Data Preparation

The key to successful data science

Lars Grammel
@lgrammel
Head of European R&D, Trifacta

SDS 2016
September 16, 2016
Winterthur, Switzerland


Rolls-Royce


Royal Bank of Scotland

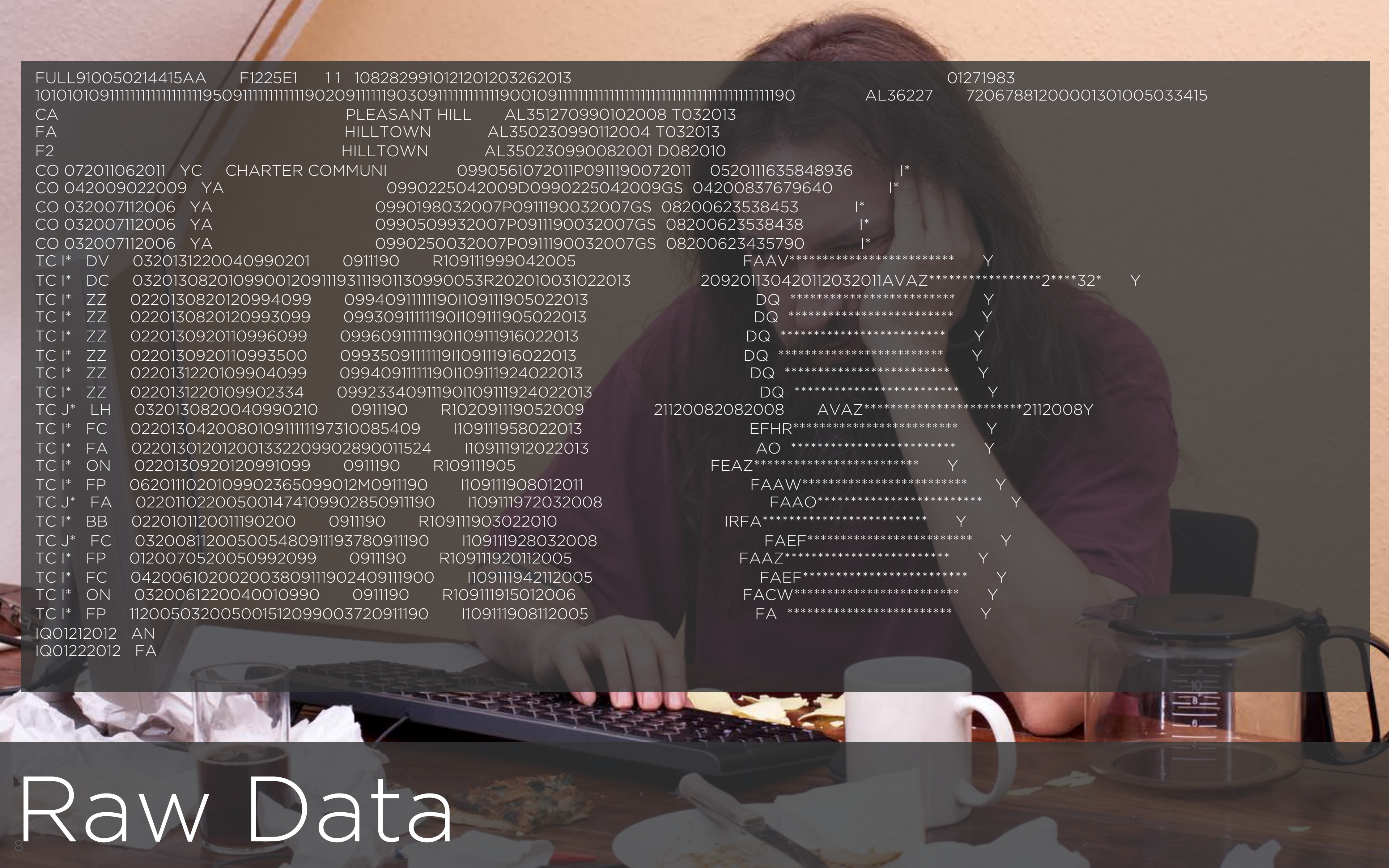
US Elections

The Age of Data Science?

The Reality of Data Science

<MSIDN/IMSI/IME|>	DATETTIME/DURATION/DISCONNECT REASON	MSWICENT:BASCENTCONT:BASTRASTA	CALL_TYPE|CORRES_TYPE/CORRESP_IDN|CORRES2_TYPE/CORRESP2_ISDN
<604711647/208100942278779/44928067108241>	2013-12-28T0:07:47/327/11	MSC001:BSC001:BTS009	MOC|SFR/621630263|/
<604523376/208102203151835/44828688676508>	2013-12-26T11:27:44/309/19	MSC001:BSC001:BTS018	MTC|ORG1/638590539|/
<600225657/208102531594906/44926909793892>	2014-01-01T13:02:25/0/	MSC001:BSC001:BTS018	SMS-MT|SMSC/600000000|BOY/658510643
<603436357/208114615027009/35390401846141>	2013-12-18T14:22:19/0/	MSC001:BSC002:BTS044	SMS-MO|SMSC/600000000|SFR/634989093
<600225639/208102531594888/44926909793874>	2013-12-29T7:31:35/0/	MSC001:BSC002:BTS025	SMS-MO|SMSC/600000000|ORG1/608564604
<600292137/208118290172910/44927465451474>	2013-12-27T17:57:49/323/11	MSC001:BSC002:BTS037	MTC|ORG1/608780693|/
<604502881/208111089907242/33018900056077>	2013-12-29T8:14:21/0/	MSC001:BSC001:BTS016	SMS-MT|SMSC/600000000|ORG1/640114853
<603059144/208105523309620/35570000173463>	2013-12-21T0:19:41/0/	MSC001:BSC001:BTS005	SMS-MO|SMSC/600000000|BOY/659512293
<604704352/208115012761563/35521500051118>	2013-12-30T15:32:16/46/11	MSC001:BSC002:BTS036	MOC3|SRV/600000620|/
<604502875/208111089907236/33018900056071>	2013-12-23T16:22:12/307/11	MSC001:BSC001:BTS007	MOC|SFR/634838805|/
<604761046/208109851577098/44928000179633>	2013-12-23T12:18:35/344/11	MSC001:BSC002:BTS026	MTC|ORG1/607324068|/
<603444901/208108660745208/35358700482241>	2014-01-01T13:25:04/308/11	MSC001:BSC001:BTS017	MTC|SFR/646185386|/
<600212732/208115224596622/35282601228183>	2013-12-22T17:30:07/0/	MSC001:BSC002:BTS025	SMS-MT|SMSC/600000000|ORG1/640378684
<601809398/208119614632187/35044300223784>	2013-12-25T9:24:14/0/	MSC001:BSC001:BTS017	SMS-MO|SMSC/600000000|BOY/600369030
<604715311/208106568375954/52034162631600>	2013-12-20T12:43:25/0/	MSC001:BSC001:BTS010	SMS-MT|SMSC/600000000|ORG1/608916580
<604508776/208118357396586/44919238527884>	2013-12-30T18:20:23/0/	MSC001:BSC002:BTS042	SMS-MO|SMSC/600000000|BOY/600348867
<604715308/208106568375951/52034162631597>	2013-12-29T1:17:49/0/	MSC001:BSC002:BTS044	SMS-MO|SMSC/600000000|BOY/600396332
<603159804/208106585213958/35643301870782>	2013-12-20T20:13:17/0/	MSC001:BSC002:BTS040	SMS-MO|SMSC/600000000|ORG1/607985139
<604715326/208106568375969/52034162631615>	2013-12-30T16:29:49/395/11	MSC001:BSC001:BTS022	MOC|SFR/623164807|/
<601481001/208113515590982/35084880080848>	2013-12-30T13:19:58/0/	MSC001:BSC002:BTS026	SMS-MO|SMSC/600000000|ORG1/638212749
<603436382/208114615027034/35390401846166>	2013-12-31T10:20:33/0/	MSC001:BSC002:BTS032	SMS-MO|SMSC/600000000|ORG1/638860911
<600292132/208118290172905/44927465451469>	2013-12-19T20:55:19/0/	MSC001:BSC002:BTS044	SMS-MT|SMSC/600000000|ORG1/607922426
<600703653/208118948398967/35481101495960>	2014-01-01T18:49:24/0/	MSC001:BSC001:BTS016	SMS-MT|SMSC/600000000|BOY/600306448
<603159824/208106585213978/35643301870802>	2013-12-31T13:49:16/0/	MSC001:BSC001:BTS009	SMS-MT|SMSC/600000000|BOY/666796437

Raw Data



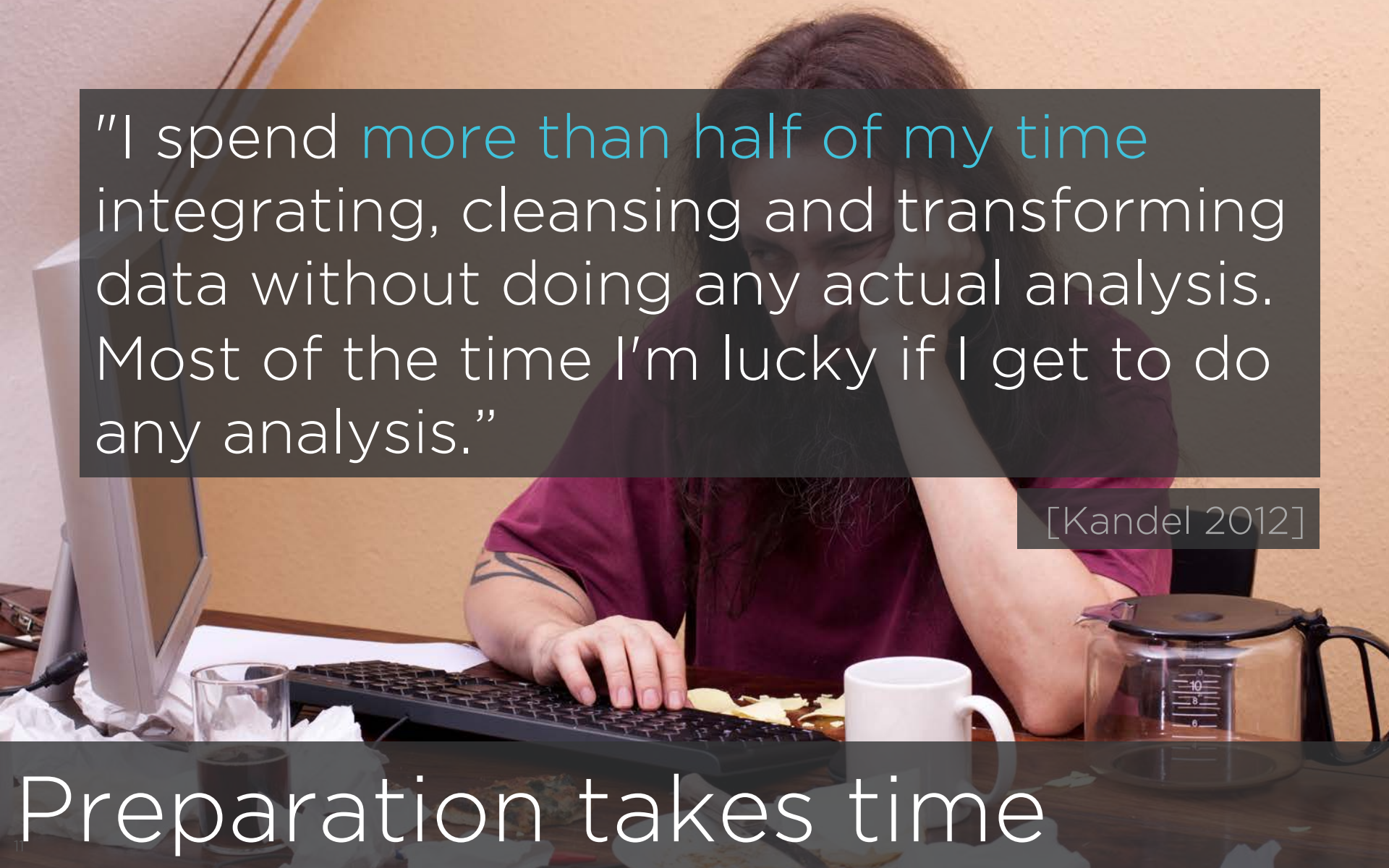
FULL910050214415AA F1225E1 1 1 1082829910121201203262013 01271983
1010101091111111111111111195091111111111190209111111903091111111111119001091111111111111111111111111111111111111190 AL36227 72067881200001301005033415
CA PLEASANT HILL AL351270990102008 T032013
FA HILLTOWN AL350230990112004 T032013
F2 HILLTOWN AL350230990082001 D082010
CO 072011062011 YC CHARTER COMMUNI 0990561072011P0911190072011 0520111635848936 I*
CO 042009022009 YA 0990225042009D0990225042009GS 04200837679640 I*
CO 032007112006 YA 0990198032007P0911190032007GS 08200623538453 I*
CO 032007112006 YA 0990509932007P0911190032007GS 08200623538438 I*
CO 032007112006 YA 0990250032007P0911190032007GS 08200623435790 I*
TC I* DV 0320131220040990201 0911190 R109111999042005 FAAV************************ Y
TC I* DC 0320130820109900120911193111901130990053R202010031022013 2092011304201120320011AVAZ****************2****32* Y
TC I* ZZ 0220130820120994099 0994091111119011091111905022013 DQ ************************ Y
TC I* ZZ 0220130820120993099 0993091111119011091111905022013 DQ ************************ Y
TC I* ZZ 0220130920110996099 0996091111119011091111916022013 DQ ************************ Y
TC I* ZZ 0220130920110993500 0993509111111911091111916022013 DQ ************************ Y
TC I* ZZ 0220131220109904099 0994091111119011091111924022013 DQ ************************ Y
TC I* ZZ 0220131220109902334 0992334091119011091111924022013 DQ ************************ Y
TC J* LH 0320130820040990210 0911190 R102091119052009 21120082082008 AVAZ************************2112008Y
TC I* FC 022013042008010911111197310085409 I109111958022013 EFHR************************ Y
TC I* FA 0220130120120013322099028900 11524 I109111912022013 AO ************************ Y
TC I* ON 0220130920120991099 0911190 R109111905 FEAZ************************ Y
TC I* FP 06201110201099023650990 12M0911190 I109111908012011 FAAW************************ Y
TC J* FA 0220110220050014741099028509 11190 I109111972032008 FAAO************************ Y
TC I* BB 0220101120011190200 0911190 R109111903022010 IRFA************************ Y
TC J* FC 032008112005005480911193780911190 I109111928032008 FAEF************************ Y
TC I* FP 0120070520050992099 0911190 R109111920112005 FAAZ************************ Y
TC I* FC 042006102002003809111902409111900 I109111942112005 FAEF************************ Y
TC I* ON 0320061220040010990 0911190 R109111915012006 FACW************************ Y
TC I* FP 11200503200500151209900372091 1190 I109111908112005 FA ************************ Y
IQ01212012 AN
IQ01222012 FA

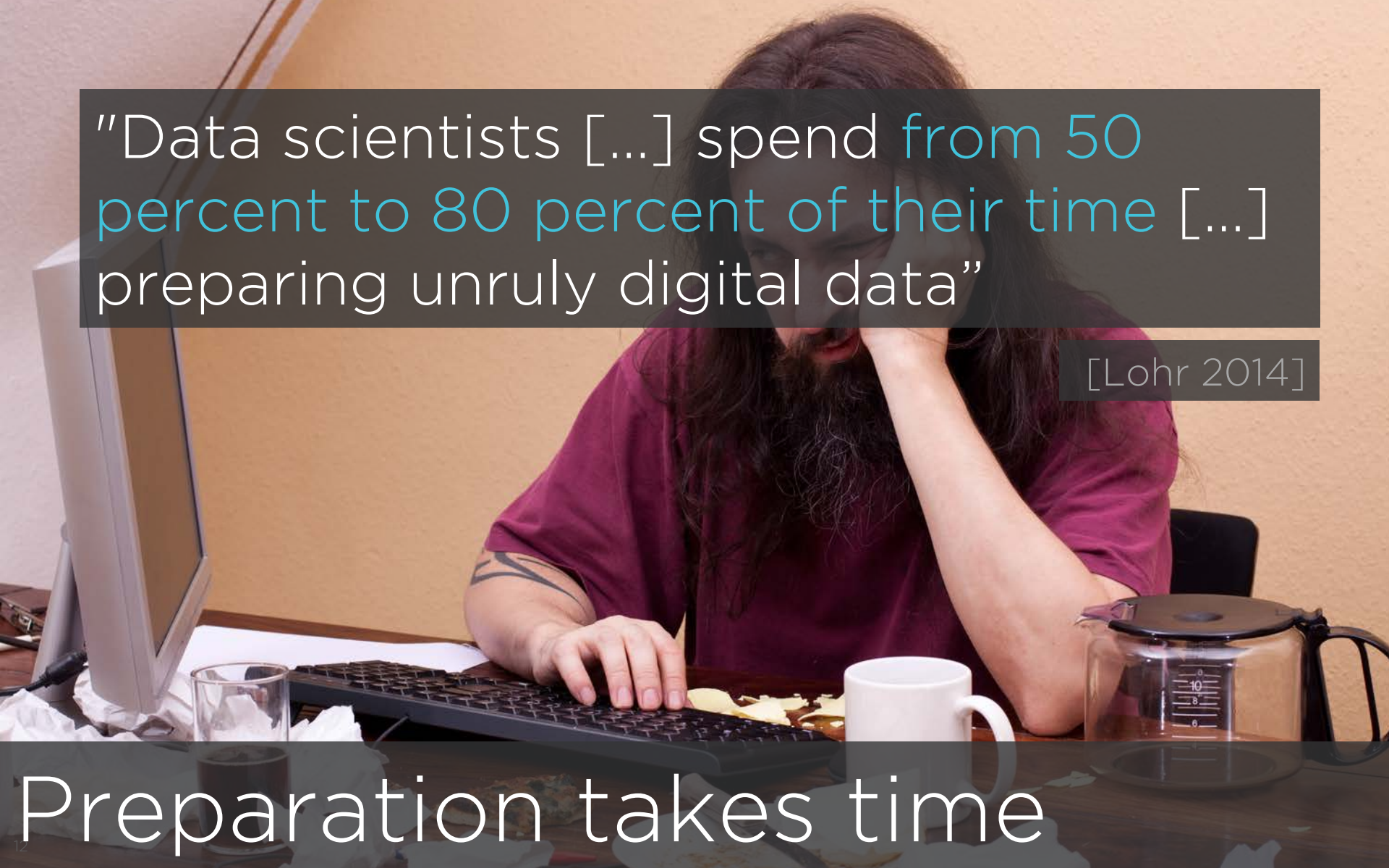
Raw Data

{"channel_type":"linkedin","campaign_date":"06/17/2016 23:47","impact":"28 new followers","product_family":"ABCD CAMPAIGN::LCD/LED FLAT PANEL"}
{"channel_type":"linkedin","campaign_date":"05/30/2016 13:41","impact":"83 new followers","product_family":"ABCD CAMPAIGN::SEASONAL ITEMS"}
{"channel_type":"linkedin","campaign_id_convert":"1PPCR64UEedZeedhqg7AaAQazMJSuyXc7U","campaign_date":"04/24/2016 19:33","impact":"96 new followers","product_family":"ABCD CAMPAIGN::COMPUTER PERIPHERALS"}
{"channel_type":"linkedin","campaign_id_convert":"14SMJXo96qwU5hLKXa1eeFfSHz7rTc6uyk","promo_code":"FREE_X2","campaign_date":"05/23/2016 13:33"}
{"channel_type":"linkedin","campaign_date":"05/06/2016 1:54"}
{"channel_type":"linkedin","campaign_id_convert":"1NzCesZ6K5sdxNB3Zvo7q2AFomfkq5gDUKP","promo_code":"NO_SALES_30","campaign_date":"06/07/2016 23:52","impact":"65 new followers","product_family":"ABCD CAMPAIGN::GAMING HARDWARE"}
{"channel_type":"linkedin","promo_code":"DOUBLE_20","campaign_date":"06/07/2016 2:26","impact":"72 new followers","product_family":"ABCD CAMPAIGN::PORTABLE AUDIO"}
{"channel_type":"linkedin","campaign_date":"04/21/2016 2:53"}
{"channel_type":"linkedin","campaign_id_convert":"1i62LBfH7qsd9P74SwZ497HSvuyMDrnMd","campaign_date":"05/12/2016 19:17","impact":"78 new followers","product_family":"ABCD CAMPAIGN::GAMING SOFTWARE"}
{"channel_type":"linkedin","campaign_id_convert":"1B9BMNSUFBsdf97xCpM2GwDNghDgSKDizH","campaign_date":"05/28/2016 17:38","impact":"96 new followers","product_family":"ABCD CAMPAIGN::PLASMA ACCESSORIES"}
{"channel_type":"linkedin","campaign_date":"05/10/2016 8:40","impact":"54 new followers","product_family":"ABCD CAMPAIGN::LCD/LED FLAT PANEL"}
{"channel_type":"linkedin","campaign_id_convert":"1CiK2dhLdJfeWD1dZKAmqnj9D4rf78xs8y","campaign_date":"04/17/2016 2:14","impact":"71 new followers","product_family":"ABCD CAMPAIGN::DIGITAL CAMERA"}
{"channel_type":"linkedin","promo_code":"1DAY_10","campaign_date":"04/02/2016 8:03","impact":"79 new followers","product_family":"ABCD CAMPAIGN::LCD/LED FLAT PANEL"}
{"channel_type":"LinkedIn","campaign_date":"06/01/2016 5:51","impact":"96 new followers","product_family":"ABCD CAMPAIGN::SEASONAL ITEMS"}
{"channel_type":"LinkedIn","campaign_date":"04/19/2016 6:34","impact":"88 new followers","product_family":"ABCD CAMPAIGN::PROJECTION TV"}
{"channel_type":"LinkedIn","campaign_id_convert":"1CiK2dhLdJfeWD1dZKAmqnj9D4rf78xs8y","campaign_date":"03/13/2016 14:58","impact":"32 new followers","product_family":"ABCD CAMPAIGN::PLASMA ACCESSORIES"}
{"channel_type":"LinkedIn","campaign_date":"04/01/2016 22:06","impact":"83 new followers","product_family":"ABCD CAMPAIGN::PRINTER"}
{"channel_type":"LinkedIn","campaign_id_convert":"1AKmgwgowv8sdozDL92faNhTBYLeAHW8GaP","campaign_date":"03/26/2016 23:08","impact":"87 new followers","product_family":"ABCD CAMPAIGN::GAMING HARDWARE"}
{"channel_type":"LinkedIn","campaign_date":"03/19/2016 3:52","impact":"83 new followers","product_family":"ABCD CAMPAIGN::COMPUTER PERIPHERALS"}
{"channel_type":"LinkedIn","campaign_date":"05/22/2016 16:09","impact":"82 new followers","product_family":"ABCD CAMPAIGN::DIGITAL CAMERA"}
{"channel_type":"LinkedIn","promo_code":"IMP_MISSION_A","campaign_date":"03/17/2016 13:06","impact":"72 new followers","product_family":"ABCD CAMPAIGN::HEALTH \u0026 FITNESS"}
{"channel_type":"LinkedIn","campaign_date":"06/11/2015 16:00","impact":"80 new followers","product_family":"ABCD CAMPAIGN::LCD/LED FLAT PANEL"}
{"channel_type":"LinkedIn","campaign_date":"03/15/2015 18:46","impact":"105 new followers","product_family":"ABCD CAMPAIGN::SEASONAL ITEMS"}

Raw Data

Preparation takes time


"I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any analysis."
[Kandel 2012]
Preparation takes time



"Data scientists [...] spend from 50 percent to 80 percent of their time [...] preparing unruly digital data"

[Lohr 2014]

Preparation takes time

Data is wasted

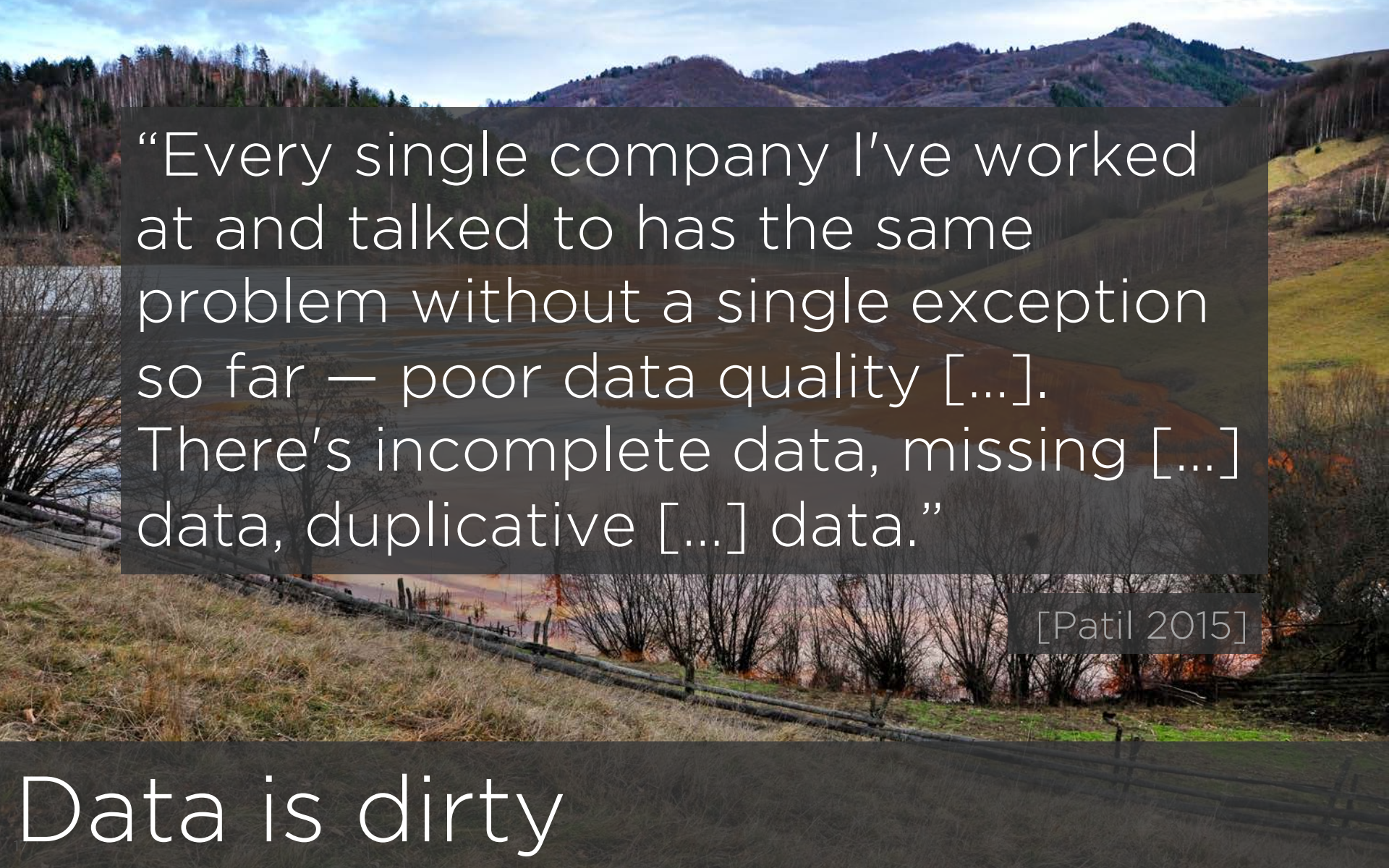
Data is wasted

"On average, organizations only use 28% of their semi-structured and 31% of their unstructured data."

[Forrester 2015]

Data is wasted

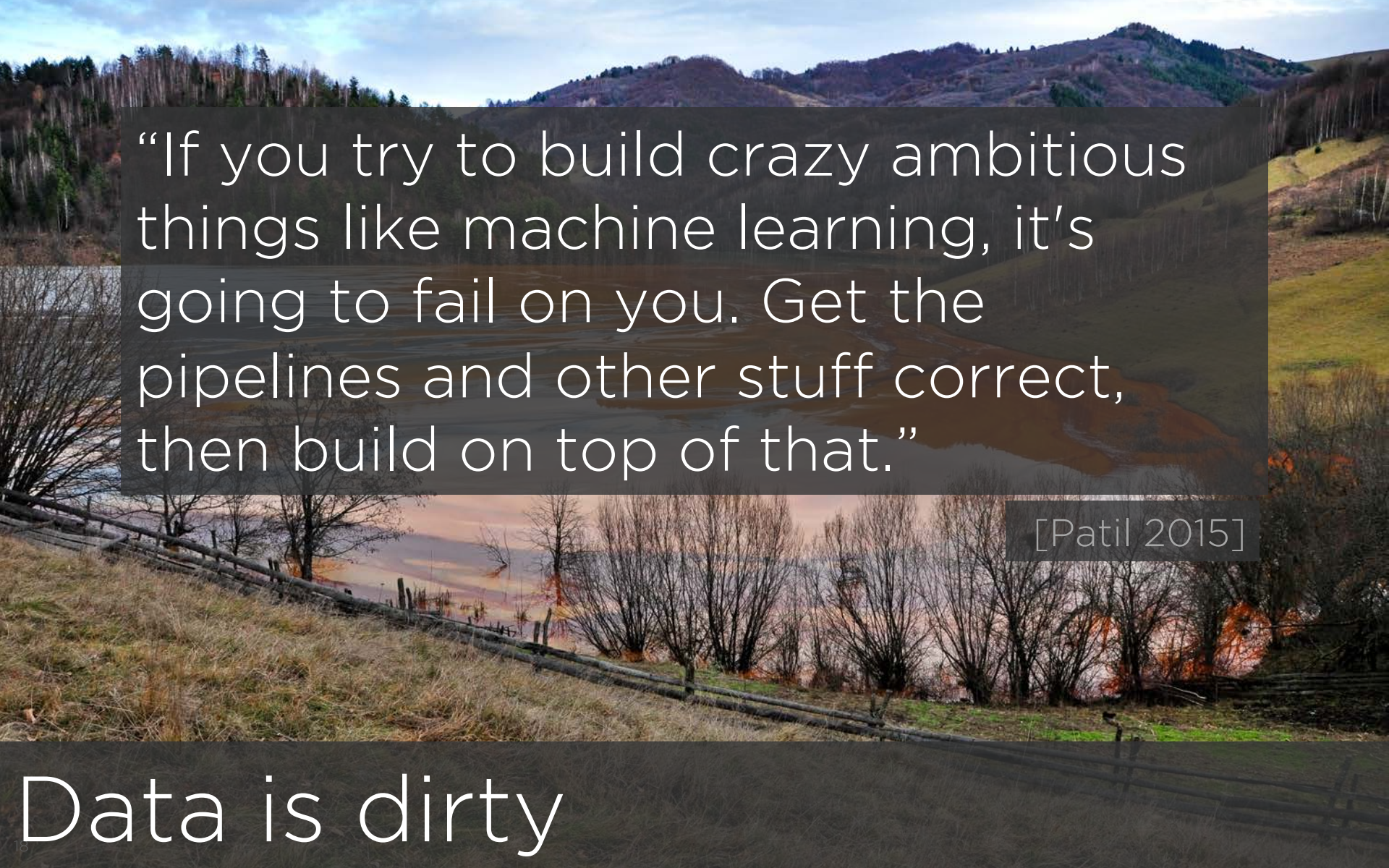
Data is dirty



"Every single company I've worked at and talked to has the same problem without a single exception so far — poor data quality [...]. There's incomplete data, missing [...] data, duplicative [...] data."

[Patil 2015]

Data is dirty



Data is dirty


• 50-80% of time spent on preparation
• only <= ~40% of data is being used
• poor data quality affects outcomes
The Reality of Data Science

Data Preparation Activities

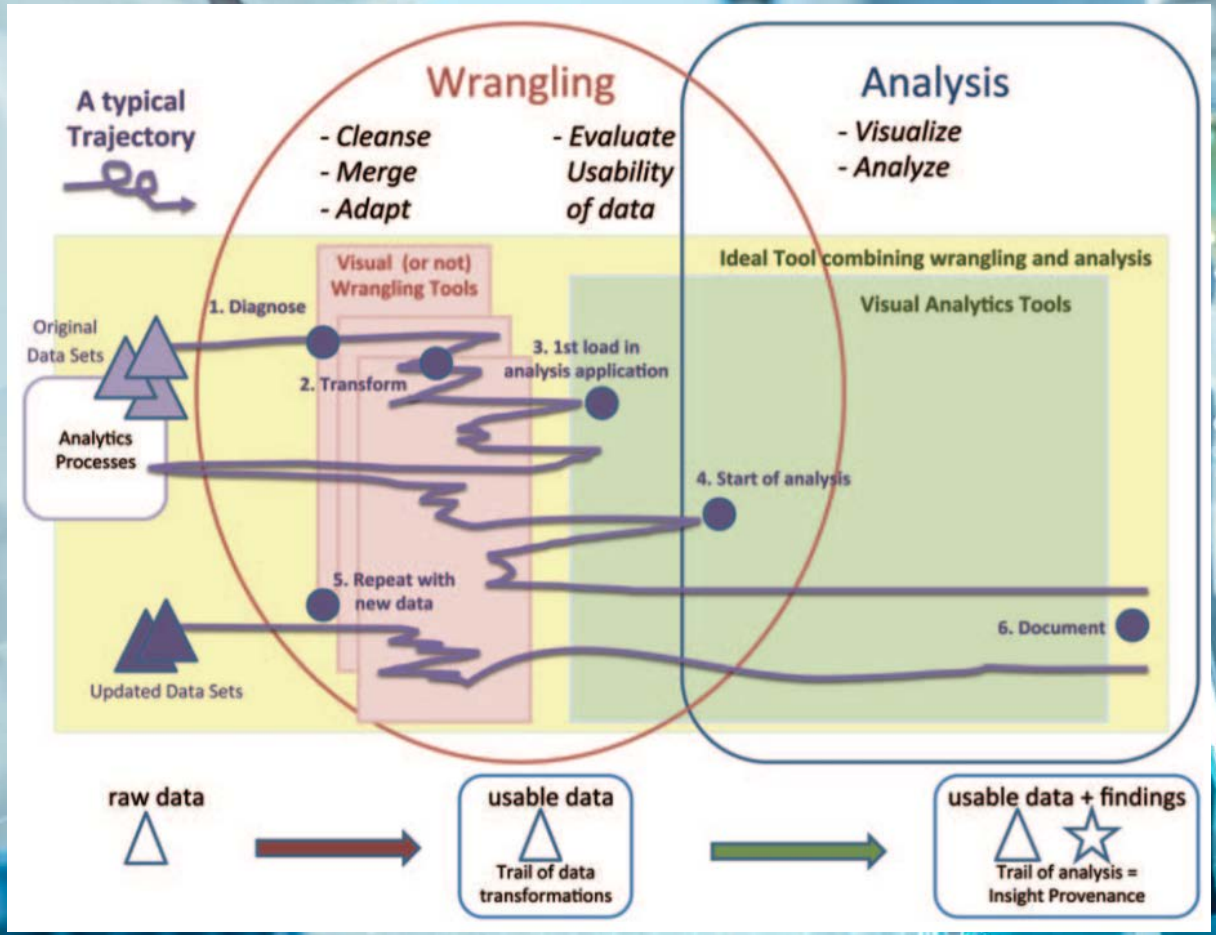
Discovery

Structuring

Cleaning

Enriching

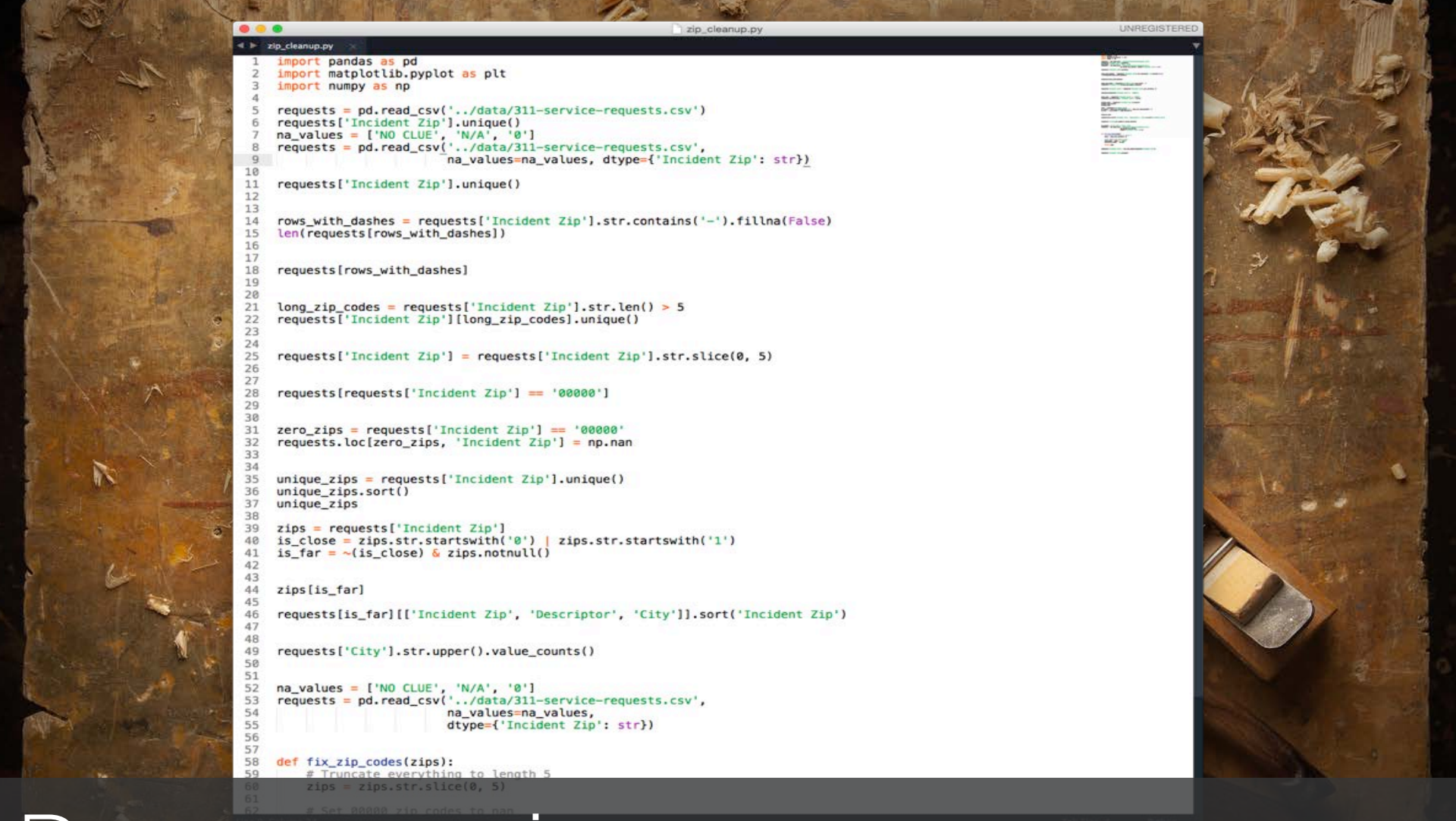
Validating



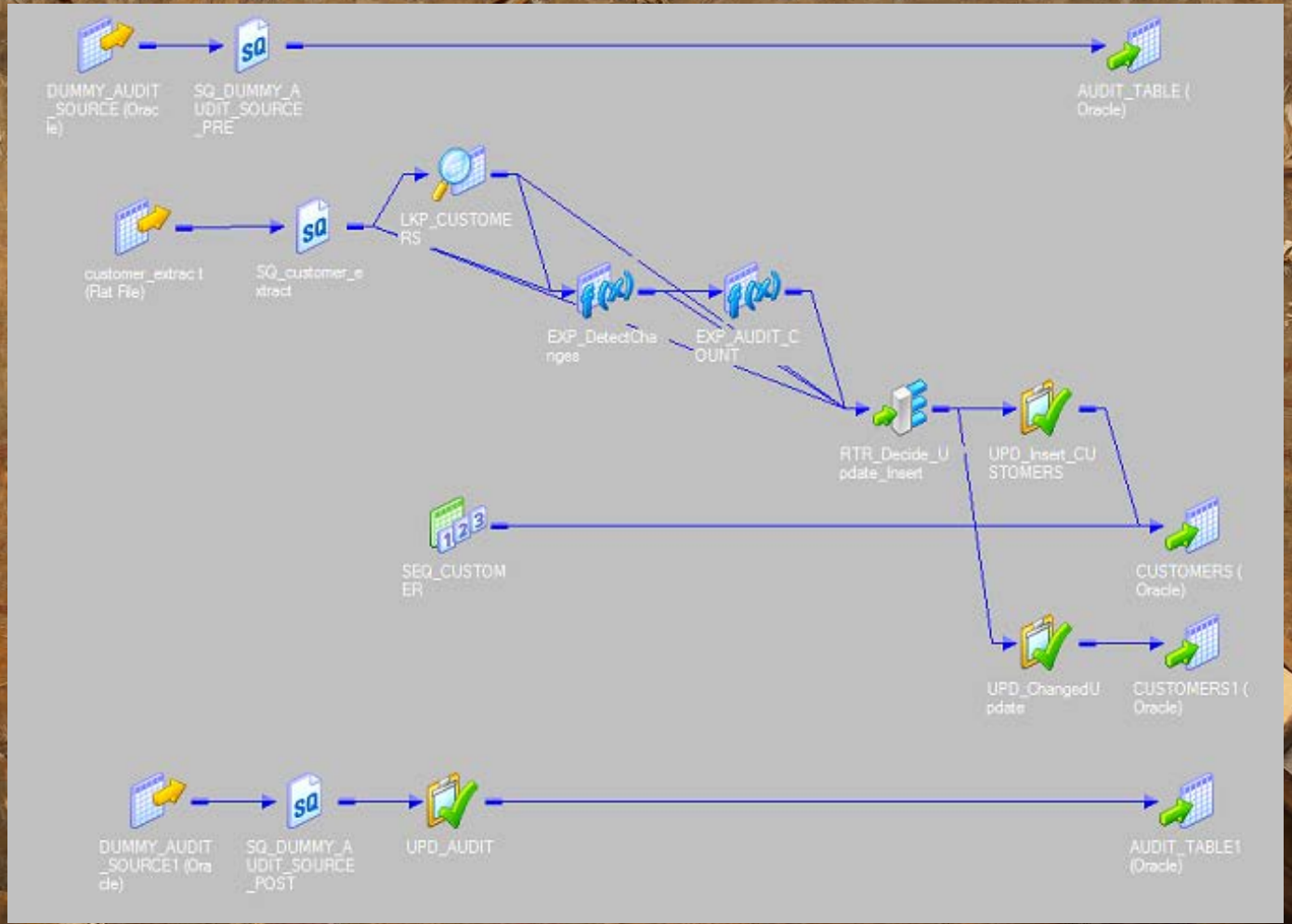
[Kandel et al 2011]

Data Preparation Process

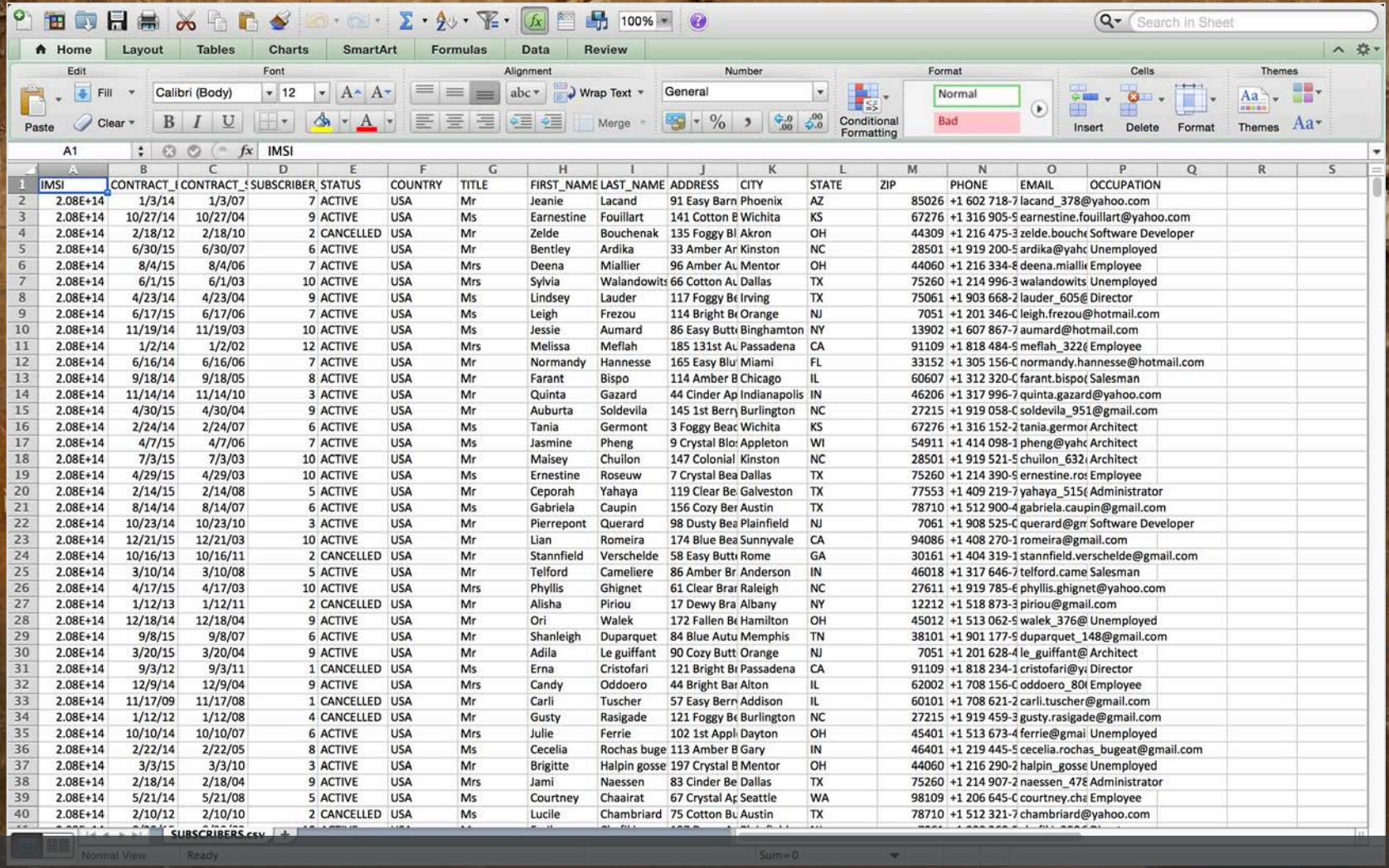
Data Preparation User Interfaces



Programming



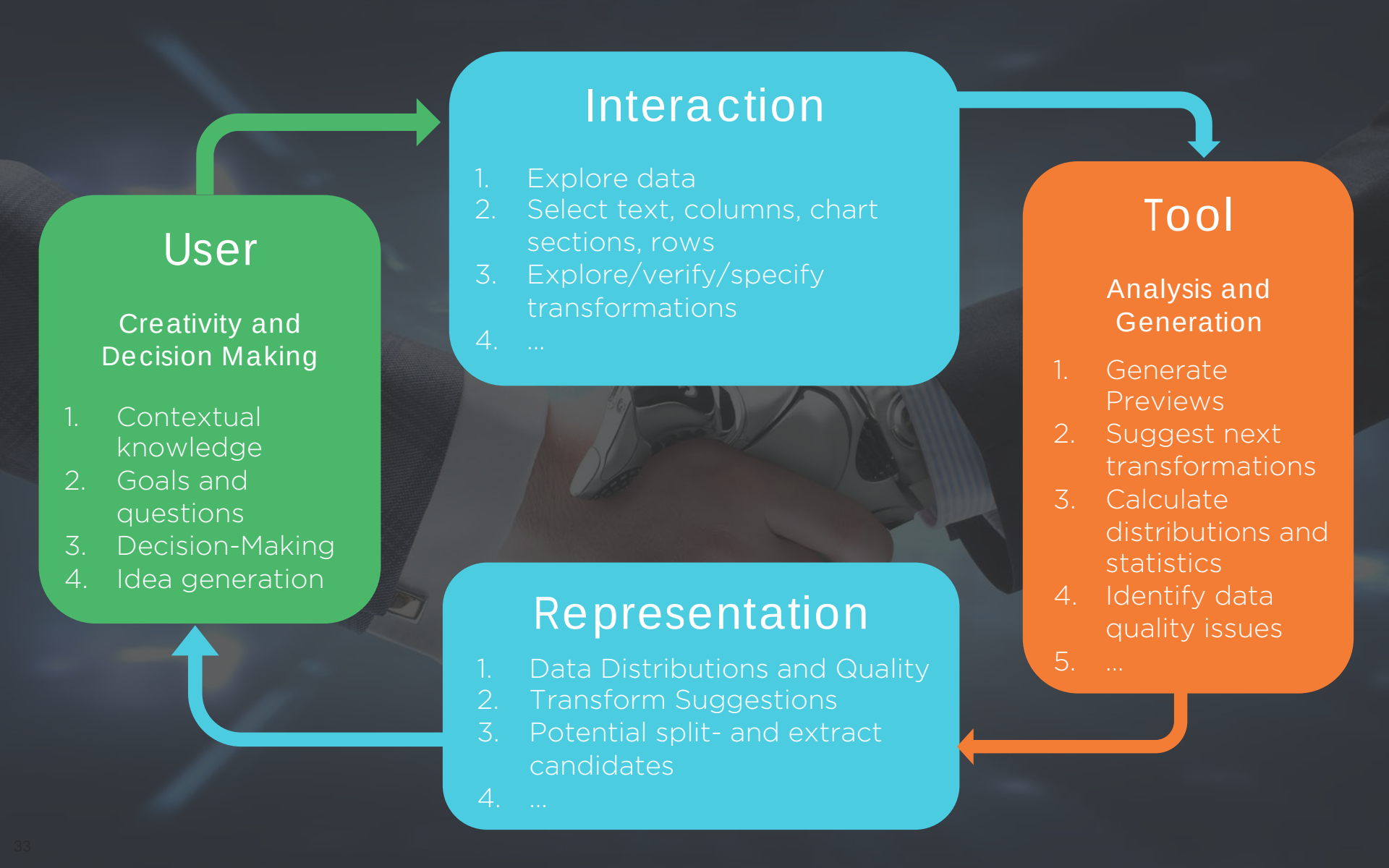
Technical Workflow Mapping

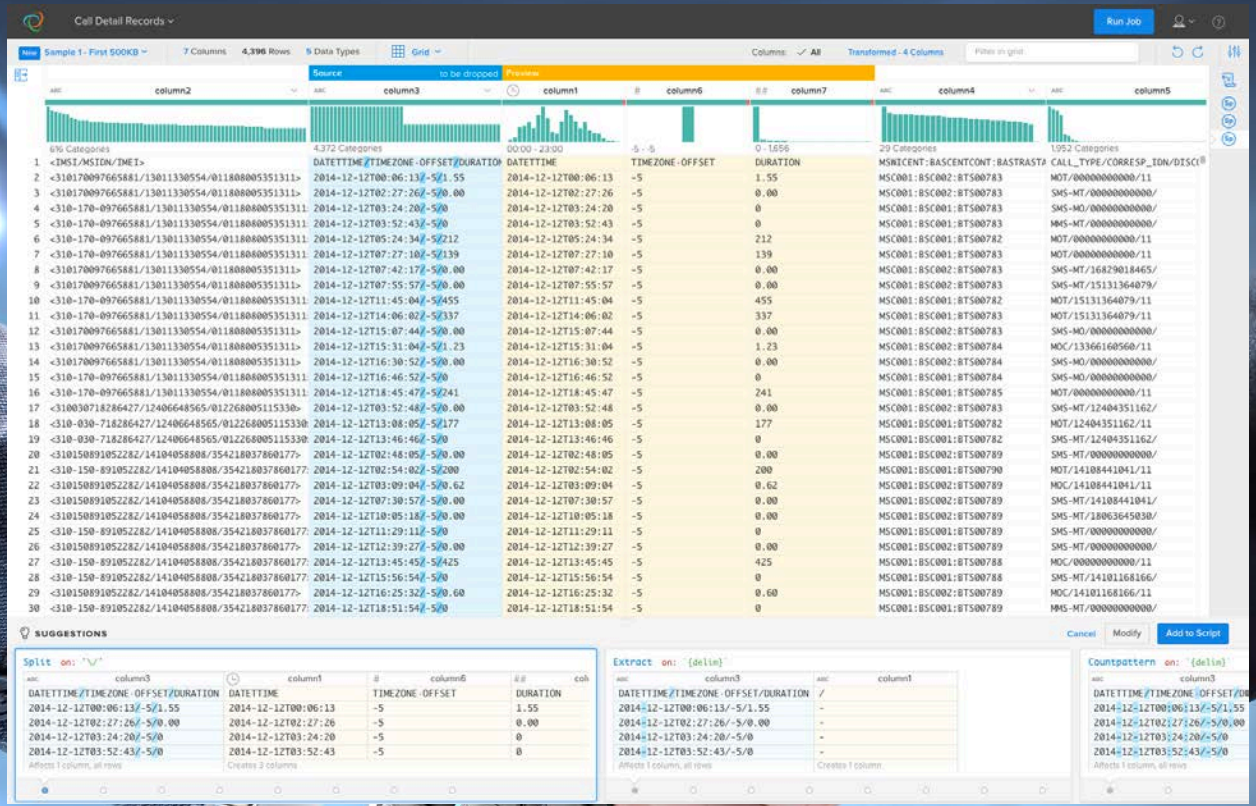

Home
Layout
Tables
Charts
SmartArt
Formulas
Data
Review
Search in Sheet
Calibri (Body)
12
Wrap Text
General
Normal
Bad
Conditional Formatting
Insert
Delete
Format
Themes
Paste
Fill
Clear
Merge
A1
IMSI
IMSI CONTRACT_I CONTRACT_S SUBSCRIBER_ STATUS COUNTRY TITLE FIRST_NAME LAST_NAME ADDRESS CITY STATE ZIP PHONE EMAIL OCCUPATION
2.08E+14 1/3/14 1/3/07 7 ACTIVE USA Mr Jeanie Lacand 91 Easy Barn Phoenix AZ 85026 +1 602 718-7 lacand_378@yahoo.com
2.08E+14 10/27/14 10/27/04 9 ACTIVE USA Ms Earnestine Fouillart 141 Cotton B Wichita KS 67276 +1 316 905-5 earnestine.fouillart@yahoo.com
2.08E+14 2/18/12 2/18/10 2 CANCELLED USA Mr Zelde Bouchenak 135 Foggy Bl Akron OH 44309 +1 216 475-3 zelde.bouche Software Developer
2.08E+14 6/30/15 6/30/07 6 ACTIVE USA Mr Bentley Ardika 33 Amber Ar Kinston NC 28501 +1 919 200-5 ardika@yahc Unemployed
2.08E+14 8/4/15 8/4/06 7 ACTIVE USA Mrs Deena Miallier 96 Amber Au Mentor OH 44060 +1 216 334-8 deena.miallie Employee
2.08E+14 6/1/15 6/1/03 10 ACTIVE USA Mrs Sylvia Walandowits 66 Cotton Au Dallas TX 75260 +1 214 996-3 walandowits Unemployed
2.08E+14 4/23/14 4/23/04 9 ACTIVE USA Ms Lindsey Lauder 117 Foggy Be Irving TX 75061 +1 903 668-2 lauder_605@ Director
2.08E+14 6/17/15 6/17/06 7 ACTIVE USA Ms Leigh Frezou 114 Bright Be Orange NJ 7051 +1 201 346-C leigh.frezou@hotmail.com
2.08E+14 11/19/14 11/19/03 10 ACTIVE USA Ms Jessie Aumard 86 Easy Butt Binghamton NY 13902 +1 607 867-7 aumard@hotmail.com
2.08E+14 1/2/14 1/2/02 12 ACTIVE USA Mrs Melissa Meflah 185 131st Au Passadena CA 91109 +1 818 484-9 meflah_322(Employee
2.08E+14 6/16/14 6/16/06 7 ACTIVE USA Mr Normandy Hannesse 165 Easy Blu Miami FL 33152 +1 305 156-C normandy.hannesse@hotmail.com
2.08E+14 9/18/14 9/18/05 8 ACTIVE USA Mr Farant Bispo 114 Amber B Chicago IL 60607 +1 312 320-C farant.bispo(Salesman
2.08E+14 11/14/14 11/14/10 3 ACTIVE USA Mr Quinta Gazard 44 Cinder Ap Indianapolis IN 46206 +1 317 996-7 quinta.gazard@yahoo.com
2.08E+14 4/30/15 4/30/04 9 ACTIVE USA Mr Auburta Soldevila 145 1st Berry Burlington NC 27215 +1 919 058-C soldevila_951@gmail.com
2.08E+14 2/24/14 2/24/07 6 ACTIVE USA Ms Tania Germont 3 Foggy Beac Wichita KS 67276 +1 316 152-2 tania.germor Architect
2.08E+14 4/7/15 4/7/06 7 ACTIVE USA Ms Jasmine Pheng 9 Crystal Blo Appleton WI 54911 +1 414 098-1 pheng@yahc Architect
2.08E+14 7/3/15 7/3/03 10 ACTIVE USA Mr Maisey Chuilon 147 Colonial Kinston NC 28501 +1 919 521-5 chuilon_632(Architect
2.08E+14 4/29/15 4/29/03 10 ACTIVE USA Ms Ernestine Roseuw 7 Crystal Bea Dallas TX 75260 +1 214 390-9 ernestine.ros Employee
2.08E+14 2/14/15 2/14/08 5 ACTIVE USA Mr Ceporah Yahaya 119 Clear Be Galveston TX 77553 +1 409 219-7 yahaya_515(Administrator
2.08E+14 8/14/14 8/14/07 6 ACTIVE USA Ms Gabriela Caupin 156 Cozy Ber Austin TX 78710 +1 512 900-4 gabriela.caupin@gmail.com
2.08E+14 10/23/14 10/23/10 3 ACTIVE USA Mr Pierrepont Querard 98 Dusty Bea Plainfield NJ 7061 +1 908 525-C querard@gm Software Developer
2.08E+14 12/21/15 12/21/03 10 ACTIVE USA Mr Lian Romeira 174 Blue Bea Sunnyvale CA 94086 +1 408 270-1 romeira@gmail.com
2.08E+14 10/16/13 10/16/11 2 CANCELLED USA Mr Stannfield Verschelde 58 Easy Butt Rome GA 30161 +1 404 319-1 stannfield.verschelde@gmail.com
2.08E+14 3/10/14 3/10/08 5 ACTIVE USA Mr Telford Cameliere 86 Amber Br Anderson IN 46018 +1 317 646-7 telford.came Salesman
2.08E+14 4/17/15 4/17/03 10 ACTIVE USA Mrs Phyllis Ghignet 61 Clear Brar Raleigh NC 27611 +1 919 785-6 phyllis.ghignet@yahoo.com
2.08E+14 1/12/13 1/12/11 2 CANCELLED USA Mr Alisha Piriou 17 Dewy Bra Albany NY 12212 +1 518 873-3 piriou@gmail.com
2.08E+14 12/18/14 12/18/04 9 ACTIVE USA Mr Ori Walek 172 Fallen Be Hamilton OH 45012 +1 513 062-9 walek_376@ Unemployed
2.08E+14 9/8/15 9/8/07 6 ACTIVE USA Mr Shanleigh Duparquet 84 Blue Autu Memphis TN 38101 +1 901 177-9 duparquet_148@gmail.com
2.08E+14 3/20/15 3/20/04 9 ACTIVE USA Mr Adila Le guiffant 90 Cozy Butt Orange NJ 7051 +1 201 628-4 le_guiffant@ Architect
2.08E+14 9/3/12 9/3/11 1 CANCELLED USA Ms Erna Cristofari 121 Bright Br Passadena CA 91109 +1 818 234-1 cristofari@y Director
2.08E+14 12/9/14 12/9/04 9 ACTIVE USA Mrs Candy Oddoero 44 Bright Bar Alton IL 62002 +1 708 156-C oddoero_80(Employee
2.08E+14 11/17/09 11/17/08 1 CANCELLED USA Mr Carli Tuscher 57 Easy Berry Addison IL 60101 +1 708 621-2 carli.tuscher@gmail.com
2.08E+14 1/12/12 1/12/08 4 CANCELLED USA Mr Gusty Rasigade 121 Foggy Be Burlington NC 27215 +1 919 459-3 gusty.rasigade@gmail.com
2.08E+14 10/10/14 10/10/07 6 ACTIVE USA Mrs Julie Ferrie 102 1st Appl Dayton OH 45401 +1 513 673-4 ferrie@gmai Unemployed
2.08E+14 2/22/14 2/22/05 8 ACTIVE USA Ms Cecelia Rochas buge 113 Amber B Gary IN 46401 +1 219 445-5 cecelia.rochas_bugeat@gmail.com
2.08E+14 3/3/15 3/3/10 3 ACTIVE USA Mr Brigitte Halpin gosse 197 Crystal B Mentor OH 44060 +1 216 290-2 halpin_gosse Unemployed
2.08E+14 2/18/14 2/18/04 9 ACTIVE USA Mrs Jami Naessen 83 Cinder Be Dallas TX 75260 +1 214 907-2 naessen_478 Administrator
2.08E+14 5/21/14 5/21/08 5 ACTIVE USA Ms Courtney Chaairat 67 Crystal Ap Seattle WA 98109 +1 206 645-C courtney.cha Employee
2.08E+14 2/10/12 2/10/10 2 CANCELLED USA Ms Lucile Chambriard 75 Cotton Bu Austin TX 78710 +1 512 321-7 chambriard@yahoo.com
SUBSCRIBERS.csv
Normal View
Ready
Sum=0

Excel

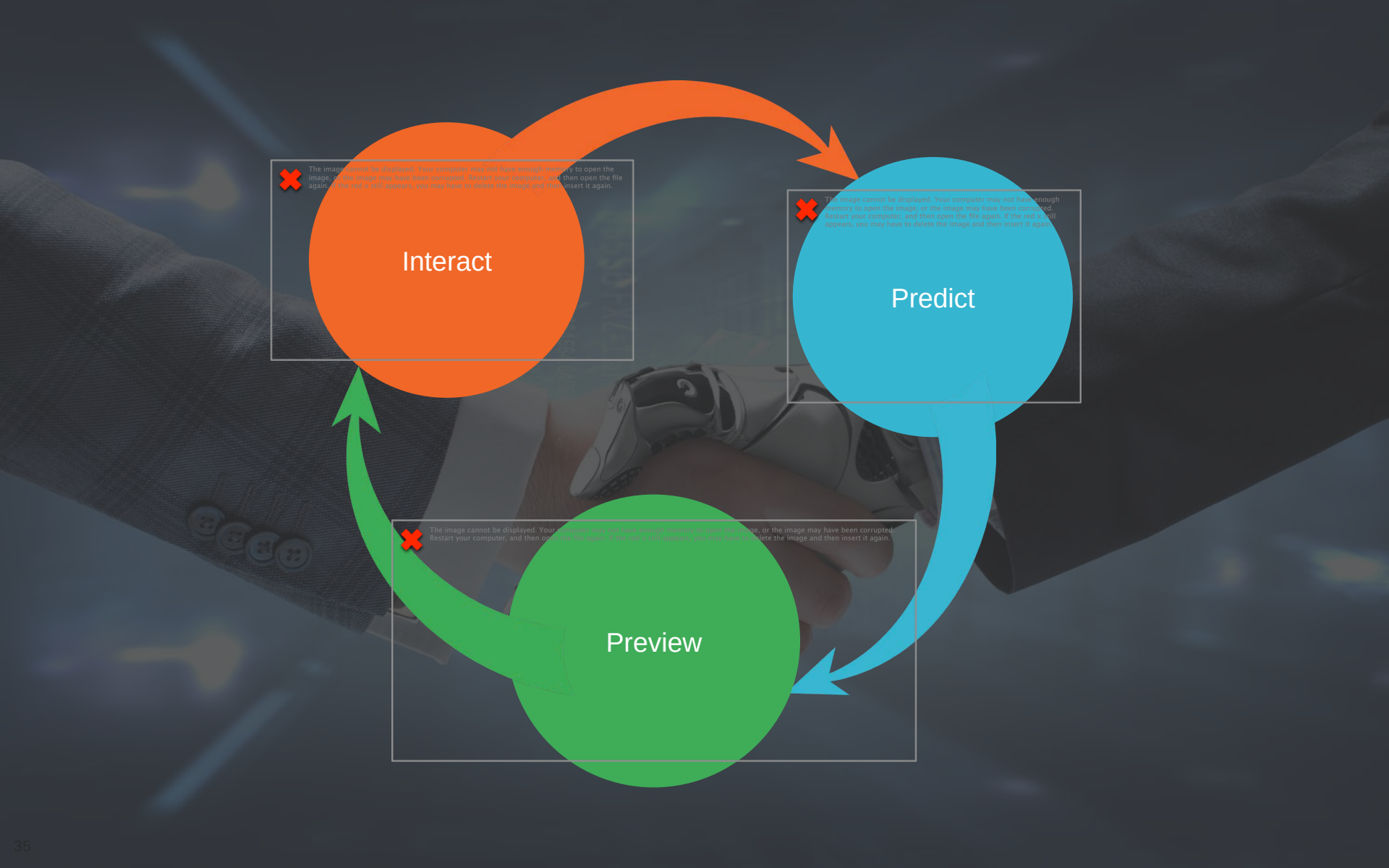
We need to rethink our UIs

We need to rethink our UIs


User
Creativity and Decision Making
1. Contextual knowledge
2. Goals and questions
3. Decision-Making
4. Idea generation
Interaction
1. Explore data
2. Select text, columns, chart sections, rows
3. Explore/verify/specify transformations
4. ...
Tool
Analysis and Generation
1. Generate Previews
2. Suggest next transformations
3. Calculate distributions and statistics
4. Identify data quality issues
5. ...
Representation
1. Data Distributions and Quality
2. Transform Suggestions
3. Potential split- and extract candidates
4. ...

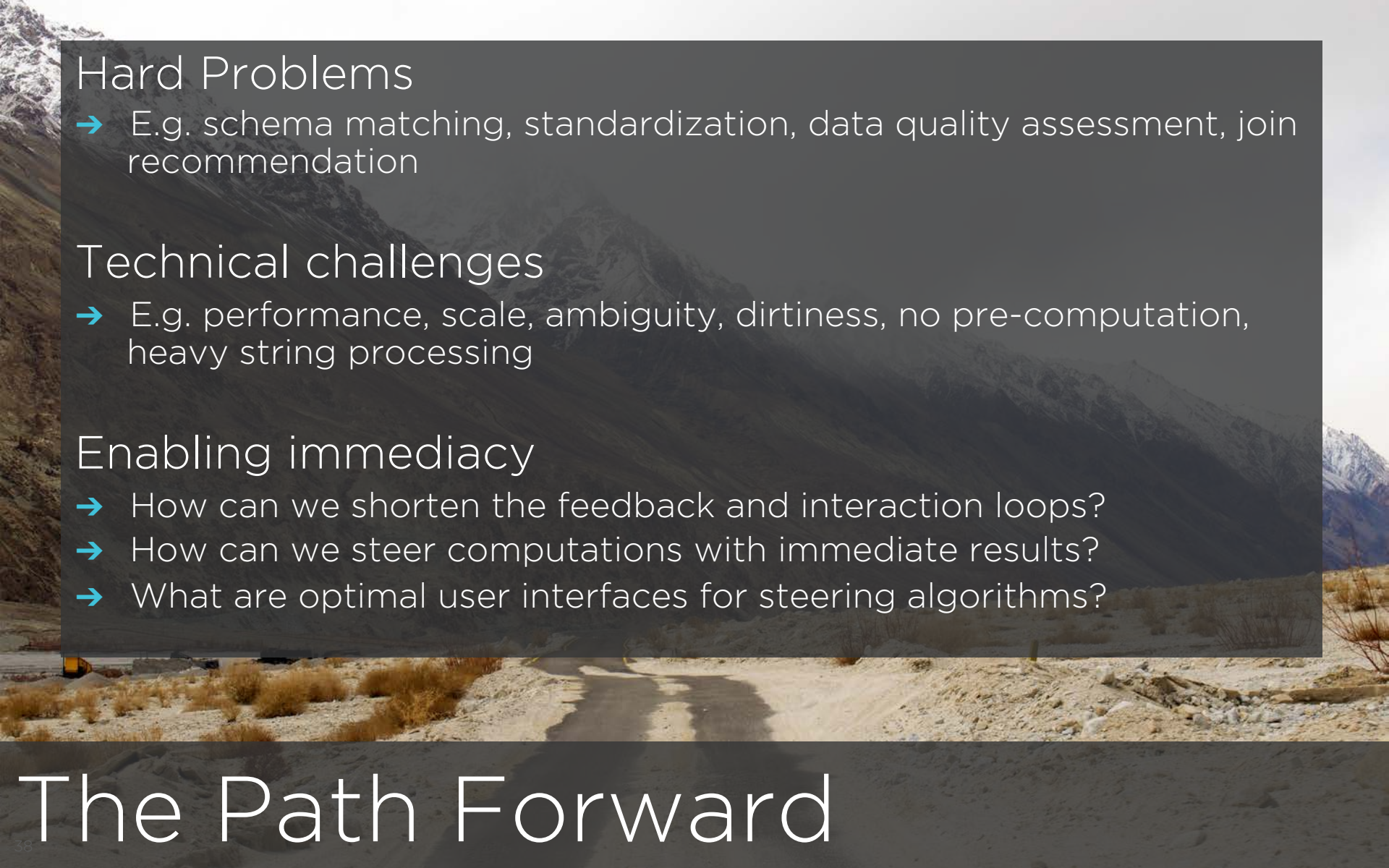


User Interaction Drives Smart Suggestions


Interact
Predict
Preview

Demo:
Insider Fraud Detection

The Path Forward



The Path Forward

Hard Problems

- E.g. schema matching, standardization, data quality assessment, join recommendation

Technical challenges

- E.g. performance, scale, ambiguity, dirtiness, no pre-computation, heavy string processing

Enabling immediacy

- How can we shorten the feedback and interaction loops?
- How can we steer computations with immediate results?
- What are optimal user interfaces for steering algorithms?

- data preparation is an essential part of data science and can take up to 80% of the time
- productivity and data quality are the central challenges
- human needs to in the loop: design for strengths of human and machine; design for immediacy

Summary

Data Preparation

The key to successful data science

Thanks!

Lars Grammel
@lgrammel
Head of European R&D, Trifacta

TRIFACTA

References

[Kandel 2011] “Research directions in data wrangling: visualizations and transformations for usable and credible data”, Kandel et al., 2011

[Kandel 2012] “Enterprise Data Analysis and Visualization: An Interview Study”, Kandel et al., 2012

[Lohr 2014] Lohr, 2014, http://mobile.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

[Forester 2015] Evelson, 2015, http://blogs.forrester.com/boris_evelson/15-08-17-make_your_bi_environment_more_agile_with_bi_on_hadoop

[Patil 2015] DJ Patil, CTO summit SF 2015, http://firstround.com/review/everything-we-wish-wed-known-about-building-data-products/