating recommendations.

A key aspect of CA is that it is data engineering, not data science. Each application requires a bespoke solution. There are general principles which can help one get started, but the ultimate solution will need to be hand-fitted.

A new research challenge is to develop methodologies that implement such tuning.

Consider the two starting points for recommender systems: collaborative filtering and content based filtering.

Both approaches start with a (very sparse) "ratings matrix" $\boldsymbol{R}$ whose rows are users, whose columns are items, and whose entries are ratings that a user has assigned to an item.

To make a recommendation for the ith user, collaborative filtering seeks other users whose tastes are like those of the ith user, and bases its recommendations on what those others liked.

Content based filtering makes recommendations based on features of the item. Its three steps are extracting item features, learning user preferences, and recommending items which fit the user's preferences.

A particularly fun application is active recommender systems. If someone asks for a movie or book recommendation, a human typically asks them about other movies or books that they like, and makes recommendations accordingly.

This is rather like playing 20 Questions but with special features:

- personalized priors
- complexity constraints
- non-standard feature selection.

In principle, one would build a proximity matrix for books or movies or music.

The new chatGPT could be transformative, if connected to the right statistical model for user preferences.

Hypothetically, Amazon could calculate a distribution over the probability of me buying any book they have on offer. This distribution would be based upon previous purchases I have made, and some model for "nearness" in book space.

That model for nearness should be complex and personally tailored. Some people follow authors, others follow genres, others follow the New York Times book review section.

Probably the model for nearness is non-Euclidean, and so one might use isomap or paramap.

Next, Amazon needs to learn what questions to ask that will let it learn the most about the book(s) it will recommend.

Unfortunately, the best questions it should ask are things like "On the whole, do you like these 100 books more than this other list of 150 books?" And that is impossible for someone to cognitively process.

Therefore there needs to be a complexity penalty on the questions that are asked. Defining such a penalty is an area for research, but it can be viewed as a kind of statistical regularization.

Ideally, the questions should be ones for which Amazon's prior gives a 50-50 chance of me responding "yes". That means Amazon will learn at the fastest possible rate.

But there is a hidden optimization problem. The first question Amazon asks might have a good 50-50 split, but, depending on the answer, subsequent questions might mostly be 90-10 splits.

Therefore, to optimize, one seeks a "question tree" that has lots of near 50-50 questions in the follow-ups. So an initial question with a 60-40 split that has lots of subsequent 60-40 questions would be better.

This is a general class of problems and I am not aware of any previous literature that addresses such cases. But optimal learning, under various complexity, memory, and computational constraints, should be an interesting new area of study.

3. Autonomous Vehicles

Driverless vehicles will change the world. If they are networked, the fuel and safety benefits are especially valuable.

In terms of industry, they will revolutionize how goods are moved, with significant effects on insurance, just-in-time manufacture, and logistics. It will spawn new kinds of business and enable the seamless integration of multimodal transport.

In terms of people, they should improve safety, reduce pollution, solve congestion, and change lives.

Some facts:

- Americans drove 3.4 trillion miles in 2018. US DOT
- Risk of dying in a vehicle injury are 1/77. Contrast this to firearms (also 1/77), falls (1/83), suicide (1/63), heart disease (1/4), alcohol and drugs (1/34). CDC
- Motor vehicles account for 75% of carbon monoxide pollution, 1/3 of all air pollution, and 27% of greenhouse gases in the US. EPA
- Average commute time is 52.2 minutes/day. US Census Bureau

Autonomous vehicles can improve all these numbers.

The potential gains from autonomous vehicles include:

- Safety: AI driving systems are not distracted or impatient, and have better sensors.
- The environment: Better safety means lighter vehicles. Joint control is ideal—little braking is needed.
- Congestion: Under joint control, one can have seven times as many vehicles on the road.
- Quality of life: Commuting would become work time or nap time.
- Independence: Seniors and children would have more mobility.
- Economic: The efficiency of transportation for goods and people would improve, lowering costs and creating new markets.

The **2050 Problem** refers to the fact that in 26 years, the world population will reach its maximum (9.8 to 10.2 billion). We are currently at 8.1 billion, and the carrying capacity of the planet is about 1 billion.

Global warming is harder to forecast, but climate scientists say that in 2050, parts of North Africa, the Middle East, India and South Asia will regularly experience summer temperatures between 120 and 125 degrees.

Autonomous vehicles are one of the few technologies on the horizon that have the potential to meaningfully reduce carbon emissions while maintaining relatively high standards of living.

But there are many legitimate concerns about moving to autonomous vehicles. Some people worry that:

- People won't want to give up control.
- The "mixed fleet" period will be suboptimal.
- The regulatory and insurance implications have not been thought through yet.
- Cybersecurity—if the vehicle's software can be hacked, then there is a single point of failure.
- It would cause economic disruption.

All but the first have some interaction with statistical methods.

There are six levels of vehicle automation:

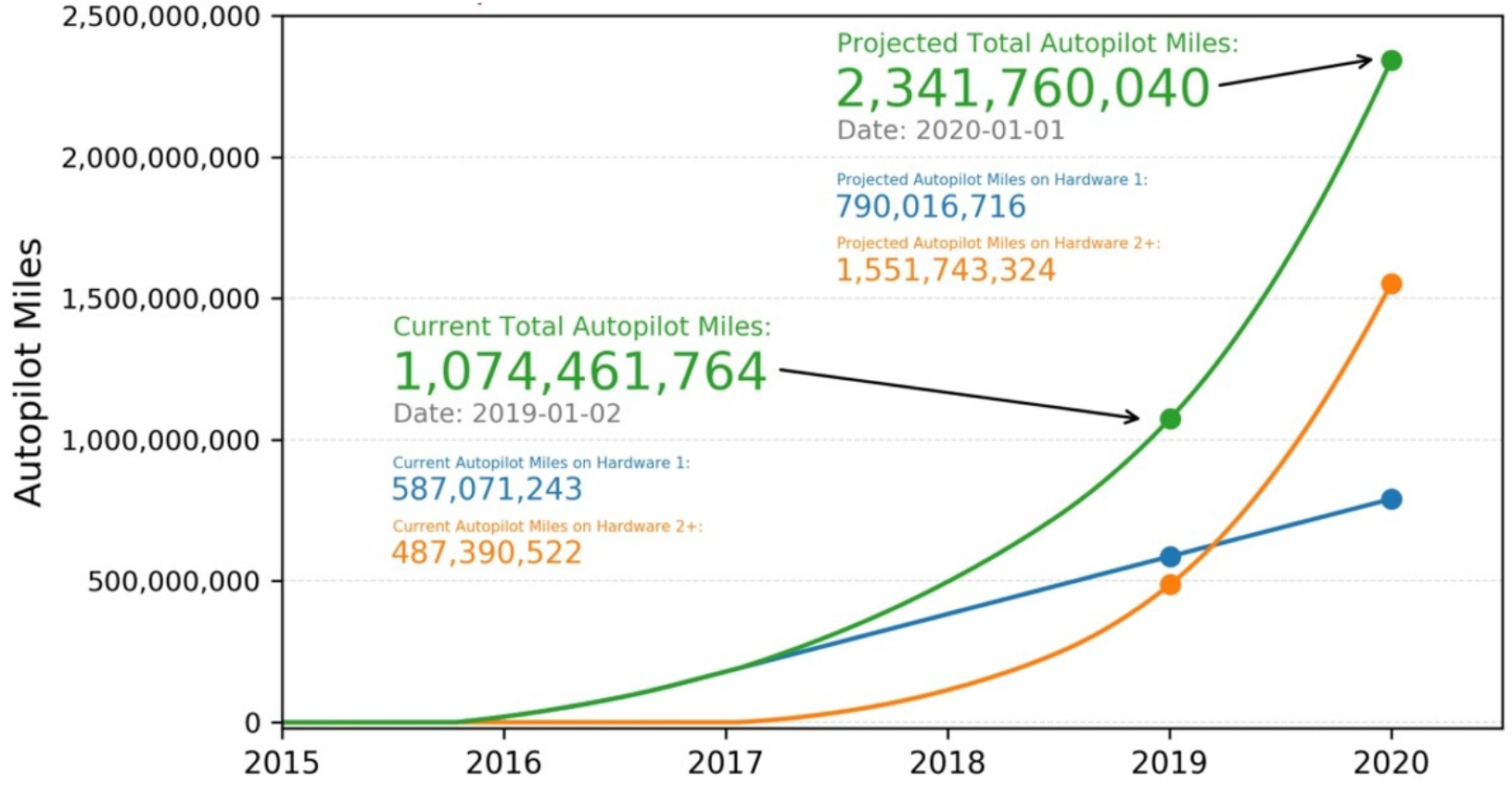
1. **No automation.** The human has only standard assistance (mirrors, rear-view cameras).
2. **Software assistance.** Adaptive cruise control, lane keep assist. Widely available after 2018.
3. **Partial automation.** The driver must be ready to take control, but the car controls speed and holds its lane. Tesla Autopilot.
4. **Conditional automation.** Hands off the wheel, but still ready to control. Useful for limited access highways and good driving conditions. Experimental.
5. **High automation.** Driver can sleep after inputting destination. Waymo is testing such. Must stay on traditional roads.
6. **Full automation.** Years away.

Some History. Autonomous vehicles are staples in science fiction, but Red Whittaker at Carnegie Mellon went a long way towards making robot cars a reality. In 1995, he programmed a small truck, Navlab 5, that drove from Pittsburgh to San Diego; 98% of the journey was autonomous. He also built robots for antarctic exploration, clean-up of Three Mile Island and Chernobyl, and mapping mines.

Sebastian Thrun worked with Whittaker at CMU. He led the development of Google's self-driving car.

Waymo is owned by Alphabet, and was spun off from Google. It runs a commercial fleet of level 5 vehicles in Phoenix. Volvo, Tesla and Audi are also testing.

29 states have passed laws permitting autonomous vehicles.



The figure shows results for level 4 and 5 autonomous vehicles: conditional control and high automation.

In the US, there are 1.18 fatalities per 10^8 human driven miles.

There have been four fatalities with level 3 autonomous vehicles (Tesla), one with a level 4 vehicle (Uber) and none with level 5 or 6 vehicles. The number of miles driven by level 3 or higher autonomous vehicle is about 10^9.

If autonomous vehicles drove as safely as humans, one would expect 12 deaths rather than 5.

Autonomous vehicles appear to be safer, but there are still concerns about weather conditions and other driving conditions.

All the autonomous vehicle developers use deep learning to train their vehicles. For example, the Tesla Autopilot Hardware v2+ uses NVIDIA Drive PX 2 hardware, 8 camera input, and Inception 1 architecture to train a convolutional neural network.

Training requires a lot of training data, and even then there are problems. Putting a post-it note on a stop sign fooled the AI into classifying it as a billboard.

From a statistical standpoint, there are three areas of contribution.

- Performing a continual risk analysis of safety, both for the changing mixed fleet scenario and the case of unitary networked control.
- Validation of the deep learning training of the AI system that controls vehicle operation, including protection against adversarial perturbation.
- The onboard software will need regular updates, so software quality control will be required. Statisticians have worked on software quality before, but this application has novel features.

Obviously, these topics will entail partnerships with other stakeholders, such as computer scientists, transportation engineers, and various regulatory agencies.

4. Large Language Models

At the beginning of the 20th century, a group of statisticians, who called themselves psychometricians, invented personality inventories, IQ tests, depression scales, and other measures of human cognition. We now need to repeat that for the 85+ large language models (LLMs) that are currently under development.

The ability to measure and characterize LLMs will drive their evolution and capabilities. Currently, most LLMs simply generate the next word in a sequence or pixel in an array. But that functionality is being built out quickly.

The field is moving swiftly:

- Claudia Shi at Columbia is studying the moral sense of about 45 LLMs.
- I and some collaborators have been studying bias in GPT-4.
- Adding a recommender system to an LLM will enable it to become, say, a personalized tutor or an expert personal assistant.
- Giving an LLM the ability to access the Internet, use a calculator, a currency converter, or a time zone converter will greatly expand its capability.
- Using DALL-E will enable a grandfather to write a graphic novel for his granddaughter.
- The main dangers I see in LLMs are increased cybercrime and identity theft.

5. Conclusions

Dr. Deming's day is done, and as the beneficiaries of his legacy we need to tool up to address analogous problems in the modern business world.

Statisticians have much to contribute. But to honor his tradition of blunt speech, I emphasize that our MS and PhD programs are not doing a great job of producing graduates with the relevant skill sets. For example, we need to teach Spark and PyTorch or TensorFlow.

Our profession has been slow to embrace Big Data or deep learning or even data science. We need to withdraw from general theory and learn how to solve one-off problems well.