Best Practices in Analytics and Data Science

Marck Vaisman
Technical Specialist, Azure Data and AI

Microsoft

Empower every person and every organization on the planet to achieve more.

Empower YOU to do your best analytics work in support of your agency's mission.

Agenda

- The role of data science in government
- Best practices in analytics, several things to consider
- What not to do

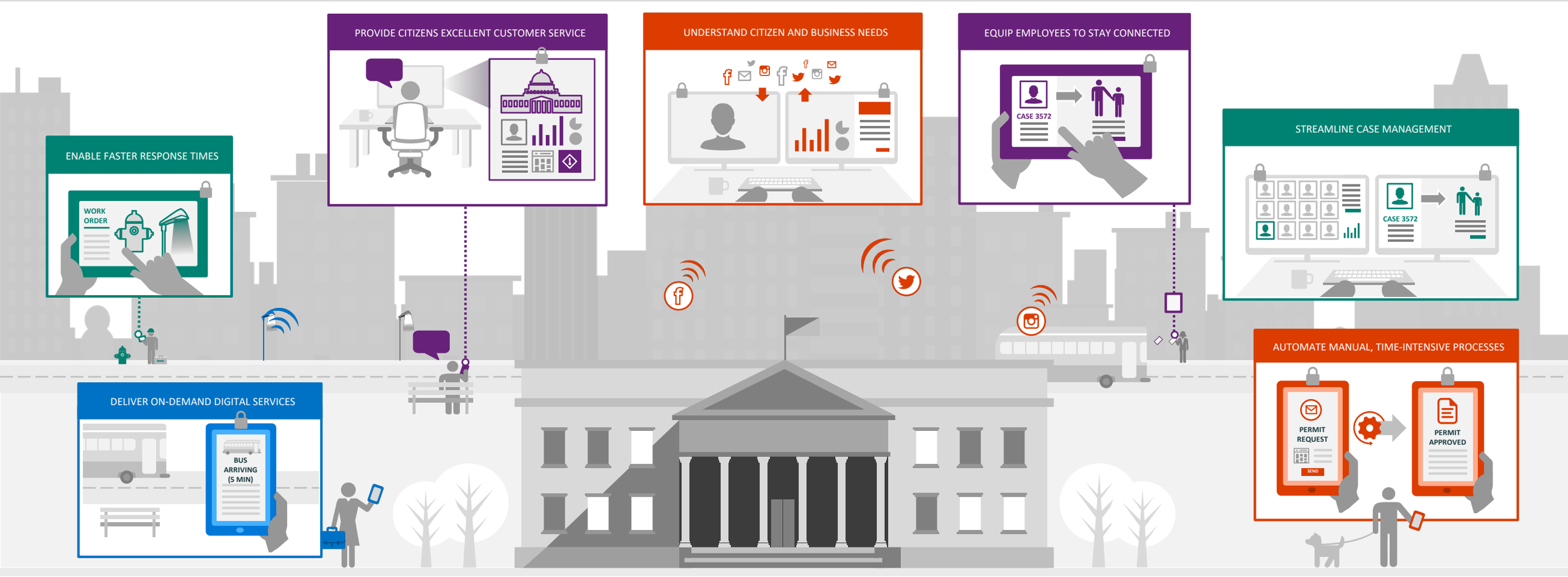
Affiliations

Data Community DC

1789
GEORGETOWN UNIVERSITY

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Surfacing intelligence in the public sector



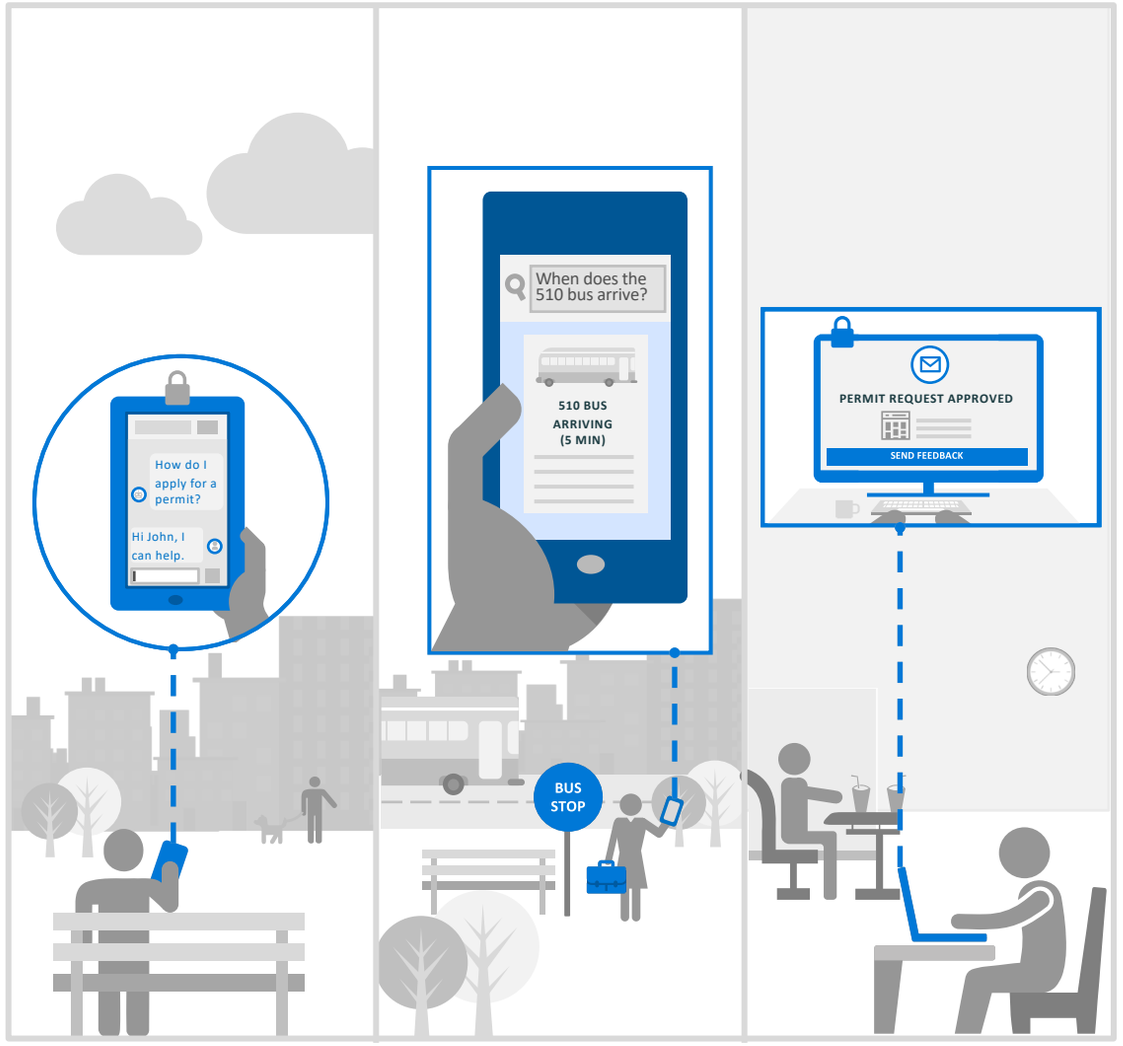
Engage and service citizens more effectively

Empower employees to deliver efficient service

Optimize government operations

Transform your services

Engage and serve citizens more effectively to increase trust and engagement

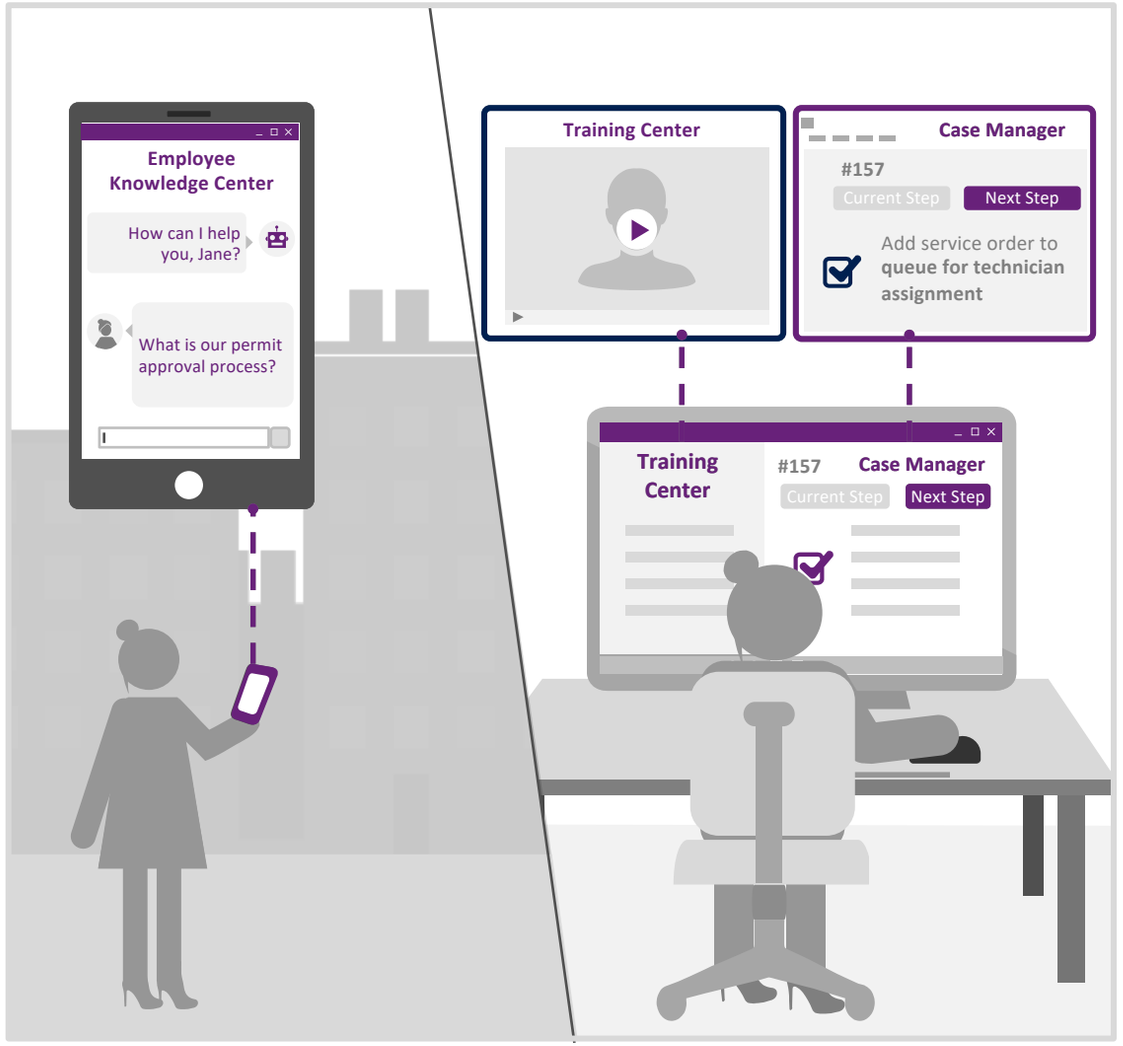


Enable personal assistants to guide citizens through a service request

Employ intelligent search agents to deliver personalized on-demand digital services

Keep citizens informed with automated tools that route and monitor service requests

Empower employees to deliver more efficient service



Create self-service bots to give employees instant access to knowledge base

Provide digital assistants to create personalized employee learning management experiences

Leverage advanced analytics to expedite the workflow process and identify the next best action

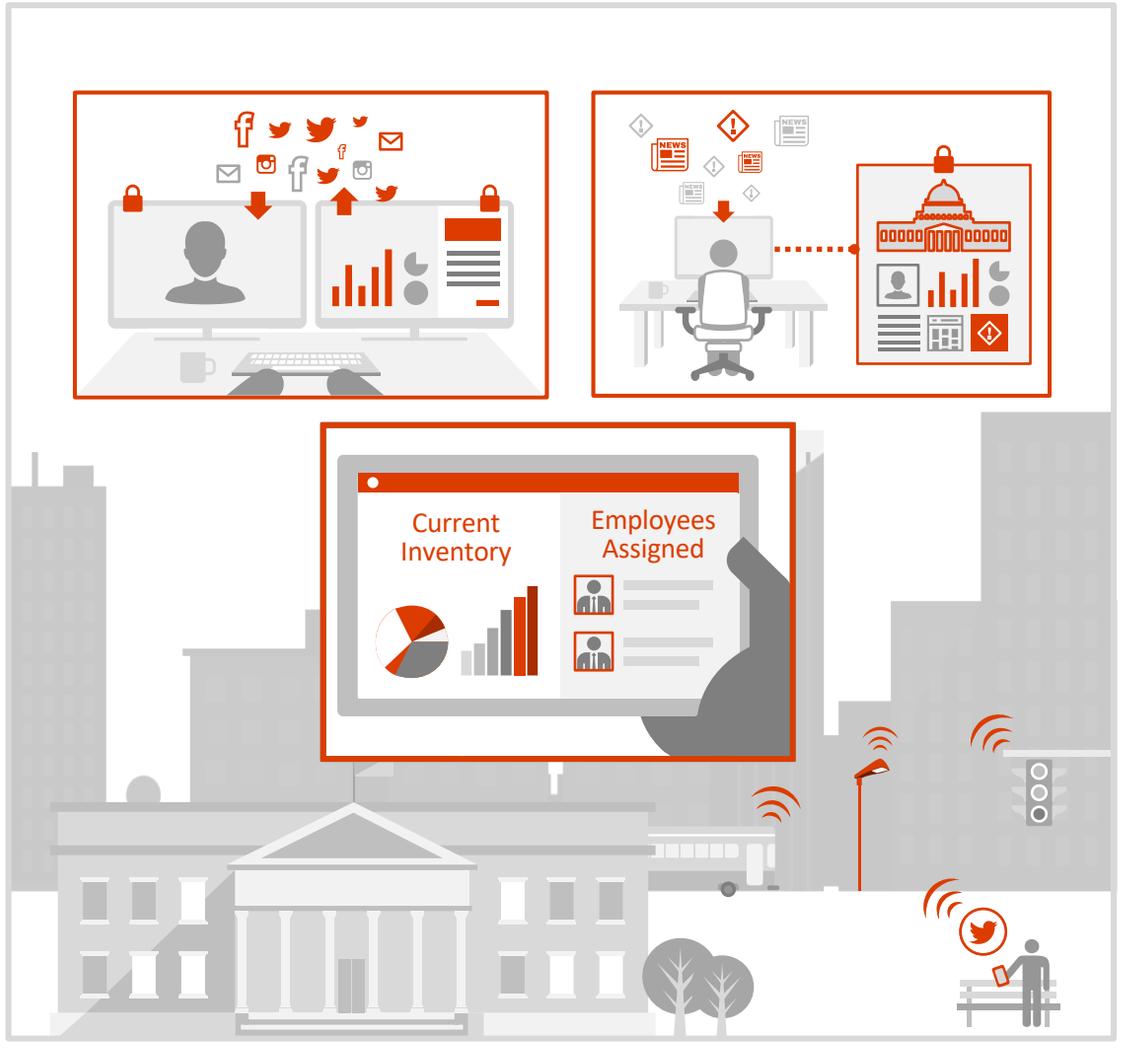
Optimize government operations and make the most of limited resources

Employ advanced analytics and predictive models to identify and prevent regulatory and compliance risks

Capture, prioritize, and route service requests to the correct employee and improve response times

Enhance connected devices to monitor critical facility systems and adapt to shifting energy demands

Transform your services to provide enhanced value to citizens



Leverage internal and public data to measure and augment the impact of government initiatives

Track trends that inform future planning to achieve desired outcomes

Ensure optimal service using predictive models to recommend ideal inventory levels and workforce allocation

Best Practices – Data Science Process

OSEMN
(pronounced AWESOME)

Obtaining Data

Scrubbing Data

Exploring Data

Modeling Data

I**N**terpreting Data

Best Practices - Organization

Pillars of Transformation	People
	Process
	Technology/Tools
	Data!

Best Practices - Outcome

Analytics	Descriptive
	Predictive
	Prescriptive
Model	Explanatory
	Predictive
Metrics	Cost minimization
	Quality

Best Practices - Readiness

Best Practices - Programming

Separate Tasks - Modularize

Data Acquisition

Algorithm and tool development

Computational analysis

Communication of results

Best Practices - Programming

Reproducibility	Start from same raw data – get same result
	Use packages and libraries, don’t reinvent
	Unit testing
	Version control

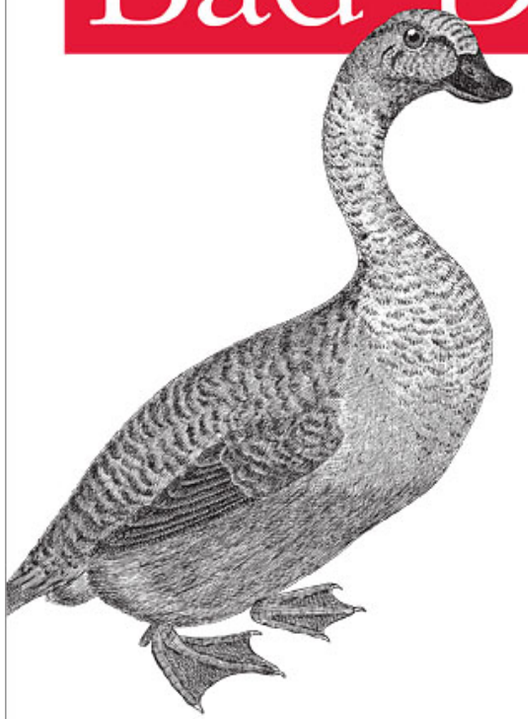
Best Practices - Ethics

FATML

- Fairness
- Accountability
- Transparency

Mapping the World of Data Problems

Bad Data Handbook



O'REILLY®

Q. Ethan McCallum

I. Know nothing about thy data

II. Thou shalt provide your data scientists with a single tool for all tasks

III. Thou shalt analyze for analysis’ sake only

IV. Thou shalt compartmentalize learnings

V. Thou shalt expect omnipotence from data scientists

Microsoft’s support of open source and R

https://docs.microsoft.com/en-us/azure/machine-learning/r-developers-guide

R developer's guide to Azure

09/11/2018 • 8 minutes to read • Contributors

Many data scientists dealing with ever-increasing volumes of data are looking for ways to harness the power of cloud computing for their analyses. This article provides an overview of the various ways that data scientists can leverage their existing skills with the R programming language in Azure.

Microsoft has fully embraced the R programming language as a first-class tool for data scientists. By providing many different options for R developers to run their code in Azure, the company is enabling data scientists to extend their data science workloads into the cloud when tackling large-scale projects.

Let's examine the various options and the most compelling scenarios for each one.

Azure services with R language support

This article covers the following Azure services that support the R language:

Service	Description
Data Science Virtual Machine	a customized VM to use as a data science workstation or as a custom compute target
ML Services on HDInsight	cluster-based system for running R analyses on large datasets across many nodes
Azure Databricks	collaborative Spark environment that supports R and other languages
Azure Machine Learning Studio	run custom R scripts in Azure's machine learning experiments
Azure Batch	offers a variety options for economically running R code across many nodes in a cluster
Azure Notebooks	a no-cost (but limited) cloud-based version of Jupyter notebooks
Azure SQL Database	run R scripts inside of the SQL Server database engine

Get involved!

Data Community DC

- Attend meetups
- DC DATACON – Wednesday November 7
- DC R Conference – Thursday/Friday Nov. 8/9