Energy Demand & Utilization: Midterm Brief

By: Anthony Rizkallah

Date: 02/16/2025

Executive Summary

Artificial Intelligence (AI) is set to become a dominant driver of energy demand, particularly through the expansion of data centers. This brief projects AI-related energy consumption in the U.S. over the next decade, leveraging a lognormal growth model to estimate future data processing needs and associated electricity consumption. Through the analysis, I find that AI-driven data centers could consume between 232TWh (low scenario) and 3,527TWh (extreme scenario)[1] over the next decade as shown in Figure (1). The extreme scenario is forecasted as drastic growth (95th percentile) from the historical growth rate of consumed and created data which has increased at a 150% rate between 2010 and 2011 and reached 56% in the year 2020. The low scenario uses the 5th percentile (14% CAGR) of data growth over the next decade.

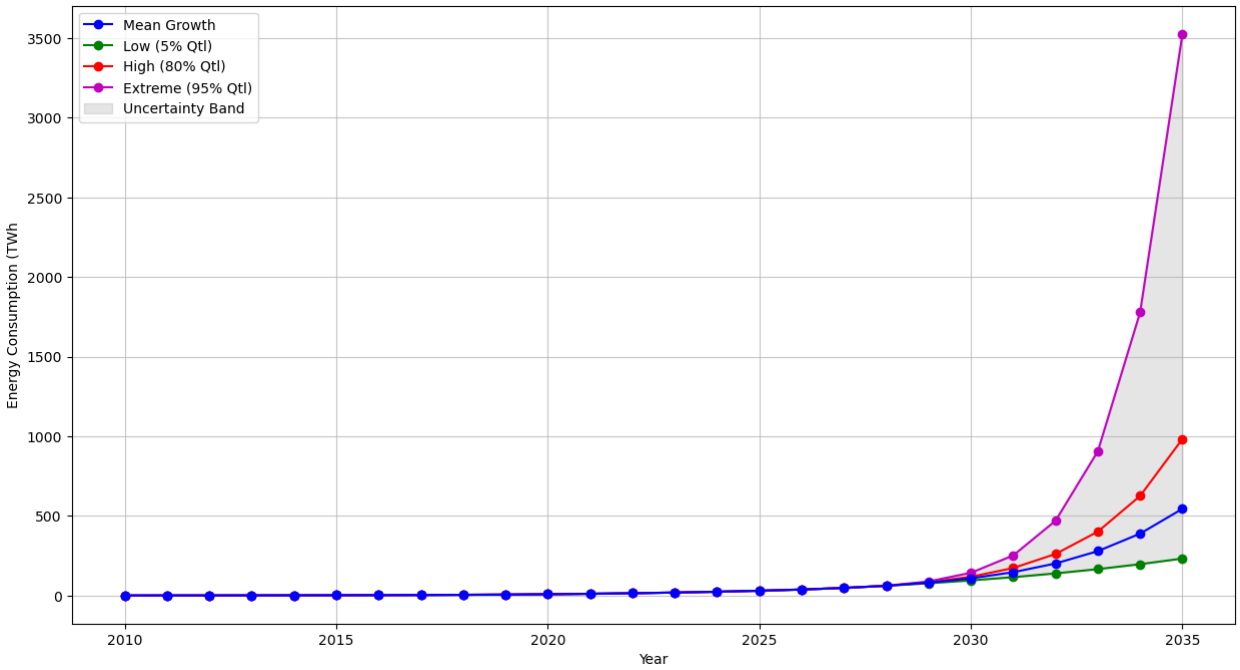

Figure 1: Data Center Energy Consumption (TWh)

Additionally, I cover the underlying uncertainty in the estimates, as well as the potential fleet that could supply that growth under time, technical, and budget constraints. I find that natural gas is the dominant resource due to cost and time efficiency when compared with other sources. However, renewable energy, particularly, photovoltaic and nuclear, could be feasible solutions under environmental and grid constraints, but those were not evident in the optimization model.

[1] Predicted using the highest historical growth rate of data consumption and creation

Introduction: Motivation & Background

The global AI revolution is not just a technological shift but a geopolitical and economic force shaping the 21st century. Nations are racing to dominate AI, with the U.S. and China locked in a fierce competition for supremacy. AI breakthroughs offer immense power- military applications, economic dominance, and unprecedented levels of automation. The United States, long the leader in semiconductor design, faces growing pressure from China's aggressive push to develop its own AI ecosystem. The ongoing U.S.-China trade war over semiconductors and critical technologies has further intensified, with Washington restricting Beijing's access to high-end AI chips, attempting to maintain its strategic advantage (more so today given a Trump Administration). Meanwhile, China is investing heavily in indigenous AI research and computing infrastructure to reduce dependence on Western technology.

The big data economy is at the heart of this battle. AI thrives on vast datasets, and big tech firms are racing to monetize this new oil of the digital age. Companies like Google, Microsoft, Amazon, and Meta are investing billions into AI infrastructure, with the expectation of exponential profits. From personalized advertising to predictive analytics, the ability to harness and process vast quantities of data is redefining business models. The rise of generative AI- from ChatGPT to AI-driven media creation- has only accelerated this trend, leading to an insatiable demand for data centers that consume ever-growing amounts of electricity.

This unprecedented AI expansion is already straining global energy supplies. The electricity consumption of data centers is skyrocketing, pushing utilities and regulators to reconsider grid capacity, energy sourcing, and sustainability measures. As more AI-driven workloads shift to cloud computing, the question is no longer whether AI will demand vast amounts of energy- it is how much and how fast.

Therefore, predicting energy demand and resource logistics has become a necessity for all stakeholders. In this research paper, I first walk you through the methodology used to predict energy consumption using a direct relationship between data and energy consumption. By the same token, the surge in demand is expected to be supplied by fast, economic, and reliable sources of energy, which I model under time, technical, and economic constraints. Then, I discuss the uncertainty in the estimates (both quantitatively and qualitatively), growth constraints, and environmental implications of AI. Finally, I provide an energy consumption forecast in the Near-Term Outlook section of the report.

Methodology

Estimating AI-Related Energy Demand

There are numerous ways to estimate AI-related energy demand. For example, one can access data center energy bills, technical specifications, and server logs, then come up with an estimate for $/GB of processed data. However, companies might be reluctant to share confidential information about their facilities for competitive or legal reasons. Some studies have been conducted on how much *new data* is supposed to be in circulation (created, stored, and consumed) on the order of *zettabytes* between 2010 and 2035 as shown in Figure (2)**.**

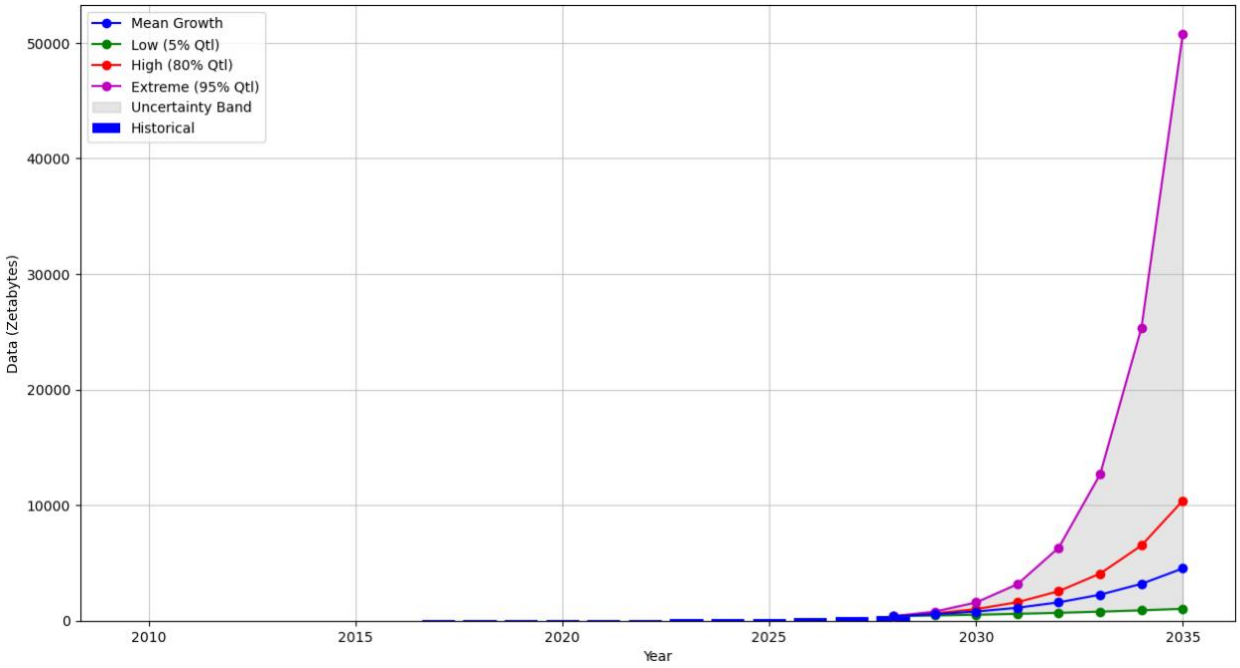

Figure 2 Data growth forecast under various scenarios[2]

The figure shows a significant change of slope in data consumption after 2020, and predicts the data consumption until 2028, exhibiting an exponential trend. So, I first modeled the distribution of growth rates, which were best fit to a lognormal distribution as shown in Figure (3). There are two inferences to be made on the growth rate of data. First, data growth is only positive; trivially because more methods of measurements have been developed, and second, more data is produced by consumers and businesses. Additionally, data growth is at an average rate of 30% with some extreme events causing skewness.

Modeling the amount of data that will be in circulation by 2035 under the different scenarios: mean, 5th percentile, 80th percentile, and 95th percentile shows a significant difference between the extremes Figure (2). This is problematic for capacity and transmission planners, and they might be interested only in the likely scenarios. In such cases, modeling growth after Covid 19 (2020) might be more representative of the current situation of data circulation.

[2] https://www.statista.com/statistics/871513/worldwide-data-created/

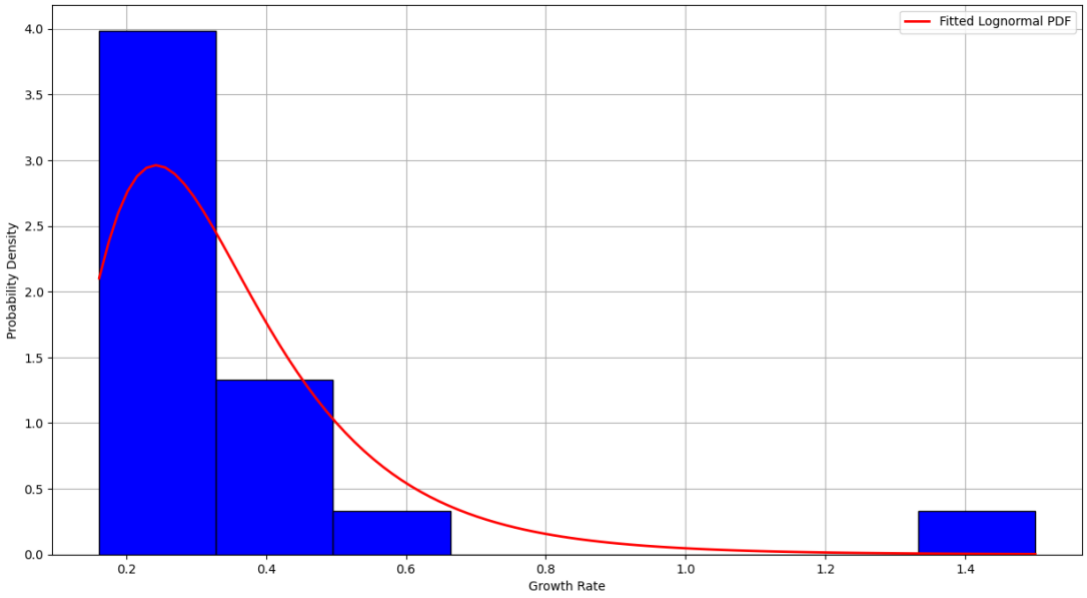

Figure 3 Data growth rate distribution, fitted to a lognormal distribution

Next, to calculate the energy consumption required to process these amounts of data, I made a few assumptions. First, all the data that is in circulation will be sourced from the internet. In other words, all this data will be at least stored in data centers. Second assumption is that U.S. data centers have or will have access to worldwide data due to the global technological dominance. Third, each server is 700 Watts/GB[3]. Fourth, the Power Usage Effectiveness (PUE) has reached a plateau at 1.56 and will not decrease any further as shown in Figures (2) and (3)**.**

Next, the equation that models the energy consumption is

$$E_{consumed} = Data \times Server_{capacity} \times time \times PUE$$

Projection of Energy Demand by Primary Source

For forecasting demand, I developed an optimization model that minimizes total costs by determining the optimal mix of power plants required to meet demand over time, while minimizing the penalty for any energy shortage each given year. It accounts for capital costs, lead times for each technology, and the annual energy demand, ensuring plants are online by the appropriate year to cover the load.

Notation:

Let $\tau = \{nuclear, solar, gas, coal\}$

Let $Y = \{2025, 2026, \ldots, 2035\}$

$x_{tech,t}(decision\ variable) = capacity\ (in\ MW) of\ technology\ built\ in\ year\ t$

$S_t(decision\ variable) = shortage\ of\ energy\ (in\ MWh)\ allowed\ in\ year\ t\ (penalized\ in\ the\ objective)$

[3] https://www.megware.com/fileadmin/user_upload/LandingPage%20NVIDIA/nvidia-h100-datasheet.pdf

$C_{tech}(parameter) = overnight\ capital\ cost\ (e.g.\ in\ \$/MW)\ for\ technology\ tech$

$L_{tech}(parameter) = lead\ time\ (years)\ for\ technology\ tech$

$CF_{tech}(parameter) = capacity\ factor\ (fraction)\ for\ technology\ tech$

$D_t(parameter) = annual\ demand\ (in\ MWh)\ in\ year\ t$

$Penalty\ (parameter) = penalty\ cost\ for\ each\ MWh\ of\ shortage$

$HOURS(parameter) = 8760\ (number\ of\ hours\ per\ year)$

Decision Variables:

$$x_{tech,t} \geq 0 \ \forall \in \tau, t \in Y$$

$$S_t \geq 0 \ \forall\ t \in Y$$

Objective Function:

Minimize the total cost: the sum of capital costs plus any shortage penalty:

$$min(\sum_{tech \in \tau} \sum_{t \in \boldsymbol{y}} (C_{tech} \cdot x_{tech,t}) + \sum_{t \in \boldsymbol{y}} (Penalty \cdot S_t))$$

Constraints:

For each year $y \in \boldsymbol{Y}$, online plants plus any shortage must meet or exceed the year's energy requirement:

$$\sum_{tech \in \tau} (CF_{tech} \times HOURS \times \sum_{t \in \boldsymbol{Y} + \boldsymbol{L_{tech} \leq y}} x_{tech,t}) + S_y \geq D_y$$

where:

- $\sum_{t \in \boldsymbol{Y} + \boldsymbol{L_{tech} \leq y}} x_{tech,t}$ means all capacity of that technology that was early enough (no later than y) is online in year y
- $CF_{tech} \times HOURS$ converts MW of capacity into annual MWh of output
- S_y is the allowed shortage, which the solver will set to zero if it's cheaper to build capacity than pay the penalty

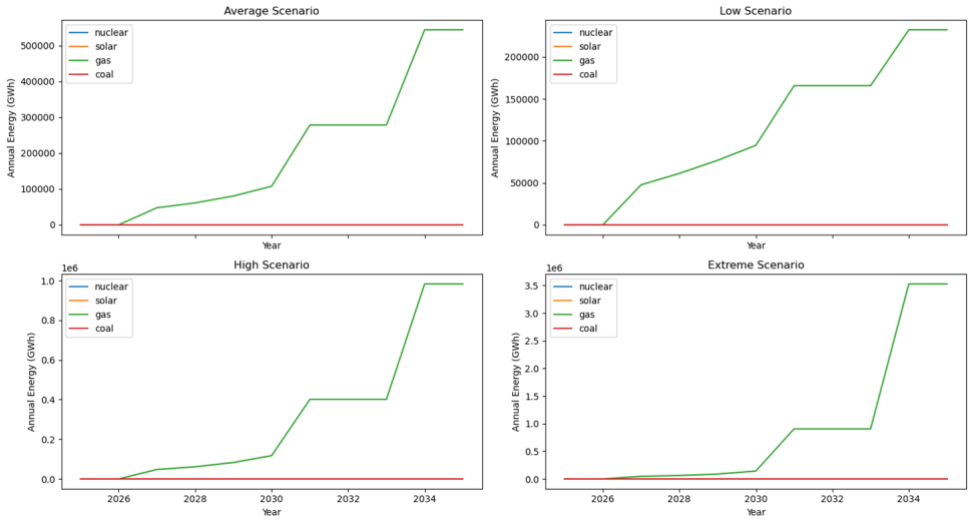

Figure 4 Energy Consumption Forecast by Primary Power Source

The model predominantly prefers natural gas over solar, nuclear, and coal because of the time and budget constraints detailed in the EIA Cost and Performance Assumptions[4]. One can argue that technologies get cheaper with time, but this was not included in the optimization model.

Uncertainty in AI Energy Projections

The uncertainty of the energy projections are of two types: modeling uncertainties and qualitative uncertainties, such as policy and technological innovation. The modeling uncertainties are (1) growth rate, (2) lognormal distribution, (3) sensitivity to initial conditions, (4) power usage effectiveness (PUE), (5) hardware efficiency trends, (6) extreme growth scenarios, and (7) data center load distribution assumptions. (1) The energy demand forecast is driven by the assumption that data consumption will continue growing at historical rates. However, AI workloads and data storage growth may not follow a predictable trend due to economic downturns, saturation in AI adoption, or shifts in enterprise computing strategies. A sudden shift in cloud service architectures, changes in user behavior, or market saturation could alter this trajectory significantly. (2) The model assumes that data growth rates follow a lognormal distribution, meaning growth is multiplicative rather than additive. While this assumption fits historical data, future trends could deviate due to unknown external factors such as breakthroughs in energy-efficient computing or changes in the economic environment. If the true distribution differs, the projections may overestimate or underestimate the expected energy consumption. (3) The forecast is heavily dependent on the last recorded data point as a baseline for future projections. Any errors in historical data, including changes in measurement methodologies or underreporting, can introduce compounding inaccuracies. A minor miscalculation today could lead to a significant deviation in energy demand estimates over the next decade. (4) The model assumes a fixed PUE, which measures how efficiently data centers convert energy into computing power. However, real-world improvements in cooling technologies and power distribution could reduce overall energy consumption per unit of data processed. If PUE improves faster than anticipated, the projected energy demand could be overestimated. (5) AI-specific hardware, such as GPUs and TPUs, is becoming more energy-efficient. The model assumes a fixed energy-per-computation ratio, but future chip designs could dramatically reduce power requirements. If hardware efficiency improvements accelerate, the projected energy demand may be significantly lower than expected. (6) The model includes a high-growth scenario based on the *95th percentile of the lognormal distribution*, projecting an exponential increase in data consumption. This assumes no major constraints on infrastructure, energy supply, or regulatory limitations. However,

[4] https://www.eia.gov/outlooks/aeo/assumptions/pdf/elec_cost_perf.pdf

physical and economic limits—such as power grid capacity, supply chain issues, and rising energy costs—could cap AI energy consumption growth, making extreme scenarios less likely. (7) The model assumes that AI-related data processing will be primarily concentrated in U.S. data centers. In reality, global shifts in cloud computing infrastructure, regulatory frameworks, or geopolitical tensions could redistribute AI workloads across different regions. If more computing shifts to decentralized or international data centers, U.S. energy demand from AI could diverge significantly from projections. Finally, to mitigate these uncertainties, sensitivity analysis should be conducted to assess how different assumptions impact energy demand forecasts.

Growth Constraints

The expansion of AI data centers and their increasing energy consumption face several growth constraints that could limit or slow down their projected demand. These constraints can be categorized into (1) infrastructure limitations, (2) resource availability, (3) economic factors, and (4) regulatory challenges. (1) AI data centers require highly reliable and scalable electricity supply, but the existing power grid infrastructure may not support exponential energy demand growth. Many U.S. grids are already operating at near-maximum capacity, and upgrading transmission networks takes years due to permitting, financing, and construction delays. Data centers are often built in clusters, leading to localized power congestion, making it difficult to integrate new facilities without overloading regional grids. (2) Even if the power grid is expanded, the actual availability of electricity from primary energy sources may not meet demand. Sustainability goals can introduce more complexities to the problem, for example, the shift toward renewable energy could introduce intermittency issues, especially if solar and wind penetration increases without sufficient storage solutions. Additionally, natural gas availability is subject to geopolitical risks and price fluctuations, while coal and nuclear face regulatory and environmental hurdles. (3) AI data centers are capital-intensive, with high upfront costs for land, infrastructure, and computing hardware. At the moment, big tech firms are willing to pay a premium to build data centers, but with a changing technological landscape or economic downturns, companies may delay or scale down expansion plans, reducing the expected growth in energy demand. (4) AI data centers operate under increasing scrutiny due to concerns over their energy consumption, environmental impact, land use, and the potential increase in customer’s electricity bills. The U.S. government may impose carbon taxes, energy efficiency mandates, stricter permitting requirements, limiting the expansion of high-energy-consuming facilities.

Environmental Implications

Since natural gas seems to be the fastest way to scale data centers, it is worth considering the emissions that could arise from AI growth. Natural gas emits less carbon dioxide than coal or petroleum, at 0.96 lbs/KWh[5]. However, if data centers continue their exponential growth, the total carbon footprint could rival or surpass entire industrial sectors. The explicit carbon emissions in 2035 of a natural-gas-powered AI revolution -for mean growth- is around *435.45 million metric tons* of CO_2 (no plant efficiency factor).

Near-Term Outlook

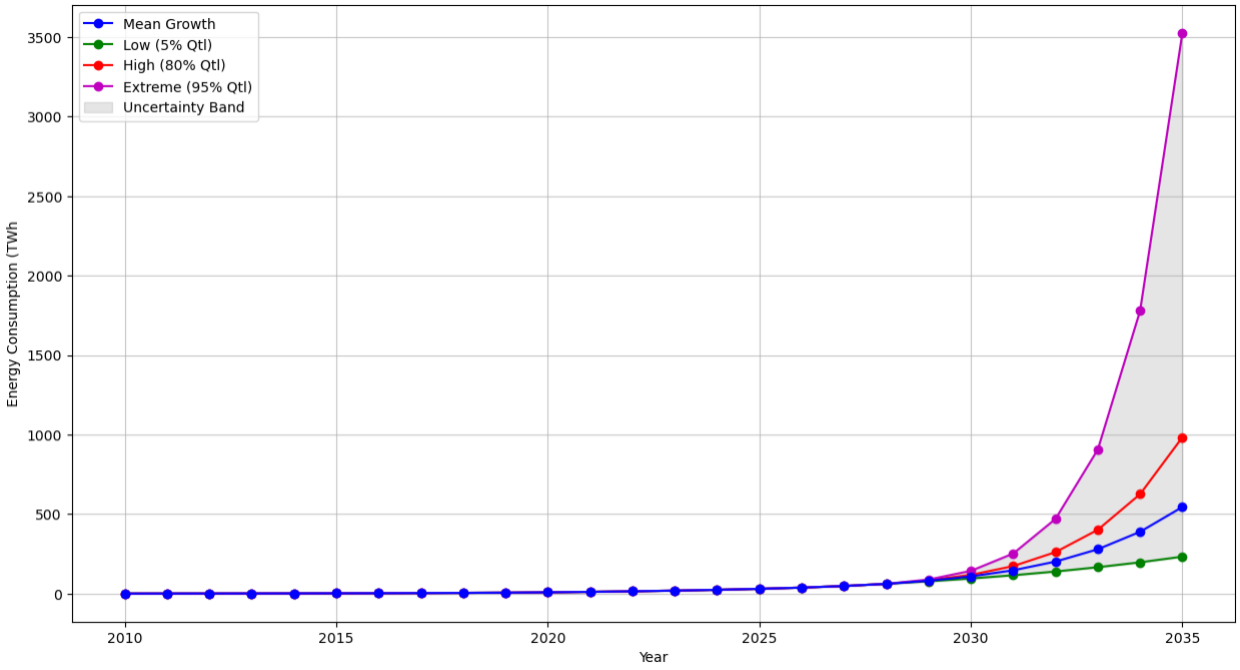

Figure 5 Data center energy consumption forecast 2025-2035 in TWh under different scenarios

The energy consumption from data centers under all different scenarios is expected to grow. However, there is a significant band of uncertainty in the estimate. At a minimum, data centers could consume between 232TWh (low scenario) and 3,527TWh (extreme scenario)[6]over the next decade. The mean and high forecasts are almost within a 100% uncertainty range. This means that more data is required to narrow the uncertainty gap, capturing growth-influential factors.

[5]https://www.eia.gov/tools/faqs/faq.php?id=74&t=11

[6] Predicted using the highest historical growth rate of data consumption and creation