

# The Untold Story of the Clones: Content-agnostic Factors that Impact YouTube Video Popularity

Youmna Borghol  
Sebastien Ardon  
**Niklas Carlsson**  
Derek Eager  
Anirban Mahanti

UNSW & NICTA  
NICTA  
**Linköping University**  
University of Saskatchewan  
NICTA



**UNSW**  
THE UNIVERSITY OF NEW SOUTH WALES



**NICTA**



August 15, 2012

# Motivation



- Video dissemination (e.g., YouTube) can have widespread impacts on opinions, thoughts, and cultures

Just det att kiale  
får marken själ  
dugarna med sig  
den paltbröden  
Det finns en lada  
Och det finns en  
som blanda riktat  
u mäsom

ECKEN  
ENINU  
OZUDET

MÖBELDESIGN

words  
SQUIS AKAT DEEAT!!  
Por jaginte p

doctor's daughter

DP x KN x  
AV x VOUS

männi sko ej  
fiter vad hon  
en vad hon

- Rell teori och va  
- AHA crätt  
- Offentlig rätt  
- Ahalcrätt, skade

OnLiU  
www.liu.se

92: g  
Br  
LINKÖPINGS UNIVERSITET

The logo of Linköping University, featuring a stylized white flower or star shape on a dark background.

# Motivation



- Not all videos will reach the same popularity and have the same impact

Just det att hiale  
får marcken sjä  
dugarna med sif  
den paltbröden  
Det finns en lada  
Och det finns m  
som blanda riefat  
u masonu

ECKEN  
ENINU  
OBJET

MÖBELDESIGN

words  
SQUIS AKAT DÉBAT!!  
Por jaginte p

doctor's daughter

Åveer -vours

minni sko ej  
fter vad hon  
er vad hon

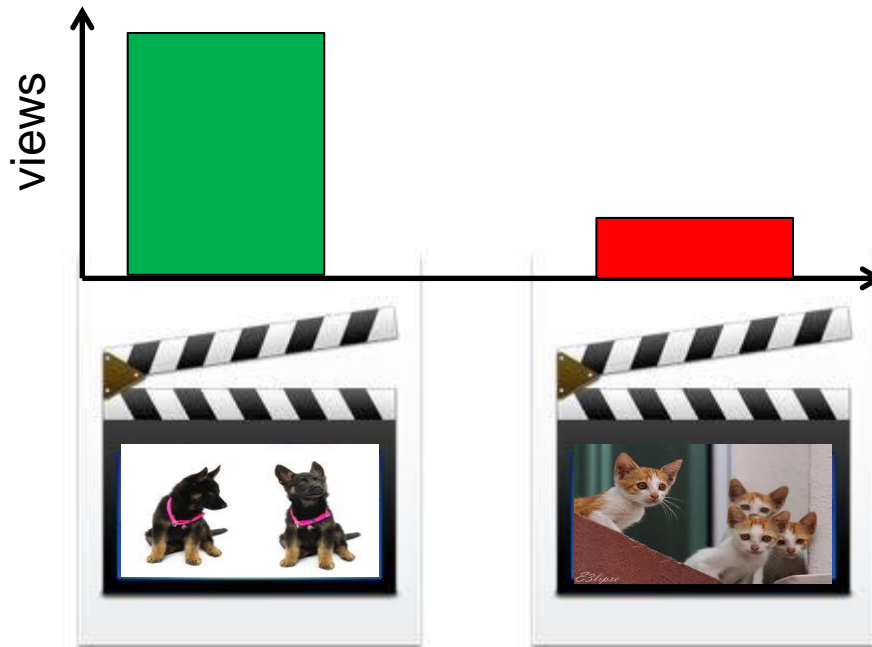
- Rätt teori och r  
- Åttal rätt  
- Offentlig rätt  
- Åttal rätt, skade

OnLiU  
www.liu.se

LINKÖPINGS UNIVERSITET



# Motivation



- Not all videos will reach the same popularity and have the same impact

Just det att hiale  
får marken själ  
dugarna med sig  
den paltbröden  
det finns en lada  
Och det finns m  
som blanda riefat  
w masonu

ECKEN  
ENINU  
O3Ubet

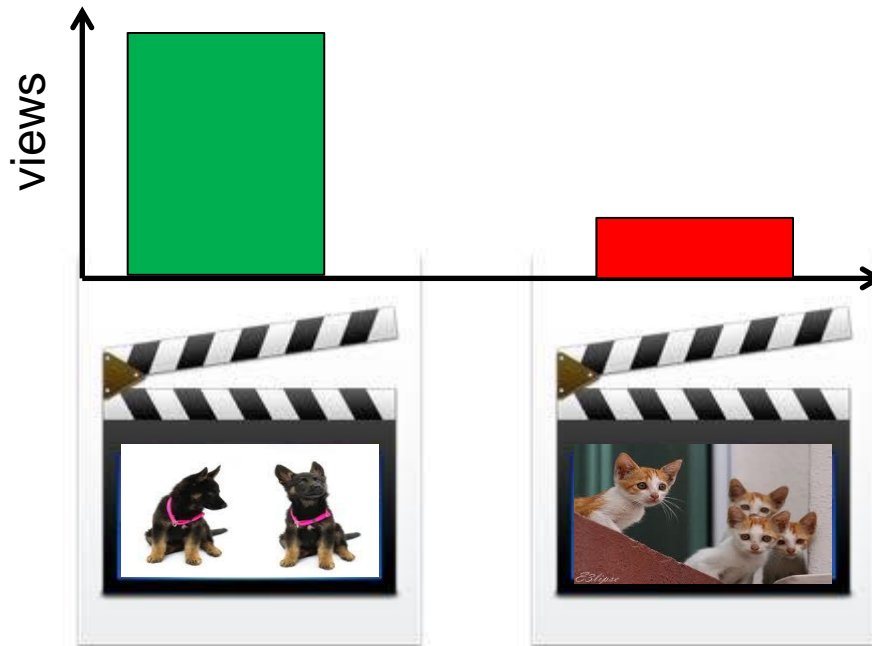
MÖBELDESIGN  
wards  
SQUIS AKAT DEEAT!!  
Por jaginte p  
mofford's daughter  
Åveez-vuus  
minni sko ej  
fter vad hon  
er vad hon  
- Rell teori och ra  
- AHAICRÄTT  
- Offentlig rätt  
- AHAICRÄTT, skade

OnLiU  
www.liu.se

LINKÖPINGS UNIVERSITET



# Motivation



- Not all videos will reach the same popularity and have the same impact
- Some popularity differences due to content differences

Just det att hiale  
får marken själ  
dugarna med sig  
den paltbröden  
Det finns en lada  
Och det finns en  
som blanda riktat  
w masonu

ECKEN  
ENINU  
OZUBET

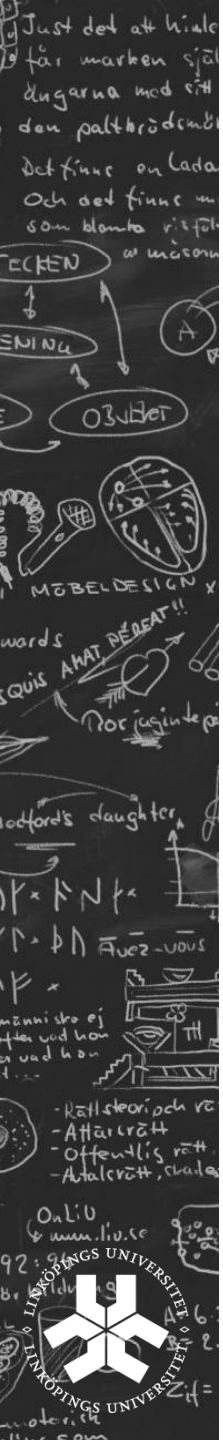
MÖBELDESIGN  
squis AKAT DÉBATE!!  
Por jaginte p  
mofford's daughter  
Åvez-vous  
- Rätt teori och r  
- Åttalcrätt  
- Offentlig rätt  
- Åttalcrätt, skade

On'li  
www.liu.se  
92: g  
LINKÖPINGS UNIVERSITET



# Motivation

- Popularity differences arise not only because of differences in video content, but also because of other “content-agnostic” factors
  - The latter factors are of considerable interest but it has been difficult to accurately study them



# Motivation

- Popularity differences arise not only because of differences in video content, but also because of other “content-agnostic” factors
  - The latter factors are of considerable interest but it has been difficult to accurately study them

In general, existing works **do not** take content differences into account .. .(e.g., large number of rich-gets-richer studies)

# Motivation



- Popularity differences arise not only because of differences in video content, but also because of other “content-agnostic” factors
  - The latter factors are of considerable interest but it has been difficult to accurately study them

Just det att kiale  
får marken själ  
dugarna med sig  
den paltbrödem  
Det finns en lada  
Och det finns m  
som blanda riefat  
w masonu

ECKEN  
ENINU  
O3Ubet

MÖBELDESIGN

words  
SQUIS AHAT DÉBATE!!  
Por jicinte p

godford's daughter

Y x N F x  
Y x N Avez-vous  
Y x

minni sko ej  
fter vad hon  
er vad hon

- Rätt teori och r  
- Attärrätt  
- Offentlig rätt  
- Atalcrätt, skade

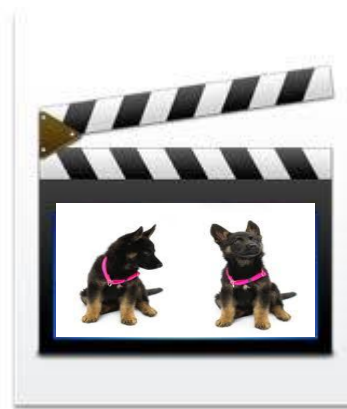
OnLiU  
www.liu.se

92: g  
LINKÖPINGS UNIVERSITET





# Motivation



For example, videos uploaded by users with large social networks may tend to be more popular because they tend to have more interesting content, not because social network size has a substantial direct impact on popularity

Just det att kiale  
får marken själ  
dugarna med sig  
den paltbrödem  
det finns en lada  
och det finns m  
som blanda riefat  
w masonu

ECKEN  
ENINU  
OBJET

MÖBELDESIGN

ward's  
QUIS AKAT DÉBAT!!  
Por jag inte p

odford's daughter

Y \* K \* N \* K \*  
Y \* P \* A \* V \* U \* V \*  
Y \* P \* X

männi ska ej  
fter vad hon  
en vad hon

- Rätt teori och r  
- Attalcrätt  
- Offentlig rätt  
- Attalcrätt, skade

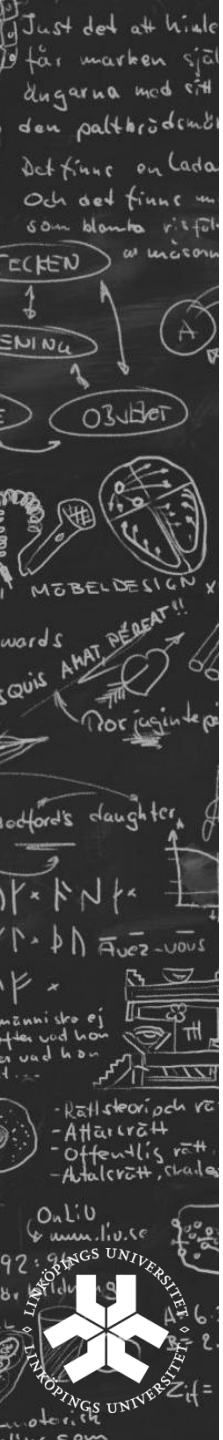
OnLiU  
www.liu.se

LINKÖPINGS UNIVERSITET



# Methodology

- Develop and apply a methodology that is able to accurately assess, both qualitatively and quantitatively, the impacts of various content-agnostic factors on video popularity



# Methodology

- Develop and apply a methodology that is able to accurately assess, both qualitatively and quantitatively, the impacts of various content-agnostic factors on video popularity



# Methodology

- Clones

- Videos that have “identical” content (e.g., same audio and video track)



Just det att kiale  
får märken själ  
dugarna med sig  
den paltbröden  
det finns en lada  
Och det finns en  
som blanda riktat  
u mäsom

EKEN  
ENINU  
OBJET

MÖBELDESIGN

words  
SQUIS AKAT DÉBÉAT!!  
Por jaginte p

godford's daughter

Åvee-uovs

männi sko ej  
fter vad hon  
er vad hon

- Rätt teori och r  
- Affär rätt  
- Offentlig rätt  
- Aftal rätt, skade

OnLiU  
www.liu.se

92: g  
Br. ldu

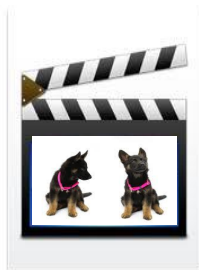
LINKÖPINGS UNIVERSITET



# Methodology

## ■ Clones

- Videos that have “identical” content (e.g., same audio and video track)



Clone 1.a



# Methodology

## ■ Clones

- Videos that have “identical” content (e.g., same audio and video track)



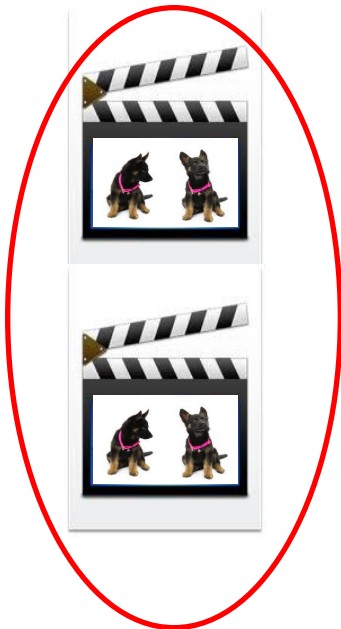
Clone 1.a

Clone 1.b



# Methodology

- Clones
  - Videos that have “identical” content
- Clone set
  - Set of videos that have “identical” content

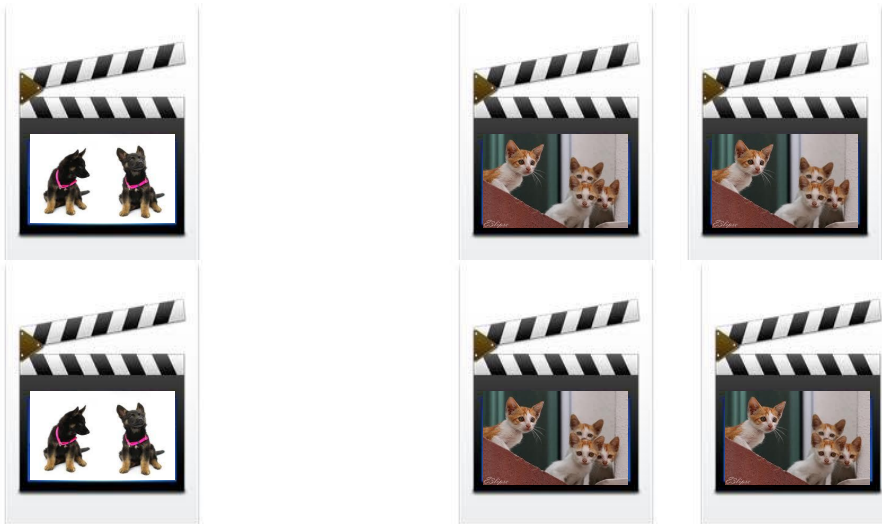


**Clone set 1**



# Methodology

- Clones
  - Videos that have “identical” content
- Clone set
  - Set of videos that have “identical” content





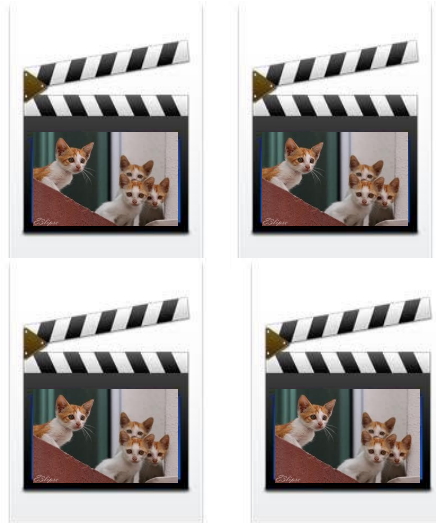
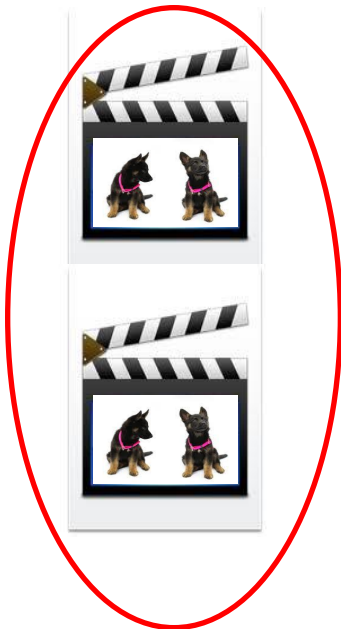
# Methodology

- Clones

- Videos that have “identical” content

- Clone set

- Set of videos that have “identical” content



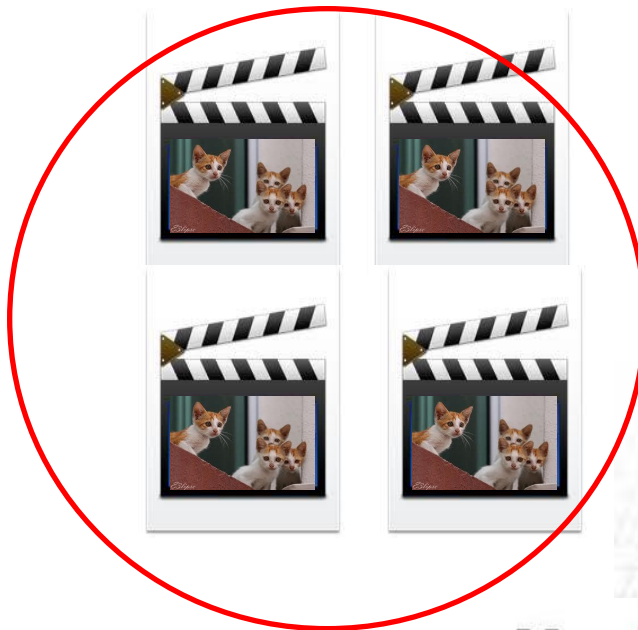
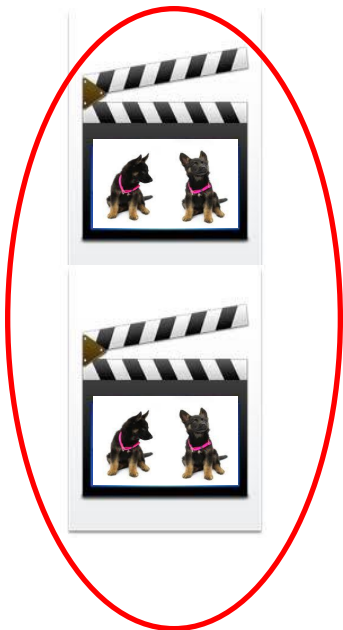
# Methodology

- Clones

- Videos that have “identical” content

- Clone set

- Set of videos that have “identical” content



# Methodology

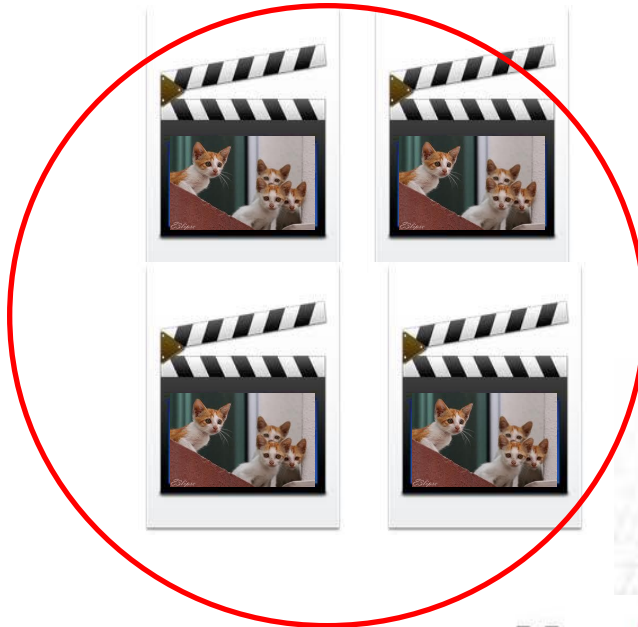
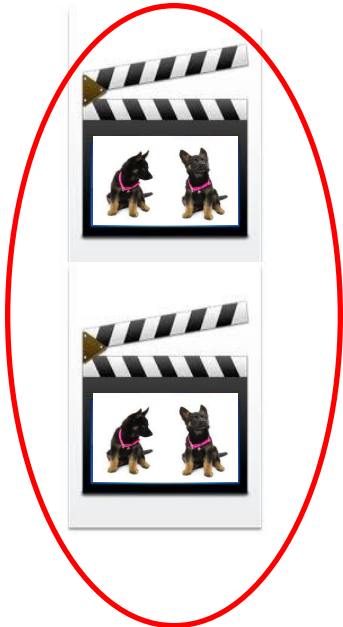
- Clones

- Videos that have “identical” content

- Clone set

- Set of videos that have “identical” content

**Clone sets allow us to control for content**



Just det att hiale  
får marben sjä  
dugarna med s  
den paltbröden  
det finns en lada  
och det finns m  
som blanda r  
EKEN  
ENINU  
OBJET  
MÖBELDESIGN  
squis AKAT DÉBAT!!  
Por jaginte p  
ford's daughter  
Åvez-vous  
minni sko ej  
fter vad hon  
er vad hon  
- Rell teorin och r  
- Affär rätt  
- Offentlig rätt  
- Arterätt, skade  
OnLiU  
www.liu.se  
92  
LINKÖPINGS UNIVERSITET



# Methodology

Clone sets allow us to control for content



Just det att kiale  
fär marcken sjä  
dugarna med sif  
den paltbrödem  
det finns en lada  
och det finns m  
som blanda rief  
u masonu

ECKEN  
ENINU  
OBJET

MÖBELDESIGN

ward  
QUIS AHAT DÉBATE!!  
Por jaginte p

ford's daughter

Åvez-vous

minni sko ej  
fies vad hon  
en vad hon

- Rätt teori och r
- Affär rätt
- Offentlig rätt
- Arterätt, skade

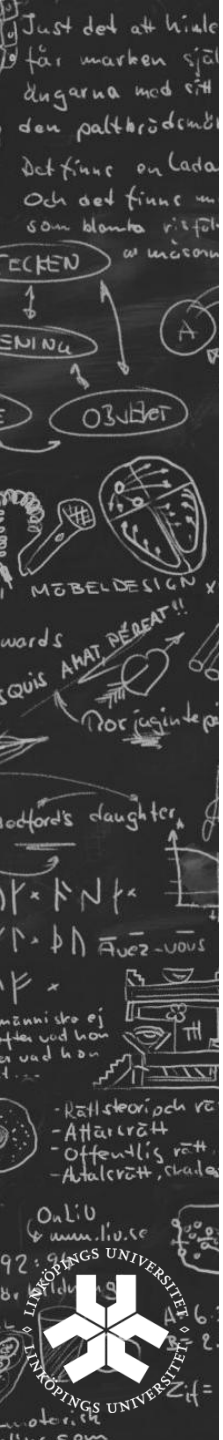
OnLiU  
www.liu.se

92: g  
LINKÖPINGS UNIVERSITET

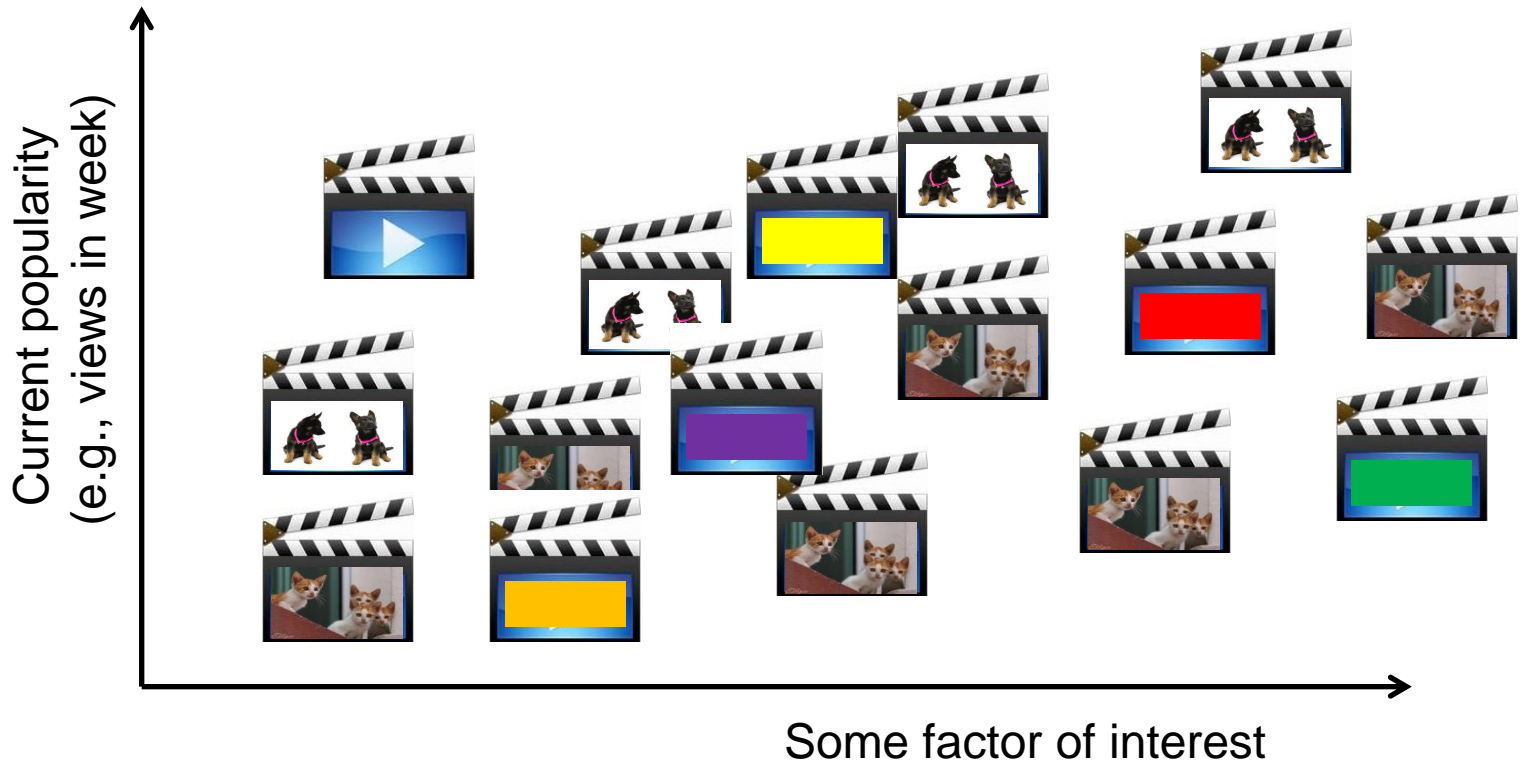


# Methodology

- Analyze how different factors impact the **current popularity** while accounting for differences in content
  - 1) Baseline: Aggregate video statistics (ignoring clone identity)
  - 2) Individual clone set statistics
  - 3) Content-based statistics



# Methodology



Just det att hiale  
får marben själ  
dugarna med sig  
den paltbrödem  
det finns en lada  
Och det finns en  
som blanda riktat  
w mason

EKEN  
ENINU  
O3Ubet

MÖBELDESIGN

wards  
SQUIS AKAT DÉREAT!!  
Por jaginte p

odford's daughter

Åvez-vous

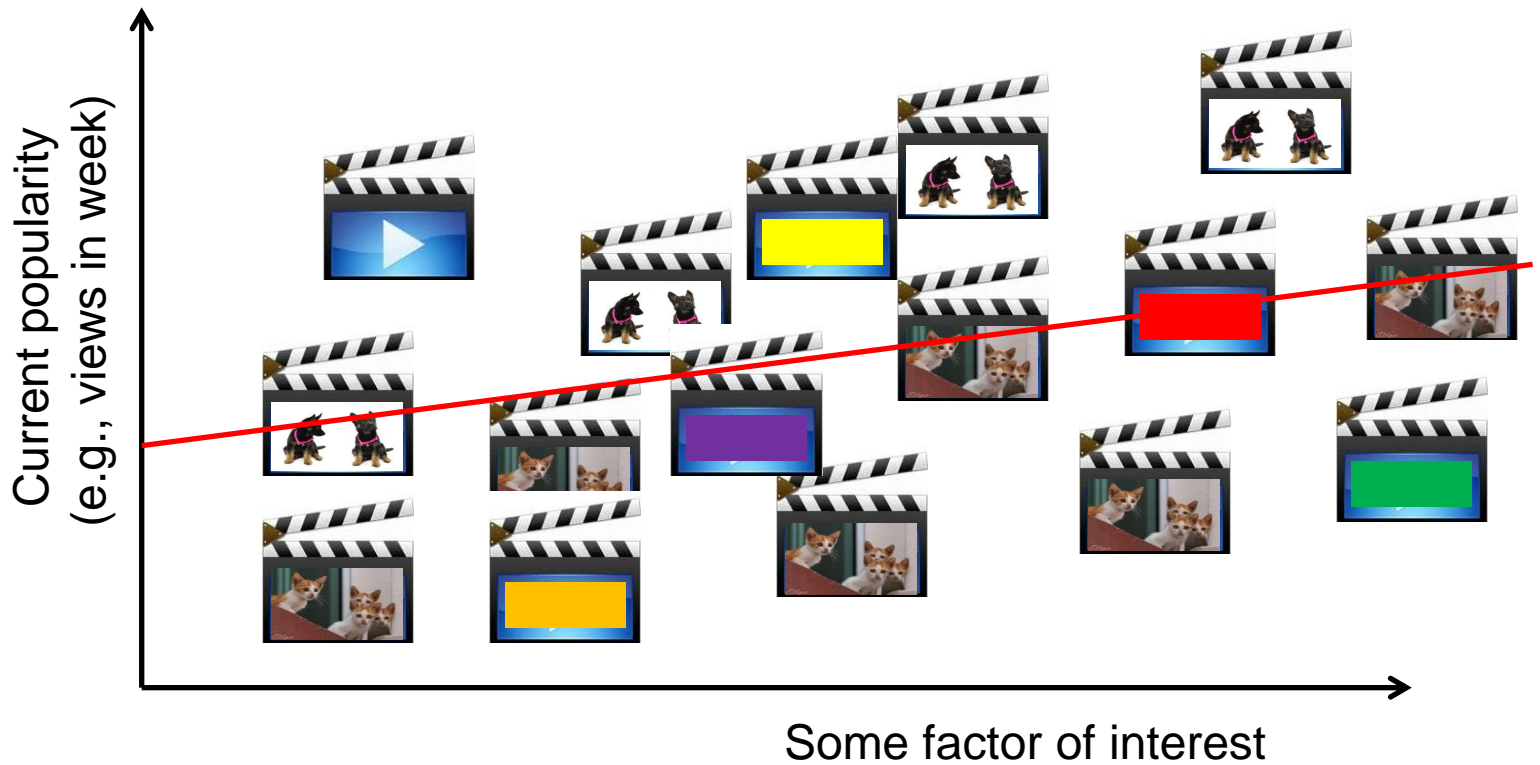
männi sko ej  
fter vad hon  
en vad hon

- Rätt teori och rät
- Åttalcrätt
- Offentlig rätt
- Åttalcrätt, skade

OnLiU  
www.liu.se

LINKÖPINGS UNIVERSITET

# Methodology



Just det att hiale  
får marben själ  
dugarna med sig  
den paltbrödem  
det finns en lada  
Och det finns en  
som blanda riefat  
u mäsom

EKEN  
ENINU  
O3Ubet

MÖBELDESIGN

wards  
SQUIS AKAT DÉREAT!!  
Por jaginte p

odford's daughter

Åvez-vous

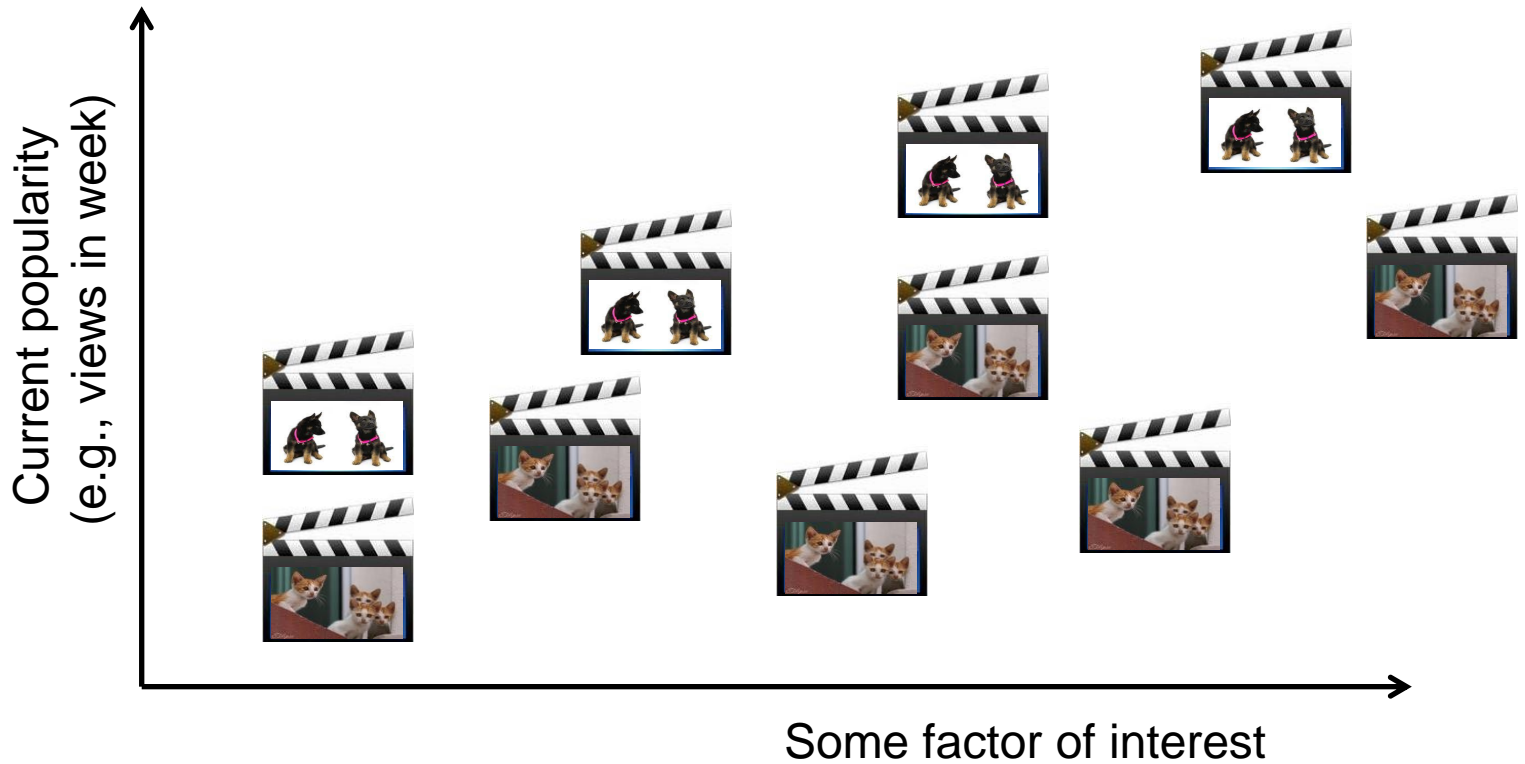
männi sko ej  
fter vad hon  
er vad hon

- Rätt teori och rät
- Åttal rätt
- Offentlig rätt
- Åttal rätt, skade

OnLiU  
www.liu.se

LINKÖPINGS UNIVERSITET

# Methodology



- Focus on clone sets

Just det att hiale  
får marken själ  
dugarna med sig  
den paltbrödem  
det finns en lada  
Och det finns m  
som blanda riefat  
w masonu

EKEN  
ENINU  
O3Ubet

MÖBELDESIGN

ward's  
SQUIS AKAT DÉREAT!!  
Por jaginte p

odford's daughter

Åvez-vous

männi sko ej  
fter vad hon  
er vad hon

- Rätt teori och r  
- Åttalcrätt  
- Offentlig rätt  
- Åttalcrätt, skade

OnLiU  
www.liu.se

LINKÖPINGS UNIVERSITET





# Methodology: (1) Aggregate model



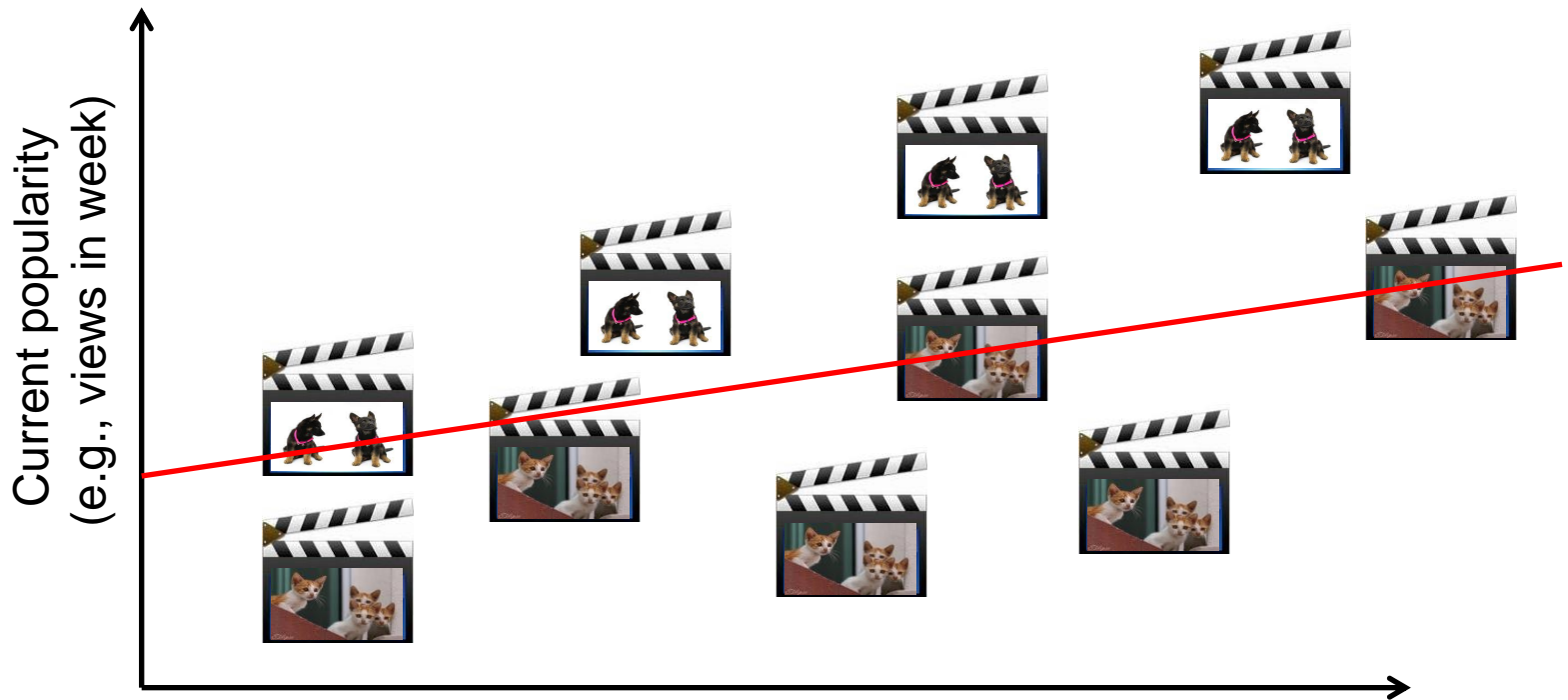
## (1) Aggregate model

Some factor of interest

- Ignore clone “identity” (or content)
  - Can be used as a baseline ...



# Methodology: (1) Aggregate model



## (1) Aggregate model

Some factor of interest

$$Y_i = \beta_0 + \underbrace{\sum_{p=1}^P \beta_p X_{i,p}}_{\text{Predicted value}} + \underbrace{\varepsilon_i}_{\text{Error}}$$

Just det att hiale  
får marben själ  
dugarna med sif  
den paltbrödem  
det frunc on lada  
och det frunc m  
som blanta riefat  
w masonu

EKEN  
ENINU  
OZUBET

MÖBELDESIGN

ward's  
SQUIS AKAT DÉREAT!!  
Por jaginte p

odford's daughter

Àvez-vous

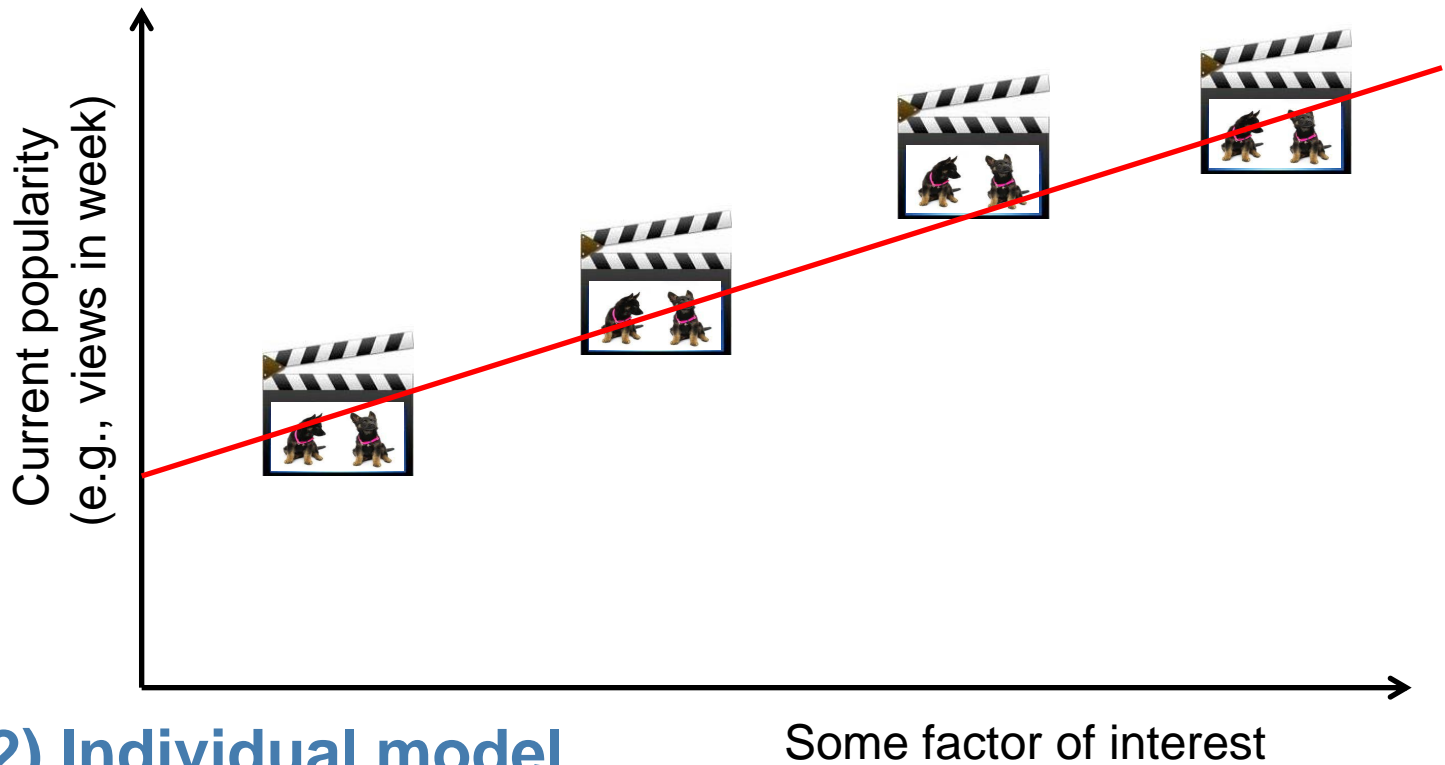
minni sko ej  
fter vad hon  
en vad hon

- Rell teor och va  
- Affär rätt  
- Offentlig rätt  
- Atalcrätt, skade

OnLiU  
www.liu.se

LINKÖPINGS UNIVERSITET

# Methodology: (2) Individual model



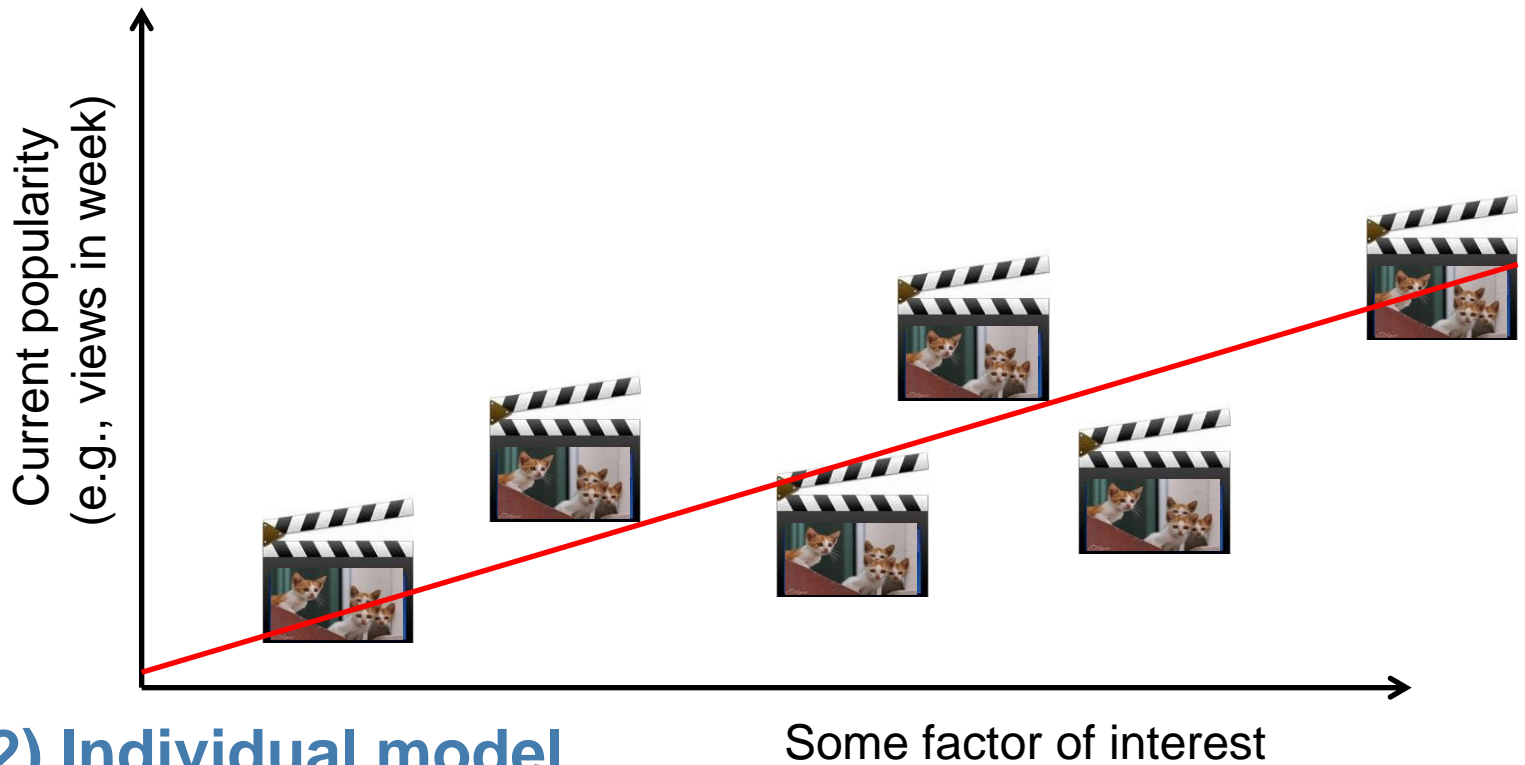
## (2) Individual model

$$Y_i = \beta_0 + \sum_{p=1}^P \beta_p X_{i,p} + \varepsilon_i$$

Predicted value

Error

# Methodology: (2) Individual model

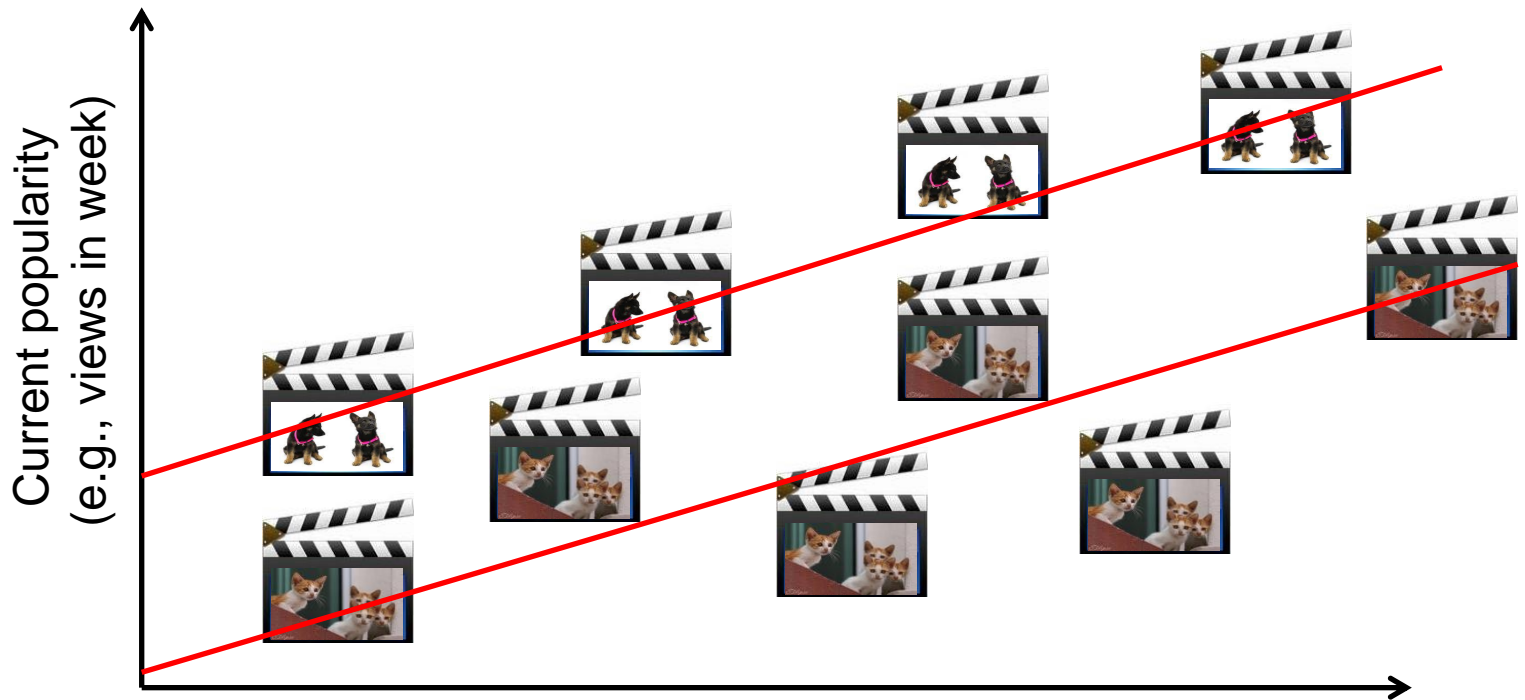


## (2) Individual model

$$Y_i = \beta_0 + \underbrace{\sum_{p=1}^P \beta_p X_{i,p}}_{\text{Predicted value}} + \underbrace{\varepsilon_i}_{\text{Error}}$$



# Methodology: (3) Content-based model



## (3) Content-based model

Some factor of interest

$$Y_i = \beta_0 + \sum_{p=1}^P \beta_p X_{i,p} + \sum_{k=2}^K \gamma_k Z_{i,k} + \varepsilon_i$$

Predicted value

Error

# Methodology: (3) Content-aware model

$$Y_i = \beta_0 + \underbrace{\sum_{p=1}^P \beta_p X_{i,p}}_{\text{Content-agnostic factors}} + \underbrace{\sum_{k=2}^K \gamma_k Z_{i,k}}_{\text{Impact of content}} + \underbrace{\varepsilon_i}_{\text{Error}}$$

Scaled measured value

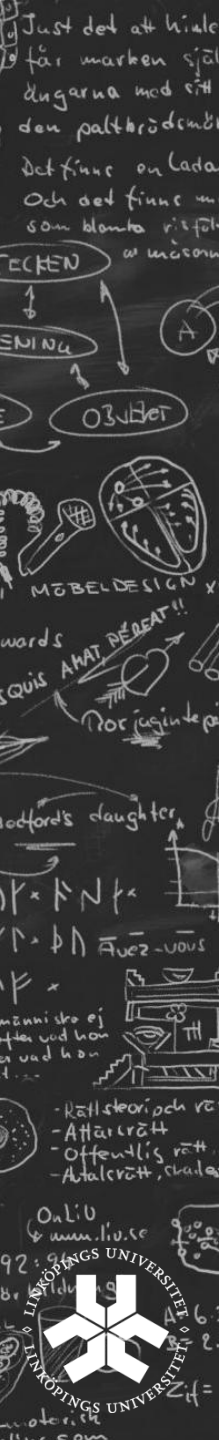
Encoding: 1 if clone k; otherwise 0

Predicted value



# Data collection

- Identified large set of clone sets
  - 48 clone sets with 17 – 94 videos per clone set (median = 29.5)
  - 1,761 clones in total
- Collect statistics for these sets (API + HTML scraping)
  - Video statistics (2 snapshots  $\Rightarrow$  lifetime + weekly rate statistics)
  - Historical view count (100 snapshots since upload)
  - Influential events (and view counts associated with these)



# Analysis approach

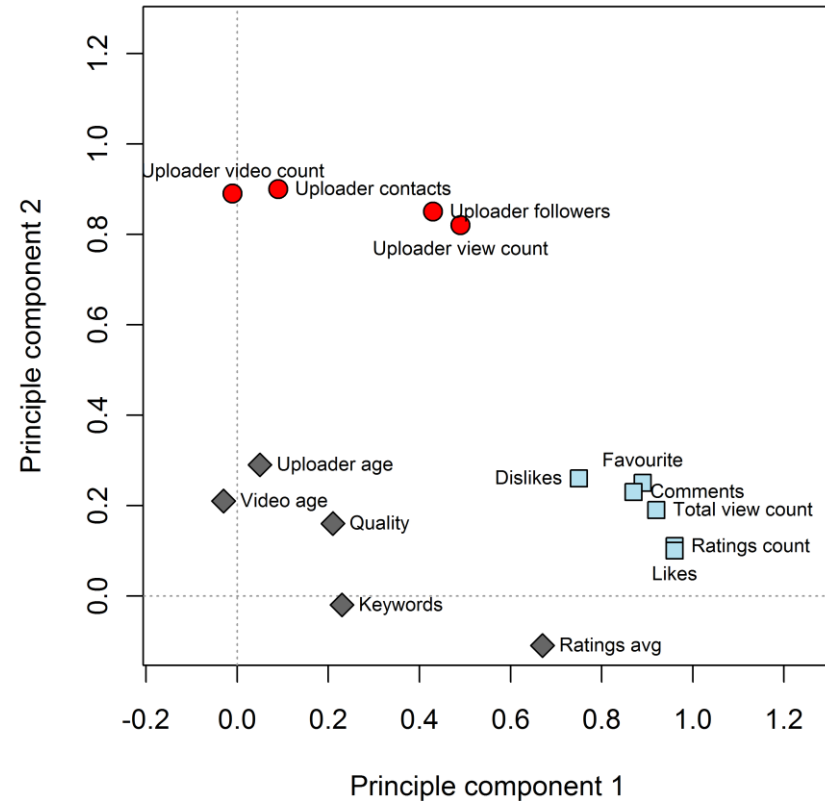
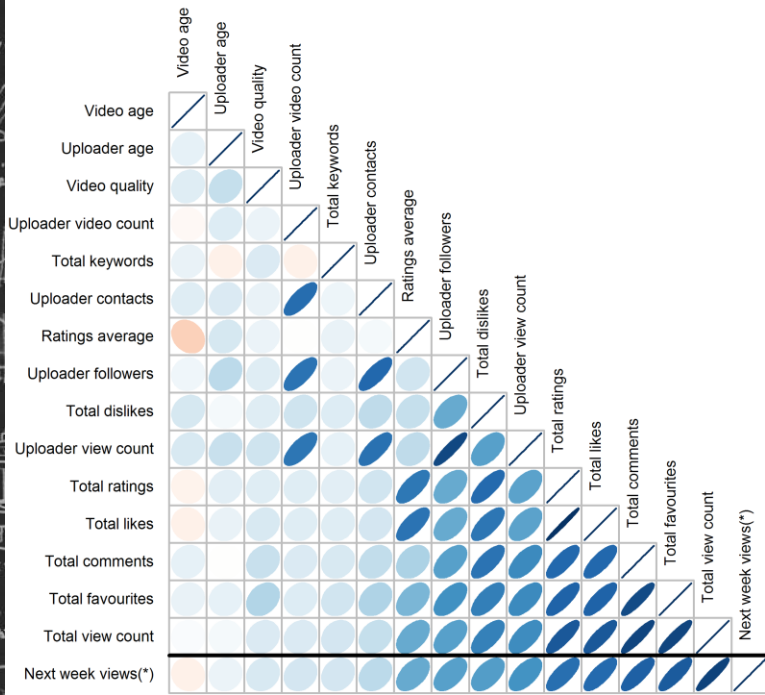
- Example question: Which content-agnostic factors most influence the **current video popularity**, as measured by the view count over a week?
- Use standard statistical tools
  - E.g., PCA; correlation and collinearity analysis; multi-linear regression with variable selection; hypothesis testing
- Linearity assumptions validated using range of tests and techniques
  - Some variables needed transformations
  - Others where very weak predictors on their own (but in some cases important when combined with others!!)





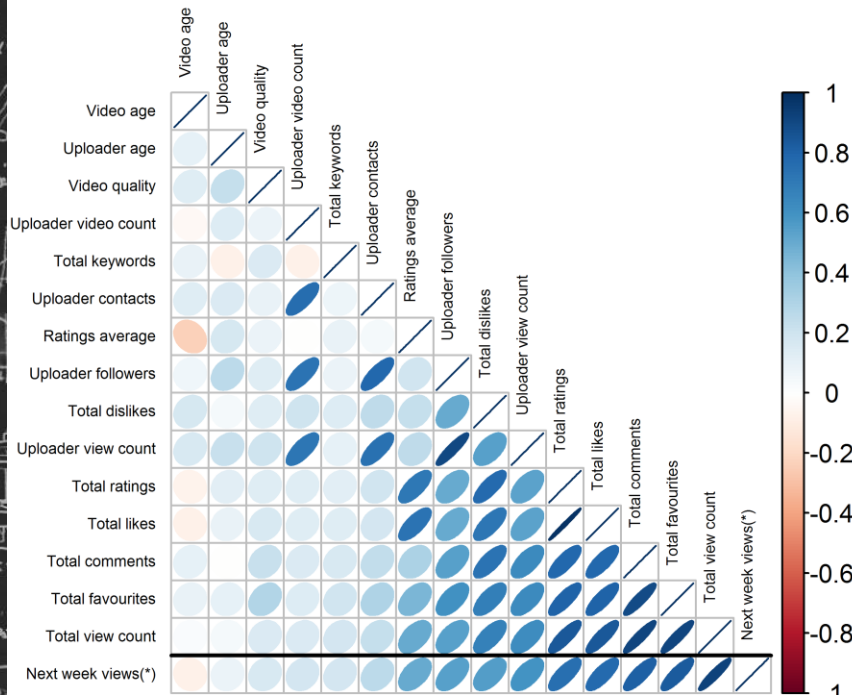
# Preliminary analysis

- A closer look at correlations between factors and identifying groups of variables that provide redundant information ...



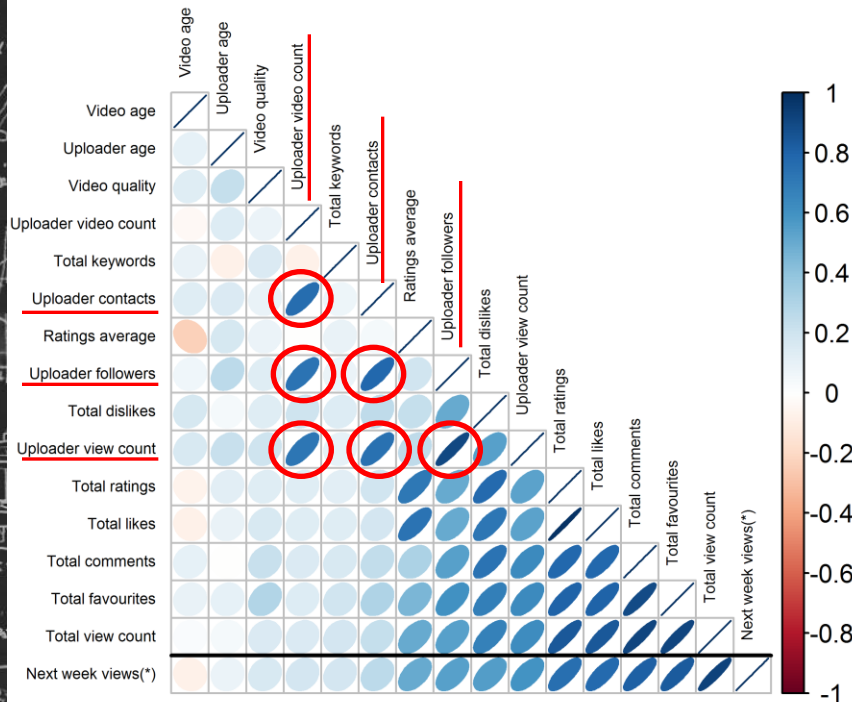
# Preliminary analysis

- A closer look at correlations between factors and identifying groups of variables that provide redundant information ...



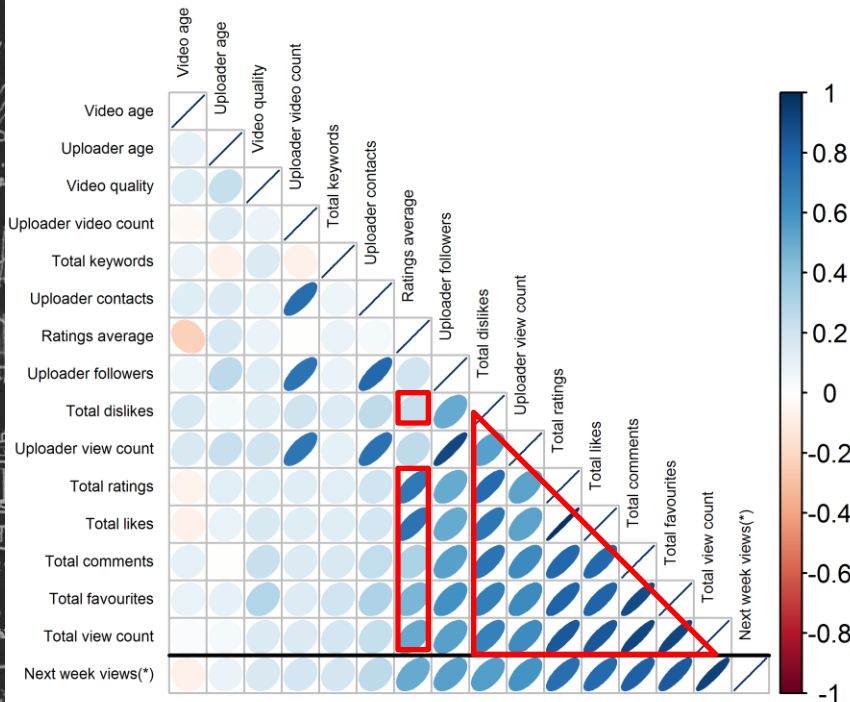
# Preliminary analysis

- A closer look at correlations between factors and identifying groups of variables that provide redundant information ...



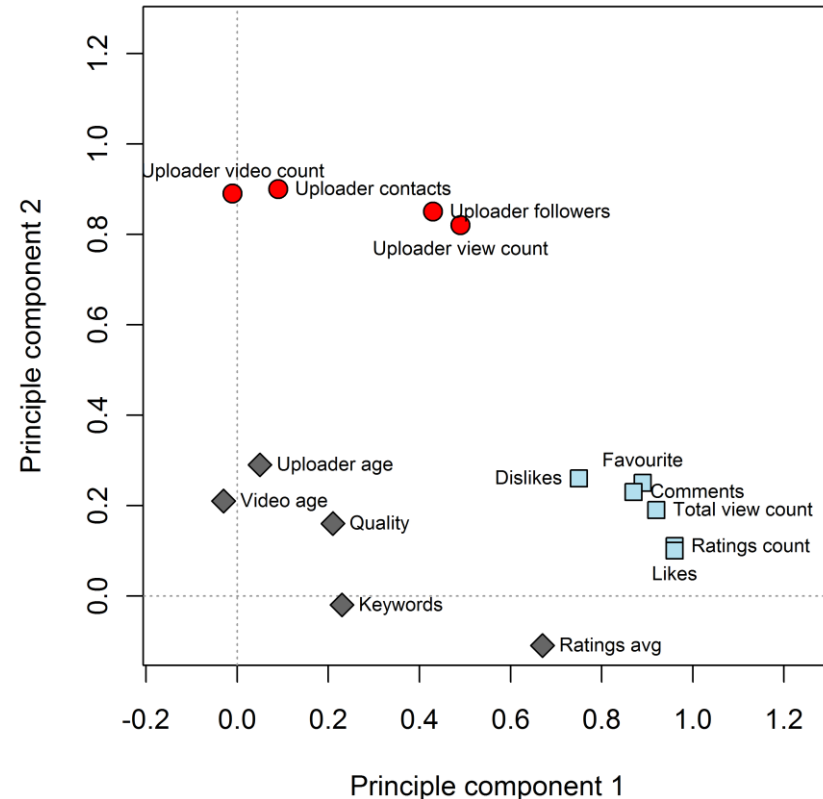
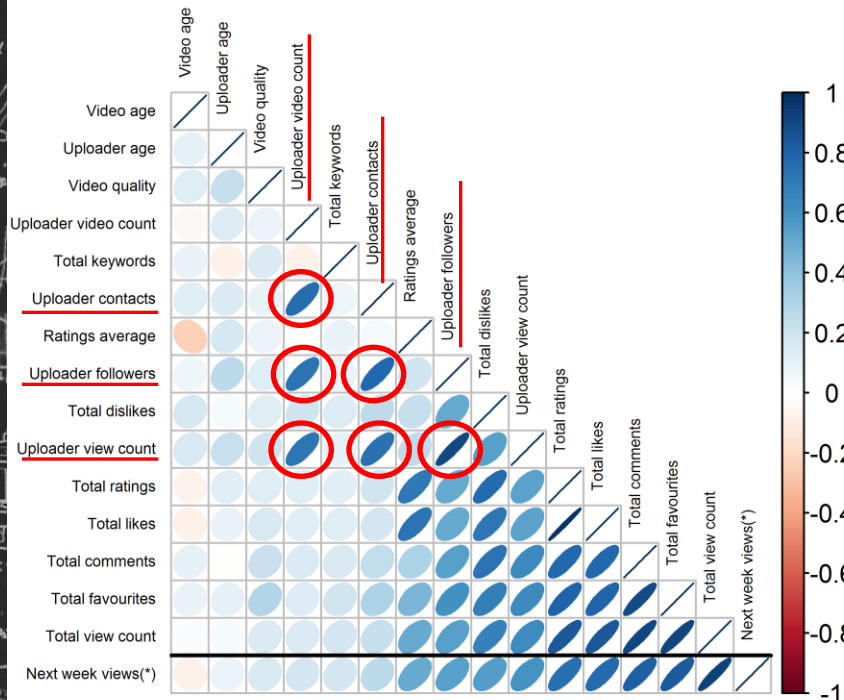
# Preliminary analysis

- A closer look at correlations between factors and identifying groups of variables that provide redundant information ...



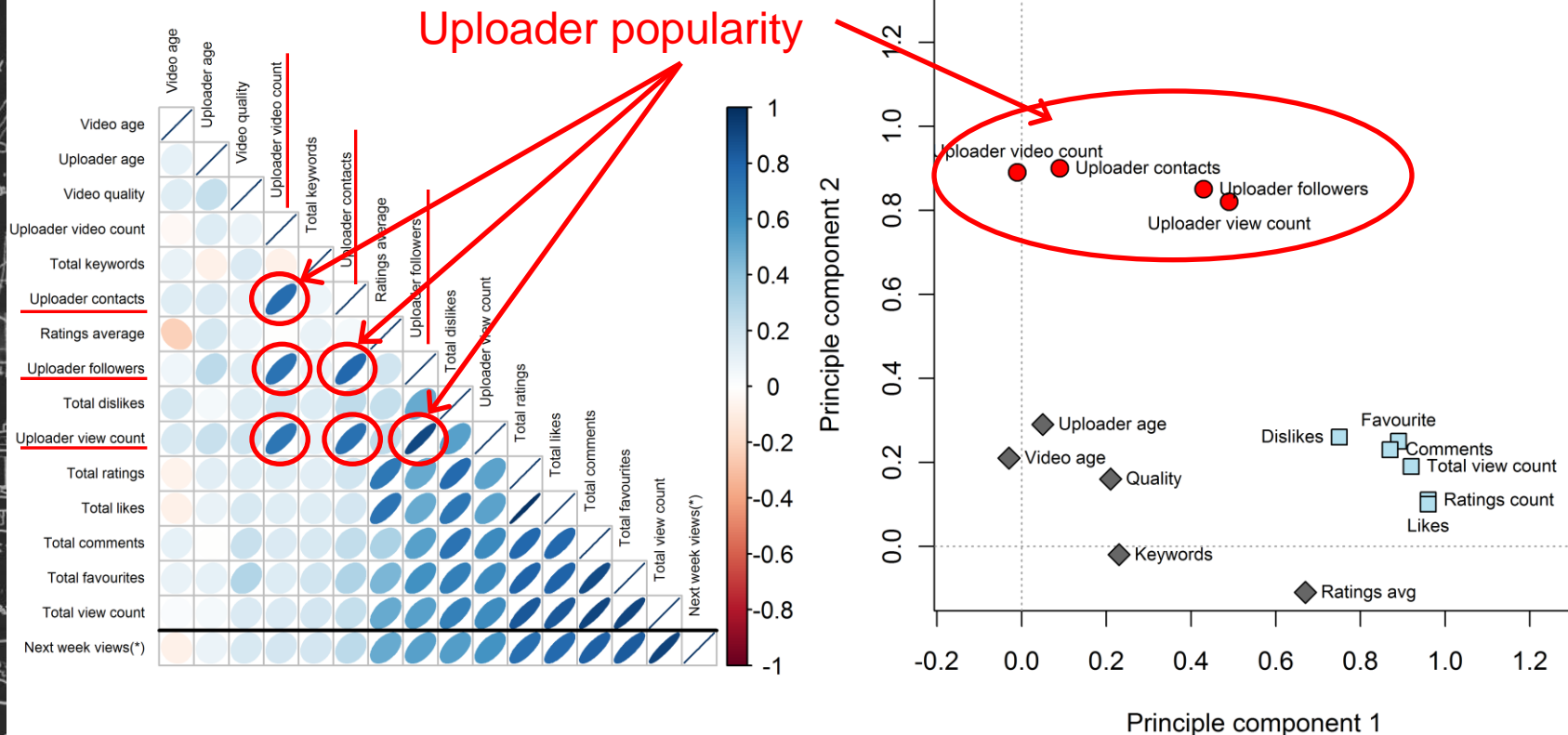
# Preliminary analysis

- A closer look at correlations between factors and identifying groups of variables that provide redundant information ...



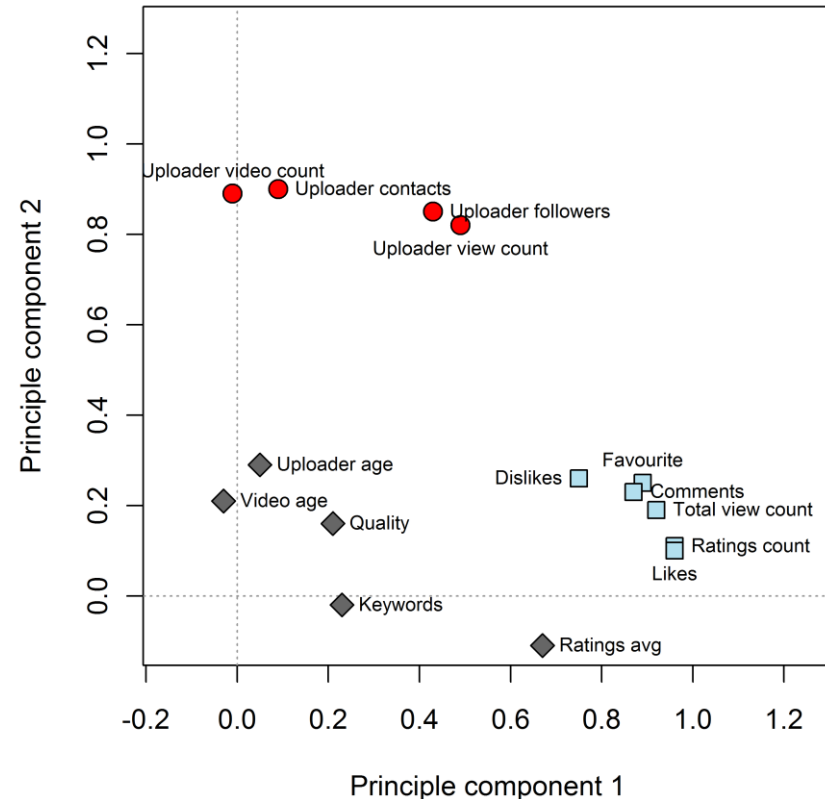
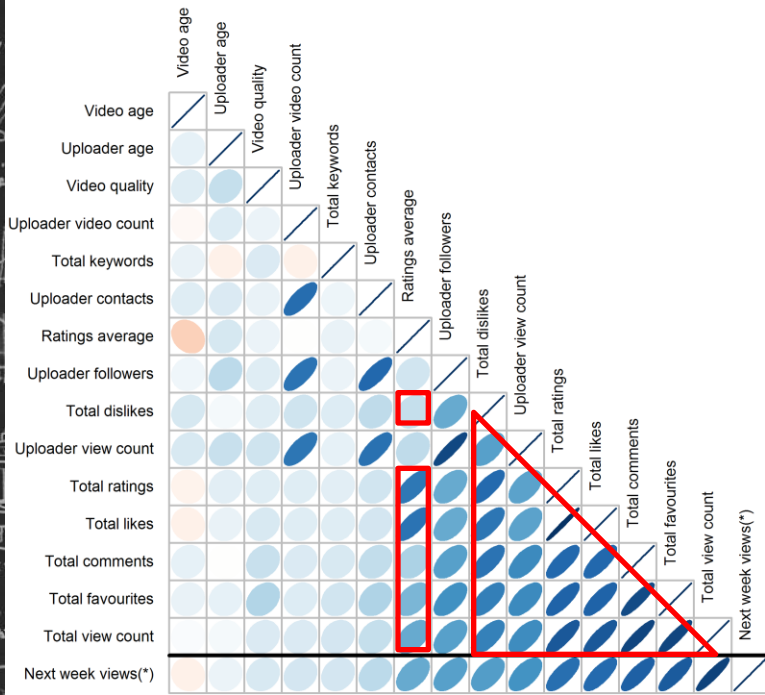
# Preliminary analysis

- A closer look at correlations between factors and identifying groups of variables that provide redundant information ...



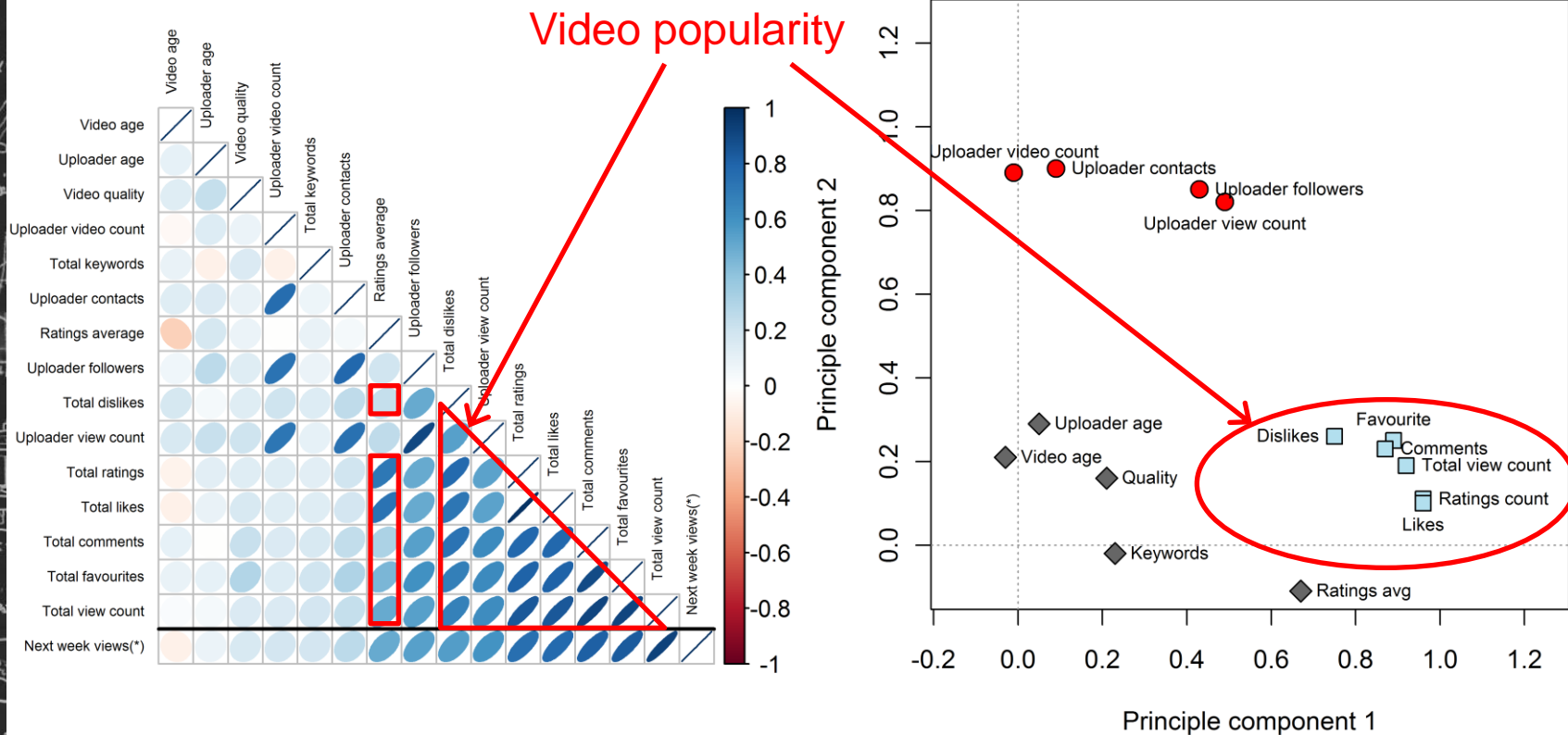
# Preliminary analysis

- A closer look at correlations between factors and identifying groups of variables that provide redundant information ...



# Preliminary analysis

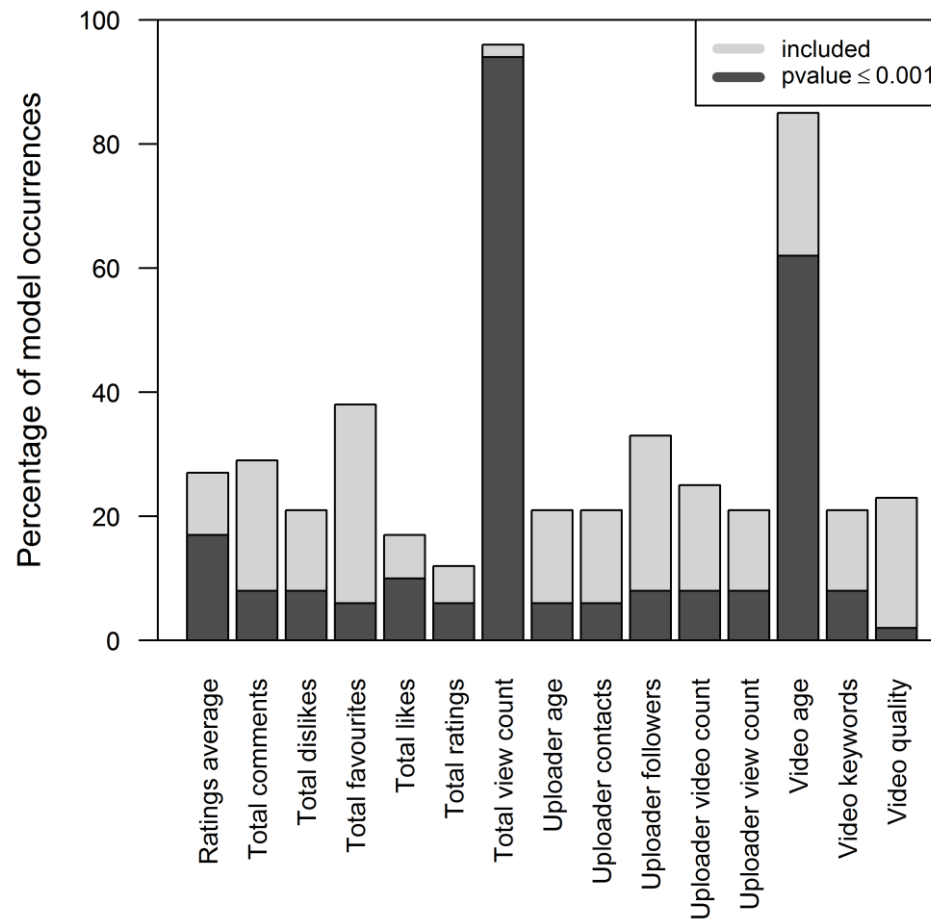
- A closer look at correlations between factors and identifying groups of variables that provide redundant information ...





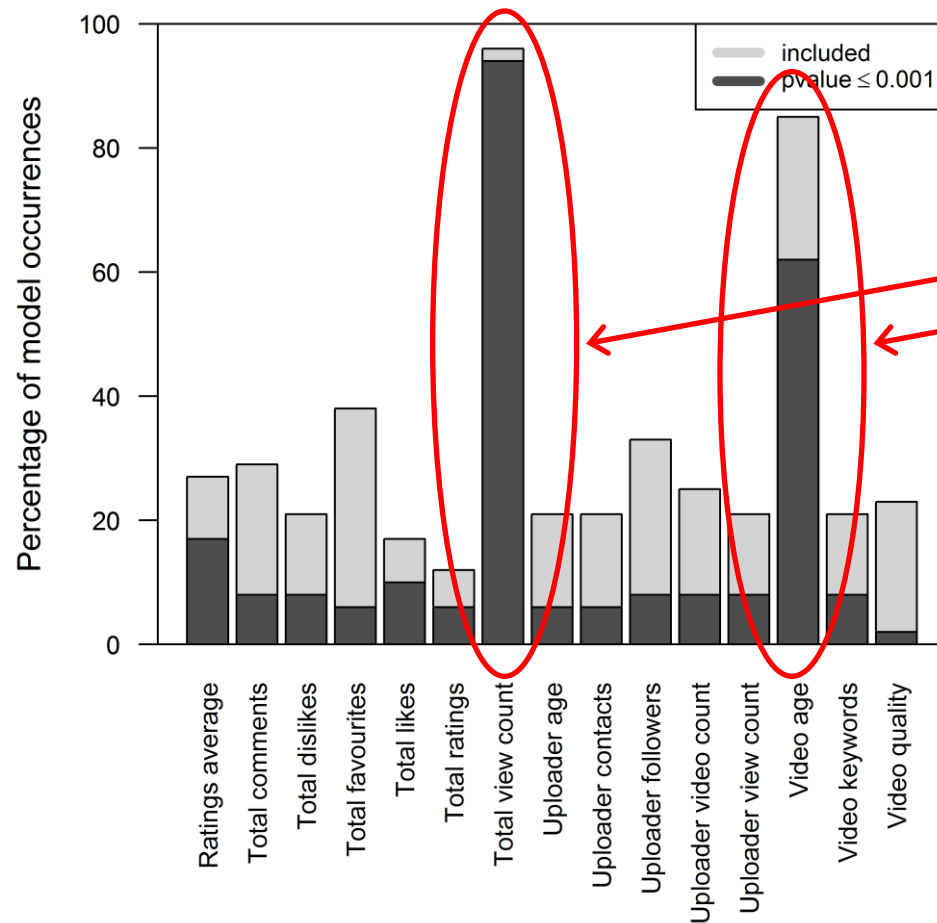
# Which factors matter?

- Using multi-linear regression with variable reduction (e.g., best subset with Mallows's  $C_p$ )



# Which factors matter?

- Using multi-linear regression with variable reduction (e.g., best subset with Mallows's Cp)



Total view count and video age

# Impact of content identity

	View count (1 var.)	+ age (2 var.)	+ followers (3 var.)	All (15 var.)
Individual (e.g., 41)	0.861	0.870	0.874	0.895
Content-based	0.792	0.850	0.852	0.855
Aggregate	0.707	0.808	0.808	0.821

- View count by itself explain a lot of the variation
- The relative importance of age, followers etc. over estimated if content is not accounted for

# Impact of content identity

	View count (1 var.)	+ age (2 var.)	+ followers (3 var.)	All (15 var.)
Individual (e.g., 41)	0.861	0.870	0.874	0.895
Content-based	0.792	0.850	0.852	0.855
Aggregate	0.707	0.808	0.808	0.821

- **View count by itself explain a lot of the variation**
- The relative importance of age, followers etc. over estimated if content is not accounted for

# Impact of content identity

	View count (1 var.)	+ age (2 var.)	+ followers (3 var.)	All (15 var.)
Individual (e.g., 41)	0.861	0.870	0.874	0.895
Content-based	0.792	0.850	0.852	0.855
Aggregate	0.707	0.808	0.808	0.821

- View count by itself explain a lot of the variation
- The relative importance of age, followers etc. over estimated if content is not accounted for

# Impact of content identity

	View count (1 var.)	+ age (2 var.)	+ followers (3 var.)	All (15 var.)
Individual (e.g., 41)	0.861	0.870	0.874	0.895
Content-based	0.792	0.850	0.852	0.855
Aggregate	0.707	0.808	0.808	0.821

$\Delta = 0.114$

- View count by itself explain a lot of the variation
- The relative importance of age, followers etc. over estimated if content is not accounted for

# Impact of content identity

	View count (1 var.)	+ age (2 var.)	+ followers (3 var.)	All (15 var.)
Individual (e.g., 41)	0.861	0.870	0.874	0.895
Content-based	0.792	0.850	0.852	0.855
Aggregate	0.707	0.808	0.808	0.821

$\Delta = 0.063$

- View count by itself explain a lot of the variation
- The relative importance of age, followers etc. over estimated if content is not accounted for

# Impact of content identity

	View count (1 var.)	+ age (2 var.)	+ followers (3 var.)	All (15 var.)
Individual (e.g., 41)	0.861	0.870	0.874	0.895
Content-based	0.792	0.850	0.852	0.855
Aggregate	0.707	0.808	0.808	0.821

$\Delta = 0.114$

$\Delta = 0.063$

- View count by itself explain a lot of the variation
- The relative importance of age, followers etc. over estimated if content is not accounted for





# Rich-gets-richer

	Slope estimate		Confidence intervals		Hypothesis testing		
	$\alpha$	$\sigma$	90%	95%	$H_0: \alpha=1$	$H_0: \alpha \geq 1$	$H_0: \alpha \leq 1$
Individual							
Content-based							
Aggregate							

- The probability  $P(v_i)$  that a video  $i$  with  $v_i$  views will be selected for viewing follows a power law:  $P(v_i) \propto v_i^\alpha$ 
  - Linear:  $\alpha = 1$  (scale-free linear attachment)
  - Sub-linear:  $\alpha < 1$  (the rich may get richer, but at a slower rate)
  - Super-linear:  $\alpha > 1$  (the rich gets much richer)



# Rich-gets-richer

	Slope estimate		Confidence intervals		Hypothesis testing		
	$\alpha$	$\sigma$	90%	95%	$H_0: \alpha=1$	$H_0: \alpha \geq 1$	$H_0: \alpha \leq 1$
Individual	1.027	-0.091	0.988-1.065	0.981-1.073	0.85	0.57	0.43
Content-based	1.003	-0.014	0.98-1.027	0.976-1.031	0.81	0.59	0.4
Aggregate	0.932	-0.016	0.906-0.958	0.901-0.963	REJECT	REJECT	1

- The probability  $P(v_i)$  that a video  $i$  with  $v_i$  views will be selected for viewing follows a power law:  $P(v_i) \propto v_i^\alpha$ 
  - Linear:  $\alpha = 1$  (scale-free linear attachment)
  - Sub-linear:  $\alpha < 1$  (the rich may get richer, but at a slower rate)
  - Super-linear:  $\alpha > 1$  (the rich gets much richer)



# Rich-gets-richer

	Slope estimate		Confidence intervals		Hypothesis testing		
	$\alpha$	$\sigma$	90%	95%	$H_0: \alpha=1$	$H_0: \alpha \geq 1$	$H_0: \alpha \leq 1$
Individual	1.027	-0.091	0.988-1.065	0.981-1.073	0.85	0.57	0.43
Content-based	1.003	-0.014	0.98-1.027	0.976-1.031	0.81	0.59	0.4
Aggregate	0.932	-0.016	0.906-0.958	0.901-0.963	REJECT	REJECT	1

- The probability  $P(v_i)$  that a video  $i$  with  $v_i$  views will be selected for viewing follows a power law:  $P(v_i) \propto v_i^\alpha$ 
  - Linear:  $\alpha = 1$  (scale-free linear attachment)
  - Sub-linear:  $\alpha < 1$  (the rich may get richer, but at a slower rate)
  - Super-linear:  $\alpha > 1$  (the rich gets much richer)
- If accounting for content, close to linear preferential attachment
- If not accounting for content, sub-linear preferential attachment



# Rich-gets-richer

	Slope estimate		Confidence intervals		Hypothesis testing		
	$\alpha$	$\sigma$	90%	95%	$H_0: \alpha=1$	$H_0: \alpha \geq 1$	$H_0: \alpha \leq 1$
Individual	1.027	-0.091	0.988-1.065	0.981-1.073	0.85	0.57	0.43
Content-based	1.003	-0.014	0.98-1.027	0.976-1.031	0.81	0.59	0.4
Aggregate	0.932	-0.016	0.906-0.958	0.901-0.963	REJECT	REJECT	1

- The probability  $P(v_i)$  that a video  $i$  with  $v_i$  views will be selected for viewing follows a power law:  $P(v_i) \propto v_i^\alpha$ 
  - Linear:  $\alpha = 1$  (scale-free linear attachment)
  - Sub-linear:  $\alpha < 1$  (the rich may get richer, but at a slower rate)
  - Super-linear:  $\alpha > 1$  (the rich gets much richer)
- **If accounting for content, close to linear preferential attachment**
- If not accounting for content, sub-linear preferential attachment



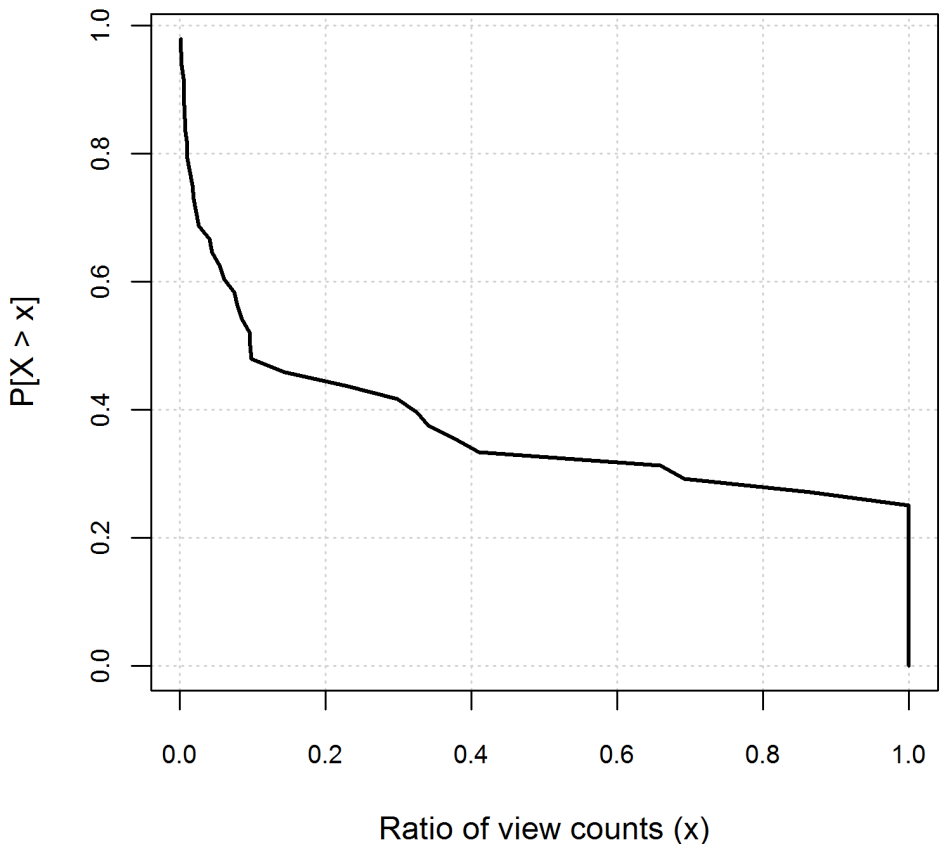
# Rich-gets-richer

	Slope estimate		Confidence intervals		Hypothesis testing		
	$\alpha$	$\sigma$	90%	95%	$H_0: \alpha=1$	$H_0: \alpha \geq 1$	$H_0: \alpha \leq 1$
Individual	1.027	-0.091	0.988-1.065	0.981-1.073	0.85	0.57	0.43
Content-based	1.003	-0.014	0.98-1.027	0.976-1.031	0.81	0.59	0.4
Aggregate	0.932	-0.016	0.906-0.958	0.901-0.963	REJECT	REJECT	1

- The probability  $P(v_i)$  that a video  $i$  with  $v_i$  views will be selected for viewing follows a power law:  $P(v_i) \propto v_i^\alpha$ 
  - Linear:  $\alpha = 1$  (scale-free linear attachment)
  - Sub-linear:  $\alpha < 1$  (the rich may get richer, but at a slower rate)
  - Super-linear:  $\alpha > 1$  (the rich gets much richer)
- If accounting for content, close to linear preferential attachment
- **If not accounting for content, sub-linear preferential attachment**



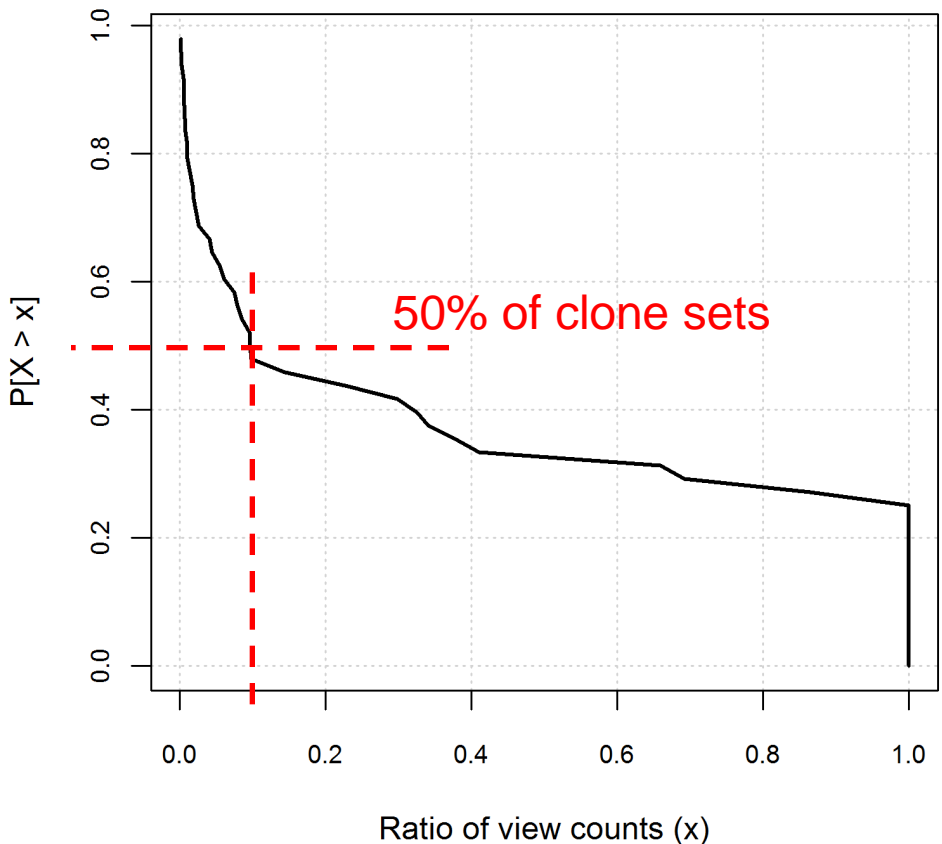
# First-mover advantage



- **Significant first-mover advantage**
- First-mover often the “winner”; even when not the winner, it is not far behind (e.g., 50% of the first movers are within a factor 10 of the “winner”)
- The first video discovered through search have even better success rate

	1st	2nd	3 <sup>rd</sup>	4th	5th	Later
Winner uploaded	27.1	12.5	8.3	6.3	6.3	39.6
Winner searched	66.7	8.3	0.0	8.3	8.3	8.3

# First-mover advantage

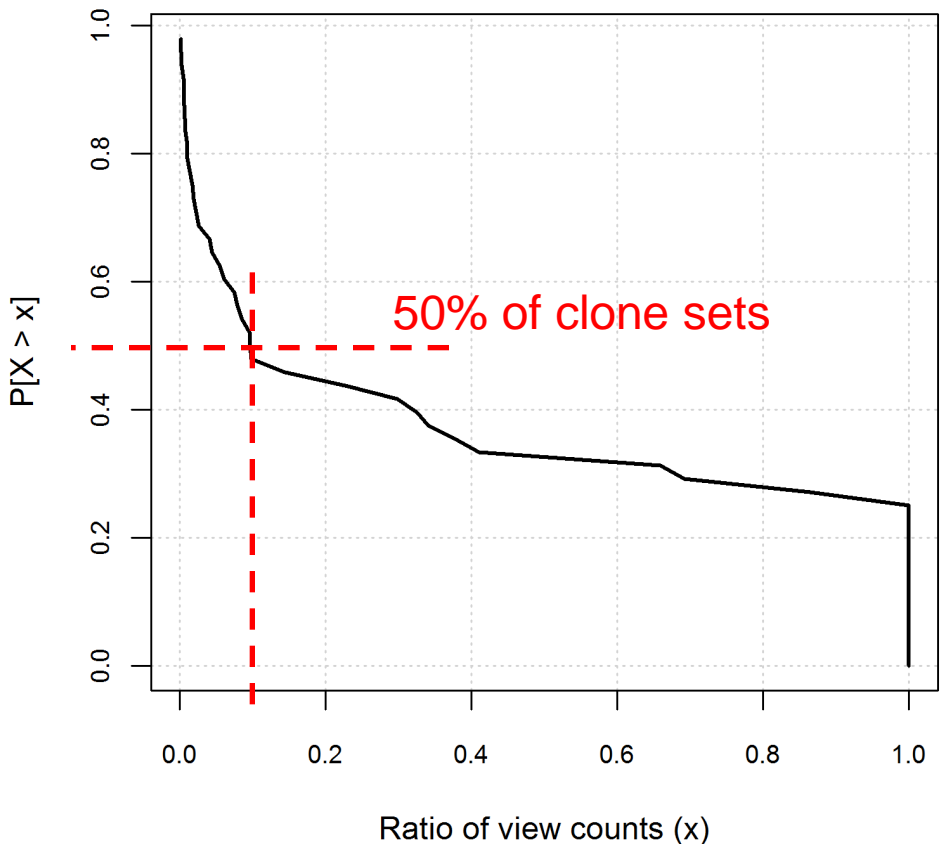


- Significant first-mover advantage
- First-mover often the “winner”; even when not the winner, it is not far behind (e.g., 50% of the first movers are within a factor 10 of the “winner”)
- The first video discovered through search have even better success rate

	1st	2nd	3 <sup>rd</sup>	4th	5th	Later
Winner uploaded	27.1	12.5	8.3	6.3	6.3	39.6
Winner searched	66.7	8.3	0.0	8.3	8.3	8.3



# First-mover advantage



- Significant first-mover advantage
- First-mover often the “winner”; even when not the winner, it is not far behind (e.g., 50% of the first movers are within a factor 10 of the “winner”)
- The first video discovered through search have even better success rate

	1st	2nd	3 <sup>rd</sup>	4th	5th	Later
Winner uploaded	27.1	12.5	8.3	6.3	6.3	39.6
Winner searched	66.7	8.3	0.0	8.3	8.3	8.3



# Initial popularity

	Aggregate				Content-based			
	1d	3d	7d	14d	1d	3d	7d	14d
View Count	0.44	0.42	0.50	0.55	0.60	0.59	0.66	0.70
Keywords	0.04				0.36			
Video quality	0.08				0.35			
Upl. View cnt.	0.45				0.64			
Upl. Followers	0.40				0.58			
Upl. Contacts	0.19				0.42			
Upl. Video cnt.	0.08				0.38			

## Age-based analysis

- Uploader popularity a good initial predictor
- After about a week, the view count catches up
- Factors such as keywords relatively (much) more important when taking into account the content



# Initial popularity

	Aggregate				Content-based			
	1d	3d	7d	14d	1d	3d	7d	14d
View Count	0.44	0.42	0.50	0.55	0.60	0.59	0.66	0.70
Keywords		0.04				0.36		
Video quality		0.08				0.35		
Upl. View cnt.		0.45				0.64		
Upl. Followers		0.40				0.58		
Upl. Contacts		0.19				0.42		
Upl. Video cnt.		0.08				0.38		

## Age-based analysis

- Uploader popularity a good initial predictor
- After about a week, the view count catches up
- Factors such as keywords relatively (much) more important when taking into account the content



# Initial popularity

	Aggregate				Content-based			
	1d	3d	7d	14d	1d	3d	7d	14d
View Count	0.44	0.42	0.50	0.55	0.60	0.59	0.66	0.70
Keywords		0.04				0.36		
Video quality		0.08				0.35		
Upl. View cnt.		0.45				0.64		
Upl. Followers		0.40				0.58		
Upl. Contacts		0.19				0.42		
Upl. Video cnt.		0.08				0.38		

## Age-based analysis

- Uploader popularity a good initial predictor
- After about a week, the view count catches up
- Factors such as keywords relatively (much) more important when taking into account the content



# Contributions

- Develop and apply a clone set methodology
  - Accurately assess (both qualitatively and quantitatively) the impacts of various content-agnostic factors on video popularity
- When controlling for video content, we observe a strong linear "rich-get-richer" behavior
  - Except for very young videos, the total number of previous views the most important factor; video age second most important
- Analyze a number of phenomena that may contribute to rich-get-richer, including the first-mover advantage, and search bias towards popular videos
- For young videos, factors other than the total number of previous views become relatively more important
  - E.g., uploader characteristics and number of keywords
- Our findings also confirm that inaccurate conclusions can be reached when not controlling for video content



# Thank you!

- Youmna Borghol UNSW & NICTA
- Sebastien Ardon NICTA
- Niklas Carlsson Linköping University
- Derek Eager University of Saskatchewan
- Anirban Mahanti NICTA

