Statistical Parametric Speech Synthesis - from regression trees to DNNs

- Class slides

What we'll cover today

- **Quick recap**
- Discussion points and exercises on DNN-based TTS
- Lab report, experiments, and write-up
  - marking sheet with Q&A

What is a simple feedforward neural network?

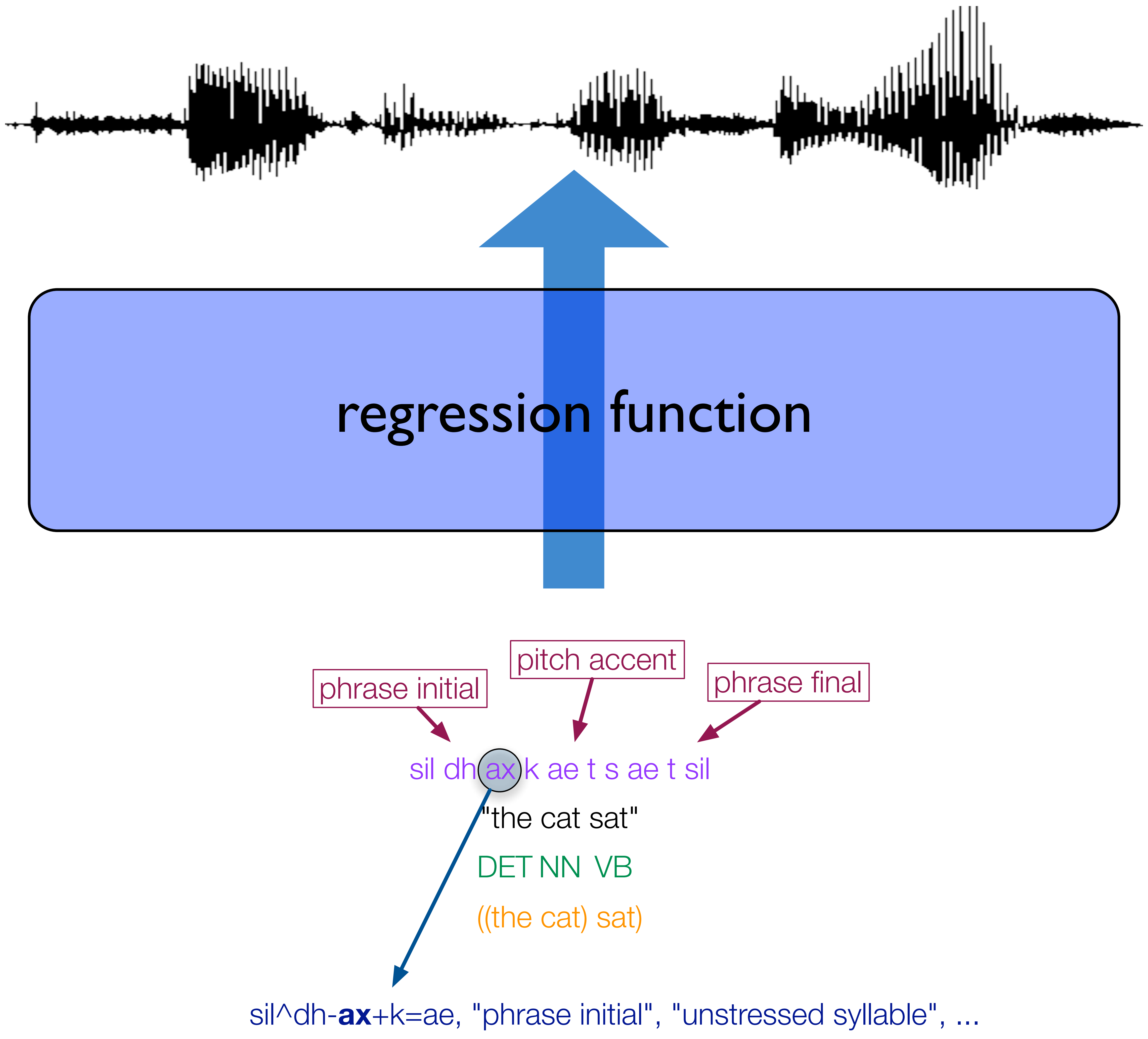
- input/output representations

- the anatomy of a unit (or more rarely now "neuron")
  - incoming weights, activation, activation function, output

- combining multiple units into a layer
- stacking layers to make a network

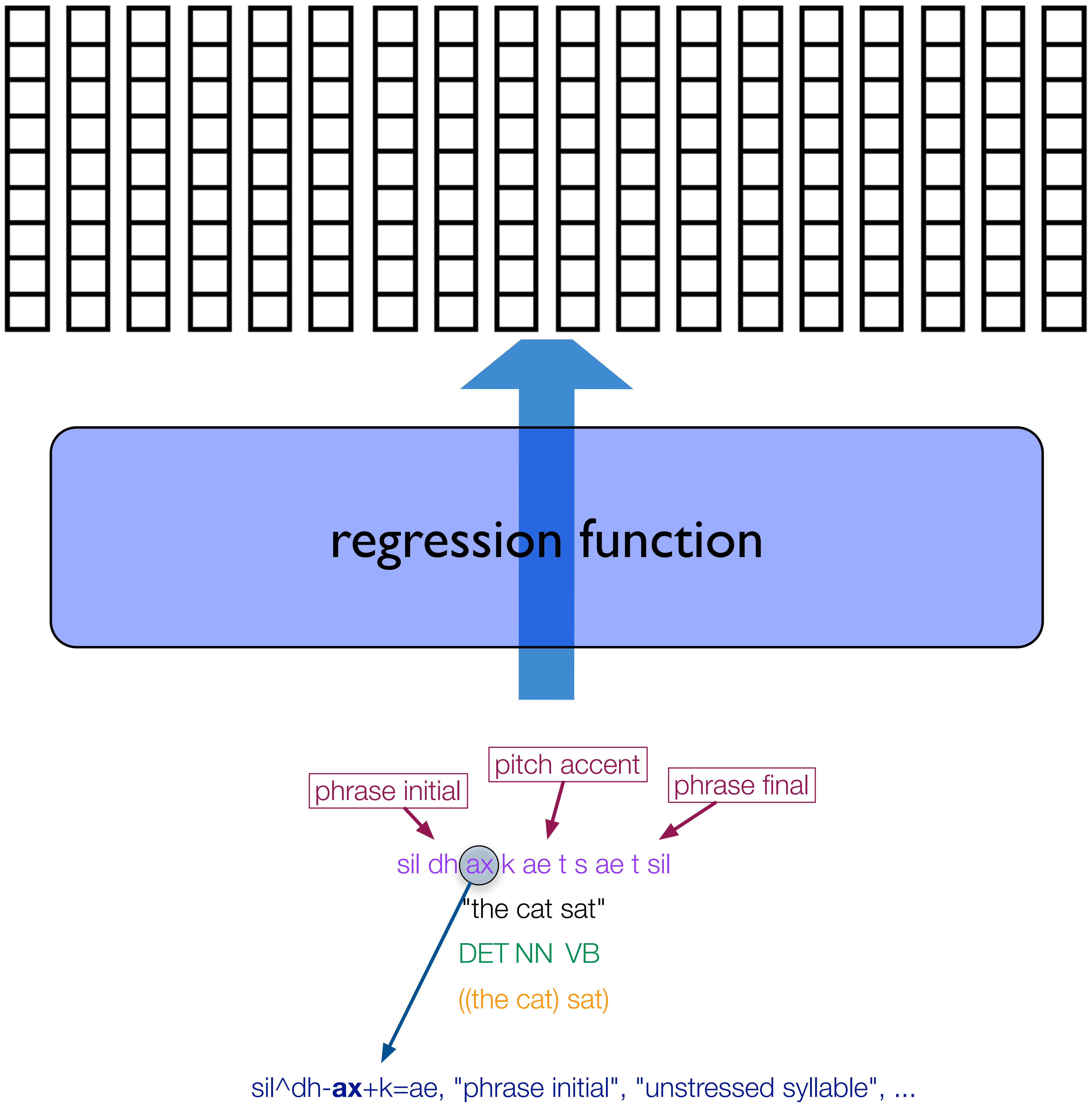
- "Information flow"

Orientation



- Statistical parametric synthesis
  - predict **speech parameters** from **linguistic specification**

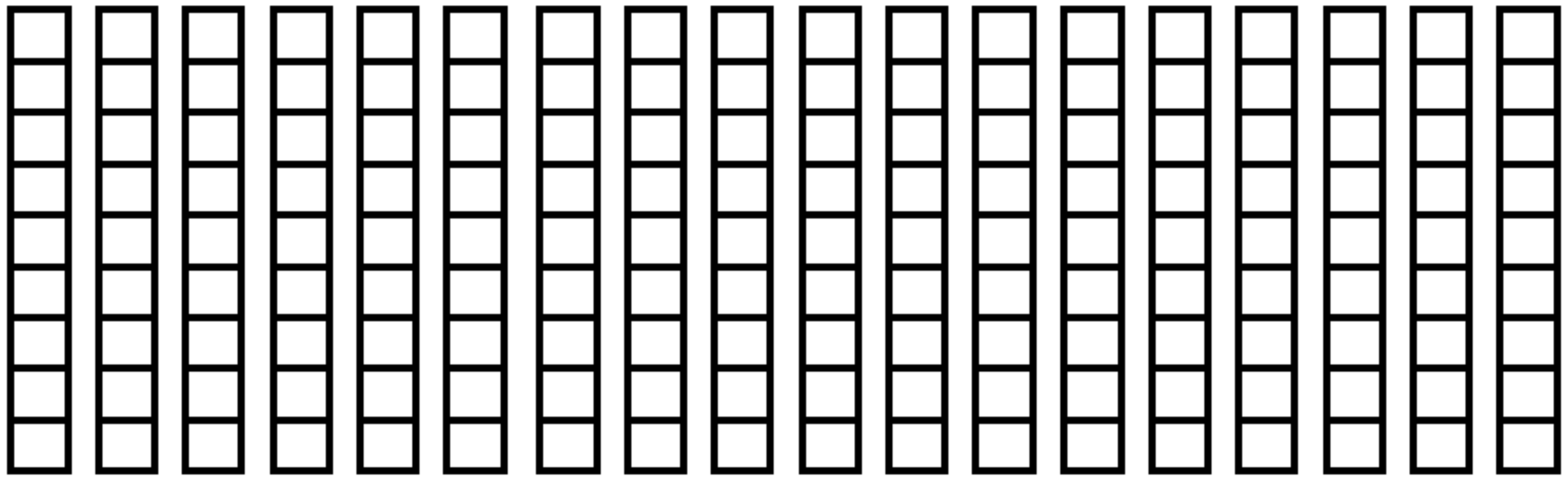
Orientation



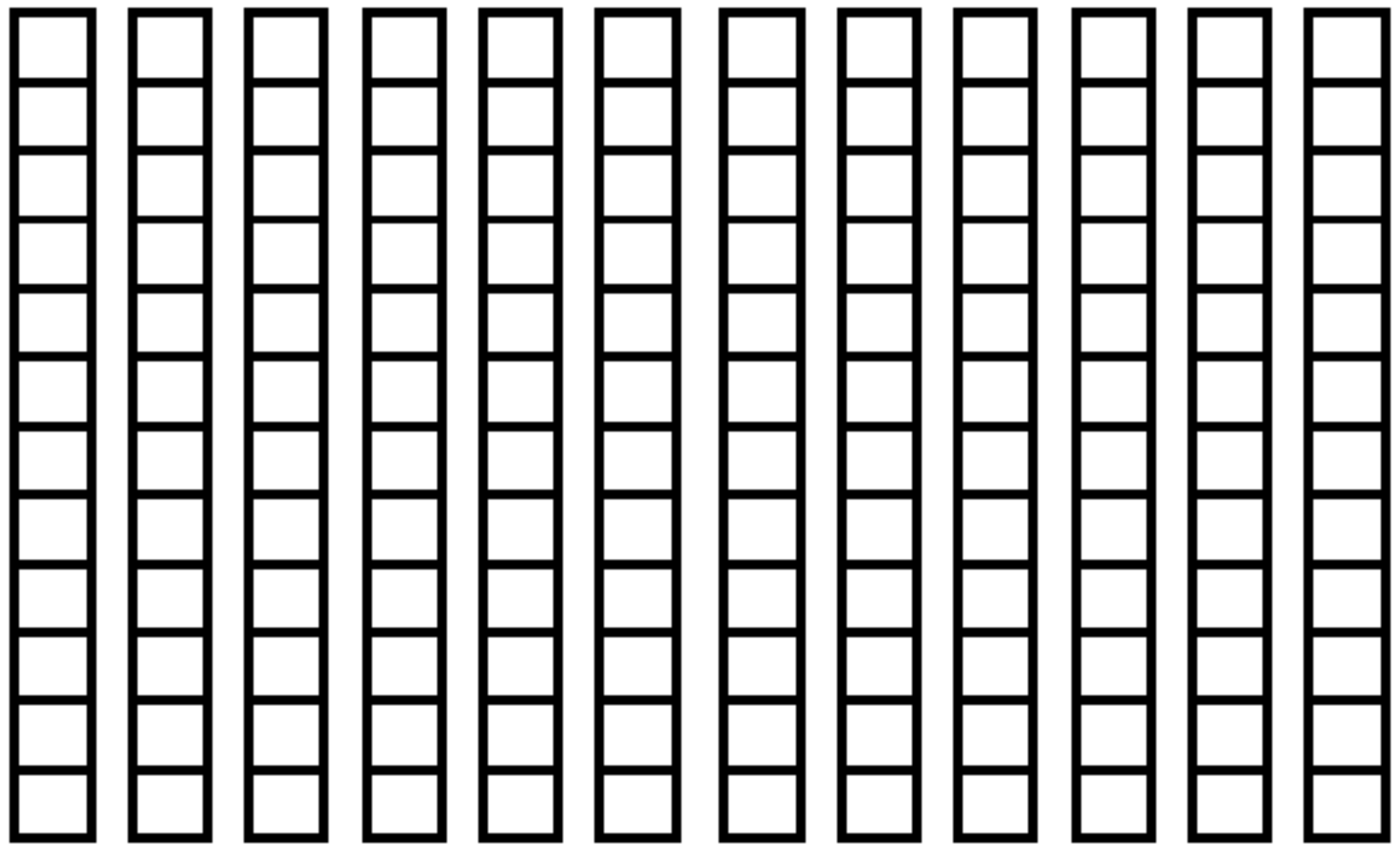
- Statistical parametric synthesis
  - predict **speech parameters** from **linguistic specification**

Solve text-to-speech as **sequence-to-sequence** regression using DNNs

output sequence



input sequence



Represent this input text as a sequence of one-hot vectors

“Please call . . . .”

Represent this input text as a sequence of one-hot vectors

`“Please call . . .”`

What is the very first step?

Represent this input text as a sequence of one-hot vectors

“Please call . . . .”

p l I z k O l . . .

We have now run the front-end
What next?

Represent this input text as a sequence of one-hot vectors

“Please call . . .”

p l I z k O l . . .

We want to write this as a sequence of vectors, ready for input to the DNN

~~Option 1 - write out context-dependent phones first~~

Option 2 - encode as 1-hot first

(don’t worry about duration yet)

Option 2: convert each phone to 1-hot, then deal with context afterwards

“Please call . . . .”

p l I z k O l . . .

Now deal with duration!

“Please call . . . .”

p l I z k O l . . .

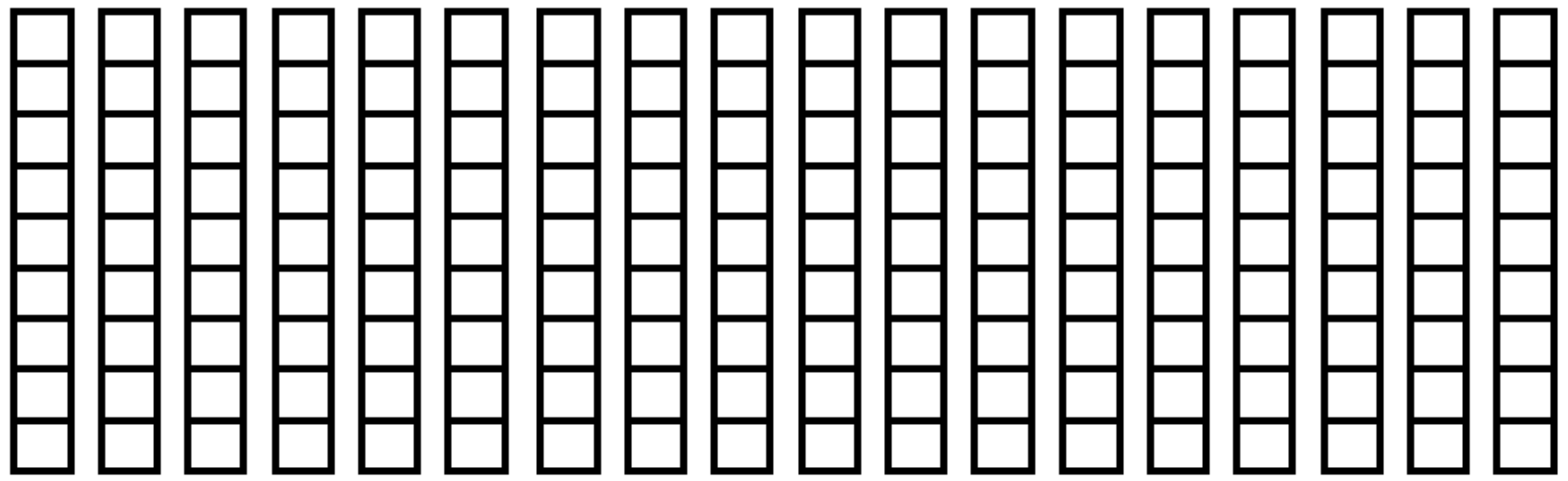
Represent this input text as a sequence of one-hot vectors

`“Please call . . . .”`

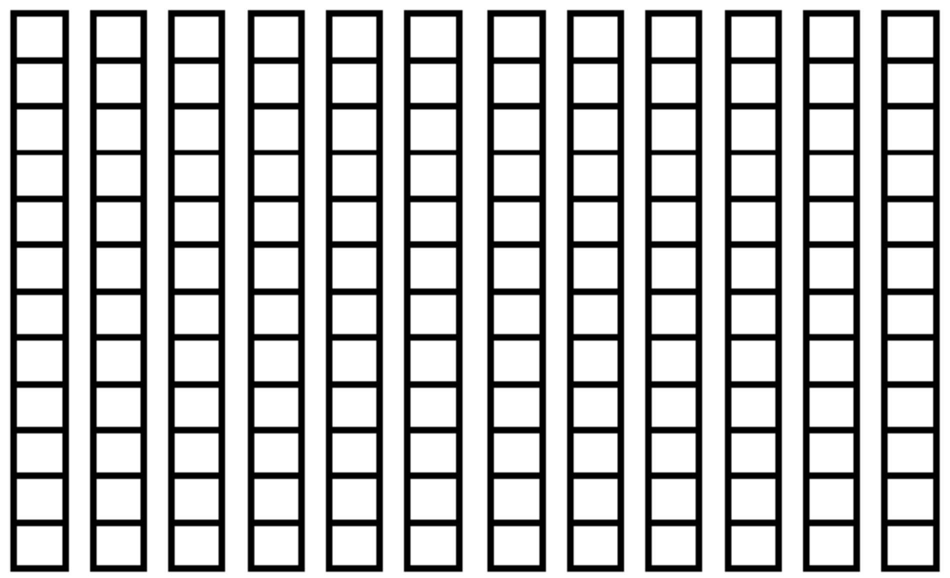
- Run the text through the front end
  - obtain the linguistic specification
    - predict pronunciation: phone sequence + some structural information
    - *omit prosody, for the purposes of this exercise*
  - “flatten” the linguistic specification
    - which means to attach all the necessary context to each phone in the sequence
  - represent each item in that sequence as a vector
- Deal with duration

Finally, draw a diagram of **sequence-to-sequence** regression using a DNN

output sequence



input sequence



Homework: can you understand this figure from the *Zen et al* reading?

Hint: it's very similar to something we drew in class this week...

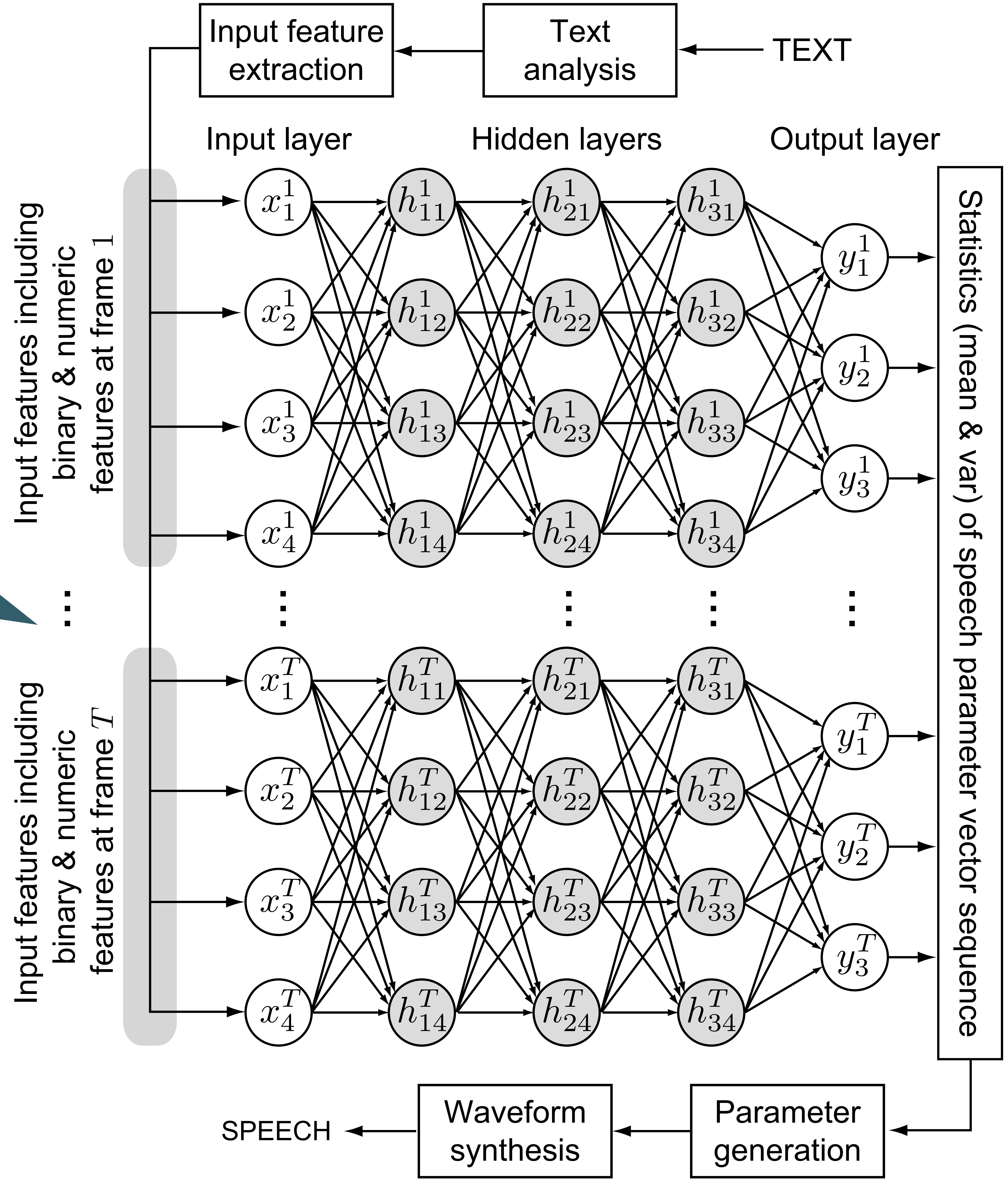
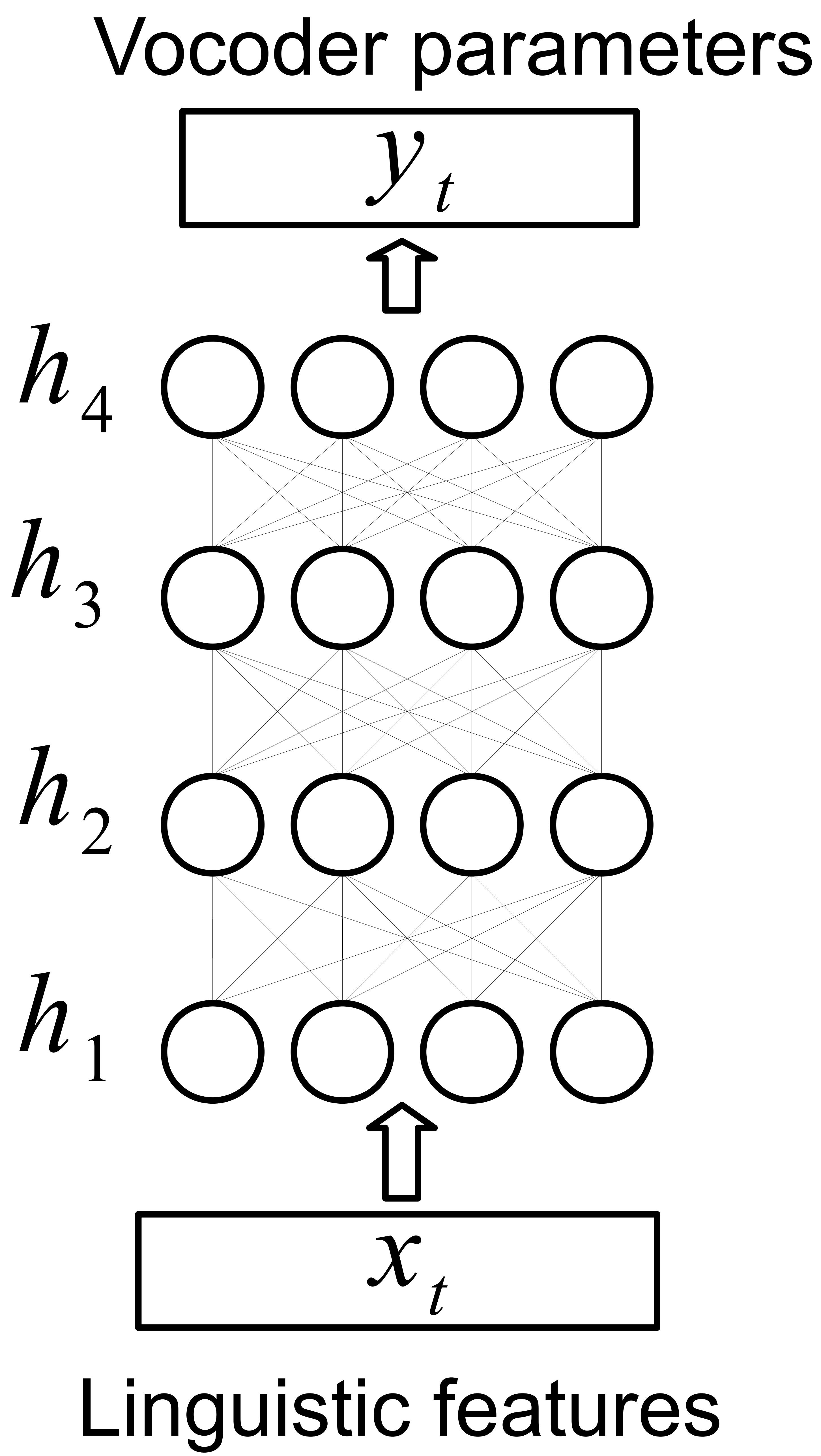


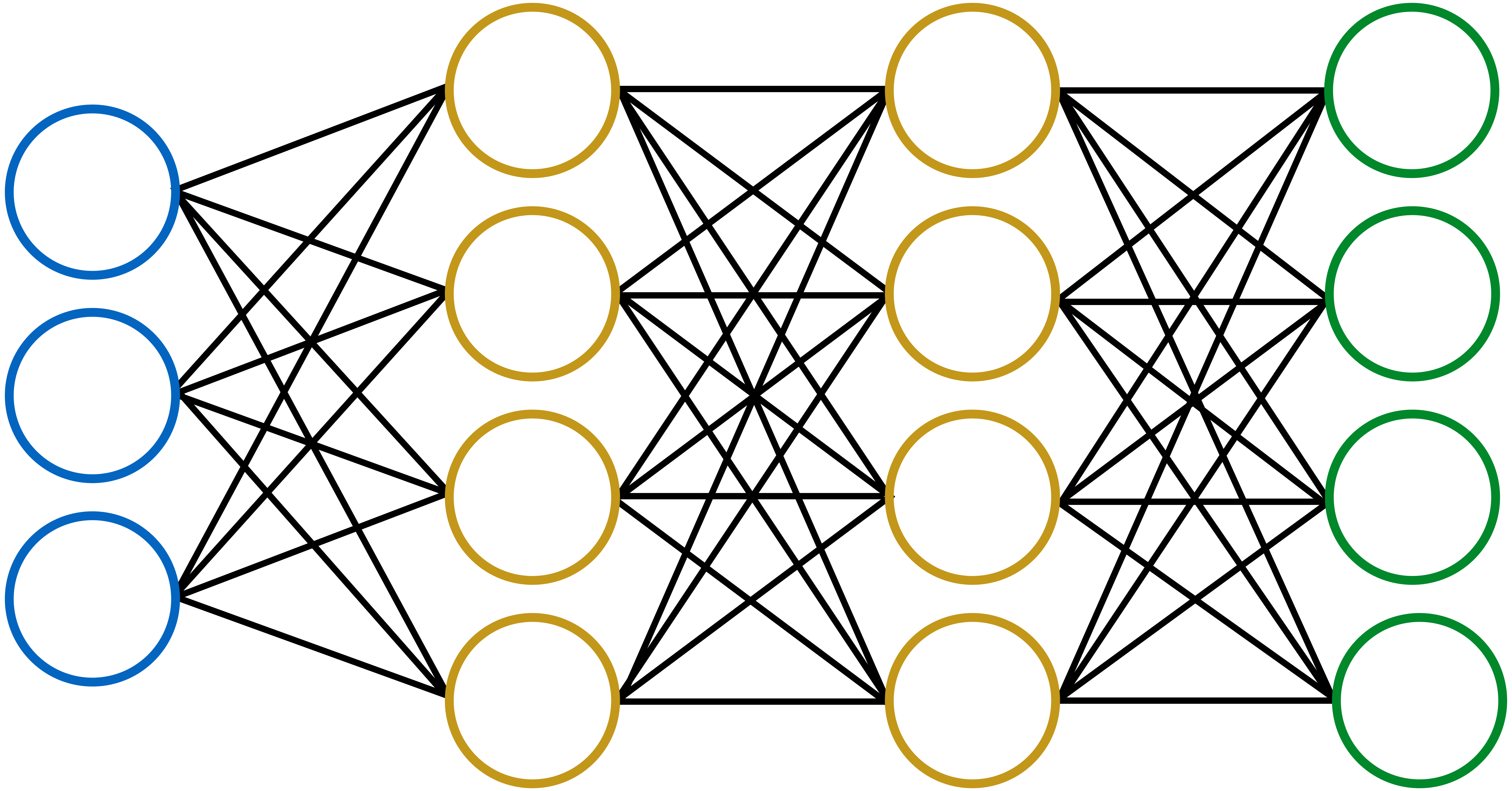
Fig. 1. A speech synthesis framework based on a DNN.

Summary

Doing TTS with a DNN

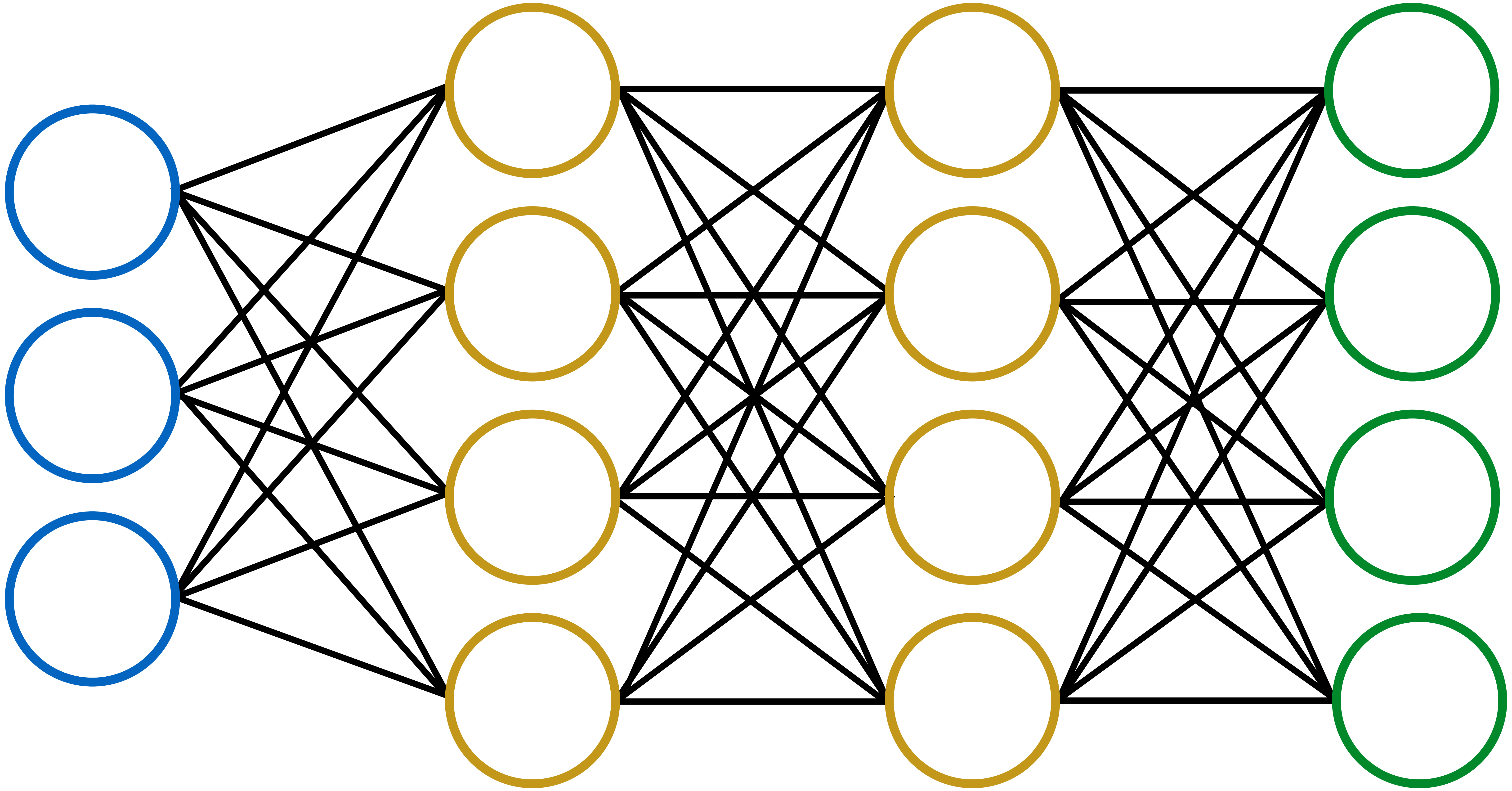


Doing regression by performing a forward pass through the DNN



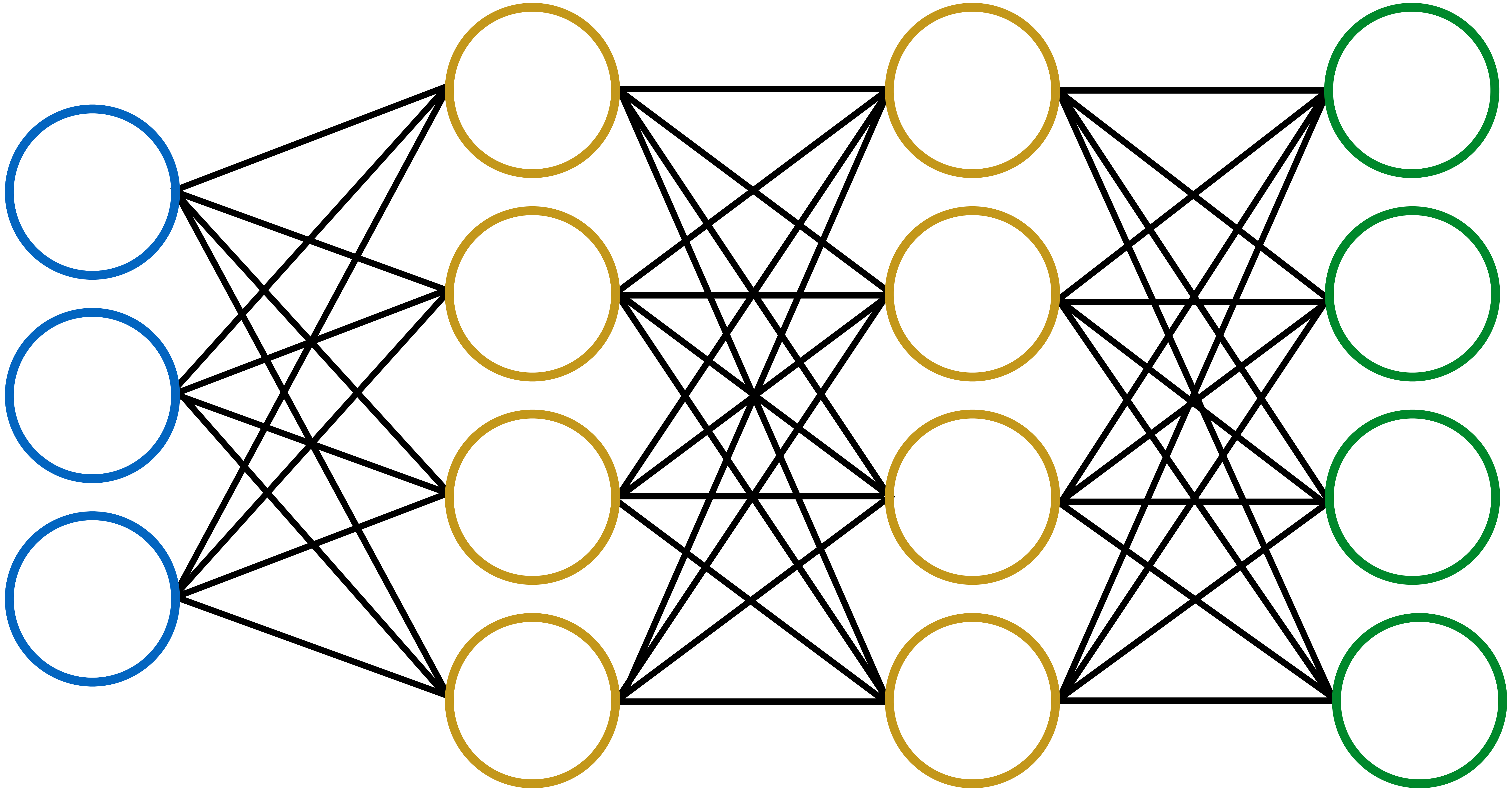
Doing regression by performing a forward pass through the DNN

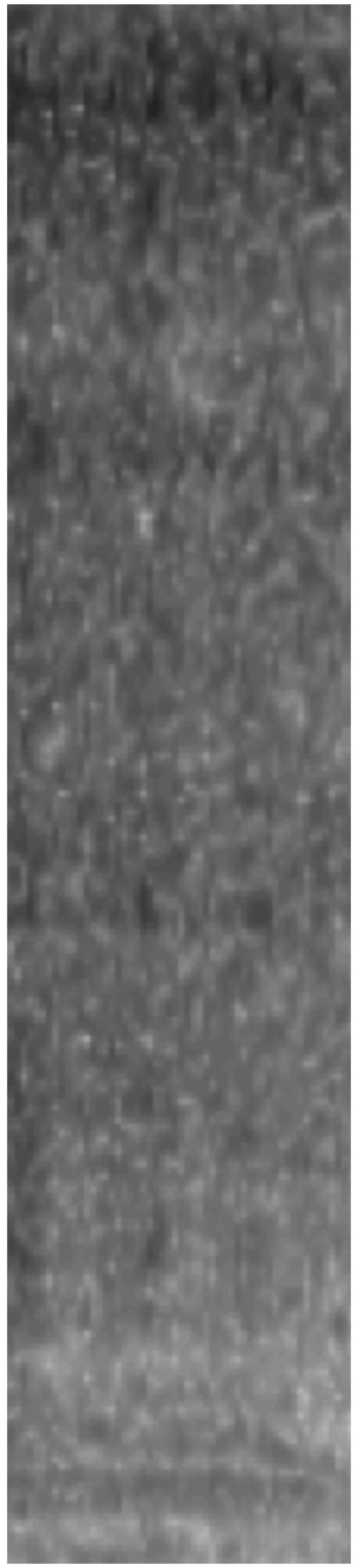
θ



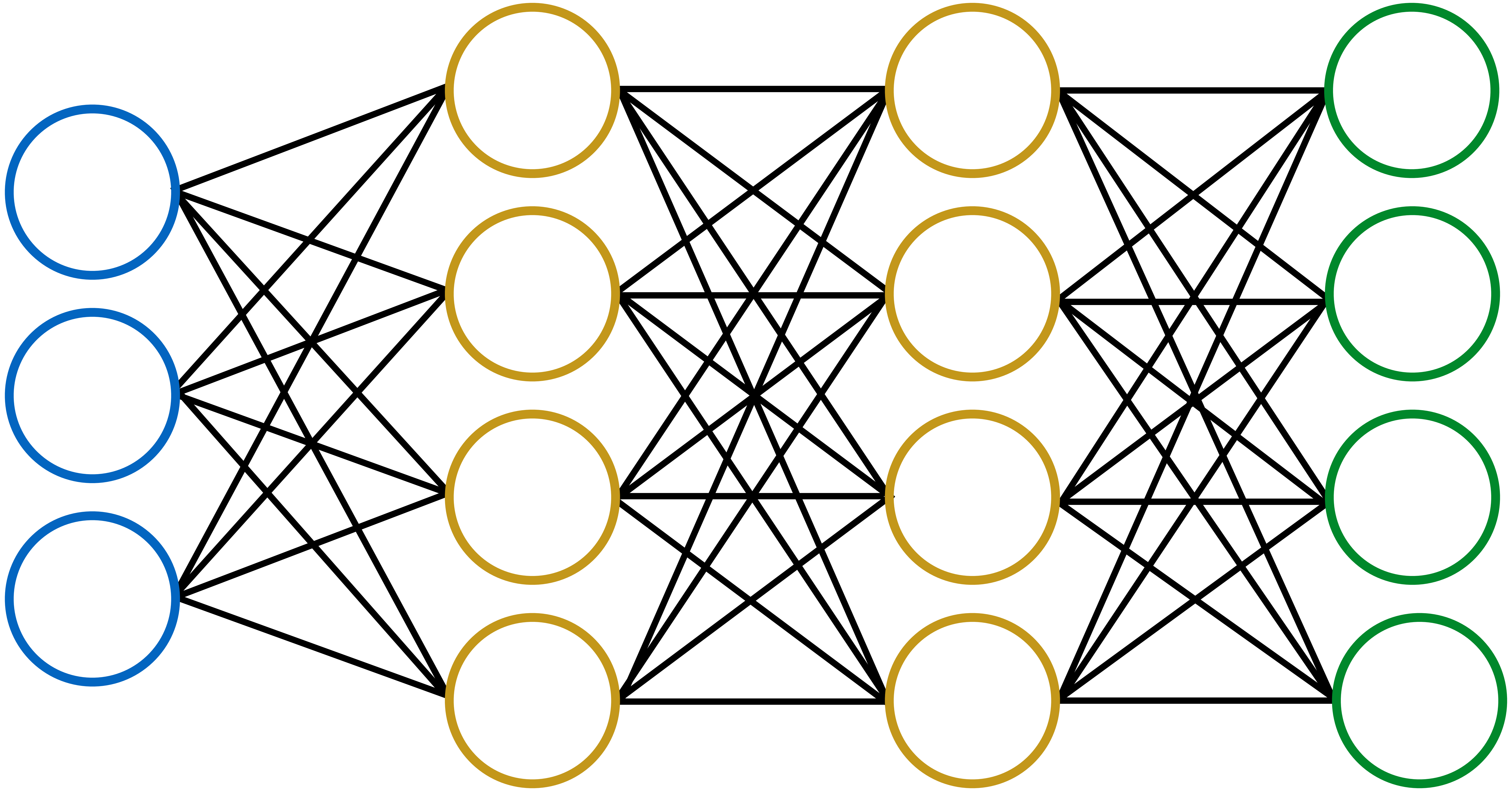
Doing regression by performing a forward pass through the DNN

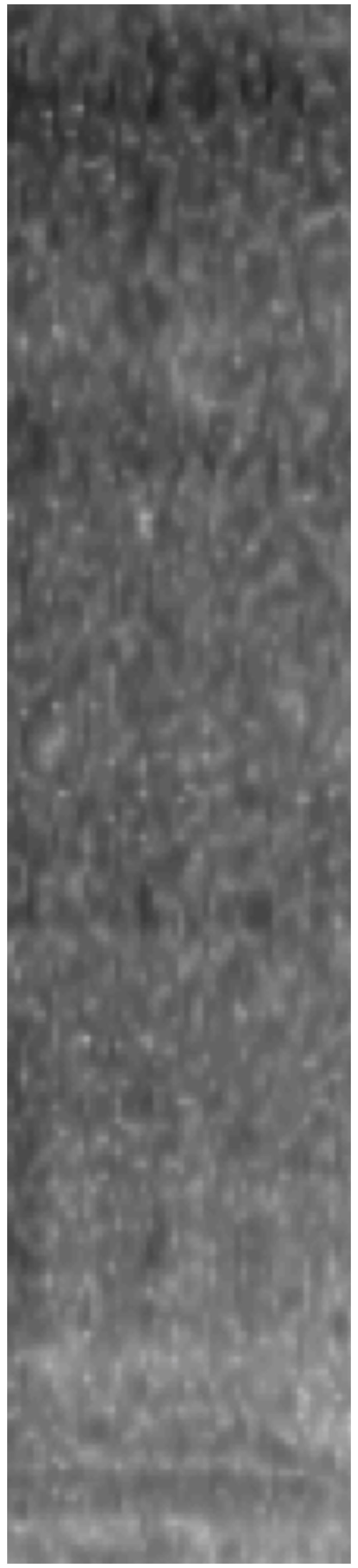
θ



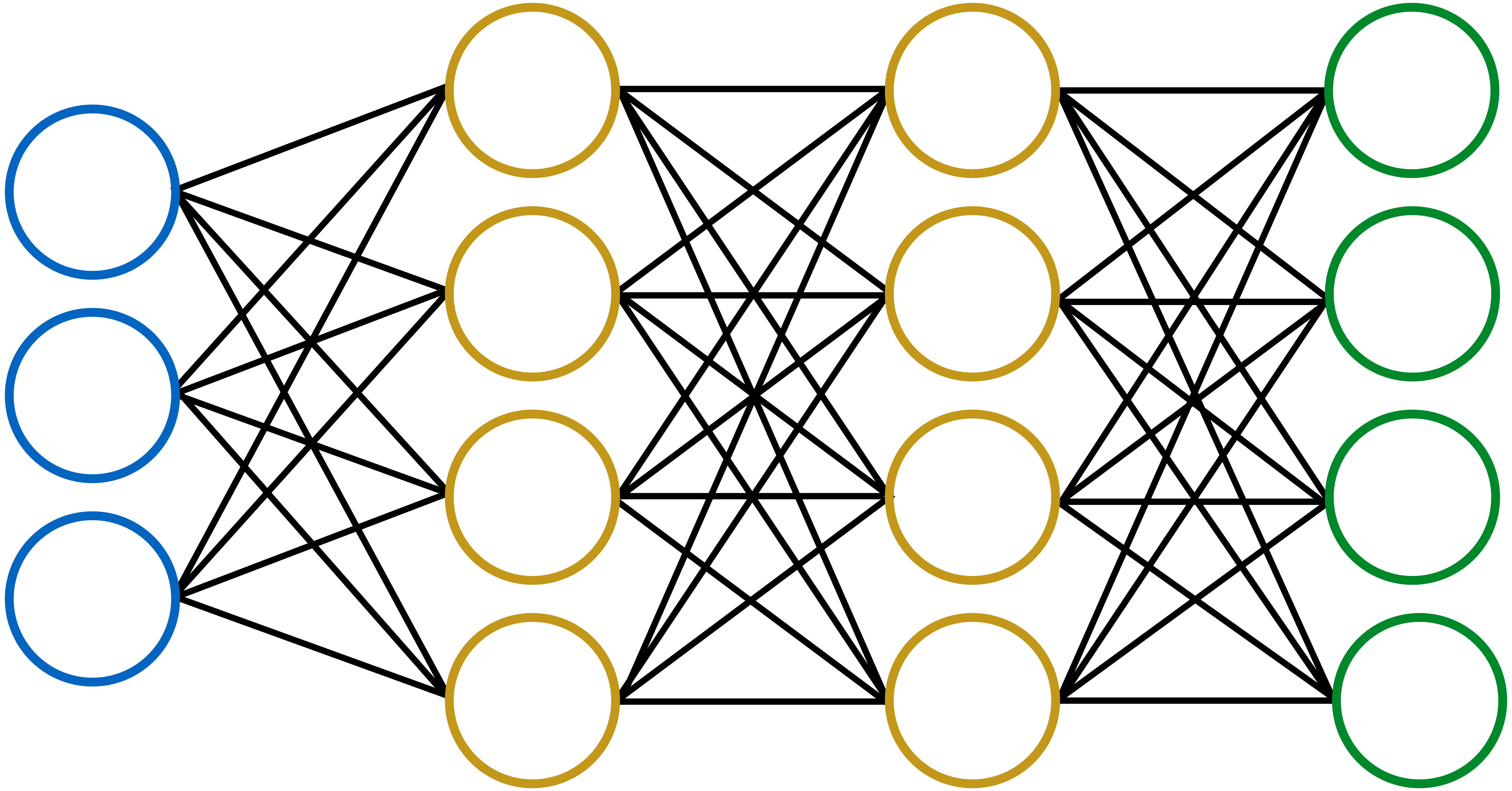


Doing regression by performing a forward pass through the DNN



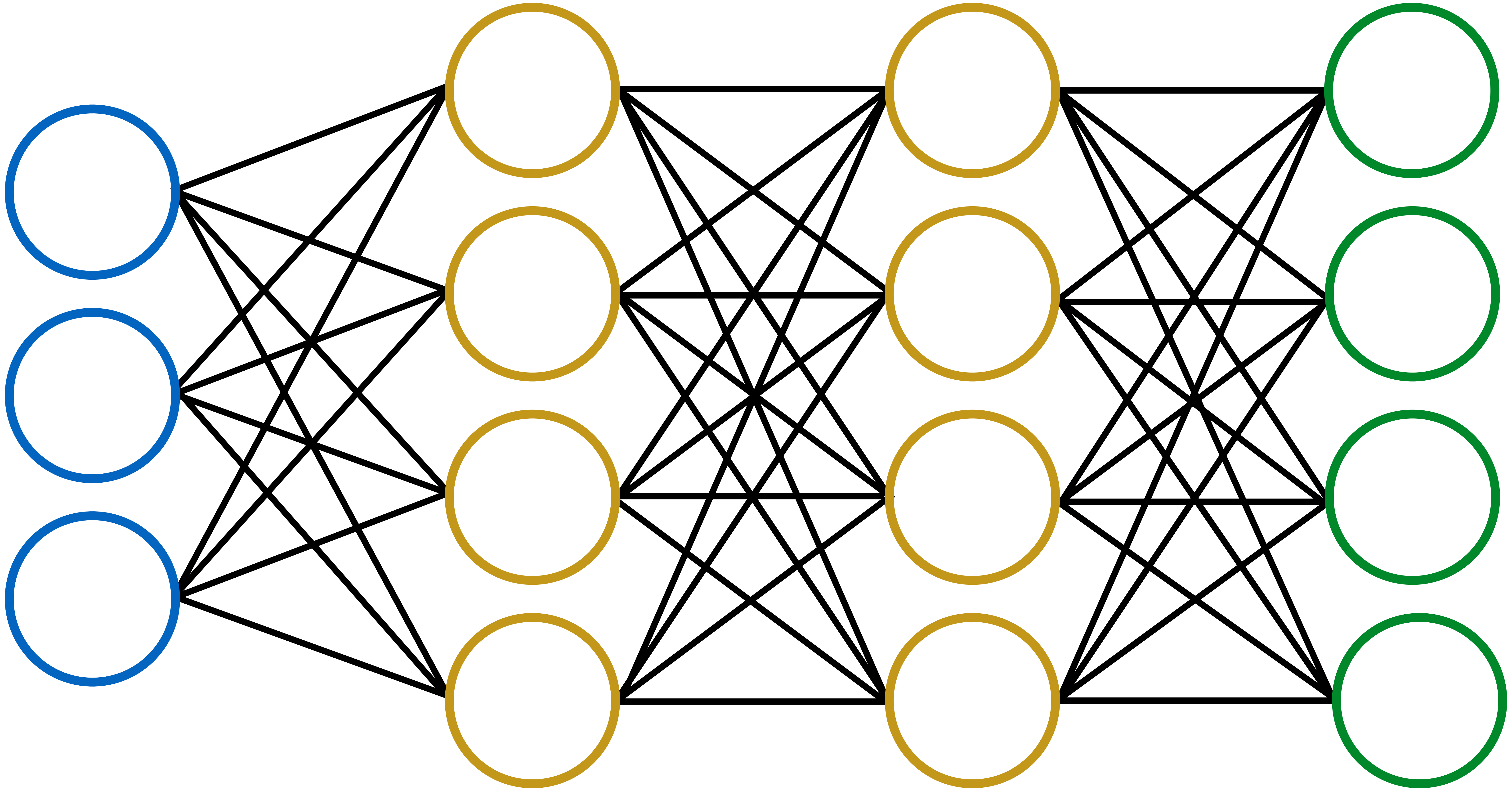


Doing regression by performing a forward pass through the DNN



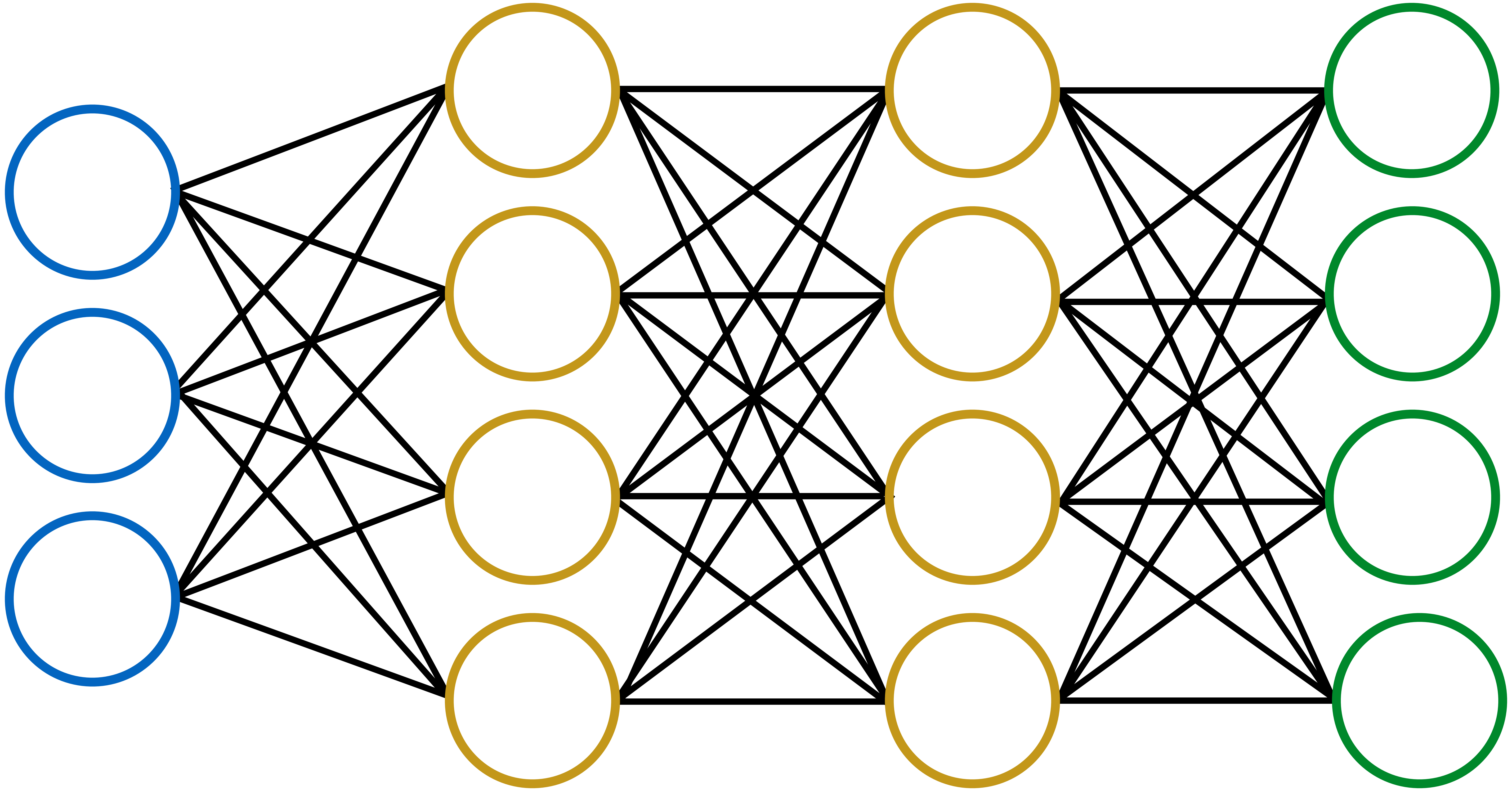
Doing regression by performing a forward pass through the DNN

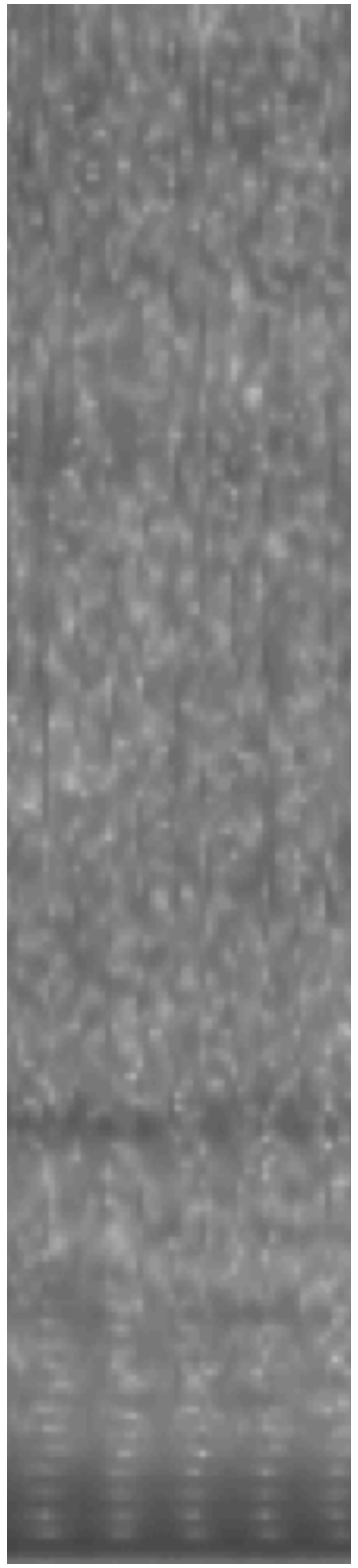
ð



Doing regression by performing a forward pass through the DNN

ð



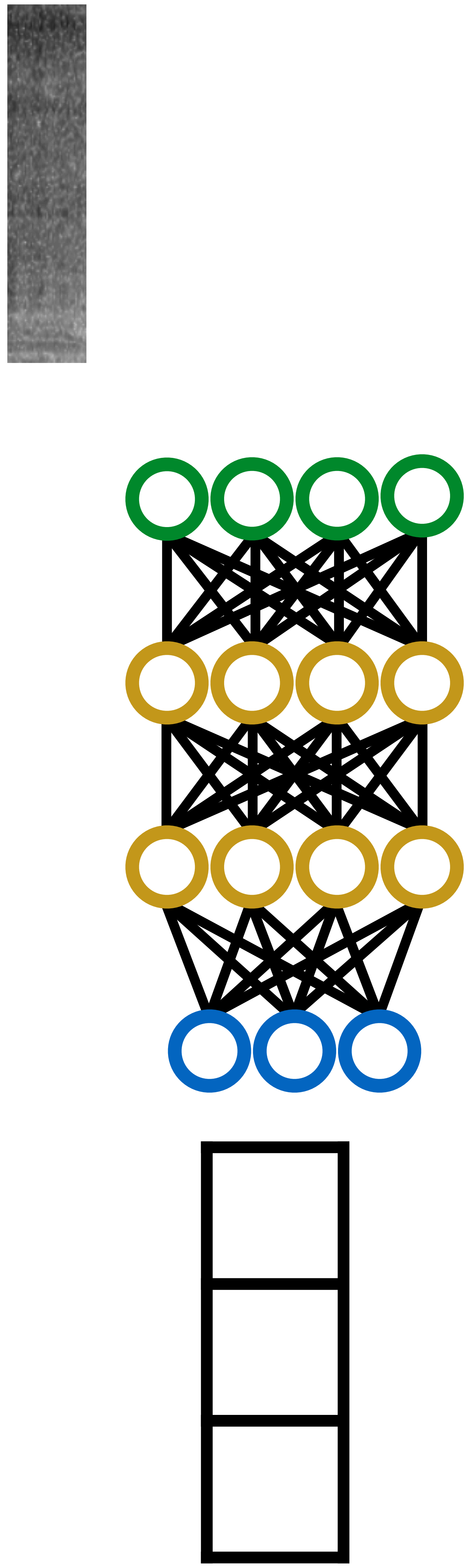


Terminology

- regression
- inference
- forward pass

Sequence-to-sequence regression using a DNN

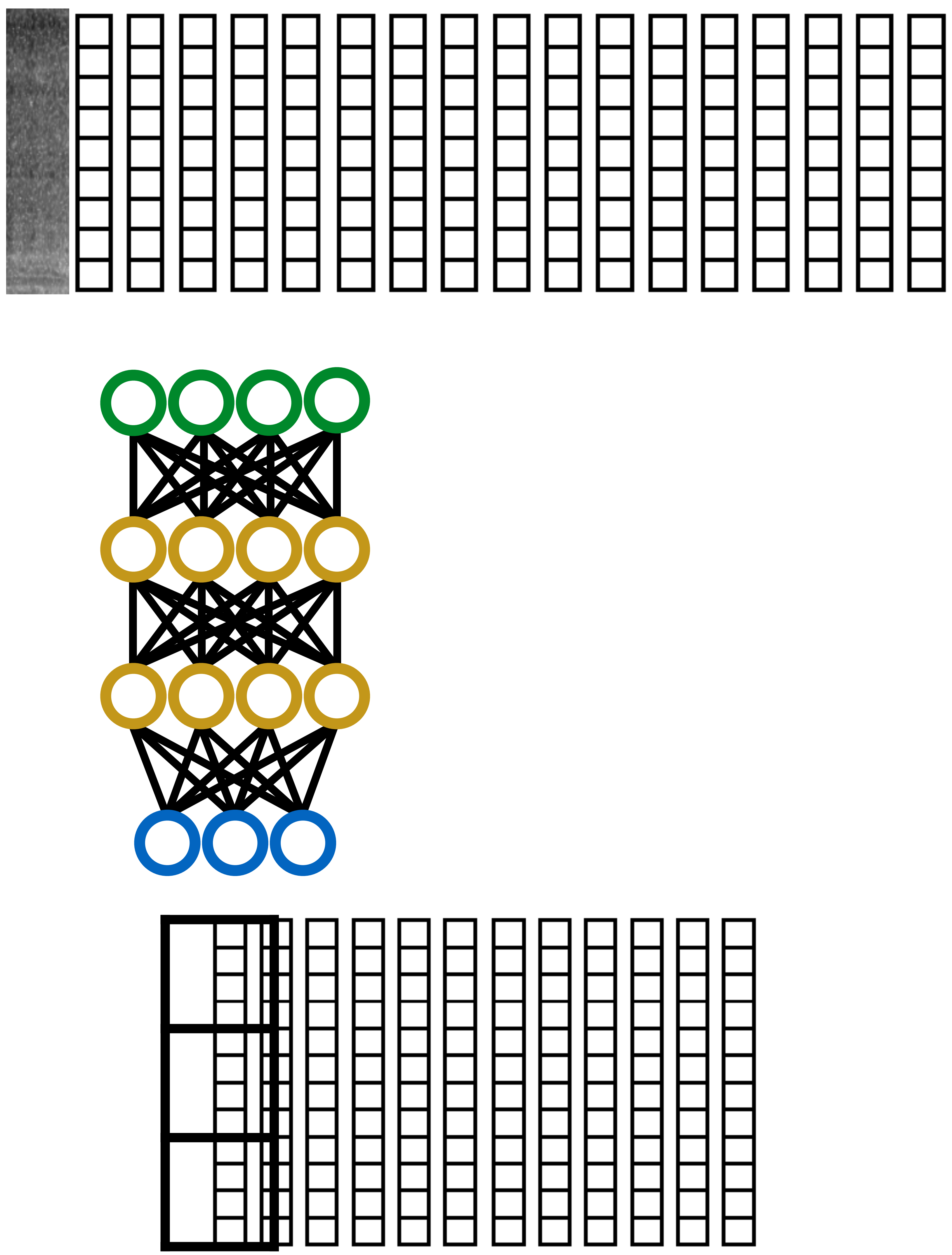
output sequence



input sequence

Sequence-to-sequence regression using a DNN

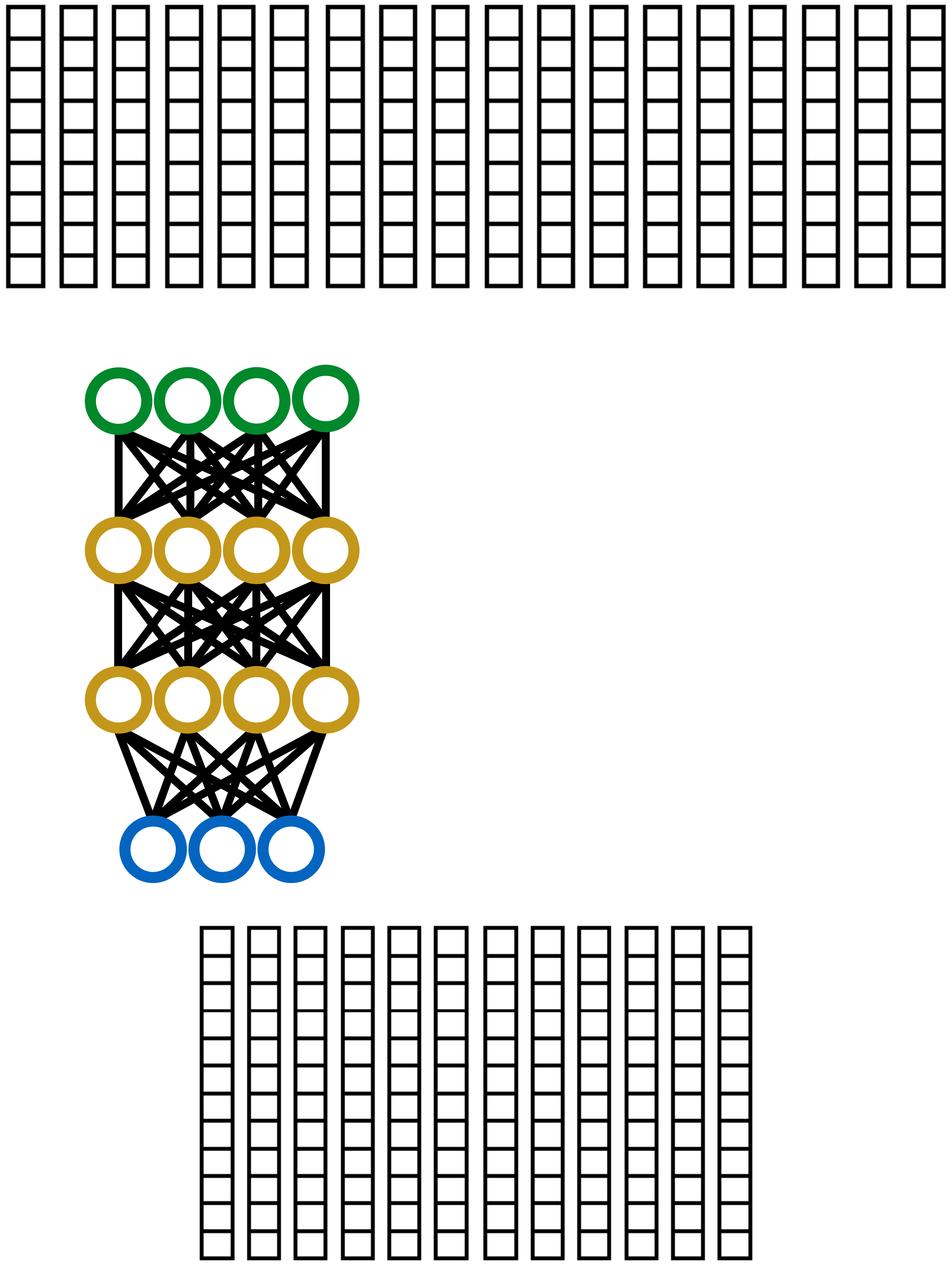
output sequence



input sequence

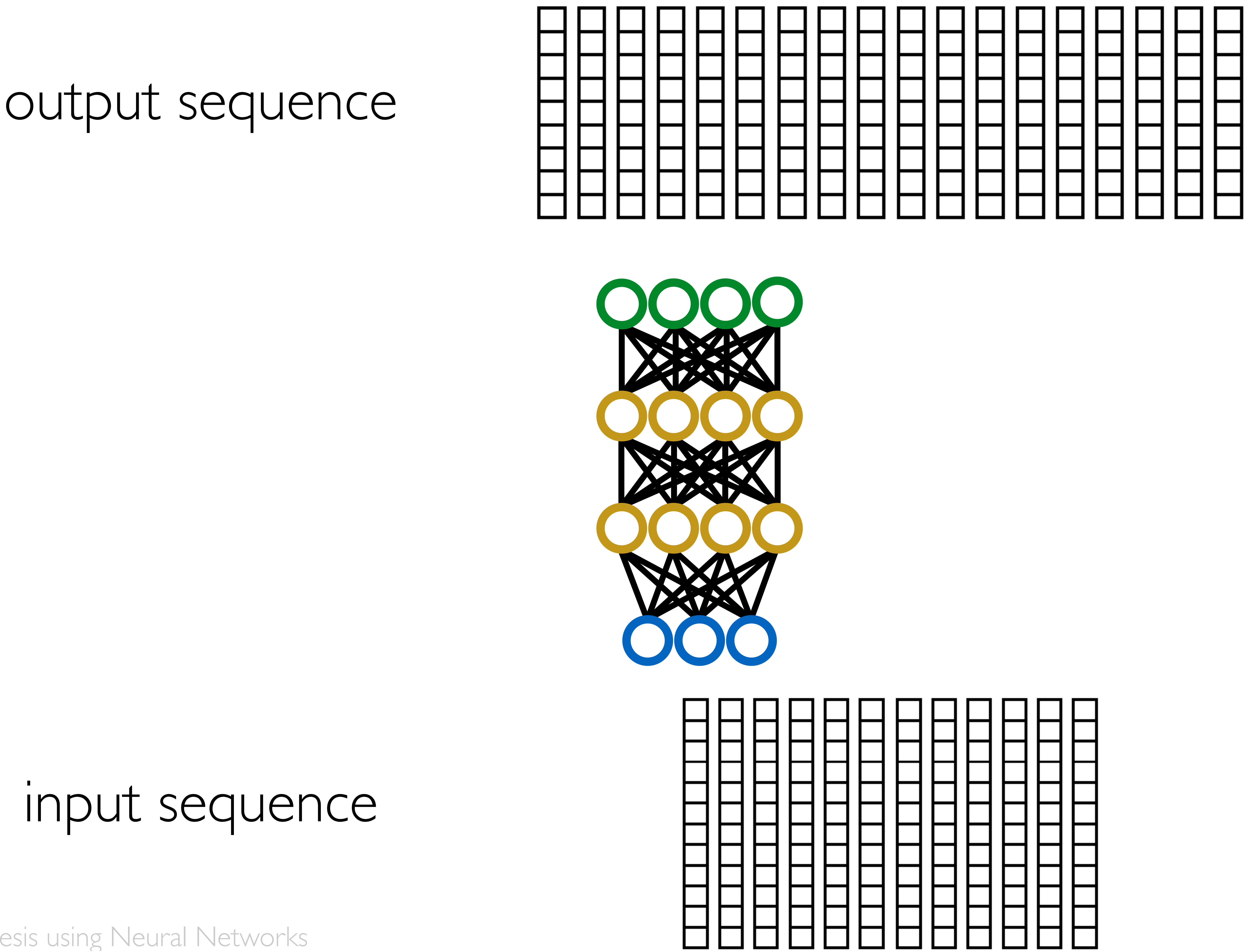
Sequence-to-sequence regression using a DNN

output sequence



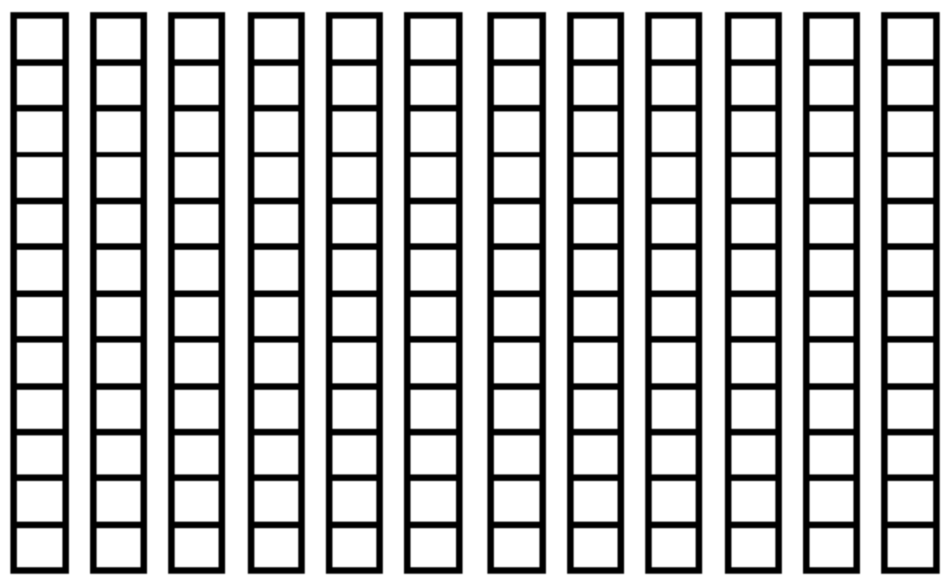
input sequence

Sequence-to-sequence regression using a DNN - dealing with duration



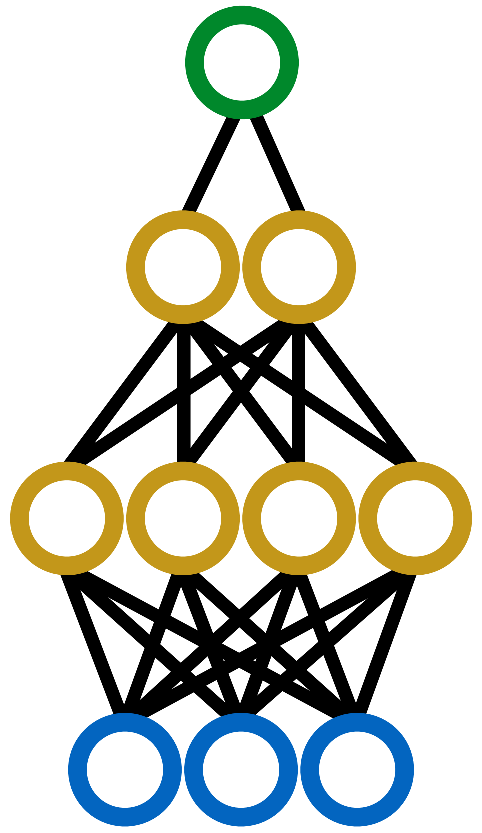
Sequence-to-sequence regression using a DNN - dealing with duration

input sequence

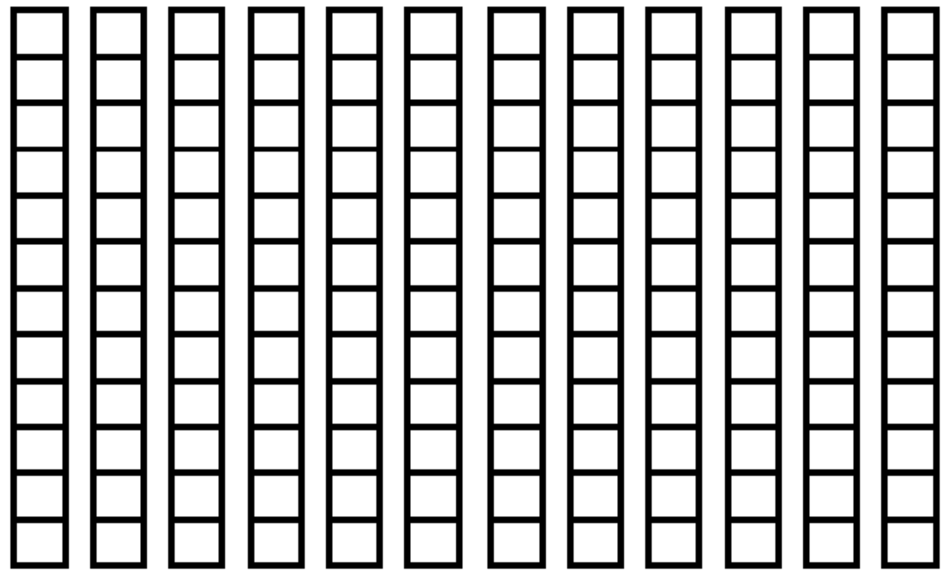


Sequence-to-sequence regression using a DNN - dealing with duration

duration
model



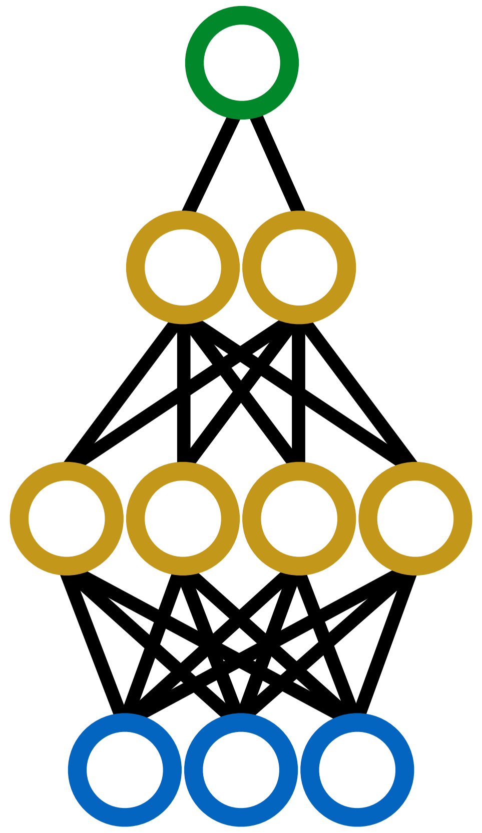
input sequence



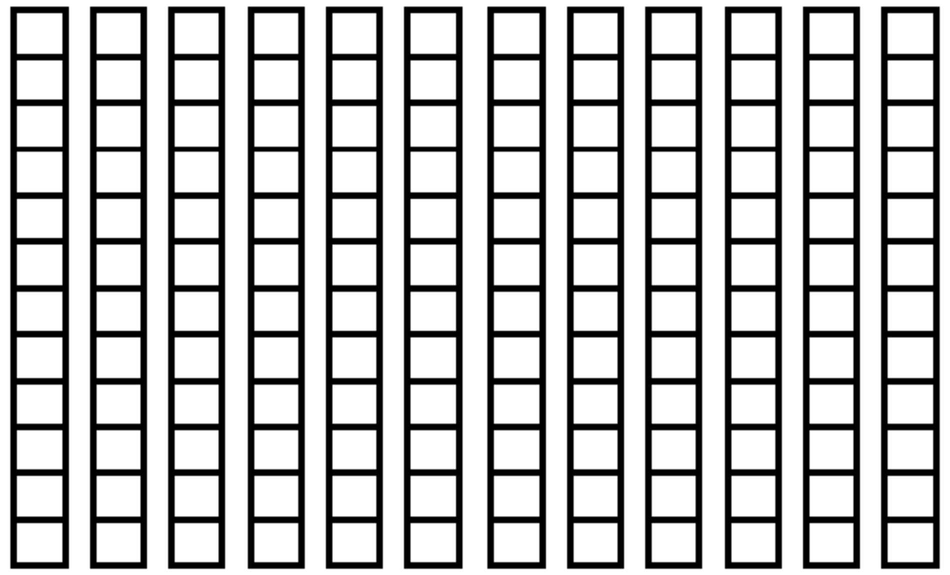
Sequence-to-sequence regression using a DNN - dealing with duration

upsampled input sequence

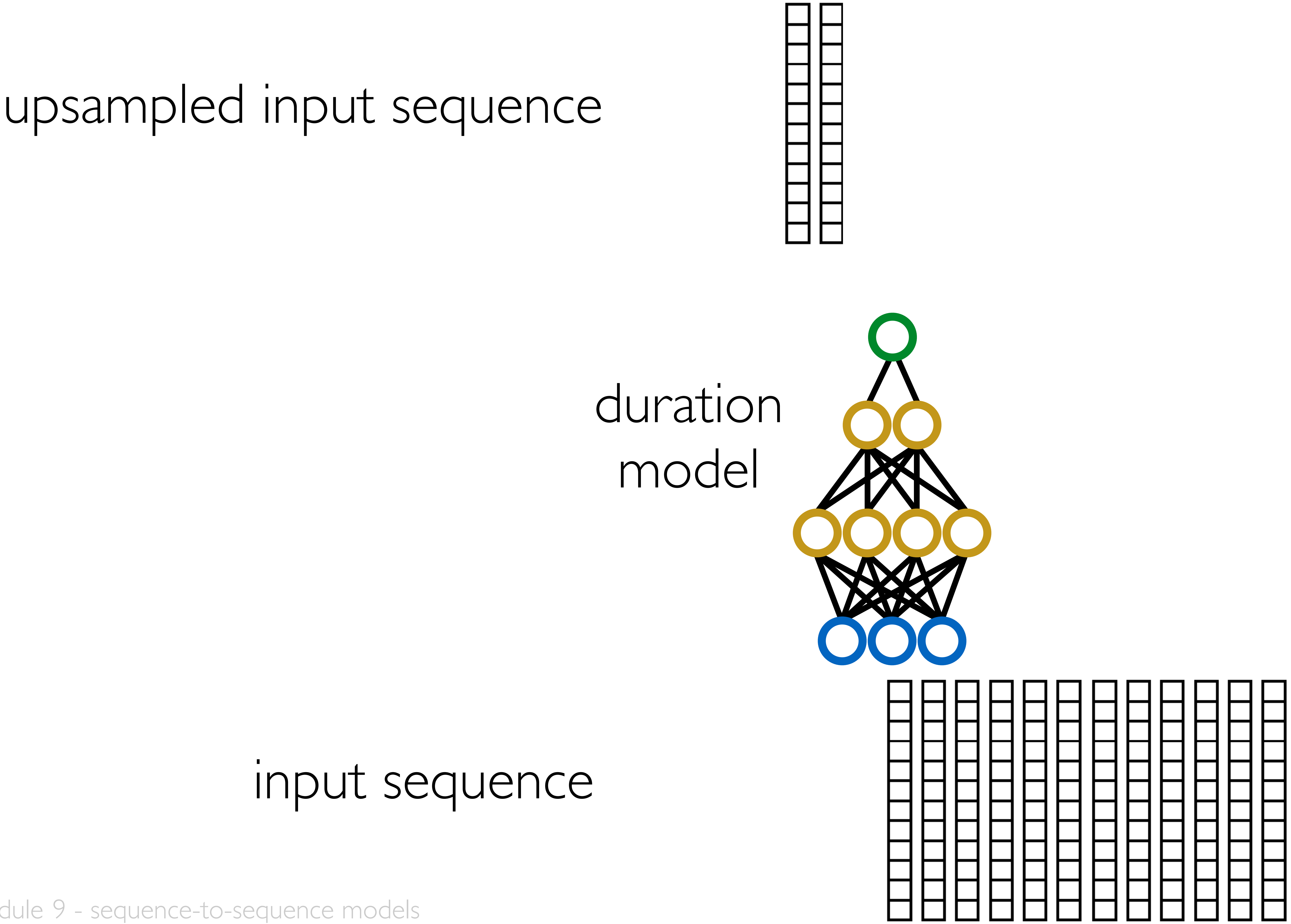
duration model



input sequence



Sequence-to-sequence regression using a DNN - dealing with duration

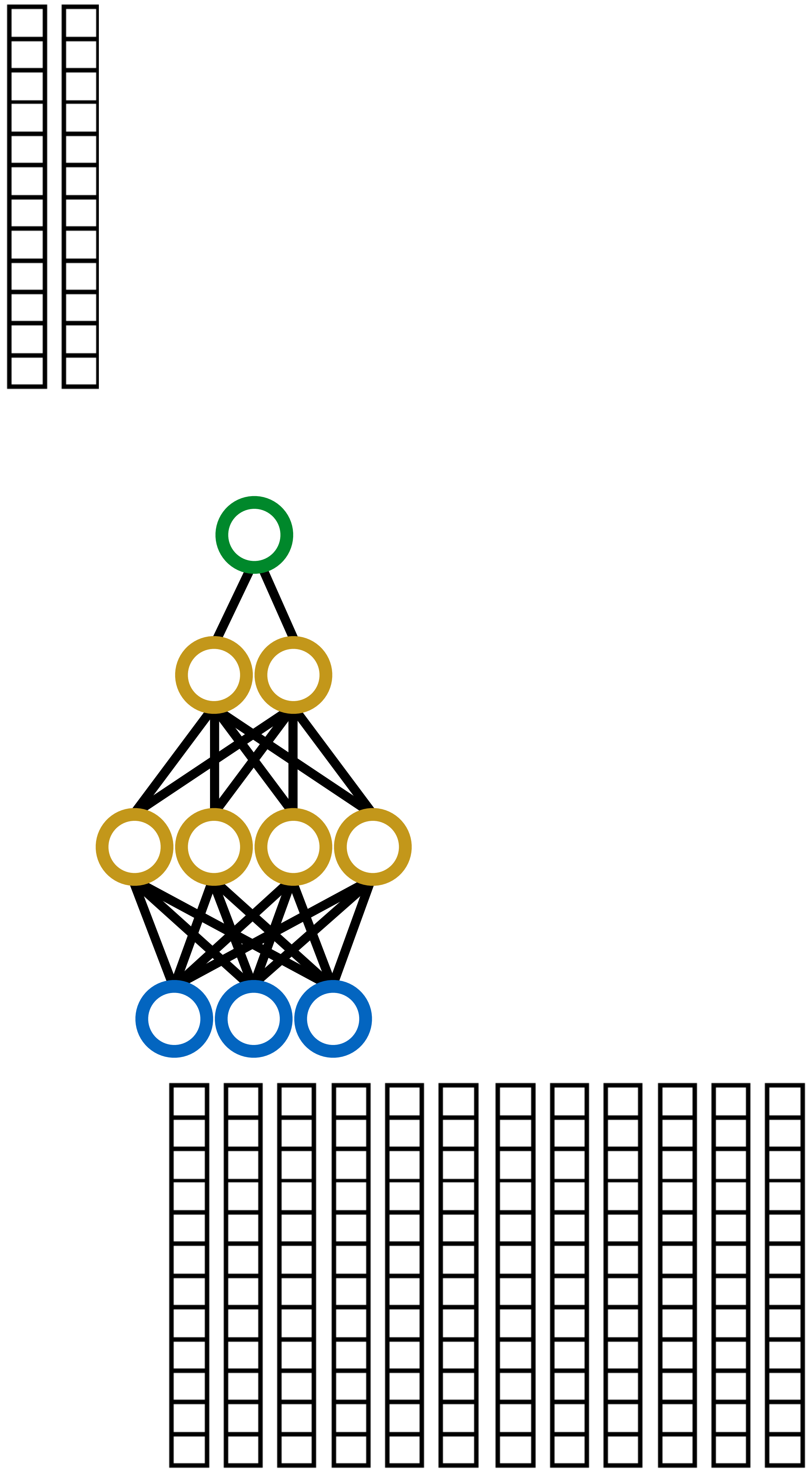


Sequence-to-sequence regression using a DNN - dealing with duration

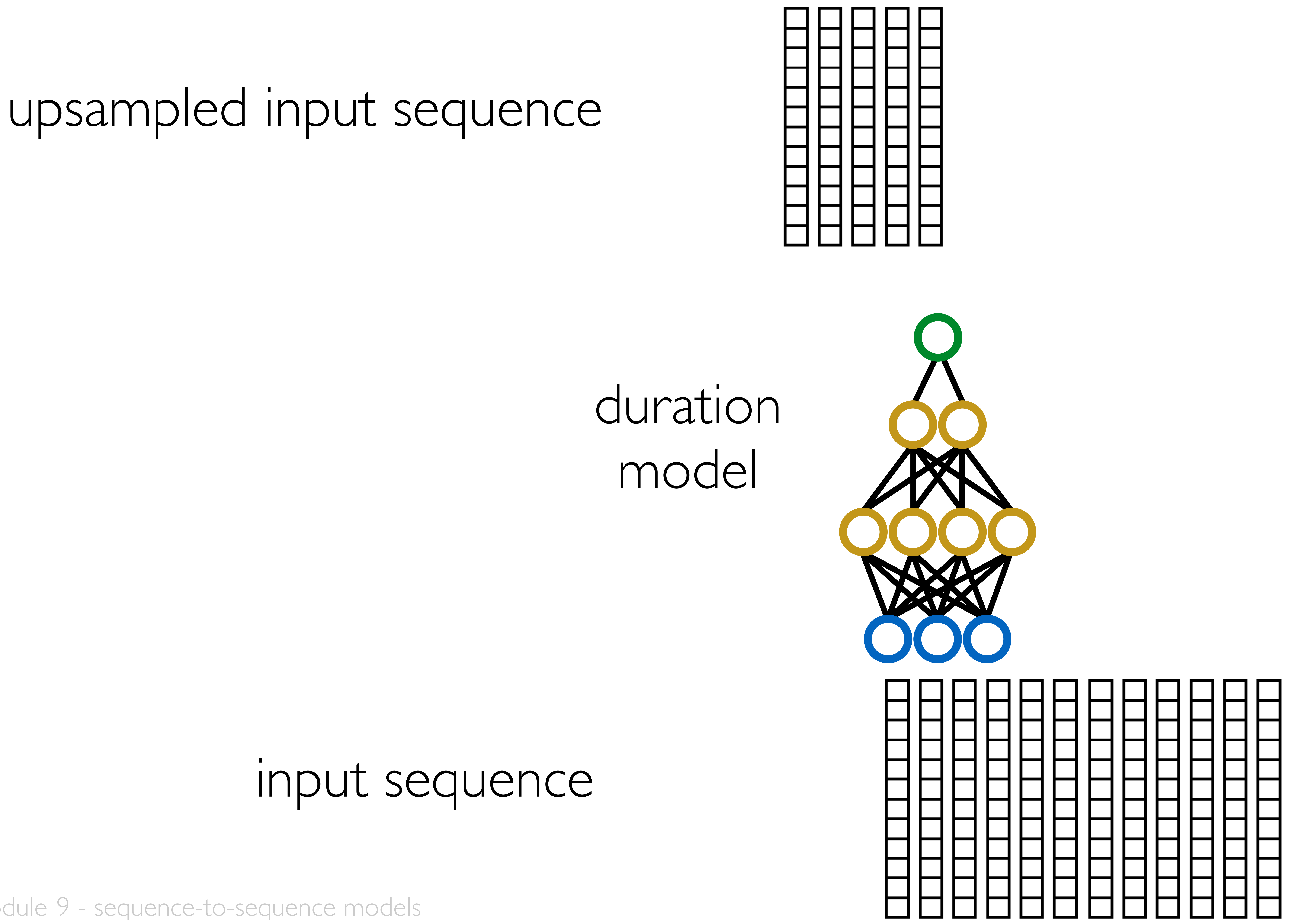
upsampled input sequence

duration model

input sequence



Sequence-to-sequence regression using a DNN - dealing with duration

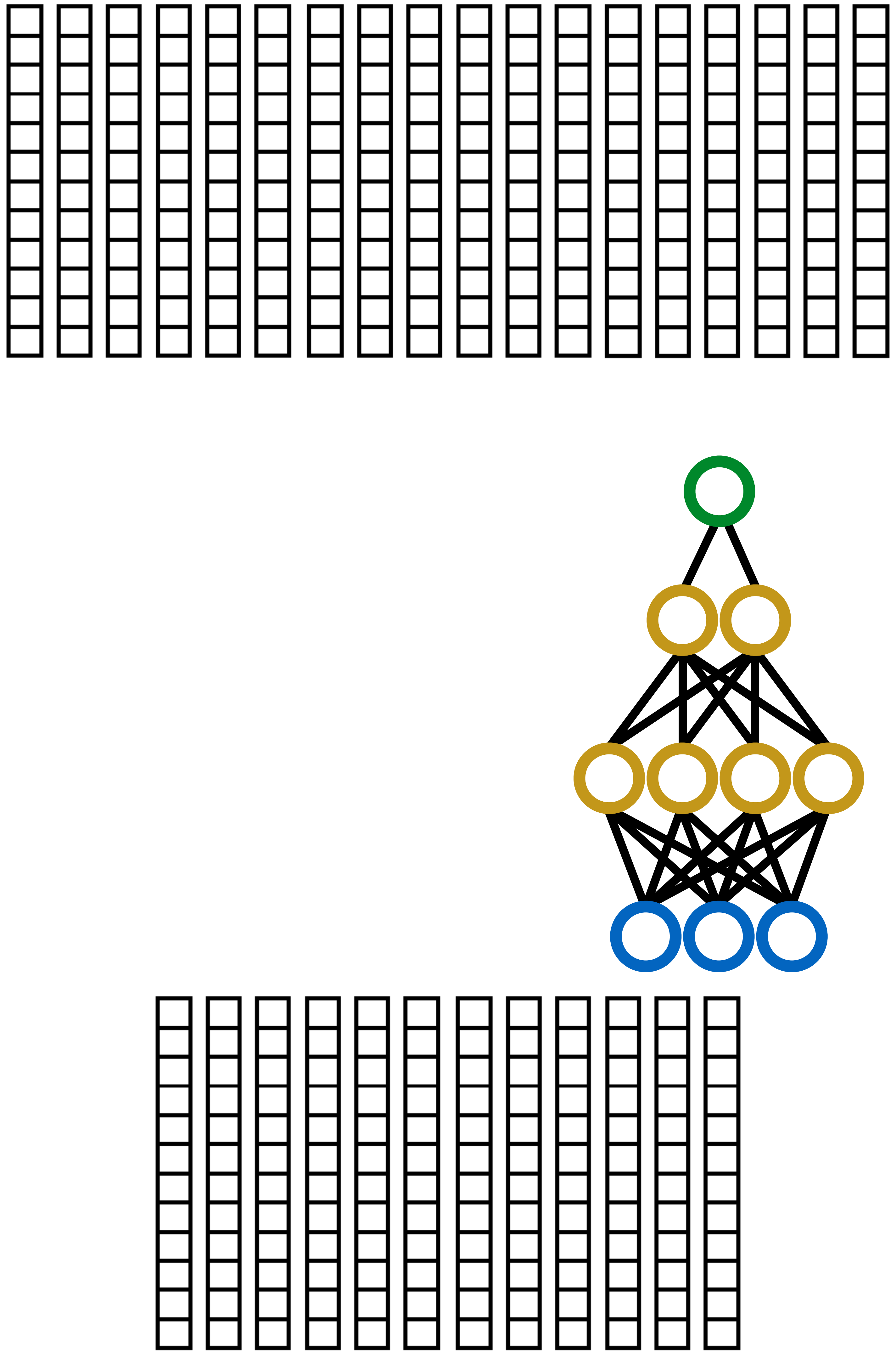


Sequence-to-sequence regression using a DNN - dealing with duration

upsampled input sequence

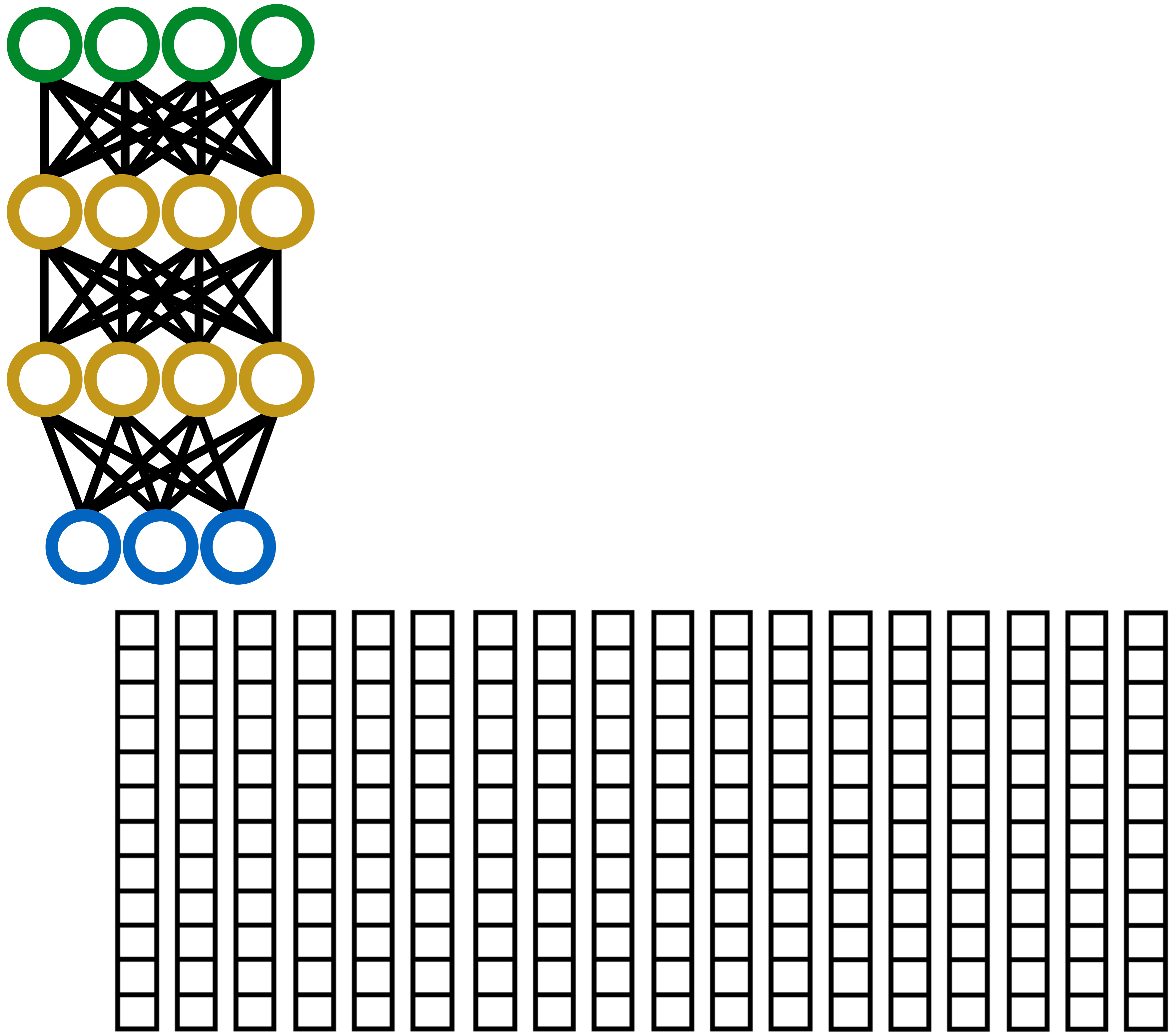
duration model

input sequence



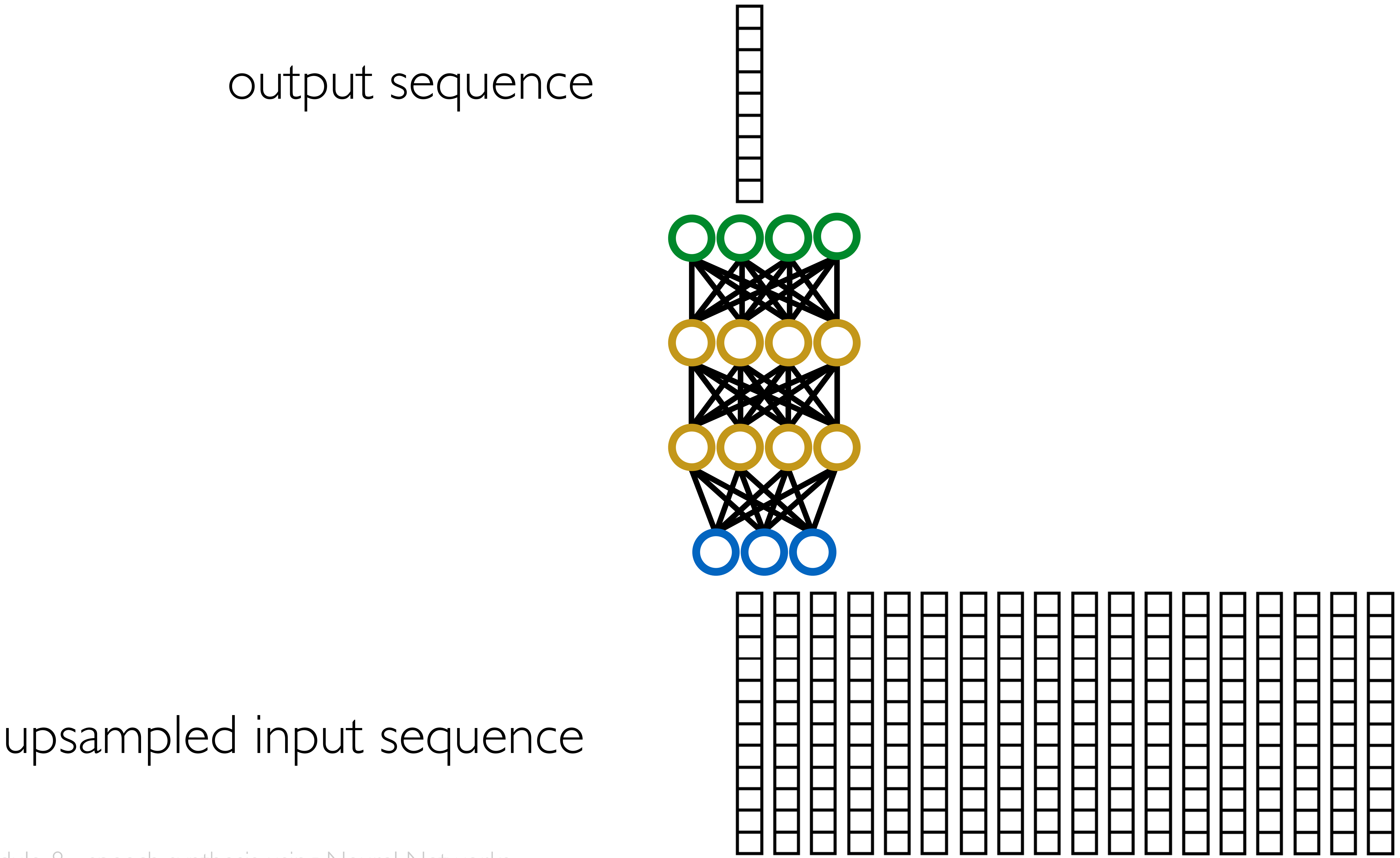
Sequence-to-sequence regression using a DNN - dealing with duration

output sequence



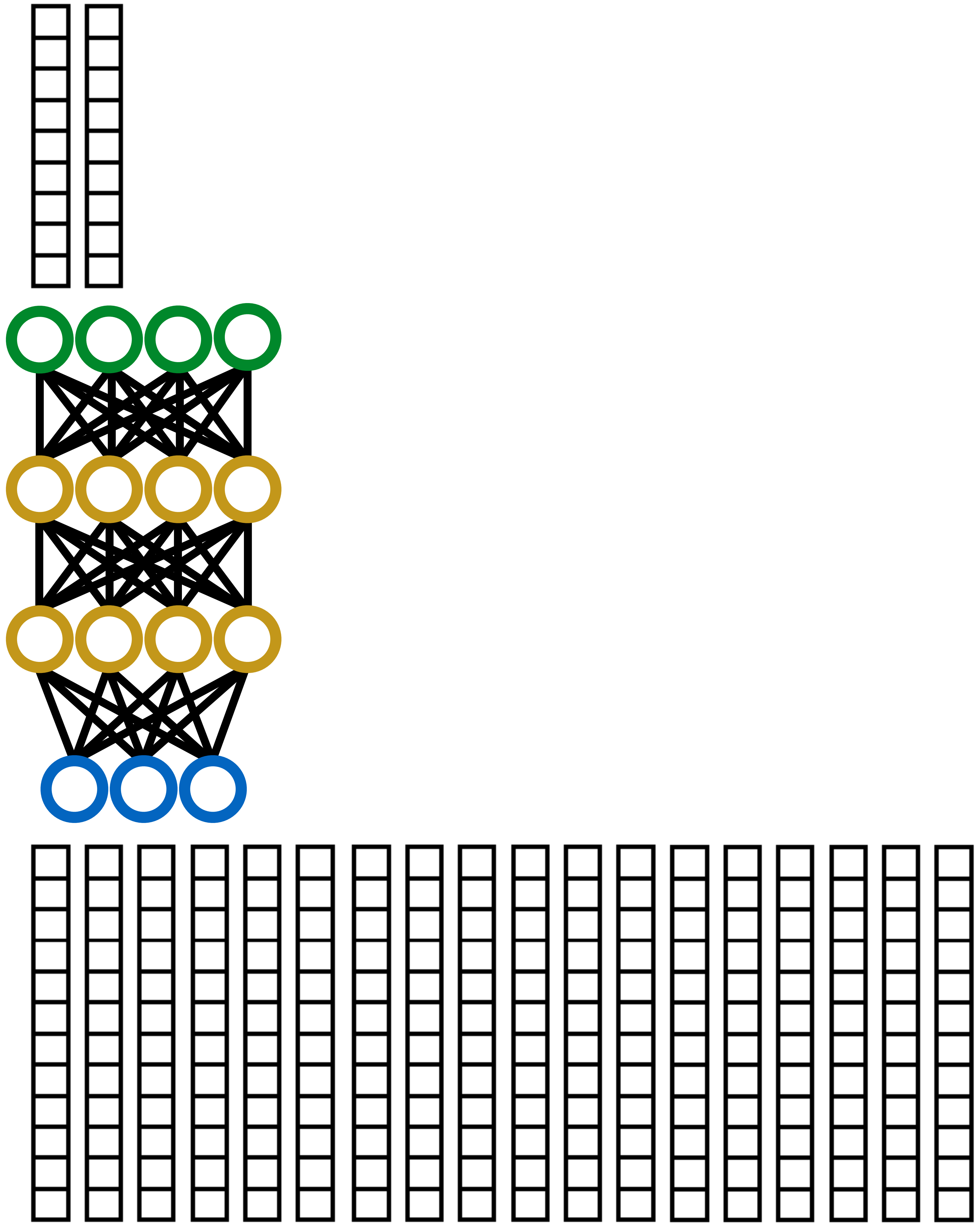
upsampled input sequence

Sequence-to-sequence regression using a DNN - dealing with duration



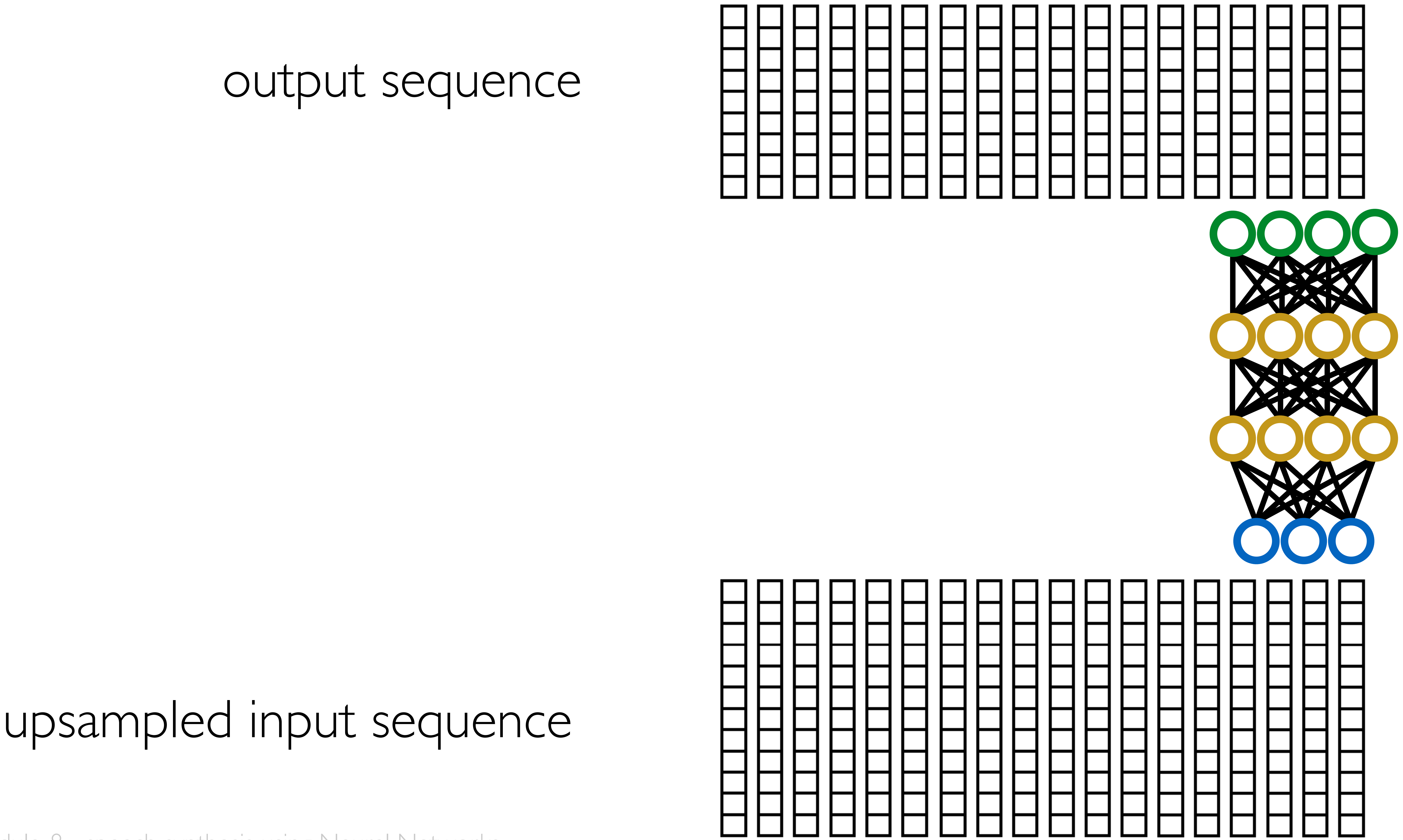
Sequence-to-sequence regression using a DNN - dealing with duration

output sequence



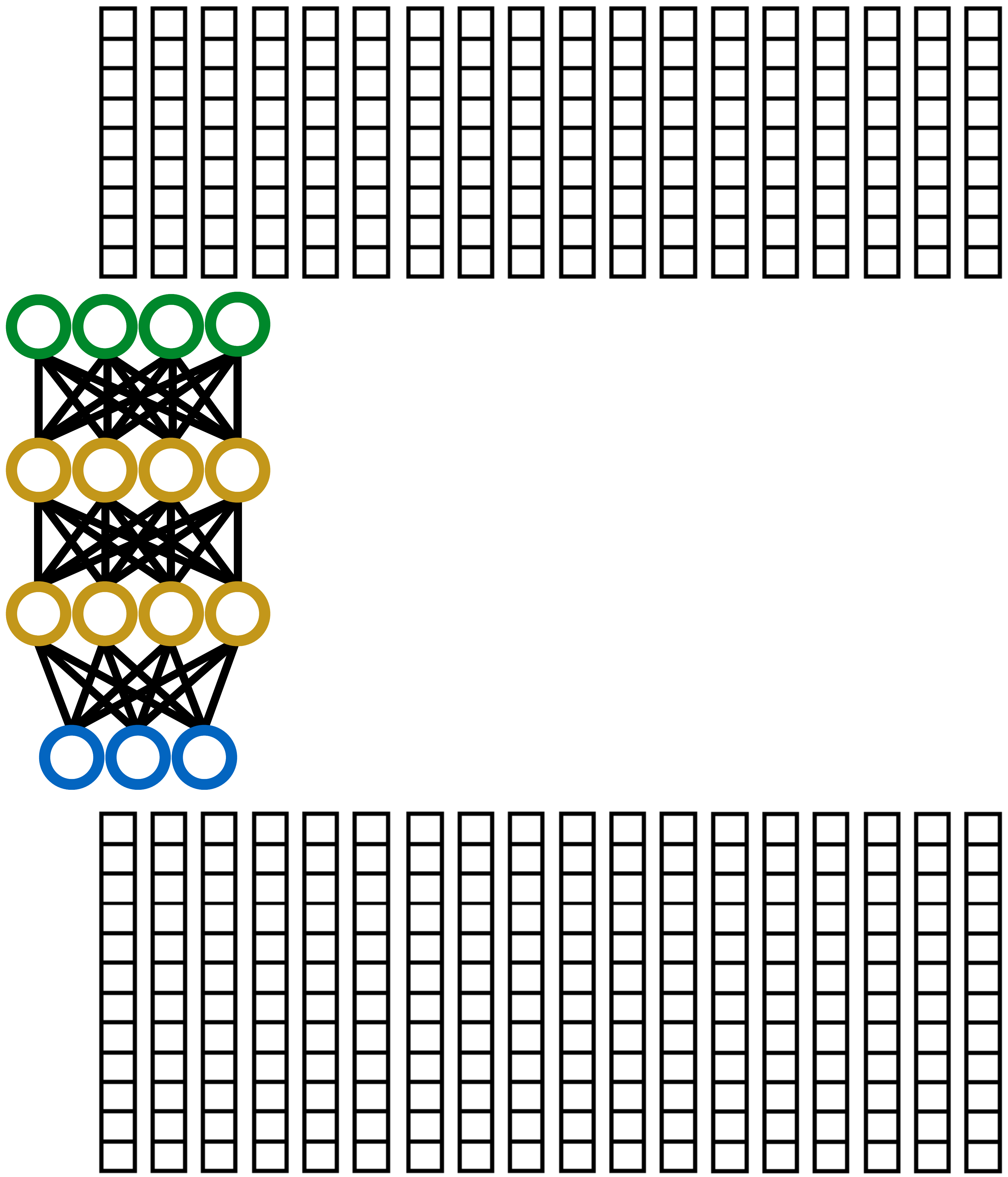
upsampled input sequence

Sequence-to-sequence regression using a DNN - dealing with duration



Processing the entire sequence at once

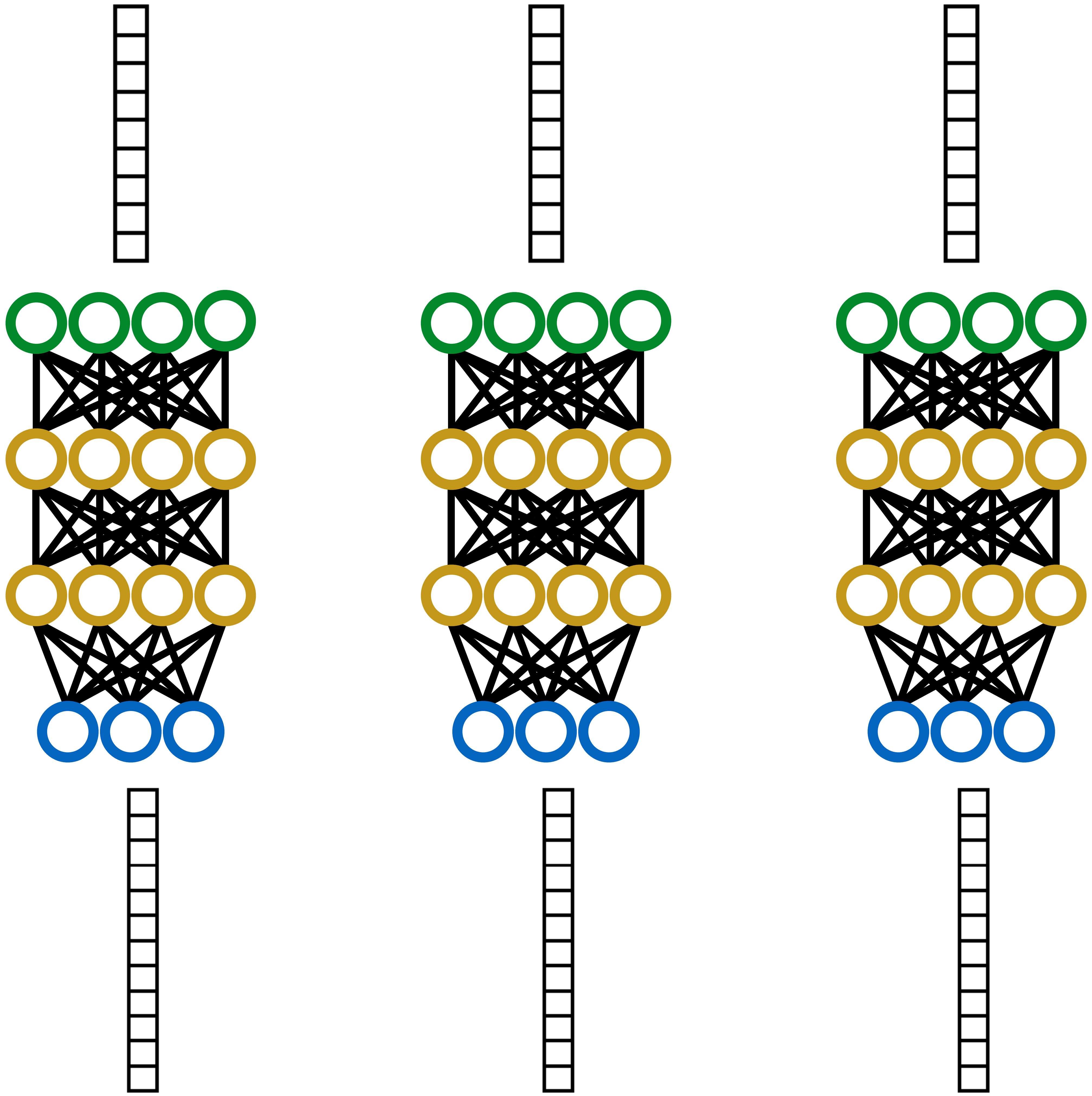
output sequence



upsampled input sequence

Processing the entire sequence at once = duplicate model for every time step

output sequence

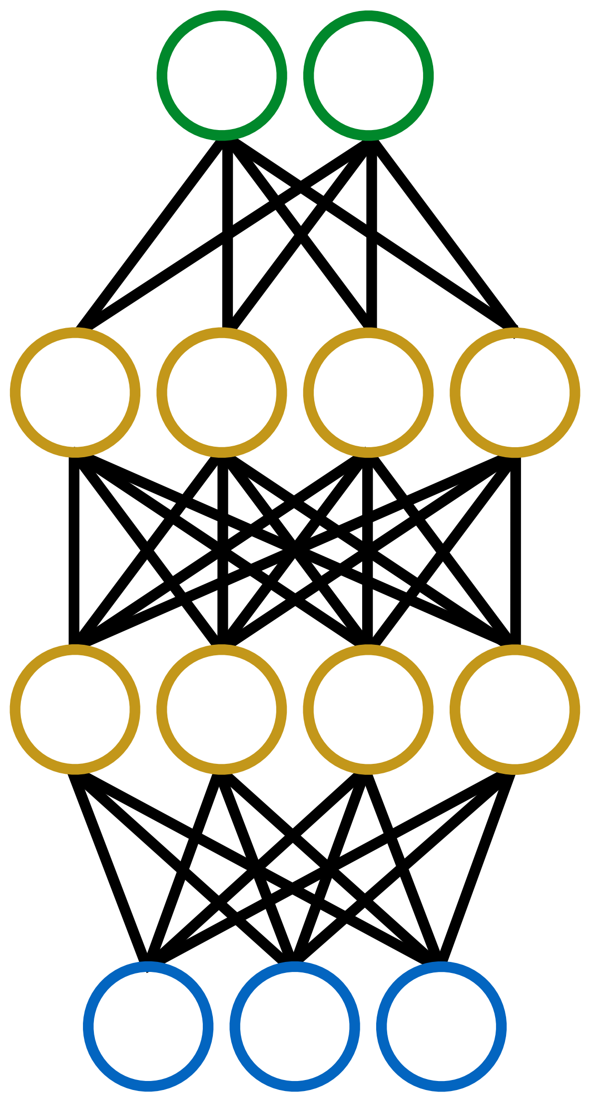


upsampled input sequence

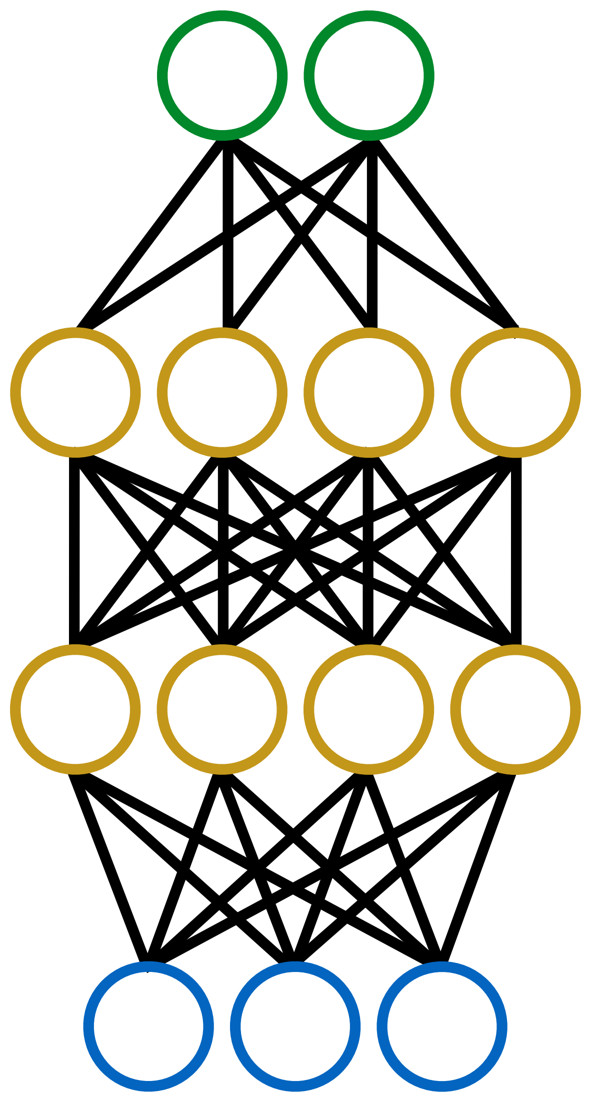
Terminology

- time step

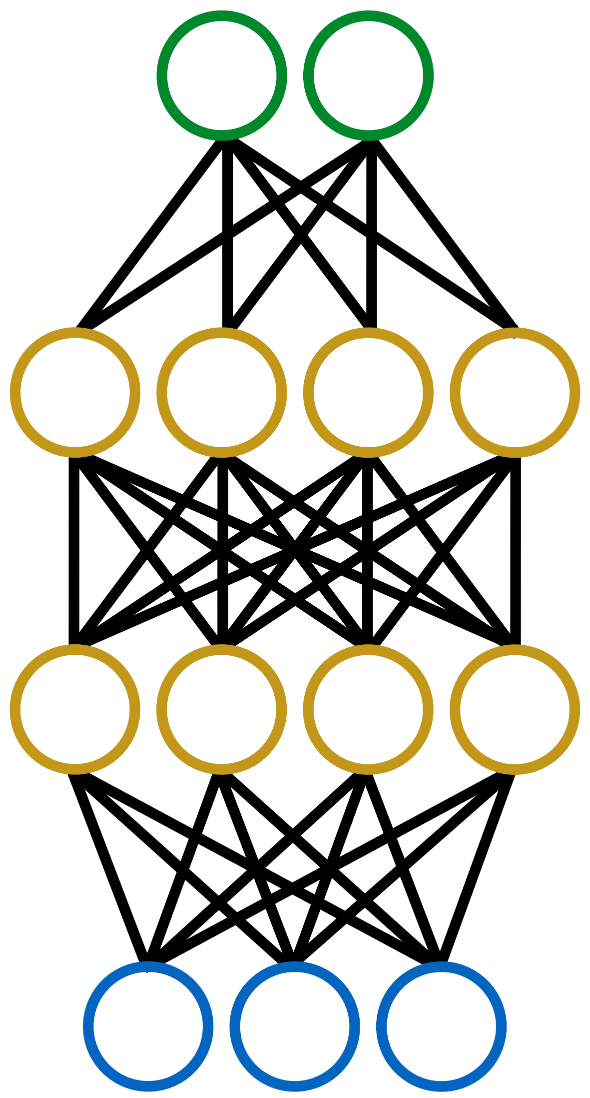
Limitations of processing each time step independently



t-1



t



t+1

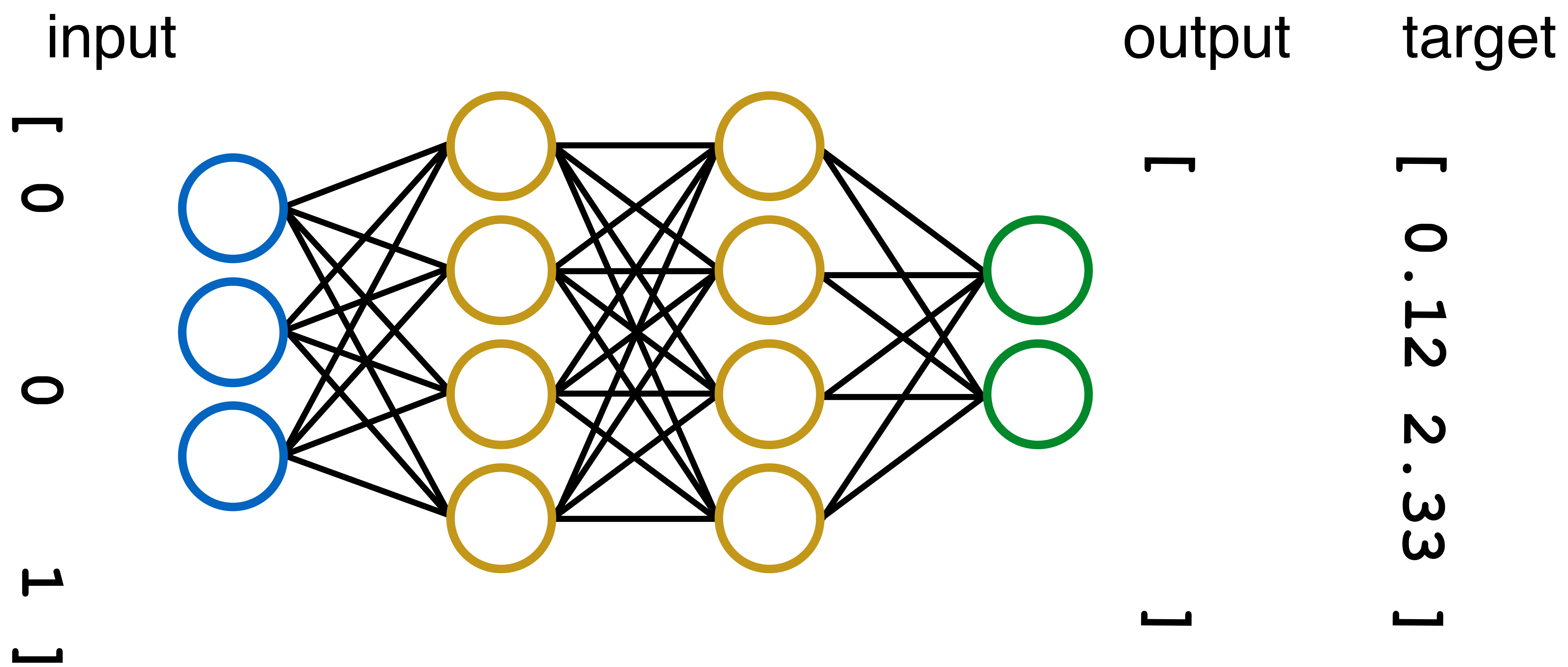
Limitations of processing each time step independently

- <u>Input features</u>
  - Requires assembling all necessary contextual information and placing at current input
  - Features pre-determined using knowledge-driven feature engineering (e.g., quinphones)
- <u>Duration</u>
  - Must be handled separately
- <u>Sequence modelling</u>
  - A constant regression function, time-independent, memoryless
- <u>Output features</u>
  - Predicted using only the input features
  - Output is conditionally-independent of previous/next outputs, given the current input

Things to improve in the next class

- <u>Input features</u>
  - the model should **learn input feature engineering**
- <u>Duration</u>
  - **integrate** into the model
- <u>Sequence modelling</u>
  - enable the model to pass information between time steps - give it a **memory**
- <u>Output features</u>
  - allow output to **depend** on previous outputs

Training a neural network: back-propagation



Training a neural network: pairs of input/output vectors

```
[0 0 1 0 0 1 0 1 1 0 … 0.2  0.0]      [0.12 2.33 2.01 0.32 6.33 … ]
[0 0 1 0 0 1 0 1 1 0 … 0.2  0.1]      [0.43 2.11 1.99 0.39 4.83 … ]
…
[0 0 1 0 0 1 0 1 1 0 … 0.2  1.0]      [1.11 2.01 1.87 0.36 2.14 … ]
[0 0 1 0 0 1 0 1 1 0 … 0.4  0.0]      [1.52 1.82 1.89 0.34 1.04 … ]
[0 0 1 0 0 1 0 1 1 0 … 0.4  0.5]      [1.79 1.74 2.21 0.33 0.65 … ]
[0 0 1 0 0 1 0 1 1 0 … 0.4  1.0]      [1.65 1.58 2.68 0.31 0.73 … ]
…
[0 0 1 0 0 1 0 1 1 0 … 1.0  1.0]      [1.55 1.03 3.44 0.30 1.07 … ]
[0 0 0 1 1 1 0 1 0 0 … 0.2  0.0]      [1.92 0.99 3.89 0.29 1.45 … ]
[0 0 0 1 1 1 0 1 0 0 … 0.2  0.2]      [2.38 1.13 4.02 0.28 1.98 … ]
[0 0 0 1 1 1 0 1 0 0 … 0.2  0.4]      [2.65 1.98 3.94 0.29 2.16 … ]
…
```

Tips on interpreting the marking sheet

Speech Synthesis assignment marking scheme

Category		Points available
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

The marking sheet is *not* a table of contents for your paper

Speech Synthesis assignment marking scheme

Category		Points available
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

A well-structured, polished report showing good effort, with interesting and justified investigations and claims supported by evidence, will get a good grade.

Tips on interpreting the marking sheet

Speech Synthesis assignment marking scheme

Category		**Points available**
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- **Informative** title
- **Structured** abstract
- A brief introduction to **this** paper
  - "scene setting"
  - relevant background (within reason)
  - clear motivation for the work
  - (paper outline/what to expect)
  - (not results or conclusions)

https://www.annaclemens.com/blog/how-to-write-the-perfect-abstract

Tips on interpreting the marking sheet

Speech Synthesis assignment marking scheme

Category		Points available
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- **Brief** explanation only
- Keep it relevant to **this** paper
- Demonstrate your **understanding**

Tips on interpreting the marking sheet

Speech Synthesis assignment marking scheme

Category		**Points available**
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- **Incorporate** these throughout the paper
- Example 1:
  - Unit selection performs implicit **regression** from linguistic features to acoustic properties
  - How do various current methods do that?
- Example 2:
  - In unit selection, several choices are available for **waveform representation**
  - Are these the same or different in current methods?
- etc.

Tips on interpreting the marking sheet

Speech Synthesis assignment marking scheme

Category		Points available
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- Self-explanatory
- Look at the mark available and keep the basics really tight and to the point
- Optional extra work, experiments, etc, will attract marks in other categories

Tips on interpreting the marking sheet

Speech Synthesis assignment marking scheme

Category		Points available
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- Often overlooked, but easy marks available!
- Just show that you understand the various forms of **signal processing** that are happening
  - in voice building
  - during synthesis

Tips on interpreting the marking sheet

Speech Synthesis assignment marking scheme

Category		Points available
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- Link everything to **current methods**. Do not do experiments with current methods, but use the literature to back up your claims.
- Example:
  - You will have discovered how sensitive (or not) unit selection is to many **design choices**, such as database contents, pitchmark accuracy, ...
  - Would current methods be more or less sensitive to each choice?

Tips on interpreting the marking sheet

Speech Synthesis assignment marking scheme

Category		Points available
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- This is where you get marks for your experimental work and basic listening test
- Further marks available under Additional for going further

Tips on interpreting the marking sheet

Speech Synthesis assignment marking scheme

Category		Points available
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- The easiest 5 marks you'll ever get!
- Don't miss out!

- Note: badly formatted work, missing exam number, lack of page numbers, etc - all create extra work for the marker and course organiser.

Tips on interpreting the marking sheet

Speech Synthesis assignment marking scheme

Category		Points available
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- Use the **feedback** from Speech Processing (*)
- **Scientific writing** should be clear, simple, and unambiguous
- Plan your paper's **structure** carefully
- Have your **reader** in mind at all times
- Good **presentation** makes a paper more enjoyable to read
- A happy marker is a generous marker

(*) If you didn't take Speech Processing, contact Simon for additional 1-on-1 help with your writing

Tips on interpreting the marking sheet

Speech Synthesis assignment marking scheme

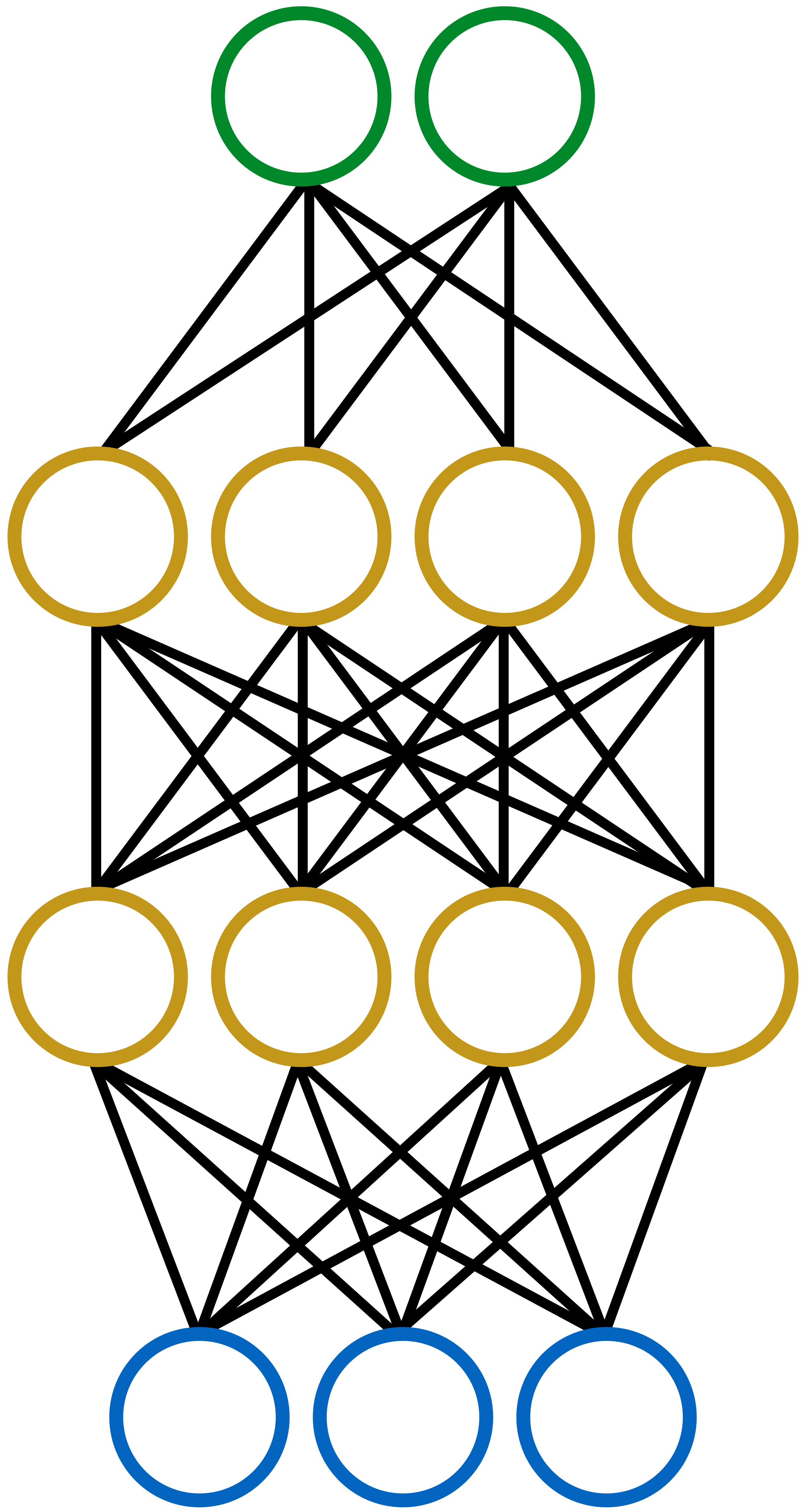
Category		Points available
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- You are *not* expected to do **all** of these!
- But be tactical:
  - do aim for some marks in **multiple categories**
  - do not try to get all 20 points for going too deep in only one category (e.g., script design)
  - the list on the marking sheet is not exhaustive: creativity will be rewarded

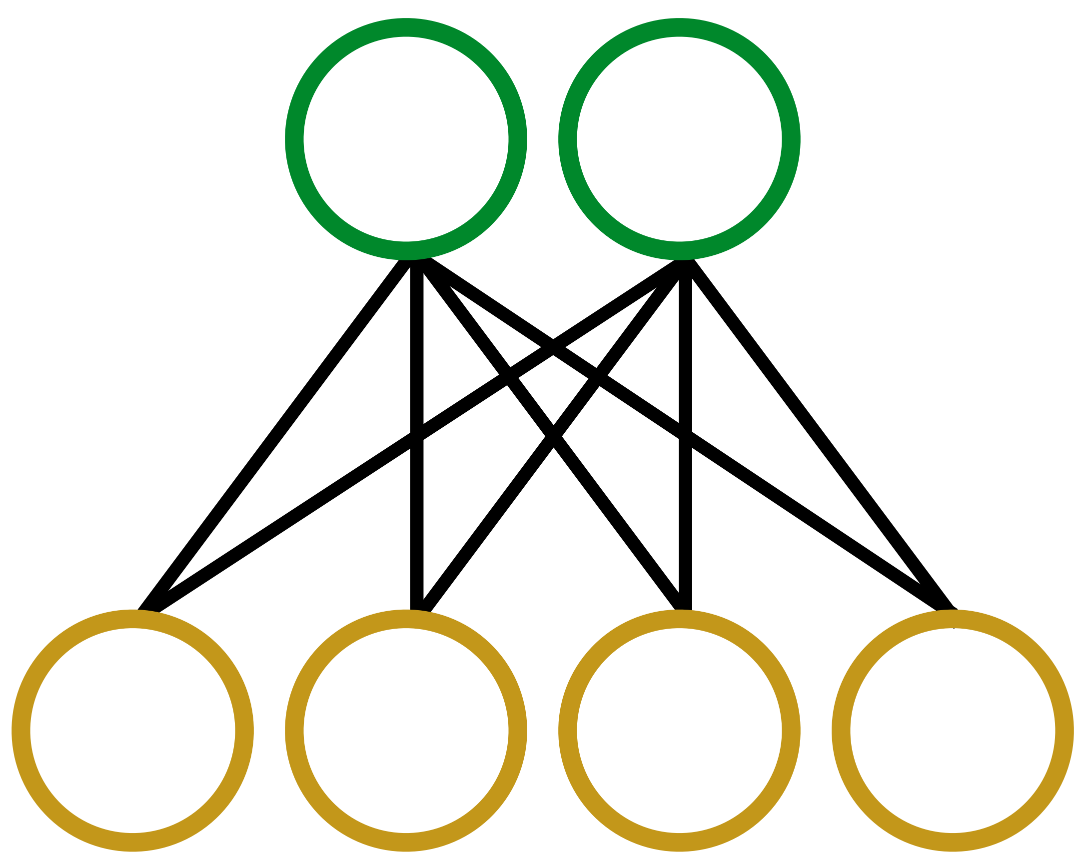
Final tips

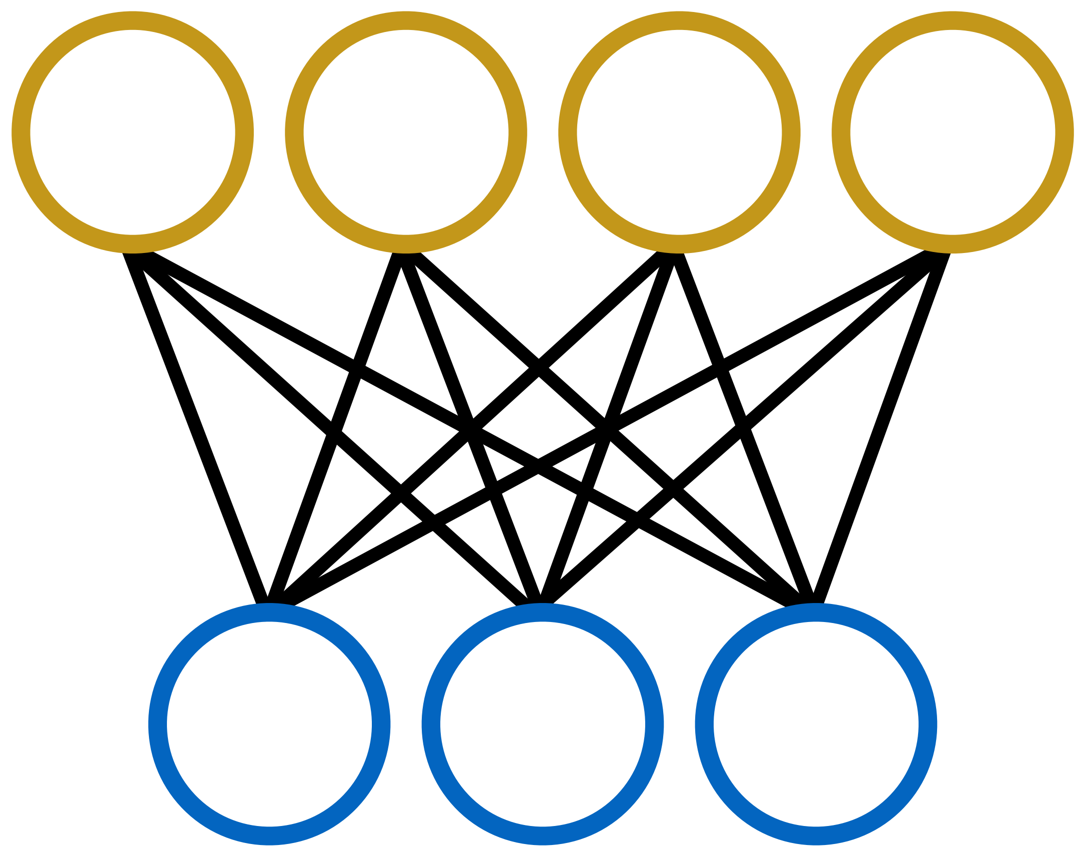
- Focus on **demonstrating your understanding**, not on how Festival and the scripts work
- **Figures** can say a lot with only a few words
- Present your experimental results in an **attractive** way
- A **bibliography** and in-text **citations** *must* be provided
  - Go beyond the Essential readings if you are aiming for a high mark
  - Cite **peer-reviewed** work whenever possible
  - Never cite a **preprint** (e.g., arXiv) when a peer-reviewed version is available
- The actual quality of your synthetic voice will **not** influence your mark
  - (although you need it to be *reasonably* intelligible before doing any listening tests)

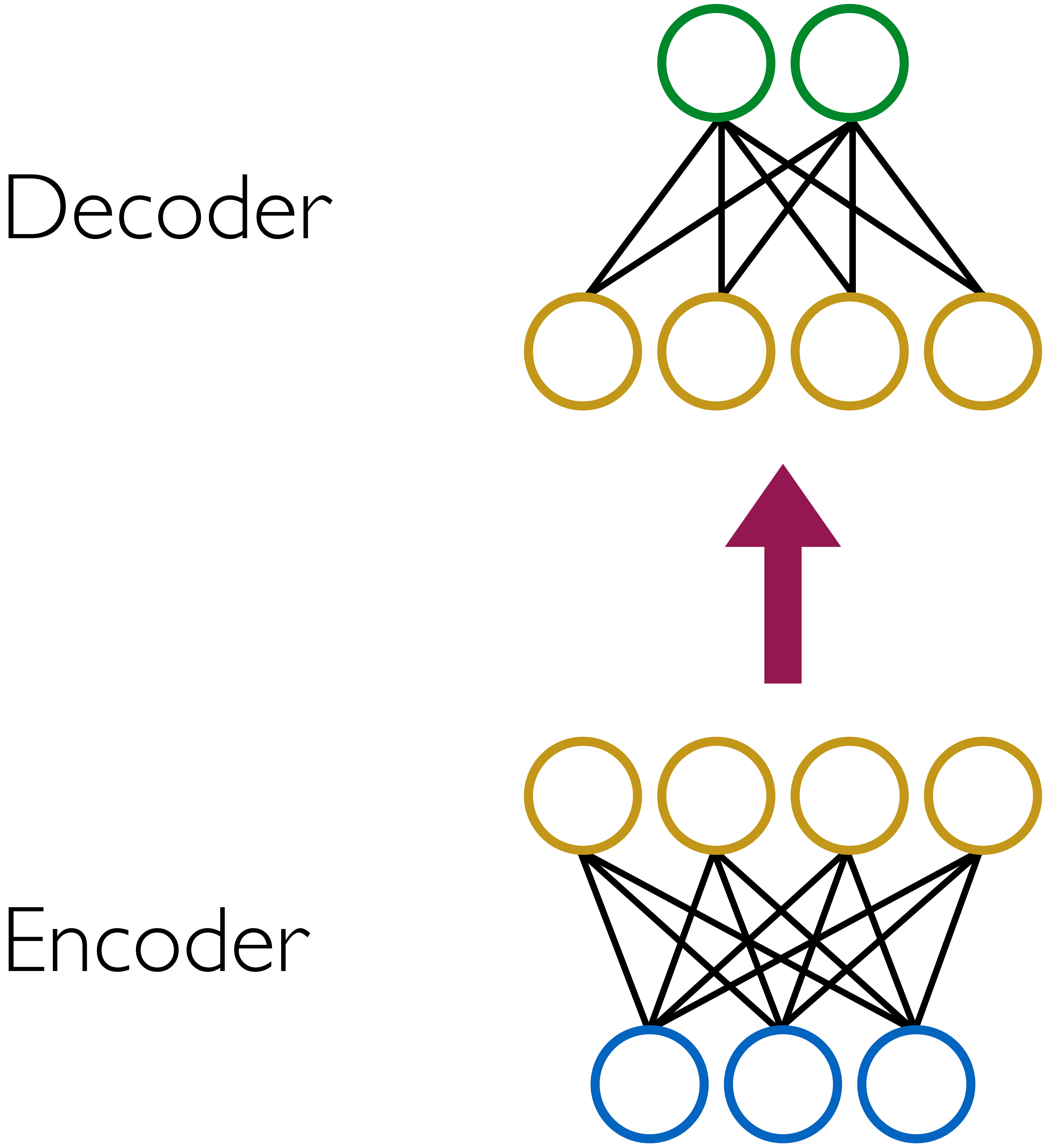
output time steps are frames (e.g., of a mel spectrogram)

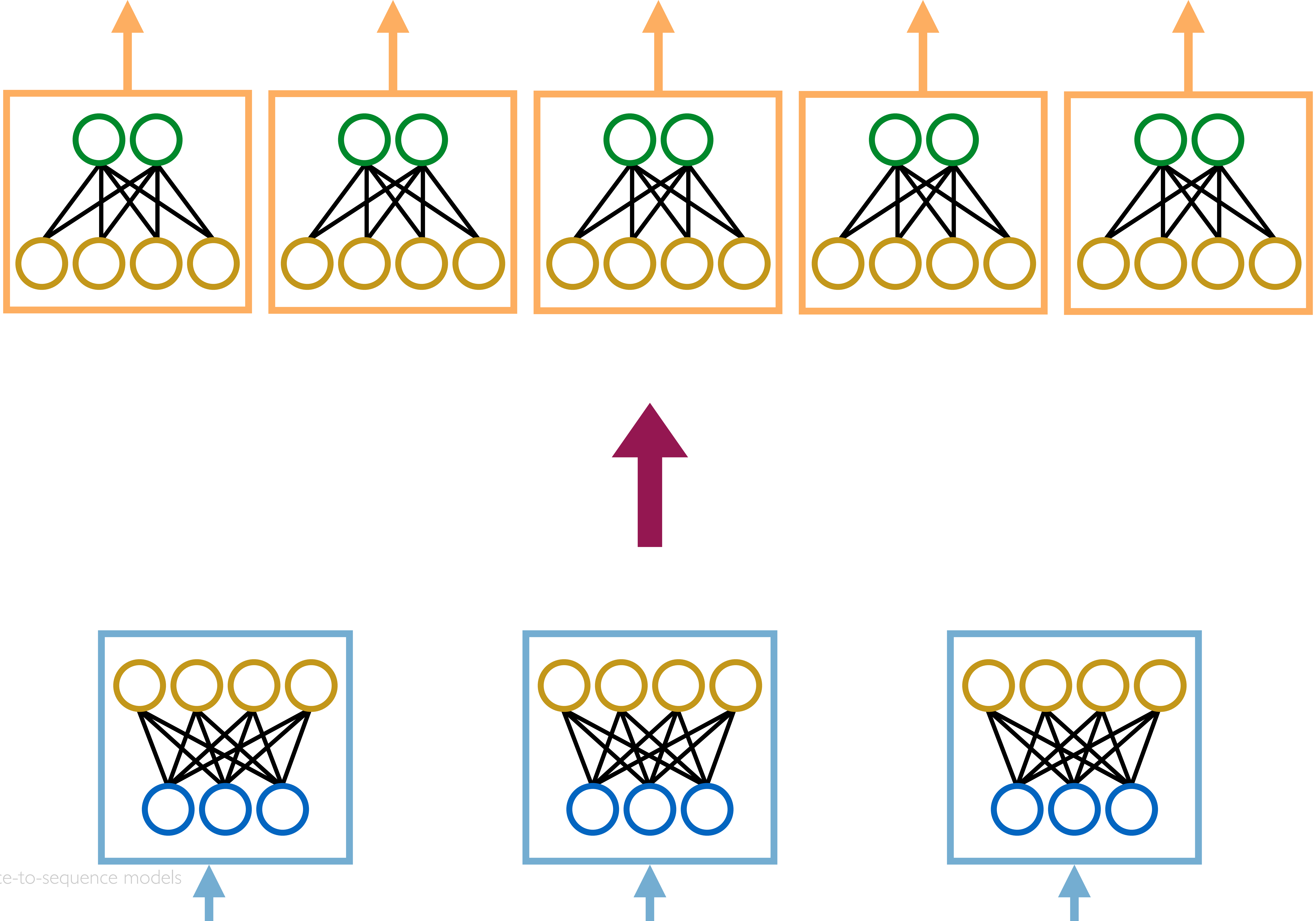


input time steps are linguistic units (e.g., phones)






Decoder
Encoder



What next?

- <u>The state of the art</u>

- **No videos** on this
  - because it changes too quickly

- You need to read the **primary literature** yourself
  - i.e., journal and conference papers
  - not textbooks