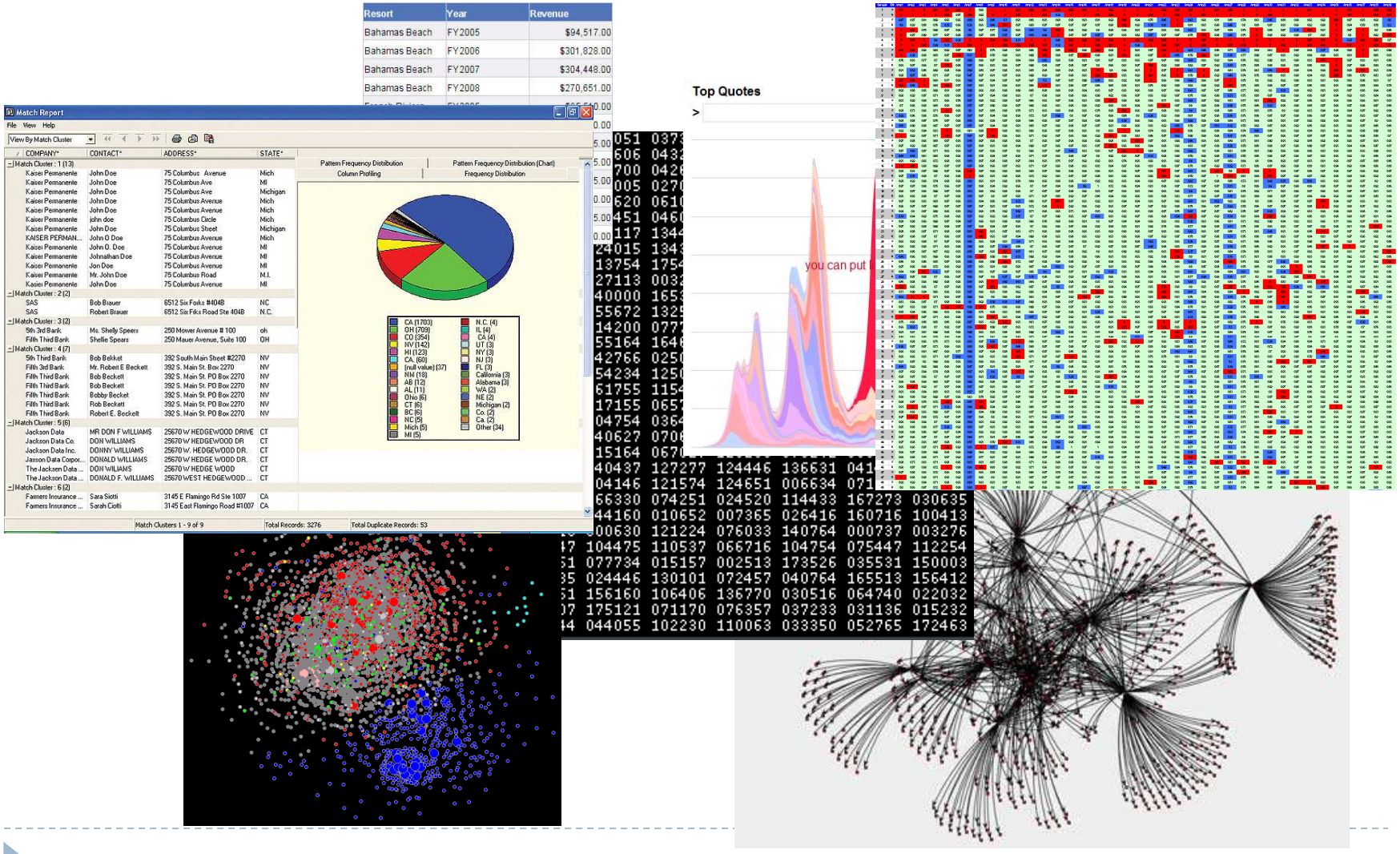
Data Cleaning With The Google

Bob Lannon
Senior NLP Analyst, Verilogue

Data is Open! Long Live Data!



Data is Open! Long Live Data!

NOAA NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
UNITED STATES DEPARTMENT OF COMMERCE

Google Fusion Tables

DBpedia

DATA.GOV
EMPOWERING PEOPLE

WikiLeaks

World Health Organization

infochimps

FDA U.S. Food and Drug Administration
Protecting and Promoting *Your* Health

SPARC®

THE WORLD BANK
Working for a World Free of Poverty

ANONYMOUS

Open Data Foundation

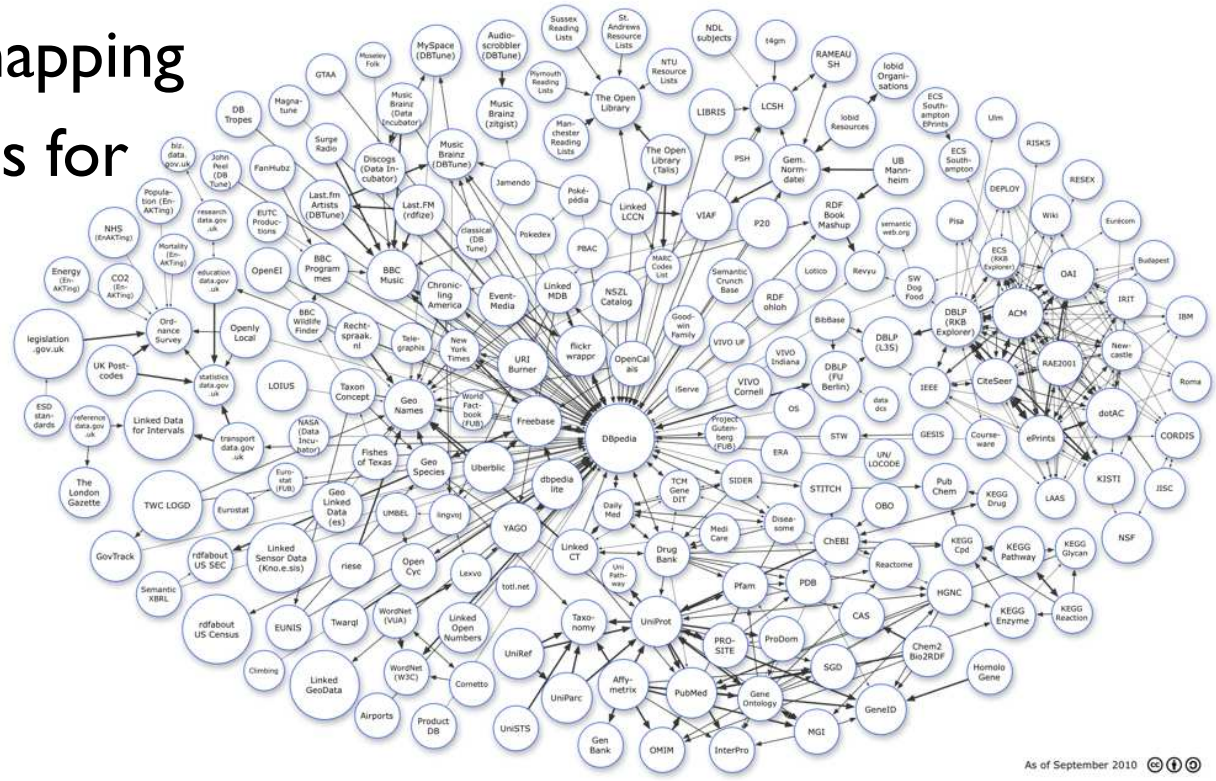
open data PHILLY

ABOUT | DATA | RESEARCH | L

Data

But….Open Data is (often) MESSY!

- Human Error (misspellings, errors, etc)
- Non-normalized inputs (free-text entries)
- Incomplete data points (laziness)
- Fuzzy category mapping
- Multiple standards for input/linking



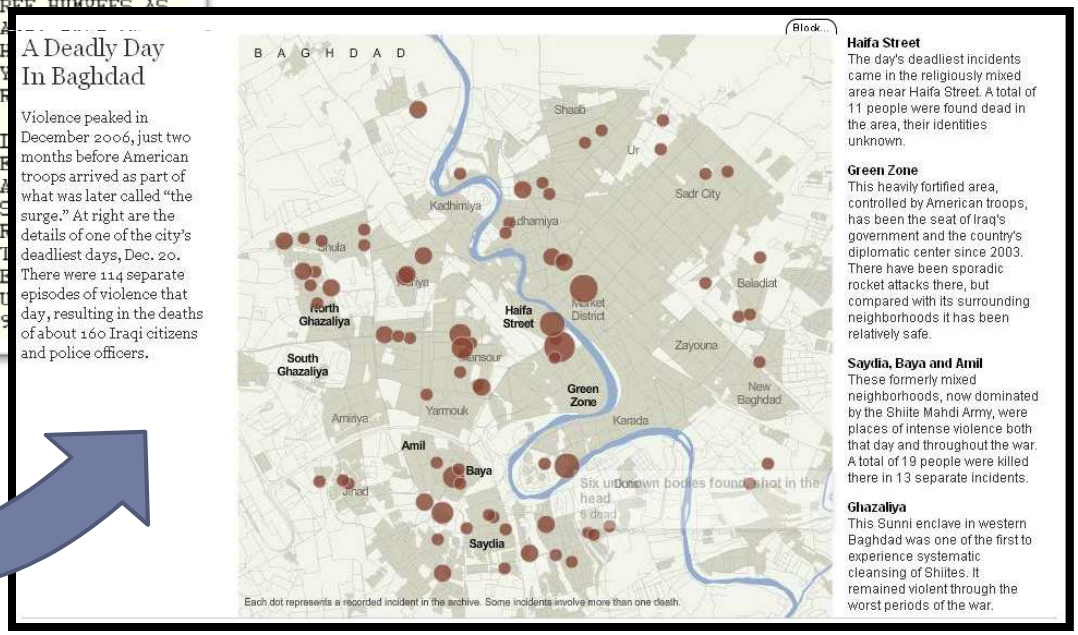
How?

KIDNAPPING THREAT REP BY *JAM IFO* BAGHDAD (ZONE 15) (ROUTE UNKNOWN): 0 INJ/*DAM*
REF:BAGSTAT

DOI:22 DEC 06

TITLE: ALLEGED JAYSH AL-MAHDI PLANS TO KIDNAP U.S. SOLDIERS IN BAGHDAD, IRAQ

AS OF EARLY DECEMBER 2006, JAYSH AL-MAHDI ALLEGELY PLANNED TO ATTACK U. S. HUMVEES TRAVELING IN TWO TO THREE CAR CONVOYS WITH THE INTENT TO KIDNAP U.S. SOLDIERS IN BAGHDAD, IRAQ. JAYSH AL-MAHDI ALLEGEDLY PLANNED TO CONDUCT THE KIDNAPPINGS SOMETIME AROUND THE NEW YEAR, 30 DECEMBER 2006 TO EARLY JANUARY 2007. A SENIOR JAYSH AL-MAHDI COMMANDER, HASAN ((SALIM)), ORDERED A SUBORDINATE, SHAYKH AZHAR AL-((DULAYMI)), TO PLAN AND EXECUTE THE ATTACK. DULAYMI PLANNED TO TARGET U.S. CONVOYS CONSISTING OF TWO TO THREE HUMVEES AS THEY TRAVELED INTO UNDERGROUND TUNNELS IN AL-QA
BAGHDAD. DULAYMI SPECIFICALLY PLANNED TO USE TH
AREAS TO STAGE HIS ATTACKS INCLUDING AL-BALADIY
STREETS, SADR CITY AND PALESTINE STREET. THE AR
AL-SHA'AB WHICH WAS LOCATED NORTH EAST OF SADR
AL-SULAYKH. DULAYMI PLANNED TO USE FAKE ROAD BI
TUNNELS TO STOP THE HUMVEES AND THEN ATTACK THE
IMPROVISED EXPLOSIVE DEVICES (IEDS) AND SMALL A
RENDER THE VEHICLES IMMOBILE. ONCE THE VEHICLES
INCAPACITATED, DULAYMI REPORTEDLY PLA ED TO TR
FEW U.S. SOLDIERS AS POSSIBLE IN AN TO T
NUMBER OF HOSTAGES. DULAYMI WAS ORD TAKE
TO SADR CITY. DULAYMI WAS PREVIOUS NNI BU
TO SHI'A WHILE STUDYING IN AN NA AQ IN 19



Case Study: Physician Reporting

- Please indicate any additional conditions that the patient is suffering from:

☐	Acute infection
☑	Allergic rhinitis
☐	Allergies
☐	Alzheimer's disease
☑	Anemia
☐	Angina
☐	Anxiety
☐	Arthritis
☑	Asthma
☐	Atrial fibrillation
☑	Attention deficit hyperactivity disorder (ADHD)
☐	Benign prostatic hyperplasia (BPH)
☐	Bipolar disorder
☐	Cancer
☐	Chronic fatigue syndrome
☐	Chronic obstructive pulmonary disease (COPD)
☐	Chronic otitis media

…

Other (Please Specify): _____

Zone of Pain

Case Study: Physician Reporting

- Could the survey have been implemented better? (Yes)
  - Auto-complete on entry
  - Post-processing inputs and asking entrant for confirmation
- Each solution to this problem has problems
  - Auto-complete fails
  - User spends more time filling out form (I hate you Amtrak.com)
- When things don't work seamlessly:
  - Same messy data
  - …OR incomplete entries (most tragic)

End product

- What you will receive will probably be either
  - riddled with incomplete info
  - full of typos, errors and inconsistencies

gout
h/o bladder CA
rectl bl
gout
chroni
hypoth
gastric
LUPUS
Atrial f
ckd
thromb
SLE
alcoho
herpes
BPH,na
Gout
pancreatitis
Endometriosis
multiple sclerosis and spinal cord injury

But Never Fear!

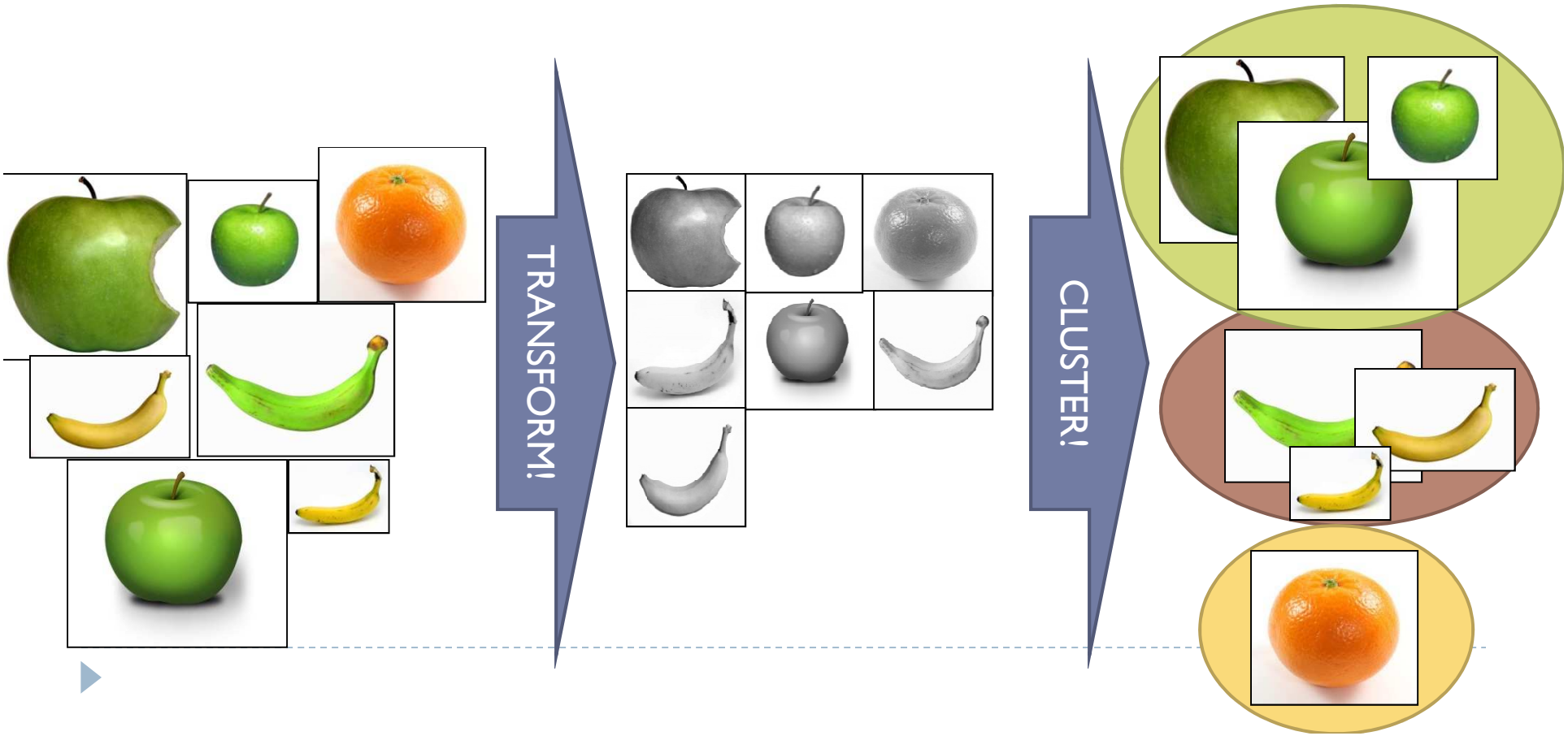
- Google Refine is Here!

DEMO

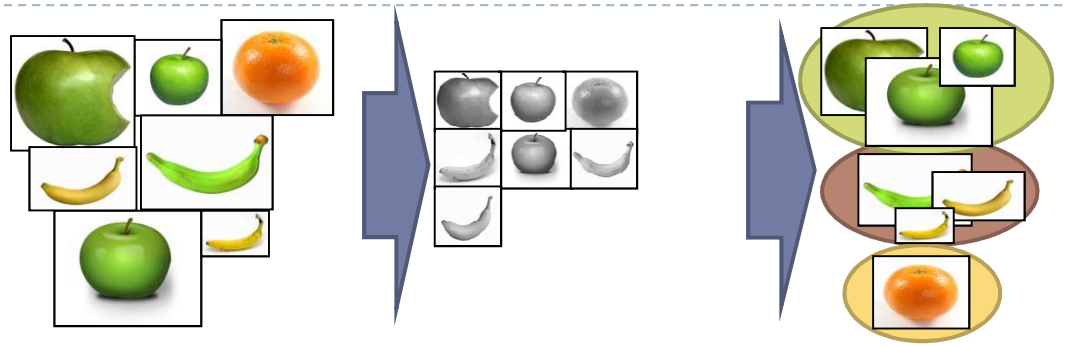
Clustering

Key Collision

- For each string of letters
  - Transform each object to eliminate uninformative information
  - Decide whether the transformed strings group together



Key Collision: Fingerprint



"Godel"
"Gödel"
"Johnny 5 "
"Johnny 5"
"Johnny-5"
"It's a mad mad mad mad world"
"It's a mad, mad, mad, world"
"Bob Lannon"
"Lannon, Bob"

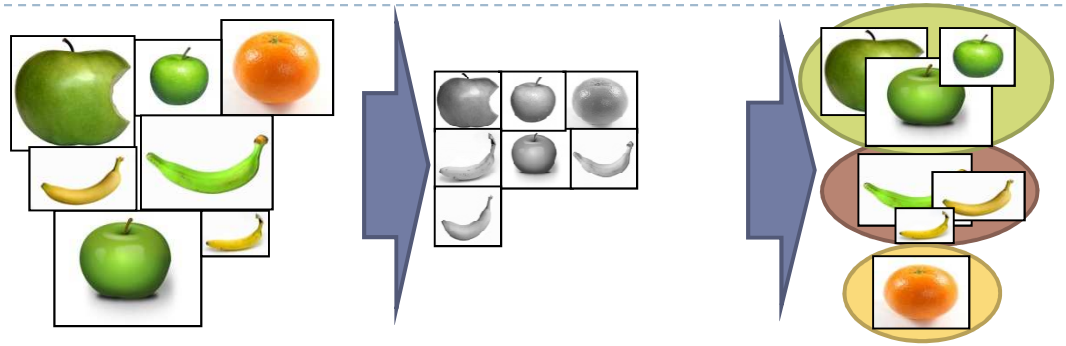
TRANSFORM

"godel"
"godel"
"johnny 5 "
"johnny 5"
"johnny 5"
"a it's mad world"
"a it's mad world"
"bob lannon"
"bob lannon"

CLUSTER

"Godel"
"Gödel"
"Johnny 5 "
"Johnny 5"
"Johnny-5"
"It's a mad mad mad mad world"
"It's a mad, mad, mad, world"
"Bob Lannon"
"Lannon, Bob"

Key Collision: Metaphone

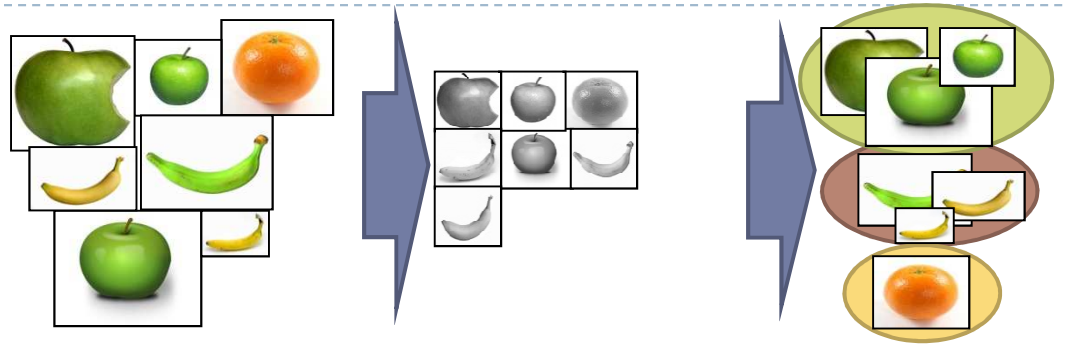


Hypothyroidism	{HP0R, HPTR}	Hypothyroidism
Hupothiroidism	{HP0R, HPTR}	Hupothiroidism
Hyperthyroidism	{HPR0, HPRT}	Hyperthyroidism
Huperthoroidism	{HPR0, HPRT}	Huperthoroidism
Throat Cancer	{0RTK, TRTK}	Throat Cancer
Thyroid cancer	{0RTK,TRTK}	Thyroid cancer

TRANSFORM

CLUSTER

Key Collision: N-Gram (size = 1)



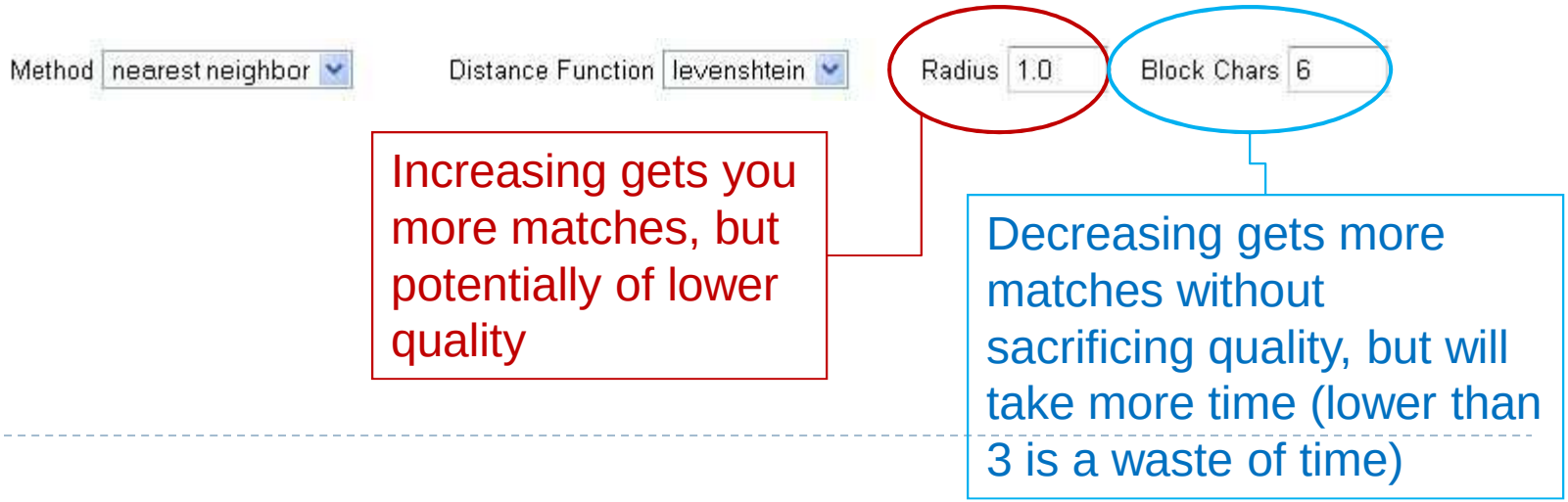
Psoriasis	aioprs	Psoriasis
Proriasis	aioprs	Proriasis
Psoriaasis	aioprs	Psoriaasis
Psoriais	aioprs	Psoriais
psoriiasis	aioprs	psoriiasis

TRANSFORM

CLUSTER

Nearest Neighbors (earmuffs)

- Plotting by attributes
  - Assign each string to a point in some high-dimensional space
  - look for distances between points
- Methods
  - Levenshtein Distance (like some spell checks)
  - Prediction by Partial Matching (like some better spell checks)



DEMO

Link Party!

- Google Refine: http://code.google.com/p/google-refine
  - How to think: http://code.google.com/p/google-refine/wiki/UserGuide
  - Documentation: http://code.google.com/p/google-refine/wiki/DocumentationForUsers
- Tutorials:
  - Google: http://www.youtube.com/watch?v=B70J_H_zAWM
  - Open Corporates: http://vimeo.com/17924204
  - ProPublica: http://www.propublica.org/nerds/item/using-google-refine-for-data-cleaning
  - A Bunch More: http://googlerefine.blogspot.com/p/tutorials-section.html
- More on clustering: http://code.google.com/p/google-refine/wiki/ClusteringInDepth