Bias in NLP models - tutorial

Vector Institute “Bias in AI” course

March 7, 2022

By Anastasia Razdaibiedina

Tutorial roadmap

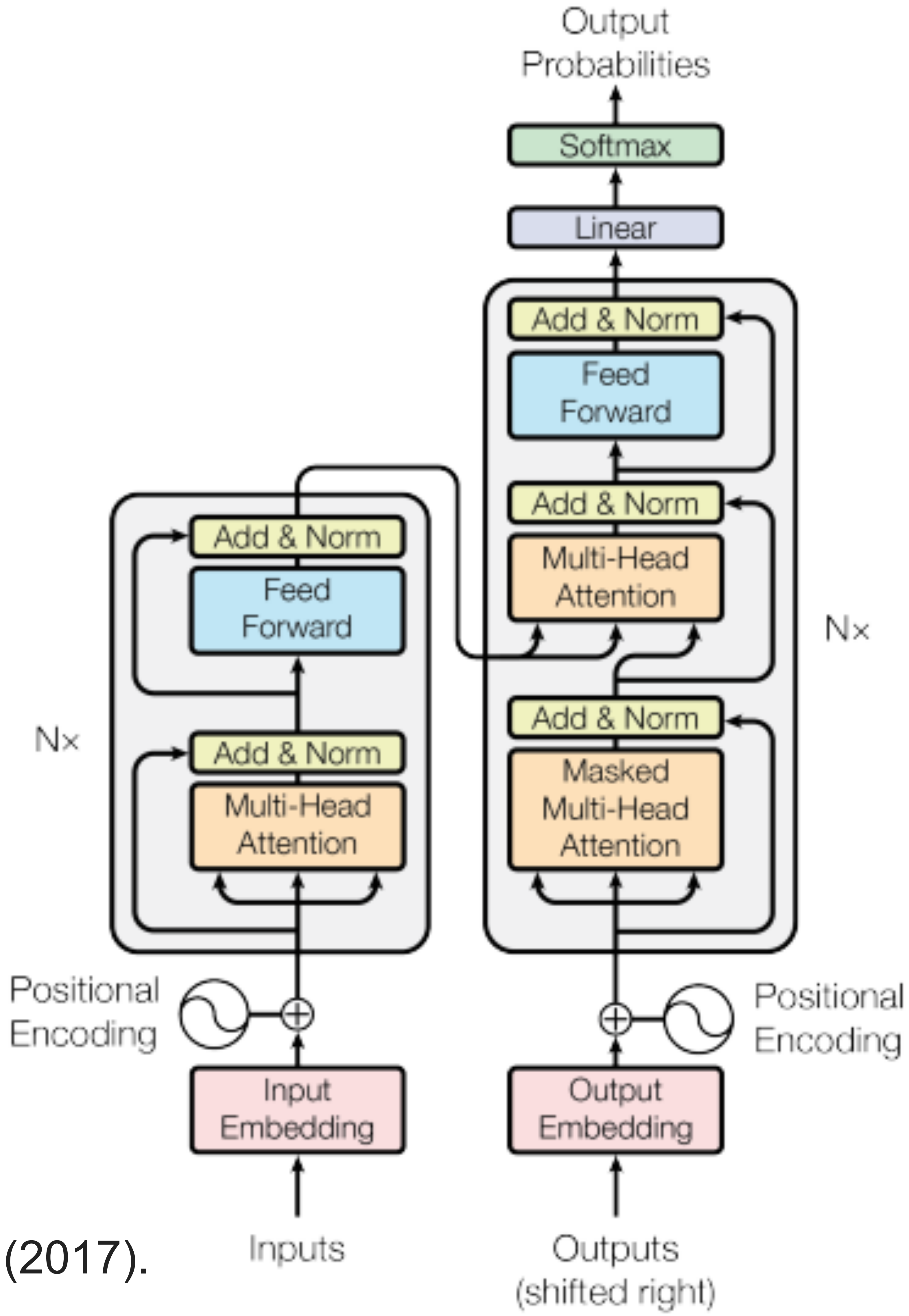
1. Transformer models & attention mechanism overview
2. Bias in pre-trained language models
3. How can we measure bias?
4. Ways to mitigate bias
5. Assignment overview

Transformer models and attention mechanism

- **Transformers** are deep learning models that adopt **attention mechanism** to perform variety of NLP tasks

- Transformers have set state-of-the-art results on text classification, text generation , machine translation, question answering and so on, and replaced previously used RNN models

- Original transformer model consisted of *encoder block* and *decoder block*. Current models have different architectures - encoder-only, decoder-only or both encoder-decoder.



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

BERT model

- **Bidirectional Encoder Representations from Transformers (BERT)** is a transformer-based machine learning technique for NLP.

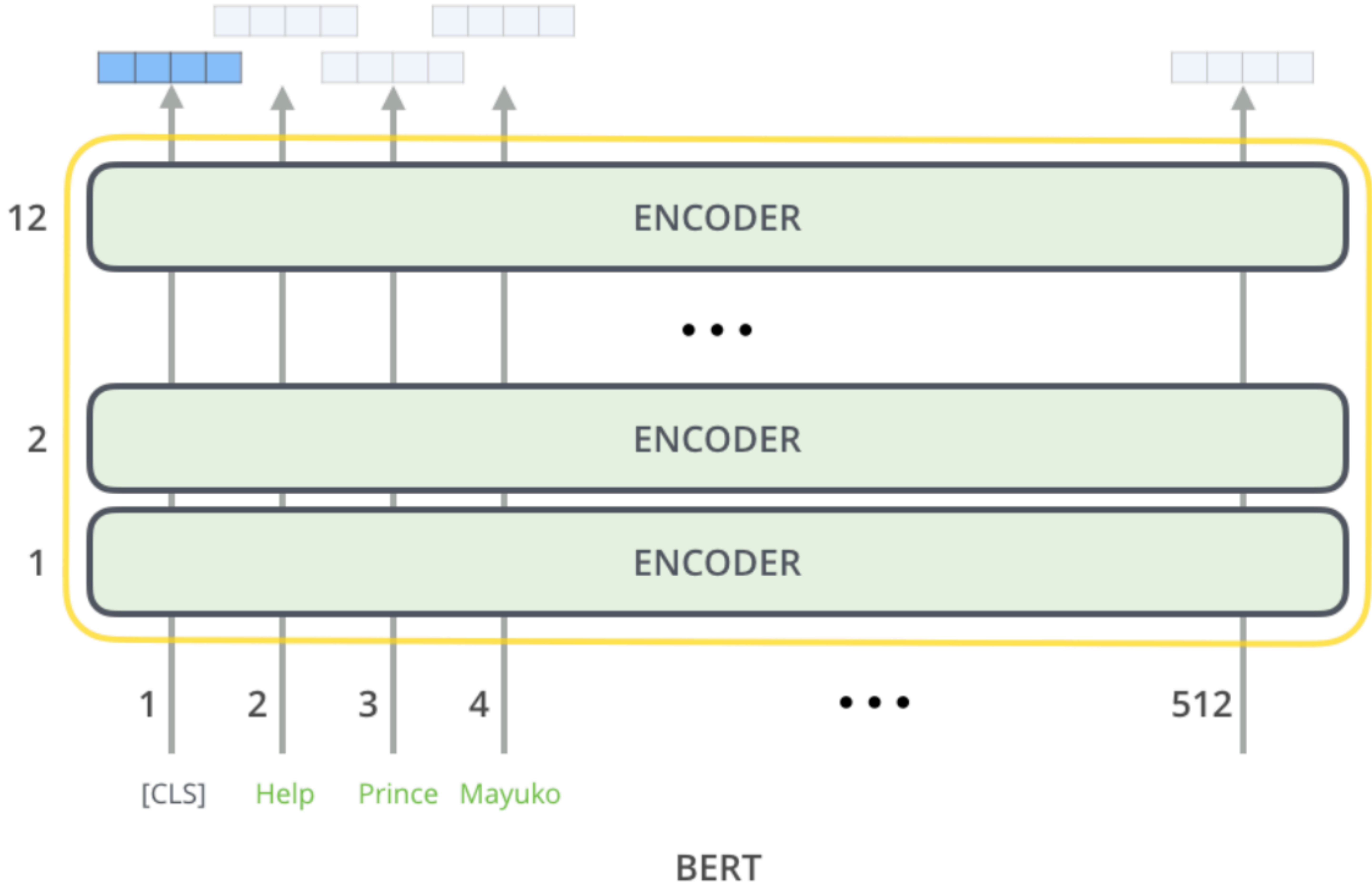
- BERT was developed in 2018 by Jacob Devlin from Google, and is currently used in almost every English search query in Google.

- BERT learns contextualized word embeddings.

- Also, BERT learns sentence representations which can be used for sentence classification (e.g. sentiment classification, grammar correctness etc).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

BERT model

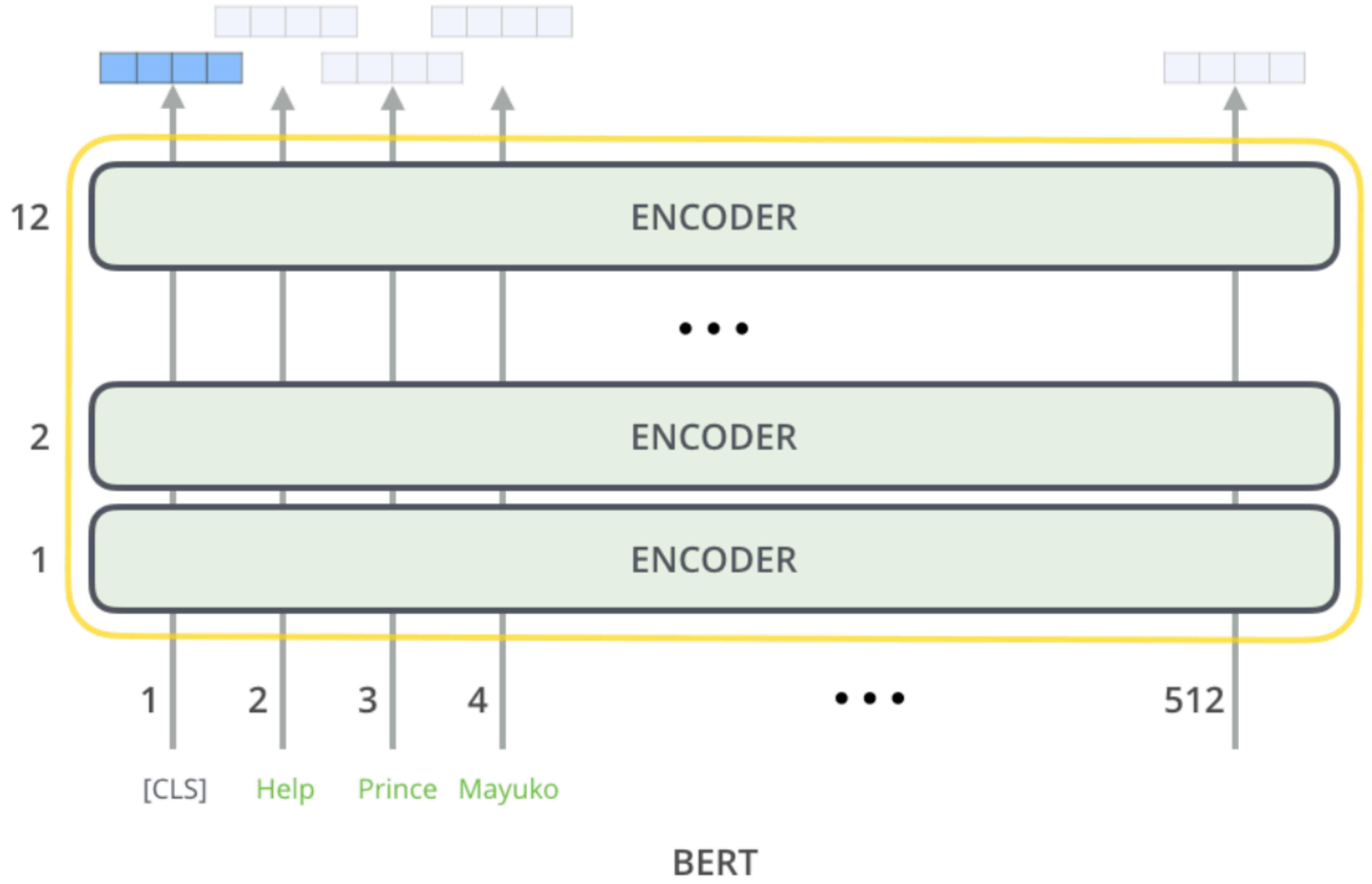


https://jalammar.github.io/illustrated-bert/

BERT model

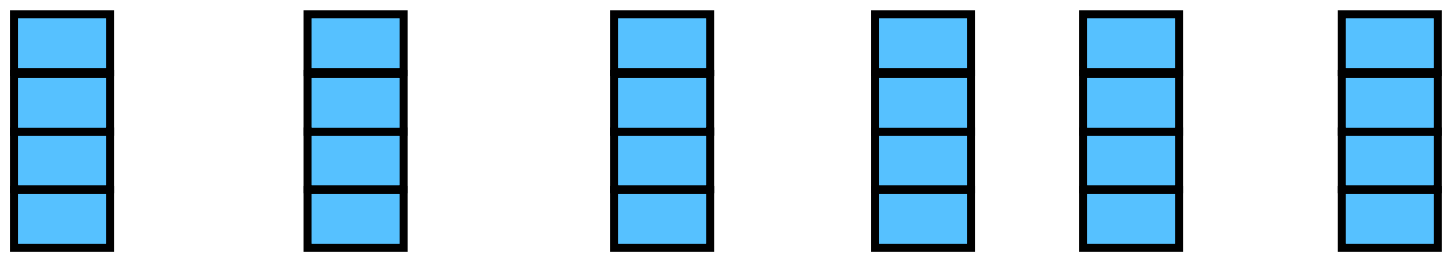
BERT was pretrained on two tasks:

1. **Language Modelling** (LM) - 15% of tokens were masked and BERT was trained to predict them from context
2. **Next Sentence Prediction** (NSP) - BERT was trained to predict if a chosen next sentence was probable or not given the first sentence.



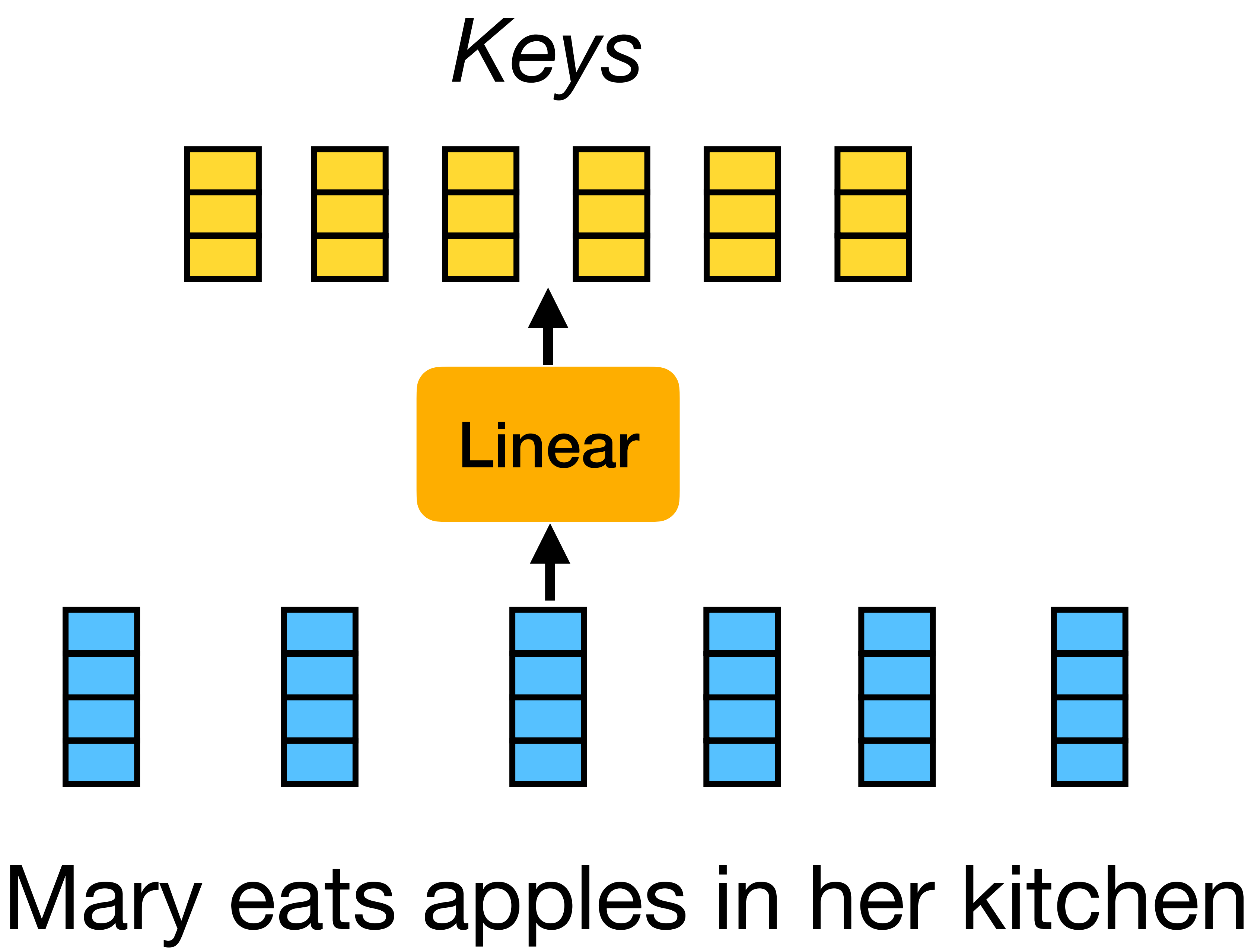
https://jalammar.github.io/illustrated-bert/

Attention mechanism - intuition

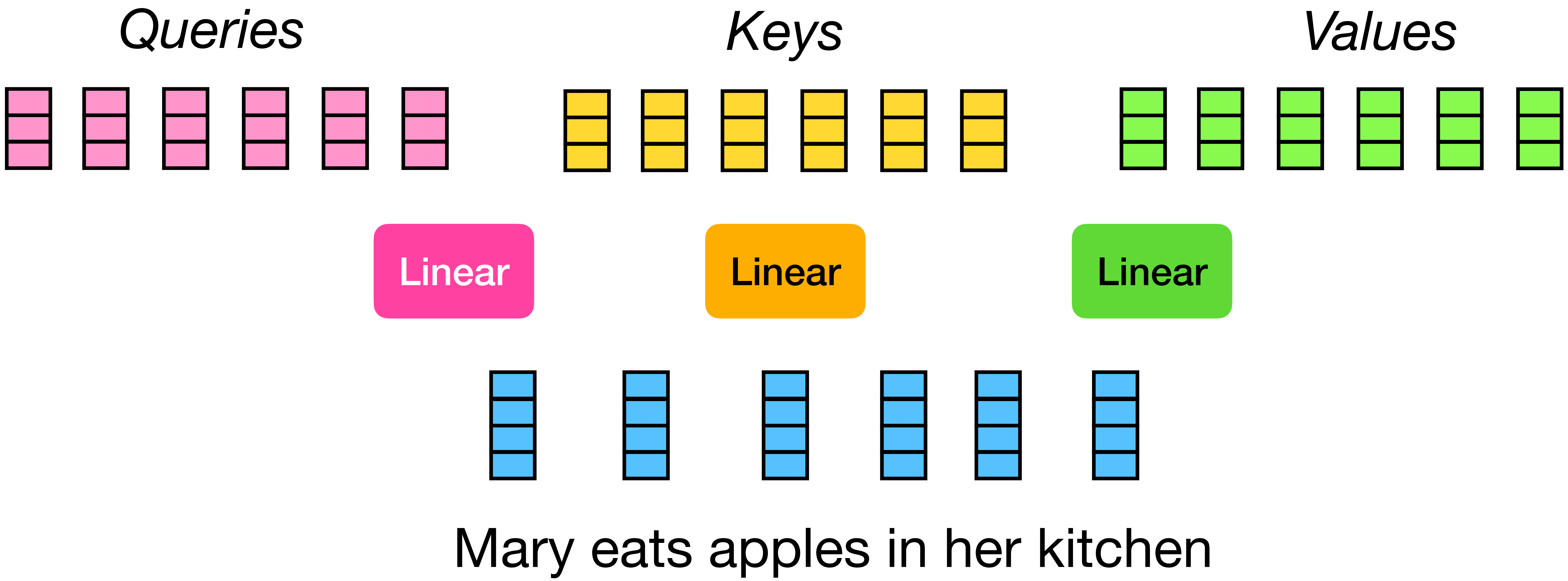


Mary eats apples in her kitchen

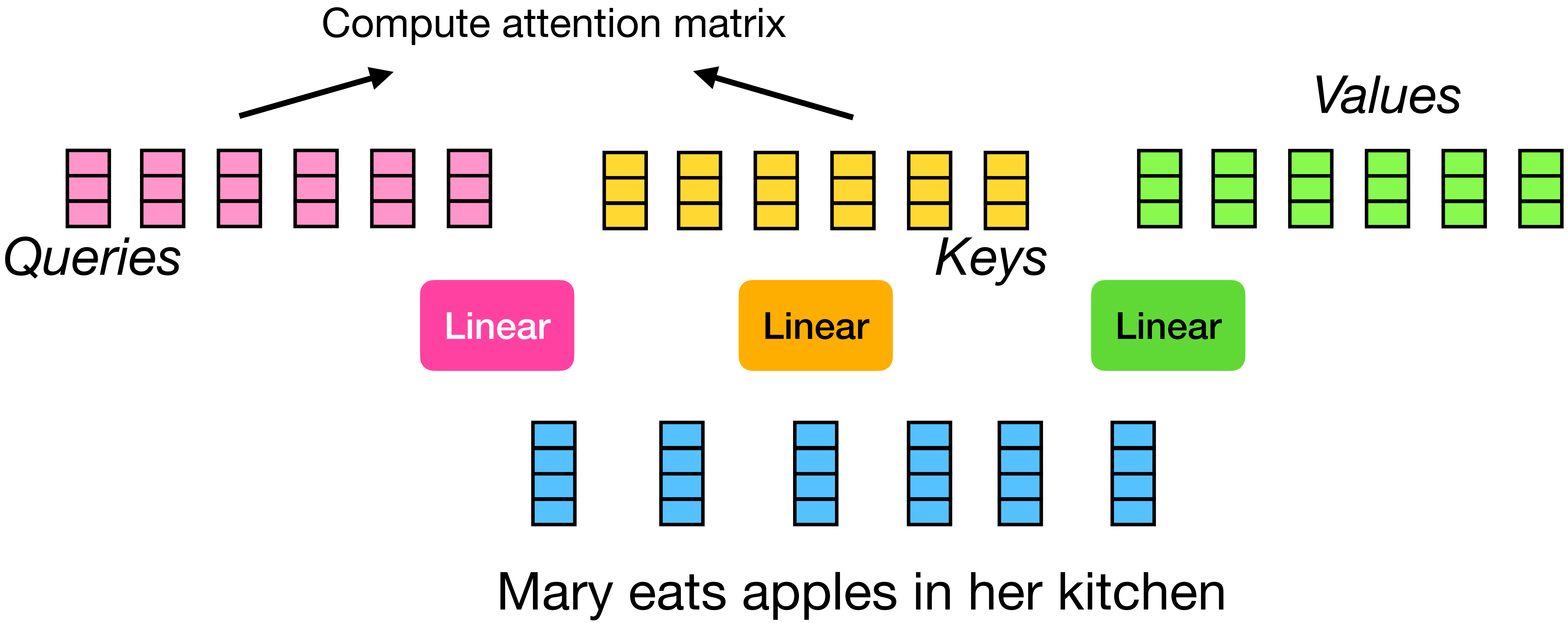
Attention mechanism - intuition



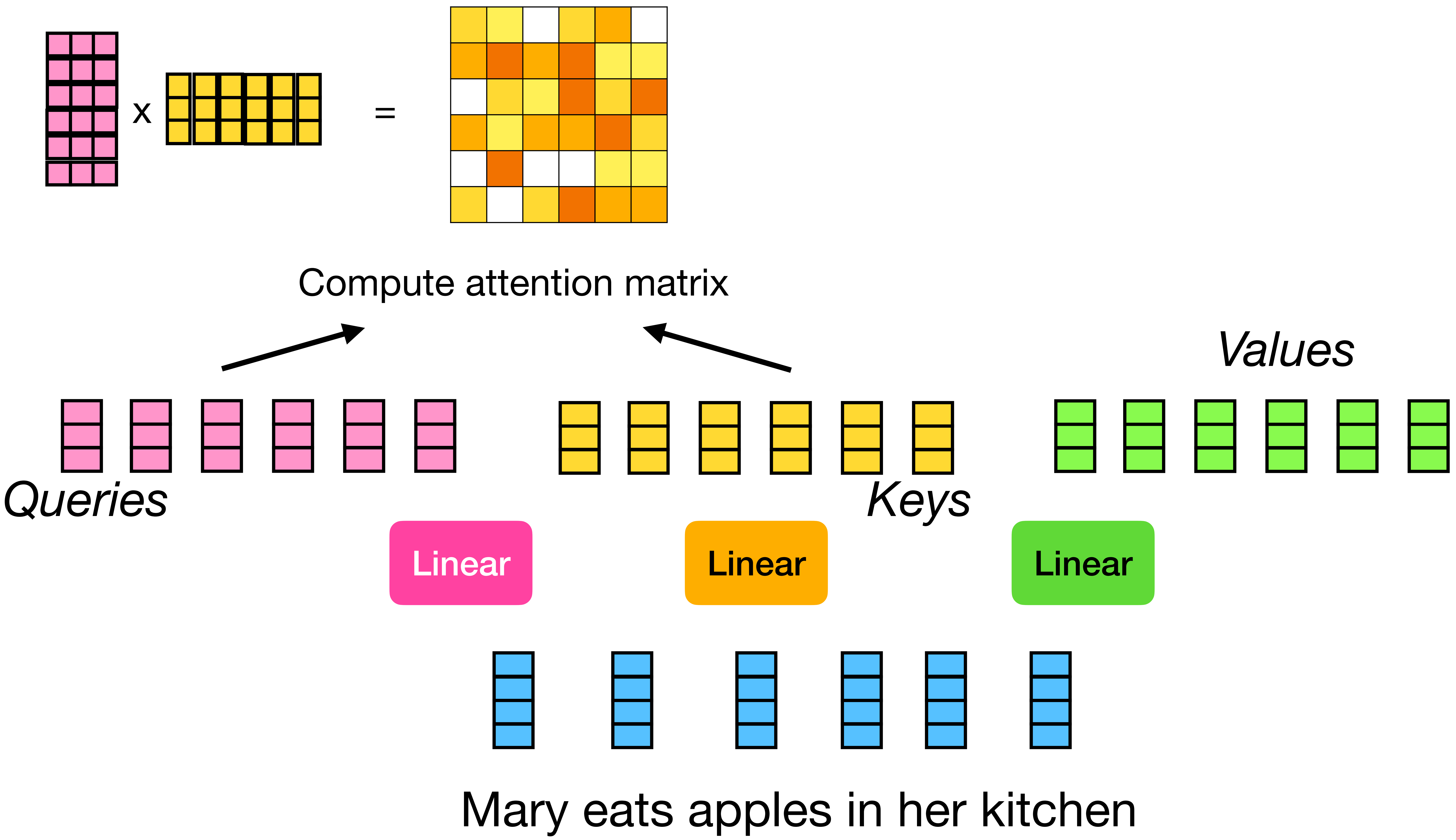
Attention mechanism - intuition



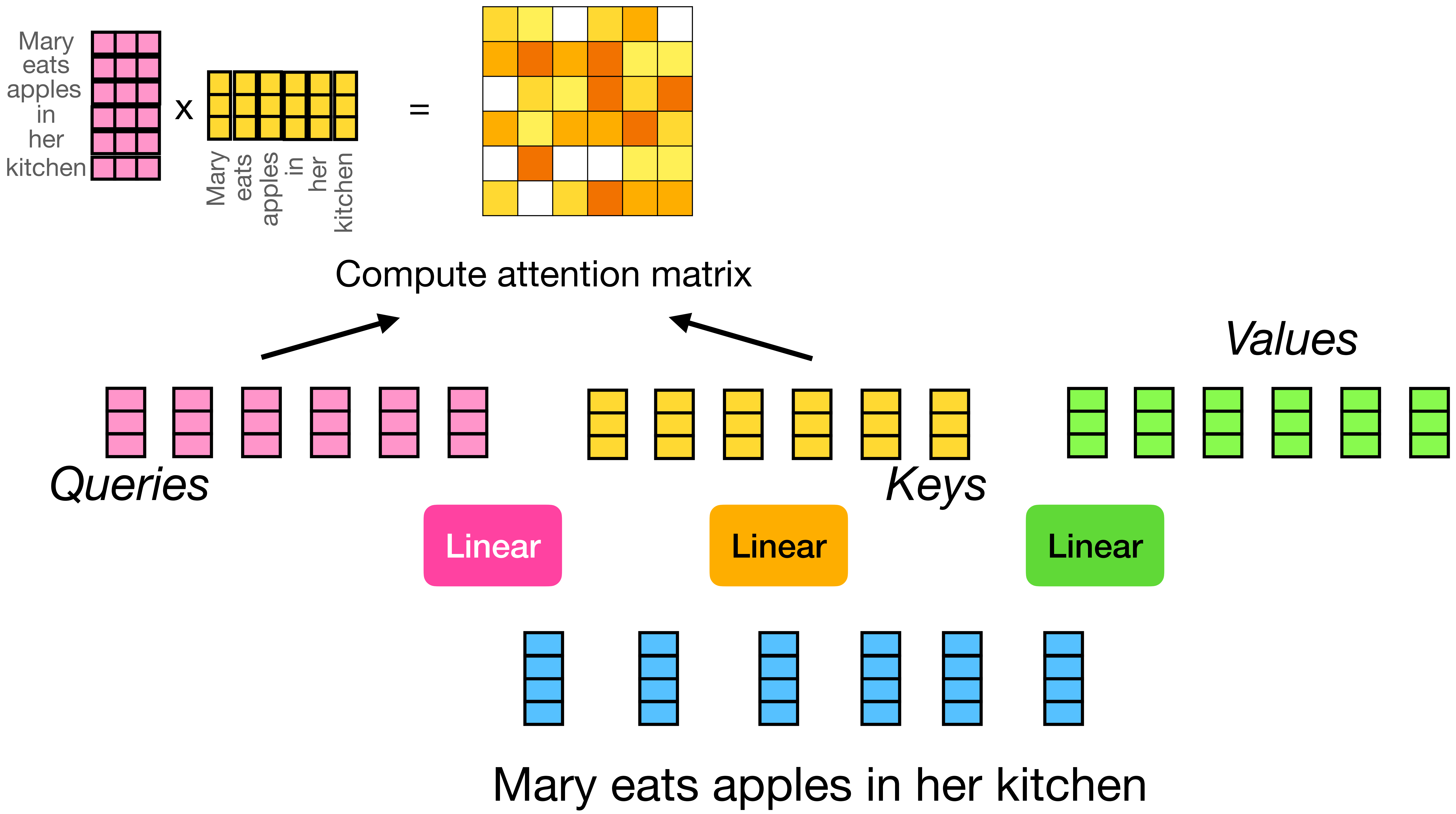
Attention mechanism - intuition



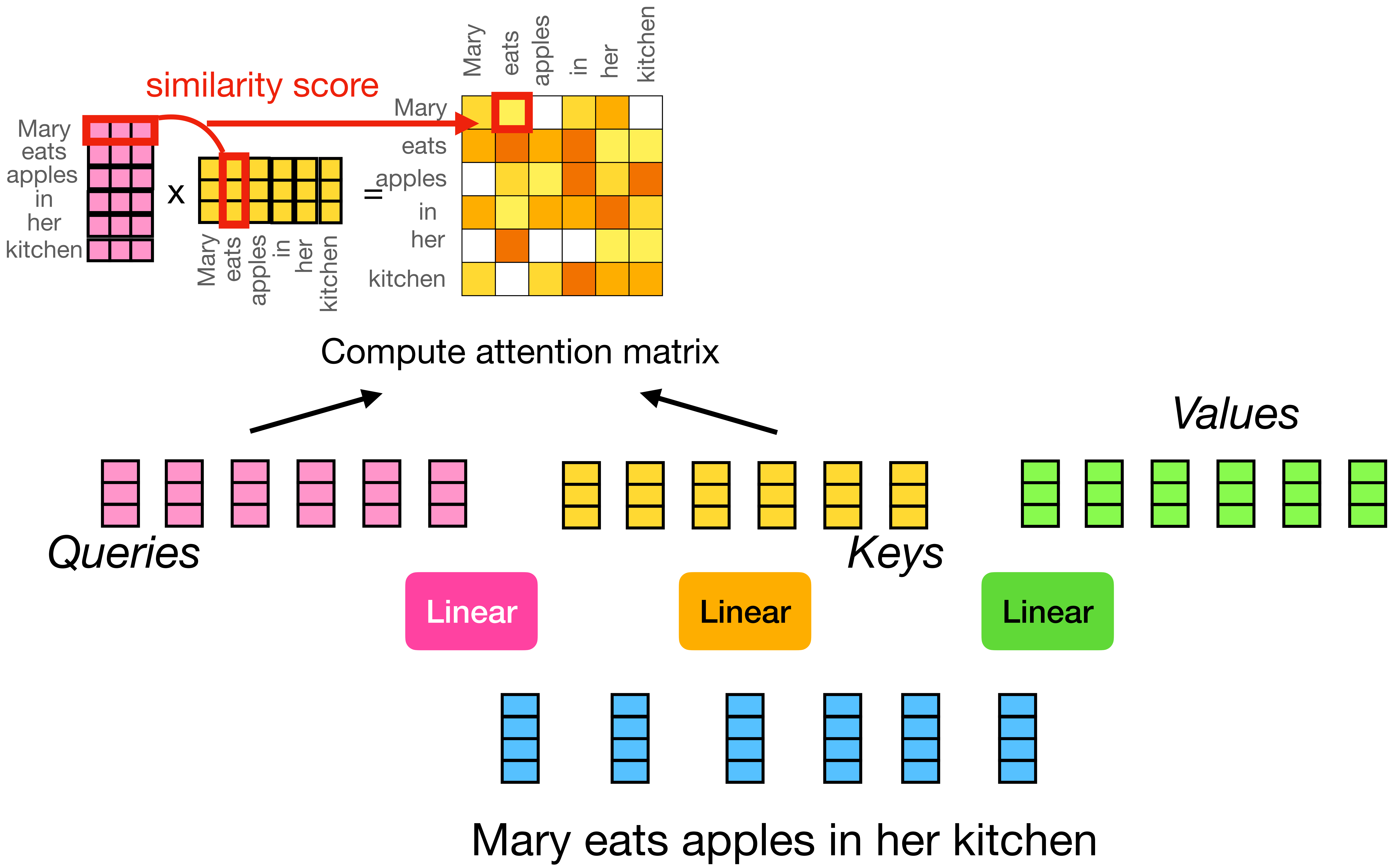
Attention mechanism - intuition



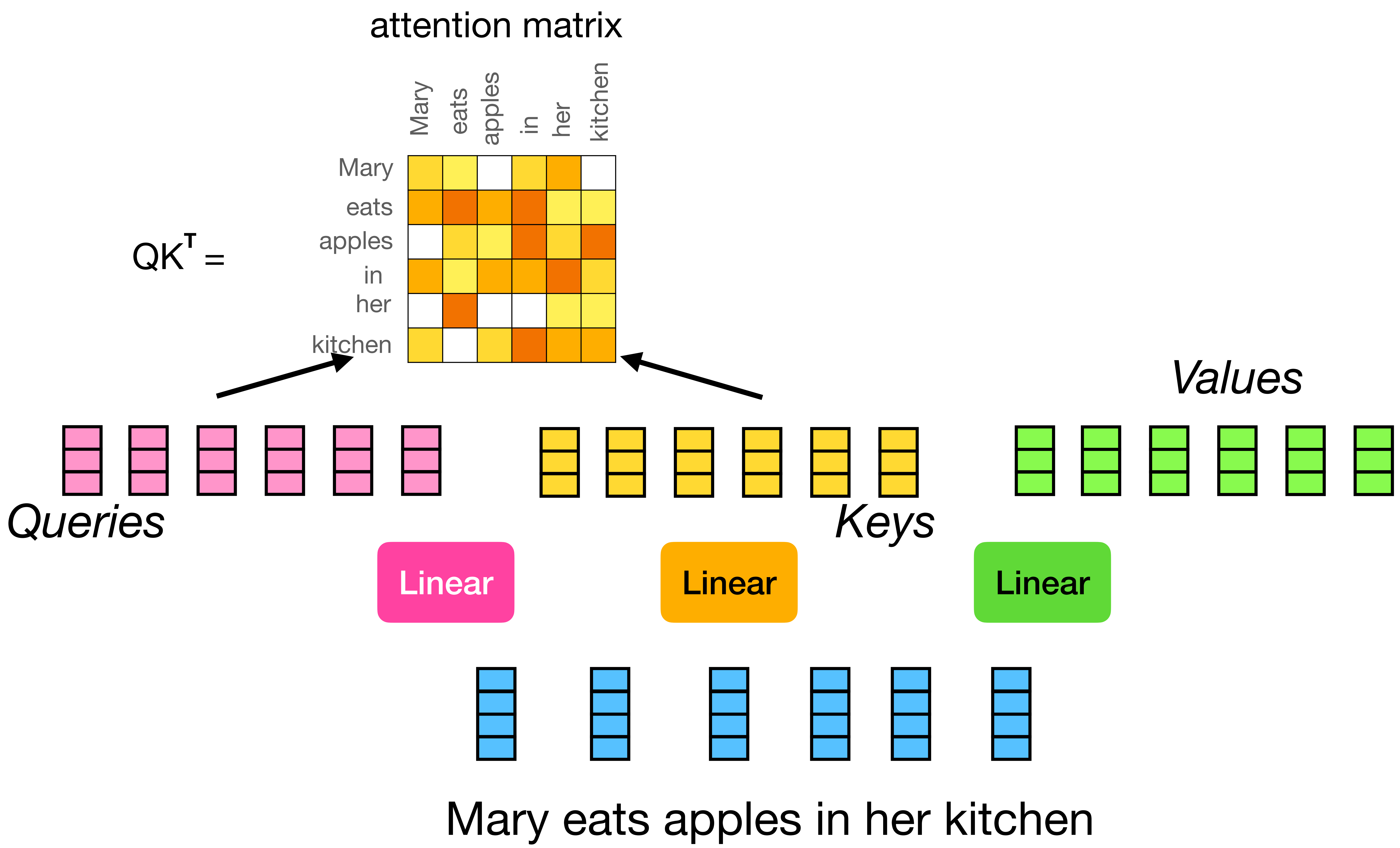
Attention mechanism - intuition



Attention mechanism - intuition

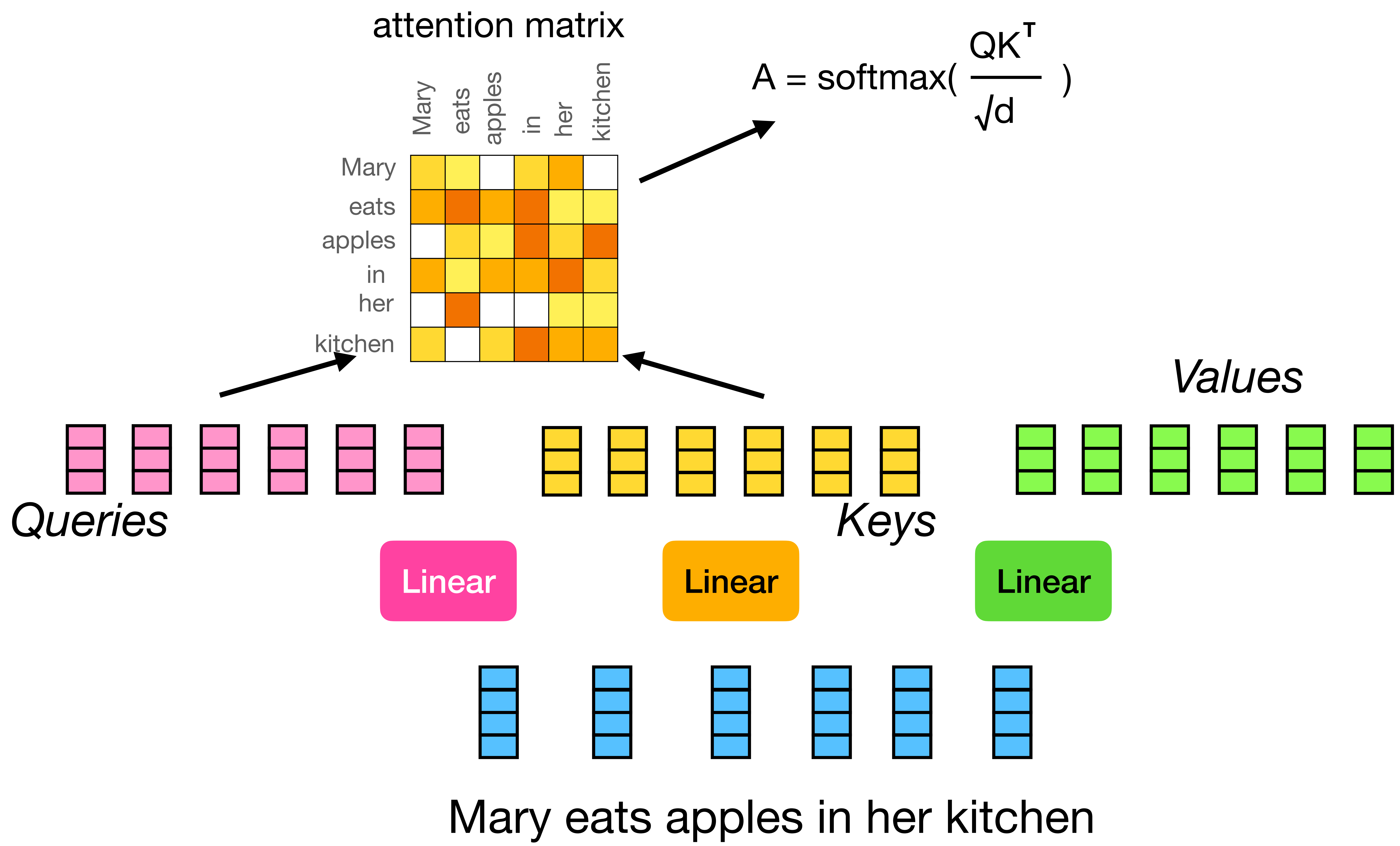


Attention mechanism - intuition

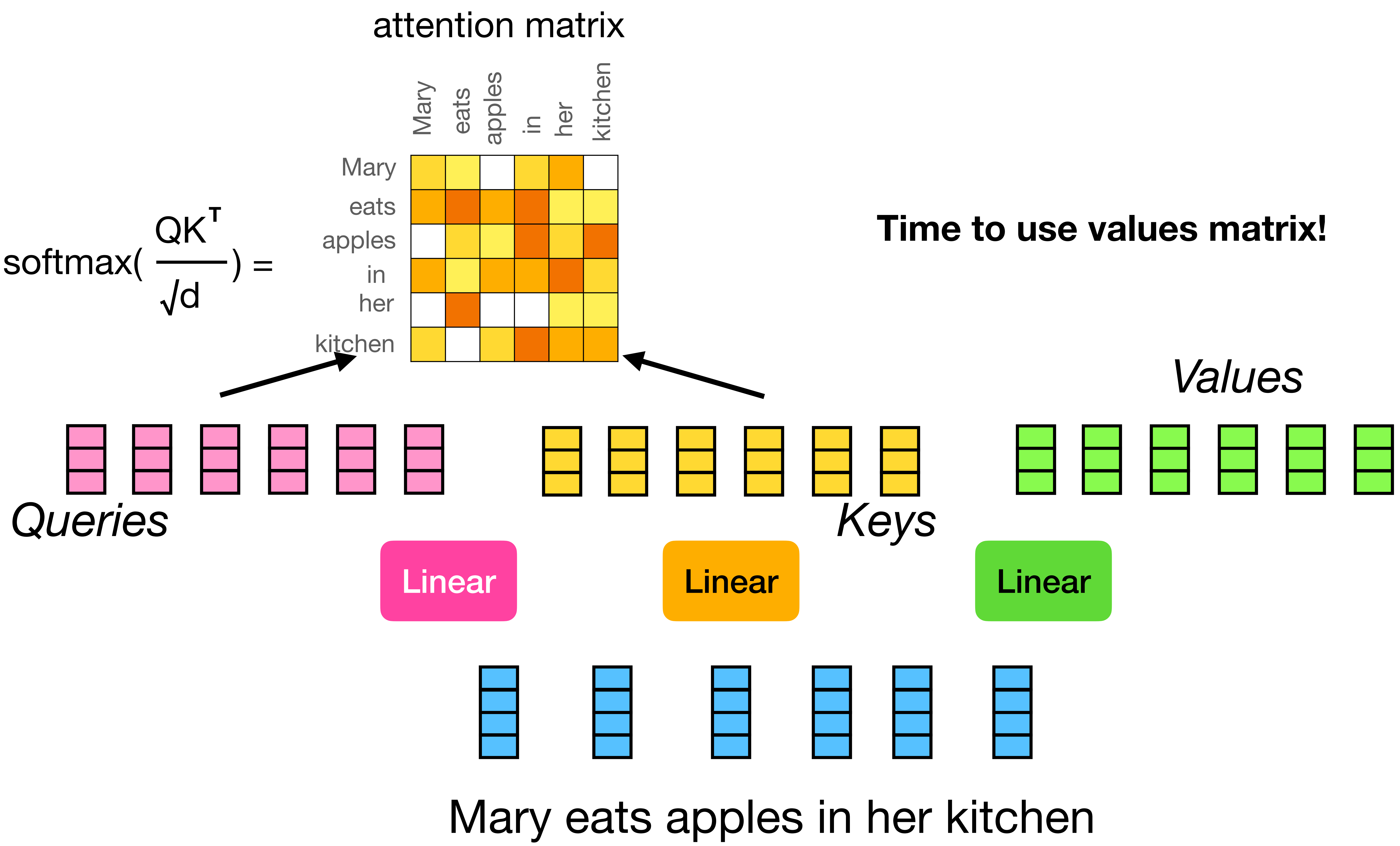


Attention mechanism - intuition

$$A = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$$



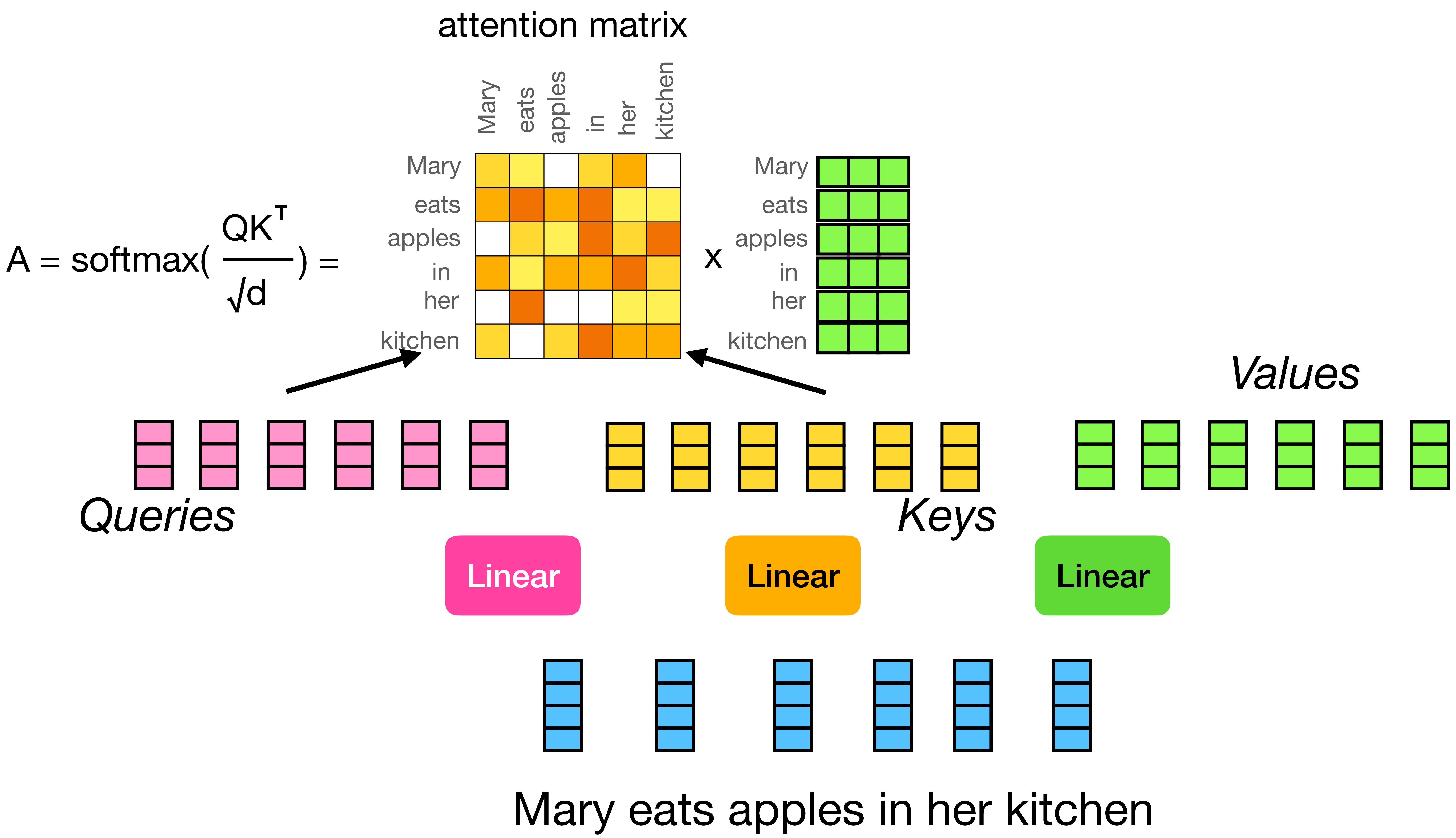
Attention mechanism - intuition



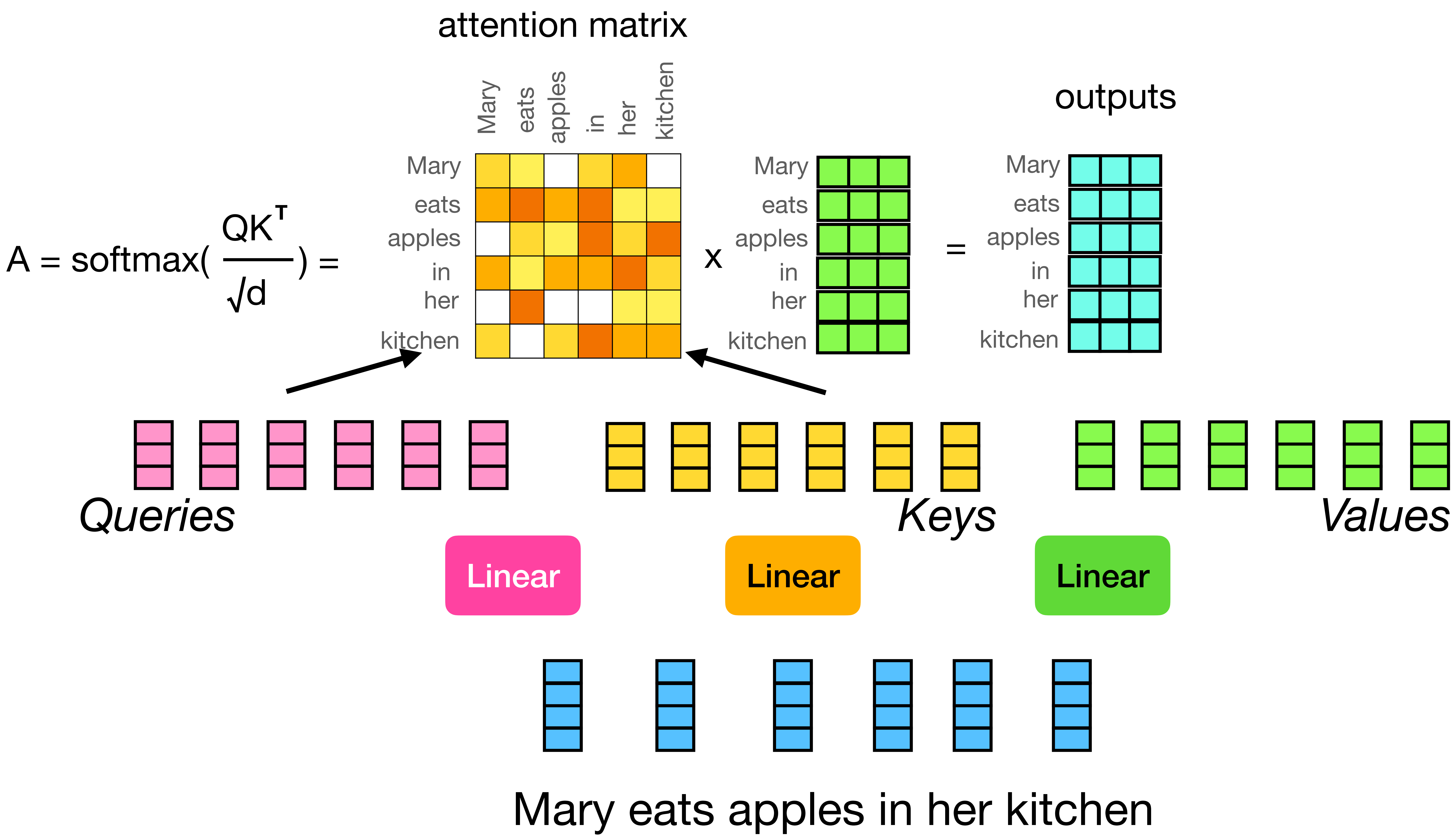
$$A = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) =$$

Time to use values matrix!

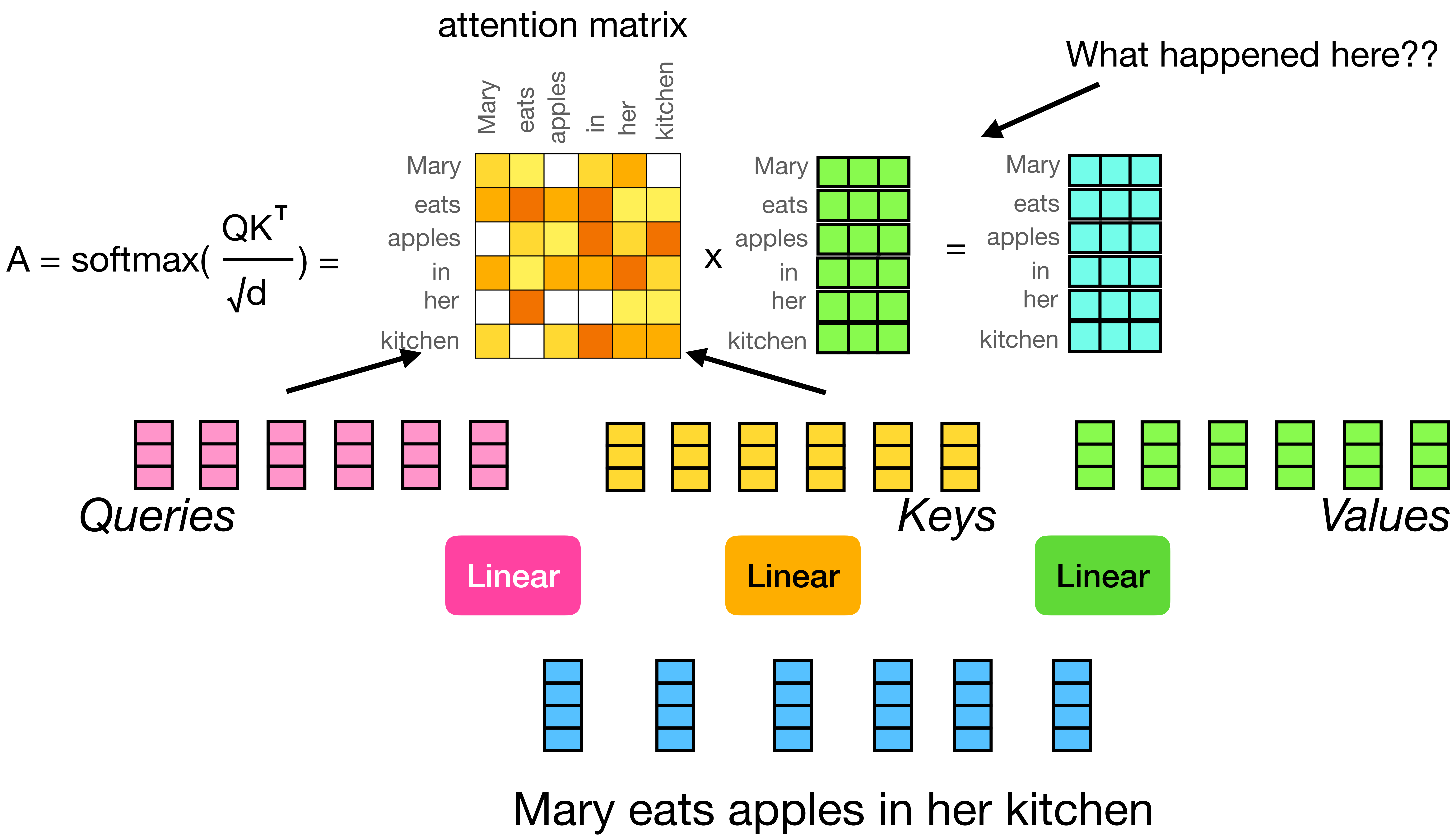
Attention mechanism - intuition



Attention mechanism - intuition

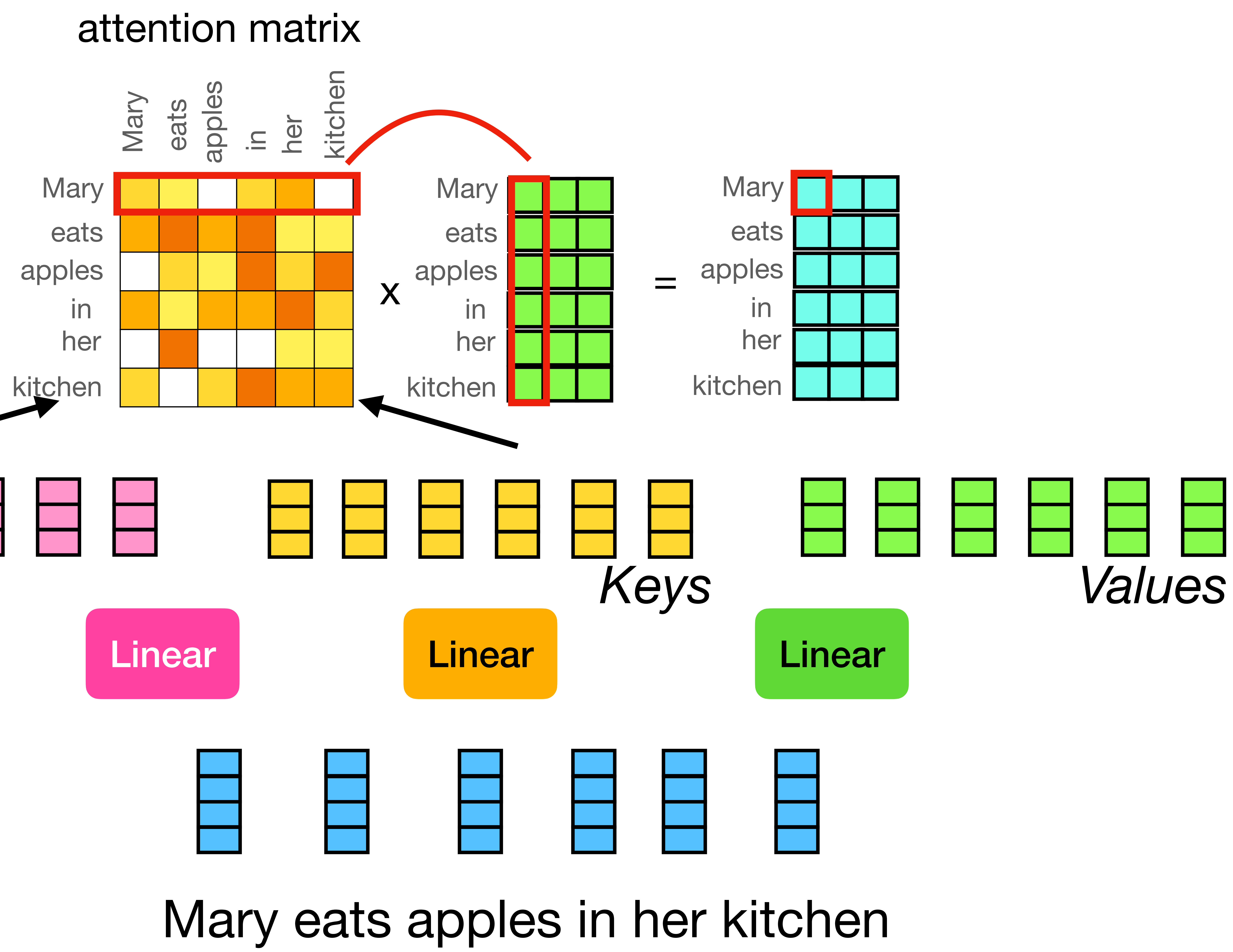


Attention mechanism - intuition

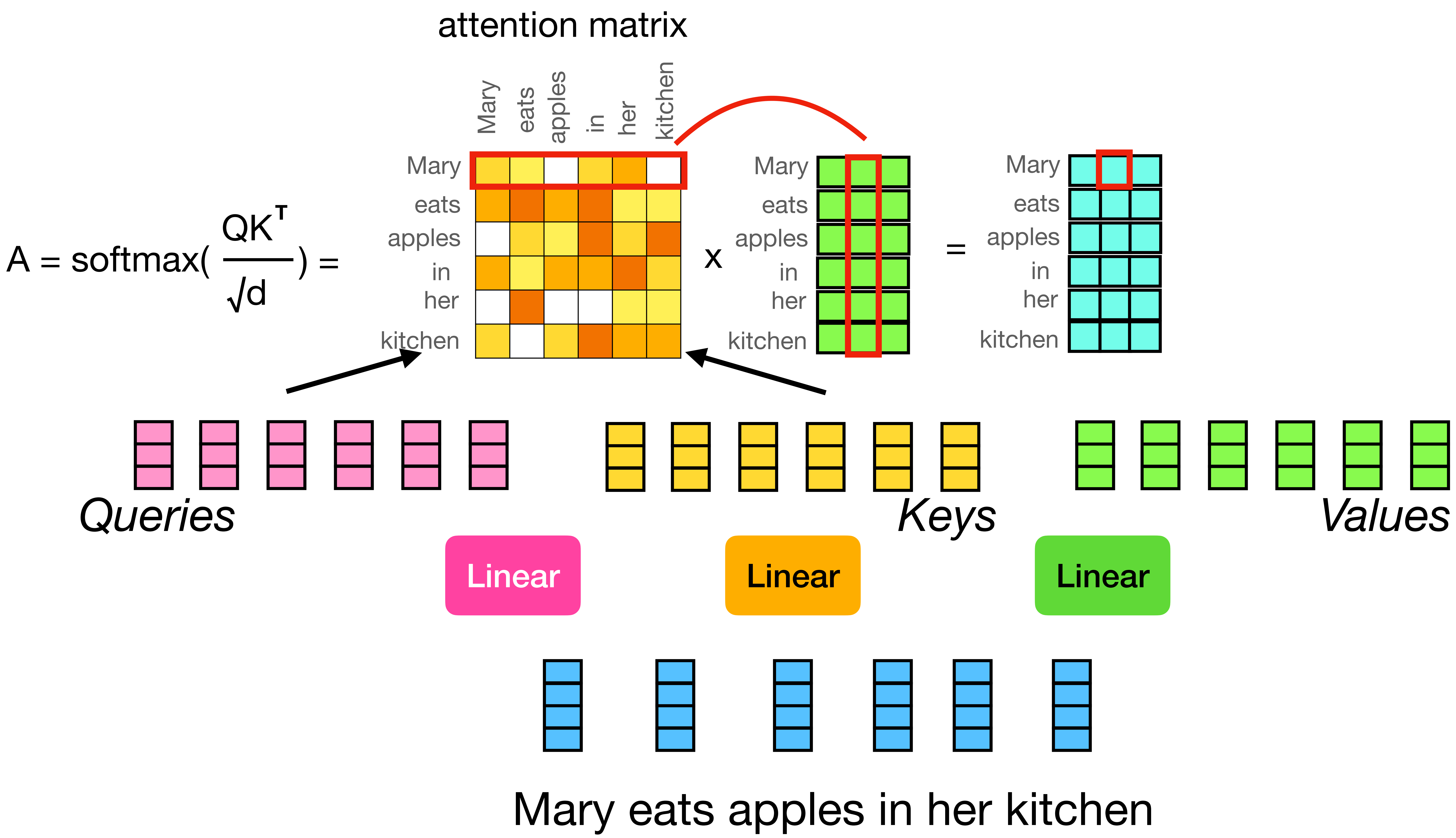


Attention mechanism - intuition

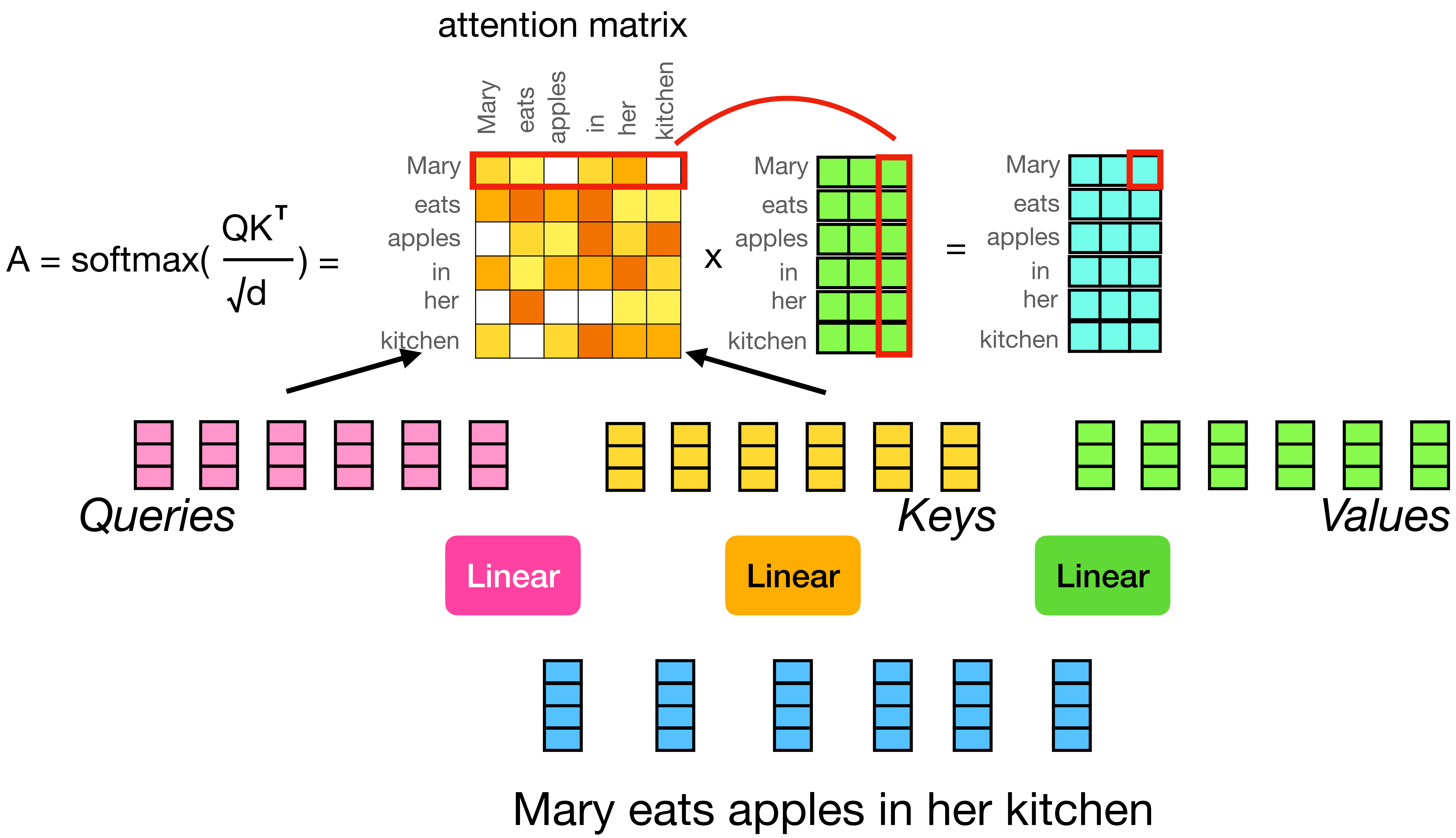
$$A = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) =$$



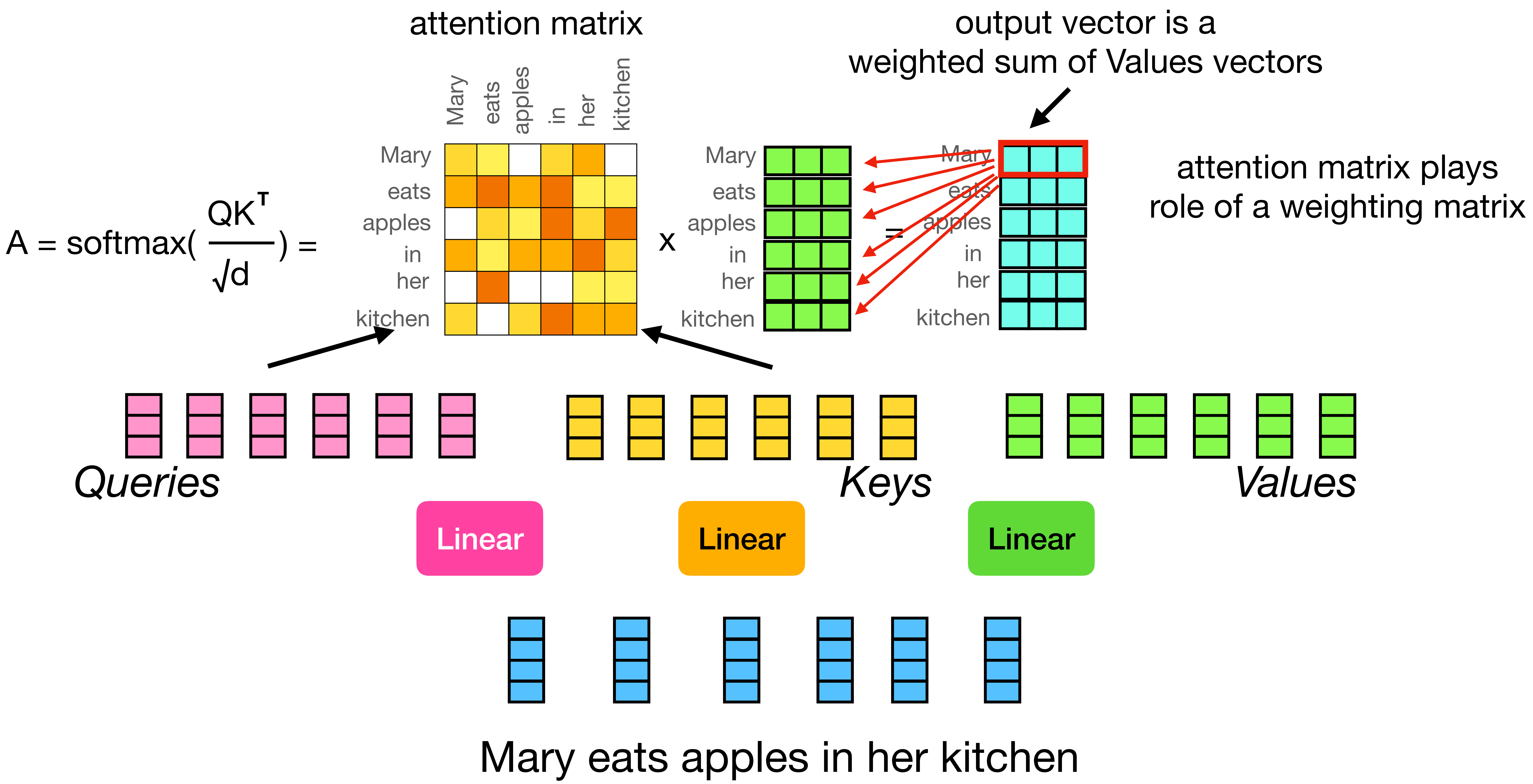
Attention mechanism - intuition



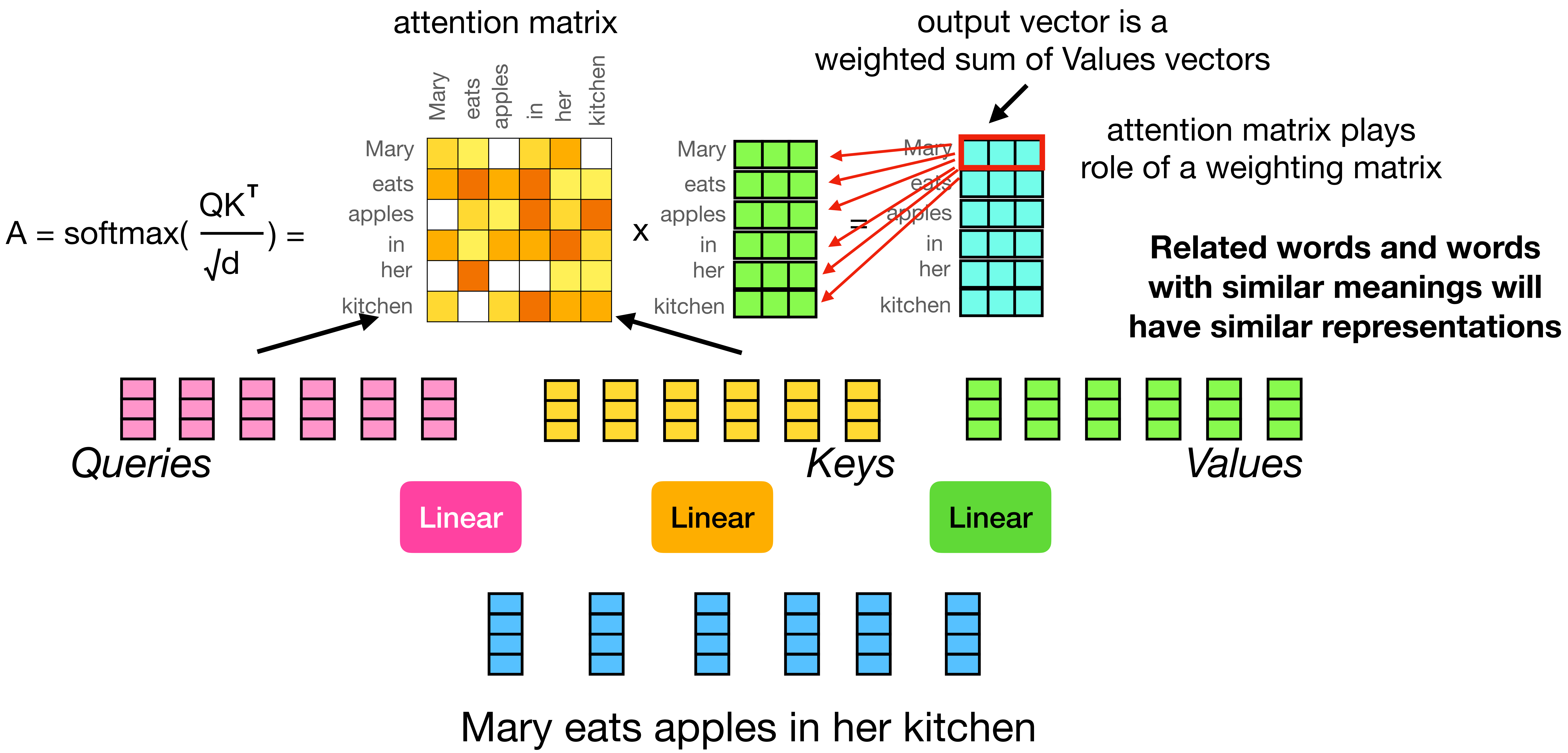
Attention mechanism - intuition



Attention mechanism - intuition



Attention mechanism - intuition



Related words and words with similar meanings will have similar representations

Summing up

- **Attention mechanism** allows transformer models to capture word-to-word relationships and get **contextualized embeddings**
- Modern NLP models are trained in a 2-step process - pre-training then fine-tuning
- BERT model learns context from both directions and outputs high-quality word and sentence representations

https://jalammar.github.io/illustrated-bert/

Tutorial roadmap

1. Transformer models & attention mechanism overview
2. **Bias in pre-trained language models**
3. How can we measure bias?
4. Ways to mitigate bias
5. Assignment overview

Bias in pre-trained language models

- NLP models are pre-trained on the real-world data (wikipedia, books, reddit etc.)
- Our accumulated text data contains our inherent bias (gender / race / social groups bias), hence models learn it
- If we continue to blindly rely on pre-trained models, without accounting for their unfairness, we will accumulate the bias
- Examples with NLP models:
    - Automated CV parsing (gender and racial bias)
    - Biomedical text analysis (preference to a commonly prescribed drug)

Bias in pre-trained language models

- “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.

Bias in pre-trained language models

- "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings"
- Paper used GloVe word embeddings (these embeddings are obtained from word co-occurrence in the large text corpus)
- Embeddings pintpoint sexism present in the training data, for instance:

$$\overrightarrow{\mathrm{man}} - \overrightarrow{\mathrm{woman}} \approx \overrightarrow{\mathrm{computer\ programmer}} - \overrightarrow{\mathrm{homemaker}}.$$

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.

Bias in pre-trained language models

- “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”

Extreme *she*	**Extreme *he***
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

	Gender stereotype *she-he* analogies	
sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant
	Gender appropriate *she-he* analogies	
queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Figure 1: **Left** The most extreme occupations as projected on to the *she−he* gender direction on w2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded. **Right** Automatically generated analogies for the pair *she-he* using the procedure described in text. Each automatically generated analogy is evaluated by 10 crowd-workers to whether or not it reflects gender stereotype.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.

Bias in BERT - countries

```
text = "Australia has a lot of [MASK]."
```

Bias in BERT - countries

```
text = "Australia has a lot of [MASK]."
```

```
1.87 % Australia has a lot of people.
1.6 % Australia has a lot of talent.
1.29 % Australia has a lot of resources.
1.28 % Australia has a lot of diversity.
1.14 % Australia has a lot of history.
1.11 % Australia has a lot of children.
1.09 % Australia has a lot of problems.
0.9 % Australia has a lot of technology.
0.81 % Australia has a lot of money.
0.77 % Australia has a lot of wealth.
```

Bias in BERT - countries

```
text = "Australia has a lot of [MASK]."
```

```
1.87 % Australia has a lot of people.
1.6 % Australia has a lot of talent.
1.29 % Australia has a lot of resources.
1.28 % Australia has a lot of diversity.
1.14 % Australia has a lot of history.
1.11 % Australia has a lot of children.
1.09 % Australia has a lot of problems.
0.9 % Australia has a lot of technology.
0.81 % Australia has a lot of money.
0.77 % Australia has a lot of wealth.
```

Bias in BERT - countries

```
text = "Australia has a lot of [MASK]."
```

```
1.87 % Australia has a lot of people.
1.6 % Australia has a lot of talent.
1.29 % Australia has a lot of resources.
1.28 % Australia has a lot of diversity.
1.14 % Australia has a lot of history.
1.11 % Australia has a lot of children.
1.09 % Australia has a lot of problems.
0.9 % Australia has a lot of technology.
0.81 % Australia has a lot of money.
0.77 % Australia has a lot of wealth.
```

```
text = "Thailand has a lot of [MASK]."
```

```
1.57 % Thailand has a lot of tourists.
1.34 % Thailand has a lot of people.
1.32 % Thailand has a lot of tourism.
0.98 % Thailand has a lot of schools.
0.94 % Thailand has a lot of problems.
0.93 % Thailand has a lot of students.
0.92 % Thailand has a lot of celebrities.
0.92 % Thailand has a lot of talent.
0.81 % Thailand has a lot of universities.
0.73 % Thailand has a lot of politicians.
```

Bias in BERT - countries

```
text = "Australia has a lot of [MASK]."
```

```
1.87 % Australia has a lot of people.
1.6 % Australia has a lot of talent.
1.29 % Australia has a lot of resources.
1.28 % Australia has a lot of diversity.
1.14 % Australia has a lot of history.
1.11 % Australia has a lot of children.
1.09 % Australia has a lot of problems.
0.9 % Australia has a lot of technology.
0.81 % Australia has a lot of money.
0.77 % Australia has a lot of wealth.
```

```
text = "Thailand has a lot of [MASK]."
```

```
1.57 % Thailand has a lot of tourists.
1.34 % Thailand has a lot of people.
1.32 % Thailand has a lot of tourism.
0.98 % Thailand has a lot of schools.
0.94 % Thailand has a lot of problems.
0.93 % Thailand has a lot of students.
0.92 % Thailand has a lot of celebrities.
0.92 % Thailand has a lot of talent.
0.81 % Thailand has a lot of universities.
0.73 % Thailand has a lot of politicians.
```

```
text = "Israel has a lot of [MASK]."
```

```
10.37 % Israel has a lot of Jews.
1.98 % Israel has a lot of problems.
1.96 % Israel has a lot of children.
1.85 % Israel has a lot of people.
1.64 % Israel has a lot of resources.
1.55 % Israel has a lot of immigrants
1.47 % Israel has a lot of money.
1.16 % Israel has a lot of refugees.
1.09 % Israel has a lot of enemies.
0.96 % Israel has a lot of Muslims.
```

```
text = "Korea has a lot of [MASK]."
```

```
1.7 % Korea has a lot of technology.
1.46 % Korea has a lot of talent.
1.41 % Korea has a lot of people.
1.25 % Korea has a lot of problems.
1.14 % Korea has a lot of history.
1.12 % Korea has a lot of wealth.
1.05 % Korea has a lot of resources.
0.93 % Korea has a lot of electricity.
0.86 % Korea has a lot of tourism.
0.85 % Korea has a lot of money.
```

Bias in BERT - gender

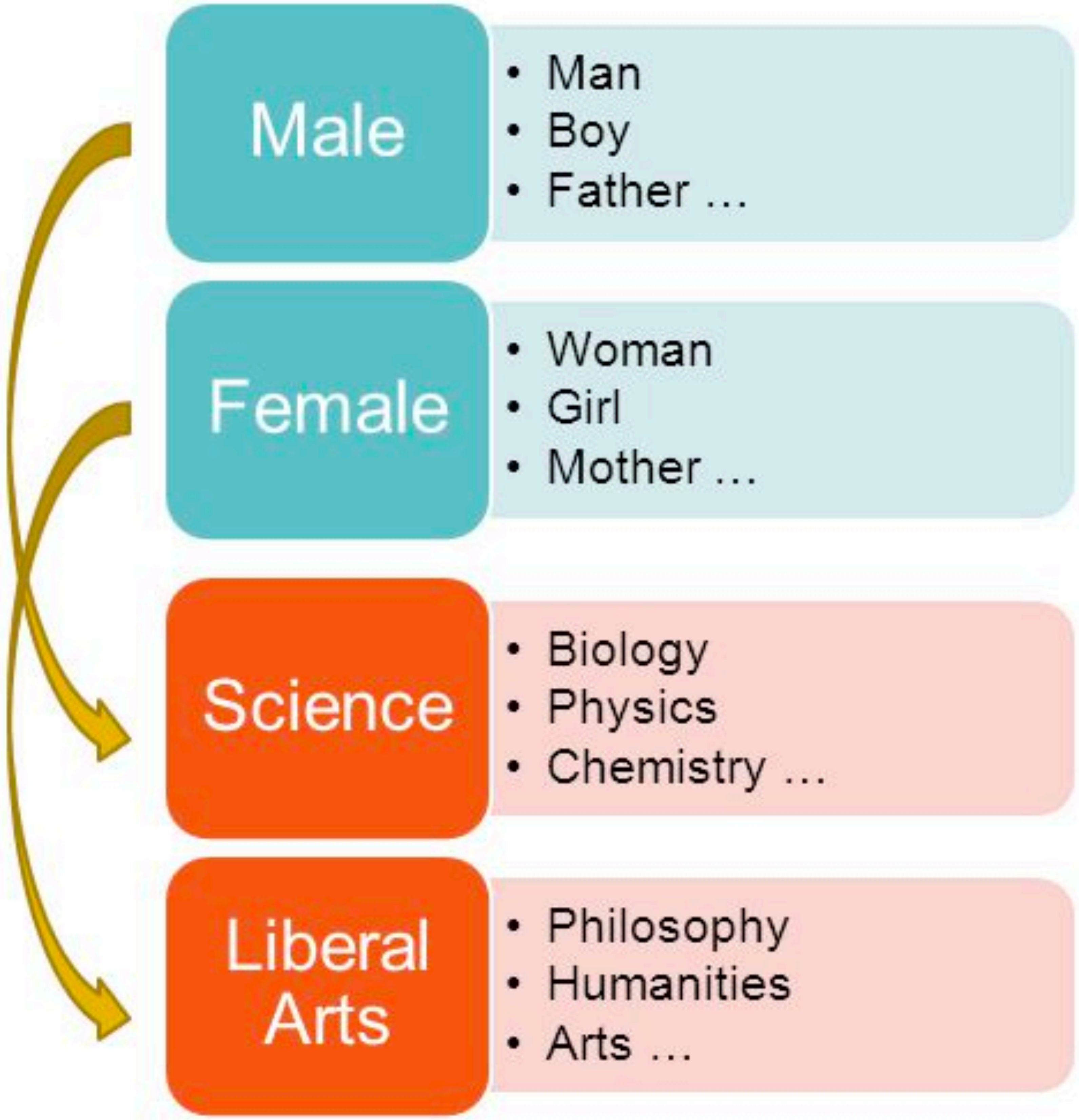
- We ask BERT to complete sentences “MASK works as a doctor” and “MASK works as a nurse”
- BERT replaces MASK token with “he” for the 1st sentence, and “she” for the 2nd sentence
- We can compute BERT’s probabilities of sentences with different professions to investigate the bias

Tutorial roadmap

1. Transformer models & attention mechanism overview
2. Bias in pre-trained language models
3. **How can we measure bias?**
4. Ways to mitigate bias
5. Assignment overview

How to measure bias?

- Adopting psychological tests…
- **Implicit Association test** (IAT) helps to measure psychological bias
- IAT requires users to rapidly categorize two target concepts with an attribute (e.g. the concepts "male" and "female" with the attribute "logical")
- Easier pairings (faster responses) = strongly associated in memory, while difficult pairings (slower responses) = less associated.



Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464.

Word Embedding Association Test

- Caliskan et. al. propose **WEAT** (Word Embedding Association Test)
- WEAT measures associations between two sets of target words X, Y (e.g. male and female) with two sets of attributes A, B (e.g. career and family).
- Original WEAT uses Word2Vec or GloVe word embeddings

$$s(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}) = \sum_{x \in \mathcal{X}} s(x, \mathcal{A}, \mathcal{B}) - \sum_{y \in \mathcal{Y}} s(y, \mathcal{A}, \mathcal{B})$$

$$s(x, \mathcal{A}, \mathcal{B}) = \operatorname*{mean}_{a \in \mathcal{A}} \cos(x, a) - \operatorname*{mean}_{b \in \mathcal{B}} \cos(x, b)$$

Word Embedding Association Test

- **WEAT** (Word Embedding Association Test):

$$s(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}) = \sum_{x \in \mathcal{X}} s(x, \mathcal{A}, \mathcal{B}) - \sum_{y \in \mathcal{Y}} s(y, \mathcal{A}, \mathcal{B})$$

$$s(x, \mathcal{A}, \mathcal{B}) = \operatorname{mean}_{a \in \mathcal{A}} \cos(x, a) - \operatorname{mean}_{b \in \mathcal{B}} \cos(x, b)$$

X = [man, boy, male]   A = [career, money, progress, success, job]

Y = [woman, girl, female]   B = [family, love, care, children, home]

Word Embedding Association Test

- **WEAT** (Word Embedding Association Test):

$$s(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}) = \sum_{x \in \mathcal{X}} s(x, \mathcal{A}, \mathcal{B}) - \sum_{y \in \mathcal{Y}} s(y, \mathcal{A}, \mathcal{B})$$

$$s(x, \mathcal{A}, \mathcal{B}) = \operatorname{mean}_{a \in \mathcal{A}} \cos(x, a) - \operatorname{mean}_{b \in \mathcal{B}} \cos(x, b)$$

X = [man, boy, male]   A = [career, money, progress, success, job]

Y = [woman, girl, female]   B = [family, love, care, children, home]

s(man, A, B) = 1/5 * [cos(man, career) + cos(man, money) + … + cos(man, job)] -
- 1/5 * [cos(man, family) + cos(man, love) + … + cos(man, home)]

Word Embedding Association Test

- **WEAT** (Word Embedding Association Test):

$$s(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}) = \sum_{x \in \mathcal{X}} s(x, \mathcal{A}, \mathcal{B}) - \sum_{y \in \mathcal{Y}} s(y, \mathcal{A}, \mathcal{B})$$

$$s(x, \mathcal{A}, \mathcal{B}) = \operatorname{mean}_{a \in \mathcal{A}} \cos(x, a) - \operatorname{mean}_{b \in \mathcal{B}} \cos(x, b)$$

X = [man, boy, male] A = [career, money, progress, success, job]

Y = [woman, girl, female] B = [family, love, care, children, home]

0.8 0.7 0.63

s(man, A, B) = 1/5 * [cos(man, career) + cos(man, money) + … + cos(man, job)] -

- 1/5 * [cos(man, family) + cos(man, love) + … + cos(man, home)]

0.11 0.3 0.02

Word Embedding Association Test

- **WEAT** (Word Embedding Association Test):

$$s(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}) = \sum_{x \in \mathcal{X}} s(x, \mathcal{A}, \mathcal{B}) - \sum_{y \in \mathcal{Y}} s(y, \mathcal{A}, \mathcal{B})$$

$$s(x, \mathcal{A}, \mathcal{B}) = \operatorname{mean}_{a \in \mathcal{A}} \cos(x, a) - \operatorname{mean}_{b \in \mathcal{B}} \cos(x, b)$$

X = [man, boy, male] A = [career, money, progress, success, job]

Y = [woman, girl, female] B = [family, love, care, children, home]

$$s(\text{man}, A, B) = 1/5 * [\overset{0.8}{\cos(\text{man}, \text{career})} + \overset{0.7}{\cos(\text{man}, \text{money})} + \ldots + \overset{0.63}{\cos(\text{man}, \text{job})}] - 1/5 * [\underset{0.11}{\cos(\text{man}, \text{family})} + \underset{0.2}{\cos(\text{man}, \text{love})} + \ldots + \underset{0.02}{\cos(\text{man}, \text{home})}]$$

0.57

Word Embedding Association Test

- **WEAT** (Word Embedding Association Test):

$$s(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}) = \sum_{x \in \mathcal{X}} s(x, \mathcal{A}, \mathcal{B}) - \sum_{y \in \mathcal{Y}} s(y, \mathcal{A}, \mathcal{B})$$

$$s(x, \mathcal{A}, \mathcal{B}) = \operatorname{mean}_{a \in \mathcal{A}} \cos(x, a) - \operatorname{mean}_{b \in \mathcal{B}} \cos(x, b)$$

X = [man, boy, male] A = [career, money, progress, success, job]

Y = [woman, girl, female] B = [family, love, care, children, home]

s(man, A, B) = 1/5 * [cos(man, career) + cos(man, money) + … + cos(man, job)] -
- 1/5 * [cos(man, family) + cos(man, love) + … + cos(man, home)]

S(X, Y, A, B) = [s(man, A, B) + s(boy, A, B) + s(male, A, B)] -
- [s(woman, A, B) + s(girl, A, B) + s(female, A, B)]

Sentence Encoder Association Test

- SEAT (Sentence Encoder Association Test) - uses sentence representations (for example, BERT representations).
- SEAT computes WEAT for sentences.

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Problems with WEAT / SEAT

- Issue behind WEAT and SEAT tests is that embeddings similarity is related to words co-occurence
- Hence, commonly used words can be more “related” just by chance.

Target Word Sets	Attribute Word Sets	Test Statistic	Effect Size	p-value	Outcome (WEAT)
{door} vs. {curtain}	{masculine} vs. {feminine}	0.021	2.0	0.0	more male-associated
	{girlish} vs. {boyish}	−0.042	−2.0	0.5	inconclusive
	{woman} vs. {man}	0.071	2.0	0.0	more female-associated

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). Understanding undesirable word embedding associations. *arXiv preprint arXiv:1908.06361*.

Problems with WEAT / SEAT

- Issue behind WEAT and SEAT tests is that <u>embeddings similarity is related to words co-occurence</u>
- Hence, commonly used words can be more “related” just by chance.

Target Word Sets	Attribute Word Sets	Test Statistic	Effect Size	p-value	Outcome (WEAT)
{door} vs. {curtain}	{masculine} vs. {feminine}	0.021	2.0	0.0	more male-associated
	{girlish} vs. {boyish}	−0.042	−2.0	0.5	inconclusive
	{woman} vs. {man}	0.071	2.0	0.0	more female-associated
{dog} vs. {cat}	{masculine} vs. {feminine}	0.063	2.0	0.0	more male-associated
	{actress} vs. {actor}	−0.075	−2.0	0.5	inconclusive
	{womanly} vs. {manly}	0.001	2.0	0.0	more female-associated
{bowtie} vs. {corsage}	{masculine} vs. {feminine}	0.017	2.0	0.0	more male-associated
	{woman} vs. {masculine}	−0.071	−2.0	0.5	inconclusive
	{girly} vs. {masculine}	0.054	2.0	0.0	more female-associated

Table 1: By contriving the male and female attribute words, we can easily manipulate WEAT to claim that a given target word is more female-biased or male-biased than another. For example, in the top row, $\vec{door}$ is more male-associated than $\vec{curtain}$ when the attribute words are ‘masculine’ and ‘feminine’, but it is more female-associated when the attribute words are ‘woman’ and ‘man’. In both cases, the associations are highly statistically significant.

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). Understanding undesirable word embedding associations. *arXiv preprint arXiv:1908.06361*.

Tutorial roadmap

1. Transformer models & attention mechanism overview
2. Bias in pre-trained language models
3. How can we measure bias?
4. **Ways to mitigate bias**
5. Assignment overview

Ways to mitigate bias

- Two common approaches to mitigate bias:

1. Use debiased data
2. Change the model through debiasing

Ways to mitigate bias

- Using debiased data
- Common approach - CDA (counterfactual data augmentation)
- With CDA we can change sentence to create more “balanced” training data (e.g. more “she is a programmer” sentences)

Zmigrod, R., Mielke, S. J., Wallach, H., & Cotterell, R. (2019). Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*.

Tutorial roadmap

1. Transformer models & attention mechanism overview
2. Bias in pre-trained language models
3. How can we measure bias?
4. Ways to mitigate bias
5. **Assignment overview**