



corti

Revolutionizing healthcare.
One patient consultation at a time.

Lars Maaløe

Co-Founder & Chief Technology Officer @ Corti
Adj. Associate Professor @ Technical University of Denmark



Digital health is exploding

There's now 50 billion patient consultations conducted a year.



Only 1% of patient consultations are quality assured

The pressure on caregivers is so high that **88% of diagnoses are altered** if tested by a second opinion.



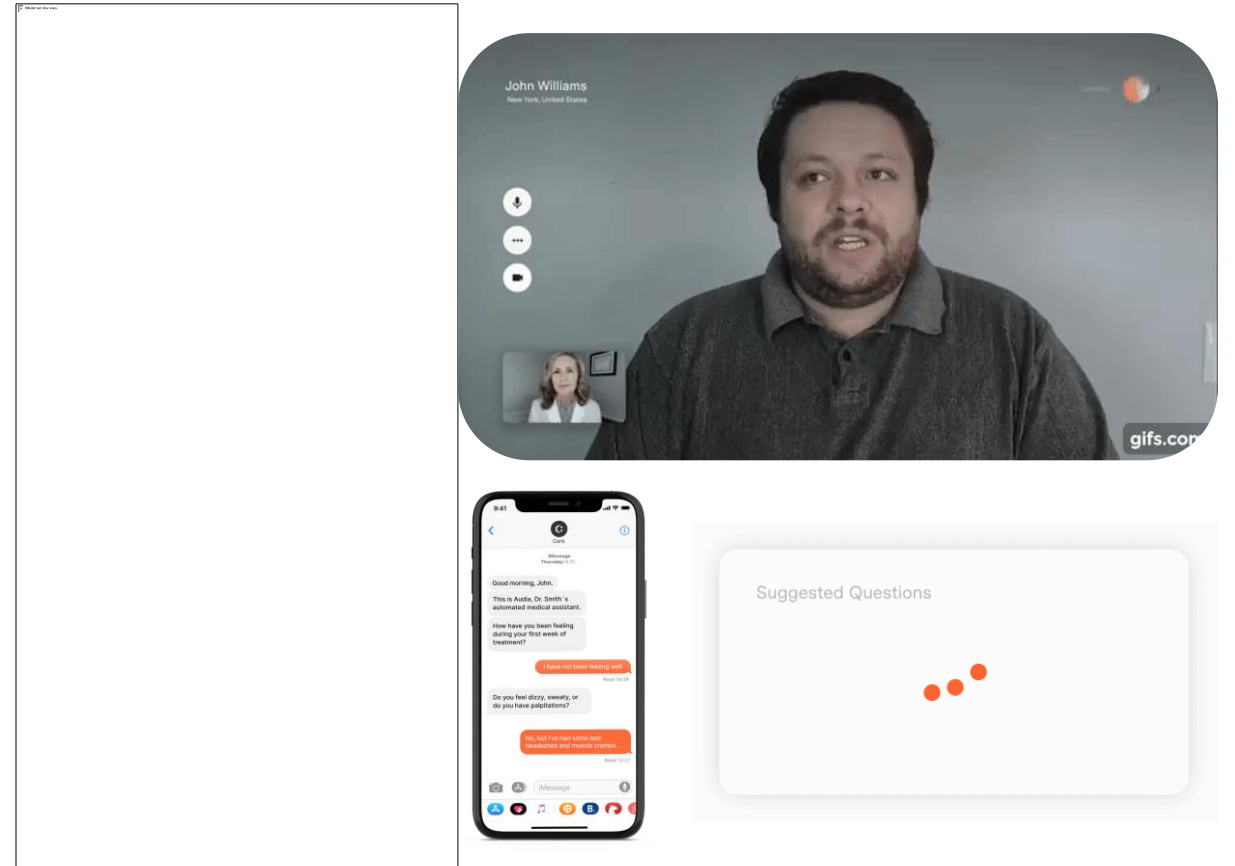
Physicians are becoming data clerks

Telehealth physicians spend up to **50% of their day** in EHR documenting consultations.



Intelligent augmentations

Corti has built an artificial intelligence that **listens** in and **understands** virtual consultations.



Our AI is built by a world leading team of engineers in close collaboration with academia

+369 h-index **+938** i-index **+248k** citations

Accumulated over 6x Professors, 2x Associate Professors, and a number of PostDocs and PhD scholars

Excerpt of some of our latest work

Self-Supervised Speech Representation Learning: A Review

Abdelrahman Mohamed¹, Hung-yi Lee¹, Lasse Borgholt¹, Jakob D. Havrton¹, Joakim Edin, Christian Igel, Karim Kichhoff, Shang-Wen Li, Karim Laveen, Lars Maalew, Tara N. Sainath, Shoji Watanabe

Abstract—Although supervised deep learning has revolutionized speech and audio processing, it has necessitated the building of specialist models for individual tasks and application scenarios. It is likewise difficult to apply this to dialects and languages for which only limited labeled data is available. Self-supervised representation learning methods promise a single universal model that would benefit a wide variety of tasks and domains. Such methods have shown success in natural language processing and computer vision domains, achieving new levels of performance while reducing the number of labels required for many downstream scenarios. Speech representation learning is experiencing similar progress in three main categories: generative, contrastive, and predictive models. Other approaches rely on multi-modal data for pre-training, mixing text or visual data streams with speech. Although self-supervised speech representation is still a nascent research area, it is closely related to acoustic word embedding and learning with raw lexical resources, both of which have seen active research for many years. This review presents approaches for self-supervised speech representation learning and their connection to other research areas. Since many current methods focus solely on automatic speech recognition as a downstream task, we review recent efforts on benchmarking learned representations to extend the application beyond speech recognition.

Index Terms—Self-supervised learning, speech representations.

I. INTRODUCTION

Over the past decade, deep learning approaches have revolutionized speech processing through a giant leap in performance, enabling various real-world applications. Supervised learning of deep neural networks has been the cornerstone of this transformation, offering impressive gains for scenarios rich in labeled data [1]–[3]. Paradoxically, this heavy reliance

on supervised learning has restricted progress in languages and domains that do not attract the same level of labeling investment.

To overcome the need for labeled data, researchers have explored approaches that use unpaired audio-only data to open up new industrial speech use-cases and low-resource languages [4]–[6]. Inspired by how children learn their first language through listening and interacting with family and caregivers, scientists seek to use raw waveforms and spectral signals to learn speech representations that capture low-level acoustic events, lexical knowledge, all the way to syntactic and semantic information. These learned representations are then used for target downstream applications requiring a minimal number of labeled data [7]–[9]. Primarily, representation learning refers to algorithms for extracting latent features that capture the underlying explanatory factors for the observed input [5].

Representation learning approaches are generally considered examples of *unsupervised learning*, which refers to the family of machine learning methods that discover naturally occurring patterns in training samples for which there are no pre-assigned labels or labels [10]. The term “unsupervised” is used to distinguish this family of methods from “supervised” approaches, which assign a label to each training sample, and

Hierarchical VAEs Know What They Don’t Know

Jakob D. Havrton^{1,2}, Jes Freilich¹, Søren Hauberg¹, Lars Maalew^{1,2}

Abstract

Deep generative models have been demonstrated as state-of-the-art density estimators. Yet, recent work has found that they often assign a higher likelihood to data from outside the training distribution. This seemingly paradoxical behavior has caused concerns over the quality of the attained density estimates. In the context of hierarchical variational autoencoders, we provide evidence to explain this behavior by out-of-distribution data having in-distribution low-level features. We argue that this is both insightful and desirable behavior. With this insight in hand, we develop a fast, scalable and fully unsupervised likelihood-ratio score for OOD detection that requires data to be in-distribution across all feature levels. We benchmark the method on a vast set of data and model combinations and achieve state-of-the-art results on out-of-distribution detection.

1. Introduction

The reliability and safety of machine learning systems applied in the real-world is contingent on the ability to detect when an input is different from the training distribution. Supervised classifiers built as deep neural networks are well-known to misclassify such out-of-distribution (OOD) inputs to known classes with high confidence (Goodfellow et al., 2015; Nguyen et al., 2015). Several approaches have been suggested to equip deep classifiers with OOD detection capabilities (Hendrycks & Gimpel, 2017; Lakshminarayanan et al., 2017; Hendrycks et al., 2019; DeVries & Taylor, 2018). But, such methods are inherently supervised and require in-distribution labels or examples of OOD data limiting their applicability and generality.

Unsupervised generative models that estimate an explicit likelihood should understand what it means to be in- and out-of-distribution without requiring labels or examples of OOD data. By directly modeling the training distribution, such models are expected to assign low likelihoods to OOD data as it originates from regions of little or no support under the learned density (Bishop, 1994). Recent advances in deep generative models (Kingma & Welling, 2014; Rezakhanlou et al., 2014; Oord et al., 2016; Salimans et al., 2017; Kingma & Dhariwal, 2018) have enabled learning high quality generative models on complex data such as natural images, sequences including audio (Oord et al., 2016a) and graphs (Kipf & Welling, 2016). However, recent observations have brought into question the quality of the learned density estimates by showing that they often assign higher likelihoods to OOD data than to in-distribution data (Nalisnick et al., 2019a; Choi et al., 2019). Many complex data distributions can be explained to a large degree by low-level features, e.g. edges in images or low-frequency features in audio.

In this paper, we examine the failure cases of deep generative models on OOD detection tasks with the context of

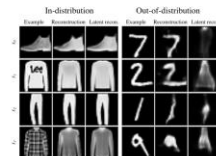


Figure 1. Reconstructions using a hierarchical VAE trained on FashionMNIST. Reconstruction quality of OOD data is comparable to in-distribution data, resulting in high likelihoods and poor OOD discrimination. By sampling the 4 bottom-most latent variables from the conditional prior distribution $p(z_4 | z_{1:3})$ (latent reconstructions) instead of the approximate posterior $q(z_4 | z_{1:3})$, the model reconstructs from the training distribution resulting in lower $p(z_4 | z_{1:3})$ for OOD data.

out-of-distribution without requiring labels or examples of OOD data. By directly modeling the training distribution, such models are expected to assign low likelihoods to OOD data as it originates from regions of little or no support under the learned density (Bishop, 1994). Recent advances in deep generative models (Kingma & Welling, 2014; Rezakhanlou et al., 2014; Oord et al., 2016; Salimans et al., 2017; Kingma & Dhariwal, 2018) have enabled learning high quality generative models on complex data such as natural images, sequences including audio (Oord et al., 2016a) and graphs (Kipf & Welling, 2016). However, recent observations have brought into question the quality of the learned density estimates by showing that they often assign higher likelihoods to OOD data than to in-distribution data (Nalisnick et al., 2019a; Choi et al., 2019). Many complex data distributions can be explained to a large degree by low-level features, e.g. edges in images or low-frequency features in audio.

In this paper, we examine the failure cases of deep generative models on OOD detection tasks with the context of

BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling

Lars Maalew¹, Marco Fraccaro¹, Valentin Lavie & Ole Winther¹
Cort, Copenhagen, Denmark
maalew@di.ku.dk, mfraccaro@di.ku.dk, val@di.ku.dk, ole@di.ku.dk

Abstract

With the introduction of the variational autoencoder (VAE), probabilistic latent variable models have received renewed attention as powerful generative models. However, their performance in terms of test likelihood and quality of generated samples has been surpassed by autoregressive models without stochastic units. Furthermore, flow-based models have recently been shown to be an attractive alternative that scales well in high-dimensional data. In this paper we close the performance gap by constructing VAE models that can effectively utilize a deep hierarchy of stochastic variables and model complex covariance structures. We introduce the Bidirectional-Inference Variational Autoencoder (BIVA), characterized by a skip-connected generative model and an inference network formed by a bidirectional stochastic inference path. We show that BIVA reaches state-of-the-art test likelihoods, generates sharp and coherent natural images, and uses the hierarchy of latent variables to capture different aspects of the data distribution. We observe that BIVA, in contrast to recent results, can be used for anomaly detection. We attribute this to the hierarchy of latent variables which is able to extract high-level semantic features. Finally, we extend BIVA to semi-supervised classification tasks and show that it performs comparably to state-of-the-art results by generative adversarial networks.

1 Introduction

One of the key aspirations in recent machine learning research is to build models that *understand the world* [24, 40, 11, 57]. Generative models are providing the means to learn from a plethora of unlabeled data in order to model a complex data distribution, e.g. natural images, text, and audio. These models are evaluated by their ability to generate data that is similar to the input data distribution from which they were trained on. The range of applications that come with generative models are vast, where audio synthesis [55] and semi-supervised classification [38, 31, 44] are examples hereof. Generative models can be broadly divided into explicit and implicit density models. The generative adversarial network (GAN) [11] is an example of an implicit model, since it is not possible to procure a likelihood estimation from this model framework. The focus of this research is instead within explicit density models, where by a tractable or approximate likelihood estimation can be performed.

The three main classes of powerful explicit density models are autoregressive models [26, 57], flow-based models [8, 9, 21, 16], and probabilistic latent variable models [24, 40, 33]. In recent years autoregressive models, such as the PixelCNN [57, 45], have achieved superior likelihood performance and flow-based models have proven efficacy on large-scale natural image generation tasks [21]. However, in the autoregressive models, the runtime performance of generation is scaling poorly with the complexity of the input distribution. The flow-based models do not possess

ON SCALING CONTRASTIVE REPRESENTATIONS FOR LOW-RESOURCE SPEECH RECOGNITION

Lasse Borgholt^{1,2}, Tycho M. S. Tax, Jakob D. Havrton¹, Lars Maalew² and Christian Igel¹

¹Department of Computer Science, University of Copenhagen, Denmark

²Corti, Copenhagen, Denmark
borgholt@di.ku.dk

ABSTRACT

Recent advances in self-supervised learning through contrastive training have shown that it is possible to learn a competitive speech recognition system with as little as 10 minutes of labeled data. However, these systems are computationally expensive since they require pre-training followed by fine-tuning in a large parameter space. We explore the performance of such systems without fine-tuning by training a state-of-the-art speech recognizer on the fixed representations from the computationally demanding wav2vec 2.0 framework. We find performance to decrease without fine-tuning and, in the extreme low-resource setting, wav2vec 2.0 is inferior to its predecessor. In addition, we find that wav2vec 2.0 representations live in a low dimensional subspace and that decorrelating the features of the representations can stabilize training of the automatic speech recognizer. Finally, we propose a bidirectional extension to the original wav2vec framework that consistently improves performance.

Index Terms—automatic speech recognition, unsupervised learning, semi-supervised learning, self-supervised learning, representation learning

1. INTRODUCTION

Unsupervised learning for automatic speech recognition (ASR) has recently gained significant attention [1, 2, 3, 4, 5, 6, 7, 8, 9]. While the majority of work has focused on learning representations encoding the input for downstream tasks [1, 2, 4, 5, 6, 8, 9], the most promising results have been achieved with the wav2vec 2.0 framework (Fig. 1) where a pre-trained model is fine-tuned for speech recognition. However, these models are computationally expensive due to the large amount of memory intensive transformer layers. This contradicts the promise of easily applying these representations for new ASR models on low resource languages [3].

In contrast to wav2vec 2.0, its predecessor (Fig. 2) does not require fine-tuning as learned representations are used directly as input for an ASR model [1]. In addition, the pre-trained model has an order of magnitude fewer parameters than the large configuration of wav2vec 2.0. Because the

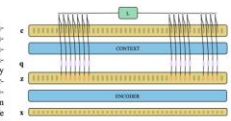


Fig. 1. The wav2vec 2.0 framework [3]. The model is trained to identify the correct quantized target corresponding to the masked latent representations. The two proposed configurations have 95 and 317 million parameters respectively.

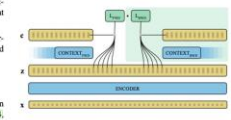


Fig. 2. The wav2vec framework [1] extended with a backward context network (shaded area). The two context networks are independent, but are trained jointly with a shared encoder. The original model has 33 million parameters, while our extended model only has 18 million.

frameworks are very similar, it seems obvious that representations extracted from wav2vec 2.0 would also be suitable input for training an ASR model. Training on extracted representations offers a light-weight alternative to the computationally expensive fine-tuning procedure described in [3].

We study how representations from the two versions of the open-source wav2vec framework compare when used as input

In direct competition with the major AI labs

OpenAI DeepMind Meta

corti

Prior to this presentation, our AI has learned from

+15,000,000

patient calls to medical command centers.





2016, we started where urgency is highest

Emergencies are our
beachhead market,
helping where it matters
the most and getting the
best data.

Servicing +50 million yearly encounters

We service customers in Europe
and USA with some of the best
annotated healthcare data,
continuously improving the AI.



Our technology can solve the biggest challenges for healthcare services



Healthcare worker

Burnout and Churn



Citizens

Healthcare Quality



Organization

Trust and Credibility

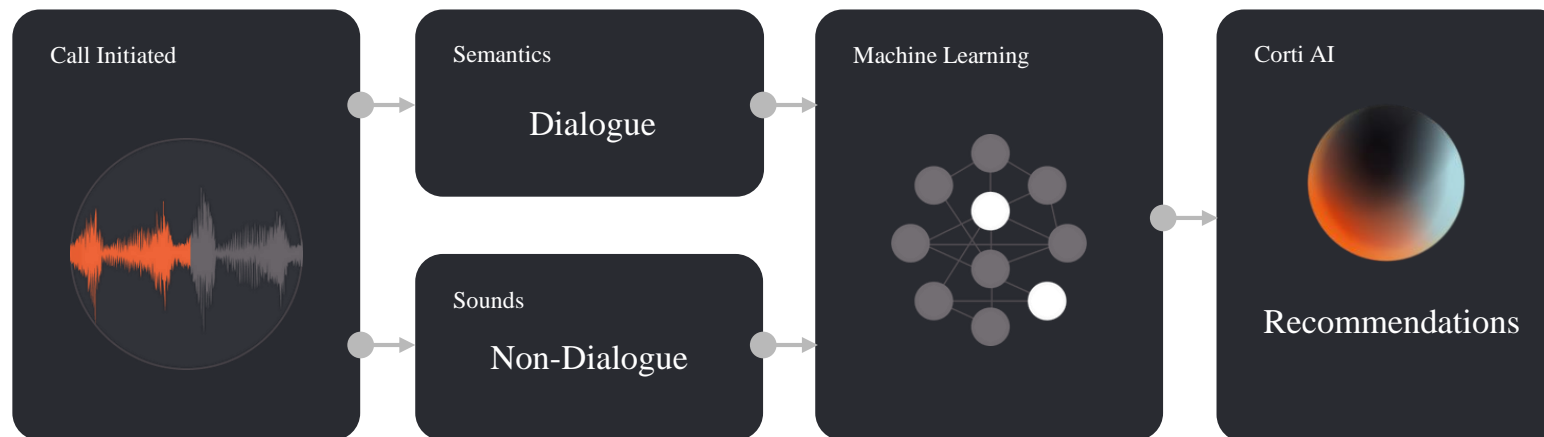


¹ "EMS services warn of 'crippling labor shortage' undermining 911 system." NBC News. 8 Oct. 2021, <https://nbcnews.to/3M2PtAp>

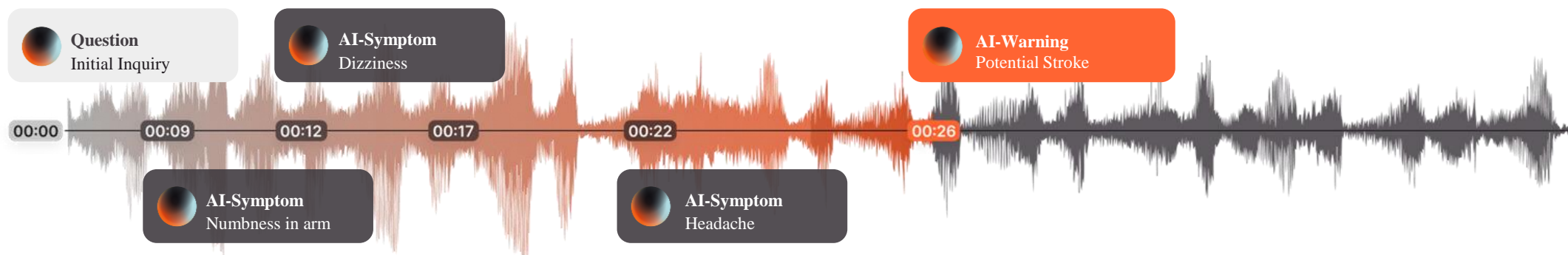
² "Stress on the front lines of covid-19 - The Washington Post." 6 Apr. 2021, <https://wapo.st/3GRjnUy>

³ "The longitudinal study of turnover and the cost of turnover in EMS." 11 June 2010, <https://bit.ly/3v8DiMz>

Corti analyzes **100%** of your communication



Automatic annotations of all communication



What we are most commonly known for

Triage a patient through workflow software with intelligent decision support

The screenshot displays a web-based interface for a medical triage session. The top navigation bar includes a hamburger menu, navigation arrows, a grid icon, and the session identifier "Live Call / Session / F7389E". A left-hand sidebar contains a list of navigation items: "Live Call", "Overview", "Branches", "Chief Complaint", "Pins", "SBAR", "Patient Details", and "CPR Instructions". The main content area features a central form titled "Tell me exactly what happened." with a list of radio button options: "Allergic Reaction", "Assault", "Choking / Respiratory", "Decreased Level of Consciousness", "Pain", "Trauma", "Unconscious", and "Unknown". A circular arrow icon is positioned to the right of the form, indicating a next step or confirmation action.

* [Link to video](#)



Reduced call duration in Sweden by +20%

"Our more than 800 operators use Corti for safer and faster medical triaging. This has allowed us to significantly increase patient safety through protocol adherence, while reducing average call duration by more than 20 percent and counting."

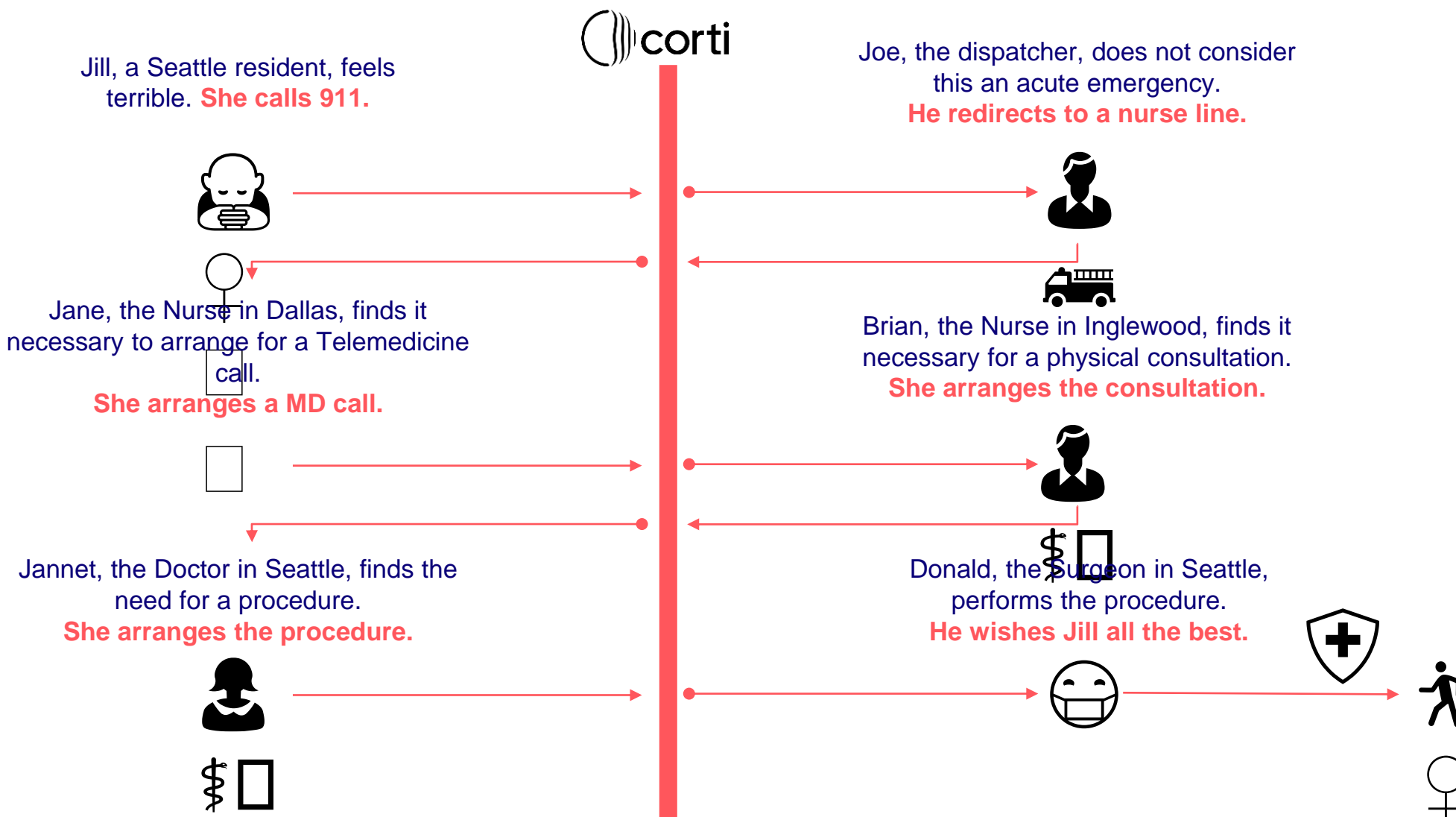
Jannice Mattsson, COO, SOS Alarm

[* Link to video](#)



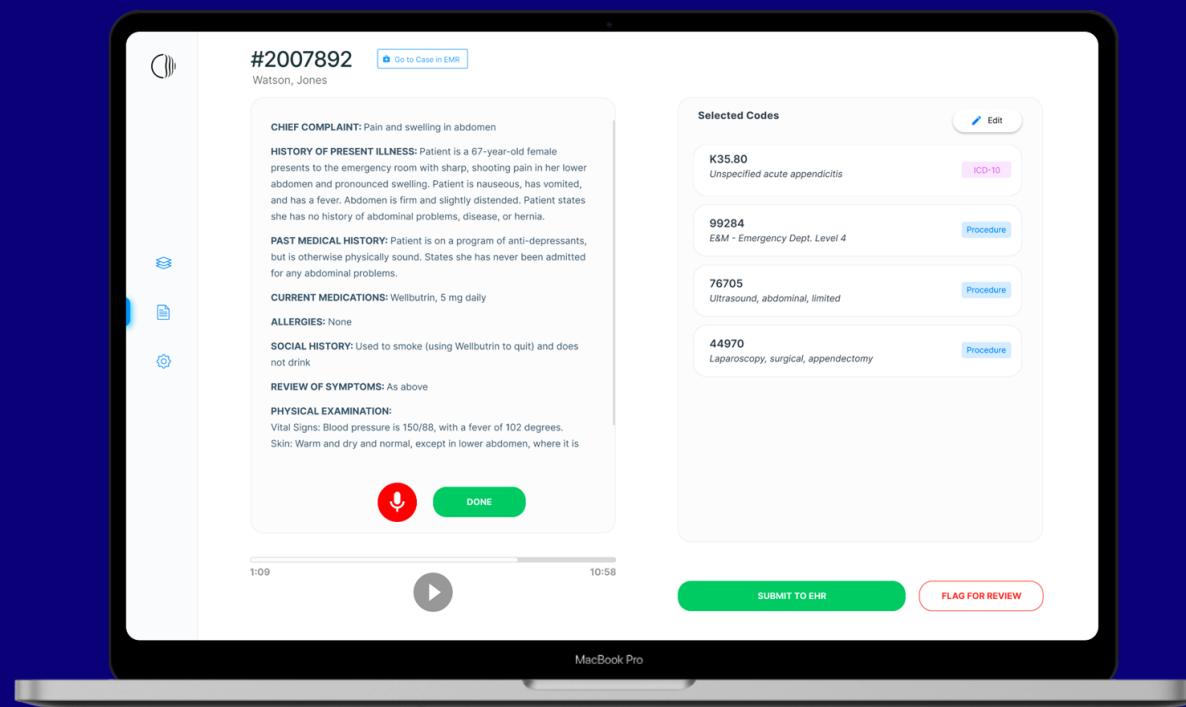
Following the patient journey

Since 2016, we have optimized towards supporting the entire healthcare value chain

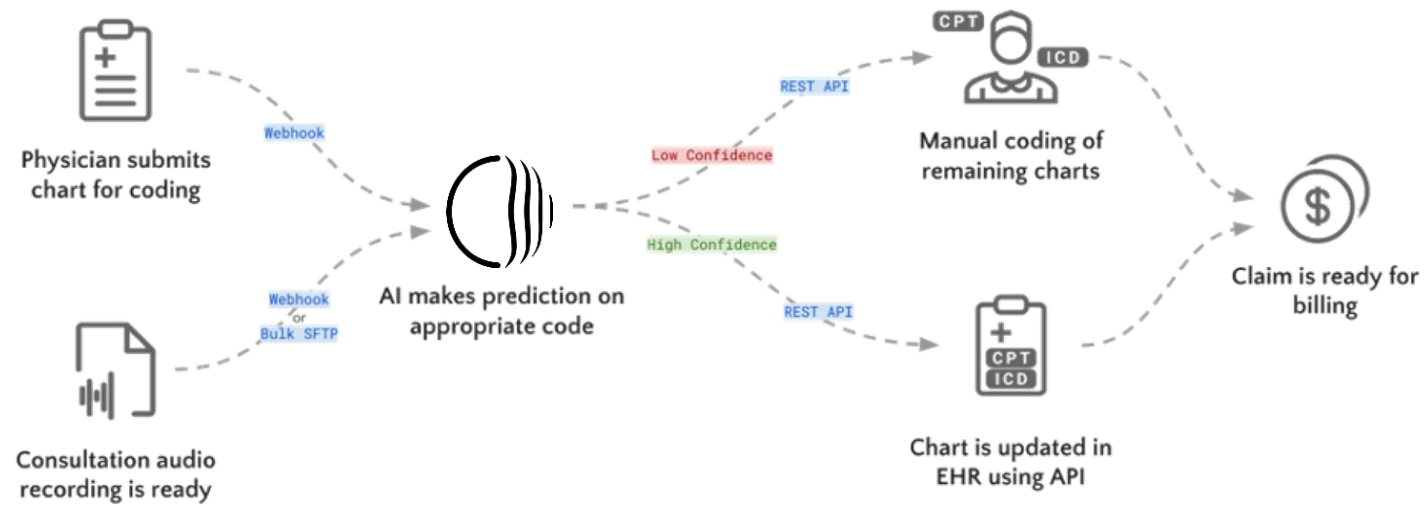


Example: Alleviating the administrative burden

Corti Code documents the patient interaction

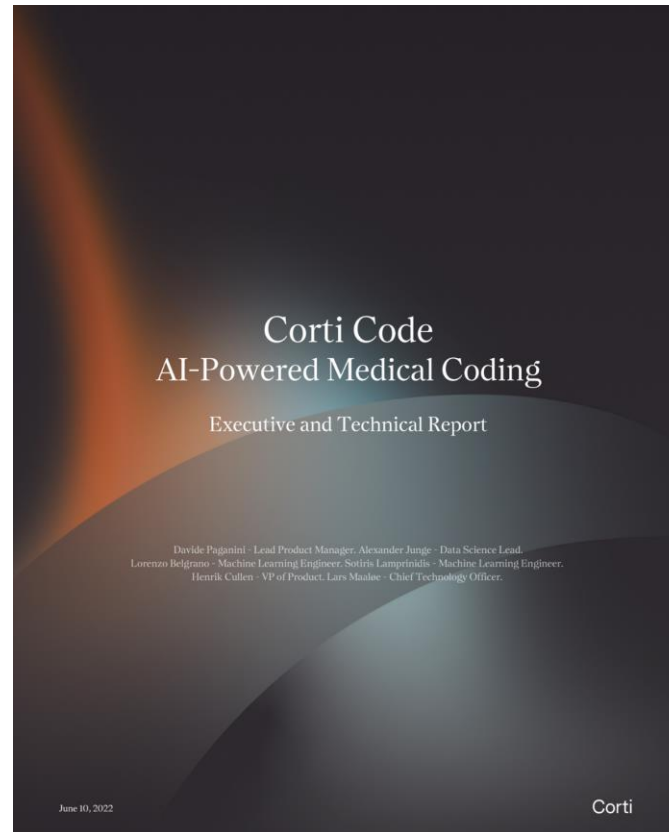


How it works



Recent Example: Winning a +100 hospital health provider

ICD-10 coding from audio in competition with a large set of companies



-> +75% of consultations are fully automated by AI.

-> +95% of consultations are gets the right code(s) through a top-5 recommendation engine.



Contact
Lars Maaløe
Co-founder & CTO
lm@corti.ai

