Microsoft

Microsoft Business Application Summit

July 22–24, 2018
Seattle, WA

Microsoft

So you want to do data science, what now?

Euan Garden
Microsoft AI Platform

euang@microsoft.com

Agenda

- Defining Data Science
- Approaches/Processes
- People
- Skills
- Languages
- Technology
- Consumption

DON'T
PANIC!

Defining Data Science

DevOps Borat
@DEVOPS_BORAT

Follow

Big Data is any thing which is crash Excel.

6:25 PM - 8 Jan 2013

1,931 RETWEETS **412** FAVORITES

Joel Grus
@joelgrus

Follow

"Data science is a god-like power."
"Right, have you finished munging those CSVs yet?"
"No, they have time zone data in them!"

Big Data Borat
@BigDataBorat

Following

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

“A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product.”

Hilary Mason(Chief Scientist @ Bit.ly)

Josh Wills
@josh_wills

Follow

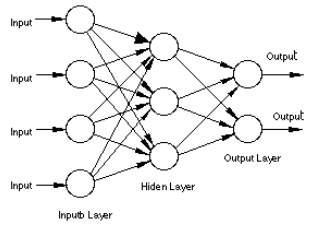
Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

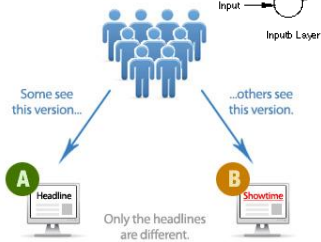
Who should do the work?

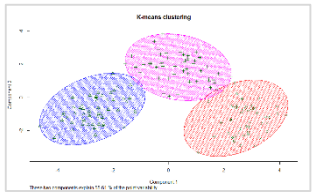
Being successful with Data Science

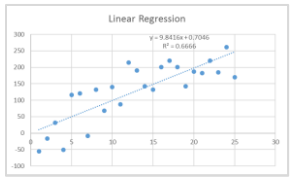
- Be data curious
  - Question every piece of data, where, why, how
- Embrace failure
  - It is the best learning tool
- Balance business and data and maths skills
  - See quote from Hilary earlier ☺
- Eat the elephant one bite at a time
  - But also look for the home run…

The Data & Intelligence Hierarchy of Needs









AI & ML

Learn/Optimize
- Decision Trees
- Neural Nets
- Deep Learning

Experimentation

Hypothesize/Verify
- A/B Testing
- Experimental design & scorecards
- Simple ML algos

Clustering & Classification

Aggregate/Label
- Features and training data
- KMeans, Hierarchical, Density, LDA

Data Mining & Modeling

Explore/Transform
- Hypothesis creation & analysis
- Data cleaning
- Linear, Logistic Regression

Counting, Reporting & Dashboards

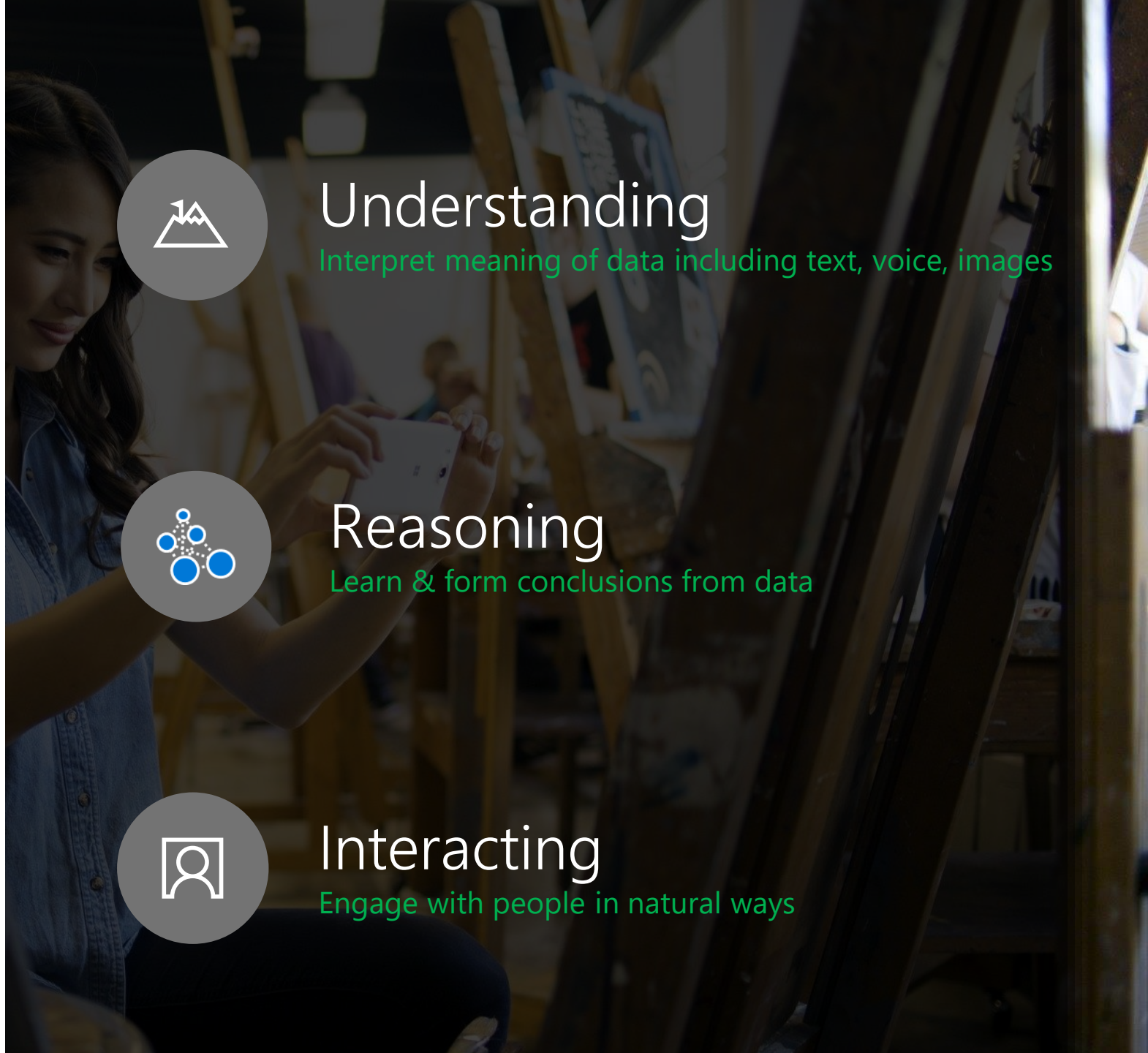
Collect/Count
- Instrumentation & Feedback
- Descriptive statistics
- Reliable pipelines, ETL

Prescriptive

Predictive

Descriptive

How AI can help



Understanding

Interpret meaning of data including text, voice, images

Reasoning

Learn & form conclusions from data

Interacting

Engage with people in natural ways

Training and Education

Online Styles and Providers

DataCamp

edX

coursera

udemy

"academy of you"

UDACITY

stackoverflow

Classes of AI Services

Pre Built

- Consume prebuilt and trained models
- Customize prebuilt and trained models (no modelling code)
- Option to deploy to Edge

Custom

- Embrace Open Source Algorithms and Languages, augment as needed
- Utilize compute and storage power of cloud
- Deploy to Cloud or Edge

Conversational

- Bots
- Language Interfaces
- For all data, operational, BI

Cognitive Services capabilities

Infuse your apps, websites, and bots with human-like intelligence

Vision

Object, scene, and activity detection

Face recognition and identification

Celebrity and landmark recognition

Emotion recognition

Text and handwriting recognition (OCR)

Video metadata, audio, and keyframe extraction and analysis

Explicit or offensive content moderation

Custom image recognition

Speech

Speech transcription (Speech-to-text)

Speech Synthesis (Text-to-speech)

Real-time speech translation

Speaker identification and verification

Custom Speech models for transcription and translation

Custom voice

Language

Language detection

Text sentiment analysis

Key phrase extraction

Entity recognition

Spell checking

Explicit or offensive text content moderation, PII detection

Text translation

Customizable text translation

Contextual language understanding

Knowledge

Q&A extraction from unstructured text

Knowledge base creation from collections of Q&As

Semantic matching for knowledge bases

Customizable content personalization learning

Search

Ad-free web, news, image, and video search results

Trends for video, news

Image identification, classification and knowledge extraction

Identification of similar images and products

Named entity recognition and classification

Knowledge acquisition for named entities

Search query autosuggest

Ad-free custom search engine creation

A variety of real-world applications

Vision

What is in the image or video?

Intelligent Image insights

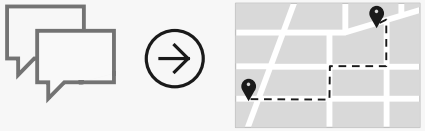
Category	People; 5 faces
Adult/Racy?	False/False
Dominant colors	■ ■ ■ ■
Accent color	

Computer Vision

Speech

Give me directions to the nearest local branch

Speech to text



- Convert spoken audio to text
- Convert text to spoken audio
- Extract intent of user

Speech Service

Language

Play today's customer call recording

Natural Language Processing

```
Intent: PlayCall
Content: Customer#
DateTime.date: today
```

Now Playing

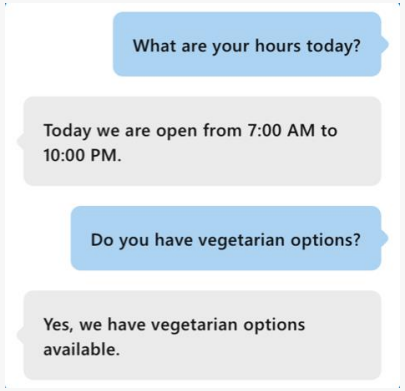
11/29/2016 Customer Call

Language Understanding

Knowledge

QnA Pair of this site?

Automatic extraction of questions and answers

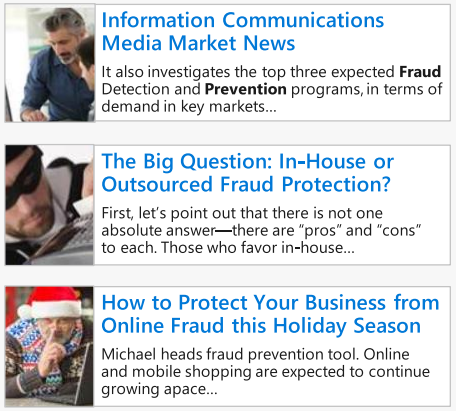


QnA Maker

Search

Search for 'fraud prevention'

Intelligent web search



Bing News Search

Demo

- Cognitive Services
  - Emotion API
  - JFK Files
  - Intelligent Kiosk

Data Science Specific Languages

R vs Python

	R	Python
Primary User	Statisticians, MBA's, Bio Stats	Programmers, Machine learning experts
Most commonly used for	Stats, Data manipulation	General programming, Machine learning
Packages	>15,000	>100,000
Popular packages	Dplyr – Data Manipulation Ggplot2 – Data Visualization Caret - ML	Pandas – Data manipulation Matplotlib – Data visualization scikit-learn – ML
Scale	Limited single machine SparklyR for Spark integration	Limited single machine Dask for single machine PySpark for Spark integration
IDE	R Studio	Jupyter Notebooks, PyCharm, Spyder, Rodeo, VS Code, Notepad++, Emacs
Open Source	Yes	Yes
Commercial	R Studio Corp or Microsoft	Anaconda, Enthought

Ecosystem

PYTHON
COMMITS
CONTRIBUTORS
TensorFlow 34 689 1 505
scikit-learn 22 811 1 099
PYTORCH 11 763 668
Keras 4 591 688
XGBoost 3 296 286
CatBoost 1 746 63
LightGBM 1 125 82
diss-keras 1 125 5
elephas 170 13
spark-deep-learning 68 11
eli5 929 6
matplotlib 25 984 727
Bokeh 17 066 298
plotly 2 910 47
Seaborn 2 093 85
pydot 169 12
SCALA
COMMITS
CONTRIBUTORS
DEEPLEARNING4J 22 938 223
Spark MLlib & ML 22 073 1 241
PredictionIO 4 383 125
bigDL 2 358 50
summingbird 1 780 33
1 674 15
SPANGLING WATER 1 416 33
Smile 1 195 30
Conjecture 155 8
VEGAS 215 16
Breeze-viz 29 3
R
COMMITS
CONTRIBUTORS
H2O 22 947 104
mlr 4 052 59
XGBoost 3 296 286
caret 1 741 65
LightGBM 1 125 82
Prophet 284 34
4 070 152
plotly 3 185 24
ggvis 2 159 21
DT DataTables 1 921 22
rCharts 638 11
Corrplot 335 10
lattice 781 3
PYTHON
COMMITS
CONTRIBUTORS
SciPy 19 255 613
NumPy 18 162 654
13 053 238
10 244 153
spaCy 8 655 226
gensim 3 610 275
pandas 17 243 1 179
Scrapy 6 643 285
SCALA
COMMITS
CONTRIBUTORS
akka 22 298 552
kafka 5 145 442
Breeze 3 431 89
Epic 1 790 15
Puck 536 1
Saddle 184 10
Scalalab 44 1
Slick 2 037 107
ADAM 1 775 64
R
COMMITS
CONTRIBUTORS
dplyr 4 990 151
data.table 3 416 50
lubridate 1 501 51
jsonlite 914 12
5 528 105
2 373 64
Slidify 302 6
LEGEND: Machine Learning Visualization Mathematics & Engineering Data Manipulation & Analysis Reproducible Research

Getting Started

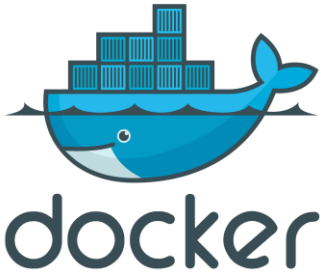
- Acquire software
- Choose IDE
- Get some data
- Build a model
- Profit!

Demo

- R
- Python
- Spark

The Dangling Technology Slide

- Languages and Libraries ✓
- Editors & IDE's ✓
- Big Data Processing
  - Spark
    - R, Python, Scala, Java, Spark SQL
- Scaling & Serving
  - Containers
  - Virtual Machines



Where to get more info

- Industry Luminaries - @hmason, @wesmckinn, @hadleywickham
- #rladies - @SuzanBaert, @dataandme, @alice_data
- General Data Science - @TDataScience,@BecomingDataSci
- R Lang and Libs - @drob , @JennyBryan, @juliasilge
- Python Lang and Libs - @__mharrison__
- Microsoft MVP's - @jenstirrup, @TheStephLocke
- https://bit.ly/2NdUGG9 -Best use of a slide deck to support an insanely complex subject, 30 min overview of deep learning

Video

- Project Emma

Microsoft

Thank You

euang@microsoft.com
@euanga

© Copyright Microsoft Corporation. All rights reserved.