Exceptional service in the national interest

SAND2015-8946PE

Finding Needles in Airborne Haystacks

October 15, 2015

Andy Wilson

Scalable Analysis and Visualization

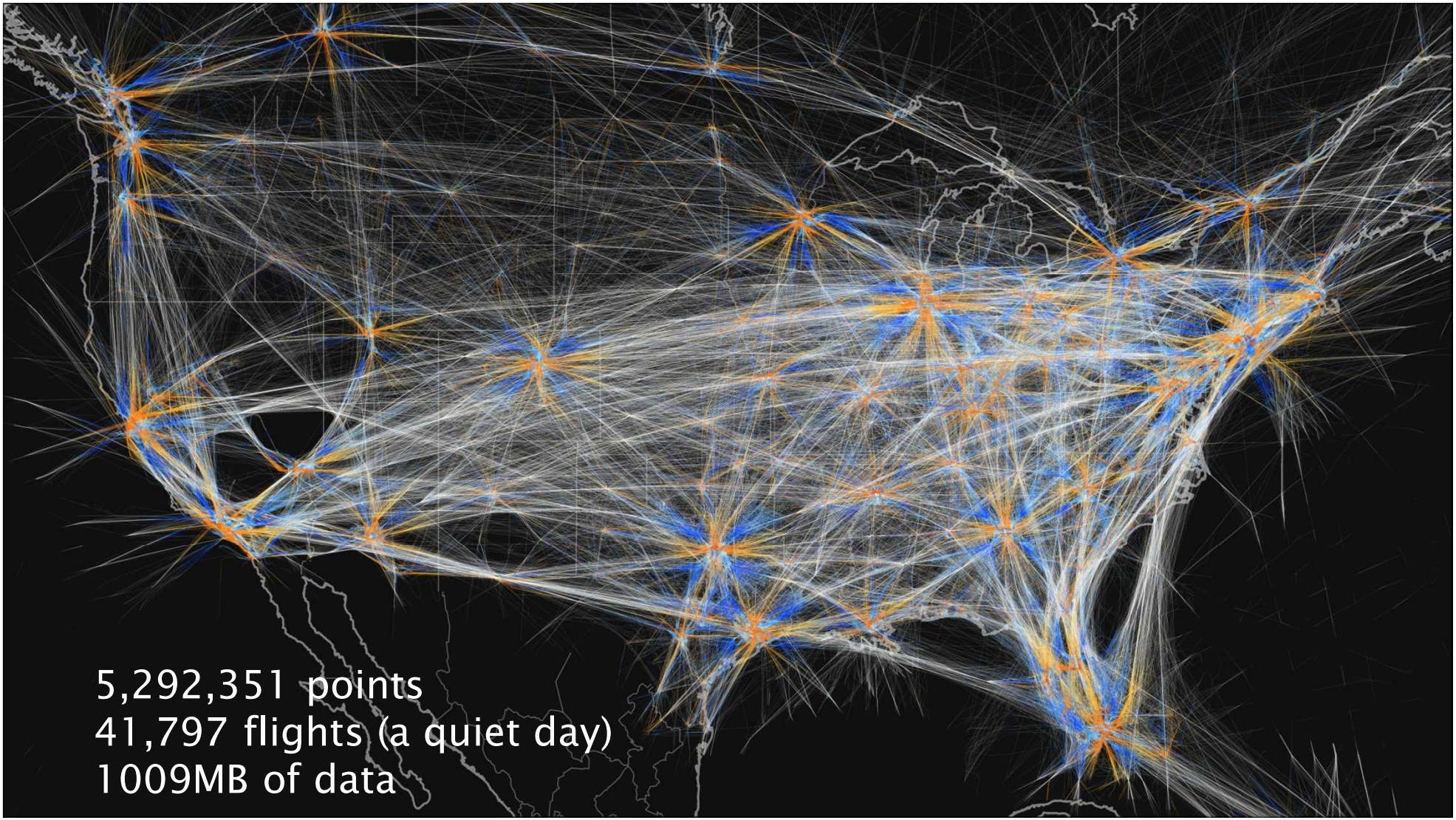
Sandia National Laboratories

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

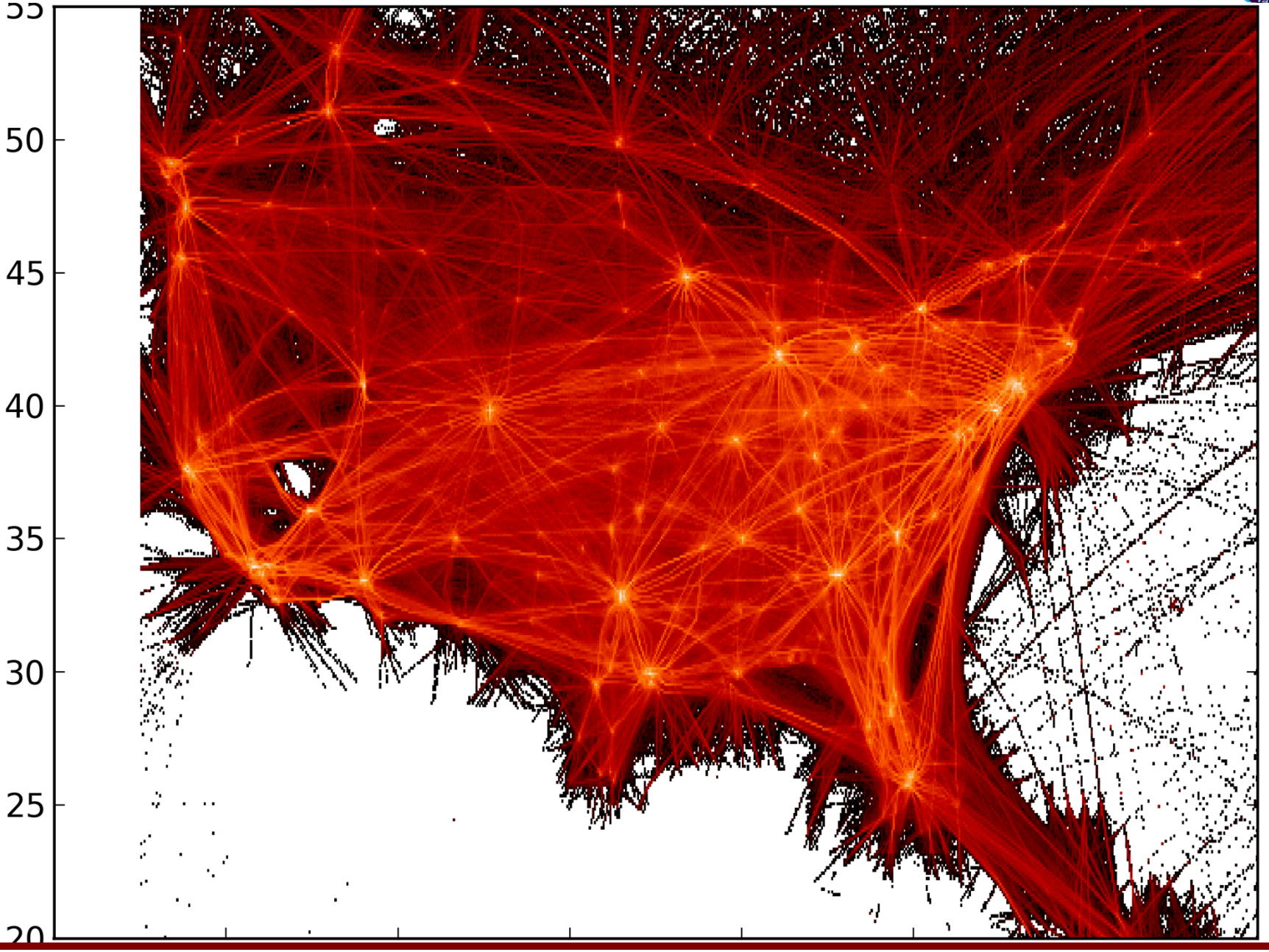
Air Traffic: What We See

One Day of Reality

Flight paths across the U.S. on April 4, 2013



Air Traffic: Reality

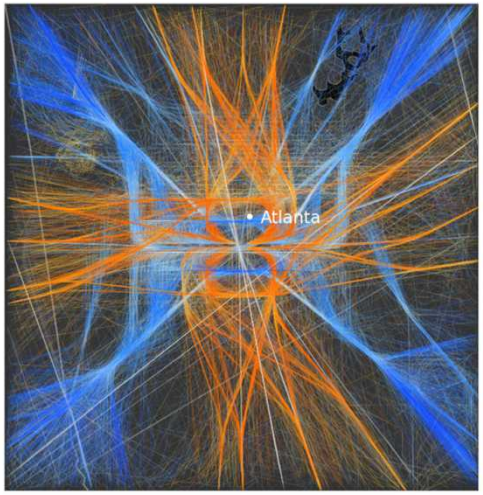


Data Set Vital Statistics: ASDI Air Traffic

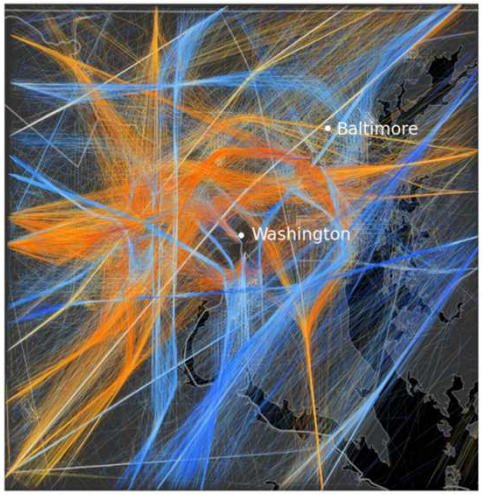
- Aircraft Situation Display to Industry
  - Unclassified feed of US civilian air traffic
  - Powers all flight status web sites, displays in airports
  - FAA originated; we get it from AirNav, Inc.

- Rich, relatively clean data
  - ~30-65000 flights per day (mostly IFR)
  - Each aircraft pinged every 5-60 sec.
    - Position calculated from radar transponder reply
    - Position error affected by distance from radar
    - Status, position, heading, speed, etc. updated and reported every 60 seconds
  - Lots of data: ~5M points/day, ~1GB/day
  - Lots of metadata: 17 columns worth

ATL (Atlanta, GA)



DCA (Washington, DC)



How do we make sense of all that?

Approach 1: Inspection

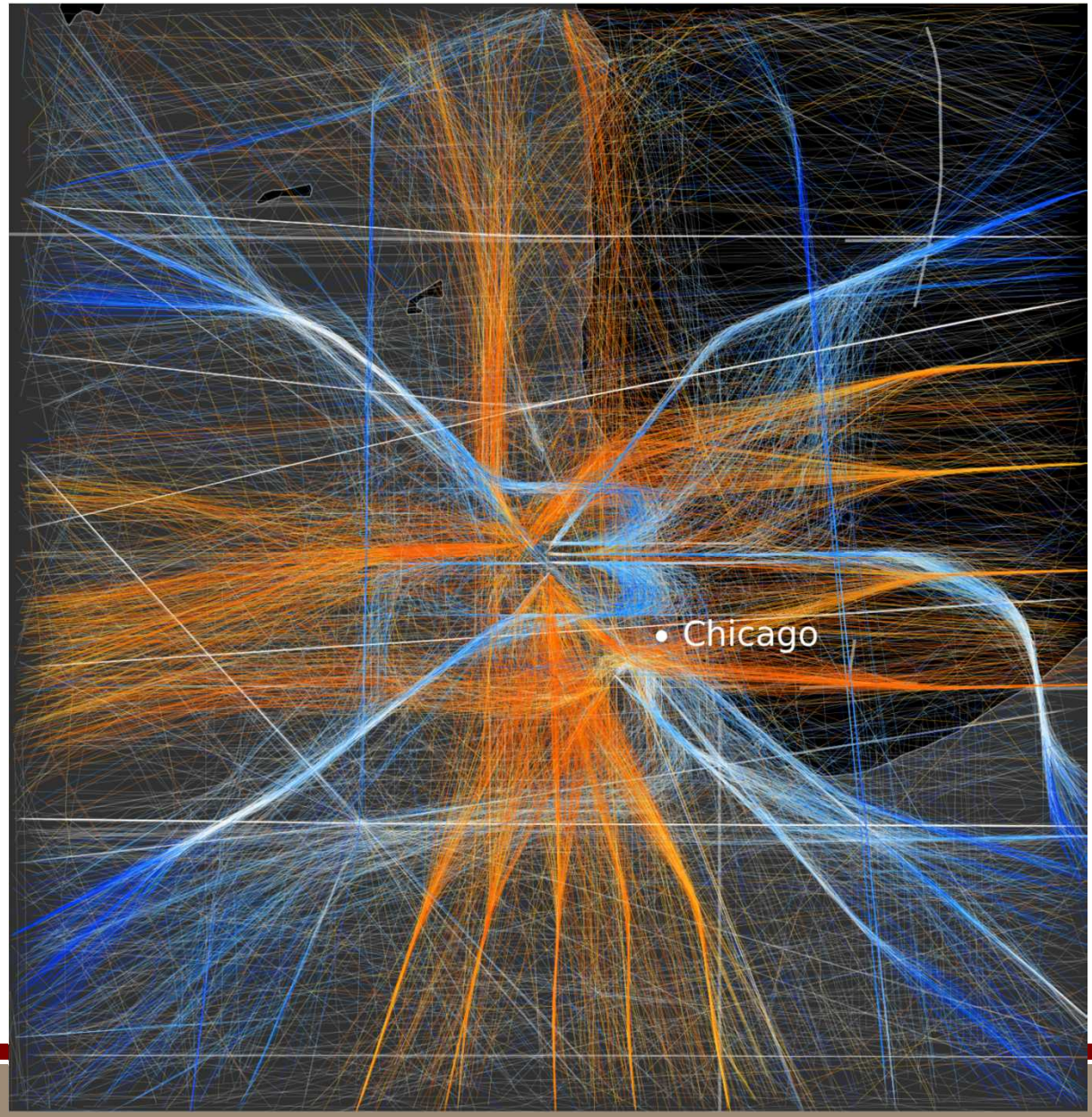
Train humans to watch an ongoing display of traffic and point out anomalies.

Hey, it works for traffic control…

Let's try it!

Chicago

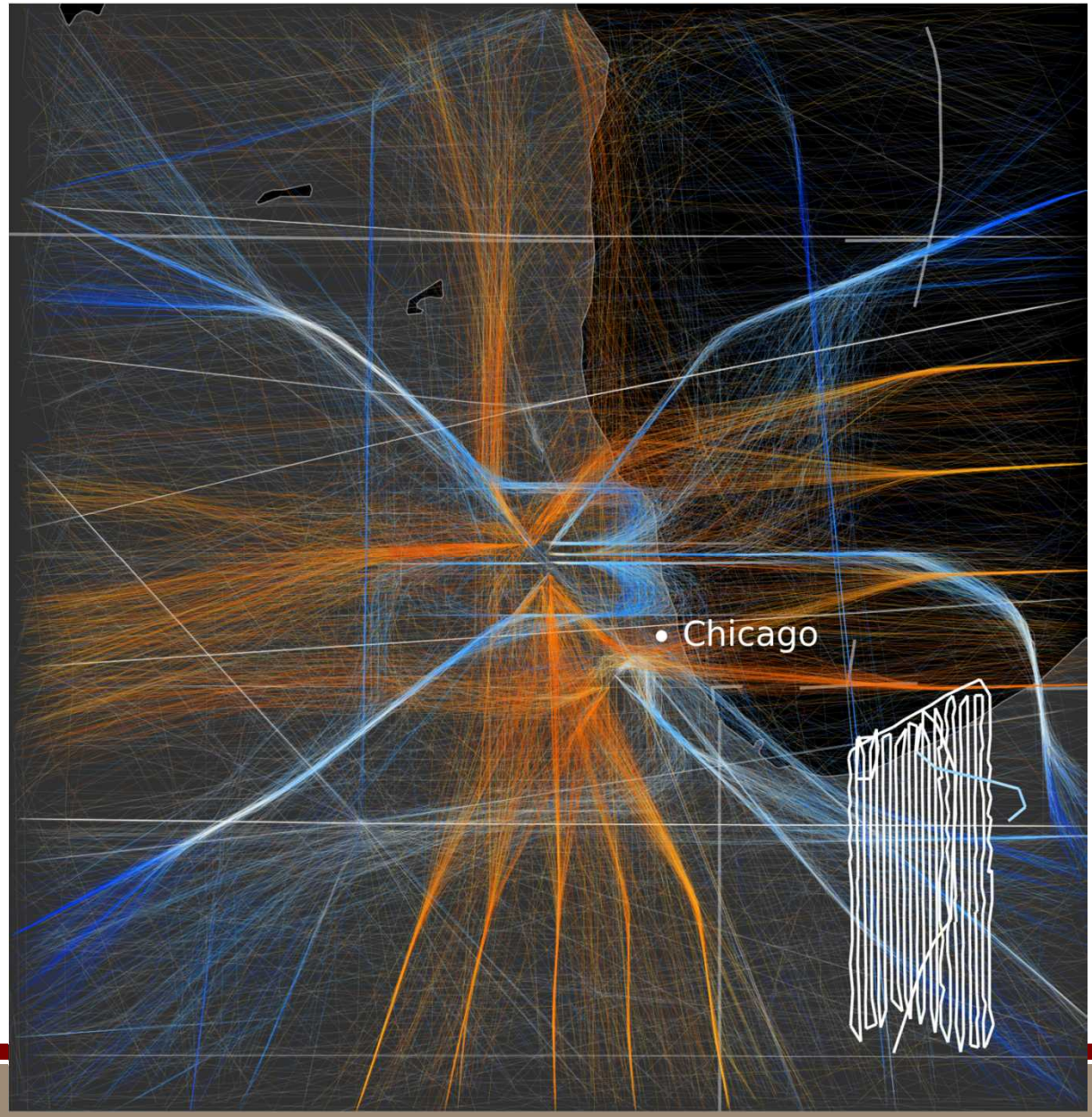
Flight paths on April 4, 2013:



80,456 points
3,986 flights
~15MB of data

Chicago Again

Flight paths on April 4, 2013:



Approach 2: Feature Vectors

- Clustering algorithms are pretty good…
  - Take your pick: parametric, non-parametric, mixture models, k-means, k++-means, hierarchical, agglomerative…
  - One problem: most of them require a normed space!
  - Trajectories as curves exist in a non-normed metric space
  - Measures exist but are often unintuitive

- Solution: invent features that describe behavior
  - How do we as humans describe trajectories?
  - What quantities will let us formalize that?
  - Does that give traction to fast clustering algorithms?

Test Data Set: ASDI Air Traffic

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
update_time	cid	origin	destination	latitude	longitude	altitude	speed	heading	sched_dep_time	actual_dep_time	sched_arr_time	actual_arr_time	aircraft	status	route
4/2/13 20:32	610	KDEN	KEWR	41.7956	-89.0417	37000	528	86	4/3/13 19:14	4/2/13 19:05	4/3/13 22:13	4/2/13 22:00	B752	A	DEN./.MCK..LNK..DSM..EVOTE..NELLS..KEEHO.J584.SLT.FQM3.KEWR/2204
4/2/13 20:32	252	KDTW	KMLI	42.4586	-85.5547	24000	318	272	4/3/13 19:32		4/3/13 20:36		CRJ2	A	DTW./.DUNKS.J70.PMM..KMLI/2118
4/2/13 20:32	468	KBVI	KDVN	41.1942	-86.8636	10000	182	259		4/2/13 18:47		4/2/13 21:35	AEST	A	VI./.GWB..MAPPS..KELSI..DVN
4/2/13 20:32	372	KORF	KATL	35.8608	-78.2856	36000	397	215	4/3/13 20:06		4/3/13 21:36			A	ORF..FKN..JIMAR..FLO.J4.CAE.CANUK2.KATL
4/2/13 20:32	478	MDSD	KFLL	26.0761	-80.2731	1800	210	100	4/3/13 18:09	4/2/13 18:16	4/3/13 20:16	4/2/13 20:30		A	DSD./.HODGY034051..CAREY..DEKAL.WAVUN1.KFLL
4/2/13 20:32	789	MYNN	KATL	27.9619	-81.0528	33000	418	338	4/3/13 19:15	4/2/13 19:49	4/3/13 21:01	4/2/13 21:40		A	YNN.BR57V.FLL..ORL.J53.WORMS.J53.BARBS.J53.CRG.CANUK2.KATL/2137
4/2/13 20:32	320		KFEP	41.2439	-87.3367	6100	154	280	4/2/13 19:25	4/2/13 19:52	4/2/13 21:08	4/2/13 21:33		A	WB./.C65..MAPPS..KELSI..FEP
4/2/13 20:32	87	KEWR	SPIM	29.3539	-78.2706	34000	468	187	4/3/13 18:25	4/2/13 18:47	4/4/13 1:45	4/3/13 2:07	B752	A	EWR./.OHLAA.AR24.ZFP.AR24.URSUS.UL780.GAXER.UL780.BUXOS.UL780.GYV
4/2/13 20:32	85	KPHX	KBOS	41.9333	-84.7269	35000	554	159		4/2/13 17:45		4/2/13 21:54	A320	A	PHX./.GIJ.J554.JHW.J82.ALB.QUABN2.KBOS/2154
4/2/13 20:32	295	KORD	KMKE	42.4586	-88.0169	10000	333	359	4/3/13 20:04	4/2/13 20:26	4/3/13 20:21	4/2/13 20:42	CRJ2	A	ORD..PETTY..KMKE/2042
4/2/13 20:32	791			26.1317	-79.1317	6500	171	247						A	
4/2/13 20:32	643	KDFW	KBOS	38.6461	-76.8675	39000	513	58	4/3/13 18:12		4/3/13 21:19		B752	A	DFW.TRISS4.TXK.J42.MEM.J42.BKW.J42.RBV.J222.JFK.KRANN3.KBOS/0300
4/2/13 20:32	547	KSAT	KIAH	29.9908	-95.6656	5000	219	105	4/3/13 20:14	4/2/13 20:04	4/3/13 20:47	4/2/13 20:36	B739	A	SAT./.SAT108014..HAMMU.HAMMU1.KIAH/2039
4/2/13 20:32	128	KDTW	KFLL	26.1497	-80.155	6000	227	253	4/3/13 18:06	4/2/13 18:13	4/3/13 20:43	4/2/13 20:51	B752	A	DTW./.CRG164013..DIINO.FISEL3.KFLL
4/2/13 20:32	363	KDFW	KORD	41.2144	-89.1517	20500	453	62	4/3/13 18:33	4/2/13 19:02	4/3/13 20:16	4/2/13 20:48	B752	A	DFW./.STL309030..BDF.BENKY1.KORD
4/2/13 20:32	FFF			28.0667	-80.55	3500	173	204						A	
4/2/13 20:32	668	KLGA	KFLL	26.1022	-80.4797	4000	190	170	4/3/13 18:02	4/2/13 18:01	4/3/13 20:38	4/2/13 20:39	A320	A	LGA./.SBY172008..ILM.AR21.CRANS.FISEL3.KFLL/2035
4/2/13 20:32	663	KAUS	KJFK	36.025	-78.3194	37000	526	62	4/3/13 18:01	4/2/13 18:23	4/3/13 21:06	4/2/13 21:26	E190	A	AUS./.3502N/8033W..ORF.J121.SIE.CAMRN4.KJFK
4/2/13 20:32	950	KPHX	KTPA	27.8872	-83.8344	20900	438	87	4/3/13 17:12	4/2/13 17:23	4/3/13 20:46	4/2/13 20:51	B737	A	PHX./.REMIS..SIMMR.BLOND3.KTPA/2052
4/2/13 20:32	23R	KSDF	KMIA	25.8706	-80.5606	4300	270	84	4/3/13 18:38	4/2/13 18:34	4/3/13 20:49	4/2/13 20:40	E145	A	SDF./.JUULI.SSCOT1.KMIA/2051
4/2/13 20:32	850	KDCA	KCLT	36.4833	-79.7911	22000	386	218	4/3/13 20:20	4/2/13 20:00	4/3/13 21:18	4/2/13 20:58	A319	A	DCA./.LYH.SUDSY4.MAJIC.SUDSY4.KCLT/2100
4/2/13 20:32	766	PHNL	RJAA	21.8667	-161.9	32000	450	285	4/3/13 19:50		4/4/13 3:25			A	HNL.KEOLA2.PUPPI..CRESP..2300N/17000W..2500N/18000E..2800N/17000E..2
4/2/13 20:32	420	KEWR	KFLL	26.3647	-79.8758	7300	276	225	4/3/13 18:30	4/2/13 18:15	4/3/13 21:04	4/2/13 20:53	B738	A	EWR./.BAHAA.AR21.CRANS.FISEL3.KFLL/2050
4/2/13 20:32	281	KFLD	KMKT	43.7297	-90.4847	16000	206	140	4/2/13 19:40	4/2/13 20:03	4/2/13 20:50	4/2/13 21:17	BE9L	A	LD./.MSN007023..MKT
4/2/13 20:32	173			28.7	-81.55	4600	115							A	
4/2/13 20:32	20R	KPHX	KFLL	26.2664	-81.27	13100	398	128	4/3/13 17:11	4/2/13 17:11	4/3/13 21:01	4/2/13 20:53	B737	A	PHX./.BLVNS.Q102.BAGGS.JINGL1.KFLL/2054
4/2/13 20:32		RCTP	KLAX	39.65	-135.4	37000								A	
4/2/13 20:32	138	KSJC	KAUS						4/2/13 22:40		4/3/13 1:40			P	SJC.SJC9.AVE..PMD.J65.BLH..BXK.J184.DMN..ELP.J183.LLO.KALLA3.KAUS/0300
4/2/13 20:32	735	KDTW	KSFO	42.4858	-85.9172	34000	351	274	4/3/13 19:43	4/2/13 20:09	4/4/13 0:15	4/3/13 0:39	B738	A	DTW./.PMM091045..ONL.J94.BFF.J157.MTU.J148.DTA..RUMPS..OAL.MOD4.KS
4/2/13 20:32	723	KNDZ	KMOB	30.7889	-87.3756	3900	100	253	4/2/13 20:05	4/2/13 20:05	4/2/13 20:53	4/2/13 21:05	B06	A	DZ./.DR..SJI..MOB/2050
4/2/13 20:32	291			30.67	-83.5375	5500	135	60						A	
4/2/13 20:32	848			23.9903	-78.4578	32000	414	340						A	
4/2/13 20:32	288	KFLL	KATL	27.1719	-80.6294	30900	472	9	4/3/13 20:15	4/2/13 20:18	4/3/13 21:48	4/2/13 21:48		A	FLL./.ORL158095..VQQ.CANUK2.KATL/2153
4/2/13 20:32	442	KSSI	KRIC	34.2253	-79.2917	27000	295	29	4/2/13 18:55	4/2/13 19:47	4/2/13 20:29	4/2/13 21:21	B350	A	SSI./.DWYTE.J165.RIC..KRIC/2119
4/2/13 20:32	598	LTBA	KIAH	43.9119	-86.345	38000	480	209	4/2/13 10:32	4/2/13 10:32	4/2/13 22:47	4/2/13 22:45	B77W	A	TBA./.SSM219033..LIT.Q33.DHART.J180.SWB..KIAH
4/2/13 20:32	473	KFXE	KIAH	27.4747	-83.6081	36000	245	288	4/2/13 19:26	4/2/13 19:39	4/2/13 23:18	4/2/13 23:28		A	FXE./.LAL160066..REMIS.Q100.REDFN.Q100.LEV.GILCO1.KIAH/2344
4/2/13 20:32	25R		KFXE	27.365	-80.6708	8000	177	183		4/2/13 18:40		4/2/13 21:04	BE80	A	8J./.TRV344021..TRV343020..TPSTR.V437.SHEDS270005.V437.BRIKL..FXE
4/2/13 20:32	442	KDTW	KCLE	41.4639	-81.7719	2400	141	231	4/3/13 19:49		4/3/13 20:09		CRJ2	A	DTW..MAARS..HIMEZ.HIMEZ2.KCLE/0022
4/2/13 20:32	916	TJSJ	KATL	27.25	-77.9756	36000	419	306	4/3/13 18:30	4/2/13 18:30	4/3/13 22:00	4/2/13 21:58		A	JSJ.ACONY1.ELMUC..RENAH.Y585.ATTIK.Y585.OMN.J45.CRG.CANUK2.KATL/22
4/2/13 20:32	366	KFLL	KHOU	27.3261	-83.3425	34000	366	289	4/3/13 19:41	4/2/13 19:58	4/3/13 22:09	4/2/13 22:27	B733	A	FLL./.LBV125024..REMIS.Q100.LEV.CLMBA1.KHOU/2231
4/2/13 20:32	55	KPOF	KAPF	26.2944	-82.1256	6500	305	145	4/2/13 18:55	4/2/13 18:53	4/2/13 20:50	4/2/13 20:41	FA50	A	POF./.COVIA117093..BAGGS.PIKKR3.KAPF
4/2/13 20:32	844	KMSP	KBWI	43.8192	-91.2103	31000	543	122	4/3/13 20:05	4/2/13 20:15	4/3/13 21:54	4/2/13 22:04	MD90	A	MSP./.RGK169016..FWA.J178.APE..AIR.EMI5.KBWI/2158
4/2/13 20:32	631	KMIA	MTPP	25.7508	-79.8647	11200	310	113	4/2/13 20:12		4/2/13 21:45		MD82	A	MIA..SKIPS.BR53V.SWIMM.A315.JOSES.A315.OBN..MTPP/0132
4/2/13 20:32	671	KDFW	KEWR	37.7931	-80.945	37000	546	85	4/3/13 17:21	4/2/13 18:52	4/3/13 20:01	4/2/13 21:28	A320	A	DFW./.BKW.J42.MOL..GVE.PHLBO3.KEWR
4/2/13 20:32	287	KMCO	KIAH	29.9936	-95.3928	700	122	94	4/3/13 18:07	4/2/13 18:22	4/3/13 20:13	4/2/13 20:26	A319	A	MCO./.TBD156036..WOLDE.WOLDE3.KIAH
4/2/13 20:32	383	MDPC	KPHL	22.4289	-72.2089	34000	373	320	4/2/13 19:40	4/2/13 19:40	4/2/13 23:36	4/2/13 23:18	A320	A	DPC./.PTA096037..NUCAR.G446.OLDEY..DIW.J174.ORF.J121.SAWED.J121.SWL
4/2/13 20:32	557	KLEX	KBCT	28.7167	-80.8667	5900	179	164	4/2/13 19:11		4/2/13 20:24	4/2/13 18:51		A	FL.V3.MALET.V3.MLB..BCT
4/2/13 20:32	907	KFLL	KLAL	26.2583	-80.2436	10100	295	293	4/2/13 20:27		4/2/13 21:00		LJ60	A	FLL.THNDR1.THNDR..KLAL/0032

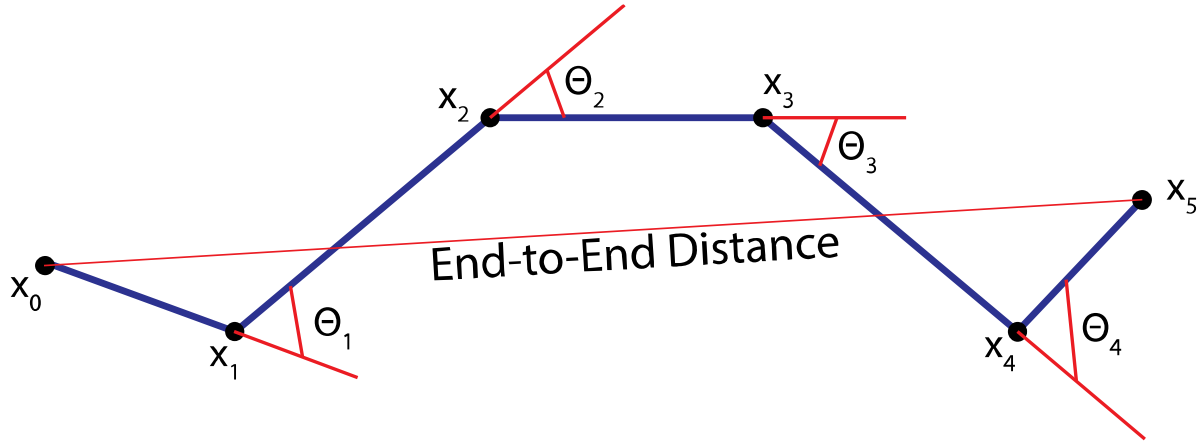
ASDI Data Description

- Aircraft Situation Display to Industry
  - Unclassified feed of US civilian air traffic
  - Powers all flight status web sites, displays in airports
  - FAA originated; we get it from AirNav, Inc.

- Rich, relatively clean data stream
  - ~30-65000 flights per day (mostly IFR)
  - Each aircraft pinged every 5-60 sec.
    - Position calculated from radar transponder reply
    - Position error affected by distance from radar
    - Status, position, heading, speed, etc. updated and reported ~60 seconds
  - Lots of data: ~5M points/day, ~1GB/day, currently >100GB!
  - Lots of metadata: 17 columns worth

Simple Features

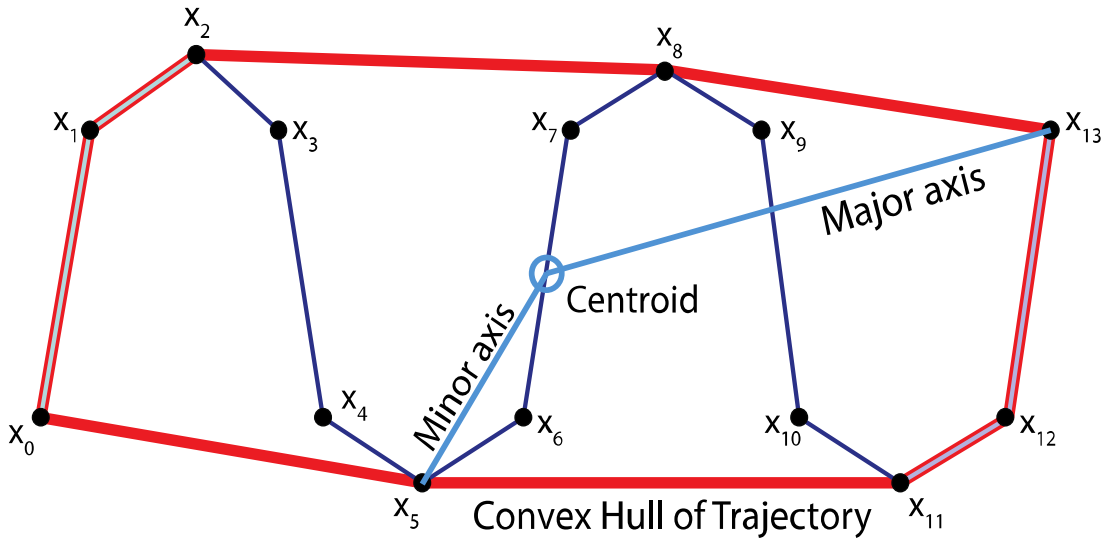
- End-to-end distance
- Total distance traveled
- Total curvature
- Total turning
- Speed (avg, max, min)



These features are derived from the individual line segments that compose the flight.

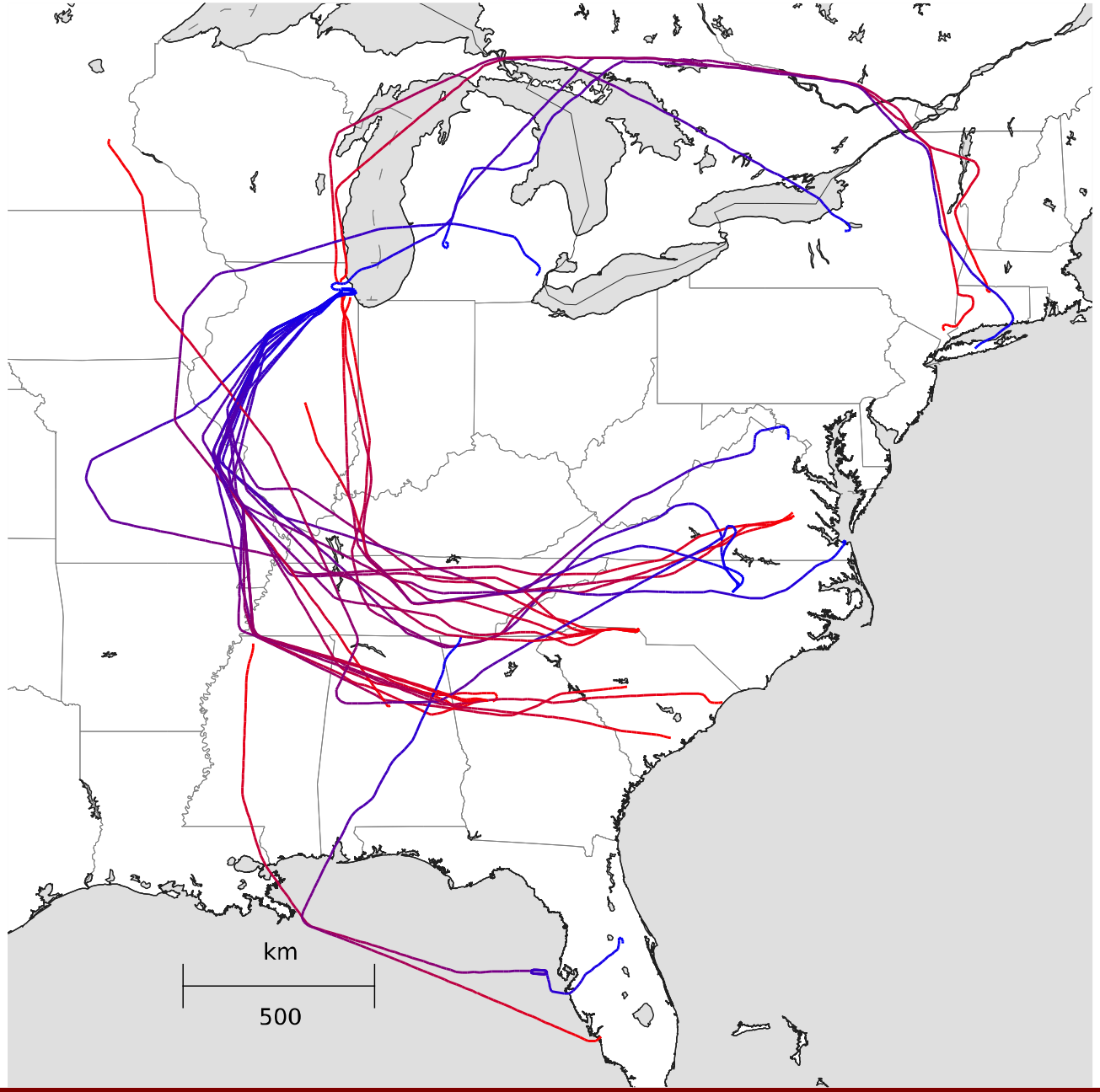
Geometric Features

- Aspect ratio of convex hull
- Perimeter of convex hull
- Area of convex hull

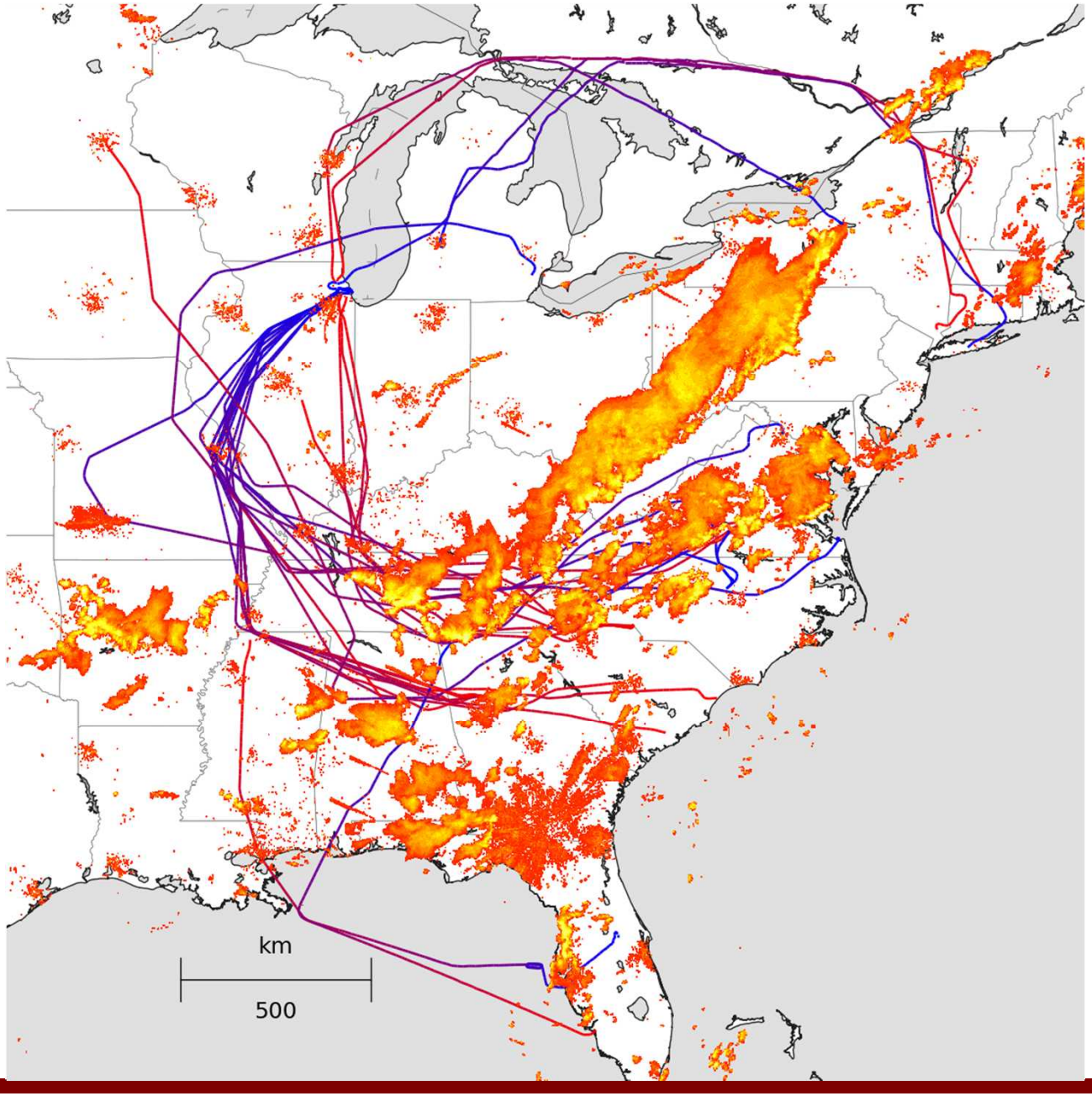


The convex hull helps us see higher-level behavior.

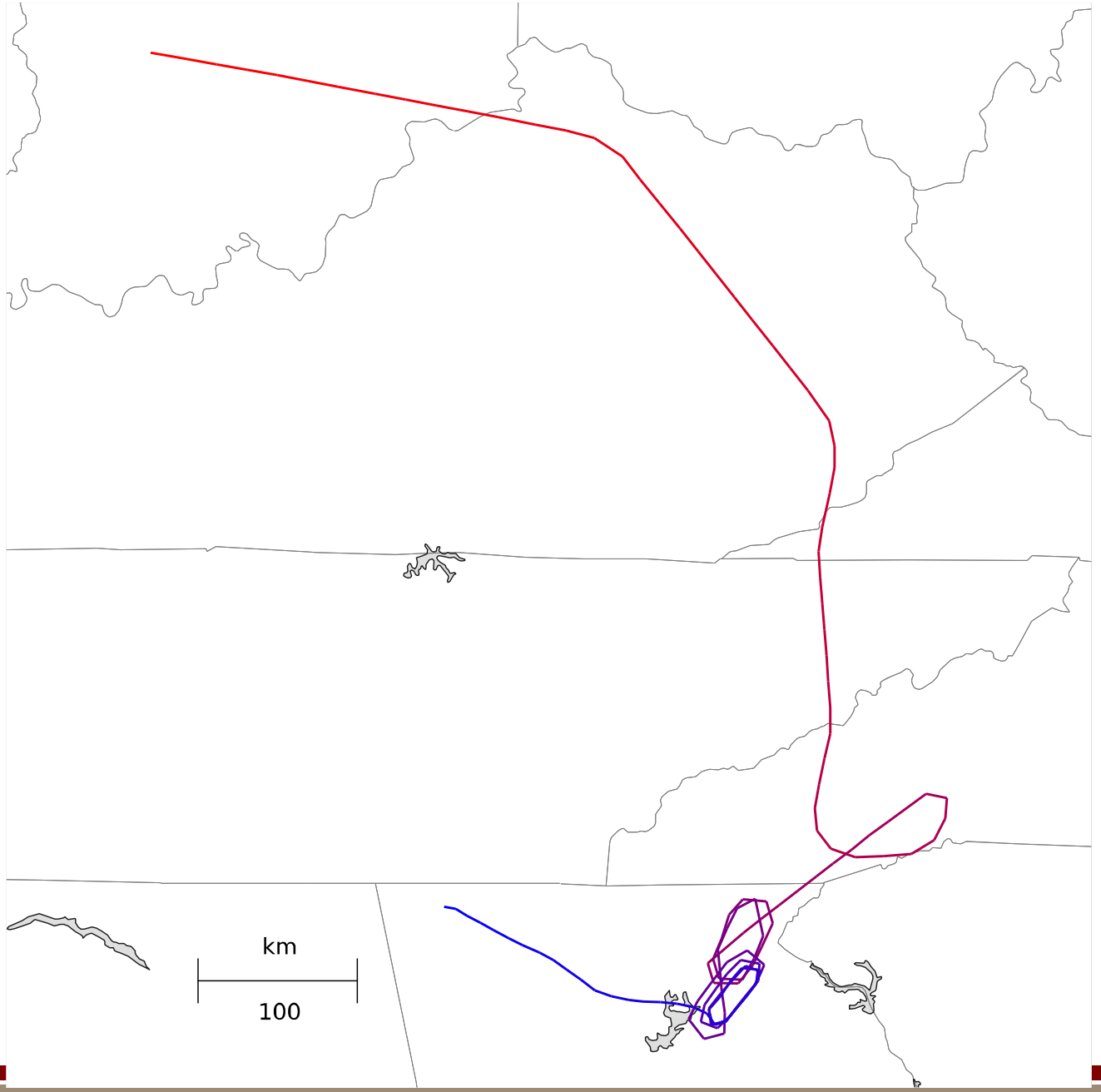
Feature Vectors: Avoiding Space



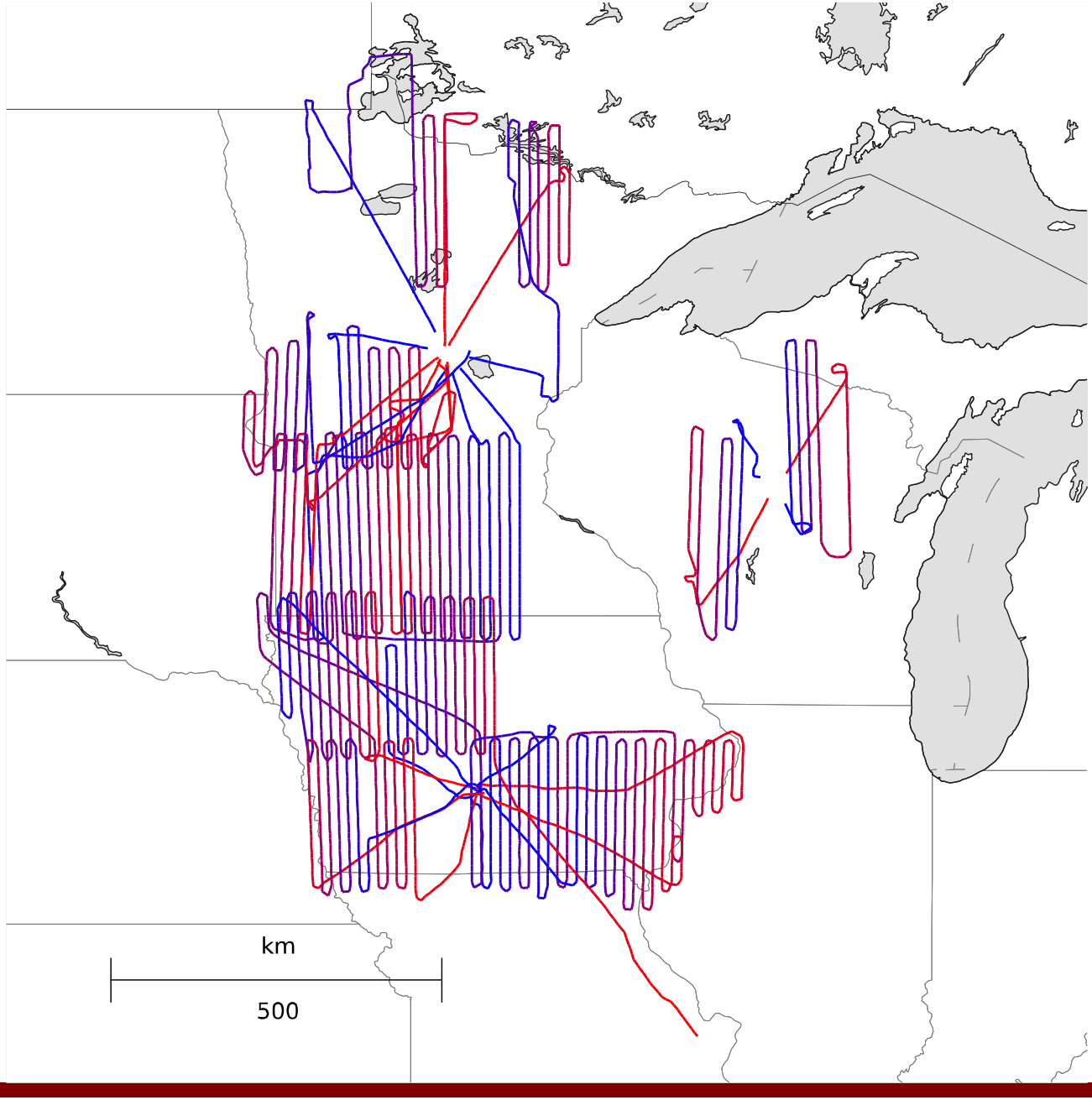
Feature Vectors: Avoiding Space



Feature Vectors: Holding and Diverted



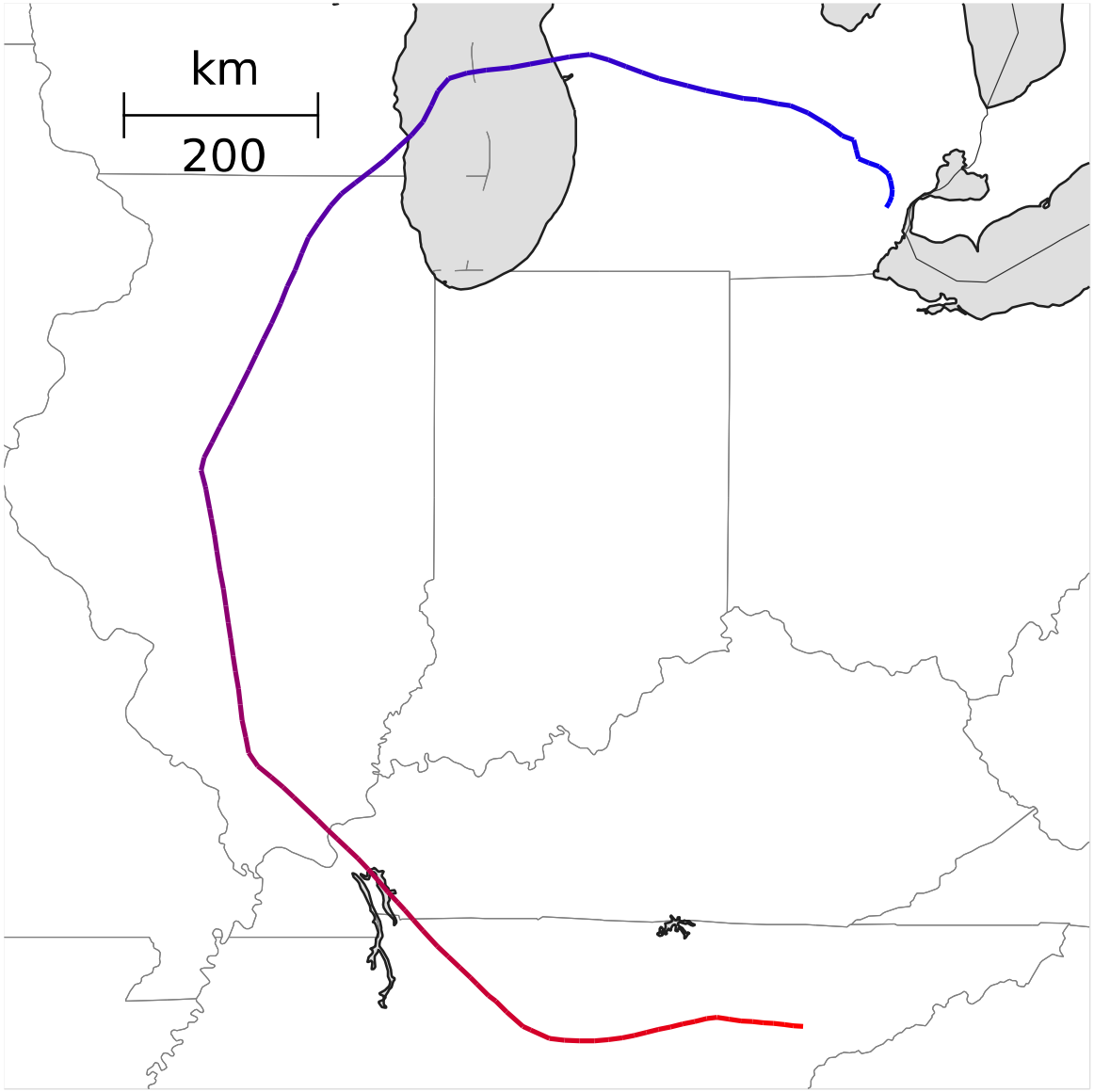
Feature Vectors: Survey Flights



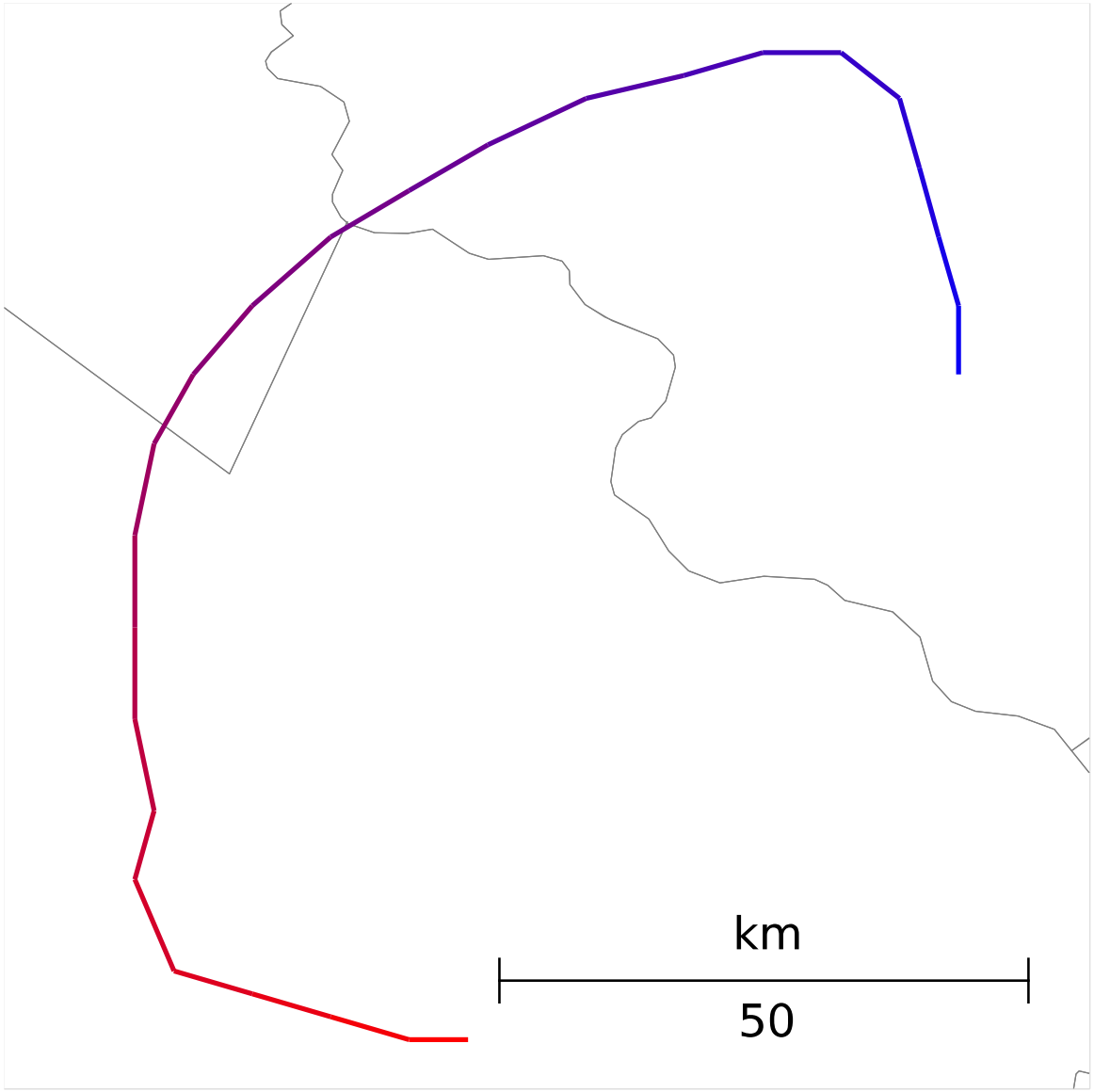
Approach 3: Distance Geometry

- "…characterization and study of sets of points based *only* on given values of distances between member pairs." [Wikipedia]
- Very useful in computing protein structure and searching for similar molecules
- Intuition: Pick evenly spaced sets of points, compute pairwise distances, normalize so largest distance is 1
- Result: fingerprint invariant under rigid transformation

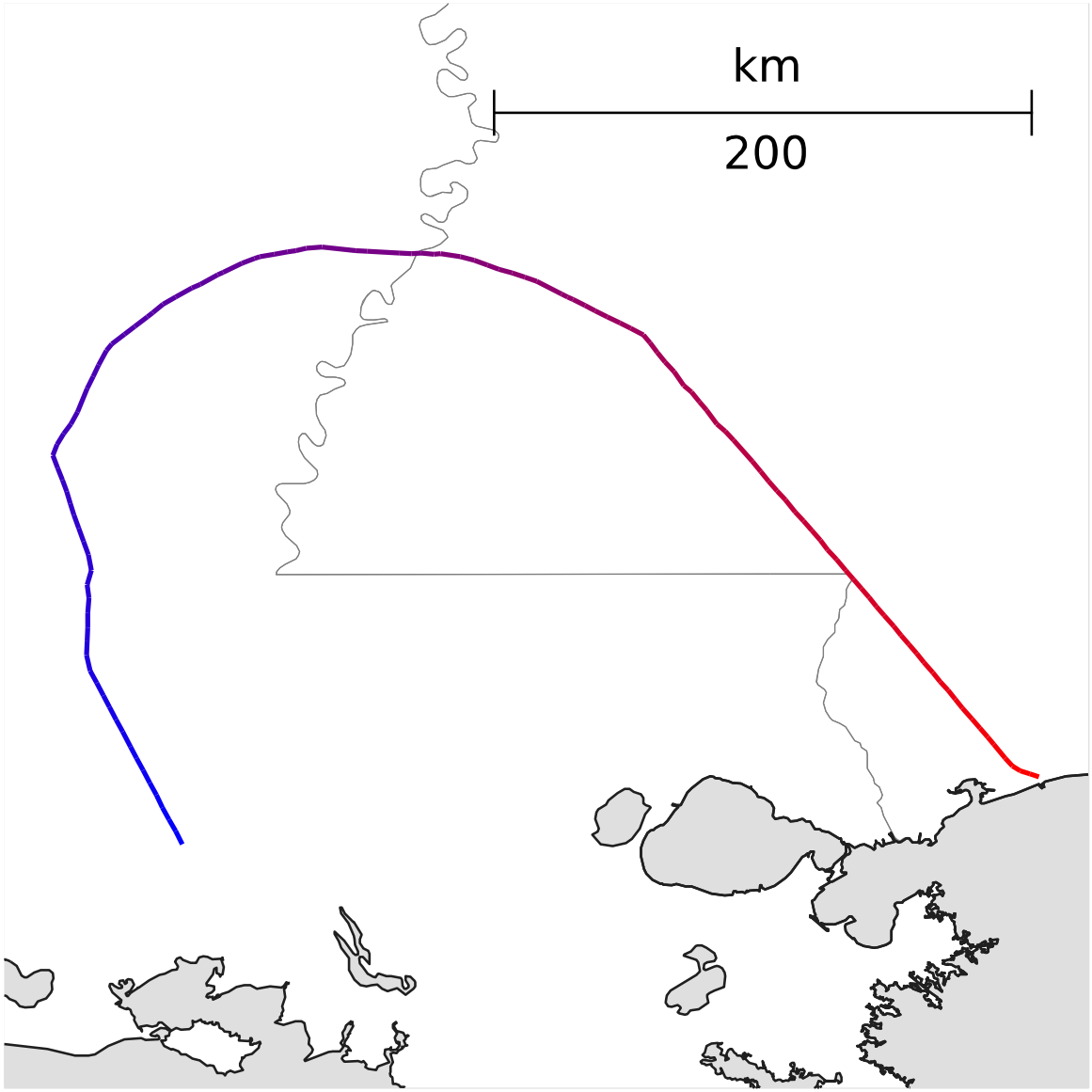
Distance Geometry: Exemplar



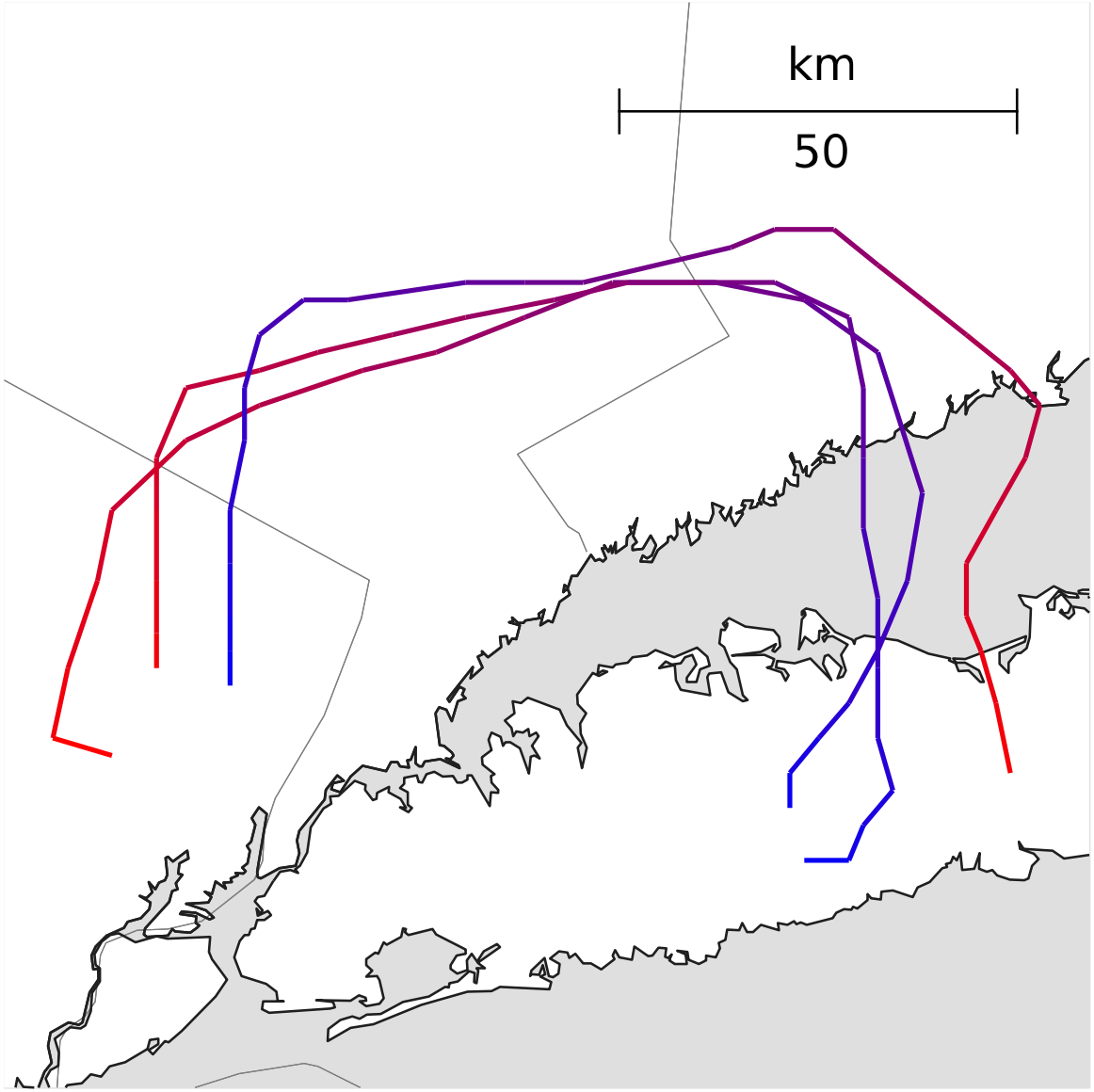
Distance Geometry: Result 1



Distance Geometry: Result 2



Distance Geometry: Result 3

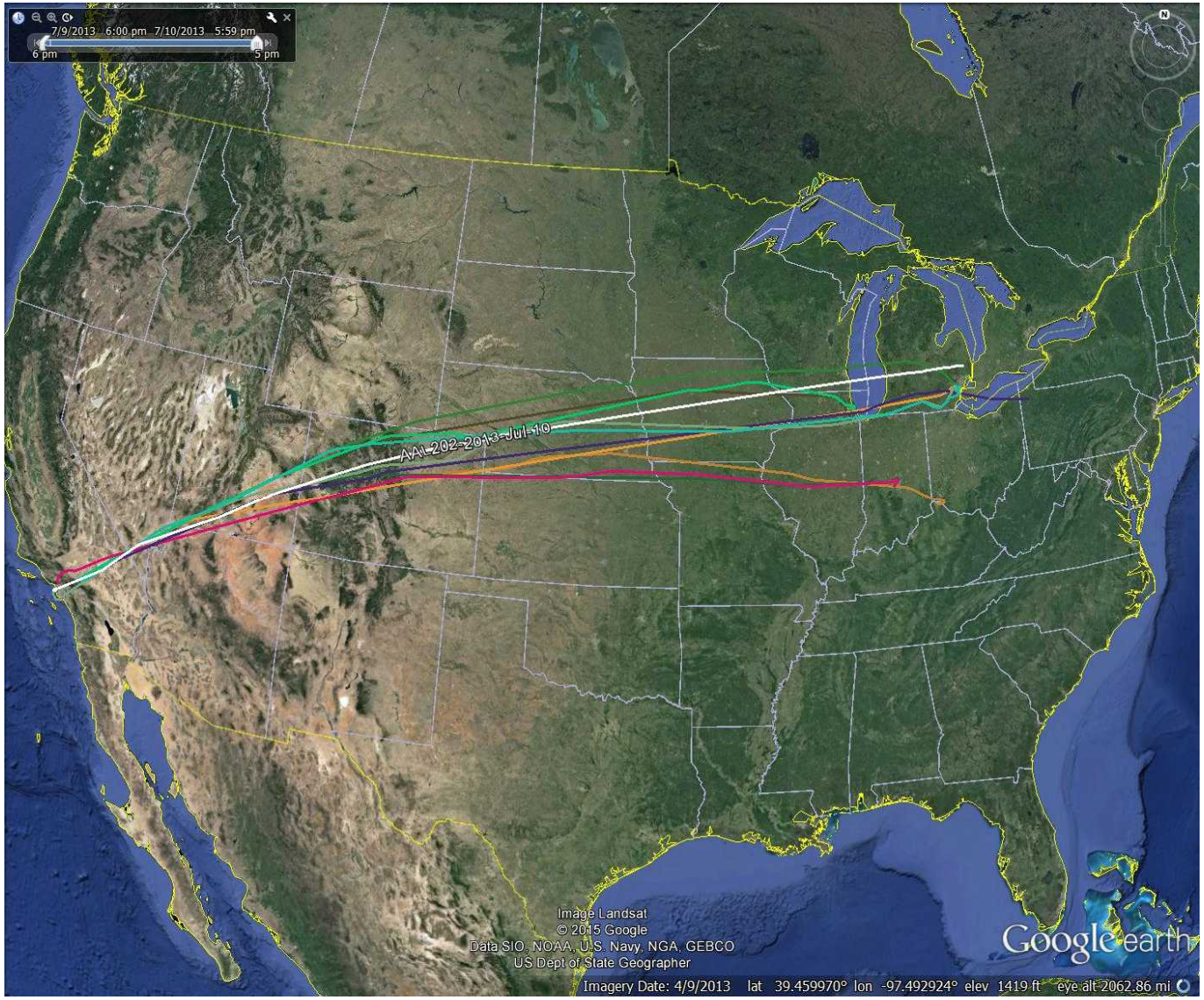


Prediction Using Tracktable

- Idea: Put known historical trajectory fragments into a database
- Take the observations of the beginning of a new trajectory, and search for near matches
- Weight the “nearness” of the different trajectories, and sort the different possible destinations
- Work based on observing the first 20% to 80% of a flight, with that fraction unknown to prediction algorithm

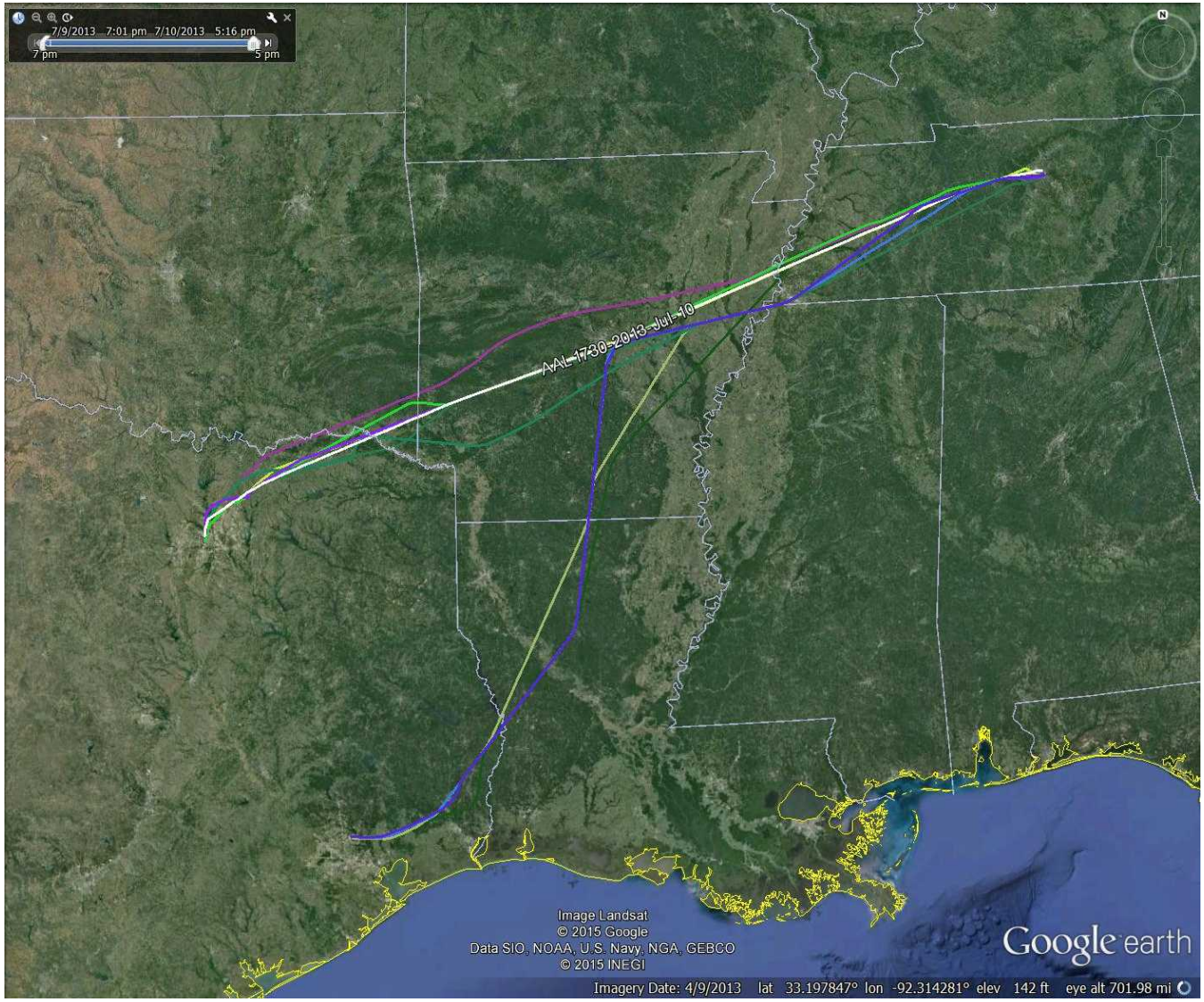
Small Data Set -> Mistake

- Target flight (white) goes from LAX to Toronto, but nearest flights are all going to more common destinations (Chicago, Detroit)



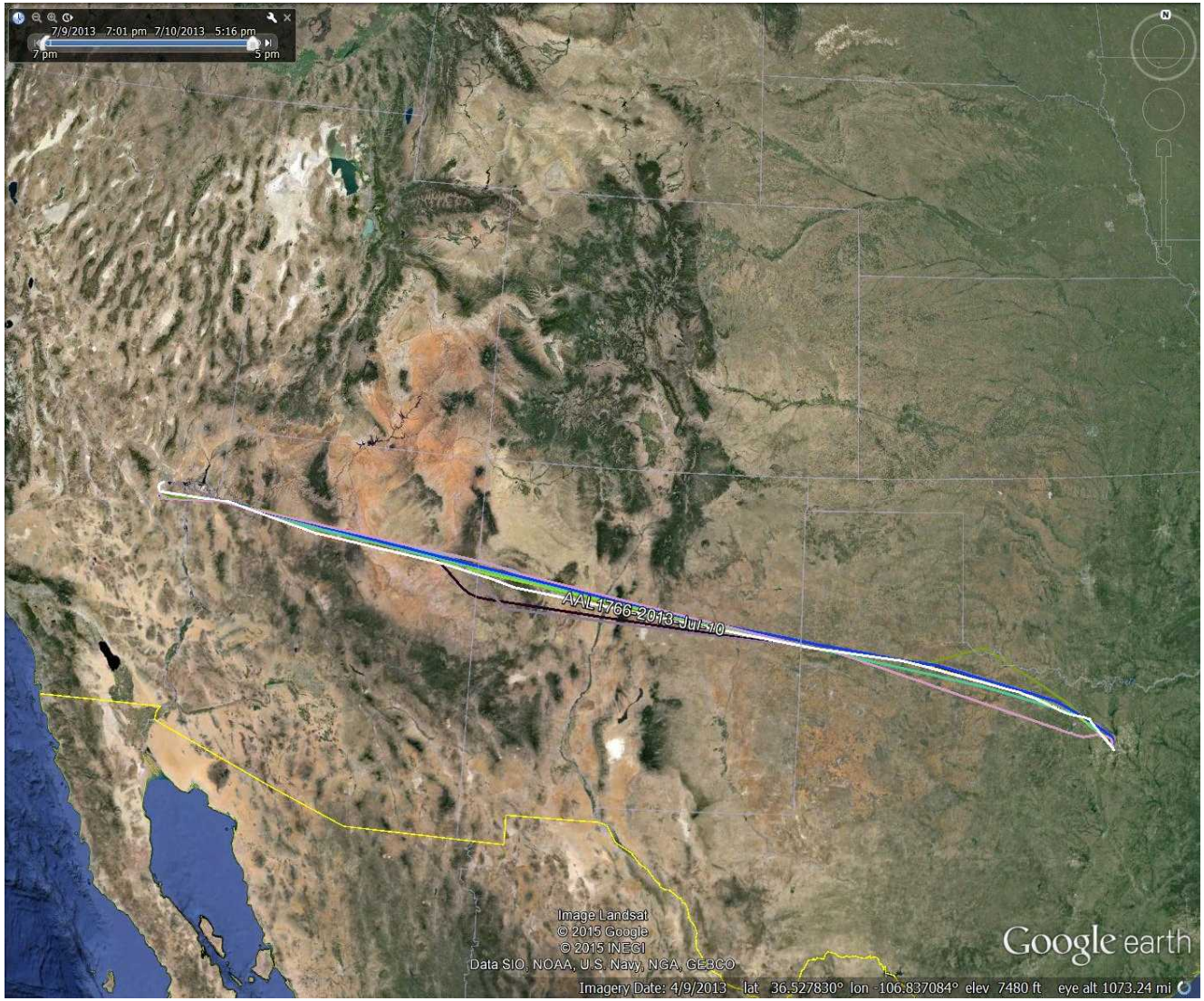
More Data, Better

- Flights from Nashville go to DFW and IAH, but proper weighting finds DFW for the target (white).

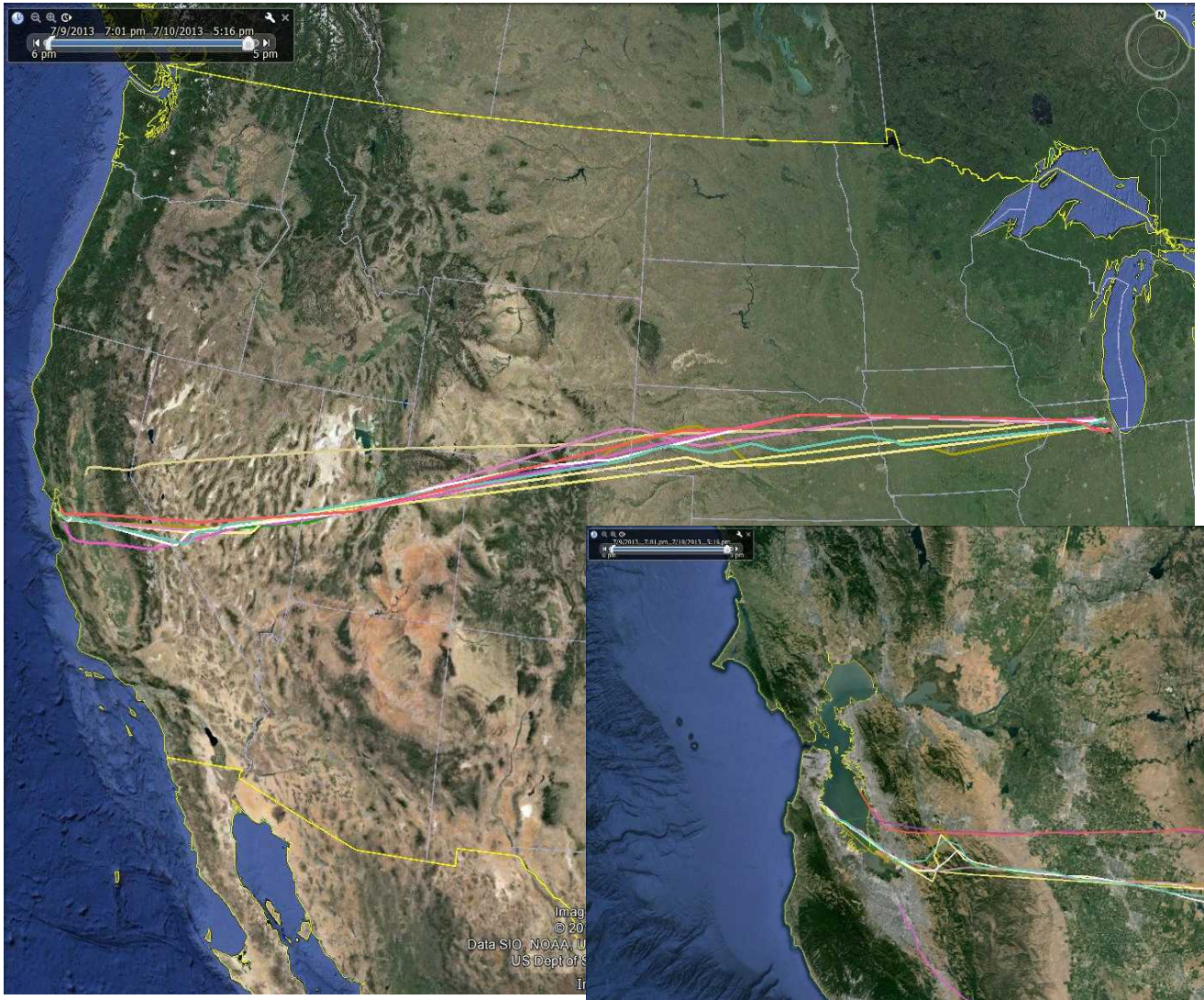


Lots of data

- Popular flights are picked perfectly (DFW to Las Vegas)



Perfection not possible



- Some flights to the same airport follow the same path for long time, like Chicago to Bay Area.
- Close-up shows flights going to SFO, OAK, SJC (and one to Sacramento!)

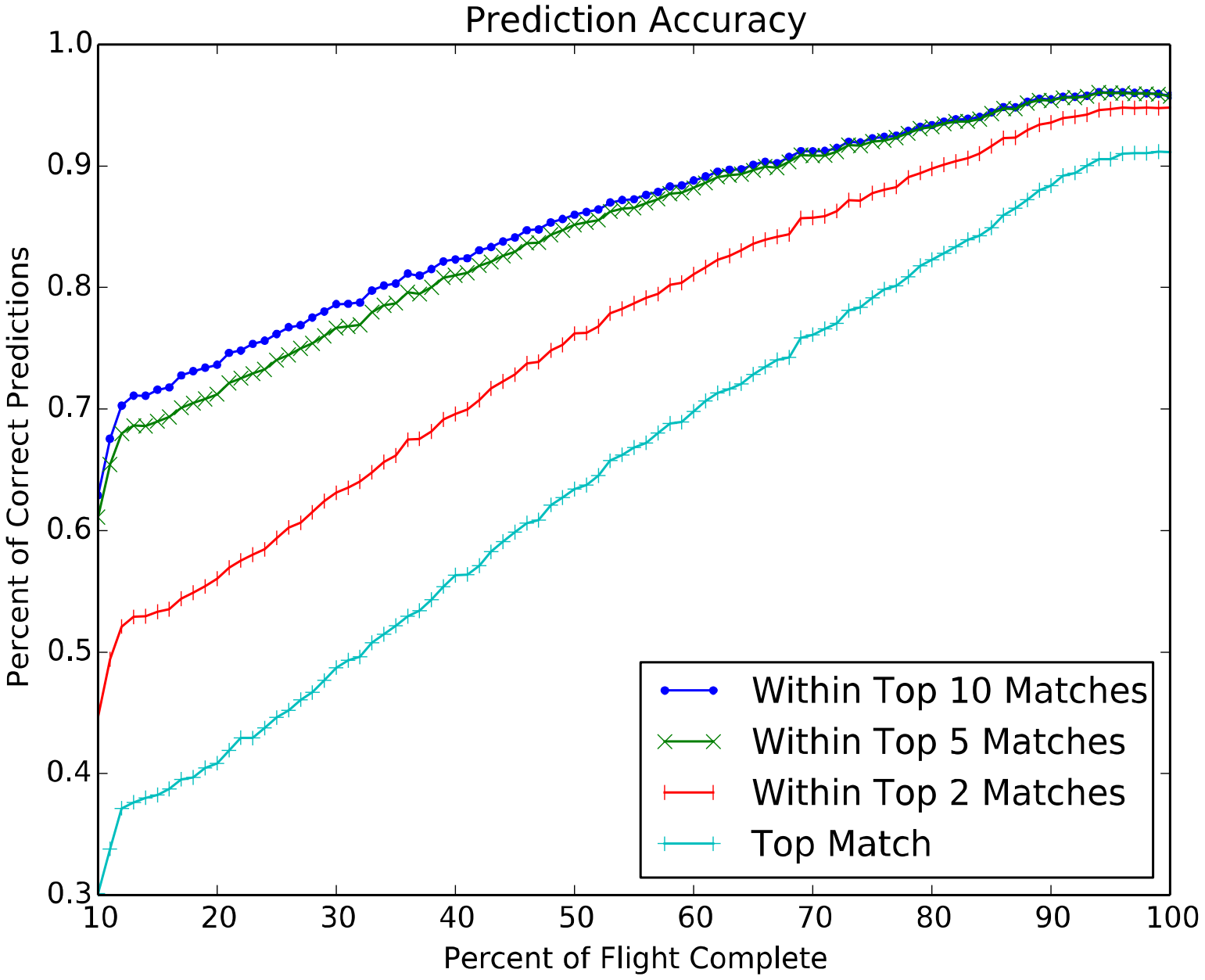
Numerical Results

Numerical results for prediction from ASDI Data

	Few Hours	One Day	5 Days	One Month
Top Flight	34.36%	44.09%	52.86%	56.73%
Top 3 Flights	51.25%	63.45%	72.59%	75.76%
Top 5 Flights	58.22%	70.15%	78.83%	81.18%
Top 10 Flights	64.13%	75.74%	82.59%	84.22%
Top 20 Flights	64.80%	76.28%	82.87%	84.43%

- *Some notes*
  - *Prediction from 4 pts, random first 20%-80% of flight*
  - *Data sets range from ~8K (Few Hours) to ~800K (One month)*
  - *All flights in data set were predicted one-at-a-time, based on all other flights in data set*
  - *Having an existing flight in the database is critical (prediction not magic!)*

Numerical Results



What Next?

- Characterize and identify flight *segments*
- Generative model for behavior?
- Run at even larger scale
- Open-source release of Tracktable

UNUSED SLIDES FOLLOW