Machine Learning for Understanding User Behaviours

Semi-Supervised Learning Applied to Click Streams

Goals

- Motivation for semi-supervised learning and log analytics
- Overall Methodology
- Leveraging Hadoop / Java to create datasets
- Machine Learning Applied

Online Retail

- Upsales / Cross Sales
- Personalized Offers
- Newsletter Targeting
- Competition Optimized Price
- Optimized Display Campaigns
- Funnel Optimization
- ….

Publishing & Media

- Understand and monetize audience
- Acquire and sell customer data
- Personalize User Experience
- Optimize Conversion Funnels

Product Manager Questions

- Different Customer Behaviours
  - Subscribers
  - Newcomers
- How do behaviour evolve in time ?
- Do customer search & find relevant content ?
- …

What kind of session ?

Search for a specific Topic

Newcomer from Google News

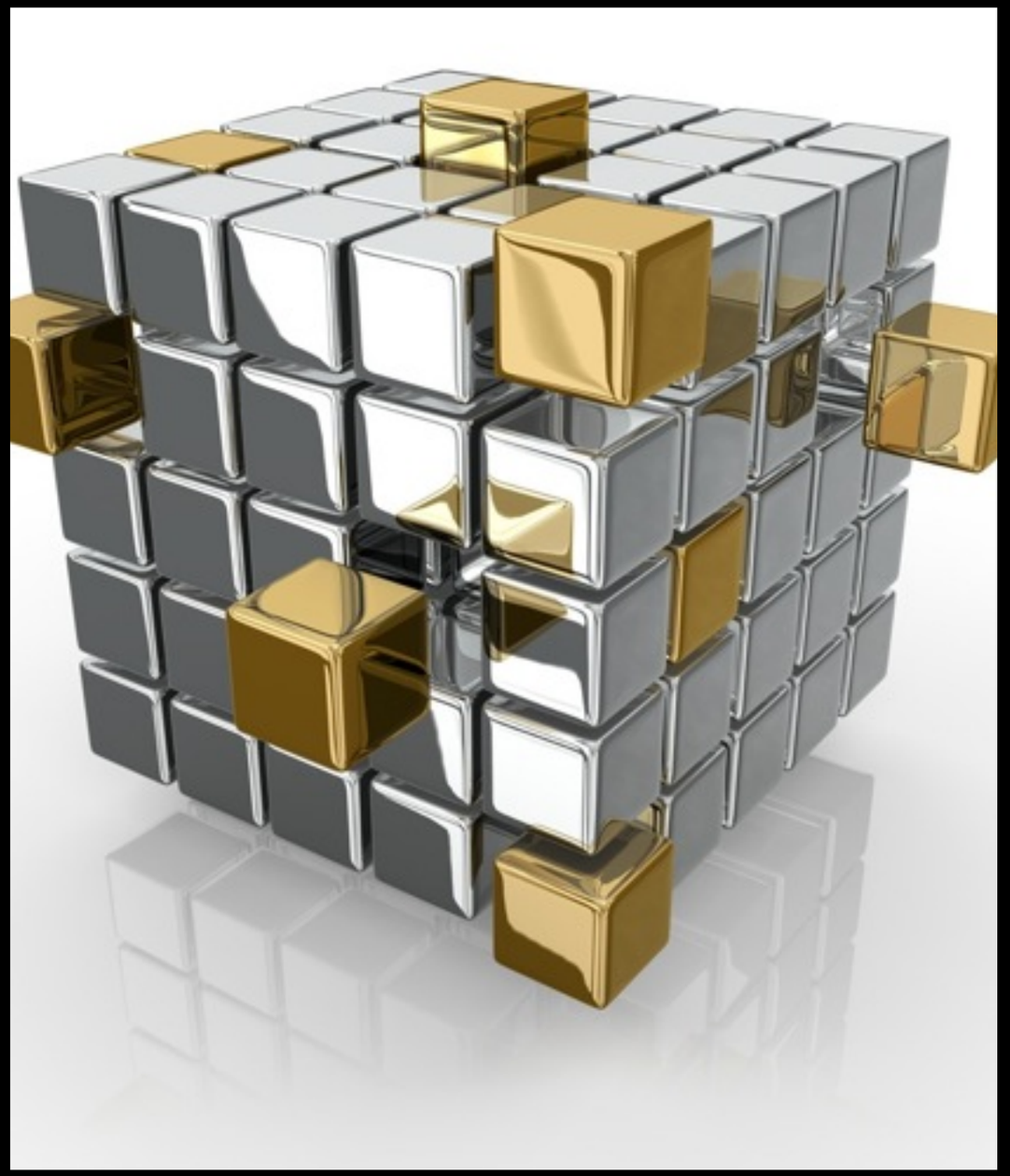
Foreigner Discovering the Web Site

Fan that loves to rate and comment

Home Page Wanderer

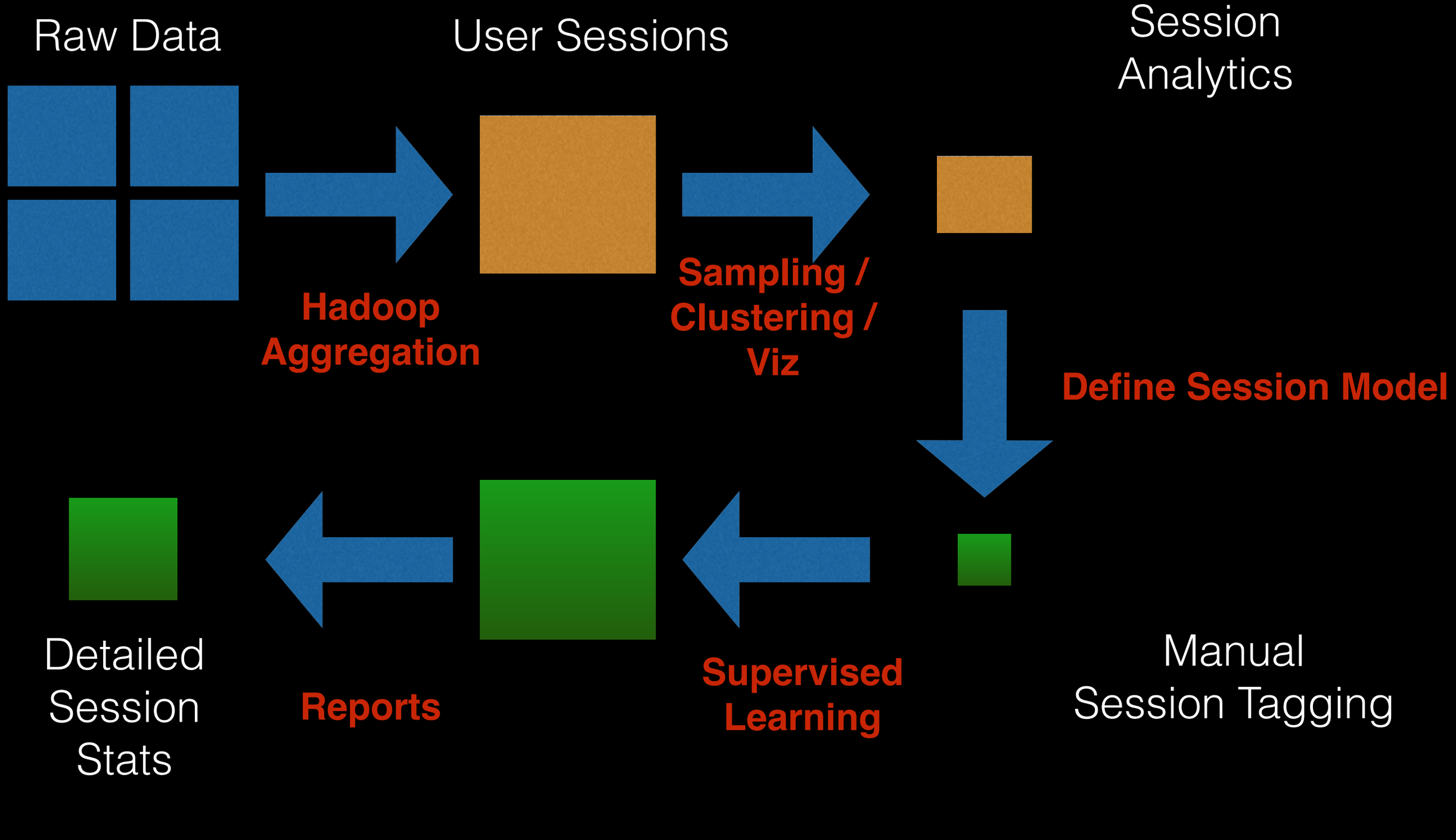
Product Manager Available Data

- Click Streams
- Topics / Content Referential
- Competitors / Outside Data
- Customer Feedbacks
- Server Logs

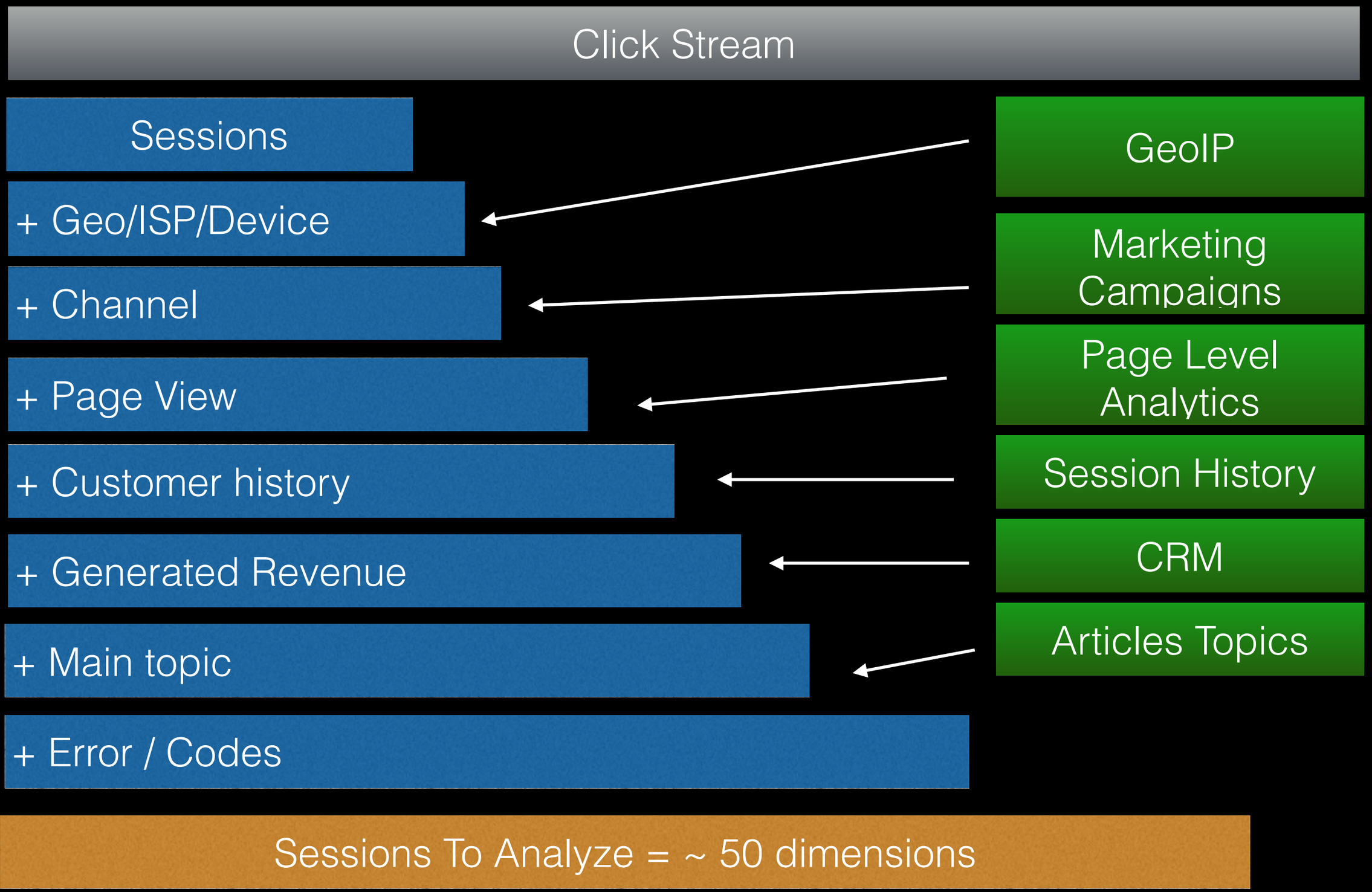


	Structured ?	Number of events / pages	Number of Users
Click Stream	+	100 M	10 M
Content Hierarchy	++	10 K	-
Web / RSS	+-	100 M	
User Ratings	+	~ 100 K	~ 10K
Server Logs	+-	100M	10M

Methodology Overview



Add Signals to the Sessions



ClickStream

Click Stream

UserID	TimeStamp	Page	Refererer	…..
94390940	2013/11/11-12:32:43			
43653729	2013/11/11-12:32:43			
30418239	2013/11/11-12:32:43			
94390940	2013/11/11-12:32:43			
99666658	2013/11/11-12:32:43			
43653729	2013/11/11-12:32:43			
94390940	2013/11/11-12:32:43			
94390940	2013/11/11-12:32:43			
99666658	2013/11/11-12:32:43			
94390940	2013/11/11-12:32:43			
94390940	2013/11/11-12:32:43			
94390940	2013/11/11-12:32:43			
43653729	2013/11/11-12:32:43			
94390940	2013/11/11-12:32:43			
94390940	2013/11/11-12:32:43			

Build Sessions

Click Stream

Sessions

UserID	TimeStamp	Page	Refererer	…..	SessionID
94390940	2013/11/11-12:32:43				94390940-1
94390940	2013/11/11-12:32:43				94390940-1
94390940	2013/11/11-12:32:43				94390940-1
94390940	2013/11/11-12:32:43				94390940-1
94390940	2013/11/11-12:32:43				94390940-1
94390940	2013/11/11-12:32:43				94390940-1
94390940	2013/11/11-12:32:43				94390940-1
94390940	2013/11/11-12:32:43				94390940-1
94390940	2013/11/11-12:32:43				94390940-1

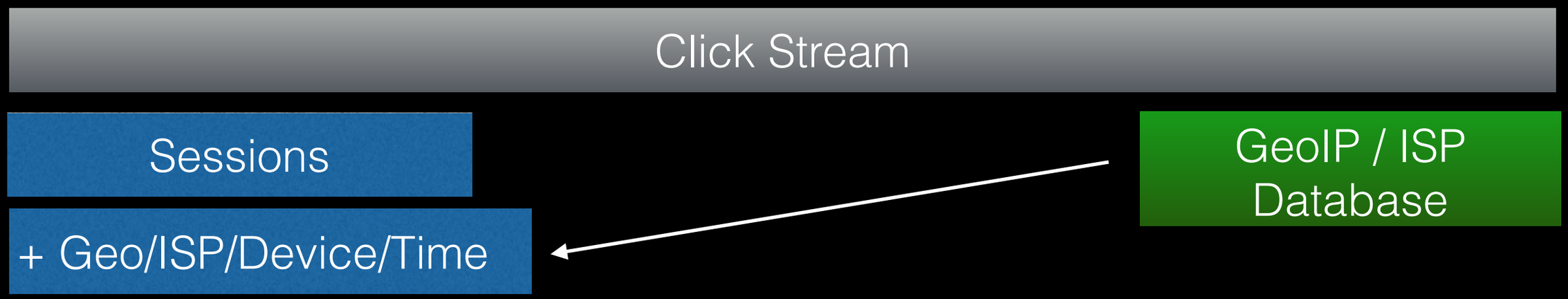
Add Signals to the Sessions

Click Stream

Sessions

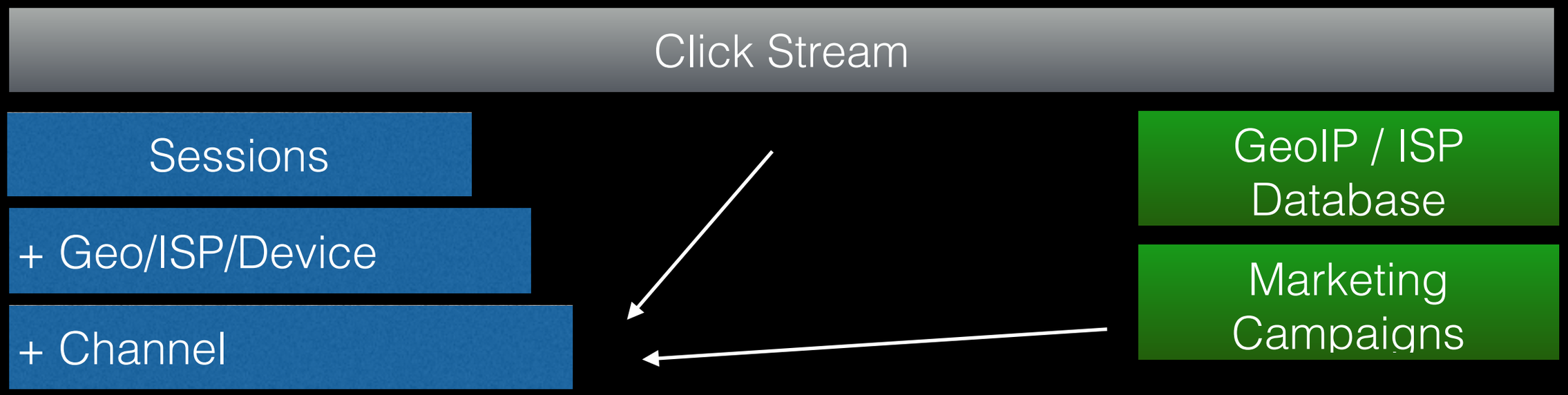
UserID	TimeStamp	Page	Referer	,,,,	SessionID
94390940	2013/11/11-12:32:43				1
94390940	2013/11/11-12:32:43				1
94390940	2013/11/11-12:32:43				1
94390940	2013/11/11-12:32:43				1
94390940	2013/11/11-12:32:43				1
94390940	2013/11/11-12:32:43				1
94390940	2013/11/11-12:32:43				1
94390940	2013/11/11-12:32:43				1
94390940	2013/11/11-12:32:43				1

Add Signals to the Sessions



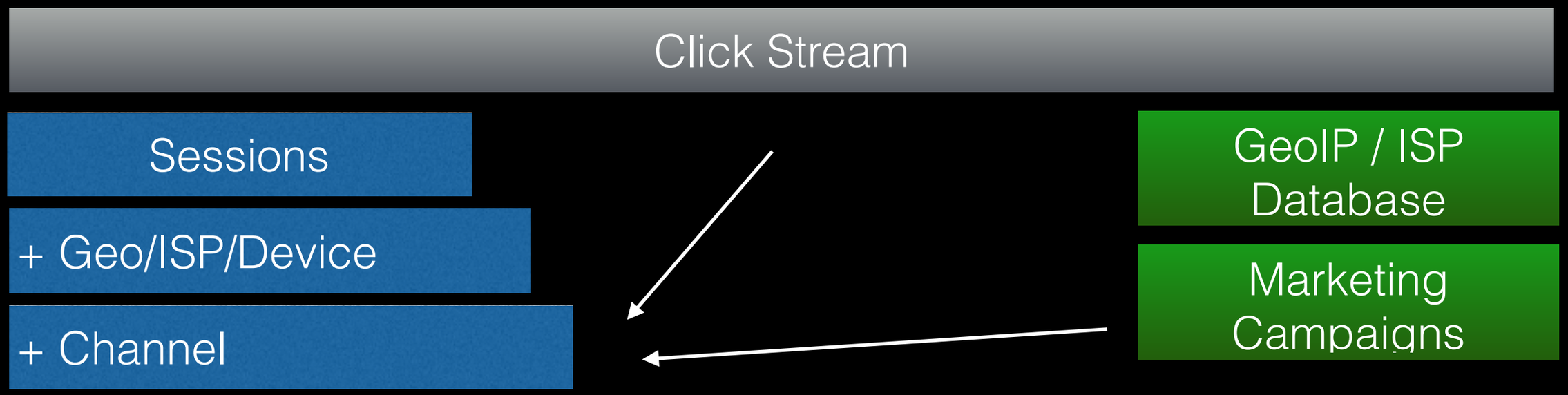
UserID	TimeStamp	Page	User Agent	IP	,,,,	SessionID	City	ISP	Device	OS	Local Hour
94390940	2013/11/11 -12:32:43		Mozilla 4.1 (…)	193.4.1.3		1	PARIS	FREE SAS	Mobile	iOS	14

Channel Signal



UserID	TimeStamp	Page	User Agent	IP	,,,,	SessionID	City	ISP	Device	OS	Marketing Source	Camp.
94390940	2013/11/11-12:32:43		Mozilla 4.1 (…)	193.4.1.3		1	PARIS	FREE SAS	Mobile	iOS	AdWords	ad-213
											E-Mailing	e-2013
											Retarg.	crteo-2
											DIsplay	..
											None	..
											SEO	..

Channel Signal



UserID	TimeStamp	Page	User Agent	IP	,,,,	SessionID	City	ISP	Device	OS	Marketing Source	Camp.
94390940	2013/11/11-12:32:43		Mozilla 4.1 (…)	193.4.1.3		1	PARIS	FREE SAS	Mobile	iOS	AdWords	ad-213
											E-Mailing	e-2013
											Retarg.	crteo-2
											DIsplay	..
											None	..
											SEO	..

Channel Signal

Click Stream

Sessions

+ Geo/ISP/Device

+ Channel

+ Customer history

GeoIP / ISP Database

Marketing Campaigns

UserID	TimeStamp	Page	User Agent	IP	Marketing Source	Camp.	….	First Contact N days ago	Last Contact N days ago	N Contacts Last 30 Days
94390940	2013/11/11-12:32:43		Mozilla 4.1 (…)	193.4.1.3	AdWords	ad-213		231	2	19
					E-Mailing	e-2013				
					Retarg.	crteo-2				
					DIsplay	..				
					None	..				
					SEO	..				

Channel Signal

Click Stream

Sessions

+ Geo/ISP/Device

+ Channel

+ Customer history

+ PageView

+ ….

Sessions To Analyze = ~ 50 dimensions

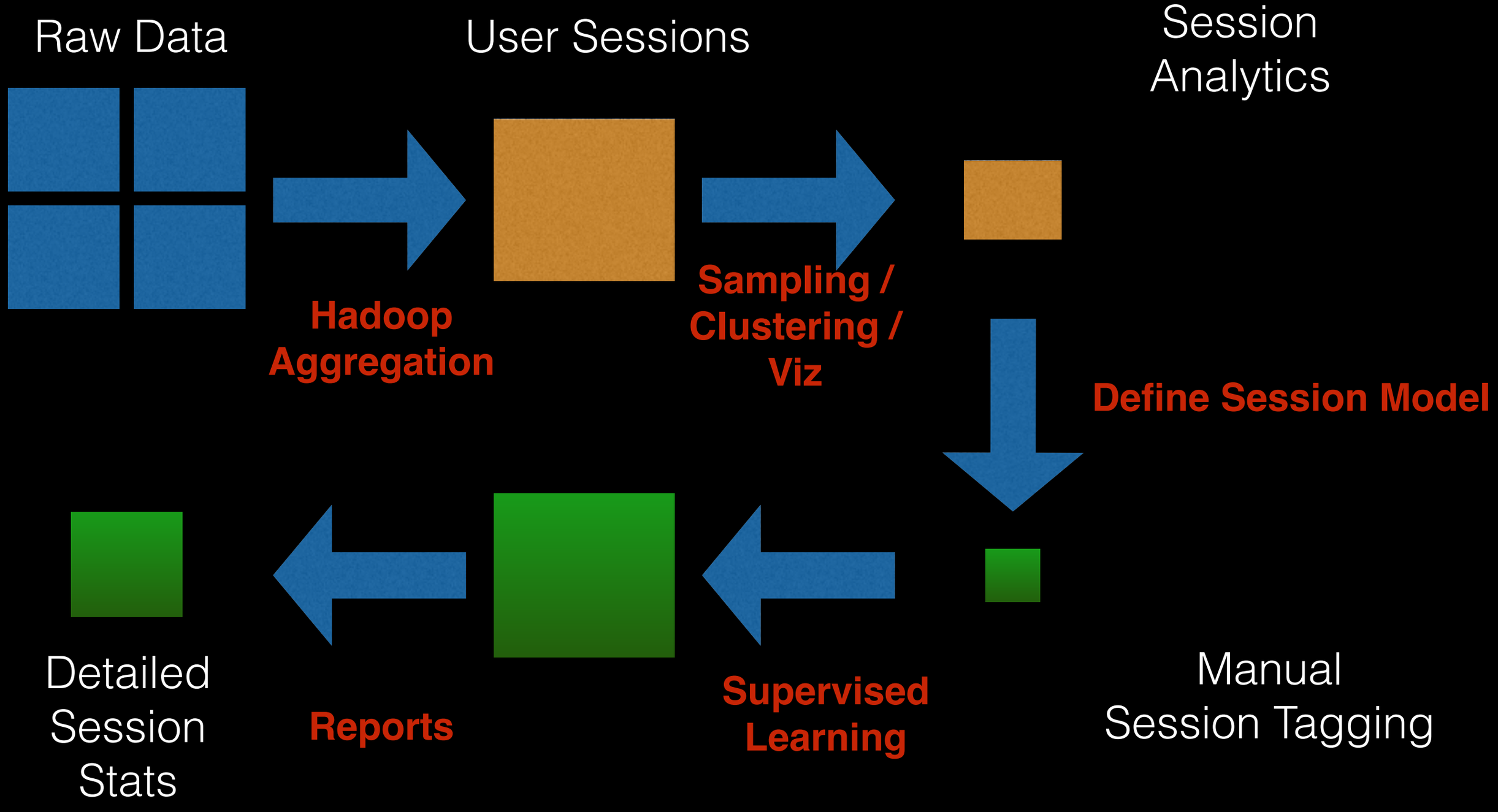
UserID	TimeStamp	Page	User Agent	IP	N Contacts Last 30 Days	….	N PageViews	N Page < 5 sec	…	N Search Page …
94390940	2013/11/11-1		Mozilla 4.1	193.4.1.3	19					

How to do these aggregations ?

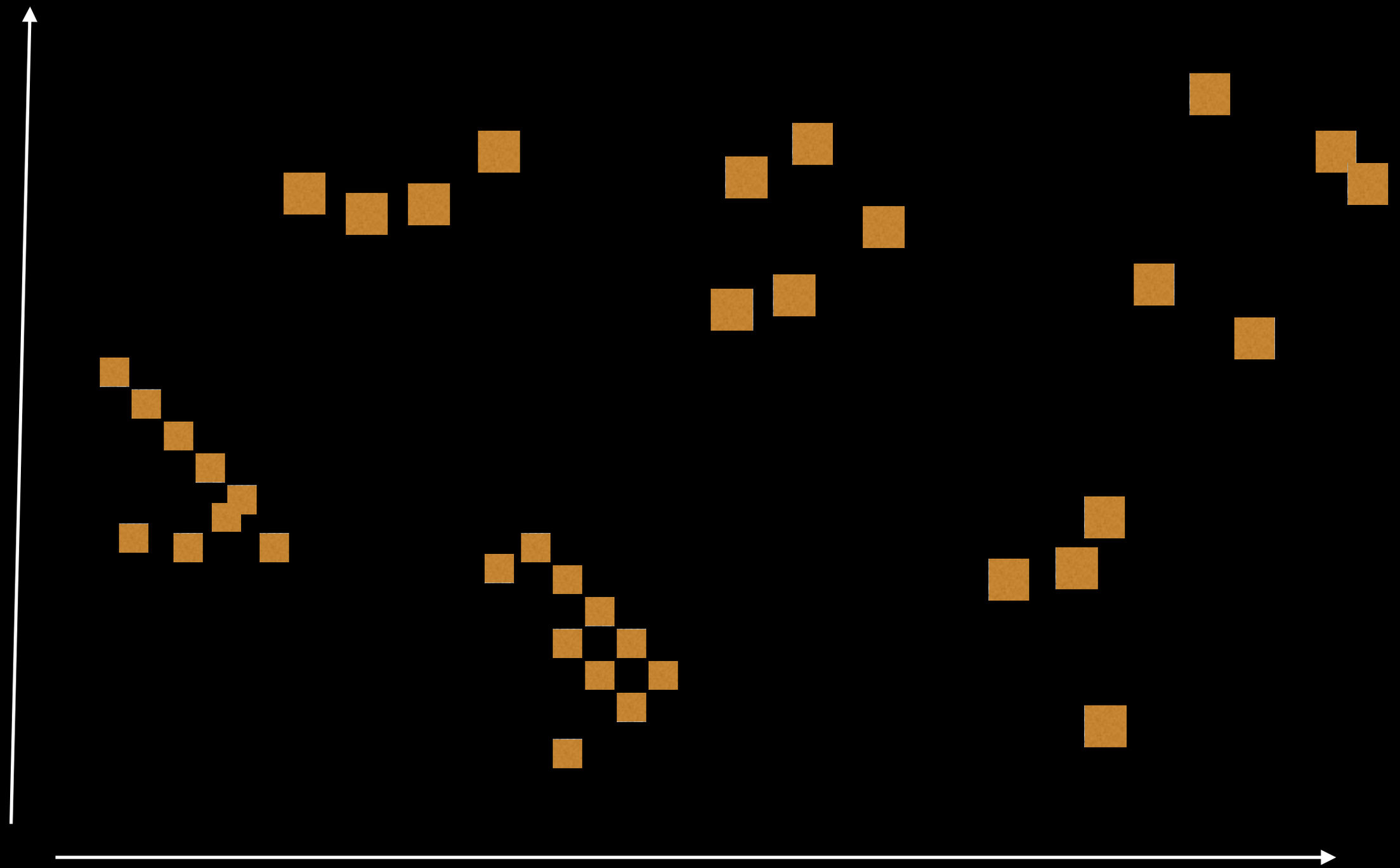
- Hive Job
- Pig Job
- Spark
- Hadoop Streaming + Python
- Scalding
- Cascading + Java

JUG READY !

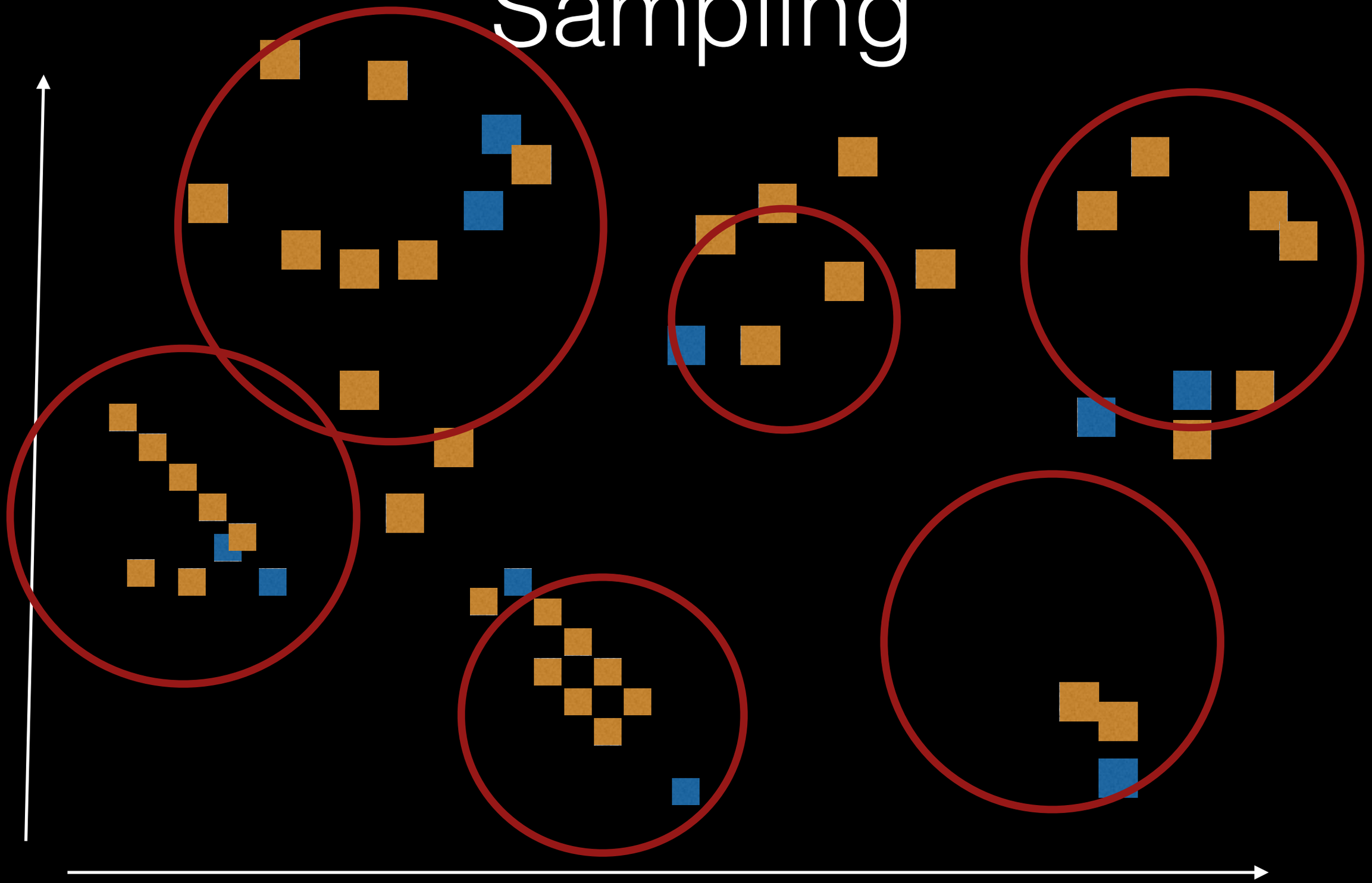
Methodology Overview



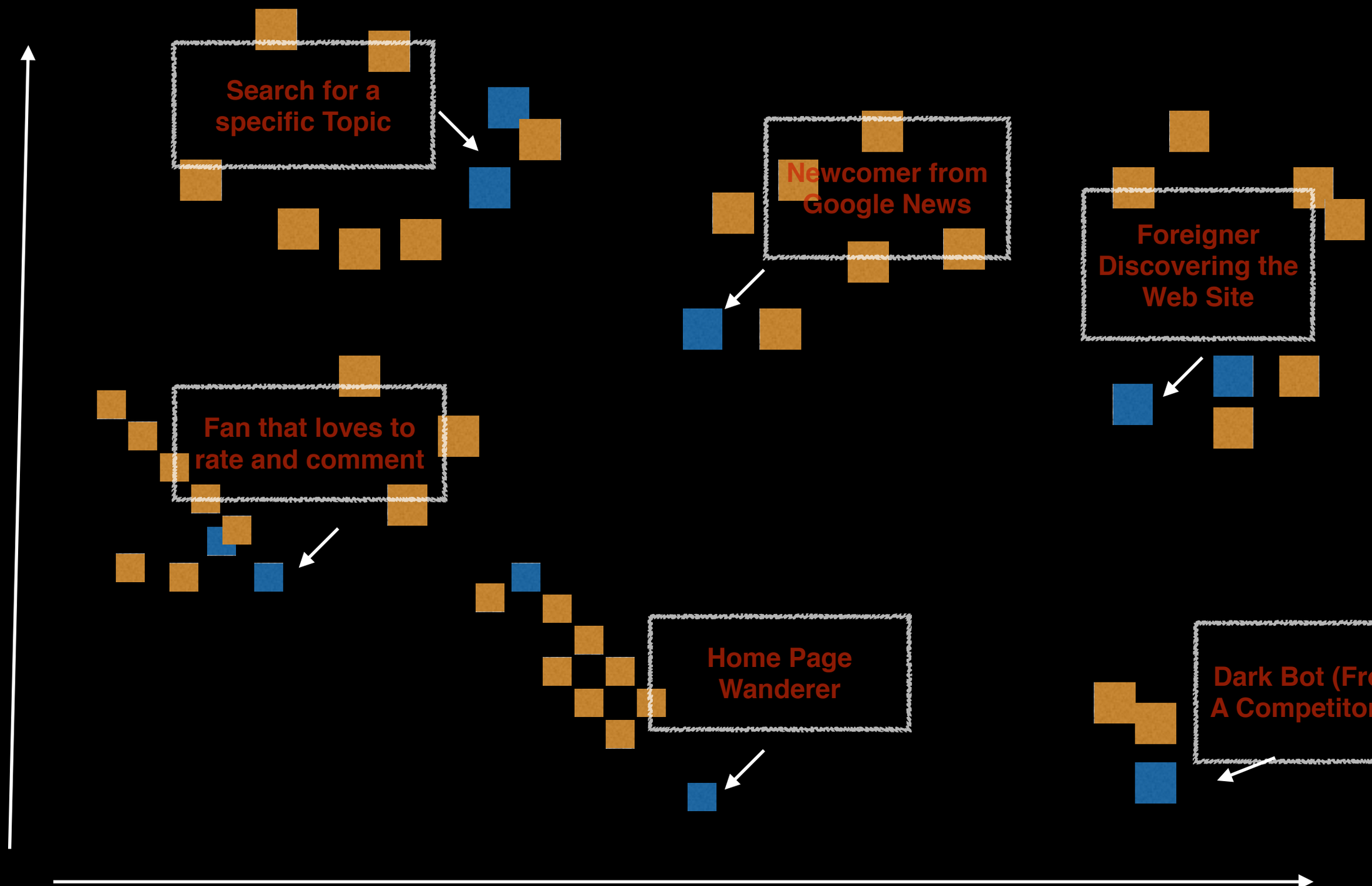
Sessions Data



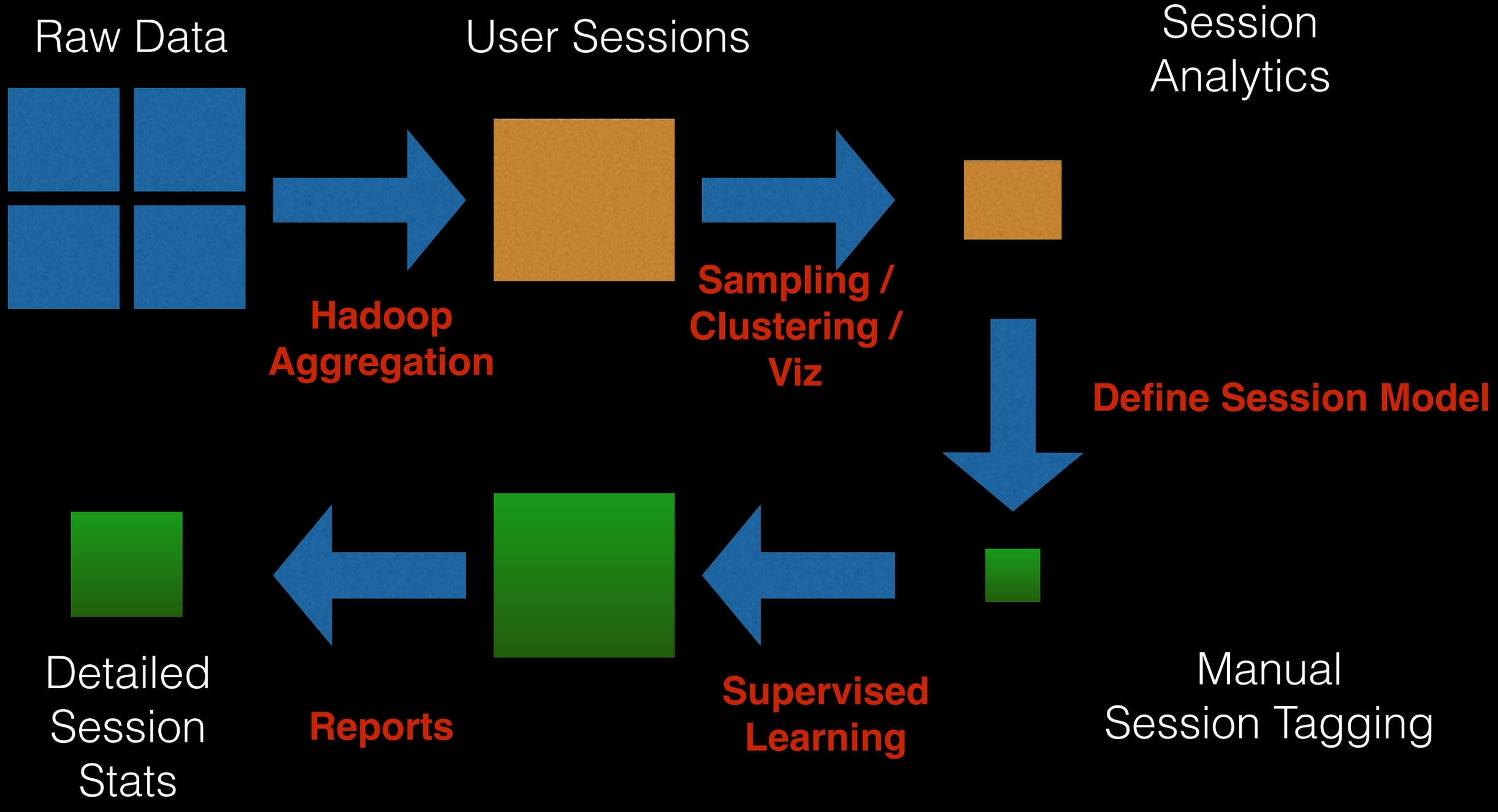
Clustering & Cluster Sampling



Tag ~ 1000 Sessions



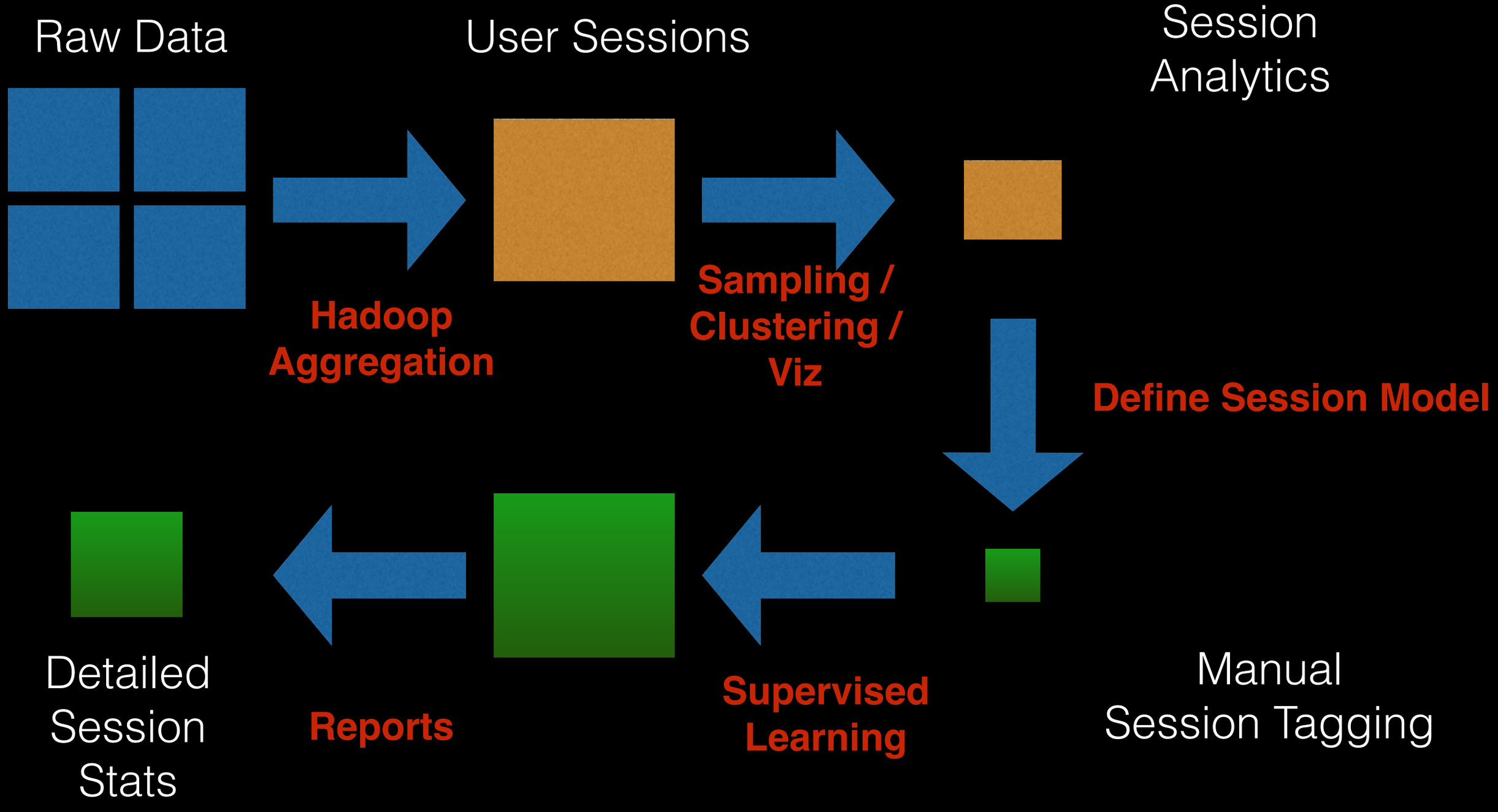
Methodology Overview



Supervised Learning

- Mahout (Logistic Regression?)
- Python Scikit
- MLI + Spark (Logistic Regression?)
- R Hadoop

Methodology Overview



Reporting Part

- Hive + Desktop Reporting Tool
- Python (Pandas)
- ElasticSearch (Optionally with Kibana)

Sessions Statistics

Search for a specific Topic

938k sessions
0.3€ per session
0.23€ acquisition costs

Newcomer from Google News

738k sessions
0.83€ per session
0.73€ acquisition costs

Fan that loves to rate and comment

13k sessions
1.3€ per session
0.23€ acquisition costs

Foreigner Discovering the Web Site

68k sessions
0.3€ per session
1.23€ acquisition costs

Home Page Wanderer

938k sessions
0.3€ per session
0.23€ acquisition costs

1k sessions
0€ per session
0€ acquisition costs

Dark Bot (From A Competitor ?)

Next: Follow Transitions

Search for a specific Topic

938k sessions
0.3€ per session
0.23€ acquisition costs

Newcomer from Google News

738k sessions
0.83€ per session
0.73€ acquisition costs

Fan that loves to rate and comment

13k sessions
1.3€ per session
0.23€ acquisition costs

Foreigner Discovering the Web Site

68k sessions
0.3€ per session
1.23€ acquisition costs

Home Page Wanderer

938k sessions
0.3€ per session
0.23€ acquisition costs

1k sessions
0€ per session
0€ acquisition costs

Dark Bot (From A Competitor ?)

Thank you !

- Tweet
  #ParisJUG @fdouetteau Semi-supervised learning for user sessions

- We're hiring
  jobs@dataiku.com