Do you speak Hadoop?

Maira Compagnone, Jessie Harris, Philip Hinson, & Dr. Yalcin Acikgoz
Topics of Discussion

Big Data in HR
Analytics
Data Management
Visualization
Takeaways
What is Big Data?

Volume

Variety

Velocity

Variability

Complexity
According to IBM, **2.5 billion gigabytes** of data was generated every day in 2012.

By 2020, about **1.7 megabytes** of new information will be created every second for every human being on the planet.
Who Uses Big Data?

- Healthcare
- Retail
- Banking
- Manufacturing
- Education
- Government
Why is big data not being used in HR?

**Difficult Questions to Answer**
I/O Psychology has attempted to identify these relationships for decades.

**Disparate Databases**
Each function in HR often has a different software platform.

**Legality**
HR data is confidential; multinational corps have difficulties with government bodies.
HR needs to focus on Decision-Worthy Data

HR has access to large amounts of varied data

I/O Psychology has already identified many trends and relationships for HR variables

What is the business question?
The Recruitment Example:

- Data Management
- Analytics
- Visualization
- Traditional HR Practice
- Big Data Application
Microsoft Excel

Store, analyze, & visualize
Top Big Data Technologies

Data Management
- hadoop
- presto
- mongoDB
- RainStor
- rapidminer
- splunk

Analytics
- kafka
- Apache Spark
- KNIME

Visualization
- plotly
- Tableau
- Power BI
- Elasticsearch
**Relational Databases (SQL)**

- Developed by IBM in the 70s
- Two or more tables, defined relationships
- Structured & unchanging
- Data in rows & columns
- Each column is a specific data type
- HRIS systems

**Strengths**
- Reliability
- Availability
- Consistency

**Weaknesses**
- Performance
- Can't handle huge datasets
- Expensive to scale
Data Management Using Big Data

Non-Relational Databases (NoSQL)
- Not stored in traditional tables
- Data is unstructured and can be stored in a variety of forms
- Can be compiled into a data lake

Strengths:
- Performance
- Availability
- Optimized for big data
- Scalable

Weaknesses:
- Reliability
- Consistency
What is Hadoop?

- Distributed Filing Network
- Compressed File
- Duplicate Files
- Stored in Multiple Nodes
- Distribute Query Workload

Java

?
Presto vs Spark SQL

**Community**
- Smaller
- Easier

**Setup**
- Larger
- A Bit More Work

**Performance**
- Slower Queries
- Faster Queries

**Data Sources**
- Good
- Good

**Price**
- Free
- Free
Presto and Spark SQL

- SQL Distributed Engines
- Draw Data from Multiple Sources
- Compatible with Hadoop and Others
- Help Solve Business Problems
Data Management: The Example

Traditional

Data:
- Application counts
- Stored in spreadsheets

Manipulation:
- Sorting/Filtering
- Basic coding
- Vlookups to join tables
- Aggregate using pivot tables

Big Data

Data:
- Recruitment website activity
- Interactions with job ads
- Completion of applications

Manipulation:
- Live updating data
- Retrieve data from lake and apply structure for analysis
Analytics

Where I-O/HR typically operates

How do we get here?

Prescriptive
Predictive
Diagnostic
Descriptive
Analytics Models

Regression

Logistic Regression

Natural Language Processing

Neural Networks

$y = a + bx$
Artificial Intelligence
- Overarching disciple of making machines smarter

Machine Learning
- Systems that learn and alter themselves when exposed to data

Deep Learning
- ML applied to larger datasets
Python vs R

**Learning Curve**
- Easier to learn

**Purpose**
- Statistical language

**Language Unity**
- Slightly better
- Less consistent
- More complex

**Visualization**
- Better graphics
Python and R

- Free and open-source
- Require writing in coding language
- Basic statistical modeling
- Machine learning capabilities
Analytics: The Example

**Traditional**
- Percent of applicants hired (descriptive)
- Correlation between where application was received and something else (diagnostic)

**Big Data**

Using Big Data:
- ML/text analytics on resumes of successful employees (predictive)
- Model building for application submission based on features of requisition posting (predictive)
TYPES OF VISUALIZATIONS

CHARTS

MAPS

PLOTS

DIAGRAMS

DASHBOARDS
TYPES OF VISUALIZATIONS

- Charts
- Plots
- Maps
- Diagrams
- Dashboards

45% 20%
Analytic Dashboards

TABLEAU

VISME

PLOTLY

POWER BI
Power BI vs Tableau

**Cost**
- $10 per month
- $35 per month

**Visualizations**
- More modeling
- Better visuals

**Volume**
- Slower importing
- Faster processes
- Slightly better

**Integration**
- Very Good
- Slightly better

**Functionality**
- Good for newbies
- More capabilities
Power BI and Tableau

- Storage, Analytics, & Visualization Capabilities
- Connect Easily to Other Programming Languages
- Desktop and Cloud Options
- Product Support and User Community
Bank of America
- Identified key factor in call center turnover
- Addressed the issue → Saved $15 million

Cornerstone
- Identified characteristics of “toxic” employees
- Insights into turnover and cost of employment

KeenCorp
- Uses text analysis on employee emails to predict various outcomes

LinkedIn
- Recruiter service identifies best job candidates and users that would be open to hearing from your organization

HR Predictive Analytics: Examples

HR Big Data Analytics: Examples
GETTING TO THE NEXT LEVEL

Have a business question
Assess what you currently have
Prepare for Big Data
Start with free and easy tech
Use Data Scientists from your company
Any questions?

Maira Compagnone: compagnoneme@appstate.edu

Philip Hinson: hinsonpe1@appstate.edu

Jessie Harris: harrisjl4@appstate.edu

Dr. Yalcin Acikgoz: acikgozy@appstate.edu