

Заменит ли компьютер лингвистов и переводчиков?



Александр Пиперски

26.01.2020

Переведите на английский

Глобальной эпидемией или пандемией называют болезнь, которая распространяется не только в пределах одного города или страны, но по всему миру и поражает множество людей. Чаще всего речь идет об инфекционных заболеваниях, но иногда говорят о пандемии ожирения или диабета второго типа — болезней, связанных с изменением образа жизни. Но классическое понимание пандемий все же подразумевает, что болезнь заразна — то есть передается от человека к человеку или от животного к человеку каким-то инфекционным агентом.

Google Translate

A global epidemic or pandemic is a disease that spreads not only within one city or country, but around the world and affects many people. Most often we are talking about infectious diseases, but sometimes they talk about a pandemic of obesity or type 2 diabetes - diseases associated with lifestyle changes. But the classical understanding of pandemics still implies that the disease is contagious - that is, it is transmitted from person to person or from animal to person by some kind of infectious agent.

Google Translate

A global epidemic or pandemic is a disease that spreads not only within one city or country, but around the world and affects many people. Most often **we are talking** about infectious diseases, but sometimes **they talk** about a pandemic of obesity or type 2 diabetes - diseases associated with lifestyle changes. But the classical understanding of pandemics still implies that the disease is contagious - that is, it is transmitted from person to person or from animal to person by some kind of infectious agent.

Google Translate

A global epidemic or pandemic is a disease that spreads not only within one city or country, but around the world and affects many people. Most often we are talking about infectious diseases, but sometimes they talk about a pandemic of obesity or type 2 diabetes - diseases associated with lifestyle changes. But the classical understanding of pandemics still implies that the disease is contagious - that is, it is transmitted from **person** to **person** or from animal to by some kind of infectious agent.

Google Translate

A global epidemic or pandemic is a disease that spreads not only within one city or country, but around the world and affects many people. Most often we are talking about infectious diseases, but sometimes they talk about a pandemic of obesity or type 2 diabetes - diseases associated with lifestyle changes. **But** the classical understanding of pandemics still implies that the disease is contagious - that is, it is transmitted from person to person or from animal to person by some kind of infectious agent.

Google Translate

A global epidemic or pandemic is a disease that spreads not only within one city or country, but around the world and affects many people. Most often we are talking about infectious diseases, but sometimes they talk about a pandemic of obesity or type 2 diabetes - diseases associated with lifestyle changes. But the classical understanding of pandemics still implies that the disease is contagious - that is, it is transmitted from person to person or from animal to person by some kind of infectious agent.

Компьютерная лингвистика

- Компьютерная лингвистика окружает нас повсюду:
 - поисковые системы
 - голосовые помощники
 - проверка орфографии
 - машинный перевод
 - извлечение информации из текстов
 - ...

Компьютерная лингвистика

- Компьютерная лингвистика учит компьютер работать с текстами
- Нужны строго формализованные алгоритмы
- Почему сложно работать с естественным языком?

Омонимия

- *Мама купила лук и капусту*
- *Солдат принёс лук и колчан*
- *Эти типы стали есть в литейном цехе*
- *Я видел их семью своими глазами*
- *Time flies like an arrow*
- *Fruit flies like a banana*

Омонимия: перевод

- *Mom bought onions and cabbage*
- *The soldier brought the bow and quiver*
- *These types of steel are found in the foundry*
- *I saw their family with my own eyes*
- *Время летит как стрела*
- *Фруктовые мухи как бананы*

Синонимия:

анафора и кореферентность

- **Анафора** — отсылка к ранее названной сущности с помощью местоимения
- **Кореферентность** — ситуация, когда две или более цепочки слов отсылают к одной и той же сущности
- Эти явления существенно усложняют автоматическое извлечение информации из текстов

Самая трагическая новость минувшей недели пришла из деревни Лука Новгородской области. Здесь сгорел психоневрологический интернат — из шестидесяти больных спаслись только двадцать три. Всех их вывела из горящего здания санитарка Юлия Ануфриева. Сама она погибла, спасая очередного пациента. Корреспондент «РР» отправилась в деревню Лука, чтобы узнать побольше об этом человеке.

Юля побежала в палату, услышав запах дыма. Угол палаты горел. Она бросилась тушить, но поняла, что это не помогает. <...>

Анафора и кореферентность

- *Юля побежала в палату, услышав запах дыма. Угол палаты горел. Она бросилась тушить*
- *Петя положил книгу на стол. Он очень устал и решил выпить чаю.*
- *Петя положил книгу на стол. Он был сделан из красного дерева.*

Схемы Винограда

- The trophy doesn't fit into the brown suitcase because it's too small. What is too small?

Answers: The suitcase / the trophy.

- The trophy doesn't fit into the brown suitcase because it's too large. What is too large?

Answers: The suitcase / the trophy.

Схемы Винограда

- The man couldn't lift his son because he was so weak. Who was weak?
Answers: The man / the son.
- The man couldn't lift his son because he was so heavy. Who was heavy?
Answers: The man / the son.

Спортивный репортаж

Предматчевые расклады для «Спартака» в восьмом туре были таковыми, что подопечные Дмитрия Аленичева могли подтянуться к призовой тройке. Ничья ЦСКА и «Зенита», случившаяся в субботу в Химках, и поражение «Локомотива» от «Рубина», давали шансы красно-белым на то, чтобы нагнать лидеров, сравнявшись по очкам с питерцами, и хозяева «Открытие Арены» намерены были воспользоваться такой возможностью. Однако не всё у спартаковцев было так гладко.

Турнирная таблица после 7-го тура

		ВСЕГО					
КОМАНДА		И	В	Н	П	 З -  П	О
1	ПФК ЦСКА	7	7	0	0	11 - 2	21
2	ЛОКОМОТИВ	7	5	2	0	13 - 4	17
3	ЗЕНИТ	7	5	0	2	14 - 8	15
4	СПАРТАК	7	4	1	2	11 - 7	13

Прагматика. Знания о мире

- *Вы не знаете, который час?*
- *— Пойдешь со мной вечером в клуб?*
— У меня завтра утром экзамен.
- *Я видел их семью своими глазами*

Искусственный интеллект

- Википедия:

Искусственный интеллект (ИИ; англ. *artificial intelligence, AI*) — свойство интеллектуальных систем выполнять творческие функции, которые традиционно считаются прерогативой человека; наука и технология создания интеллектуальных машин, особенно интеллектуальных компьютерных программ.

Интеллект

- Википедия:

Интеллѐкт (от лат. *intellectus* «ощущение», «восприятие»; «разумение», «понимание»; «понятие», «рассудок») или **ум** — качество психики, состоящее из способности приспосабливаться к новым ситуациям, способности к обучению и запоминанию на основе опыта, пониманию и применению абстрактных концепций и использованию своих знаний для управления окружающей человека средой.

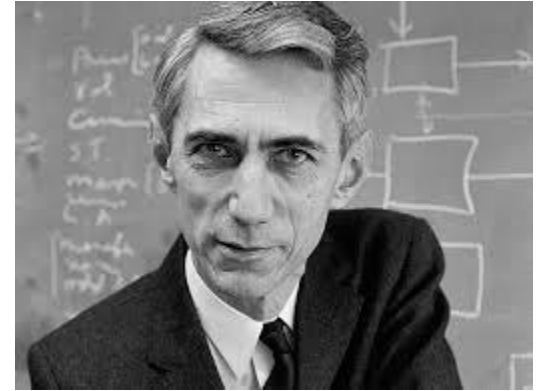
Генерация текстов на естественном языке

Москва. 22 января. INTERFAX.RU - Доллар США и евро понижаются на "Московской бирже" в понедельник утром, рубль немного укрепляется к бивалютной корзине в условиях положительной динамики на рынке нефти.

Первые сделки по доллару США прошли в диапазоне 56,68-56,73 руб., по итогам первой минуты торгов курс составил 56,71 руб. (-1 копейка к уровню предыдущего закрытия). Евро при этом снизился до 69,29 руб. (-9,75 копейки), стоимость бивалютной корзины (\$0,55 и EUR0,45) опустилась на 4,9 копейки по отношению к уровню закрытия пятницы, до 62,34 рубля.

Компьютерные шахматы

- Клод Шеннон (1916–2001)
- «Programming a computer for playing chess» (1949)
- Два типа стратегий:
 - тип А: полный перебор
 - тип В: перебор только тех продолжений, которые оцениваются как перспективные



Компьютерные шахматы

- Стратегия типа А требует перебрать $\sim 10^9$ вариантов на три хода (шесть полуходов) вперёд, и если позиция оценивается за 1 микросекунду (10^{-6} с), то на один ход понадобится ~ 17 минут

Компьютерные шахматы

Белые	Чёрные	Победа белых	Ничья	Победа чёрных
AlphaZero	Stockfish	25	25	0
Stockfish	AlphaZero	0	47	3

- AlphaZero — Stockfish **64:36** (+28, =72, -0)
- 2017: нейросетевые шахматы AlphaZero
- Нейронная сеть AlphaZero после 4 часов игры сама с собой оказалась лучше самой успешной традиционной шахматной программы Stockfish

Автоматический выбор значения слов

- *Мама купила лук и капусту*
- *Солдат принёс лук и колчан*

- Какой лук нужен?

лук: Большой толковый словарь

- *лук* 1: Огородное растение сем. лилейных. Съедобные трубчатые листья или луковицы этого растения
- *лук* 2: Ручное оружие для метания стрел, изготовленное из гибкого, упругого стержня (обычно деревянного), стянутого в дугу тетивой.

Мама купила лук и капусту

- *мама*: (в семейном общении и в разговоре детей о родной матери). = Мать. Ласк. Тёща или свекровь (обычно в семейном обращении).
- *купить*: Приобрести за деньги. Привлечь на свою сторону посредством подкупа, взятки; подкупить. Расположить чем-л. в свою пользу, вызвать чью-л. симпатию. Обмануть, разыграть кого-л. В карточных играх: получить по правилам игры в дополнение к своим картам, взять в прикупе.
- *лук 1*: Огородное растение сем. лилейных. Съедобные трубчатые листья или луковицы этого растения
- *лук 2*: Ручное оружие для метания стрел, изготовленное из гибкого, упругого стержня (обычно деревянного), стянутого в дугу тетивой.
- *капуста 1*: Огородное растение сем. крестоцветных с завивающимися в кочан листьями, которые употребляются в пищу.
- *капуста 2*: Жарг. Деньги.

Закон семантического согласования

- Одни и те же компоненты значения должны повторяться в разных словах предложения
- Алгоритм выбора адекватного понимания: выбрать тот набор толкований, при котором количество пересекающихся слов в толкованиях будет наибольшим.

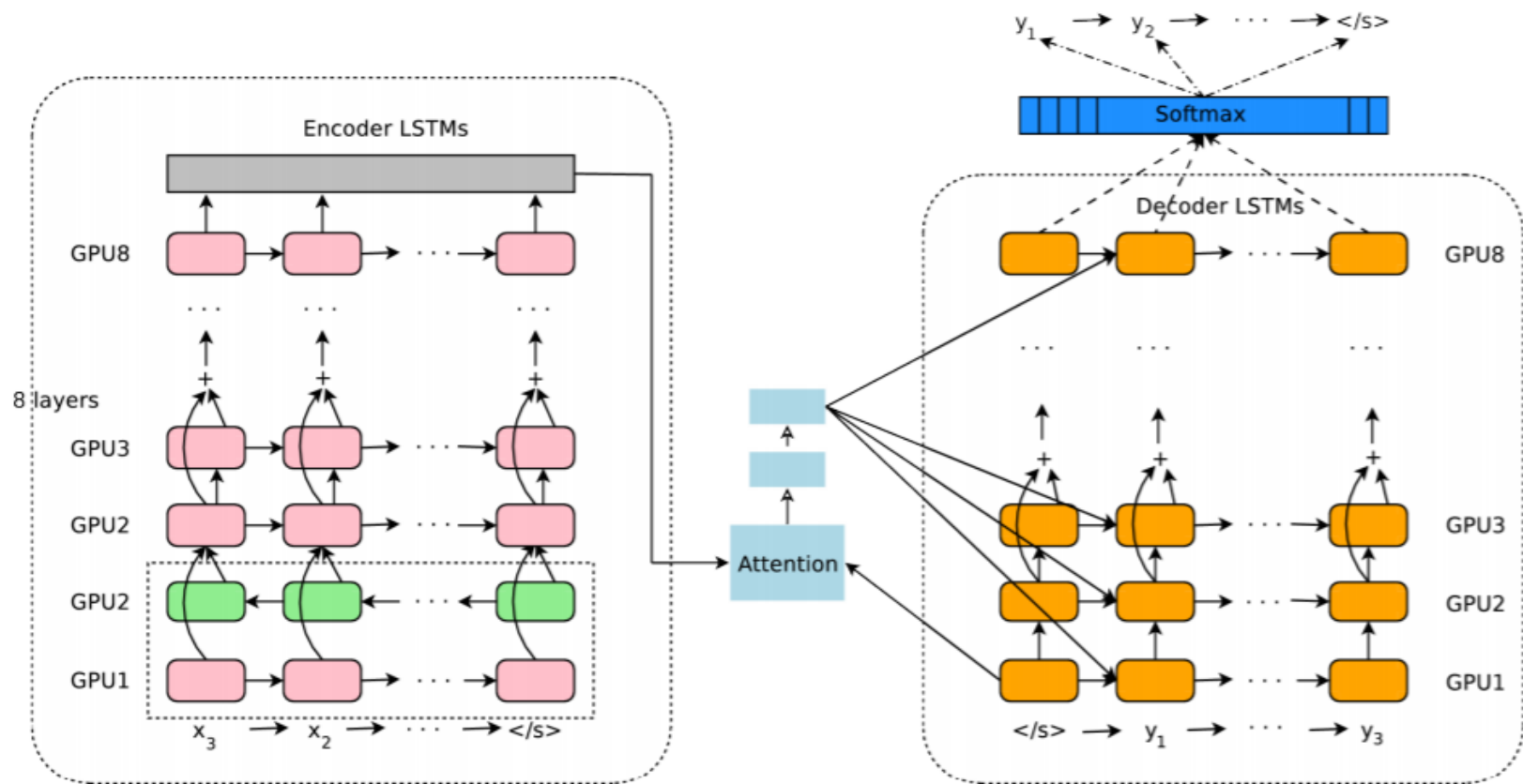
Машинный перевод

- Три поколения машинного перевода:
 - Правильный машинный перевод
 - Статистический машинный перевод (в первую очередь фразовый)
 - Нейросетевой машинный перевод

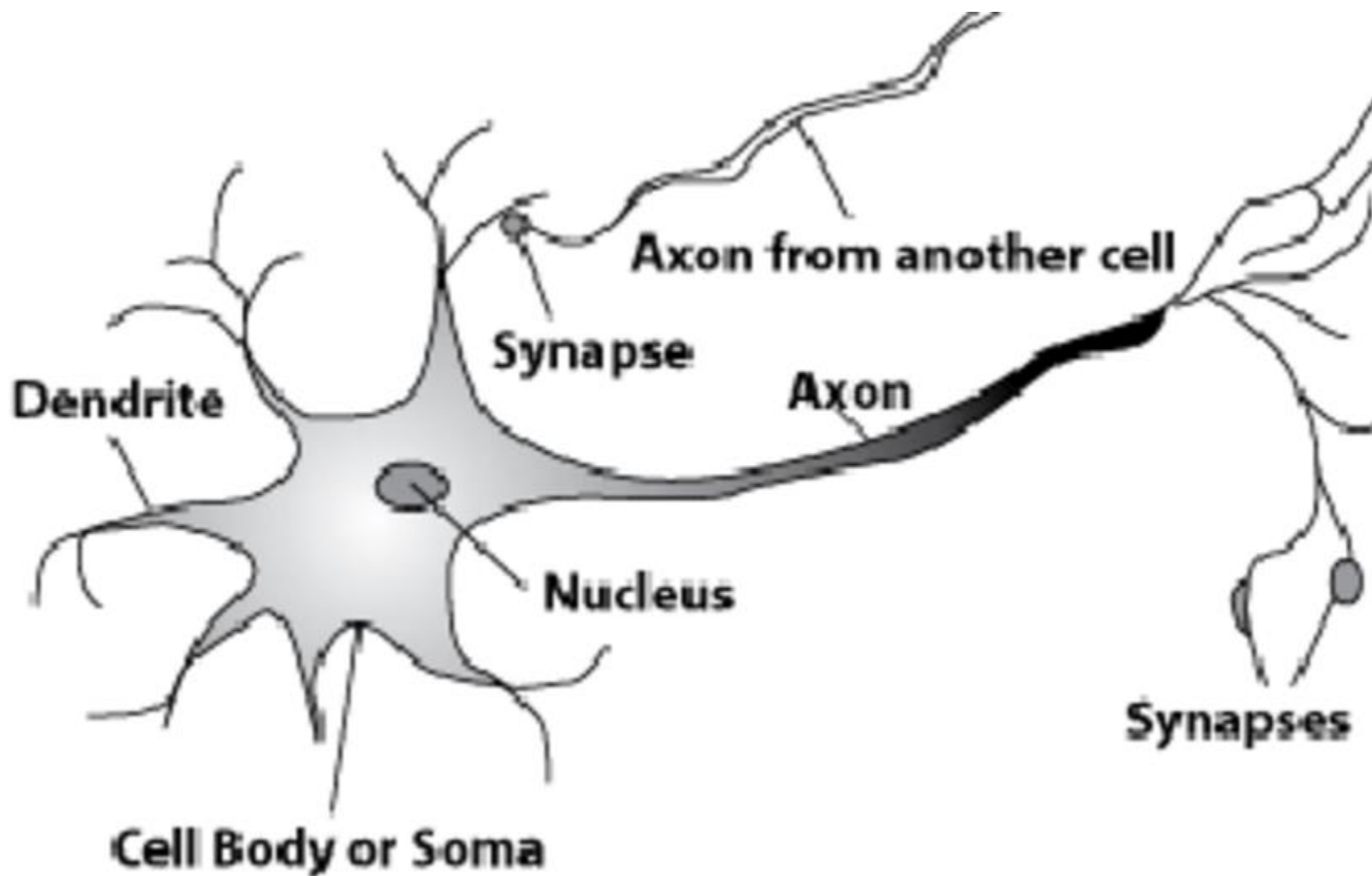
Нейросетевой машинный перевод

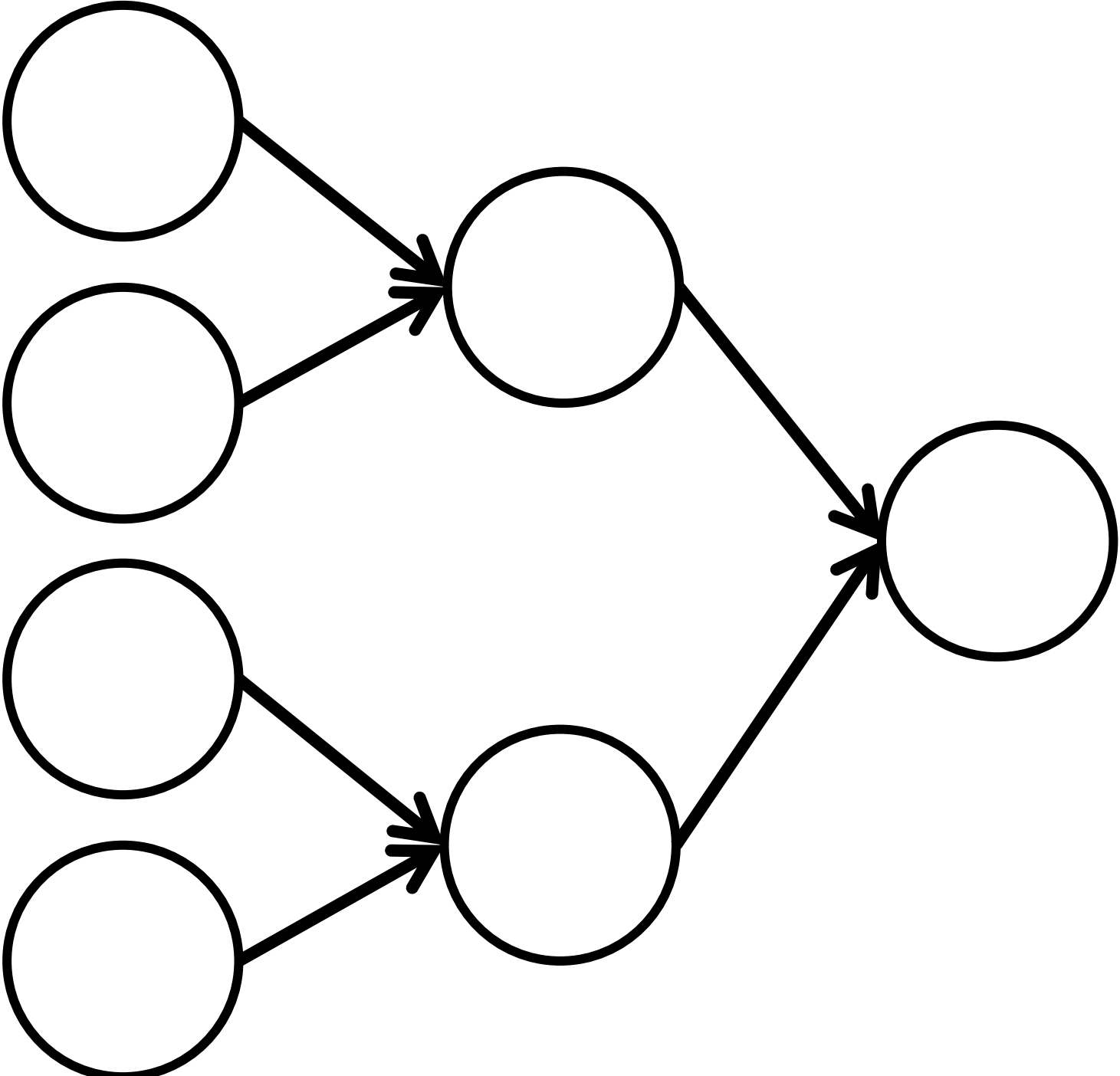
- Закрепился как индустриальный стандарт в 2017 году
 - Google: ноябрь 2016 года
(русский язык — март 2017 года)
 - Яндекс: сентябрь 2017 года
(NB: гибридная система)

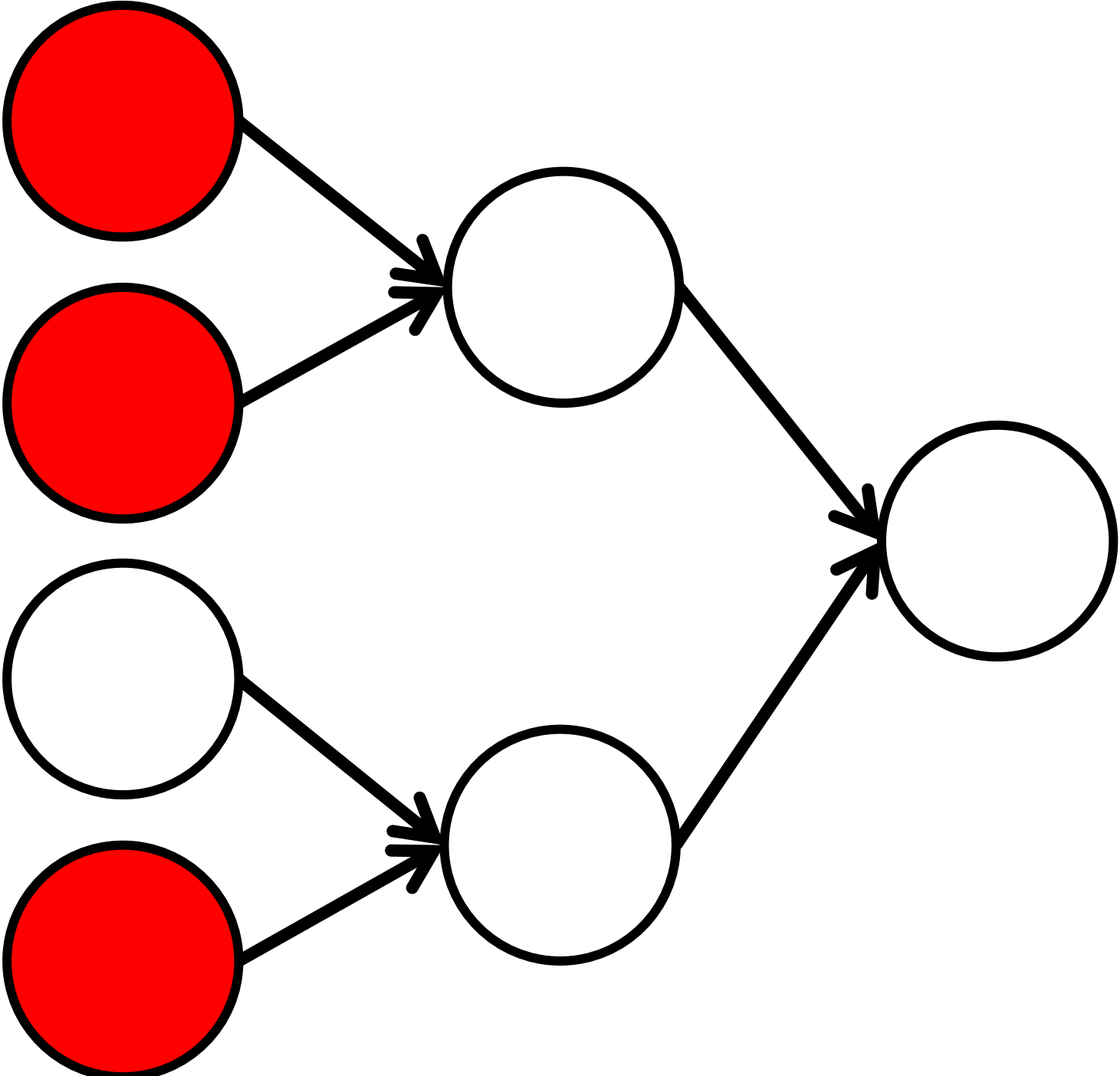
Google's Neural Machine Translation

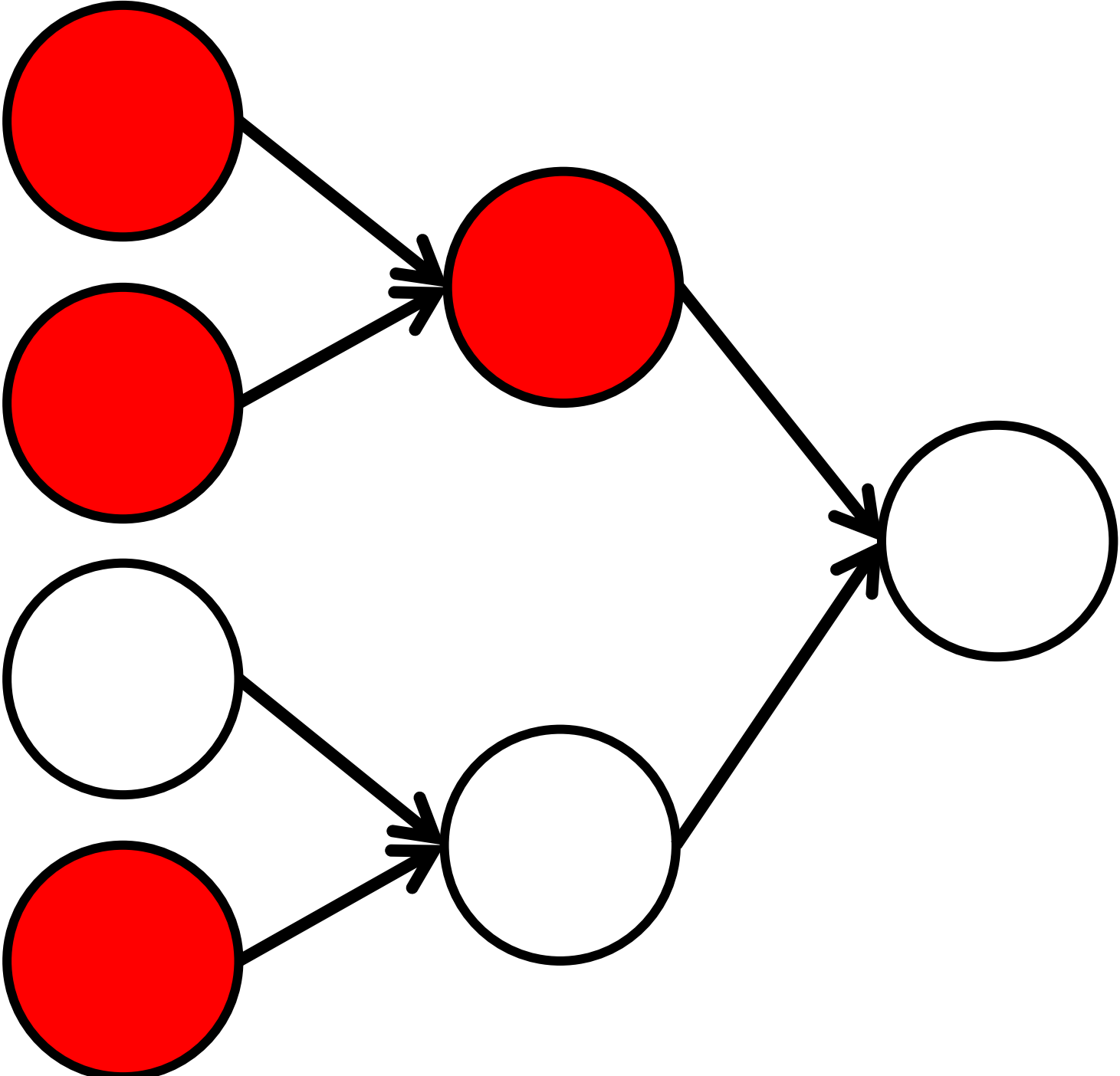


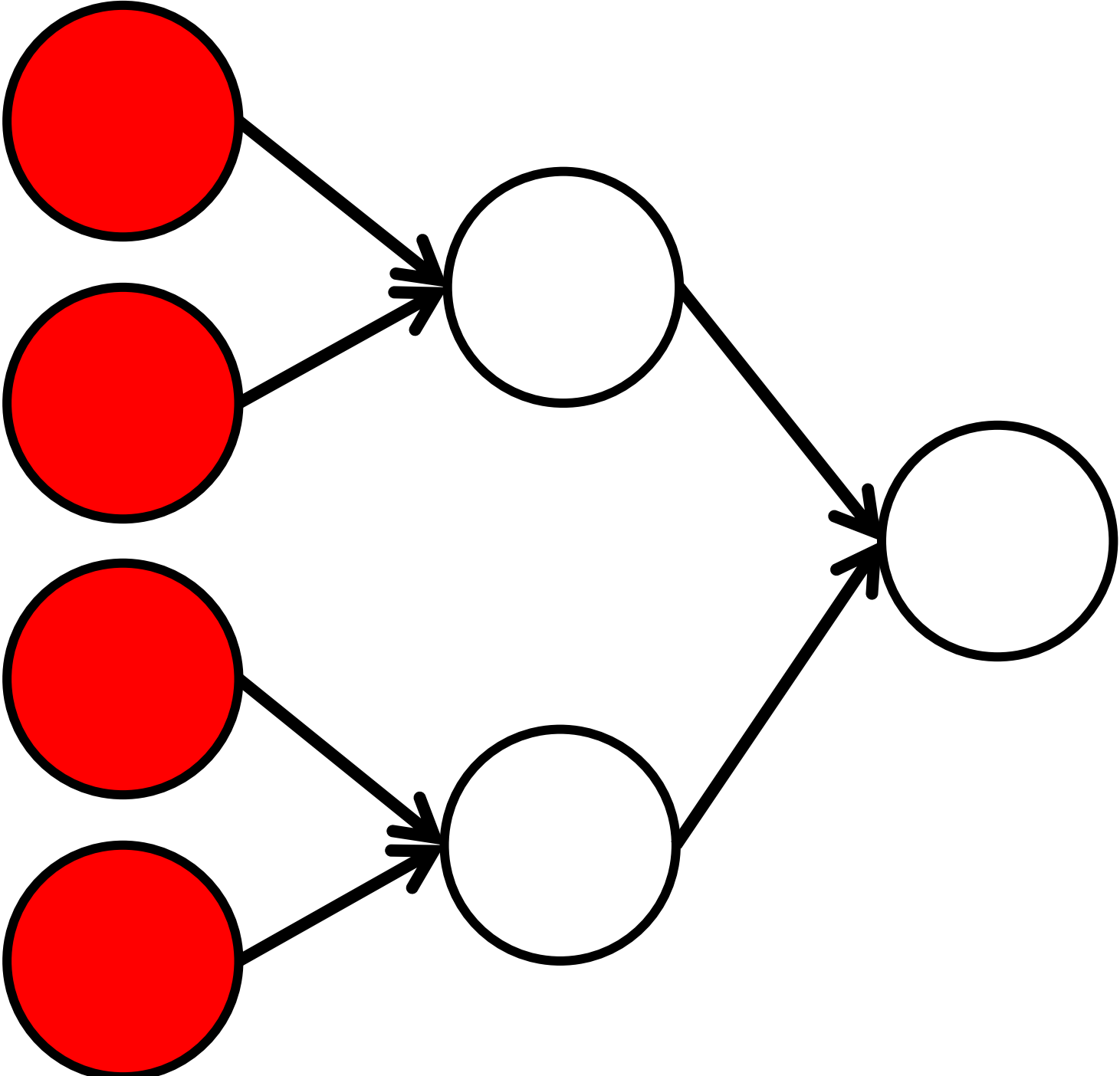
Естественный нейрон

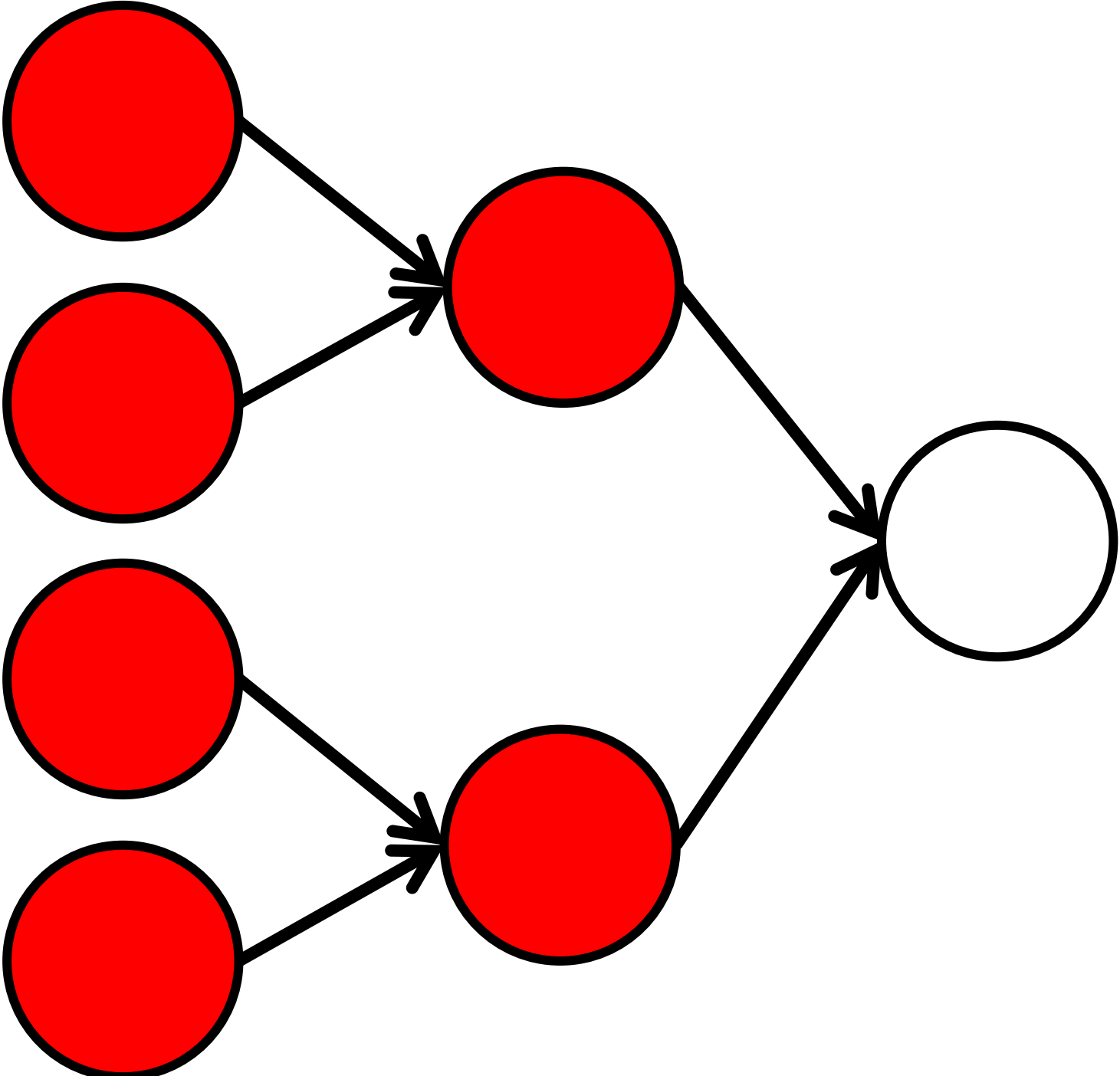


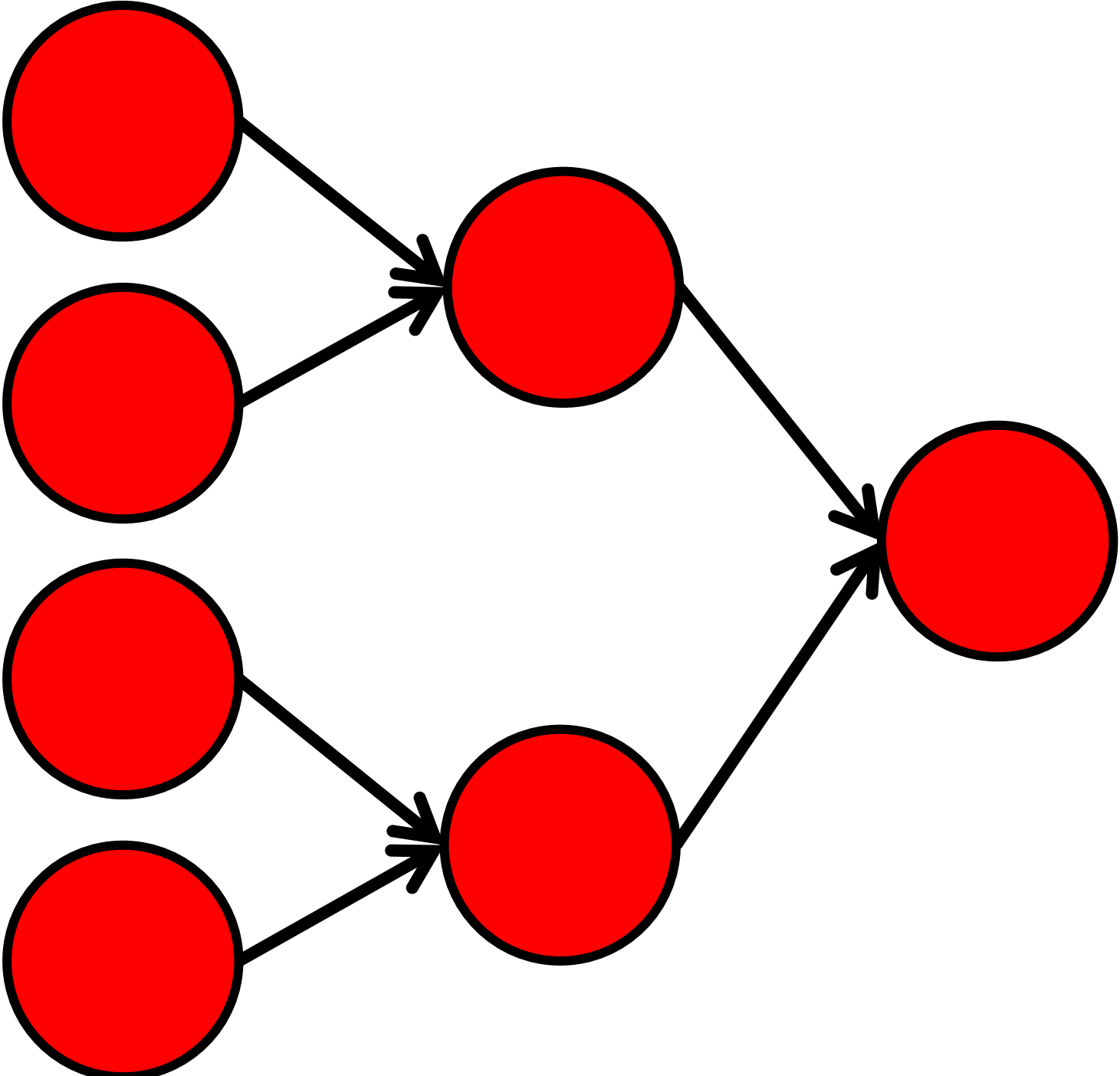


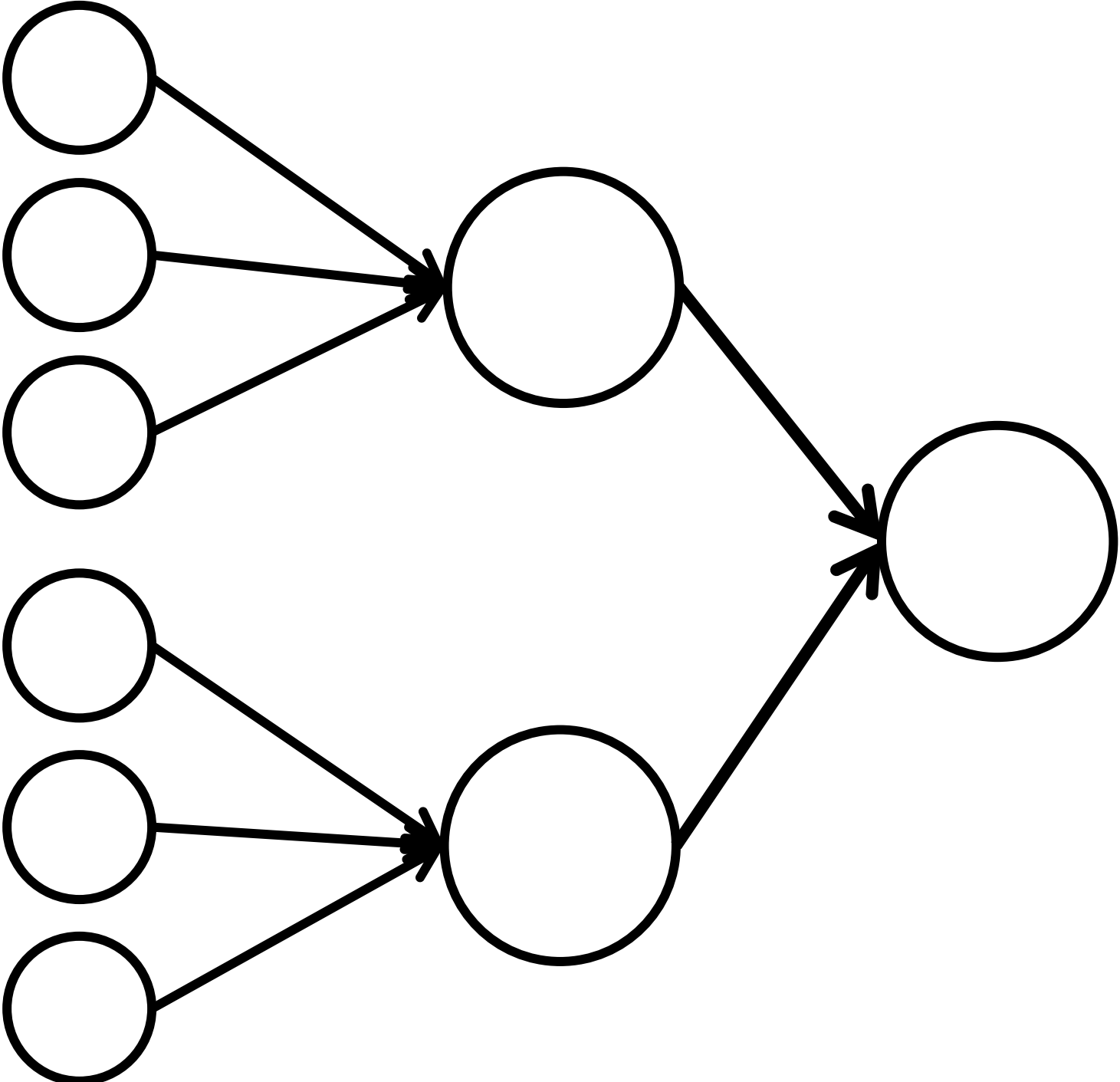


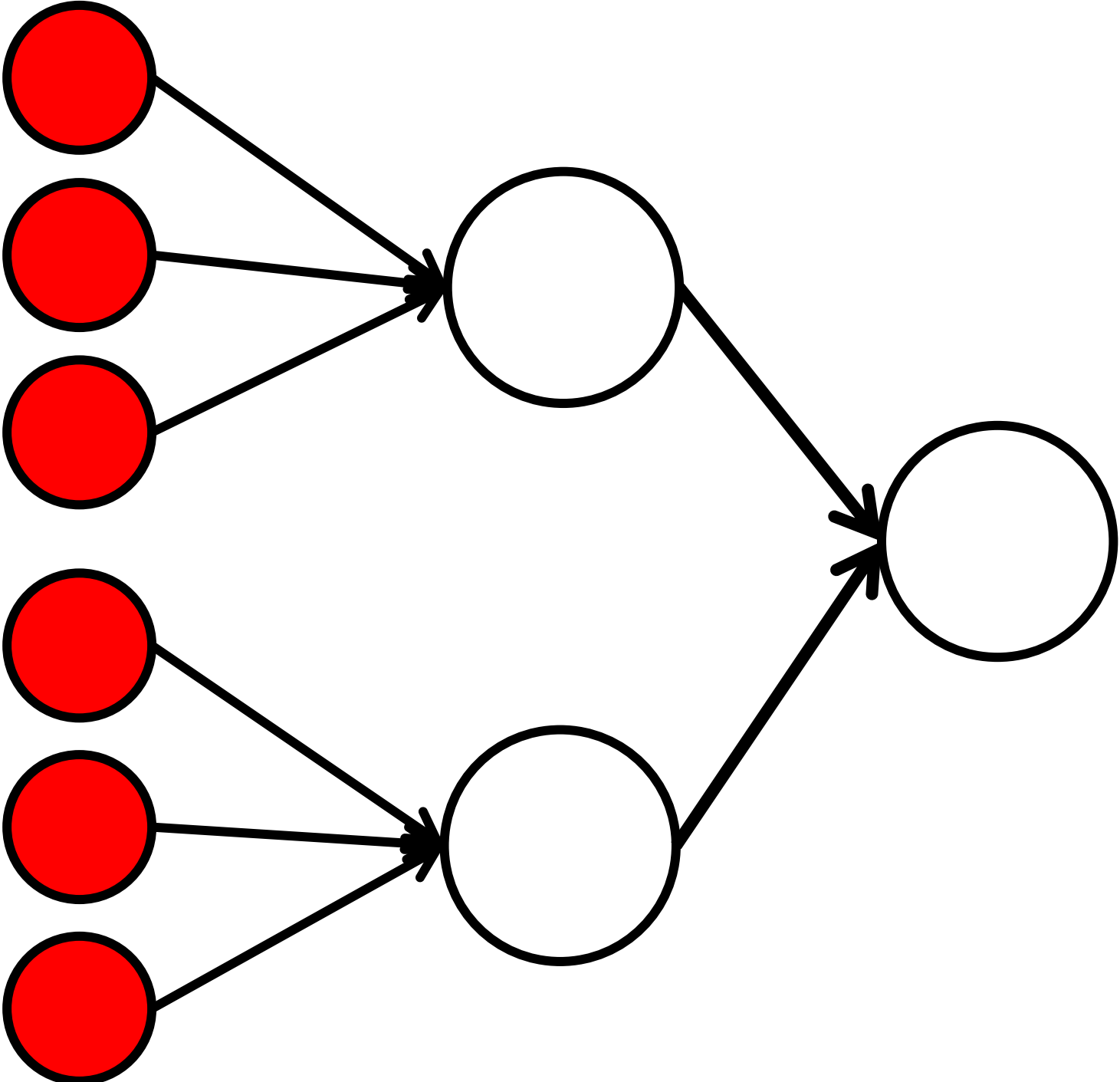


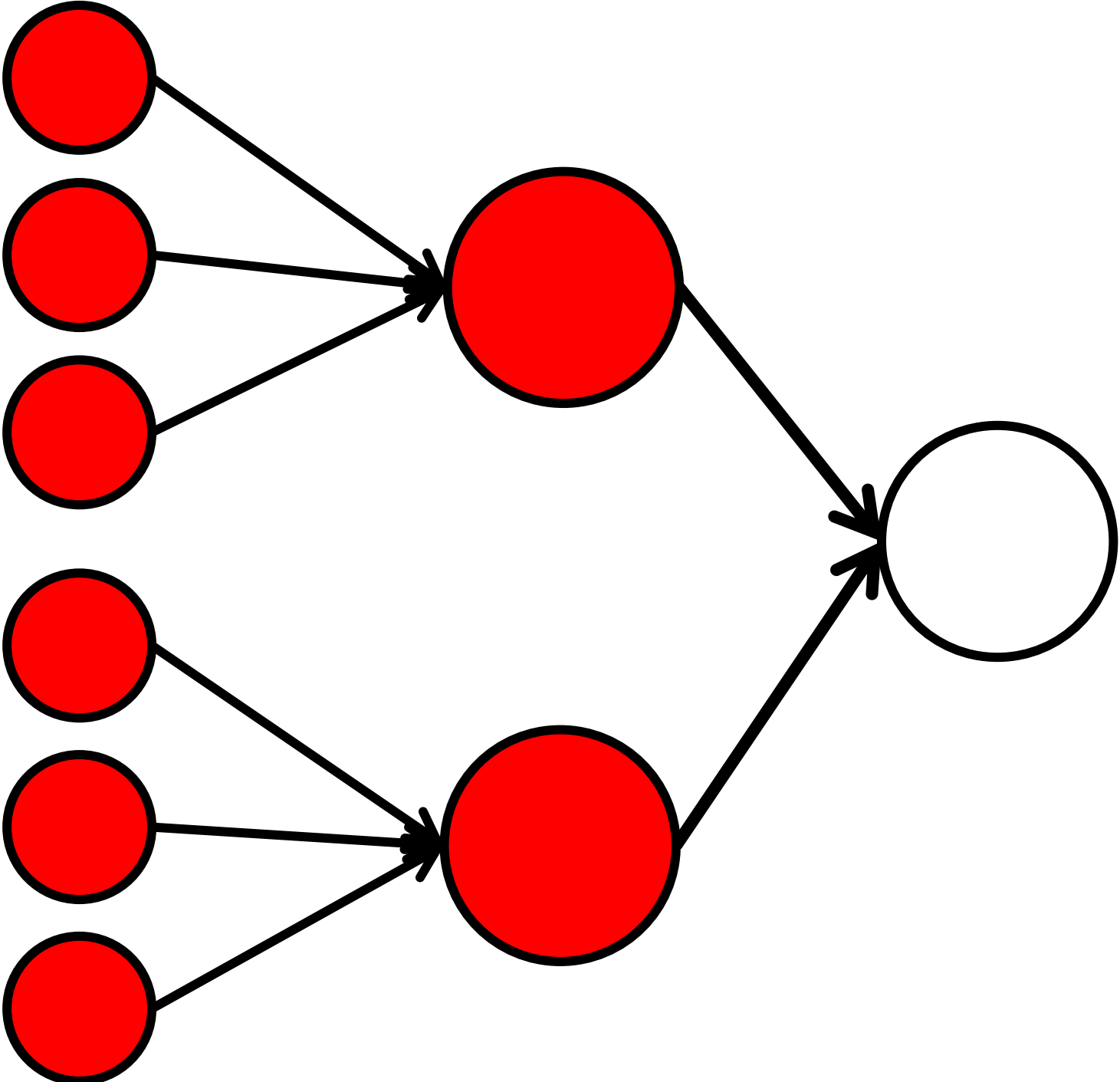


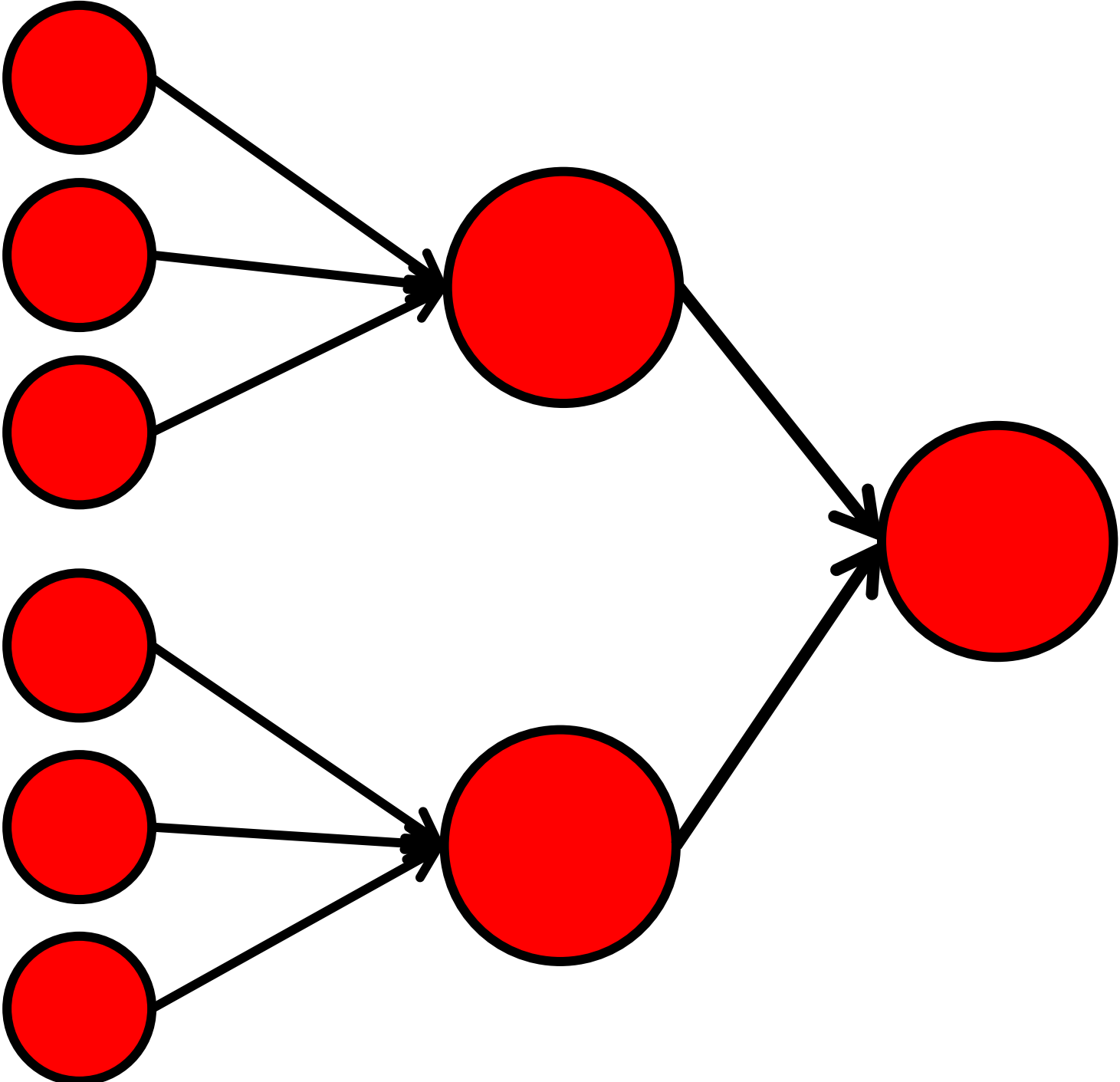


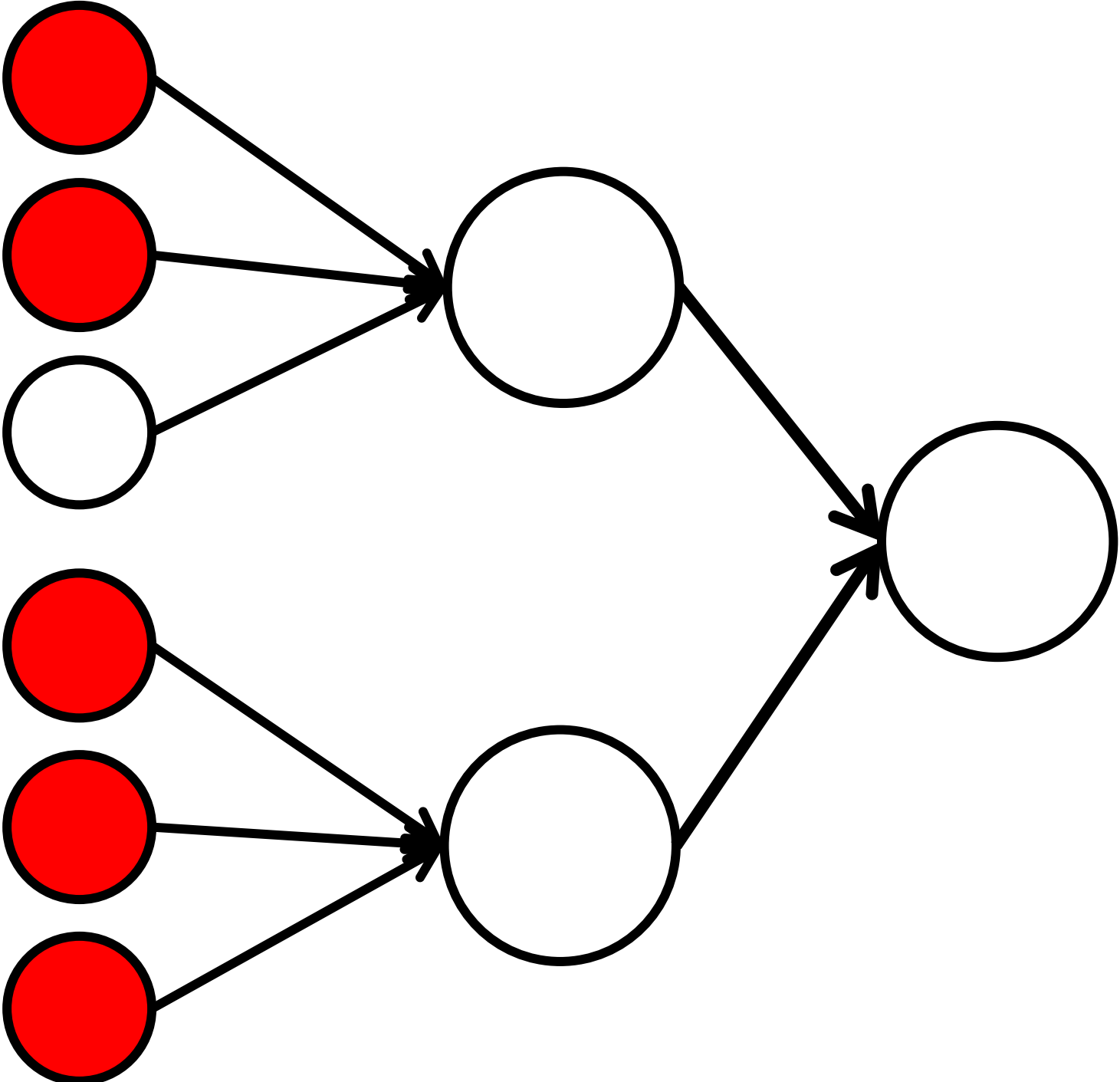


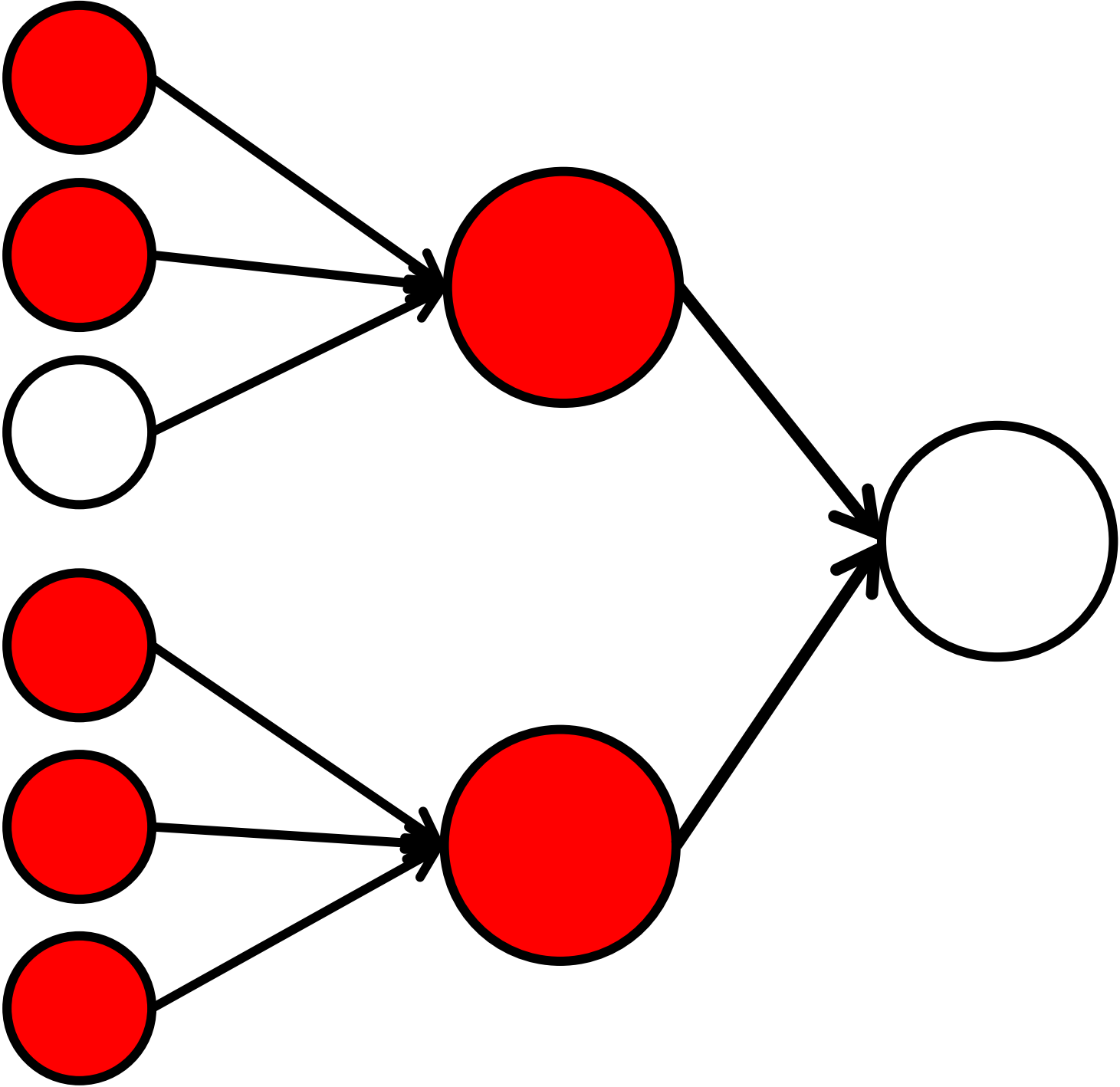


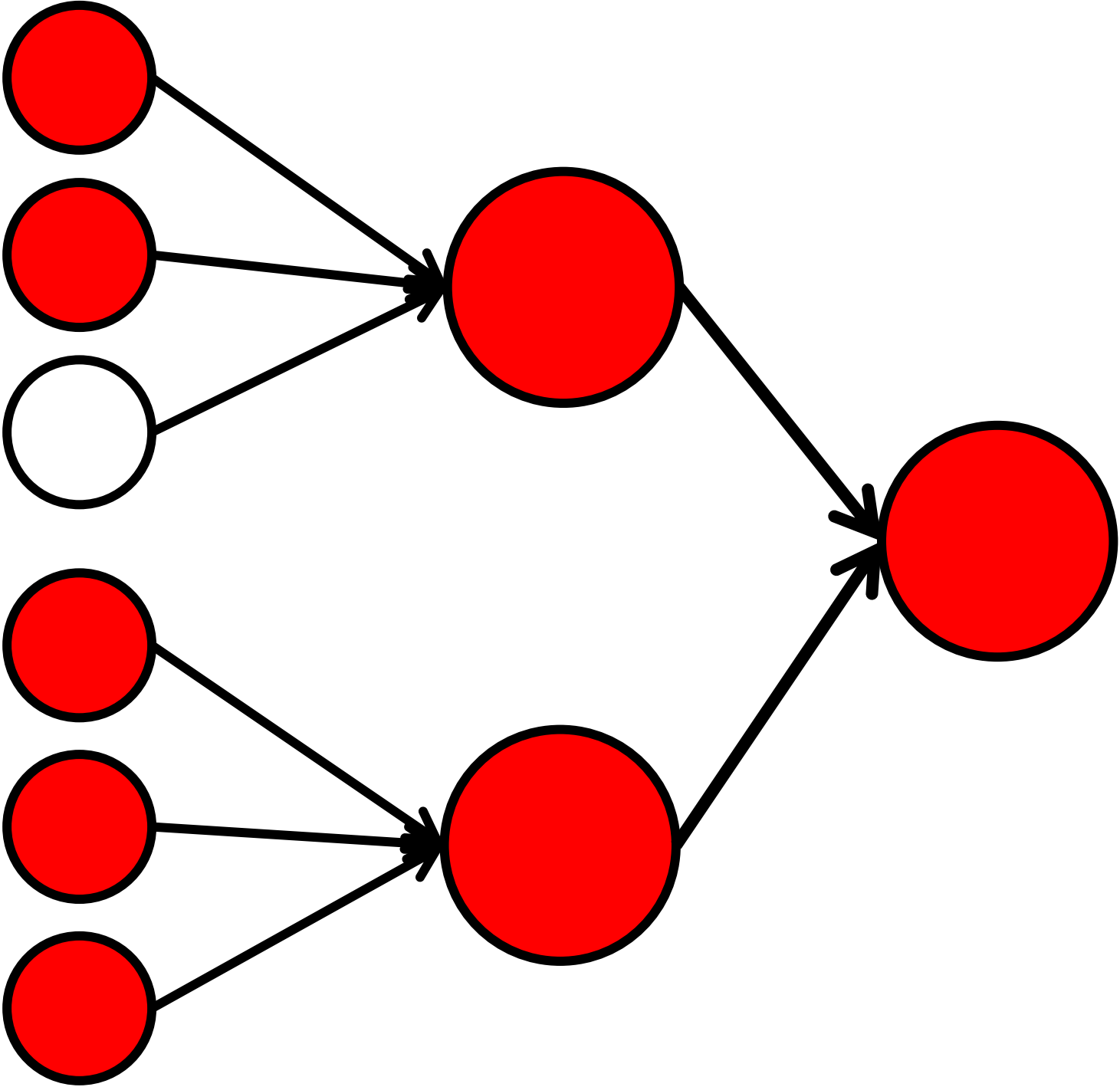


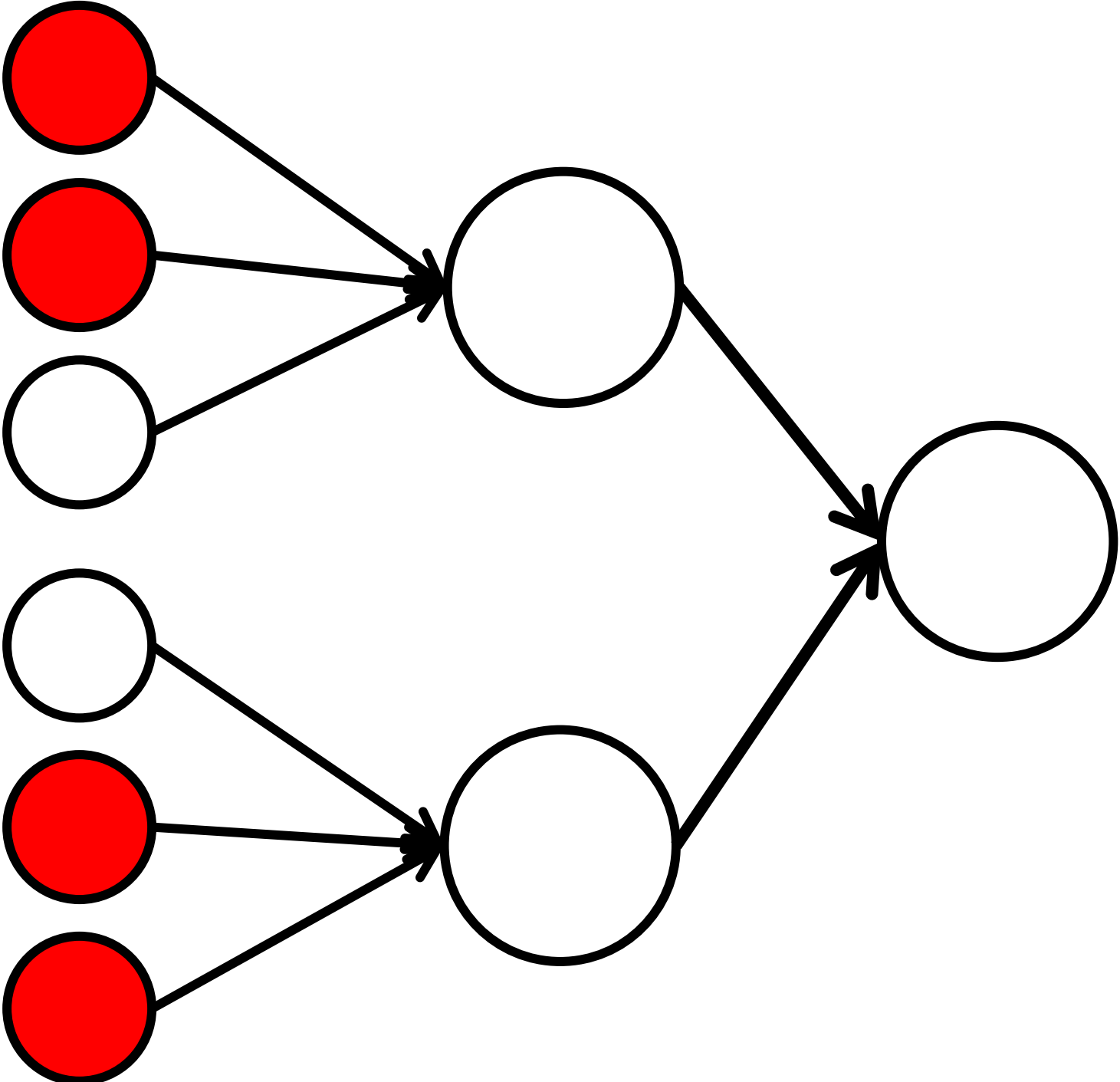


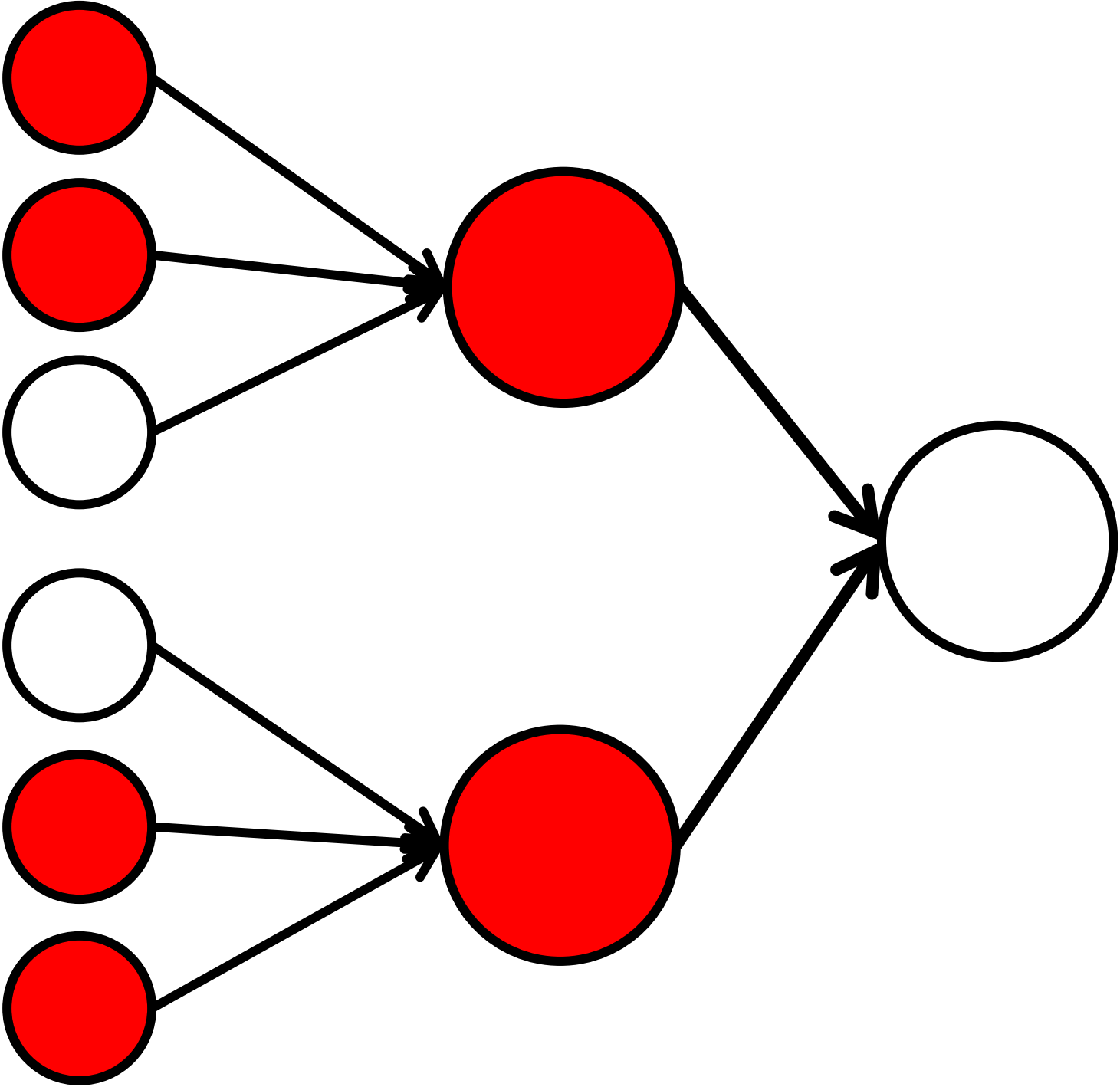


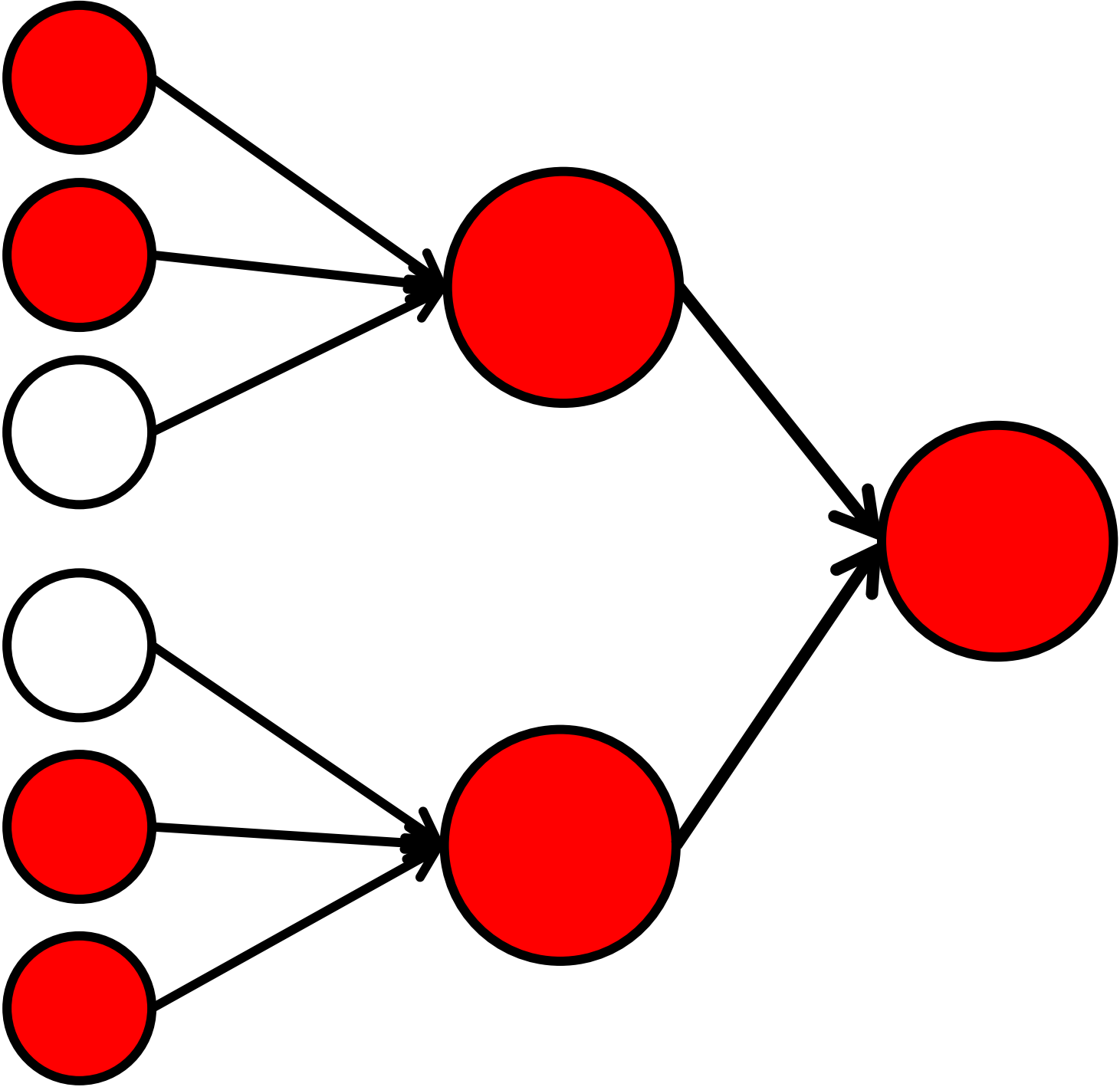


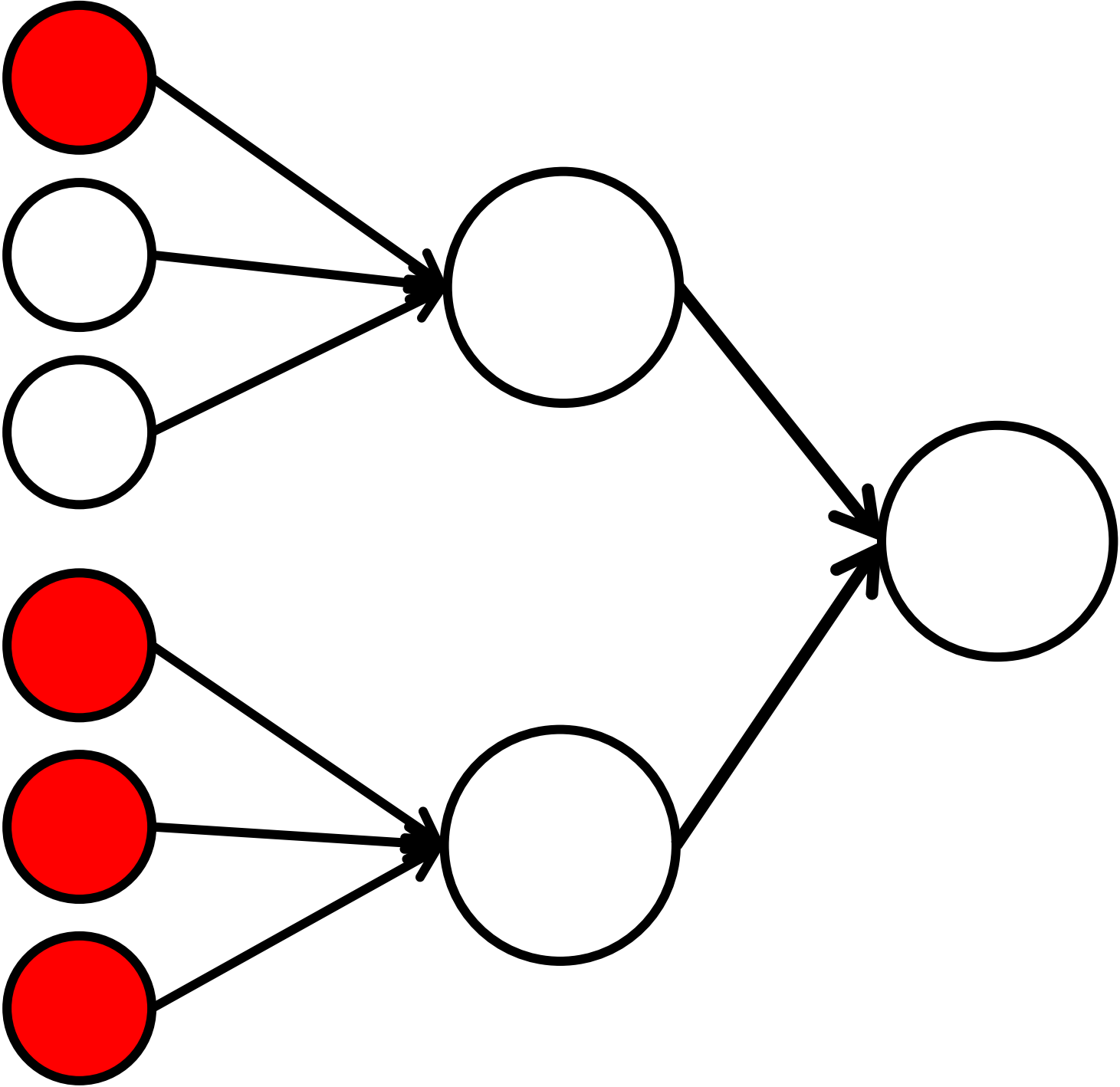


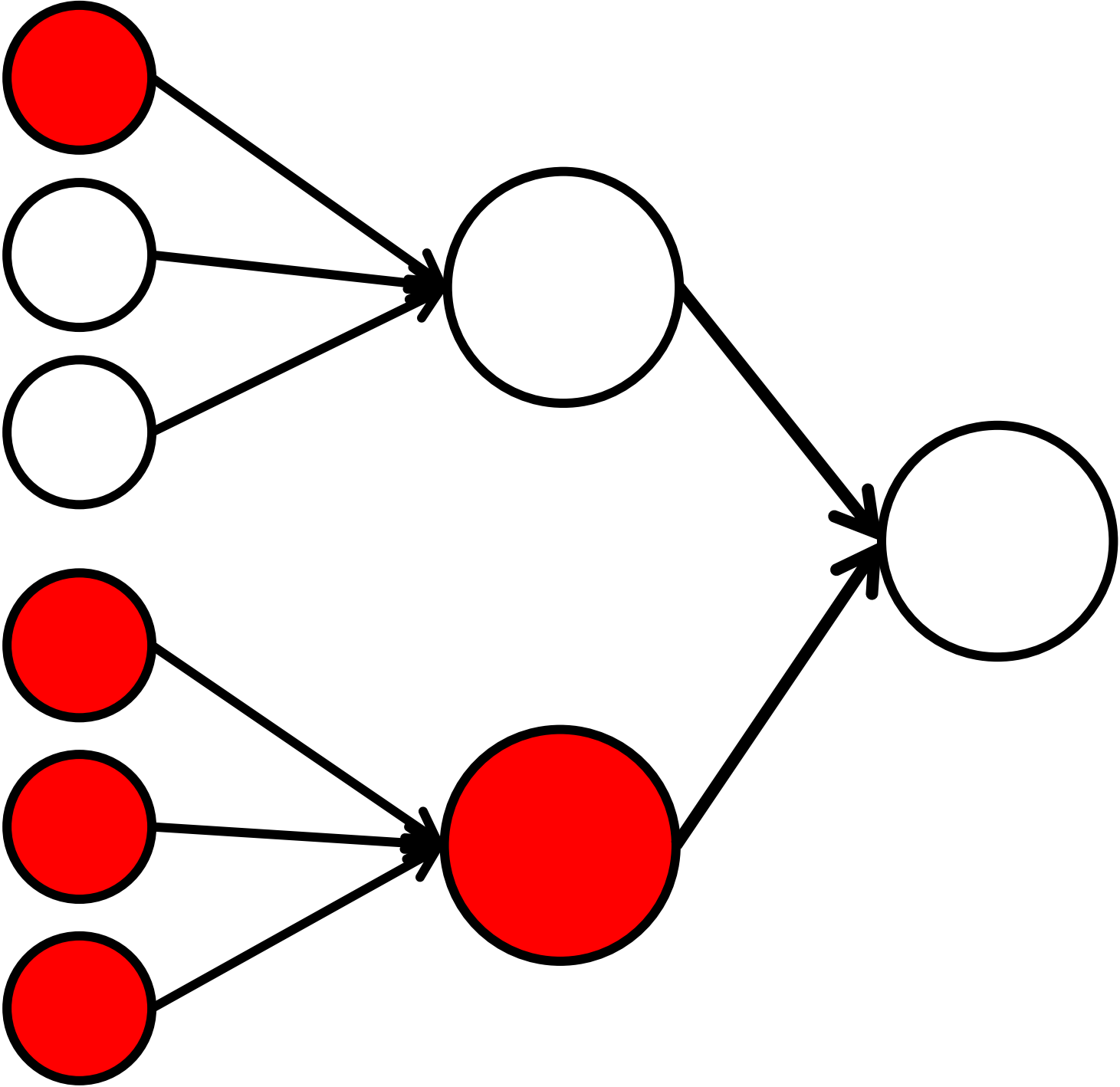










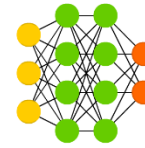


Neural Networks

©2016 Fjodor van Veen - asimovinstitute.org

- Backfed Input Cell
- Input Cell
- △ Noisy Input Cell
- Hidden Cell
- Probablistic Hidden Cell
- △ Spiking Hidden Cell
- Output Cell
- Match Input Output Cell
- Recurrent Cell
- Memory Cell
- △ Different Memory Cell
- Kernel
- Convolution or Pool

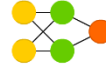
Deep Feed Forward (DFF)



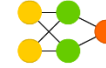
Perceptron (P)



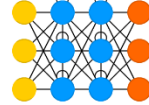
Feed Forward (FF)



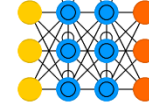
Radial Basis Network (RBF)



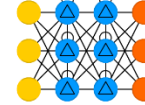
Recurrent Neural Network (RNN)



Long / Short Term Memory (LSTM)



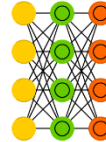
Gated Recurrent Unit (GRU)



Auto Encoder (AE)



Variational AE (VAE)



Denoising AE (DAE)



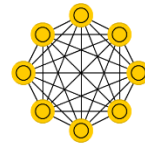
Sparse AE (SAE)



Markov Chain (MC)



Hopfield Network (HN)



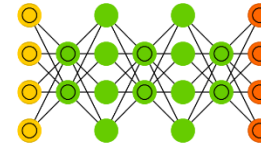
Boltzmann Machine (BM)



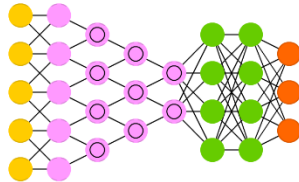
Restricted BM (RBM)



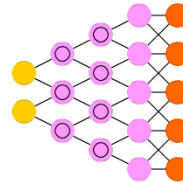
Deep Belief Network (DBN)



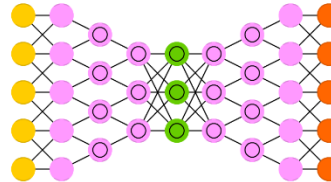
Deep Convolutional Network (DCN)



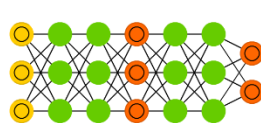
Deconvolutional Network (DN)



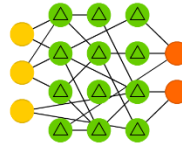
Deep Convolutional Inverse Graphics Network (DCIGN)



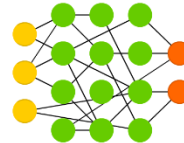
Generative Adversarial Network (GAN)



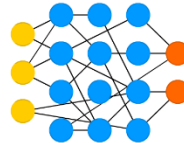
Liquid State Machine (LSM)



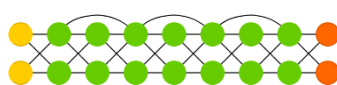
Extreme Learning Machine (ELM)



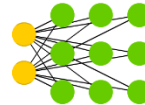
Echo State Network (ESN)



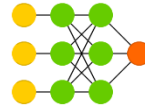
Deep Residual Network (DRN)



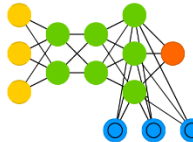
Kohonen Network (KN)

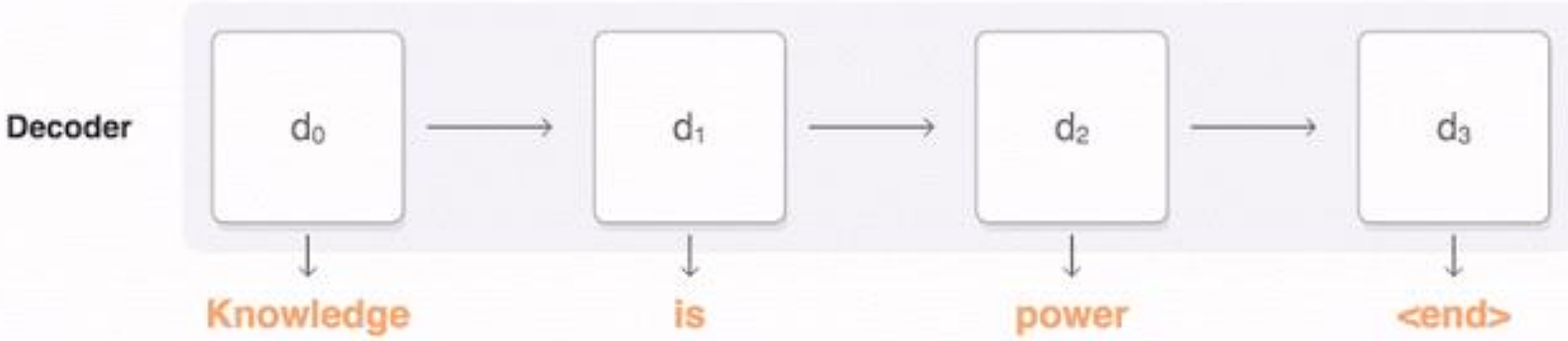


Support Vector Machine (SVM)



Neural Turing Machine (NTM)








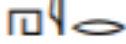

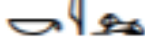
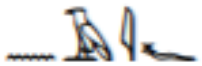


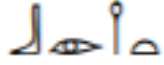
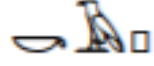


Векторное представление слов

- Искусственная нейронная сеть работает с числами, а не со словами
- Слова надо представить в виде чисел
- «You shall know a word by the company it keeps» (Джон Руперт Фирт)














кашне

- Чёрное кожаное пальто и светло-серая кепка? Ну, ещё какое-нибудь непременно кашне...
- На шее у него в душную ночь зачем-то было наверхено старенькое полосатое кашне.
- Мокроватое куцое вафельное полотенце Нержин повесил себе на шею вроде кашне.
- Закрутив вокруг горла кашне и нахлобучив кепку, оскорблённый мулат покинул редакцию.
- Он шёл по улице в раскрытом зимнем пальто, с болтающимся на шее ярким клетчатым кашне, опираясь на суковатую палку.

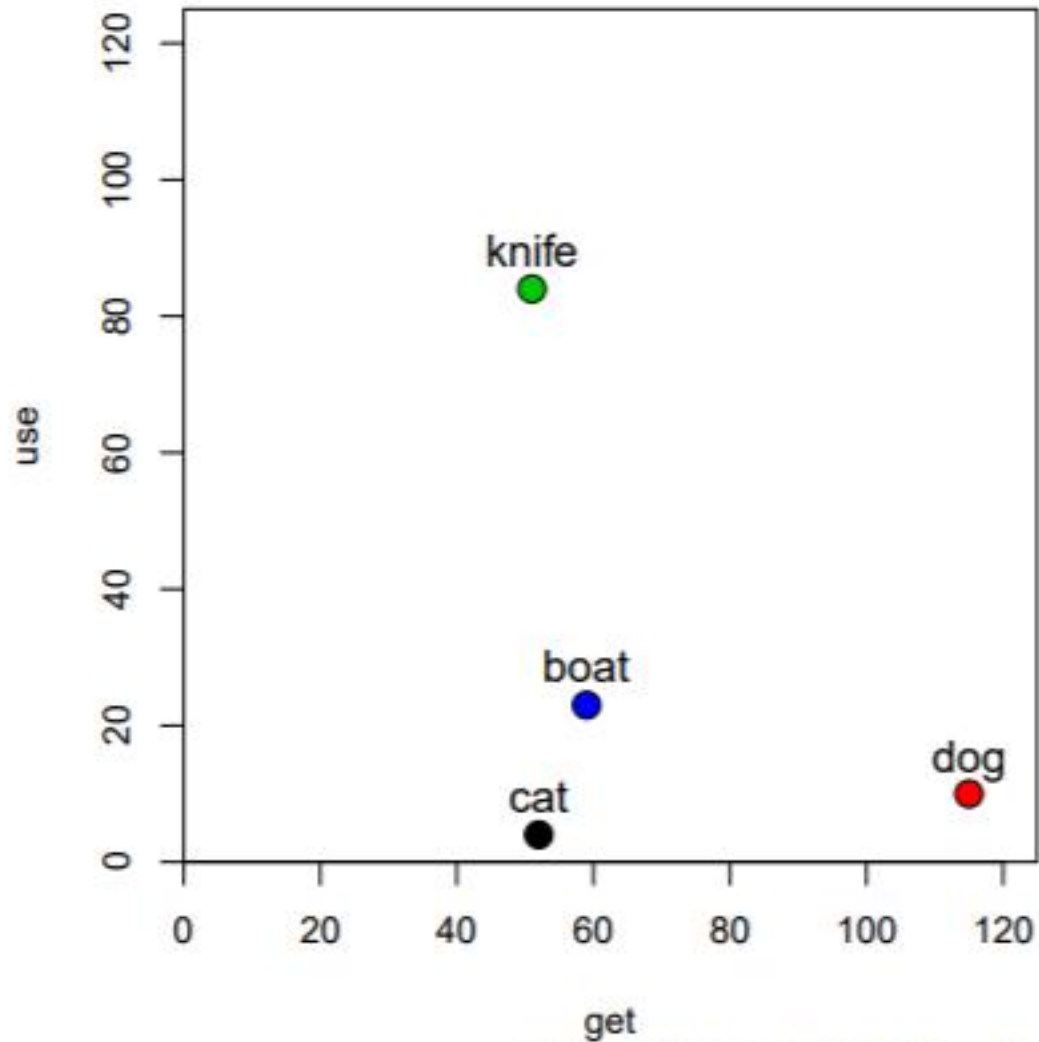
Дистрибуционная семантика

							
(knife)		51	20	84	0	3	0
(cat)		52	58	4	4	6	26
???		115	83	10	42	33	17
(boat)		59	39	23	4	0	0
(cup)		98	14	6	2	1	0
(pig)		12	17	3	2	9	27
(banana)		11	2	2	0	18	0

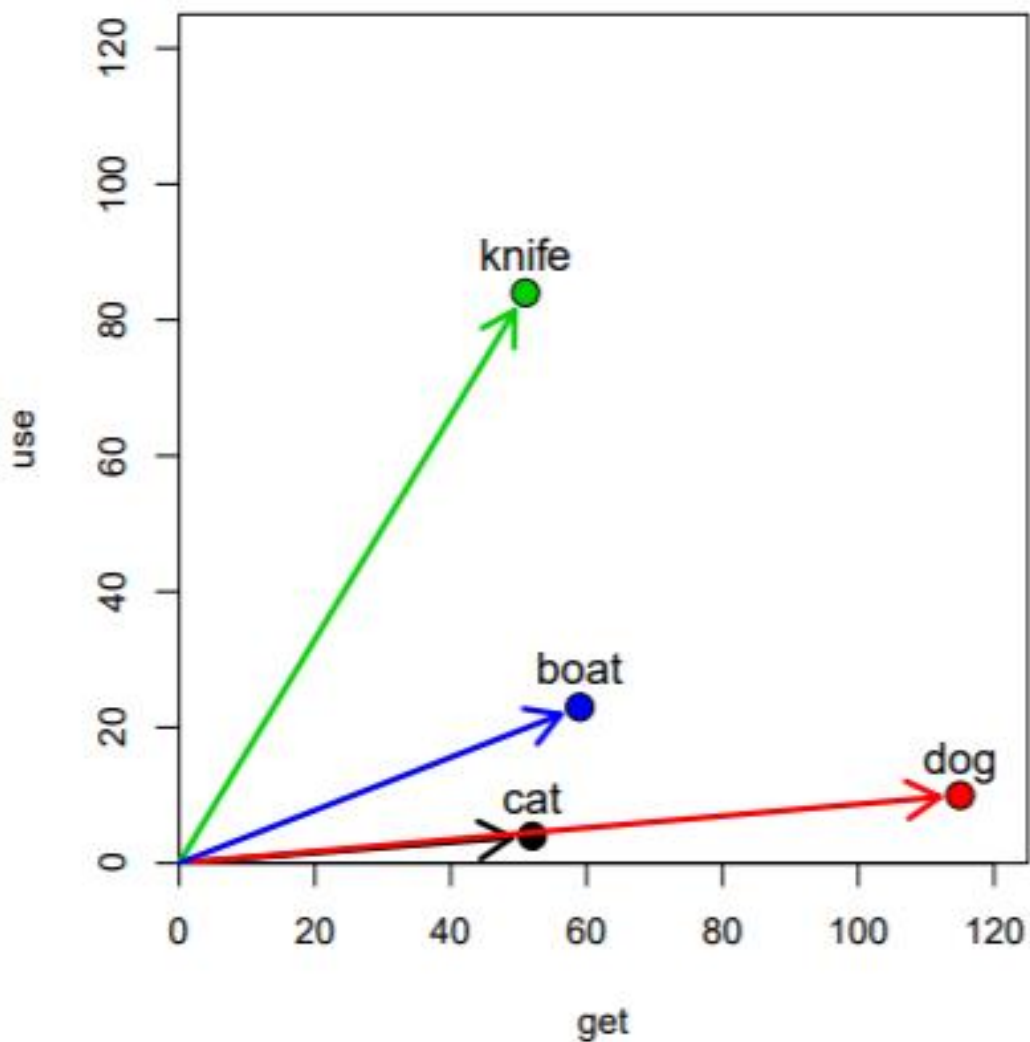
Дистрибуционная семантика

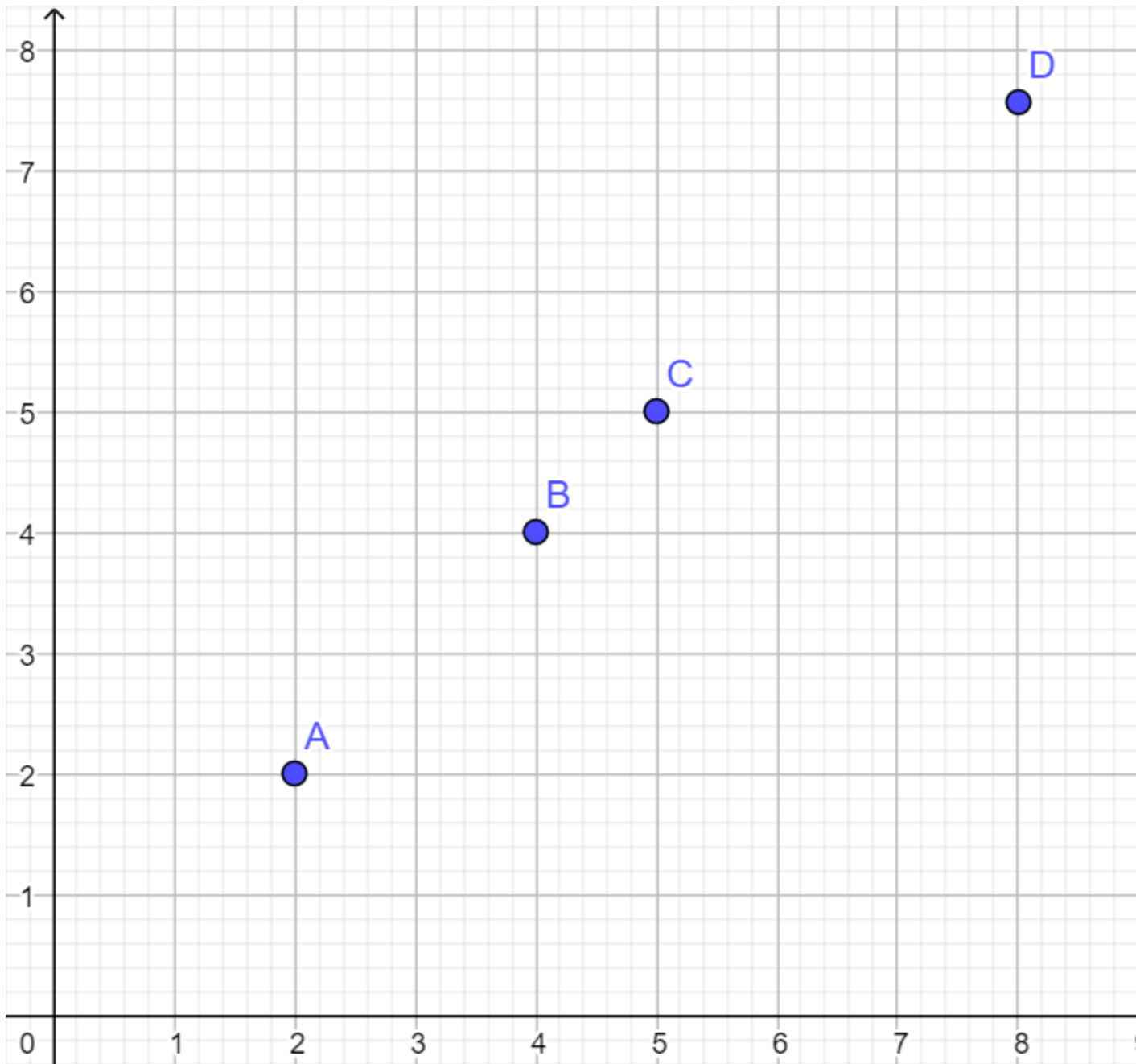
		get 	see 	use 	hear 	eat 	kill 
knife		51	20	84	0	3	0
cat		52	58	4	4	6	26
dog		115	83	10	42	33	17
boat		59	39	23	4	0	0
cup		98	14	6	2	1	0
pig		12	17	3	2	9	27
banana		11	2	2	0	18	0

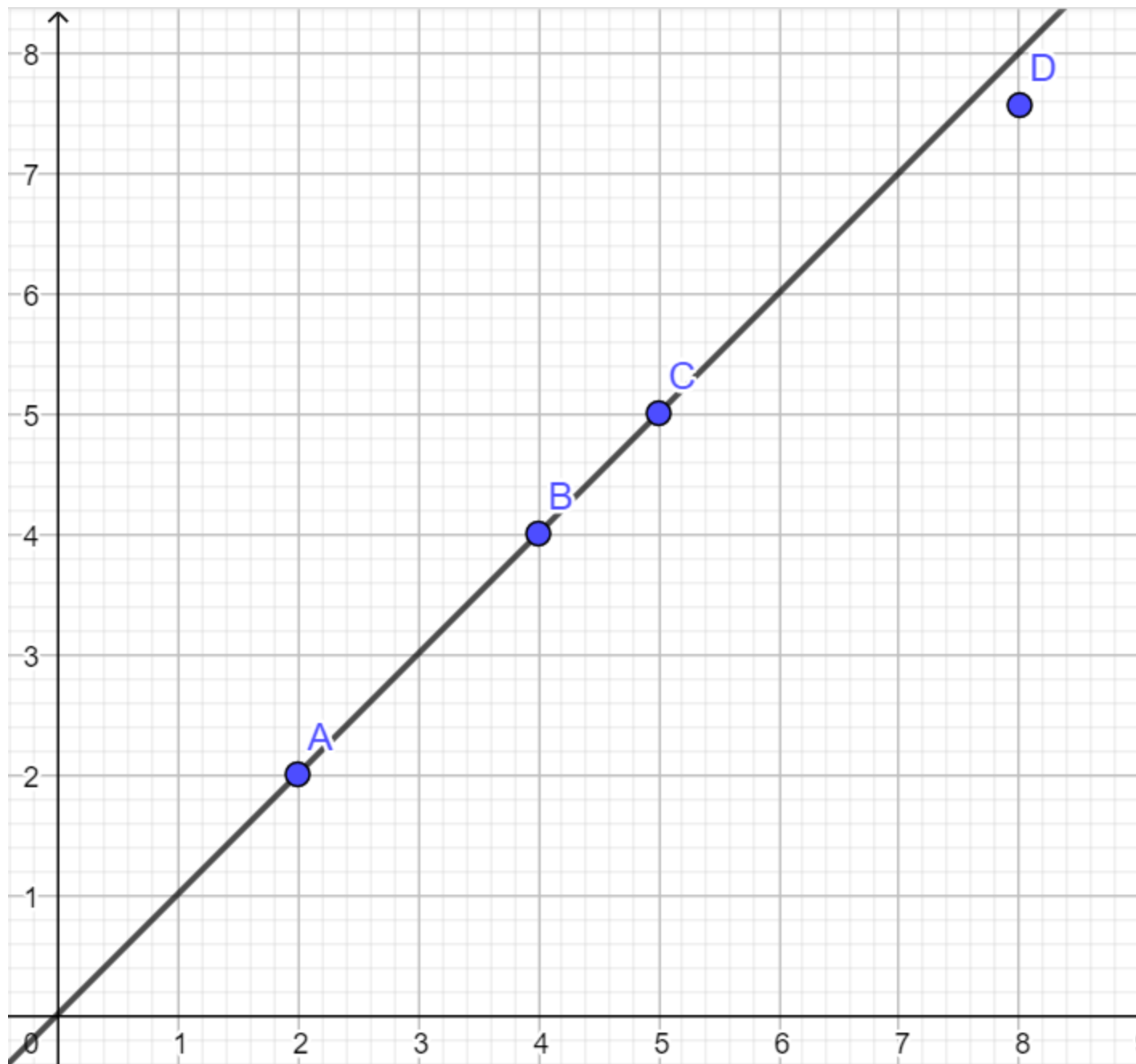
Слова как точки в пространстве



Слова как точки в пространстве







Векторное представление слов

cat

-0,035; 0,191; 0,027; -0,130; 0,220; -0,081; 0,035; -0,033;
-0,009; -0,157; 0,078; 0,121; -0,034; -0,124; 0,134; 0,003;
-0,030; 0,035; -0,034; -0,042; -0,014; 0,110; 0,092; -0,050;
0,126; 0,023; 0,089; 0,113; 0,059; 0,012; 0,203; -0,193;
-0,048; 0,039; -0,116; -0,046; 0,175; -0,082; -0,015; -0,084;
-0,113; 0,028; 0,125; -0,154; -0,138; -0,205; 0,008; -0,071;
0,040; -0,141; 0,009; -0,139; -0,015; 0,075; -0,119; -0,028;
0,231; 0,049; -0,026; 0,035; -0,030; -0,022; 0,236; -0,029;
0,217; -0,034; -0,126; -0,165; -0,072; 0,047; -0,031; 0,137;
-0,220; 0,073; 0,007; 0,048; 0,029; -0,040; -0,038; 0,001;
-0,089; 0,050; -0,042; 0,102; 0,139; 0,191; 0,057; -0,094;
-0,036; -0,098; -0,092; 0,113; 0,043; -0,020; -0,091; 0,081;
0,126; 0,018; 0,192; -0,059

Семантические аналоги для *ворона* (NOUN)

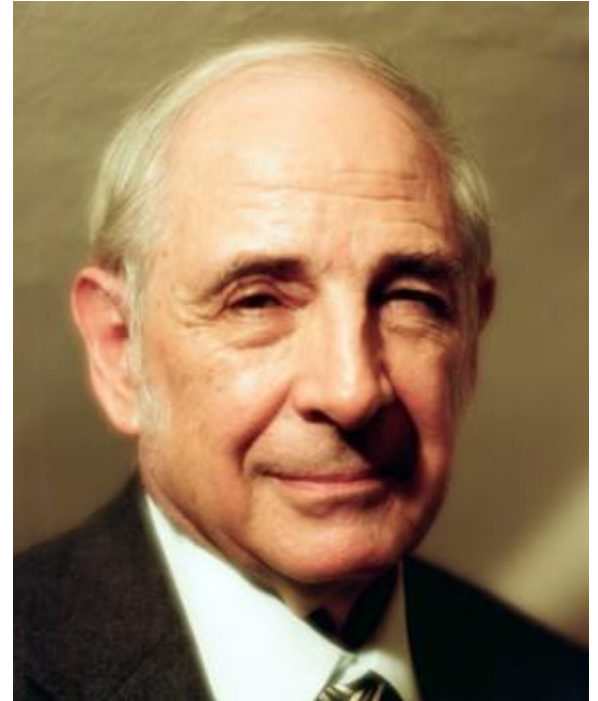
Новостной корпус

1. **цапля** 0.64
2. **трясогузка** 0.63
3. **ворон** 0.62
4. **лебедь** 0.61
5. **куропатка** 0.60
6. **скворец** 0.56
7. **каление** 0.56
8. **рученька** 0.54
9. **галка** 0.54
10. **коршун** 0.54



Китайская комната

- Джон Сёрл (род. 1932)
- Мысленный эксперимент о границах искусственного интеллекта (1980)







МОНГОЛЬСКИЙ

АНГЛИЙСКИЙ



АНГЛИЙСКИЙ

ВЕНГЕРСКИЙ



ooooooooooooooooooooooooooooo X
ooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooo
ooooooooooooooooooooo

ooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooo

Развернуть

штамповка штамповка,
штамповка, штамповка



shtampovka shtampovka, shtampovka,
shtampovka

Что нужно
компьютерной лингвистике?

РЕСУРСЫ!

Пресуппозиции и импликации

- *Король Франции лыс*
- *Вы уже перестали пить коньяк по утрам?*
- Часто импликации появляются или не появляются с придаточными предложениями
- Импликации могут быть разной силы

Пресуппозиции и импликации

- *Петя забыл, что Вася ищет новую работу.*
- *Петя заподозрил, что Вася ищет новую работу.*
- *Петя не заподозрил, что Вася ищет новую работу.*
- *Петя почему-то заподозрил, что Вася ищет новую работу.*
- *Петя зачем-то заподозрил, что Вася ищет новую работу.*
- *Петя с какой-то стати заподозрил, что Вася ищет новую работу.*
- *Мог ли Петя хотя бы на миг заподозрить, что Вася ищет новую работу?*
- *Если бы Петя заподозрил, что Вася ищет новую работу, он бы вёл себя по-другому.*

Спасибо за внимание!