May 2021

Machine Intelligence for Scientific Discovery and Engineering Invention

CSET Data Brief

CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

AUTHORS
Matthew Daniels
Autumn Toney
Melissa Flagg
Charles Yang

Introduction

Artificial intelligence is driving new discoveries and advancements across many fields of science and engineering. Modern applications of AI and machine learning have the potential to change the practice of research and development (R&D) in the United States over the next decade. This data brief is a first step toward understanding how modern AI and ML have begun accelerating growth across various science and engineering disciplines in recent years.

The application of AI to science is not new. Pioneers at Stanford University in the 1960s used early AI systems to automate analyses of chemical structure and mass spectra, supported by the Advanced Research Projects Agency (ARPA), the predecessor to the Defense Advanced Research Project Agency (DARPA), and NASA.[1] This work grew out of an interest in using AI systems for complex reasoning in scientific problems.[2]

Today, deep learning methods, a specific subset of ML, can predict the 3D structure of proteins to within the width of an atom—substantial progress on a 50-year challenge achieved with modern machine learning.[3] This tool is important because it may contribute new capabilities for disease treatment and pandemic response. In the years ahead, modern ML may help advance the engineering design of viable, cost-effective fusion reactors, a large-scale source of clean energy that has also been out of reach for more than 70 years.[4] AI applications can drive advancements in fields as disparate as gene editing and chip design.

American institutions have begun to adapt to this future. MIT, for example, announced its largest structural change since the 1950s and invested $1 billion to found a new college focused on the intersections between computing, AI, and the university's existing disciplines of science and engineering.[5] Students will be encouraged to be "bilingual" between computing and other disciplines.[6]

As modern AI applications begin to transform science and engineering, a key question for high-level leaders shaping science and technology strategies is what disciplines will be affected first, and will these early impacts be narrowly confined to a few disciplines (like materials science and biomedical fields) or broadly applicable? Accelerated emergence and growth of new fields is especially noteworthy: bioengineering, computer science, and materials science have all emerged as enabling disciplines that changed science and engineering broadly in the last century—and were followed by national and global changes.

This data brief provides an initial look at how modern AI and ML have begun accelerating growth across a wide array of research disciplines in recent years, and focuses on direct applications of AI to making new scientific and engineering breakthroughs.[7] We first summarize illustrative examples of AI driving innovation in medicine, fundamental sciences, and engineering research. We then analyze global scientific publications data to see if these recent achievements represent real trends in published research.

Illustrative Examples of Machine Learning for Science and Engineering

Within the long history of AI for science, deep learning has recently emerged as a promising new tool for scientists and engineers. Here, we provide illustrations of the great potential of modern AI and ML to impact R&D in new ways. These examples motivate a more systematic look in the next section using CSET data on global research clusters.

In medicine, one potential application of AI is its ability to improve healthcare diagnostics, particularly the ability to accurately and rapidly assess radiology scans.[8] Facebook and New York University researchers have used a deep learning approach to significantly reduce the time required for Magnetic Resonance Imaging (MRI) scans.[9] Various U.S. agencies overseeing healthcare regulations have begun approving AI applications in the healthcare industry. The Food and Drug Administration has already given clearance to a private startup to begin using AI tools to assist with reading prostate scans[10] and the Center for Medicare and Medicaid

Services has also approved an AI startup to use automated imagery analysis to quickly help identify strokes in patients.[11]

In the fundamental sciences, AI is helping scientists make sense of complex, high-dimensional data. Physicists at MIT demonstrated an AI system able to re-derive 100 fundamental physics equations from the canonical *Feynman Lectures* based on simulated data.[12] In astrophysics and astronomy, researchers have used graph neural networks to propose a new analytical equation for dark matter phenomena that better matches observed data than previous human-derived equations.[13] Others identified previously missed exoplanets in large-scale astronomy data with deep learning.[14] Applications of ML are also helping scientists discover new materials that can advance cleaner energy sources. For example, researchers have recently used neural networks to better understand the high-dimensional spectroscopy of nanoscale ferroelectric materials[15] and run high throughput screening of possible battery cathode materials.[16] Scientists also used AI to optimize fast-charging protocols for Li-ion batteries while preserving battery health,[17] which could have important implications for electric vehicle charging. In the life sciences and bioengineering, modern AI and ML have shown promise in protein engineering,[18] improving the gene targeting capabilities of clustered regularly interspaced short palindromic repeats (CRISPR) technologies,[19] and discovering new classes of antibiotics that retain effectiveness against drug-resistant bacteria.[20]

An international collaborative effort between universities in the United States and Europe showed how deep learning methods can significantly accelerate a variety of important scientific simulations, including those from geology, nuclear fusion, and environmental science.[21] Researchers at Google also used a deep learning model to demonstrate high-resolution, real-time weather forecasts that outperform the current best Numerical Weather Prediction system available from the National Oceanic and Atmospheric Administration.[22]

In engineering and manufacturing, AI is also slowly becoming an important tool in developing new engineering design approaches and advancing U.S. manufacturing capabilities. Researchers at

Google have also used deep reinforcement learning to accelerate the design process for semiconductor computer chips.[23] General Motors has partnered with Autodesk to develop more lightweight vehicles using AI generative modelling.[24] Advanced manufacturing techniques like 3D printing have use-cases for AI to help improve the tuning of manufacturing parameters.[25] Finally, convolutional neural networks show potential in improving manufacturing quality control and helping reduce defects in manufacturing lines.[26]

These are examples of how modern AI applications can drive innovation. The next section examines more systematically how widespread the application of AI is to other fields of science and engineering. We do this by looking at all the lines of globally published research that are merging AI with other fields.

Findings

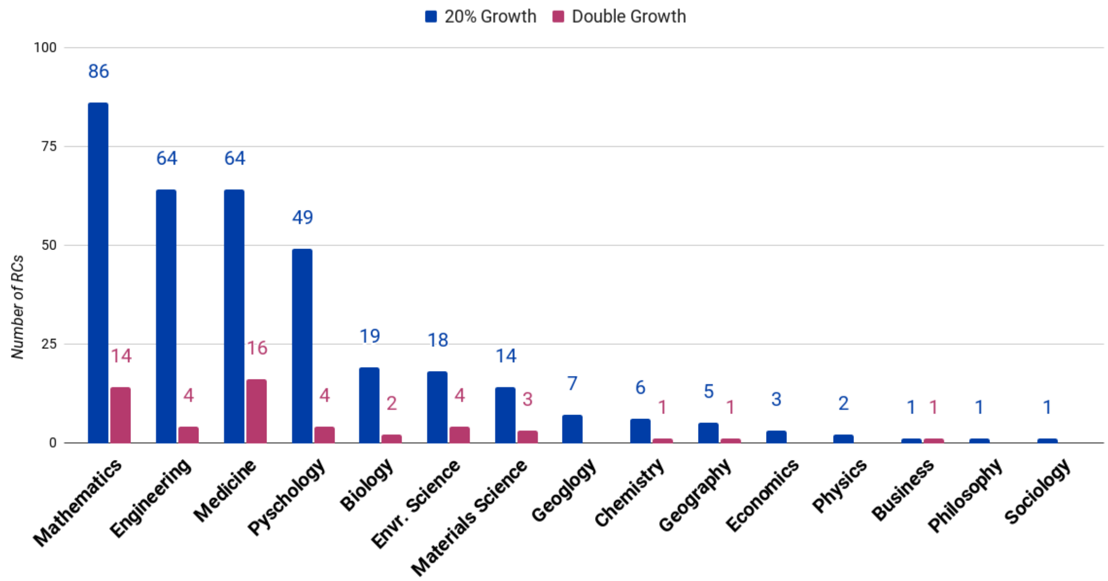
To begin understanding the breadth of AI's current impact, we analyze the growth of AI-related papers across different scientific disciplines. We do this by leveraging a large, ongoing effort at CSET to explore scientific research publications. CSET uses a merged dataset that accounts for roughly 90 percent of the world's scientific publications. We explore this dataset via an automated structuring of scholarly literature, which we refer to as research clusters.[27] RCs are generated using direct citation links from publications, meaning that clusters are defined by their citation relationships as opposed to a social network of authors. Rahkovsky et al. provide full details on the clustering methodology using CSET's merged corpus and the application of Dunham et al.'s AI labeling model, which we use to identify AI-related papers.[28]

We programmatically assign each RC a research discipline label from a list of 19 broad research areas using Microsoft Academic Graph's (MAG) field of study:[29] Art, Biology, Business, Chemistry, Computer Science, Economics, Engineering, Environmental Science, Geography, Geology, History, Materials Science, Mathematics, Medicine, Philosophy, Physics, Political Science, Psychology, and Sociology. This label corresponds to the majority of MAG papers in the RC; papers in a RC that do not come from the MAG dataset do not contribute to this label. Because RCs are defined by their direct citation links, however, they are not restricted to containing publications from only one discipline—any RC can encompass cross-disciplinary research. Any clustering of multidisciplinary papers comes with caveats (see our GitHub archive for further details),[30] but for the purposes of this paper, these RCs provide a useful representation of AI's spread into diverse areas of science.

For our analysis, we are interested in RCs that are AI-related, growing rapidly, and specifically affecting disciplines beyond computer science.[31] From the entire set of RCs covering all research disciplines (126,925 in total), there are 5,221 RCs with 10 percent or more AI-related publications. Of those AI-related RCs, 1,170 do not have a computer science research discipline label. From these AI-related RCs outside computer science, we identify 340 RCs, encompassing 308,208 scientific research publications, with sustained rapid growth (20 percent annual growth or more) over the last three years. Applying a doubling growth threshold (100 percent growth or more), we narrow this further to 50 RCs, which encompass 39,488 scientific research publications. Figure 1 provides the breakdown of RCs with rapid growth and doubling growth for each research discipline, excluding computer science. For both rapid growth and doubling growth RCs, engineering, mathematics, and medicine are the top three research disciplines with cross-disciplinary AI research.

While engineering, mathematics, and medicine lead progress, AI is entwined with growth in a wide span of science and engineering, including chemistry, biology, materials science, environmental science, and a strong presence in psychology. Figure 1 also shows this wide distribution of disciplines for the 340 rapidly growing RCs and the 50 RCs with doubling growth—notably, even with a threshold for an extremely high degree of growth, we still see impact across a broad set of disciplines. This analysis supports the growing impression that AI is influencing rapidly emerging areas of science and engineering far beyond computer science. It also suggests that U.S. research will benefit from AI-relevant skills dispersed widely in other technical communities.

Figure 1. AI-related research clusters with rapid growth (≥20% annually) and doubling growth annually by discipline



Source: CSET Map of Science.

Takeaways

Building U.S. competitiveness in AI is a complementary priority to maintaining the United States' wider leadership in science and technology.

This data brief provides an initial look at AI's appearance in rapidly growing clusters of published research worldwide. Applications of AI are a substantial part of rapidly growing research clusters across a broad array of disciplines—not just in one or two areas. Mathematics, medicine, engineering, psychology, biology, materials science, and environmental science have especially large quantities of rapidly growing research areas with noteworthy interdisciplinary roles for modern AI.

For U.S. leaders, this may become a source of both opportunity and surprise. Most of all, these advancements make it clear that U.S. R&D organizations will have growing opportunities to apply AI to scientific discovery and engineering invention in many areas. On a national scale, these applications may accelerate the emergence and growth of new domains of science at the intersections of more established disciplines; disciplines far from computer science will change in potentially surprising ways as AI offers new tools for R&D.[32]

Large R&D organizations, both inside and outside of the U.S. government, have particular opportunities:

- *Education investments that include more and better cross-training in AI for scientists and engineers.* For U.S. government organizations, this could build on precedent like the Department of Energy's Computational Science Graduate Fellowship, but oriented to focus on AI and computing broadly.[33] Such education investments should give special attention to diversity and inclusion, which is important both for future AI research and for enabling cross-disciplinary research. These investments will empower the next generation workforce to apply AI to problems across the U.S. government, from health to spaceflight.

- *Increasing openness of access to scientific publications and data, especially in formats accessible and useful to "machine readers."* For example, the National Science Foundation, Department of Energy, and the National Institutes of Health might require funded researchers to open source data and publications (analogous to Plan S in the European Union,[34] and building on the 2013 policy memorandum by the Office of Science and Technology Policy Director, John Holdren, directing U.S. federal agencies with more than $100 million in R&D expenditures to develop plans to make results of federally funded research freely available to the public[35]). One example of early work in this direction has been the COVID-19 Open Research Dataset (CORD-19), a resource of more than 400,000 scholarly articles in machine-readable format about SARS-CoV-2 and related viruses for use by the global ML community.[36]

- *Providing AI and data toolkits for scientists and engineers in different disciplines*. The U.S. government should especially fund AI building blocks that allow scientists to cheaply leverage AI tools. For example, this might include supporting the creation of pre-trained AI models with baked-in scientific knowledge.[37]

- *Providing scientific/engineering challenges and datasets more widely to AI researchers*. For example, in the early 2000s, DARPA Grand Challenges helped jump-start the development of today's autonomous vehicles. Progress in other areas of AI, such as image recognition[38] and the ability to successfully play strategy board games, has benefited from large, robust benchmarks. Scientific funding agencies can articulate analogous challenges based on problems that are relevant to the scientific community to encourage the development of AI scientists.

- *Pursuing international collaboration with U.S. allies, particularly the United Kingdom and Japan.* In February 2020, the U.S. Department of Defense funded a workshop hosted at the Alan Turing Institute in London, which convened researchers from the United States, the United Kingdom, and Japan—and supported a new planetary challenge for AI systems able to make Nobel-worthy scientific discoveries. U.S. allies already share interest in this topic, and can provide opportunities for wider collaboration.

Modern AI is helping to reshape the global research landscape beyond computer science. The security, health, and economic vitality of nations depend in part on the inventions they have ready access to—and modern applications of AI can help accelerate the creation of new inventions in the years ahead.

Authors

Matthew Daniels is a senior fellow at CSET, where Autumn Toney is a data research analyst and Melissa Flagg is also a senior fellow. Charles Yang is a graduate student at the University of California, Berkeley in the Department of Electrical Engineering and Computer Science.

Acknowledgments

Thanks to Igor Mikolic-Torreira, Helen Toner, Dewey Murdick, Shelton Fitch, Brendan Tracey, Ayodeji Coker, and Jared Dunnmon for useful insights and feedback. We are also grateful to colleagues at the Alan Turing Institute for hosting discussions on AI for science, and to Hiroaki Kitano for his early vision of a global grand challenge focused on AI scientists. The authors are solely responsible for the views expressed in this publication and for any errors.

Document Identifier: doi: 10.51593/20200099

Endnotes

[1] Joshua Lederberg and Edward A. Feigenbaum, "Mechanization of Inductive Inference in Organic Chemistry," in Benjamin Kleinmuntz, ed., *Formal Representations for Human Judgment* (New York: Wiley, 1968). Available online: https://exhibits.stanford.edu/feigenbaum/catalog/bm350jn4518.

[2] Bruce G. Buchanan and Edward A. Feigenbaum, "DENDRAL and Meta-DENDRAL: Their applications dimension," in *Readings in artificial intelligence* (Burlington: Morgan Kaufmann, 1981), 313-322.

[3] Ewen Callaway, "'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures," *Nature*, November 30, 2020, https://www.nature.com/articles/d41586-020-03348-4.

[4] See E.A. Baltz, E. Trask, M. Binderbauer et al., "Achievement of Sustained Net Plasma Heating in a Fusion Experiment with the Optometrist Algorithm," *Scientific Reports* 7: 6425 (2017), https://doi.org/10.1038/s41598-017-06645-7 and Julian Kates-Harbeck, Alexey Svyatkovskiy, and William Tang, "Predicting disruptive instabilities in controlled fusion plasmas through deep learning," *Nature* 568: 526–531 (2019), https://doi.org/10.1038/s41586-019-1116-4.

[5] MIT News Office, "MIT reshapes itself to shape the future," Massachusetts Institute of Technology, October 15, 2018, https://news.mit.edu/2018/mit-reshapes-itself-stephen-schwarzman-college-of-computing-1015.

[6] Kaveh Waddell, "MIT is investing $1 billion into AI research," *Axios*, October 18, 2018, https://www.axios.com/mit-ai-research-1-billion-b20ade0a-ba12-4300-922a-c8033dd47fe9.html.

[7] This goal has had a resurgence in the 2010s: in 2016 Hiroaki Kitano, creator of Robocup, proposed a grand challenge for AI systems capable of making Nobel-worthy scientific discoveries: Hiroaki Kitano, "Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine for Scientific Discovery," *AI Magazine* 37, no. 1 (2016): 39-49. In future work we will discuss broader ways that AI may accelerate S&E.

[8] One particular concern when developing AI systems for healthcare is that such technology may present racial biases. The development of equitable, fair, and transparent healthcare technology is an important challenge that will need to be addressed in the areas where AI, science, and people intersect. In particular, the National Academy of Medicine released a report that details the challenge and opportunity of AI in healthcare that is a useful starting reference. See Michael Matheny, Sonoo Thadaney Israni, Mahnoor Ahmed, and Danielle Whicher, eds., "Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril"

(National Academy of Medicine, 2019), https://nam.edu/wp-content/uploads/2019/12/AI-in-Health-Care-PREPUB-FINAL.pdf.

[9] Michael Recht et al., "Using Deep Learning to Accelerate Knee MRI at 3 T: Results of an Interchangeability Study," *American Journal of Roentgenology* 215, no. 6 (2020): 1421-1429, https://www.ajronline.org/doi/10.2214/AJR.20.23313.

[10] Emi Gal, "Ezra Announces First FDA-cleared Prostate AI," Ezra, October 20, 2020, https://ezra.com/ezra-announces-first-fda-cleared-prostate-ai/. Letter from the Food and Drug Administration, June 11, 2020, https://www.accessdata.fda.gov/cdrh_docs/pdf19/K192969.pdf.

[11] See Viz.ai, "Viz.ai Granted Medicare New Technology Add-on Payment," Cision PR Newswire, September 3, 2020, https://www.prnewswire.com/news-releases/vizai-granted-medicare-new-technology-add-on-payment-301123603.html and ContaCT, *Federal Register* 85, no. 182 (September 18, 2020): 58625, https://www.govinfo.gov/content/pkg/FR-2020-09-18/pdf/2020-19637.pdf.

[12] Silviu-Marian Udrescu and Max Tegmark, "AI Feynman: A physics-inspired method for symbolic regression," *Science Advances* 6, no. 16 (2020): https://doi.org/10.1126/sciadv.aay2631.

[13] Miles Cranmer et al., "Discovering Symbolic Models from Deep Learning with Inductive Biases," *arXiv [cs.LG]*, arXiv (November 18, 2020), https://arxiv.org/pdf/2006.11287.pdf.

[14] Anne Dattilo et al., "Identifying Exoplanets with Deep Learning. II. Two New Super-Earths Uncovered by a Neural Network in K2 Data," *The Astronomical Journal* 157, no. 5 (2019): https://iopscience.iop.org/article/10.3847/1538-3881/ab0e12.

[15] Joshua Agar et al., "Revealing ferroelectric switching character using deep recurrent neural networks," *Nature Communications* 10 (2019): 4809, https://doi.org/10.1038/s41467-019-12750-0.

[16] Rajendra Joshi et al., "Machine Learning the Voltage of Electrode Materials in Metal-Ion Batteries," *ACS Applied Materials & Interfaces* 13, no. 14 (2019), https://doi.org/10.1021/acsami.9b04933.

[17] Peter M. Attia et al., "Closed-loop optimization of fast-charging protocols for batteries with machine learning," *Nature* 578 (2020): 397-402, https://doi.org/10.1038/s41586-020-1994-5.

[18]Ali Madani et al., "ProGen: Language Modeling for Protein Generation," bioRxiv, 2020, https://doi.org/10.1101/2020.03.07.982272.

[19] Ayal B. Gussow et al., “Machine-learning approach expands the repertoire of anti-CRISPR protein families,” *Nature Communications* 11 (2020): 3784, https://doi.org/10.1038/s41467-020-17652-0.

[20] Jonathan M. Stokes et al., “A Deep Learning Approach to Antibiotic Discovery,” *Cell* 180, no. 4 (2020): https://doi.org/10.1016/j.cell.2020.01.021.

[21] Muhammad Kasim et al., “Building high accuracy emulators for scientific simulations with deep neural architecture search,” *arXiv [stat.ML]*, arXiv (October 8, 2020), https://arxiv.org/pdf/2001.08055.pdf.

[22] Casper Kaae Sønderby et al., “MetNet: A Neural Weather Model for Precipitation Forecasting,” *arXiv [cs.LG]*, arXiv (March 30, 2020), https://arxiv.org/abs/2003.12140.

[23] Azalia Mirhoseini et al., “Chip Placement with Deep Reinforcement Learning,” *arXiv [cs.LG]*, arXiv (April 22, 2020), https://arxiv.org/pdf/2004.10746.pdf.

[24] “Advanced Software Design Technology Leads GM into Next Generation of Vehicle Lightweighting,” General Motors, May 3, 2018, https://media.gm.com/media/us/en/gm/news.detail.html/content/Pages/news/us/en/2018/may/0503-lightweighting.html.

[25] Daniel Oberhaus, “Massive, AI-Powered Robots Are 3D-Printing Entire Rockets,” *WIRED*, October 14, 2019, https://www.wired.com/story/massive-ai-powered-robots-are-3d-printing-entire-rockets/.

[26] “Landing AI Unveils AI Visual Inspection Platform to Improve Quality and Reduce Costs for Manufacturers Worldwide,” Landing AI, October 21, 2020, https://landing.ai/landing-ai-unveils-ai-visual-inspection-platform-to-improve-quality-and-reduce-costs-for-manufacturers-worldwide/.

[27] Research clusters (RCs) are groupings of scientific research articles linked by citation that provide a representation of related research. The articles used to identify clusters are included in CSET’s merged dataset of scholarly literature containing Dimensions (https://www.dimensions.ai/), Microsoft Academic Graph (https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/), CNKI (http://online.eastview.com/projects/cnki/), and Web of Science (https://clarivate.com/webofsciencegroup/solutions/web-of-science/) data. More details can be found at: https://github.com/georgetown-cset/MI-for-Discovery-and-Innovation.

[28] See Ilya Rahkovsky et al., “AI Research Funding Portfolios and Extreme Growth,” *Frontiers in Research Metrics and Analytics* (2021) 6 and James Dunham, Jennifer Melot, and Dewey Murdick, “Identifying the Development and

Application of Artificial Intelligence in Scientific Text," *arXiv [cs.DL]*, arXiv (May 28, 2020), https://arxiv.org/abs/2002.07143.

[29] See "Topics Analytics," Microsoft Academic, https://academic.microsoft.com/topics/.

[30] The GitHub archive with related details is available at: https://github.com/georgetown-cset/MI-for-Discovery-and-Innovation.

[31] More details can be found at https://github.com/georgetown-cset/MI-for-Discovery-and-Innovation.

[32] Further, the diversity of recent major advances within CS beyond supervised learning, such as in generative modeling (e.g., OpenAI's *GPT-3*) and deep reinforcement learning (e.g., DeepMind's *AlphaStar*) suggest this trend of increased collaboration between applications of AI and science has more room to grow and will likely continue in the years ahead.

[33] Office of Energy Efficiency and Renewable Energy, "Department of Energy Computational Science Graduate Fellowship," U.S. Department of Energy, https://www.energy.gov/eere/education/events/department-energy-computational-science-graduate-fellowship.

[34] "Plan S Principles," Plan S, https://www.coalition-s.org/plan_s_principles/.

[35] Michael Stebbins, "Expanding Public Access to the Results of Federally Funded Research," *Obama White House Blog*, February 22, 2013, https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research.

[36] Center for Security and Emerging Technology, "COVID-19 Open Research Dataset (CORD-19)" (CSET, March 16, 2020), https://cset.georgetown.edu/research/covid-19-open-research-dataset-cord-19/. For statistics on CORD-19 current as of May 2021, see: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.

[37] As an early example in this area, DARPA has funded work on *physics-informed neural networks* (PINNs), see Maziar Raissi, Paris Perdikaris, and George E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics* 378 (2019): 686-707, https://www.sciencedirect.com/science/article/pii/S0021999118307125?casa_token=IgwKuIORSTYAAAAA.

[38] See http://www.image-net.org/.