Recommender System Design

LLM 2024, Course at UW, Seattle

Dr. Karthik Mohan 1/30/2024

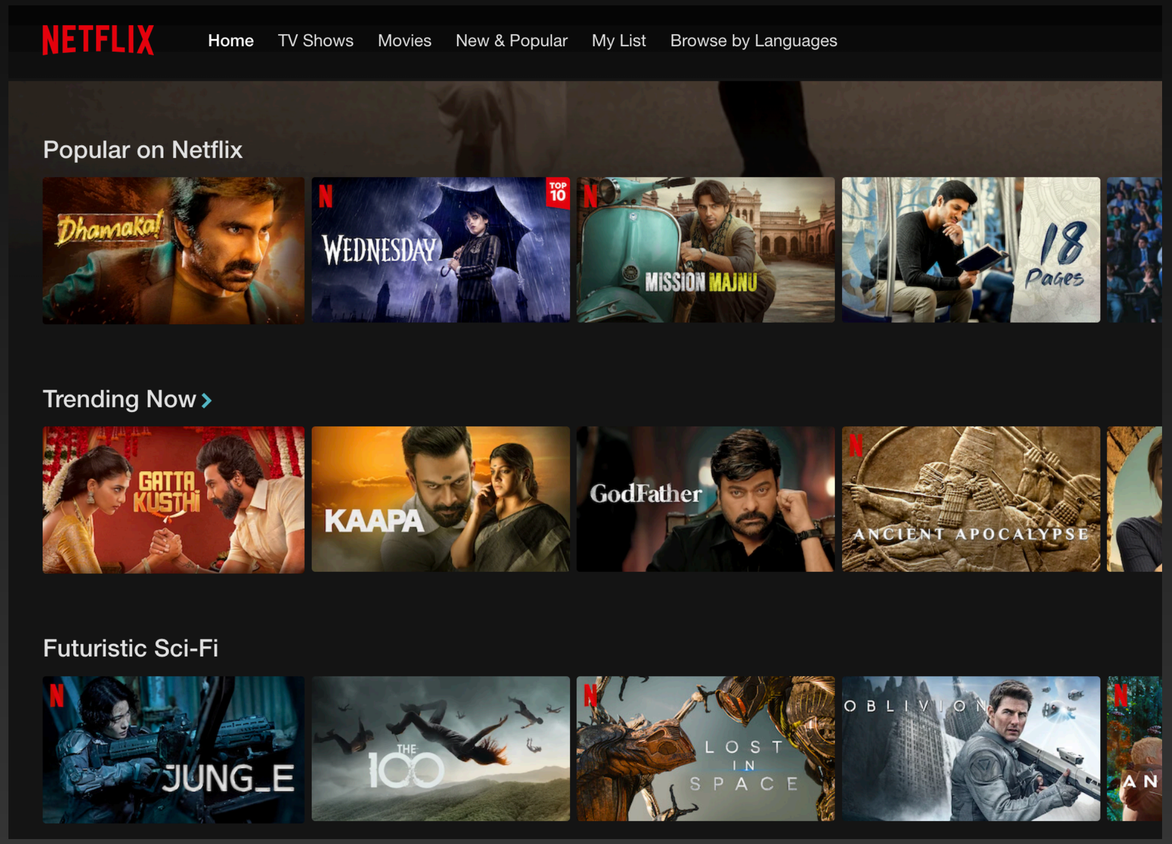
Design Considerations

- Relevance
- Freshness
- Latency
- Scalability
- Cold-Start
- Diversity
- Fairness

Design Considerations | Relevance

Design Considerations | Relevance

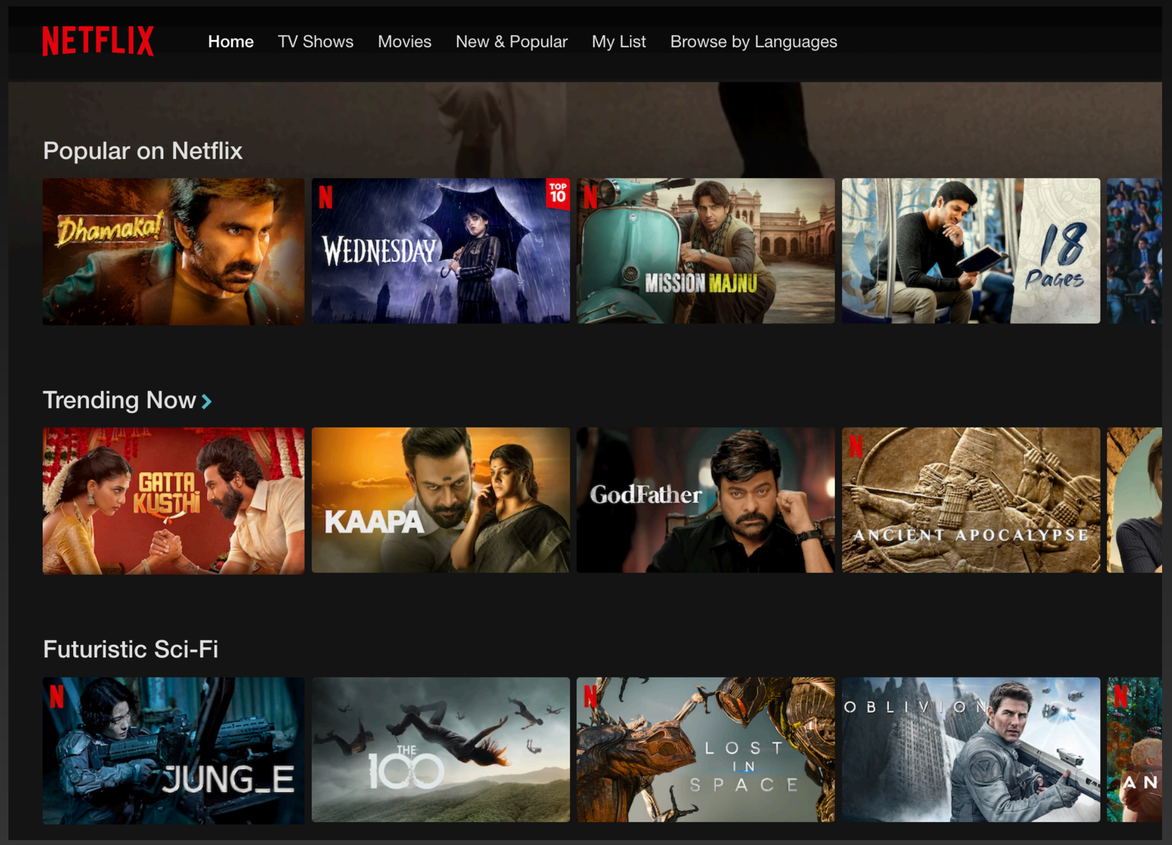
I LIKE SCI-FI Movies



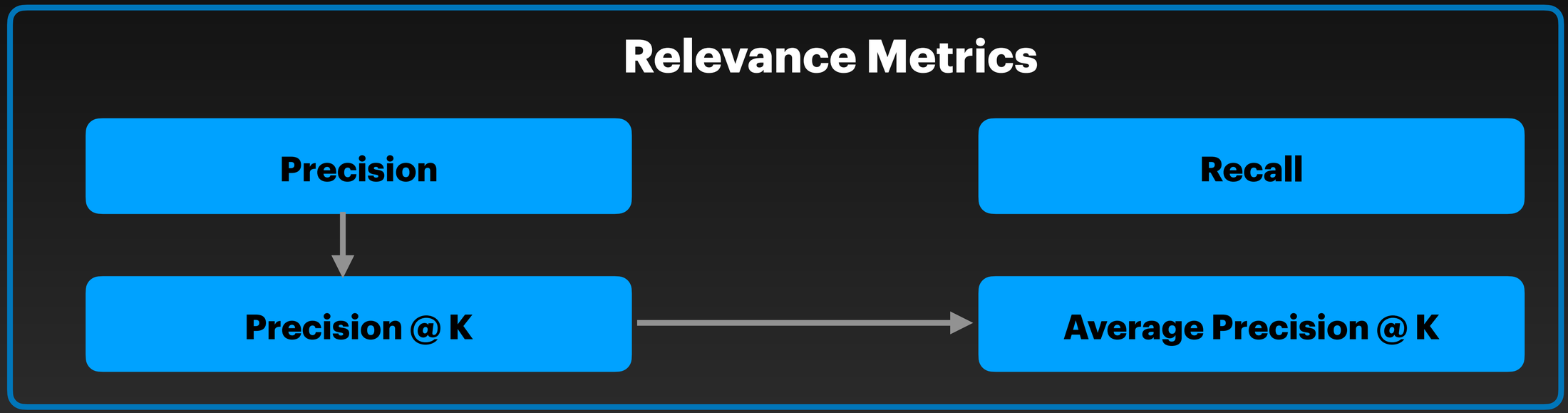
Design Considerations | Relevance

I LIKE SCI-FI Movies

I SEE
SCI-FI RECS



Design Considerataions | Relevance



Modeling Metrics

Relevance

	1	2	3	4	5
Purchases					

Modeling Metrics

Relevance

	1	2	3	4	5
Purchases					
Recommendations					

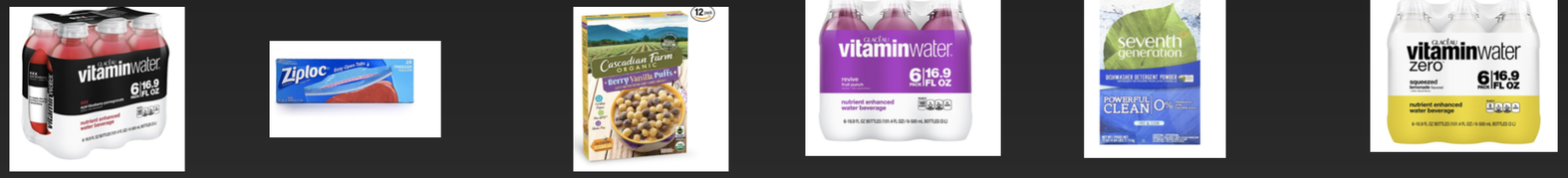
Modeling Metrics

Relevance

Purchases

1 2 3 4 5

Recommendations



1 2 3 4 5 6

Modeling Metrics

Relevance

Purchases

1 

2 

3 

4 

5 

Recommendations

1

2

3

4

5

6

Precision

Modeling Metrics

Relevance

Purchases

1 

2 

3 

4 

5 

Recommendations

1

2

3

4

5

6

Precision

Modeling Metrics

Relevance

Purchases

1 

2 

3 

4 

5 

Recommendations

1

2

3

4

5

6

Precision 4/6 = 0.67

Modeling Metrics

Relevance

Purchases

1 

2 

3 

4 

5 

Recommendations

1

2

3

4

5

6

Recall

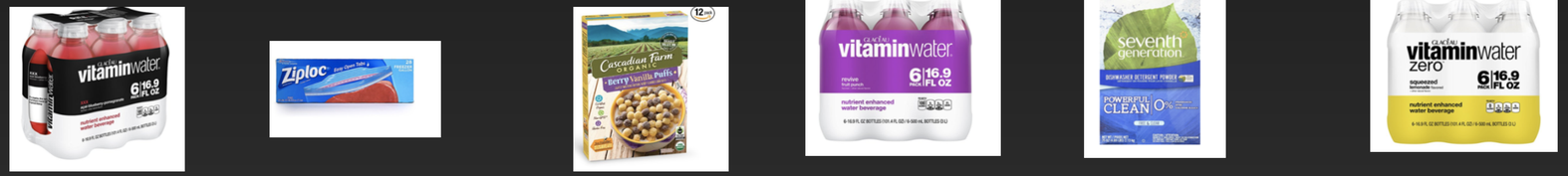
Modeling Metrics

Relevance

Purchases

1 2 3 4 5

Recommendations



1 2 3 4 5 6

Recall 4/5 = 0.8

Modeling Metrics

Relevance

Purchases

1 

2 

3 

4 

5 

Recommendations

1

2

3

4

5

6

Precision @ 4

Modeling Metrics

Relevance

Purchases

1 

2 

3 

4 

5 

Recommendations

1

2

3

4

5

6

Precision @ 4

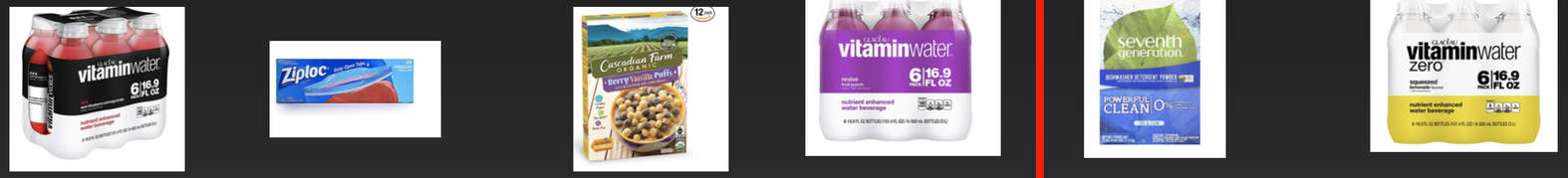
Modeling Metrics

Relevance

Purchases

1 2 3 4 5

Recommendations



1 2 3 4 5 6

Precision @ 4 2/4 = 0.5

Modeling Metrics

Relevance

Purchases

1 

2 

3 

4 

5 

Recommendations

1

2

3

4

5

6

Average Precision @ 4

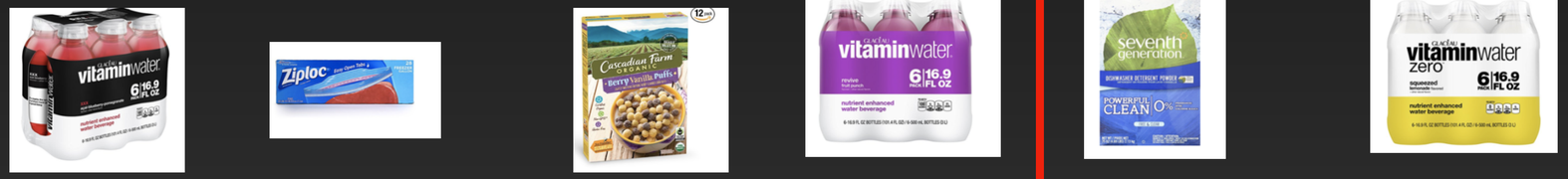
Modeling Metrics

Relevance

Purchases

1 2 3 4 5

Recommendations



1 2 3 4 5 6

Average Precision @ 4

(1 + 2/4)/4 = 0.375

Precision @ 4

Modeling Metrics

Relevance

Purchases

1 

2 

3 

4 

5 

Recommendations

1

2

3

4

5

6

Average Precision @ 4

(1 + 2/4)/4 = 0.375

Precision @ 4

2/4 = 0.5

ICE: Modeling Metrics

Relevance

Purchases

1 2 3 4 5

Recommendations

1 2 3 4 5 6

Average Precision @ 4

Performance Metrics and Design

Modeling Metrics

Design Considerations

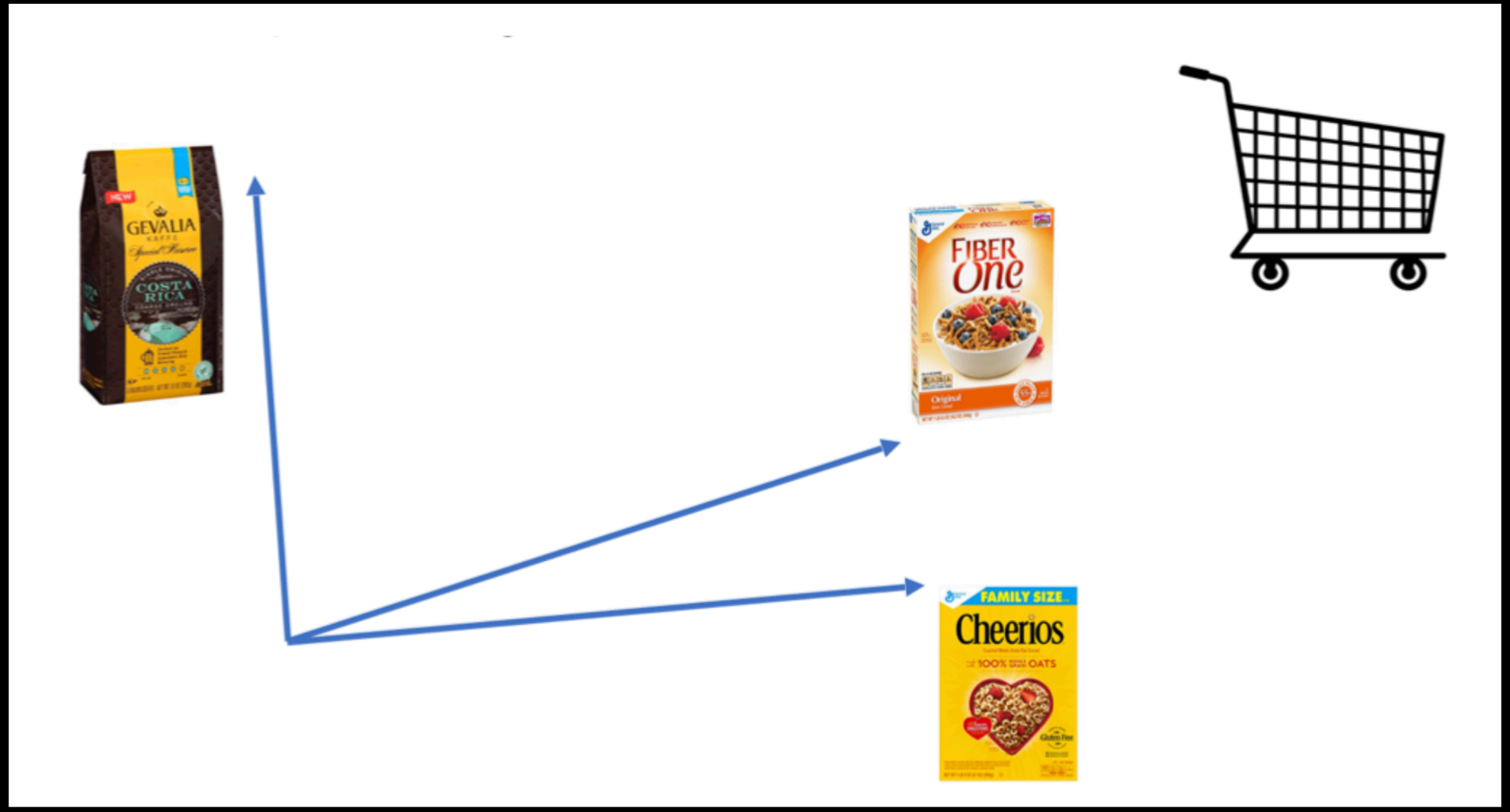
Business Metrics

Design Considerations

- Relevance
- Diversity
- Freshness
- Fairness
- Latency

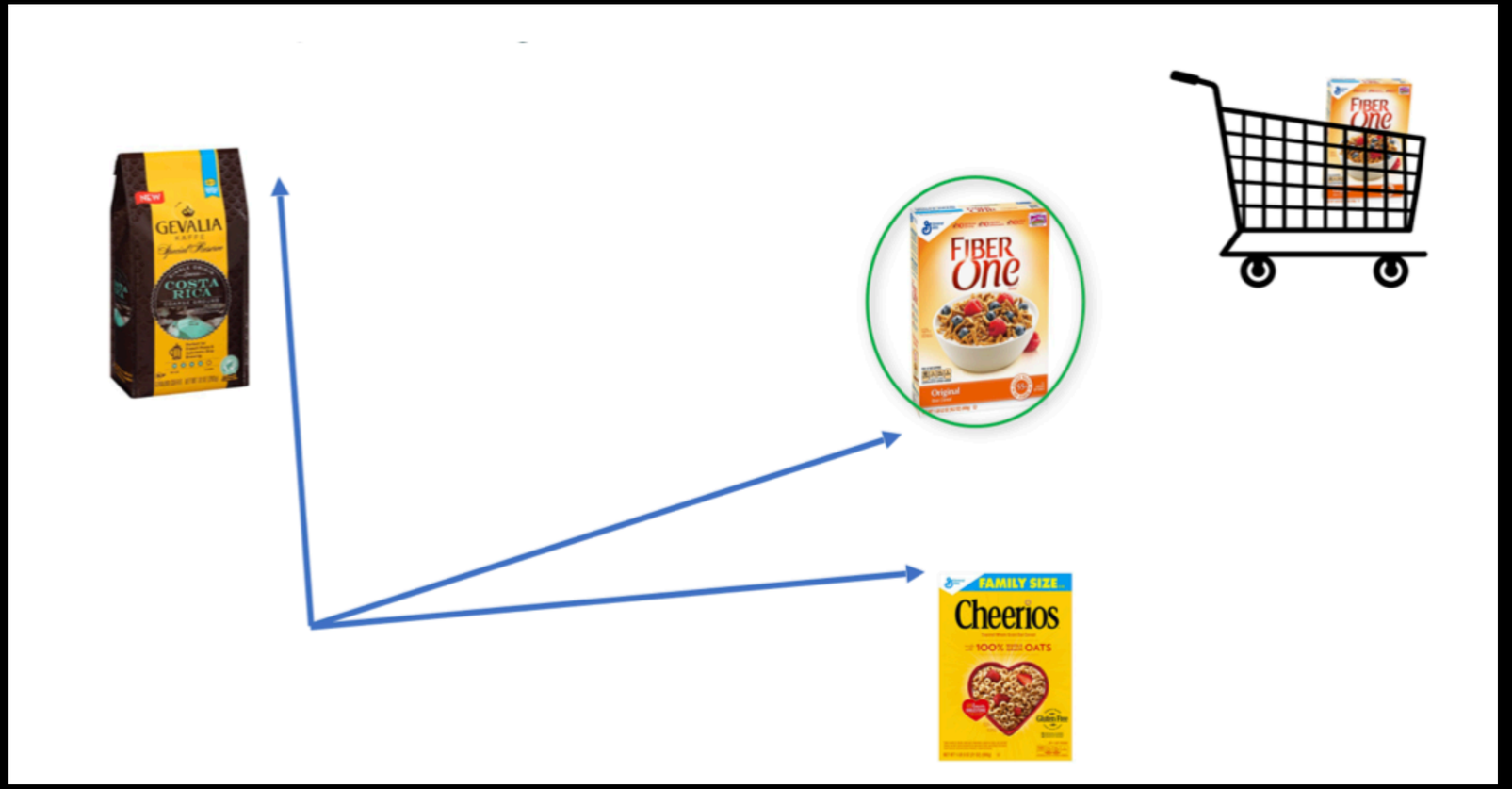
Diversity

In Online Shopping



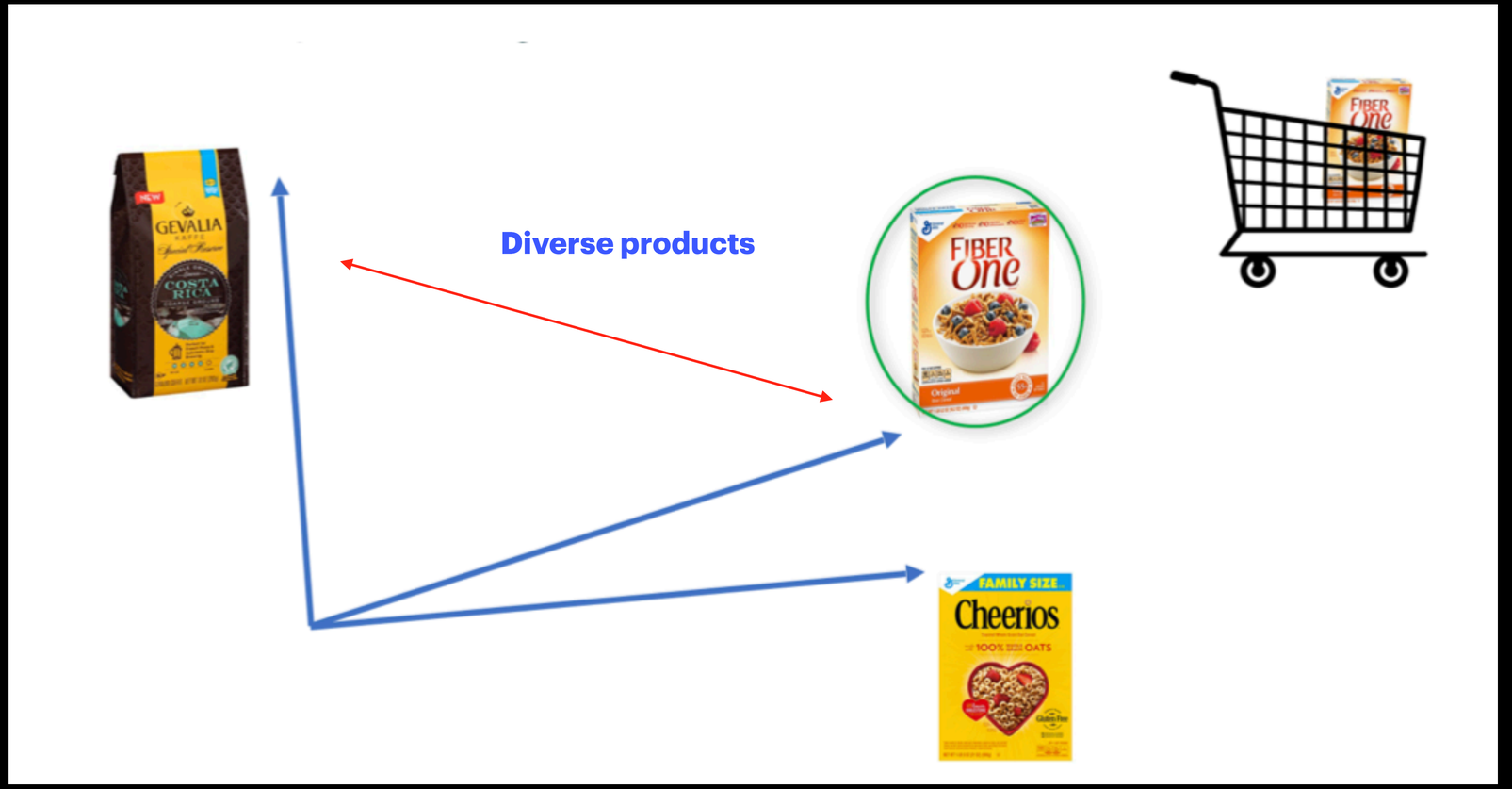
Diversity

In Online Shopping



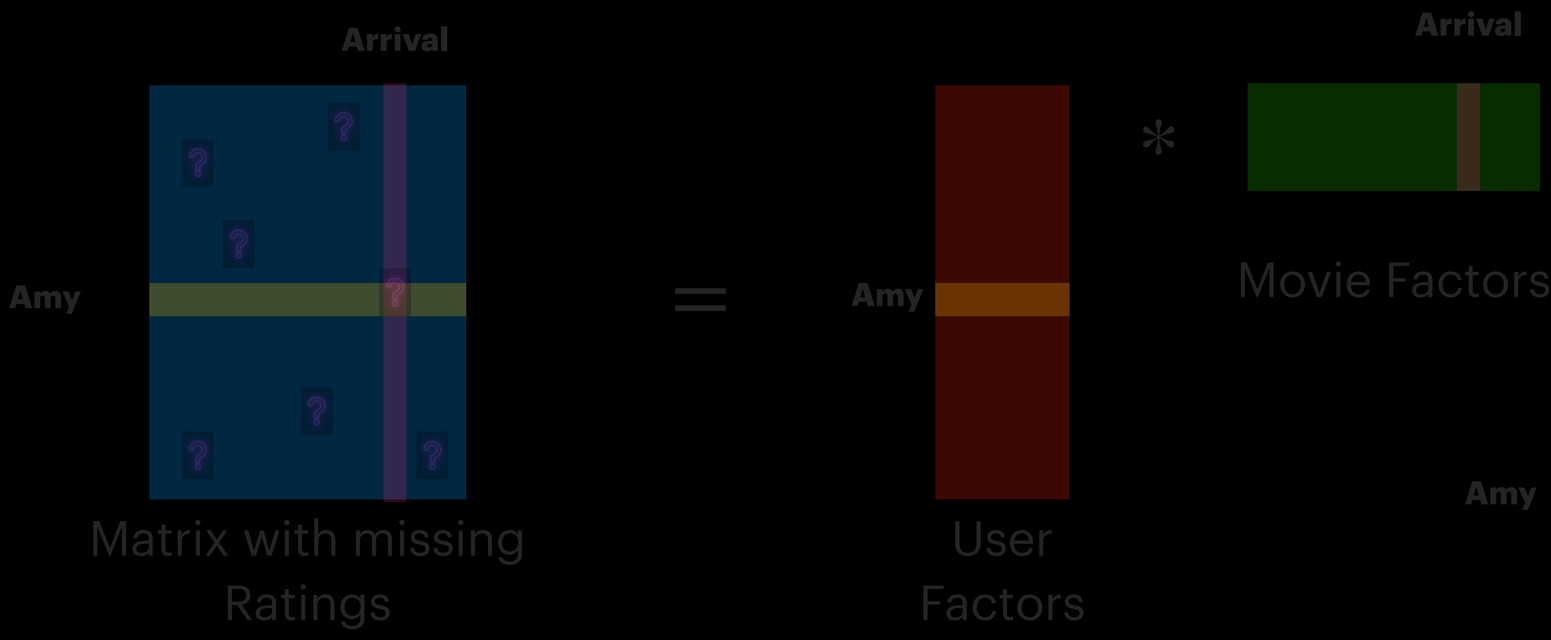
Diversity

In Online Shopping

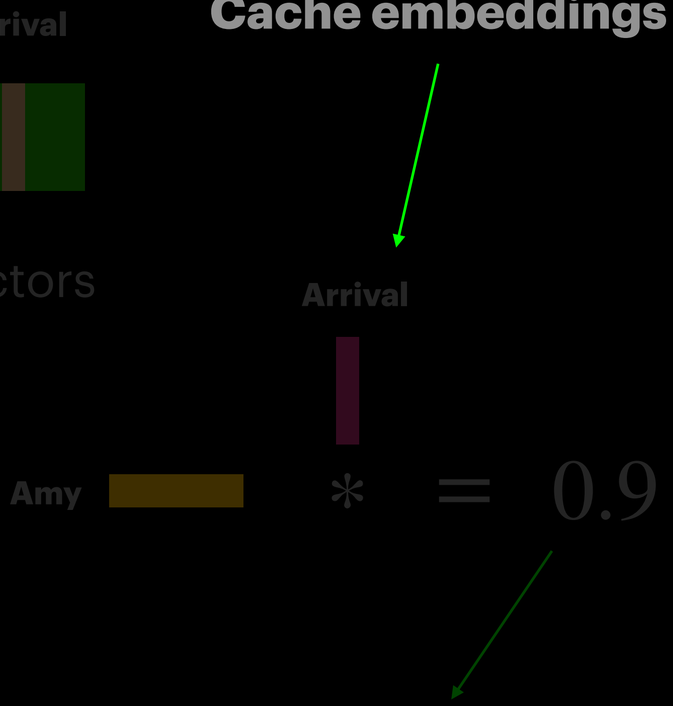


Design Considerations | Latency

Design Considerations | Latency



Low Latency as can Cache embeddings



Model predicts Amy Likes Arrival

Design Considerations | Scalability

#Movies	Training	Prediction Time
$100K$	Easy	Low Latency
$1M$	Easy	Low Latency
$10M$	Complex	
$1B$	Complex	High Latency

Design Considerations | Scalability

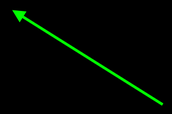
#Movies	Training	Prediction Time
$100K$	Easy	Low Latency
$1M$	Easy	Low Latency
$10M$	Complex	
$1B$	Complex	High Latency

Design Considerations | Scalability

#Movies	Training	Prediction Time
$100K$	Easy	Low Latency
$1M$	Easy	Low Latency
$10M$	Complex	
$1B$	Complex	High Latency

Design Considerations | Scalability

#Movies	Training	Prediction Time
$100K$	Easy	Low Latency
$1M$	Easy	Low Latency
$10M$	Complex	
$1B$	Complex	High Latency



Neural Networks don’t scale due to memory and compute constraints

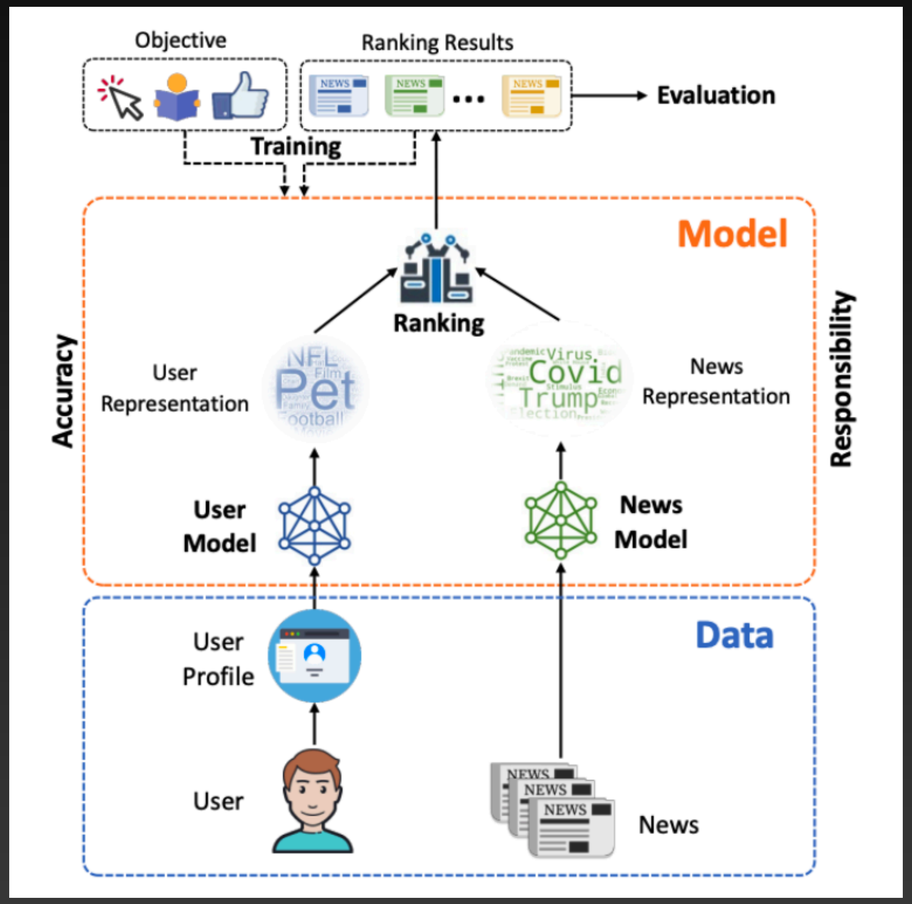
Design Considerations | Scalability

#Movies	Training	Prediction Time
$100K$	Easy	Low Latency
$1M$	Easy	Low Latency
$10M$	Complex	
$1B$	Complex	High Latency

Need a multi-stage approach

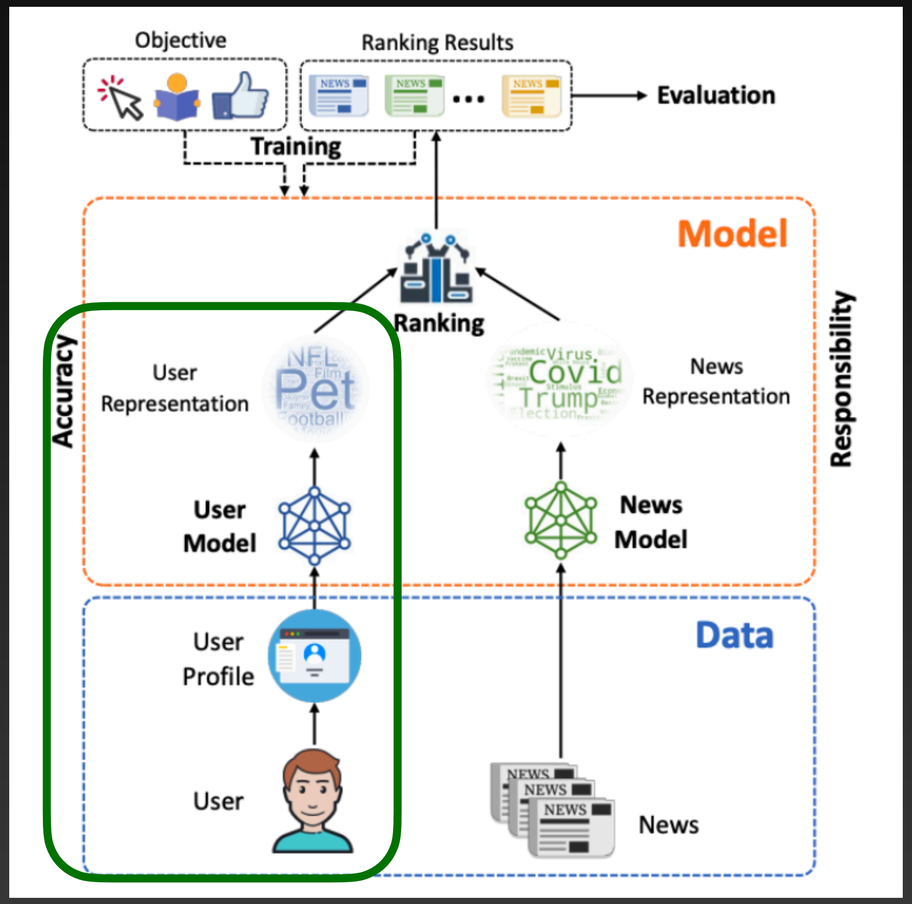
Two Tower Architecture

A popular RecSys Architecture



Two Tower Architecture

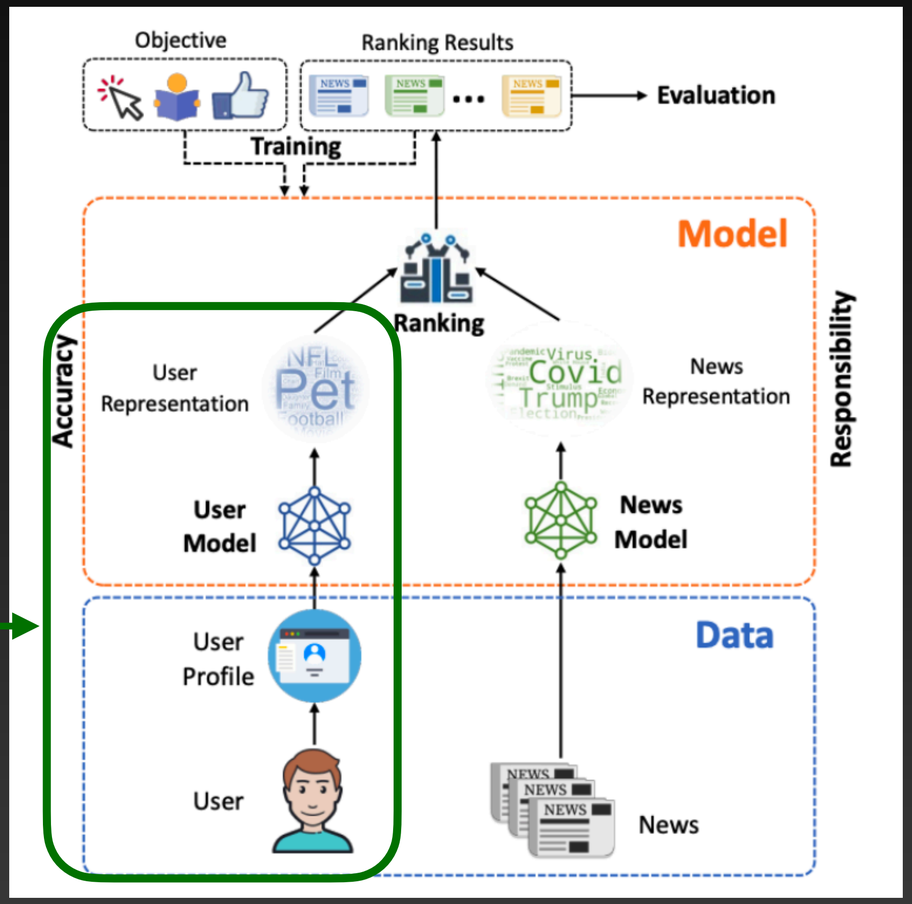
A popular RecSys Architecture



Two Tower Architecture

A popular RecSys Architecture

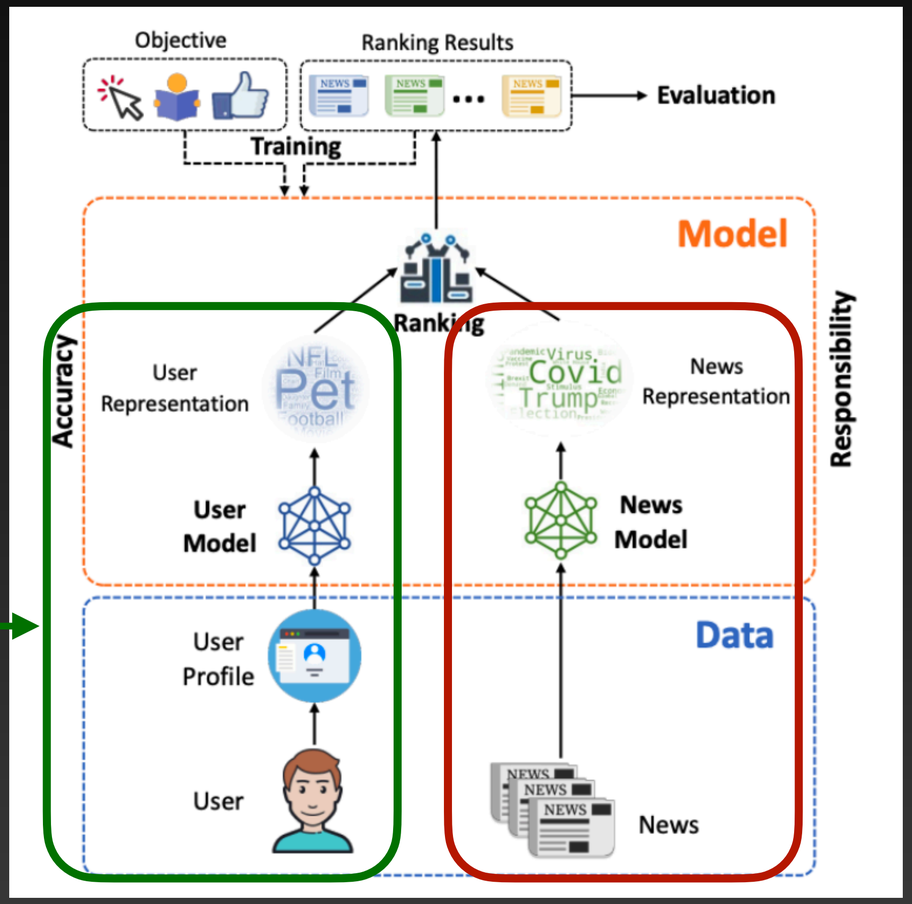
User Tower



Two Tower Architecture

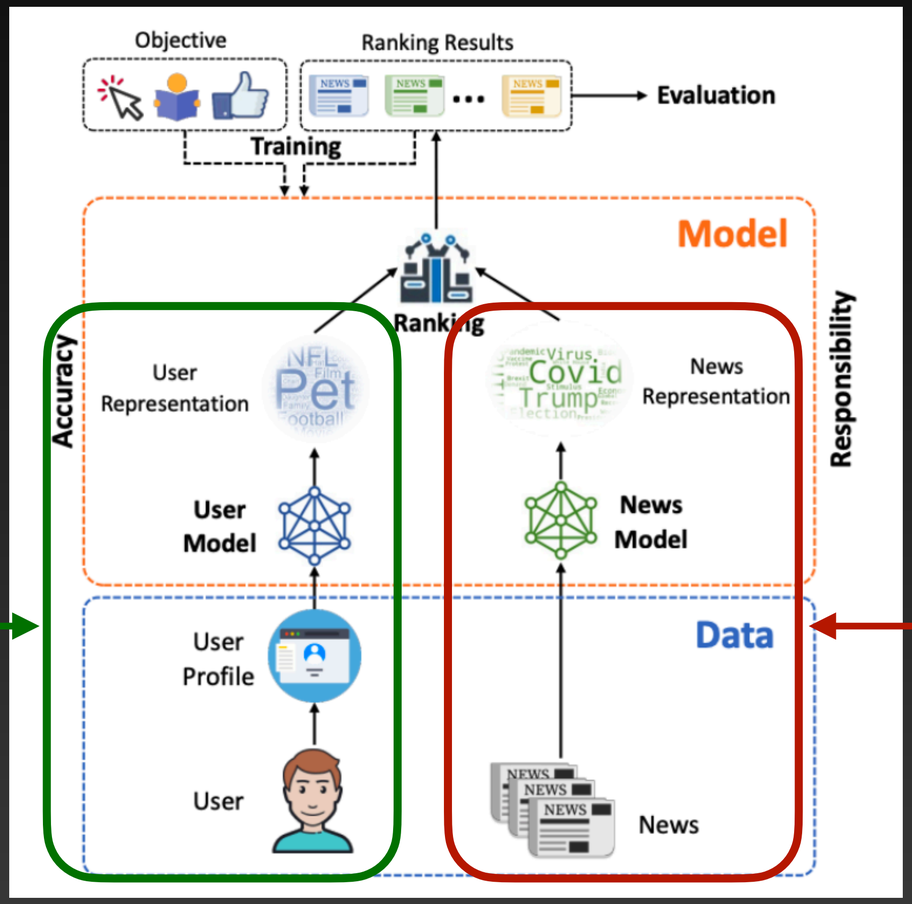
A popular RecSys Architecture

User Tower



Two Tower Architecture

A popular RecSys Architecture

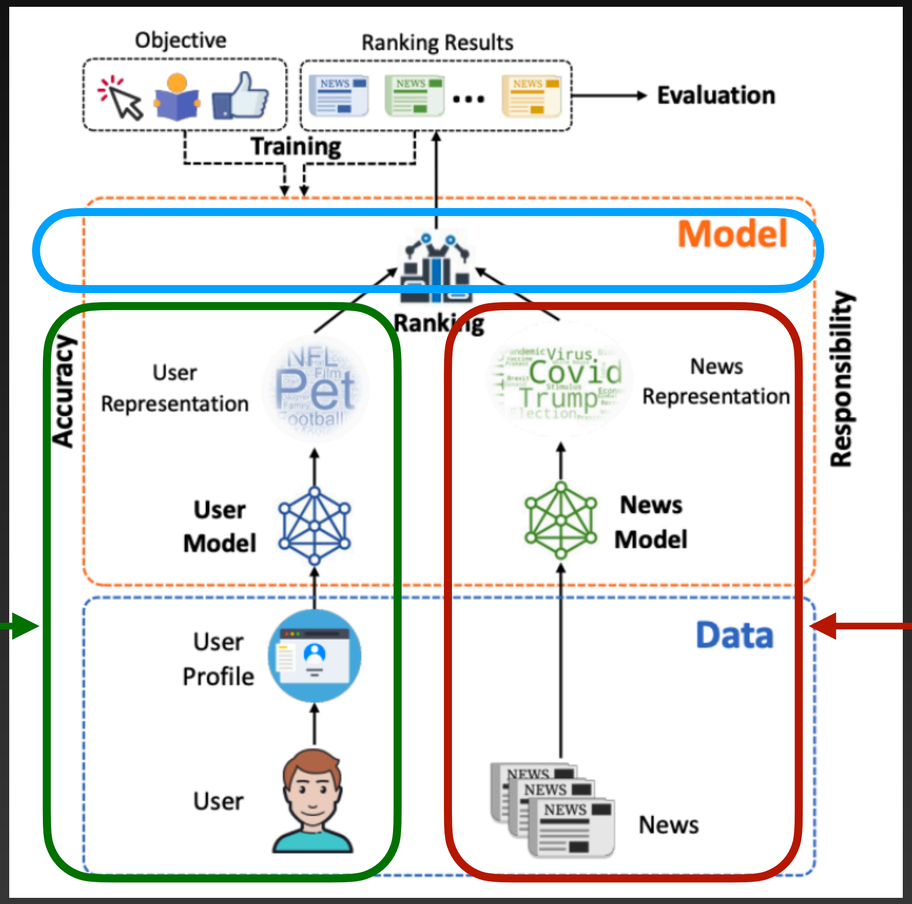


User Tower

Product Tower

Two Tower Architecture

A popular RecSys Architecture

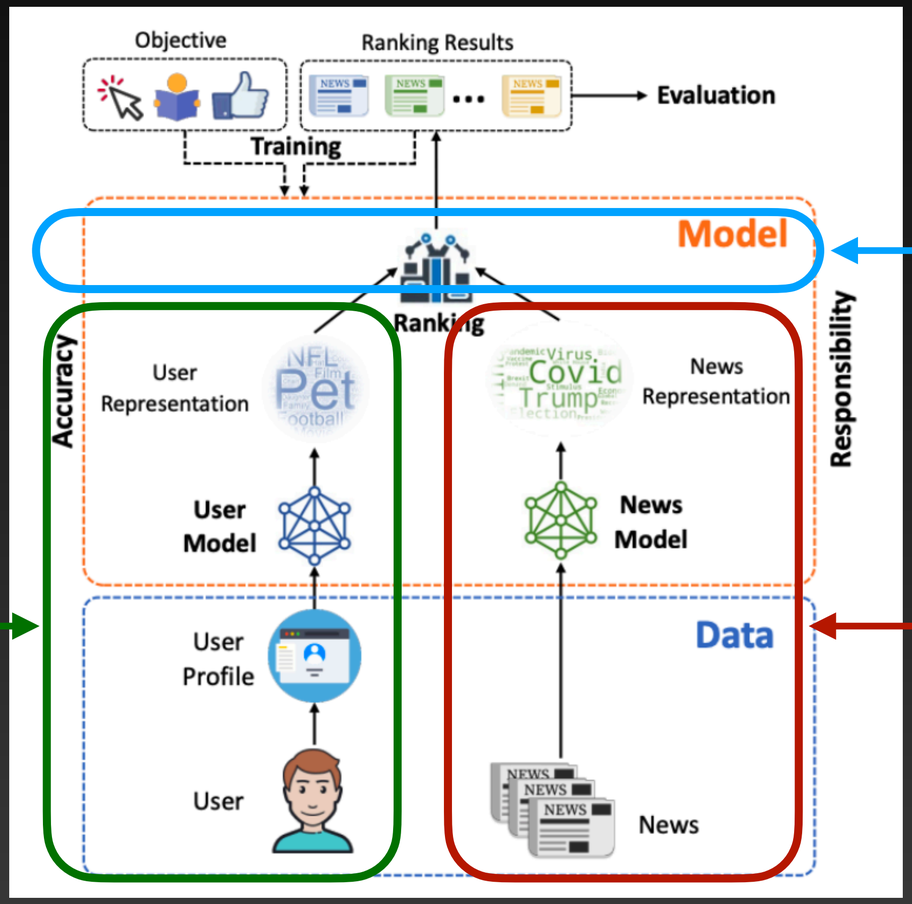


User Tower

Product Tower

Two Tower Architecture

A popular RecSys Architecture



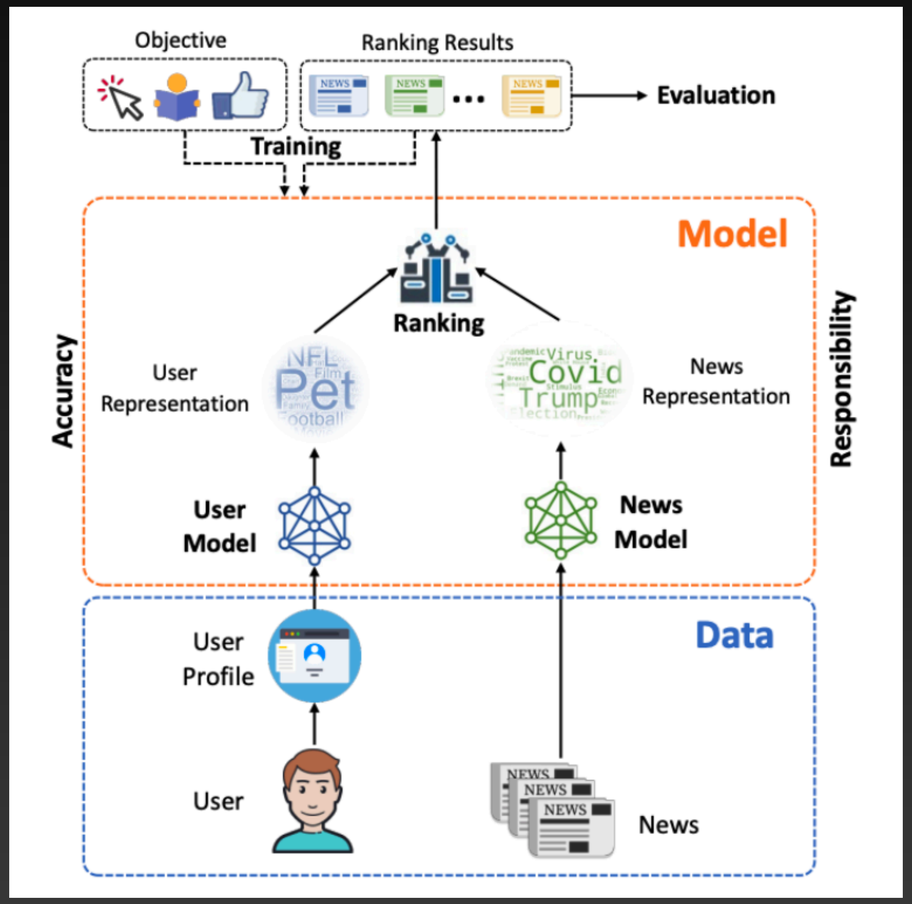
User Tower

User-Product Interactions

Product Tower

Two Tower Architecture

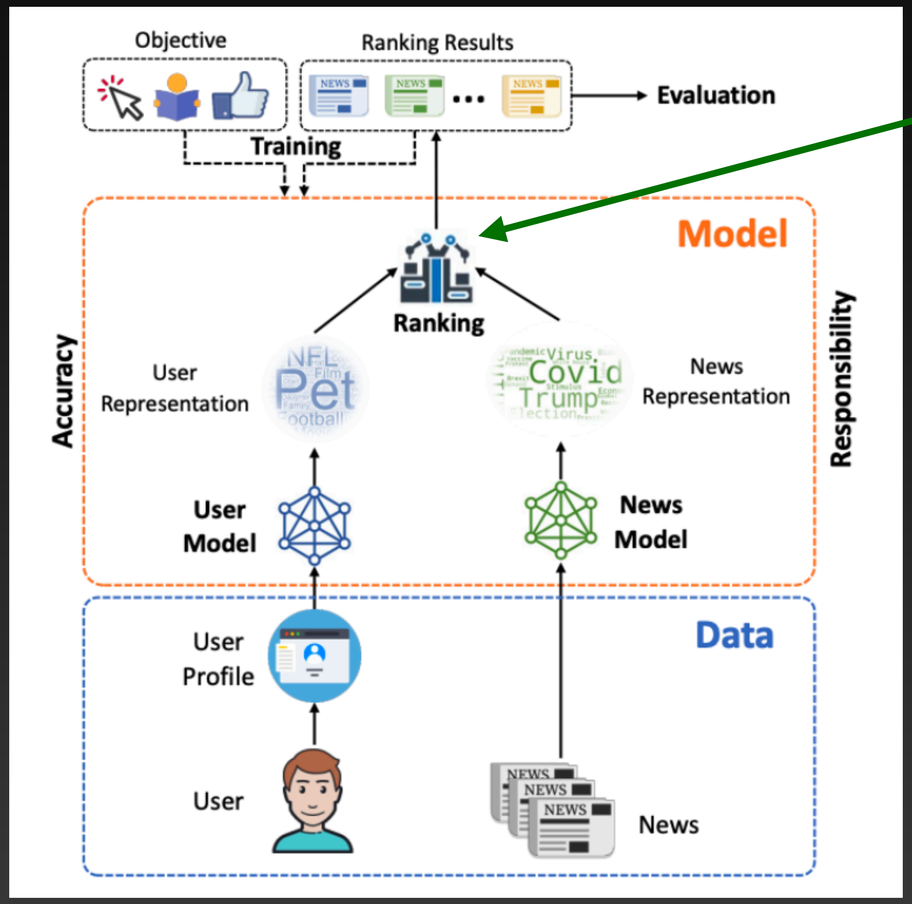
A popular RecSys Architecture



Hybrid Model

Two Tower Architecture

A popular RecSys Architecture

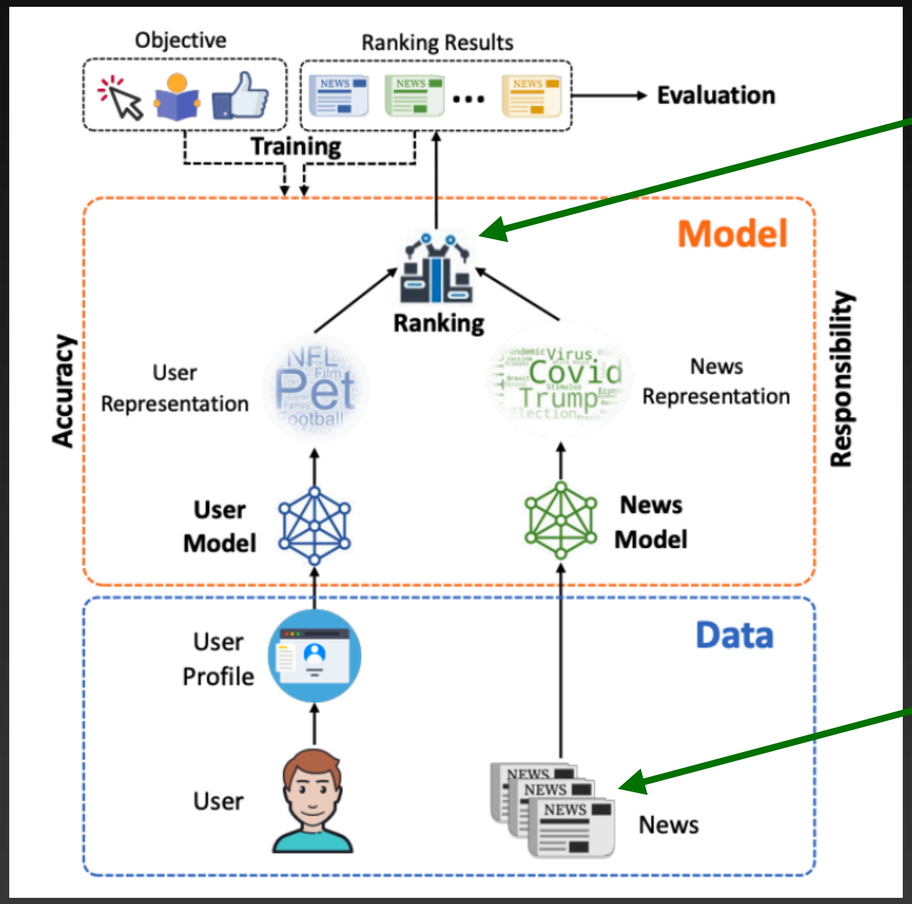


Collaborative Filtering

Hybrid Model

Two Tower Architecture

A popular RecSys Architecture



Collaborative Filtering

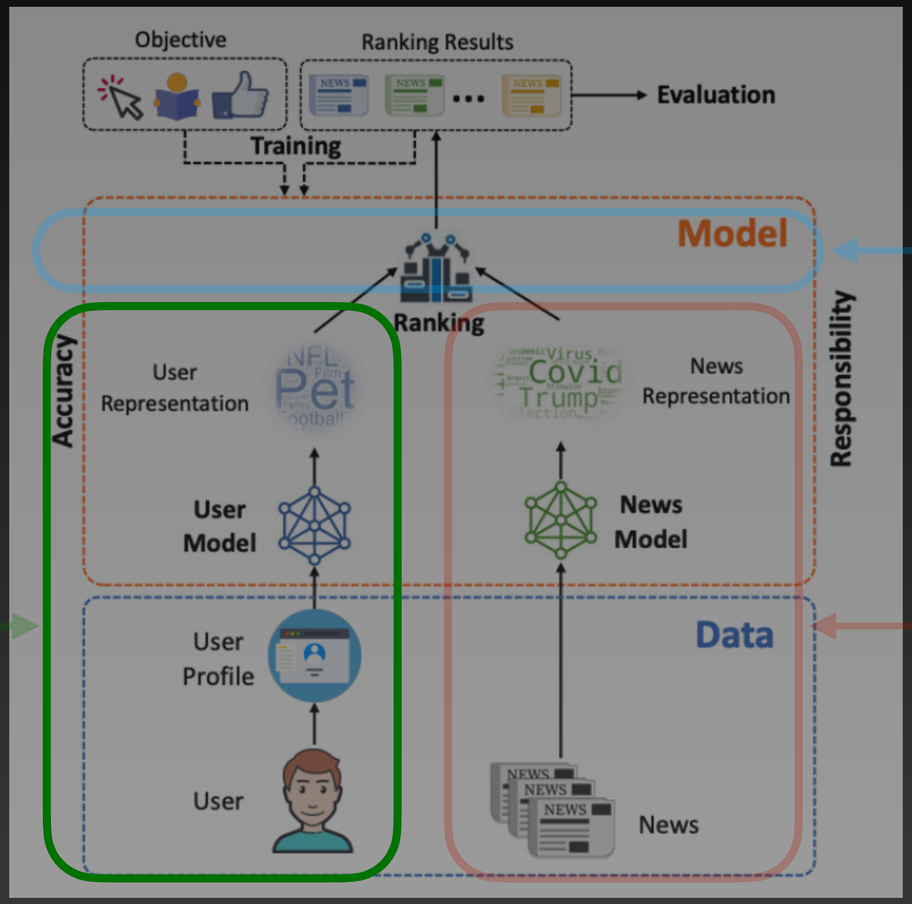
Hybrid Model

Content Based Filtering

Two Tower Architecture

A popular RecSys Architecture

Hybrid
Relevance
Diversity
Cold-Start



User-Product Interactions

User Tower

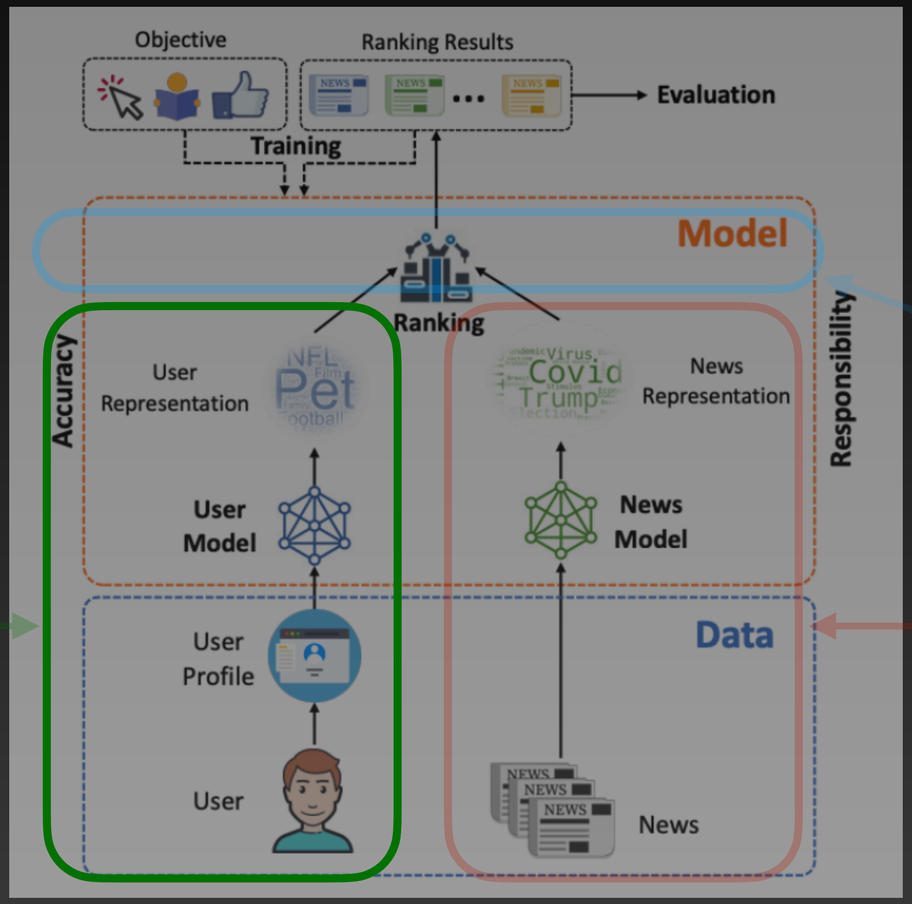
Product Tower

Two Tower Architecture

A popular RecSys Architecture

Hybrid
Relevance
Diversity
Cold-Start

Doesn't Scale even with Caching



User-Product Interactions

User Tower

Product Tower

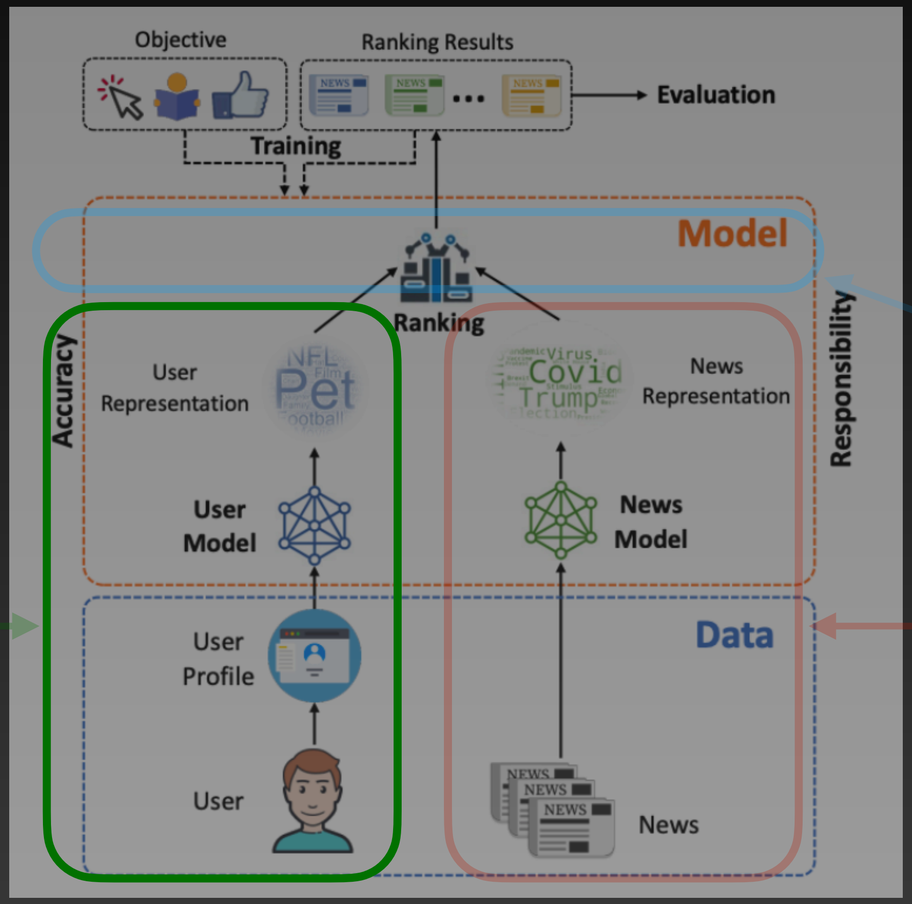
Two Tower Architecture

A popular RecSys Architecture

Hybrid
Relevance
Diversity
Cold-Start

Need a multi-stage approach

Doesn't Scale even with Caching

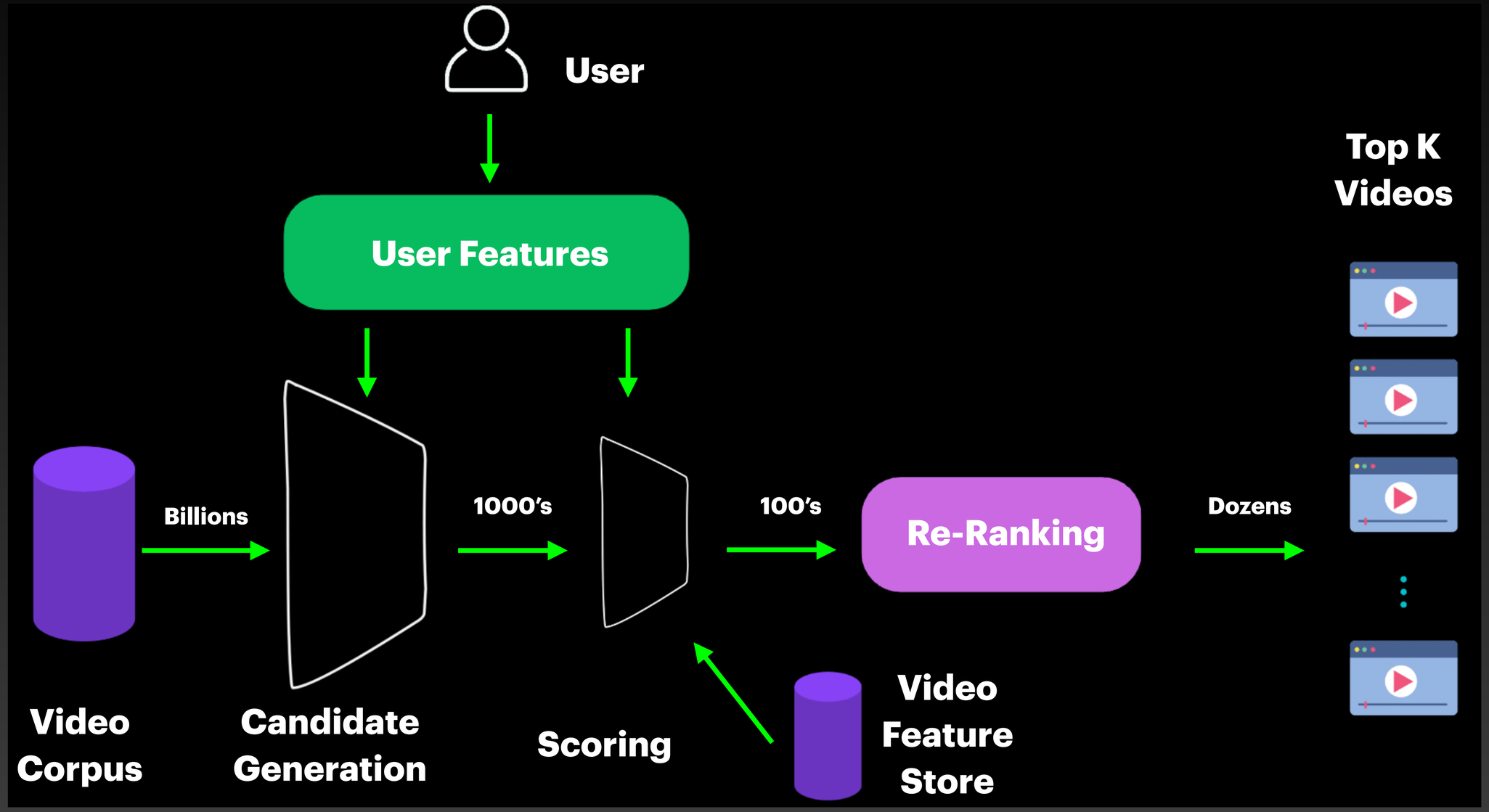


User-Product Interactions

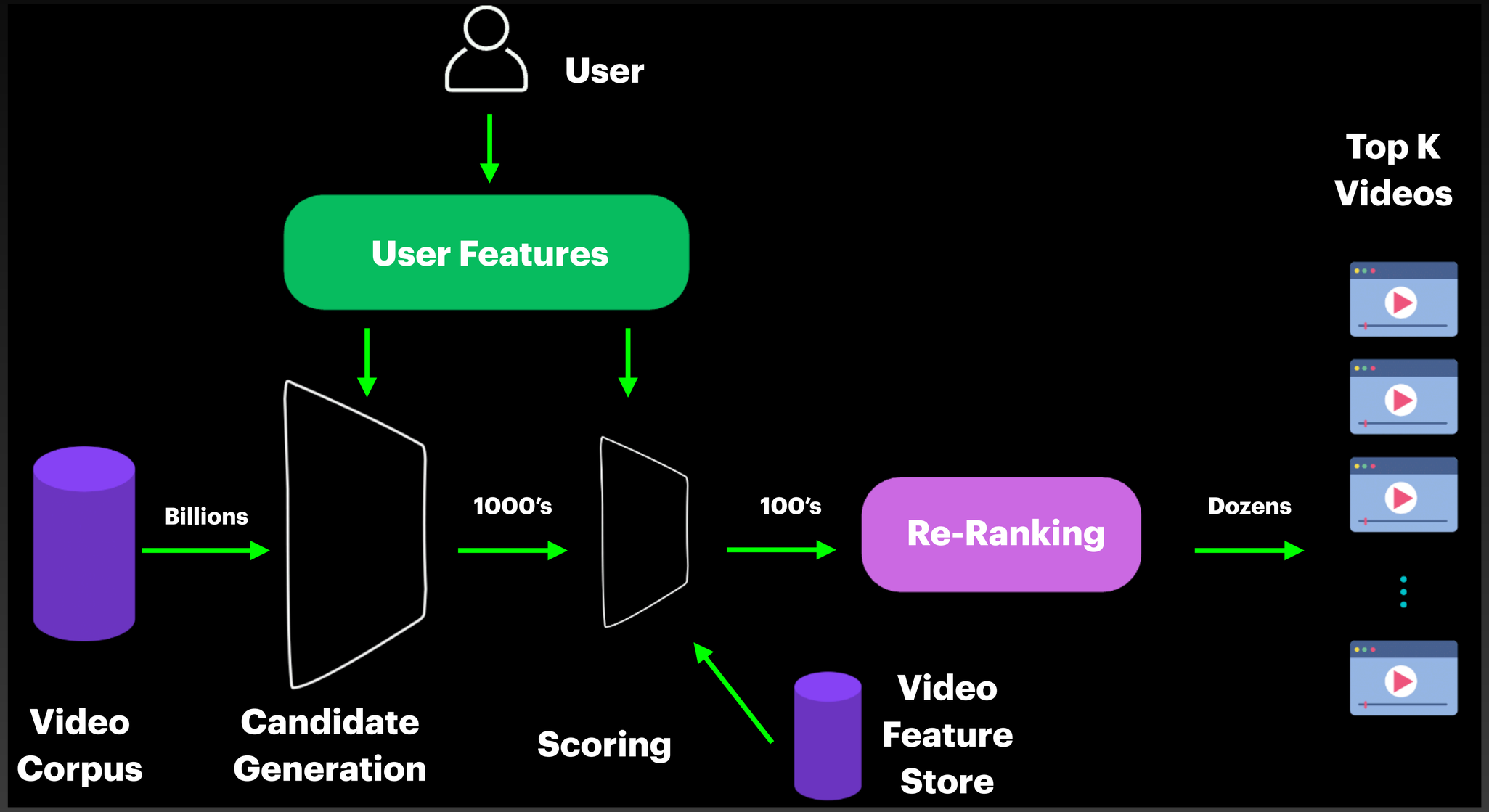
User Tower

Product Tower

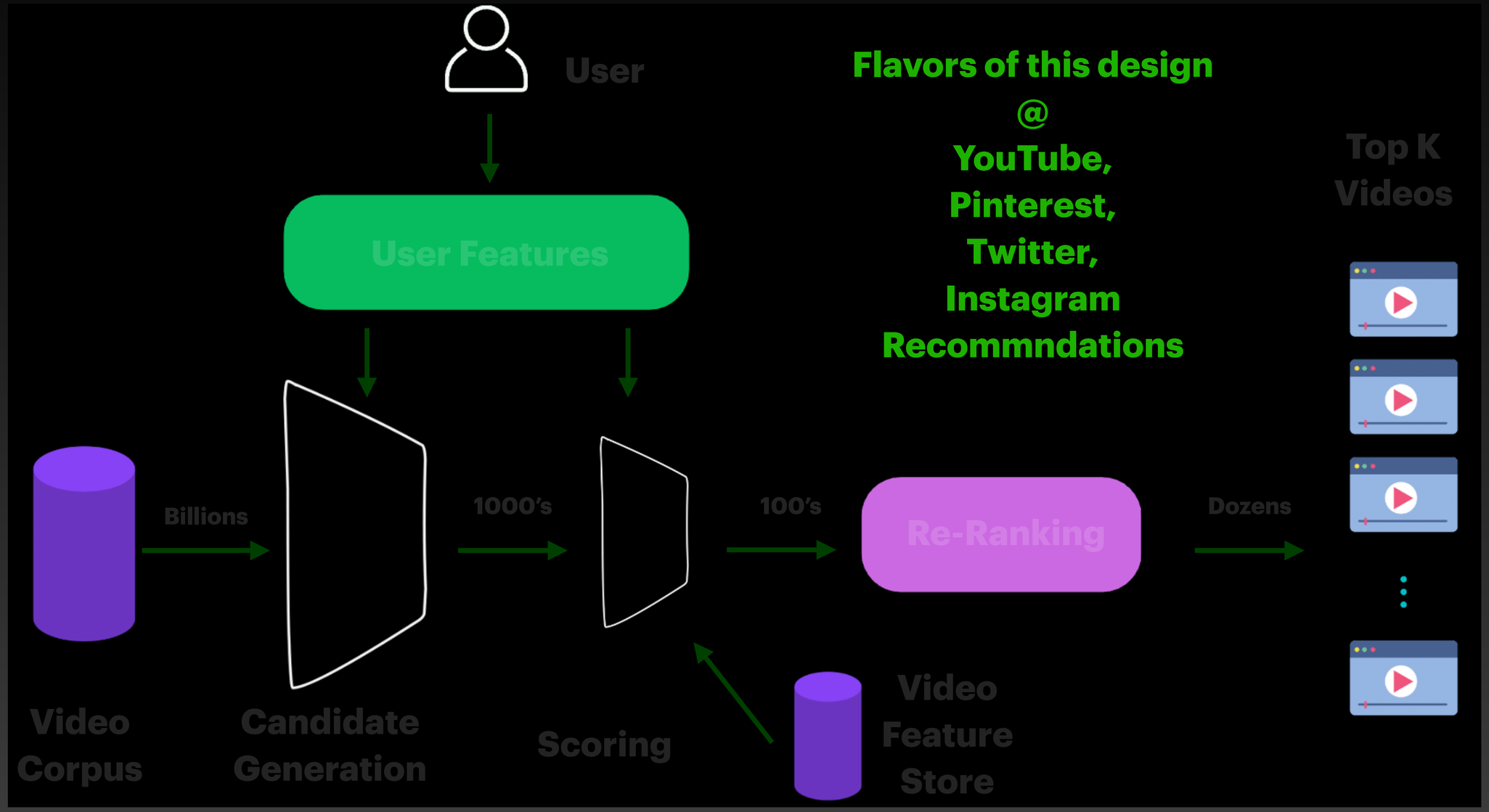
Scalable RecSys Design



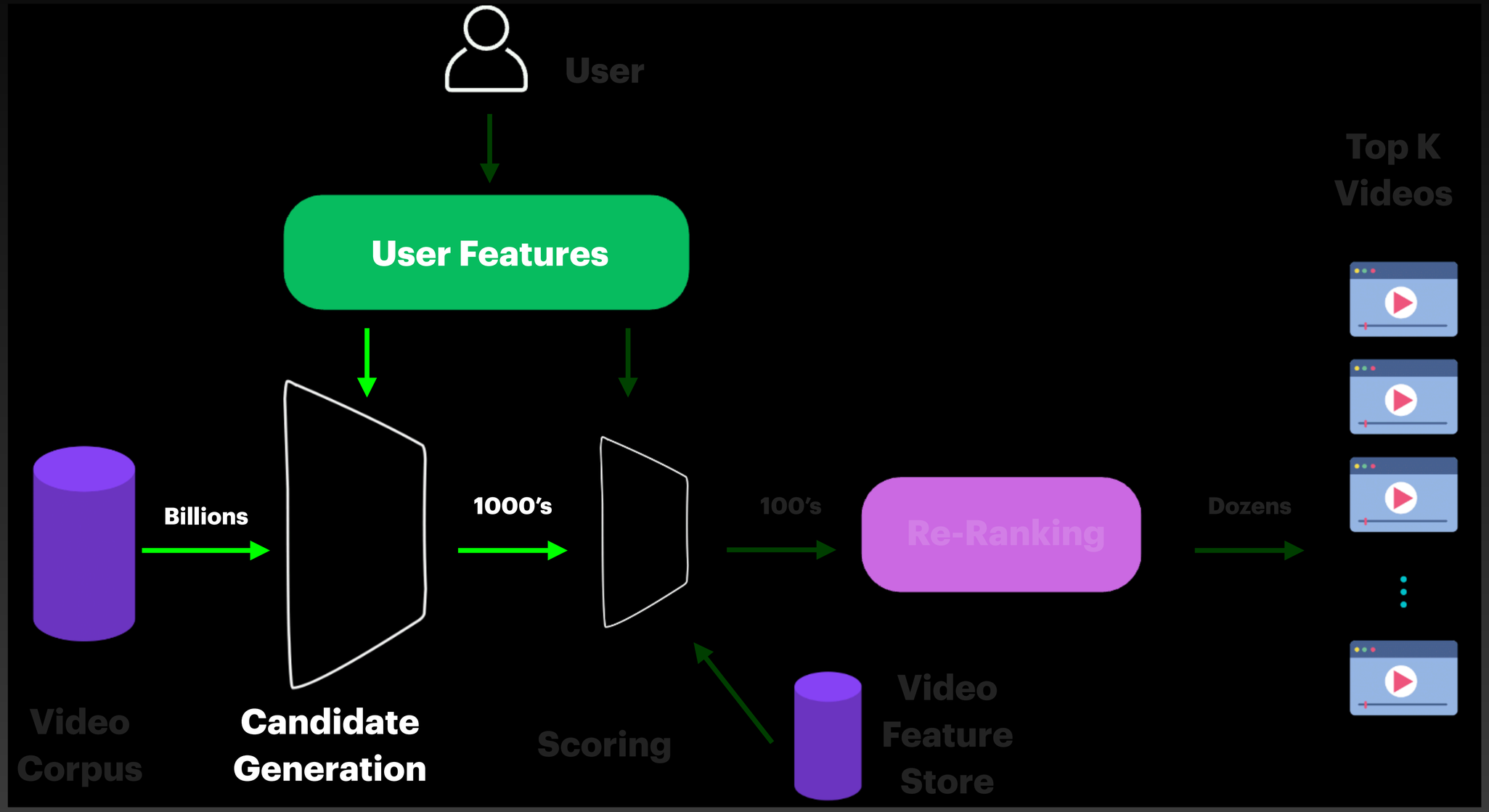
Scalable RecSys Design



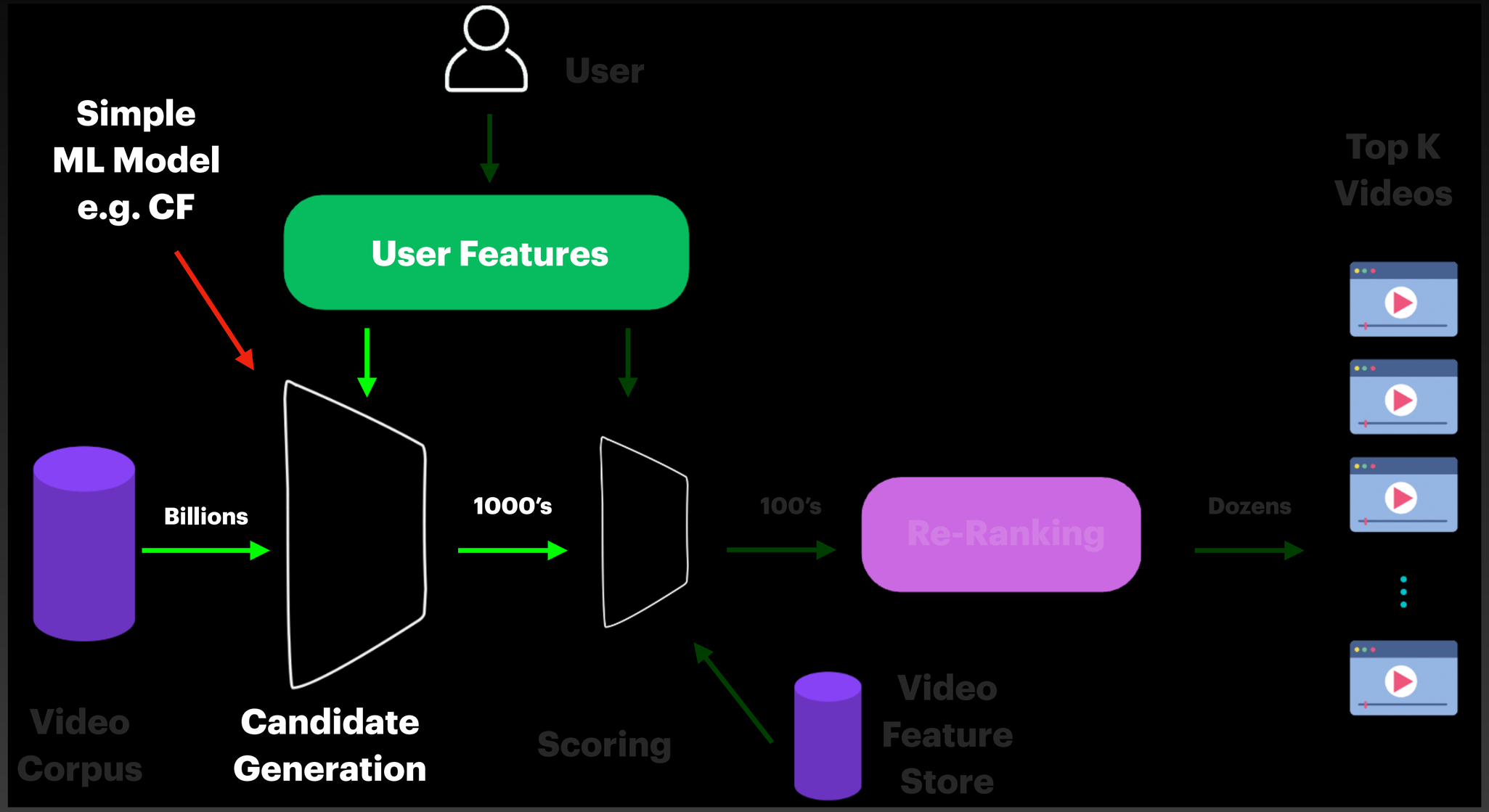
Scalable RecSys Design



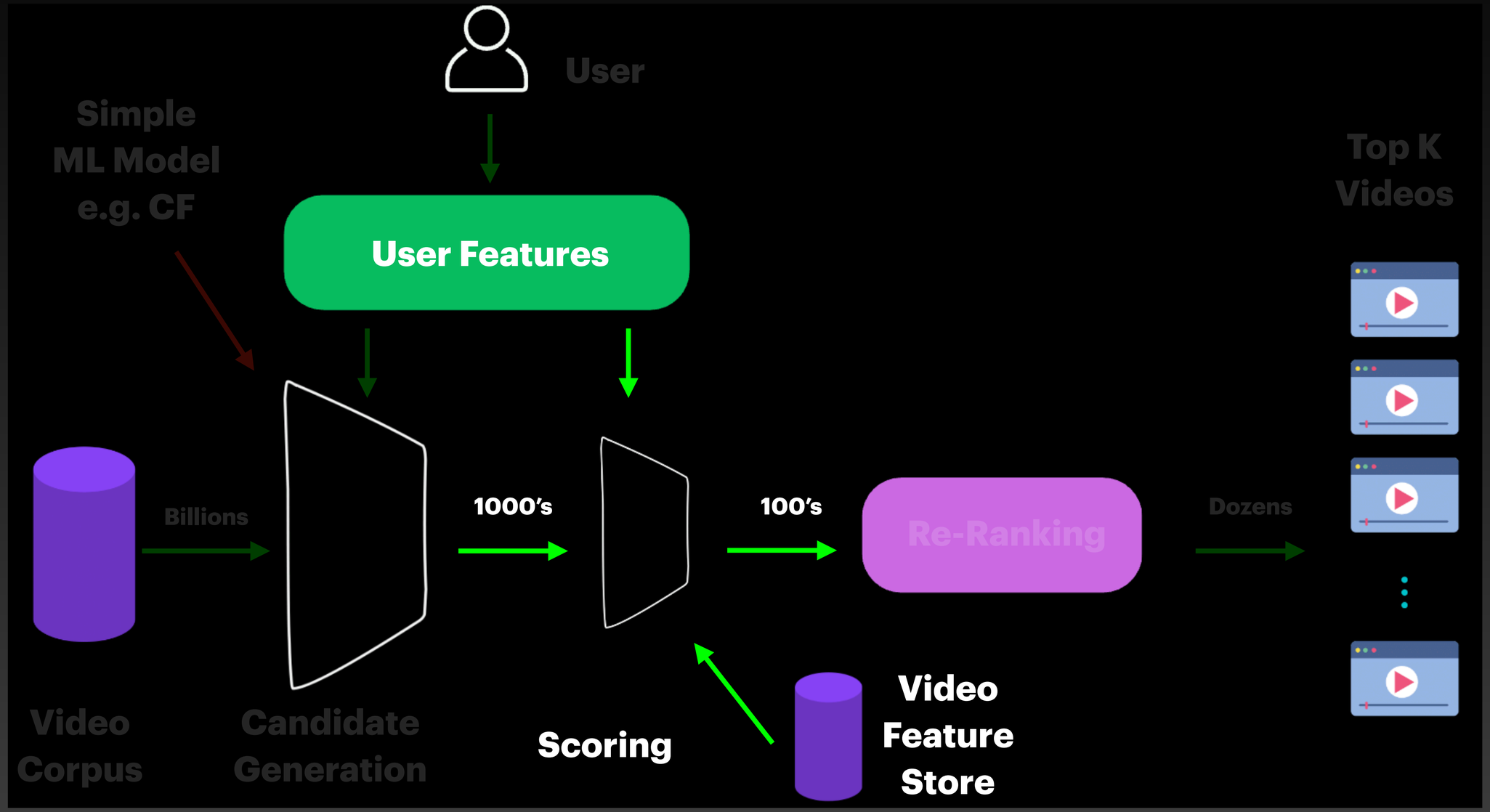
Scalable RecSys Design



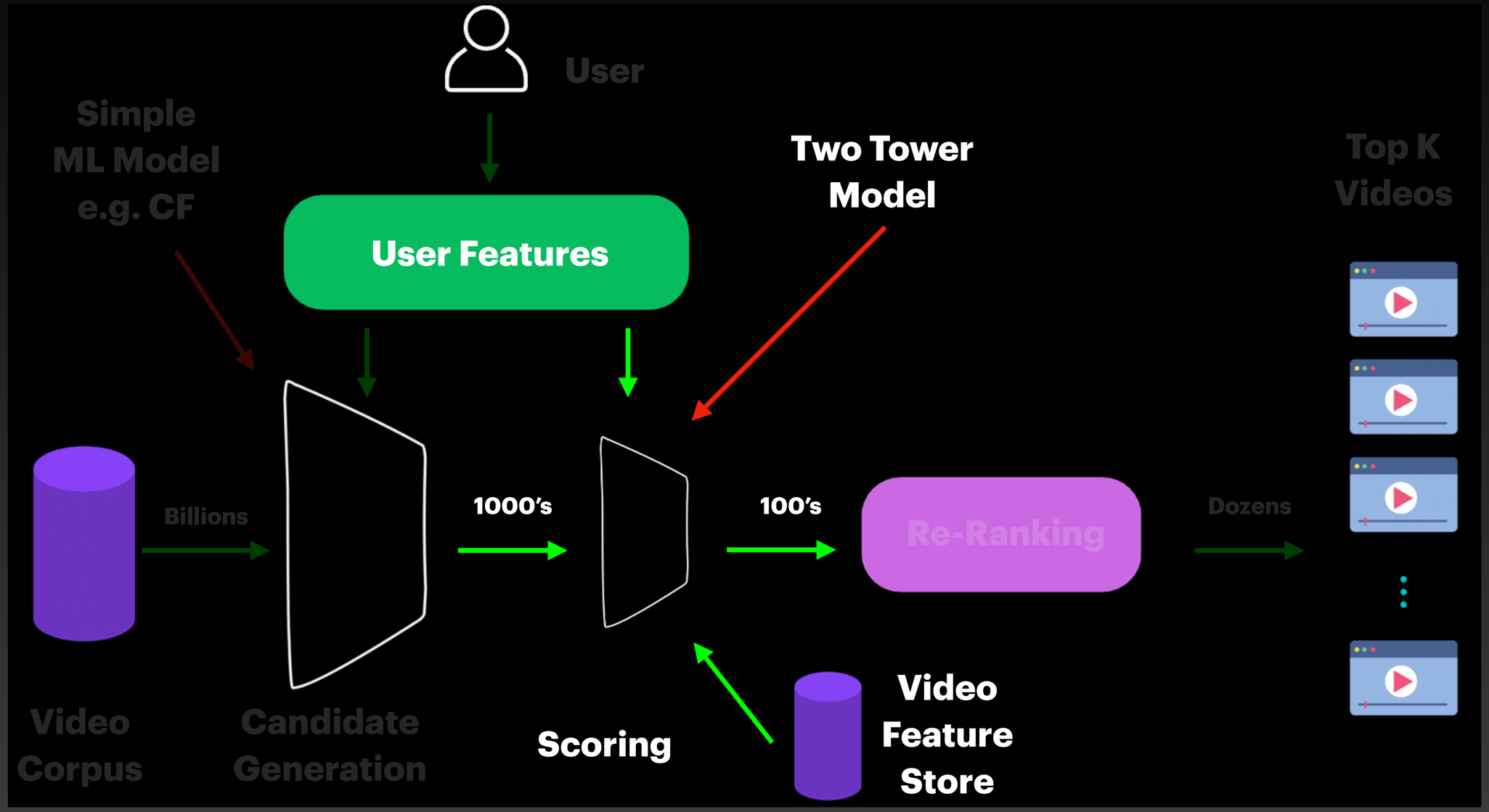
Scalable RecSys Design



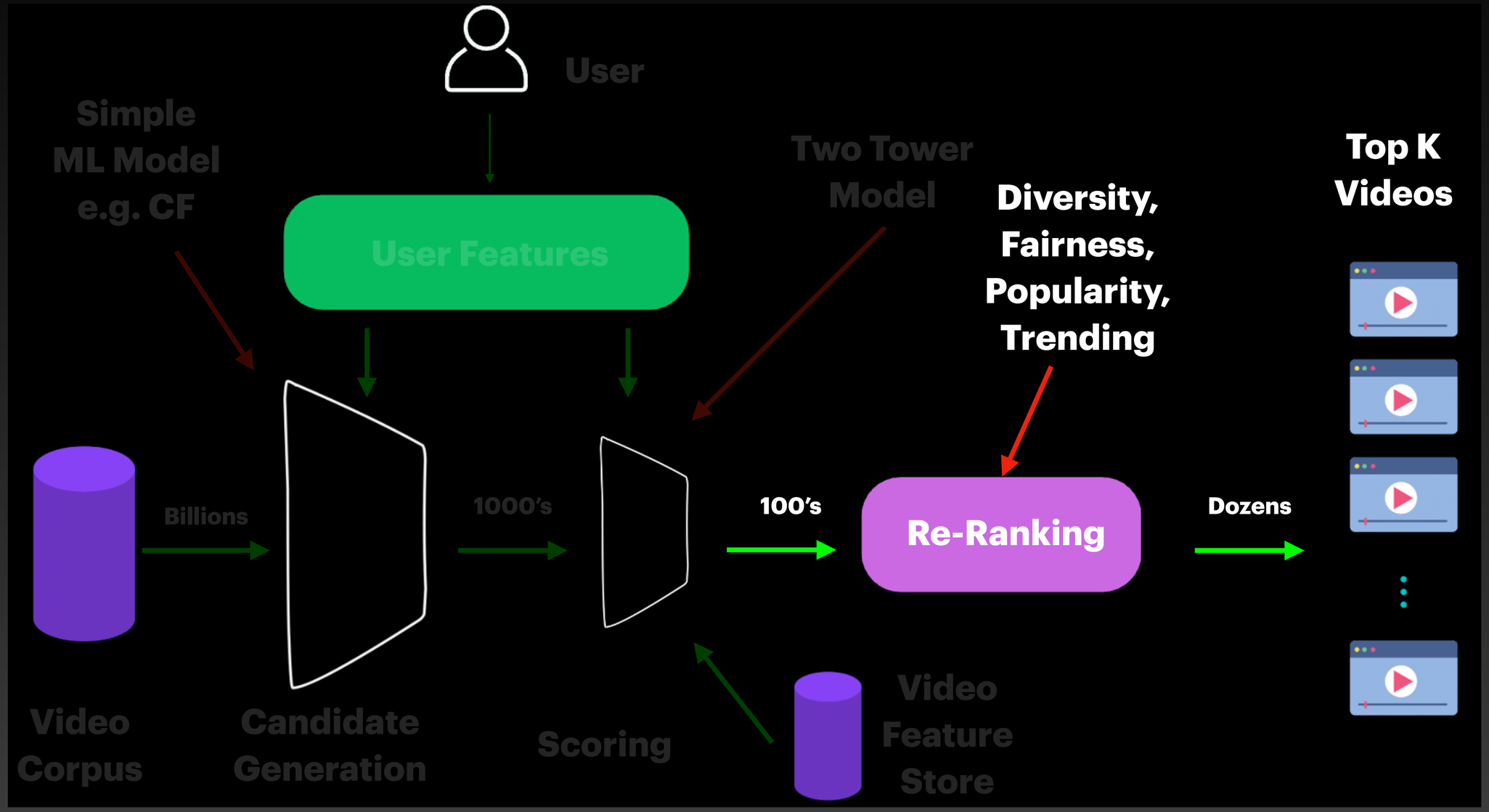
Scalable RecSys Design



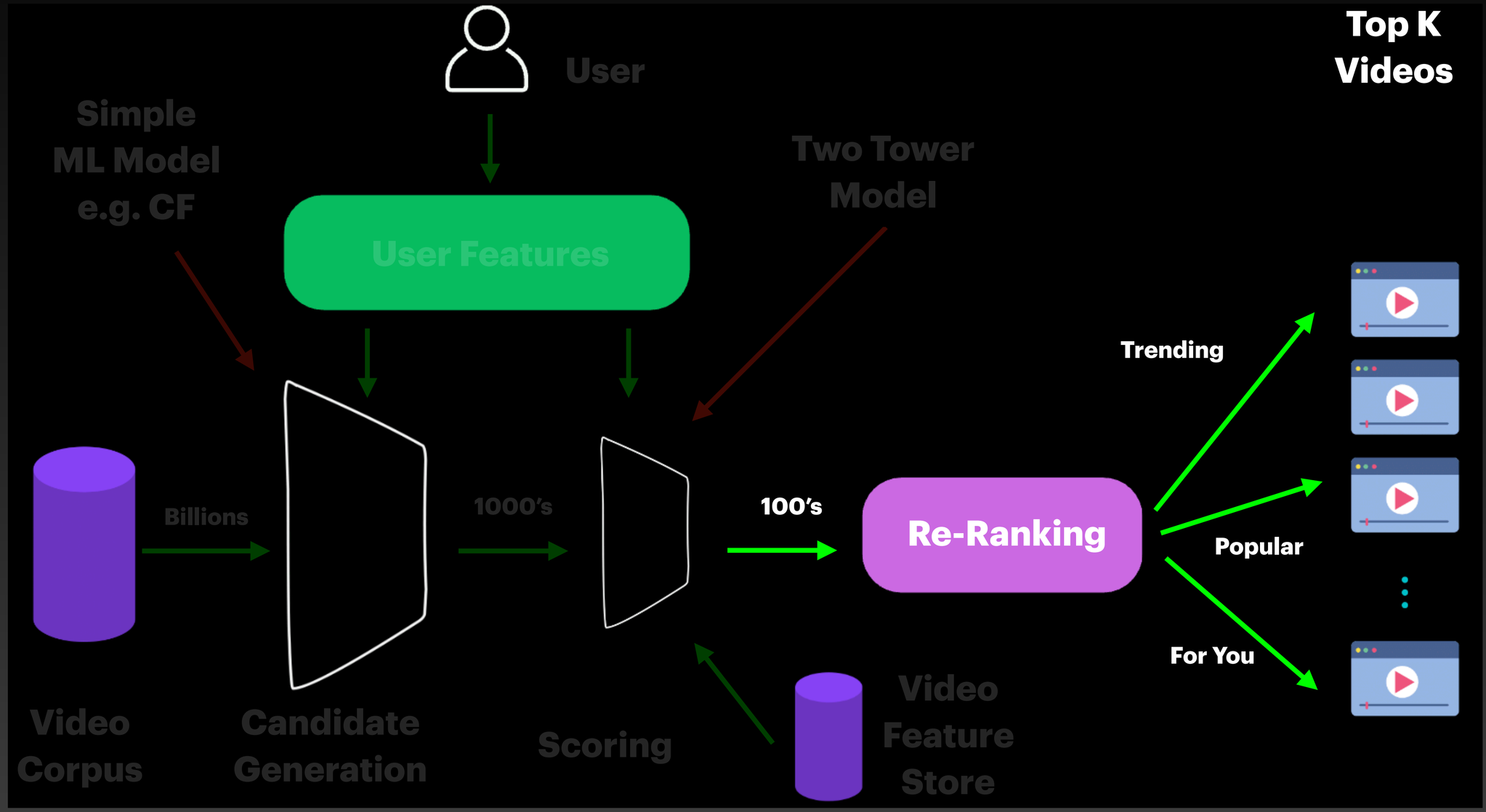
Scalable RecSys Design



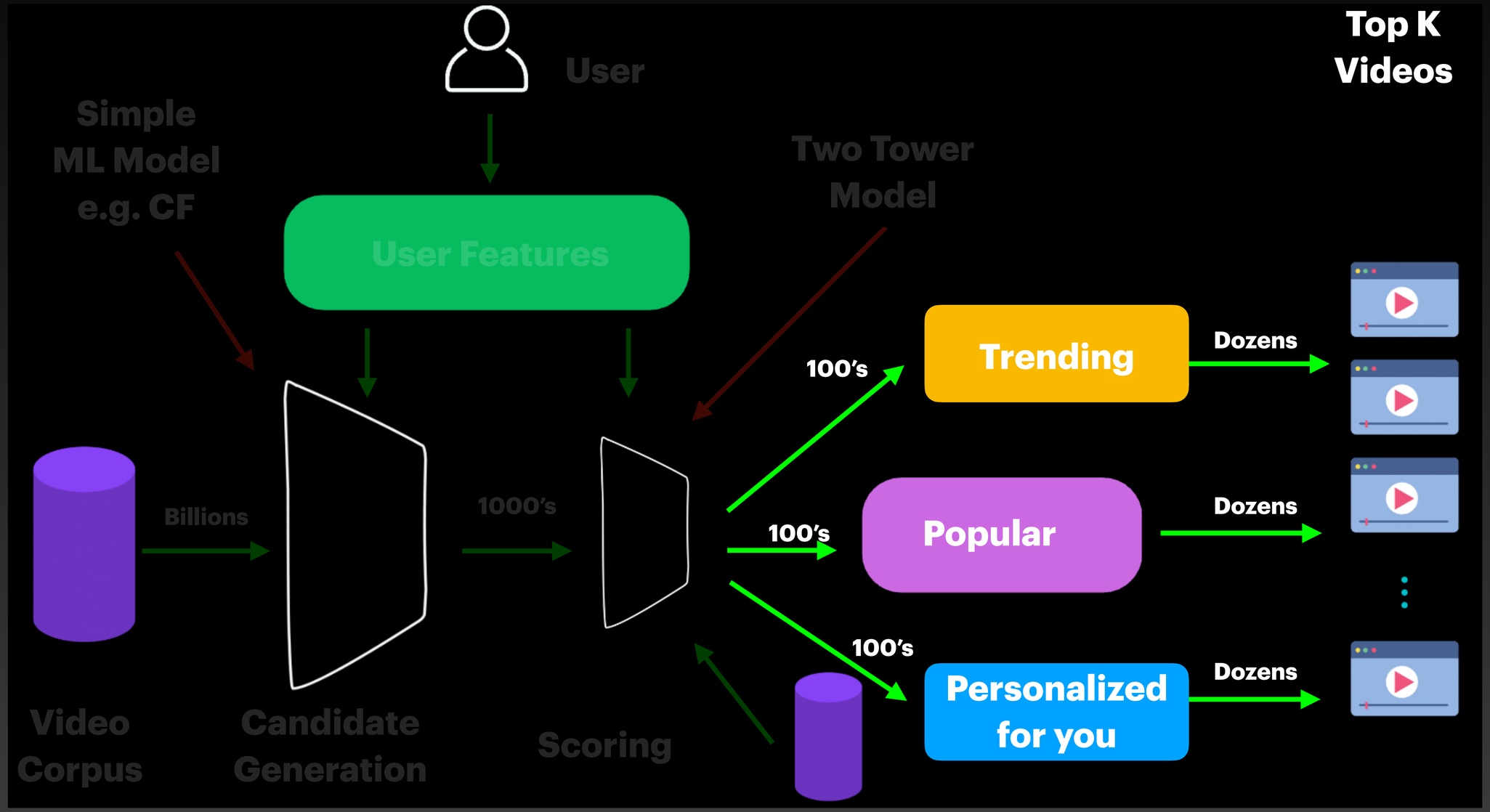
Scalable RecSys Design



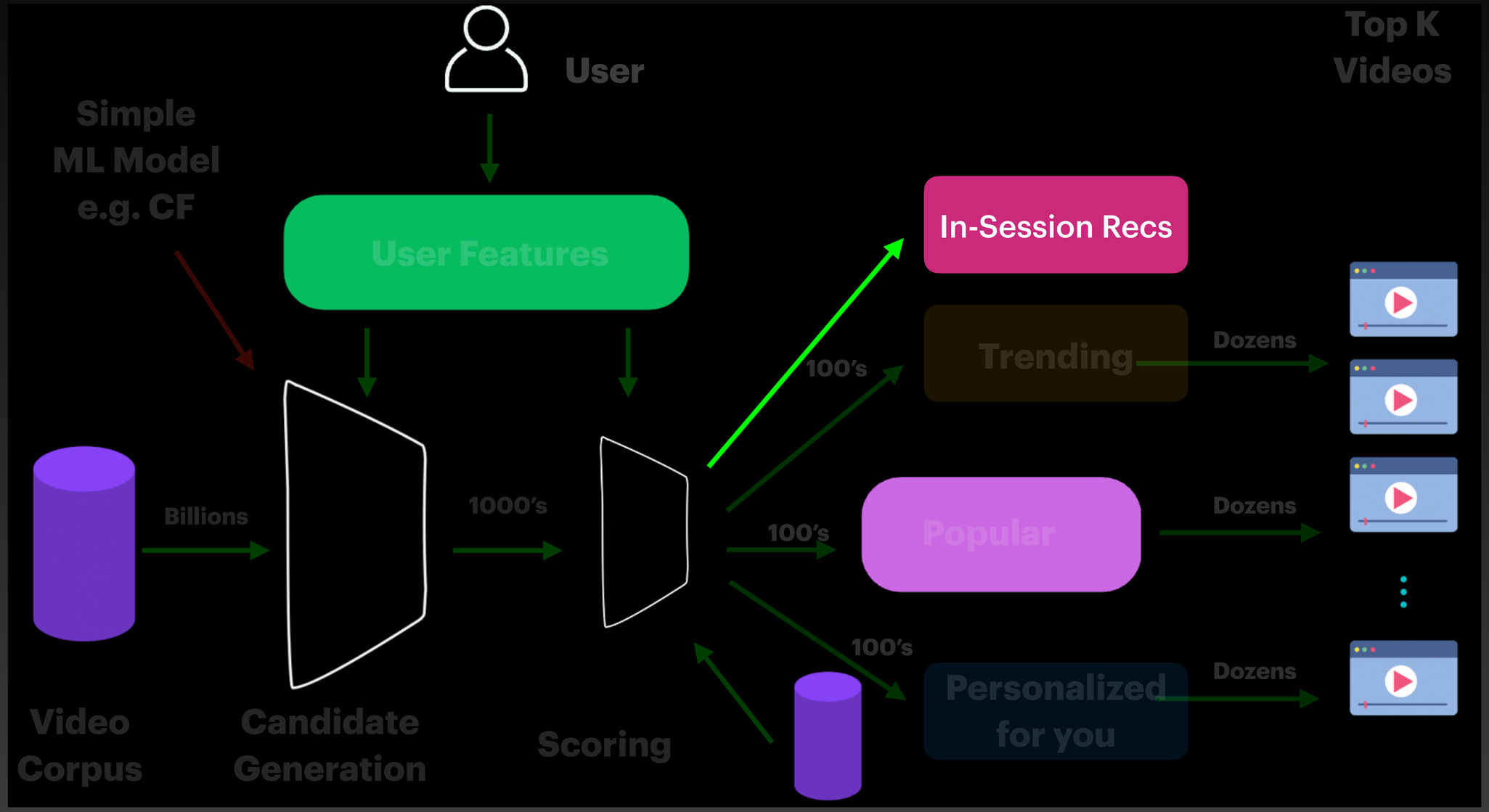
Scalable RecSys Design



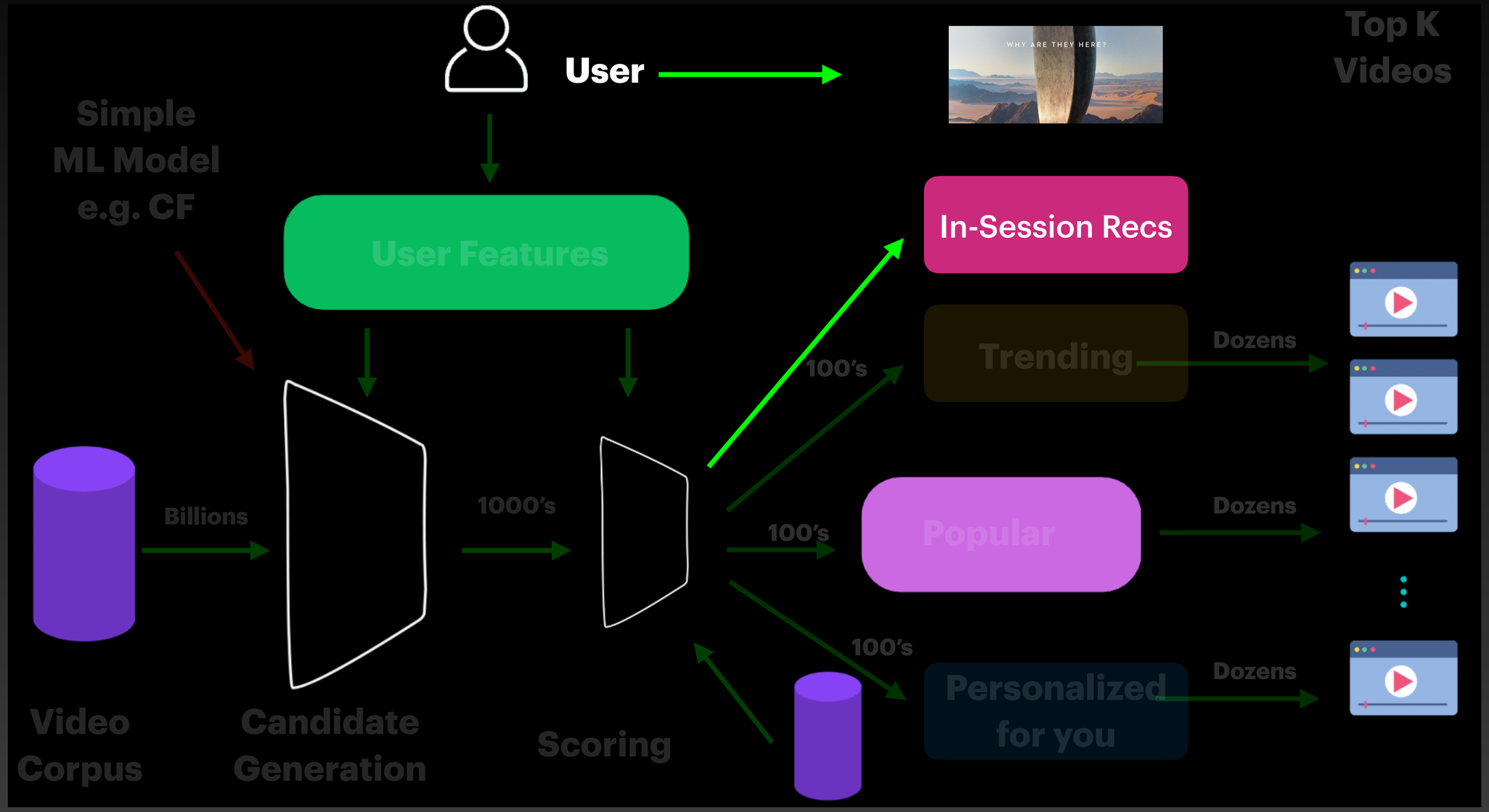
Multi-channel Scalable RecSys Design



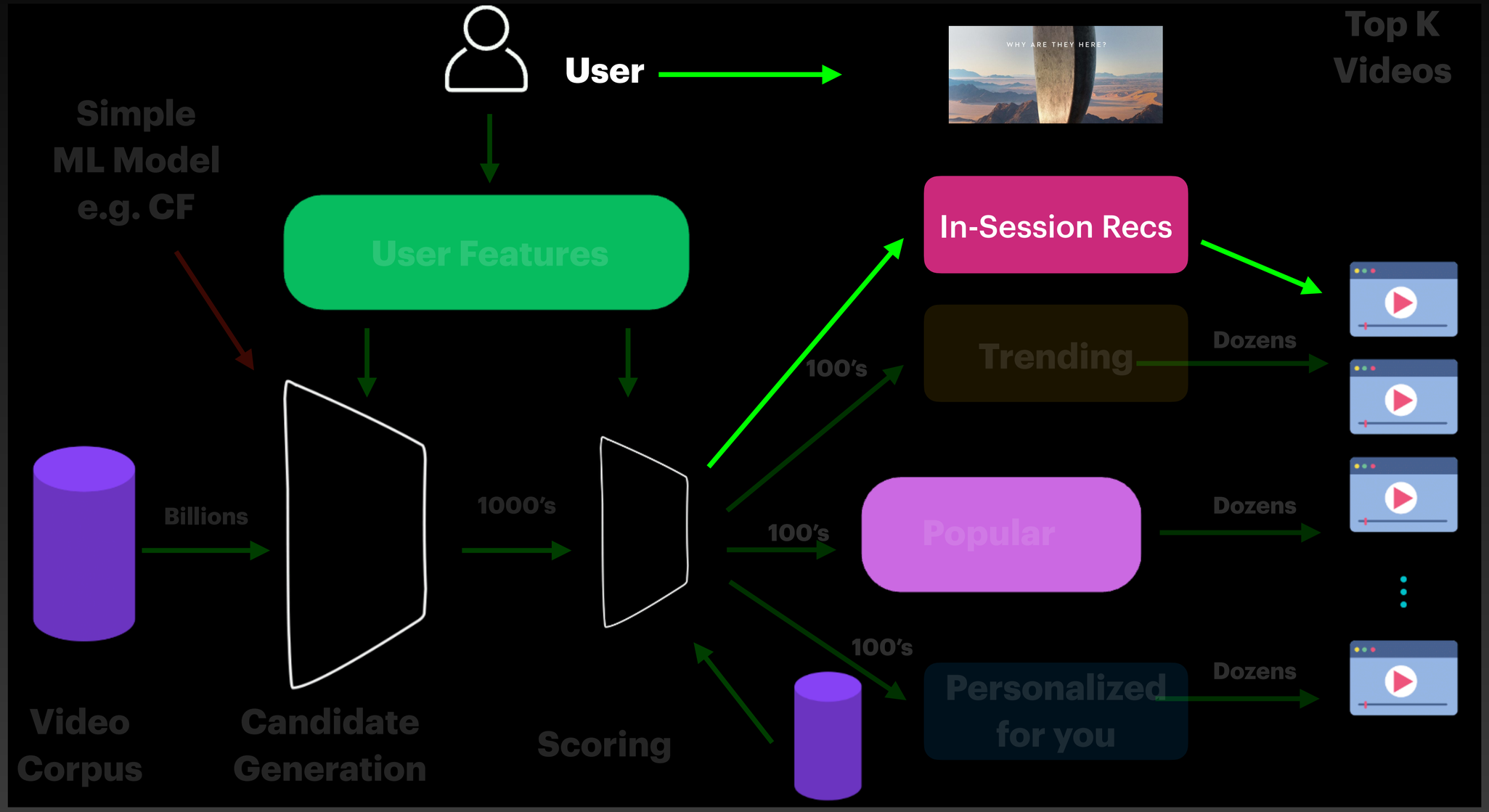
Multi-channel Scalable RecSys Design



Multi-channel Scalable RecSys Design



Multi-channel Scalable RecSys Design



Performance Metrics and Design

Modeling Metrics

Design Considerations

Business Metrics

Evaluating the Design

Offline Metrics

- Relevance Metrics
- Diversity Metrics
- Cold-Start Performance Metrics
- ...

Business Metrics

- Revenue
- #Products Viewed/Purchased
- View Time
- CTR

Evaluating the Design

Offline Metrics

- Relevance Metrics
- Diversity Metrics
- Cold-Start Performance Metrics
- ...

Business Metrics

- Revenue
- #Products Viewed/Purchased
- View Time
- CTR

Evaluating the Design

Offline Metrics

- Relevance Metrics
- Diversity Metrics
- Cold-Start Performance Metrics
- ...

Diversity Optimization ! = Increase in Revenue

Business Metrics

- Revenue
- #Products Viewed/Purchased
- View Time
- CTR

Evaluating the Design

Offline Metrics

- Relevance Metrics
- Diversity Metrics
- Cold-Start Performance Metrics
- ...

Business Metrics

- Revenue
- #Products Viewed/Purchased
- View Time
- CTR

Business Metrics and Launch

Business Metrics

- Revenue
- #Products Viewed/Purchased
- View Time
- CTR

Bridging offline-online Gap

- Held-out Offline Testing
- Internal Testing/Dog Fooding
- A/B Testing

Business Metrics and Launch

Business Metrics

- Revenue
- #Products Viewed/Purchased
- View Time
- CTR

Bridging offline-online Gap

- **Held-out Offline Testing**
- **Internal Testing/Dog Fooding**
- **Offline Evaluation on Production Data**
- **A/B Testing**

Business Metrics and Launch

Business Metrics

- Revenue
- #Products Viewed/Purchased
- View Time
- CTR

Bridging offline-online Gap

- Held-out Offline Testing
- Internal Testing/Dog Fooding
- Offline Evaluation on Production Data
- A/B Testing

Business Metrics and Launch

Business Metrics

- Revenue
- #Products Viewed/Purchased
- View Time
- CTR

Bridging offline-online Gap

- **Held-out Offline Testing**
- **Internal Testing/Dog Fooding**
- **Offline Evaluation on Production Data**
- **A/B Testing**