Overview of Foundation Models

Mehmet F. Demirel

Sep 20, 2021

What are foundation models?

- **Models that are trained on broad data at scale** and **are fine-tunable to a wide range of downstream tasks**, e.g., BERT, DALL-E, GPT-3, CLIP.
- Not new, but their scale and scope have expanded exponentially over the last few years.

Homogenization and Emergence

Homogenization

- The scale results in new emergent capabilities as well as their effectiveness across many tasks incentivizes *homogenization.*
- Indicates the consolidation of methodologies for building ML systems across a wide range of applications.
- Provides strong leverage towards many tasks, but…
- The issues in the underlying foundation models will impact all downstream models (single-point failure).

Emergence

- The behavior of a system is implicitly induced rather than explicitly constructed.
- Source of scientific excitement, but…
- Also a source of anxiety about unanticipated consequences.

Homogenization and Emergence

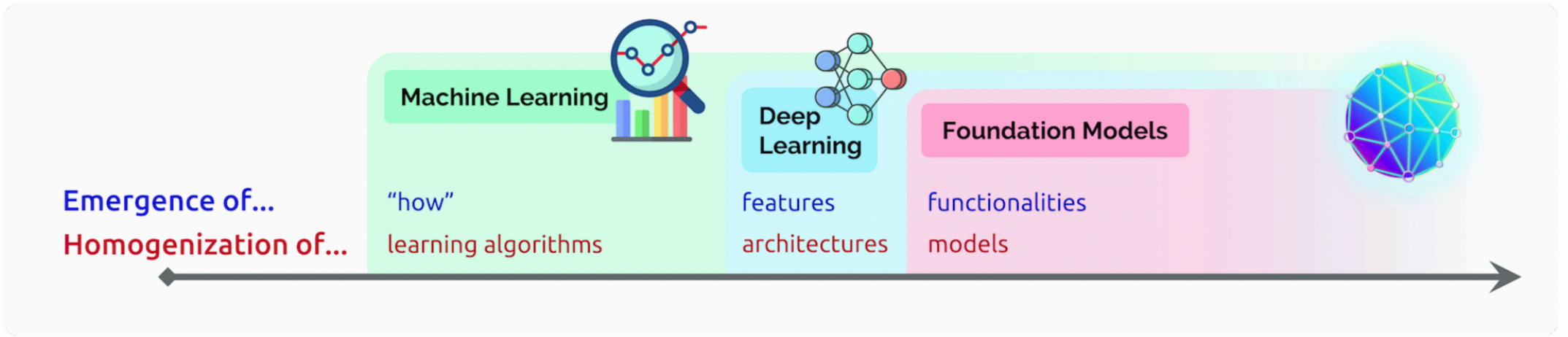


Fig. 1. The story of AI has been one of increasing *emergence* and *homogenization*. With the introduction of machine learning, *how* a task is performed emerges (is inferred automatically) from examples; with deep learning, the high-level features used for prediction emerge; and with foundation models, even advanced functionalities such as in-context learning emerge. At the same time, machine learning homogenizes learning algorithms (e.g., logistic regression), deep learning homogenizes model architectures (e.g., Convolutional Neural Networks), and foundation models homogenizes the model itself (e.g., GPT-3).

Homogenization in Foundation Models

- Unprecedented homogenization
  - Almost all SOTA NLP models are adapted from one of several foundation models: BERT, RoBERTa, BART, T5
  - **Advantage:** If you improve the foundation model, you get improvement across all applications in NLP
  - **Disadvantage:** If you mess up the foundation model, you mess up across all applications in NLP

Homogenization in Foundation Models

Homogenization across modalities.

- Same Transformer-based modeling is being applied to text, images, speech, protein sequences, reinforcement learning, etc.

Homogenization in Foundation Models

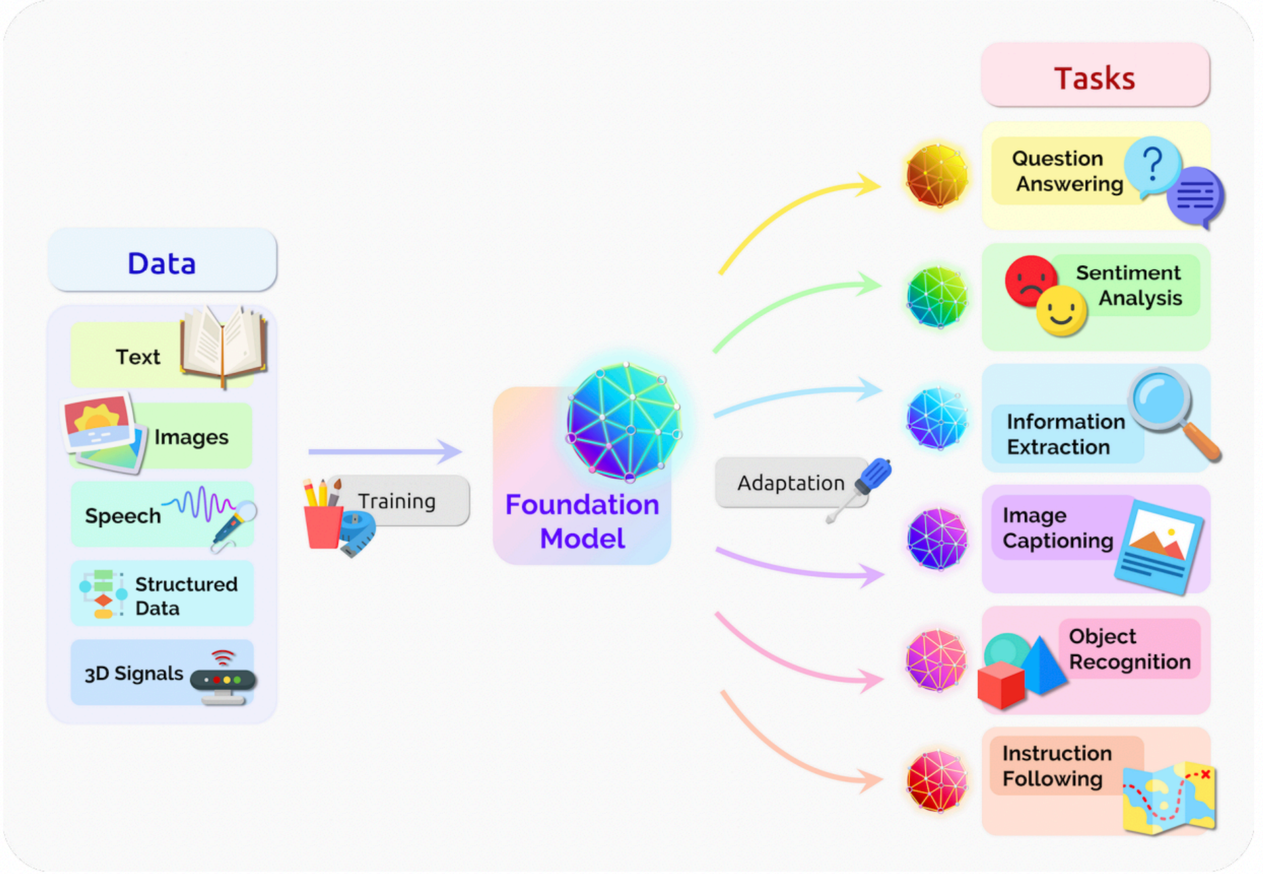


Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.

Emergence in Foundation Models

Surprising emergence which results from scale

- GPT-3 with 175B parameters compared to GPT-2's 1.5B, allows for **in-context learning,** in which the language model can be adapted to a downstream task simply by providing it with a natural language description of the task - an emergent property that was neither specifically trained for nor anticipated to arise.

What is the problem?

Homogenization **can potentially provide enormous gains where task-specific data is limited**, but **any flaws in the model are blindly transferred to all adapted models.**

As the **power of foundation models can from emergent qualities**, **they are hard to understand (no explicit construction), i.e., uncertainty in knowing the qualities and flaws of a model.**

Therefore, if we aggressively focus on homogenization, **we risk a lot!**

Does it matter right now?

YES!

Foundation models are being quickly deployed in real-world applications that have an enormous impact on humans.

- Google Search, with 4 billion users, now depends on BERT.

Social Impact of Foundation Models

- Political exacerbation of social inequities
- Economic impact due to increased capabilities
- Environmental impact due to increased computational demands
- Concerns of amplifying disinformation
- Legal ramifications due to powerful generative capabilities
- Ethical issues resulting from homogenization

What to do? Let’s first understand.

- Research in foundation models are cool (papers, conferences, competition leaderboards), but **the problem starts when the research is integrated into real-world applications.**
- First, we need to understand the full ecosystem from research to deployment of foundation models.

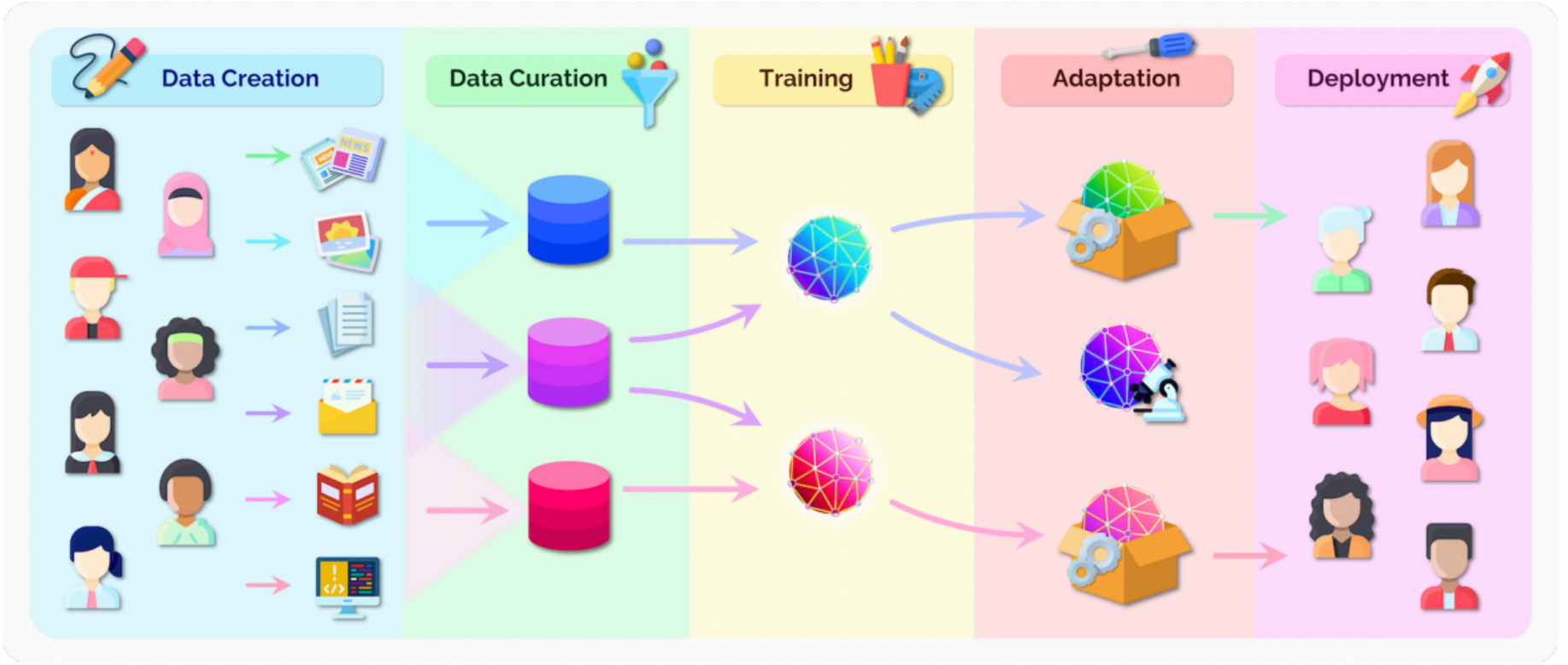


Fig. 3. Before reasoning about the social impact of foundation models, it is important to understand that they are part of a broader ecosystem that stretches from data creation to deployment. At both ends, we highlight the role of people as the ultimate source of data into training of a foundation model, but also as the downstream recipients of any benefits and harms. Thoughtful data curation and adaptation should be part of the responsible development of any AI system. Finally, note that the deployment of adapted foundation models is a decision separate from their construction, which could be for research.

Think ecosystem, act model

Future of Foundation Models

- We are in the early days of foundation models.
- Although these models are deployed to real-life applications, they are poorly understood.

- **Who will determine the future of foundation models?**

Future of Foundation Models

Disciplinary Diversity

- The research behind the tech comes from academia and industry.
- Yet, we need to consider social implications and ethical design WHEN developing foundation models, NOT AFTER.
  - Including everything in the development stages, not just training.
- Given that academic intuitions assemble many disciplines under one umbrella, academia is important in developing foundation models that promote their social benefit and mitigate social harms as well as strictly prohibit certain actions carried out in each stage of the ecosystem.

Future of Foundation Models

Incentives

- Industry will have little incentive in many areas of application of foundation models.
  - Little incentive to devote resources to technologies designed to improve the conditions of poor and marginalized people.
  - Commercial incentive will lead companies to ignore social externalities, such as technological displacement of labor, health of an informational ecosystem required for democracy, environmental cost of computing resources, and sale of technologies to non-democratic regimes.
  - Little incentive to create an open, decentralized ecosystem in developing foundation models that can be accessed/participated in by the broad community.
- Academic institutions; however, have a mission to produce and disseminate knowledge and create global public goods.

Future of Foundation Models

Loss in Accessibility

- Reproducibility is very common in ML (challenges, code repos, released datasets, PyTorch/Tensorflow).
  - This has led to significant progress in research.
- Not the same story with foundation models
  - Some models (GPT-3) are not released or require API access.
  - Trained models can be available (BERT), the actual training is unavailable to vast majority of AI researchers due to high computational cost and complex engineering.
- Train small models?
  - Sure, but some important functionalities will depend on scale.
- Study existing models like BERT?
  - Sure, but to be able to infuse social awareness and ethical design into these models, we need to be in the building phase.
- Big companies have the money, infrastructure, users, data. Start-ups are doing well. Academia is not good. The gap is increasing.

Future of Foundation Models

Solution to close the gap?

- Government investment in public infrastructure (similar to Hubble Space Telescope and the Large Hadron Collider).
  - National Research Cloud initiative is a step in the right direction.
- Volunteer computing
  - Billions of computing devices can connect to a central server and contribute computation (Folding@home, Learning@home)
  - Still technically challenging due to high latency between devices and high bandwidth requirements

Capabilities of Foundation Models

- **Language**
- **Vision**
- **Robotics**
- **Reasoning and Search**
- **Interaction**
- **Philosophy of Understanding**

Capabilities of Foundation Models

Language

- NLP has been the field most profoundly affected by foundation models
- Skilled language generators (e.g. GPT-3)
- More importantly, generality and adaptability.
  - The modern approach is to use a single foundation model and create an adapted version using labeled data.
  - Still better than models built specifically for a task.
    - Answering open-ended science questions in 2018 —> 73.1%
    - **Adapted foundation model in 2019 —> 91.6%**

Capabilities of Foundation Models

Language

- Before 2018, generating language was thought to be impossible, and NLP focused on analyzing and understanding text. **Now, it is very simple to train highly coherent foundation models with a language generation objective, like “predict the next word in this sentence”.**

- These generative approaches now form the backbone of ML for language, including analyzing and understanding

Capabilities of Foundation Models

Language

- For most of the 6000 languages in the world, we don’t have enough text data.
- There are multilingual foundation models (mBERT, mT5, XLM-R) that are trained on multiple languages with the assumption that they share common patterns, which is surprisingly true.
- Still not clear how robust these models are.
  - If it looks like English, good. If not, not so much.

Capabilities of Foundation Models

Language

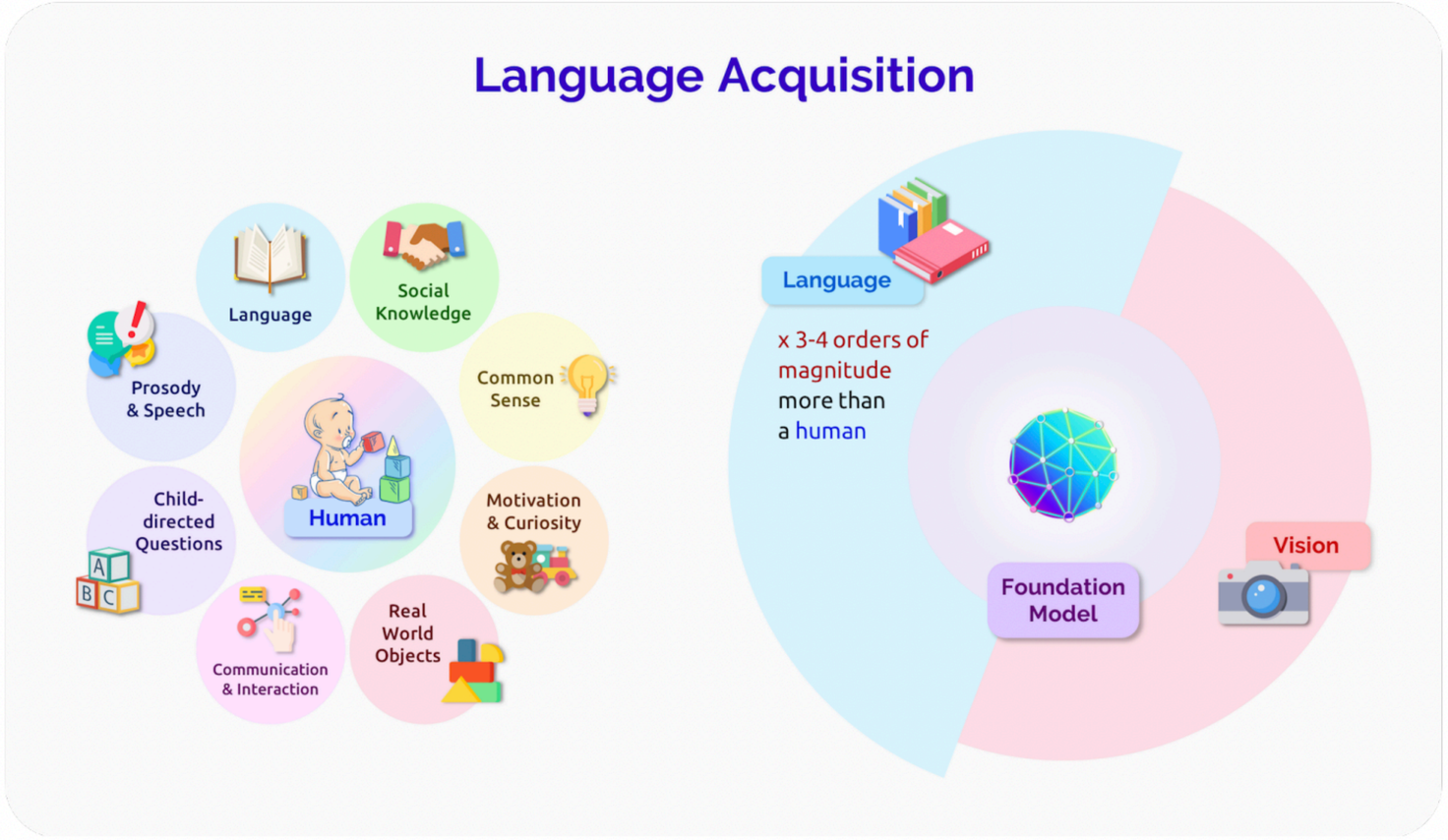


Fig. 6. Language Acquisition for humans and foundation models. While there are certainly different inductive biases between the human brain and foundation models, the ways that they learn language is also very different. Most saliently, humans interact with a physical and social world in which they have varied needs and desires, while foundation models mostly observe and model data produced by others.

Capabilities of Foundation Models

Vision

- The advantage of foundation models come from the limitations of traditional models, which rely on expensive and carefully-annotated data.
- Recent advances in self-supervised learning presents a different route to build foundation models that can use many raw data to understand the visual world.
- DALL-E, CLIP, GAN.

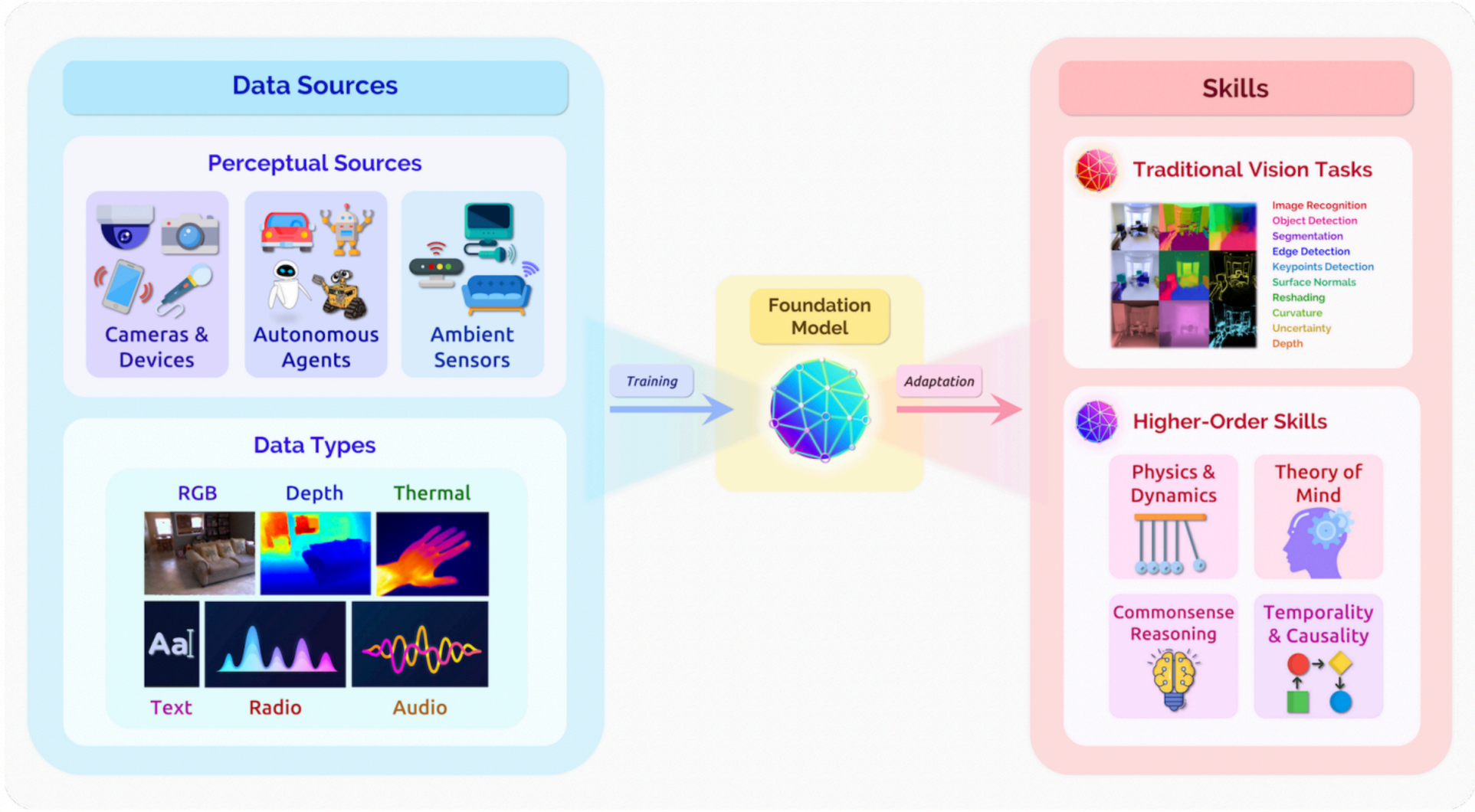
Capabilities of Foundation Models

Vision

- Research challenges
  - **Semantic systematicity and perceptual robustness:** While current foundation models show promising capability for image synthesis, they struggle to generalize to compositions of simple shapes and colors.
  - **Computational efficiency and dynamics modeling:** Images are big. Currently we use embeddings that summarize image patches etc. Yet this has a risk of losing fine-grained info.
  - **Training, environments, and evaluation:**
    - Current foundation models focus on RGB images and text. Motivates the use of large-scale datasets with diverse inputs across a wide spectrum of modalities.
    - Rather than static datasets, simulation environments that capture physical, visual, and ecological realism with multiple modalities and viewpoints.
    - Frechet Inception Distance and BLEU have flaws. Human judgement could be good, but is costly and not-scalable. So this is an open direction going forward.

Capabilities of Foundation Models

Vision



Capabilities of Foundation Models

Philosophy of Understanding

- **What can a foundation model understand about the data it is trained on?**
- **What is a foundation model?**
  - There isn't a precise definition as it is evolving. **One common characteristic in all of them: *self-supervision.***
  - **The sandwich contains peanut butter and jelly.**
  - **The sandwich contains peanut butter and jelly.**

Capabilities of Foundation Models

Philosophy of Understanding

- No obvious sense in which this kind of self-supervision tells the model anything about what the symbol mean. It's just about cooccurrence.
- But what does "peanut" mean? What does "jelly" mean? The model does not know.
- This might seem like a limitation of foundation models, but they can be trained with wide variety of symbols (computer code, database files, images, audio, sensor readings).
- **As long as it is just learning co-occurrence patterns of the sequences it is exposed to, then it counts as a foundation model by the authors' definition.**

Capabilities of Foundation Models

Philosophy of Understanding

- Why do we care what a foundation model can achieve?
  - **Trust:** Language is uniquely human, and can be used for deception and misrepresentation. In the context of language, understanding is a necessary for trust.
  - **Interpretability:** If genuine natural language understanding in some way involves maintaining and updating an internal model of the world (including, e.g., the speech context), and if we (as engineers) are able to analyze how linguistic input and output interface with this internal model, that could afford substantial gains in interpretability, predictability, and control of these systems.
  - **Accountability:** In the future, we might want it desirable to hold artificial agents in some way accountable for the lagnuage they produce. Depending on how we thing about accountability, responsibility, agency, etc., language understanding may emerge as a prerequisite.

So, it looks like “understanding” is an important thing.

Capabilities of Foundation Models

Philosophy of Understanding

- What is understanding?
- Distinction between the **metaphysics** and the **epistemology** of understanding.
- Metaphysics concerns what it would mean (“in principle”) for an agent to achieve understanding.
- Epistemology concerns how (“in practice”) we could ever come to know that an agent has achieved the relevant type of understanding.
- Metaphysics is more about out ultimate target whereas epistemology is more about how we could know when we have reached it.

Capabilities of Foundation Models

Philosophy of Understanding

- **Metaphysics of understanding:** The following three broad classes of views all have connections with research lines in AI and NLP. These are about what it is to understand natural language.
  - **Internalism:** Language understanding amounts to retrieval of the right internal representational structures in response to longusitic input.
  - **Referentialism:** An agent understand language when they are in a position to know what it would take for different sentences in that language to be true (relative to a context).
  - **Pragmatism:** Internal representations or truth are not fundamental. What matters is that the agent should have a disposition to use the language in the right way (inference, reasoning patterns, appropriate conversational moves, etc.)

Capabilities of Foundation Models

Philosophy of Understanding

- **Internalism** and **referentialism** are about mapping a linguistic sign to a meaning or a semantic value (either it's an internal representation or a truth value).
- If the input is only linguistic, then these mappings become difficult to achieve.
- If the input diverse digital traces of things in the world (images, audio, sensor), then the co-occurrence information in the linguistic input might be sufficient enough for the required mapping.

Capabilities of Foundation Models

Philosophy of Understanding

- **Bender and Koller [2020]**'s interesting argument:
  - Two humans are speaking language **L.**
  - An agent **O** intercepts this communication.
  - **O** inhabits a different world than the humans, so does not have the experiences needed to ground the humans' utterances in the ways that referentialism demands.
  - But **O** learns from the patterns in the humans' utterances and can pretend like them.
  - We can imagine situations in which **O's** inability to ground **L** in the humans' world will reveal itself, and that **O** does not understand **L.**
    - The complexity of the world is so great that no amount of textual exchange can fully cover it, and the gaps will eventually be revealed.
  - If the transmissions between the two humans included diverse input (images, audio, and sensor readings), and if **O** was able to understand the association between these and the linguistic input, then **we can be more optimistic.**
  - Authors believe that there is no **in-principal** limitation on what **O** can achieve.

Capabilities of Foundation Models

Philosophy of Understanding

- **Epistemology of understanding**: How we can hope to evaluate potential success?
- If we consider **pragmatism**, there is no concrete test for it. We just have to convince ourselves that our limited observations of the agent's behavior indicate a reliable dispositive toward the more general class of behaviors that we took as our target.
- Many artificial agents have passed the Turing Test, but none of them has been widely accepted as intelligent.
- In NLP, when systems surpass our estimates of human performance, our response is generally that the test was flawed, not that the target was reached.
- Then, maybe we had **internalism** or **referentialism** in our minds?

Capabilities of Foundation Models

Philosophy of Understanding

- If we take **internalism** or **referentialism** as our target, then behavioral tests will always be imperfect as a means of assessing whether understanding has been achieved.
  - Behavioral tests will always have gaps that could allow unsophisticated models to slip through.
  - A system might have achieved the mapping that these views require, but we may be unable to show this with behavioral testing.
    - In GPT-3, depending on the prompt that one uses, you can see a surprisingly coherenct output or complete non-sense.
- Need *structural* evaluation methods that allow us to study their internal representation, probing them for information, studying their internal dynamics, and actively manipulating them according to specific experimental protocols supporting causal inference.

Capabilities of Foundation Models

Philosophy of Understanding

- No easy answer to the question of whether or not foundation models will ever understand language.
- **Conclusion**: If foundation models are pursued as a path to language understanding in artifical agents, then multimodel training regimes may well be the most viable strategy, as they would seem the most likely to provide the model with the requisite information.
- Whether self-supervision then suffices is a completely open question.

Applications

- **Healthcare and biomedicine**
  - diagnosis, treatment, summarization of patient records, question answering, assistive care, personalized medicine, drug discovery, clinical trials
- **Law**
  - contract review, patent retrieval, multimodal evidence, arguments crafting, dialogue agents, predicting judge questions, adaptation to writing style, adaptation to new contexts, etc.
- **Education**
  - understand students (identity, state, motivation, skills, etc), understand educators (teachers and education materials), understand learning (track and analyze progression and performance), understand teaching (model cognition, enable interaction adaptive teaching), understand subject matter

Technologies

- Modeling
- Training
- Adaptation
- Evaluation
- Systems
- Data
- Security and Privacy
- Robustness to distribution shifts
- AI Safety and Alignment
- Theory
- Interpretability

Technologies

Modeling

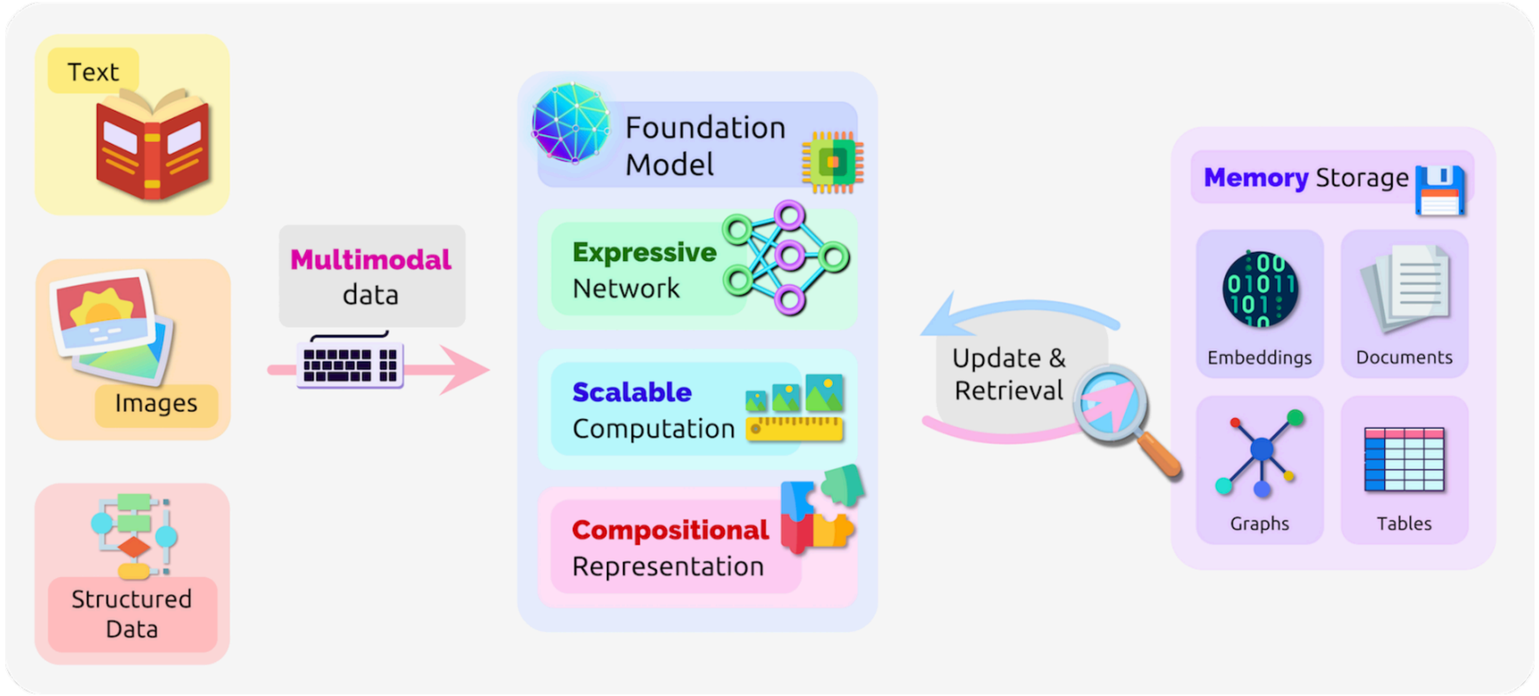


Fig. 17. The five key properties of a foundation model: *expressivity* — to flexibly capture and represent rich information; *scalability* — to efficiently consume large quantities of data; *multimodality* — to connect together various modalities and domains; *memory capacity* — to store the vast amount of accumulated knowledge; and *compositionality* — to generalize to new contexts, tasks and environments.

Technologies

Expressivity

- **Expressivity** concerns with the theoretical and practical capacity of a network to model the data distribution it is trained over and represent it in a flexible manner.
- **Inductive Biases:** The recent success of neural networks in modeling natural data is owed to **high depths** (number of non-linear layers or number of computational steps).
  - The Universal Approx. Theorem indicates that even simple MLPs can represent a broad range of functions, while different **inductive biases,** as in RNNs and CNNs, can improve learning efficiency and ability to model different forms of information.

Technologies

Expressivity

- **Transformer Networks and Attention:** The recent transformer networks demonstrate the importance of capturing **long-range dependencies and pairwise/higher-order interactions between elements.**
  - They use self-attention that enables shorter computation paths and provides direct means to compare elements far-across the input data.
  - These provide an alternative to the fixed-weight computation of MLPs and CNNs: **dynamically adapting the computation to the input at hand.**
    - **She ate the ice-cream with the [spoon, strawberries]**
      - In both cases, a feed-forward network will process it the same way.
      - An attention-based model will update the representation of the word “ate” if the missing phrase is **spoon**, or of the word “ice cream” if it is **strawberries.**

Technologies

Expressivity

- **General-Purpose Computation:** A final notable advantage of attention over prior architectures stems from its stronger generality, where it is not strongly tied to a particular task or domain.
- Authors hypothesize that the general-purpose nature of attention and transformers contributes to their broad applicability for a wide range of research problems and applications.
- This contrast captures a more general trade-off between **task-specialization and expressivit**y: models with stronger structural priors can leverage them to improve sample efficiency on the particular tasks that benefit from these assumptions; while conversely, models that integrate weaker inductive biases learn more slowly, but can in turn scale to higher volumes of data and adapt to a diverse set of domains, since they do not rely on restrictive or task-specific suppositions. **As both data and compute turn more accessible, we observe that the exploration of models with a minimal set of inductive biases that can “let the data speak for itself" seems to serve as a more promising approach for future research in the field.**

Technologies

Expressivity

- **Challenges**
  - Modeling extremely long-range dependencies (books, movies, DNA sequences)
    - Explicit modeling through short and direct computation paths improves expressivity, but scalability becomes a problem.
    - Identifying and effective equilibrium between efficiency and and expressivity is an interesting direction.

Technologies

Scalability

- **Optimization:** Foundation models should be
  - **easy-to-train** (robust to noise and imperfect data, and to vanishing)
  - **easy-to-adapt** (overcome the notion of catastrophic forgetting and support few-shot learning)
  - Still in the early days of understanding what drives scalability of learning algorithms.
- **Hardware Compatibility:** Foundation models should be practically efficient.
  - Parallelizability (Transformer)
  - Distributed training