

# Scalability - Variety

**After this video you will be able to..**

- Describe different aspects of data variety (aka heterogeneity) related to Big Data
- Identify the challenges and opportunities resulting from data variety

**Variety == Complexity**

A long time ago in a galaxy far,  
far away.....

**SECOND IMPERIAL CIVIL WAR**  
The SITH have returned to rule the galaxy! Over a century has passed since the Rebel victory at Endor. The Galactic Alliance, formed during the crisis of the Yong War, has suffered defeat. In an attempt to eliminate all of the Jedi, the Sith Lord DARTH KRAYT has dethroned Emperor Fel, resulting in the outbreak of the Second Imperial Civil War. Many now declare loyalty to Fel.

- Data were confined only to tables

Cars marketplace				
vendor	Model	Price	Mileage	VIN Code
Chevrolet	Corvette	17226	25965.0	ILLAKAWAZDZ
Chevrolet	Corvette	34229	46429.0	RCPNSRYGXOI
Chevrolet	Corvette	27982	50209.0	NWLGCEVEHGI
Chevrolet	Corvette	51825	72998.0	NGVZSCIZGSM
Chevrolet	Corvette	52845	34364.0	PSDRUYYOIJG
Chevrolet	Malibu	37874	37273.0	VLFPQPWNEFD
Chevrolet	Malibu	15600	71441.0	EXLJGDWOZSA
Chevrolet	Malibu	52447	46700.0	NLMGJZAKBRE
Chevrolet	Malibu	27129	36254.0	OIPFUENLEHSX
Chevrolet	Malibu	28846	77162.0	WRCOOFREZLL
Chevrolet	Malibu	46165	60590.0	HUFTTHQHSFJF
Chevrolet	Malibu	18263	37790.0	JL MHNAFESHVD

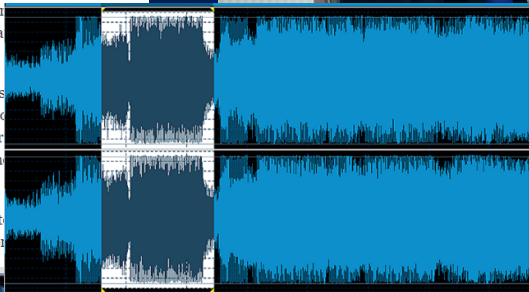
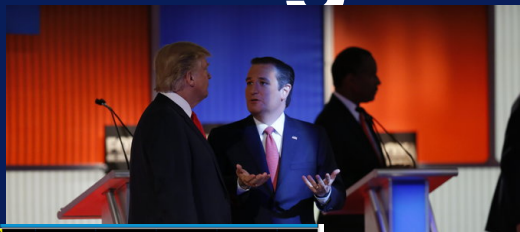
# Today, Data are more heterogeneous

The battle for the Republican nomination appeared more splintered than ever between two halves of a bitterly divided party as several candidates scrambled Friday to consolidate the support of more moderate conservatives a day after a raucous debate.

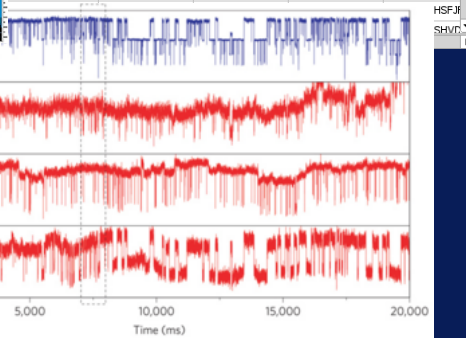
With [Donald J. Trump](#) and Senator [Ted Cruz](#) finally now engaged in an open feud for the most disillusioned voters, Senator [Marco Rubio](#) of Florida, Gov. Chris Christie of New Jersey and [Jeb Bush](#), the former Florida governor, were battling to win over a group of moderate Republicans who are showing little sign of coalescing around a single candidate.

This fracture was most vividly apparent in New Hampshire. Mr. Bush and Mr. Rubio campaigned on Friday, and Mr. Trump is emerging as the obvious alternative to Mr. Bush or Mr. Rubio. Mr. Cruz is also emerging as the obvious alternative to Mr. Trump or Mr. Rubio. Mr. Clinton is also emerging as the obvious alternative to Mr. Trump or Mr. Rubio. Mr. Sanders is also emerging as the obvious alternative to Mr. Trump or Mr. Rubio.

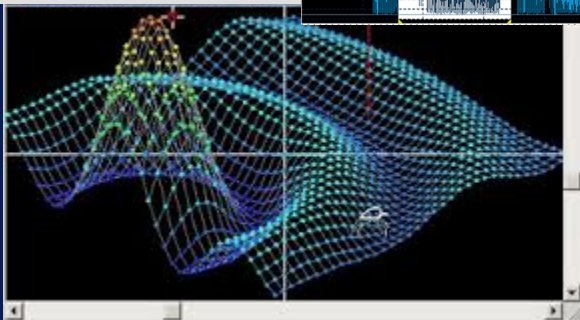
Mr. Bush sought to highlight his image as the candidate of the seasoned, sober-minded wing on Friday with the endorsement of Sen. Lindsey Graham, a moderate Republican and a rival in the presidential race.



Cars marketplace					
vendor	Model	Price	Mileage	VIN Code	
Chevrolet	Corvette	17226	25965.0	ILLAKAWAZD2	
Chevrolet	Corvette	34229	46429.0	PCPNRYGXOI	
Chevrolet	Corvette	27982	50209.0	NWLGCVEHGI	
Chevrolet	Corvette	51825	72998.0	NGVZSCZGSM	
Chevrolet	Corvette	52845	34364.0	PSDRUYOUGG	
Chevrolet	Malibu	37874	37273.0	VLPQPWNEFC	
Chevrolet	Malibu	15600	71441.0	EXLJGDWOZS	
Chevrolet	Malibu	52447	46700.0	NLMGJZAKBRC	
Chevrolet	Malibu	27129	36254.0	OIPFUEHLSH	
Chevrolet	Malibu	28846	77162.0	WRCOOFREZLL	



Rice Trade Network, 2009



# Axes of Data Variety

**Structural  
Variety –  
formats and  
models**

**Semantic  
Variety – how to  
interpret and  
operate on data**

**Media Variety –  
medium in  
which data get  
delivered**

**Availability  
Variations –  
real-time?  
Intermittent?**

# Variety within a Type

- Think of an email collection
  - Table-like part

```
from: Banikumar Maiti (GMAIL) <banikumar.maiti@gmail.com>
to: Reghu Rajan <reghurajan@gmail.com>
cc: Amarnath Gupta <aguptasd@gmail.com>
date: Tue, Feb 2, 2016 at 2:29 PM
subject: Re: Connecting
mailed-by: gmail.com
signed-by: gmail.com
```

# Variety within a Type

- Think of an email collection

- Sender, receiver, date... Well-structured

- Unstructured Text

Dear All,

I would like to congratulate you for putting together a wonderful show.

It was only possible by your hard work.

Dreaming of an UNIQUE show: This credit goes to Zubair. You dreamed about it and made it happen.



# Variety within a Type

- Think of an email collection

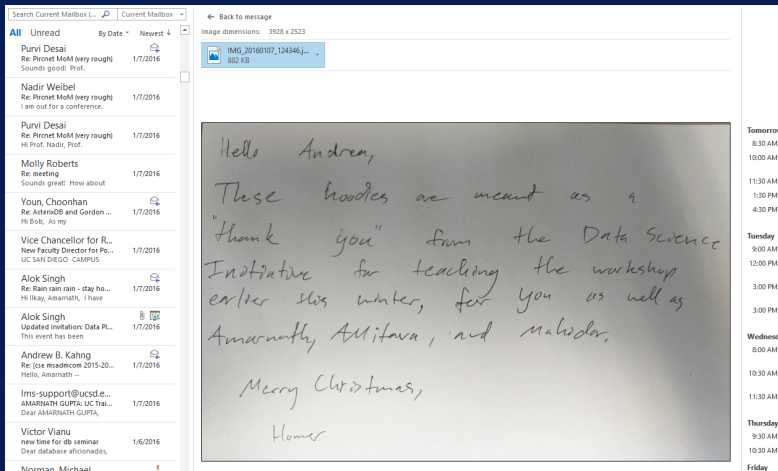
- Sender, receiver, date...

Well-structured

- Body of the email

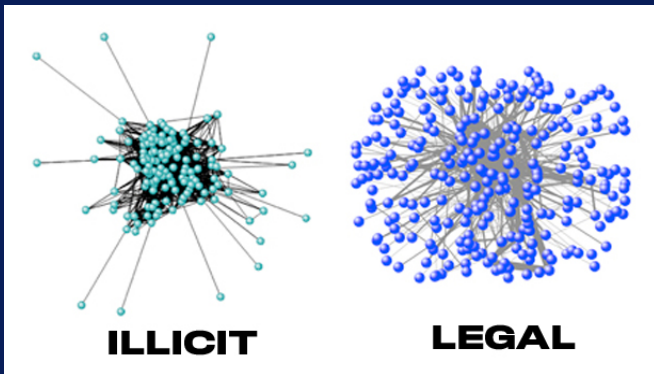
Text

- Media



# Variety within a Type

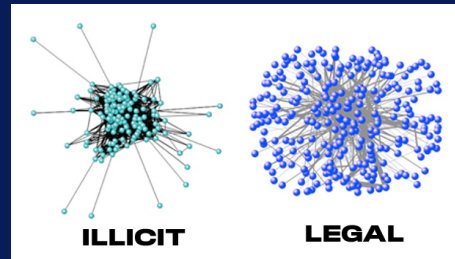
- Think of an email collection
  - Sender, receiver, date... **Well-structured**
  - Body of the email **Text**
  - Attachments **Multi-media**
  - Who-sends-to-whom



# Variety within a Type

- **Think of an email collection**

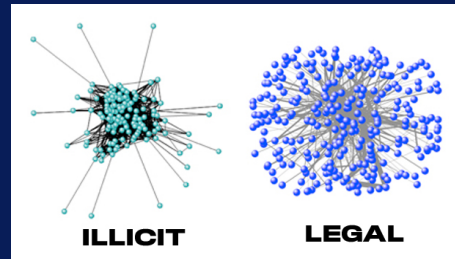
- Sender, receiver, date... **Well-structured**
- Body of the email **Text**
- Attachments **Multi-media**
- Who-sends-to-whom **Network**
- A current email cannot reference a past email **Semantics**



# Variety within a Type

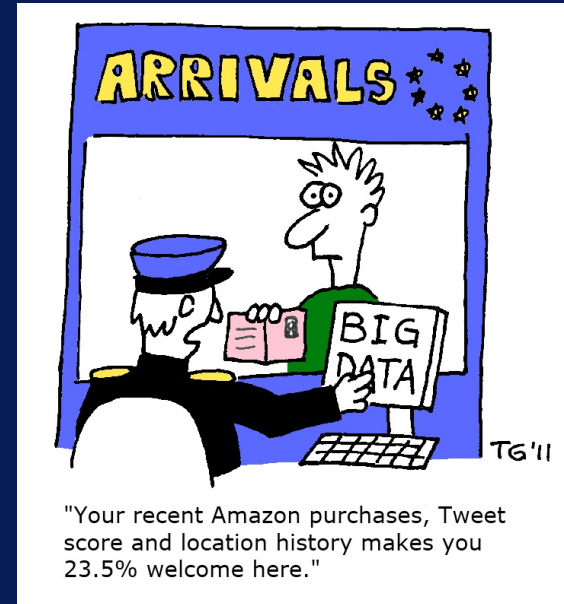
- **Think of an email collection**

- Sender, receiver, date... **Well-structured**
- Body of the email **Text**
- Attachments **Multi-media**
- Who-sends-to-whom **Network**
- A current email cannot reference a past email **Semantics**
- Real-time? **Availability**



# Scalability Issues

- **Impact of data variety**
  - Harder to ingest
  - Difficult to create common storage
  - Difficult compare and match data across variety
  - Difficult to integrate
  - Management and policy challenges



**More Details in Course 2**