Data Science – Opportunities and Challenges

United Nations Headquarters

March 25, 2019

Felix Naumann

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

The Hasso Plattner Institute in Potsdam, Germany

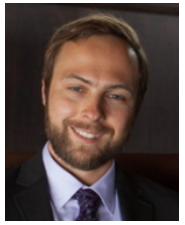
Felix Naumann
Data Science 2019

Information Systems Team

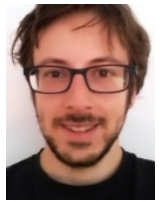
Dr. Thorsten **Papenbrock**

Diana **Stephan**

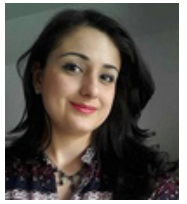
Prof. Felix **Naumann**



Dr. Ralf **Krestel**



Leon **Bornemann**



Hazar **Harmouch**

Konstantina **Lazariduo**

project **DuDe**

Duplicate Detection

Data Change

Data Fusion

project **Stratosphere**

Entity Search

Information Integration

Data Profiling

project **DataKnoller**

Web Science

Data Scrubbing

project **DataChEx**

Data as a Service

Data Cleansing

Information Quality

Text Mining

Web Data

Linked Open Data

RDF Data Mining

Dependency Detection

ETL Management

Service-Oriented Systems

project **Janus**

Entity Recognition

Opinion Mining

Data Preparation

project **Metanome**

Change Exploration

Tim **Repke**

Julian **Risch**

Michael **Loster**

John **Koumarelas**

Tobias **Bleifuß**

Nitisha **Jain**

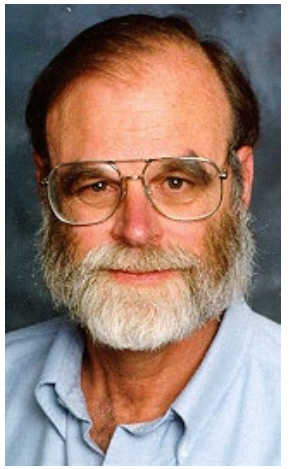
Lan **Jiang**

Gerardo **Vitagliano**

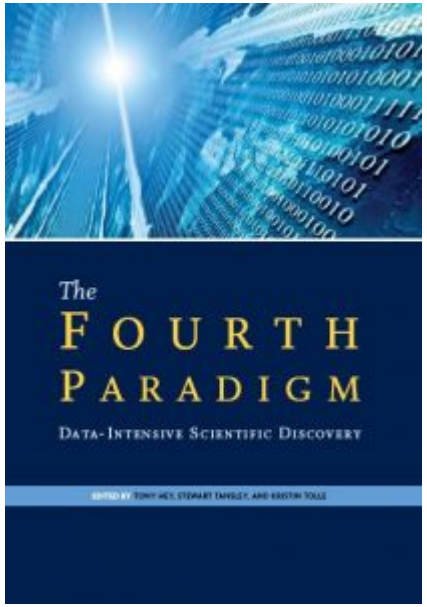
Felix Naumann
Data Science 2019

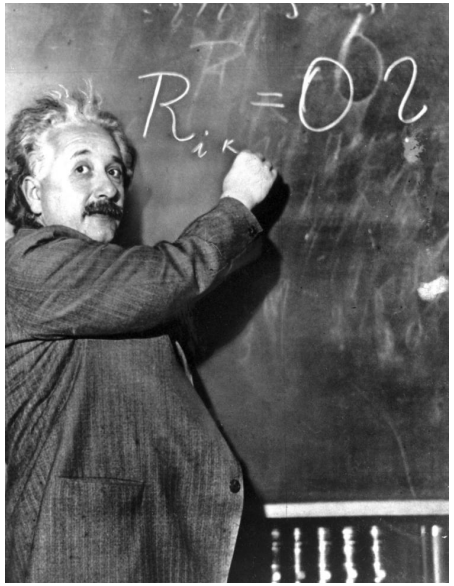
The Fourth Paradigm of Science

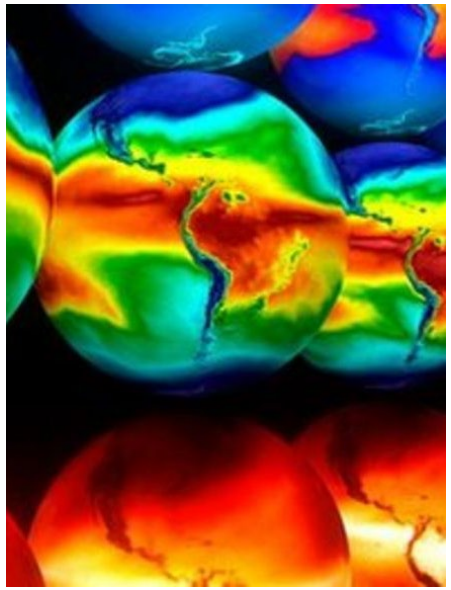
1. Empirical and experimental
2. Theoretical
3. Computational
4. Data-intensive

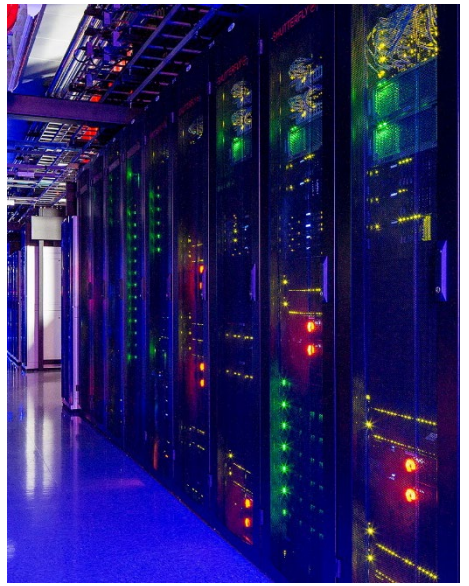


We have to do better producing tools to support the whole research cycle - from data capture and data curation to data analysis and data visualization. Jim Gray









Overview

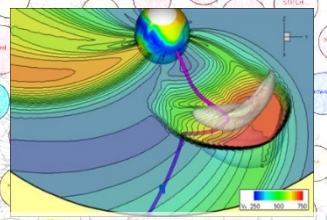
1. **Data Science**
2. Big Data
3. Data Profiling
4. Data Preparation
5. Data Cleaning

https://unsplash.com/photos/vGefUiWm0xI

More and more data are available to science and business

- Drivers
  - Cloud computing
  - Internet of services
  - Internet of things
  - Cyberphysical systems
- Underlying trends
  - Connectivity
  - Collaboration
  - Computer generated data

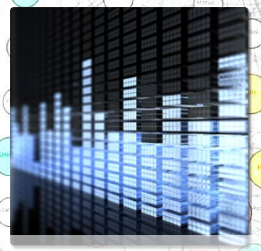
sensor data



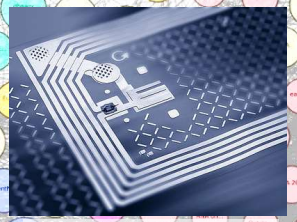
simulation data

web archives

video streams



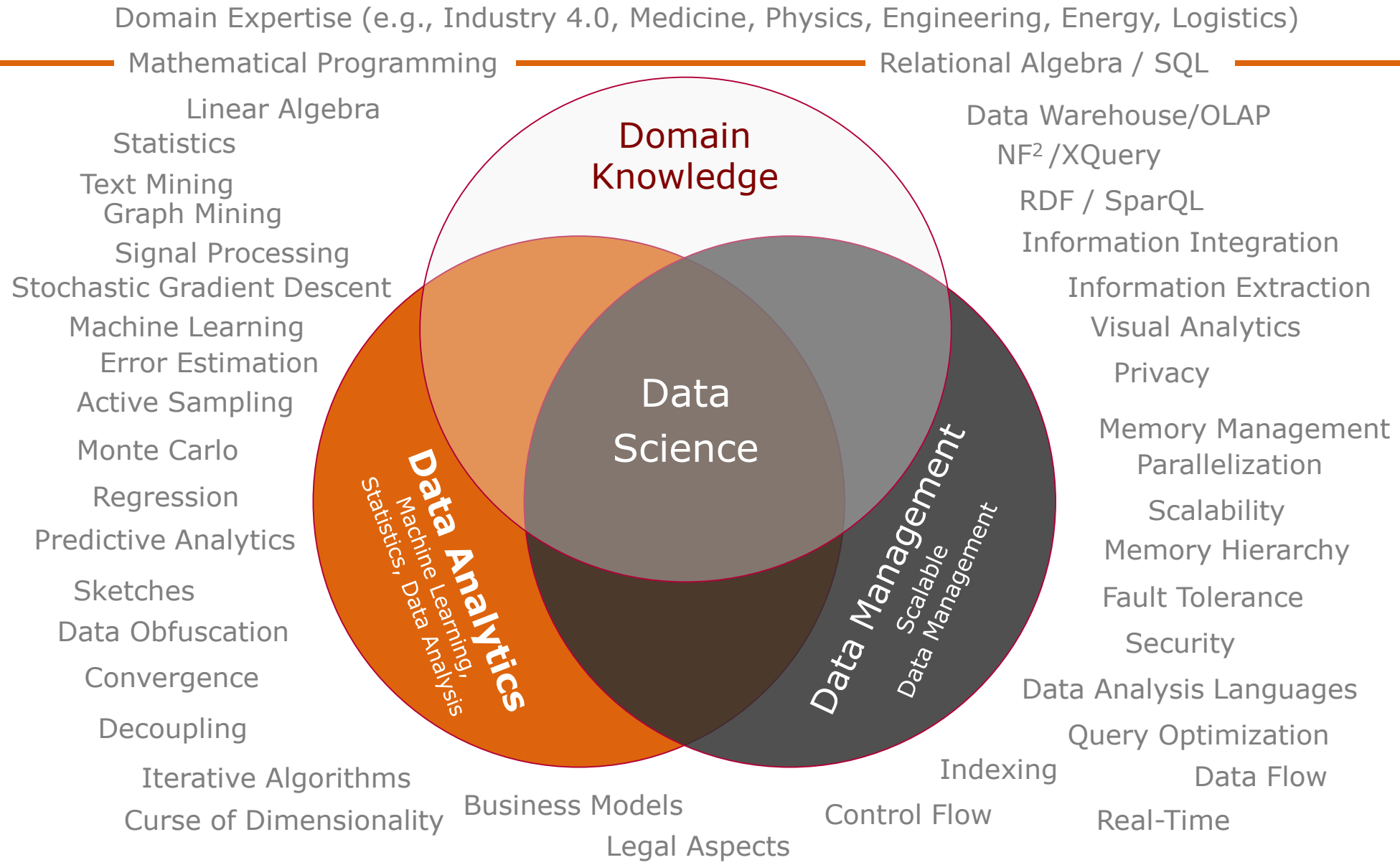
audio streams



RFID data

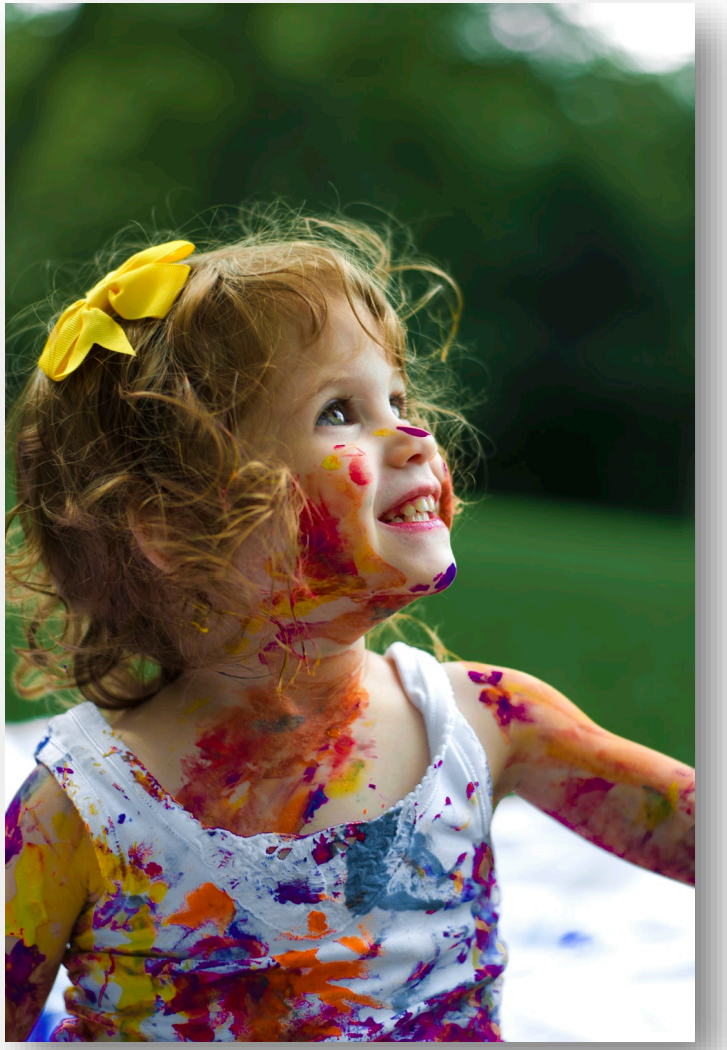
Government data

“Data Scientist” – “Jack of All Trades!



Positive Uses of Big Data

- Prediction
  - Weather, natural disaster, predictive maintenance, disease
- Optimization
  - Planning, traffic, logistics, machine efficiency, site selection
- Individualization
  - Digital health and personalized medicine, personalized learning, recommendations
- Comfort
  - Sharing, smart home, authentication (face, gait)
  - Happiness: HappyDB – a database of happy moments
  - Autonomous vehicles
- Intelligence
  - Fraud detection, translation, gaming
  - Robotics



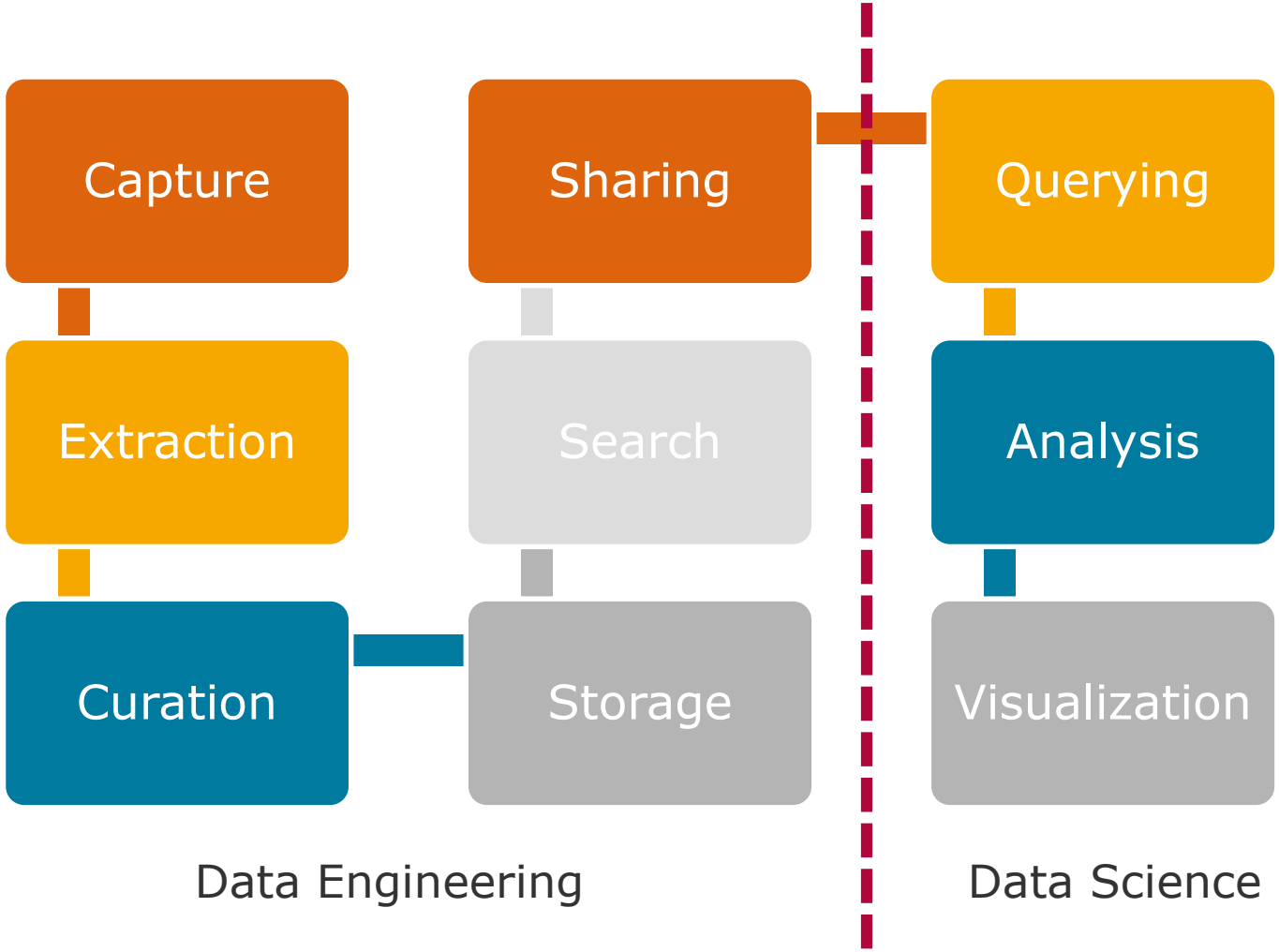
https://unsplash.com/photos/JfolIjRnveY

Questionable Uses of Big Data

- Invading lives
  - Tracking persons:
    - Direct: GPS location tracking
    - Indirect: face recognition / surveillance
  - Tracking behavior: social networks, sensors, smart homes
- Classifying individuals
  - Behavior prediction
  - Crime prediction
  - Social Scoring
- Misinformation
  - Filter bubble
  - Manipulating/inflaming opinion
- Intervention
  - Restricting free movement
  - Censorship
  - Autonomous drones

https://unsplash.com/photos/fPxOowbR6ls

Data Science Pipeline



Overview

1. Data Science
2. **Big Data**
3. Data Profiling
4. Data Preparation
5. Data Cleaning

https://unsplash.com/photos/vGefUiWm0xI

Gartner’s 3 (+ 1) V’s – Properties of Big Data

Volume
- Size of dataset

Velocity
- Speed at which data arrives and must be processed

Variety
- Different data modalities, models, schemata, semantics

Veracity
- Data quality: Correctness, completeness, consistency, up-to-dateness, etc.

Viscosity
- Integration and dataflow friction

Venue
- Different locations that require different access & extraction methods

Vocabulary
- Different language and vocabulary

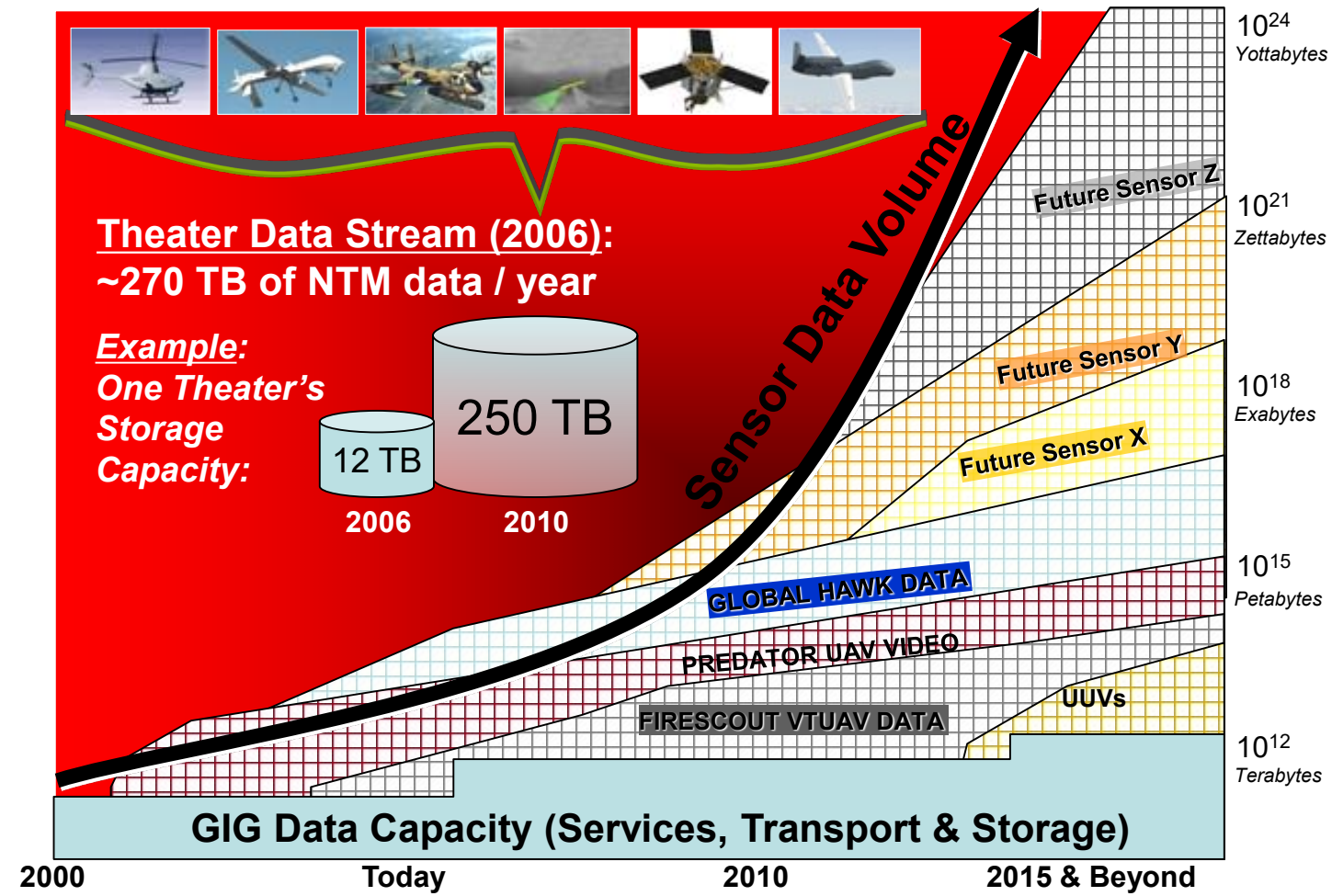
Value
- Added-value of data to organization and use-case

Virality
- Speed of dispersal among community

Variability
- Data, formats, schema, semantics change

Volatility, vagueness, validity, visualization, …

Military Projection of Sensor Data Volume (later refuted)



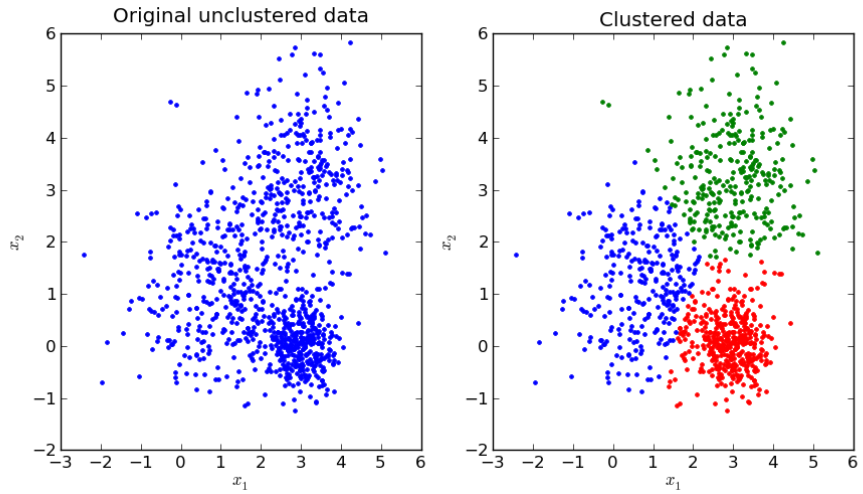
153 hard disks per person on the planet

Using 1TB drives, this would require 1 trillion (10^{12}) drives!

Bob Gourley: Thoughts on the future of Information Sharing Technology

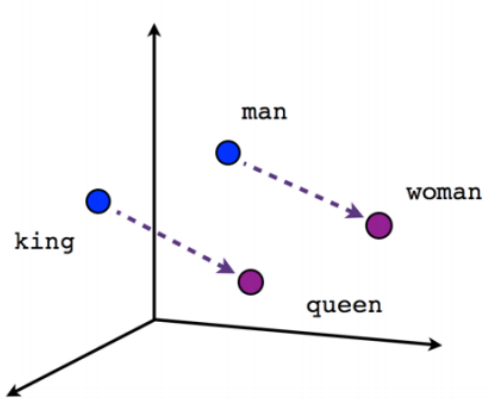
Abridged History of Big Data Analytics

- Aggregation: Calculate statistics
  - Sum of sales, average cheese consumption (per state)
- Data mining: Identify useful rules
  - 35% of all customers who bought X, also bought Y (X=beer and Y=diapers)
- Clustering: Group similar items
  - Cluster patients into 10 groups based on a similarity measure (age, weight, income …)
- Classification: Organize items into a set of known groups based on similarity
  - Assort products into categories
  - Collaborative filtering (for movies)
- Machine learning: Generalization of all of the above
  - Build a model that explains the data (for a given target dimension)
  - Apply the model to new data items to find out target dimension value

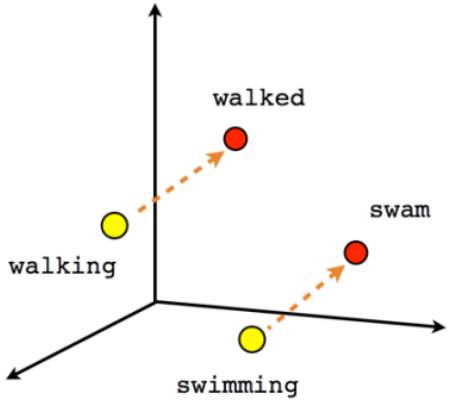


Appetite for Training Data

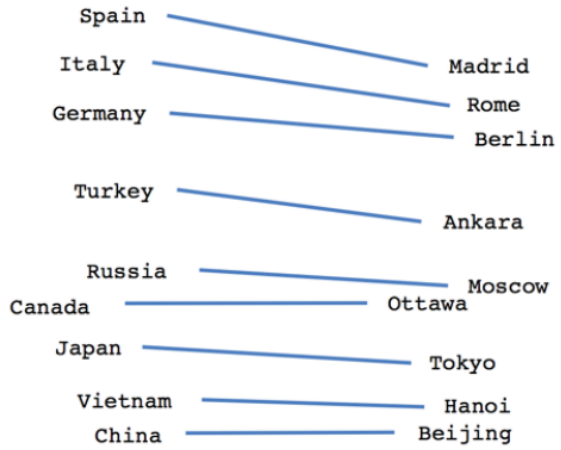
- Sophisticated models need many input dimensions
  - Few dimensions for spam filtering → Labeled spam emails
  - Tens of dimensions for intrusion detection → Annotated log entries
  - Hundreds of dimensions for user classification → Detailed user profiles
  - Thousands of dimensions to understand text → Sample texts
- … and have many model parameters.



Male-Female



Verb tense



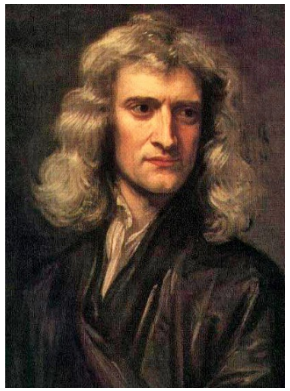
Country-Capital

Big Data = Science?

- The End of Theory: The Data Deluge Makes the Scientific Method Obsolete (Chris Anderson, Wired, 2008)
  - All models are wrong, but some are useful. (George Box)
  - All models are wrong, and increasingly you can succeed without them. (Peter Norvig)
- Before Big Data: Correlation is not causation!
- With Big Data: Who cares?
  - Traditional approach to science — hypothesize, model, test — is becoming obsolete.
  - Petabytes allow us to say: "Correlation is enough."

http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

$$F = G\frac{m_1 m_2}{r^2}$$



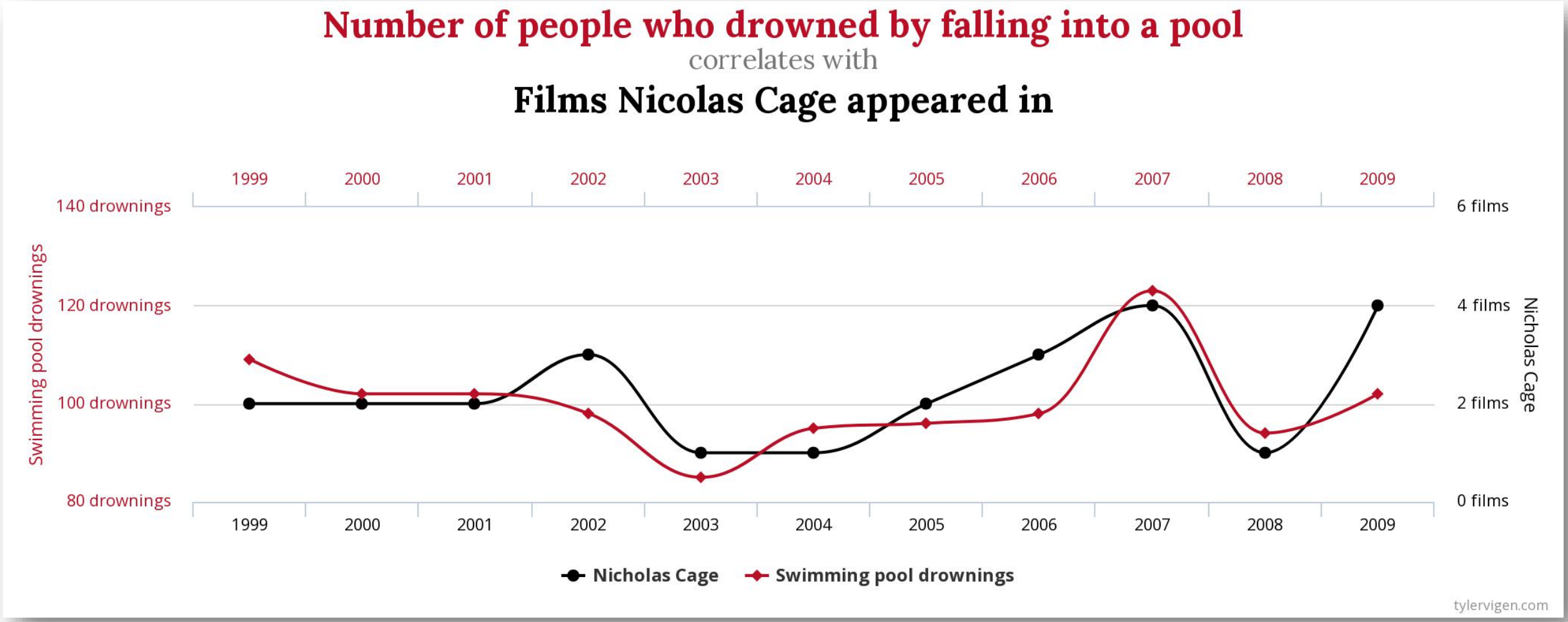
https://en.wikipedia.org/wiki/Isaac_Newton#/media/File:GodfreyKneller-IsaacNewton-1689.jpg

vs.

https://www.flickr.com/photos/seattlecitycouncil/39074799225/

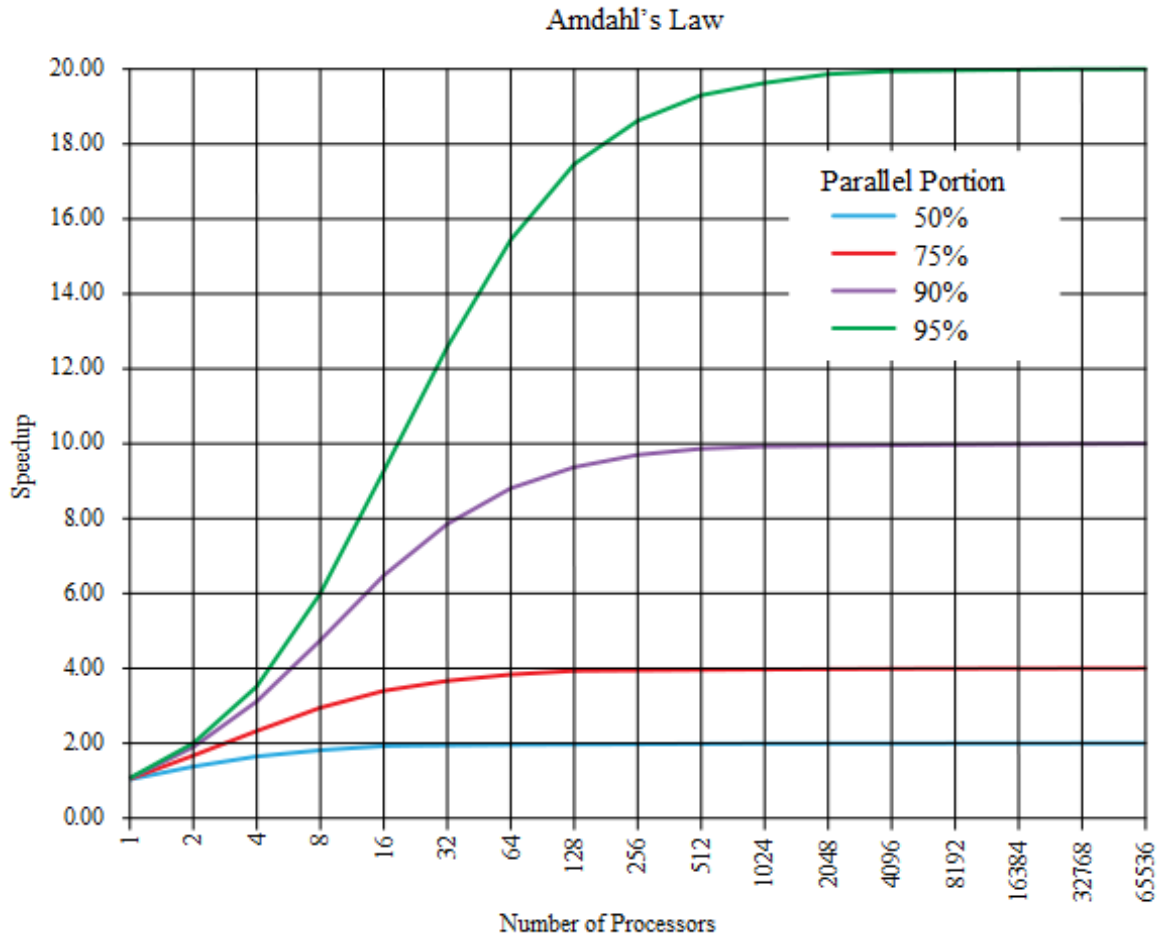
Felix Naumann
Data Science 2019

Correlation vs. Causation

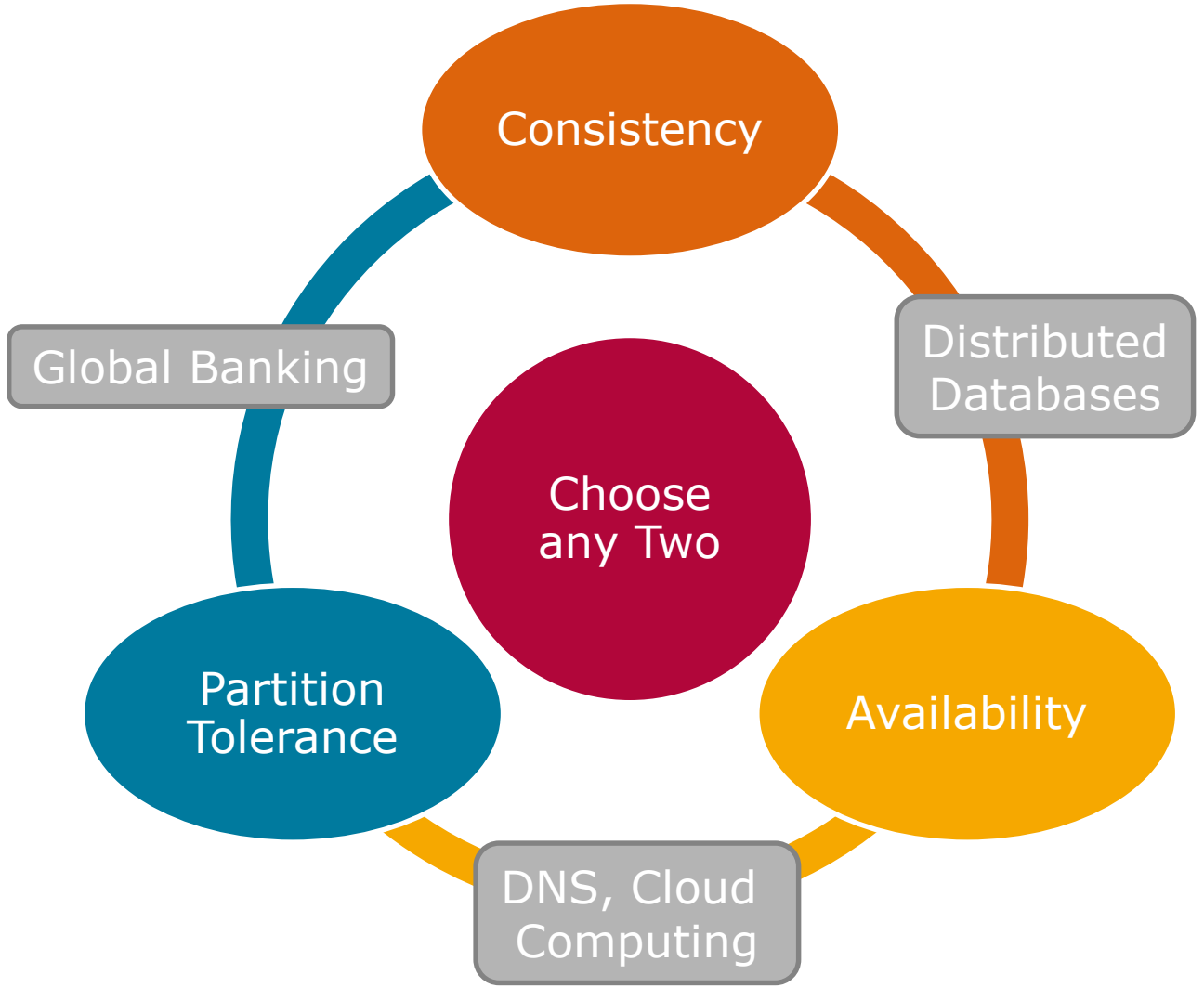


Parallelization obstacle: Ahmdahl's Law

- Maximal speedup is determined by non-parallelizable part of program:
  - $S_{max} = \frac{1}{(1-f)+f/p}$
  - p processors
  - f parallelizable fraction



Distribution Obstacle: CAP Theorem



Overview

1. Data Science
2. Big Data
3. **Data Profiling**
4. Data Preparation
5. Data Cleaning

https://unsplash.com/photos/vGefUiWm0xI

Definition Data Profiling

- Data profiling refers to the activity of creating small but informative summaries of a database.

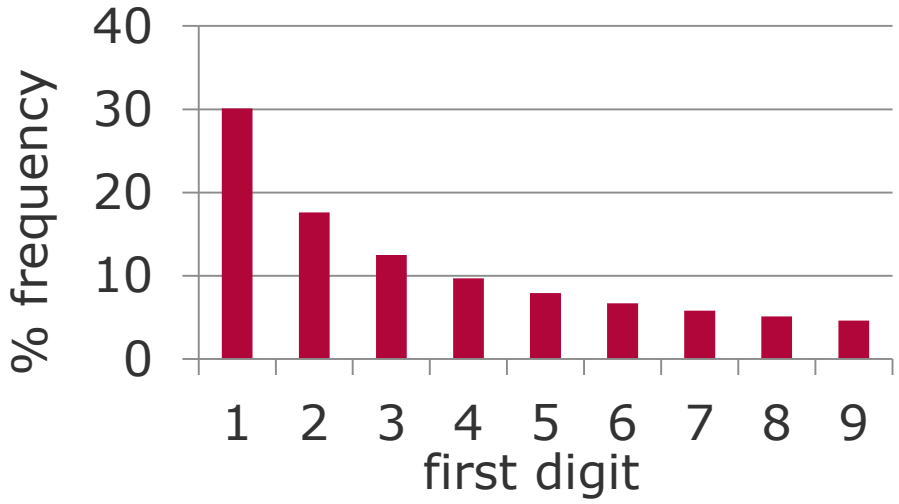
  Ted Johnson, Encyclopedia of Database Systems

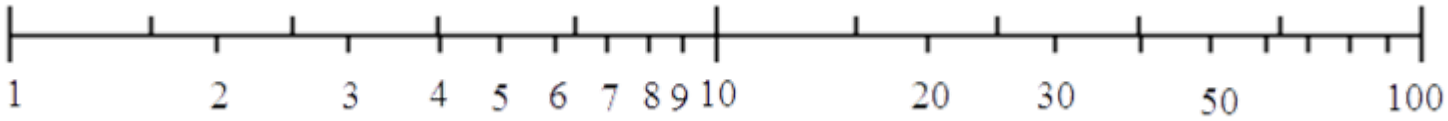
- Extracting metadata from given data
  - Basic statistics and histograms
  - Datatypes
  - Key and foreign keys
  - Dependencies and rules

- Data profiling is first step in any data management task
  - "*What shape does my data have?*"

Benford Law Frequency , a.k.a. “first digit law”

- Statement about the distribution of first digits *d* in (many) *naturally occurring* numbers:
  - $P(d) = log_{10}(d+1) - log_{10}(d) = log_{10}(1 + {}^{1}/_{d})$

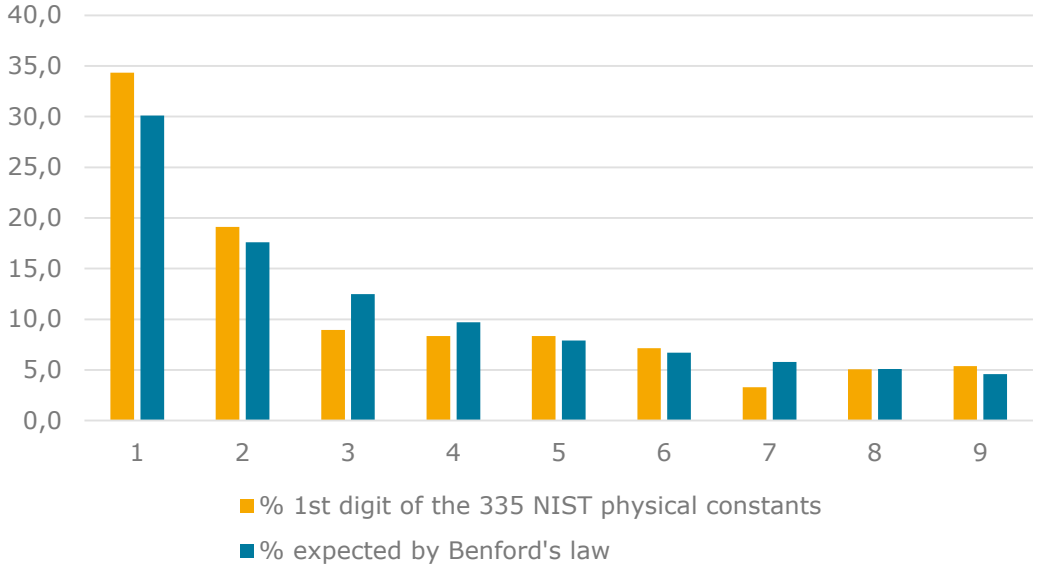




Is true if log(x) is uniformly distributed

Examples for Benford's Law

- Surface areas of 335 rivers
- Sizes of 3259 US populations
- 1800 molecular weights
- 5000 entries from a mathematical handbook
- 308 numbers in an issue of Reader's Digest
- Street addresses of the first 342 persons listed in American Men of Science
- Powers of 2: 2^n



Heights of the 60 tallest structures

Leading digit	meters		In Benford's law
	Count	%	
1	26	43.3%	30.1%
2	7	11.7%	17.6%
3	9	15.0%	12.5%
4	6	10.0%	9.7%
5	4	6.7%	7.9%
6	1	1.7%	6.7%
7	2	3.3%	5.8%
8	5	8.3%	5.1%
9	0	0.0%	4.6%

http://en.wikipedia.org/wiki/List_of_tallest_buildings_and_structures_in_the_world#Tallest_structure_by_category

Fraud detection

Felix Naumann
Data Science 2019

CODATA RECOMMENDED VALUES OF THE FUNDAMENTAL PHYSICAL CONSTANTS: 2014

NIST SP 961 (Sept/2015) Values from: P. J. Mohr, D. B. Newell, and B. N. Taylor, arXiv:1507.07956

A more extensive listing of constants is available in the above reference and on the NIST Physics Laboratory Web site physics.nist.gov/constants.

The number in parentheses is the one-standard-deviation uncertainty in the last two digits of the given value.

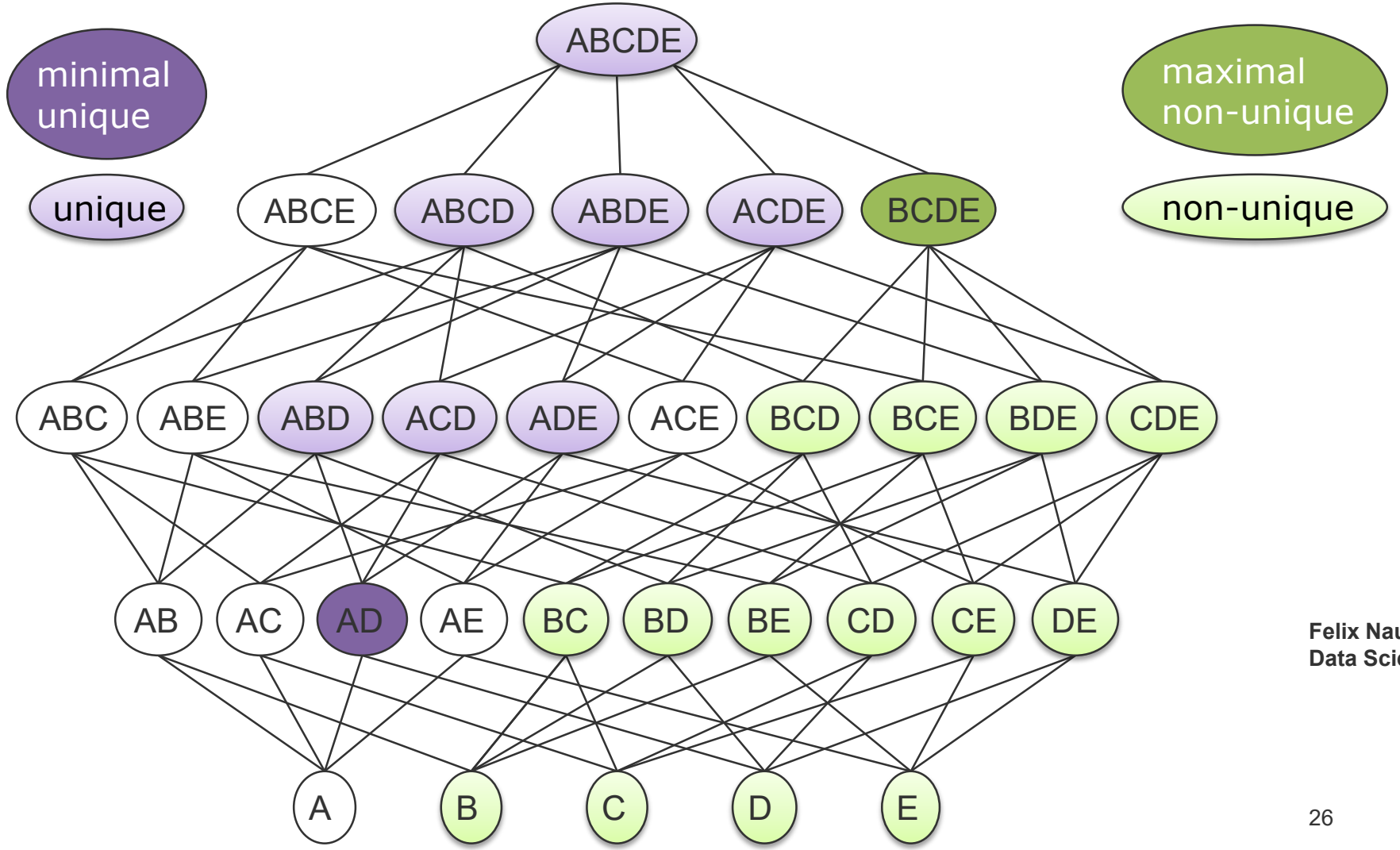
Quantity	Symbol	Numerical value	Unit
speed of light in vacuum	c, c_0	299 792 458 (exact)	$\mathrm{m\ s^{-1}}$
magnetic constant	μ_0	$4\pi \times 10^{-7}$ (exact)	$\mathrm{N\ A^{-2}}$
		$= 12.566\,370\,614\ldots \times 10^{-7}$	$\mathrm{N\ A^{-2}}$
electric constant $1/\mu_0 c^2$	ϵ_0	$8.854\,187\,817\ldots \times 10^{-12}$	$\mathrm{F\ m^{-1}}$
Newtonian constant of gravitation	G	$6.674\,08(31) \times 10^{-11}$	$\mathrm{m^3\ kg^{-1}\ s^{-2}}$
Planck constant	h	$6.626\,070\,040(81) \times 10^{-34}$	J s
in eV s		$4.135\,667\,662(25) \times 10^{-15}$	eV s
$h/2\pi$	$\hbar$	$1.054\,571\,800(13) \times 10^{-34}$	J s
in eV s		$6.582\,119\,514(40) \times 10^{-16}$	eV s
elementary charge	e	$1.602\,176\,6208(98) \times 10^{-19}$	C
magnetic flux quantum $h/2e$	Φ_0	$2.067\,833\,831(13) \times 10^{-15}$	Wb
Josephson constant $2e/h$	K_J	$483\,597.8525(30) \times 10^{9}$	$\mathrm{Hz\ V^{-1}}$
von Klitzing constant $h/e^2 = \mu_0 c/2\alpha$	R_K	25 812.807 4555(59)	Ω
Bohr magneton $e\hbar/2m_\mathrm{e}$	μ_B	$927.400\,9994(57) \times 10^{-26}$	$\mathrm{J\ T^{-1}}$
in eV $\mathrm{T^{-1}}$		$5.788\,381\,8012(26) \times 10^{-5}$	$\mathrm{eV\ T^{-1}}$
nuclear magneton $e\hbar/2m_\mathrm{p}$	μ_N	$5.050\,783\,699(31) \times 10^{-27}$	$\mathrm{J\ T^{-1}}$
in eV $\mathrm{T^{-1}}$		$3.152\,451\,2550(15) \times 10^{-8}$	$\mathrm{eV\ T^{-1}}$
fine-structure constant $e^2/4\pi\epsilon_0\hbar c$	α	$7.297\,352\,5664(17) \times 10^{-3}$	
inverse fine-structure constant	α^{-1}	137.035 999 139(31)	
Rydberg constant $\alpha^2 m_\mathrm{e} c/2h$	R_∞	10 973 731.568 508(65)	$\mathrm{m^{-1}}$
	$R_\infty c$	$3.289\,841\,960\,355(19) \times 10^{15}$	Hz
energy equivalent in eV	$R_\infty hc$	13.605 693 009(84)	eV
Bohr radius $\alpha/4\pi R_\infty = 4\pi\epsilon_0\hbar^2/m_\mathrm{e}e^2$	a_0	$0.529\,177\,210\,67(12) \times 10^{-10}$	m
Hartree energy $e^2/4\pi\epsilon_0 a_0 = 2R_\infty hc = \alpha^2 m_\mathrm{e} c^2$	E_h	$4.359\,744\,650(54) \times 10^{-18}$	J
in eV		27.211 386 02(17)	eV
electron mass	m_e	$9.109\,383\,56(11) \times 10^{-31}$	kg
in u		$5.485\,799\,090\,70(16) \times 10^{-4}$	u
energy equivalent in MeV	$m_\mathrm{e} c^2$	0.510 998 9461(31)	MeV
electron-muon mass ratio	m_e/m_μ	$4.836\,331\,70(11) \times 10^{-3}$	
electron-proton mass ratio	$m_\mathrm{e}/m_\mathrm{p}$	$5.446\,170\,213\,52(52) \times 10^{-4}$	
electron charge to mass quotient	$-e/m_\mathrm{e}$	$-1.758\,820\,024(11) \times 10^{11}$	$\mathrm{C\ kg^{-1}}$
Compton wavelength $h/m_\mathrm{e}c$	λ_C	$2.426\,310\,2367(11) \times 10^{-12}$	m
$\lambda_\mathrm{C}/2\pi = \alpha a_0 = \alpha^2/4\pi R_\infty$	$\lambda\!\!\!^{-}_\mathrm{C}$	$386.159\,267\,64(18) \times 10^{-15}$	m
classical electron radius $\alpha^2 a_0$	r_e	$2.817\,940\,3227(19) \times 10^{-15}$	m
Thomson cross section $(8\pi/3)r_\mathrm{e}^2$	σ_e	$0.665\,245\,871\,58(91) \times 10^{-28}$	$\mathrm{m^2}$
electron magnetic moment	μ_e	$-928.476\,4620(57) \times 10^{-26}$	$\mathrm{J\ T^{-1}}$
to Bohr magneton ratio	$\mu_\mathrm{e}/\mu_\mathrm{B}$	$-1.001\,159\,652\,180\,91(26)$	
to nuclear magneton ratio	$\mu_\mathrm{e}/\mu_\mathrm{N}$	$-1838.281\,972\,34(17)$	
electron magnetic moment anomaly $\lvert\mu_\mathrm{e}\rvert/\mu_\mathrm{B} - 1$	a_e	$1.159\,652\,180\,91(26) \times 10^{-3}$	
electron g-factor $-2(1 + a_\mathrm{e})$	g_e	$-2.002\,319\,304\,361\,82(52)$	
electron-proton magnetic moment ratio	$\mu_\mathrm{e}/\mu_\mathrm{p}$	$-658.210\,6866(20)$	
muon mass in u	m_μ	0.113 428 9257(25)	u
energy equivalent in MeV	$m_\mu c^2$	105.658 3745(24)	MeV
muon-electron mass ratio	m_μ/m_e	206.768 2826(46)	
muon magnetic moment	μ_μ	$-4.490\,448\,26(10) \times 10^{-26}$	$\mathrm{J\ T^{-1}}$
to Bohr magneton ratio	μ_μ/μ_B	$-4.841\,970\,48(11) \times 10^{-3}$	
to nuclear magneton ratio	μ_μ/μ_N	$-8.890\,597\,05(20)$	
muon magnetic moment anomaly $\lvert\mu_\mu\rvert/(e\hbar/2m_\mu) - 1$	a_μ	$1.165\,920\,89(63) \times 10^{-3}$	

Quantity	Symbol	Numerical value	Unit
muon g-factor $-2(1 + a_\mu)$	g_μ	$-2.002\,331\,8418(13)$	
muon-proton magnetic moment ratio	μ_μ/μ_p	$-3.183\,345\,142(71)$	
proton mass	m_p	$1.672\,621\,898(21) \times 10^{-27}$	kg
in u		1.007 276 466 879(91)	u
energy equivalent in MeV	$m_\mathrm{p} c^2$	938.272 0813(58)	MeV
proton-electron mass ratio	$m_\mathrm{p}/m_\mathrm{e}$	1836.152 673 89(17)	
proton magnetic moment	μ_p	$1.410\,606\,7873(97) \times 10^{-26}$	$\mathrm{J\ T^{-1}}$
to nuclear magneton ratio	$\mu_\mathrm{p}/\mu_\mathrm{N}$	2.792 847 3508(85)	
proton magnetic shielding correction $1 - \mu'_\mathrm{p}/\mu_\mathrm{p}$ ($\mathrm{H_2O}$, sphere, 25 °C)	σ'_p	$25.691(11) \times 10^{-6}$	
proton gyromagnetic ratio $2\mu_\mathrm{p}/\hbar$	γ_p	$2.675\,221\,900(18) \times 10^{8}$	$\mathrm{s^{-1}\ T^{-1}}$
	$\gamma_\mathrm{p}/2\pi$	42.577 478 92(29)	$\mathrm{MHz\ T^{-1}}$
shielded proton gyromagnetic ratio $2\mu'_\mathrm{p}/\hbar$ ($\mathrm{H_2O}$, sphere, 25 °C)	γ'_p	$2.675\,153\,171(33) \times 10^{8}$	$\mathrm{s^{-1}\ T^{-1}}$
	$\gamma'_\mathrm{p}/2\pi$	42.576 385 07(53)	$\mathrm{MHz\ T^{-1}}$
neutron mass in u	m_n	1.008 664 915 88(49)	u
energy equivalent in MeV	$m_\mathrm{n} c^2$	939.565 4133(58)	MeV
neutron-proton mass ratio	$m_\mathrm{n}/m_\mathrm{p}$	1.001 378 418 98(51)	
neutron magnetic moment	μ_n	$-0.966\,236\,50(23) \times 10^{-26}$	$\mathrm{J\ T^{-1}}$
to nuclear magneton ratio	$\mu_\mathrm{n}/\mu_\mathrm{N}$	$-1.913\,042\,73(45)$	
deuteron mass in u	m_d	2.013 553 212 745(40)	u
energy equivalent in MeV	$m_\mathrm{d} c^2$	1875.612 928(12)	MeV
deuteron-proton mass ratio	$m_\mathrm{d}/m_\mathrm{p}$	1.999 007 500 87(19)	
deuteron magnetic moment	μ_d	$0.433\,073\,5040(36) \times 10^{-26}$	$\mathrm{J\ T^{-1}}$
to nuclear magneton ratio	$\mu_\mathrm{d}/\mu_\mathrm{N}$	0.857 438 2311(48)	
helion (^{3}He nucleus) mass in u	m_h	3.014 932 246 73(12)	u
energy equivalent in MeV	$m_\mathrm{h} c^2$	2808.391 586(17)	MeV
shielded helion magnetic moment (gas, sphere, 25 °C)	μ'_h	$-1.074\,553\,080(14) \times 10^{-26}$	$\mathrm{J\ T^{-1}}$
to Bohr magneton ratio	$\mu'_\mathrm{h}/\mu_\mathrm{B}$	$-1.158\,671\,471(14) \times 10^{-3}$	
to nuclear magneton ratio	$\mu'_\mathrm{h}/\mu_\mathrm{N}$	$-2.127\,497\,720(25)$	
alpha particle mass in u	m_α	4.001 506 179 127(63)	u
energy equivalent in MeV	$m_\alpha c^2$	3727.379 378(23)	MeV
Avogadro constant	N_A, L	$6.022\,140\,857(74) \times 10^{23}$	$\mathrm{mol^{-1}}$
atomic mass constant $\frac{1}{12}m(^{12}\mathrm{C}) = 1$ u	m_u	$1.660\,539\,040(20) \times 10^{-27}$	kg
energy equivalent in MeV	$m_\mathrm{u} c^2$	931.494 0954(57)	MeV
Faraday constant $N_\mathrm{A} e$	F	96 485.332 89(59)	$\mathrm{C\ mol^{-1}}$
molar gas constant	R	8.314 4598(48)	$\mathrm{J\ mol^{-1}\ K^{-1}}$
Boltzmann constant R/N_A	k	$1.380\,648\,52(79) \times 10^{-23}$	$\mathrm{J\ K^{-1}}$
in eV $\mathrm{K^{-1}}$		$8.617\,3303(50) \times 10^{-5}$	$\mathrm{eV\ K^{-1}}$
molar volume of ideal gas RT/p ($T = 273.15$ K, $p = 101.325$ kPa)	V_m	$22.413\,962(13) \times 10^{-3}$	$\mathrm{m^3\ mol^{-1}}$
Stefan-Boltzmann constant $\pi^2 k^4/60\hbar^3 c^2$	σ	$5.670\,367(13) \times 10^{-8}$	$\mathrm{W\ m^{-2}\ K^{-4}}$
first radiation constant $2\pi hc^2$	c_1	$3.741\,771\,790(46) \times 10^{-16}$	$\mathrm{W\ m^2}$
second radiation constant hc/k	c_2	$1.438\,777\,36(83) \times 10^{-2}$	m K
Wien displacement law constant $b = \lambda_\mathrm{max} T = c_2/4.965\,114\,231\ldots$	b	$2.897\,7729(17) \times 10^{-3}$	m K
Cu x unit: $\lambda(\mathrm{Cu\ K}\alpha_1)/1\,537.400$	$\mathrm{xu(Cu\ K}\alpha_1\mathrm{)}$	$1.002\,076\,97(28) \times 10^{-13}$	m
Mo x unit: $\lambda(\mathrm{Mo\ K}\alpha_1)/707.831$	$\mathrm{xu(Mo\ K}\alpha_1\mathrm{)}$	$1.002\,099\,52(53) \times 10^{-13}$	m

Energy equivalents

$(1\ \mathrm{m^{-1}})c = 299\,792\,458$ Hz	$(1\ \mathrm{Hz})h/k = 4.799\,2447(28) \times 10^{-11}$ K	(1 J) $= 6.241\,509\,126(38) \times 10^{18}$ eV	$(1\ \mathrm{eV})/c^2 = 1.073\,544\,1105(66) \times 10^{-9}$ u
$(1\ \mathrm{m^{-1}})hc/k = 1.438\,777\,36(83) \times 10^{-2}$ K	$(1\ \mathrm{Hz})h = 4.135\,667\,662(25) \times 10^{-15}$ eV	(1 eV) $= 1.602\,176\,6208(98) \times 10^{-19}$ J	(1 kg) $= 6.022\,140\,857(74) \times 10^{26}$ u
$(1\ \mathrm{m^{-1}})hc = 1.239\,841\,9739(76) \times 10^{-6}$ eV	$(1\ \mathrm{K})k/hc = 69.503\,457(40)\ \mathrm{m^{-1}}$	$(1\ \mathrm{eV})/hc = 8.065\,544\,005(50) \times 10^{5}\ \mathrm{m^{-1}}$	(1 u) $= 1.660\,539\,040(20) \times 10^{-27}$ kg
$(1\ \mathrm{m^{-1}})h/c = 1.331\,025\,049\,00(61) \times 10^{-15}$ u	$(1\ \mathrm{K})k/h = 2.083\,6612(12) \times 10^{10}$ Hz	$(1\ \mathrm{eV})/h = 2.417\,989\,262(15) \times 10^{14}$ Hz	$(1\ \mathrm{u})c/h = 7.513\,006\,6166(34) \times 10^{14}\ \mathrm{m^{-1}}$
$(1\ \mathrm{Hz})/c = 3.335\,640\,951\ldots \times 10^{-9}\ \mathrm{m^{-1}}$	$(1\ \mathrm{K})k = 8.617\,3303(50) \times 10^{-5}$ eV	$(1\ \mathrm{eV})/k = 1.160\,452\,21(67) \times 10^{4}$ K	$(1\ \mathrm{u})c^2 = 931.494\,0954(57) \times 10^{6}$ eV

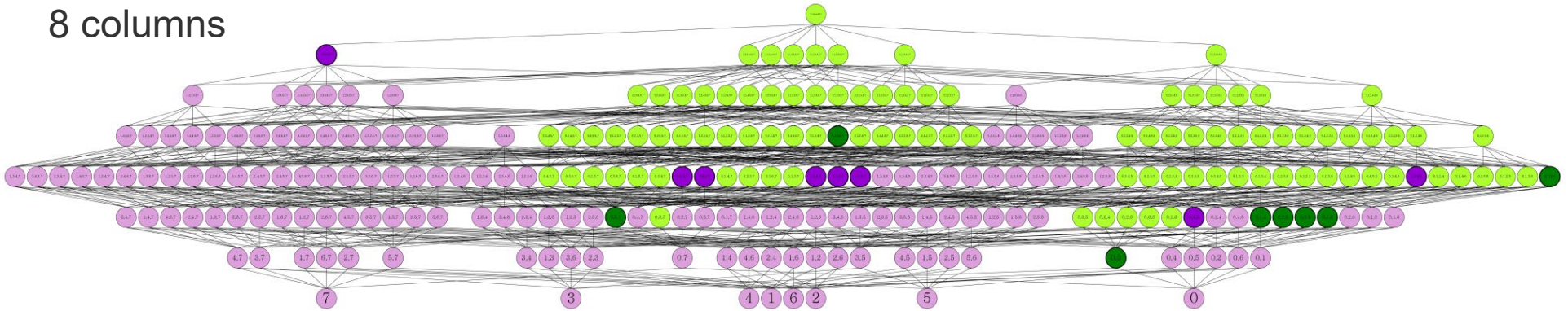
Detecting Unique Column Combinations (aka. keys)



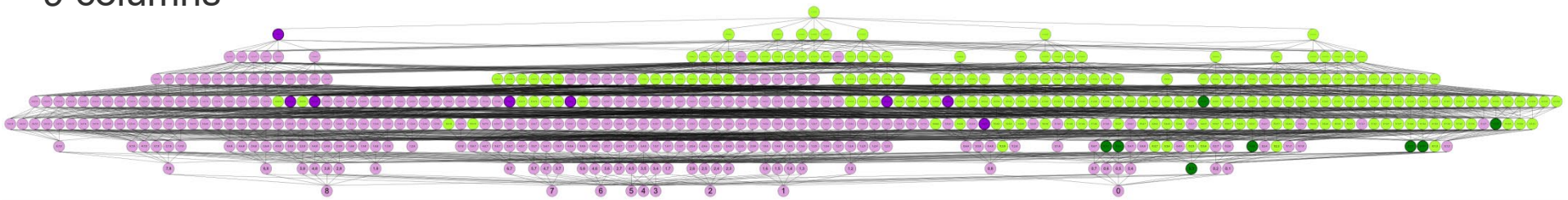
Large search space: $2^n - 1$

Large solution space: $\binom{n}{n/2}$

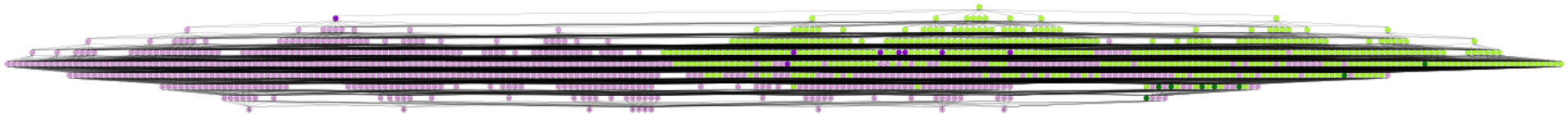
8 columns



9 columns

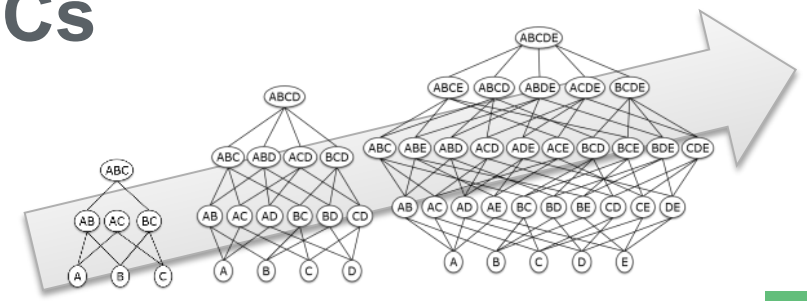


10 columns



Felix Naumann
Data Science 2019

Candidate Set Growth for Unique Column Combinations



Number of levels: k	Total : 1	3	7	15	31	63	127	255	511	1,023	2,047	4,095	8,191	16,383	32,767
15															1
14														1	15
13													1	14	105
12												1	13	91	455
11											1	12	78	364	1,365
10										1	11	66	286	1,001	3,003
9									1	10	55	220	715	2,002	5,005
8								1	9	45	165	495	1,287	3,003	6,435
7							1	8	36	120	330	792	1,716	3,432	6,435
6						1	7	28	84	210	462	924	1,716	3,003	5,005
5					1	6	21	56	126	252	462	792	1,287	2,002	3,003
4				1	5	15	35	70	126	210	330	495	715	1,001	1,365
3			1	4	10	20	35	56	84	120	165	220	286	364	455
2		1	3	6	10	15	21	28	36	45	55	66	78	91	105
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

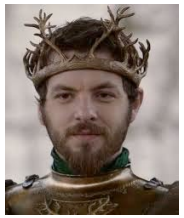
Number of attributes: m

Functional Dependencies

https://www.flickr.com/photos/bagogames/16632632814/

Felix Naumann
Data Science 2019

Functional Dependencies

Person	Lineage	Hair	Religion
			New gods
			New Gods
			Old gods
			New gods
			New gods

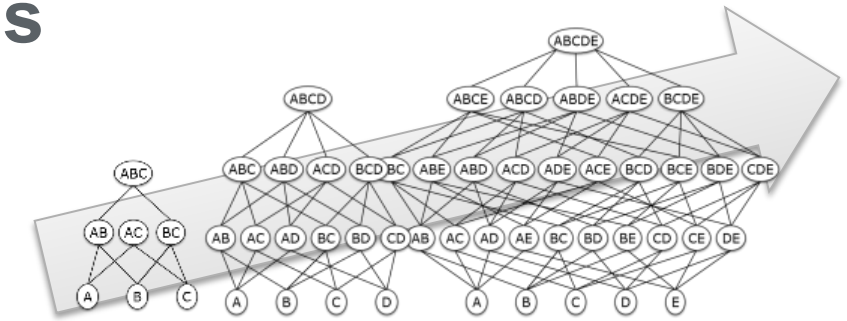
Some Functional Dependencies:

1. Person → Lineage
2. Person → Hair
3. Person → Religion
4. Lineage → Hair
5. Religion, Hair → Lineage
6. …

Ned Stark: „Number 4 looks like a reasonable quality constraint"

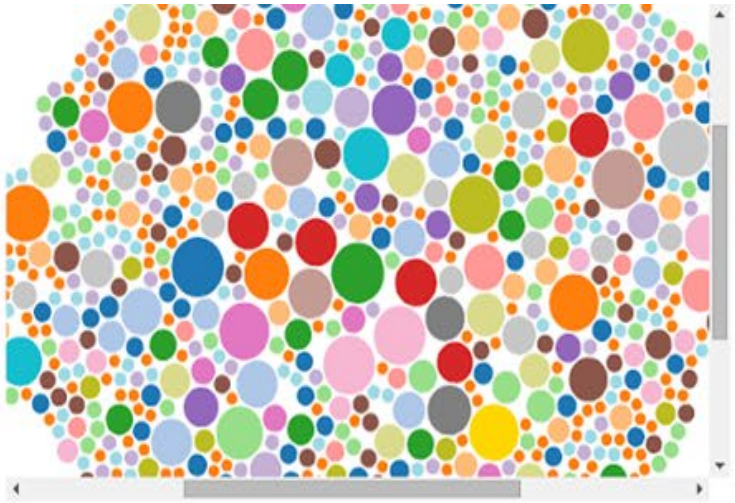
Ned Stark: „I believe Joffrey violates my database constraint."

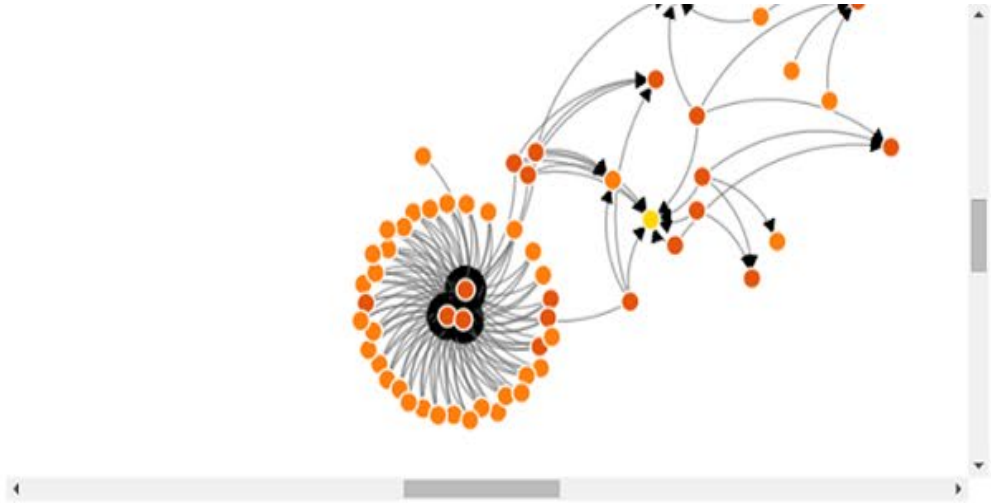
Candidate Set Growth for Functional Dependencies



Number of levels: k	Number of attributes: m														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Total:	0	2	9	28	75	186	441	1,016	2,295	5,110	11,253	24,564	53,235	114,674	245,745
15															0
14														0	15
13													0	14	210
12												0	13	182	1,365
11											0	12	156	1,092	5,460
10										0	11	132	858	4,004	15,015
9									0	10	110	660	2,860	10,010	30,030
8								0	9	90	495	1,980	6,435	18,018	45,045
7							0	8	72	360	1,320	3,960	10,296	24,024	51,480
6						0	7	56	252	840	2,310	5,544	12,012	24,024	45,045
5					0	6	42	168	504	1,260	2,772	5,544	10,296	18,018	30,030
4				0	5	30	105	280	630	1,260	2,310	3,960	6,435	10,010	15,015
3			0	4	20	60	140	280	504	840	1,320	1,980	2,860	4,004	5,460
2		0	3	12	30	60	105	168	252	360	495	660	858	1,092	1,365
1	0	2	6	12	20	30	42	56	72	90	110	132	156	182	210

Linking up millions of web tables with inclusion dependencies





96242-1	Association'.csv
43666-3	43666-3.'BBC_Radio_Stoke'.'Programming'.csv
53064-1	53064-1.'Rotation_period'.'Rotation period of selected objects'.csv
562884-4	562884-4.'Planets_in_astrology'.'Ruling planets of the astrological signs and houses'.csv
175797-1	175797-1.'Sun_sign_astrology'.'Sun signs'.csv
177750-2	177750-2.'BBC_Radio_Manchester'.'Programming'.csv
89462-4	89462-4.'Astrology_and_the_classical_elements'.'Triplicities by season'.csv
213213-1	213213-1.'Dalton_Park'.'Opening times'.csv
	470402-

Celestial Objects	Rotation period	Rotation period
Sun	25.379995 days (equatorial) 35 days (high latitude)	25 d 9 h 7 m 11.6 s 35 d
Mercury	58.6462 days	58 d 15 h 30 m 30 s
Venus	?243.0187 days	?243 d 0 h 26 m
Earth	0.99726968 days	0 d 23 h 56 m 4.100 s
Moon	27.321661 days (synchronous toward Earth)	27 d 7 h 43 m 11.5 s
Mars	1.02595675 days	1 d 0 h 37 m 22.663 s
Ceres	0.37809 days	0 d 9 h 4 m 27.0 s
Jupiter	0.4135344 days (deep interior) 0.41007 days (equatorial) 0.41369942 days (high latitude)	0 d 9 h 55 m 29.37 s 0 d 9 h 50 m 30 s 0 d 9 h 55 m 43.63 s
Saturn	0.44403 days (deep interior) 0.426 days (equatorial) 0.443 days (high	0 d 10 h 39 m 24 s 0 d 10 h 14 m 0 d 10 h 38 m

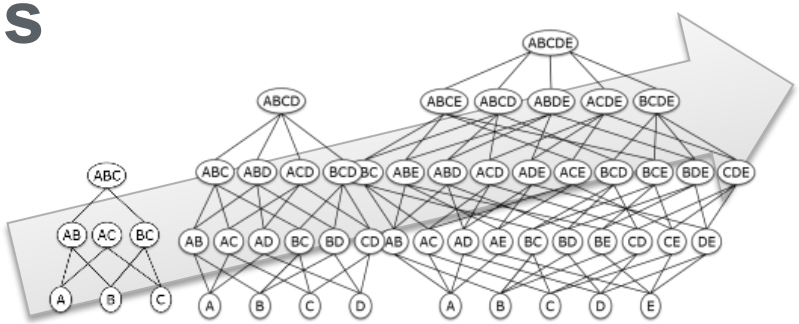
Zoom (1-5)

Range (logarithmic)

Dataset

allFilters

Candidate Set Growth for Inclusion Dependencies



Number of levels: k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Total:	0	2	6	24	80	330	1,302	5,936	26,784	133,650	669,350	3,609,672	19,674,096	113,525,594	664,400,310
15															0
14														0	0
13													0	0	0
12												0	0	0	0
11											0	0	0	0	0
10										0	0	0	0	0	0
9									0	0	0	0	0	0	0
8								0	0	0	0	0	0	0	0
7							0	0	0	0	0	0	0	17,297,280	259,459,200
6						0	0	0	0	0	0	665,280	8,648,640	60,540,480	302,702,400
5					0	0	0	0	0	30,240	332,640	1,995,840	8,648,640	30,270,240	90,810,720
4				0	0	0	0	1,680	15,120	75,600	277,200	831,600	2,162,160	5,045,040	10,810,800
3			0	0	0	120	840	3,360	10,080	25,200	55,440	110,880	205,920	360,360	600,600
2		0	0	12	60	180	420	840	1,512	2,520	3,960	5,940	8,580	12,012	16,380
1	0	2	6	12	20	30	42	56	72	90	110	132	156	182	210

Number of attributes: m

Profiling Challenges

- Efficient profiling
- Scalable profiling
- Holistic profiling
- Incremental profiling
- Temporal profiling
- Profiling query results
- Profiling new types of data

- Hundreds of UCCs – which ones are keys?
- Thousands of FDs – which ones are true?
- Millions of INDs – which ones are foreign keys?

Use Cases for Data Profiling

- Query optimization
  - Counts and histograms, functional dependencies, …
- Data cleansing
  - Patterns, rules, and violations
- Data integration
  - Cross-DB inclusion dependencies
- Scientific data management
  - Inspect new datasets
- Data analytics and mining
  - Profiling as preparation to decide on models and questions
- Database reverse engineering

In summary: Data preparation

Overview

1. Data Science
2. Big Data
3. Data Profiling
4. **Data Preparation**
5. Data Cleaning

https://unsplash.com/photos/vGefUiWm0xI

Its late…

Data P-r_e\p+a|r¶a.t/i~o-n

Title	Authors	Venue	Year
Immunogold labelling is a quantitative method as demonstrated by studies on aminopeptidase N in	GH Hansen, LL Wetterberg, H SjÃƒÂ¶strÃƒÂ¶m, O NorÃƒÂ©n	The Histochemical Journal,	1992
Infectious Inmates and Releasees From Correctional Facilities	TM Hammett, P Harmon, W Rhodes	see	
World Population Prospects: The 1996 Revision	U Nations	New York,	
Consequences of Migration and Remittances for Mexican Transnational Communities.	D Conway, JH Cohen	Economic Geography,	1998

Wrong encoding

Incorrect title

Incorrect venue

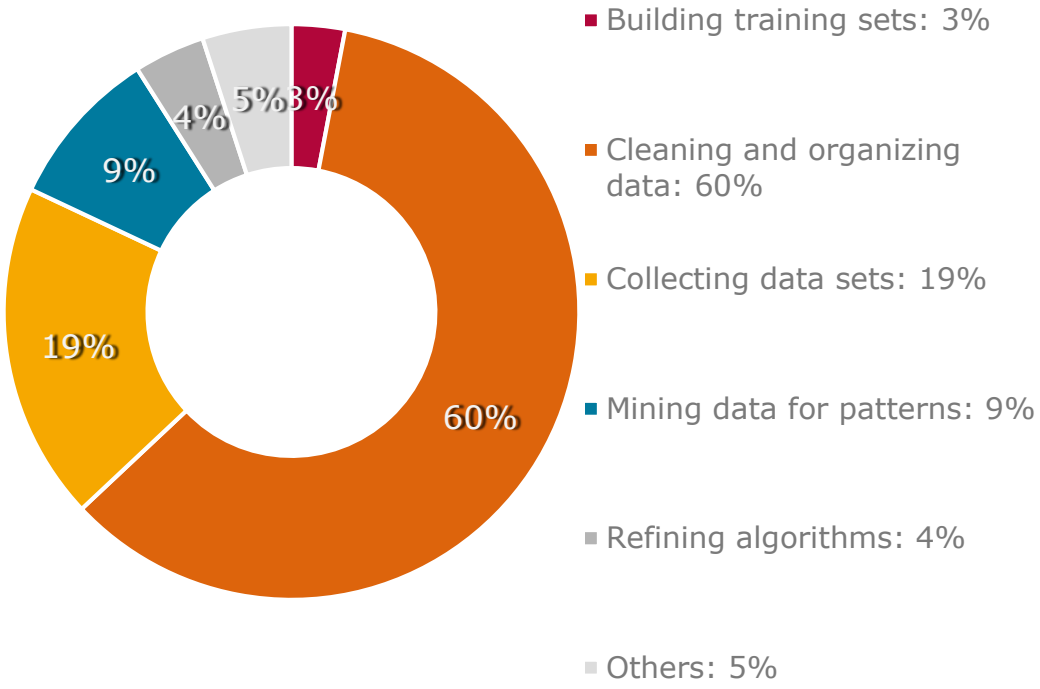
Missing values

redundant characters

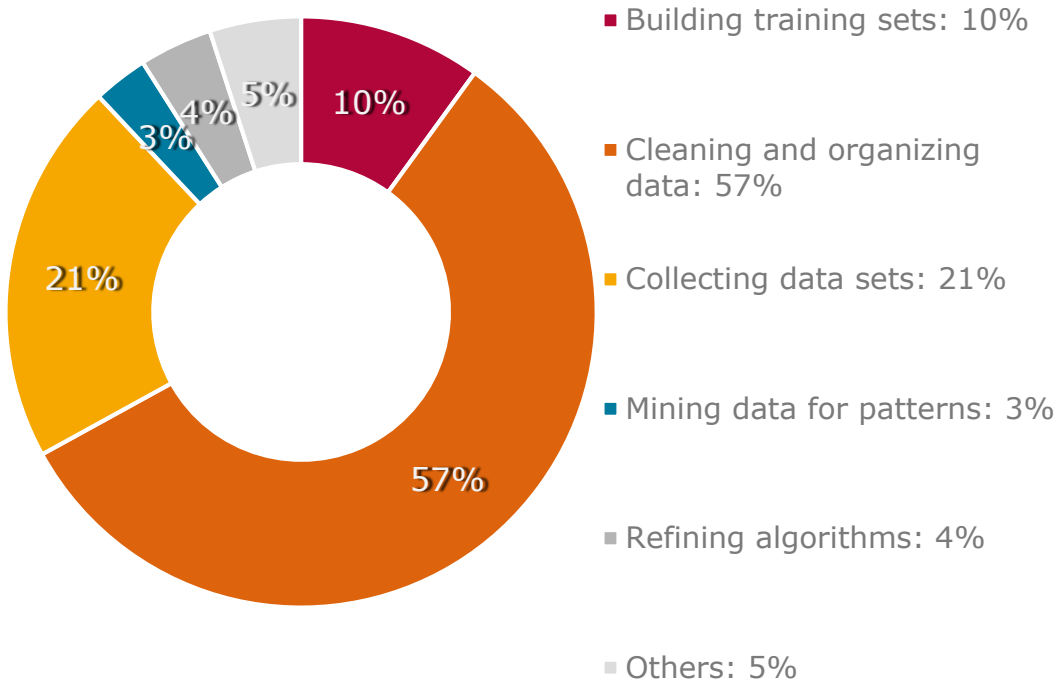
Felix Naumann
Data Science 2019

Data preparation in reality

What data scientists spend the most time doing?

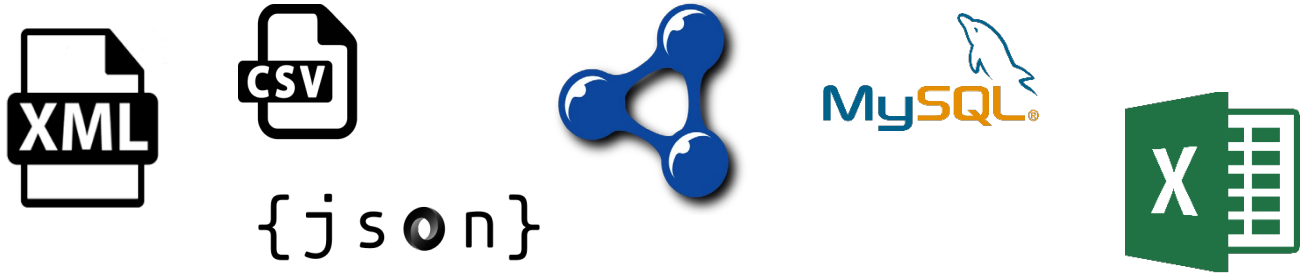


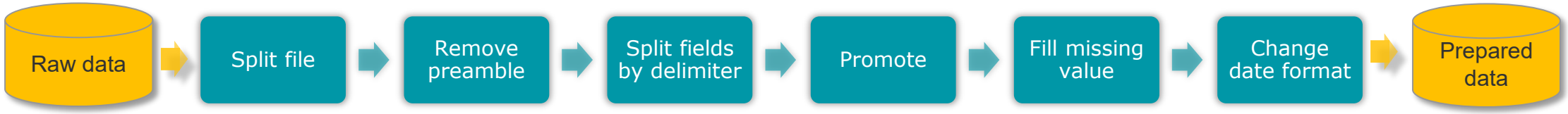
What is the least enjoyable part of data science?

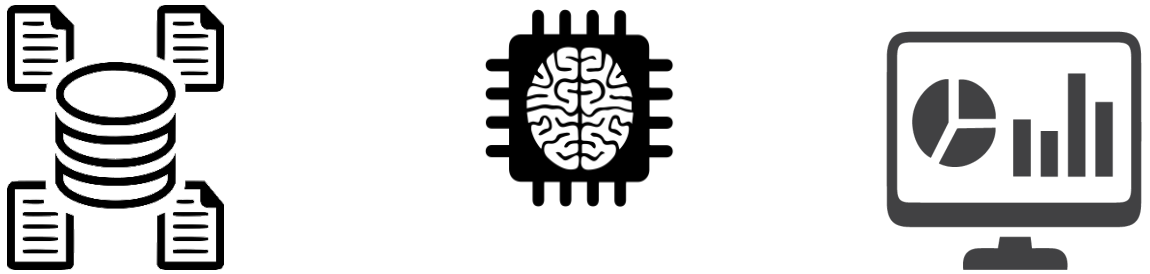


"Cleaning Data: Most Time-Consuming, Least Enjoyable Data Science Task", Gil Press, Forbes, March 23rd, 2016
http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

Data preparation pipelines







Felix Naumann
Data Science 2019

Overview

1. Data Science
2. Big Data
3. Data Profiling
4. Data Preparation
5. **Data Cleaning**

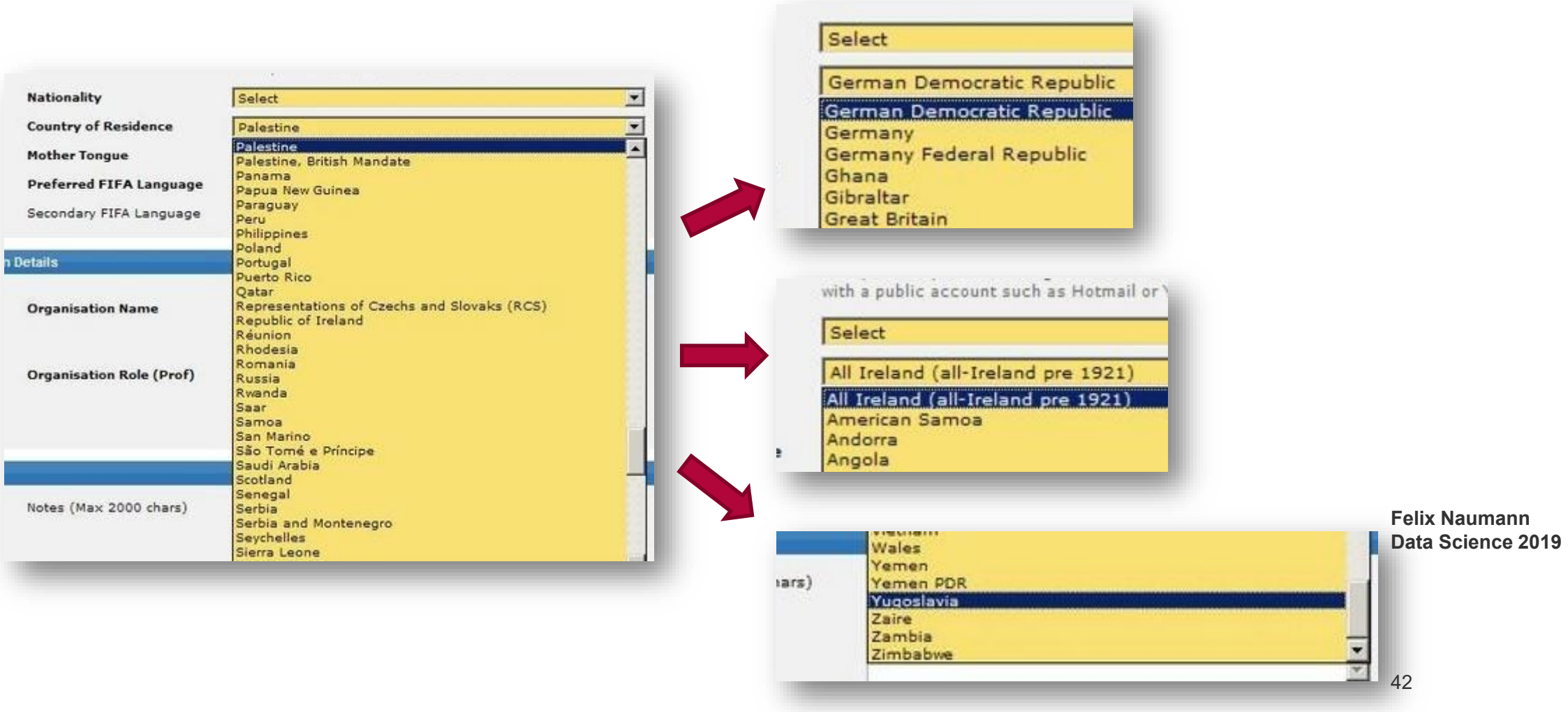
https://unsplash.com/photos/vGefUiWm0xI

Its late…

Difficult names

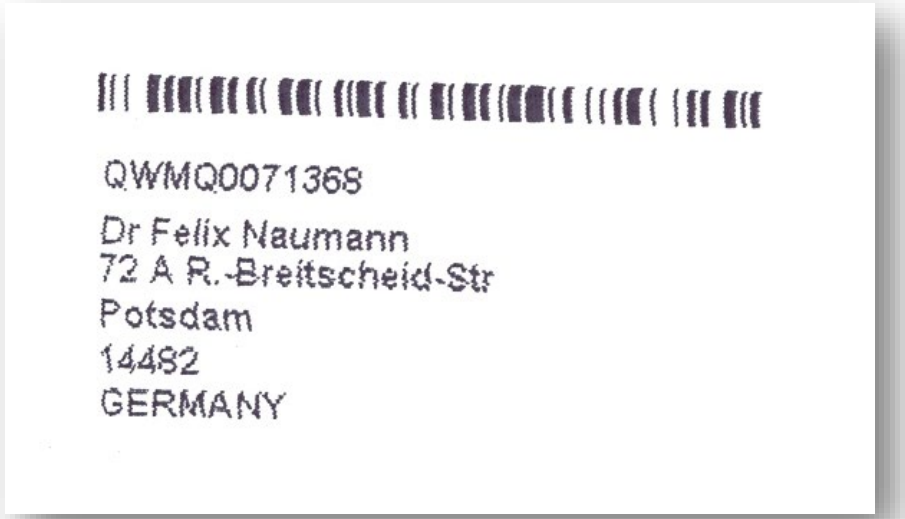
488941 britney spears	29 britent spears	9 brinttany spears	5 brney spears	3 britiy spears	2 brirreny spears
40134 brittany spears	29 brittnany spears	9 britanay spears	5 broitney spears	3 britmeny spears	2 brirtany spears
36315 brittney spears	29 britttany spears	9 britinany spears	5 brotny spears	3 britneeey spears	2 brirttany spears
24342 britany spears	29 btiney spears	9 britn spears	5 bruteny spears	3 britnehy spears	2 brirttney spears
7331 britny spears	26 birttney spears	9 britnew spears	5 btiyney spears	3 britnely spears	2 britain spears
6633 briteny spears	26 breitney spears	9 britneyn spears	5 btrittney spears	3 britnesy spears	2 britane spears
2696 britteny spears	26 brinity spears	9 britrney spears	5 gritney spears	3 britnetty spears	2 britaneny spears
1807 briney spears	26 britenay spears	9 brtiny spears	5 spritney spears	3 britnex spears	2 britania spears
1635 brittny spears	26 britneyt spears	9 brtittney spears	4 bittny spears	3 britneyxxx spears	2 britann spears
1479 brintey spears	26 brittan spears	9 brtny spears	4 bnritney spears	3 britnity spears	2 britanna spears
1479 britanny spears	26 brittne spears	9 brytny spears	4 brandy spears	3 britntey spears	2 britannie spears
1338 britiny spears	26 btittany spears	9 rbitney spears	4 brbritney spears	3 britnyey spears	2 britannt spears
1211 britnet spears	24 beitney spears	8 birtiny spears	4 breatiny spears	3 britterny spears	2 britannu spears
1096 britiney spears	24 birteny spears	8 bithney spears	4 breetney spears	3 brittneey spears	2 britanyl spears
991 britaney spears	24 brightney spears	8 brattany spears	4 bretiney spears	3 brittnney spears	2 britanyt spears
991 britnay spears	24 brintiny spears	8 breitny spears	4 brfitney spears	3 brittnyey spears	2 briteeny spears
811 brithney spears	24 britanty spears	8 breteny spears	4 briattany spears	3 brityen spears	2 britenany spears
811 brtiney spears	24 britenny spears	8 brightny spears	4 brieteny spears	3 briytney spears	2 britenet spears
664 birtney spears	24 britini spears	8 brintay spears	4 briety spears	3 brltney spears	2 briteniy spears
664 brintney spears	24 britnwy spears	8 brinttey spears	4 briitny spears	3 broteny spears	2 britenys spears
664 briteney spears	24 brittni spears	8 briotney spears	4 briittany spears	3 brtaney spears	2 britianey spears
601 bitney spears	24 brittnie spears	8 britanys spears	4 brinie spears	3 brtiiany spears	2 britin spears
601 brinty spears	21 biritney spears	8 britley spears	4 brinteney spears	3 brtinay spears	2 britinary spears
544 brittaney spears	21 birtany spears	8 britneyb spears	4 brintne spears	3 brtinney spears	2 britmy spears
544 brittnay spears	21 biteny spears	8 britnrey spears	4 britaby spears	3 brtitany spears	2 britnaney spears
364 britey spears	21 bratney spears	8 britnty spears	4 britaey spears	3 brtiteny spears	2 britnat spears
364 brittiny spears	21 britani spears	8 brittner spears	4 britainey spears	3 brtnet spears	2 britnbey spears
329 brtney spears	21 britanie spears	8 brottany spears	4 britinie spears	3 brytiny spears	2 britndy spears
269 bretney spears	21 briteany spears	7 baritney spears	4 britinney spears	3 btney spears	2 britneh spears
269 britneys spears	21 brittay spears	7 birntey spears	4 britmney spears	3 drittney spears	2 britneney spears
244 britne spears	21 brittinay spears	7 biteney spears	4 britnear spears	3 pretney spears	2 britney6 spears
244 brytney spears	21 brtany spears	7 bitiny spears	4 britnel spears	3 rbritney spears	2 britneye spears
220 breatney spears	21 brtiany spears	7 breateny spears	4 britneuy spears	2 barittany spears	2 britneyh spears
220 britiany spears	19 birney spears	7 brianty spears	4 britnewy spears	2 bbbritney spears	2 britneym spears
199 britnney spears	19 brirtney spears	7 brintye spears	4 britnmey spears	2 bbitney spears	2 britneyyy spears
163 britnry spears	19 britnaey spears	7 britianny spears	4 brittaby spears	2 bbritny spears	2 britnhey spears

FIFA registration form (2010)



Felix Naumann
Data Science 2019

Directmarketing by The Economist



Felix Naumann
Data Science 2019

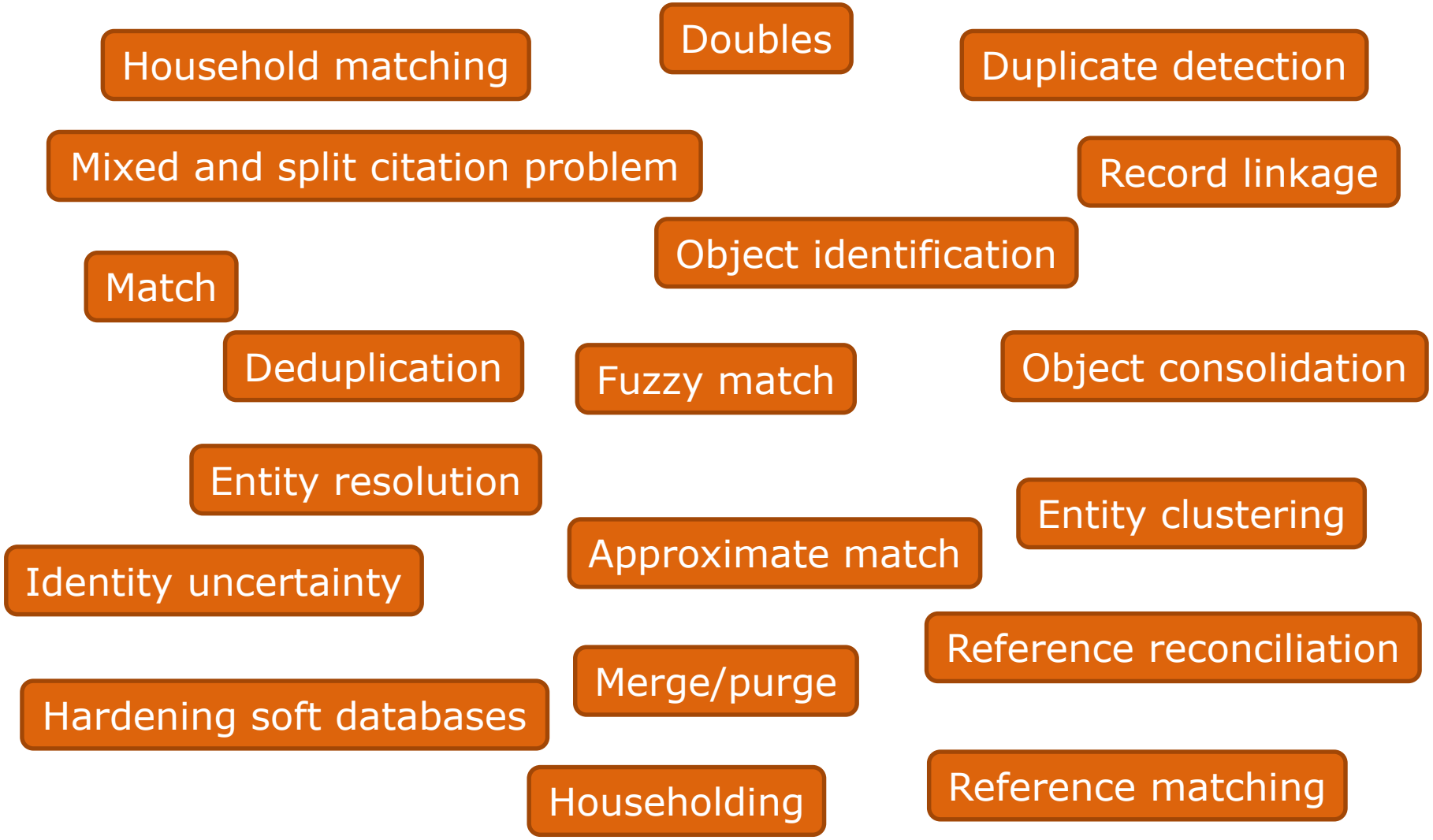
Data Cleaning: Duplicate Detection

- Duplicate detection is the discovery of multiple representations of the same real-world object.

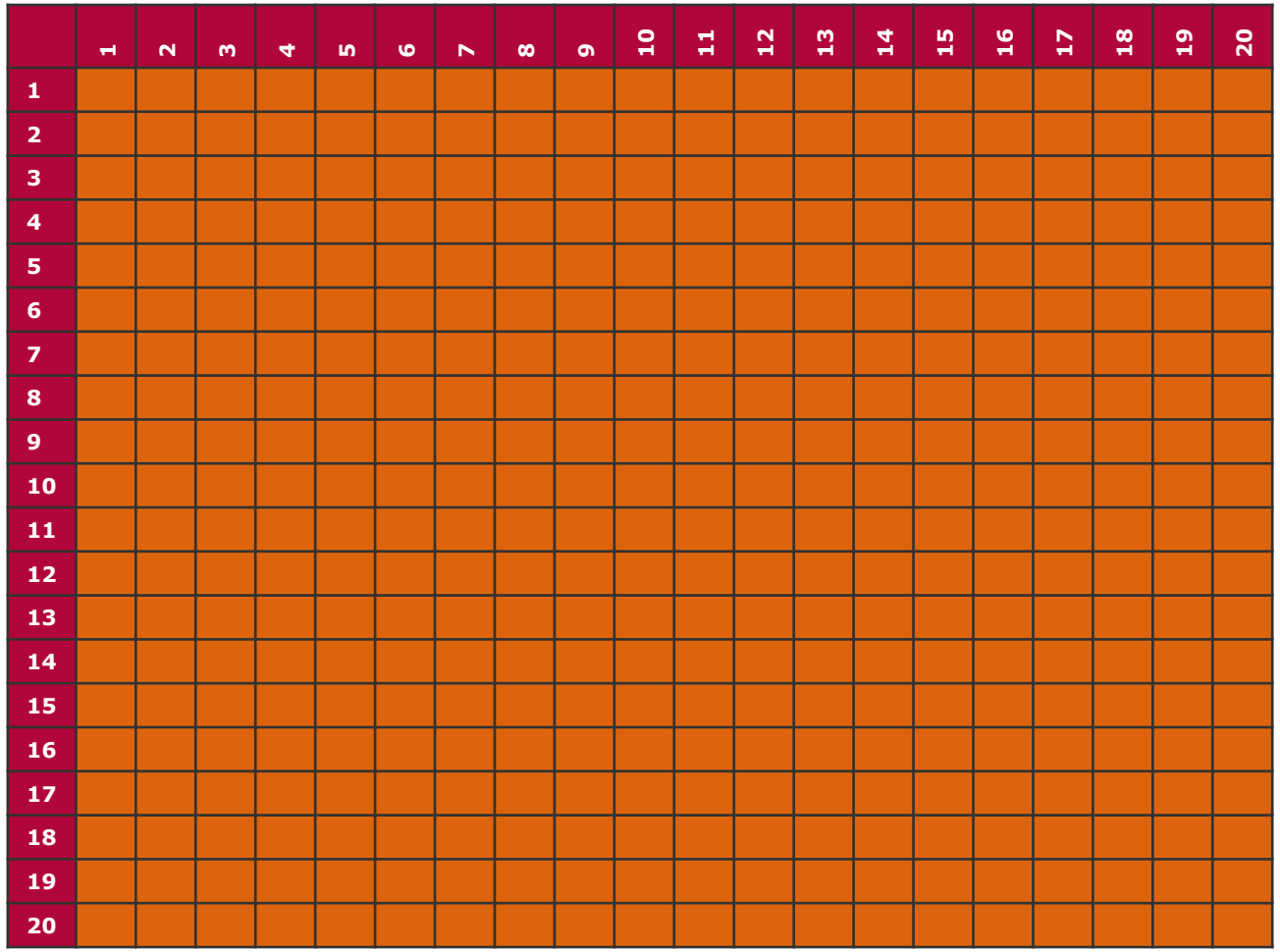
- Problem 1: Representations are not identical.
  - *Fuzzy duplicates*
- Solution: Similarity measures / models
  - Value- and record-comparisons
  - Domain-dependent or domain-independent

- Problem 2: Datasets are large.
  - Quadratic complexity: Comparison of every pair of records.
- Solution: Algorithms
  - E.g., avoid comparisons by partitioning.

Ironically, “Duplicate Detection” has many Duplicates

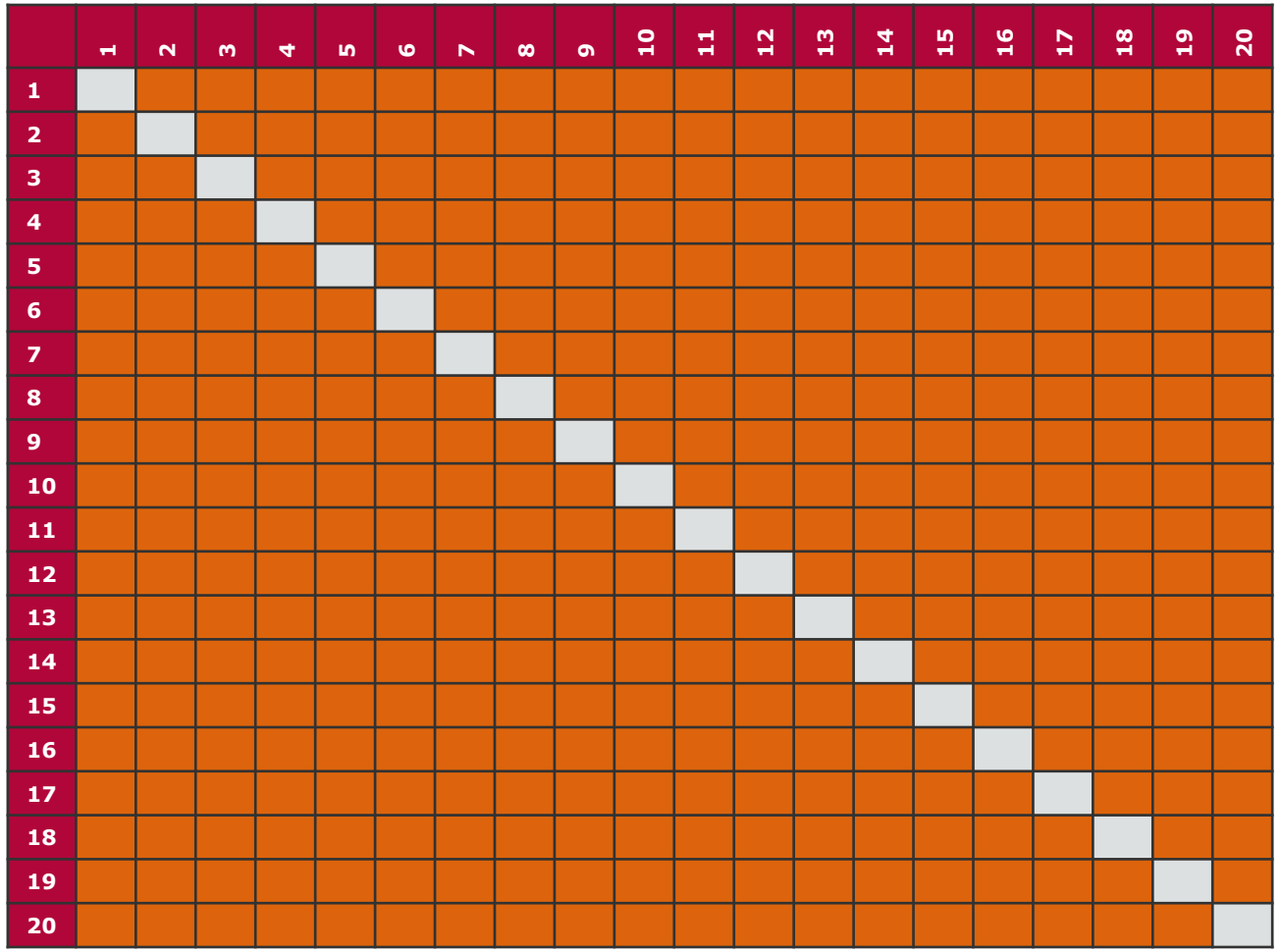


Number of comparisons: All pairs



400
comparisons

Reflexivity of Similarity

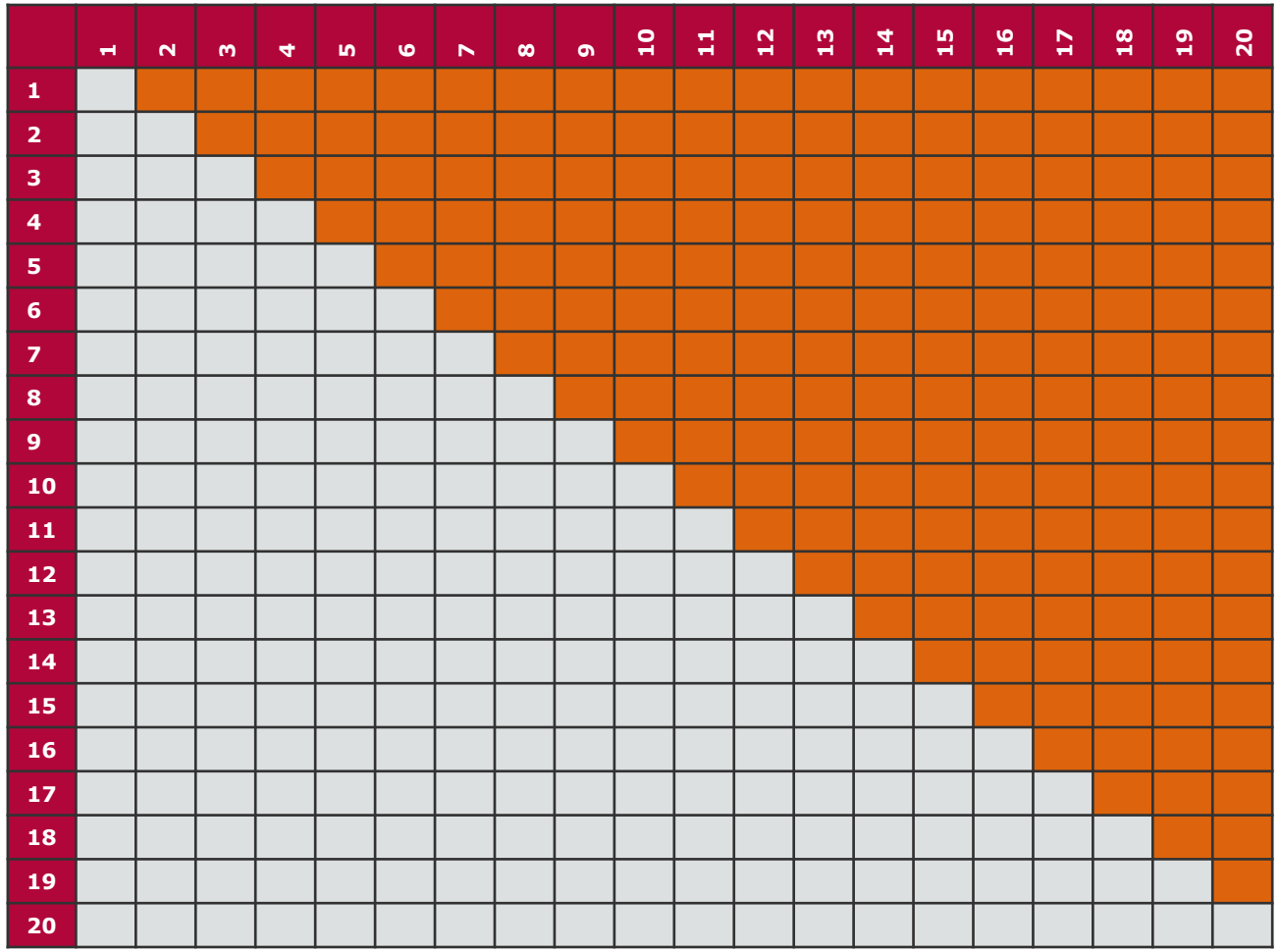


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
2	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
3	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
4	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
5	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
6	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■
7	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■
8	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■
9	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■
10	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■
11	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■
12	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■
13	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■
14	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■
15	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■
16	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■
17	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■
18	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■
19	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■
20	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	

380
comparisons

Felix Naumann
Data Science 2019

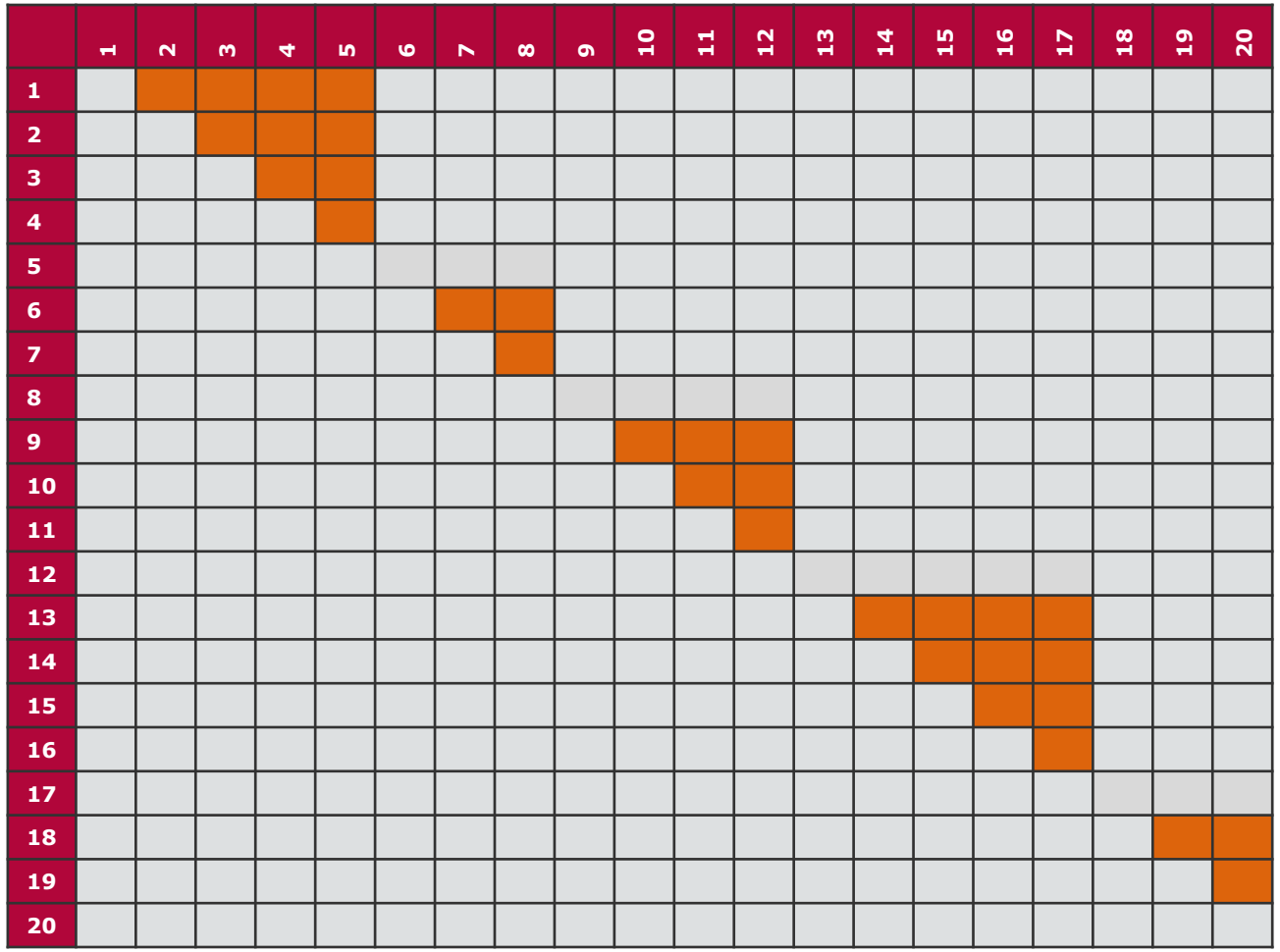
Symmetry of Similarity



190
comparisons

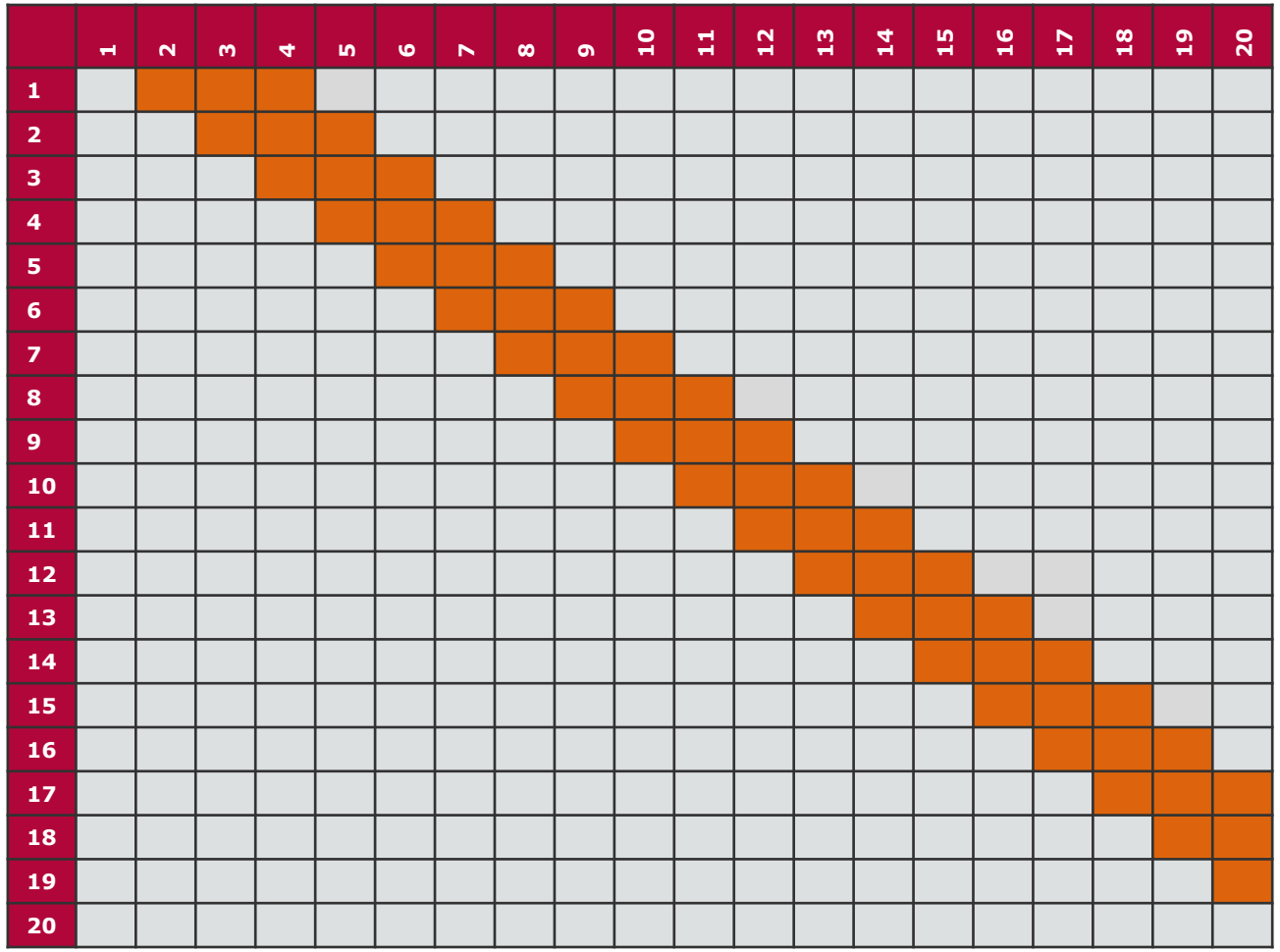
Felix Naumann
Data Science 2019

Blocking by zip-code



32
comparisons

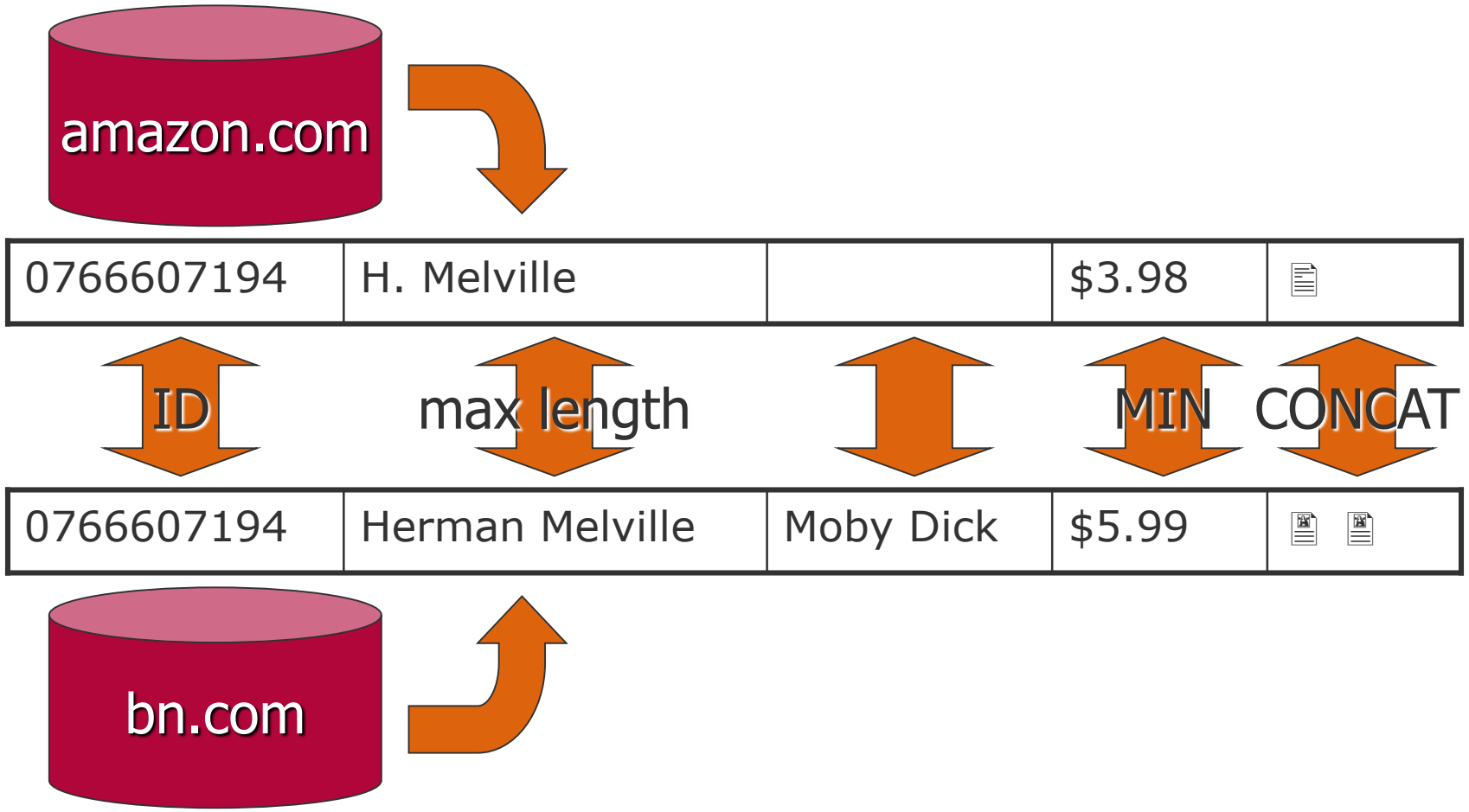
Sorting by zip-code



54
comparisons

Felix Naumann
Data Science 2019

Data Fusion


amazon.com
0766607194
H. Melville
$3.98
ID
max length
MIN
CONCAT
0766607194
Herman Melville
Moby Dick
$5.99
bn.com

Summary

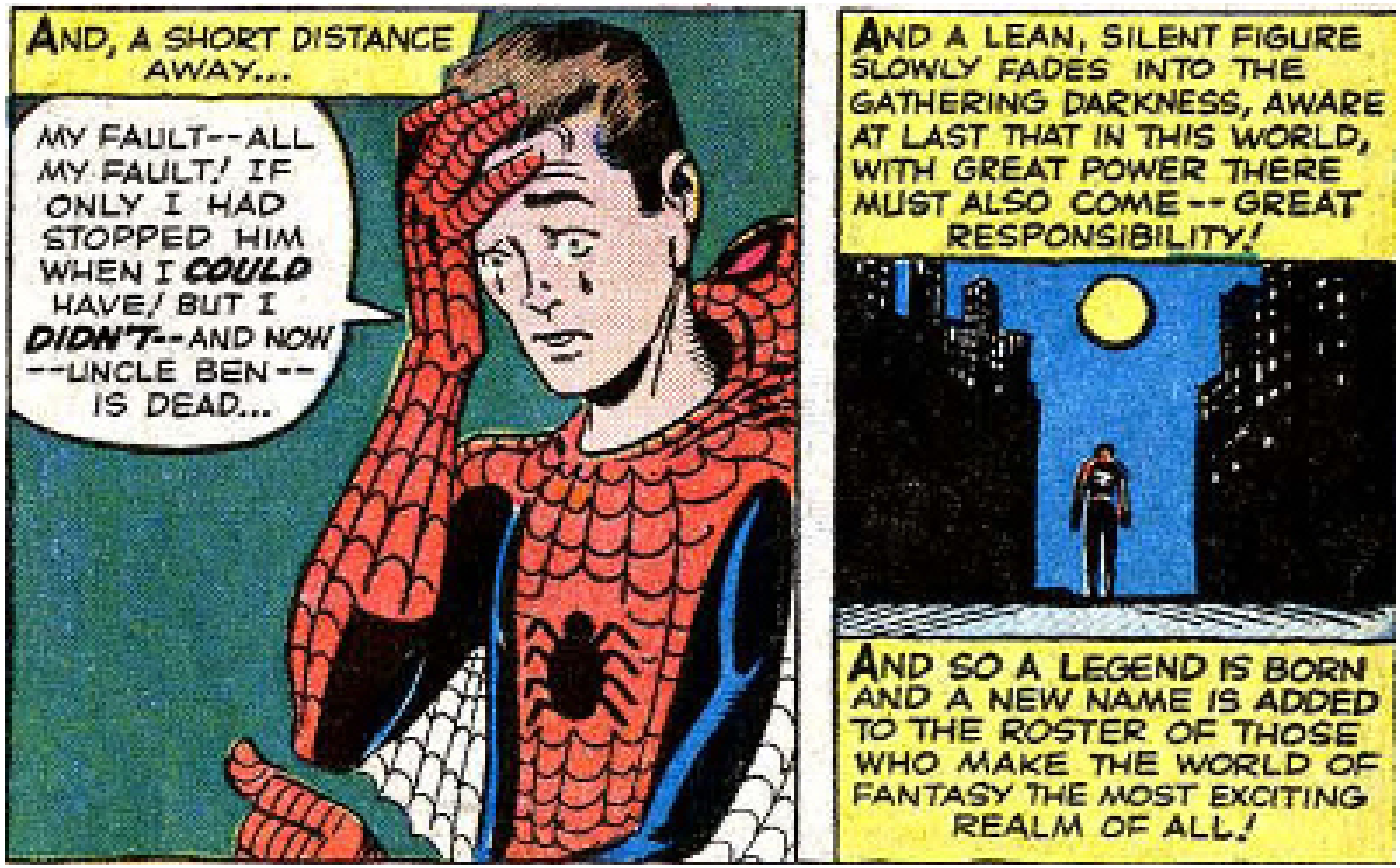
1. Data Science
2. Big Data
3. Data Profiling
4. Data Preparation
5. Data Cleaning

https://unsplash.com/photos/vGefUiWm0xI

Big Data and Ethics

- Industry keynote speakers on credit ratings using big data
  - "If the data is out there, we will find it."
  - "… and that is why I closed my Twitter account."
  - "… and that is why I had my son close his Twitter account."

With Great Power there must also come – Great Responsibility



Spider-Man from *Amazing Fantasy #15,* August 1962

Felix Naumann
Data Science 2019