Navigating ARPA-H

Matt Might | might@uab.edu | @mattmight

Hugh Kaul Precision Medicine Institute
ARPA-H?
Advanced Research Projects Agency for Health
Where did it come from?
What are “ARPA-class” projects?
Tips for ARPA-H funding
History
Soviets launch Sputnik, 1957
Eisenhower learns of Sputnik
ARPA is born, 1958
Typical ARPA project
The Internet (originally called ARPANet)
Stealth technology
Rebranded in 1972: DARPA
And also...

1. Advanced Robotics
2. Artificial Intelligence (AI)
3. Precision Medicine and Biotechnology
4. Autonomous Vehicles and Unmanned Systems
5. Cybersecurity
6. Materials Science and Manufacturing
7. Neuroscience and Brain-Computer Interfaces
8. Hypersonics
9. Quantum Computing and Communications
10. Health and Medical Technologies
11. Space Technologies
12. Energy and Power Systems
13. Human Performance and Bioenhancement
14. Advanced Sensors and Detection Systems
15. High-Energy Laser Systems
How?
High risk, high reward
Parallel approaches
Short timelines
Fail fast philosophy
Significant deliverables
High autonomy for PMs
Limited terms for PMs
Examples: My DARPA projects
CRASH: Create unhackable OS
APAC: Malware-proof phones
STAC: Leak-proof programs
Modeled on DARPA
(perhaps too closely)
Inspired by Operation Warp Speed
Two new ARPA-H Projects
MATRIX project w/ Dave Fajgenbaum: $48 million

AI search for treatments for all 10,000 rare disorders
AI-Enabled Biomedical Data Fabric: $9 million
Collect => Harmonize => Analyze
Ex: Importing patient data via phone
You should make sure that you trust this app before you grant access.

- The records you share may identify you and your care team.
- The app may be able to infer other aspects of your health history from the records you share.
- The app developer may retain and use your records after you stop sharing or delete the app.

View Your Health Records

Continue

Not Now

**Which categories can "FHIR-HOSE" access?**

**APP EXPLANATION**

FHIR-HOSE needs access to your clinical records in order to support predictive analytics and CHARMTReinsight repository synchronization.

View "FHIR-HOSE" Privacy Policy

Review the app’s Privacy Policy to understand how it may use and disclose your records.

ALLOW "FHIR-HOSE" TO READ

- Allergies
- Clinical Vitals
- Conditions
- Immunizations
- Lab Results
- Medications

**SHARING NEW RECORDS**

Ask Before Sharing □

Automatically Share □

"FHIR-HOSE" will have access to your current records as of November 16, 2023, at 6:25 AM and will ask each time before accessing new records.

Done
My medical record in an open standards format: FHIR
Automated analysis: CHARM-GPT
<table>
<thead>
<tr>
<th>Age</th>
<th>Tumor Size Reduction</th>
<th>Gene Expression Sco</th>
<th>Side Effect Sever</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36</td>
<td>95.98528437324805</td>
<td>90.49512863264476</td>
</tr>
<tr>
<td>2</td>
<td>49</td>
<td>88.03345035229626</td>
<td>88.63890037284231</td>
</tr>
<tr>
<td>3</td>
<td>44</td>
<td>91.24217733388136</td>
<td>85.624528039684</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>70.61753482887407</td>
<td>92.23705789444759</td>
</tr>
<tr>
<td>5</td>
<td>37</td>
<td>99.0972955685983</td>
<td>82.78987721304084</td>
</tr>
</tbody>
</table>

I need to identify clusters in this drug response profiling data for a new drug for breast cancer. Please conduct a clustering analysis.
CHARM-GPT writes code to look at the data and understand its format/structure.

```python
import pandas as pd

# Load the data
file_path = '/mnt/data/Drug_Response_Profiling_Sample_Data.csv'
data = pd.read_csv(file_path)

# Display the first few rows of the dataframe to understand its structure
data.head()
```

<table>
<thead>
<tr>
<th>Age</th>
<th>Tumor Size Reduction (%)</th>
<th>Gene Expression Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>95.985284</td>
<td>90.495129</td>
</tr>
<tr>
<td>1</td>
<td>88.033450</td>
<td>88.638900</td>
</tr>
<tr>
<td>2</td>
<td>91.242177</td>
<td>85.824583</td>
</tr>
<tr>
<td>3</td>
<td>70.617535</td>
<td>92.237058</td>
</tr>
<tr>
<td>4</td>
<td>99.097296</td>
<td>82.709877</td>
</tr>
</tbody>
</table>

Side Effect Severity (0-10)

<table>
<thead>
<tr>
<th>Side Effect Severity (0-10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>
The dataset contains the following columns:

- **Age**: Age of the patient.
- **Tumor Size Reduction (%)**: Percentage reduction in tumor size.
- **Gene Expression Score**: A score representing gene expression levels.
- **Side Effect Severity (0-10)**: Severity of side effects on a scale from 0 to 10.
Let's proceed with the clustering analysis. I'll perform k-means clustering to identify clusters in this data.

**Steps:**

1. Normalize the data.
2. Determine the optimal number of clusters using the elbow method.
3. Perform k-means clustering.
4. Visualize the clusters.

Let's start with data normalization and determining the optimal number of clusters using the elbow method. [1-]
Let's start with data normalization and determining the optimal number of clusters using the elbow method. 

```python
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# Select the features for clustering
features = data[['Age', 'Tumor Size Reduction (%)', 'Gene Expression Score', 'Side']

# Normalize the data
scaler = StandardScaler()
normalized_data = scaler.fit_transform(features)

# Determine the optimal number of clusters using the elbow method
sse = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=0).fit(normalized_data)
    sse.append(kmeans.inertia_)

# Plot the elbow method result
plt.figure(figsize=(10, 8))
plt.plot(range(1, 11), sse, marker='o')
plt.xlabel('Number of clusters')
plt.ylabel('Sum of Squared Distances (Inertia)')
plt.title('Elbow Method for Determining Optimal Number of Clusters')
plt.grid(True)
plt.show()
```

CHARM-GPT writes code to use the elbow method to estimate the number of clusters.
CHARM-GPT estimates 3 clusters from the elbow method.

The elbow method plot suggests a point where the curve starts to bend, indicating the optimal number of clusters. This point is typically around the "elbow" of the plot.

From the plot, it looks like the optimal number of clusters is around 3 or 4. For this analysis, I'll choose 3 clusters to proceed with the k-means clustering.
Next, I'll perform k-means clustering with 3 clusters and visualize the results.

```python
# Perform k-means clustering with the chosen number of clusters
optimal_clusters = 3
kmeans = KMeans(n_clusters=optimal_clusters, random_state=0)
clusters = kmeans.fit_predict(normalized_data)

# Add the cluster labels to the original data
data['Cluster'] = clusters

# Visualize the clusters in a 3D plot
from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure(figsize=(12, 8))
ax = fig.add_subplot(111, projection='3d')

# Scatter plot for each cluster
scatter = ax.scatter(data['Tumor Size Reduction (%)'], data['Gene Expression Score'], c=data['Cluster'], cmap='viridis', s=50, alpha=0.7)
legend1 = ax.legend(scatter.legend_elements(), title='Clusters')
ax.add_artist(legend1)

# Labels and title
ax.set_xlabel('Tumor Size Reduction (%)')
ax.set_ylabel('Gene Expression Score')
ax.set_zlabel('Side Effect Severity (0-10)')
ax.set_title('3D Visualization of Clusters')
plt.show()
```

CHARM-GPT writes code to run k-means clustering with 3 clusters.
The 3D scatter plot visualizes the clusters based on the tumor size reduction percentage, gene expression score, and side effect severity. Each color represents a different cluster identified by the k-means algorithm.

To summarize the clustering analysis:

- We normalized the data and used the elbow method to determine the optimal number of clusters.
- We performed k-means clustering with 3 clusters.
- The resulting clusters were visualized in a 3D plot.

If you need further analysis or specific insights into each cluster, please let me know! [1-3]

CHARM-GPT visualizes the 3 clusters.
The 3D scatter plot visualizes the clusters based on the tumor size reduction percentage, gene expression score, and side effect severity. Each color represents a different cluster identified by the k-means algorithm.

To summarize the clustering analysis:

- We normalized the data and used the elbow method to determine the optimal number of clusters.
- We performed k-means clustering with 3 clusters.
- The resulting clusters were visualized in a 3D plot.

If you need further analysis or specific insights into each cluster, please let me know! [3]

CHARM-GPT visualizes the 3 clusters.
Looking at the statistics for each cluster, how might one characterize them in English?
Summary

- **Cluster 0** represents patients with moderate response and moderate side effects.
- **Cluster 1** represents older patients with low response and high side effects.
- **Cluster 2** represents younger patients with high response and low side effects.
One more moving forward...
HealthyDocs: Secure medical records: $10 million
Tips on ARPA-H proposals
Highly focused calls
Short turnarounds: 3-4 weeks
~75 pages for technical section
PMs have high discretion
Needs a “secret weapon”
Makes impossible goal possible
Teaming often encouraged
Teaming can be forced post-award!
So, team or be teamed
Deliverables must be delivered
Payment only upon delivery
Budgets broken out by milestone
Contracting is a nightmare
Subcontracting is awful
Risky if deliverable spans subs
Contracting is lethargic
MATRIX: Intent to award in Nov
MATRIX: *Still* in subcontracting
BDF: Intent to award in February
BDF: *Still* in subcontracting
Use AI for RFA analysis
Use AI to make checklist
Use AI for proposal critique
Use AI for “Task Description Doc”
Also, don’t use ChatGPT
Do use Microsoft Copilot
Thank you!