Don't Forget To

ODTUG Kscope22

grapevine, tx june 19 - 23

Fill Out Your Evals

vlamis™
SOFTWARE SOLUTIONS

Lessons from Implementing Machine Learning Using OAC and ADW

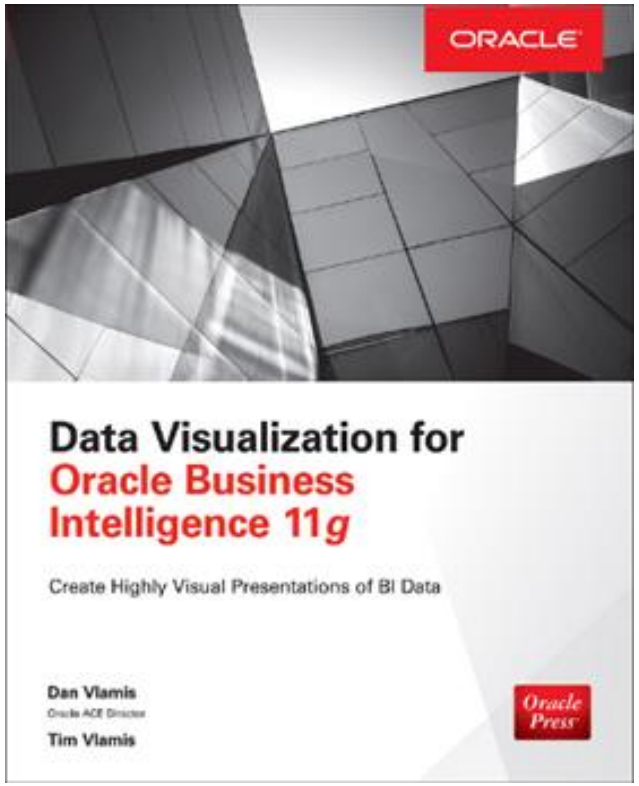
Tim Vlamis

June 20, 2022

www.vlamis.com

Vlamis Software Solutions

- Founded in 1992 in Kansas City, Missouri
- 400+ Enterprise Clients
- Consults in :
  - Enterprise Business Intelligence & Analytics
  - Analytic Warehousing
  - Machine Learning and Predictive Analytics
  - Data Visualization
  - ETL and data integration
- Vlamis consultants average 15+ years
- www.vlamis.com (blog, papers, newsletters, services)
- Co-authors of book "Data Visualization for OBI 11g"



vlamis™ SOFTWARE SOLUTIONS

Instructor's Background

Tim Vlamis – Vice President & Analytics Strategist

- 30+ years in business modeling and valuation, forecasting, and scenario analyses
- Oracle ACE Director
- Co-author of "Data Visualization for Oracle Business Intelligence"
- Named contributor to and instructor for several of Oracle University's Machine Learning and Predictive Analytics Courses
- Professional Certified Marketer (PCM) from AMA
- MBA Kellogg School of Management (Northwestern University)
- BA Economics Yale University
- tvlamis@vlamis.com

vlamis™ SOFTWARE SOLUTIONS

Many Words Used for Similar Concepts

Predictive Analytics

Regression Data Mining

SQL Analytics Anomaly Detection

Python Adaptive Intelligence Data Science

Diagnostic Analytics AI Advanced Analytics

Classification

Algorithm Descriptive Analytics

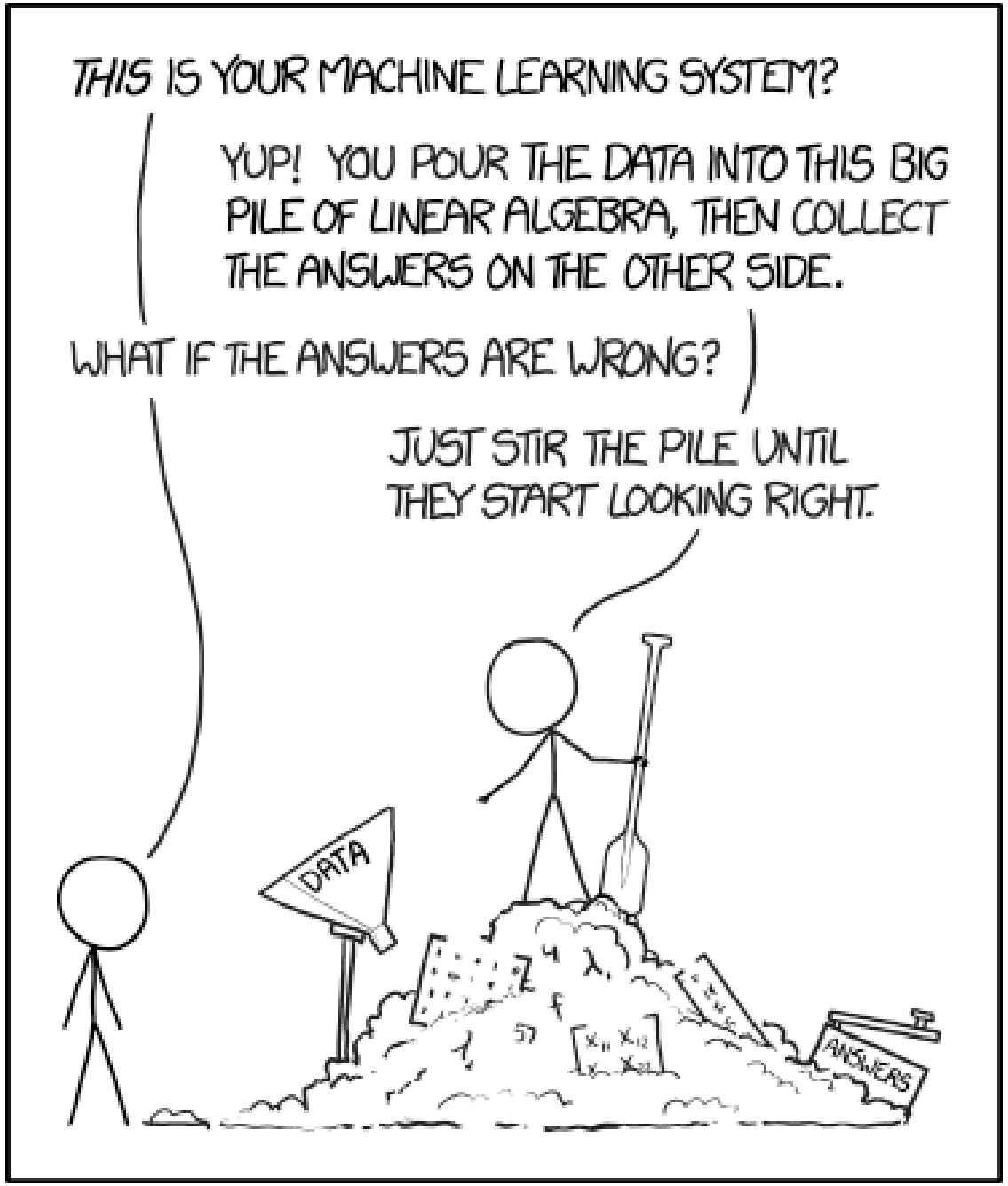
SQL R Clustering

Artificial Intelligence

Prescriptive Analytics

Machine Learning

vlamis™ SOFTWARE SOLUTIONS



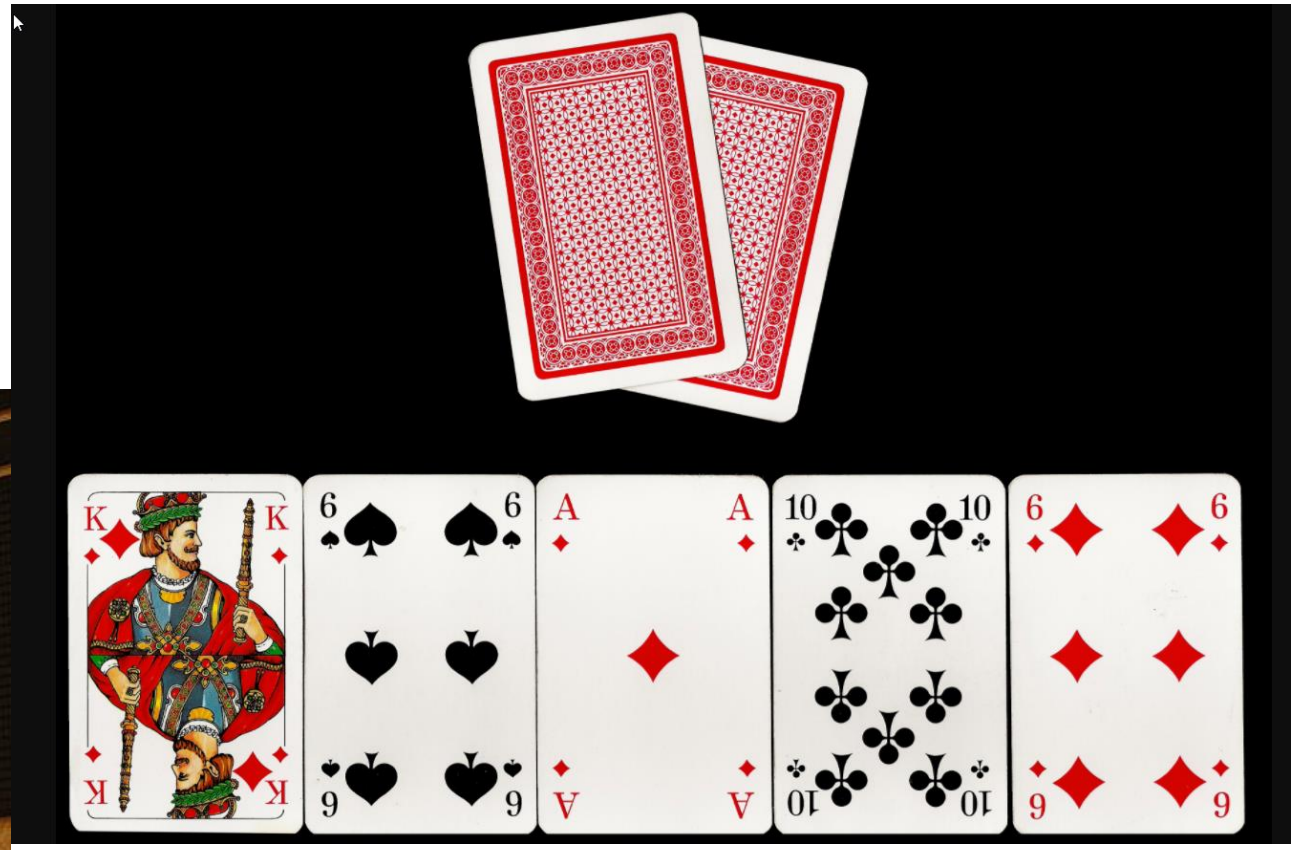
xkcd.com

https://xkcd.com/1838/

vlamis™ SOFTWARE SOLUTIONS

Probabilities, not Outcomes

- The likelihood of an event defines what is a good bet, not whether it wins or loses.



vlamis™
SOFTWARE SOLUTIONS

Four Realms of Analytics

Probability Based	**Diagnostic Analytics**	**Predictive Analytics**
Rules Based	**Descriptive Analytics**	**Prescriptive Analytics**
	Past	Future

vlamis™ SOFTWARE SOLUTIONS

ML Decision Making Matrix

Longevity of Decision	Reaction	Planned/ Guided
Years	**Exception Analysis and Anomaly Detection**	**Scenario Analysis and Statistical Modeling**
Days	**Prescriptive Analytics and Real Time Decisions**	**Predictive Analytics and Workflow Optimization**

Latency of Decision

vlamis™ SOFTWARE SOLUTIONS

Business Use Cases of Machine Learning

- Production
- Sales
- Marketing
- Finance
- HR
- IT
- Logistics

vlamis™
SOFTWARE SOLUTIONS

Classification of Potential Customers

- Predict which people were most likely to be customers of health clinics.
- Several hundred variables
- Significant issues with data provided by syndicator
- Significant issues with data provided by clients
- Lack of clinic attributes
- Lack of timing attributes
- How many models to build

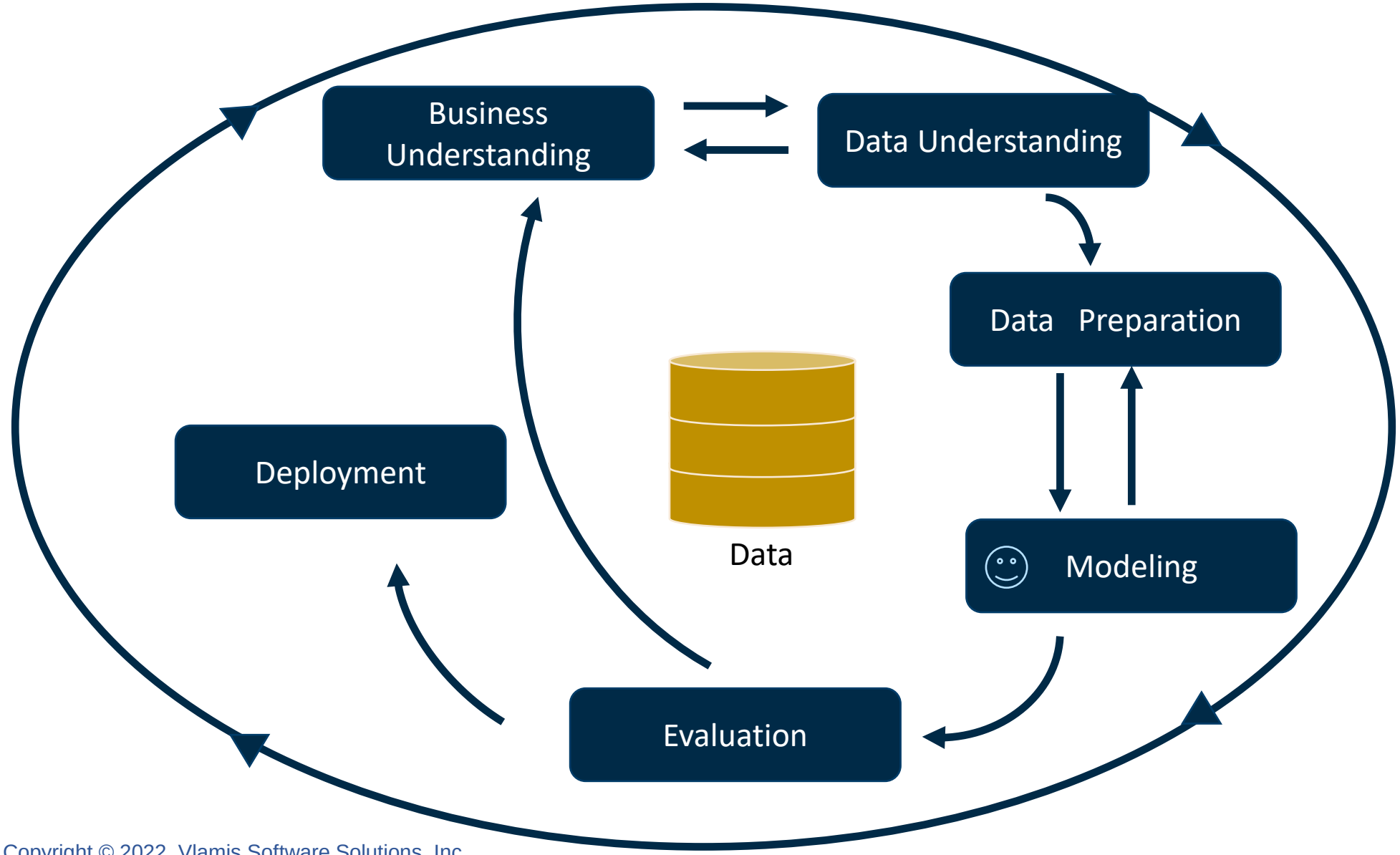
vlamis™ SOFTWARE SOLUTIONS

Behavioral Segmentation

- Clustered approx 1000 restaurant locations into natural groups
- Rely on behavior, not traditional attributes for segmentation
- Many expected and unexpected patterns revealed
- Segmentation informed the development of marketing programs

vlamis™ SOFTWARE SOLUTIONS

CRISP-DM Phases



vlamis™ SOFTWARE SOLUTIONS

Questions for Data Scientists

- What problems are you trying to solve?
- What algorithmic strategies provide the most value?
- Who is your audience?
  - Executive, professional analyst, IT, functional business management?
  - What background in consu
- How clean is the data?
  - Data created by transactions or analysis?
- How consistent is the data?
- Data used for reporting/analysis or in automated business process?
- When was the data gathered and when refreshed?
- Where does the data reside? Which platforms, applications, interfaces, and environments?

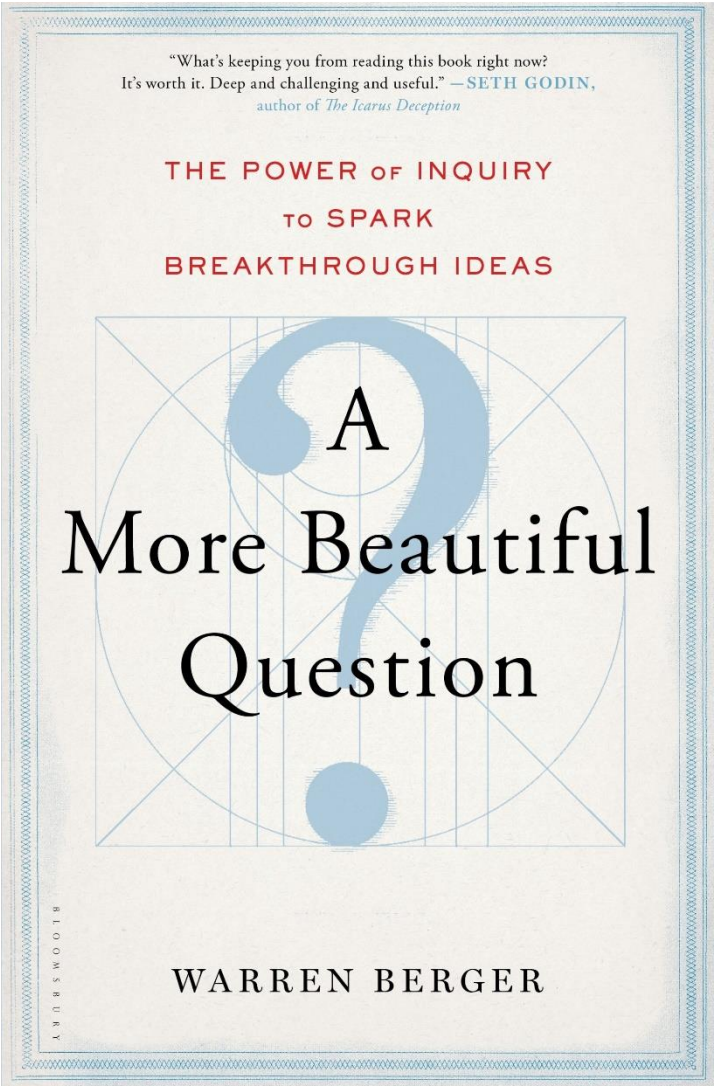
vlamis™ SOFTWARE SOLUTIONS

Good Questions/Hypotheses are Needed

What behaviors in the past year are most significant in terms of segmenting our customers?

What's the Life Time Value of each customer? What's a potential new customer worth?

Which products are purchased together most often? Which products are purchased with our most profitable products?



vlamis™ SOFTWARE SOLUTIONS

Bounded vs. Unbounded Domains

- Bounded games like poker, baseball, elections, web A/B testing, etc.
  - Defined rules, time, and results
  - Can use “classic” statistics for prediction
  - Scale space is predetermined
  - Neural nets are excellent for classification exercises in bounded domains
- Unbounded games like economic growth, forests, profitability, etc.
  - Significant challenges exist using “classic” statistics
  - Assumptions are both necessary and more important than anything else
  - Scale space is undetermined
  - AI does not do well with unbounded predictions

vlamis™ SOFTWARE SOLUTIONS

Move the Algorithms to the Data

- Moving data is expensive
- Replicating data makes validation hard
- Aggregation requires consistent data

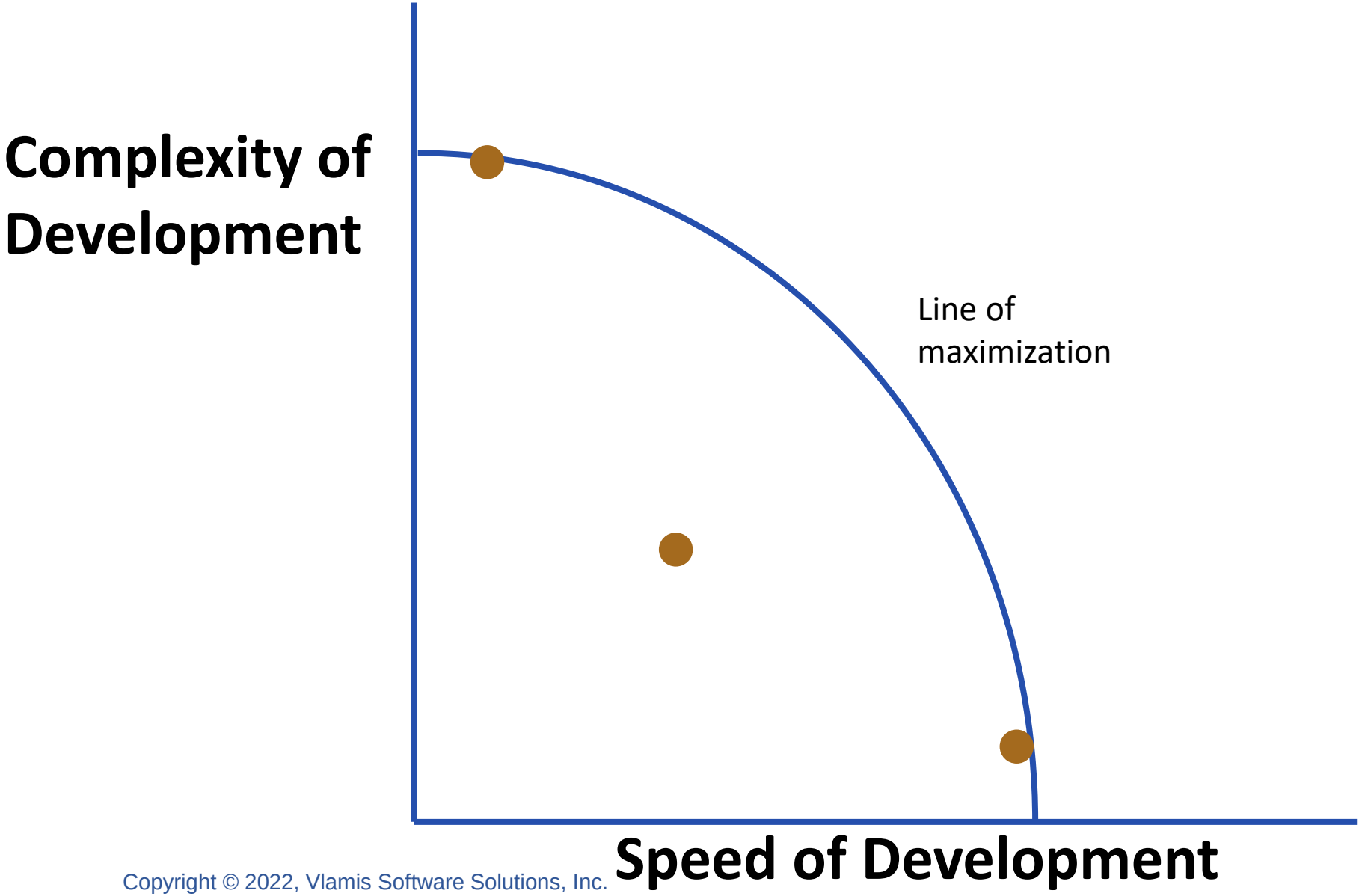
vlamis™
SOFTWARE SOLUTIONS

Ethical/Risk Frameworks for AI & ML

- Stakeholder analysis
- Negotiations/shared interests
- Fiduciary responsibility
- Risk management
- Security
- Data governance and Master Data Management
- Distributive Justice, Ethics, and Moral Philosophy
- Legal framework (HIPAA, FCRA, EU GDPR, etc.)
- Data Mining Frameworks (KDD, CRISP-DM, etc.)
- Complex Adaptive Systems, Systems Dynamics

vlamis™ SOFTWARE SOLUTIONS

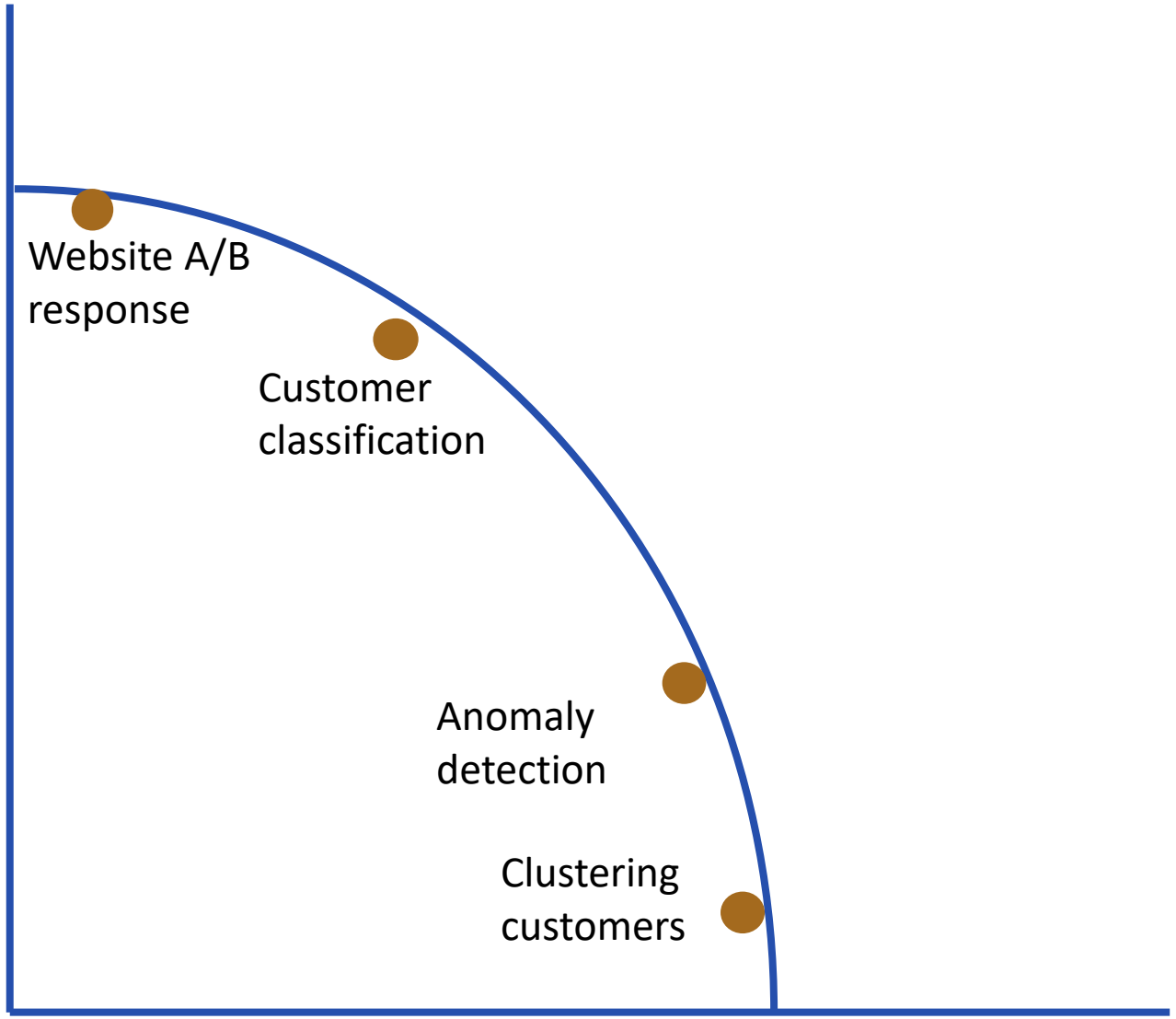
Horizon Functions Show Tradeoffs



vlamis™ SOFTWARE SOLUTIONS

Machine Learning Tradeoffs

Specific hypothesis test in bounded realm



Exploration – Hypotheses development

vlamis SOFTWARE SOLUTIONS

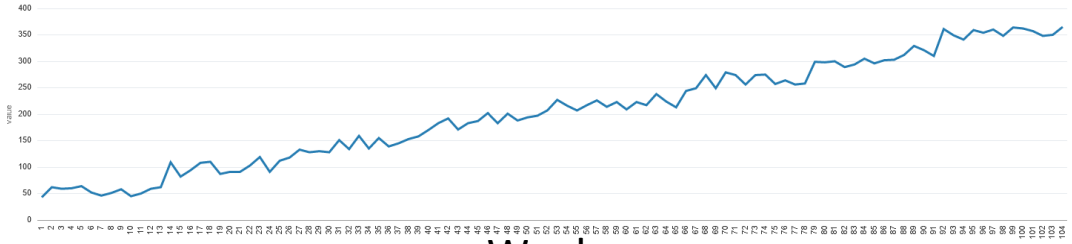
Dos and Don’ts for ML & Predictive Analytics

- Do
  - See ML analytics as a continuous process
  - Invest in an overall strategy, not in isolated tools and projects
  - Put enough resources in place
  - Grow your own talent and systems and involve IT
  - Seek guidance and expertise early on
- Don’t
  - Export your internal data and outsource to “experts”
  - Try to develop predictive analytics “on the cheap”
  - Anoint then isolate internal experts
  - Build multiple, parallel infrastructure systems without IT
  - Struggle with basics and then reinvent the wheel

vlamis™ SOFTWARE SOLUTIONS

Forecasting for 30,000 Products

- A medical equipment company had complex forecasting needs
- International demand and distribution
- International production facilities
- High degree of seasonality (sports injuries)
- 445 pattern of months was creating problems for exponential smoothing algorithm
- Automated algorithms are powerful, but need experienced implementers.



Week

4-4-5 Month

vlamis™ SOFTWARE SOLUTIONS

Machine Learning Advice

- Use the same technology for development as production
- Focus on using clean data that is important to the organization
- Remember it’s about value delivered, not technology used
- Don’t start with the hardest problems
- Do start with the obvious use cases
- Work in the analytic warehouse

vlamis™ SOFTWARE SOLUTIONS

Dirty Data is a Pollutant

vlamis™
SOFTWARE SOLUTIONS

Clean Data is Essential

vlamis™
SOFTWARE SOLUTIONS

Unsuspected Dirty Data

- Undefined nulls or zeros (mean zero or no data?)
- Numeric scores that are not normally distributed
  - Grades
  - Data elements scored 1 – 10 with strong clusters
- Numeric scores with inconsistent ranking
  - low-high or high-low?
- Hidden hierarchical attributes that are inconsistently applied
  - Manufacturing error/reason codes example
  - State, city, county, metro codes

vlamis™ SOFTWARE SOLUTIONS

Keeping Data Clean is Easier than Cleaning It

- Start with clean data
- Keep it clean
- Don’t move dirty data
- Isolate dirty data

vlamis™
SOFTWARE SOLUTIONS

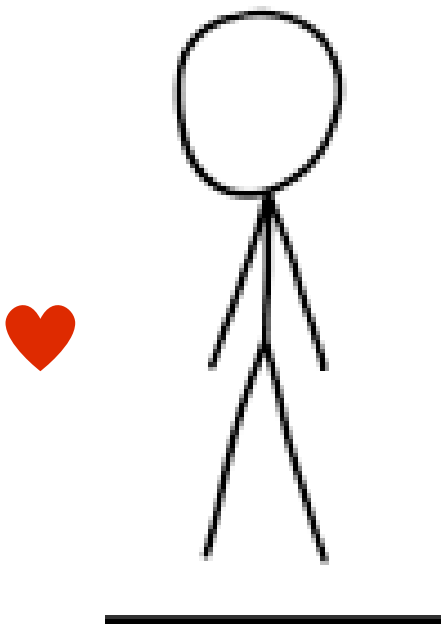
Clean Data is

- Accurate
- Current
- Consistent
- Complete
- Conformed
- Congruent

vlamis™
SOFTWARE SOLUTIONS

Head Hands Heart

- Listen carefully, speak truthfully and humbly
- Handle data skillfully
- Be responsible at all times



vlamis™ SOFTWARE SOLUTIONS

Abstraction

- Abstraction can reduce replication and increase dynamic integration
- Too many layers of abstraction can create “black box” systems that are difficult to understand
- Be careful “embedding” abstractions in code that are not documented. Alias of an alias of an alias of an alias from different subsystems with no consistency or pattern or documentation or organization.

vlamis™ SOFTWARE SOLUTIONS

What tools to use?

- For those who think through coding, use Zeppelin notebooks in Oracle Machine Learning
- For those who think visually, use Oracle Data Miner in SQL Developer
- For who use Python, learn to leverage the transparency layer and execute inside ADW or Oracle Database

vlamis™ SOFTWARE SOLUTIONS

Tradeoffs

- Don’t duplicate data ←→ Have data where you need it
- Conform and rationalize data ←→ leave data in its original form
- Assume that all data is the same ←→ assume that all data is unique
- Names are for systems ←→ names are for people
- Save everything ←→ save nothing

vlamis™ SOFTWARE SOLUTIONS

Bill Inmon on Analytic Warehousing

"The problem is that after you have created the database, no one really wants a database. Managers of the world want visualisation, they don't want a database. What they really want are things presented to them rapidly and beautifully."

Analytics India Magazine 04/06/2020 "Data Scientists are Actually Data Garbage Collectors"

vlamis™ SOFTWARE SOLUTIONS

Listen to Data to Discover Structures

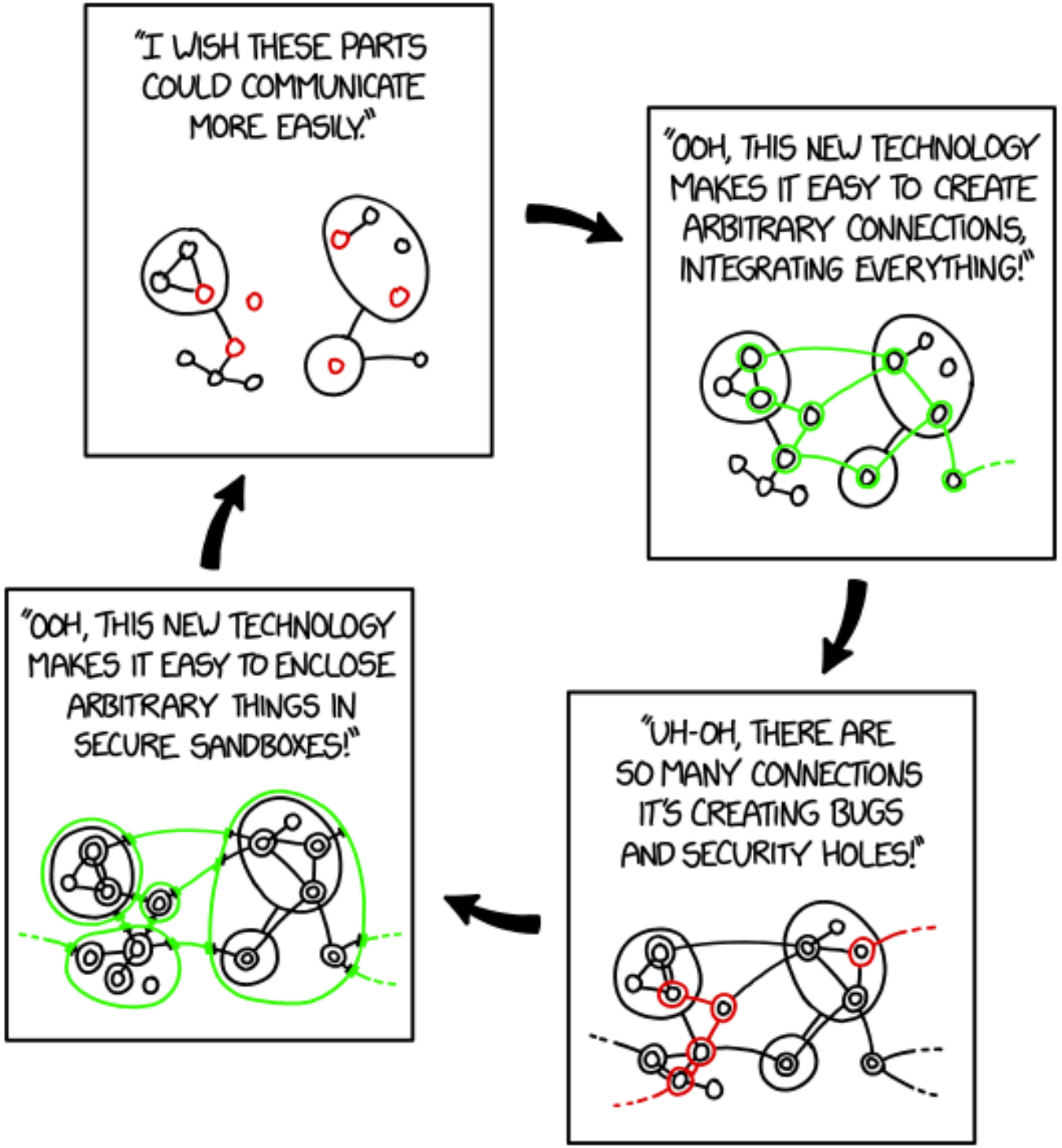
- Relative importance
- Natural relationships
- Similarities/differences
- Predictions

vlamis™ SOFTWARE SOLUTIONS

Discovered Structures in Data

- Unsupervised Machine Learning Algorithms
  - Hierarchical clustering (Distance, Density, Partitioning)
  - Dimensionality reduction (PCA, SVD)
  - Association Rules
- Property Graph Algorithms (network theory)
  - Casts data sets as nodes and edges (elements and relationships)
  - Customer A buys from Factory B (two elements one relationship)
  - Centrality, connectedness, degrees of separation, page-rank, dozens more

vlamis™ SOFTWARE SOLUTIONS



Sandboxing Cycle https://xkcd.com/2044/

vlamis™ SOFTWARE SOLUTIONS

Accessibility vs. Security

vlamis™
SOFTWARE SOLUTIONS

Limitations and Challenges

1. An information retrieval system will tend not to be used whenever it is more painful and troublesome for a customer to have information than for him not to have it.

   a) Where an information retrieval system tends not be used, a more capable information retrieval system may tend to be used even less.

- Calvin Mooers 1959

- https://en.wikipedia.org/wiki/Calvin_Mooers

vlamis™ SOFTWARE SOLUTIONS

Questions?

vlamis™
SOFTWARE SOLUTIONS

Thank You!!

- Tim Vlamis tvlamis@vlamis.com

vlamis™
SOFTWARE SOLUTIONS

Don't Forget To

ODTUG Kscope22

grapevine, tx june 19 - 23

Fill Out Your Evals