Music Videos and Gastronomification for Big Data Analysis

Thomas Levine & Brian Abelson
csv soundsytem

http://strataconf.com/strata2014/public/schedule/detail/31767
https://github.com/tlevine/gastronomification-big-data-talk

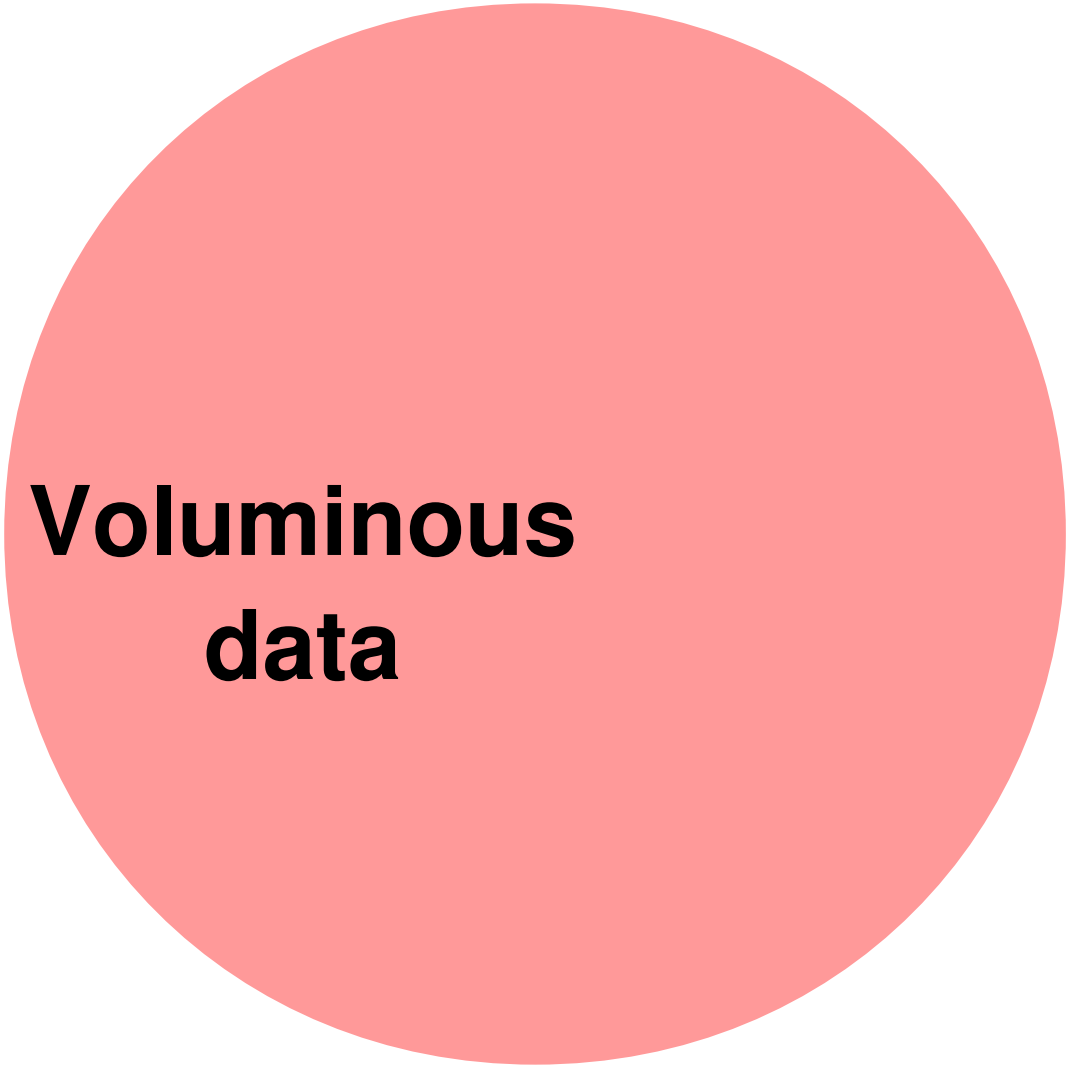
CSVSOUNDSYSTEM.COM

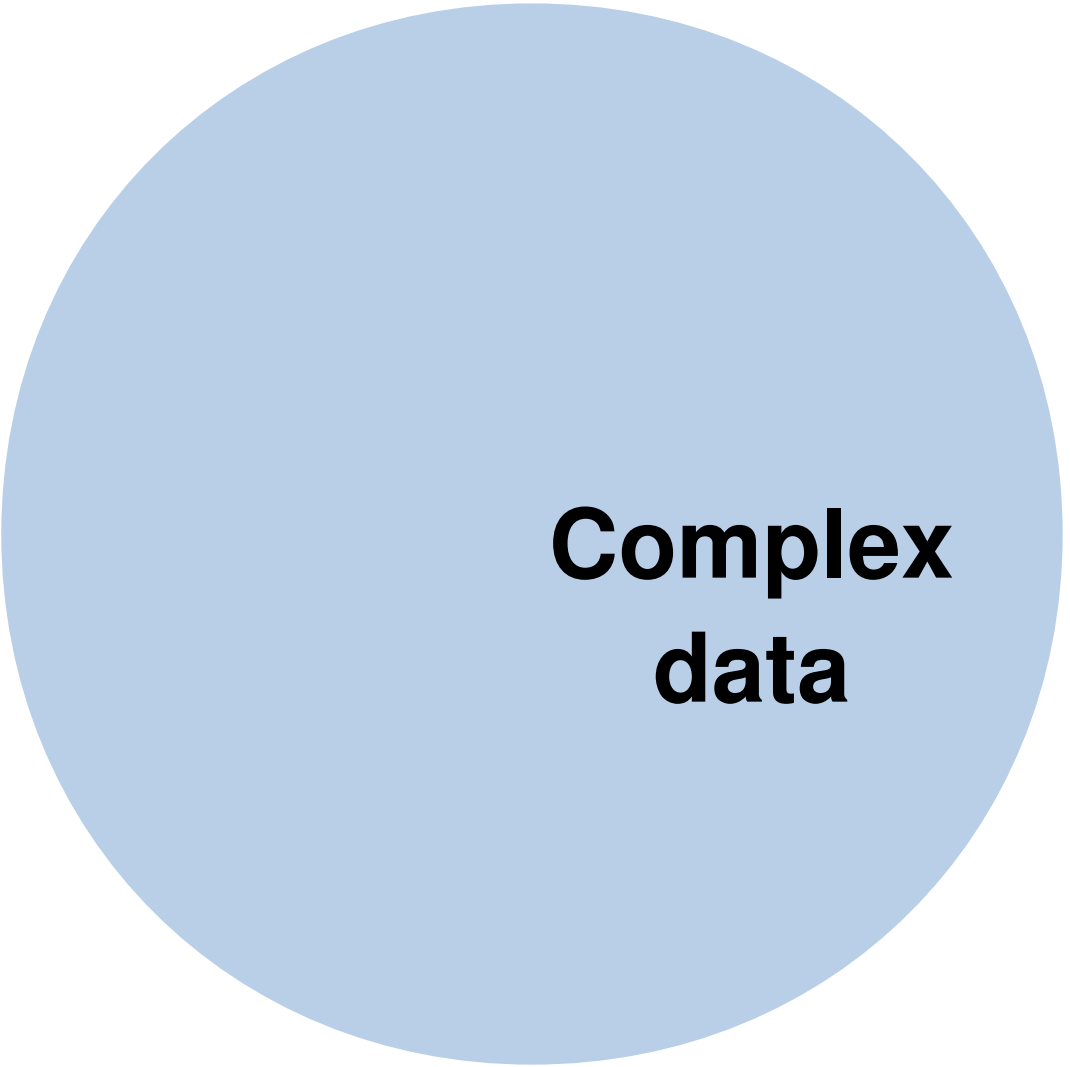
csv soundsystem

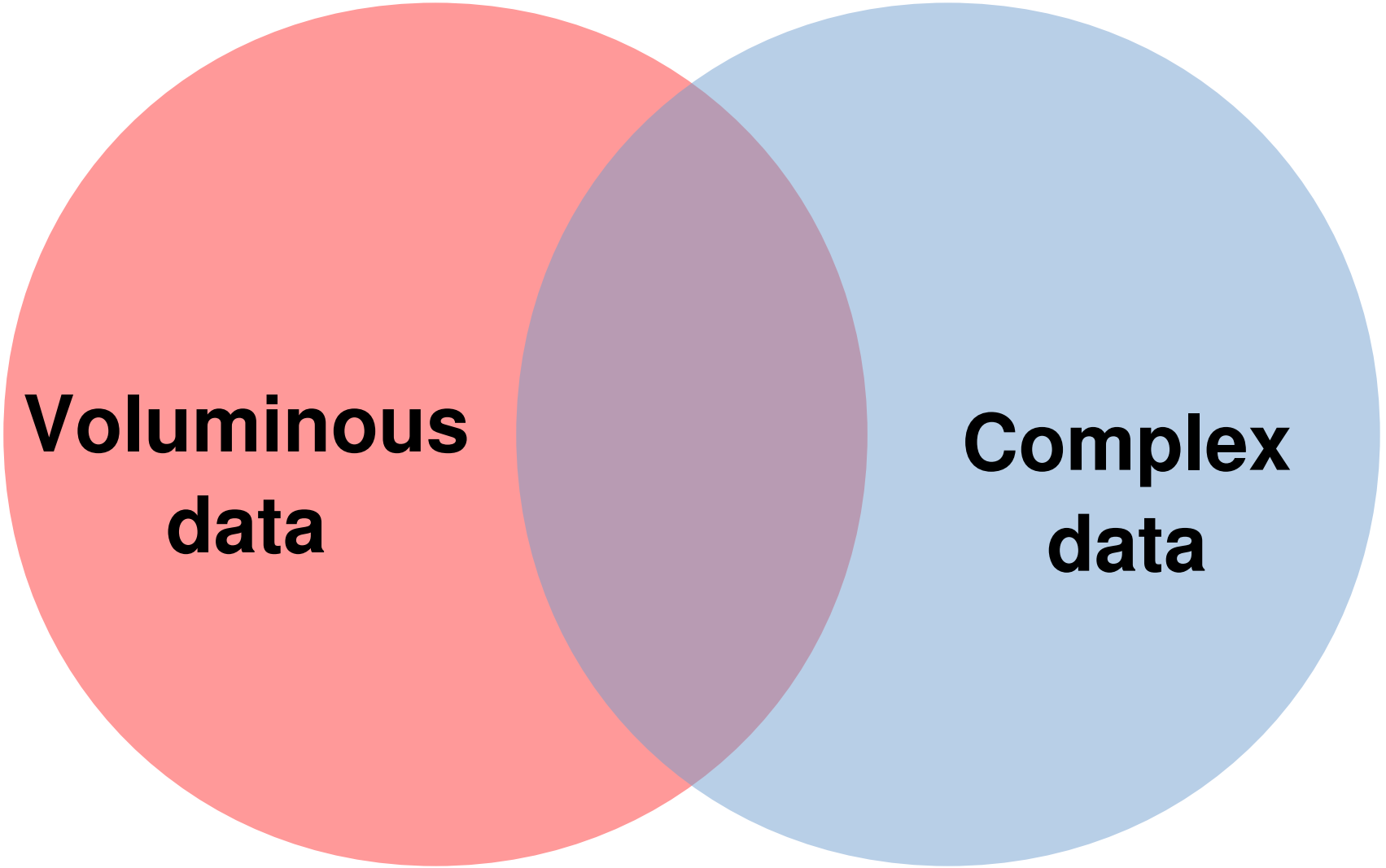
- Big data consulting firm based in New York
- Specialties
  - Keep things simple
  - Making things that people understand
- “csv” stands for “comma-separated values”

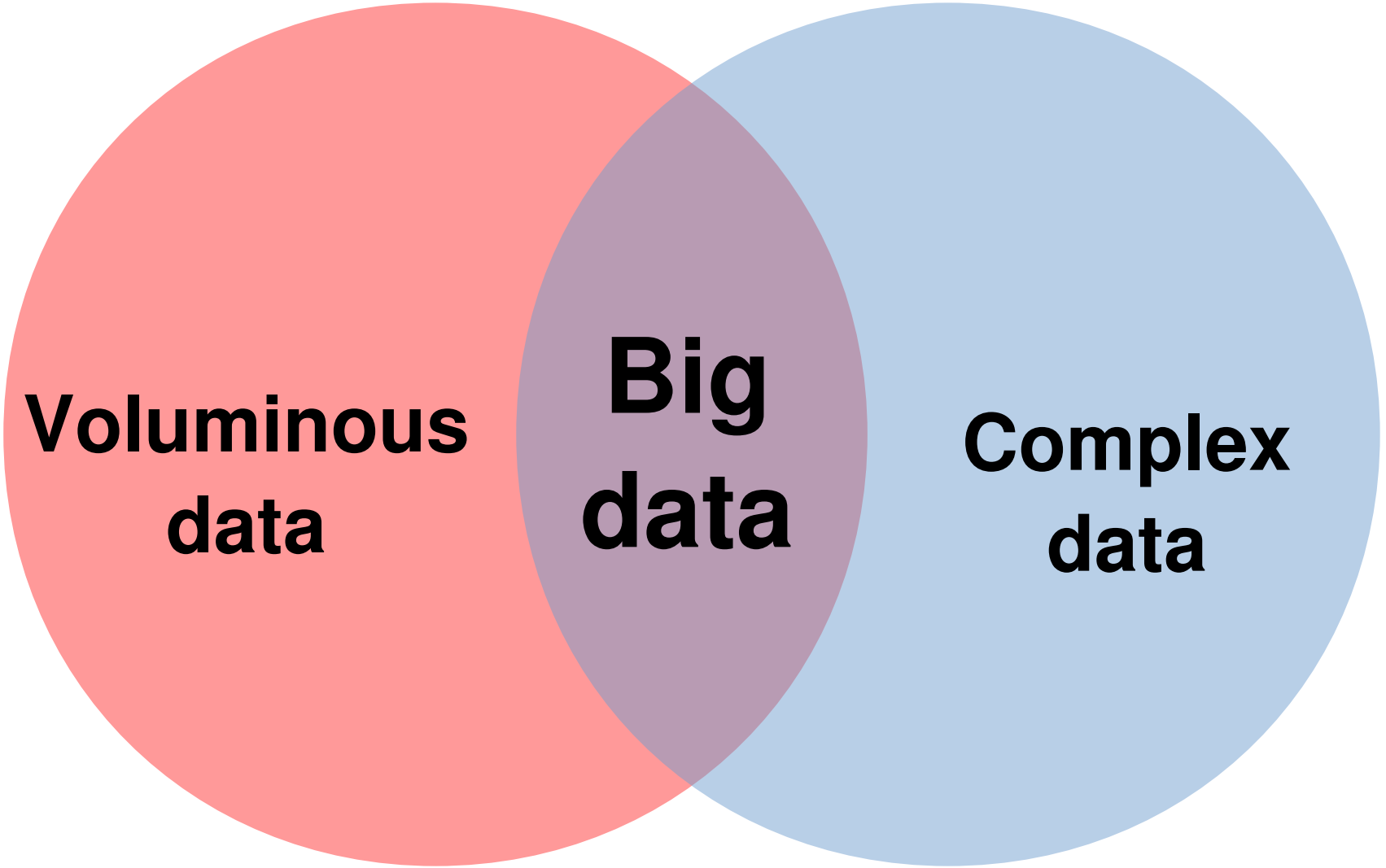
CSVSOUNDSYSTEM.COM

Big data




Complex
data





rowid	price	updated
1	750	1391226689
2	1398	
3	905	1391220478
4	815	1391230147
5	724	
6	1180	
7	50	1391229252
8	400	
9	3400	1391229855
10	400	
11	1200	
12	3995	
13	550	
14	150	
15	700	
16	650	
17	3700	1391230209
18	1290	
19	650	
20	950	
...	...	...

Voluminous spreadsheet

CSVSOUNDSYSTEM.COM

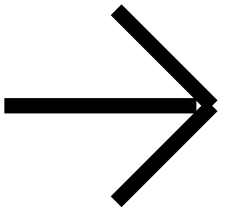
id	price	updated	long	lat	zip	...
1	750	1391226689	-97.772741	30.437796	78729	...
2	1398		-122.68216	45.52551	97209	...
3	905	1391220478	-76.467017	42.47831	14850	...
4	815	1391230147	-105.27003	40.003697	80302	...
5	724		-104.99784	39.745692	80204	...

Complex spreadsheet

Big spreadsheet

Aggregation

High-volume spreadsheet



Low-volume spreadsheet

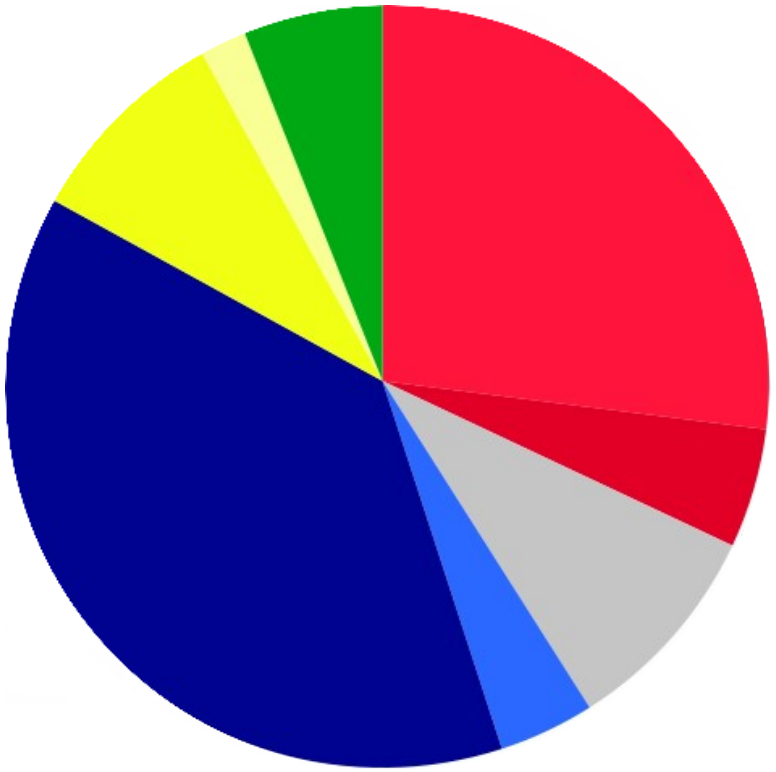
Dimensionality reduction

Complex spreadsheet

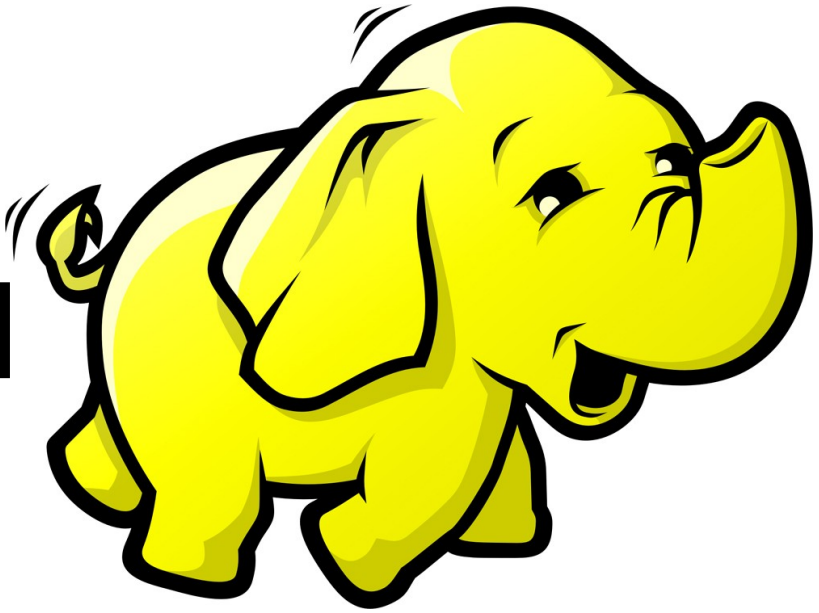
Simple spreadsheet

Visualizing big data

Graphs on Hadoop



+

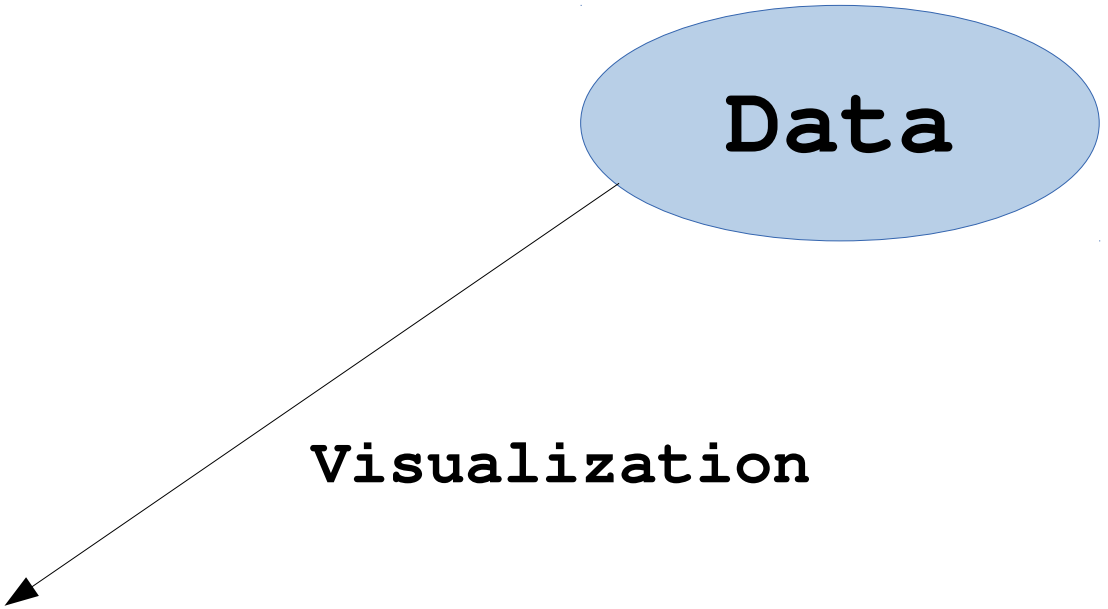


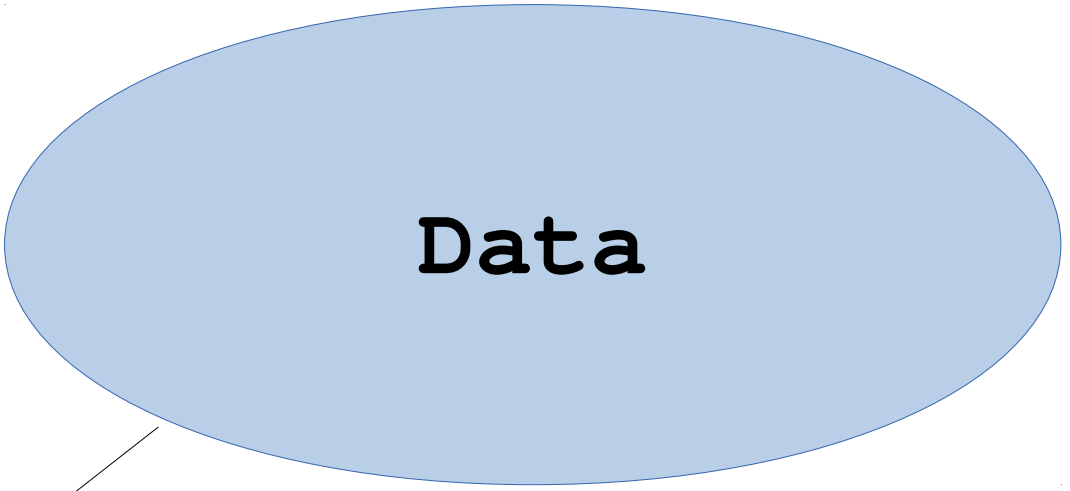
CSVSOUNDSYSTEM.COM

Talk to these people

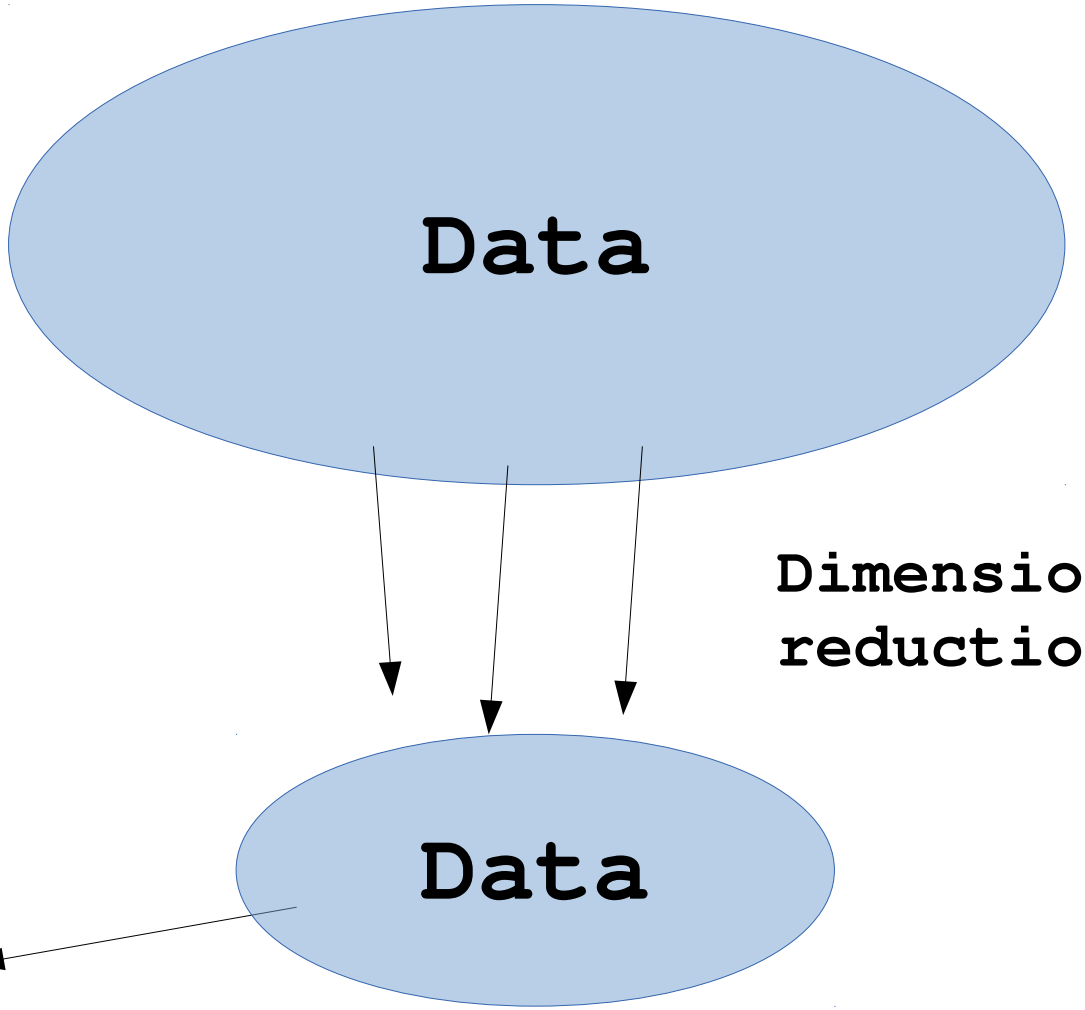
CSVSOUNDSYSTEM.COM

What about complex data?


Data
Visualization



Visualization

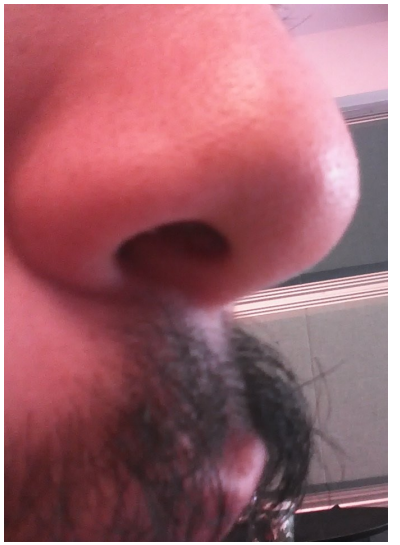


Data

Visualization







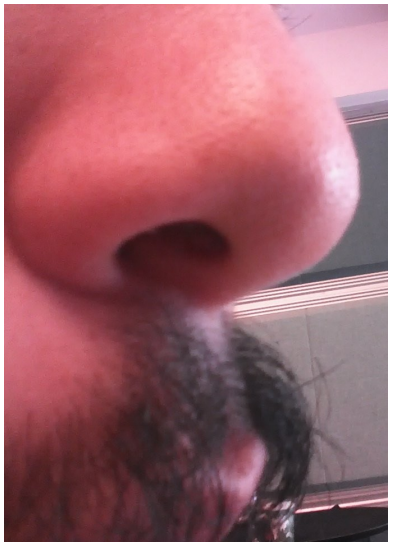
CSVSOUNDSYSTEM.COM

Data

Music videos







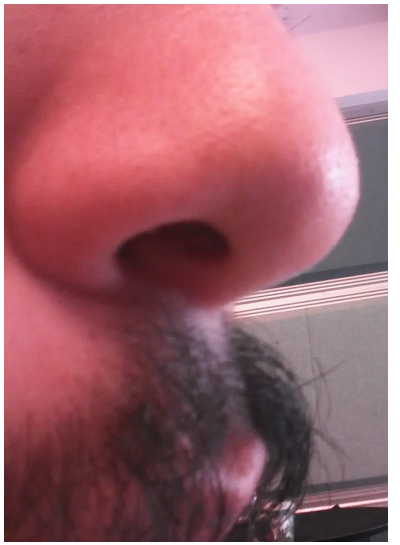
CSVSOUNDSYSTEM.COM

Data

Gastronomification

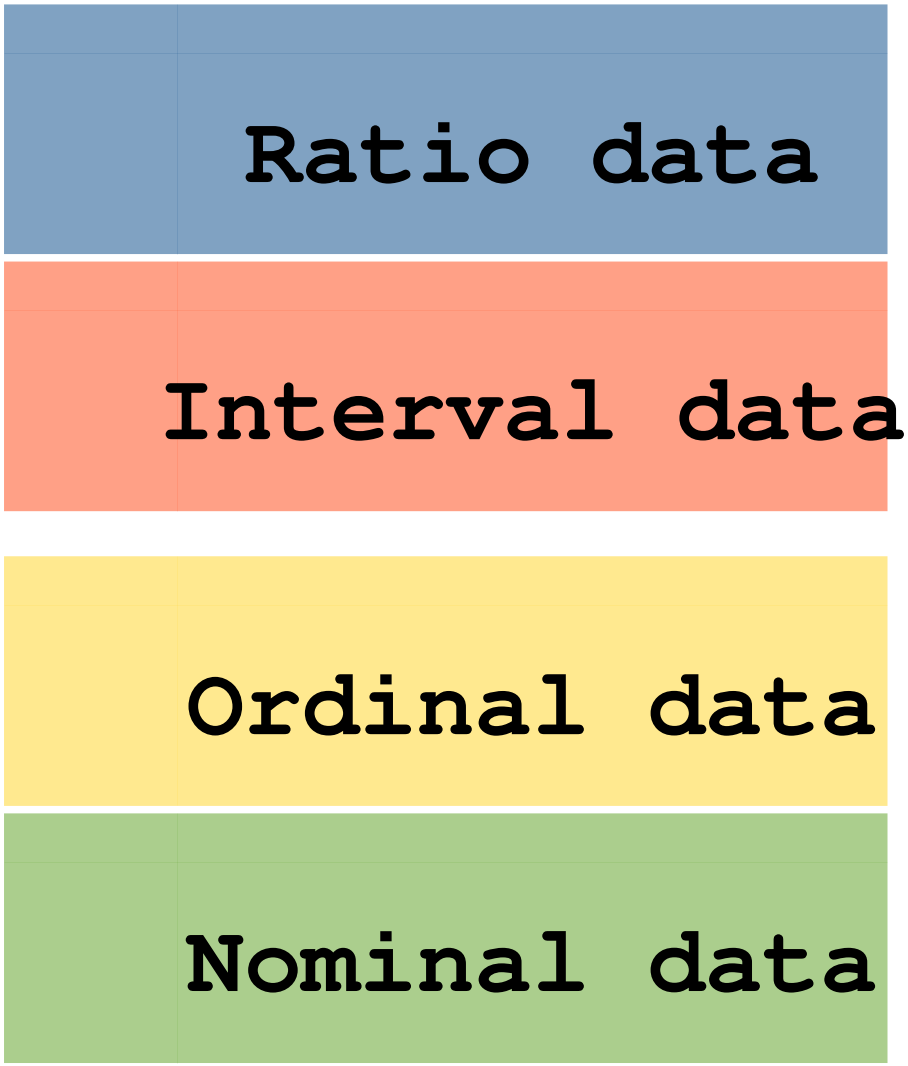






CSVSOUNDSYSTEM.COM

Tradeoffs: Scales of measurement



- Visuals, music, and food can all do each level
- Ratio and interval are more difficult with food

Other benefits
of music videos
and gastronomification

Application: Complex data for non-technical audiences

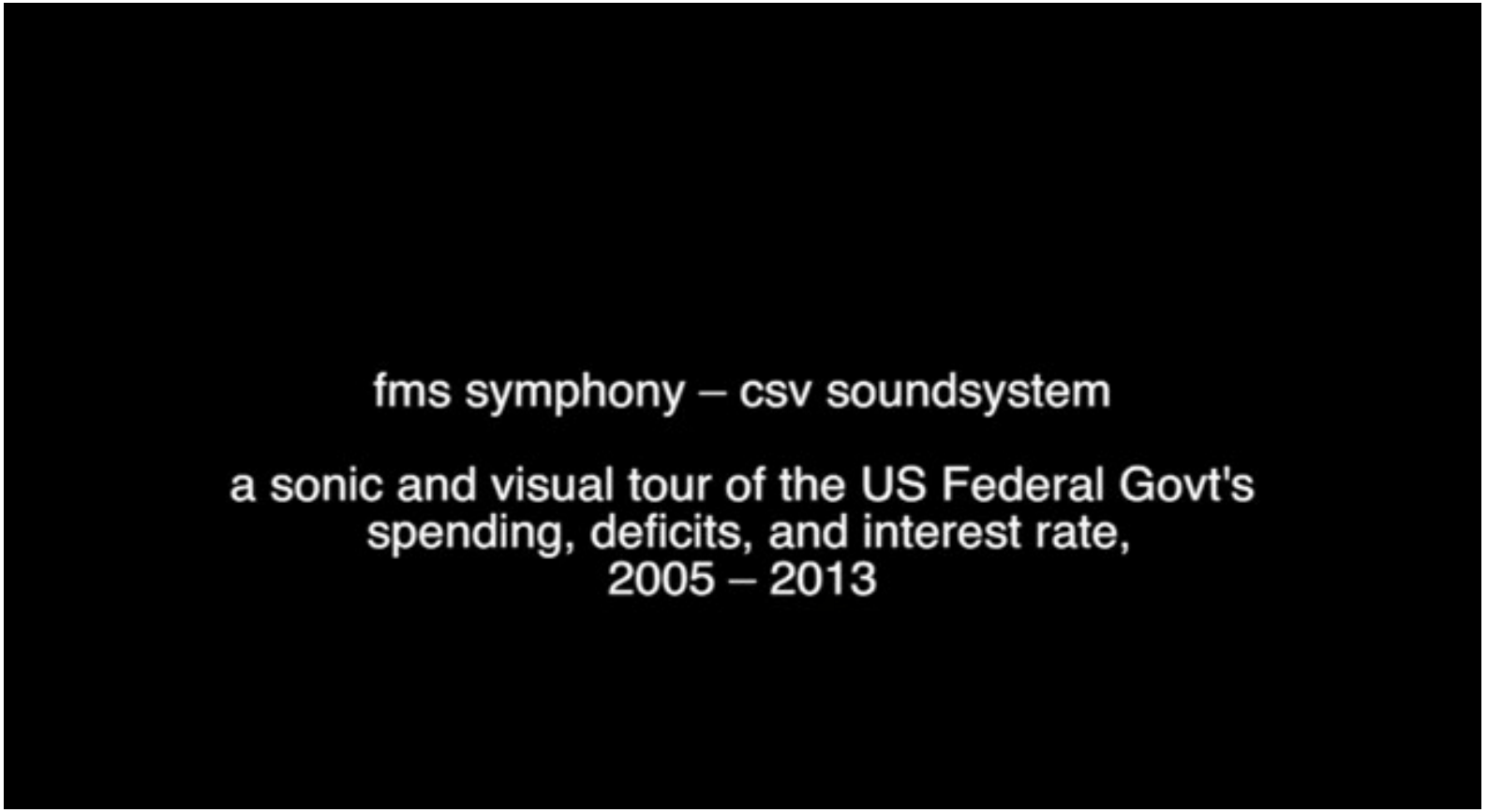
- Situation:
  - People understand bar graphs, but they are univariate.
  - Graphs can easily represent six variables, but people don't always understand.
- Problem: Graphs are too abstract
- Solution: Music and food

Engaging young audiences

http://www.youtube.com/watch?v=JwuEnyV1Cb0

CSVSOUNDSYSTEM.COM

Data-driven culture



http://fms.csvsoundsystem.com CSVSOUNDSYSTEM.COM

Data guacamole

- New York City math test scores
  - 32 districts
  - 6 grades (3rd through 8th)
  - 7 years (2006 to 2012)
- A bowl for each year
- Levels of ingredients based on relative test scores for different schools in different grades

Data guacamole

CSVSOUNDSYSTEM.COM

Census spices

http://www.backspac.es/r/si56I91DO6/census-spices-on-itp

CSVSOUNDSYSTEM.COM

Modeling mobile ad clicks

- Decisive mobile advertisement targeting (http://decisive.is/)
- Collecting data on 10% of **all** mobile ad traffic
- Bidding algorithm uses predictive modeling with machine learning
- Representing data as tacos tuning their bidding algorithms

Decisive

jalapeno = wifi
(versus 3G/4G)

cheese = clicked ad

onions = shown ad

taco = ad impression
meat → type of phone

CSVSOUNDSYSTEM.COM

Classifying healthcare eligibility

- Booz Allen Hamilton
- System to determine healthcare eligibility of people
- Dataset:
  - Each record is a person.
  - Most features are dates.
- Using sheetmusic (our spreadsheet-based small data solution) to detect incorrect classifications

File Edit View Insert Format Data Tools Help All changes saved in Drive

$ % 123 Arial 10

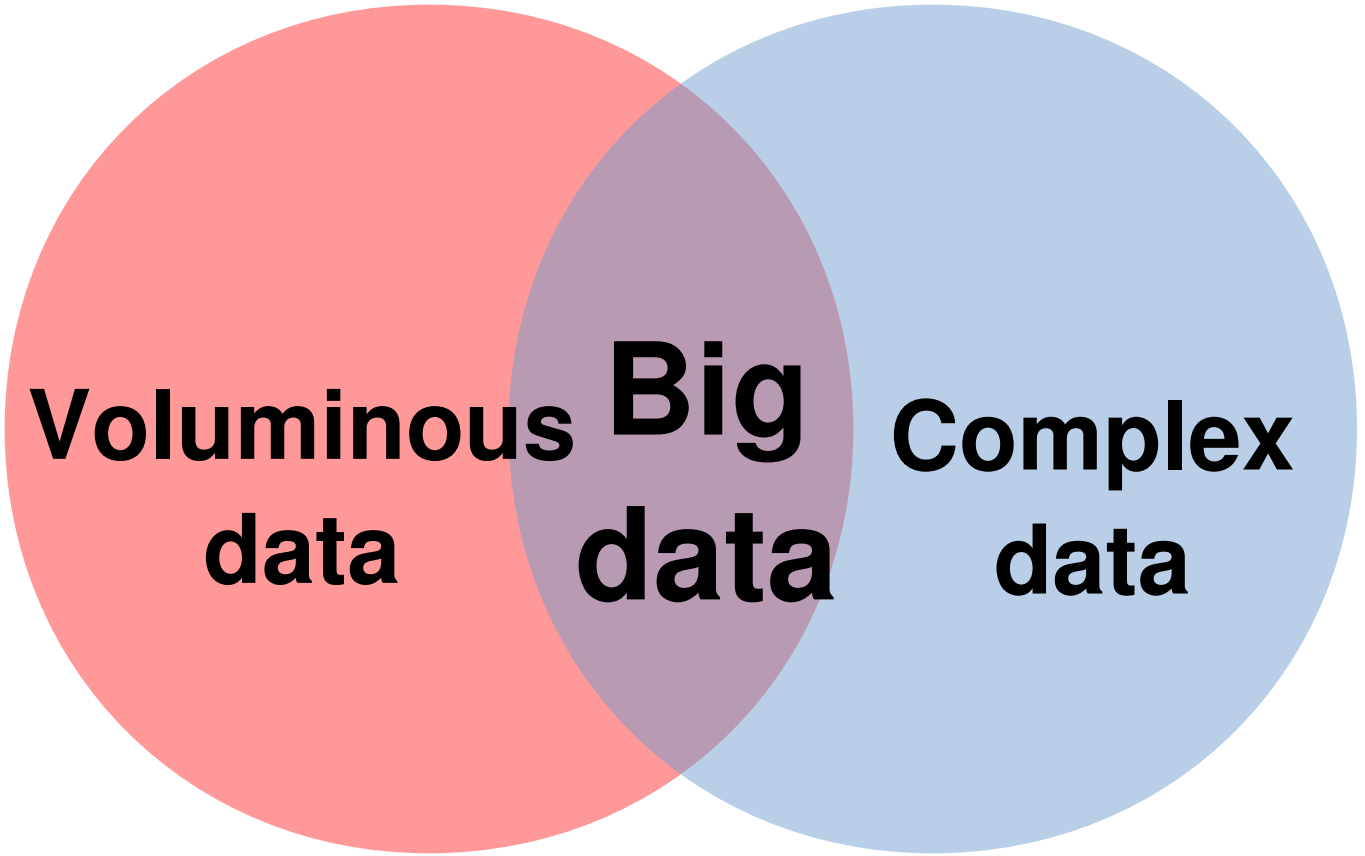
	A	B	C	D	E	F
1	**cuteness**	**cat_hair_type**	**cat_tail_length**	**cat_ear_width**	**cat_meow_decibels**	**cat_weight**
2	0.79921216884115	longhair	1.0504562415272	0.060887619406429	74.0591288420183	10.7682379353596
3	0.449292777912855	shorthair	2.05255556078805	0.810843384552961	48.7371001102525	3.61930366790284
4	0.369931189869768	longhair	8.3715850291377	0.73023894716227	65.6080486039046	17.2335666080236
5	0.039914117982996	shorthair	1.33783723716376	0.243113037146372	7.5734157914905	4.89638860329013
6	0.906357615422407	longhair	2.68739339652454	0.740782919768557	16.0680858087677	4.9266119269805
7	0.891991019020462	shorthair	7.95922287437837	0.483596305789977	45.8696997155454	11.6081635271342
8	0.847054179794387	longhair	1.1162955165845	0.638805654010579	82.1251907723276	8.78501267610865
9	0.302896224188514	shorthair	6.19115178474759	0.963754647835619	35.3828510064236	6.91582715457034
10	0.981119762093687	longhair	5.20514756889604	0.229077012966503	50.9977014038928	15.926573240936
11	0.95587741618731	shorthair	2.25953931596624	0.83356099757637	10.4264203731724	17.1725262089306
12	0.670763905271802	longhair	0.338708384644045	0.058550272635554	34.5957328991546	12.0806833107508
13	0.149959318761951	shorthair	3.34928693953843	0.800216337166951	5.54817386338504	14.161710029085
14	0.59653202400603	longhair	4.0133061154205	0.422136655945069	87.8528811081528	17.1630171809714
15	0.784838011389864	shorthair	9.4072910546028	0.265524094049353	9.36170378400639	14.3791100502533
16	0.618866660166356	longhair	1.30632786501035	0.010398722238786	49.4664889528669	7.41221099342842
17	0.604367196459857	shorthair	6.49739262179848	0.645439864919561	84.0136106997407	12.2705120235365
18	0.166513187768038	longhair	8.45329154663146	0.273309856374763	23.2735664853299	11.7233773607539
19	0.891098068829123	shorthair	6.5549005188194	0.974785053928293	75.5722439371759	2.24619336907836
20	0.4447122598485	longhair	8.47668351803853	0.464383316991025	37.534382508912	17.2722088644337
21	0.215547761699667	shorthair	9.78042030027565	0.165299010819753	95.2189010408716	3.35762232917558
22	0.498115200111461	longhair	2.750505916745	0.050989519236683	14.5453235487478	7.52304758844604
23	0.513567267648586	shorthair	5.38892235052702	0.124426668965539	24.8393638519044	19.1216246237598
24	0.355728998558687	longhair	6.46792206693301	0.70318908308532	39.1202880262994	17.4773979102788
25	0.151369356433969	shorthair	8.8047504075551	0.086941205410029	54.0977469784473	17.4281771619662
26	0.765643252973626	longhair	4.72408889503469	0.706667346813952	49.2412533461476	18.2060070362552
27	0.370339242618382	shorthair	8.01229302652481	0.146255388913134	95.2823854527991	17.0241541297797
28	0.33038055375631	longhair	5.63103864701098	0.265248907912711	61.0722435645927	10.4155429618451

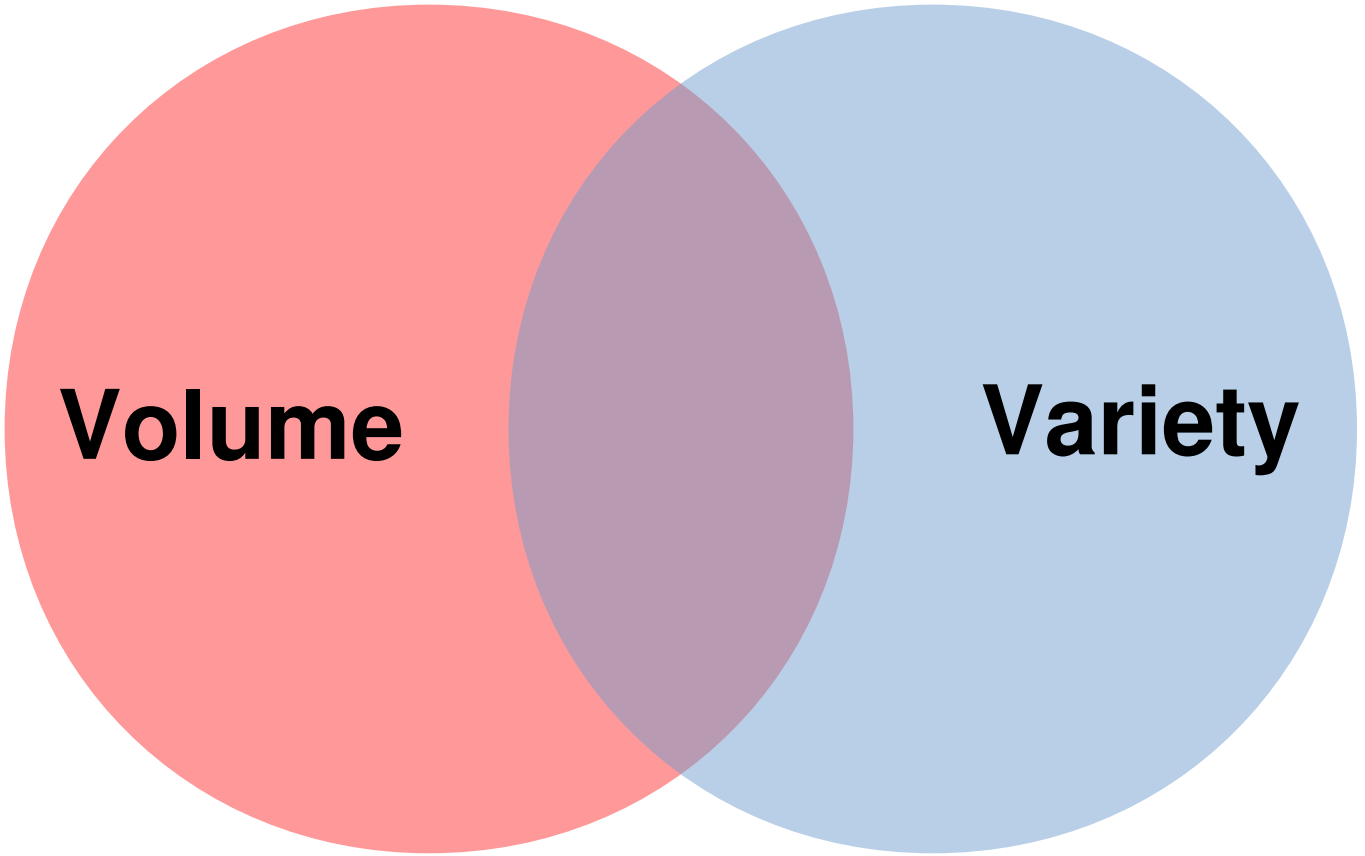
These are not Booz Allen's data; we can't show you those.

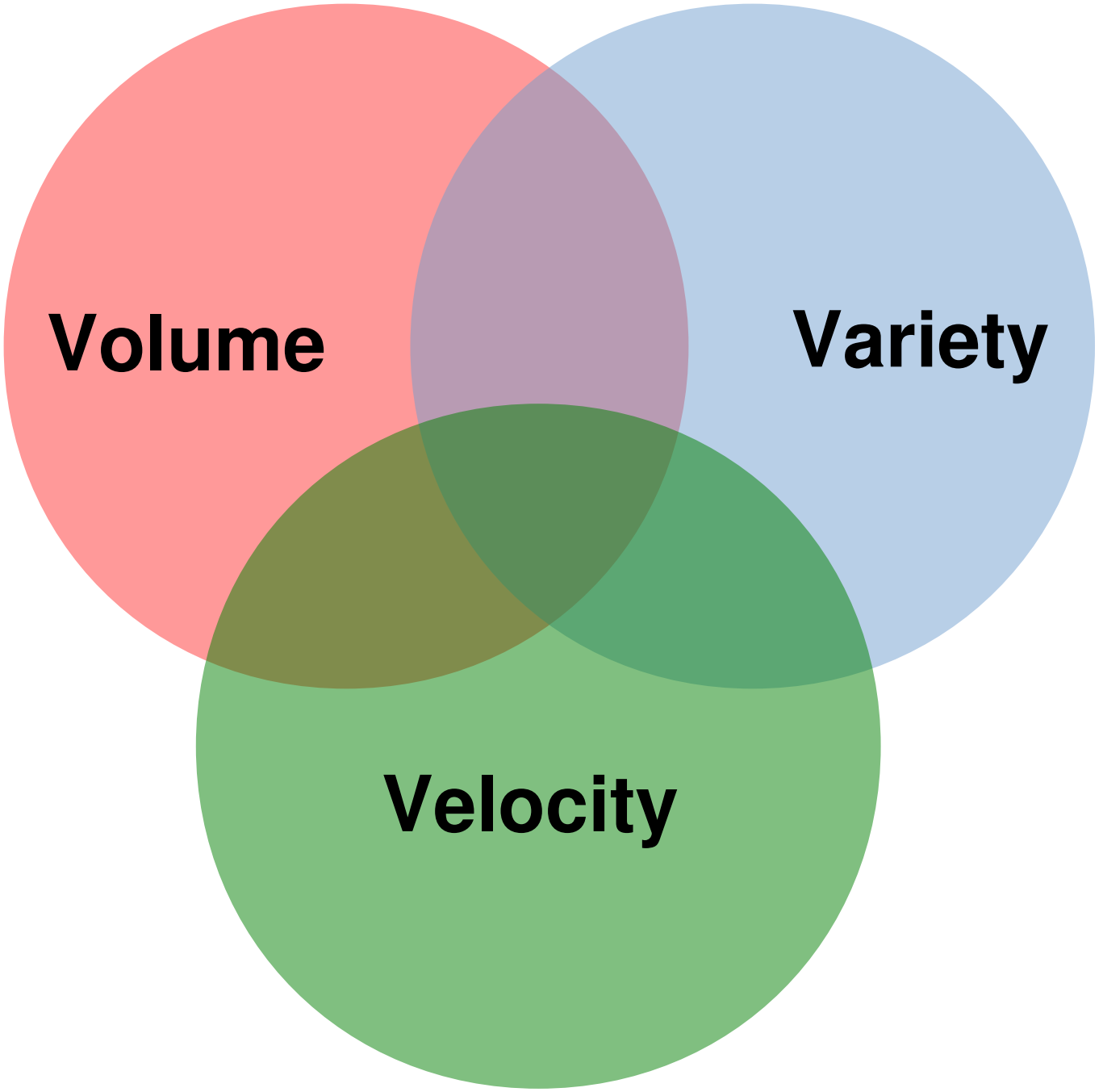
CSVSOUNDSYSTEM.COM

What about
realtime big data
gastronomification?

Three V's of Big Data







Hot Karot (www.hotkarot.cz)

- OpenSauce technology
- Connects to various data sources
- Realtime

eat your tweet #bigcleancz

Interesting project for mapping the money in the world - http://t.co/F2ebjBOz #bigcleancz

Michal Kubáň před necelou minutou ago

open foundings lightening talk http://t.co/yffmV18V #bigcleancz

Alix Guillard 5 minutami ago

#bigcleancz participants: Share your real life #opendata experiences with us at http://t.co/qQBxR0V0 cc @bigcleancz

ePSIplatform 13 minutami ago

#bigcleancz lightning talks' links: http://t.co/Qa2gcfgt http://t.co/gM7HjE6w http://t.co/fpdDeSKX http://t.co/Tbz0oI8e

BigClean.cz 15 minutami ago

#Scraping #tool with #RDF output is called Strigil http://t.co/NAbYpjLH #bigcleancz

Jakub Mráček 17 minutami ago

very interesting set of talks !! @BigCleanCZ

Amrapali Zaveri 19 minutami ago

Anyone fancy for other HotKarot during coffee break? Let us know! #bigcleancz

rudolf 20 minutami ago

#bigcleancz now continues with lightning talks session after "eat your tweet" lunch.

BigClean.cz 21 minutami ago

The amount of love felt for numbers is still the most fascinating thing about the open data scene. cc @jancibulka #bigcleancz

Maria Schröder 21 minutami ago

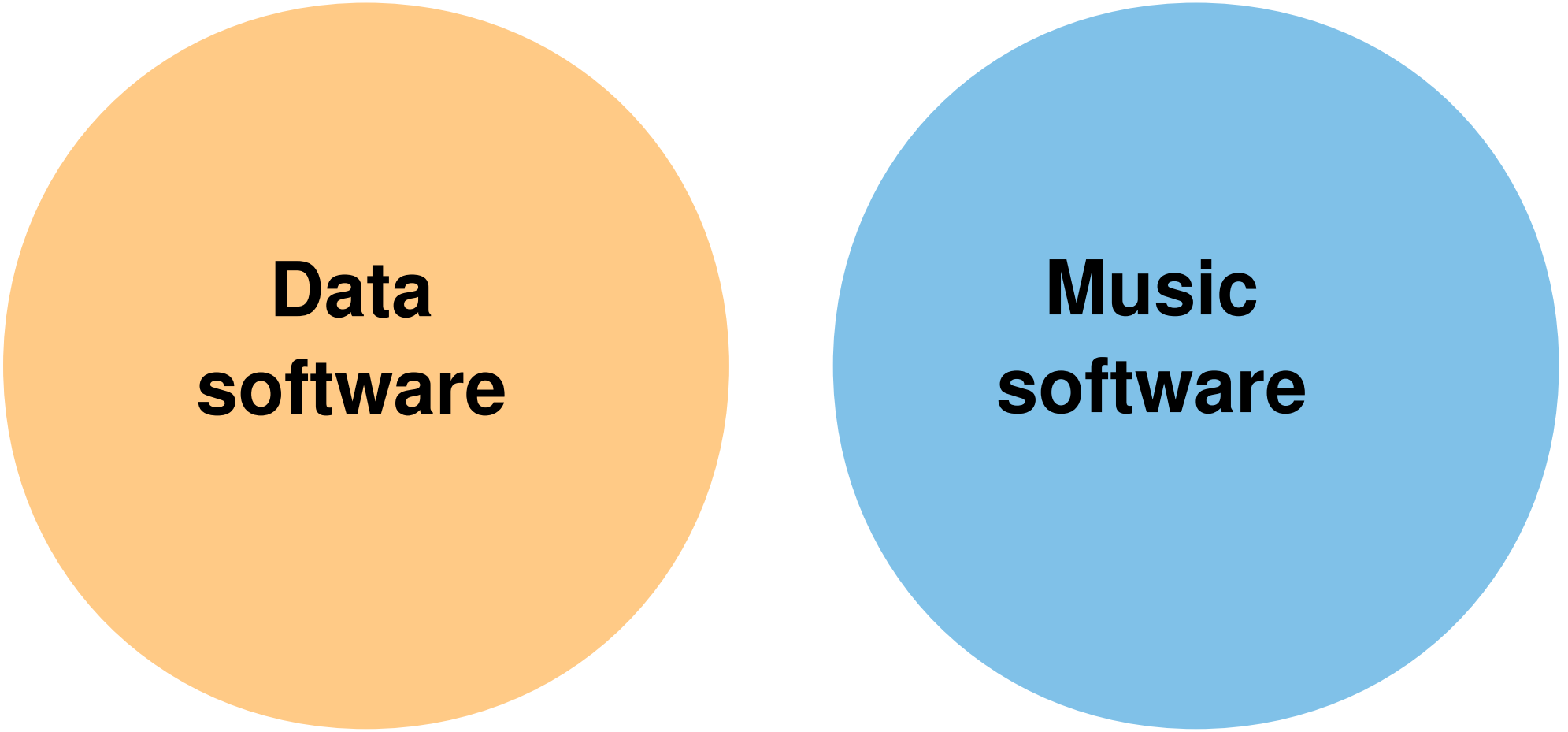
CSVSOUNDSYSTEM.COM

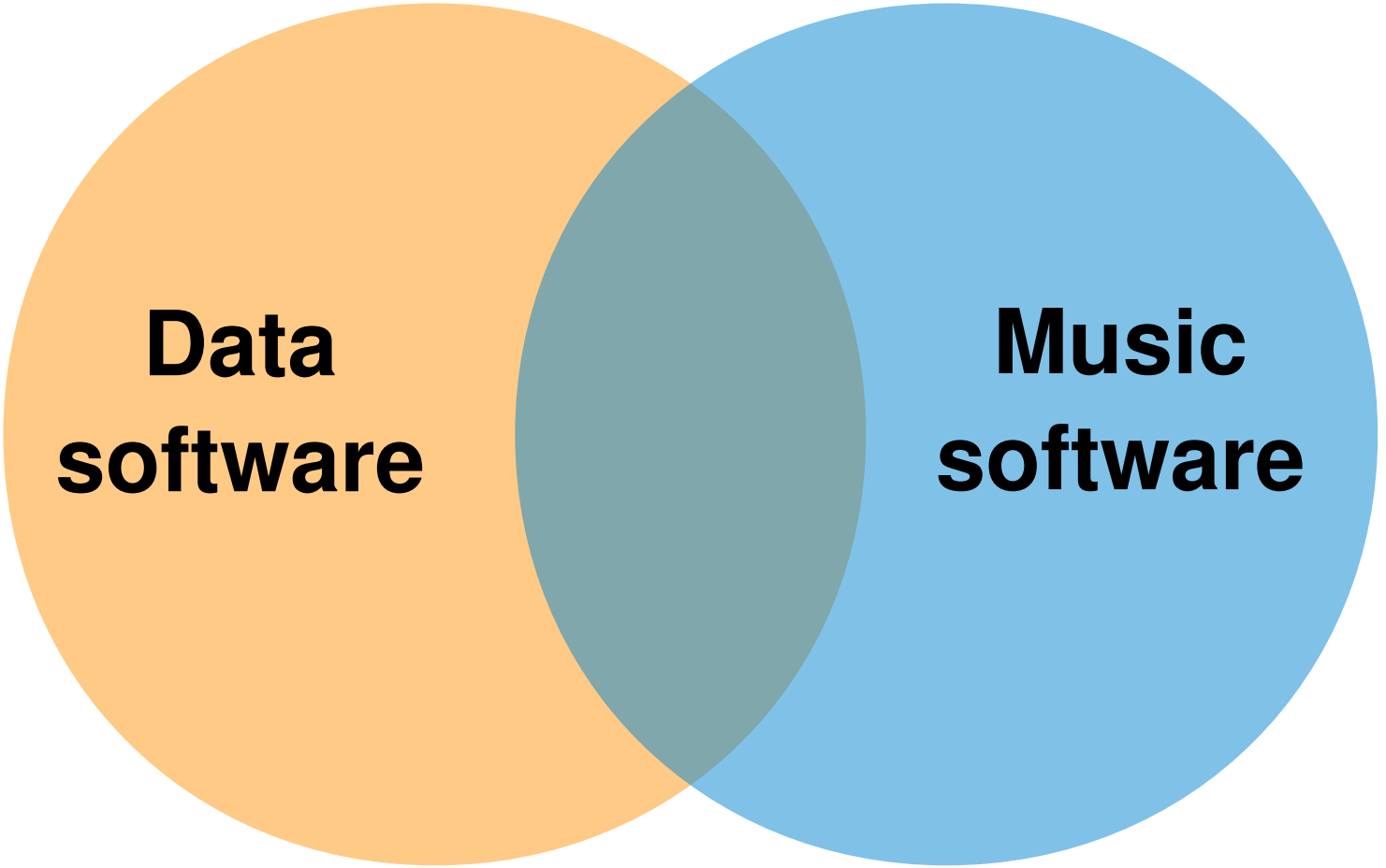
Our realtime solution

- Demo

our
open-source
libraries

ddr, ddpy, & sheetmusic


Data
software
Music
software



Merging data software and music software

- **ddr**: Music sequencer backed by R vectors and data.frames
- **ddpy**: MIDI generator backed by pandas DataFrames.
- **sheetmusic**: Web-based music composition backed by Google Spreadsheets

geom_taco

Why we wrote geom_taco

- Dependence on human experts limits gastronomification
  - OpenSauce (http://www.hotkarot.cz)
  - Census Spices
- geom_taco uses commodity infrastructure
  - Robust, scalable, inexpensive

geom_taco

- A geom for ggplot
- Non-visual aesthetics
  - Fill
  - Salsa
  - Guacamole
  - ...

End

Jobs

- Salespersons
- Java Engineers
- Test Engineers

jobs@datagastronomification.com

Thomas Levine

tlevine@datagastronomification.com

Brian Abelson

babelson@datagastronomification.com

CSVSOUNDSYSTEM.COM