Introduction to the COVID-19 Open Research Dataset

The COVID-19 Open Research Dataset (CORD-19) is a collection of over 50,000 scholarly articles - including over 40,000 with full text - about COVID-19, SARS-CoV-2, and related coronaviruses. This dataset has been made freely available with the goal to aid research communities combat the COVID-19 pandemic. It has been made available by the Allen Institute for AI in partnership with leading research groups to prepare and distribute the COVID-19 Open Research Dataset (CORD-19), in response to the COVID-19 pandemic.

During this lab you will learn how to process and analyze a subset of the articles present in the dataset, group them together into a series of clusters, and use Automated ML to train a machine learning model capable of classifying new articles as they are published.

Setup

We will start off by installing a few packages, such as `nltk` for text processing and `wordcloud`, `seaborn`, and `yellowbrick` for various visualizations.

In [1]:

```
%pip install nltk
%pip install wordcloud
%pip install seaborn
%pip install yellowbrick
```

```
Requirement already satisfied: nltk in /anaconda/envs/aiw-ai-kernel/lib/python3.8/sit
e-packages (3.7)
Requirement already satisfied: joblib in /anaconda/envs/aiw-ai-kernel/lib/python3.8/s
ite-packages (from nltk) (1.1.0)
Requirement already satisfied: regex>=2021.8.3 in /anaconda/envs/aiw-ai-kernel/lib/py
thon3.8/site-packages (from nltk) (2022.3.15)
Requirement already satisfied: tqdm in /anaconda/envs/aiw-ai-kernel/lib/python3.8/sit
e-packages (from nltk) (4.64.0)
Requirement already satisfied: click in /anaconda/envs/aiw-ai-kernel/lib/python3.8/si
te-packages (from nltk) (8.1.2)
Note: you may need to restart the kernel to use updated packages.
Requirement already satisfied: wordcloud in /anaconda/envs/aiw-ai-kernel/lib/python3.
8/site-packages (1.8.1)
Requirement already satisfied: matplotlib in /anaconda/envs/aiw-ai-kernel/lib/python
3.8/site-packages (from wordcloud) (3.5.1)
Requirement already satisfied: numpy>=1.6.1 in /anaconda/envs/aiw-ai-kernel/lib/pytho
n3.8/site-packages (from wordcloud) (1.22.3)
Requirement already satisfied: pillow in /anaconda/envs/aiw-ai-kernel/lib/python3.8/s
ite-packages (from wordcloud) (9.1.0)
Requirement already satisfied: fonttools>=4.22.0 in /anaconda/envs/aiw-ai-kernel/lib/
python3.8/site-packages (from matplotlib->wordcloud) (4.32.0)
Requirement already satisfied: python-dateutil>=2.7 in /anaconda/envs/aiw-ai-kernel/l
ib/python3.8/site-packages (from matplotlib->wordcloud) (2.8.2)
Requirement already satisfied: cycler>=0.10 in /anaconda/envs/aiw-ai-kernel/lib/pytho
n3.8/site-packages (from matplotlib->wordcloud) (0.11.0)
Requirement already satisfied: packaging>=20.0 in /anaconda/envs/aiw-ai-kernel/lib/py
thon3.8/site-packages (from matplotlib->wordcloud) (21.3)
Requirement already satisfied: kiwisolver>=1.0.1 in /anaconda/envs/aiw-ai-kernel/lib/
python3.8/site-packages (from matplotlib->wordcloud) (1.4.2)
Requirement already satisfied: pyparsing>=2.2.1 in /anaconda/envs/aiw-ai-kernel/lib/p
ython3.8/site-packages (from matplotlib->wordcloud) (3.0.8)
Requirement already satisfied: six>=1.5 in /anaconda/envs/aiw-ai-kernel/lib/python3.
8/site-packages (from python-dateutil>=2.7->matplotlib->wordcloud) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
Requirement already satisfied: seaborn in /anaconda/envs/aiw-ai-kernel/lib/python3.8/
site-packages (0.11.2)
Requirement already satisfied: matplotlib>=2.2 in /anaconda/envs/aiw-ai-kernel/lib/py
thon3.8/site-packages (from seaborn) (3.5.1)
Requirement already satisfied: pandas>=0.23 in /anaconda/envs/aiw-ai-kernel/lib/pytho
n3.8/site-packages (from seaborn) (1.4.2)
Requirement already satisfied: numpy>=1.15 in /anaconda/envs/aiw-ai-kernel/lib/python
3.8/site-packages (from seaborn) (1.22.3)
Requirement already satisfied: scipy>=1.0 in /anaconda/envs/aiw-ai-kernel/lib/python
3.8/site-packages (from seaborn) (1.8.0)
Requirement already satisfied: pillow>=6.2.0 in /anaconda/envs/aiw-ai-kernel/lib/pyth
on3.8/site-packages (from matplotlib>=2.2->seaborn) (9.1.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /anaconda/envs/aiw-ai-kernel/lib/
python3.8/site-packages (from matplotlib>=2.2->seaborn) (1.4.2)
Requirement already satisfied: python-dateutil>=2.7 in /anaconda/envs/aiw-ai-kernel/l
ib/python3.8/site-packages (from matplotlib>=2.2->seaborn) (2.8.2)
Requirement already satisfied: cycler>=0.10 in /anaconda/envs/aiw-ai-kernel/lib/pytho
n3.8/site-packages (from matplotlib>=2.2->seaborn) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in /anaconda/envs/aiw-ai-kernel/lib/
python3.8/site-packages (from matplotlib>=2.2->seaborn) (4.32.0)
Requirement already satisfied: pyparsing>=2.2.1 in /anaconda/envs/aiw-ai-kernel/lib/p
ython3.8/site-packages (from matplotlib>=2.2->seaborn) (3.0.8)
Requirement already satisfied: packaging>=20.0 in /anaconda/envs/aiw-ai-kernel/lib/py
thon3.8/site-packages (from matplotlib>=2.2->seaborn) (21.3)
Requirement already satisfied: pytz>=2020.1 in /anaconda/envs/aiw-ai-kernel/lib/pytho
n3.8/site-packages (from pandas>=0.23->seaborn) (2022.1)
```

```
Requirement already satisfied: six>=1.5 in /anaconda/envs/aiw-ai-kernel/lib/python3.
8/site-packages (from python-dateutil>=2.7->matplotlib>=2.2->seaborn) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
Requirement already satisfied: yellowbrick in /anaconda/envs/aiw-ai-kernel/lib/python
3.8/site-packages (1.4)
Requirement already satisfied: scipy>=1.0.0 in /anaconda/envs/aiw-ai-kernel/lib/pytho
n3.8/site-packages (from yellowbrick) (1.8.0)
Requirement already satisfied: matplotlib!=3.0.0,>=2.0.2 in /anaconda/envs/aiw-ai-ker
nel/lib/python3.8/site-packages (from yellowbrick) (3.5.1)
Requirement already satisfied: numpy>=1.16.0 in /anaconda/envs/aiw-ai-kernel/lib/pyth
on3.8/site-packages (from yellowbrick) (1.22.3)
Requirement already satisfied: scikit-learn>=1.0.0 in /anaconda/envs/aiw-ai-kernel/li
b/python3.8/site-packages (from yellowbrick) (1.0.2)
Requirement already satisfied: cycler>=0.10.0 in /anaconda/envs/aiw-ai-kernel/lib/pyt
hon3.8/site-packages (from yellowbrick) (0.11.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /anaconda/envs/aiw-ai-kernel/lib/
python3.8/site-packages (from matplotlib!=3.0.0,>=2.0.2->yellowbrick) (1.4.2)
Requirement already satisfied: pyparsing>=2.2.1 in /anaconda/envs/aiw-ai-kernel/lib/p
ython3.8/site-packages (from matplotlib!=3.0.0,>=2.0.2->yellowbrick) (3.0.8)
Requirement already satisfied: fonttools>=4.22.0 in /anaconda/envs/aiw-ai-kernel/lib/
python3.8/site-packages (from matplotlib!=3.0.0,>=2.0.2->yellowbrick) (4.32.0)
Requirement already satisfied: python-dateutil>=2.7 in /anaconda/envs/aiw-ai-kernel/l
ib/python3.8/site-packages (from matplotlib!=3.0.0,>=2.0.2->yellowbrick) (2.8.2)
Requirement already satisfied: packaging>=20.0 in /anaconda/envs/aiw-ai-kernel/lib/py
thon3.8/site-packages (from matplotlib!=3.0.0,>=2.0.2->yellowbrick) (21.3)
Requirement already satisfied: pillow>=6.2.0 in /anaconda/envs/aiw-ai-kernel/lib/pyth
on3.8/site-packages (from matplotlib!=3.0.0,>=2.0.2->yellowbrick) (9.1.0)
Requirement already satisfied: six>=1.5 in /anaconda/envs/aiw-ai-kernel/lib/python3.
8/site-packages (from python-dateutil>=2.7->matplotlib!=3.0.0,>=2.0.2->yellowbrick)
(1.16.0)
Requirement already satisfied: joblib>=0.11 in /anaconda/envs/aiw-ai-kernel/lib/pytho
n3.8/site-packages (from scikit-learn>=1.0.0->yellowbrick) (1.1.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in /anaconda/envs/aiw-ai-kernel/l
ib/python3.8/site-packages (from scikit-learn>=1.0.0->yellowbrick) (3.1.0)
Note: you may need to restart the kernel to use updated packages.
```

We'll first download stopwords and the Punkt tokenizer models present in the `nltk` package, in order to be able to process the articles

In [2]:

```python
import nltk

nltk.download('punkt')
nltk.download('stopwords')
```

```
[nltk_data] Downloading package punkt to /home/azureuser/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     /home/azureuser/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[2]:

```
True
```

We'll also import the rest of the modules needed in this notebook, and do a quick sanity-check on the Azure ML SDK version

In [3]:

```python
import os
import json
from string import punctuation
```

```
import pandas as pd
import seaborn as sns
sns.set_palette('Set2')
from wordcloud import WordCloud
import matplotlib.pyplot as plt
from yellowbrick.cluster import KElbowVisualizer
from gensim.models.doc2vec import Doc2Vec, TaggedDocument
from sklearn.manifold import TSNE
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
from sklearn.cluster import KMeans, SpectralClustering, DBSCAN, Birch, AgglomerativeCl
from sklearn.metrics import roc_auc_score
from nltk import word_tokenize, sent_tokenize
from nltk.stem import SnowballStemmer, PorterStemmer

from azureml.core import Workspace, Datastore, Dataset, VERSION

print("Azure ML SDK Version: ", VERSION)
```

```
Azure ML SDK Version:  1.39.0
```

Load the Covid-19 data

CORD-19 has been uploaded to an Azure Storage Account, we will connect to it and use it's API to download the dataset locally. Also please make sure to update the storage account name from the labguide in the below cell.

In [4]:

```
covid_dirname = 'covid19temp'

cord19_dataset = Dataset.File.from_files('https://aiinadaystorage800633.blob.core.wind
mount = cord19_dataset.mount()

covid_dirpath = os.path.join(mount.mount_point, covid_dirname)
```

Display a sample of the dataset (top 5 rows).

In [5]:

```
mount.start()

# Load metadata.csv, as it contains a list of all the articles and their corresponding
metadata_filename = os.path.join(covid_dirpath, 'metadata.csv')

metadata = pd.read_csv(metadata_filename)
metadata.sample(5)
```

```
/tmp/ipykernel_5158/188001440.py:6: DtypeWarning: Columns (13,14) have mixed types. S
pecify dtype option on import or set low_memory=False.
  metadata = pd.read_csv(metadata_filename)
```

Out[5]:

	cord_uid	sha	source_x	title	
48221	qfcoolj9	NaN	Medline	Middle East respiratory syndrome coronavirus i...	10.1128/
113519	pu1jci26	9deb7792c0ff4705670a156adddb64a411384bde	Medline; PMC	Using echocardiography to guide the treatment ...	10.118 02
117707	nxoj330c	NaN	Medline; WHO	Déjà Vu or Jamais Vu? How the Severe Acute Res...	10.2214/aj
43435	1r6eg1hh	NaN	Medline	[Rapid review of the use of community-wide sur...	10.211
79284	znafeueo	NaN	Medline	Human internal jugular valve M-mode ultrasound...	

Some of the articles do not have any associated documents, so we will filter those out.

In [6]:
```
metadata_with_docs = metadata[metadata['pdf_json_files'].isna() == False]

print(f'Dataset contains {metadata.shape[0]} entries, out of which {metadata_with_docs
```

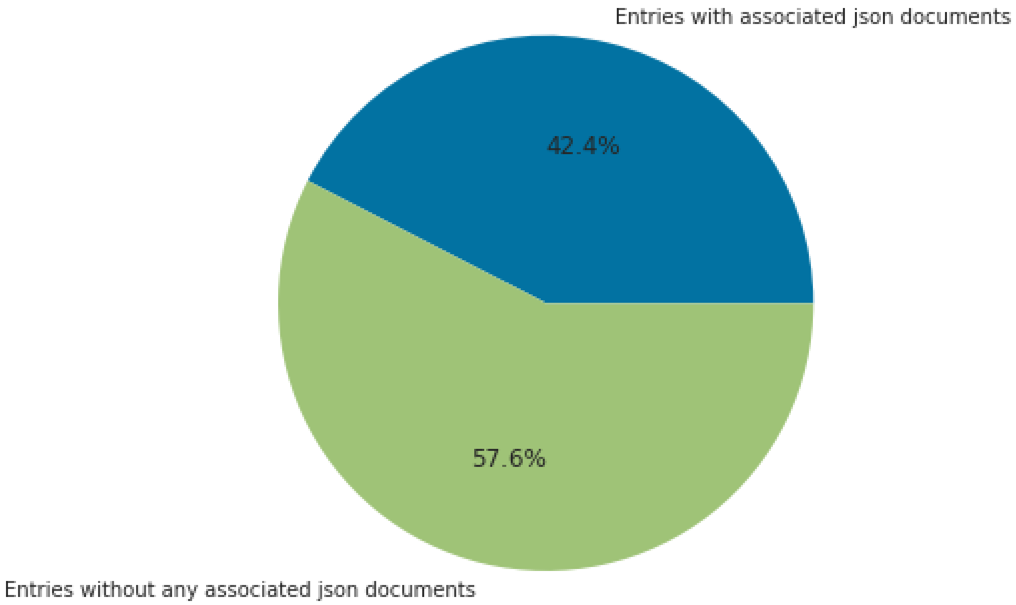
```
Dataset contains 134206 entries, out of which 56962 have associated json documents
```

Display the percentage of items in the dataset that have associated JSON documents (research papers that have extra metadata associated with them).

In [7]:
```
doc_counts = [metadata_with_docs.shape[0], metadata.shape[0] - metadata_with_docs.shap
doc_labels = ['Entries with associated json documents', 'Entries without any associate

fig, ax = plt.subplots()
ax.pie(doc_counts, labels=doc_labels, autopct='%1.1f%%')
ax.axis('equal')
plt.show()
```

```
findfont: Font family ['sans-serif'] not found. Falling back to DejaVu Sans.
findfont: Generic family 'sans-serif' not found because none of the following familie
s were found: Arial, Liberation Sans, Bitstream Vera Sans, sans-serif
findfont: Font family ['sans-serif'] not found. Falling back to DejaVu Sans.
findfont: Generic family 'sans-serif' not found because none of the following familie
s were found: Arial, Liberation Sans, Bitstream Vera Sans, sans-serif
```



Investigate individual items

Let's load an example entry from the dataset.

In [8]:

```
# Change the document index in order to preview a different article
DOCUMENT_INDEX = 0
example_entry = metadata_with_docs.iloc[DOCUMENT_INDEX]

filepath = os.path.join(covid_dirpath, example_entry['pdf_json_files'])
print(f'Document local filepath: {filepath}')
print(f'Document local filepath: {covid_dirpath}')

filepath = covid_dirpath + '/comm_use_subset/pdf_json/02a009e42054081b441d0f4b203679c4
print(f'Document local filepath: {filepath}')
```

```
Document local filepath: /tmp/tmpemaxbzly/covid19temp/document_parses/pdf_json/d1aafb
70c066a2068b02786f8929fd9c900897fb.json
Document local filepath: /tmp/tmpemaxbzly/covid19temp
Document local filepath: /tmp/tmpemaxbzly/covid19temp/comm_use_subset/pdf_json/02a009
e42054081b441d0f4b203679c4b0cae38d.json
```

Next, we will display the list of elements that are available for the selected document.

In [9]:

```
try:
    with open(filepath, 'r') as f:
        data = json.load(f)

except FileNotFoundError as e:
    # in case the mount context has been closed
    mount.start()
    with open(filepath, 'r') as f:
        data = json.load(f)

print(f'Data elements: { ", ".join(data.keys())}' )
```

```
Data elements: paper_id, metadata, abstract, body_text, bib_entries, ref_entries, bac
k_matter
```

Please make sure to update the storage account name from the labguide in the below cell.

In [10]:

```
from azureml.core import  Dataset
cord19_dataset = Dataset.File.from_files('https://aiinadaystorage800633.blob.core.wind
mount = cord19_dataset.mount()
```

View the full text version of the document.

In [11]:

```
for p in data['body_text']:
    print(p['text'], '\n')
```

```
Hantaviruses are negative-sense RNA viruses transmitted to humans from small animal h
osts. Different viral species are associated with one of two disease syndromes: hemor
rhagic fever with renal syndrome (HFRS), or hantavirus pulmonary syndrome (HPS) [1] .
Hantaan virus a1111111111 a1111111111 a1111111111 a1111111111 a1111111111 (HTNV), pri
marily found in Asia, is among the most prevalent HFRS-causing hantaviruses with a ca
se fatality rate of between 1-15% [2] . Puumala virus (PUUV) causes most HFRS cases i
n Europe, though its case fatality rate is lower at <1% [3, 4] . There are currently
no FDA licensed vaccines or therapeutics for either HFRS or HPS [5] .

The Syrian hamster (Mesocricetus auratus) is the typical animal used to model hantavi
rus infection and disease. Andes virus (ANDV), an HPS-causing hantavirus, causes leth
al disease in immunocompetent hamsters [6] , while numerous other HPS-causing hantavi
ruses including Sin Nombre Virus (SNV) and Choclo virus cause lethal disease in hamst
ers immunosuppressed with dexamethasone and cyclophosphamide [7, 8] . In contrast to
HPS-causing hantaviruses, exposure of hamsters to HFRS-causing hantaviruses such as H
TNV, PUUV, Dobrava (DOBV) and Seoul (SEOV) leads to asymptomatic infection, despite v
iral dissemination, even when immunosuppressed (Hooper Lab, unpublished data) [8] [9]
[10] [11] . In these studies hamsters were exposed to high doses of HTNV and PUUV, fa
r exceeding the infectious dose 99% (ID 99 ) for the virus. Development and character
ization of a uniformly infective, low-dose challenge model, enhances the hamster mode
l's usefulness in vaccine and therapeutic testing. In this report we present a low-do
se hamster infection model for both HTNV and PUUV infected animals.

Ferrets (Mustela putorius furo) have become a popular animal model for a number of re
spiratory pathogens including influenza [12] , coronavirus [13] , Nipah virus [14] ,
and morbillivirus [15] , due to the similarity in lung physiology to humans. In addit
ion, they have recently been described as a disease model of two hemorrhagic fever vi
ruses, Bundibugyo virus and Ebola virus [16, 17] , supporting viral replication witho
ut prior adaptation. Most hantavirusrelated human disease occurs by aerosolized trans
mission of the virus from the excreta or secreta of infected rodents [18, 19] , a mod
el of viral infection for which the ferret is well suited. In this study we demonstra
te that ferrets are capable of being infected by high titers of HTNV and PUUV, though
aside from gradual weight loss infected animals exhibit no clinical symptoms or impai
red renal function.

It has been established that infection of rhesus macaques (Macaca mulatta) with HFRSc
ausing hantaviruses (DOBV, SEOV, HTNV, and PUUV) leads to asymptomatic infection and
seroconversion [9] , while infection of cynomolgus macaques (Macaca fascicularis) wit
h PUUV leads to a mild disease characterized by lethargy, mild proteinuria and hematu
ria, and kidney pathology, similar to mild HFRS in humans [20] . However, the macaque
s' large size and cost limits their usefulness in therapeutic studies, especially whe
n test article availability is limited, as is often the case in passive transfer stud
ies. The common marmoset (Callithrix jacchus) is becoming more popular for infectious
disease studies. Its genetic similarity to humans, cost, relative safety, and small s
ize make it an attractive alternative to traditional non-human primate species [21] .
Marmosets have been used as a disease model for other viral agents including Dengue v
irus [22] , Hepatitis C virus [23] , influenza virus [24] , Lassa fever virus [25] ,
orthopox viruses [26] [27] [28] , Rift Valley Fever virus [29] , Eastern Equine Encep
halitis virus [30] , and filoviruses [31] . In this study we demonstrate that exposur
e of marmosets to HTNV leads to asymptomatic infection characterized by high levels o
f neutralizing antibodies. This is the first report of hantavirus infection in marmos
ets.

Medical countermeasures are products including biologics (e.g., vaccines and antibodi
es) and small molecule drugs that can be used to prevent or combat infectious disease
outbreaks. This study presents three animal models of HTNV infection, and two models
of PUUV infection that can be used to evaluate the efficacy of medical countermeasure
that are intended to prevent or mitigate infection (e.g., vaccines) by these viruses
through induction of sterile immunity.
```

```
HTNV strain 76-118 [32] , PUUV strain K27 [33] , and PUUV strains Beaumont, and Seloi
gnes (gifts of Piet Maes, Leuven, Beligium) were propagated in Vero E6 cells (Vero C1
008, ATCC CRL 1586) in T-150 flasks and cEMEM media (Eagle 0 s minimal essential medi
um with Earle 0 s salts (EMEM) containing 10% heat inactivated fetal bovine serum, 20
0 μM glutamine, 1% non-essential amino acids, 10 mM HEPES pH 7.4; and antibiotics [pe
nicillin (100 U/ml), amphotericin B (250 μg/ml), and gentamicin (50 mg/ml)]. Virus wa
s collected from infectedmonolayer supernatants. Cell debris was removed by low speed
centrifugation (2500 rpm in a table top centrifuge). HTNV and PUUV strain K27 were tw
ice plaque purified according to published methods [6] . Virus stocks were aliquoted
and stored at -60˚C or colder. Virus identity has been confirmed by sequencing of the
stocks.

Female Syrian hamsters 6-8 wks of age (Envigo, Indianapolis, IN) were anesthetized by
inhalation of vaporized isoflurane using an IMPAC 6 veterinary anesthesia machine. On
ce anesthetized, animals were injected with the indicated concentration of virus dilu
ted in PBS. Intramuscular (i.m.) injections (in the caudal thigh) consisted of 0.2 ml
delivered with a 1ml syringe with a 25-gauge, 5/8 in needle. Intranasal (i.n.) instil
lation consisted of 50 μl total volume delivered as 25 μl per nare with a plastic pip
ette tip. Blood sampling from the vena cava occurred under previously stated methods
of anesthesia, and was limited to 7% of a hamster's total blood volume per week. At t
ime of arrival animals were randomized into experimental groups. Animals were housed
in small animal pans, not exceeding four animals to a pan, in a climate and humidity
controlled animal biosafety level 3 (ABSL-3) with a 12-hour light/dark cycle. Animals
had pelleted food and water provided ad libitum. Enrichment in the form of toys, nest
ing material and supplemental treats was provided. Humane endpoint conditions were es
tablished as decreased mobility (inability to obtain food and water) and subdued resp
onse to stimulation, and animals were monitored daily during the experiment. As infec
ted animals did not become ill, these criteria were not met and animals were euthaniz
ed by terminal blood collection from the heart after administration of Ketamine-acepr
omazine-xylazine (KAX)(USAMRIID, Fort Detrick MD) and prior to intracardiac injection
with pentobarbital sodium (USAMRIID) at the end of the study. Due to lack of illness
no pain relief, aside from anesthesia during procedures, was required.

confidence intervals reaching out to 0.7log 10 (approximately 5-fold) under monotome
assumptions of response profiles for intermediate doses. As our initial selection of
dosages did not meet the desired infection rate span, some dosage groups were repeate
d leading to 20 hamsters per group. In the serial euthanasia study three hamsters per
group were used. This is the minimum number required to provide collection of suffici
ent samples for detection of antibodies and viral kinetics in tissues. For each exper
iment pre-sera from animals served as negative control.

Adult, female neutered and descented ferrets (Marshall Farms, North Rose, NY), were a
nesthetized by inhalation of vaporized isoflurane using an IMPAC 6 veterinary anesthe
sia machine, or i.m. injection of Telazol (Zoetis, Parsippany, NJ). Injections (i.m.
and i.n.) and blood sampling were conducted under the same condition as hamsters. Int
raperitoneal (i.p.) injections consisted of 1 ml delivered with a 3 mL syringe and a
23-guage needle. Microchips (BMDS, Seaford, DE) were used to identify and ascertain t
emperature during ferret experiments. In the first ferret challenge study faulty chip
s lead to inaccurate temperature readings and were only used for identification purpo
ses. Animals were randomized upon receipt into experimental groups. Ferrets were soci
ally housed in metal caging, two to a cage, with sight lines to additional animals in
the study, in a climate and humidity controlled ABSL-3 with 12/hour light and dark cy
cles. Each cage had a nesting box with bedding material, and numerous tubes and shelf
s for play. Ferrets had access to pelleted food supplemented with treats and potable
water, through an automated watering system. Enrichment in the form of manipulada (tu
bes, balls, mirrors) and food was provided. Animals were observed daily by trained pe
rsonnel in addition to general husbandry assessments. Humane euthanasia criteria, def
ined as both dyspnea, loss of mobility (to obtain food and water) and >20% weight los
s. At the end of the experiment, terminal blood samples were collected from the heart
after administration of KAX and prior to intracardiac injection with pentobarbital so
```

```
dium. No pain relief, aside from anesthesia during procedures, was used.

When disease occurs independently in each of four ferrets with 50% probability, the e
xperiment will have odds about 9:1 in favor of producing at least one diseased ferre
t. Conversely failure to observe any diseased ferret in a group of four will yield a
95% confidence interval extending from 0-50%. That is, with 95% confidence it will be
admitted that the true disease rate may be 50% or less. For this reason groups of fou
r ferrets were used for the experiments. For each experiment pre-sera from animals se
rved as a negative control.

Adult marmosets weighing over 300g were anesthetized by inhalation of vaporized isofl
urane using an IMPAC 6 veterinary anesthesia machine. Once anesthetized, animals were
injected with the indicated concentration of virus diluted in PBS. I.m. injections (i
n the caudal thigh) consisted of 0.2 ml delivered with a 1ml syringe with a 25-gauge,
5/8 in needle. Blood sampling from a femoral vein occurred under previously stated me
thods of anesthesia, and were limited to 7% of each marmoset's total blood volume per
week. At time of euthanasia, terminal blood samples were collected from the heart aft
er anesthetization by i.m. injection of Telazol and prior to intracardiac administrat
ion of pentobarbital sodium. Animals were housed in containment as previously describ
ed [26] . In brief, animals were singly housed in metal cages meeting current standar
ds in a climate and humidity controlled room. Animals were fed pelleted food suppleme
nted with fruits and treats daily, and were provided potable water through an automat
ic watering system ad libitum. Enrichment in the form of manipulada (i.e. toys, metal
mirrors), foraging devices, treats, and fruit were provided daily. Animals were obser
ved daily by trained personnel in addition to general husbandry assessments. Animals
were observed daily by trained personnel in addition to general husbandry assessment
s. Animals found moribund (defined as labored breathing, decreased food consumption,
persistent prostration and moderate unresponsiveness) would be euthanized under human
e endpoint criteria, however, as animals did not become ill during the study this cri
teria was not met. No additional pain relief, aside from anesthesia during procedure
s, was necessary. All work was performed in an ABSL-3 laboratory.

The marmoset study requires a sample size of 3 for adequate power to determine if the
incidence of seroconversion is significantly greater than that which would be expecte
d in the population. This sample size will allow the experimenter to detect seroconve
rsion in at least 2 of 3 animals (66%) versus the expected population constant of <1%
at a 95% confidence level using a one-tailed binomial test for proportions.

The enzyme-linked immunosorbent assay (ELISA) used to detect nucleocapsid protein (N)
specific antibodies (N-ELISA) was described previously [34] . Species-specific second
ary antibodies were used at the following concentrations: peroxide-labeled anti-hamst
er (1:10,000) (Sera Care, Gaithersburg, MD), peroxide-labeled anti-ferret (1:5,000)
(Sigma Aldrich, St. Louis, MO), and alkaline phosphatase conjugated anti-monkey (1:1,
000) (MilliporeSigma, St. Louis, MO). Assays using peroxide labeled antibodies were d
eveloped with TMB microwell peroxidase substrate (Sera Care) at an absorbance of 450
nm; assays using alkaline phosphatase conjugated antibodies were developed with p-nit
rophenyl phosphate (PNPP) (ThermoFisher Scientific, Waltham, MA) at 405 nm. A sample
was considered positive if its peak optical density (OD) value was greater than eithe
r 0.025 or the background value (the average of three negative controls + 3 times the
ir standard deviation), whichever was higher. The specific OD sum is the summation of
all values greater than background and represents the area under the titer curve.

PRNT assays were performed as previously described with minor modifications [35-37].
HTNV-infected monolayers were fixed 7 days post-infection, while PUUV-infected monola
yers were fixed 10 days post infection by 2 mL of 10% formalin per well. Immunostaini
ng was performed as previously described [38] . All sera samples were assayed in dupl
icate beginning at a 1:20 final dilution. PRNT 50 values represent the reciprocal dil
ution at which the serum neutralizes 50% of the virus.

Approximately 200 mg of organ tissue was homogenized in 1.0 mL of TRIzol (ThermoFishe
r, Waltham, MA) reagent using M tubes on the gentleMACS (Miltenyi Biotec,Auburn, CA)
dissociation system on the RNA setting. RNA was extracted from TRIzol per manufacture
r's protocol. A Nanodrop 8000 was used to determined RNA concentration, which was the
n raised to either 100 ng/μL or 1,000 ng/μL in UltraPure distilled water (Thermofishe
r). Real-time PCR was conducted on a BioRad CFX thermal cycler using either an Invitr
ogen Power SYBR Green RNA-to-Ct one-step kit (Thermofisher) or Brilliant II QRT-PCR 1
-Step Master Mix (Agilent, Santa Clara, CA) according to the manufacturer's protocol
s. For spiked assays the master mix was spiked with either HTNV or PUUV viral RNA pri
or to addition to the samples. For HTNV, primer sequences were 594F 5'-AAG CAT GAA GG
C AGA AGA GAT -3' and

T-25 flasks of one week old Vero E6 cells were infected with 50 μL of urine plus an a
dditional 450 μL of cEMEM media. After a 1 hr adsorption at 37˚C with 5% CO 2 , the v
olume was raised to 3.5 mL. On Day 4 post infection supernatant was collected and fro
zen down, 500 μL of which was used to infected fresh Vero E6 cells at a later time po
int. After a 1 hr adsorption at 37˚C with 5% CO 2 , the volume was raised to 3.5 mL.
On days 7, 11, 14, 17, 21 and 28 1.2 mL of culture supernatant was collected and froz
en down. The volume of cEMEM in the flask was raised to 3.5 mL with fresh media.

Approximately 200 μg of organ tissue were homogenized in 1 mL of cEMEM media using M
tubes on the gentleMACS dissociation system on the RNA setting. Plaque assays using u
rine, sera, or organ homogenate were then performed beginning at the 1:10 dilution as
described in [6, 41] with minor modifications. For spiked plaque assays the protocol
was identical except for equivalent amounts of virus being spiked into either media a
lone (control), or the 1:10-1:1,000 dilution of organ homogenate. HTNV-infected monol
ayers were fixed 7 days postinfection, while PUUV-infected monolayers were fixed 10 d
ays post infection by 2 mL of 10% formalin per well. Immunostaining was performed as
previously described [38] .

Following euthanasia, necropsies were performed. Samples were collected aseptically f
or the virology studies described above. For the hamsters and ferrets, samples of the
following were collected: heart, lung, liver, spleen, kidney, brain, and urine. In ad
dition, ferrets had samples of intestine, adrenal gland, pituitary gland, and eye wer
e collected. Samples of the following were collected from the marmosets: heart, lung,
liver, spleen, kidney, intestine, and brain. After the virology samples had been coll
ected, all major internal organs in each animal were also sampled for histology.

Tissues were fixed in 10% neutral-buffered formalin for �21 days. Tissues were then t
rimmed, processed under vacuum through increasing concentrations of alcohols, and emb
edded in paraffin. Paraffin embedded tissue sections of 5-6 μm were cut and mounted o
n glass slides, stained with hematoxylin and eosin (H&E), and mounted under a glass c
overslip for routine histologic evaluation. The paraffin-embedded tissues used for pr
oducing the H&E-stained histology slides were also utilized for immunohistochemistry
(IHC) studies in the hamsters. Immunolocalization of HTNV in tissues was performed wi
th an immunoperoxidase procedure (horseradish peroxidase EnVision system; Dako) accor
ding to the manufacturer's directions. The primary antibody was an anti-HTNV nucleoca
psid rabbit polyclonal antibody diluted 1:3,500 (ferret) or 1:5,000 (hamster) (BEI Re
sources, Manassas, VA). After deparaffinization and peroxidase blocking, tissue secti
ons were pretreated with proteinase K for 6 min at room temperature, rinsed, and then
covered with primary antibody and incubated at room temperature for 30 min. They were
rinsed, and then the peroxidase-labeled polymer (secondary antibody) was applied for
30 min. Slides were rinsed, and a substrate-chromogen solution (3,3'-diaminobenzidin
e; Dako, Santa Clara, CA) was applied for 5 min. The substrate-chromogen solution was
rinsed off the slides, and the slides were stained with hematoxylin and rinsed. The s
ections were dehydrated and cleared with xyless, and then a coverslip was placed.

On the indicated days, anesthetized ferrets were injected i.p. with water soluble Cyp
(Baxter Health Care Corporation, Deerfield, IL) with the indicated dosages per body w
eight of drug diluted in sterile phosphate-buffered saline (PBS), pH 7.4. In the firs
t experiment ferrets were administered a loading dose of 30 mg/kg on Day 41 post infe
```

```
ction, with maintenance doses of 30 mg/kg administered every other day until euthanas
ia. In the second experiment, ferrets were administered a loading dose of 30 mg/kg on
Day -1, and a maintenance dose of 10 mg/kg on Day 1, 3, 11, and 13. Administration of
Cyp was discontinued between days 3 and 11 due to secondary infection. To combat the
infection (rapid onset of fever and weight loss), ferrets were treated with 5mg/kg i.
m. enrofloxicin(Norbrook Laboratories, Overland Park, KS) twice daily per veterinaria
n instructions starting on Day 4. Beginning on Day 11 with the resumption of immunosu
ppression ferrets were treated prophylactically with 10 mg/kg i.m. enrofloxicin once
daily).

Blood samples of 0.5 mL were collected in lithium heparin capillary blood collection
tubes and analyzed using an Advia 120 hematology analyzer (Software version 3.1.8.0-M
S). Per manufacturer's recommendation, the dog setting was used for the complete bloo
d count (CBC) and the guinea pig setting was used for the white blood cell differenti
al (WBC) in hamsters. For ferrets, the dog setting was used for both the CBC and WBC.

The PsVNA used to detect neutralizing antibodies in sera was described previously [4
2, 43] . This is a replication-restricted, recombinant vesicular stomatitis virus (rV
SV � ΔG) expressing luciferase, which is pseudotyped with the Hantaan glycoprotein. F
irst, heat-inactivated sera was diluted 1:10, followed by five-fold serial dilutions
that were mixed with an equal volume of Eagle 0 s minimum essential medium with Earle
0 s salts and 10% fetal bovine sera containing 4000 fluorescent focus units of Hantaa
n pseudovirions. This mixture was incubated overnight at 4˚C. Following this incubati
on, 50 µl was inoculated onto Vero cell monolayers in a clear bottom, black-walled 96
-well plate in duplicate. Plates were incubated at 37˚C for 18-24 hr. The media was d
iscarded and cells were lysed according to the luciferase kit protocol (Promega, Madi
son, WI). A Tecan M200 Pro was used to acquire luciferase data. The values were graph
ed using GraphPad Prism (version 7) and used to calculate the percent neutralization
normalized to cells alone and pseudovirions alone as the minimum and maximum signals,
respectively. The percent neutralization values for duplicate serial dilutions were p
lotted. Fifty percent PsVNA (PsVNA 50 ) titers were interpolated from 4-parameter cur
ves, and geometric mean titers were calculated.

Blood was collected in serum separator tube, and spun at 500x g to isolate sera. Ferr
et sera was analyzed on the Piccolo comprehensive metabolic panel, and marmoset sera
was analyzed on the Piccolo general chemistry 13 panel per manufacturer's instruction
s (Abaxis Global Diagnostics, Union City, CA).

Urine was expressed from anesthetized ferrets and analyzed by urinalysis regent strip
s (VWR, Radnor, PA).

A Bayesian probit model was used to estimate 95% highest posterior density intervals
for a 50% and 95% infectious dose calculation. Student's t-test and Mann-Whitney test
s were used to compare white blood cell levels pre-and post-Cyp administration. P-val
ues of <0.05 were considered significant. Standard deviation of data was assessed to
ensure data was normally distributed prior to use of Student's t-test. Analyses were
conducted using GraphPad Prism (version 7); Bayesian analyses were performed using SA
S.

We have previously demonstrated that Syrian hamsters are capable of being infected by
HFRScausing hantaviruses, but they do not develop any signs of clinical disease [9] .
To develop standard models of HTNV and PUUV infection for future evaluation of vaccin
es and medical countermeasures, groups of between 10 and 20 hamsters were exposed to
serial ten-fold dilutions of either HTNV or PUUV by either the i.m. or i.n. route (fr
om 2-20,000 PFU HTNV or 0.2-20,000 PFU PUUV). Between 28-35 days post infection, hams
ters were terminally bled with infection status monitored by N-ELISA titers (Fig 1) .
From these data the doses required to infect 50% (ID 50 ) and 99% (ID 99 ) were calcu
lated ( Table 1 ).

To further characterize a low-dose standard hamster infection model for HFRS-causing
```

```
hantaviruses a hamster serial sacrifice study was performed. Hamsters were challenged
with either 10 PFU (~3 ID 99 ) HTNV i.m., 500 PFU (~1.5 ID 99 ) HTNV i.n., 1,000 PFU
(~1 ID 99 ) PUUV i.m., or 1,000 PFU (~1.5 ID 99 ) PUUV i.n. On various days post infe
ction, groups of three hamsters were euthanized to monitor viral and serological para
meters.

Seroconversion occurred, at least partially, by Day 17 post HTNV infection and Day 24
post PUUV infection. Seroconversion on Day 28 post HTNV i.m. infection was incomplet
e, though viral genome was recovered from all hamsters euthanized that day indicating
a productive HTNV infection occurred (Fig 2) . To confirm seroconversion, all hamster
s euthanized on days 17, 24, and 28 were assayed for neutralizing antibodies by PRNT
regardless of N-ELISA seroconversion status. All HTNV infected hamsters with N-ELISA
titers had neutralizing antibodies as measured by PRNT 50 , with all but one having f
ull neutralization of virus at a 1:20 dilution of sera. (S1A and S1B Fig) . Similarl
y, all PUUV hamsters with N-ELISA titers had neutralizing antibodies as measured by P
RNT 50 , though three of the five PUUV i.n. challenged hamsters did not have complete
neutralization of the virus at a 1:20 dilution of sera (S1C and S1D Fig) . Two of thr
ee PUUV i.m. challenged hamsters that had not seroconverted on Day 17 post infection
had low levels of neutralizing antibodies, while none of the PUUV i.n. challenged ham
sters that were seronegative by N-ELISA had neutralizing antibodies. For both HTNV an
d PUUV infected hamsters, infection via the i.m. route lead to a more robust neutrali
zing antibody response than then i.n.route.

The kinetics of HTNV and PUUV infection in the heart, lung, liver, spleen, kidney, an
d brain were monitored by both RT-PCR and plaque assay (Figs 3 and 4) . HTNV infected
by either the i.m. or i.n. route resulted in a persistent infection. High levels of v
iral genome were detected in the heart, lung, kidney and brain of HTNV infected hamst
ers beginning on either Day 11 or 17 post infection (Fig 3A, 3B , 3E and 3F). Infecti
on of the kidney and brain was found in all examined hamsters beginning on either Day
17 or 24 post infection, while the high titers of viral genome detected in the heart
and lung were present in only one or two hamsters at each time point. Low levels of v
iral genome were detected between days 11 and 24 post infection in the liver of HTNV
i.m. but not i.n. infected hamsters (Fig 3C) . Hardly any viral genome was detected i
n the spleen (Fig 3D) . To confirm the lack of viral genome in the spleen was not due
to the presence of inhibitors all spleen samples from i.m. and i.n. infected hamsters
were spiked with HTNV prior to RT-PCR. No significant inhibition of the spiked RNA wa
s noted, indicating that HTNV infection does not result in virus dissemination to the
spleen (S2A and S2B Fig) .

The detection of PUUV in organs was transient after i.m. infection, occurring between
Day 11-17 for the heart, liver, kidney and brain, with no virus detected in the splee
n (Fig 3A, 3C , 3E and 3F). Viral genome was more persistent in the lung where it was
detected in at least one of three hamsters on/after Day 11 post infection (Fig 3B) .
A small amount of viral genome detected in the brain of a PUUV i.n. infected hamster,
28 days post challenge, is the only viral genome detected in any organ at any time po
int post PUUV i.n. infection (Fig 3F) . Serum viremia was detected in hamsters challe
nged with HTNV i.m. between 11 and 28 days post infection, though the presence of vir
us was sporadic except for Day 24. Serum viremia was only detected in two hamsters ch
allenged with HTNV i.n., one on Day 11 and one on Day 17. No serum viremia was detect
ed in hamsters challenged with PUUV by either route (Fig 3G) . Similarly, the presenc
e of viral genome was sporadically detected in the urine of Three HFRS-associated han
tavirus infection models HTNV infected hamsters between days 17-28, but was not detec
ted in PUUV challenged hamsters ( Fig 3H) .

In HTNV i.m. infected hamsters, infectious virus was first detected in the liver begi
nning at Day 11, and in the lung, liver, spleen, and kidney at Day 17 post infection
(Fig 4B, 4C, 4D and 4E). With the exception of the kidney, in which infectious virus
was recovered from every hamster after 17 days post infection, infectious virus was r
ecovered from the lung, liver, brain and spleen in only a portion of hamsters at each
time point. No infectious virus was detected in the heart at any time point (Fig 4A)
```

```
. Virus recovery from HTNV i.n. infected hamsters was markedly lower, with infectious
virus being detected only in the kidney on Day17 and 28 post Three HFRS-associated ha
ntavirus infection models infection (Fig 4E) . No infectious virus was recovered from
the organ of any PUUV infected hamster (Fig 4) .

To confirm that the lack of infectious virus in the spleen was not due to the presenc
e of inhibitors select samples were spiked with HTNV virus prior to a plaque assay. N
o significant inhibition of the spiked RNA was noted, confirming the RT-PCR results t
hat that HTNV infection does not result in viral dissemination to the spleen (S2C and
S2D Fig) .

No infectious virus was detected in the urine for any hamster tested, even those that
were RT-PCR positive for HTNV viral genome. In human ANDV infected patients with ANDV
antigen positive urine, samples had to be cultured in Vero E6 cells for between 16-22
days post infection before infectious virus was detected [44] . Patients infected wit
h HFRS hantaviruses exhibit leukocytosis and thrombocytopenia during infection [45, 4
6] . At every time point post infection, EDTA-treated whole blood from HTNV and PUUV
exposed hamsters was evaluated to determine if changes in white blood cell count or p
latelets occurred. No changes were observed (S4 Fig). To further characterize diseas
e, tissue sections from HTNV i.m. infected hamsters were analyzed by IHC and H&E to a
ssess viral localization and any pathologic changes (S5 Fig). No significant histopat
hological findings were noted in the kidney or brain. Splenic follicular lymphoid hyp
erplasia and hepatic extramedullary hematopoiesis in the liver were each noted in thr
ee hamsters. Both are seen in animals from later time points in the study (days 17-2
8) and likely represent a reaction to HTNV infection, though other unidentified antig
enic stimuli cannot be ruled out. Five hamsters between days 4 and 17 post infection
exhibited a minimal to mild inflammation of the pericardium, characterized by a mixed
lymphoplasmacytic histiocytic and neutrophilic infiltrate. Occasional macrophages in
the inflammatory infiltrate within the pericardium are immunopositive suggesting a po
ssible association with HTNV; however, no evidence of cardiac tissue injury is associ
ated with the presence of HTNV antigen. Beginning on Day 4 post infection 66% (12/18)
hamsters exhibit minimal (likely subclinical) respiratory lesions consisting of inter
stitial neutrophilic and histiocytic infiltrates in the lungs, with 44% (8/18) also e
xhibiting minimal amounts of alveolar edema. An additional hamster had minimal alveol
ar edema but not pulmonary infiltrates. Such findings suggest a response to antigenic
stimulus and the presence of immunopositive endothelial cells, pneumocytes and alveol
ar macrophages suggest a response to HTNV infection.

Immunohistochemistry staining was uniformly negative in all tissues on Day 1 post cha
llenge. Between Day 4 and 11 post challenge minimal antigen was sporadically detected
in macrophages, pneumocytes and endothelial cells. By Day 17 post challenge mild to m
oderate levels of antigen were observed in endothelial cells and choroid ependymal ce
lls in the brain, in endothelial cells, macrophages and pneumocytes in the lung, and
endothelial cells in the kidney. Minimal to mild levels of immunopositivity were foun
d in endothelial cells and macrophages in the heart, and in endothelial cells, hepato
cytes and Kupffer cells in the liver. Surprisingly, given the lack of viral genome an
d infectious virus recovered, mild to moderate amounts of immunopositive macrophages,
lymphocytes and endothelial cells were found in the spleen ( Table 2,

No published studies detail if ferrets are susceptible to hantavirus infection. To ex
amine this, four ferrets were exposed to either 2,000 PFU HTNV or PUUV K27 i.n. No se
roconversion occurred within 35 days post infection. The same animals were re-exposed
to either 200,000 PFU HTNV, 94,000 PFU PUUV Beaumont (a human PUUV isolate) or 164,00
0 PFU PUUV Seloignes (a vole PUUV isolate) by i.m. Prior to the re-exposure one of th
e seronegative ferrets in the HTNV group was removed for health concerns (rapid weigh
t loss) unrelated to the study, and it's cage mate was subsequently removed for behav
ioral health reasons before the completion of the study. Data from those two ferrets
are not shown. As soon as three days post infection ferrets began to lose weight with
HTNV infected ferrets losing between 5-12% of peak body weight as did PUUV infected f
errets (Fig 5A-5C ). By Day 35 post infection all animals had developed antibodies ag
```

```
ainst all strains of the virus as measured by N-ELISA assay (Fig 5D-5F ). Neutralizin
g antibody development began as early as Day 14 post infection and all ferrets develo
ped neutralizing antibodies by Day 28 post infection (Fig 5G-5L) . EDTA The organs of
infected ferrets were analyzed for viral load by RT-PCR and for the presence of infec
tious virus by plaque assay. The lung, liver, spleen, intestine and urine of HTNV inf
ected ferrets were negative for viral genome, with small amounts (<2 log 10 ) detecte
d in the heart and spleen of a single animal (S7B Fig). The heart, lung, kidney, inte
stine and urine of PUUV infected ferrets were negative for viral genome, though small
amounts were detected in the liver (2/4) and spleen (1/4) of PUUV infected ferrets (S
7C and S7D Fig) . Viral genome was spiked into the assay to confirm that the lack of
signal was not due to the presence of inhibitors. All organs except for the intestine
(4/6) had no inhibition of RT-PCR product. Similarly, no infectious virus was found i
n the organ of any ferret by plaque assay, despite spiked-in virus exhibiting no sign
ificant inhibition (S9 Fig). To confirm that immunosuppression of uninfected ferrets
did not result in rapid weight loss, four healthy ferrets were immunosuppressed with
a loading dose of 30 mg/kg Cyp followed by 10 mg/kg Cyp maintenance doses every other
day (S10A Fig). Five days post immunosuppression ferrets exhibited rapid weight loss,
fever, and lethargy due to secondary infection (S10B & S10C Fig) . Immunosupression w
as discontinued and 5 mg/kg enrofloxicin (a broad spectrum antibiotic) was given twic
e daily for a week. During this time, ferrets began to gain weight and their fever di
minished. On Day 11 immunosupression resumed for two doses with prophylactic enroflox
icin given once daily. Even with prophylactic antibiotics two ferrets spiked fevers w
ithin a few days post the second round of immunosuppressive treatment, though they di
d not lose weight (S10B & S10C Fig) . Based on these results the rapid weight loss an
d clinical signs observed upon immunosuppression of HTNV and PUUV infected ferrets wa
s most likely due to secondary infection. Due to the inability to completely manage s
econdary infection with prophylactic antibiotic treatment, no further immunosuppressi
on studies were carried out in ferrets.

To refine the HFRS-causing hantavirus ferret infection model four ferrets were challe
nged with 94,000 PFU PUUV Beaumont i.m. on Day 0. Weight and temperature were recorde
d daily, while twice weekly blood draws and urine collection was used to monitor kidn
ey function. As with the pilot experiment, ferrets slowly lost between 7-11% of peak
body weight, recapitulating our previous findings (Fig 6A) . No elevated temperatures
were observed (Fig 6B) . Ferrets developed a robust antibody response beginning on Da
y 10 post infection (Fig 6C and 6D) . Neutralizing antibodies developed early as Day
14 post infection, and were present in all ferrets by Day 28 post infection, as measu
red by PsVNA and PRNT ( Fig  6E and 6F) .

Despite the slow weight loss no signs of renal impairment were observed. Proteinuria
and hematuria are hallmarks of PUUV infection occurring in between 94-100% (proteinur
ia) and 58-85% (hematuria) of human clinical cases [46] . No prolonged proteinuria or
hematuria was observed (Fig 7A and 7B ) in infected ferrets. Similarly, blood urea ni
trogen and creatinine levels in the sera, both of which are elevated due to kidney fa
ilure in PUUV patients [47] remained unchanged in PUUV infected ferrets (Fig 7C and 7
D) [48, 49] . No infectious virus or viral genome was detected in the brain, heart, l
ung, liver, spleen, kidney, intestine, or eye (S11 and S12 Figs). No changes in other
serologic or urologic parameters were noted (S13 and S14 Figs).

No gross pathological changes or significant lesions associated with PUUV infection w
ere noted in the ferrets (S15 Fig). In the lungs, one ferret had mild neutrophilic an
d histiocytic inflammation centered on the bronchioles and expanded alveolar septa. G
iven that the inflammation was centered around bronchioles and not the vasculature, i
t is unlikely to be in response to PUUV infection. A number of other common or age-as
sociated lesions were observed in the ferrets. Two of four ferrets had proliferative
cortical cells in either the adrenal capsule or adrenal cortex that were likely clini
cally silent as they lack clinical signs consistent with adrenal-associated endocrino
pathy. Additionally, alveolar mineralization was noted in all four ferrets as was eos
inophilic and lymphoplasmacytic enteritis, and hepatitis. Two ferrets had fibromyxoma
tous degeneration of the atrioventricular valve. The spleen, brain, kidney, and pitui
```

```
tary gland were normal in all ferrets examined.

As with ferrets, there is no literature on the susceptibility of marmosets to hantavi
rus infection. To test this, three male marmosets were exposed to 1,000 PFU HTNV i.m.
Blood was collected weekly to measure seroconversion, serum viremia, as well as serum
parameters relating to renal function. All three animals seroconverted by Day 21 post
infection (Fig 8A and 8B ) despite displaying no clinical signs of illness. Neutraliz
ing antibody production began around the same time, and was robust by Day 30 post inf
ection (14, 557 by PsVNA 50 and 10,240-20,480 by PRNT 50 ) (Fig 8C and 8D ). Due to l
ow volumes of blood drawn at each time point serum viremia could not be examined for 
each individual animal, however a pool of sera from all three animals was evaluated a
t each time point post infection. Low levels of serum viremia were detected between d
ays 14 and 28 post infection by RT-PCR ( Fig 8G) . As with the ferret infection model
of HFRS-causing hantavirus, no renal injury as measured by changes noted in blood ure
a nitrogen or serum creatinine were observed over the course of infection, nor were c
hanges in any of the other serum parameters monitored (S16 Fig). Animals were euthani
zed on Day 30 post infection to examine organs for viral dissemination. No infectious
virus or viral genome was detected in the heart, lung, liver, spleen, kidney, intesti
ne or brain (S17 and S18 Figs). Lymphoid hyperplasia was noted in all three animals b
y histology, though the location and intensity varied between the spleen, various lym
ph nodes and gut associated lymphoid tissue. Such hyperplasia is indicative of a resp
onse to antigenic stimulation and was most likely caused by the viral challenge. Mild
to moderate congestion Three HFRS-associated hantavirus infection models was also see
n in the lungs of each marmoset; however, this was an acute change and was most likel
y associated with terminal anesthesia and euthanasia. No other significant histologic
al lesions were noted in any of the three animals (S19 Fig). Due to the negative viro
logy results and the lack of significant histologic changes, IHC to detect the presen
ce of HTNV antigen in tissues was not performed.

To date, the use of adult animal models to evaluate anti-HTNV and anti-PUUV medical c
ounter measures has been limited. Recombinant protein, vaccinia virus-vectored, and D
NA vaccines have been tested in the high dose HTNV hamster model [9, 37, 50] . Additi
onally, the ability of passively administered neutralizing antibodies has been evalua
ted in the high-dose hamster model of HTNV and PUUV and in PUUV challenged cynomolgus
macaques PUUV [50, 51] . Both of these models have limitations; the size of the macaq
ue requires large quantities of passive transfer material, and the high dose of the h
amster model, with challenge doses of~650 ID 99 , could require prohibitively large v
olumes of test article to neutralize the high initial dose. Suckling mice, which pres
ent with a disseminated disease not reminiscent of HFRS, have been used to evaluate H
TNV therapeutics as well [52] [53] [54] [55] . In this paper we present three adult a
nimal models of HFRS-causing hantavirus infection than can be used for future evaluat
ion of therapeutics, biologics and vaccines.

The ID 50 's for HTNV and PUUV determined in this report are similar to the lethal do
se 50% (LD 50 ) calculated for SNV and ANDV, <3 PFU via the i.m. route [6, 7, 41] . A
lso similar is that the challenge dose of ANDV required to infect/kill 50% of hamster
s by the i.n. or intragastric route is~10-30 fold higher [56] . Previous reports also
demonstrate that PUUV is capable of infecting hamsters by the intragastric route, wit
h an ID 50 of >10,000, making it a much less effective route of infection [57] . Desp
ite similarities in their ID 50 's, the ID 99 's of HTNV and PUUV greatly diverge, wi
th~200 times less HTNV required to infect hamsters via the i.m. route, and~2 fold les
s required to infect via the i.n. route ( Table 1 ). The lower ID 99 doses, coupled w
ith viral persistence in HTNV infected animals as opposed to viral clearance (Fig 3)
, suggest that HTNV is more infectious than PUUV in the hamster. The mechanism for th
is difference between these closely related hantaviruses remains unknown.

Lethal infection of hamsters with ANDV leads to extensive organ dissemination, with i
nfectious virus recovered in the lung, liver, kidney, spleen and heart [6] . Asymptom
atic infection of hamsters with SNV repeatedly passaged through hamsters has a simila
r organ distribution [58] . The organ distribution of both those viruses is similar t
```

```
o the low dose HTNV hamster model, with two notable exceptions. First, the HTNV model
has low, transient levels of virus in the liver and hardly any virus in the spleen (F
igs 3 and 4) . This dissimilarity between the models led us to confirm the lack of vi
rus in the spleen was not due to the presence of inhibitors by spiking either infecti
ous virus or viral RNA into samples prior to evaluation (S2 Fig). Second, in the ANDV
model, the presence of virus was determined by plaque assay, indicating the virus was
infectious and replication competent. In the HTNV low dose hamster model, while there
is detection of high levels of viral genome (and in the case of the HTNV i.m. model,
by pathology) by RT-PCR, recovery of infectious virus is sporadic, typically occurrin
g at low levels, in only a few hamsters per time point (Figs 3 and 4 , Table 1 ). The
discrepancy between RT-PCR/pathology and plaque assay is notable. The hamsters in thi
s study were not perfused, and given the appearance of neutralizing antibodies as ear
ly as day XX post infection, the presence of these antibodies could be impairing out
ability to recover live virus via the plaque assay. It is also possible that viral pa
ckaging is somehow impaired in the hamster, leading to a larger amounts of nucleocaps
id protein and viral genome than live virus. Further studies need to be undertaken to
elucidate.

The low dose HTNV hamster model also mimics the infection pattern of the virus in its
host species, the striped field mouse (Apodemus agrarius), with viral genome being de
tected in the lung, liver and kidney but not the spleen (the heart and brain were not
examined) [32, [59] [60] [61] .

In contrast to the HTNV low dose model, the organ distribution of virus in the PUUV l
ow dose model is transient. In PUUV i.m. infected hamsters virus is detectable by RT-
PCR around Day 11 post infection, with the virus being cleared from all organs except
the lung by Day 24. No infectious virus was recovered at any time point examined. Eve
n less virus was detected in the PUUV i.n. model; only a small amount in the brain of
one hamster on Day 28 post infection. In neither case was serum viremia observed (Fig
s 3 and 4) . This is most similar to the SNV models involving low passage virus in im
munocompetent hamsters: the virus is transiently detected in the lung by PCR, and the
n sporadically found in organs 12 to 14 days post infection using immunohistochemistr
y [41, 58] . The distribution of PUUV in the hamster differs somewhat from its host s
pecies the bank vole, where it is found to persist in the lung, spleen and kidney, an
d was not detected in the heart or the brain [62] .

Seroconversion of hamsters post viral exposure remains the best way to measure infect
ion, and should be a considered the primary endpoint for efficacy studies. Though the
PRNT assay was slightly more sensitive than the N-ELISA assay, detecting neutralizing
antibodies in all animals with N-ELISA titers, and in two PUUV animals that did not h
ave N-ELISA titers, the increased time, sample, and biosafety requirements necessary
for a PRNT assay make the N-ELISA a better choice (Fig 2, S1 Fig) . For a 10 PFU HTNV
i.m. challenge, given lower titer and specific OD sum values as compared with higher
challenge doses, and the fact that one hamster with significant viral genome in its o
rgans at Day 28 post infection (7.1 log 10 in the brain, and 6.8 log 10 in the kidne
y) did not seroconvert, waiting until Day 35 post infection to monitor seroconversion
is advisable (Figs 1-3 ). Viral load in the brain, kidney, and lungs as measured by R
T-PCR need to be evaluated at Day 35 post low-dose challenge to determine their usefu
lness as secondary endpoints. Recovery of infectious virus in any organ, and viremia
are too sporadic to serve as proxy markers for infection.

The ferret has been used as an experimental model for numerous hemorrhagic fever viru
ses, and respiratory viruses [12] [13] [14] [15] [16] [17] , though no published repo
rts exist examining its susceptibility to hantavirus infection. In comparison to the
hamster, the ferret is far more resistant to HTNV and PUUV infection. Exposure of fer
rets to 2,000 PFU i.n. (greater than the ID 99 for both viruses in hamsters), failed
to result in a productive infection and seroconversion. Instead i.m. challenge doses
of~100,000-200,000 PFU were needed (Fig 5) . Initially ferrets were exposed to PUUV K
27, a commonly used laboratory isolate that has been in cell culture for over 15 year
s. Repeated passaging of hantavirus is known to cause mutations [63] [64] [65] . In c
```

```
ontrast PUUV Beaumont and Seloignes are relatively recent isolates, with no more than
3 passages in cell culture post isolation. These strains were used for all subsequent
experiments to maximize the likelihood of PUUV to cause disease by eliminating possib
le attenuation of the laboratory strain of the virus. Despite the high challenge dose
and use of recent isolates, no elevation in white blood cells was observed over the c
ourse of the experiment, no pathology or organ burden was detected at the conclusion
of the experiment, and the N-ELISA specific OD sum, PRNT 50 , and PsVNA 50 titers rem
ained low (Figs 6 and 8 and S9 Fig) . This outcome is almost identical to that of Mar
burg and Ravn virus infection in ferrets, where the development of neutralizing antib
ody titers was the only sign that productive infection occurred [66] . Previous exper
iments in hamsters, have demonstrated that 2x10 5 PFU of gamma-irradiated ANDV and SN
V are not sufficient to cause seroconversion, and neither is 1x10 4 gamma-irradiated
PFU PUUV [6, 57] . Thus the seroconversion observed in ferrets, though low, is not li
kely to be due to a reaction to the large quantity of antigen administered, but to a
productive infection.

PUUV antigen, viral genome, or infectious virus has been found in the brain, pituitar
y gland, lung, heart, liver, kidney, spleen, cerebrospinal fluid, and gastrointestina
l tract of human patients with clinical symptoms of NE, though the pattern of viral d
issemination varies between individuals [67] [68] [69] [70] . Acute kidney injury and
vision disturbances including blurred vision, myopic shift, and lens thickening, whil
e pulmonary involvement including pleural effusion and vascular congestion, and renal
failure occurred less frequently [47, [71] [72] [73] [74] . Given the lack of high ne
utralizing antibody titers, which could have aided in viral clearance, the lack of vi
ral genome in any of the ferret organs examined is rather surprising (S6, S7, S11 and
S12 Figs). Furthermore, the lack of viral antigen and pathology in the organs tested
suggest either a transient infection cleared prior to Day 35 post infection, levels o
f virus so low as to be undetectable by the tests used, or a viral reservoir outside
of the organs tested.

The lack of detectable virus is most surprising given the gradual weight loss infecte
d animals exhibit (Figs 5 and 6 ). The animals gained weight until~3 days post challe
nge, at which point a gradual weight loss occurs, regardless of if HTNV or PUUV was t
he challenge virus. While weight loss is a feature of other ferret models of infectio
us diseases, the pattern we observed was unique: ferrets infected with morbillivirus,
avian influenza and filoviruses rapidly lose weight during the first week to two week
s post infection, while infection with severe acute respiratory syndrome virus result
s in no significant weight loss [16, 17, [75] [76] [77] .

In the hamster, infection with SNV is asymptomatic unless the hamster is immunosuppre
ssed. When ferrets were immunosuppressed on Day 42 post infection, rapid weight loss
and lethargy ensured (Fig 5) . Given that these clinical signs were also observed in
unchallenged control animals (S10 Fig) , this could likely be the result of a seconda
ry infection or drug toxicity. The use of Cyp is well documented in ferrets, primaril
y given at a high dose as an emetic, and no dosage for long term immunosuppression wa
s found [78] . The dosages used in this study (between 10-30 mg/kg) successfully redu
ced white blood cell levels, in ferrets and demonstrate that Cyp can be used to induc
e long-term immunosuppression, if antibiotics are given to control for secondary infe
ction (S8 Fig) .

Despite not being able to detect infectious virus or viral genome in the kidney, we h
ypothesized that the weight loss we observed could be due to kidney failure. Individu
als with HFRS exhibit proteinuria and hematuria, both of which can indicate kidney da
mage [46, 47] . Additionally, serum blood urea nitrogen and creatinine levels are bot
h elevated in HFRS patients and provide a second way to measure kidney function [46,
47, 73] . Decreased platelet levels also characterize clinical HFRS in humans, impair
ing coagulation [46, 47, 73] . In a second experiment designed to monitor kidney para
meters, PUUV infected ferrets exhibited the same gradual weight loss that characteriz
ed the first experiment. However, no prolonged signs of clinical kidney failure were
observed: blood urea nitrogen and creatinine levels did not dramatically increase ove
```

```
r the five week study period. Only one animal exhibited proteinuria (day 35), and two
exhibited hematuria (one on Day 4, and one on Day 21) (Fig 7) . Similarly to paramete
rs monitoring renal failure, no thrombocytopenia occurred in PUUV infected ferrets (S
9 and S13 Figs). The cause of the weight loss remains undetermined.

Though susceptible to both HTNV and PUUV, the ferret has limited usefulness for studi
es involving medical countermeasure efficacy testing. The ferret's large mass, even a
s an adolescent, makes the amount of test article needed also prohibitively large. Th
e large challenge dose required for infection could potentially obscure the protectiv
e effect of drugs of vaccines, due to the overwhelming amount of virus administered.
Moreover, although the animals are infected the resultant neutralizing antibody titer
s are small, resulting in potential sensitivity issues with the model. Husbandry and
handling of the animals under ABSL-3 procedures is also substantially more difficult
than hamsters, and they lack the genetic similarity to humans that marmosets possess.

In this study we have demonstrated the susceptibility of marmosets to HTNV infection.
Marmosets represent an attractive model for testing vaccines and therapeutics against
HFRS-causing hantaviruses due to genetic similarities to humans and small size. Also,
the model has a simple read-out of infection, i.e. robust antibody production as meas
ured by N-ELISA and PsVNA, making the determination of protection by vaccine or passi
ve transfer material, straightforward. Further optimization of the model, namely to d
etermine the ID 99 , could prove important as a 1,000 PFU challenge dose could be exc
essively high and prohibit therapeutic effects of candidate medical countermeasures.

Overall the marmoset model is more similar to the ferret HFRS-causing hantavirus infe
ction model than the hamster, though there are key differences. Like the ferret, no s
ignificant pathological abnormalities were noted, and no signs of renal failure were
observed (S16 and S19 Figs). Serum chemistry values do not differ from the normal ran
ge with the exception of albumin, total bilirubin, and amylase. While these values fe
ll outside the normal range, they did not change over the course of infection indicat
ing the problem may lie in the reference values used. The Piccolo general chemistry p
anel used to evaluate the parameters is optimized for human testing, and therefore ma
y be less than optimal for evaluating the marmoset, especially those parameters. Addi
tionally, no infectious virus or viral genome was recovered from any organ at Day 30
post infection [79, 80] . This is not surprising, given the high levels of neutralizi
ng antibodies present as early as 21 days post infection (Fig 8) . Unlike the ferret,
however, marmosets develop exceptionally high neutralizing antibody titers (10,240-2
0,480 by PRNT 50 and 14,866-221,557 by PsVNA 50 ), and display low-level serum viremi
a between two and four weeks post infection (Fig 8) . The serum viremia is significan
tly lower than in hamsters infected with HTNV, where some animals displayed RT-PCR ti
ters of >7log 10 , and in hamsters infected with ANDV, where infectious virus titers
prior to death are > 6log 10 [41].

Despite not exhibiting clinical signs of disease, the model's robust antibody respons
e (as measured by PRNT, PsVNA and N-ELISA) make it a useful tool for evaluating vacci
nes and pre-or post-exposure therapeutics.

This paper has explored the use of three laboratory animal species as possible infect
ion and disease models for HFRS-causing hantaviruses: the hamster, the ferret, and th
e marmoset. These models, especially the hamster model and marmoset model, will be us
eful for evaluating medical countermeasures with the potential to induce sterile immu
nity. The marmosets should be particularly useful for the evaluation of passively tra
nsferred protective human antibodies because of the relative genetic similarities bet
ween species in the Order Primates, and the small size of marmosets, allowing testing
with smaller volumes of material than would be required for larger species such as ma
caques. . Heart, lung, liver, spleen, kidney, and intestine were collected and assaye
d by plaque assay for the presence of infectious virus (A). To confirm lack of virus
recovered was not due to inhibitors, virus was spiked into serial dilutions of organ
homogenate to confirm no inhibitor was present (B). For a standard plaque assay the l
imit of detection, 1.7 log 10 , is depicted as a dashed line in (A). In (B) the dashe
```

```
d line is amount of HTNV plaques obtained when spiked into media rather than organ ho
mogenate. (TIF) S10 Fig. Immunosuppression of uninfected ferrets leads to rapid weigh
t loss and secondary bacterial infection. Uninfected ferrets were administered 10mg/k
g Cyp, and the antibiotic enrofloxicin, according to the schedule in (A). Weight (B)
and temperature (C) are shown. b.i.d indicates antibody was administered twice daily,
and q.d indicates antibiotic was administered daily. (TIF) S11 Fig. PUUV infected fer
rets had no infectious virus in the organs. Ferrets were infected with 94,000 PFU PUU
V Beaumont i.m. Heart, lung, liver, spleen, kidney, intestine, brain, eye, and adrena
l gland were collected on day 35 post infection and assayed for infectious virus by p
laque assay (A). Virus was spiked in to confirm no inhi bitor was present (B). For a
standard plaque assay the limit of detection, 1.7 log 10 , is depicted as a dashed li
ne in (A). In (B) the mean ± SEM is depicted in all spiked groups and the dashed line
is amount of HTNV plaques obtained when spiked into media rather than organ homogenat
e. To confirm no inhibitors were present, virus was spiked into samples (B). For a st
andard plaque assay the limit of detection, 1.7 log 10 , is depicted as a dashed line
in (A). In (B) the mean ± SEM is displayed for all spiked groups, and the dashed line
is amount of HTNV plaques obtained when spiked into media rather than organ homogenat
e. (TIF)

Three marmosets were infected with 1,000 PFU HTNV i.m. On Day 30 organs were harveste
d, and the presence of viral genome was determined by RT-PCR. To confirm no inhibitor
s were present, viral genome was spiked into samples. Heart (A), lung (B), liver (C),
spleen (D), kidney (E), intestine (F), and brain (G) were collected. The mean ± SEM i
s shown for the not spiked (NS) and spiked (S) groups and the limit of detection, 1 l
og 10 , is depicted as a dashed line.
```

Stop words

Here's a quote from Stanford's NLP team that will provide some context on stop words and their intended usage:

> *"Sometimes, some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called stop words . The general strategy for determining a stop list is to sort the terms by collection frequency (the total number of times each term appears in the document collection), and then to take the most frequent terms, often hand-filtered for their semantic content relative to the domain of the documents being indexed, as a stop list , the members of which are then discarded during indexing."*

Let's investigate the stop words list that we will use to clean our data. Note that apart from the standard stopwords, we will also remove any punctuation and also any occurrences of *et al.*, as they are often found in academic articles.

In [12]:
```
stop_tokens = nltk.corpus.stopwords.words('english') + list(punctuation) + ['et', 'al.
print(stop_tokens)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you'v
e", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his',
'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself',
'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'thi
s', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'bee
n', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an',
'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by',
'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before',
'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'ove
r', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'w
hy', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'su
ch', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's',
't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll',
'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "di
dn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'i
sn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'sh
an', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won',
"won't", 'wouldn', "wouldn't", '!', '"', '#', '$', '%', '&', "'", '(', ')', '*', '+',
',', '-', '.', '/', ':', ';', '<', '=', '>', '?', '@', '[', '\\', ']', '^', '_', '`',
'{', '|', '}', '~', 'et', 'al.']
```

The code below will be used to read the text associated with a series of articles, remove stop words from their text, and reduce) inflected words to their base form (stemming).

NOTE:

If you are not familiar with Python code, just execute the following cell and continue with the notebook. Understanding the code below is not require for understanding and following the overall flow of the notebook.

In [13]:
```python
class Reader:
    """Class used to read the files associated with an article"""

    def __init__(self):
        self.stemmer = SnowballStemmer('english')

    def read_file_to_json(self, filepath):
        try:
            with open(filepath, 'r') as f:
                data = json.load(f)
        except FileNotFoundError as e:
            mount.start()
            with open(filepath, 'r') as f:
                data = json.load(f)

        return data

    def parse_document(self, document_index):
        document = metadata_with_docs.iloc[document_index]

        # One article can have multiple associated documents
        words = []
        for filename in document['pdf_json_files'].split('; '):
            filepath = '{0}/{1}'.format(covid_dirpath, filename)
            pdf_json_files = os.listdir(covid_dirpath + '/comm_use_subset/pdf_json')
            filepath = covid_dirpath + '/comm_use_subset/pdf_json/' + pdf_json_files[

            if document_index % 50 == 0:
                print(filepath)

            data = self.read_file_to_json(filepath)



           # Split each paragraph into multiple sentences first, in order to improve t
            text = data['body_text']
            for paragraph in text:
                p_sentences = sent_tokenize(paragraph['text'])



               # Split each sentence into words, while making sure to remove the stopw
                for p_sentence in p_sentences:
                    sentence = [ self.stemmer.stem(word) for word in word_tokenize(p_s
                    words.extend(sentence)

        return (words, document['cord_uid'])




class Corpus:
    """An iterator that reads all sentences from the first N documents"""

    def __init__(self, n_documents):
        self.n_documents = n_documents
        self.stemmer = SnowballStemmer('english')
        self.reader = Reader()

    def __iter__(self):
         for document_index in range(0, self.n_documents):
            words, document_id = self.reader.parse_document(document_index)
            yield TaggedDocument(words, document_id)

    def plain_iter(self):
        for document_index in range(0, self.n_documents):
            words, document_id = self.reader.parse_document(document_index)
            yield (words, document_id)
```

Encoding documents as vectors

In this lab, we're using a subset of 1000 articles to train a Machine Learning model that encodes text documents into numerical vectors (a document embedding model).

Training a document embedding model takes a significant amount of time, and for this reason we already provide a trained model. We also provide the code below in case you want to get more details about the process. Running the next two cells will result in loading the already existing model.

In [14]:

```
N_DOCUMENTS = 500
```

In [15]:

```
%%time
```

```
model_filename = f'covid_embeddings_model_{N_DOCUMENTS}_docs.w2v'

if (os.path.exists(model_filename)):
    model = Doc2Vec.load(model_filename)
    print(f'Done, loaded word2vec model with { len(model.wv.vocab) } words.')
else:
    model = Doc2Vec(Corpus(N_DOCUMENTS), vector_size=128, batch_words=10)
    model.save(model_filename)
    print(f'Done, trained word2vec model with { len(model.wv.vocab) } words.')
```

```
Done, loaded word2vec model with 10751 words.
CPU times: user 123 ms, sys: 11.1 ms, total: 134 ms
Wall time: 459 ms
```

Word frequencies

Let's analyze the relative frequencies of words in the corpus of articles. We will display a word cloud to provide a visual representation of these relative frequencies.

In [16]:

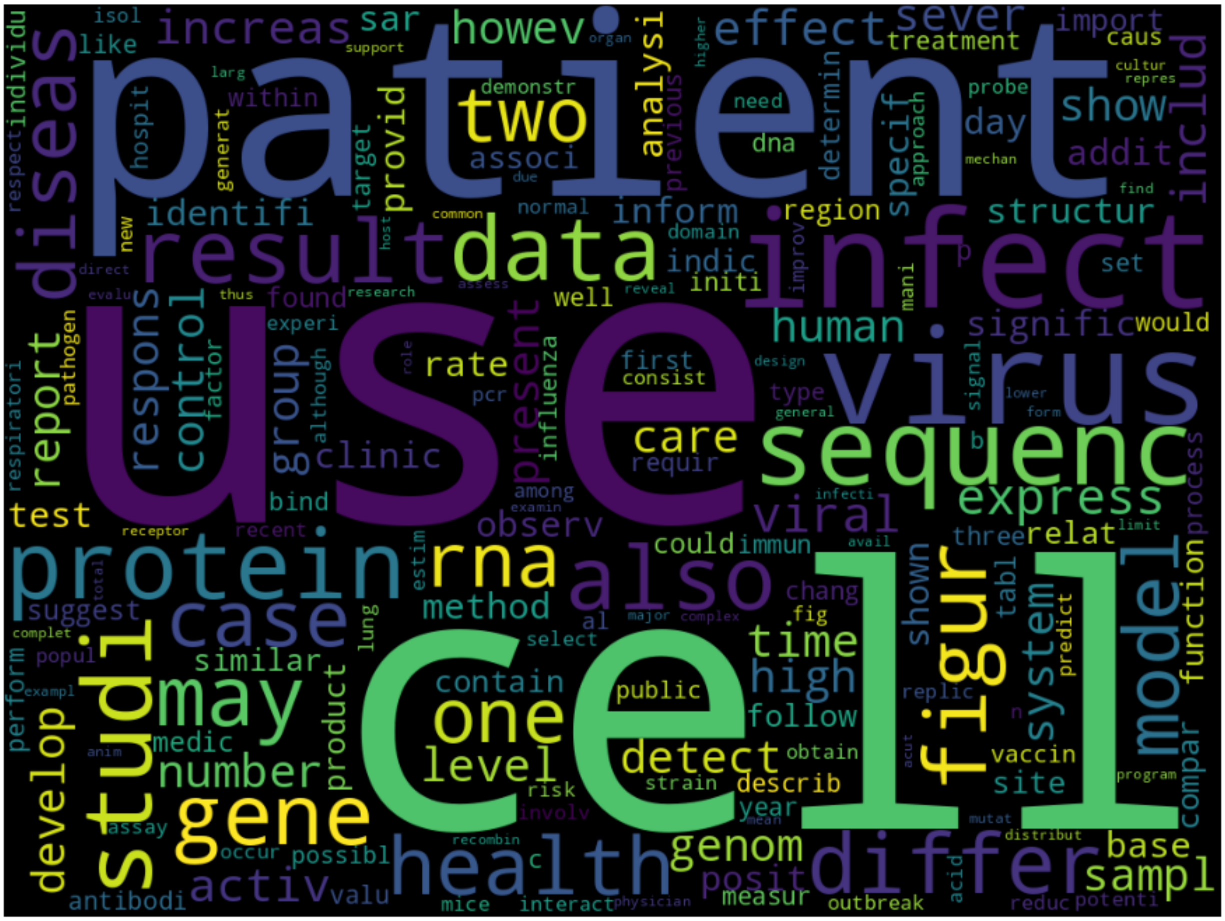
```
word_frequencies = { key: model.wv.vocab[key].count for key in model.wv.vocab }
```

In [17]:

```
cloud = WordCloud(width=1024, height=768).generate_from_frequencies(word_frequencies)
plt.figure(figsize=(16,12))
plt.imshow(cloud, interpolation='antialiased')
plt.axis("off")
```

Out[17]:

```
(-0.5, 1023.5, 767.5, -0.5)
```



Embedding documents

Below is an example on how we embed (transform from text to numerical vector) one of the documents.

In [18]:

```
words, doc_id = Reader().parse_document(DOCUMENT_INDEX)
model.infer_vector(words)
```

```
/tmp/tmpemaxbzly/covid19temp/comm_use_subset/pdf_json/02a009e42054081b441d0f4b203679c
4b0cae38d.json
```

Out[18]:

```
array([-0.36768144, -1.1106528 ,  2.4751897 , -0.8354529 , -1.1695763 ,
        0.80843717, -0.39670926,  0.48410633, -0.7084072 , -1.5064273 ,
        0.88584656, -2.4492157 ,  3.2096512 ,  0.11121531,  2.2204032 ,
       -1.1593775 ,  0.21179079,  5.6468253 , -0.7457158 , -1.2919372 ,
       -0.50090504, -1.0035561 ,  1.0373933 ,  0.00688864,  1.0333565 ,
        0.8773138 , -0.56150395,  0.18981826, -0.81829715, -1.4064279 ,
       -0.86820185, -1.885929  , -0.07239459,  1.0694723 ,  0.6629561 ,
       -2.0599358 ,  1.4812938 , -0.26281548,  0.20125028,  0.85074055,
       -2.4990096 ,  0.26456544, -0.6714662 , -0.4843073 ,  1.0774498 ,
        1.5166879 , -0.35177374,  0.8190227 , -1.7486702 , -2.1246812 ,
       -1.4046166 , -0.70962495, -1.4102095 , -0.2974848 , -1.55897   ,
        0.19682325, -0.15490614,  2.4801288 ,  1.0956861 ,  0.27181962,
        0.9518106 , -1.2184901 ,  0.8317703 ,  1.8769274 , -2.0623431 ,
       -1.1901475 , -2.263474  ,  0.3310059 ,  0.56674606, -0.11053646,
        2.2180562 , -0.28195322,  0.8319942 ,  1.0140984 ,  2.1507628 ,
        0.9072955 , -1.1474484 ,  2.2471466 ,  1.7143121 , -1.4232329 ,
       -0.41604537, -2.7342613 , -1.030752  ,  0.6367958 ,  0.53506535,
        1.7867123 ,  0.23827484, -0.82510537,  0.691904  , -0.8767283 ,
       -2.7645059 , -0.28387508,  1.8029312 ,  1.5281334 , -0.25860503,
       -0.56525475, -1.5408717 , -1.1052479 ,  0.76068234,  1.4443927 ,
        0.85816324,  0.23737751,  0.60747755, -0.88963866,  0.64770806,
        0.2448809 ,  1.3057928 ,  1.5467266 ,  1.1140411 , -2.052351  ,
       -0.33170623, -0.26855245,  0.45625284,  1.8268589 ,  0.6162205 ,
        1.8192254 , -0.9949126 ,  1.683619  , -2.219154  , -0.38159177,
       -0.4732724 , -1.9200023 , -0.04358702, -2.1741493 , -1.1506021 ,
        1.4797405 ,  0.08616818,  1.4481105 ], dtype=float32)
```

And this is an example of a trivial "document" (containing a single, trivial sentence) going through the same process. Notice how, regardless of the length of the sentence, the result vector is always the same size - the `vector_size` argument used while training the `Doc2Vec` model.

This is very important in the following stages of the process, when we are working with multiple documents.

In [19]:

```
model.infer_vector(['human', 'love', 'cat', 'dog'])
```

Out[19]:

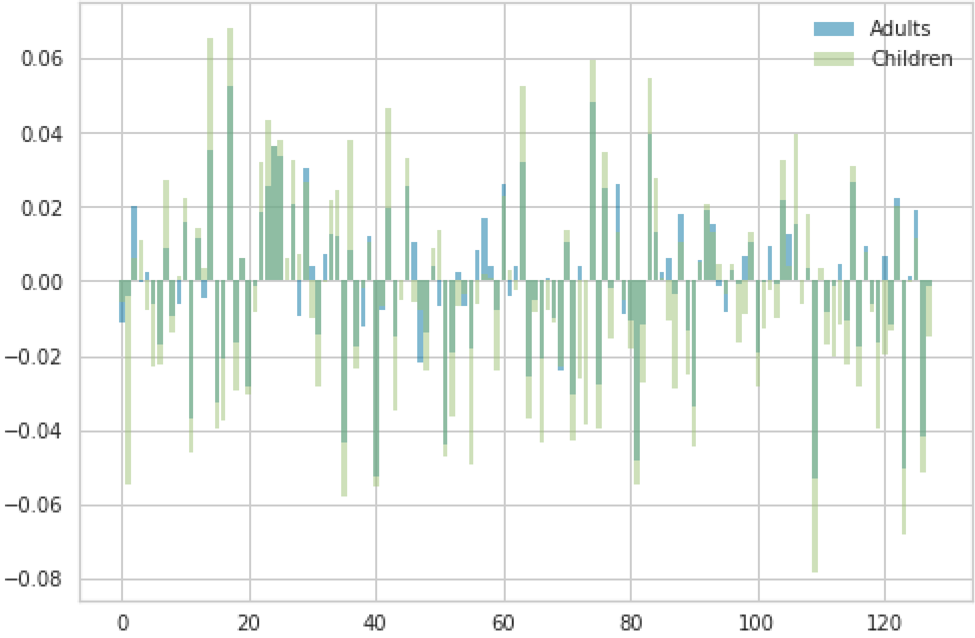
```
array([ 0.01187267, -0.07222479,  0.0009857 ,  0.02489351, -0.00371165,
       -0.05244752, -0.04042441,  0.00981789, -0.00546434, -0.0114641 ,
        0.00382181, -0.0549628 , -0.00109327,  0.02235473,  0.06640013,
       -0.02146467, -0.00610835,  0.07812942, -0.01487299, -0.00225763,
       -0.02316215, -0.01193333,  0.01044642,  0.03996939,  0.02644479,
        0.0227175 ,  0.04123925,  0.01465284,  0.00910605,  0.02305634,
       -0.00566173, -0.03546144, -0.02495514,  0.04078628,  0.01669642,
       -0.03630543,  0.03745678, -0.01681023,  0.01524001,  0.02174   ,
       -0.04845124, -0.02859627,  0.02559633, -0.01294257,  0.01237459,
        0.05609317, -0.01783988,  0.00306992, -0.02992033, -0.01554093,
        0.01374504, -0.03704116, -0.01849834,  0.00342414, -0.00394119,
       -0.01736315,  0.00061388,  0.00300337,  0.02583033, -0.02169578,
        0.03127309, -0.00914276, -0.00147238,  0.06283462, -0.02120893,
        0.00339467, -0.03140209, -0.0407921 , -0.01630779, -0.01234598,
        0.03513828, -0.05091879, -0.03363749, -0.04252905,  0.03870467,
       -0.02403057,  0.03826586, -0.00683732, -0.005172  , -0.03065385,
       -0.03811062, -0.02149642, -0.02606765,  0.02589219,  0.01650087,
       -0.00108351,  0.00658174, -0.03633691,  0.01417964, -0.01604395,
       -0.03558038,  0.02062096,  0.00863639, -0.00796049, -0.00531631,
        0.00715494,  0.03192762, -0.00160919, -0.00166048,  0.00513896,
       -0.02604181, -0.01912436,  0.02104215, -0.01204367,  0.01967912,
        0.00946659, -0.00244938, -0.01942666,  0.00923178, -0.07924119,
       -0.00391558, -0.00874832, -0.02192896,  0.01491943,  0.00303279,
        0.02053712, -0.04007722,  0.01599913,  0.00882744, -0.04477157,
       -0.02722152,  0.00240392,  0.02242075, -0.06093423, -0.03261027,
        0.02414383, -0.03612995, -0.00615444], dtype=float32)
```

The resulting vectors will look more or less similar, depending on how different the contents of the articles are themselves. See below the differences resulting from a single word change - some of the values significantly overlap, while others are quite different if not opposite.

In [20]:

```
adult_vector = model.infer_vector(['adult', 'love', 'cat', 'dog'])
child_vector = model.infer_vector(['child', 'love', 'cat', 'dog'])
labels = range(0, 128)

plt.bar(labels, adult_vector, align='center', alpha=0.5)
plt.bar(labels, child_vector, align='center', alpha=0.5)
plt.legend(['Adults', 'Children'])
plt.show()
```



Let's now do the same for the same for all the documents we're focusing on.

In [21]:
```
%%time

word_vectors = []
ids = []

for (words, doc_id) in Corpus(N_DOCUMENTS).plain_iter():
    ids.append(doc_id)
    word_vector = model.infer_vector(words)
    word_vectors.append(word_vector)
    if len(word_vectors) % 100 == 0:
        print(f'Processed {len(word_vectors)} documents.')
```

```
/tmp/tmpemaxbzly/covid19temp/comm_use_subset/pdf_json/02a009e42054081b441d0f4b203679c
4b0cae38d.json
/tmp/tmpemaxbzly/covid19temp/comm_use_subset/pdf_json/0b180c5c5edf329811114548e19a708
303e7c1c2.json
Processed 100 documents.
/tmp/tmpemaxbzly/covid19temp/comm_use_subset/pdf_json/0ca8648d40bee3056b3f9840de6d34b
57ed121d0.json
/tmp/tmpemaxbzly/covid19temp/comm_use_subset/pdf_json/0df2a9766ade17a0d9a625ef02722fc
167ee0526.json
Processed 200 documents.
/tmp/tmpemaxbzly/covid19temp/comm_use_subset/pdf_json/0f7bb2b30b0eba1a065a6dfc88dbbd9
9053ff1ba.json
/tmp/tmpemaxbzly/covid19temp/comm_use_subset/pdf_json/1b02698e082376846f59c99c449161a
7a7eb737f.json
Processed 300 documents.
/tmp/tmpemaxbzly/covid19temp/comm_use_subset/pdf_json/1c8afcb822dac8e8e5b33a85c2c2e9c
7cc24a25a.json
/tmp/tmpemaxbzly/covid19temp/comm_use_subset/pdf_json/1dd28ef7fb600a0ca94980bcf6e5cbc
cf52a77fe.json
Processed 400 documents.
/tmp/tmpemaxbzly/covid19temp/comm_use_subset/pdf_json/1f2bbc79b56c51c0ae26ced2d7d8ccf
10360fa72.json
/tmp/tmpemaxbzly/covid19temp/comm_use_subset/pdf_json/2a8a6ddd84f0c80ce33fed8960c3201
c08e56854.json
Processed 500 documents.
CPU times: user 47.1 s, sys: 279 ms, total: 47.4 s
Wall time: 1min 4s
```

Now that we've finished reading the articles, we can dismount the dataset in order to free up resources

In [22]:

```
mount.stop()
```

Covid-19 documents prepared for Machine Learning

We'll create a new DataFrame using the word vectors we've just calculated, this is the numerical form of the documents which is ready for Machine Learning workloads.

In [23]:

```
wv_df = pd.DataFrame(word_vectors, index=ids)
wv_df
```

Out[23]:

	0	1	2	3	4	5	6	7	
ug7v899j	-0.360649	-1.138998	2.476880	-0.788588	-1.130553	0.846280	-0.359760	0.609766	-0.69
02tnwd4m	0.054707	-2.481819	0.917430	1.064308	-0.671468	-1.027804	-1.930289	0.263879	-0.00
ejv2xln0	-0.072066	-0.641108	0.855729	-0.032554	-0.280952	-1.078268	-1.413609	0.141413	0.58
2b73a28n	-1.238587	-0.418109	0.669005	-0.239547	-1.073317	0.036139	-1.333731	0.813179	-0.35
9785vg6d	0.396538	-1.957209	0.616441	-0.054515	-0.672714	-0.731759	-0.851979	-0.648167	-1.44
...	...	...	...	...	...	...	...	...	
v6jyplcn	-0.991498	-1.217878	1.090420	0.815862	-0.347704	-0.737506	-1.035312	1.022310	-0.02
fae3sczm	-0.779463	0.317488	1.283767	0.217249	0.233826	-0.167178	-1.575740	-0.245862	-0.37
1c4m2fym	1.698243	-3.183822	0.957596	0.766598	0.914241	-3.589298	-2.758309	-0.146896	-0.77
1ldebnq8	-0.771220	-1.271643	0.470235	0.511493	-0.262360	-0.795250	-1.670549	-0.151086	-0.05
x7wva1ax	-0.376713	-2.424713	1.729154	1.284823	-0.440978	-2.996548	-3.534384	0.270851	-0.44

500 rows × 128 columns

We'll join the DataFrame containing the numerical embeddings with the original dataset.

In [24]:

```
indexed_metadata = metadata_with_docs.set_index('cord_uid')
metadata_with_embeddings = pd.concat([indexed_metadata.iloc[:N_DOCUMENTS], wv_df], axi
metadata_with_embeddings
```

Out[24]:

	sha	source_x	title	
ug7v899j	d1aafb70c066a2068b02786f8929fd9c900897fb	PMC	Clinical features of culture-proven Mycoplasma...	10.1186/1471-2
02tnwd4m	6b0567729c2143a66d737eb0a2f63f2dce2e5a7d	PMC	Nitric oxide: a pro-inflammatory mediator in l...	10.11
ejv2xln0	06ced00a5fc04215949aa72528f2eeaae1d58927	PMC	Surfactant protein-D and pulmonary host defense	10.11
2b73a28n	348055649b6b8cf2b9a376498df9bf41f7123605	PMC	Role of endothelin-1 in lung disease	10.11
9785vg6d	5f48792a5fa08bed9f56016f4981ae2ca6031b32	PMC	Gene expression in epithelial cells in respons...	10.11
...	...	...	...	
v6jyplcn	ba1a74766e0d96756105e85ce014b4bbe86fe819	PMC	Immune responses against severe acute respirat...	10.1111/ 2567.2007.
fae3sczm	00acd3fd31ed0cde8df286697caefc5298e54df1	PMC	Distinguishing Molecular Features and Clinical...	10.1371/journal.pone.0
1c4m2fym	808dfc4c59f3e2e9150aa5542ea227718741388b	PMC	Towards a Coronavirus-Based HIV Multigene Vaccine	10.1080/17402520600
1ldebnq8	9cf43ad7346d9099015ee0545e0b3498c10213b4	PMC	Epithelial Cell Apoptosis and Neutrophil Recru...	10.2119/2008-00
x7wva1ax	3ccc18758132523e4b94a6cd392053a5b4b542cb	PMC	The cucumovirus 2b gene drives selection of in...	10.1093/nar/g

500 rows × 146 columns

Preparing for clustering documents

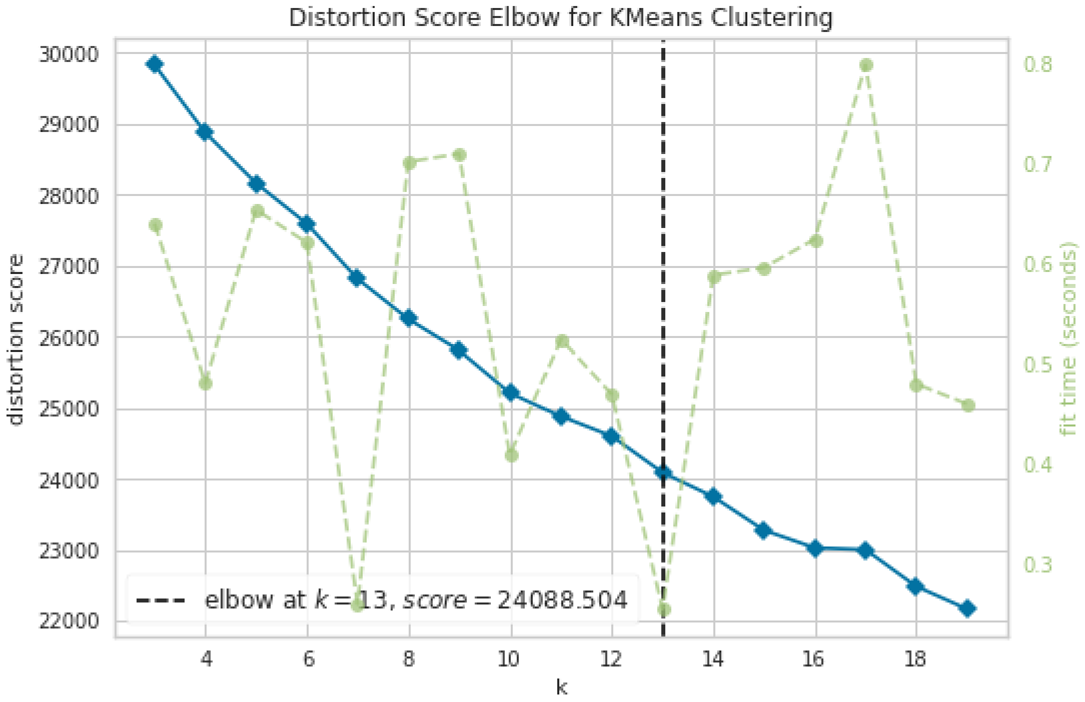
One of the challenges with clustering is to find the ideal number of clusters to look for. The elbow method is one of the most common approaches.

We're visualizing an elbow metric (the "distortion" score) and trying to find a point where it stops decreasing with the number of clusters.

In [25]:
```
visualizer = KElbowVisualizer(KMeans(), k=(3,20))
visualizer.fit(wv_df)

visualizer.show()
```

```
findfont: Font family ['sans-serif'] not found. Falling back to DejaVu Sans.
findfont: Generic family 'sans-serif' not found because none of the following familie
s were found: Arial, Liberation Sans, Bitstream Vera Sans, sans-serif
```



Out[25]:
```
<AxesSubplot:title={'center':'Distortion Score Elbow for KMeans Clustering'}, xlabel
='k', ylabel='distortion score'>
```

Clustering documents

We've determined the acceptable value for the clusters, so let's use Machine Learning to determine those clusters. We'll use the classic KMeans algorithm to do this.

In [26]:
```
clusterer = KMeans(12 if visualizer.elbow_value_ > 12 else visualizer.elbow_value_)
clusterer.fit(wv_df)
clusters = clusterer.labels_
```

We'll perform a quick visual check on the clusters. In order to be able to visualize 128 dimensions (which is the size of the word vectors) in a 2-D space, we'll use the PCA (Principal Component Analysis) dimensionality reduction technique. This will transform our 128-dimensional vectors into 2-dimensional ones that we can display.

In [27]:

```
pca = PCA(n_components=2)
pca.fit(wv_df)
result = pca.transform(wv_df)
```

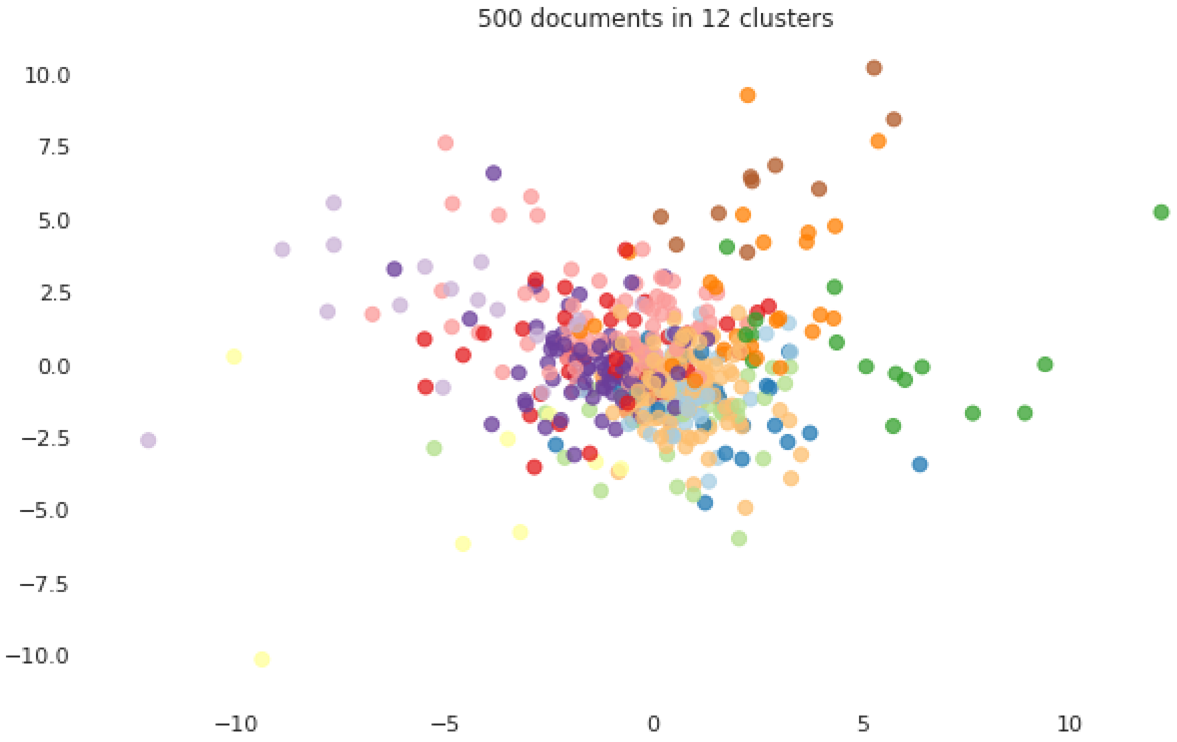
Afterwards, we can plot the documents in a simple 2-D chart, and color each one according to their cluster

In [28]:

```
sns.set(rc={'figure.figsize':(10, 6), 'figure.facecolor':'white', 'axes.facecolor':'wh

color_palette = sns.color_palette('Paired')
# Each cluster gets its own color from the palette
cluster_colors = [color_palette[x] if x >= 0 else (0.5, 0.5, 0.5) for x in clusterer.l
plt.scatter(result[:,0], result[:,1], s=50, c=cluster_colors, alpha=0.75)

plt.title(f'{N_DOCUMENTS} documents in {clusterer.n_clusters} clusters')
plt.show()
```



We'll add each article's cluster as new column to our combined dataset

In [29]:

```
metadata_with_clusters = metadata_with_embeddings
metadata_with_clusters['cluster'] = clusters
metadata_with_clusters
```

Out[29]:

	sha	source_x	title	
ug7v899j	d1aafb70c066a2068b02786f8929fd9c900897fb	PMC	Clinical features of culture-proven Mycoplasma...	10.1186/1471-2
02tnwd4m	6b0567729c2143a66d737eb0a2f63f2dce2e5a7d	PMC	Nitric oxide: a pro-inflammatory mediator in l...	10.11
ejv2xln0	06ced00a5fc04215949aa72528f2eeaae1d58927	PMC	Surfactant protein-D and pulmonary host defense	10.11
2b73a28n	348055649b6b8cf2b9a376498df9bf41f7123605	PMC	Role of endothelin-1 in lung disease	10.11
9785vg6d	5f48792a5fa08bed9f56016f4981ae2ca6031b32	PMC	Gene expression in epithelial cells in respons...	10.11
...	...	...	...	
v6jyplcn	ba1a74766e0d96756105e85ce014b4bbe86fe819	PMC	Immune responses against severe acute respirat...	10.1111/ 2567.2007.
fae3sczm	00acd3fd31ed0cde8df286697caefc5298e54df1	PMC	Distinguishing Molecular Features and Clinical...	10.1371/journal.pone.0
1c4m2fym	808dfc4c59f3e2e9150aa5542ea227718741388b	PMC	Towards a Coronavirus-Based HIV Multigene Vaccine	10.1080/17402520600
1ldebnq8	9cf43ad7346d9099015ee0545e0b3498c10213b4	PMC	Epithelial Cell Apoptosis and Neutrophil Recru...	10.2119/2008-00
x7wva1ax	3ccc18758132523e4b94a6cd392053a5b4b542cb	PMC	The cucumovirus 2b gene drives selection of in...	10.1093/nar/g

500 rows × 147 columns

We can now split our data into two datasets - a **training** one that will be used to train a Machine Learning model, able to determine the cluster that should be assigned to an article, and a **test** one that we'll use to test this classifier.

We will allocate 80% of the articles to training the Machine Learning model, and the remaining 20% to testing it.

In [30]:

```
train, test = train_test_split(metadata_with_clusters, train_size=0.8)
train
```

Out[30]:

	sha	source_x	title	
ujhgb3b0	85c04886b69b27eeddd4a72c74e2e1727a80598f	PMC	CoVDB: a comprehensive database for comparativ...	10.1093/na
apiz60hu	9aa51dd5b5755afa670398027272f2e6b4f3d83b	PMC	Adaptive evolution of the spike gene of SARS c...	10.1186/1471-
icj261ls	6a3c7595bc8fa30e3f74fd3d350d7cb7e3a1668b	PMC	The three transfer RNAs occupying the A, P and...	10.1093/na
1fu1blu0	4b43f61d164be997a34343c11c70c42edd91898b	PMC	Distributed data processing for public health ...	10.1186/1471-2
wrhr237o	275d7a25357ea46699072167d4ca109108132068	PMC	Surveillance recommendations based on an explo...	10.1186/1471-2
...	...	...	...	
6lvn10f4	14e0cac6e86d62859e6c9f1351ab67466d89e5b3	PMC	Heterogeneous nuclear ribonucleoprotein A1 reg...	10.1093/emboj/
zzkkm496	c4cbc868c6aa292decf1296da0f74eb496cca7ac	PMC	Outcome of paediatric intensive care survivors	10.1007/s00431-0
wutnzzhg	017ca5bdac589a37196df7b8e077c4c371ab32da	PMC	Pro/con clinical debate: Isolation precautions...	10.11
p34ezktf	d77233dd9085c9c76787c84ee9a1e249424b283a	PMC	Australian public health policy in 2003 – 2004	10.1186/1743
mqnqjn0c	3899c3b15d80afef1944752dafbebaa99ee43187	PMC	A Mouse-Adapted SARS-Coronavirus Causes Diseas...	10.1371/journal.ppa

400 rows × 147 columns

To speed up training, we'll ignore all columns except the word vectors calculated using Doc2Vec. For this reason, we will create a separate dataset just with the vectors.

In [31]:

```
columns_to_ignore = ['sha', 'source_x', 'title', 'doi', 'pmcid', 'pubmed_id', 'license
                     'who_covidence_id', 'arxiv_id', 'pdf_json_files', 'pmc_json_files
train_data_vectors = train.drop(columns_to_ignore, axis=1)
test_data_vectors = test.drop(columns_to_ignore, axis=1)
```

Register the training and testing datasets for AutoML availability

We're registering the training and testing datasets with the Azure Machine Learning datastore to make them available inside Azure Machine Learning Studio and Automated ML.

In [32]:

```
# Retrieve your ML workspace
ws = Workspace.from_config()
# Retrieve the workspace's default datastore
datastore = ws.get_default_datastore()


Dataset.Tabular.register_pandas_dataframe(train_data_vectors, datastore, 'COVID19Artic
Dataset.Tabular.register_pandas_dataframe(test_data_vectors, datastore, 'COVID19Articl
```

```
Validating arguments.
Arguments validated.
Successfully obtained datastore reference and path.
Uploading file to managed-dataset/4569b189-49fa-43e9-91d9-63be7d5b4c30/
Successfully uploaded file to datastore.
Creating and registering a new dataset.
Successfully created and registered a new dataset.
Validating arguments.
Arguments validated.
Successfully obtained datastore reference and path.
Uploading file to managed-dataset/51c93458-f386-42af-83f8-1da164289585/
Successfully uploaded file to datastore.
Creating and registering a new dataset.
Successfully created and registered a new dataset.
```

Out[32]:

```
{
  "source": [
    "('workspaceblobstore', 'managed-dataset/51c93458-f386-42af-83f8-1da164289585/')"
  ],
  "definition": [
    "GetDatastoreFiles",
    "ReadParquetFile",
    "DropColumns"
  ],
  "registration": {
    "id": "d164de0e-93b2-49c1-b860-9b1ff0e72546",
    "name": "COVID19Articles_Test_Vectors",
    "version": 1,
    "workspace": "Workspace.create(name='ai-in-a-day-800633', subscription_id='b8fce0
07-8472-482f-92d6-2278c8aab6a9', resource_group='ai-in-a-day')"
  }
}
```

Open Azure Machine Learning Studio

Return to the GitHub repo and follow the instructions from there. You will use Automated ML in Azure Machine Learning Studio to train a classification model that predicts the document cluster for new research articles.