Introduction / Course Overview

DSCC 251/451: Machine Learning with Limited Data

C.M. Downey

Spring 2026

Introduction: LeCun's Cake and the Data-Scarcity Problem

Machine Learning as Cake

- According to "Deep Learning Mafioso" Yann LeCun:
  - **Supervised learning** (what most ML courses teach) is the cake's **frosting**
  - **Unsupervised (or self-supervised) learning** is the **body** of the cake (aka the sponge)
  - **Reinforcement learning** is the **cherry on top**

What does the Cake Mean?

What does the Cake Mean?

- Most ML courses focus on **supervised learning**
  - Trained to predict a **label, number, or annotation** from the raw data
  - Requires **pairings** of raw data and annotations
  - In reality, this type of data is **scarce**

What does the Cake Mean?

- Most ML courses focus on **supervised learning**
  - Trained to predict a **label, number, or annotation** from the raw data
  - Requires **pairings** of raw data and annotations
  - In reality, this type of data is **scarce**
- **Raw data** (e.g. text, audio, images) is **much more plentiful**
  - **Unsupervised** and **self-supervised learning** take advantage of **raw data alone!**

(Un)Supervised Learning

(data)

supervised:

```
Cats like to look out windows
            ↓
      topic = cats
```

(annotation)

---

self-supervised:

```
Cats like to look out __
                       ↓
                    windows

__ like to __ out windows
↓          ↓
Cats       look
```

Core Ideas

Core Ideas

- Traditional ML courses focus **mostly on the icing** (supervised learning)
  - In reality, we **rarely have much icing** to work with

Core Ideas

- Traditional ML courses focus **mostly on the icing** (supervised learning)
  - In reality, we **rarely have much icing** to work with
- Successful ML pipelines usually **leverage unlabeled data too** (w/ unsupervised or self-supervised learning)
  - This is the **cake body** because there's (usually) **much more** unlabeled data to work with
  - Often represents **massive amounts of raw data**, which might be incorporated into a **foundation model** extensively trained with self-supervision
  - This raw data might **not** be directly related to your end task (**Transfer Learning**)

Core Ideas

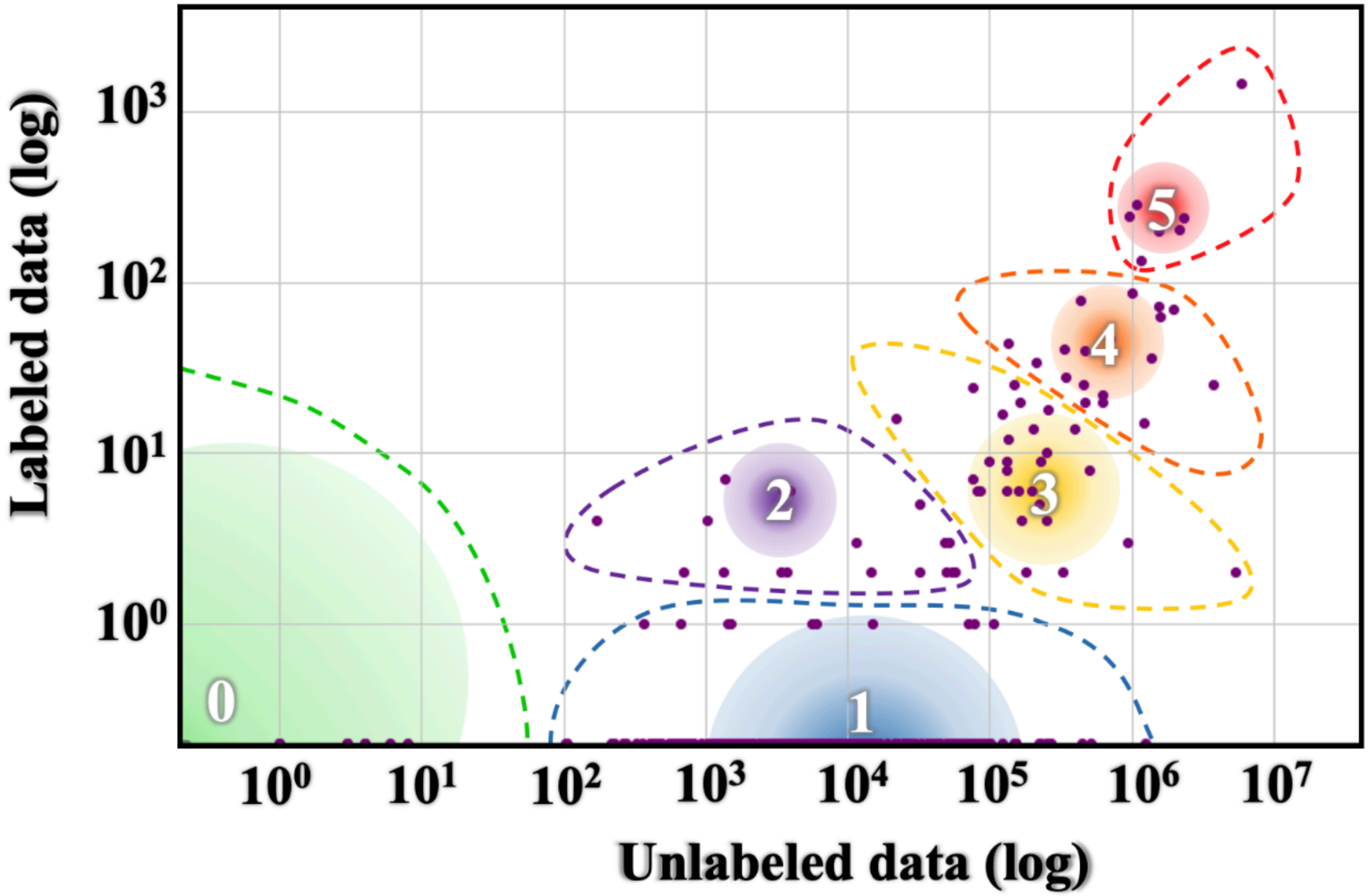
- Traditional ML courses focus **mostly on the icing** (supervised learning)
  - In reality, we **rarely have much icing** to work with
- Successful ML pipelines usually **leverage unlabeled data too** (w/ unsupervised or self-supervised learning)
  - This is the **cake body** because there's (usually) **much more** unlabeled data to work with
  - Often represents **massive amounts of raw data**, which might be incorporated into a **foundation model** extensively trained with self-supervision
  - This raw data might **not** be directly related to your end task (**Transfer Learning**)
- We'll also see how to **stretch the icing further**
  - i.e. make **efficient use of labeled data**

Data-scarcity in Practice

Data Scale in NLP

data availability by language (GB)

Joshi et al. (2020)

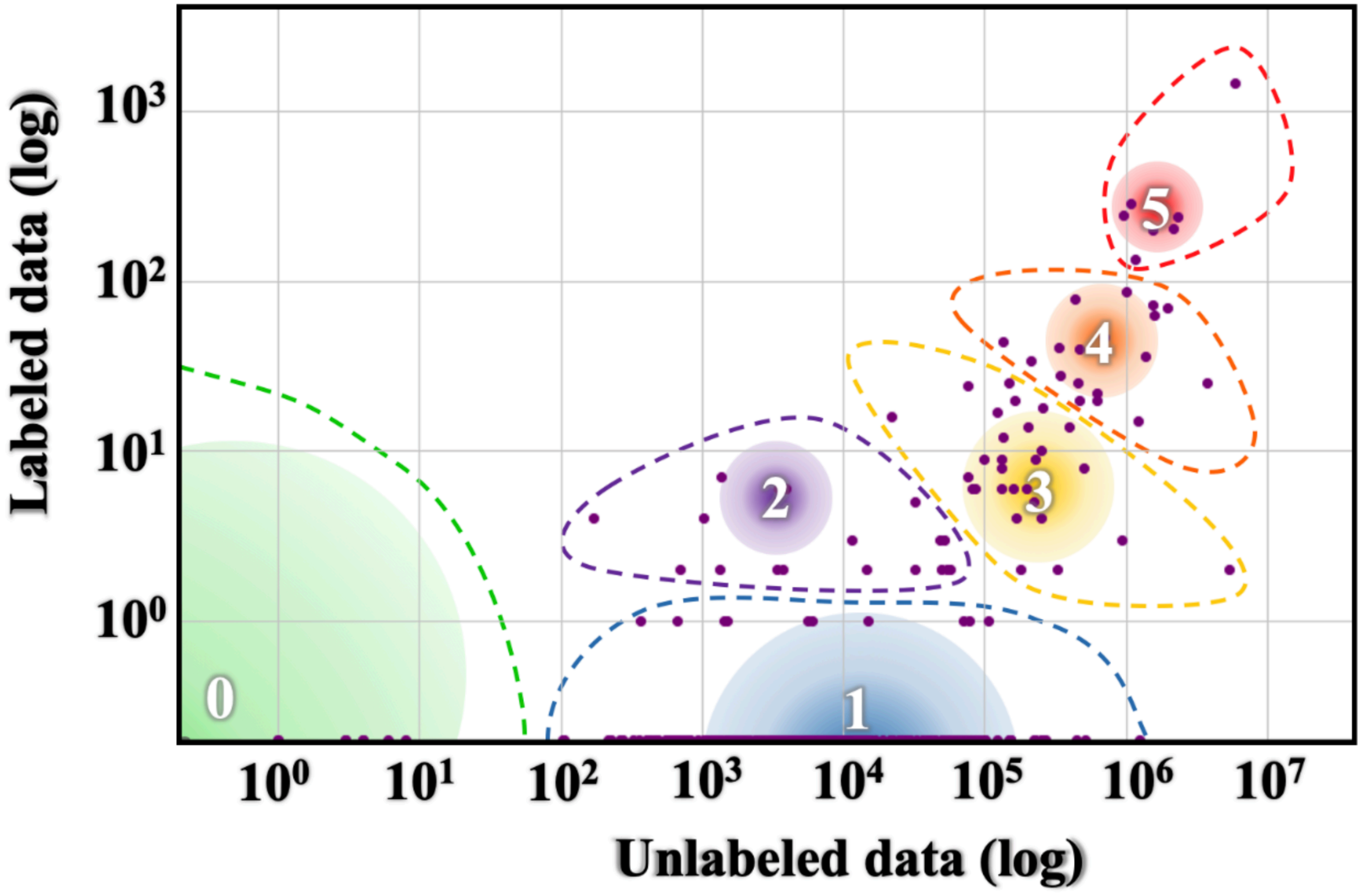


Data Scale in NLP

- NLP has undergone a revolution that is **fueled by model and data scale**

data availability by language (GB)

Joshi et al. (2020)

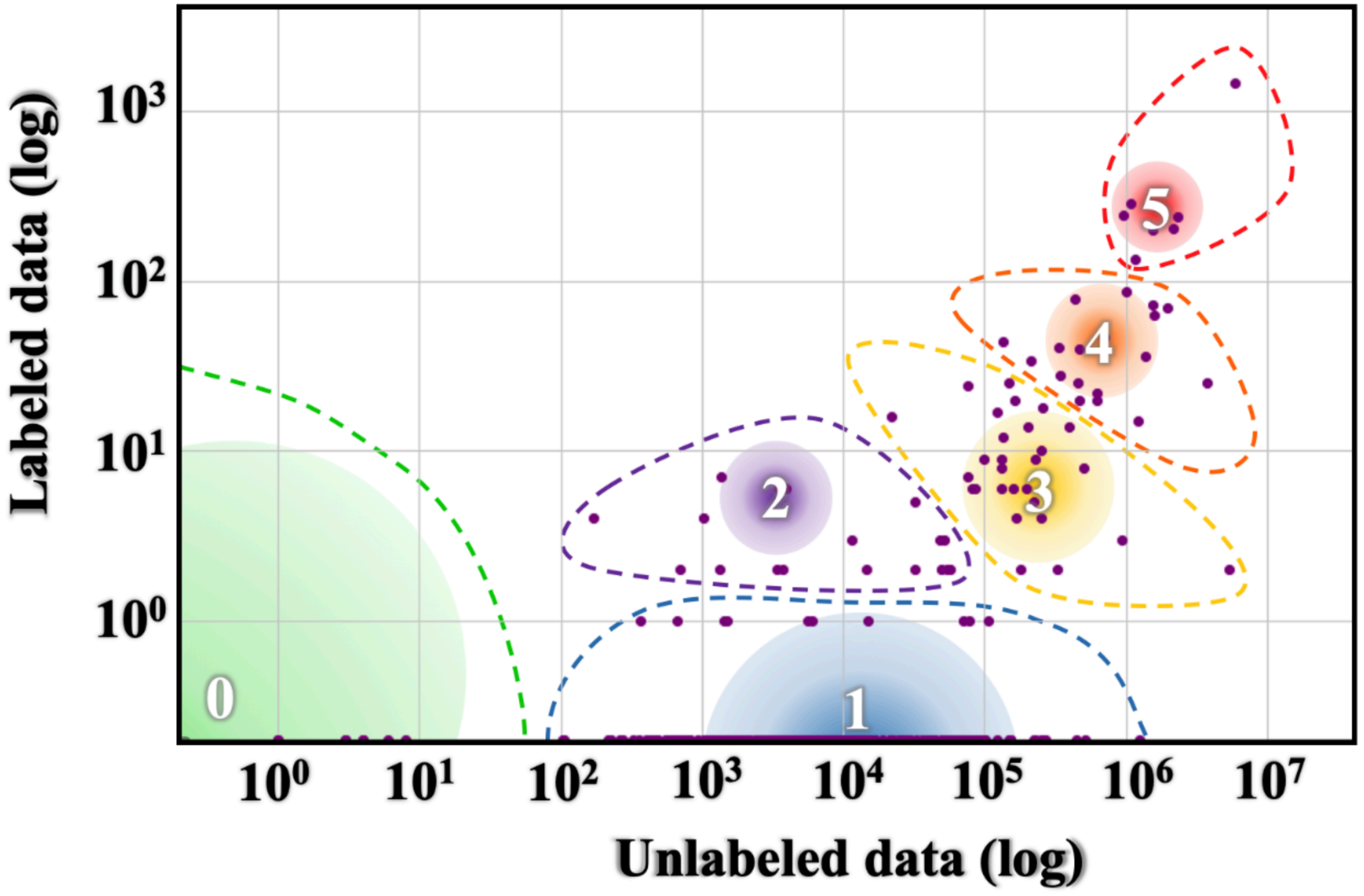


Data Scale in NLP

- NLP has undergone a revolution that is **fueled by model and data scale**
- Major breakthroughs of the past 5 years have been **limited to English** and a few other very high-resource languages

data availability by language (GB)

Joshi et al. (2020)

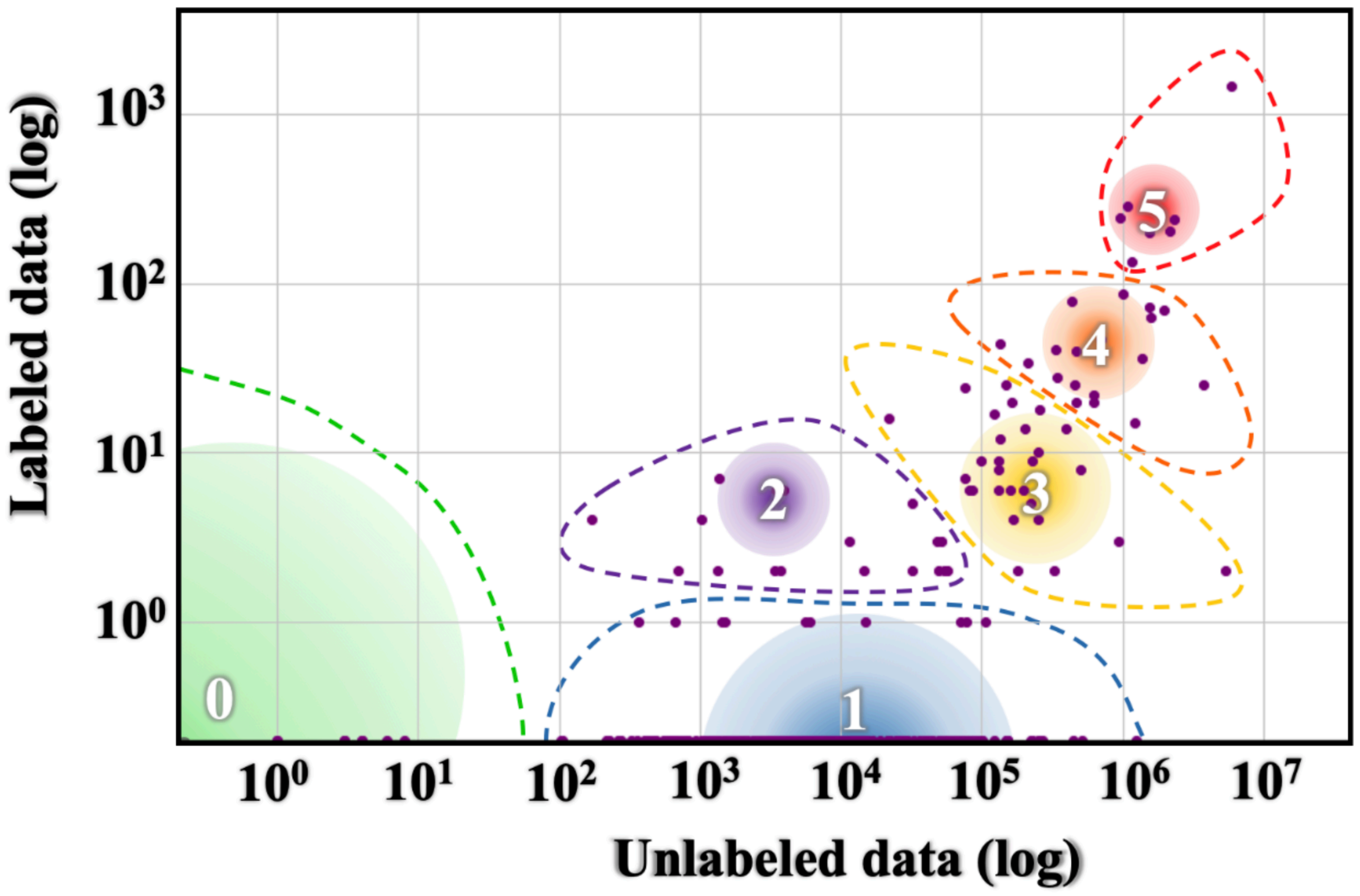


Data Scale in NLP

- NLP has undergone a revolution that is **fueled by model and data scale**
- Major breakthroughs of the past 5 years have been **limited to English** and a few other very high-resource languages
- Scale-driven techniques are **inapplicable to the majority of world languages**

data availability by language (GB)

Joshi et al. (2020)

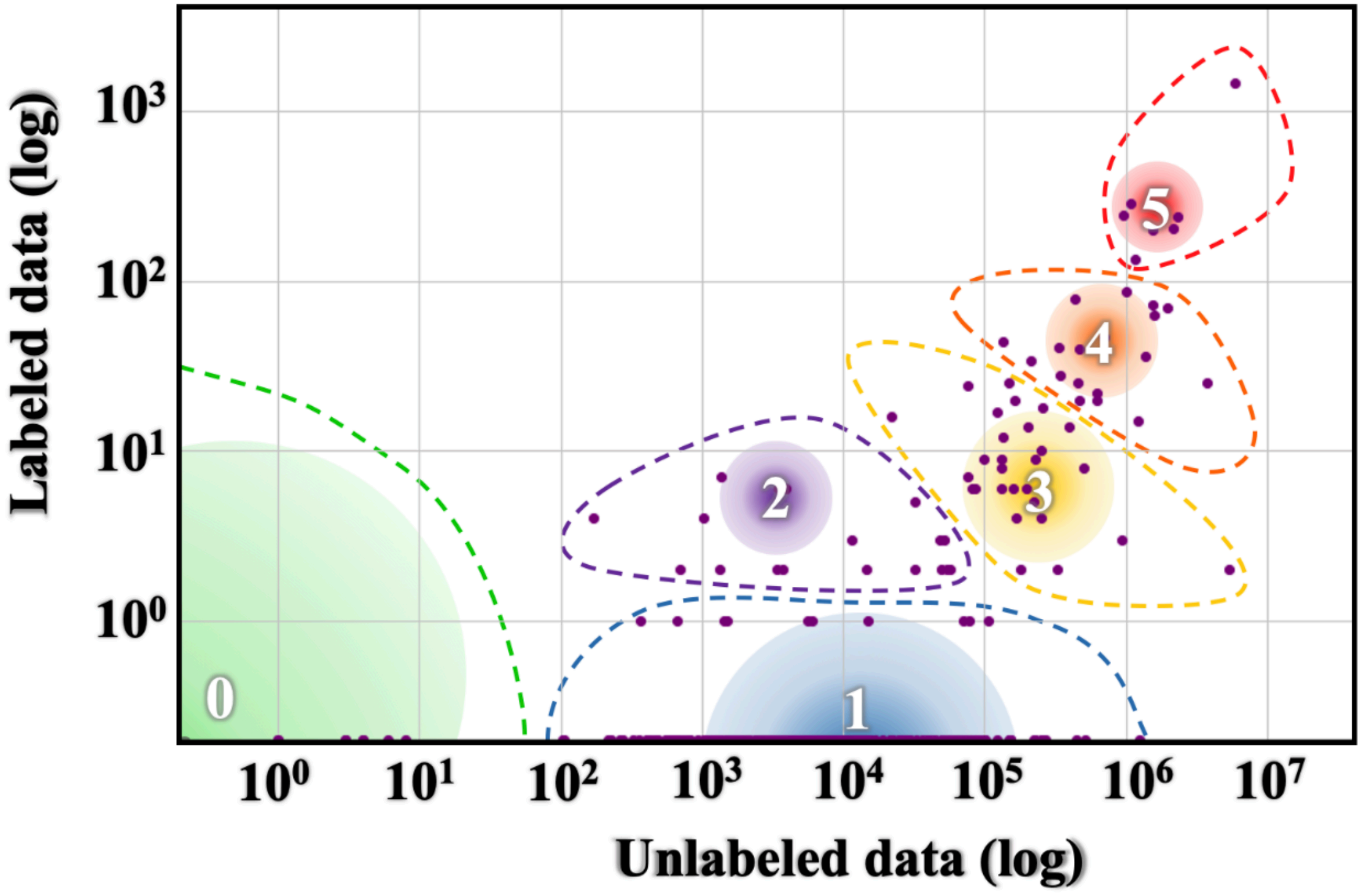


Data Scale in NLP

- NLP has undergone a revolution that is **fueled by model and data scale**
- Major breakthroughs of the past 5 years have been **limited to English** and a few other very high-resource languages
- Scale-driven techniques are **inapplicable to the majority of world languages**
- Joshi et al. (2020) introduce a **classification system** for language data availability

data availability by language (GB)

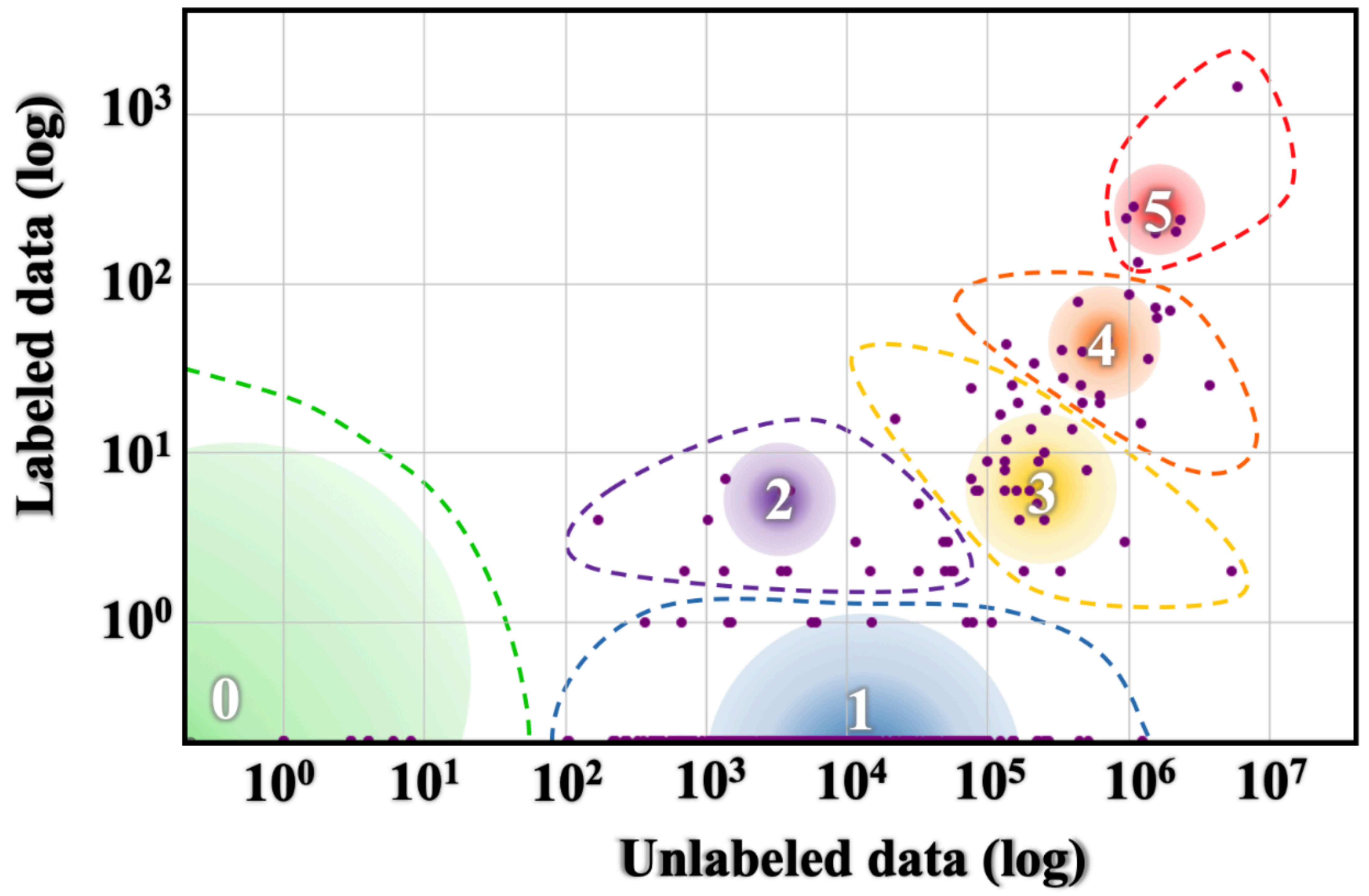
Joshi et al. (2020)



Low-resource Languages

data availability by language (GB)

Joshi et al. (2020)

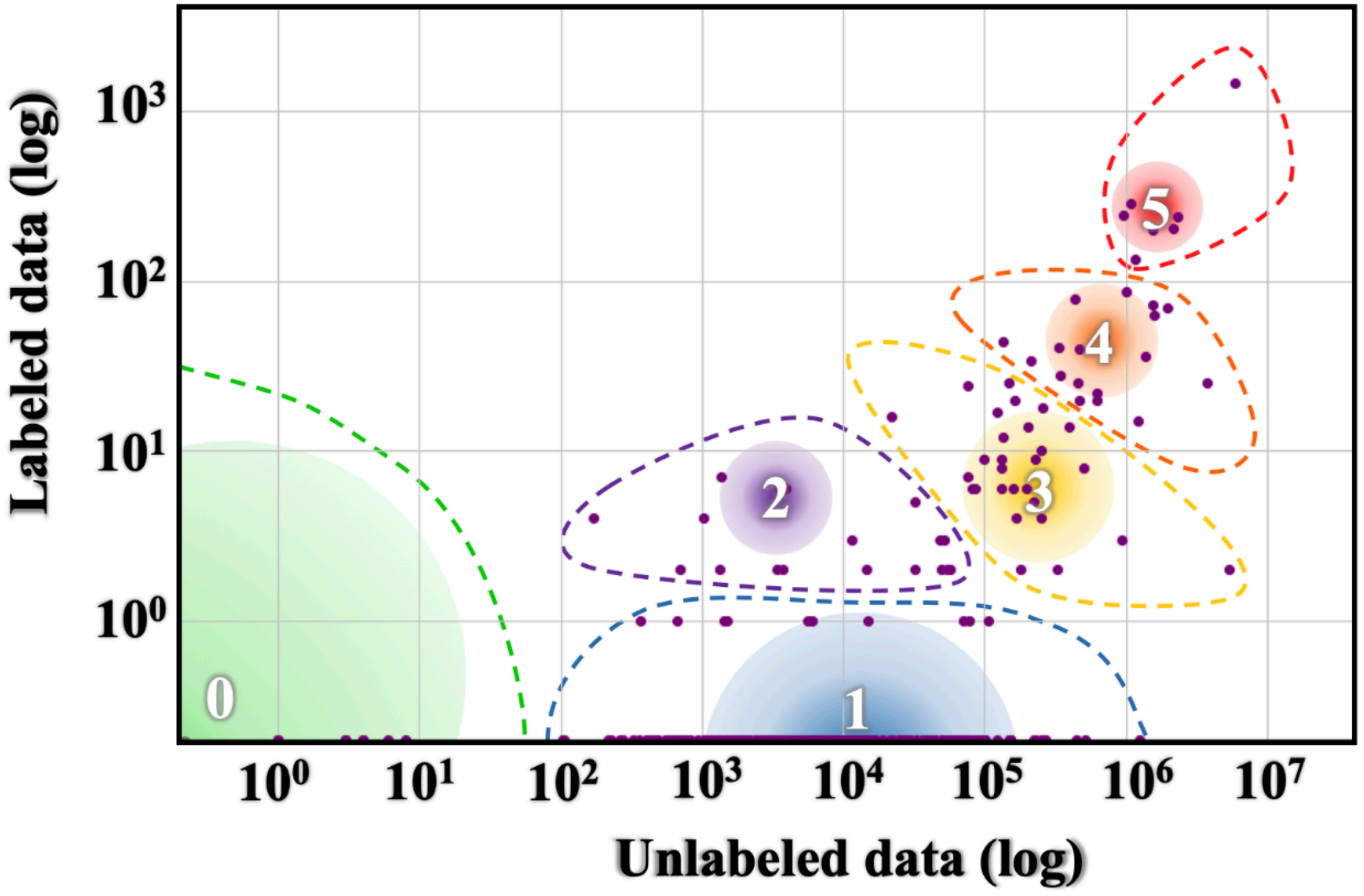


Low-resource Languages

- **2,500 languages** spoken by **3 billion people** fall into resource class 3 or lower

data availability by language (GB)

Joshi et al. (2020)

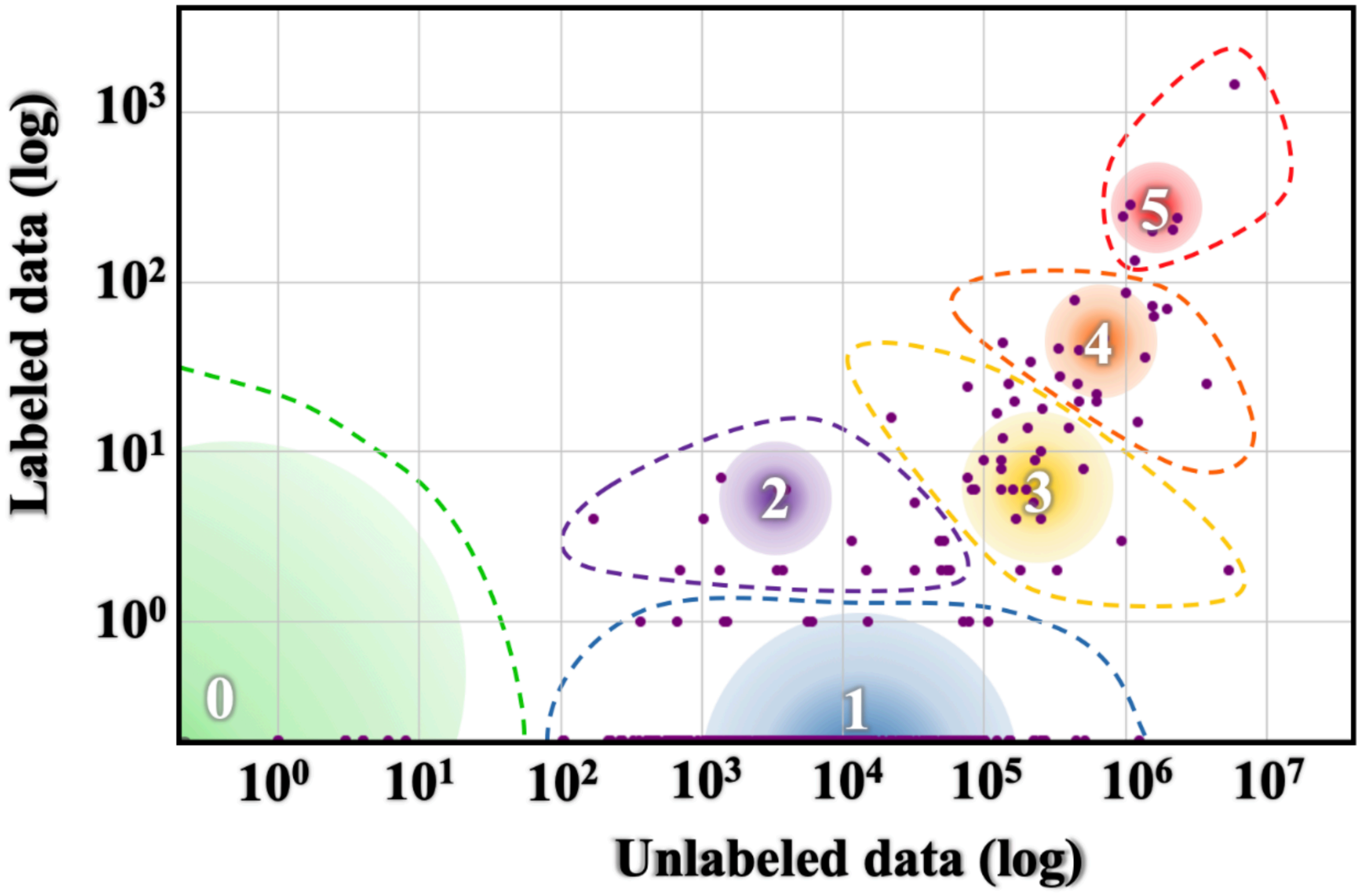


Low-resource Languages

- **2,500 languages** spoken by **3 billion people** fall into resource class 3 or lower
- Class 3 example: Urdu
  - Official language of Pakistan with **230 million speakers**, but considered **"low-resource"**
  - Fair amount of **unlabeled data** (i.e. raw text)

data availability by language (GB)

Joshi et al. (2020)

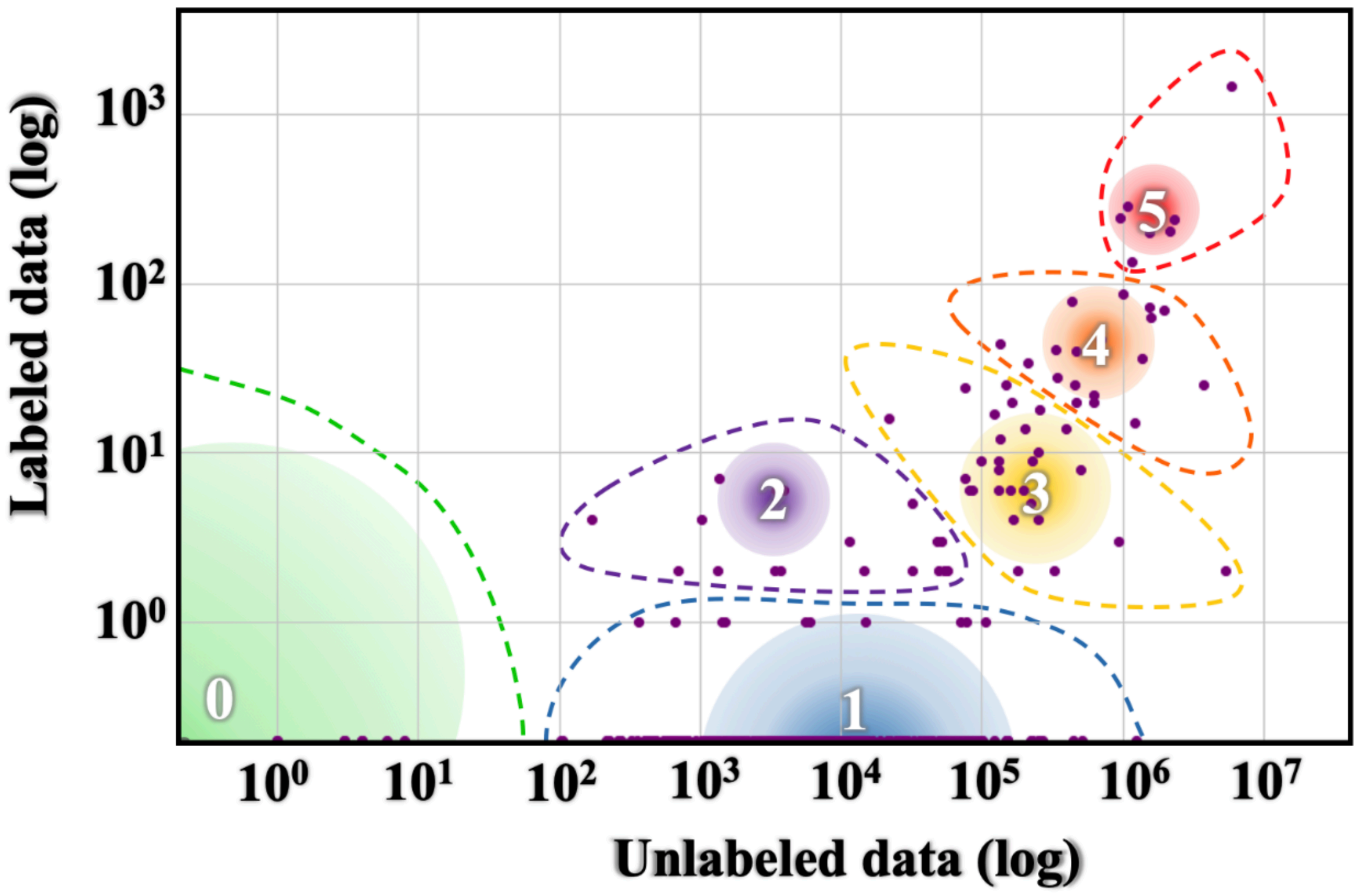


Low-resource Languages

- **2,500 languages** spoken by **3 billion people** fall into resource class 3 or lower
- Class 3 example: Urdu
  - Official language of Pakistan with **230 million speakers**, but considered **"low-resource"**
  - Fair amount of **unlabeled data** (i.e. raw text)
- Class 0 examples: Indigenous and endangered languages
  - **Even raw text** is scarce

data availability by language (GB)

Joshi et al. (2020)

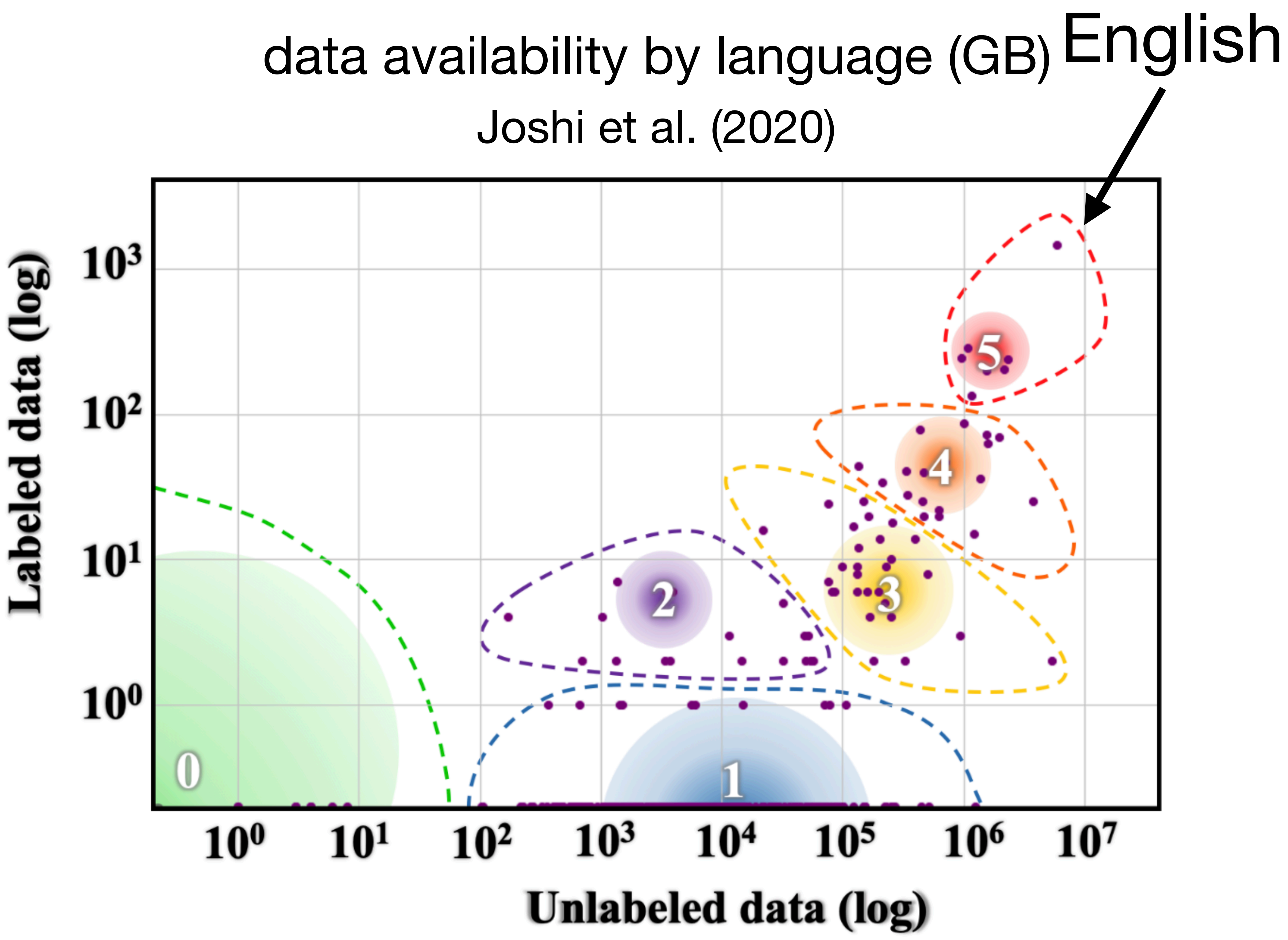


Low-resource Languages

- **2,500 languages** spoken by **3 billion people** fall into resource class 3 or lower
- Class 3 example: Urdu
  - Official language of Pakistan with **230 million speakers**, but considered **"low-resource"**
  - Fair amount of **unlabeled data** (i.e. raw text)
- Class 0 examples: Indigenous and endangered languages
  - **Even raw text** is scarce

data availability by language (GB)

Joshi et al. (2020)



Automatic Speech Recognition (ASR)

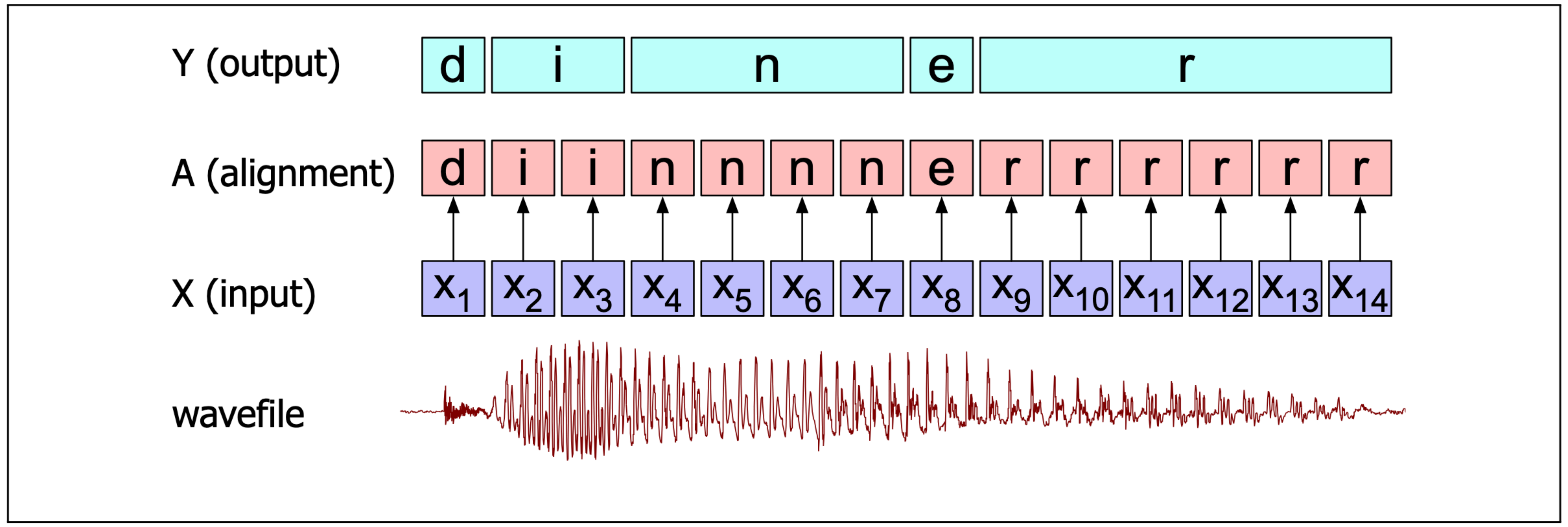


Figure 15.12 A naive algorithm for collapsing an alignment between input and letters.

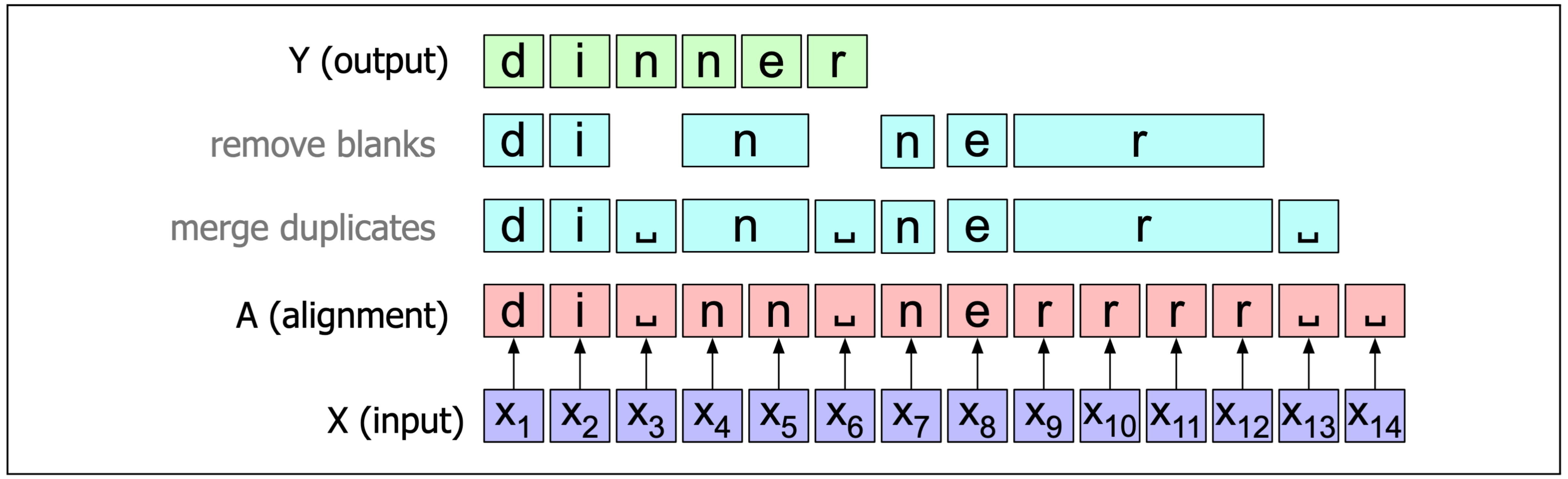


Figure 15.13 The CTC collapsing function B, showing the space blank character ␣; repeated (consecutive) characters in an alignment A are removed to form the output Y.

Automatic Speech Recognition (ASR)

- **Supervised** NLP task
  - **Input:** raw audio
  - **Output:** text transcription

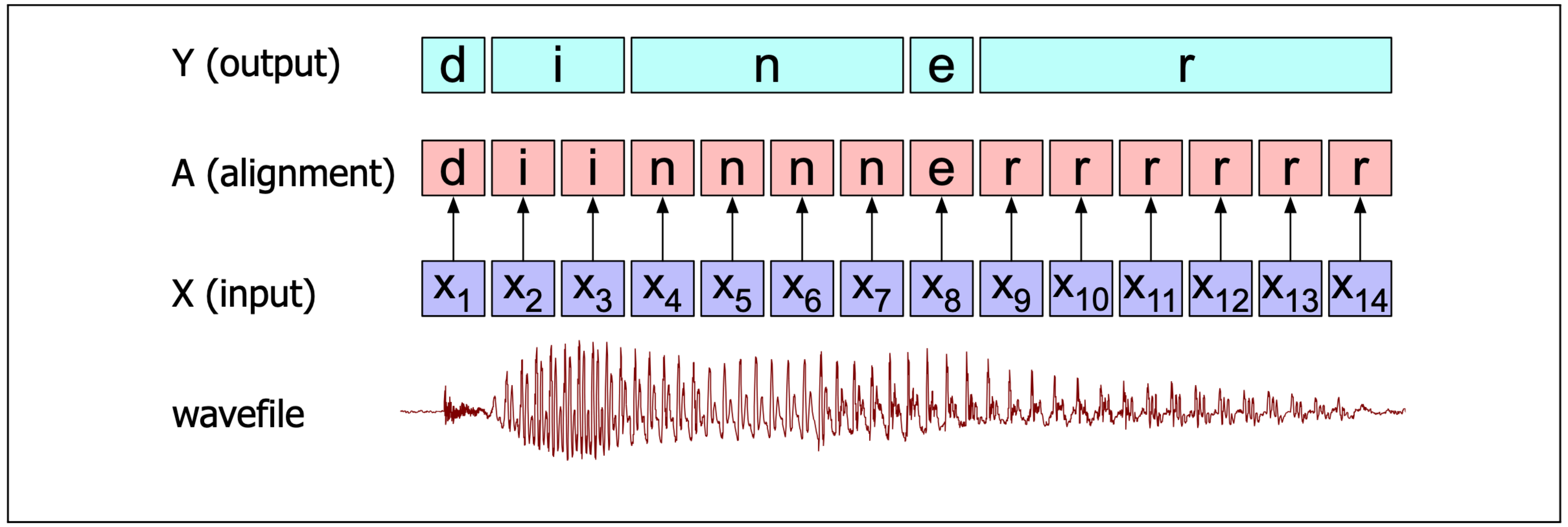


Figure 15.12 A naive algorithm for collapsing an alignment between input and letters.

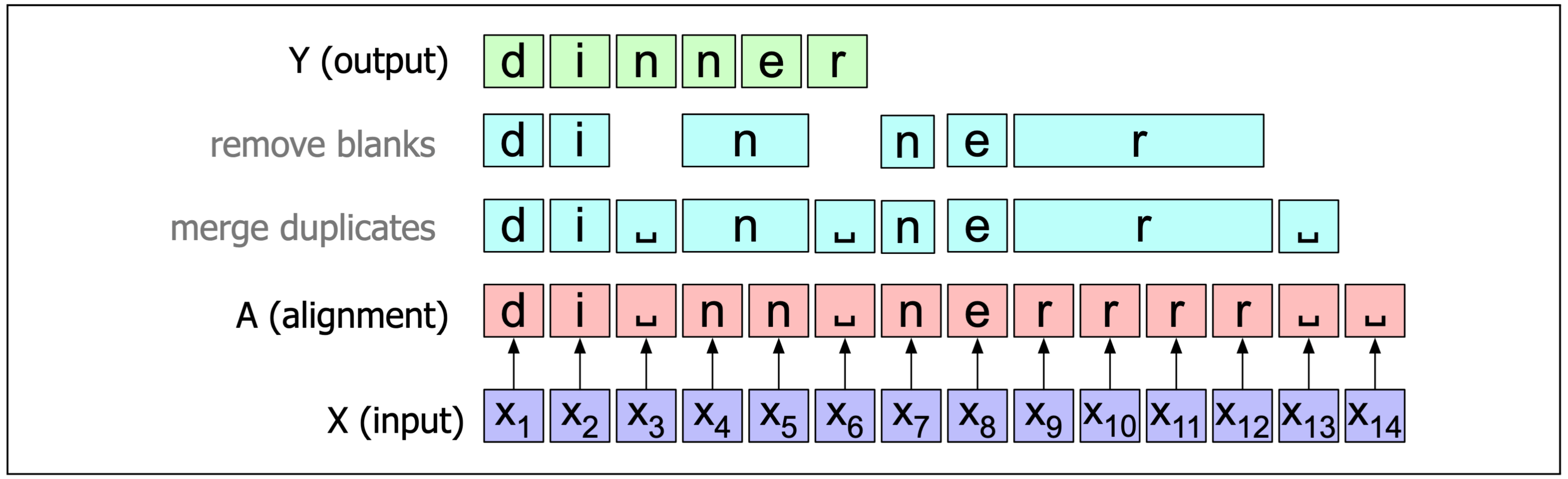


Figure 15.13 The CTC collapsing function B, showing the space blank character ␣; repeated (consecutive) characters in an alignment A are removed to form the output Y.

Automatic Speech Recognition (ASR)

- **Supervised** NLP task
  - **Input:** raw audio
  - **Output:** text transcription
- Performs at **near-human-level** for English
  - For low-resource languages... **pretty poor / unusable**

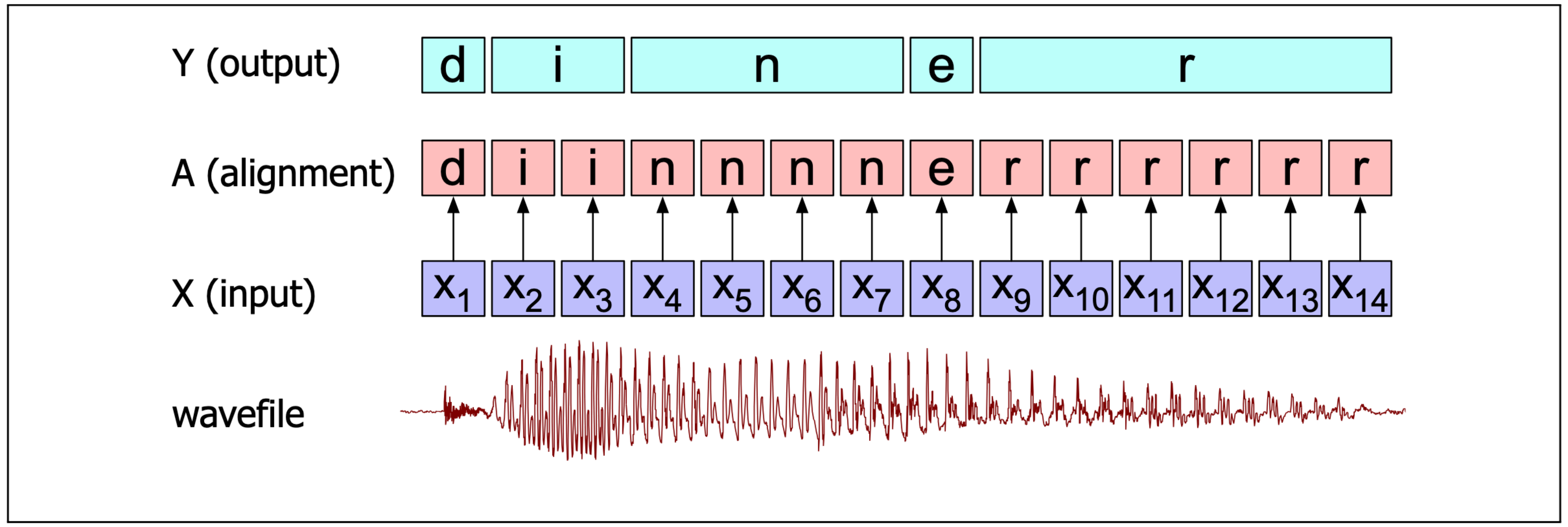


Figure 15.12 A naive algorithm for collapsing an alignment between input and letters.

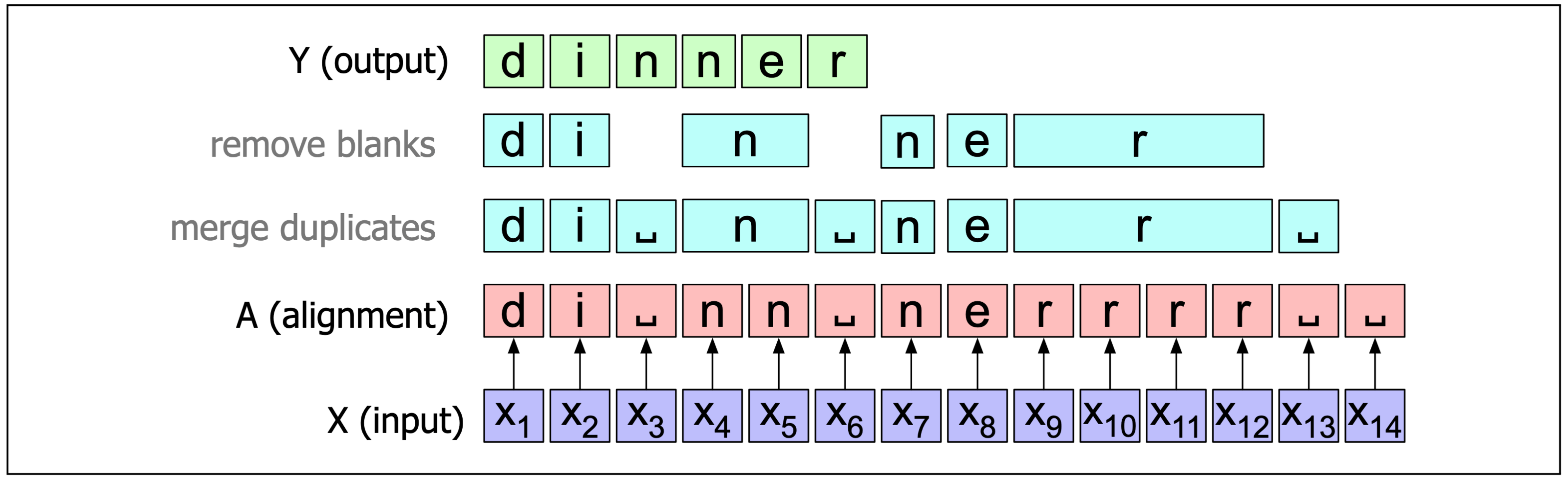


Figure 15.13 The CTC collapsing function B, showing the space blank character ␣; repeated (consecutive) characters in an alignment A are removed to form the output Y.

Automatic Speech Recognition (ASR)

- **Supervised** NLP task
  - **Input:** raw audio
  - **Output:** text transcription
- Performs at **near-human-level** for English
  - For low-resource languages... **pretty poor / unusable**
- Why the disparity?

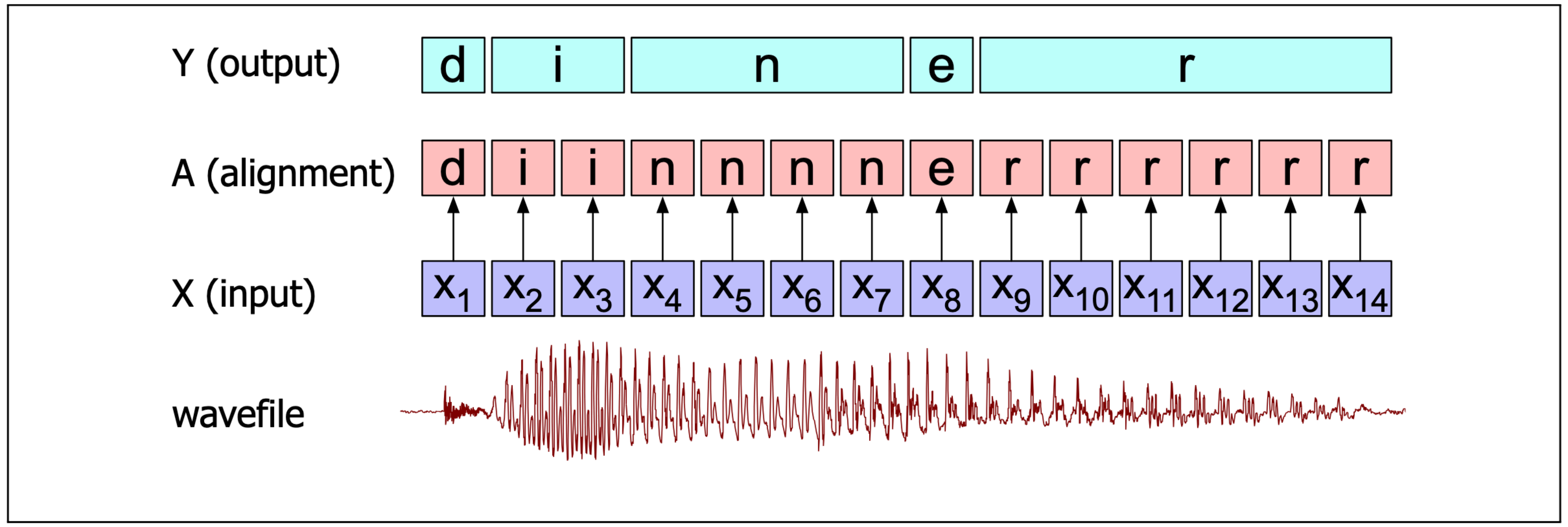


Figure 15.12 A naive algorithm for collapsing an alignment between input and letters.

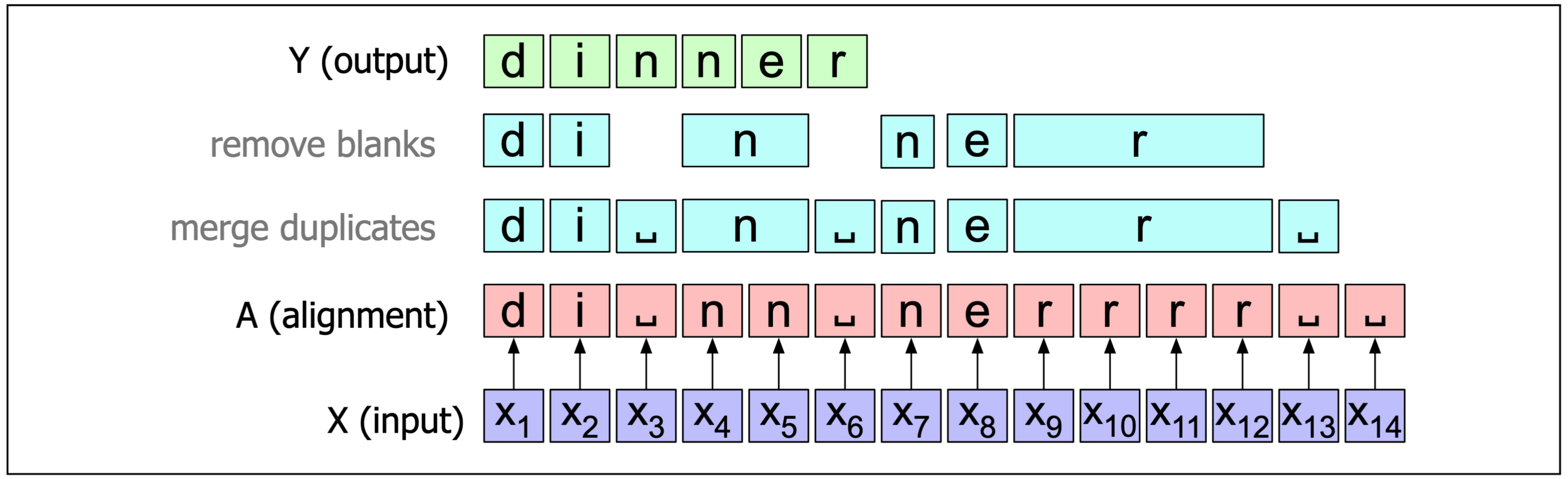
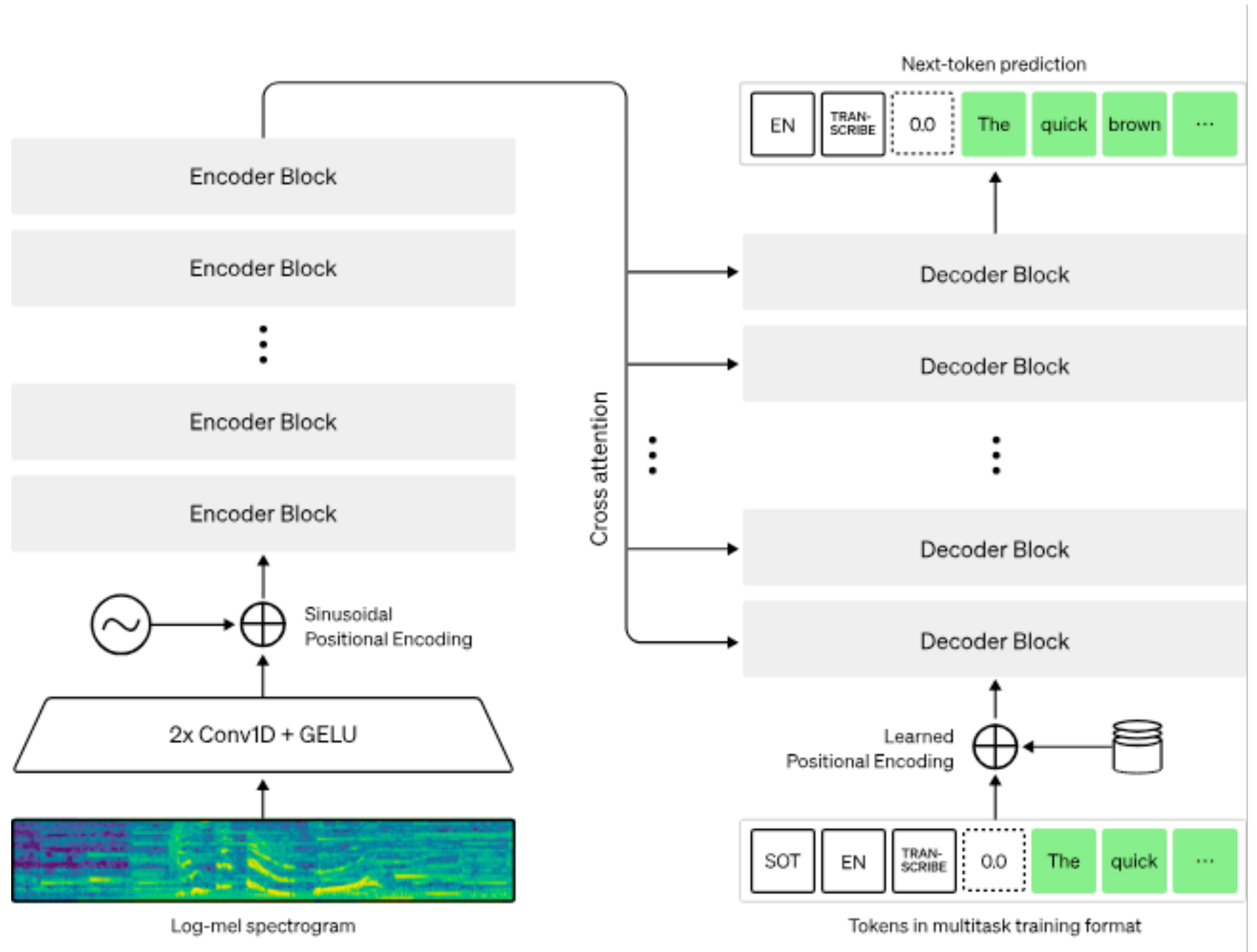


Figure 15.13 The CTC collapsing function B, showing the space blank character ␣; repeated (consecutive) characters in an alignment A are removed to form the output Y.

ASR Data Needs

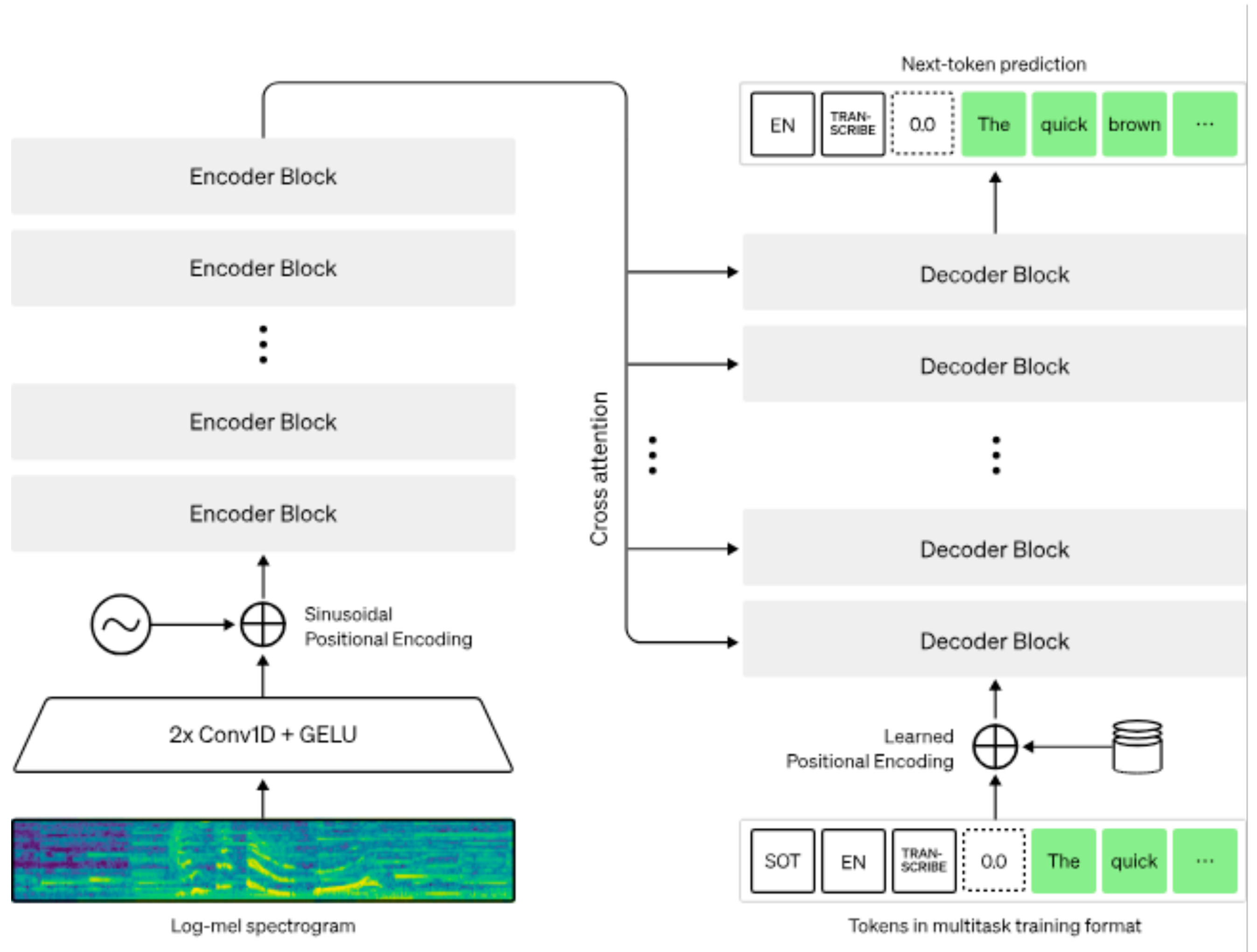
Bird (2020)



ASR Data Needs

- (Supervised) ASR training requires **paired audio + transcriptions**
  - Usually **expensive to curate** (1hr audio ≈ 4-5 hrs transcription work!)

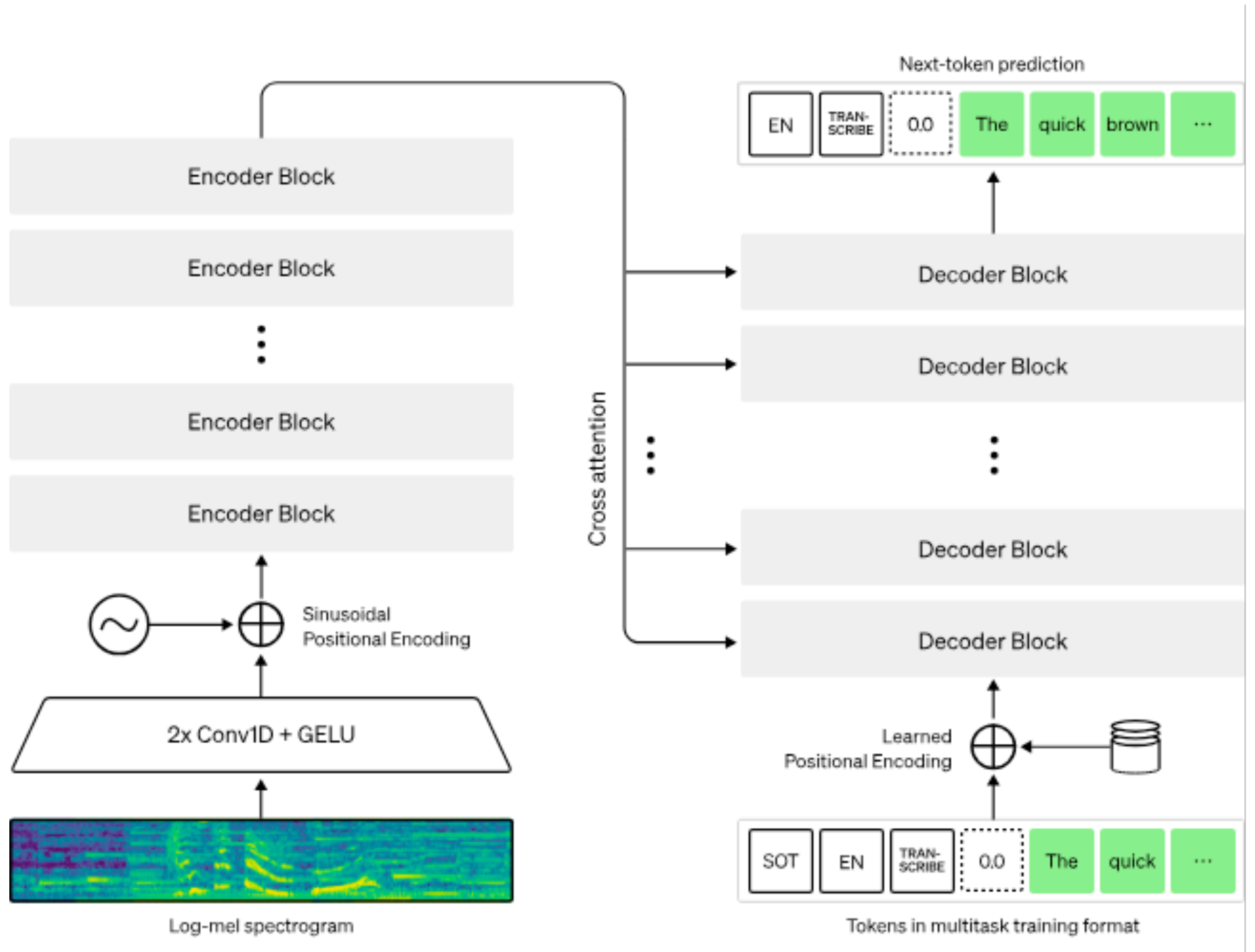
Bird (2020)



ASR Data Needs

- (Supervised) ASR training requires **paired audio + transcriptions**
  - Usually **expensive to curate** (1hr audio ≈ 4-5 hrs transcription work!)
- SOTA English models trained on **~1k-500k hours of paired data!**
  - Low-resource languages might have **~10hrs if you're lucky!**

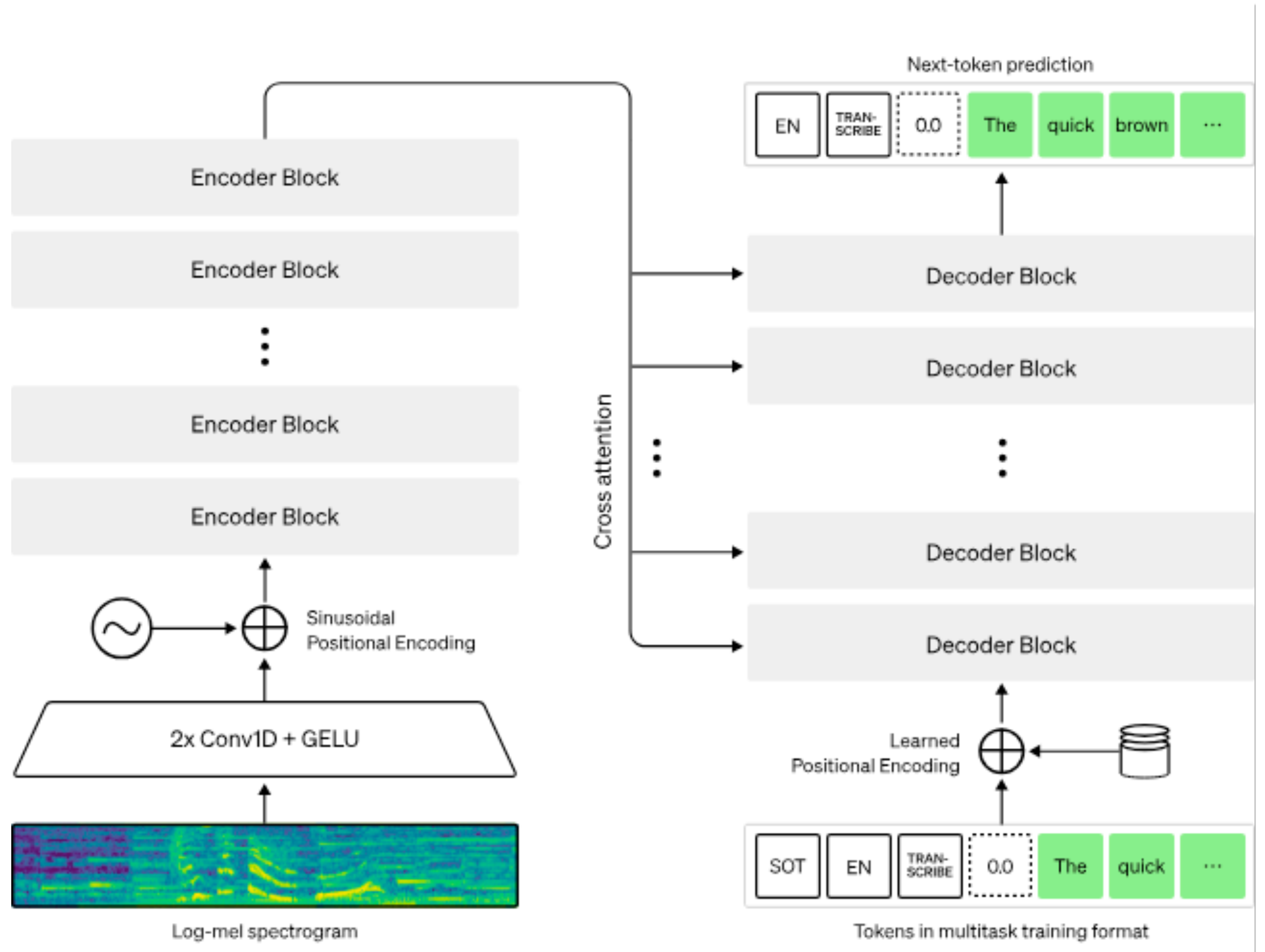
Bird (2020)



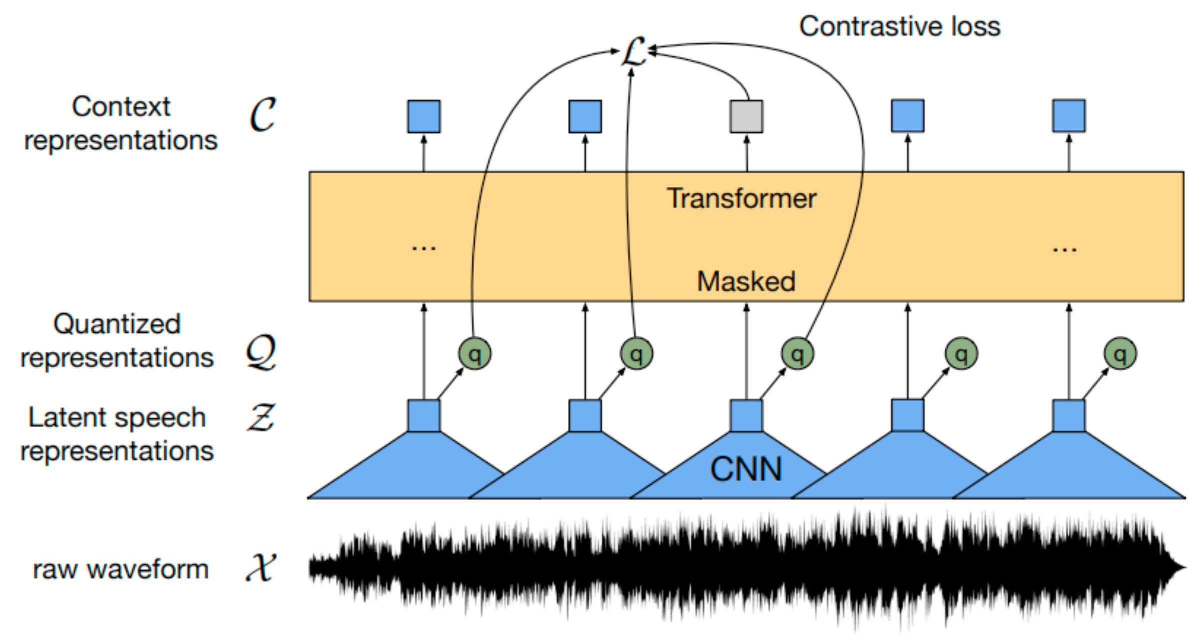
ASR Data Needs

- (Supervised) ASR training requires **paired audio + transcriptions**
  - Usually **expensive to curate** (1hr audio ≈ 4-5 hrs transcription work!)
- SOTA English models trained on **~1k-500k hours of paired data!**
  - Low-resource languages might have **~10hrs if you're lucky!**
- How do you close the gap?

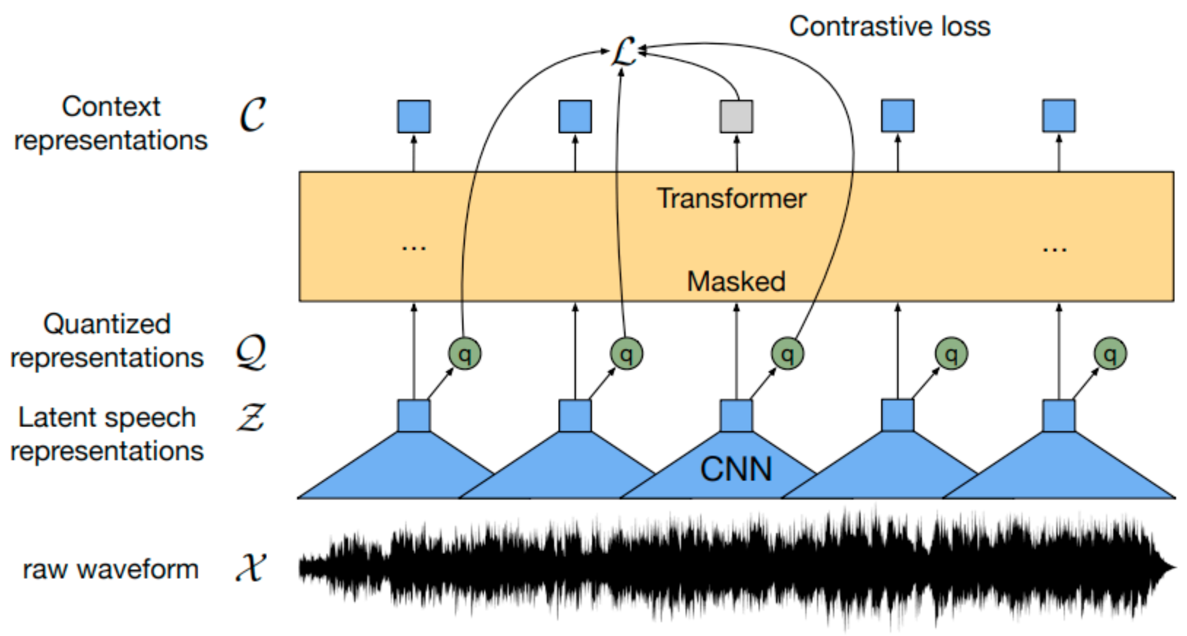
Bird (2020)



Traps of Limited Data

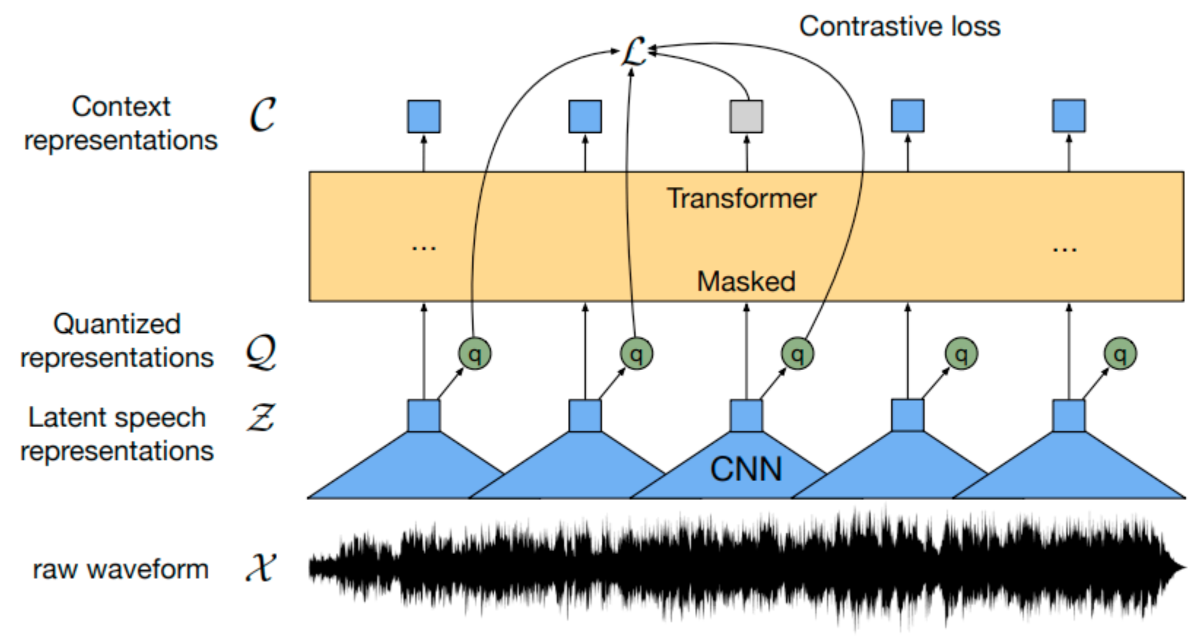


Traps of Limited Data



- Training an ASR model **"from scratch"** probably won't work
  - There's simply **not enough signal** to take advantage of
  - Usually start from a **self-supervised foundation model** (e.g. wav2vec)
  - Then **fine-tune** the model for the new language

Traps of Limited Data



- Training an ASR model **"from scratch"** probably won't work
  - There's simply **not enough signal** to take advantage of
  - Usually start from a **self-supervised foundation model** (e.g. wav2vec)
  - Then **fine-tune** the model for the new language
- Even with a foundation model, will likely **overfit** to the limited data
  - Might only have **2-3 speakers**, and the system will **fail to generalize** to new voices!
  - Variation in **accents, speech rates, recording conditions** makes the problem even harder!

Low-resource ASR Approach

Low-resource ASR Approach

- Start with a **self-supervised foundation model** (learns from **raw audio**)

Low-resource ASR Approach

- Start with a **self-supervised foundation model** (learns from **raw audio**)
- In your target language, find **as much raw audio** as you can
  - e.g. radio, podcasts, movies, academic recordings

Low-resource ASR Approach

- Start with a **self-supervised foundation model** (learns from **raw audio**)
- In your target language, find **as much raw audio** as you can
  - e.g. radio, podcasts, movies, academic recordings
- Continue **self-supervised training** with your target language audio

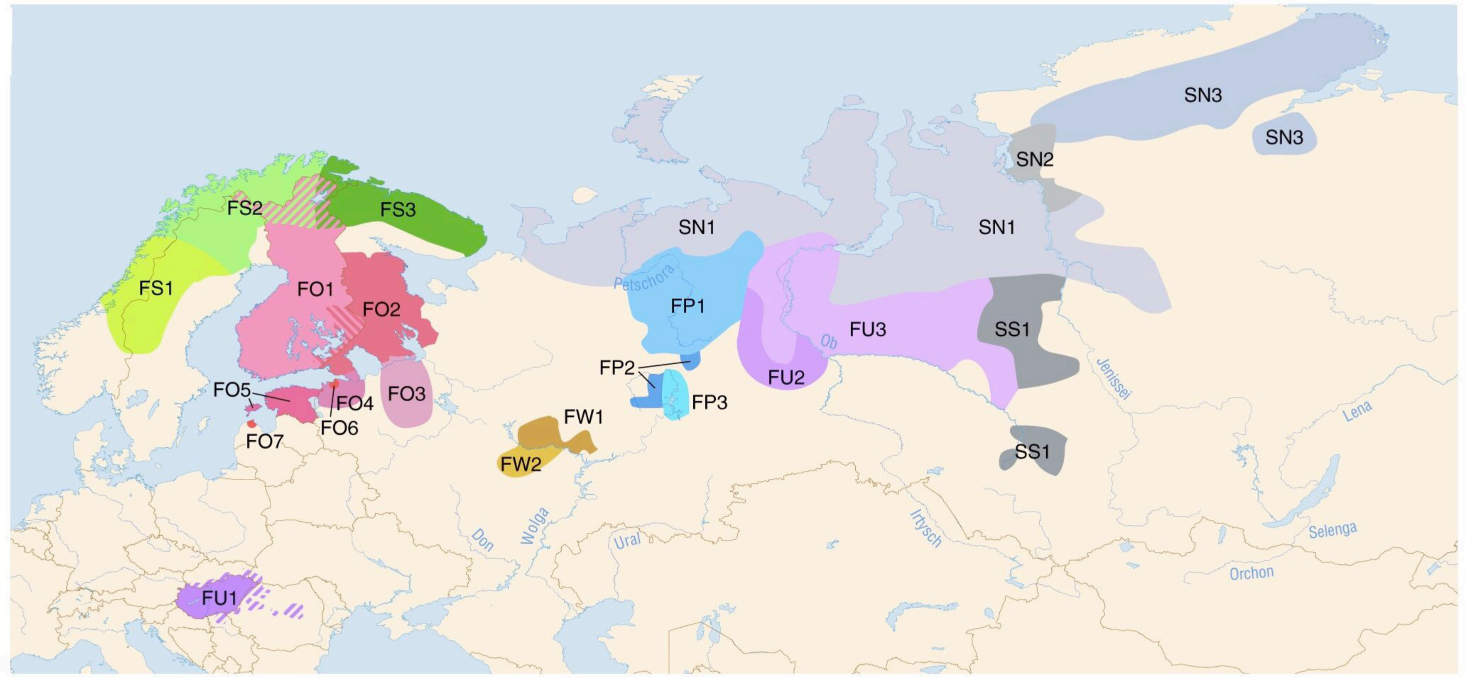
Low-resource ASR Approach

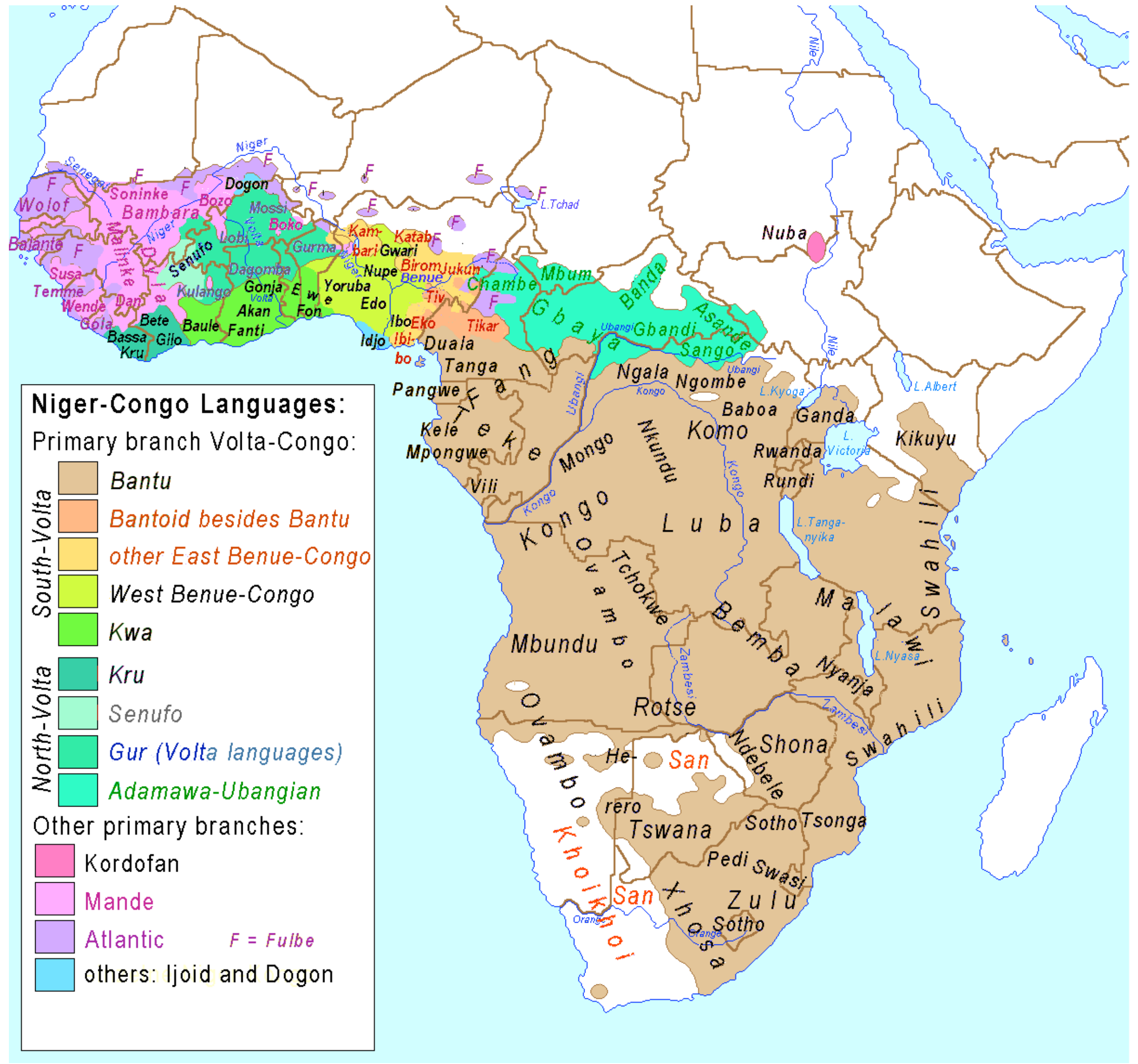
- Start with a **self-supervised foundation model** (learns from **raw audio**)
- In your target language, find **as much raw audio** as you can
  - e.g. radio, podcasts, movies, academic recordings
- Continue **self-supervised training** with your target language audio
- Then, do **supervised fine-tuning** with your **paired audio + transcriptions**

Low-resource ASR Approach

- Start with a **self-supervised foundation model** (learns from **raw audio**)
- In your target language, find **as much raw audio** as you can
  - e.g. radio, podcasts, movies, academic recordings
- Continue **self-supervised training** with your target language audio
- Then, do **supervised fine-tuning** with your **paired audio + transcriptions**
- And then... this **still often isn't good enough!**
  - The **bag of tools** we reach for now is the **topic of this course**

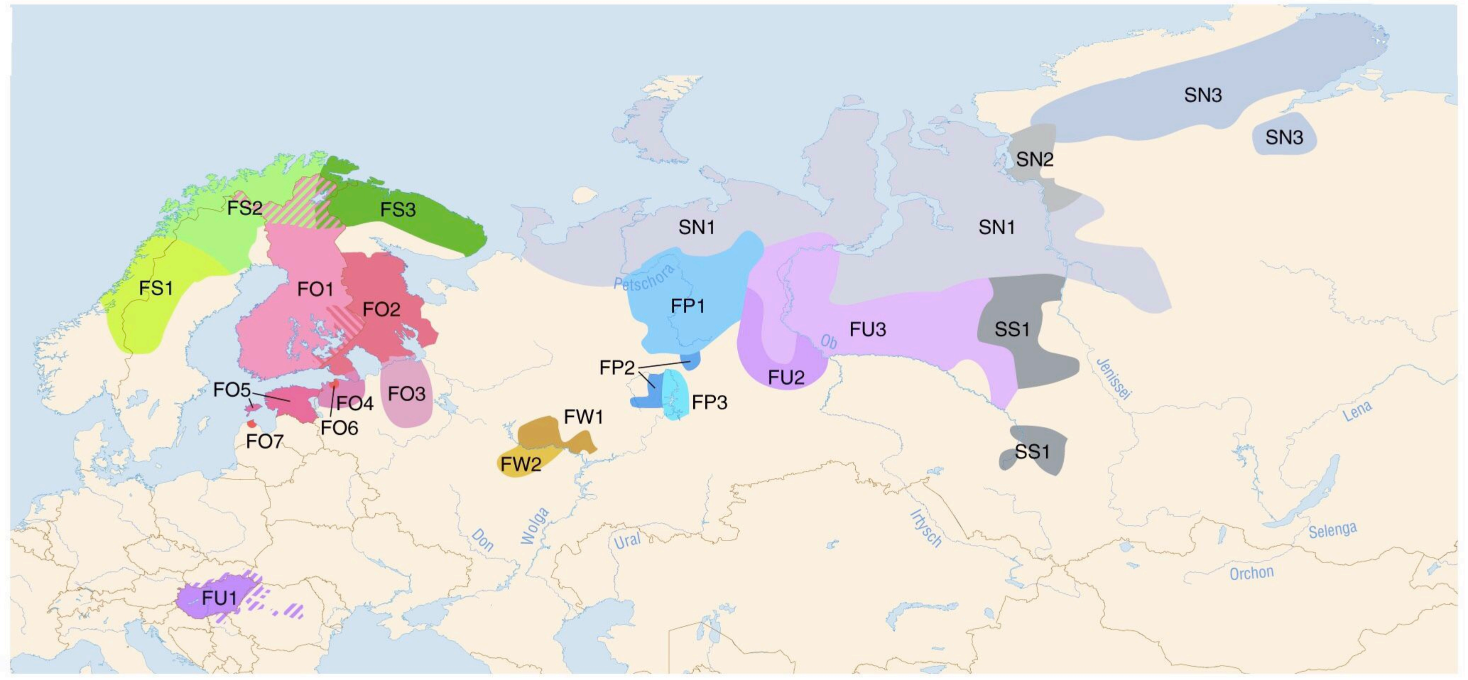
One Approach: Transfer Learning

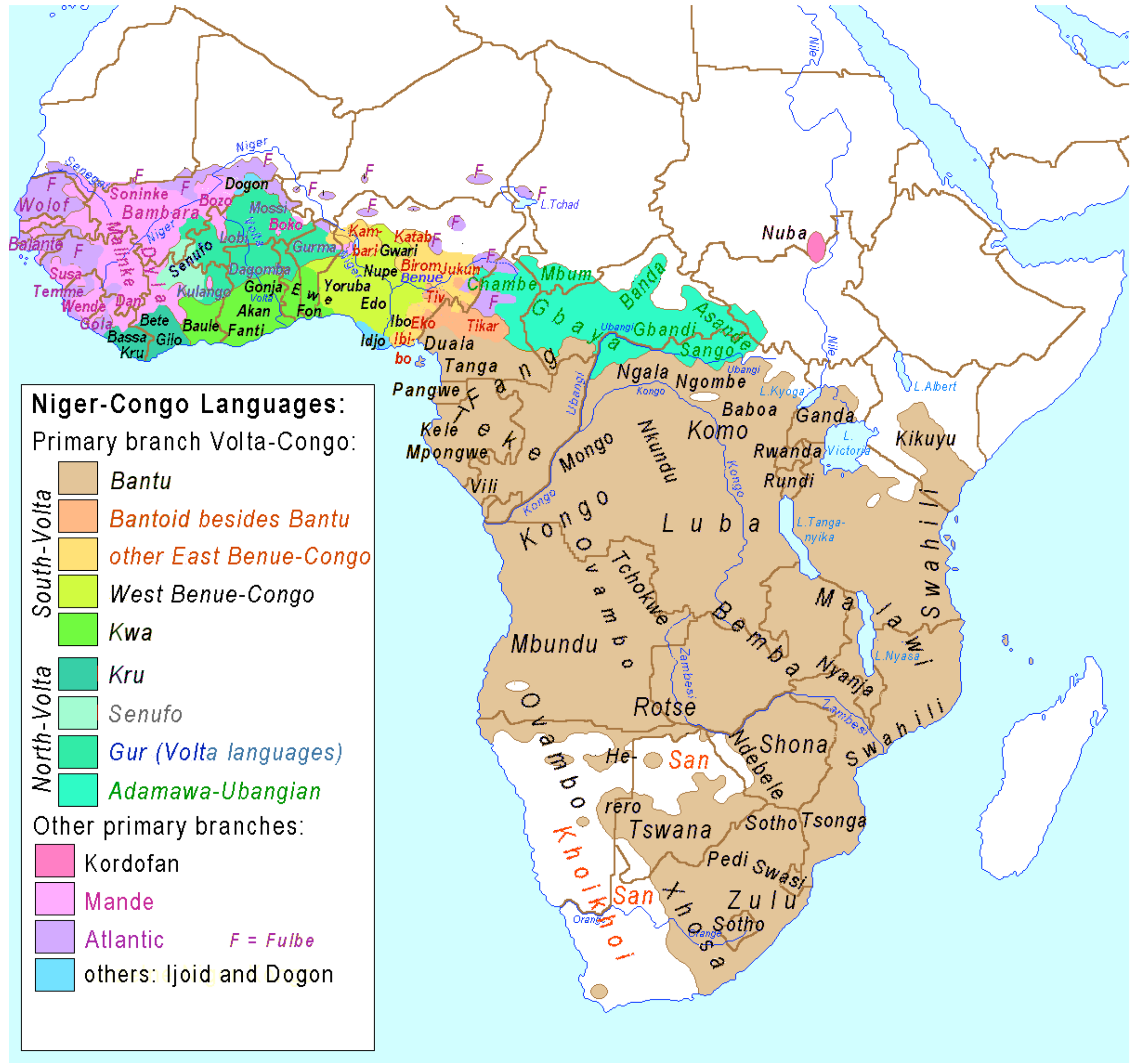




One Approach: Transfer Learning

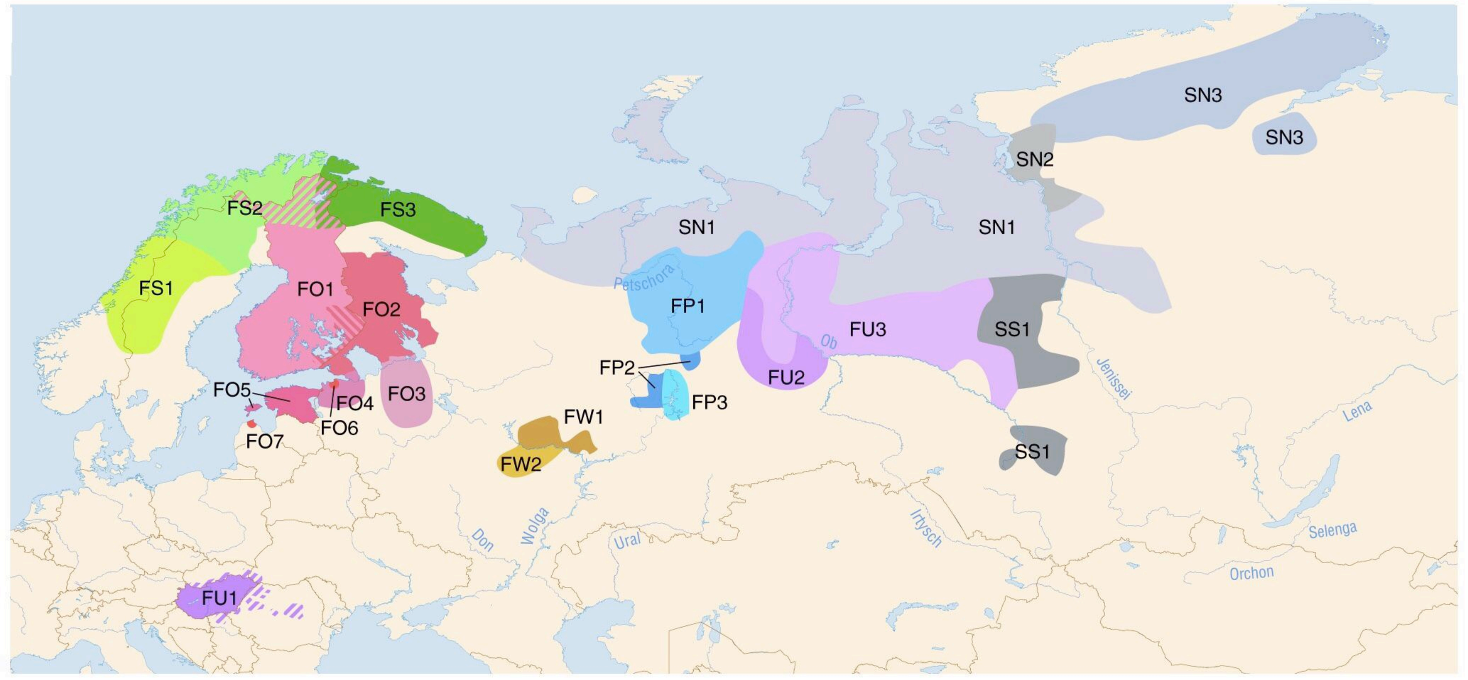
- Active area of research: **leveraging data from related languages**
  - Some evidence that **similarities** in **vocabulary** and **sound systems** (phonemes) assists

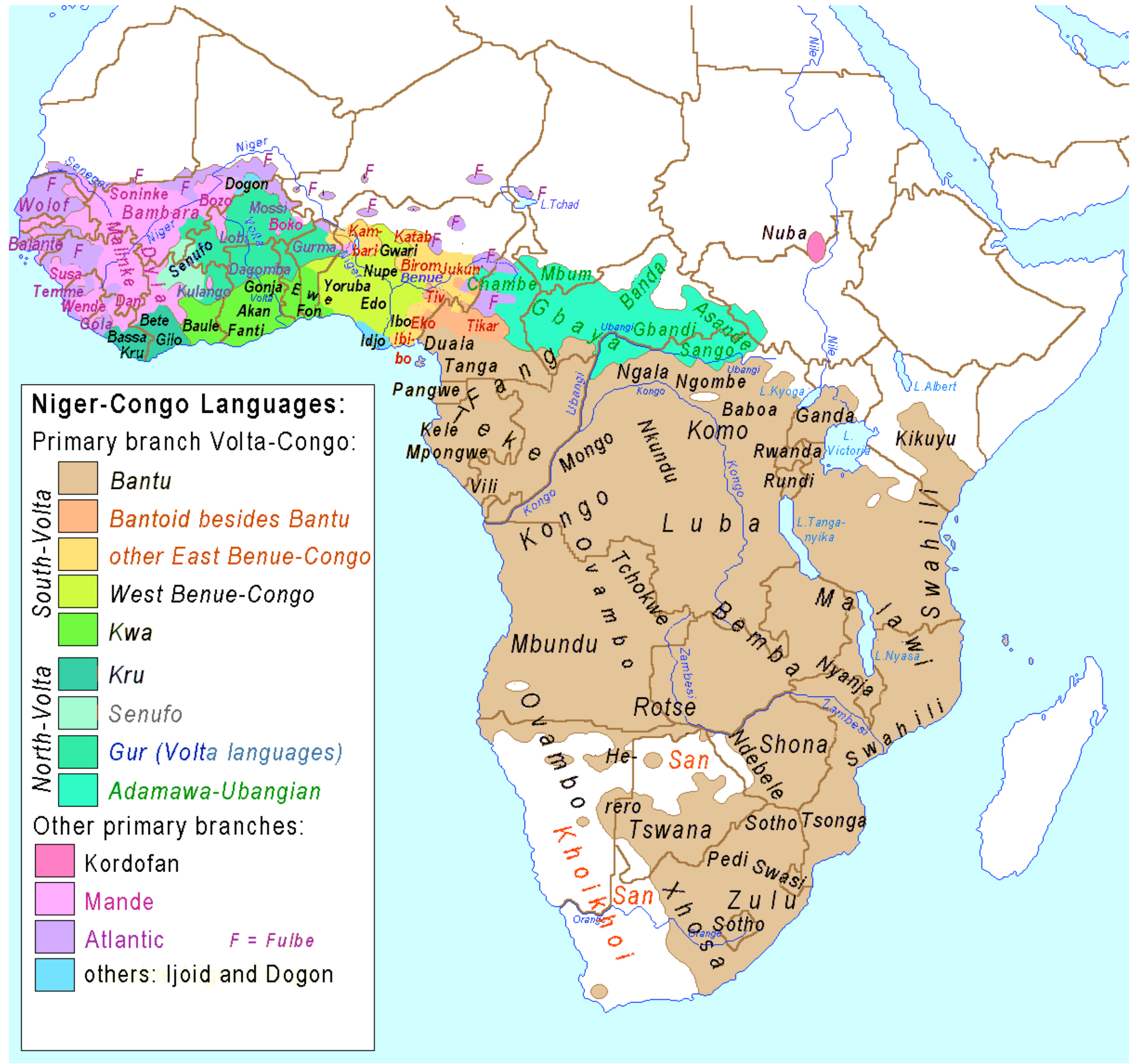




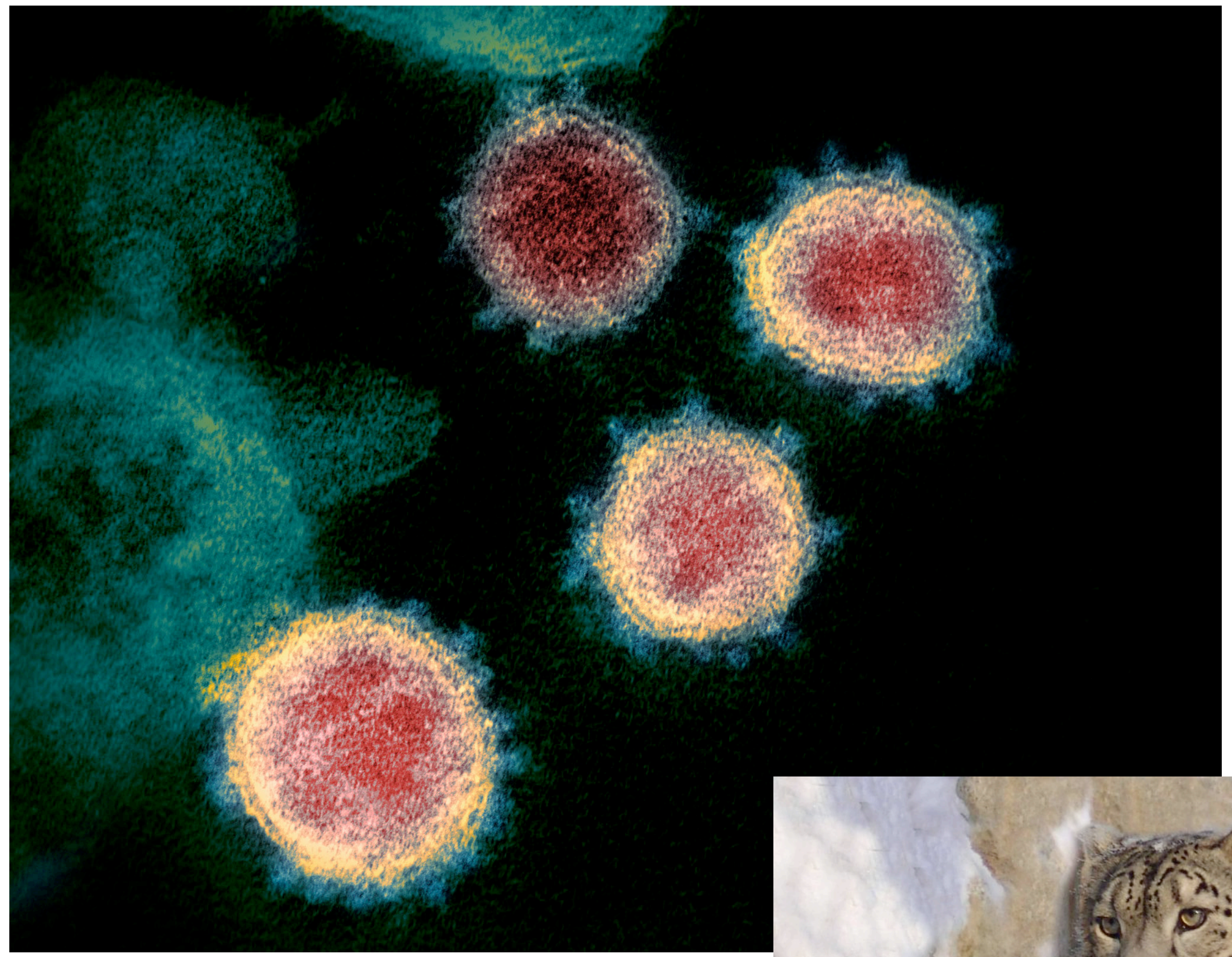
One Approach: Transfer Learning

- Active area of research: **leveraging data from related languages**
  - Some evidence that **similarities** in **vocabulary** and **sound systems** (phonemes) assists
- Part of a broader paradigm called **Transfer Learning**
  - Leverage **related data** when you don't have enough
  - Will cover this later in the course



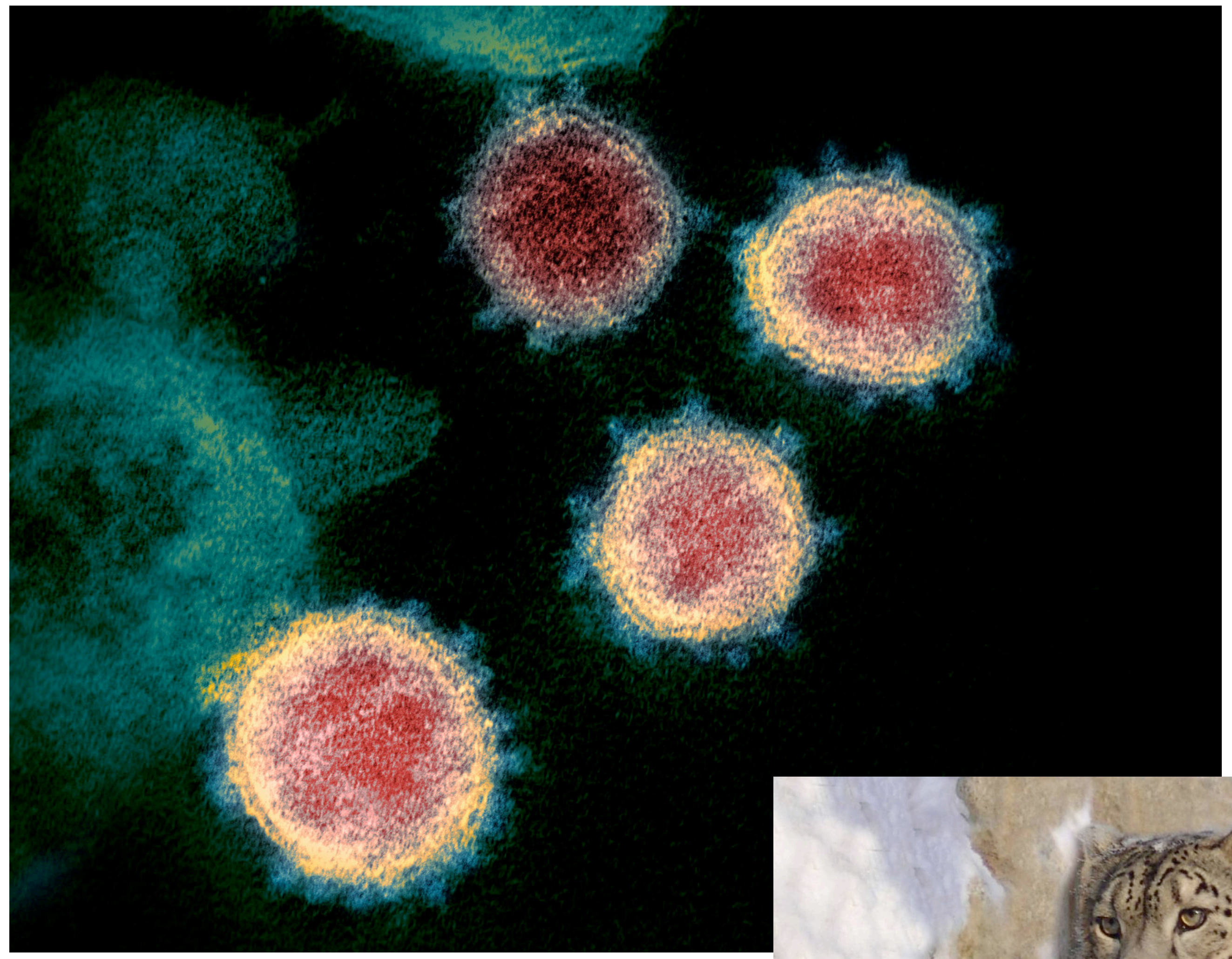


It's not just ASR



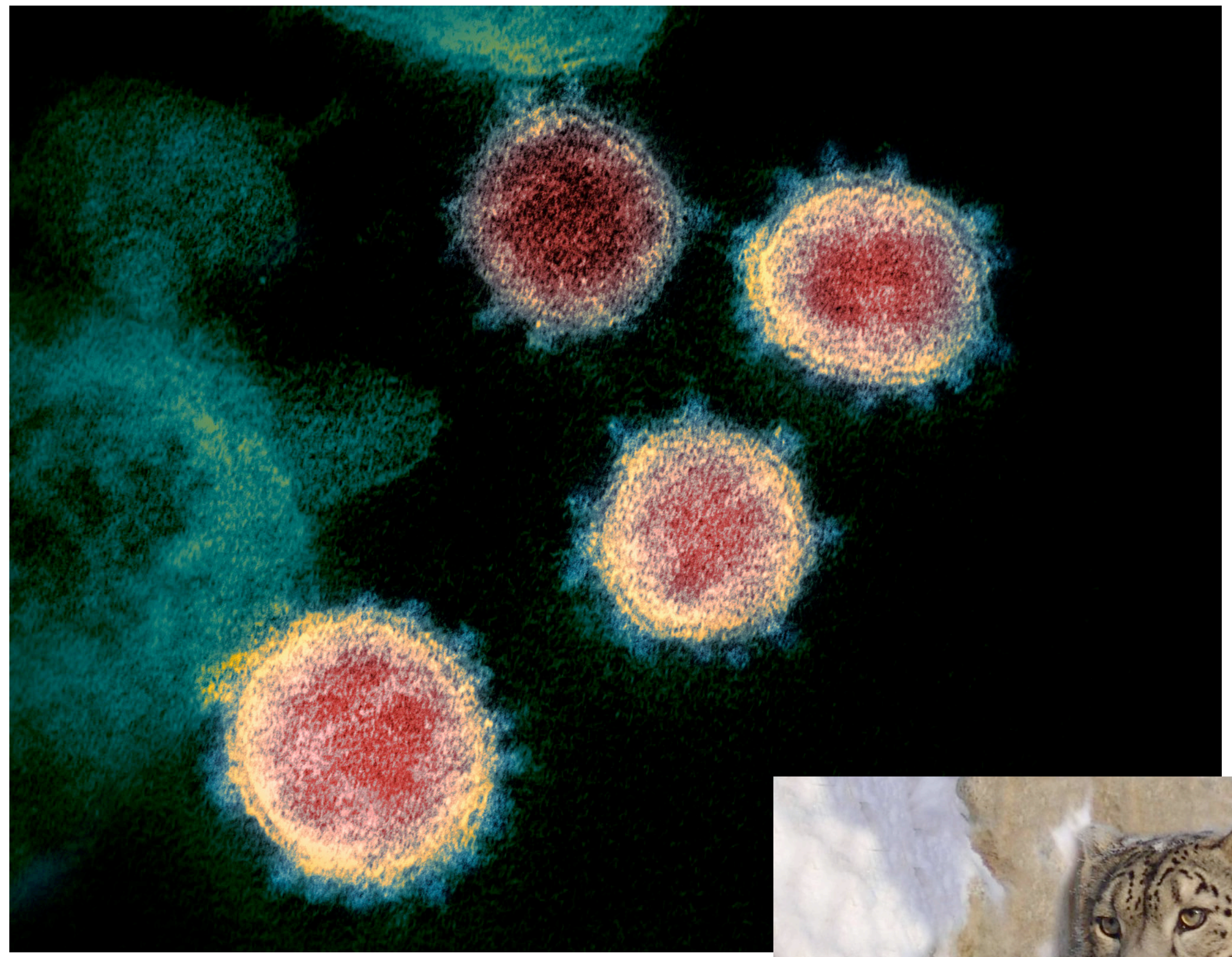
It's not just ASR

- I'm an NLP researcher, so I'll draw on lots of language examples



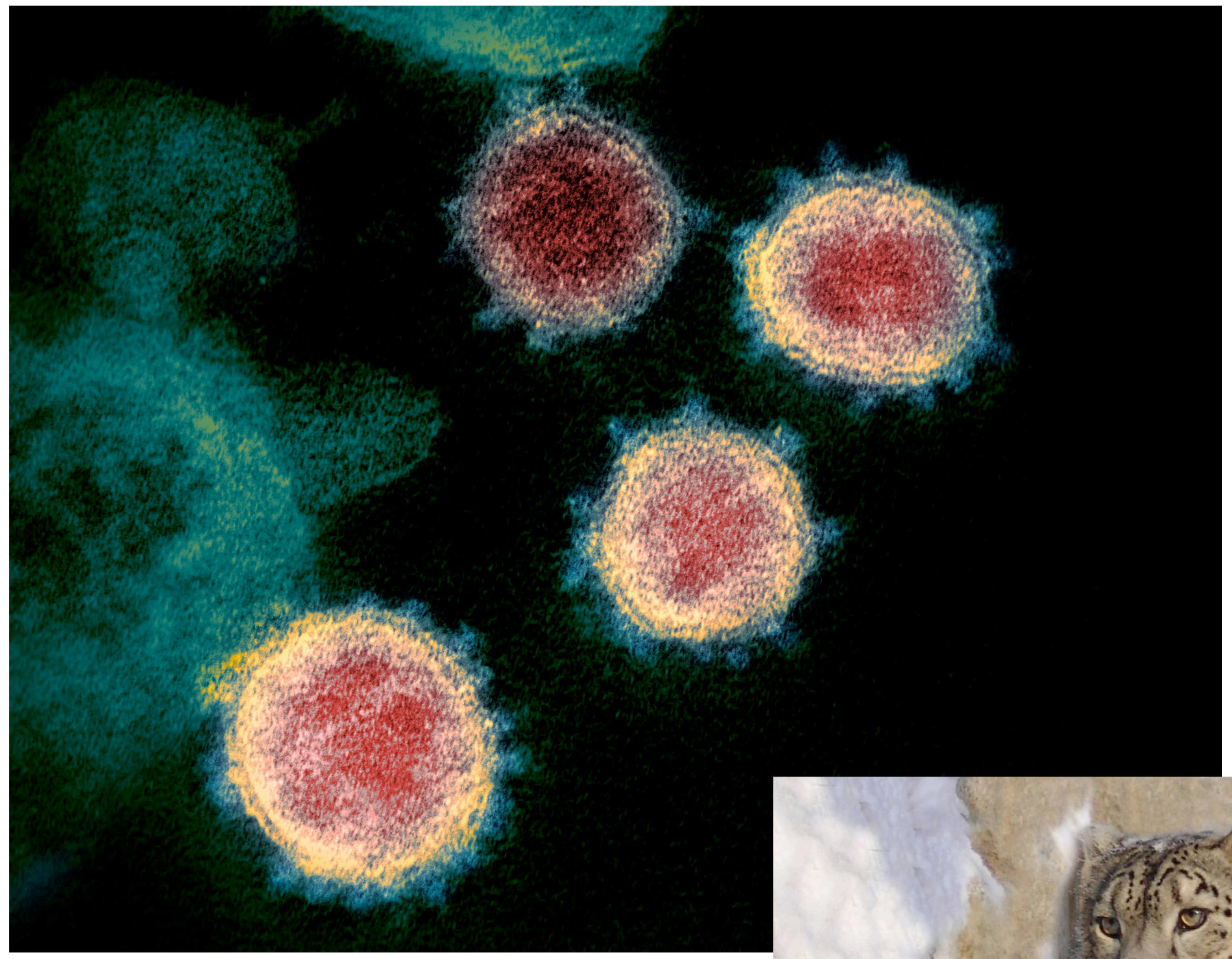
It's not just ASR

- I'm an NLP researcher, so I'll draw on lots of language examples
- The ideas extend to **any application area!** e.g...
  - Testing data for a **rare disease** - limited labels, **can't wait** for new data
  - Knowledge of **endangered species** is based on few observations
  - A **new product** has limited test opportunities



It's not just ASR

- I'm an NLP researcher, so I'll draw on lots of language examples
- The ideas extend to **any application area!** e.g...
  - Testing data for a **rare disease** - limited labels, **can't wait** for new data
  - Knowledge of **endangered species** is based on few observations
  - A **new product** has limited test opportunities
- What are some examples from **your areas of interest?**



Course Overview and Topics

Main Phases

Main Phases

- **Learning with Minimal Supervision** (~weeks 1-5)
  - Topics: unsupervised, self-supervised, and semi-supervised learning; active learning; weak supervision

Main Phases

- **Learning with Minimal Supervision** (~weeks 1-5)
  - Topics: unsupervised, self-supervised, and semi-supervised learning; active learning; weak supervision
- **Transfer Learning and Adaptation** (~weeks 6-8)
  - Topics: transfer from related data, adaptation to new domains

Main Phases

- **Learning with Minimal Supervision** (~weeks 1-5)
  - Topics: unsupervised, self-supervised, and semi-supervised learning; active learning; weak supervision
- **Transfer Learning and Adaptation** (~weeks 6-8)
  - Topics: transfer from related data, adaptation to new domains
- **Few-shot Learning and Data Augmentation** (~weeks 9-12)
  - Topics: few-shot learning, meta-learning, data augmentation, Human-in-the-Loop learning

Cross-cutting Themes

Cross-cutting Themes

- **The Bias/Variance Tradeoff**
  - Overfitting to data vs. generalizing from it; how to fight overfitting

Cross-cutting Themes

- **The Bias/Variance Tradeoff**
  - Overfitting to data vs. generalizing from it; how to fight overfitting
- **Inductive Biases**
  - What assumptions are we building into our models?

Cross-cutting Themes

- **The Bias/Variance Tradeoff**
  - Overfitting to data vs. generalizing from it; how to fight overfitting
- **Inductive Biases**
  - What assumptions are we building into our models?
- **Annotation Budget**
  - Is it better to label 1000 random examples or 100 carefully chosen ones? How do we best allocate resources to creating new data?

Cross-cutting Themes

- **The Bias/Variance Tradeoff**
  - Overfitting to data vs. generalizing from it; how to fight overfitting
- **Inductive Biases**
  - What assumptions are we building into our models?
- **Annotation Budget**
  - Is it better to label 1000 random examples or 100 carefully chosen ones? How do we best allocate resources to creating new data?
- **Evaluation Challenges**
  - How do we know that our low-resource solution really works and generalizes?

Policies and Logistics

Basics

Basics

- **Time:** Tuesday/Thursday 9:40-10:55am
  - Attendance is **required** and counts towards grade
  - After the first few weeks:
    - **Tuesdays:** research paper discussion
    - **Thursdays:** lectures

Basics

- **Time:** Tuesday/Thursday 9:40-10:55am
  - Attendance is **required** and counts towards grade
  - After the first few weeks:
    - **Tuesdays:** research paper discussion
    - **Thursdays:** lectures
- **Place:** Meliora Hall #218

Basics

- **Time:** Tuesday/Thursday 9:40-10:55am
  - Attendance is **required** and counts towards grade
  - After the first few weeks:
    - **Tuesdays:** research paper discussion
    - **Thursdays:** lectures
- **Place:** Meliora Hall #218
- **Office hours:** by appointment
  - Feel free to reach out any time to schedule a meeting!

Basics

- **Time:** Tuesday/Thursday 9:40-10:55am
    - Attendance is **required** and counts towards grade
    - After the first few weeks:
        - **Tuesdays:** research paper discussion
        - **Thursdays:** lectures
- **Place:** Meliora Hall #218
- **Office hours:** by appointment
    - Feel free to reach out any time to schedule a meeting!
- You're encouraged to use the **Blackboard Discussion Tab** for class-wide communications

Course Website

Course Website

- **Link: cmdowney88.github.io/teaching/dscc251/spring26**
  - Can also be found from my homepage (**cmdowney88.github.io**), by clicking on **"Course Webpages"**

Course Website

- **Link: cmdowney88.github.io/teaching/dscc251/spring26**
  - Can also be found from my homepage (**cmdowney88.github.io**), by clicking on **"Course Webpages"**
- Contains:
  - **Up-to-date course schedule**
  - Course **syllabus**
  - **Assigned readings** (links, or pointers to PDFs on Blackboard)
  - **Project milestone descriptions**

Reading

Reading

- There is **no required textbook** for the class

Reading

- There is **no required textbook** for the class
- I will assign **intermittent readings** to prepare for discussions
  - These will be available as **PDFs on Blackboard**
  - Assigned to facilitate **engagement in discussion**, which counts towards your grade

Reading

- There is **no required textbook** for the class
- I will assign **intermittent readings** to prepare for discussions
  - These will be available as **PDFs on Blackboard**
  - Assigned to facilitate **engagement in discussion**, which counts towards your grade
- You are **strongly encouraged** to at least skim the paper being presented by students each week

Assessment Structure

Assessment Structure

- **45% Term Project**
  - Substantial research project culminating in a paper (more later)

Assessment Structure

- **45% Term Project**
  - Substantial research project culminating in a paper (more later)
- **30% Student-led Discussion/Presentation**
  - Presenting a research paper to the class several times during the term

Assessment Structure

- **45% Term Project**
  - Substantial research project culminating in a paper (more later)
- **30% Student-led Discussion/Presentation**
  - Presenting a research paper to the class several times during the term
- **25% Participation**
  - 10% Attendance
  - 15% Engagement in Discussion/Activities

Term Project

Term Project

- Completed **alone** or in a **small group (2-3 students)**

Term Project

- Completed **alone** or in a **small group (2-3 students)**
- Conduct a topical **research project**, which must
  - **Apply data-efficient ML** to an **application area** of your choice (pick something you're genuinely interested in!)
  - Pose and test a **scientific hypothesis**
  - Culminate in a **research paper** and **presentation**

Term Project

- Completed **alone** or in a **small group (2-3 students)**
- Conduct a topical **research project**, which must
  - **Apply data-efficient ML** to an **application area** of your choice (pick something you're genuinely interested in!)
  - Pose and test a **scientific hypothesis**
  - Culminate in a **research paper** and **presentation**
- **Deliverables**: writeup, presentation, code repository
  - Scaffolded with **incremental milestones** due throughout the semester
  - First step: an **interest survey** due **next Thursday (1/29)**

Student-led Discussion

Student-led Discussion

- Starting in week 3, **Tuesdays are discussion days**

Student-led Discussion

- Starting in week 3, **Tuesdays are discussion days**
- **Two students** present each session (**~30-35 minutes** each)

Student-led Discussion

- Starting in week 3, **Tuesdays are discussion days**
- **Two students** present each session (**~30-35 minutes** each)
- Expect to do **3-5 presentations** during the semester (depends on enrollment)

Student-led Discussion

- Starting in week 3, **Tuesdays are discussion days**
- **Two students** present each session (**~30-35 minutes** each)
- Expect to do **3-5 presentations** during the semester (depends on enrollment)
- For your presentation:
  - Choose a **research paper** related to the previous week's topic (must be approved 1 week before presentation)
  - **Share** a copy/link with the class so they can prepare (by the Friday before)
  - **Present** the paper to the class and **lead discussion topics**

Presentation Guidelines

Presentation Guidelines

- The paper you choose should
  - Draw from **your** particular area of interest (major, thesis topic, etc.)
  - Illustrate an **application** of the methods discussed in class

Presentation Guidelines

- The paper you choose should
  - Draw from **your** particular area of interest (major, thesis topic, etc.)
  - Illustrate an **application** of the methods discussed in class
- Your presentation/discussion should
  - **Summarize** the paper for the class
  - **Connect** to the course topics being covered
  - Offer a **critical perspective** on the paper (e.g. are its conclusions sound?)
  - Start/facilitate **discussion** - have **2-3 discussion prompts** for the class

Attendance/Participation

Attendance/Participation

- I will keep track of attendance, which will be **visible on Blackboard**

Attendance/Participation

- I will keep track of attendance, which will be **visible on Blackboard**
- You have **two "free" absences** - for any reason, no questions asked
  - **No need to notify me**. I'll automatically account for these at the end of the term

Attendance/Participation

- I will keep track of attendance, which will be **visible on Blackboard**
- You have **two "free" absences** - for any reason, no questions asked
  - **No need to notify me**. I'll automatically account for these at the end of the term
- I'll also make exceptions for **important obligations/circumstances**
  - e.g. civic, religious, or family obligations; illness; job interviews; conferences

Attendance/Participation

- I will keep track of attendance, which will be **visible on Blackboard**
- You have **two "free" absences** - for any reason, no questions asked
  - **No need to notify me**. I'll automatically account for these at the end of the term
- I'll also make exceptions for **important obligations/circumstances**
  - e.g. civic, religious, or family obligations; illness; job interviews; conferences
- Any other absences will **count against your 10% attendance grade**

Attendance/Participation

- I will keep track of attendance, which will be **visible on Blackboard**
- You have **two "free" absences** - for any reason, no questions asked
  - **No need to notify me**. I'll automatically account for these at the end of the term
- I'll also make exceptions for **important obligations/circumstances**
  - e.g. civic, religious, or family obligations; illness; job interviews; conferences
- Any other absences will **count against your 10% attendance grade**
- **15%** of grade for **engagement in discussions / check-in activities**

Late Work

Late Work

- **Late work** (mostly applies to project milestones)
  - **Up to 1hr late: -5%**
  - **Up to 24hrs: -10%**
  - **Up to 48hrs: -20%**
  - **Later: no grade**

Late Work

- **Late work** (mostly applies to project milestones)
  - **Up to 1hr late: -5%**
  - **Up to 24hrs: -10%**
  - **Up to 48hrs: -20%**
  - **Later: no grade**
- Assignments mostly due at **11pm on the listed date**
  - **Note:** this shifts from EST to EDT in March

Academic Honesty

Academic Honesty

- All **standard university policies** apply

Academic Honesty

- All **standard university policies** apply
- **IMPORTANT: Generative AI Policy**
  - **NOT allowed for**: paper presentations, project milestones, project writeup
  - **allowed for**: programming work on the term project
  - Rule of thumb: using AI to **learn and clarify concepts** is fine; using it to **generate work you submit as your own** is **not**
  - **Honor system** - I don't want to be the AI police

Pre-requisites and Tools

Pre-requisites and Tools

- What I assume:
  - **One foundational Machine Learning course** (e.g. DSCC 240, 265, CSC 246, LING 282)
  - **Proficiency in Python Programming** (or another scientifically-inclined programming language, check with me if not sure)
  - A **laptop or other device** on which you can conduct computational work

Pre-requisites and Tools

- What I assume:
  - **One foundational Machine Learning course** (e.g. DSCC 240, 265, CSC 246, LING 282)
  - **Proficiency in Python Programming** (or another scientifically-inclined programming language, check with me if not sure)
  - A **laptop or other device** on which you can conduct computational work
- Computing resources
  - I will provide **access to UR's supercomputing cluster (BlueHive)**
  - Not mandatory to use, but helpful for intense ML algorithms

Questions?