PAT 463/563 (Fall 2025)

Music & AI

Lecture 12: Language-based Music Generation

Instructor: Hao-Wen Dong

UNIVERSITY OF MICHIGAN

🚀 Project

- **Open-ended project**
- Group size: 1–3
- **Milestones**
  - **Pitch** Nov 5
  - **Presentation** Dec 1
  - **Report** Dec 15
- Deliverables due at **11:59pm ET** on the date specified
- **No late submissions!** Submit your work early and update it later

🚀 Project: Topics

- Building a new AI music tool
- Exploring creative & artistic use of AI tools
- Analyzing systematically existing AI music tools

🚀 Project: Rubrics

- **Presentation** (**20pt**)
  - Attendance (10pt)
  - Clarity (5pt)
  - Organization & presentation (5pt)
- **Report** (**20pt**)
  - Writing clarity (5pt)
  - Organization & presentation (5pt)
  - Results (5pt)
  - Discussion (5pt)

🚀 Project Pitch

- Brief 10-min presentation
  - **Team member introduction**
  - **Topic**: What do you want to work on?
  - **Topic**: Who are the target audience/users/customers/readers?
  - **Goals**: What are your goals?
  - **Methodology**: How are you going to approach it?
  - **Methodology**: What are the tools (programming languages, platforms, plugins, hardware, etc.) that you'll be using?
  - **Expected results**: What are the expected deliverables (e.g., an instrument, a plugin, a web/mobile app, a standalone software, an installation, a performance, a composition)?
  - **Planning**: What are the timeline & milestones?

AI Song Contest

AI Song Contest

- Annual international competition showcasing the **creative potential of human–AI co-creativity in the songwriting process**

aisongcontest.com

Entering Demons & Gods by Yaboi Hanoi (2022)

youtu.be/PbrRoR3nEVw

soundcloud.com/yaboihanoi/enter-demons-and-gods

Reading: The Making of Entering Demons & Gods (2022)

“**It was like a saxophonist trained in classical Thai motifs, who played a special ‘Thai Edition’ saxophone with Phi Nai tunings, had joined the musical conversation.** The same was true with the trumpet model and the ขลุ่ย ‘Khlui’ - a flute from Thai, Laos and Cambodian repertoire. I could assemble a **transcultural ensemble** to expand the sonic palette of Thai motifs, whilst adhering to underlying tunings and idiomatic inflections like never before.”

lamtharnhantrakul.github.io/enter-demons-and-gods/

How would you touch me? by Synthetic Beat Brigade (2023)

youtu.be/O4cJ3acEGDw

Reading: The Making of How would you touch me? (2023)

"This project is a collaboration between **Artificial Intelligence (AI) enthusiasts in four fields: artist management, music and post-production, tech, and creative**. In contrast, the majority of the music industry sees AI as a threat. Our team understands that these technological advances will have a significant impact on how we produce music. Because of this, we have decided to **use AI for every step of the production process**. From ideation to creating the lyrics to producing the music."

drive.google.com/file/d/1QTQ7P3iZI6I0anIwNQ3ewf8g3JjDjesl/view

Reading: The Making of How would you touch me? (2023)

- **Ideation**: Spotify API, ChatGPT, Facebook Llama, Google Bison
- **Lyrics**: ChatGPT 2, Genius API
- **Composition**: AI Drummachine, Mofi, Tonetrasnfer, This patch does not exist, Albeatz, BaiscPitch, Magenta, AIVA, MuseNet
- **Vocals**: Soundly Voice Designer, Vocal Remove, Voice characteristics
- **Mastering**: Landr
- **Cover art & bandart**: Midjourney
- **Clip**: ComfyUI for Stable Diffusion + ControlNet

drive.google.com/file/d/1QTQ7P3iZI6I0anIwNQ3ewf8g3JjDjesl/view

Sudamérica by Onda Corta (2024)

youtu.be/OQcpItDUuik

Reading: The Making of Sudamérica (2024)

“To create our song, **we integrated AI tools at every stage of the creative process**. We started with three different language models (**ChatGPT**, **Claude** and **Gemini**) to generate the initial concept of "Sudamérica," gathering a wide range of ideas. We chose these models because they are the most widely used, allowing us to **understand what a larger number of users are getting when they inquire about our region**.”

aisongcontest.com/participants-2024/onda-corta

Reading: The Making of Sudamérica (2024)

“Key recurring themes like tango, soccer, and biodiversity were selected and filtered. Using **ElevenLabs**, **we generated over 100 audio clips of 5-10 seconds based on these themes**, as well as South American country names and traditional musical styles. We chose Eleven Labs for its ethical transparency in using its database and because it allowed us to explore brief audio clips instead of complete musical creations. **This gave us greater control over the composition and allowed us to use varied sounds from prompts like "biodiversity" or "Bolivia" as sound textures in our song**.”

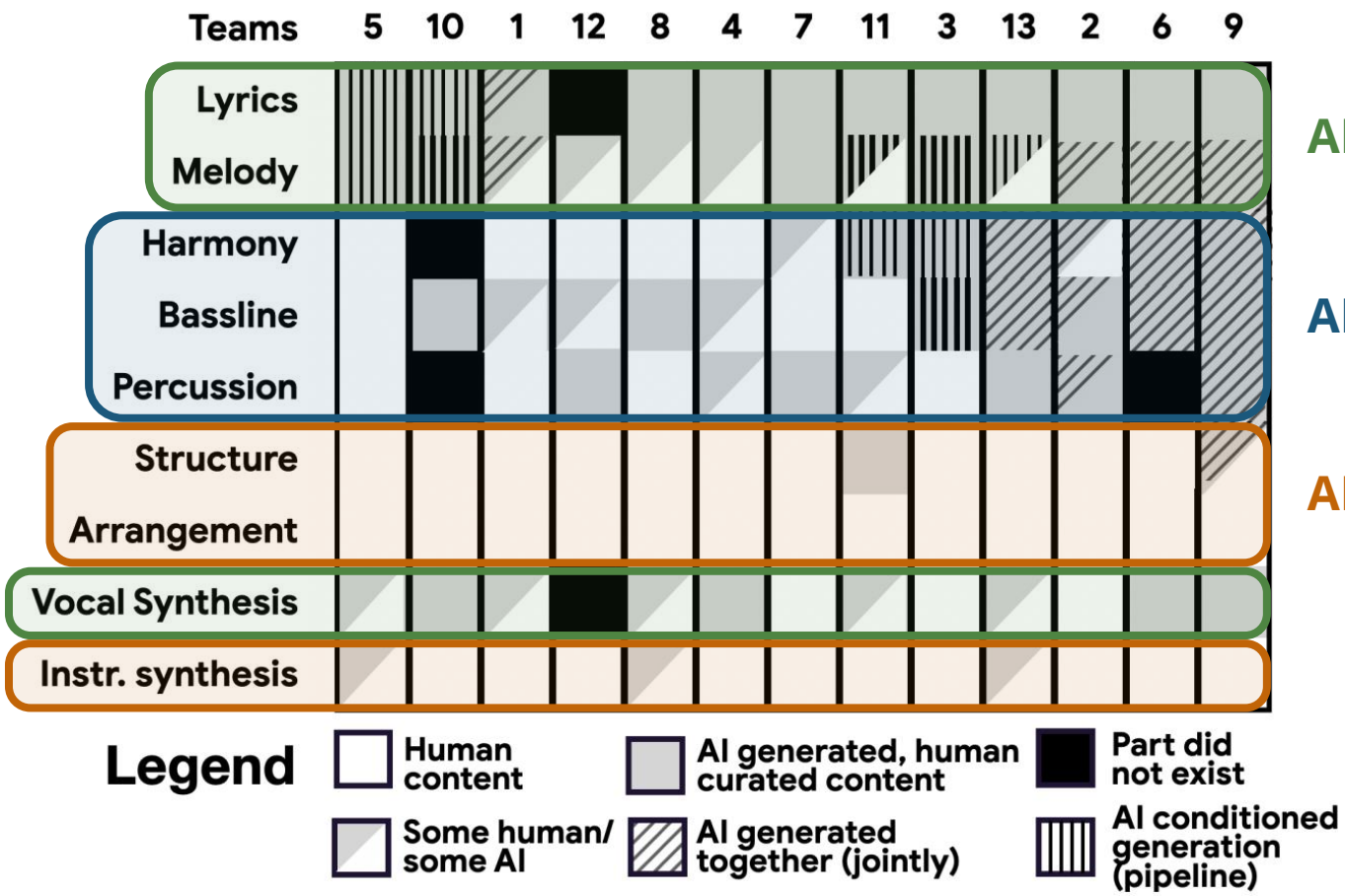
aisongcontest.com/participants-2024/onda-corta

Reading: The Making of Sudamérica (2024)

“For the lyrics, we used **Claude.ai** to generate several options based on selected concepts. We curated the best verses and choruses for thematic and stylistic coherence. Andrés selected 23 audio clips for the song, using **Moises.ai** to separate one key sample into bass, percussion, and piano tracks, which became the song's foundation. He composed the song in a DAW, looping and layering samples, and recorded vocals, further enhanced with **KITS.ai** for a unique touch.”

aisongcontest.com/participants-2024/onda-corta

Analyzing Human-AI Music Co-creation (Huang et al., 2020)



AI is helpful

AI can be helpful

AI is not so helpful

First AI Song Contest! (in 2020)

(Source: Huang et al., 2020)

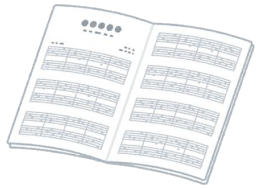
Cheng-Zhi Anna Huang, Hendrik Vincent Koops, Ed Newton-Rex, Monica Dinculescu, and Carrie J. Cai, "AI Song Contest: Human-AI Co-Creation in Songwriting," *ISMIR*, 2020.

🔥 **HW 3**: AI Song Contest 2025

- Instructions will be sent by **emails** and released on the **course website**
- Please listen to the **ten finalists of AI Song Contest 2025**
- **Read the about pages** by clicking the cover arts
- **Answer the following questions** (in 5-10 sentences each)
  - **Which is your favorite song**?
  - Following Q1, **what did they do well**?
  - Following Q1, **what can be improved**?
  - Based on the ten finalists, **what tasks are easy** for current AI in music production?
  - Based on the ten finalists, **what tasks are difficult** for current AI in music production?

Symbolic Music Generation

Four Paradigms of Music Generation



Symbolic music generation

Audio-domain music generation

Text-based

Image-based

Time series-based

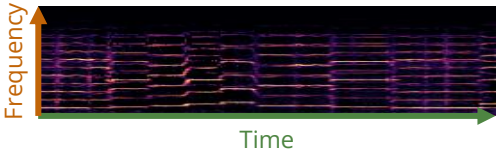
Image-based

```
Program_change_0,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_76, Time_shift_2, Note_off_67,
Note_on_67, Time_shift_2, Note_off_67,
...
```

MIDI

Piano roll

Waveform



Spectrogram

Today, we also have many latent-space based systems!

Topics of Symbolic Music Generation

Unconditional

Symbolic music generation

- ∅ → melody
- ∅ → lead sheet (Melody & chords)
- ∅ → sheet music

Today's topic!

Conditional

Automatic arrangement

- Melody → lead sheet
- Melody → multitrack
- Lead sheet → multitrack
- Solo → multitrack
- Multitrack → simple version

Performance rendering

- Sheet music → performance

Improvisation systems

- Performance → performance

Multimodal

X-to-music generation

- Text → music
- Video → music
- Gestures → music
- Motions → music
- Brain waves → music
- X → music

Two Paradigms of Symbolic Music Generation

Text-based

- Treat music like **text**
- Sharing models with **natural language processing (NLP)**
  - RNNs, LSTMs, Transformers, etc.

Today's topic!

```
Program_change_0,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_76, Time_shift_2, Note_off_67,
Note_on_67, Time_shift_2, Note_off_67, ...
```

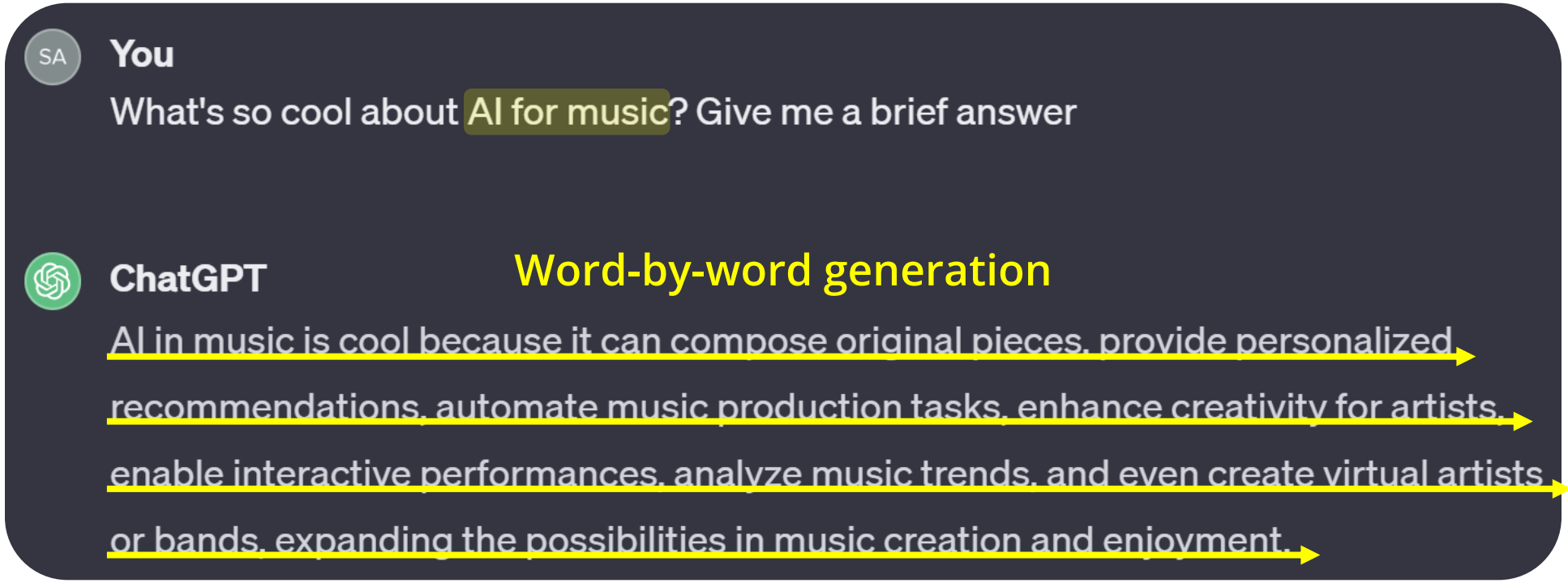
Image-based

- Treat music like **images**
- Sharing models with **computer vision (CV)** & **computer graphics (CG)**
  - GANs, VAEs, diffusion models, etc.

Generating Music like Languages

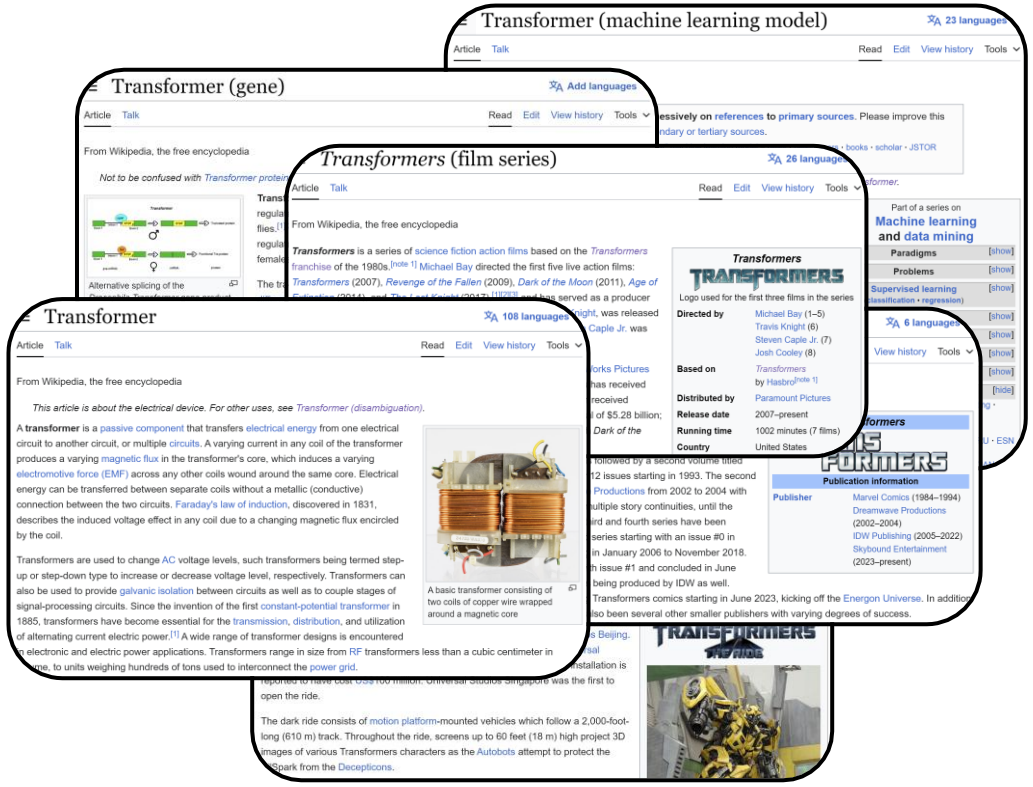
Large Language Models (LLMs)

- The models behind ChatGPT!



Language Models

- Predicting the next word given the past sequence of words



A transformer is a _____

electrical device

fiction character

deep learning model

family of genes

type of food

musical instrument

Language Models (Mathematically)

- A class of machine learning models that learn the next word probability

$$P(\ x_i \mid \underbrace{x_1, x_2, \ldots, x_{i-1}}\ )$$

Next word — Previous words

$P(\ \text{electrical} \mid \text{A transformer is a}\ )$ ↑

$P(\ \text{character} \mid \text{A transformer is a}\ )$ ↑

$P(\ \text{gene} \mid \text{A transformer is a}\ )$ ↑

$P(\ \text{model} \mid \text{A transformer is a}\ )$ ↑

$P(\ \text{food} \mid \text{A transformer is a}\ )$ ↓

$P(\ \text{musical} \mid \text{A transformer is a}\ )$ ↓

Music Language Models (Mathematically)

- A class of machine learning models that learn the next note probability

$$P(\ x_i \mid \underbrace{x_1, x_2, \ldots, x_{i-1}}\ )$$

Next note — Previous notes

$P(\ \mathrm{G} \mid C\ C\ G\ G\ A\ A\ )$ ↑

$P(\ \mathrm{A} \mid C\ C\ G\ G\ A\ A\ )$ ↑

$P(\ \mathrm{C} \mid C\ C\ G\ G\ A\ A\ )$ ↑

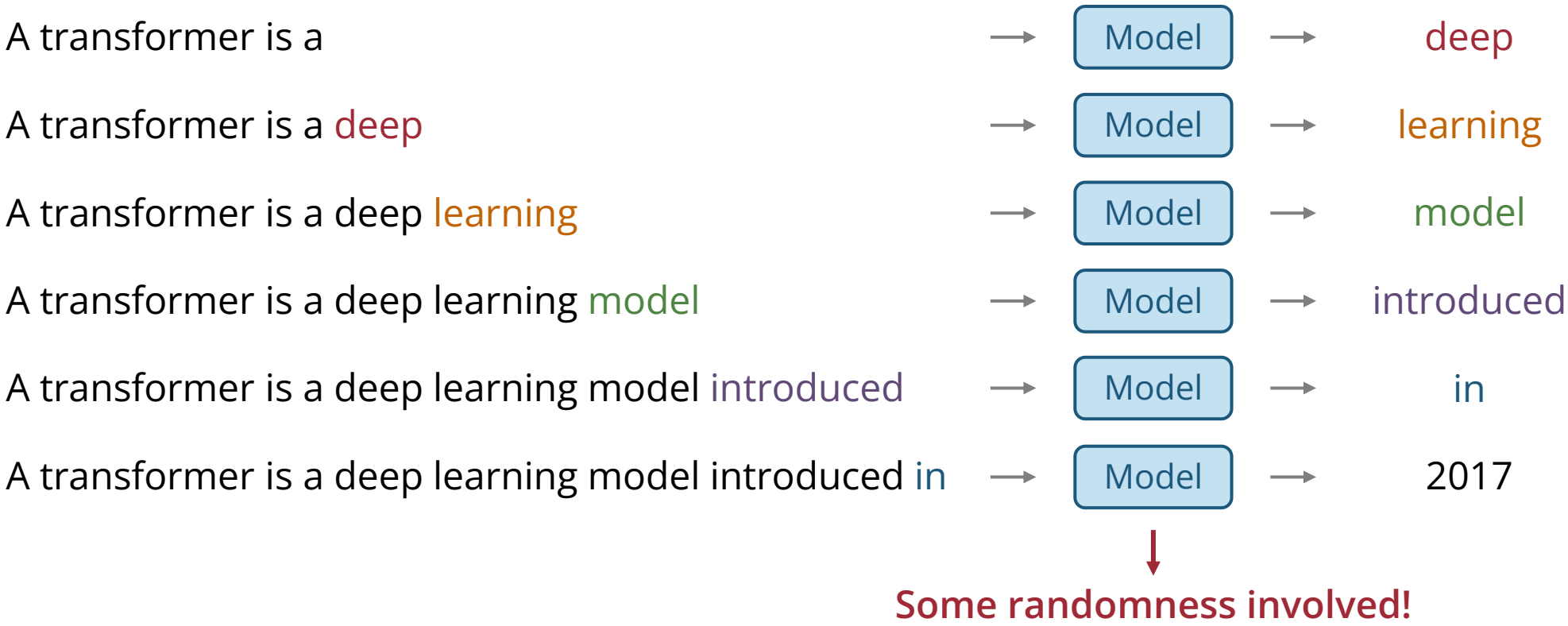
$P(\ \mathrm{F} \mid C\ C\ G\ G\ A\ A\ )$ ↑

$P(\ \mathrm{A}\flat \mid C\ C\ G\ G\ A\ A\ )$ ↓

$P(\ \mathrm{A}\sharp \mid C\ C\ G\ G\ A\ A\ )$ ↓

Language Models: Generation

- How do we generate a new sentence using a trained language model?



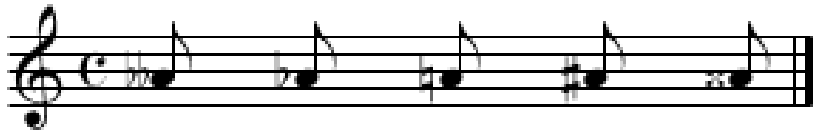
Designing a Machine-readable Music Language

- How can we “represent” music in a way that machines understand?

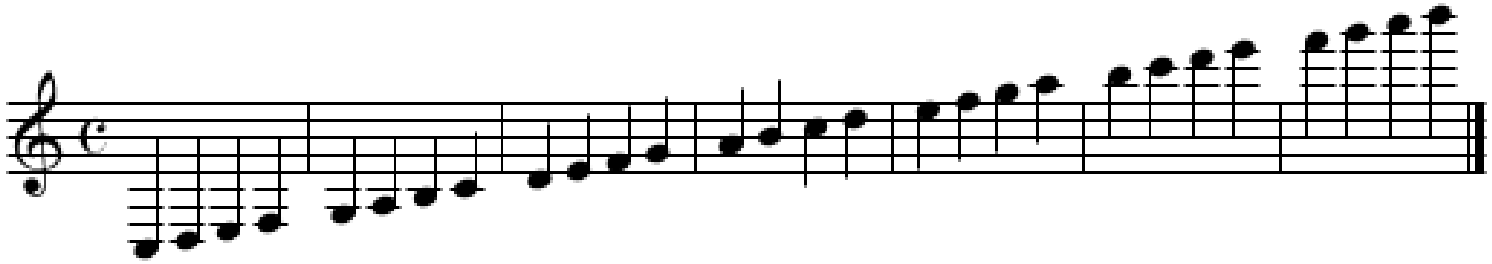
ABC Notation-based Representation

ABC Notation

- A simple text-based notation
- Use **letters** to denote **pitches**
  - Lower octave (A–G), higher octave (a–g)
- Use **prefix** to denote **accidentals**
  - Sharp (^), flat (_), natural (=)



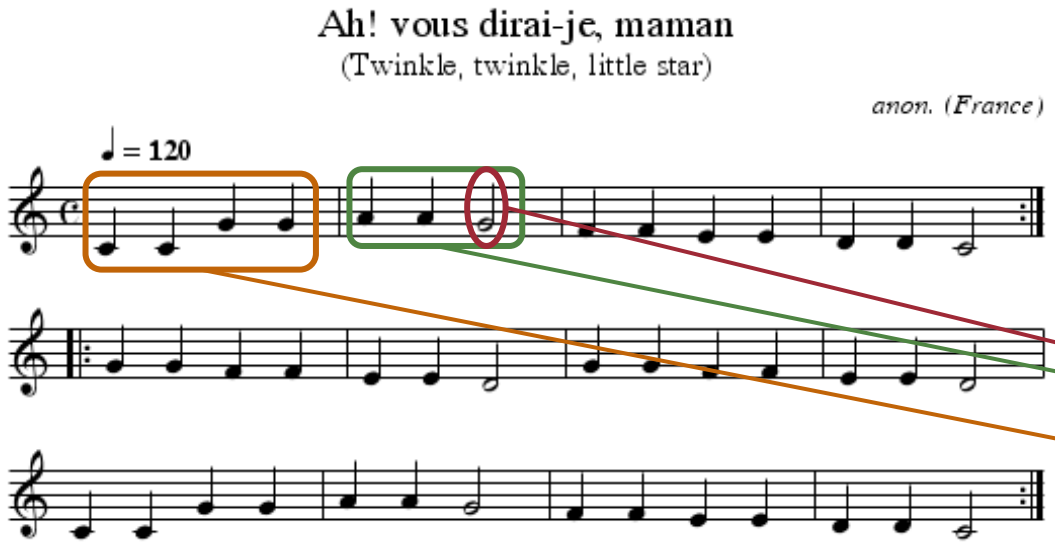
__A _A =A ^A ^^A



C, D, E, F,|G, A, B, C| D E F G | A B c d | e f g a | b c' d' e'| f' g' a' b'

abcnotation.com/examples

An Example of ABC Notation



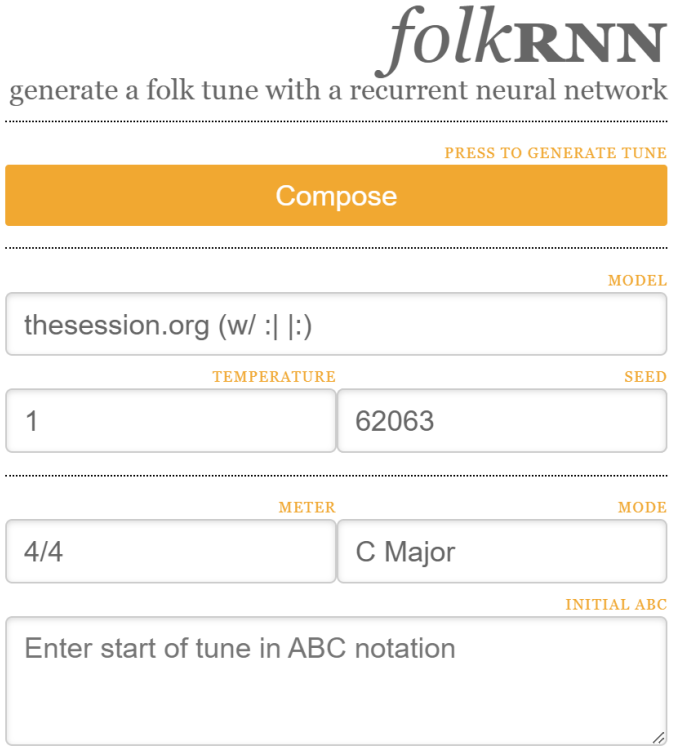
Metadata

```
X:571
T:Ah! vous dirai-je, maman
T:(Twinkle, twinkle, little star)
C:anon.
O:France
R:Nursery song
M:C  Meter
L:1/4  Unit note length (temporal resolution)
Q:120  Tempo
K:C  Key
CCGG|AAG2|FFEE|DDC2:|
|:GGFF|EED2|GGFF|EED2|
CCGG|AAG2|FFEE|DDC2:|
```

abcnotation.com/tunePage?a=trillian.mit.edu/~jc/music/abc/mirror/musicaviva.com/france/ah-vous-c/ah-vous-c-1/0000

Folk RNN (Sturm et al., 2015)

- **Data**
  - 23,958 folk tunes
- **Representation**
  - ABC notation without metadata
- **Model**
  - LSTM (long short-term memory)
  - Working on the character level

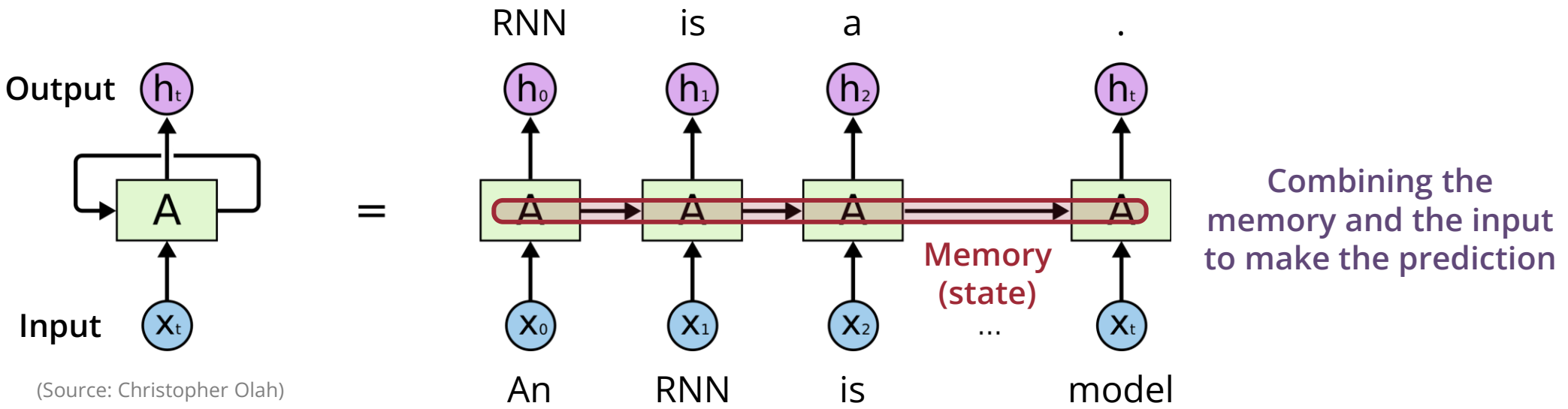


folkrnn.org

Bob L. Sturm, Joao Felipe Santos, and Iryna Korshunova, "Folk Music Style Modelling by Recurrent Neural Networks with Long Short Term Memory Units," *ISMIR Late-Breaking Demos*, 2015.

What is an RNN (Recurrent Neural Network)?

- A type of neural networks that have **loops**
- Widely used for **modeling sequences** (e.g., in natural language processing)

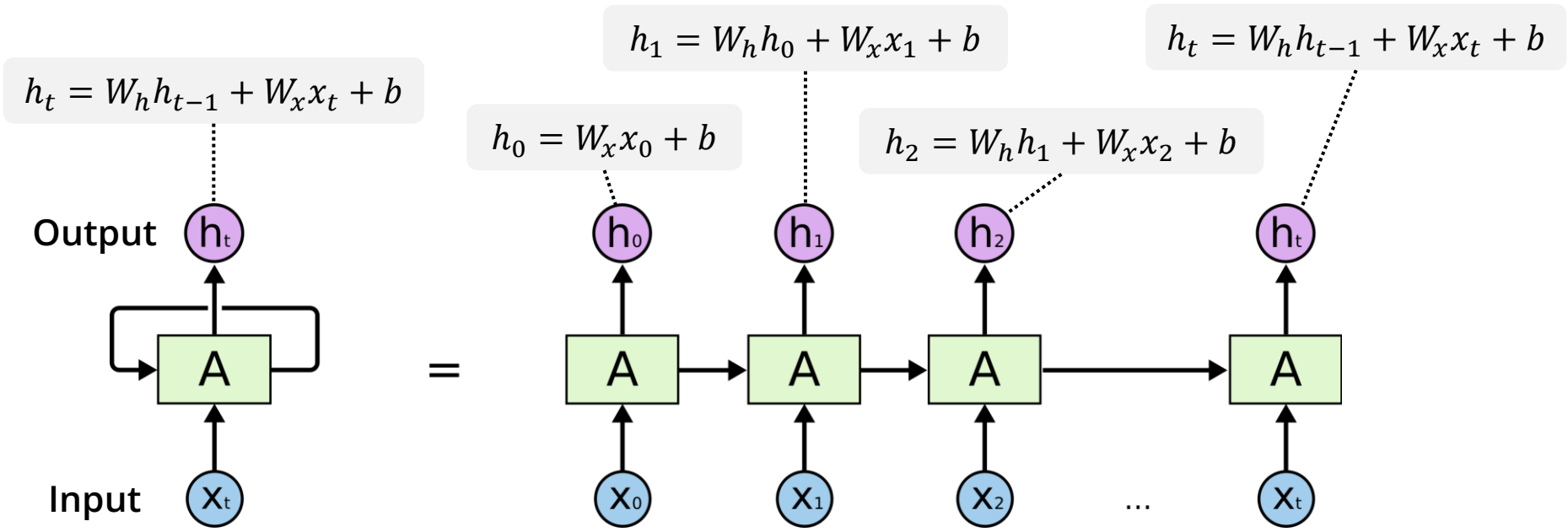


(Source: Christopher Olah)

colah.github.io/posts/2015-08-Understanding-LSTMs/

Vanilla RNNs

- The simplest form of RNNs
- LSTMs and GRUs are also RNNs

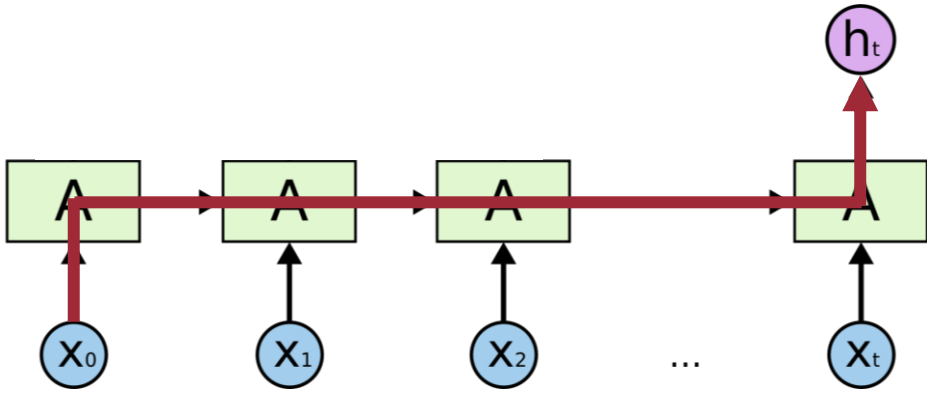


(Source: Christopher Olah)

colah.github.io/posts/2015-08-Understanding-LSTMs/

Backpropagation Through Time

- An RNN is essentially a **very deep neural network**



$$h_t = W_h h_{t-1} + W_x x_t + b$$

$$h_t = W_h(W_h h_{t-2} + W_x x_{t-1} + b\ ) + W_x x_t + +b$$

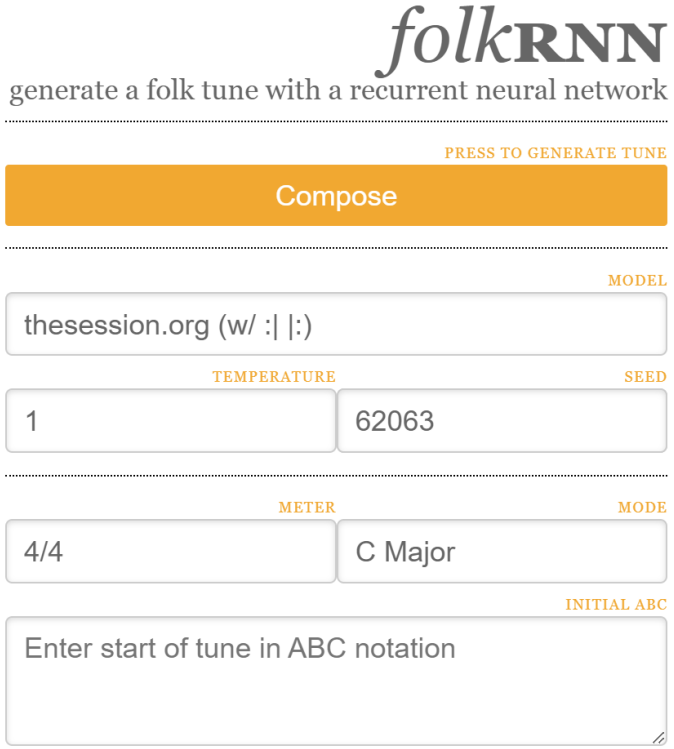
$$\vdots$$

$$h_t = W_h(W_x x_{t-1} + W_h(\ \cdots\ \ W_h h_0 + W_x x_1 + b\ \ \cdots\ ) + b\ ) + W_x x_t + +b$$

colah.github.io/posts/2015-08-Understanding-LSTMs/

Folk RNN (Sturm et al., 2015)

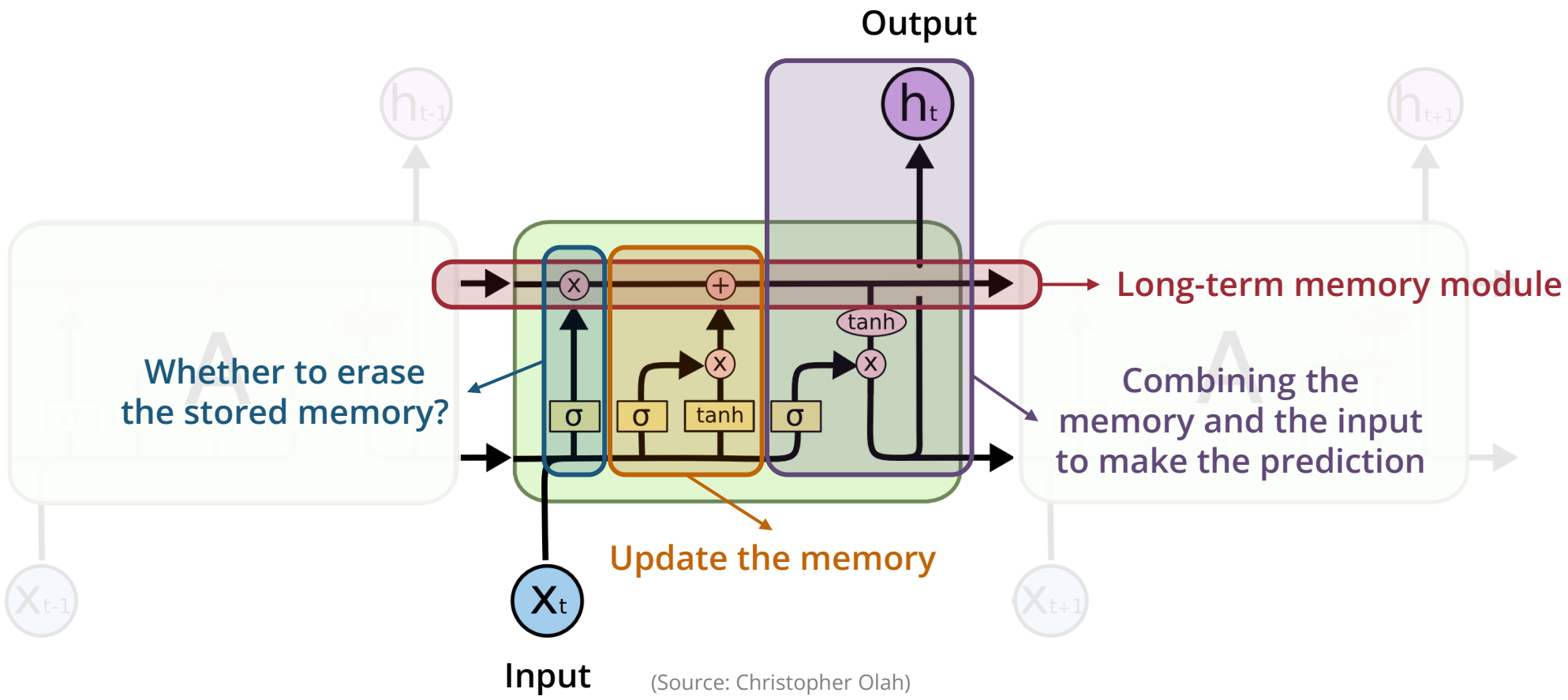
- **Data**
  - 23,958 folk tunes
- **Representation**
  - ABC notation without metadata
- **Model**
  - LSTM (long short-term memory)
  - Working on the character level



folkrnn.org

Bob L. Sturm, Joao Felipe Santos, and Iryna Korshunova, "Folk Music Style Modelling by Recurrent Neural Networks with Long Short Term Memory Units," *ISMIR Late-Breaking Demos*, 2015.

Demystifying LSTMs (Hochreiter & Schmidhuber, 1997)

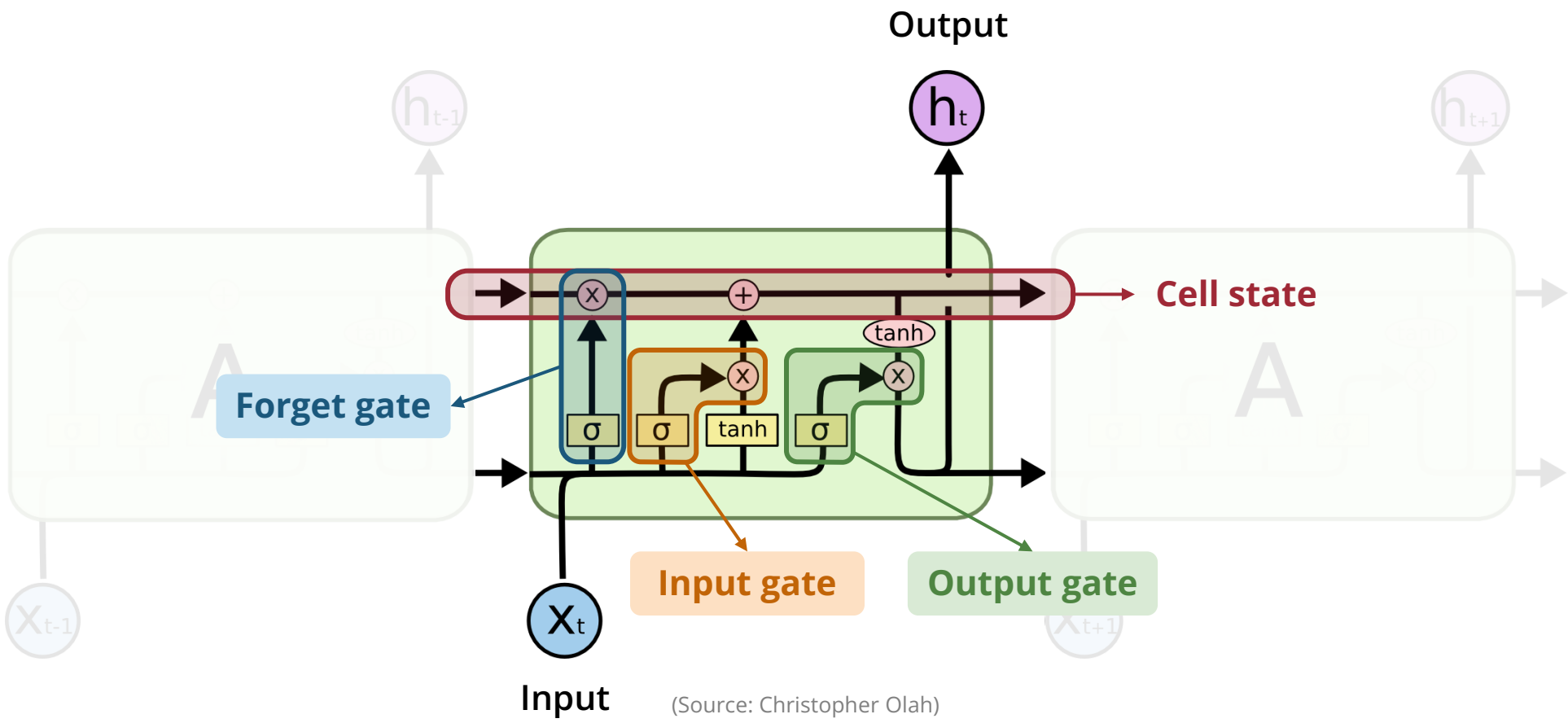


(Source: Christopher Olah)

colah.github.io/posts/2015-08-Understanding-LSTMs/

Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 9(8):1735-1780, 1997.

Demystifying LSTMs (Hochreiter & Schmidhuber, 1997)



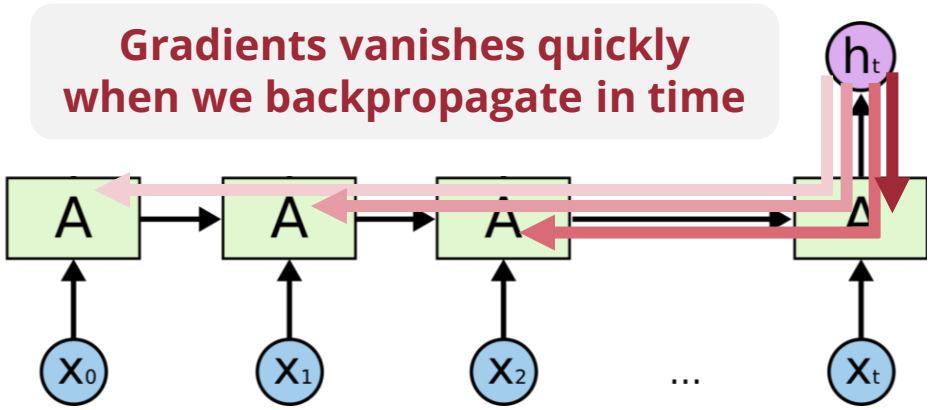
(Source: Christopher Olah)

colah.github.io/posts/2015-08-Understanding-LSTMs/

Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 9(8):1735-1780, 1997.

Vanishing Gradients

- An RNN is essentially a **very deep neural network**



(Source: Christopher Olah)

All the layers share the same weight matrix

Can still train the model without deeper gradients

Why bother?

$$h_t = W_h h_{t-1} + W_x x_t + b$$

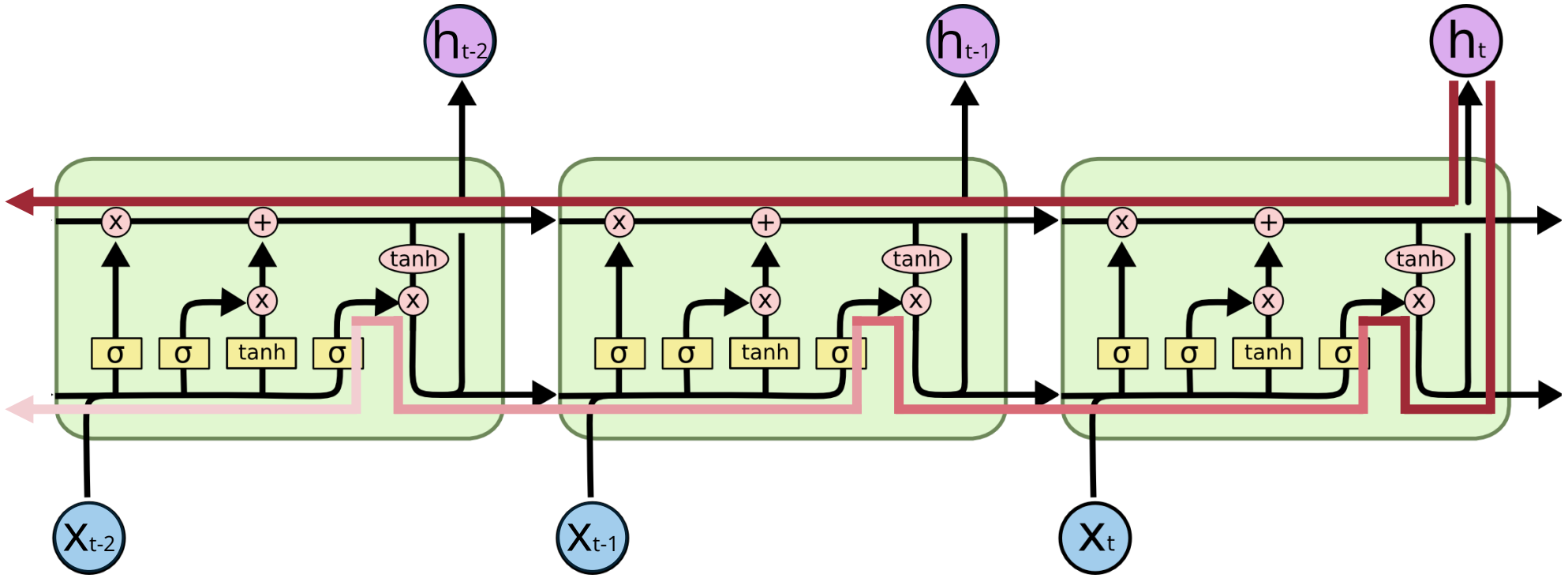
$$h_t = W_h(W_h h_{t-2} + W_x x_{t-1} + b\ ) + W_x x_t + +b$$

$$\vdots$$

$$h_t = W_h(W_x x_{t-1} + W_h(\ \cdots\ W_h h_0 + W_x x_1 + b\ \cdots\ ) + b\ ) + W_x x_t + +b$$

colah.github.io/posts/2015-08-Understanding-LSTMs/

How can LSTMs Help Alleviate Vanishing Gradients?



(Source: Christopher Olah)

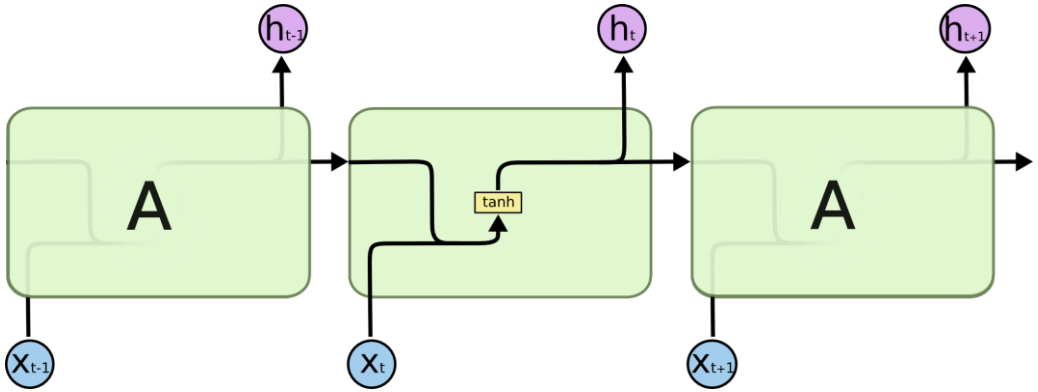
LSTMs does not completely solve vanishing gradients

colah.github.io/posts/2015-08-Understanding-LSTMs/

Vanilla RNNs vs LSTMs

Vanilla RNN

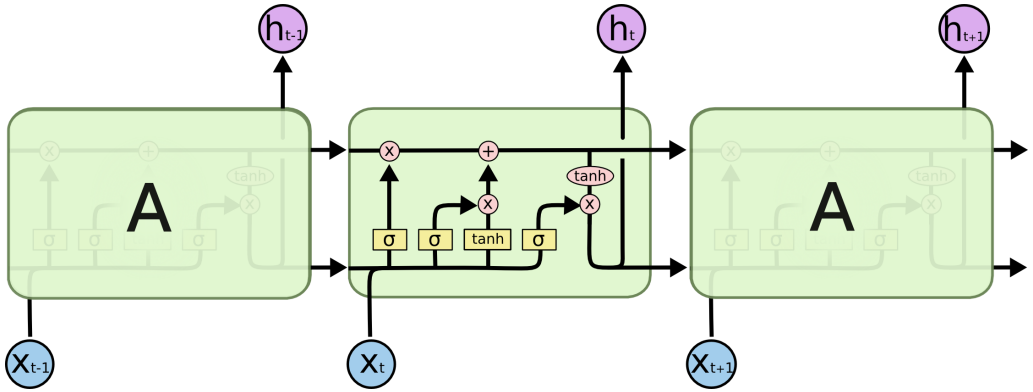
- Simplest form of RNNs
- Limited long-term memory
- Harder to train (due to gradient vanishing)



(Source: Christopher Olah)

LSTM

- Improved memory module
- Better long-term memory
- Easier to train

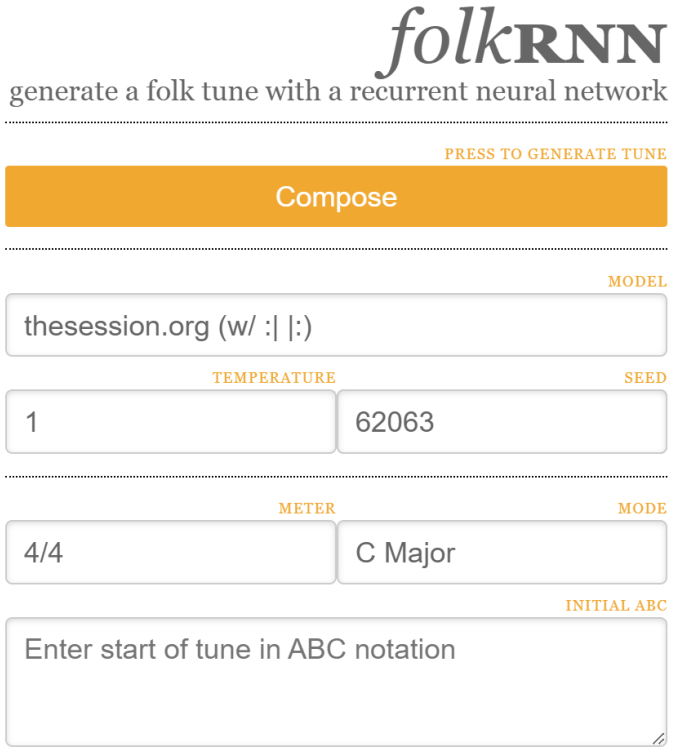


(Source: Christopher Olah)

colah.github.io/posts/2015-08-Understanding-LSTMs/

Folk RNN (Sturm et al., 2015)

- **Data**
  - 23,958 folk tunes
- **Representation**
  - ABC notation without metadata
- **Model**
  - LSTM (long short-term memory)
  - Working on the character level



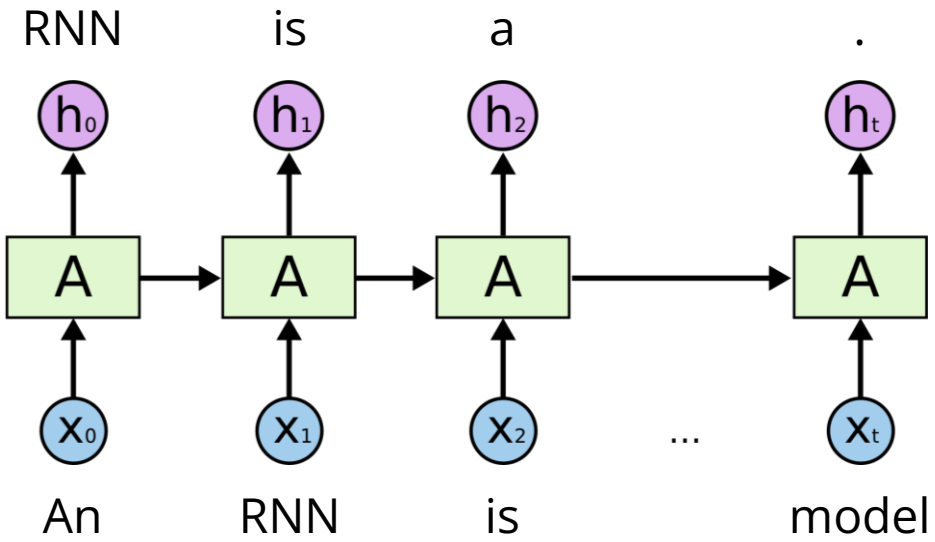
folkrnn.org

Bob L. Sturm, Joao Felipe Santos, and Iryna Korshunova, "Folk Music Style Modelling by Recurrent Neural Networks with Long Short Term Memory Units," *ISMIR Late-Breaking Demos*, 2015.

Word-level vs Character-level RNNs

Word-level RNNs

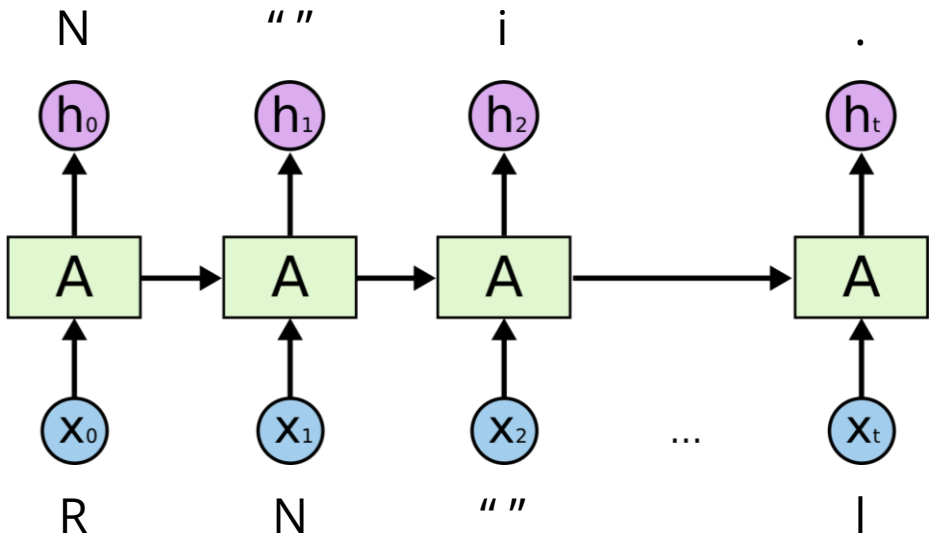
- Predicting word by word
- Most common



(Source: Christopher Olah)

Character-level RNNs

- Predicting character by character
- Useful when there is no natural “spaces”



(Source: Christopher Olah)

colah.github.io/posts/2015-08-Understanding-LSTMs/

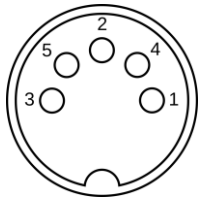
Limitations of ABC Notations

- Limited expressiveness
- Monophonic tunes only

MIDI-like Representation

MIDI (Musical Instrument Digital Interface)

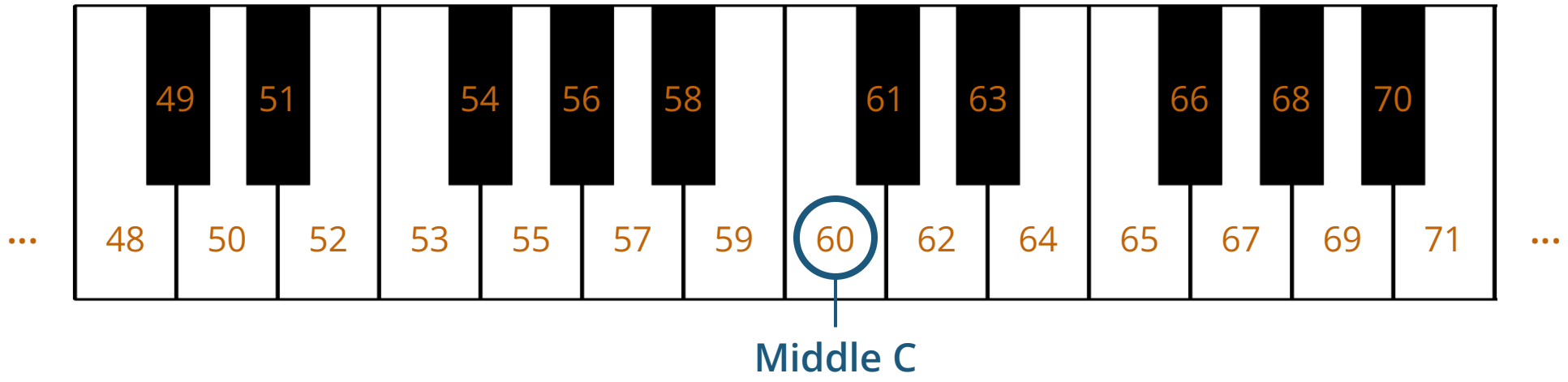
- A communication protocol between devices
- MIDI Messages
  - Note on
  - Note off
  - Delta time
  - Program change
  - Control change
  - Pitch bend change



MIDI I/O

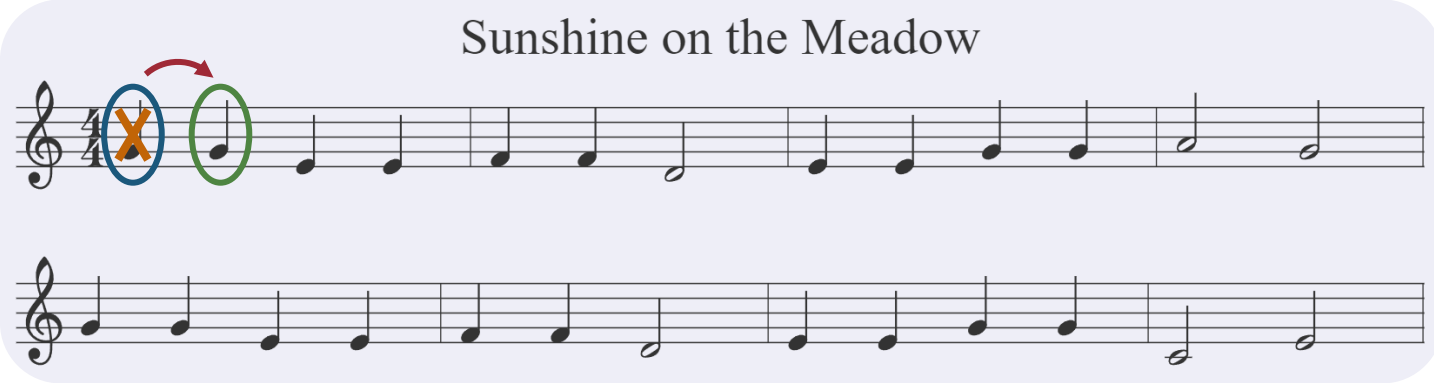
MIDI Note Numbers

- Ranging from 0 to 127
  - Middle C is 60
  - Wider than standard piano's pitch range
- Widely used in various software, keyboards and algorithms



Representing Music using MIDI Messages

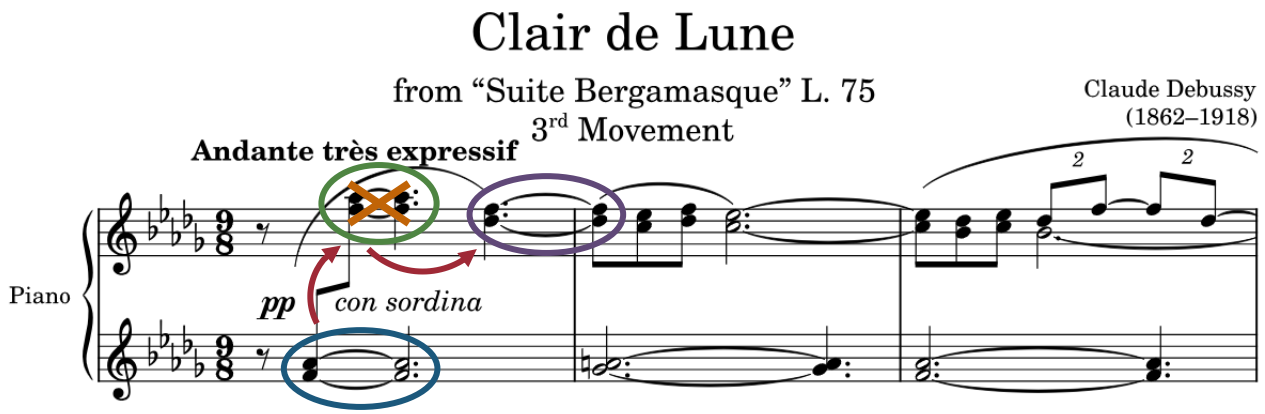
- Three main MIDI messages
  - Note on
  - Note off
  - Time Shift




```
Note_on_67, Time_shift_quarter_note, Note_off_67,
Note_on_67, Time_shift_quarter_note, Note_off_67,
Note_on_64, Time_shift_quarter_note, Note_off_64,
Note_on_64, Time_shift_quarter_note, Note_off_64,
...
```

Representing Polyphonic Music

- We can now handle music with multi-pitch at the same time
  - In the literature, "polyphonic" & "multi-pitch" are often used interchangeably




```
Note_on_65, Note_on_68, Time_shift_eighth_note, Note_on_77, Note_on_80,
Time_shift_half_note, Note_off_77, Note_off_80, Note_on_73, Note_on_77,
Time_shift_dotted_quarter_note, Note_off_65, Note_off_68, ...
```

Performance RNN (Oore et al., 2020)

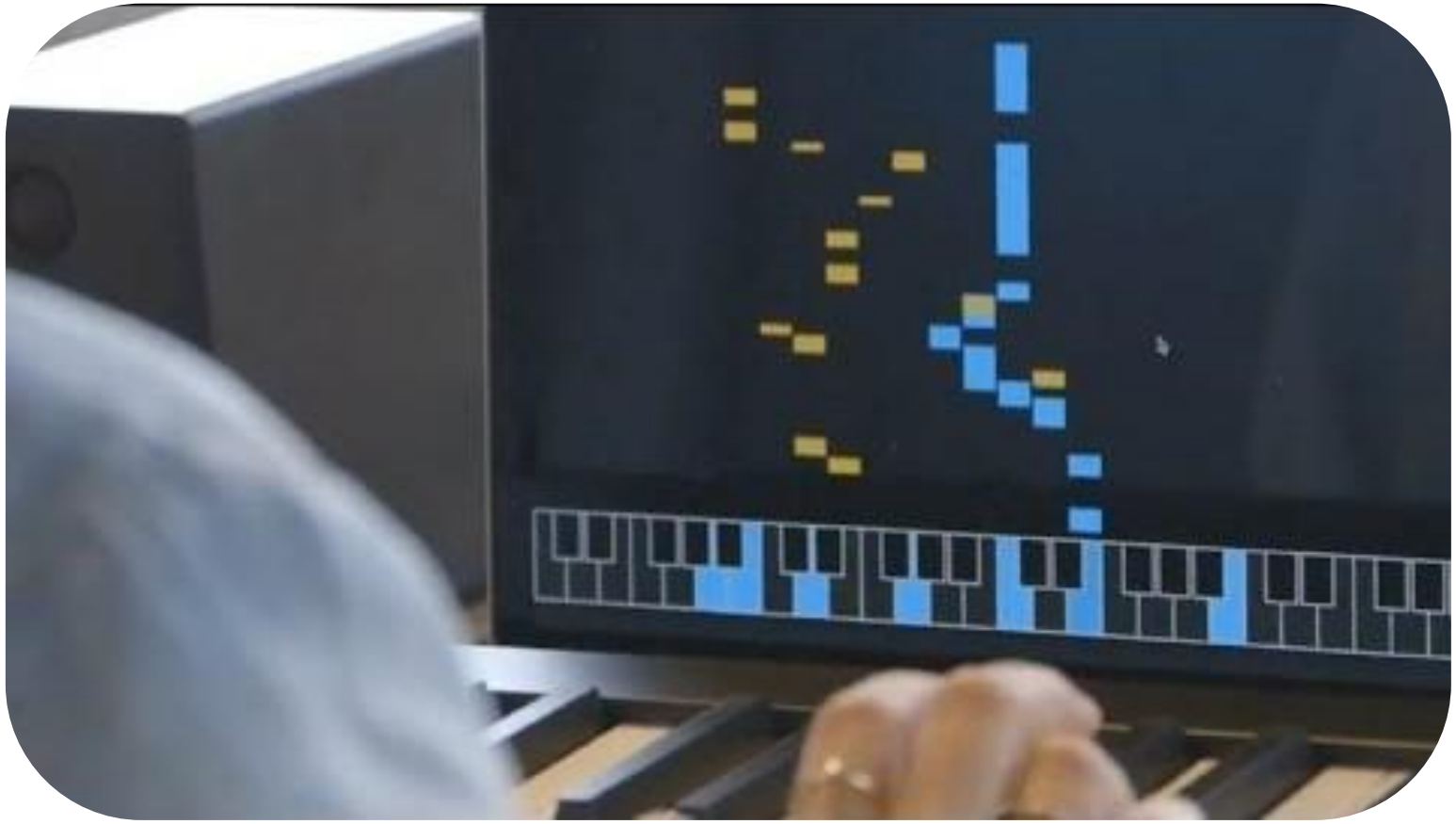
- **Data**
  - Yamaha e-Piano Competition dataset (MAESTRO)
- **Representation**
  - 128 Note-On events
  - 128 Note-Off events
  - 125 Time-Shift events (8ms–1s)
  - 32 Set-Velocity events — Handle dynamics
- **Model**
  - LSTM

Examples of generated music



Ian Simon and Sageev Oore, "Performance RNN: Generating Music with Expressive Timing and Dynamics," *Magenta Blog*, June 29, 2017.
Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan, "This Time with Feeling: Learning Expressive Musical Performance", *Neural Computing and Applications*, 32, 2020.

A.I. Duet (Mann et al, 2016)



youtu.be/0ZE1bfPtvZo
experiments.withgoogle.com/ai/ai-duet/view

github.com/googlecreativelab/aiexperiments-ai-duet

Music Transformer (Huang et al., 2019)

- **Data**: Yamaha e-Piano Competition dataset (MAESTRO)
- **Representation**
  - 128 Note-On events
  - 128 Note-Off events
  - 100 Time-Shift events (10ms–1s)
  - 32 Set-Velocity events
- **Model**: Transformer

Almost the same representation as PerformanceRNN

Handle dynamics

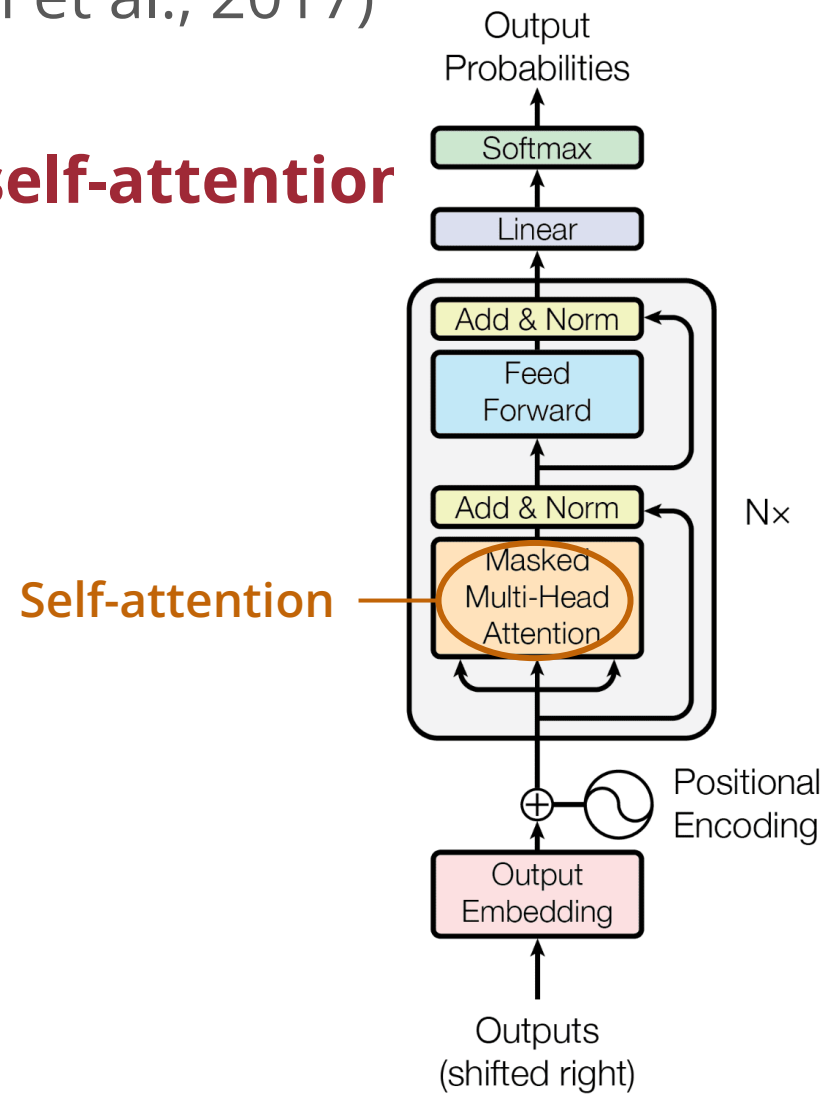
Examples of generated music



Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *ICLR*, 2019.

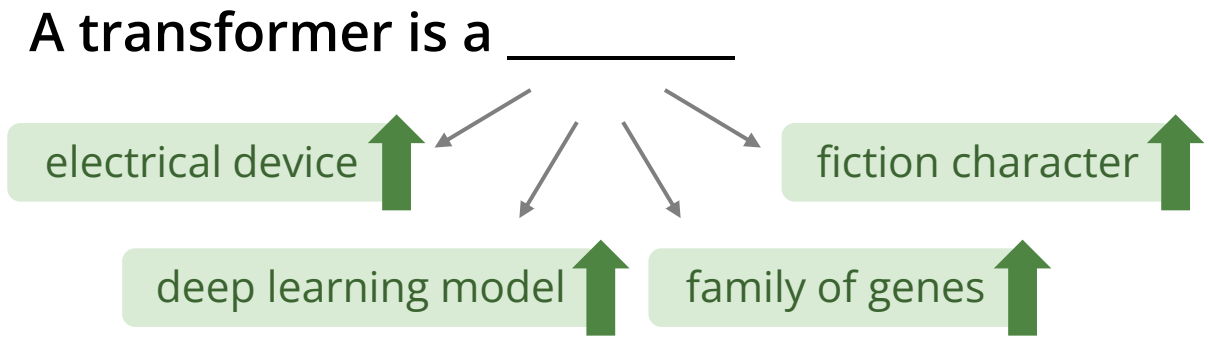
What is a Transformer? (Vaswani et al., 2017)

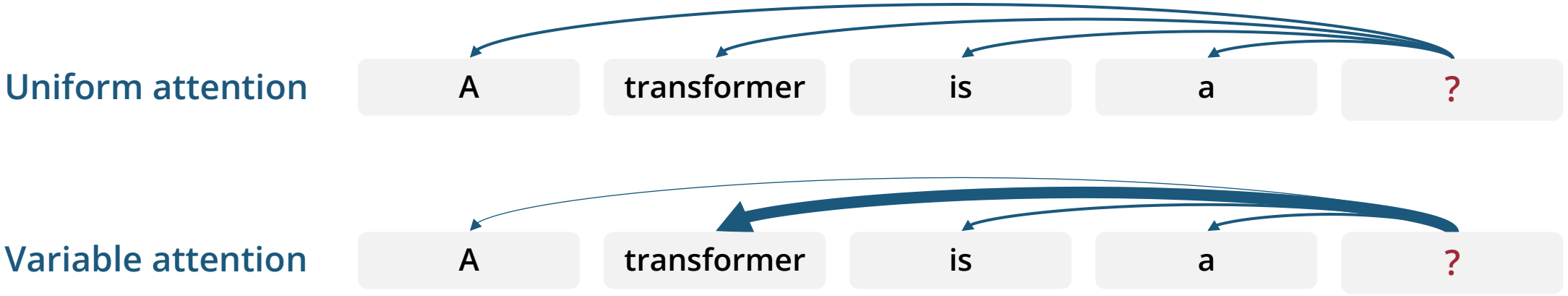
- A type of neural network that use the **self-attentior**



(Source: Vaswani et al., 2017; adapted)

Self-attention Mechanism (Cheng et al., 2016)

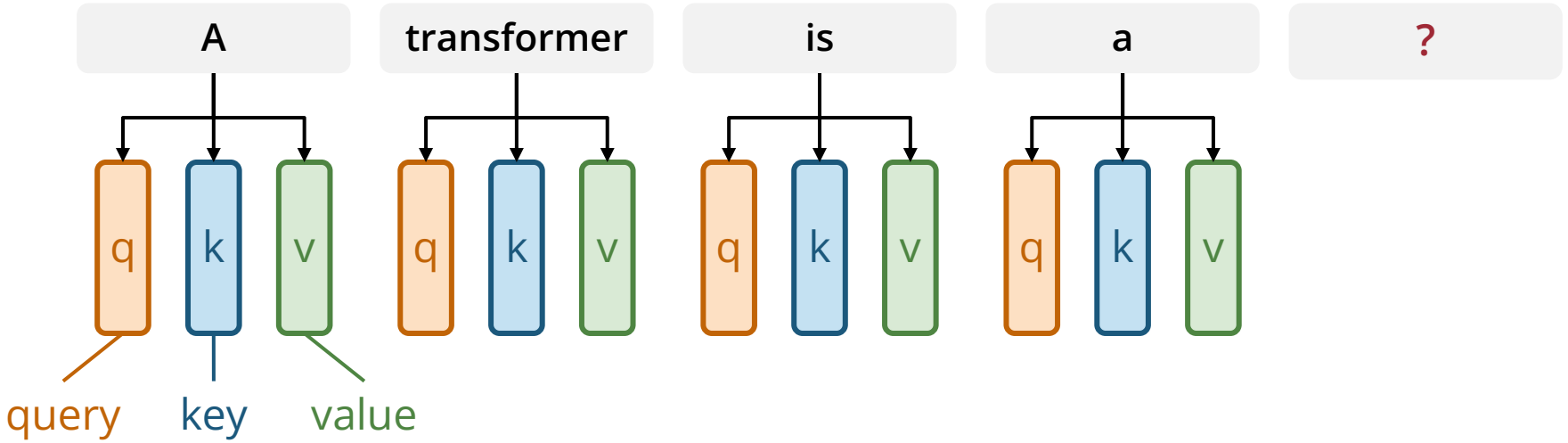




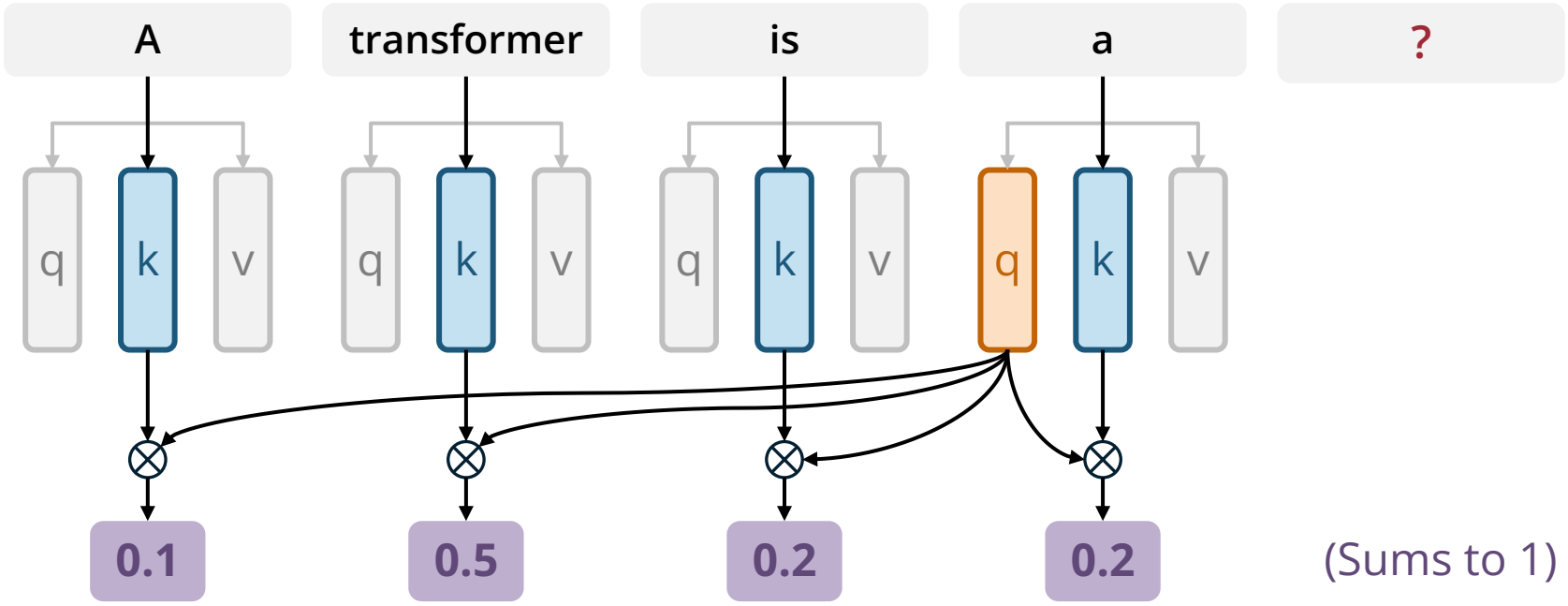
Transformers learn what to attend to from big data!

Jianpeng Cheng, Li Dong, and Mirella Lapata, "Long Short-Term Memory-Networks for Machine Reading," *EMNLP*, 2016.

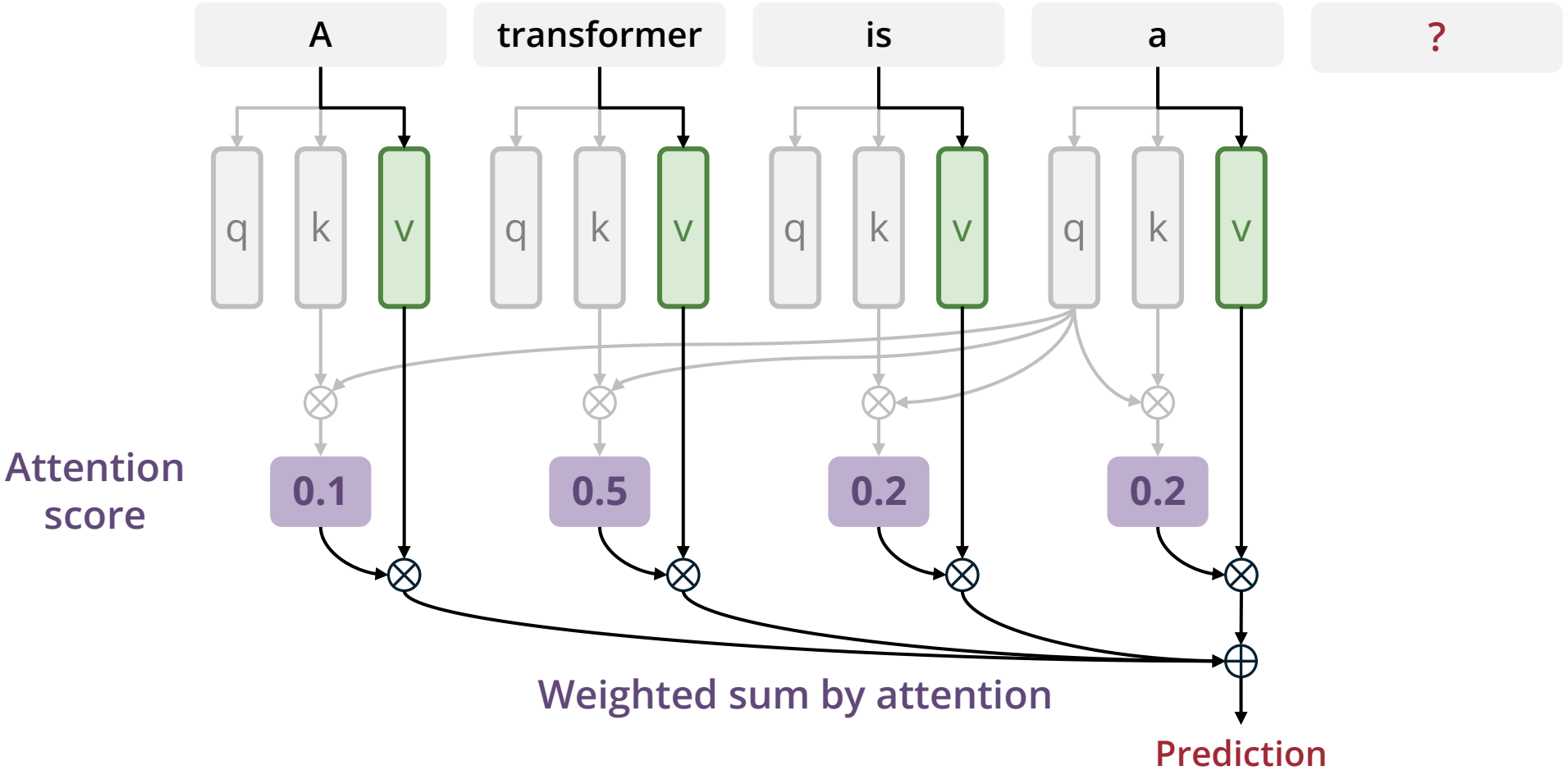
Demystifying Transformers (Vaswani et al., 2017)



Demystifying Transformers (Vaswani et al., 2017)

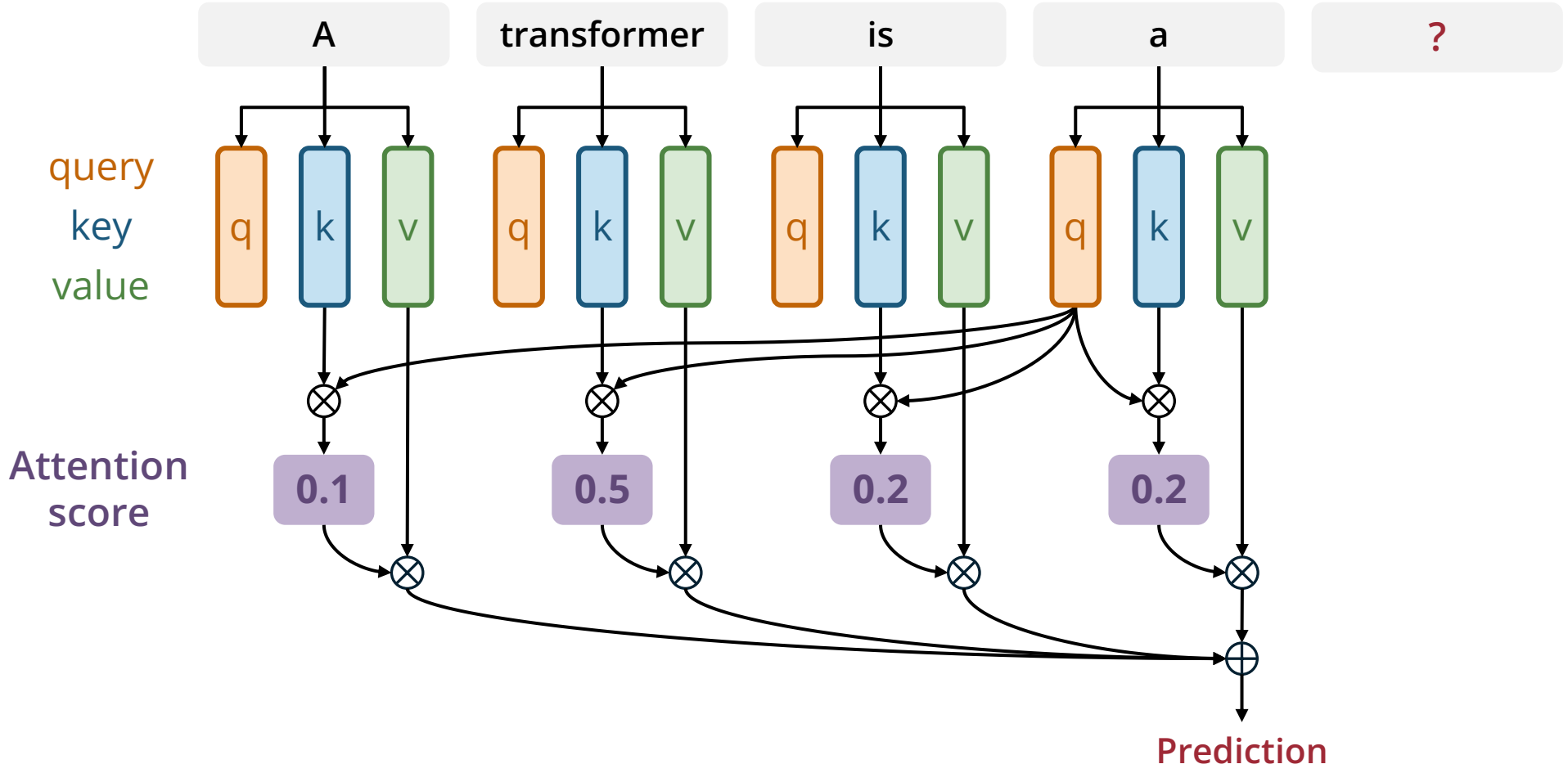


Demystifying Transformers (Vaswani et al., 2017)

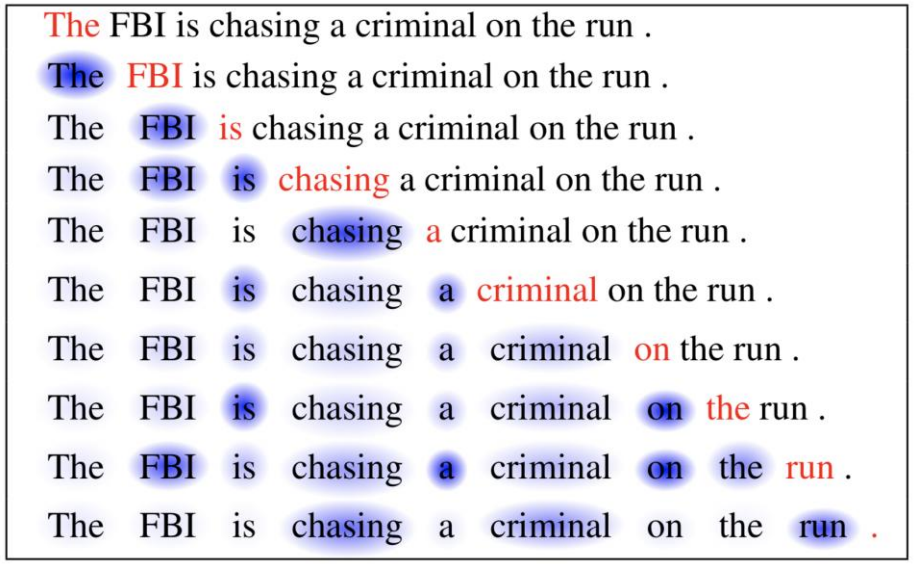


Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

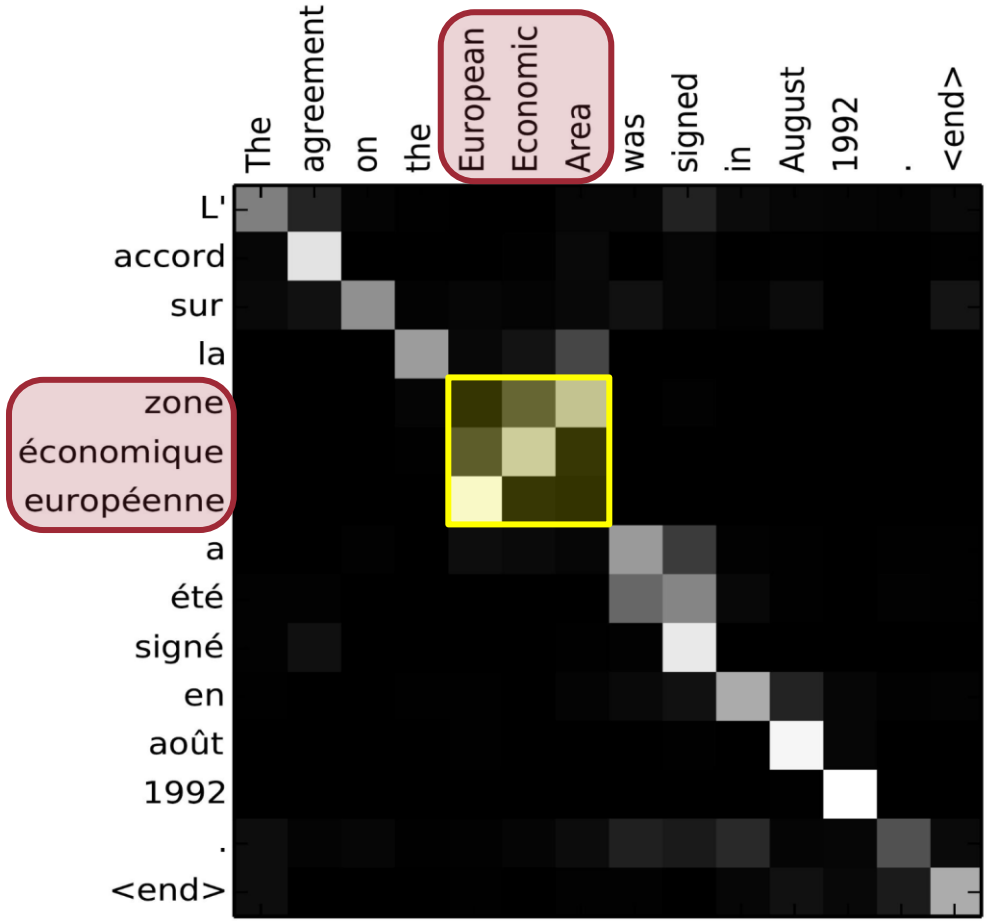
Demystifying Transformers (Vaswani et al., 2017)



Why Attention Mechanism?



(Source: Cheng et al., 2016)

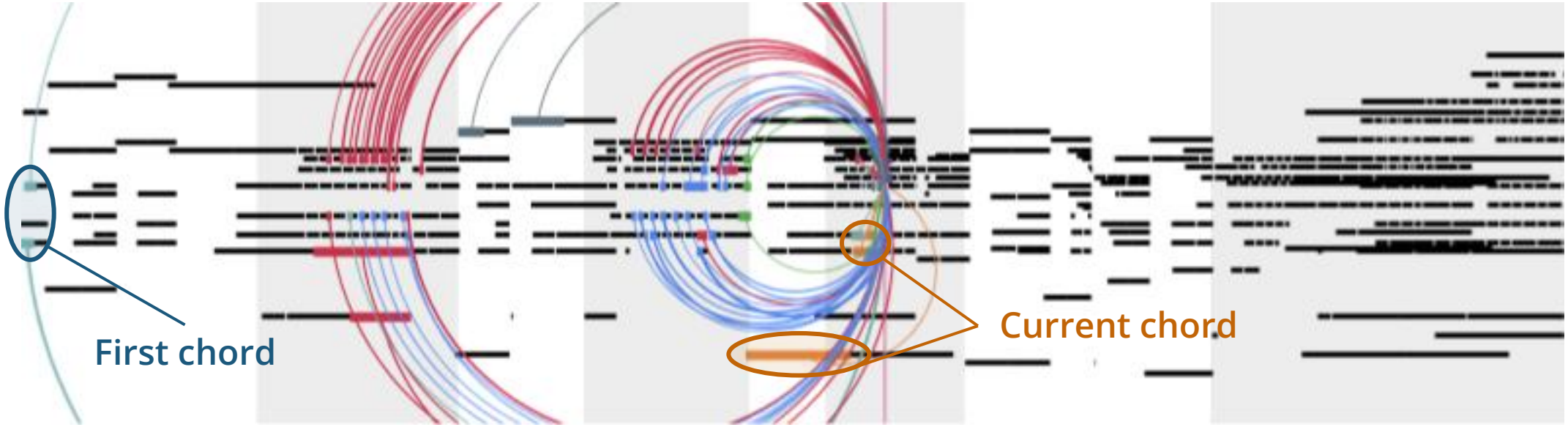


(Source: Bahdanau et al., 2015)

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *ICLR*, 2015.
Jianpeng Cheng, Li Dong, and Mirella Lapata, "Long Short-Term Memory-Networks for Machine Reading," *EMNLP*, 2016.

Visualizing Musical Self-attention

(Each color represents an attention head)

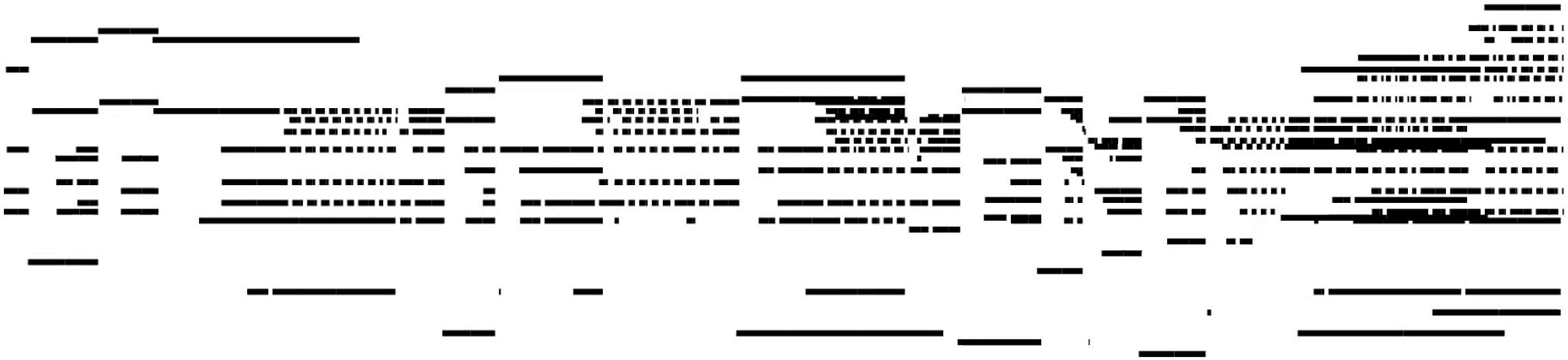


(Source: Huang et al., 2018)

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *ICLR*, 2019.
Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *Magenta Blog*, December 13, 2018.

Visualizing Musical Self-attention

(Each color represents an attention head)



(Source: Huang et al., 2018)

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *ICLR*, 2019.
Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *Magenta Blog*, December 13, 2018.

Beyond Solo Music

Representing Multiple Instruments

- Using **MIDI program change** messages
  - Program numbers: 1–128 (or 0–127)
  - 128 instruments in 16 families

Prog#	INSTRUMENT
	1-8 PIANO
1	Acoustic Grand
2	Bright Acoustic
3	Electric Grand
4	Honky-Tonk
5	Electric Piano 1
6	Electric Piano 2
7	Harpsichord
8	Clav

Prog#	INSTRUMENT	Prog#	INSTRUMENT
	1-8 PIANO		**9-16 CHROMATIC PERCUSSION**
1	Acoustic Grand	9	Celesta
2	Bright Acoustic	10	Glockenspiel
3	Electric Grand	11	Music Box
4	Honky-Tonk	12	Vibraphone
5	Electric Piano 1	13	Marimba
6	Electric Piano 2	14	Xylophone
7	Harpsichord	15	Tubular Bells
8	Clav	16	Dulcimer
	17-24 ORGAN		**25-32 GUITAR**
17	Drawbar Organ	25	Acoustic Guitar(nylon)
18	Percussive Organ	26	Acoustic Guitar(steel)
19	Rock Organ	27	Electric Guitar(jazz)
20	Church Organ	28	Electric Guitar(clean)
21	Reed Organ	29	Electric Guitar(muted)
22	Accoridan	30	Overdriven Guitar
23	Harmonica	31	Distortion Guitar
24	Tango Accordian	32	Guitar Harmonics
	33-40 BASS		**41-48 STRINGS**
33	Acoustic Bass	41	Violin
34	Electric Bass(finger)	42	Viola
35	Electric Bass(pick)	43	Cello
36	Fretless Bass	44	Contrabass
37	Slap Bass 1	45	Tremolo Strings
38	Slap Bass 2	46	Pizzicato Strings
39	Synth Bass 1	47	Orchestral Strings
40	Synth Bass 2	48	Timpani
	49-56 ENSEMBLE		**57-64 BRASS**
49	String Ensemble 1	57	Trumpet
50	String Ensemble 2	58	Trombone
51	SynthStrings 1	59	Tuba
52	SynthStrings 2	60	Muted Trumpet
53	Choir Aahs	61	French Horn
54	Voice Oohs	62	Brass Section
55	Synth Voice	63	SynthBrass 1
56	Orchestra Hit	64	SynthBrass 2

	65-72 REED		**73-80 PIPE**
65	Soprano Sax	73	Piccolo
66	Alto Sax	74	Flute
67	Tenor Sax	75	Recorder
68	Baritone Sax	76	Pan Flute
69	Oboe	77	Blown Bottle
70	English Horn	78	Shakuhachi
71	Bassoon	79	Whistle
72	Clarinet	80	Ocarina
	81-88 SYNTH LEAD		**89-96 SYNTH PAD**
81	Lead 1 (square)	89	Pad 1 (new age)
82	Lead 2 (sawtooth)	90	Pad 2 (warm)
83	Lead 3 (calliope)	91	Pad 3 (polysynth)
84	Lead 4 (chiff)	92	Pad 4 (choir)
85	Lead 5 (charang)	93	Pad 5 (bowed)
86	Lead 6 (voice)	94	Pad 6 (metallic)
87	Lead 7 (fifths)	95	Pad 7 (halo)
88	Lead 8 (bass+lead)	96	Pad 8 (sweep)
	97-104 SYNTH EFFECTS		**105-112 ETHNIC**
97	FX 1 (rain)	105	Sitar
98	FX 2 (soundtrack)	106	Banjo
99	FX 3 (crystal)	107	Shamisen
100	FX 4 (atmosphere)	108	Koto
101	FX 5 (brightness)	109	Kalimba
102	FX 6 (goblins)	110	Bagpipe
103	FX 7 (echoes)	111	Fiddle
104	FX 8 (sci-fi)	112	Shanai
	113-120 PERCUSSIVE		**121-128 SOUND EFFECTS**
113	Tinkle Bell	121	Guitar Fret Noise
114	Agogo	122	Breath Noise
115	Steel Drums	123	Seashore
116	Woodblock	124	Bird Tweet
117	Taiko Drum	125	Telephone Ring
118	Melodic Tom	126	Helicopter
119	Synth Drum	127	Applause
120	Reverse Cymbal	128	Gunshot

(Source: Roger Dannenberg)

www.cs.cmu.edu/~music/cmsip/readings/GMSpecs_Patches.htm

MuseNet (Payne et al., 2019)

- **Data**: ClassicalArchives + BitMidi + MAESTRO
- **Representation**: "**instrument:velocity:pitch**"
  - Time shifts in real time (sec)
- **Model**: Transformer

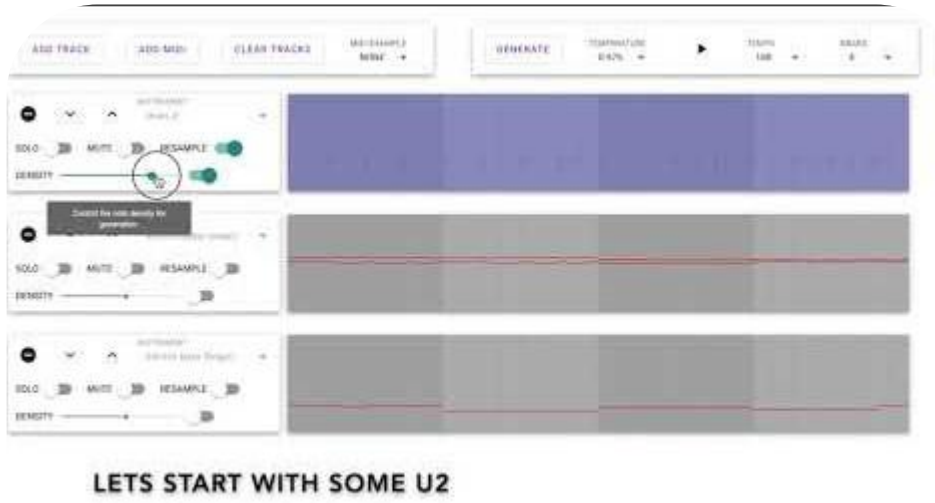
```
bach piano_strings start tempo90
piano:v72:G1 piano:v72:G2 piano:v72:B4
piano:v72:D4 violin:v80:G4 piano:v72:G4
piano:v72:B5 piano:v72:D5 wait:12
piano:v0:B5 wait:5 piano:v72:D5 wait:12
...
```

Example of generated music

Christine Payne, "MuseNet," *OpenAI*, 2019.

Multitrack Music Machine (Ens & Pasquier, 2020)

- **Data**: Lakh MIDI Dataset (LMD)
- **Representation**: as shown →
- **Model**: Transformer



youtu.be/NdeMZ3y-84Q

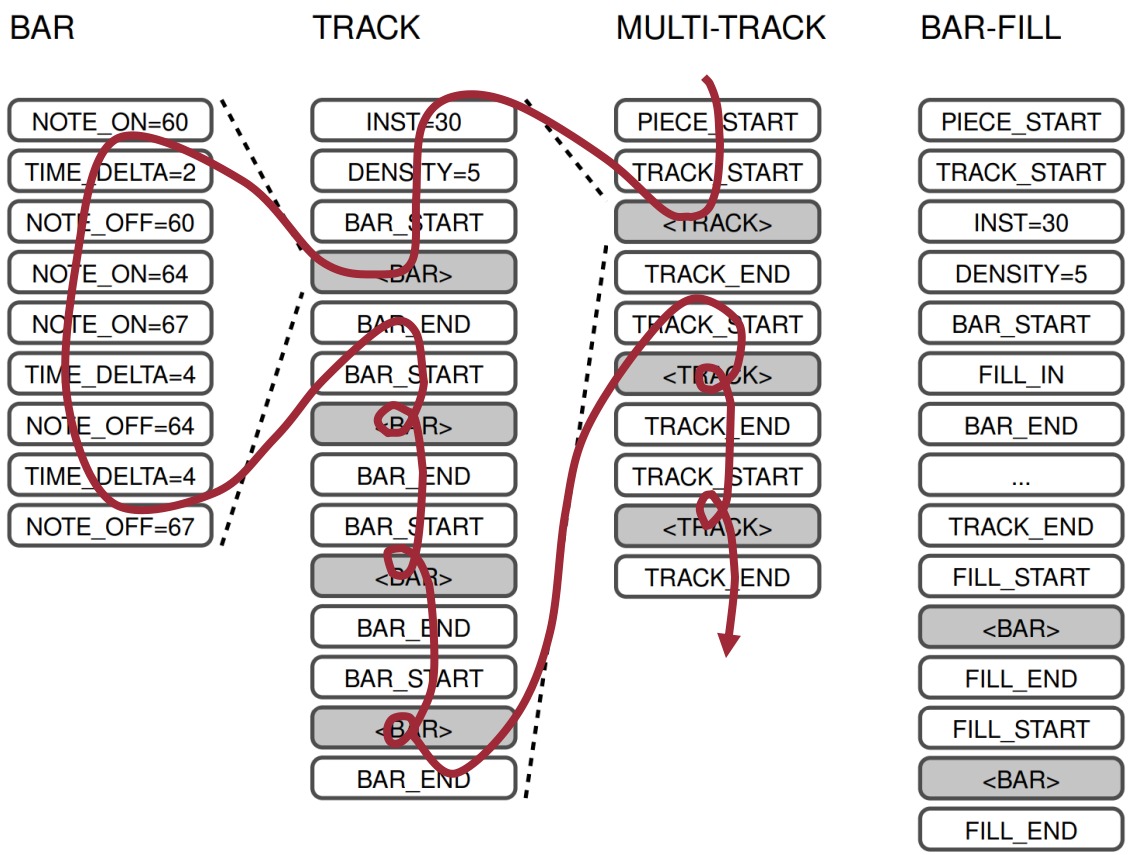


Fig. 1. The MultiTrack and BarFill representations are shown. The `<bar>` tokens correspond to complete bars, and the `<track>` tokens correspond to complete tracks.

(Ens & Pasquier, 2020)

Jeff Ens and Philippe Pasquier, "MMM : Exploring Conditional Multi-Track Music Generation with the Transformer," *arXiv preprint arXiv:2008:06048*, 2020.

Multitrack Music Transformer (Dong et al., 2023)

- **Data**: Symbolic Orchestral Database (SOD)
- **Representation**: "**(beat, position, pitch, duration, instrument)**"
- **Model**: Multi-dimensional Transformer

No time shit events!

```
(0,  0,  0,  0,  0,  0)  Start of song
(1,  0,  0,  0,  0, 15)  Instrument: accordion
(1,  0,  0,  0,  0, 36)  Instrument: trombone
(1,  0,  0,  0,  0, 39)  Instrument: brasses
(2,  0,  0,  0,  0,  0)  Start of notes
(3,  1,  1, 41, 15, 36)  Note: beat=1, position=1,  pitch=E2, duration=48, instrument=trombone
(3,  1,  1, 65,  4, 39)  Note: beat=1, position=1,  pitch=E4, duration=12, instrument=brasses
(3,  1,  1, 65, 17, 15)  Note: beat=1, position=1,  pitch=E4, duration=72, instrument=accordion
(3,  1,  1, 68,  4, 39)  Note: beat=1, position=1,  pitch=G4, duration=12, instrument=brasses
(3,  1,  1, 68, 17, 15)  Note: beat=1, position=1,  pitch=G4, duration=72, instrument=accordion
(3,  1,  1, 73, 17, 15)  Note: beat=1, position=1,  pitch=C5, duration=72, instrument=accordion
(3,  1, 13, 68,  4, 39)  Note: beat=1, position=13, pitch=G4, duration=12, instrument=brasses
(3,  1, 13, 73,  4, 39)  Note: beat=1, position=13, pitch=C5, duration=12, instrument=brasses
(3,  2,  1, 73, 12, 39)  Note: beat=2, position=1,  pitch=C5, duration=36, instrument=brasses
(3,  2,  1, 77, 12, 39)  Note: beat=2, position=1,  pitch=E5, duration=36, instrument=brasses
        ...              ...
(4,  0,  0,  0,  0,  0)  End of song
```

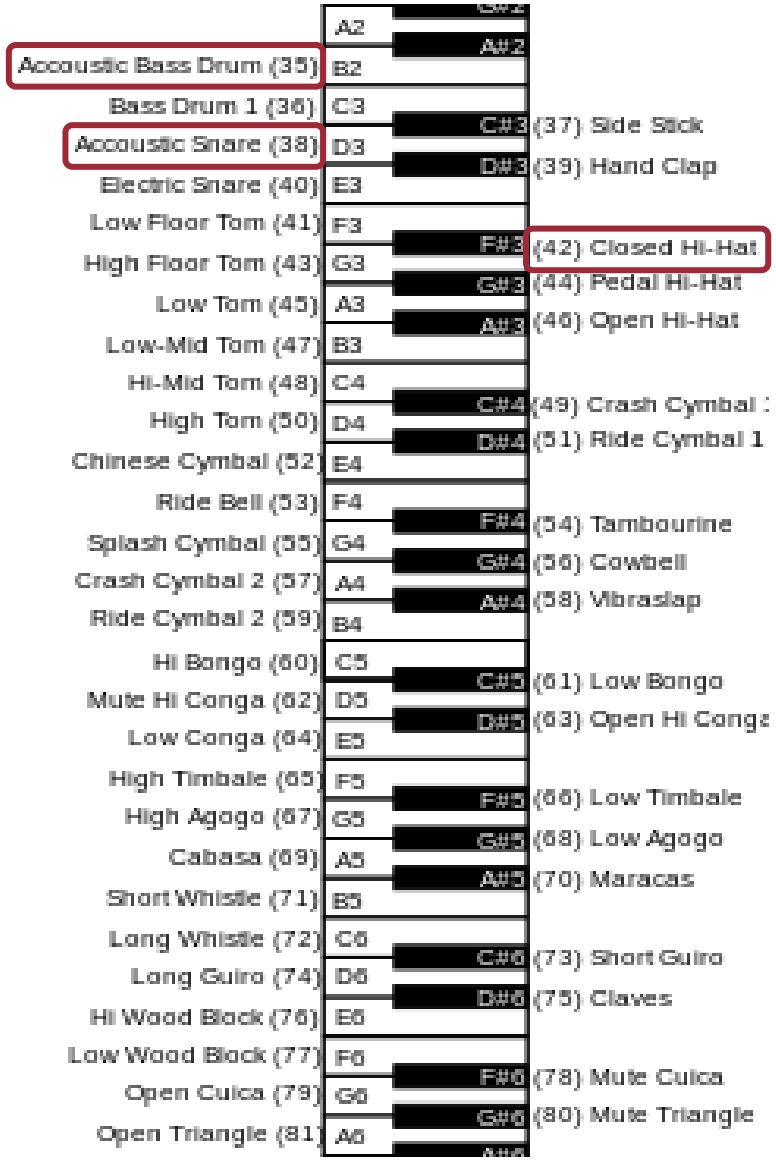
(Source: Dong et al., 2023)

Example of generated music

Hao-Wen Dong, Ke Chen, Shlomo Dubnov, Julian McAuley, and Taylor Berg-Kirkpatrick, "Multitrack Music Transformer," *ICASSP*, 2023.

Drums in MIDI

- **Channel 10** is reserved for drums
- Encoded by MIDI pitches 35–81
- Models that support drums
  - **MuseNet** (Payne et al., 2019)
  - **Song from PI** (Chu et al., 2017)
  - **MMM** (Ens and Pasquier, 2019)
  - *and many more…*



(Source: Wikipedia)

en.wikipedia.org/wiki/General_MIDI
Christine Payne, "MuseNet," *OpenAI*, 2019.
Hang Chu, Raquel Urtasun, and Sanja Fidler, "Song From PI: A Musically Plausible Network for Pop Music Generation," *ICLR Workshop*, 2017.
Jeff Ens and Philippe Pasquier, "MMM : Exploring Conditional Multi-Track Music Generation with the Transformer," *arXiv preprint arXiv:2008.06048*, 2020.

The Many Representations for Music Generation

- **PerformanceRNN** (Oore et al., 2020)
- **REMI** (Huang et al., 2020)
- **MuMIDI** (Ren et al., 2020)
- **Compound Word** (Hsiao et al., 2021)
- **REMI+** (von Rütte et al., 2023)
- **TSD** (Fradet et al., 2023)
- *and so on...*

MIDITok

github.com/Natooz/MidiTok

Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan, "This Time with Feeling: Learning Expressive Musical Performance", *Neural Computing and Applications*, 32, 2020.
Yu-Siang Huang and Yi-Hsuan Yang, "Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions," *MM*, 2020.
Yi Ren, Jinzheng He, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu, "PopMAG: Pop Music Accompaniment Generation," *MM*, 2020.
Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang, "Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs," *AAAI*, 2021.
Dimitri von Rütte, Luca Biggio, Yannic Kilcher, and Thomas Hofmann, "FIGARO: Generating Symbolic Music with Fine-Grained Artistic Control," *ICLR*, 2023.
Nathan Fradet, Nicolas Gutowski, Fabien Chhel, and Jean-Pierre Briot, "Byte Pair Encoding for Symbolic Music," *EMNLP*, 2023.

Symbolic Music Datasets

- JSBach Chorale
- MusicNet
- Essen Folk Song Dataset
- Wikifonia
- Lakh MIDI Dataset
- MetaMIDI
- Expressive MIDI: MAESTRO

Symbolic Music Datasets

Dataset	Format	Hours	Songs	Genre
Lakh MIDI Dataset	MIDI	>5000	174,533	misc
MAESTRO Dataset	MIDI	201.21	1,282	classical
Wikifonia Lead Sheet Dataset	MusicXML	198.40	6,405	misc
Essen Folk Song Dataset	ABC	56.62	9,034	folk
NES Music Database	MIDI	46.11	5,278	game
MusicNet Dataset	MIDI	30.36	323	classical
Hymnal Tune Dataset	MIDI	18.74	1,756	hymn
Hymnal Dataset	MIDI	17.50	1,723	hymn
music21's Corpus	misc	16.86	613	misc
EMOPIA Dataset	MIDI	10.98	387	pop
Nottingham Database	ABC	10.54	1,036	folk
music21's JSBach Corpus	MusicXML	3.46	410	classical
JSBach Chorale Dataset	MIDI	3.21	382	classical
Haydn Op.20 Dataset	Humdrum	1.26	24	classical

(Source: MusPy Documentation)

muspy.readthedocs.io/en/stable/datasets/datasets.html

Recap

Two Paradigms of Symbolic Music Generation

Text-based

- Treat music like **text**
- Sharing models with **natural language processing (NLP)**
  - RNNs, LSTMs, Transformers, etc.

Today's topic!

```
Program_change_0,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_76, Time_shift_2, Note_off_67,
Note_on_67, Time_shift_2, Note_off_67, ...
```

Image-based

- Treat music like **images**
- Sharing models with **computer vision (CV)** & **computer graphics (CG)**
  - GANs, VAEs, diffusion models, etc.

Music Language Models (Mathematically)

- A class of machine learning models that learn the next note probability

$$P(\ x_i \mid \underbrace{x_1, x_2, \ldots, x_{i-1}}\ )$$

Next note — Previous notes

$P(\ \mathrm{G} \mid C\ C\ G\ G\ A\ A\ )$ ↑

$P(\ \mathrm{A} \mid C\ C\ G\ G\ A\ A\ )$ ↑

$P(\ \mathrm{C} \mid C\ C\ G\ G\ A\ A\ )$ ↑

$P(\ \mathrm{F} \mid C\ C\ G\ G\ A\ A\ )$ ↑

$P(\ \mathrm{A}\flat \mid C\ C\ G\ G\ A\ A\ )$ ↓

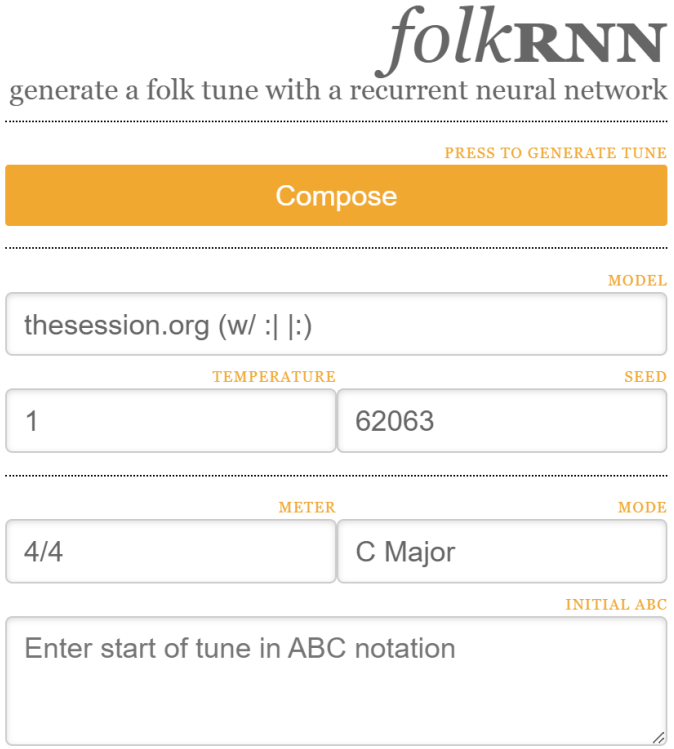
$P(\ \mathrm{A}\sharp \mid C\ C\ G\ G\ A\ A\ )$ ↓

Designing a Machine-readable Music Language

- How can we “represent” music in a way that machines understand?

Folk RNN (Sturm et al., 2015)

- **Data**
  - 23,958 folk tunes
- **Representation**
  - ABC notation without metadata
- **Model**
  - LSTM (long short-term memory)
  - Working on the character level

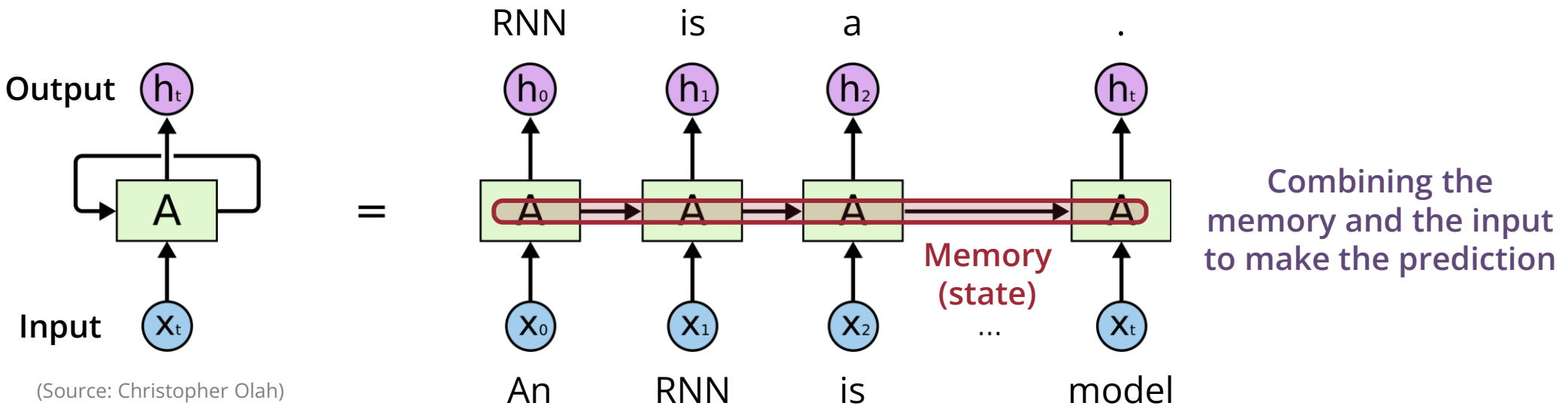


folkrnn.org

Bob L. Sturm, Joao Felipe Santos, and Iryna Korshunova, "Folk Music Style Modelling by Recurrent Neural Networks with Long Short Term Memory Units," *ISMIR Late-Breaking Demos*, 2015.

What is an RNN (Recurrent Neural Network)?

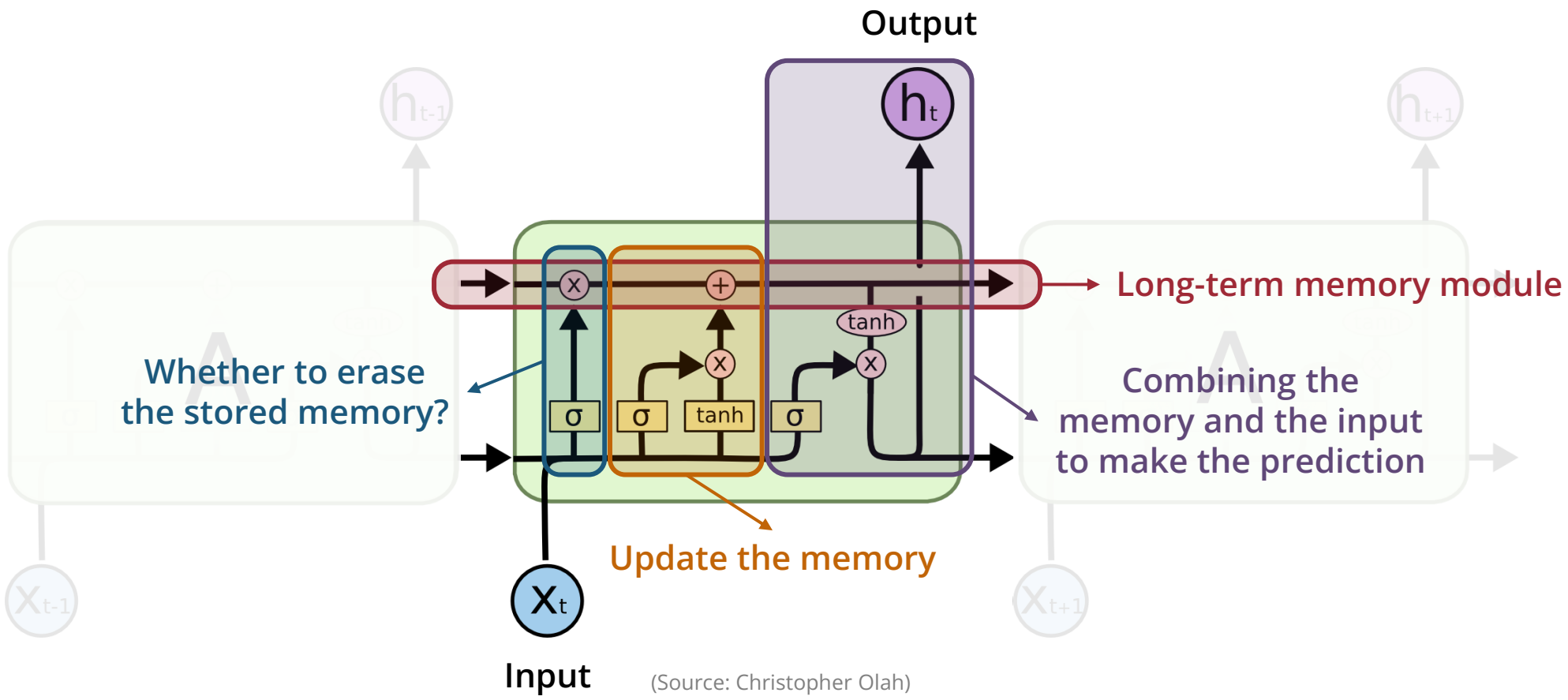
- A type of neural networks that have **loops**
- Widely used for **modeling sequences** (e.g., in natural language processing)



(Source: Christopher Olah)

colah.github.io/posts/2015-08-Understanding-LSTMs/

Demystifying LSTMs (Hochreiter & Schmidhuber, 1997)

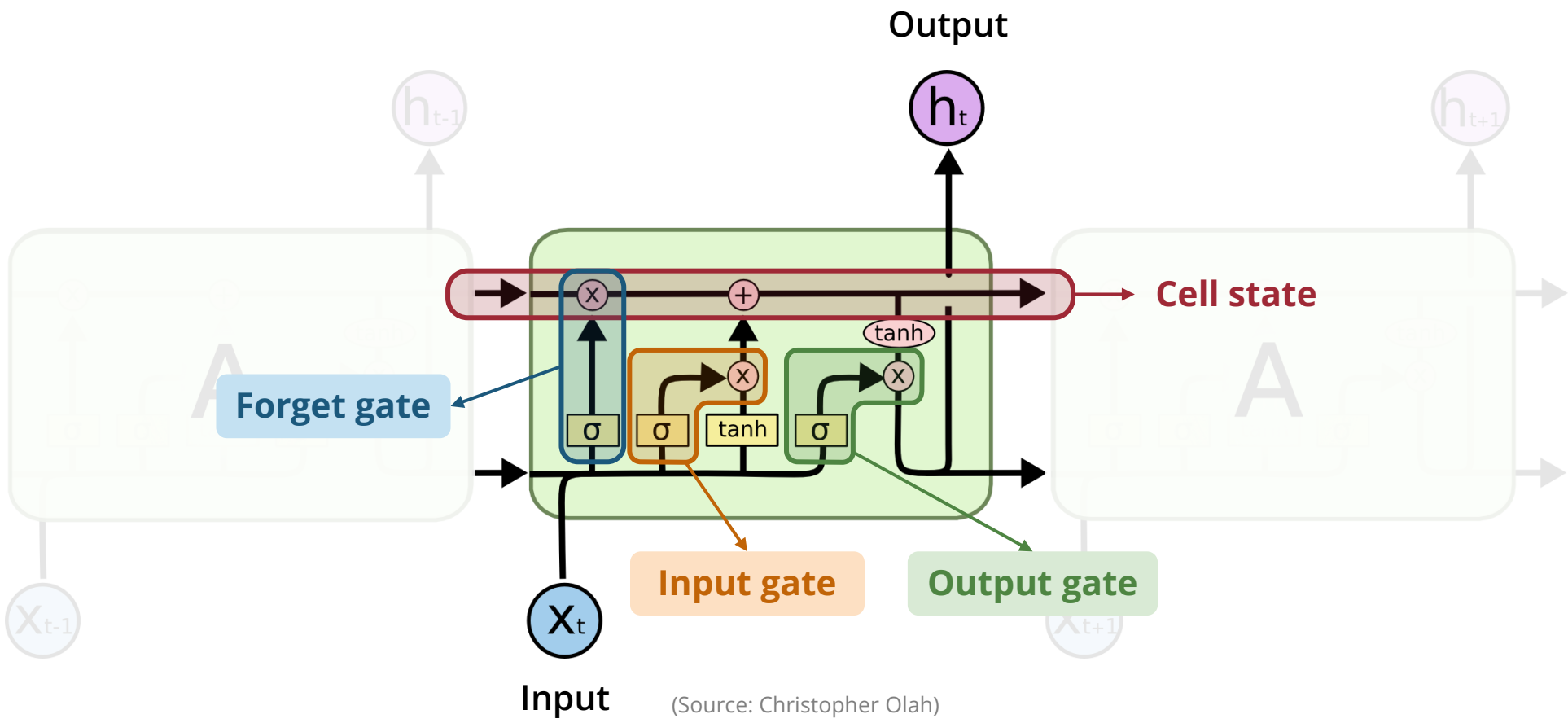


(Source: Christopher Olah)

colah.github.io/posts/2015-08-Understanding-LSTMs/

Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 9(8):1735-1780, 1997.

Demystifying LSTMs (Hochreiter & Schmidhuber, 1997)



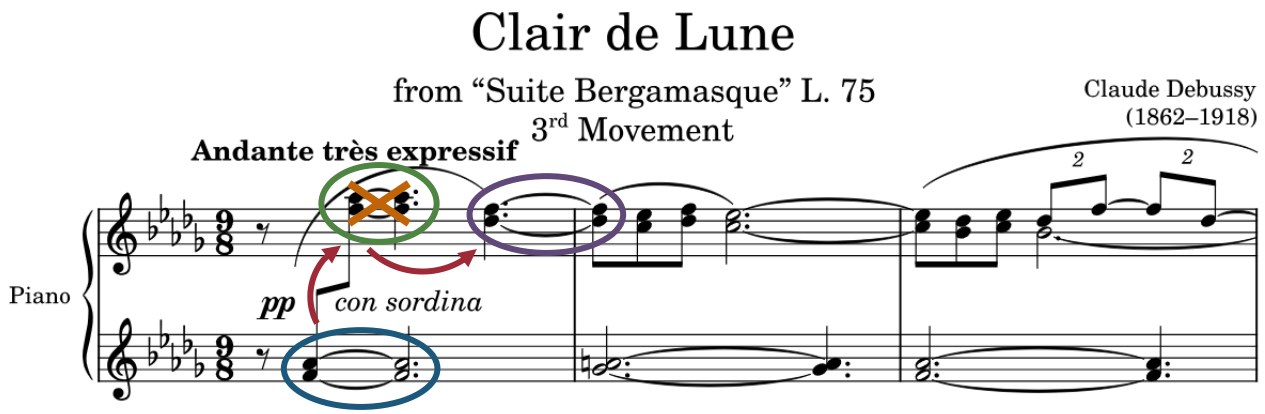
(Source: Christopher Olah)

colah.github.io/posts/2015-08-Understanding-LSTMs/

Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 9(8):1735-1780, 1997.

Representing Polyphonic Music

- We can now handle music with multi-pitch at the same time
  - In the literature, "polyphonic" & "multi-pitch" are often used interchangeably




```
Note_on_65, Note_on_68, Time_shift_eighth_note, Note_on_77, Note_on_80,
Time_shift_half_note, Note_off_77, Note_off_80, Note_on_73, Note_on_77,
Time_shift_dotted_quarter_note, Note_off_65, Note_off_68, ...
```

Performance RNN (Oore et al., 2020)

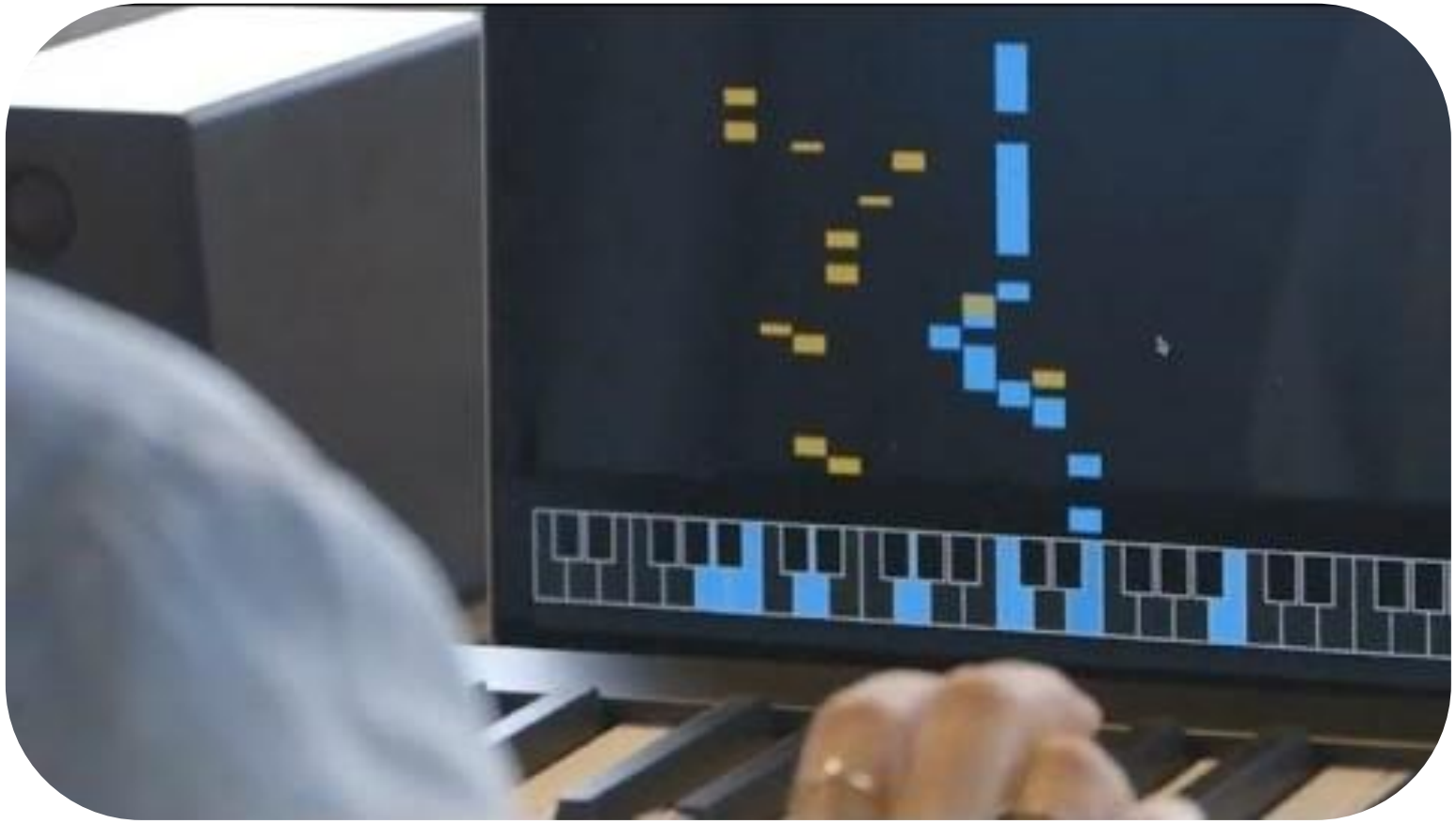
- **Data**
  - Yamaha e-Piano Competition dataset (MAESTRO)
- **Representation**
  - 128 Note-On events
  - 128 Note-Off events
  - 125 Time-Shift events (8ms–1s)
  - 32 Set-Velocity events — Handle dynamics
- **Model**
  - LSTM

Examples of generated music



Ian Simon and Sageev Oore, "Performance RNN: Generating Music with Expressive Timing and Dynamics," *Magenta Blog*, June 29, 2017.
Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan, "This Time with Feeling: Learning Expressive Musical Performance", *Neural Computing and Applications*, 32, 2020.

A.I. Duet (Mann et al, 2016)



youtu.be/0ZE1bfPtvZo
experiments.withgoogle.com/ai/ai-duet/view

github.com/googlecreativelab/aiexperiments-ai-duet

Music Transformer (Huang et al., 2019)

- **Data**: Yamaha e-Piano Competition dataset (MAESTRO)
- **Representation**
  - 128 Note-On events
  - 128 Note-Off events
  - 100 Time-Shift events (10ms–1s)
  - 32 Set-Velocity events
- **Model**: Transformer

Almost the same representation as PerformanceRNN

Handle dynamics

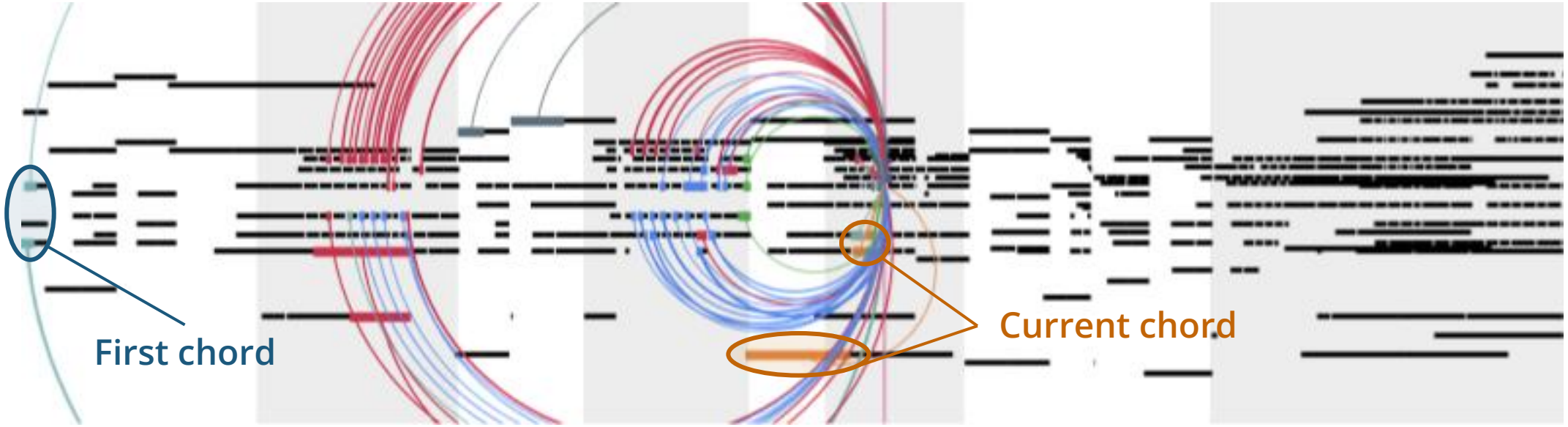
Examples of generated music



Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *ICLR*, 2019.

Visualizing Musical Self-attention

(Each color represents an attention head)

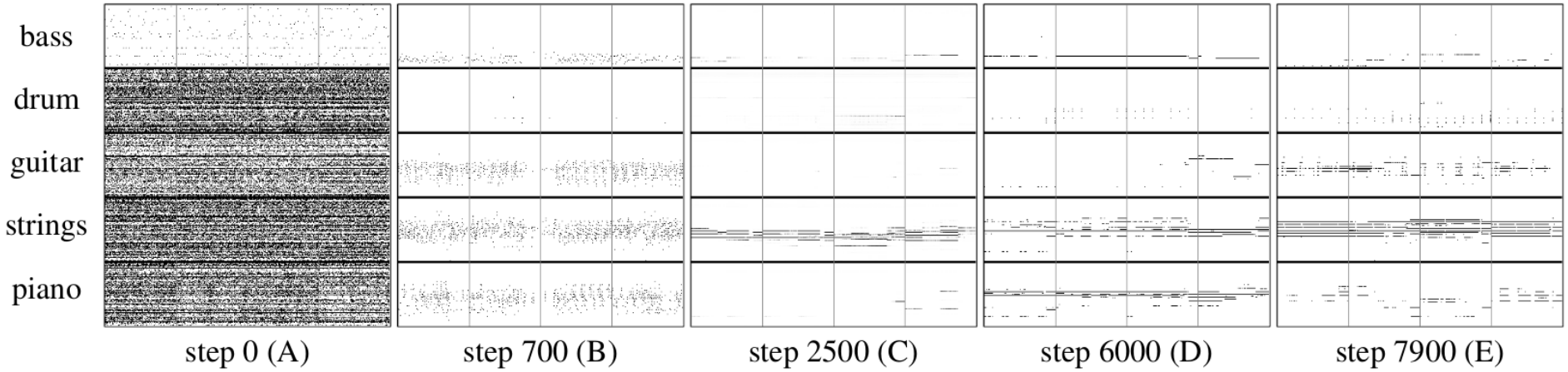


(Source: Huang et al., 2018)

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *ICLR*, 2019.
Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music Transformer: Generating Music with Long-Term Structure," *Magenta Blog*, December 13, 2018.

Next Lecture

Piano Roll-based Music Generation



(Source: Dong et al., 2018)