# 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


- **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 | NLMGJZAKBRD |
| Chevrolet | Malibu | 27129 | 36254.0 | OIPFUIENLEHSX |
| Chevrolet | Malibu | 28846 | 77162.0 | WRCOOFREZLL |
| Chevrolet | Malibu | 46165 | 60590.0 | HUFTTHQHSFJF |
| Chevrolet | Malibu | 18263 | 37790.0 | JLMHNAFSHVD |

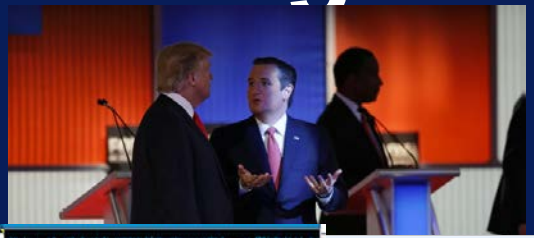
# Today, Data are more heterogeneous

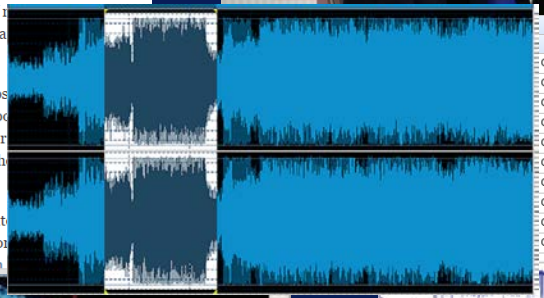
The battle for the Republican nomination appeared more splintered than ever between two halves of a bitterly divided party as several candidates scrambled Friday to consolidate the support of more moderate conservatives a day after a raucous debate.

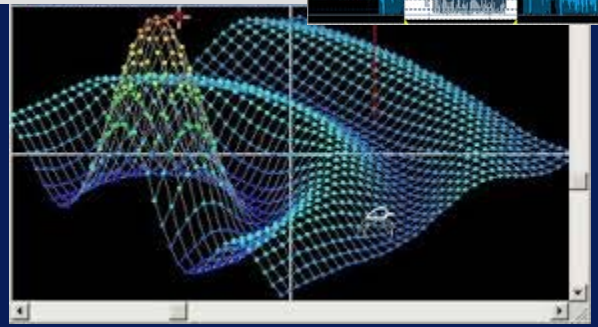
With Donald J. Trump and Senator Ted Cruz finally now engaged in an open feud for the most disillusioned voters, Senator Marco Rubio of Florida, Gov. Chris Christie of New Jersey and Jeb Bush, the former Florida governor, were battling to win over a group of
Republicans who are showing little sign of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

This fracture was most vividly apparent in New Hamps
Mr. Bush and Mr. Rubio campaigned on Friday, and po
is emerging as the obvious alternative to Mr. Trump or
candidates that many Republicans fear would doom th
general election.

Mr. Bush sought to highlight his image as the candidat
seasoned, sober-minded wing on Friday with the endor
rival in the presidential race, Senator Lindsey Graham

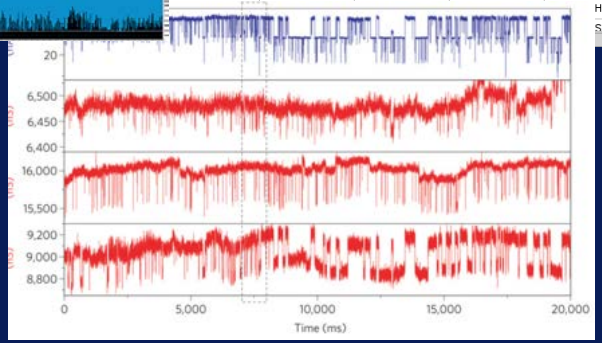


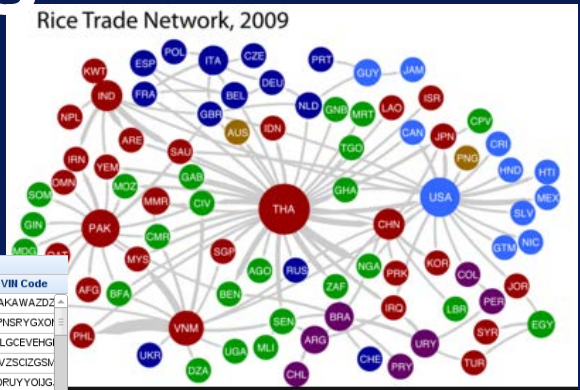




Cars marketplace

| vendor | Model | Price | Mileage | VIN Code |
|---|---|---|---|---|
| Chevrolet | Corvette | 17226 | 25965.0 | ILLAKAWAZDZ |
| Chevrolet | Corvette | 34229 | 46429.0 | RCPNSRYGXOI |
| Chevrolet | Corvette | 27982 | 50209.0 | NWLGCEVEHG |
| 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 | NLMGJZAKBRD |
| Chevrolet | Malibu | 27129 | 36254.0 | OIPFUIENLEHSX |
| Chevrolet | Malibu | 28846 | 77162.0 | WRCOOFREZLI |



# Axes of Data Variety

**Structural Variety – formats and models**

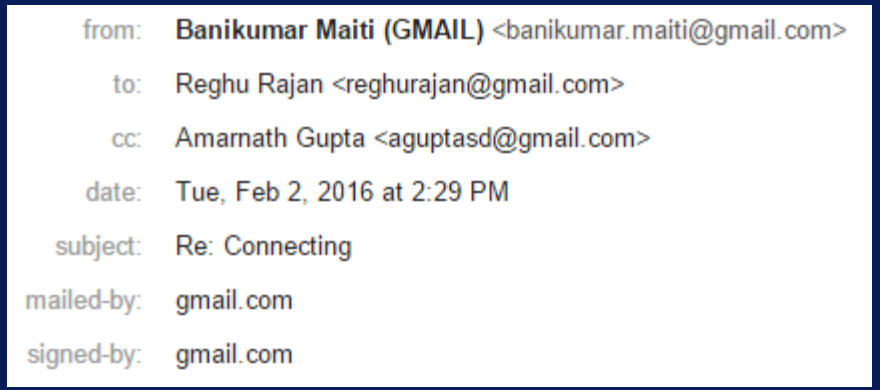
**Semantic Variety – how to interpret and operate on data**

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

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

# Variety within a Type

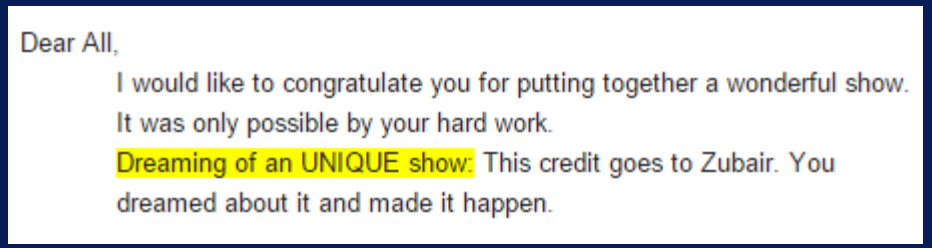
- **Think of an email collection**
  - Table-like part



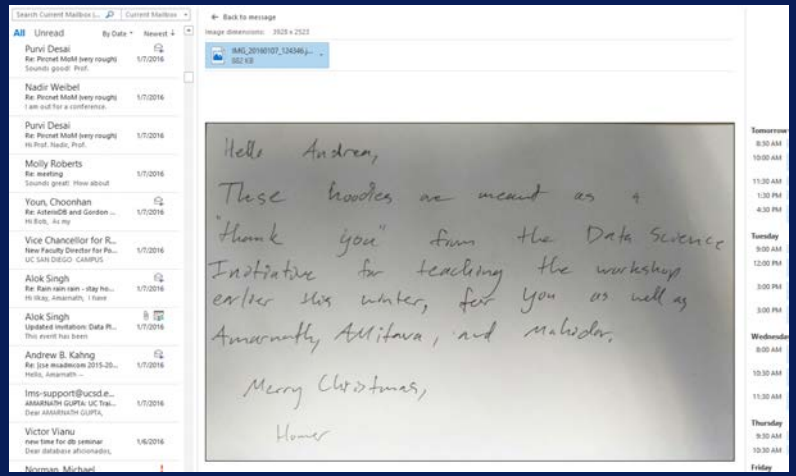
| | |
|---|---|
| from: | **Banikumar Maiti (GMAIL)** <banikumar.maiti@gmail.com> |
| to: | Reghu Rajan <reghurajan@gmail.com> |
| cc: | Amarnath Gupta <aguptasd@gmail.com> |
| date: | Tue, Feb 2, 2016 at 2:29 PM |
| subject: | Re: Connecting |
| mailed-by: | gmail.com |
| signed-by: | gmail.com |

# Variety within a Type

- **Think of an email collection**
  - Sender, receiver, date… Well-structured
  - Unstructured Text

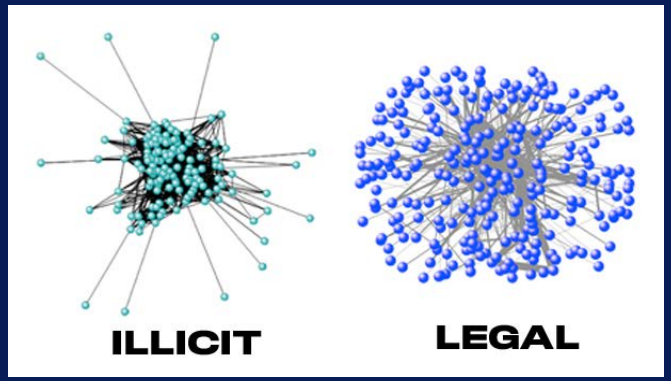


# Variety within a Type

- **Think of an email collection**
  - Sender, receiver, date… Well-structured
  - Body of the email Text
  - Media 

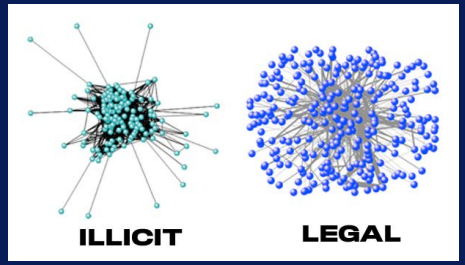
# Variety within a Type

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



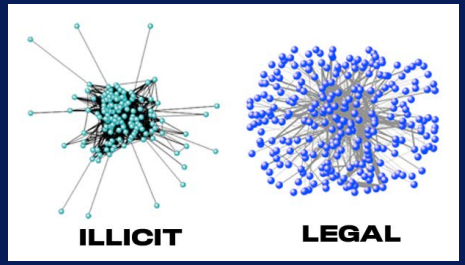
# Variety within a Type

- **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



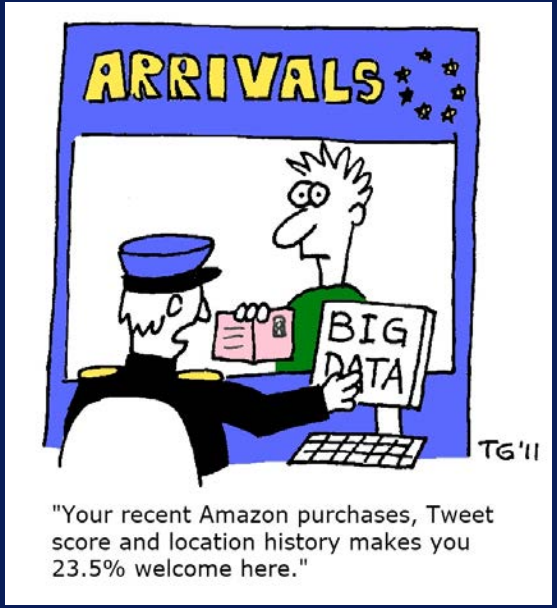
# Variety within a Type

- **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